Astronomy & Astrophysics manuscript no. output
January 9, 2020

RASSINE: An interactive tool to normalise spectra

I. Description and performance of the code

M. Cretignier[1], J. Francfort[2], R. Allart[1], X. Dumusque[1], and F. Pepe[1]

[1] Astronomy Department of the University of Geneva, 51 ch. des Maillettes, 1290 Versoix, Switzerland
e-mail: michael.cretignier@unige.ch
[2] Département de Physique Théorique, Université de Genève, 24 quai Ansermet, CH–1211 Genève 4, Switzerland
e-mail: jeremie.francfort@unige.ch

Received XXX ; accepted XXX

ABSTRACT

Aims. Provide an open-source code allowing an intuitive and robust spectra normalisation.
Methods. We developed a *Python* code to normalise *1d* spectra through the concepts of convex hulls. The code uses 6 parameters and good results can be obtained without hardly fine-tuning them. The code also provides a complete user-friendly interactive interface, including graphical feedback, helping the user to chose the parameters as easily as possible. To facilitate even further the normalisation, RASSINE can provide a first guess for the parameters which are derived directly from the spectrum thanks to calibrations already performed on the parameters.
Results. We found out that such normalisation were accurate and precise enough to provide reliable line depth for solar and α Cen B spectra. The accuracy of RASSINE was estimated to 1.9% on α Cen B spectra where this low accuracy is in part due to molecular band absorption and line density in the bluest part of the spectrum. A simple correction based on stellar template comparison allow to increase the accuracy up to 1.3%. For HARPN spectra of the Sun, an accuracy of 0.26% can be reached which is 3 times better than with the commonly used method of polynomial fit. The precision on individual spectra normalisation is estimated around 0.11% which can be considerably improved giving as input a spectra-timeseries to stabilize the derived continua. In this context, the precision is found compatible with the photon noise limit.
Conclusions. With an accuracy higher than polynomial fitting method and a precision compatible with the photon noise, RASSINE is a tool that could find applications in numerous situations as stellar parameters determination, transmission spectra of exoplanets or detection of activity-sensitive lines.

Key words. Spectrum normalisation – Alpha shape – Python code – RASSINE

1. Introduction

A spectrum is a fundamental observable used to study astronomical objects such as galaxies, stars and exoplanets. It describes the distribution of photons per wavelength bins and can either be used in term of absolute quantity to determine the luminosity of the objects or colors can be analyzed separately through photometric bands. A rich content of information is also brought by the absorption or emission lines for which the spectrum has to be normalised by a continuum. This could happen for instance in the framework of stellar abundances studies (Blanco-Cuaresma et al. 2014) or exoplanets atmospheres (Wyttenbach et al. 2015). For radial velocity, the spectrum has not to be continuum-normalised but a color correction has to be applied (Malavolta et al. 2017), itself related to the continuum of the spectrum. Spectrum normalised are also mandated to construct the binary mask used in the cross-correlation (Pepe et al. 2003).

When studying absorption lines, a non-trivial step consists in normalising the spectrum by its continuum, where the latter can differ from a blackbody curve due to Rayleigh scattering reddening. Such differential chromatic response can also be directly induced by the spectrograph itself through the optical elements and the CCD captor which cannot be modeled by any radiative transfer code. Due to the large number of high resolution spectra upcoming from ESPRESSO (Pepe et al. 2014), EXPRESS, NEID (Schwab et al. 2016), HIRES (Pasquini et al. 2010), CRIRES, NIRPS (Bouchy et al. 2017) and the huge amount of spectra already available produced by CORALIE (Queloz et al. 2000), HARPS (Pepe et al. 2002b; Mayor et al. 2003; Pepe et al. 2003) or HARPN (Cosentino et al. 2012), an effective process to normalise them in a unified and coherent way appears as an important step.

Current methods often deal with *2d* spectra from scalegrating spectrograph since they represent a narrower band of the spectrum where the continuum variation present a smaller number of inflection points. A filter is often used in order to smooth the data and remove as many stellar lines as possible. This filter can be a rolling maximum or moving average, an asymmetric sigma clipping or even a Fourier filtering. A low order polynomial can eventually be used to fit the continuum and often the blaze function simultaneously.

In the case of *1d* spectra with large wavelength coverage, high order polynomial functions are mandated to account for the numerous inflection points. Unfortunately, the minimization of high polynomial fitting is generally poorly

performed and the algorithm convergence is not guarantied. The presence of smaller variation structures in the continuum produced directly by the optical elements of the instrument could also lead to more concerns which are difficult to model.

Another approach consists to scale the stellar continuum on a stellar template or reference, before to fit the trend on the ratio spectrum. The main advantage of this method is to considerably reduce the polynomial order needed, but the intrinsic disadvantages of polynomial fit remain (Škoda 2008). Such method allow for instance to correct for color variation induced by different airmass observations (Malavolta et al. 2017). If a continuum is known for the stellar reference, all the spectra can then be normalised by the product of the reference continuum with the long trend fit on the ratio.

Using the methods cited above, most scientists can at some level perform a sufficiently good normalisation, namely through smoothing, frequency filtering and fitting often used by iterative processes (Tody 1986, 1993). However, a lot of issues are usually met by polynomial fitting (Škoda 2008) and the fine-tuning of the parameters can often appear as tricky or laborious. Also, the chain of operations necessary to normalise the spectra can suffer from a lack of consistency and is often poorly detailed by the authors.

The difficulties met in some cases are even more frustrating considering that a normalisation is a thought experiment as simple as putting a "veil" on a spectrum, which can be easily visualized by eye, but is more complex to implement. In other words, it consists in finding the best upper envelop falling on an object, in this case a spectrum. The veil is also related to a parameter of tension which will allow it to fall or not inside the lines.

RASSINE (Rolling Alpha Shape for a Spectrum Interpolating Normalisation Estimator) is a Python code written to guide the user with the different steps necessary to realise an efficient spectrum normalisation, with animated and graphical feedback provided. The algorithm is described in detail in section 2.2. The normalisation is based on an alpha shape strategy, equivalent to a convex hull algorithm. A code rather similar to ours has been published recently by Xu et al. (2019). The authors showed the higher performances of the alpha shape method compared to classical iterative methods. A description of the algorithms used in automatic mode by RASSINE are described in appendix A. This mode provides a first guess for 5 of the 6 relevant parameters. The code was also developed to deal with a spectra time-series as detailed in section 2.3. Its impact on broad absorption lines was studies in section 3.1. Finally, an evaluation of the accuracy is performed in section 3.2. The precision of RASSINE is presented for independent spectra normalisation in section 3.3 and in section 3.4 for the time-series' case.

2. Theory

2.1. Convex hull and alpha shape

The problem of finding the upper envelop of a spectrum is closely related to the concepts of convex hull and alpha shape, the latter being a generalization of the former. We briefly review those ideas here, restricting our description to the $2D$ case

The convex hull of a set of points S in the plane can be understood instinctively as follows: take an elastic band and stretch it around the set S, before letting it go. The final shape that the band will take is the convex hull of the set. Mathematically speaking, as defined by Asaeedi et al. (2013), if the set is made of points x_i, the convex hull $C(S)$ can be written as :

$$C(S) = \left\{ y_i = \sum \alpha_i x_i \in \mathbb{R}^2 \;\middle|\; \alpha_i \geq 0, \sum \alpha_i = 1 \right\}. \quad (1)$$

The convexity of the convex hull forbids the existence of *holes*. Simply speaking, already in the approximation of the blackbody radiation, a spectrum cannot be described by a convex hull since this latter contain inflection points. In order to account of such possibilities, we need to consider alpha shapes, which generalize convex hulls and can provide concave hulls.

Given a convex hull, all the internal angles are at most of 180°. In the alpha shape framework, this condition is relaxed such that each internal angle is at most $180° + \alpha$. Obviously, a value of α arbitrarily large (close to 180°) is not useful, because it will correspond to link all the points by straight lines. The upper envelop of the spectrum corresponds then to an intermediate result, between a convex hull ($\alpha = 0°$) and the full alpha shape ($\alpha = 180°$).

2.2. Outline and structure of the code

2.2.1. Brief overview

RASSINE can be subdivided in 5 steps called the SNAKE sequence:

1. Smoothing of the spectrum,
2. Neighborhood's maxima detection,
3. Alpha shape algorithm,
4. Killing outliers,
5. Envelop interpolation.

The main assumption of the code states that the local maxima of the spectrum are probing or "touching" the continuum. This assumption, relatively well satisfied for solar-type stars, is however completely violated for the coolest ones like M dwarfs for which the high stellar lines density lead to blended lines. Some questions may also occur for the blue part of the spectrum because of the strong chromatic effect of scattering and because the line density is often higher in this wavelength region.

The logic behind the SNAKE sequence suits well with the normalisation process. The smoothing is necessary to increase the signal-to-noise ratio (SNR) per bin element which allows to deal with spectra at different noise levels. This step is also relevant, since, if not performed, our continuum will follow the upper envelop of the noise which is not wanted. This stage hence increases the reliability of the detected maxima. A modified gift wrapping algorithm is performed on the local maxima to select only the maxima describing the upper envelop which remove the local maxima formed by blended lines. The rolling radius is not fixed but evolves according to a penalty law in order to "jump" above broadest absorption bands. Outliers points are then rejected according to given criteria to bias as least as possible the different interpolations performed at the end.

2.2.2. Smoothing of the spectrum

The first task of the code is to smoothen the spectrum. For this purpose, two different approaches can be used to remove the high frequency noise that can damage the final result. The easiest method remains to degrade the resolution of the spectrum after convolving it with a specified kernel. When the resolution decreases, the lines cores are becoming much weaker, but the flat parts of the continuum stay stable. Since RASSINE is only focused on the search of local maxima, the change of the line depth does not constitute an issue. Two kernels are available: a rectangular one or a Gaussian one than can be selected with the parameter ***par_smoothing_kernel***[1]. The strength of the smoothing is controlled by the length of the kernel. The associated parameter ***par_smoothing_box*** corresponds to the window width of the smoothing process.

The second filter available is a Savitzky–Golay filter of third order (Savitzky & Golay 1964), which can be described as a low order polynomial rolling fit on the data. The advantage of this filter comprises its ability to keep almost invariant the global shape of the spectrum, whereas it filters very well the oscillations on smaller scale than the window. However, this filter remains more sensitive to rude flux variation which can be due to cosmic rays, lamp contamination or gap between the CCD. To counteract this issue, an asymmetric sigma-clipping is performed on the absolute difference between the smoothed and initial spectrum and outliers points are brought back to their initial flux values.

2.2.3. Local maxima

The second task realized by the code is the search of the local maxima. A peak is called a local maxima if its value is the highest in its closest neighbourhood. A parameter ***par_vicinity*** is introduced, corresponding to the size of the half-window defining the neighbourhood. If the previous smoothing step has been made correctly, this parameter stays quite irrelevant, but it gets some interest when low SNR spectra are studied.

2.2.4. Penalty

At first order, all the lines should have the same width, this latter being determined by the projected stellar rotation, macroturbulence and instrumental resolution[2]. This line width can be for instance extracted from the cross-correlation function (CCF) and represents the typical length scale associated to a stellar spectrum observed by a specific instrument. For this reason, one of the most important parameter is the full width half maximum (FWHM) of the CCF that should be provided in $\mathrm{km\,s^{-1}}$ stored in a parameter called ***par_fwhm***. In an ideal world, the alpha shape would have a constant radius (in $\mathrm{km\,s^{-1}}$) arbitrary larger than the typical line width. However, such considerations are only true in the weak line regime. When lines begin to saturate, the Voigt profile's wings begin to dominate and the line width diverges from the typical value of the weak regime. As examples of strong absorption lines, we can mention the H_α, CaII H&K, sodium doublet NaD or triplet magnesium MgIb. This part of the code takes into account this problem.

The philosophy is the following: if we require the radius of the rolling pin to be constant, this latter will fall in broad absorption lines. To prevent that, a *penalty map* is used to increase the radius as soon as it gets close to a broad absorption. The interactive graphical interface for this stage is presented in Fig.1. First of all, we realize two approximated continua of the spectrum by a rolling maximum. The first one, S_1, with a small window about 50 times the σ-value of the CCF (red curve first panel Fig.1) and the second one S_2 using a window 10 times bigger than S_1 (black curve first panel Fig.1). Because a rolling maxima can be sensitive to anomalous flux intensity (cosmics, contamination between fibers,...), before computing these continua, a rolling median and rolling interquartile is performed in a window with a size of 10 wavelength elements. A 20σ clipping (or 20IQ clipping in our case) is performed to remove such unrealistic flux values. Then, for each wavelength, the penalty $p = (S_2 - S_1)/S_2$ is computed and normalised by the minimum and maximum value to lie between 0 and 1. Because this curve is irregular (gray curve second panel Fig.1), which tends to slow down the code afterwards, the penalty is stratified so that the final shape has a "skyline" shape (black curve second panel Fig.1). This step also allows the rolling pin to "feel" the hole before falling inside.

The next step consist to choose a function given by the parameter ***par_reg_nu*** which will map the penalty p to a new radius r (black curve bottom panel Fig.1). First of all, one should take into account the overall broadening of the lines with increasing wavelength. Indeed, since the line width is chromatic (line are broader in the red than in blue), the value for the radius has to be specified at a wavelength reference position that is considered to be by default the bluest wavelength of the spectrum, hence the λ_{min}. A generic chromatic law $C(\lambda)$ is then needed which would be strictly linear[3] if all the lines had exactly the same velocity width. The simplest expression for this law is

$$C(\lambda) = \frac{\lambda}{\lambda_{min}}. \tag{2}$$

Two penalty laws are then provided: a polynomial function and a sigmoid. For polynomial mapping (black curve right panel Fig.1), the radius is given by

$$r(p, \lambda) = C(\lambda)[R + (R_{max} - R) \cdot p^{\nu}] \tag{3}$$

where ν is a real (positive) parameter specified directly with ***par_reg_nu***, and R, R_{max} are two parameters of the model: ***par_R*** and ***par_Rmax***. They define the minimum radius and the maximum one in Å units. The values

[1] In the rest of this article, we will use the convention of the bold and cursive syntax for the parameters, where the name of the parameter is written as in the code.

[2] Note that all the lines have the same line width only in a logarithmic wavelength space since their width are the same in the velocity space and not in the wavelength space

[3] The natural line profile is a Lorentzian with a unique width Γ for each line. However, observed spectral lines are the convolution of this natural profile with a Gaussian profile $\mathcal{N}(0, \sigma)$, where the Gaussian accounts for the finite resolution of the instrument, the stellar rotational broadening, thermal broadening and macroturbulence. Since, $\Gamma \ll \sigma$, all the lines possess roughly the same width Δv. Because this line width is small, the classical Doppler formula can be applied and we found that the width in wavelength $\Delta\lambda$ is a linear function: $\Delta\lambda \propto \lambda\Delta v$.

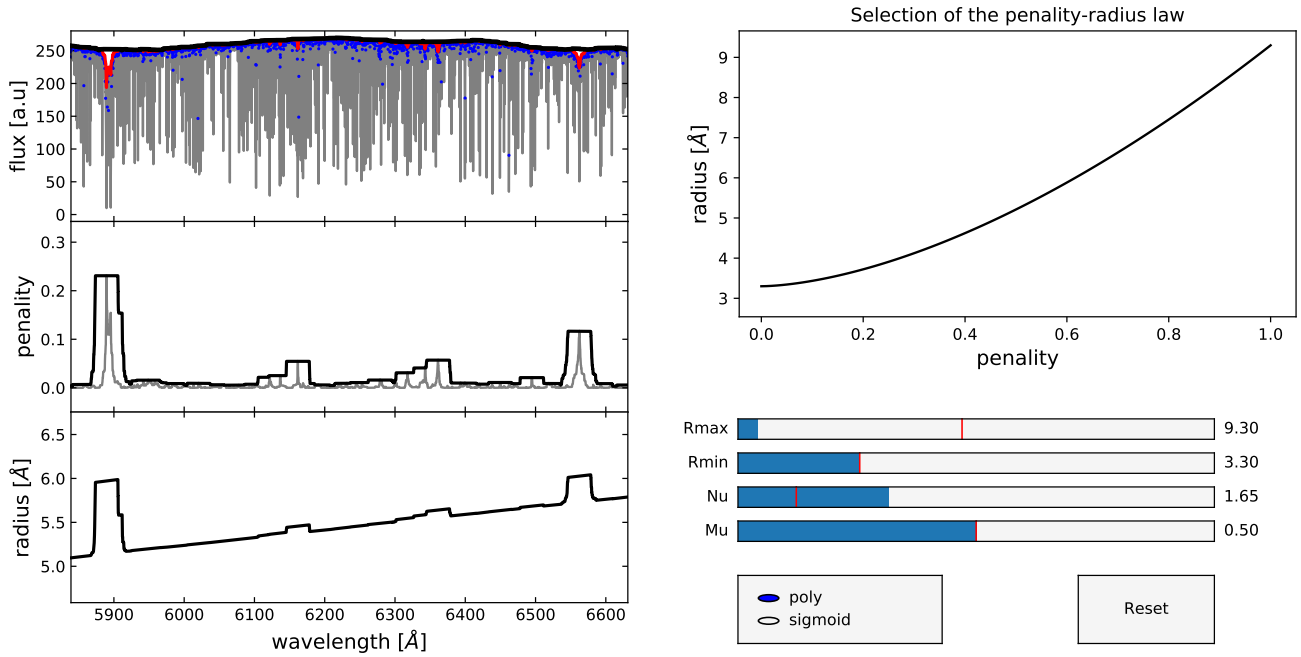

Fig. 1. Graphical interface of the second RASSINE interactive plot. In this case, the values for the parameters are bad (R_{max} too small, ν too high) but chosen to highlight the chromatic law. **Top left:** The local maxima (blue dots) extracted from the smoothed spectrum (gray curve). A continuum S_1 in a small (red curve) and S_2 in a large (black curve) window are computed with a rolling maxima on the spectrum. **Middle left:** The penalty is computed from the relative difference of S_1 and S_2 (gray curve). The penalty is rounded up and looks like a stratified box shape (black curve). **Bottom left:** Final radius values for the rolling pin along the wavelength axis. Broader absorption lines like the sodium doublet NaD around 5890 Å or the H_α at 6563 Å are penalized such that the rolling pin increases in size before reaching the line wings. The linear trend is the chromatic law $C(\lambda)$ which accounts for the natural broadening of the lines towards the red part of the spectrum. **Top right:** Penalty map relation between the radius and the penalty value. The curve is parametrised with a minimum radius R, a maximum radius R_{max} and an exponent ν. The second parameter μ is only useful when the sigmoid function is selected for the penalty map and represents the center of the sigmoid (ν being proportional to the sigmoid width).

for ***par_R*** and ***par_Rmax*** have to be provided by the user regarding the typical line width of its spectrum as well as the broadest absorption gap. If $\nu > 1$, the penalty map is convex, and roughly speaking, only high penalties will modify the radius substantially. On the opposite, if $\nu < 1$, the map is concave and only low penalties tend to avoid modifying the radius. The second available law, the sigmoid, embodies an extreme generalisation of this behaviour. Its step-like shape allows to produce a two-radius regime. Where the transition from the smallest and biggest radius is given by the sigmoid center μ and the smoothness of the transition is given by the sigmoid width ν. The best value for a given spectrum will depend on the shape on this latter. Smoothed spectra can be efficiently reduced with $\nu < 1$ whereas spectra with long oscillation due to instrumental effects or bad *1d* reconstruction are reduced better with $\nu > 1$. A good value for several cases was found with the polynomial mapping and $0.7 < \nu < 1.3$. Note that this parameter is the only one for which no automatic algorithm is available.

2.2.5. Alpha shape (rolling pin)

The next part is the main task of the code. Recall what we have now: a collection of selected local maxima, from which we would like to extract a continuum. Moreover, for each of this local maxima, a radius was assigned in the previous part (section 2.2.4). The idea is the following: the rolling pin starts at the first maxima in the lowest wavelength. It will then roll clockwise by keeping its attach to this maxima and *collide* with another maxima. This maxima is then taken to be the next anchor point, the radius is updated and the code goes on the same way. Such algorithm is called gift wrapping algorithm or Jarvis walk (Jarvis 1973) and will be presented below. Note that other algorithms of convex hull exist (Graham 1972; Eddy 1977; Phan 2007), but Jarvis walk remains the most intuitive.

A quite tricky step to understand is the scaling applied between both axis which is equivalent to change the shape of the rolling pin (or to change the distance metric). A rolling pin is represented by a circle which connects the x and y axis. However, in our case, both axes represent completely different dimensions and scales. In general, flux units are far more extended than wavelength ones. This is the equivalent of a "wall" that the rolling pin cannot overcome. For this reason, the y-axis has to be stretched relatively to the x-axis in order to provide a flatter landscape for the rolling pin. A first scaling is applied on the flux units in order to match the lengths in each directions. The flux is modified such that $f = f_i \frac{\Delta\lambda}{\Delta f_i}$, with f_i the input spectrum's flux, $\Delta\lambda$ the length of the wavelength axis and Δf_i is the height of the flux axis. In other words, after the first scaling the following equality is satisfied : $\lambda_{max} - \lambda_{min} = f_{max} - f_{min}$. The second stretching is controlled by the parameter ***par_stretching***, which is positive and bigger than 1. Namely, the flux is rescaled one

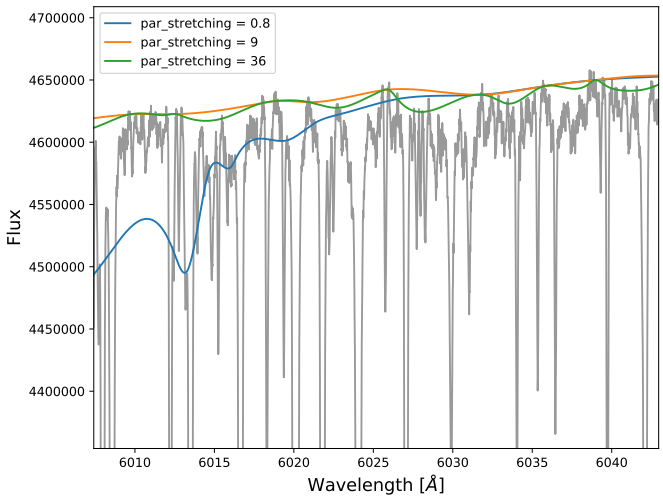

Fig. 2. Effect of different values for ***par_ stretching*** on the normalisation of the spectrum (gray curve). When ***par_ stretching*** is too low (blue curve), the rolling pin fall on the blended local maxima. As soon as ***par_ stretching*** is sufficiently high (orange curve), the rolling provides a good normalisation. A wide range of values (orange and green curves) can provide a good enough normalisation under the 1% level.

more time, leading to the relation

$$\lambda_{max} - \lambda_{min} = \boldsymbol{par\_\ stretching} \cdot (f_{max} - f_{min}). \tag{4}$$

The scaling of the axes appears like a major aspect of the process regarding the success of the alpha shape step performed later. Indeed, an inappropriate value can lead to unsuitable normalisation where the continuum is not "covering" the spectrum, but is rather going through it (see Fig.2). Hopefully, a wild range of values produces good enough results and wrong values are easily recognizable in the final product. Same conclusions were raised in Xu et al. (2019). Actually, the ***par_ stretching*** parameter is quite similar to the tension of the veil in our allegory, smaller is the value of the parameter higher is the tension. We can wonder why the veil can pierce the spectrum as on the figure (blue curve). It is logical considering that the veil is deployed from left to right and that only the local maxima exist and not the spectrum itself. If the stretching parameter is too small, the good local maxima are unreachable and the rolling pin chooses a blended local maxima in order to keep rolling.

The rolling pin process itself is only an issue of trigonometry. In more details, assuming our rolling pin is situated on a local maxima, we first obtain all the centers (to the right) such that the distance d to the current center satisfies $d < 2r$. If no such points exist, the radius is increased by a factor 1.5 until a close enough point is found. Let's call the current point P and the potential next point N (see Fig.3). First of all, we compute the vector going from P to N, called $\boldsymbol{\delta}$ whose norm is δ. We want to find the coordinates of the center C of the circle of radius r (the current value), passing by P and N. The distance h between this center and the segment δ is given by Pythagora's theorem

$$r^2 = h^2 + \frac{\delta^2}{4}. \tag{5}$$

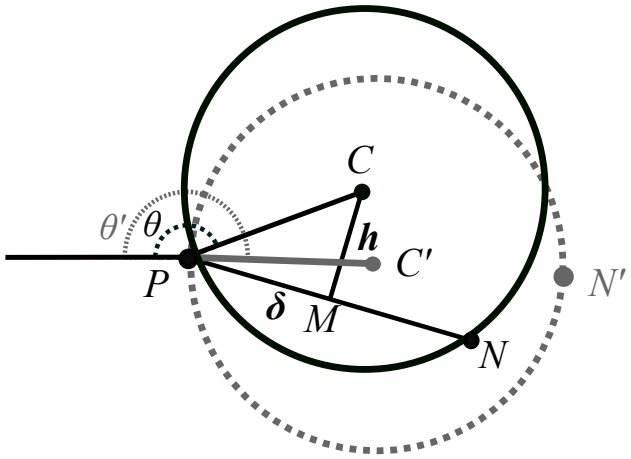

Fig. 3. The rolling pin rotates around the point P. N and N' are two local maxima. The selected one is N, because the rolling pin falls on it before falling on N'. Mathematically speaking, this is represented by the condition $\theta < \theta'$.

The (vectorial) components of the center C are given by

$$\boldsymbol{C} = \boldsymbol{P} + \frac{1}{2}\boldsymbol{\delta} + \frac{h}{\delta}\bar{\boldsymbol{\delta}}. \tag{6}$$

The second term goes from P to the middle of the segment $\boldsymbol{\delta}$, say the point M. Note that PCN is isosceles and CM is perpendicular to PN. The last term of the equation is a vector perpendicular to δ that is scaled to have norm h. In components, we have $\bar{\boldsymbol{\delta}} = (-\delta_y, \delta_x)$.

Once the coordinates of the center are found, we need to compute the rotation angle, i.e. the angle the rolling pin has to roll to touch this particular point. Let (p_x, p_y) and (c_x, c_y) be the coordinates of the current point and of the circle center respectively. Trigonometric considerations show that this angle is given by

$$\theta = \begin{cases} -\arccos\left(\frac{c_x - p_x}{r}\right) + \pi, & \text{if} \quad c_y - p_y \geq 0 \\ -\arcsin\left(\frac{c_y - p_y}{r}\right) + \pi, & \text{if} \quad c_y - p_y < 0. \end{cases} \tag{7}$$

This angle is computed for every candidate points, namely the ones that are closer than $2r$ to the current point (N and N' on the figure). The next selected point is the first candidate touched by the rolling pin, mathematically speaking the one with the smallest θ, N in this case.

This process goes on until the code reaches the end of the spectrum. All the selected points are kept in memory and represent the reliable local maxima.

2.2.6. Outliers detection

An upper envelope remains very sensitive to outliers with anomalous flux intensities, because such points are by analogy similar to needles which prevent the veil to fit the shape of the spectrum. Even if, until now, a lot of precautions have been established to suppress as many of these points as possible, we perform a last check by looking at three types of outliers: 1) border-maxima, 2) high- or low-maxima and 3) close-maxima.

1. The rolling pin will fall on the border of the spectrum on spurious maxima, simply because it starts and ends

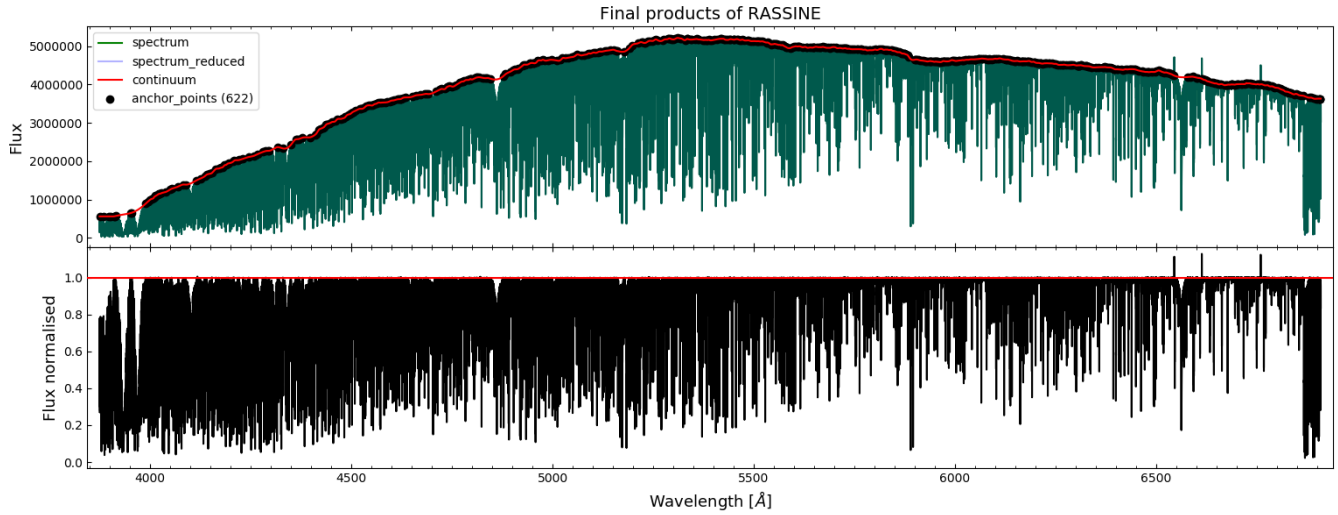

Fig. 4. Final plot interface. **Top:** The initial solar spectrum given as input to RASSINE (green curve). The continuum (red curve) as computed by RASSINE in complete automatic mode (see appendix A) based on a cubic interpolation of a selection of relevant local maxima also called anchor points (black dots). The continuum avoids three spurious cosmic peaks around 6550, 6600 and 6750 Å and jumps above the large absorption lines. **Bottom:** Output continuum normalised by RASSINE.

on them. If the spectrum was measured on an infinite wavelength grid, the continuum would touch other maxima which are actually not available because the captor has a finite range of measurement. It is necessary to prevent the continuum from falling on those spurious maxima. However, we do not want these spurious maxima to be suppressed. Instead, the values of the j first maxima are equalise to the value of the $j + 1^{\text{th}}$. At the end of this process, the first $j + 1$ maxima are then on the same height. By default, RASSINE flattens the three first and last points using the optional parameter ***count_cut_lim***.

2. *Sharp peaks* are removed by computing the left and right derivative at each selected maxima, with respect to the neighbouring ones. If a peak is too high, the left derivative will be positive and steep, while the right one will be negative and steep. The five points with the highest absolute difference between the left and right derivative can be investigated visually, and the user can choose to keep or delete them. If one point is suppressed, the next one with the highest absolute derivative difference will replace it. By default, RASSINE suppresses once the five first outliers using the optional parameter ***count_out_lim***.
3. Redundant pairs of maxima are problematic for the cubic interpolation used at the end (see section 2.2.7) since interpolating on a non-equidistant grid can sometimes lead to weird behaviours. Moreover, since these points do not bring any relevant information they are removed automatically. To detect them the code computes the distance between each pair of neighbouring maxima. Then, an asymmetric sigma clipping is performed on the distribution of the distance difference to identify the doubtful points. The algorithm removes the points in order to obtain the most equidistant residual grid.

2.2.7. Envelop interpolation

Now, we are left with several reliable local maxima, none or few of them being outliers. To generate the continuum, RASSINE interpolates linearly and cubically on them. Some points may still be too close, producing wiggles in somes cases due to the reason explained in section 2.2.6 regarding the third type of outliers.

RASSINE also provides two other normalisation procedures (similar to the former ones). Such method is necessary for noisy spectra for which the upper envelop is always more or less correlated with the noise level and thus the spectrum SNR. This tends to produce continua slightly too high as already mentioned by Xu et al. (2019). In this case, instead of taking the flux value of the local maxima to build the continuum, an average of the flux in a window determined by the parameter ***denoising_dist*** (half-window) is taken. By default, the window is made up 5 wavelength elements. In summary, RASSINE produces four continua and the user is free to decide which of them is the best according to his judgment. Nevertheless, the precision was found better in the case of the linear interpolation (see section 3.4). An example of the RASSINE final result, performed in complete automatic mode (see appendix A), is shown for a solar spectrum in Fig.4 and for four other spectra in appendix C, Fig.22.

2.3. Normalisation of a spectra time-series

When several spectra of the same star are considered, for instance during a series of consecutive nights, it is expected that the positions of the local maxima forming the continua are fixed. Hence, it is rather confusing to observe a local maxima at a specific wavelength position in only one nightly spectrum. Such issues will provide wiggles in spectra ratios or differences, which are clearly not desired (see blue curve bottom panel Fig.5). These numerical artefacts are mostly produced by the cubic interpolation (blue curve bottom panel) and can already be dumped using the linear interpolation (blue curve top panel). Nevertheless, some of them are still present and need to be cleared.

To tackle this issue, RASSINE contains a function in its library (called *intersect_all_continuum*) which detects and forms cluster of local maxima. A set of local maxima

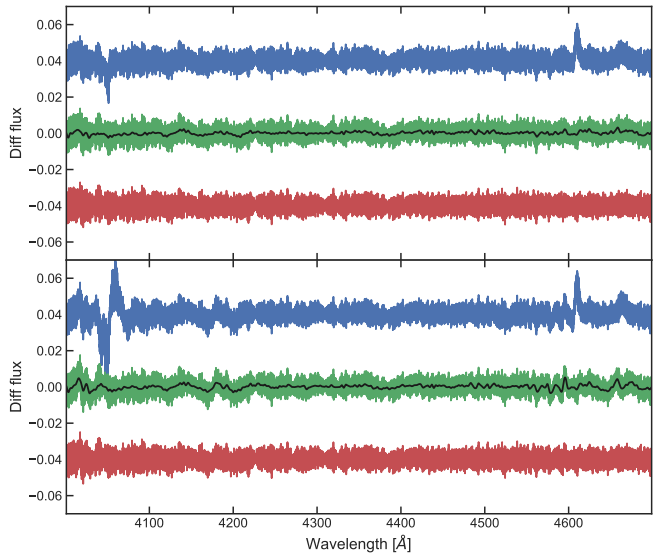

Fig. 5. Visualisation of the wiggles produced in a difference of normalised spectra (blue curve). The two spectra come from the same star and are taken one night apart. **Top:** Difference between the spectra normalised by the linear interpolated continuum. **Bottom:** Same as top with the cubic interpolation. The wiggles are more important for the cubic interpolation as expected. These are induced when a local maxima (used to build the continuum curve) is selected in one spectrum but not in the other. We resolved this issue by developing a clustering algorithm on spectra time-series which can be called with the function *intersect_ all_ continuum*. The resulting effect is a considerable improvement of the stability of the continuum (green curve). The remaining fluctuation (black curve) are under the 1% level. The residual curve after suppressing these fluctuations is indicated in red and can also be obtained from the RASSINE library function *matching_ diff_ continuum*.

with an occurence rate higher than a given threshold is selected. On the opposite, the local maxima which are selected only in a few number of nights are rejected. Also, if for one nightly spectra, two maxima are falling in the same cluster, the farthest one from the cluster center is removed. Eventually, maxima which are "missing" from the cluster are added. By doing this, we ensure that all the continua are formed with the same anchor points which considerably reduces the wiggles (see green curves Fig.5).

A second function can be used (called *matching_ diff_ continuum*) if one wishes to reduce even more the wiggles. This latter searches for the spectrum with the highest SNR and uses it as a reference. The spectra difference is computed with all the other spectra turn-by-turn, the wiggles are determined on each spectrum difference by a Savitzky-Golay filtering (black curves Fig.5) and removed from the spectra (red curve Fig.5). Note that because this step can also suppress true variations, we do not necessarily advise to use it. In case the fluctuations are still too high, it happens to be the last solution, even if rather rude. Nevertheless, this option seems to produce the most precise spectra time-series (see section 3.4 and Fig.9).

In the case where the user has to work with time-series of spectra at moderate SNR (from 30 up to 100), if the smoothing step is not efficient enough, there is a risk for the clustering algorithm to not work properly since the local maxima will be spuriously positioned. In this specific case, the user can follow this procedure. We assume that the user wants to match the continuum's anchors points of N spectra. Firstly, all the spectra should be stacked together to form a master spectrum. Some care should be taken to shift them in the same rest frame to form a high SNR spectrum before running RASSINE on this master. The user has, in one hand, to specify in the *intersect_ all_ continuum* function this master spectrum as an option, whereas in the other hand, the user has to enter the number of times M the file will be copied (typically $M > 2N$). Doing this trick gives more weights to the local maxima of the master spectrum compared to all the other moderated SNR spectra, forcing the anchor points' locations to be at the master spectrum's ones. Eventually, the user can run the *matching_ diff_ continuum* function with the N spectra and the master spectrum.

3. Results

3.1. Broad lines absorption

As a first test, we investigated if broad absorption lines were correctly normalised or if RASSINE was disturbing their profile. To do so, we chose as test cases 61 Cyg A, the Sun and HD142 which are three stars probing a wide range of spectral-types and CCF FWHM values (see Table 3 in appendix). The spectra were normalised in automatic mode (see appendix A) with intermediate tension (***par_ stretching*** = 'auto_0.5') and polynomial law (ν = 1). We focused on two broad lines of first interest which are the sodium doublet NaD and H_α. In Fig.6, we compared the spectrum normalised by the alpha shape with a more classical method which goes as follows. Two spectral windows, as much free as possible free of stellar lines, were selected on the right and left side of each broad lines, each window being different from each star. The average flux was measured in each window before removing the linear fit from the spectrum.

On 61 Cyg A, which is a K5V star, we observed that the classical method leads to a too low continuum level for both broad absorption lines. This produces a normalised spectrum with several values higher than 1. It is due to the fact that, for cool stars, the regions selected as continuum are necessarily contaminated by absorption lines due to the high lines density. The normalisation produced by RASSINE appears to be better for this star, the continuum level being higher. Note that this latter could still be too low regarding the true continuum, which is free of blended lines.

For the Sun, we observed a strong discrepancy for the sodium doublet. Indeed, whereas RASSINE selected the interline region as part of the continuum. The classical method provided a value 2% lower which is also found in other solar atlases, namely the Kitt peak (Wallace et al. 2011) and the IAG ones (Reiners et al. 2016). Hence for the Sun, the tension used in the default automatic mode was not sufficient. We solved this issue by increasing the tension (***par_ stretching*** = 'auto_0.0') and decreasing the coefficient ν down to 0.7 for the polynomial mapping. After doing so, RASSINE was in well agreement with the classical method. However, both methods led to a thinner H_α width than the IAG atlas which could be explained by different solar activity level.

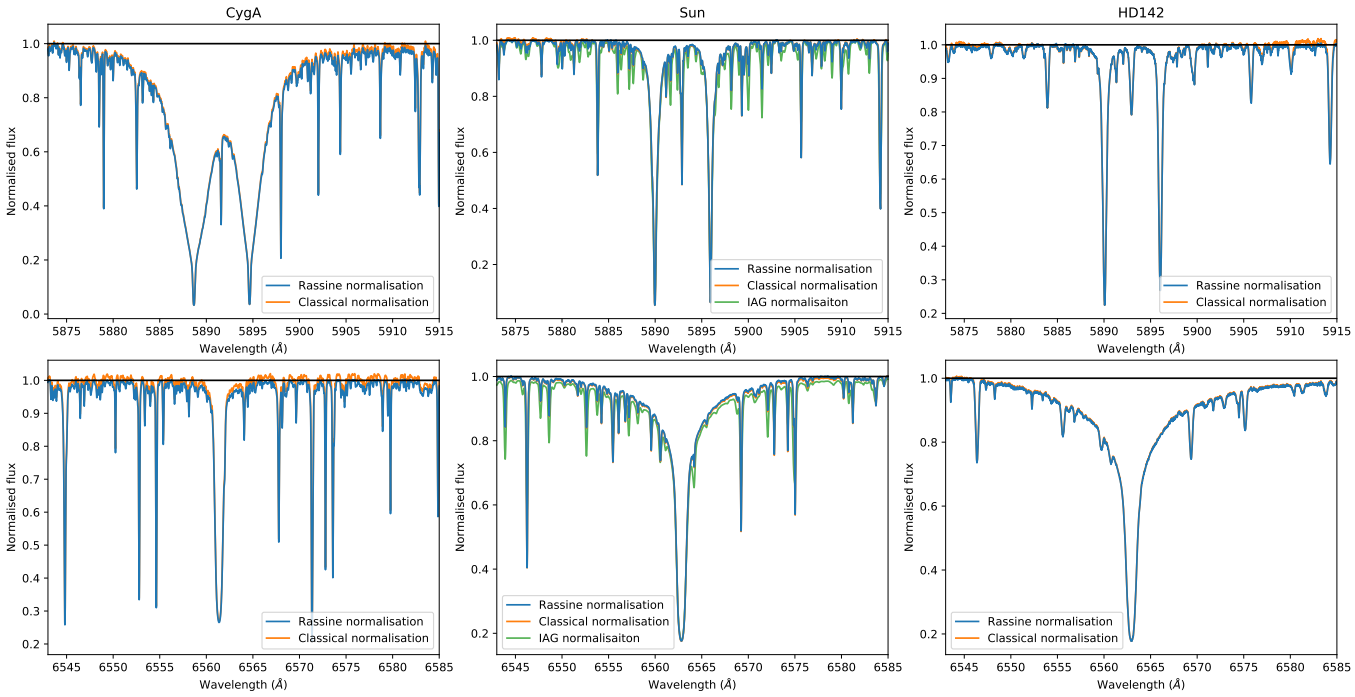

Fig. 6. Comparison between the RASSINE normalisation and classical method (linear fit) around spectral broad lines for three different spectral type stars : K5V, G2V and F7V, namely from left to right : α Cyg A, the Sun and HD142. **Top:** Sodium doublet NaD lines. **Bottom:** H_α line. Both α Cyg A and HD142 were normalised in automatic mode with intermediate tension (***par_ stretching*** = 'auto_0.5') and polynomial mapping $\nu = 1$. For the Sun, higher tension was needed in order to not select the interline region as part of the continuum. High tension (***par_ stretching*** = 'auto_0.0') and $\nu = 0.7$ appeared has satisfying. A comparison with a solar atlas (IAG, Reiners et al. (2016)) degraded at the HARPN resolution $R \sim 115'000$ is also displayed and show that ours normalisations are matching well the NaD lines, but lead to a thinner H_α width compared to the IAG. The classical method is also in better agreement with our normalisation even if the discrepancy can be due to the different instruments used or due to the different times of the observations.

For HD142, a F7 dwarf, good agreement is found for H_α. However another feature is visible in the sodium doublet by the classical method. A clear wiggle is visible in the right part of the spectrum which is not induced by the method itself, since a linear fit was performed. Such wiggles are thus inherent to the *1d* spectrum and are either due to the instruments optics or produced during the *1d* construction. This latter option being more likely since this wavelength range is precisely situated in the overlapping region of the HARPS order 56 and 57.

3.2. Evaluation of the accuracy of RASSINE

Even if not the primary interest of the RASSINE code, the accuracy of the normalisation has to be tested in order to evaluate if the derived lines depths are accurate in terms of absolute value. Since the alpha shape method is an upper envelope approach, it is known that the continuum level may be too high, depending on the SNR of the spectrum, which already introduces an incorrect absolute depth offset (Xu et al. 2019). Moreover, there is no way to avoid circular reasoning when we want to measure accuracy comparing our normalised spectra with a normalised stellar template. Indeed, to generate a stellar template, few atmospheric parameters are needed as input, which are themselves determined from spectra observations and thus depend somehow on the way the latter have been normalised. Another strong difficulty is to obtain a stellar template with atmospheric parameters as close as possible to the observed star. Indeed, due to the high dimensionality of the stellar atmospheric parameter space, templates library often mainly focus on a good coverage of the effective temperature and gravity surface parameters, at the expense of abundances diversity. Also, the resolution of a spectra is generally not uniform which could introduce wrong line depth since the stellar template is convoluted with a kernel at a fixed resolution. As an example, for HARPS, resolution R is varying typically between $95'000$ and $110'000$ (private communication). For this reason, the computation of the accuracy cannot be performed by measuring for instance a difference of line depth.

To measure the accuracy, we compared a spectrum normalised by RASSINE with a normalised synthetic spectrum of reference. First, we interpolated the normalised synthetic spectrum on the same wavelength grid than our normalised spectrum. We then extracted the flux value at the same wavelength position that the anchor points used to build our continuum. By definition our continuum will take the value of 1 at these locations and thus the standard deviation of this set of $1 - f$ values provides a good metric for the accuracy, where f is the flux value of the normalised synthetic spectrum. Note that with this approach, no conclusion can be raised about the accuracy value of the continuum between two anchor points and thus the accuracy score for the linear and cubic interpolated continuum are the same. If we consider that the continuum is a smooth function between two anchor points, which is a good approximation here with an average distance between two anchors points

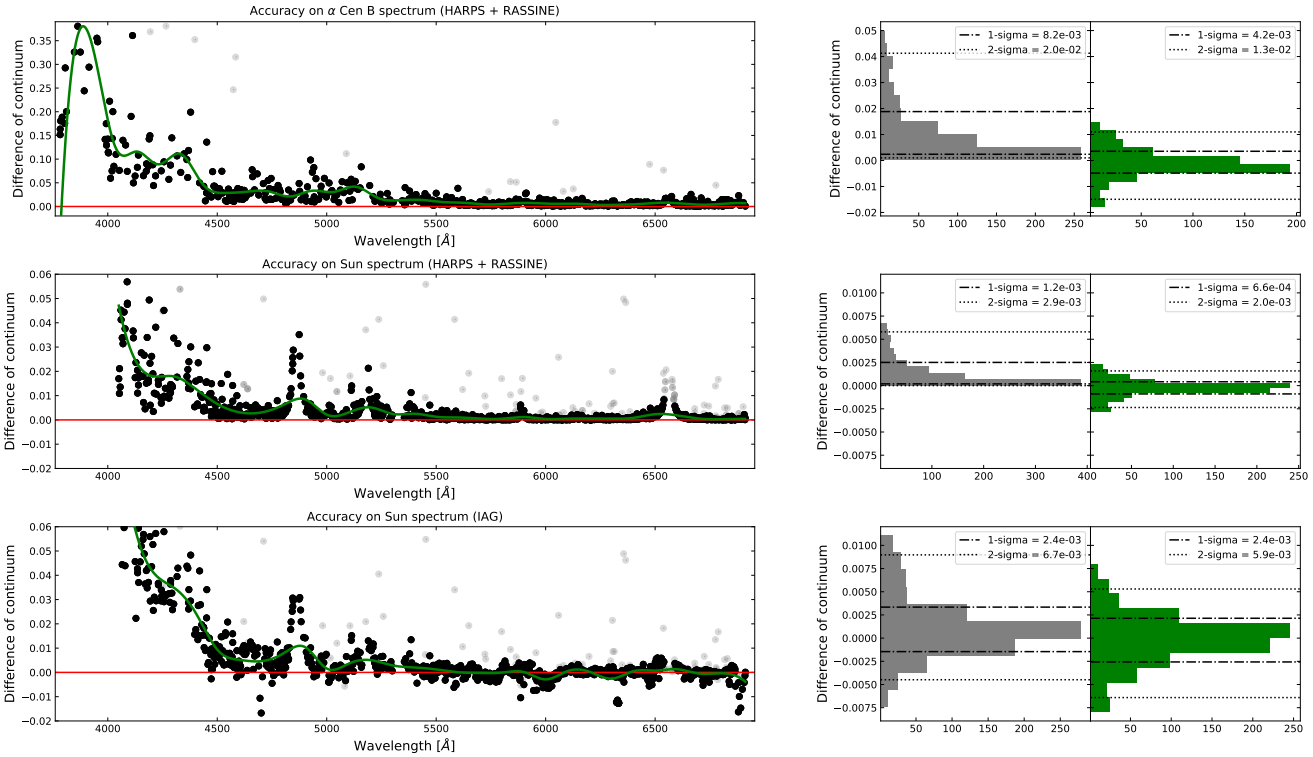

Fig. 7. Computation of the RASSINE accuracy by measuring the continuum difference level between our spectra and synthetic stellar templates. **First row:** Difference of continuum for α Cen B between the MARCS template and the normalised spectrum of RASSINE as a function of the wavelength. A few continuum levels cannot be extracted correctly in the synthetic template due to the the presence of stellar lines absent in our spectra (gray dots). The distribution of the continuum difference is plotted on the right as well as the distribution once the chromatic trend (green curve) is removed. **Second row:** Difference of continuum for the Sun between the ATLAS template and the normalised spectrum of RASSINE. **Third row:** Difference of continuum for the Sun between the ATLAS template and the IAG normalised spectrum.

of 4.5 Å, the derived accuracy can be considered as an average value representative of the alpha shape strategy for the whole spectrum between 3800 to 6700 Å. Such metric is however completely insensitive to a global offset between the normalised spectrum and the template since such artefact will affect the mean and not the standard deviation of the distribution, but they are strongly unlikely for high SNR spectra.

New spectral lines were sometimes present in the synthetic spectrum and thus produced easily identifiable outliers (see gray dots in Figure.7). We rejected those points by performing a rolling k-sigma clipping in a window of 20 adjacent points, computing in each window the median and the median absolute deviation (MAD). We rejected a point if its distance from the rolling median is greater that ten times the rolling MAD. The rolling k-sigma clipping was performed iteratively until no more outliers were detected. The accuracy was then extracted as the sigma-width of the distribution, taking from the percentile definition. The 1-sigma width being equivalent to the half distance between the 84^{th} and 16^{th} percentiles. The 2-sigma width is samely defined with the 97.5^{th} and 2.5^{th} percentiles and will be the default value when we will refer to *accuracy*.

We used the POLLUX database (Palacios et al. 2010) which is a library of spectra containing high resolution stellar templates normalised. We evaluated the RASSINE accuracy on two stars: α Cen B and the Sun. We use the spectra from the α Cen B 2008, but same results are obtained with the 2010 dataset. The atmospheric parameters chosen for the templates are given in Table 1 and were derived from MARCS (de Laverny et al. 2012) and ATLAS (Kurucz 2005) models. For the Sun, it is also possible to use the IAG solar atlas (Reiners et al. 2016) in order to compare the accuracy of the alpha shape strategy versus iterative fitting methods. Solar atlas as well as synthetic spectra were degraded at the HARPS/HARPN resolution, shifted to match the rest frame of the stellar spectra and interpolated on the same wavelength grid. Since the IAG spectra begins at 4047 Å, we removed shorter wavelength range from the HARPN spectrum. Also, because the IAG atlas was taken at a different BERV value, a few locations of flux extraction were sometimes located on tellurics lines. We removed these points by performing the same rolling k-sigma clipping that for the synthetic spectra.

Table 1. Atmospheric parameters chosen to generated our reference synthetic spectra. Spectra were generated either by a MARCS (M) or ATLAS (A) model specifying an effective temperature T_{eff} in K, a surface gravity $\log g$, a metalicity [Fe/H] and [α/Fe] value and a micro-turbulence velocity ξ_t in $\mathrm{km\,s^{-1}}$.

Star	Model	T_{eff}	$\log g$	[Fe/H]	[α/Fe]	ξ_t
α Cen B	M	5250	4.5	0.25	0	1
Sun	A	5800	4.5	0	0	2

We observed than our $1-f$ distribution is always positive, meaning that our continuum level remains always below the synthetic continuum (in absolute flux) ; an observation coherent with the alpha shape strategy. Indeed, the only possibility for our continuum level to be higher than the synthetic one would be that at least one of the local maxima that we are selecting as anchor point is similar to an "emission-peak" either induced by a cosmic or a contamination, but such outliers are already removed during the reduction (see section 2.2.6). On the opposite, RASSINE will mainly tend to select too many anchors points as part of the continuum, this effect being enhanced while the tension parameter is reduced. For instance, RASSINE considers here the inter-region of the CaII H&K line as part of the continuum, whereas in the synthetic template this region is clearly not part of the continuum because still situated in the merging zone of the two strong lines wings. This discrepancy produces the strong excursion of 0.35 around 3900 Å, against 0.20 for the solar spectrum (not visible on the figure since, HARPS spectra has been cut below 4047 Å).

The accuracy is lower for α Cen B in the blue part of the spectrum due to the CH molecular band at 4300 Å, also called G-band, and the CN violet molecular band at 3883 Å which are deeper in cool stars and sensitive to stellar activity (Berdyugina & Usoskin 2003). The accuracy for α Cen B is estimated around 1.9% which is 6 times worse than the accuracy for the Sun around 0.30%. Such a conclusion is of course well expected since the cooler the star is, the more blended the lines are. Thus, less local maxima really probe the continuum.

Opposite to the alpha shape strategy, the IAG atlas normalised by low polynomial fit can present either a higher continuum, or a lower one, with respect to the synthetic continuum. We also clearly see here a strong issue with polynomial fit which is their instability on the border of the fitting domain, only visible in the blue since the IAG spectrum was measured in the red until 10650 Å. A clear excursion of 3% around 4861 Åis visible in both IAG and HARPS spectra.

Once extracted, the chromatic trend can be subtracted on the RASSINE normalised spectrum. Here we performed a simple correction by binning the points in bins of 100 Å before interpolating with the cubic fit between the bins. By doing so, it is then possible to measure the dispersion around the chromatic trend which is the new accuracy score of the normalised spectrum. The improvement is more notable for α Cen B since the chromatic trend is more pronounced on the cool star. The new accuracy scores were respectively of 1.3% and 0.26% for α Cen B and the Sun. Finally, we can comment on the scores reached by iterative fitting method which are between 2 and 3 times lower than with the alpha shape strategy, namely with score of 0.82% and 0.47% for the IAG. The improvement is clearly visible by the lower dispersion between 5000 and 6500 Å which is much smaller with the RASSINE normalisation.

3.3. Evaluation of the precision of RASSINE

Another important score to determine is the code's precision. For this purpose, we normalised several very high SNR spectra, taken during successive nights, of α Cen B. We selected 9 consecutive nightly-binned spectra at minimum activity in 2008, from BJD = 2′454′550 to BJD = 2′454′558. We also selected 13 consecutive nights in 2010, from BJD = 2′455′288 to BJD = 2′455′301, even if this dataset is known to be contaminated by a large active region (Dumusque et al. 2015) which could change the depth of the spectral lines (Thompson et al. 2017; Wise et al. 2018). Since we are looking from night-to-night variation and that the rotational period is around 36 days (DeWarf et al. 2010), we expect the variation induced by stellar activity to be rather small. It will be confirmed later since both dataset produce the same scores.

All the binned spectra have similar SNR ranging in 2008 (resp. 2010) from 1050 up to 1369 (resp. from 1293 up to 2390) at 5500 Å. For each of them, the lines depths were measured and compared between adjacent nights. Because spectra are close in time, the line depths should not change significantly from a night to the next one and its variation is thus a direct measurement of the RASSINE precision. The spectra were reduced in complete automatic mode with a Savitzky-Golay filtering in a ***par_ smoothing_ box*** of 6 pixels, a polynomial penalty mapping ($\nu = 1.0$) and with an intermediate tension (***par_ stretching*** = 'auto_0.5').

The line depth was determined by the minimum of a parabola fitted on the core of the normalised lines inside a window of ± 2.5 km s^{-1}. The lines to be fitted were defined as in Cretignier et al 2019. The uncertainties on the line depth were derived only considering photon noise. Only lines which were not contaminated by tellurics at 2% were kept for the analysis, where the telluric spectrum was generated by Molecfit (Smette et al. 2015). Two examples of line depth difference between two adjacent nights are displayed in Fig.8. We measured the precision for the 8 and 13 pairs of adjacent nights (see blue curves Fig.9) by computing the weighted RMS of the line depth difference. Due to the poor constrain of the interpolation on the border of the spectrum, the weighted RMS was computed excluding the first and last 20 Å of the spectra. The photon noise is higher in 2008 than in 2010 due to the lower SNR of the spectra for this period. The photon noise induces an average systematics on the line depth measurement about $1.21 \cdot 10^{-1}$ % and $8.72 \cdot 10^{-2}$ % respectively. We observe that the precision is lower for the cubic interpolation than for the linear interpolation. For the former, the RMS significantly changes depending on the pair of nights studied which is related to the amount of wiggles produced in the spectra difference. In average, the precision of RASSINE for the linear interpolated continuum is found to be $1.65 \cdot 10^{-1}$ % and $1.44 \cdot 10^{-1}$ % for each year. After measuring the quadratic difference, D_2, with the photon noise, the RASSINE uncertainty is found to be of $1.13 \cdot 10^{-1}$ % and $1.14 \cdot 10^{-1}$ % which is a good agreement between both datasets.

However, RASSINE produces systematics which are not white noise but structured in wavelength (see inner plot Fig.8 or Fig.5). These structures are produced when a local maxima is selected in a zone where no local maxima were detected for the other spectrum (see section 2.3) and constitute a major limitation to achieve high precision.

3.4. Evaluation of the precision of RASSINE for a spectra time-series

Until now, we have investigated the precision and accuracy of RASSINE for the normalisation of individual spectra. Each spectrum was thus reduced independently with

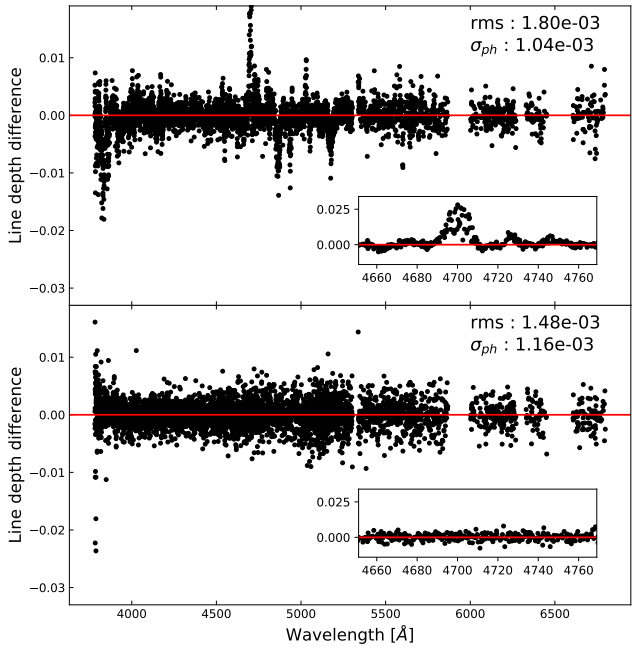

Fig. 8. Line depth difference between two spectra taken with a time delay of one day. The uncertainties coming from the photon noise are not displayed for sake of clarity but their median value is indicated as well as the weighted RMS used to measure the dispersion. **Top:** Spectra from α Cen B 2010 dataset. It matches with the fourth blue square point on the bottom panel Fig.9. Both spectra are normalised by the RASSINE cubic continuum which is the least precise one (see Fig.9). Part of this jitter is not Gaussian, as displayed in the inner plot which is a zoom at 4700 Å where the discrepancy is observed to be about 2.5%. The structures are related to the method used to interpolate on the local maxima, here the cubic. **Bottom:** Same as top for the last data point of the 2008 dataset derived from the linear interpolation and after performing the clustering and low frequency filtering (see section 2.3), it matches the 8^{th} red circle data point top panel Fig.9).

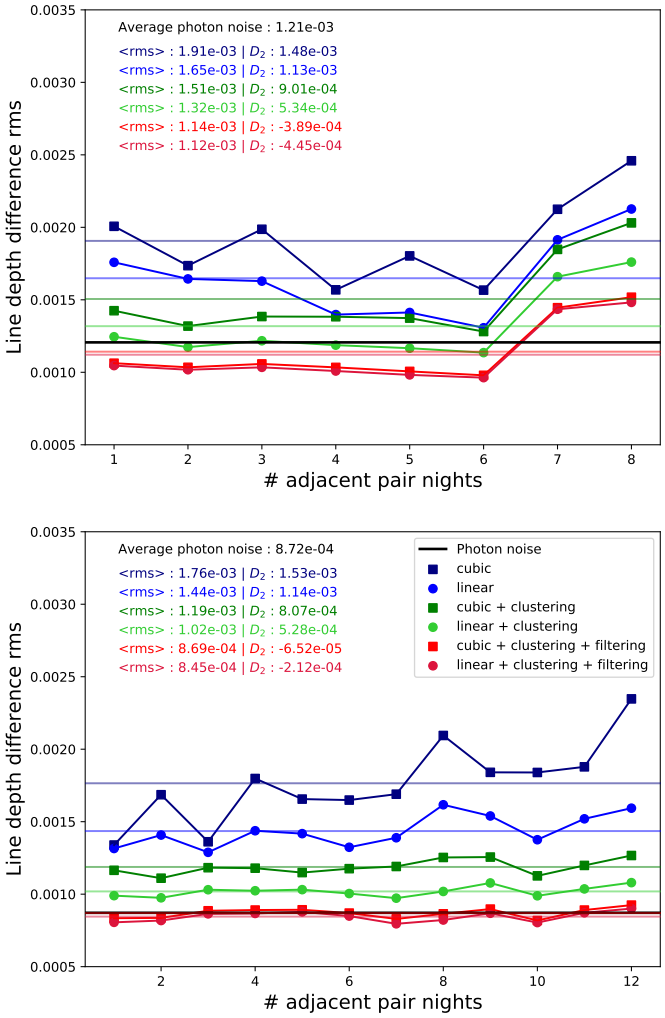

Fig. 9. Measurement of the RASSINE precision through line depth evaluation. A series of adjacent nights were selected. For each night, the line depth was computed on the full spectrum, excepted on telluric regions. The weighted RMS (color dots) was extracted from the distribution of the line depth difference where the difference is taken between two adjacent nights (see for example Fig.8). The average RMS $\langle rms \rangle$ on the night-to-night differences is measured (solid lines). The values of $\langle rms \rangle$ for each reduction as well as its quadratic difference D_2 with the average photon noise (black line) is indicated. Negatives values meaning that the precision is better than the photon noise level. **Top:** 2008 α Cen B dataset (see the text for the precise dates). **Bottom:** 2010 α Cen B dataset.

respect to the others, without sharing any information between the different nights. However, if all the spectra are coming from the same star (and thus formed a spectra time-series), it is possible to gather the information of all the continua in order to improve the code's stability. We already presented this aspect in section 2.3.

We reperformed the previous analysis of the line depth measurements on the 9 and 13 nightly adjacent spectra of the 2008 and 2010 α Cen B dataset, applying this time the *intersect_all_continuum* and *matching_diff_continuum* functions on each year separately. Recall that the former produces clusters of local maxima by identifying those with the highest occurrence rate whereas the second is a low frequency filter performed on the previous output (see Fig.5). The same measurement of line depth difference was performed on adjacent night-to-night spectra in order to extract the weighted RMS of the distribution. The precision on the reduced spectra with the *intersect_all_continuum* function are the green curves in Fig.9, whereas those from *matching_diff_continuum* are in red.

Again the linear interpolation appears as more precise than the cubic one due to the wiggles produced by the latter. Nevertheless, we note that after performing the *matching_diff_continuum* function, no difference is observed. The two datasets are again in relatively well agreement for the clustering algorithm with a precision (averaging both years and after removing the photon contribution) of about $8.54 \cdot 10^{-2}$ % and $5.31 \cdot 10^{-2}$ % for the cubic and linear interpolation. After performing the low frequency filter, the precision is compatible with the photon noise level and is even slightly smaller which is possible since a high frequency filter is applied on the spectra, at the first stage of the RASSINE reduction, to produce the continua. Remark that by construction, the red curves are necessarily the flatter ones

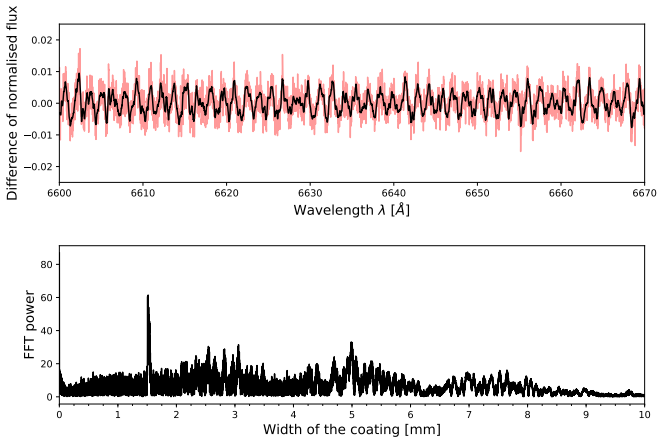

Fig. 10. Top: Zoom in the red 6640 Å region of the normalised spectra difference corresponding to the last red dot of the 2008 dataset in Fig.9. A clear pattern of interference is visible the spectra difference (red curve) even more noticeable once the difference is smoothed (black curve). **Bottom:** Fourier transform of the spectra difference. A power excess is visible for a thin layer with a width of 1.51 mm.

which does not necessarily mean that they are the best continua. The rather flat behaviour of the RMS as a function of the pairs of night is a good sight that our normalisation is now stable enough.

The only notable discrepancy is observed for the last two nights' pairs of the 2008 dataset for which the RMS seems slightly higher, regardless the algorithm used to build the continuum. After investigations, it turned out that 2008 HARPS spectra were contaminated by an interference pattern caused by a filter installer in front of the Tungsten lamp until an upgrade of the instrument in 2009 occurred. The pattern is clearly visible in the difference of the normalised spectra as highlighted in Fig.10. The peak-to-peak of the smoothed pattern at 6400 Å is about 1% in normalised flux units. The periodicity $\Delta\lambda$ of the pattern at the wavelength λ is directly related to the width w of the thin layer producing the interference according to the law :

$$w = \frac{\lambda^2}{2\Delta\lambda} \tag{8}$$

This width can be extracted from the excess of power in Fourier space domain where a broad peak was visible between 1.50 and 1.55 mm. Such patterns are also present in transmission spectra of others instruments when filters are for instance introduced in the calibration arm, producing fringing in the products of the calibrations frames like flat-field. In fact, RASSINE is offering an advantageous solution to correct them if the periodicity $\Delta\lambda$ is bigger than the distance between two anchors points used to build the continuum. Unfortunately, here, it is not the case since $\Delta\lambda = 0.69$ Å, far below the typical average distance of two consecutive anchors points around 3.93 Å. A solution could be to decrease the tension in the veil but such solution would not manage to remove the pattern in broad absorption lines and is thus not advised. However, RASSINE offers an easy way to point out such features in spectra difference once spectra are normalised, providing opportunities to correct them for instance by Fourier filtering after identifying the highest peak, but we keep such cleaning process for a future paper.

Table 2. Summarize of the different scores reached with the RASSINE automatic mode. Only the scores related to the linear interpolated continuum are presented. Accuracy and accuracy with chromatic filtering are both written. The precision values are not subtracted from the photon noise also displayed in the last row. The three values of precision are respectively for individual spectra normalisation, spectra with clustering and clustering + filtering. All the values are in %.

Scores	α Cen B 2008	α Cen B 2010	Sun	Sun (IAG)
Accuracy	2.0	2.0	0.29	0.67
Accuracy (chrom. filt.)	1.3	1.2	0.20	0.59
Precision	0.17	0.14	-	-
Precision (clust.)	0.13	0.10	-	-
Precision (clust. + filt.)	0.11	0.08	-	-
σ_{phot}	0.12	0.09	-	-

4. Conclusion

In this paper, we presented the RASSINE Python code, a tool created to facilitate the normalisation of stellar spectra. The realisation of the code was motivated by the challenge to develop a coherent and robust normalisation algorithm of *1d* spectra, allowing to deal with different spectral types and SNR, with a limited number of parameters to control. Because parameters are unavoidable to realize such a task, RASSINE proposes in one hand to display interactive graphical tools to help the user in the choice of the values of the parameters. On the other hand, the code can provide a first guess for 5 of the 6 relevant parameters, calibrated by the SNR level of the spectrum and the FWHM value of its CCF. The code has been tested with 4 instruments spanning different instrumental resolution like CORALIE, HARPS, HARPN and ESPRESSO and has produced each time visually satisfying results. In this paper, we focused on really high SNR spectra (SNR $>$ 500) which are the basic objects for which RASSINE is expected to be efficient. However, low SNR spectra can theoretically also be reduced with RASSINE if the smoothing step is correctly performed. No numerical results were however derived in this case.

In this paper, we tested the accuracy and the precision of an automatic mode available for the code on the 2008 and 2010 α Cen B dataset, reminding that no guaranty is given for this mode to provide the best values of accuracy. We showed that the linear interpolation provided better continua that the cubic interpolation in all cases since this latter is producing unwanted artefacts. The accuracy was measured on the 2008 α Cen B dataset and on solar spectra by comparing our continuum with stellar template from MARCS and ATLAS model. An accuracy of 1.9% was de-

rived for the former star, whereas a value a 0.26% was found for the Sun, so 3.3 times better than the 0.82% obtained with low polynomial fit of the IAG. Correcting a clear chromatic trend allows to increase the accuracy of the spectra up to 1.3%, 0.26% and 0.47%. The precision of the algorithm was found to be better than the accuracy. For α Cen B, the precision is about 0.11% after removing the photon noise contribution. Other algorithms were developed in order to stabilised even more the continuum, in particular if the goal of the study is to work with a spectra time-series. A clustering algorithm allowed to increase the precision up to 0.05%. The photon noise level can even be reached after applying a low frequency filtering on the residual. All the scores are summarized in Table 2.

Such tool can find applications in numerous of examples such as computations of stellar atmospheric parameters, the development of personal mask for each star in the context of radial velocities (Bourrier et al., in prep), line-by-line variability related to stellar activity (Cretignier et al., in prep) or to correct the interferences observed in transmission spectra. Such normalisation algorithm could also improve the RV derived, reducing the jitter in RV time-series by providing for instance a better color correction and facilitating further correction algorithm for the stellar activity or telluric contamination.

5. Acknowledgments

We are grateful to Nathan Hara and its positive feedback. We thanks Lila Chergui and Yannick Demets for their help regarding the English. 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. This work has been carried out within the frame of the National Centre for Competence in Research "PlanetS" supported by the Swiss National Science Foundation (SNSF). M.C, J.F and R.A. acknowledges the financial support of the SNSF. F.P. greatly acknowledges the support provided by the Swiss National Science Foundation through grant Nr. 184618.

References

Asaeedi, S., Didehvar, F., & Mohades, A. 2013, arXiv e-prints [arXiv:1309.7829]
Berdyugina, S. V. & Usoskin, I. G. 2003, A&A, 405, 1121
Blanco-Cuaresma, S., Soubiran, C., Heiter, U., & Jofré, P. 2014, A&A, 569, A111
Bouchy, F., Doyon, R., Artigau, É., et al. 2017, The Messenger, 169, 21
Cosentino, R., Lovis, C., Pepe, F., et al. 2012, in Proc. SPIE, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV, 84461V
de Laverny, P., Recio-Blanco, A., Worley, C. C., & Plez, B. 2012, A&A, 544, A126
DeWarf, L. E., Datin, K. M., & Guinan, E. F. 2010, ApJ, 722, 343
Dumusque, X., Glenday, A., Phillips, D. F., et al. 2015, ApJ, 814, L21
Eddy, W. F. 1977, ACM Trans. Math. Softw., 3, 398
Graham, L. 1972, Information processing letters
Gray, D. F. 2005, The Observation and Analysis of Stellar Photospheres
Jarvis, R. 1973, Information processing letters
Kurucz, R. L. 2005, Memorie della Societa Astronomica Italiana Supplementi, 8, 14
Malavolta, L., Lovis, C., Pepe, F., Sneden, C., & Udry, S. 2017, MNRAS, 469, 3965
Mayor, M., Pepe, F., Queloz, D., et al. 2003, The Messenger, 114, 20
Palacios, A., Gebran, M., Josselin, E., et al. 2010, A&A, 516, A13
Pasquini, L., Cristiani, S., García López, R., et al. 2010, in Proc. SPIE, Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III, 77352F
Pepe, F., Bouchy, F., Queloz, D., & Mayor, M. 2003, in Astronomical Society of the Pacific Conference Series, Vol. 294, Scientific Frontiers in Research on Extrasolar Planets, ed. D. Deming & S. Seager, 39–42
Pepe, F., Mayor, M., Galland, F., et al. 2002a, A&A, 388, 632
Pepe, F., Mayor, M., Rupprecht, G., et al. 2002b, The Messenger, 110, 9
Pepe, F., Molaro, P., Cristiani, S., et al. 2014, Astronomische Nachrichten, 335, 8
Phan, T. 2007, Annales Mathematicae et Informaticae
Queloz, D., Mayor, M., Weber, L., et al. 2000, A&A, 354, 99
Reiners, A., Mrotzek, N., Lemke, U., Hinrichs, J., & Reinsch, K. 2016, A&A, 587, A65
Savitzky, A. & Golay, M. J. E. 1964, Analytical Chemistry, 36, 1627
Schwab, C., Rakich, A., Gong, Q., et al. 2016, in Proc. SPIE, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 99087H
Smette, A., Sana, H., Noll, S., et al. 2015, A&A, 576, A77
Thompson, A. P. G., Watson, C. A., de Mooij, E. J. W., & Jess, D. B. 2017, MNRAS, 468, L16
Tody, D. 1986, in Proc. SPIE, Vol. 627, Instrumentation in astronomy VI, ed. D. L. Crawford, 733
Tody, D. 1993, in Astronomical Society of the Pacific Conference Series, Vol. 52, Astronomical Data Analysis Software and Systems II, ed. R. J. Hanisch, R. J. V. Brissenden, & J. Barnes, 173
Škoda, P. 2008, in Astronomical Spectroscopy and Virtual Observatory, ed. M. Guainazzi & P. Osuna, 97
Wallace, L., Hinkle, K. H., Livingston, W. C., & Davis, S. P. 2011, ApJS, 195, 6
Wise, A. W., Dodson-Robinson, S. E., Bevenour, K., & Provini, A. 2018, AJ, 156, 180
Wyttenbach, A., Ehrenreich, D., Lovis, C., Udry, S., & Pepe, F. 2015, A&A, 577, A62
Xu, X., Cisewski-Kehe, J., Davis, A. B., Fischer, D. A., & Brewer, J. M. 2019, arXiv e-prints, arXiv:1904.10065

Appendices

A. Description of the automatic procedure

Even if RASSINE is a code with a complete interactive Python interface taking advantage of widgets and sliders, making the choice of the parameters as easy as possible, some may argue that it could be an exhausting exercise for an unfamiliar user to find them, especially if they have to deal with several spectra of different stellar types or several SNR spectra of one star. Hopefully, these parameters can often be approximated directly from the spectra. All of them, except the vicinity parameter and the penalty law, can be replaced by the key word "auto". In this case, several algorithms define automatically a first guess for the values. The reader must still keep in mind that there is no guarantee regarding the quality of the final product. Such analysis should only be used to help the user getting a first guess, or when a lot of spectra of the same star have to be reduced. Also, even if theoretically nothing prevent RASSINE to be used on *2d* spectra, the user has to be aware that the automatic mode presented below was calibrated on *1d* spectra.

The first information one can get from a spectrum is the typical line width which is determined by the instrumental resolution, projected stellar velocity or macroturbulence values (Gray 2005). As said previously, the easiest method to determine its value consists in computing the FWHM of the CCF which represents roughly an average

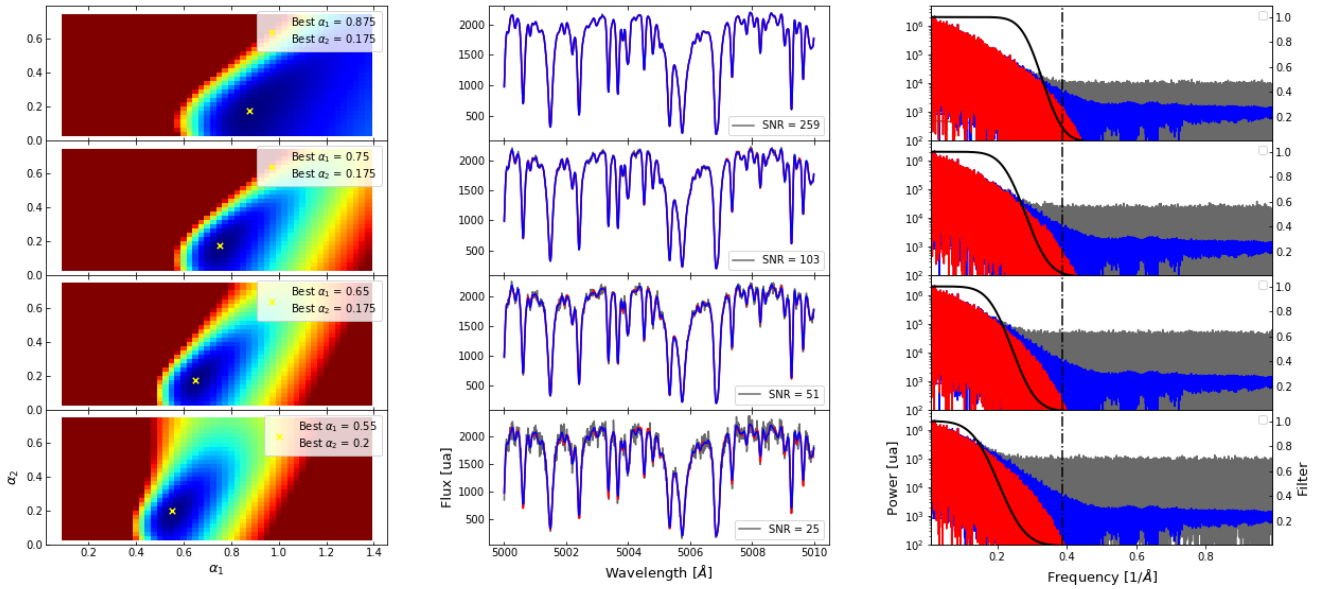

Fig. 11. Calibration of the α_1 and α_2 as a function of the spectrum SNR for the erf Fourier filter. The spectrum of α Cen B was used as a high SNR input, before injecting the photon noise. **Left:** (α_1, α_2) parameter space for different SNR realisations. The simulations were run on a grid of evenly spaced value of α_1 and α_2. The colors encode the score of the metric we used (a weighted standard deviation) to measure the quality of the smoothed spectrum. The local minimum of the parameter space is indicated by a yellow cross. Noisier is the spectrum, lower is the best α_1. The best value for α_2 appears rather constant around 0.175. **Middle:** A high SNR spectrum (blue curve) is considered to be noisy-free. Different realisations of Poisson noise are generated (gray curves) in order to determine the best couple (α_1,α_2) whose corresponding smoothed curve (red curve) will be the most similar to the original one. Only a subpart of the spectra is displayed for graphical considerations. **Right:** Fourier transform of the spectra of the middle column (blue curve). The high frequency noise (gray curve) increases significantly from $SNR = 259$ to $SNR = 25$ such that the center of the erf function is moving toward shorter frequencies. The erf filter corresponding to the best (α_1,α_2) couple is displayed (black curve). The red curve is the Fourier transform after smoothing by the function. The σ-width of the CCF is also indicated (dashed-dotted line).

line profile. A rude estimation of the spectrum continuum is first formed with a rolling maximum in a window of 30 Å. Because of the step-like shape of the continuum, this latter is smoothed with a rectangular kernel of 15 Å. If the parameter ***CCF_mask*** is put in 'master' mode, the normalised stellar spectrum is then convoluted with a binary mask where the line position is taken at the local minima position. The weights for the mask are fixed according to Pepe et al. (2002a) recommendations by the relative depth of the stellar lines computed with the same parabola fitting. The CCF is fitted with a Gaussian and the sigma width ($\sigma = 2\ln(2) \cdot \mathrm{FWHM}$) extracted. A list of bands can be given as parameters ***mask_telluric*** in order to exclude some regions of the CCF mask. A mask can also be given directly as input in the ***CCF_mask*** parameter, even if we highly recommend not to select this option. If such option is selected, the user also has to specify the RV stellar system of the star in order to match the rest frame of the star and of the binary mask.

Once the FWHM is known, the first step consists in smoothing the spectra. If "auto" is enabled for the ***par_smoothing_box***, a high frequency filter is applied in Fourier space to suppress the frequencies above the sigma width. Two filters are available: an error function and a top-hat with an exponentially decreasing tail. Their expressions are given by

$$\begin{aligned} f_{erf}(\omega) &= C\ \mathrm{erf}\left(\frac{\omega_0 - |\omega|}{\delta\omega}\right) \\ f_{hat}(\omega) &= \begin{cases} 1 & |\omega| \leqslant \omega_0 \\ e^{-\frac{|\omega|-\omega_0}{\delta}} & |\omega| > \omega_0 \end{cases} \end{aligned}, \tag{9}$$

where C is a normalisation constant (the filters should all satisfy $f(0) = 1$), ω_0 is the center of the filter which can be understood as the cut-off from which high-frequency modes are suppressed and $\delta\omega$ is the width of the filter (a small value of $\delta\omega$ will produce a sharp filter, while a high value will produce a smooth one).

By default, the error function filter is used if "auto" is used for ***par_smoothing_kernel***. Some care is then needed to get the correct values of ω_0 and $\delta\omega$. As mentioned before, a typical wavelength scale is given by the σ-width of the CCF. This implies the existence of a given frequency scale σ^{-1}. Hence, we can parametrise $\omega_0 = \alpha_1\sigma^{-1}$ and $\delta\omega = \alpha_2\sigma^{-1}$, where the α_i's are two dimensionless parameters. To adjust them to the best values, a calibration curve was constructed. A high SNR (~ 10000) spectra of α Cen B and of the Sun were formed by stacking several observations. We took care of suppressing long RV term in both cases in order to allow the stacking.

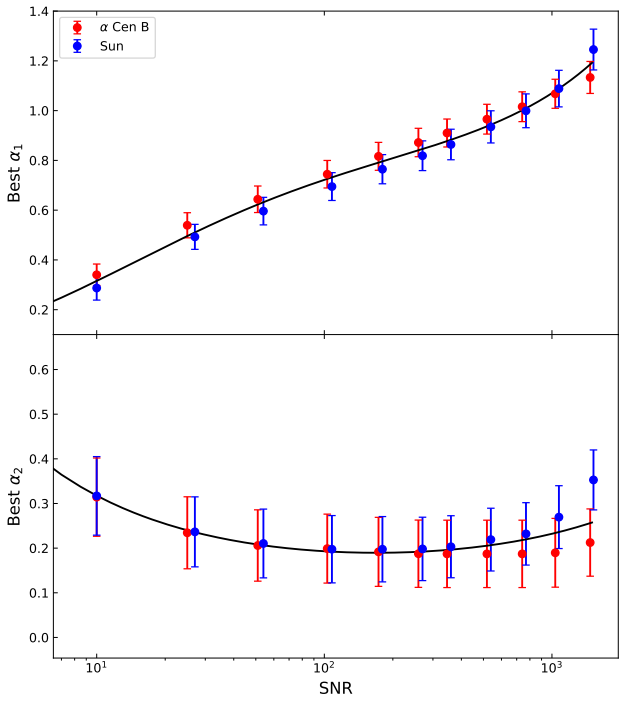

Fig. 12. Calibration curves of the α_1 and α_2 parameters depending on the SNR value of a spectrum for the erf Fourier filter. **Top:** Calibration of α_1. The dots represent the best value according to the simulations (see Fig.11) which produce the least standard deviation (weighted, as explained in the text) between the noisy-free spectrum and the smoothed one. The error bars correspond to the 5% best values of the simulations. Two stars were used for the calibration (highlighted by the color of the points). The calibration curves were fitted by a third degree polynomial function in log-log space. **Bottom:** Calibration of α_2 with the same procedure. The calibration curves were fitted by a second degree polynomial function in log-log space.

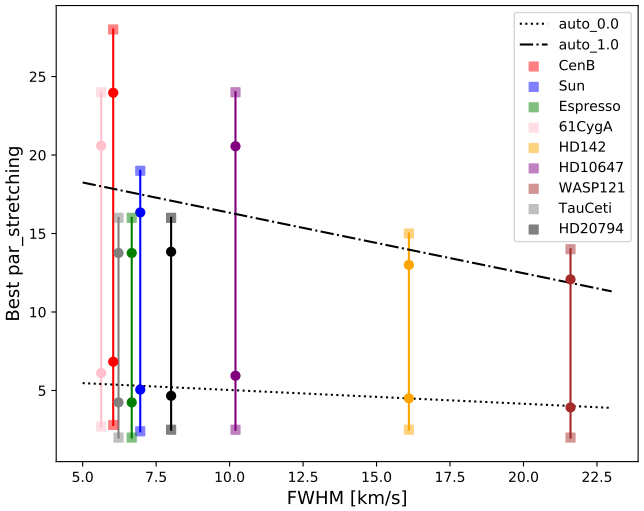

Fig. 13. Calibration curves of the ***par_ stretching*** parameter depending on the CCF FWHM value of a spectrum. The calibration has been established using 9 stars (see Table 3) probing the spectral type range from F6 to K5 (each star is indicated with a different color). The square dots indicate the minimum and the maximum values for the ***par_ stretching*** parameter (chosen by eye). The circles dots are more "conservative" (16% and 84% of the parameter range). We fit the latter on a line to get the minimum value in automatic mode (***par_ stretching*** = 'auto_0.0', dotted) and the maximum of the parameter (***par_ stretching*** = 'auto_1.0', dash-dotted) as a function of the FWHM. The final choice is then let at the user discretion. Note the near constant line for the minimal value, and the decreasing line for the maximal value.

We then simulated different SNR spectra by adding several levels of Poisonian noise in the spectrum (see Fig.11). The Fourier filter was performed in an optimisation grid of α_1 and α_2. In order to determine the best values of the α's parameters, we proceed as follows. First, we note that, even if we do not want the smoothing to affect the continuum, it is not a problem if the former modified the absorption lines (for example by changing their depth). Hence, it should not matter much if the smoothed spectrum is different from the original one in the lines. To quantify this, we introduce a coefficient function Γ, whose value is 1 if the noisy-free spectrum is equals to the continuum, and decreases when the spectrum goes further away from the continuum. The norm in the (α_1, α_2) parameter space is then given by the standard deviation between the smoothed spectrum and the noisy-free one, weighted by this function Γ. By minimizing this quantity, we find the best values for α_1 and α_2 for each SNR value, which we can extrapolate for every value of the SNR (see Fig.12). If the automatic mode for the spectrum smoothing is chosen, the flux units of the input file has to be in photons units such that the SNR value can be extracted by taking simply the square root of the flux.

Once the spectrum has been smoothed, we need to normalise it. To do so, the code will fix the ***par_ stretching*** parameter. Recall that it is used to scale the y-axis with respect to the x-axis, and to some extend, this parameter corresponds to a "tension" applied on the veil. If the stretching parameter is too small, there is too much tension on the veil and the rolling pin can go through the spectrum (recall the rolling pin considers only the local maxima, and there is no spectrum from its "point of view"). Similar issues can occur if the maximum radius for the rolling pin is too small. On the opposite, if the parameter is too large, the tension is not strong enough and the veil starts to fall inside the lines of the spectrum.

There is a priori no precise way to determine the best value for this parameter. The calibration is presented in Fig.13 and was performed by eye with high SNR spectra of 9 stars listed in Table 3. We need a typical length scale in order to calibrate the parameter of the stretching. The only such quantity is again the FWHM of the CCF. Furthermore, as raised previously, there is a large range of values working quite well to normalise the spectra, depending on the amount of tension we wish. Hence, the calibration provides, for a given value of the FWHM, a range over which the parameter can be taken. We fit two lines, one to get the minimum value in automatic mode (***par_ stretching*** = 'auto_0.0') and one for the maximum value of the parameter (***par_ stretching*** = 'auto_1.0'). The user has thus to specify after the auto keyword the level of tension in its continuum by also entering a number between 0 and 1. Note that 0 is a strong tension whereas 1 is a weak tension.

By observing Fig.13, one can notice interesting features. The minimal value of the parameter seems to be rather constant for all the stars, regardless of the value of the FWHM. This is the case because the flux of all the star are normalised in a same way, by scaling x and y axis on a same

Table 3. Table of the stars and of the master spectra used to calibrate the ***par_ stretching*** parameters. The stars were chosen to probe different instrumental resolutions and spectral types ranging respectively from $R \sim 55'000$ to $140'000$ and from F6V to K5V. The table contains from left to right: the name of the star, its spectral type, the number of exposures stacked to form the master spectrum, the SNR of the master spectrum at 5500 Å, the instrument which observed the star and the average date of the observations. At the end are displayed three parameters determined in automatic mode by RASSINE: the FHWM of the CCF in $\mathrm{km\,s^{-1}}$, the ***par_ stretching*** with intermediate tension (***par_ stretching*** = 'auto_0.5') and the ***par_ Rmax*** value in Å.

Name	Spec. type	#	SNR	Instr.	Date	FWHM	***par_ stretching***	***par_ Rmax***
WASP-121	F6V	140	406	HARPS	2018/01	21.6	7.6	138
HD142	F7V	40	595	HARPS	2004/10	16.1	9.2	50
HD10647	F9V	56	784	HARPS	2004/10	10.2	10.8	86
Sun	G2V	711	8950	HARPN	2015/08	6.96	11.7	78
HD20794	G6V	28	968	CORALIE	2016/12	8.01	11.7	76
τ Ceti	G8V	101	3036	HARPS	2017/07	6.22	11.9	76
α Cen B	K1V	1767	12244	HARPS	2010/03	6.07	11.9	72
α Hor	K2III	1	1081	ESPRESSO	X	6.67	11.8	72
61 Cyg A	K5V	129	3494	HARPS	2013/03	5.63	12.1	65

length. Recall that if the value of the parameter is too small (high tension on the veil), the rolling pin will go through the spectrum by reaching unsuitable local maxima (corresponding to blended lines). Those lines are present for most of the stars, either because of high stellar lines density or high rotational broadening. Hence a minimal value of 2 was found to be the same lower limit for all the stars. Regarding the upper value for the parameter, it decreases when the FWHM increases. If the value of the parameter is too high (small tension on the veil), the code will fall into absorption lines. Hence, if the value of the FWHM is already high, there is not much room to stretch the horizontal axis before having lines that are broad enough to make the rolling pin fall. Hence, the value of the parameter has to be smaller as the FWHM gets bigger.

The last two automatic parameters are the minimum and maximum radius of the rolling pin, R and R_{max}. The former is fixed as 50 times the σ value of the CCF (5 times the 5σ-width) converted back in Å for the bluest part of the spectrum, which prevents the rolling pin from falling inside the stellar lines. The maximum radius is computed with the same pre-continuum used before the penalty. Recall that two continua were obtained, with a small and a large window by a rolling maxima which are used for the penalty (see section 2.2.4). The small window continuum will fall inside broad absorption lines, whereas the large window continuum will be relatively insensitive to them and therefore will be on the top of the small window one at the lines locations. Outside broad lines, none of the two continua will be systematically on the top. We thus looked for the sign of the difference between both continua before performing a clustering algorithm computing the length on which the continuum difference keeps the same sign. By keeping only the clusters which are situated in a penalty zone sufficiently high, namely the zones where the big window curve is significantly bigger than the small window one, the maximum radius is fixed as the length of the longest selected cluster. By sufficiently high we mean than the algorithm proceeds as follows: the initial threshold is fixed at a penalty of 0.75. If no cluster is found, the algorithm decrease iteratively the threshold by 0.05 units until a cluster is found. In all our simulations, the cluster defining the ***par_ Rmax*** was always situated on the CaII H&K lines.

B. Graphical interfaces of RASSINE

C. Collections of RASSINE reduction in automatic mode

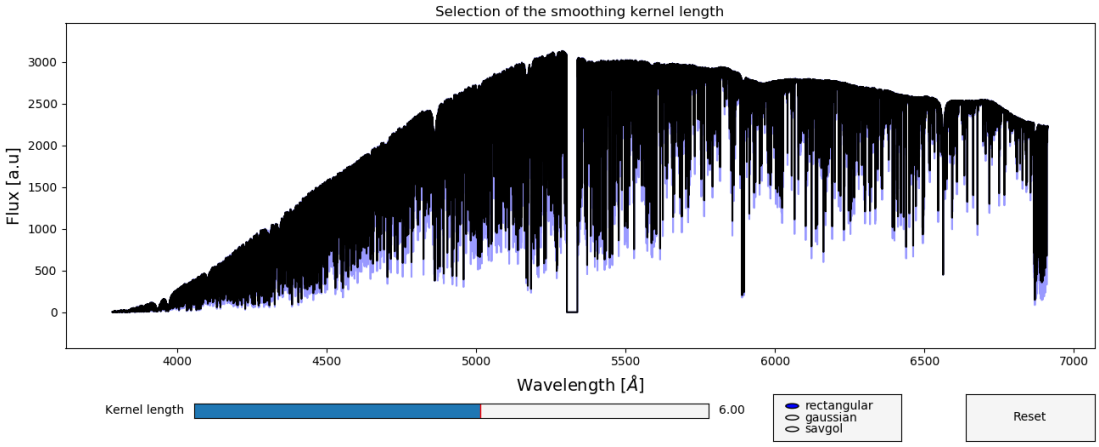

Fig. 14. First graphical interface of RASSINE to select the smoothing kernel. The user has only to select the smoothing length and kernel shape.

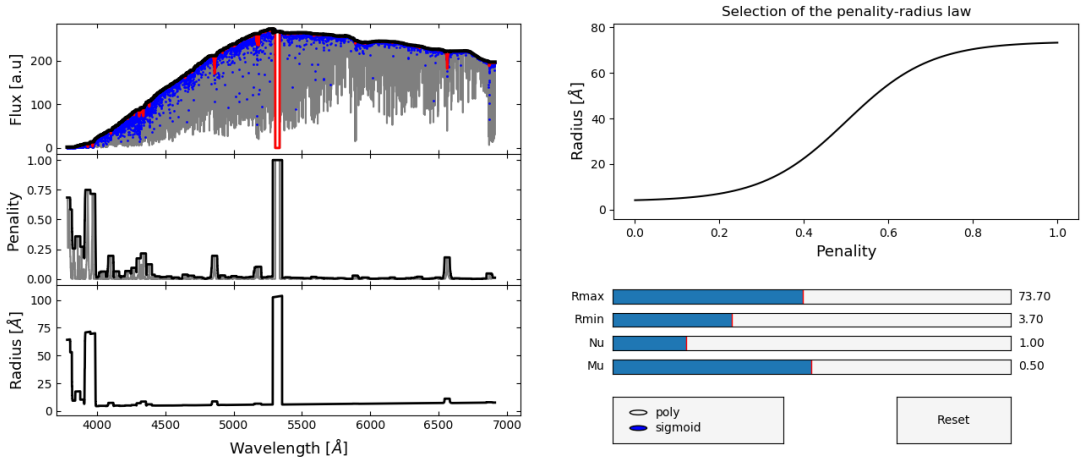

Fig. 15. Second graphical interface of RASSINE to select the penalty law. The user has to select the functional form of the penalty law (polynomial or sigmoid) as well as the minimum and maximum values.

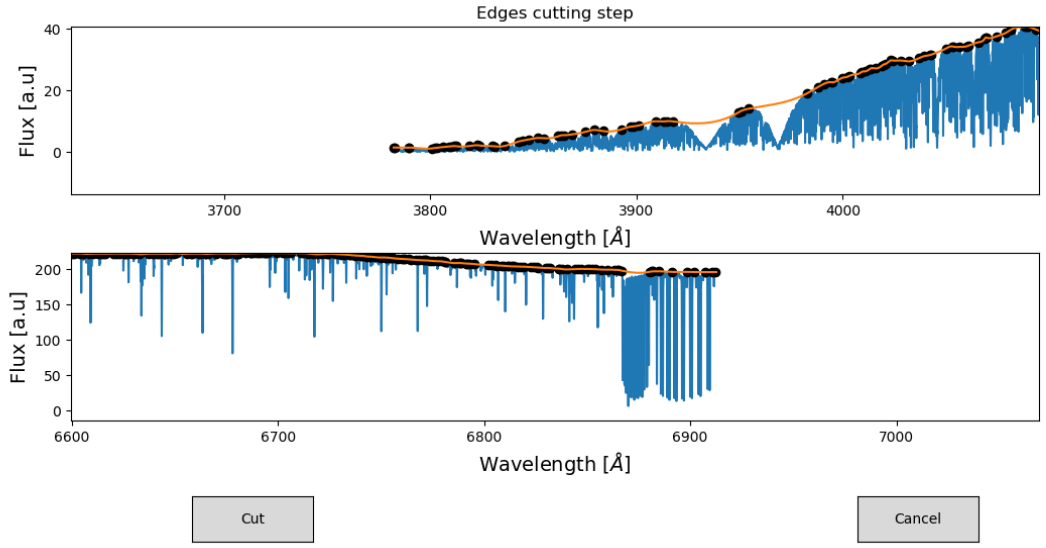

Fig. 16. Third graphical interface of RASSINE to cut the borders of the continuum. The user has to click on the "cut" button until to be visually satisfied.

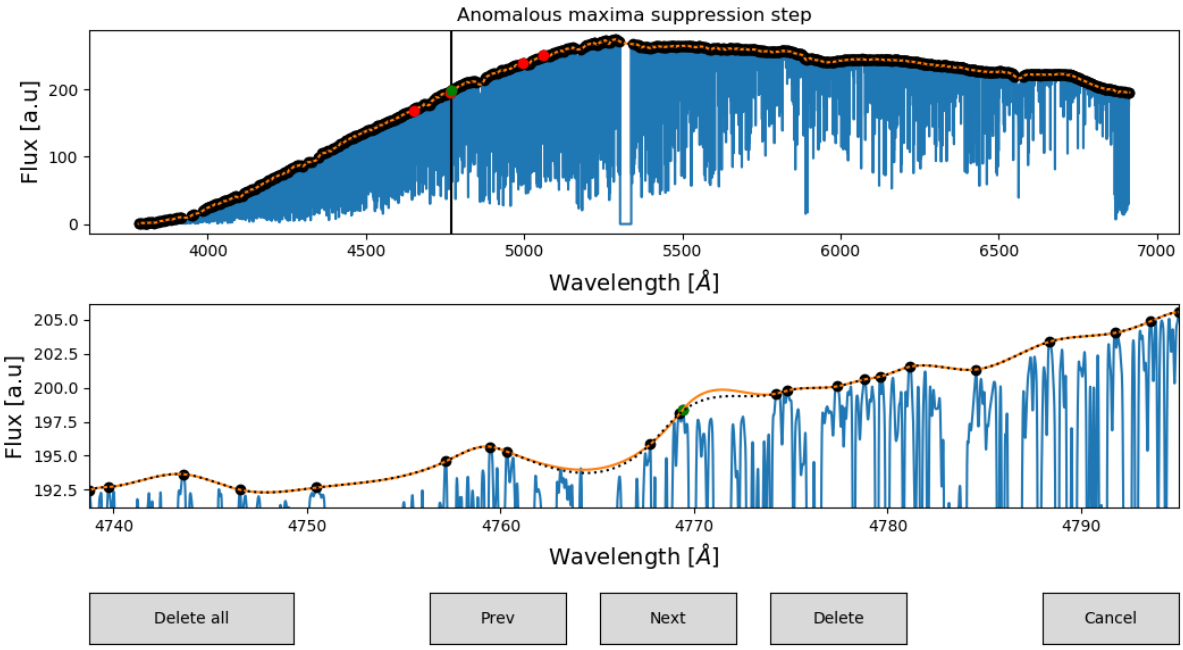

Fig. 17. Fourth graphical interface of RASSINE to suppress doubtful local maxima. The user can navigate through 5 doubtful at the same time.

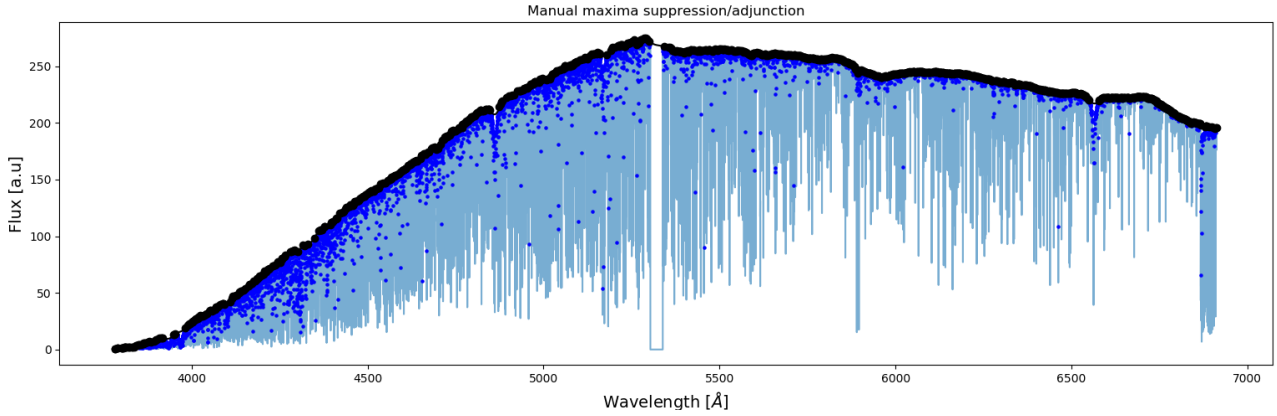

Fig. 18. Fifth graphical interface of RASSINE to supress or add manually some local maxima by clicking on the points.

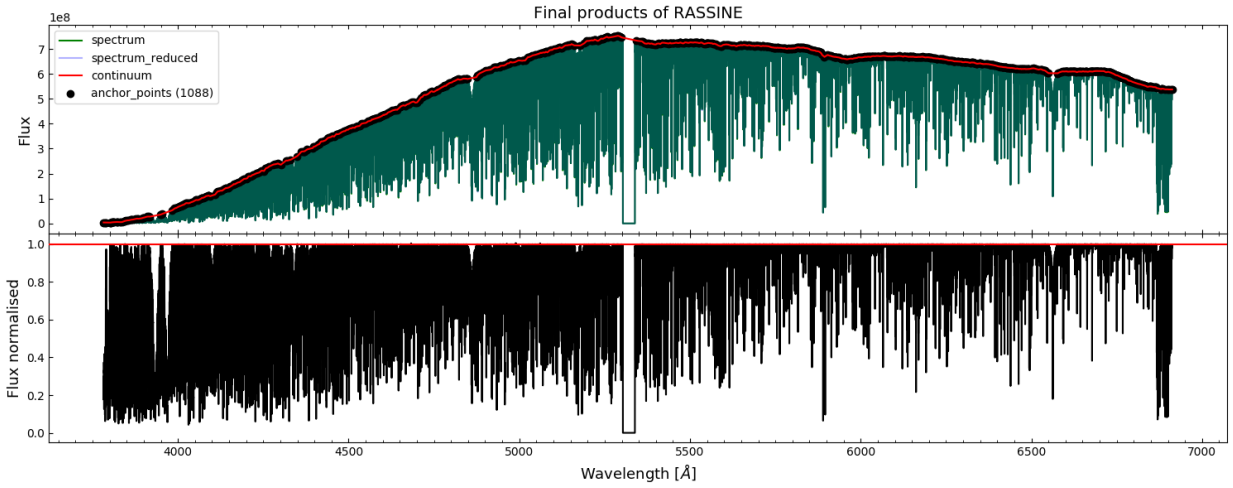

Fig. 19. Last graphical interface of RASSINE presenting the final output.

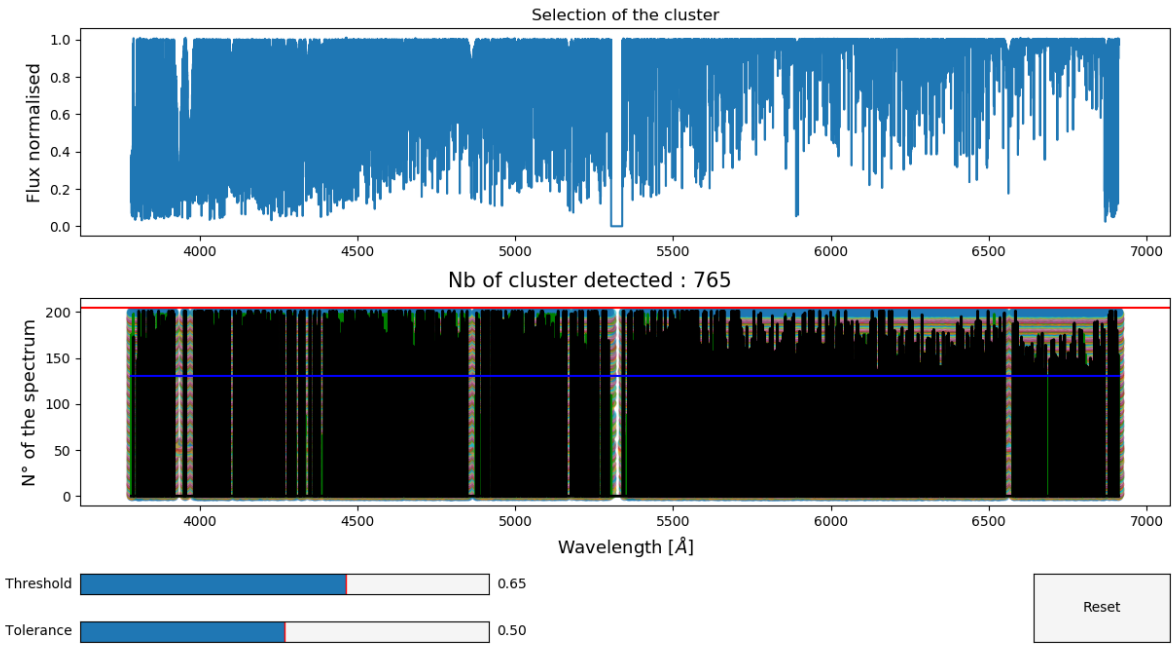

Fig. 20. First optional function of RASSINE (*intersect_ all_ continuum*) to reduce spectra timeseries and stabilize the derived continua. The anchors points of all the points are plot

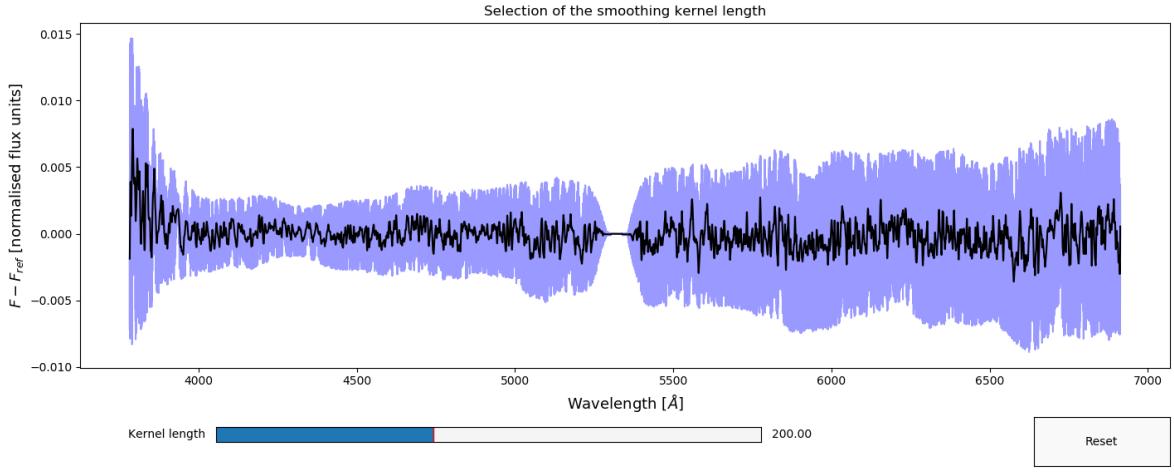

Fig. 21. Second optional function of RASSINE (*matching_ diff_ continuum*) to reduce spectra timeseries and stabilize the derived continua by applying a Savitzky–Golay filter on the spectra difference with a spectrum of reference.

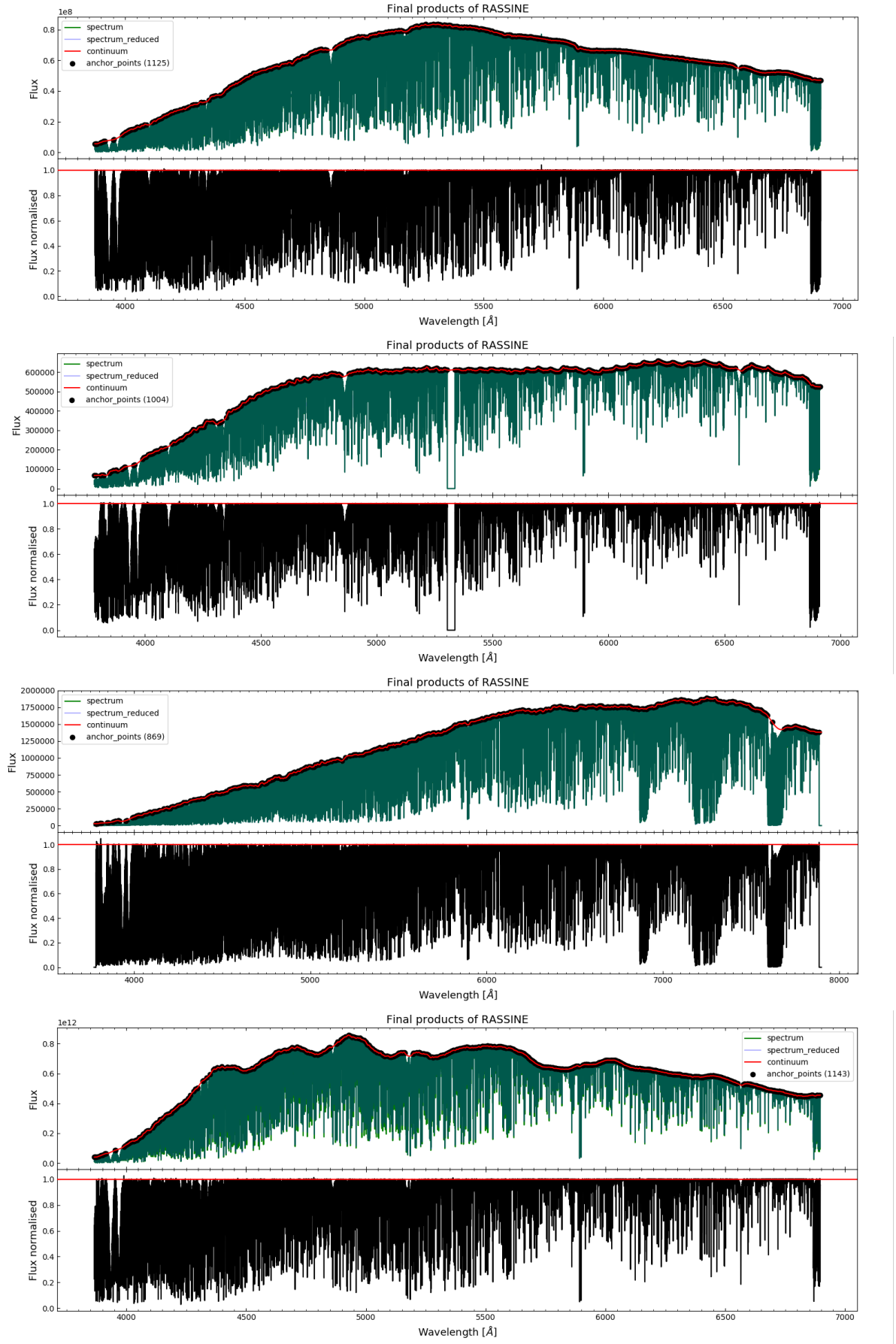

Fig. 22. A collection of spectra normalised by RASSINE in automatic mode, $\nu = 1$, intermediate tension (***par_stretching*** = 'auto_0.5') and Savitzky-Golay filtering with ***par_smoothing_box*** = 6. Each spectra is coming from a different instrument, from top to bottom : HARPN, HARPS, ESPRESSO and CORALIE.