# Bachelor Project in Physics

Jakob Harteg, wmc573

June 2022, University of Copenhagen

**todo: Make cover page**

**todo: Add abstract**

# Contents



# Note about the project

In the beginning of February 2022, Troels Petersen (KU) and Lars Buchhave (DTU) started working together to improve computational LFC calibration methods, and Troels invited me and a few other students to come along for the ride. During the project I've been exploring real data and developing a crude method for performing LFC calibrations and subsequent radial velocity extractions. With this project report it is my intent to share and explain the process of performing such analyses, thinking of the reader as a fellow physics student perhaps interested in picking up where I left off.

# 1 Introduction

**todo now: Expand**: and add illustration of the RV method.

The radial velocity method was used to discover the first exoplanets and continues to be one of the main methods for the discovery and characterization exoplanets [5]. With new extreme-precision radial velocity (EPRV) spectrographs such as the EXtreme PREcision Spectrograph (EXPRES), data from which this project is based on, we are slowly approaching the precision necessary for the discovery of Earth-sized planets around Sun-like stars. For this to succeed, it is however necessary to understand and mitigate many effects, such as the movement of the Earth relative to the center of mass of the solar system (barycentric corrections), light scattering in the Earth's atmosphere (tellurics) and light scattering inside the spectrograph (blaze). A general wavelength calibration of the spectrograph is also needed, the quality of which of course directly influences the precision of the final radial velocities that can be obtained. For that purpose, EXPRES utilizes a Laser Frequency Comb (LFC), is a rather new technique. While to extract radial velocities (RV) we measure the apparent wavelength shift between observations over a period of time, preferably more than a year, and utilize the well-known doppler effect.

The full procedure from raw data to results, also referred to as the *pipeline* in the literature, is extensive and complex, which is why I've ignored many aspects of it during this project. I've worked directly on the LFC calibration and RV extractions and tried to orient myself about the most important corrections, thus that is what I will describe in this report. The full pipeline is described in detail in [6].

# 2 Theory

todo now: Go through

## 2.1 Radial velocity method for exoplanet detection

The radial velocity method is one of the few current methods of detecting exoplanets. Two celestial bodies in orbit around each other, such as a star and a planet, orbit their common center of mass (barycenter). This means that the star, although typically much more massive than the planet, is also in movement relative to an outside observer. The larger the planet is, compared to the star, the faster the star will appear to be moving and we can measure this movement through the doppler effect: the electromagnetic spectrum of the star observed on Earth will be blue shifted when the star is moving toward us and red shifted when moving away. If there is a planet around a star, we should observe a periodic doppler shift. The method is illustrated in figure 1.

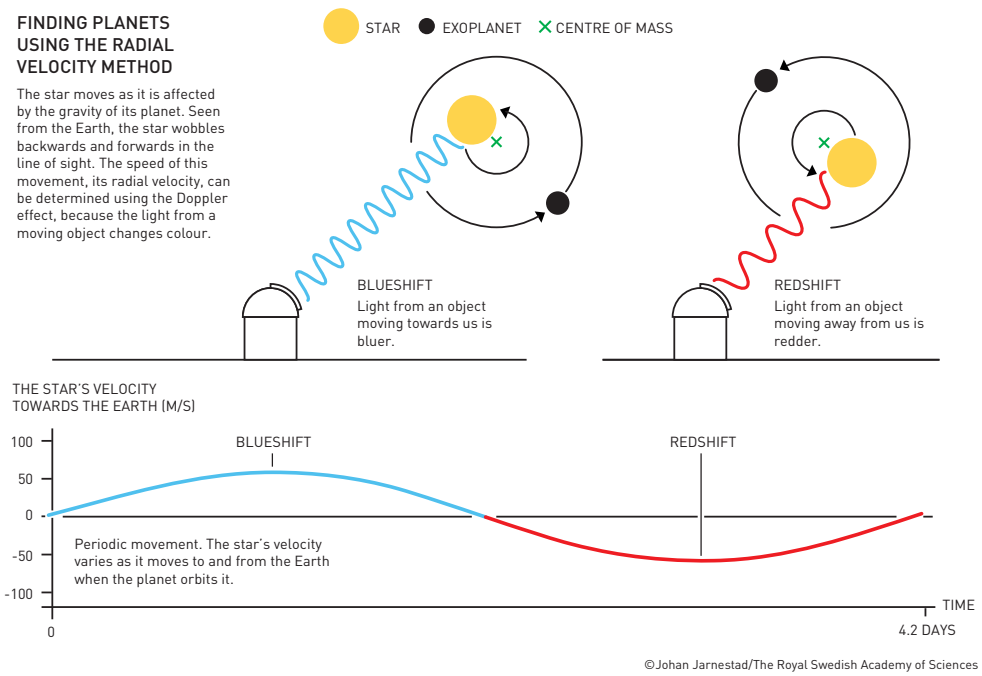


**Figure 1:** Illustration of the radial velocity method for exoplanet detection. © Johan Jarnestad/The Royal Swedish Academy of Sciences

A large planet like Jupiter induces a radial velocity (RV) in the Sun of about 12.7 m/s when observed in its plane of orbit. While a small one like Earth only induces an RV of about 9 cm/s. (p. 29, [5]).

**todo: Possibly describe the RV calculation in more detail and compute Earth K.**

### 2.1.1 Doppler shift

The radial velocity method relies on the well-known Doppler effect. Ignoring terms of $c^-4$ and higher, the general shift caused by a relative displacement betweeen the source and an observer at zero gravitatonal potential is given by

$$\lambda = \lambda_0 \frac{1 + \frac{1}{c}\mathbf{k}\cdot\mathbf{v}}{1 - \frac{\Phi}{c^2} - \frac{v^2}{2c^2}}, \tag{1}$$

which accounts for both special relativistic effects and gravitatonal doppler shift described by general relativity. Where $\lambda$ is the observed wavelength, $\lambda_0$ is the emitted wavelength, $\Phi$ is the Newtonian gravitational potential at the source ($\Phi = GM/r$ at a distance $r$ of a spherically symmetric mass $M$ ), $\mathbf{k}$ is the unit vector pointing from the observer to the source, $\mathbf{v}$ is the velocity of the source relative to the observer and $c$ is the speed of light [4].

Special relativistic effects we can safely ignore, as we are dealing with velocity shifts on the order of meters or centimeters per second, and thereby cross out the third term in

the denominator. The remaining terms $(1 - \Phi/c^2)$ evaluted for HD 34411 ($M = (1.08 \pm 0.14)M_\odot$, $R = (1.28 \pm 0.04)R_\odot$ [2]) is around 0.999998. Adding this term to the computation of RV does not change my results, so we can neglect the denominator completely. **todo: How to check analytically though?**

If the unit vector **k** were pointing directly toward us, it would mean that we were observing the system in the plane of orbit. This is however unlikely. Since we don't know the inclication angle, a possible simplification is to omit **k** is treat the resulting **v** as a minimum radial velocity.

Thus we are left with

$$\lambda = \lambda_0 \times \left(1 + \frac{v}{c}\right), \tag{2}$$

which is to say that the observed wavelength is simply the emitted wavelength scaled by a factor $(1 + v/c)$. This formula allows us to compute the minimum relative velocity shift, $v$, between two observations, $\lambda$ and $\lambda_0$.

## 2.2 Description of the instrument

The EXtreme PREcision Spectrograph or EXPRES is an extreme-precision spectrograph situated at the Lowell Observatory's 4.3m Lowell Discovery Telescope (LDT) near Flagstaff, Arizona, USA. The LDT allows for up to 280 partial nights of observation per year.

Like in many spectrographs, at the heart of EXPRES is a Charge Coupled Device (CCD). A CCD is a silicon-based multi-channel photon detector consisting of a large number of small light-sensitive areas called pixels. The CCD is EXPRES an STA1600LN CCD backside illuminated image sensor with a $10,560 \times 10,560$ array containing 9µm×9µm pixels, designed to with a wavelength range of 3800−7800Å. When a photon hits a pixel it is converted into a charge, and each pixel can thus supply independent measurements. Since a one dimensional sensor would be impractical, EXPRES is constructed in such a way, that it wrap the spectrum inside the CDD, meaning that the spectrum starts in the top row of the sensor, and continues in the second row. Short wavelengths are thus to be found in the top of the CCD and long wavelengths at the bottom. EXPRES is housed in a vacuum enclosure to minimize changes in temperature and pressure, which can otherwise cause the spectra to change position on the CCD and thus lead to errors in the RV mesaurements.

**Calibration device:** Wavelength calibrations are performed with the use of a Laser Frequency Comb (LFC), produced by Menlo Systems, which is a laser source whose spectrum consists of a series of discrete, equally spaced frequency lines. The LFC however also needs calibration for which a Thorium Argon (ThAr) lamp with known frequencies is used.

**Spectral resolution:** EXPRES has a spectral resolution of $R = 150'000$, where R is defined as $R = \lambda/\Delta\lambda$. Inverting this we get what's called the resolution element of the instrumental spread Function (or line spread function), $\Delta\lambda = \lambda/R$. For a wavelength of say of $\lambda = 5000$Å, this comes out to $\Delta\lambda = 5000\text{Å}/150,000 \approx 0.033$Å, and it describes the "blurring" of monochromatic beams on the detector. Absorption features narrower than $\Delta\lambda$ can, in a well-behaved spectrograph, be approximated as a normalized, symmetric Gausssian function with FWHM = $\Delta\lambda$. Specifically for EXPRES, a super-gauss is however a better fit[7]. LFC lines being monochromatic and thus very narrow will appear on the detector as a super-gaussian with $\Delta\lambda$ ranging from 3.9-5 pixels across the detector. By fitting, the center of the peak can be identified to a fraction of a pixel. The LFC lines are seperated by about 10 pixels to remain distinct after this blurring. For star-spectra however some emission and absorption lines will be too close together and will appear "blended" on the detector.

**Barycentric correction:** Barycentric corrections are derived from the EXPRES exposure-meter, which is essentially a smaller, less precise spectrograph. Described in detail in [1]. EXPRES as a whole is described in technical detail in [3].

## 2.3 Description of the data

EXPRES data are meant to serve as an example of the data being produced by next-generation spectrographs.

The data used in this project was supplied by Lily Zhao and is by no means raw data, but data that has already gone through a lot of processing.

For development of RV extraction method, observations from four stars were used:

- HD 101501 (45 observations, 22 nights, Feb. 10, 2019 - Nov. 26, 2020)
- HD 26965 (114 observations, 37 nights, Aug. 20, 2019 - Nov. 27, 2020)
- HD 10700 (174 observations, 34 nights, Aug. 15, 2019 - Nov. 27, 2020)
- HD 34411 (188 observations, 58 nights, Oct. 08, 2019 - Nov. 27, 2020)

Most days have 3-4 observations, and there are significant gaps in the data as well. LFC exposure files were provided by Lars A. Buchhave.

### 2.3.1 Data structure

The data used in this project consists of already packaged FITS (Flexible Image Transport System) files, which is a portable file standard widely used in the astronomy community to store images and tables. There is a FITS file for each observation, containing a variety of mesaurements for each pixel on the CCD. The rows of the CCD data are referred to as orders. There are 86 orders each of which has values from 7920 pixels. Drawing a coordinate system on the CCD, we are thus moving through pixels as we go along the x-axis and through orders as we go along the y-axis.

This would give the CCD the very elonganted dimensions of $86 \times 7920$, but as mentioned earlier, the CCD is actually square. The orders however hit the CCD at an angle and for this reason *order tracing* is necessary. Order tracing reduces each order from 2d array to a 1d array, which means the final image comes out much shorter in the vertical/order dimension. Described in detail in section 3.2.1 of [6].

Furthermore, the CCD is not equally sensitive everywhere, and there are areas along the edges that are deemed useless. The data comes with a mask which shows which pixels are should be used.

### 2.3.2 Noise and corrections

Photon noise and read noise are the two largest contributors to the noise on a given pixel on the EXPRES CCD. These two quantities are measured and summed in quadrature for each pixel. Photon noise is assumed to be poisson distributed and the standard deviation is then the square root of photon counts. Read noise is calculated emperically and is assumed to be consistent throughout each night of observation. [6].

Although manufactures have tried their best to limit it, the CCD still gets hits by scattering light, being the strongest in the center of the detector. This has been modeled and subtracted from the spectrum by measureing the photon count in between orders. The blaze function is available in the data file and allows for recovering the original counts for each pixel by multiplying the blaze with the spectrum.

Tellurics in the context of spectrographs refers to the contamination that ground based spectrographs must cope with, which occurs as the light passes through our atmosphere encountering molecules such as oxygen and water vapor on the way. The I use is already corrected for this, with a technique called SELENITE[7].

The barycentric correction is vital, as this is what removes the movement of the Earth around the center-of-mass of the solar system from the data. How it is done exactly I've done delved into during this project.

At the end, the largest source of error comes from stellar activity. Variations in the light from the star in quesiton due to various physical processes. Dark spots, granulation, "starquakes", rotation are a few examples.

# 3 Data Analysis

In this section I will explain crude methods for performing LFC calibrations and radial velocity extractions. Although planned, I did not end up having access to star spectra with associated LFC exposures, such that I could calibrate the star spectra myself before computing the radial velocities. As a result, part one of the data analysis does not directly carry into the second, although, with the right data, it could.

## 3.1 Calibration

The calibration is needed to map each pixel on the CCD to a specific wavelength. Such a map is referred to as a wavelength solution. To do this, we first shine a light source with known frequencies, preferably many discrete lines, onto the CCD. The exact location on the CCD where each line appeared can then be mapped to the known wavelength of the given line. EXPRES uses a Thorium Argon (ThAr) lamp for an initial rough wavelength solution, which has 4,000 lines across 82 orders. These values have been linearly interpolated to provide a solution for all pixels across the CCD and were included in the data file that I was provided with for this project.

For a better wavelength solution EXPRES uses a laser frequency comb (LFC), which generates about 20,000 lines across 50 orders, with frequencies given by the relation

$$\nu_n = \nu_{\text{rep}} \times n + \nu_{\text{offset}} \tag{3}$$

for integers $n$. The repetition rate $\nu_{\text{rep}}$ and offset frequency $\nu_{\text{offset}}$ are referenced against a GPS-disciplined quartz oscillator, providing calibration stability corresponding to a fractional uncertainty of less than $8 \times 10^{-12}$ for integration times greater than 1s. [6]. The values I have used, $\nu_{\text{rep}} = 14e9$ and $\nu_{\text{offset}} = 6.19e9$, were provided by Lars A. Buchhave.

The following steps are followed to perform the LFC calibration: 1) determine the CCD pixel position of each LFC line, 2) using ThAr wavelength solution find the corresponding mode $n$ in equation 3 to know the true wavelength of the line, and finally 3) find way to estimate a wavelength solution in between LFC lines.

Figure 2 is an illustration (not real data) of this process. The left panel shows LFC lines appearing in different locations on the CCD (with exaggerated variations and errors for clarity) and the corresponding wavelength solution given by equation 3. The middle and right panel illustrate two ways of modelling a wavelength solution in between LFC lines, polynomial fitting and cubic-spline interpolation respectively. With such a wavelength solution, you have a direct map between any location on the CCD (x-axis) and a wavelength solution (y-axis). This is to be performed in every order of the CCD.

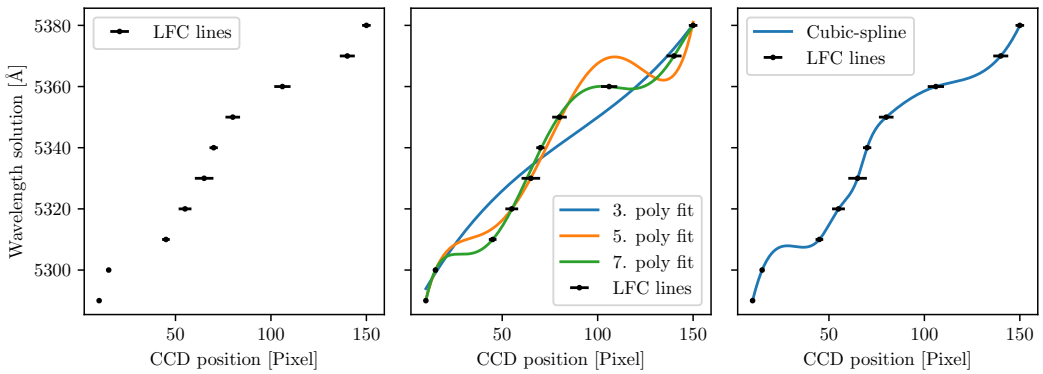


**Figure 2:** Illustration of the LFC calibration process (not real data). See text for explanation.

### 3.1.1 Determining LFC line locations

See figure 3 right side for a plot of the intensities measured across the CCD.

The right side of figure 3 shows an LFC exposure plotted in its entirety. It's clear that the LFC does not cover all orders of the CCD, but starts shortly before order 40 and stops around 76.

To determine the location of the LFC lines I follow this procedure: 1) Find peaks using scipy peak finder, 2) make data slices around each peak with the size of the average distance

between peaks, 3) using iminuit do a chi2 minimisation fit to each peak with a super-gauss plus a linear background. See figure 3 left side.

A super-gauss, defined in equation 4, is a regular gaussian but with an extra parameter, here denoted $P$, that allows the top of the gaussian to be flattened. The last two terms here add a linear background and an offset.

$$f(x; A, B, C, P, \mu, \sigma) = A \exp\left(-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)^P\right) + B(x-\mu) + C \tag{4}$$

The fit is then a standard chi2 minimization:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x; A, B, C, P, \mu, \sigma)}{\sigma_i}\right]^2 \tag{5}$$

Where $N$ is the number of data points, $x$ is pixel-space, $y_i$ and $\sigma_i$ is the measured intensity and uncertainty respectively. The fit returns the values and uncertainties for the parameters $A, B, C, P, \mu, \sigma$ when the chi2 is minimized.

We are mainly interested in $\mu$, which gives the position of the LFC peak on the CCD (in pixel-space). With the intial rough wavelength solution derived from the ThAr lamp I can determine what the approximate wavelength of the LFC peak should be. To find the better wavelength solution I then go look up the closest frequency given by eq. 3. And we now have a map of 20,000 points on the CCD with a good wavelength solution. We would however like to have a wavelength solution for all pixels on the CCD, so we need to find a way to estimate the wavelength solution in between LFC lines. To do that I have explored two approaches: polynomial fitting and cubic-spline interpolation. Before moving on though, it is worthwhile to study the chi2 values from the many fits we just performed, as they can help up evaluate whether the errors of the original data are correct.

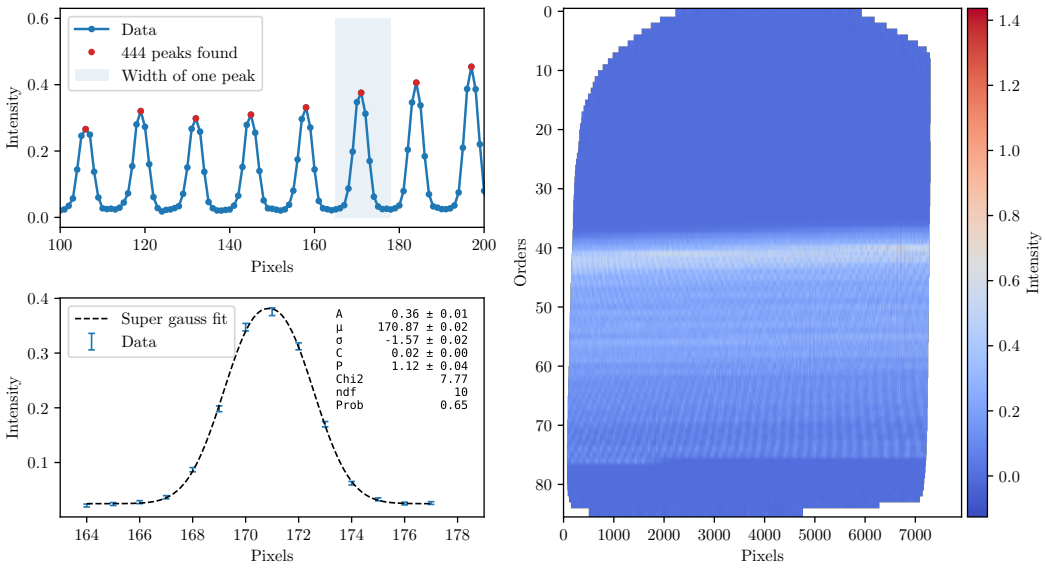


**Figure 3:** Right: Measured intensities for the LFC across the CCD (unitless). Upper left: illustration of a few LFC lines in order 65. Peaks are identified with scipy peak finder. Lower left: each peak is fitted with a super gauss to find the exact top of the peak with uncertainties.

### 3.1.2 Errors in the calibration data

According to Lily Zhao et al, the line-spread-function of EXPRES can best be represented by a super-gaussian [7], and so as we fit a super-gaussian to the peaks, the chi2 will grow if the measured values do not agree with the predicted values within the uncertainties. For a given data point, if the measured value differs by exactly the uncertainty, that data point would contribute 1 to the sum. However, fitting with the super-gaussian (including a linear background) we also have 6 fitting parameters that allow some wiggle room for the model to fit the data, which we must compensate for. For a given data series, such as an LFC peak, we can expect the chi2 roughly equal the number of data points in the sum *minus* the number of parameters in the fit, i.e. the number of the degrees of freedom:

$$N_{\text{dof}} = N_{\text{data-points}} - N_{\text{fit-parameters}} = 13 - 6 = 7, \tag{6}$$

as I use roughly 13 data points in each fit (the LFC line spacing does vary a bit across the CCD).

So if we plot all chi2 values we get from fitting the LFC lines, as I've done in figure 4, we should see a typical chi2 distribution with a peak around 7. Using the uncertainties as provided in the data file, the chi2 distribution is however very flat (red curve). It peaks somewhere around 25, which suggests that the uncertainties are about a factor $\sqrt{3}$ too small (square-root because the chi2 of course is squared and $25/3 \sim 8$), and scalling up the uncertainties by $\sqrt{3}$ and fitting the peaks again does in fact give a chi2-distribution with a peak roughly around 8. additionally we can also look at the p-value distribution. With proper uncertainties, we should expect neither too many awful fits (p=0) nor too many perfect fits (p=1), but rather a roughly flat distribution. Although the chi2-distribution check is more clear-cut, the p-value distribution for uncertainties scaled by $\sqrt{3}$ (green) is better. Scaling by $\sqrt{6}$ is too much: the chi2 peaks around 4 and the p-dstribution shows many more perfect fits.

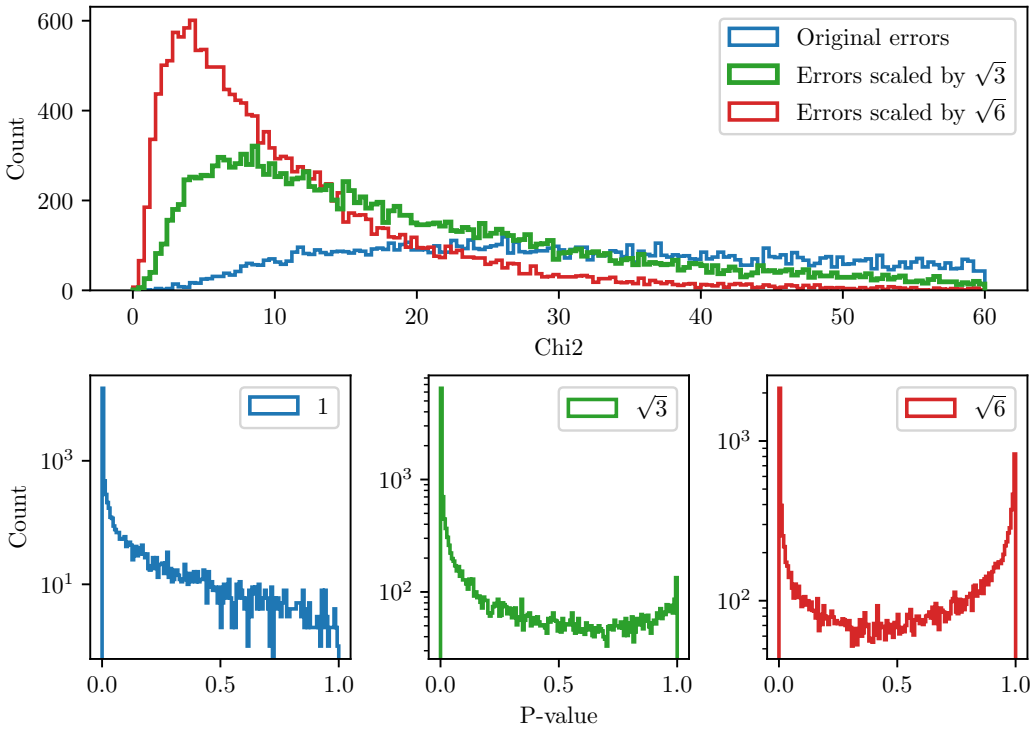


**Figure 4:** Chi2-values and p-values from individual LFC peak super-gauss fits with photon count (spectrum) errors multiplied by different scale-factors (1, $\sqrt{3}$ and $\sqrt{10}$). See text for more details.

### 3.1.3 Poly-fit calibration

Since the LFC peak positions, as seen in the right plot in figure 3, appear to exhibit a certain periodic behavour, my initial approach to compute a wavelength solution across the whole CCD was to fit the LFC peak positions with a polynomial. Looking at the residuals of fitting the LFC line positions with polynomials of increasing degree revealed smaller and smaller periodic variations, until reaching 5th degree, see figure 5. The p-value drops to 0 for 6th degree, and the per line rms (given by eqation 7) explodes. Only LFC lines with a chi2 (from the super-gauss fit) smaller than 100 was used for the analysis and errors were scaled by $\sqrt{3}$.

The residuals plotted show the difference between the theoretical wavelength for each LFC line (equation 3) and the wavelength that the calibration predicts for that location. You can get a feel for the quality of the calibration by looking at the residuals, but a more formal quantification is to compute the per-line rms for a given LFC exposure as:

$$\text{RMS/line } [\text{ms}^{-1}] = \sqrt{\sum_{n=1}^{N} \frac{\left[\frac{(\lambda_{n,\text{pred.}} - \lambda_{n,\text{theory}})}{\lambda_{n,\text{theory}}} \times c\right]^2}{N}} \tag{7}$$

for $N$ lines, where $c$ is the speed of light in meters.

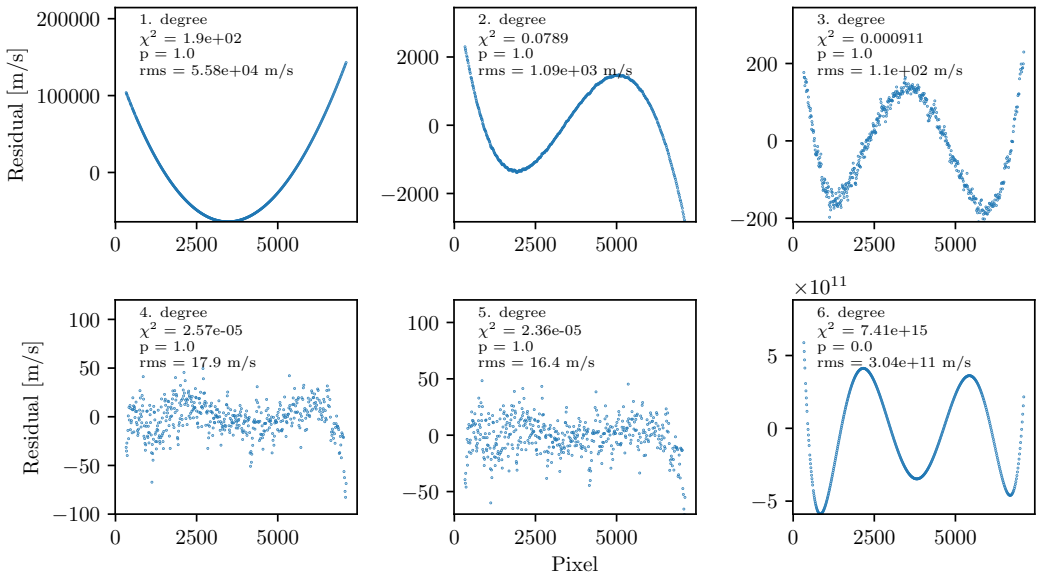


**Figure 5:** Example of residuals from fitting LFC line locations with polynomials of increasing degree in one order (44). The x-axis shows pixels and the y-axis residuals in m/s in all plots. Errors used for determining line locations have been scaled by $\sqrt{3}$. The noted rms in each plot is the per-line rms as defined in eqation 7.

### 3.1.4 Interpolation calibration

A cubic-spline interpolation would force all the residuals to be zero, so in order to evaluate the quality of the method, we can omit for instance every second peak from the interpolation and then compute the residual between the omitted peaks and the resulting interpolation function. Then we can flip it around and interpolate the peaks we left out before and compute the remaining residuals.

Figure 6 shows a comparison of the residuals from a 5th degree poly-fit and the cubic-spline. For the cubic-spline I get a per-line rms of 10.0 m/s, while the poly-fit gives $5.73 \times 10^8$ m/s, which suggest that the cubic-spline appraoch is the better one. It is also worth noting that because the interpolation was done on only half the data points at a time, it will be even better when performed on all data points, as it would be, when used for calibrating data before an RV analysis.

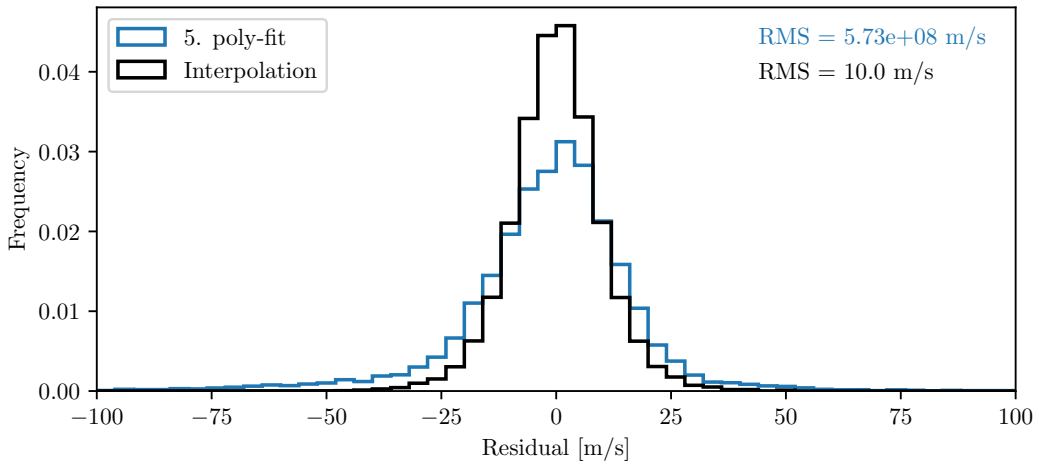


**Figure 6:** Residuals from calibrations performed through a 5th degree poly-fit and a cubic-spline interpolation. The per-line rms as defined in eqation 7 is given in the top-right corner in the color of the corresponding method.

## 3.2 RV extraction

To extract radial velocities we need to measure the doppler shift between spectra from different days of observation. One way to do that is to compute the cross-correlation, which is a measure of the similarity of two data series as a function of the displacement of one relative to the other.

We can do this either for individual absorption features, chunks of the spectrum a few angstroms wide or entire orders at a time. I've chosen primarily to work with the individual features.

Due to a lack of access to data consisting of star spectra with associated LFC captures, I've worked on RV extractions using already calibrated data provided by Lily Zhao. This data has been calibrated using a technique called excalibur [8].

The data is visualized in 7, where the top left shows an extract of wavelength vs. intensity data from an observation of HD 34411. The file also includes a model of the continuum function, with which we can normalize the spectrum through division, shown in the bottom left. On the right side is plotted all continuum normalized data within the EXCALIBUR mask, i.e. data marked as having a proper calibration.

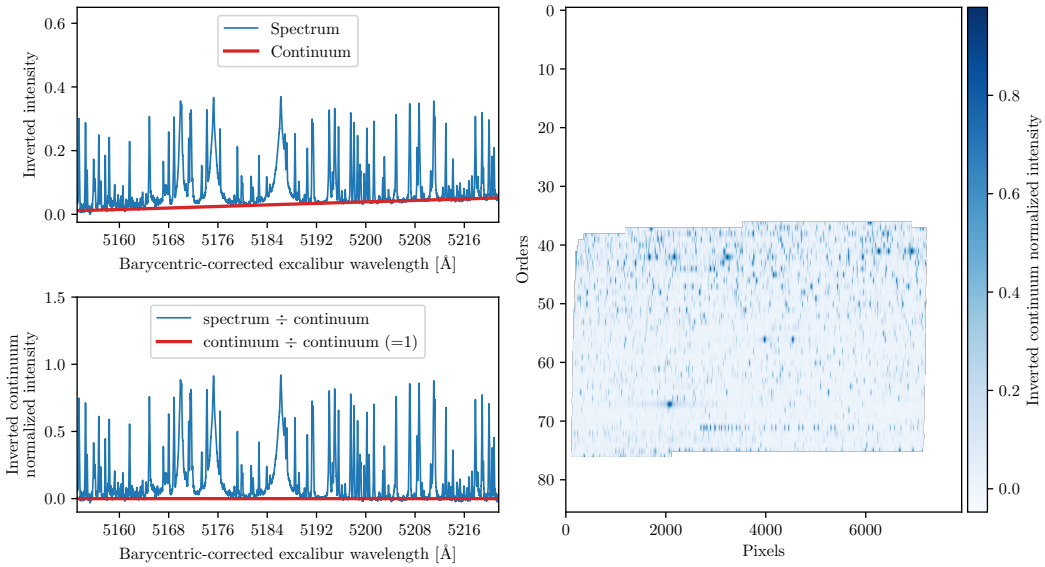


**Figure 7:** Overview of excalibur calibrated data from an observation of HD 34411. Upper left: extract of wavelength solution vs. intensity. Lower left: continuum normalized spectrum. Right: all continuum normalized data within the excalibur mask.

### 3.2.1 Finding and matching features across observations

In order to measure how much individual absorption features move in between observations the first challenge is to find the "same" features in both observations. What follows is a quick rundown of the procedure I've devised:

- Load intensities from the data column `"spectrum"` and errors from `"uncertainty"` as well as excalibur calibrated barycentric-corrected wavelengths from `"bary_excalibur"`, all masked by `"excalibur_mask"`.
- Normalize intensities and errors with the continuum function from `"continuum"`.
- Invert intensities to turn absorption features into positive peaks by $y = 1 - y$.
- Locate peaks using `find_peaks` from `scipy.signal` with minimum peak distance of 5 pixels and a minimum peak prominence of 0.25 (unitless).
- Finally slice data around each peak with a width of 30 data points.

And then to match features/peaks between two observations:

- Iterate through the peaks of observations1 and find the closest peak in observation2. With excalibur calibrated data, peaks should not shift so much that they overlap. However the algorithm laid out so far does sometime match peaks that are far apart or do not resemble each other in shape at all. To bypass such matches we can add two filters:
  - Maximum peak distance: We could filter out all matches where the distance between the peaks is equivalant to a radial velocity greater than 12.5 m/s (the RV Jupiter induces in the Sun). However, since we are dealing with discrete data, the difference sometimes comes out much larger than it actually is and a narrow cut of 12.5 m/s would remove many good matches. Instead setting a very generous cut of 0.5 Å, equivalant to about 20-30 km/s depending on the wavelength, filters out the few very bad matches, but leaves the rest. When analyzing non-barycentric-corrected data, I set this up to 1Å.

- Maximum difference between the areas under the graph of two features (the sum of the intensity values in the feature): Peaks with similar shapes will give a low difference. This filter is most useful when analyzing non-barycentric-corrected, where features often move so much that they overlap. In figure 8 is plotted an example of a bad and a good match, where the bad match can be avoided by setting the max area difference down to 0.1 (unitless).

- I've not yet devised any formal way to determine the best choice of these filters.

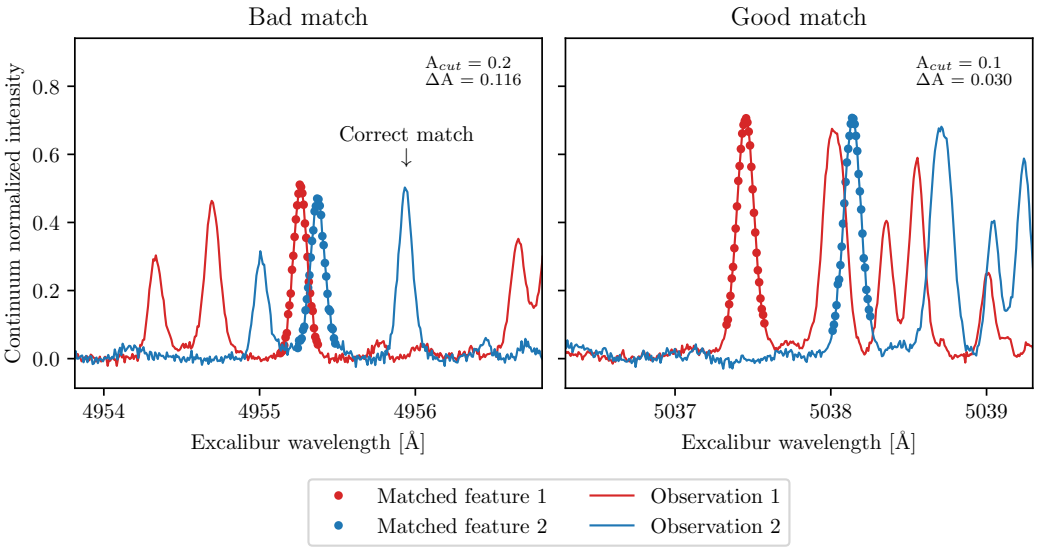


**Figure 8:** Example of match filtering with non-barycentric-corrected data, where features move a lot. The maximum area difference cut was set to 0.2 and 0.1 one the left and right plot repsectively. The cut on the left plot was not strict enough to avoid the bad match that was made, with a an area difference of 0.116. While on the right plot, the correct feature was selected spite another feature being closer in wavelength.

### 3.2.2 Computing velocity shift as the cross correlation

At this point, I have a list of matching features in different observations. The cross-correlations will be performed as a chi2 minimization fit to find the radial velocity that for each match correctly shifts one feature onto the other one.

Before moving on though, we have to cubic-spline interpolate the spectra data. I do this for two reasons: 1) the shifts we are looking for are much smaller than the individual pixels on the CCD, so we need to be able to shift by sub-pixel amounts, and 2) in order to compute the difference in intensity values between peaks, the intensity values must have the same wavelength solution, but, since EXPRES is calibrated independently for each observation, the wavelength solutions are different.

So I cubic-spline interpolate the spectra data from the first observation in the match, but before interpolating the second observation, I shift the wavelength solution by multiplying the shift factor from equation 2. Now I can evaluate the two interpolation functions on a common wavelength range, using $N = 1000$ steps[1], and I am ready to compute the chi2:

$$\chi^2 = \sum_{i=1}^{N} \left[ \frac{y_i - f(x; v)}{\sigma_i} \right]^2 \tag{8}$$

where $y_i$ are the unshifted interpolated intensity from the first observation, $\sigma_i$ are the errors on the intensity of the first observation also sampled through a cubic-spline interpolation, and the function $f(x; v)$ is the cubic-spline interpolation function given by interpolating the intensity values of the second observation with wavelength values shifted by equation 2, evaluated on the wavelength range common to both features:

$$f(x; v) = \textbf{interp}[x \times (1 + v/c), y]\,(x_{\text{common}}) \tag{9}$$

I then compute the cross-correlation and obtain the radial velocity, $v$, as a minimization of eq (8) using iminuit for one feature.

[1] 1000 steps appeared as the best balance between run time and resulting uncertainty. See figure 13 in appendix A:

### 3.2.3 Computing velocity shifts for all features

We can now find and compute the cross-correlation for all feature matches between two observations. In general I find between 500-1000 matches between two observations, which results in a fair amount of statistics. The computed radial velocities do build up a normal distribution around a central value, but there are many outliers. This is due to primarily two things: 1) Bad matches making it through the selection filters, and 2) stellar activity; processes on the star's surface that cause absorption features to shift independently of the velocity of the star.

A better matching algorithm could supposedly be developed, but for now, I've just experimented with setting different cuts both in radial velocity and chi2 to weed out the bad ones. What I've found to work the best however is simply taking the median instead of the weighted average or the mean, as the median is much less affected by outliers. And as there is no reason to believe that stellar activity would cause more RV outliers in one particular direction, the spread can be assumed to be symmetric and thus the median is a good approximation. In figure 9 is plottet an example, which also contains at least one bad match on the far left and a constellation of features likely to be affected by stellar activity on the far right.

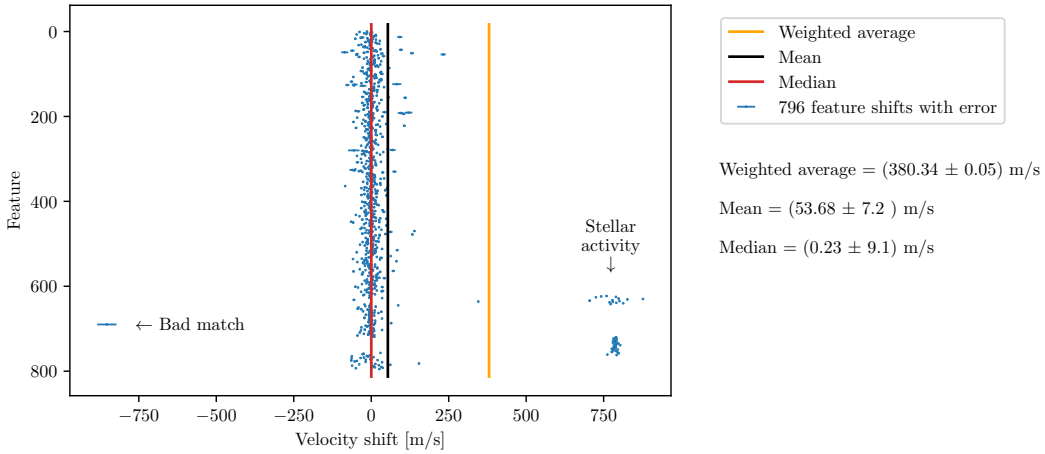


**Figure 9:** Comparison of weighted average, mean and median for computed rv shifts for two observations of HD 34411. The feature on the left side turns out to be a very broad feature, for which the standard data slice is too narrow and therefor yields bad results, (see figure 14 in appendix A:) for details. While the the features on the right side are likely to be caused by some kind of stellar activity.

The errors however is where my method is falting. Also plotted in figure 9 are the respective errors on the weighted average, the mean and the median. As it is the median that I've decided to work with, it would make sense to use the median error

$$\sigma_{\text{median}} = \frac{\sigma}{\sqrt{N-1}} \times \sqrt{\pi/2}$$

for large $N$ where $\sigma$ is the standard deviation,however, it not only comes out very large, but also breaks the propergation of the original errors from the spectrograph, as it essentially is a scaled standard deviation. For the sake of continuing with the analysis I go on with the weighted error, although that in constrast is very small. More on errors in the discussion.

### 3.2.4 Extracting relative shifts from over an overconstrained system

With the devised method we can now compute the relative readial velocity shift between two observations and get out one number with an uncertainty. The next most obvious step would be to compute the shift between observations 1 and 2, 2 and 3, etc. Doing this however leads to correlated results, as the difference between say observation 1 and 10 will depend on all the observations in between, and if there is one bad one, this will affect all the rest. To circumvent this, we can first compute the relative shift between all observations. This will give us an *overconstrained system*, in the sense that there is more information than necessary. All these differences can then be reduced down to a single array, where each shift is relative to all the rest, not only the neighbor.

Computing the shifts between all observations yields an $N \times N$ upper triangular matrix, where each cell is the shift between observations $i$ and $j$, and thus with a diagonal of zero. I will call this matrix $\Delta V_r^{ij}$, see figure 10 for an example.

To reduce the matrix to one array, we can perform another chi2 minimization fit, defined bellow in equation 10, in which we fit an array of parameters we can call $V_r^i$ of length $N$

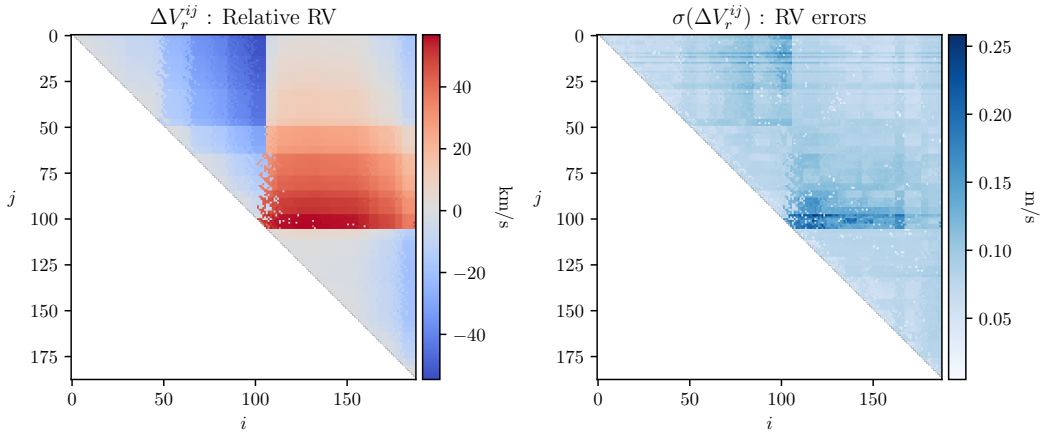


**Figure 10:** Radial velocity shifts matrix computed for 188 observations for HD34411 using excalibur calibrated but non-barycentric-corrected data (column `excalibur`). Each cell shows the median radial velocity shift for all features found between observations $i$ and $j$.

(the number of observations), initialized to zero. The chi2 will be at its minimum when it has found an array of velocities $V_r^i$ that best describe all the differences in the matrix $\Delta V_r^{ij}$ and each of the resulting velocities are thus relative not only to its neighbors but to all the other observations as well, thereby avoiding the correlation.

$$\chi^2 = \sum_{i,j=0}^{N} \left[ \frac{\Delta V_r^{ij} - (V_r^i - V_r^j)}{\sigma(\Delta V_r^{ij})} \right]^2 \quad : \quad i < j. \tag{10}$$

For the sake of illustrating the method, I've analysed non-barycentric-corrected data, that is to say, data in which we should be able to see the movement of the Earth around the center-of-mass of the solar system. That is the data plotted in the matrix in figure 10 and the resulting extracted relative radial velocities (the fit parameters $V_r^i$) are plotted in figure 11 (black), where we see a clear signal of Earth's movement. I've fitted the data with a periodic function and found a period of $(366.482 \pm 3\text{e}{-6})$ days and an amplitude (the orbital speed of the Earth really) of $(28.132 \pm 9\text{e}{-7})$ km/s. Here there are three things to notice: 1) the errors of the period and amplitude do not cover the discrepancy with the actual values of 365.24 days and 29.78 km/s, 2) the very high chi2 value, and 3) the p-value of zero. These suggest very clearly that my errors are wrong. Nevertheless, the signal is clear and the period and orbital speed found are definitely in the right order of magnitude, from which I confirm that my method in general is working. Comparing with the direct differences between observations 1 and 2, 2 and 3, etc, plotted in blue in figure 11, it is also clear that the last step of computing the relative shift between all observatons is vital.

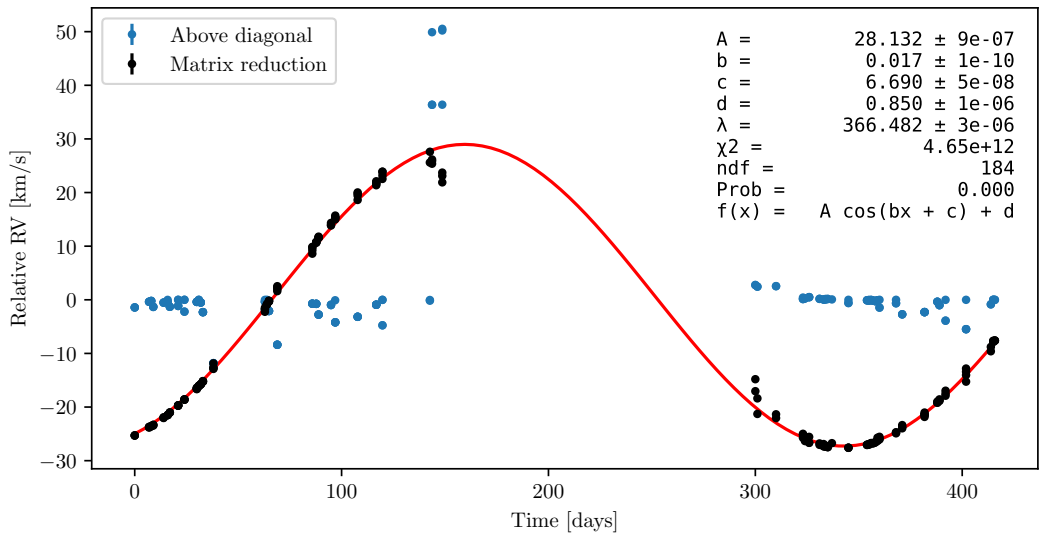


**Figure 11:** Final relative radial velocity results for 188 observations of HD 34411 using excalibur calibrated but non-barycentric-corrected data (data column `excalibur`). Black: values computed through the overconstrained system approach. Blue: above diagonal of $V_r^{ij}$.

Although the amount of computations necessary for the described analysis is high, it is possible to execute for 188 observations on a personal computer in about 2 hours. More details are listed in appendix A:.

# 4 Results

Figure 12 shows my extracted relative radial velocity results for HD 34411 compared to those of Lily Zhao et al (also available upon request [7]). As the Earth's movement has been removed, we are now on the scale of m/s instead of km/s. The method used by Lily Zhao et al is a "chunk-by-chunk" method, where each spectrum is split into 2Å chunks for each of which an RV is found by shifting a template spectrum to match the observed spectrum. That is to say, a different method from mine. Our results however coincide within the same order of magnitude and with a similar rms of 1.67 m/s and 1.78 m/s, mine and hers respectively. The residuals however do jump arund quite a bit, and, of course, my mean error of 0.8 cm/s is not correct.

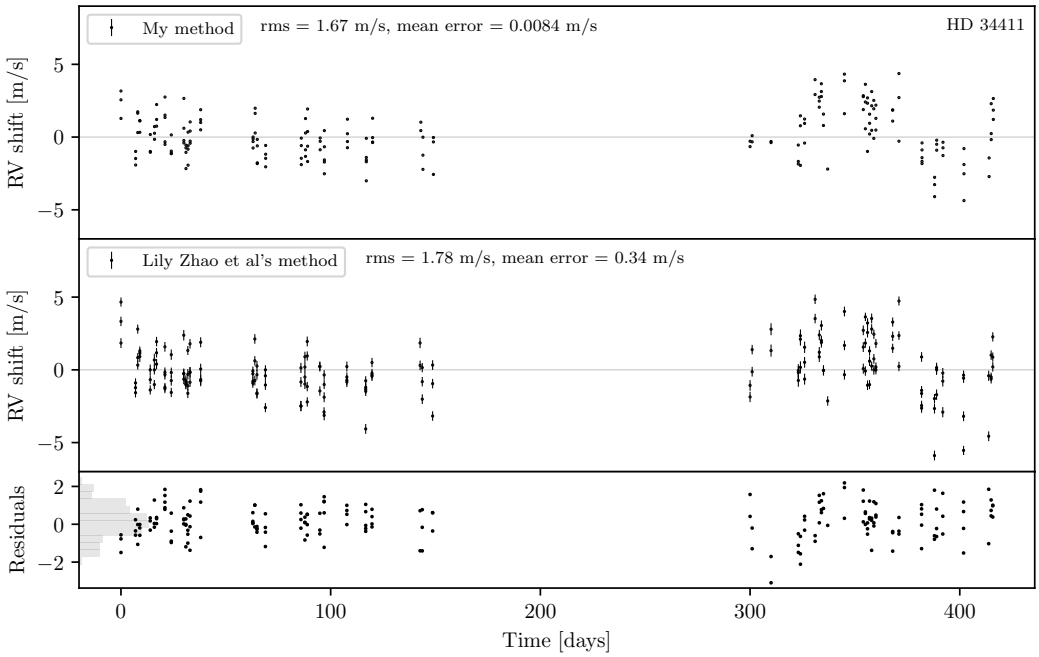


**Figure 12:** Top: my final extracted RV shifts for HD 34411 using excalibur calibrated, barycentric-corrected data (data column `bary_excalibur`). Center: Lily Zhao et al's results using a chunk-by-chunk (CBC) analysis technique [7]. Bottom: residuals from subtracting Lily's results from mine. 188 observations made between 2019-10-8 and 2020-11-27.

I've analysed data from three more stars and again compared the results with those of Lily et al. As with HD 34411, my rms values come within a few m/s, the residuals are not small, and my errors are too small. See figures 15, 16 and 17 in appendix B: for plots.

# 5 Discussion

For the calibration my best solution was the cubic-spline interpolation, with which I got a per-line rms of 10.0 m/s. This is not to much help in the pursuit of rv precision to centimetres per second and nowhere near that of excalibur, which, obvisouly using more advanced techniques, achieve an rms of 3.3 cm/s [8]. **todo: Try calib with 90% random selection**

The results from analyzing the non-barycentric-corrected data were not on point. I measured one earth-year to be ∼366.5 days with a very small uncertainty. One day off. And the orbit speed was also off by more than 1 km/s. If EXPRES should be capable of measuring stellar RV to a precision of a dozen centimetres per second, I think I should be able to get these values right. However, considering the crudeness of my method and especially the resulting very large spread of the RV computed for individual features (as seen in figure 9), it is perhaps not surprising, but rather justifies the decade long development of much more advanced RV extraction methods.

Nevertheless, the signal is definitely there. And, comparing my results for the barycentric-corrected data to that of Lily Zhao et al is also quite positive. RVs on the same order and many patterns in her results coincide with mine. The residuals from subtracting her results from mine are however quite large. For all four stars analysed my rms also comes out smaller than hers. This might be due to my match filtering, which is biased toward selecting features that are closer together if those features do not differ too much in shape.

My final RV errors are wrong. It is the pursuit of many scientists to reach centimetre per second precision. I obvisouly did not just achieve a precision of 8 millimetres per second. But I also suspect I know where the breaking point is. As shown in figure 13 in appendix A:, the error that iminuit gives on the radial velocity when computing the cross-correlation for a given feature gets smaller the more data points I sample from the calibration function. This is a pretty clear sign that I am doing something wrong. Perhaps a solution could be to only interpolate one of the peaks at a time. The chi2 would then be a sum, not of 1000 sampled points, but of the number of actual data points, so around 13. As there might be some bias from interpoloating only one of the peaks, one could then flip it around and interpolate the other instead, finally taking the mean of the two results. This would however lead to twice as many computations, but also with a lot fewer parameters. Depending on the overhead of iminuit I imagine it could even out. At a first attempt however, I often got statistically incompatible results when flipping whose turn it was to be interpolated. I would deem it worthwhile to pursue this, as it could be a breaking point for the errors.

Having spent much more time during this project on exploring the data and devising solutions with my supervisor than reading papers, there is obvisouly much inspiration to gather from digging down more carefully in the literature. Feature lines could for instance be be weighted according to their stability over time. It is also conceivable that through the application of knowledge of stellar activity along with machine learning and a lot of data, features could be grouped and categorized according to the cause of their shift, which would be another way to sort out features are unstable over time.

# 6 Conclusion

# 7 Acknowledgements

# References


[1] Ryan T Blackman, JM Joel Ong, and Debra A Fischer. The measured impact of chromatic atmospheric effects on barycentric corrections: Results from the extreme precision spectrograph. *The Astronomical Journal*, 158(1):40, 2019.

[2] John M Brewer, Debra A Fischer, Jeff A Valenti, and Nikolai Piskunov. Spectral properties of cool stars: extended abundance analysis of 1,617 planet-search stars. *The Astrophysical Journal Supplement Series*, 225(2):32, 2016.

[3] C Jurgenson, D Fischer, T McCracken, D Sawyer, A Szymkowiak, Allen Davis, G Muller, and F Santoro. Expres: a next generation rv spectrograph in the search for earth-like worlds. In *Ground-based and Airborne Instrumentation for Astronomy VI*, volume 9908, page 99086T. International Society for Optics and Photonics, 2016.

[4] Lennart Lindegren and Dainis Dravins. The fundamental definition of "radial velocity". *Astronomy & Astrophysics*, 401(3):1185–1201, 2003.

[5] Christophe Lovis, Debra Fischer, et al. Radial velocity techniques for exoplanets. *Exoplanets*, pages 27–53, 2010.

[6] Ryan R Petersburg, JM Joel Ong, Lily L Zhao, Ryan T Blackman, John M Brewer, Lars A Buchhave, Samuel HC Cabot, Allen B Davis, Colby A Jurgenson, Christopher Leet, et al. An extreme-precision radial-velocity pipeline: First radial velocities from expres. *The Astronomical Journal*, 159(5):187, 2020.

[7] Lily Zhao et al. Expres stellar-signals project.

[8] Lily L Zhao, David W Hogg, Megan Bedell, and Debra A Fischer. Excalibur: A nonparametric, hierarchical wavelength calibration method for a precision spectrograph. *The Astronomical Journal*, 161(2):80, 2021.

# Appendix A: Additional details

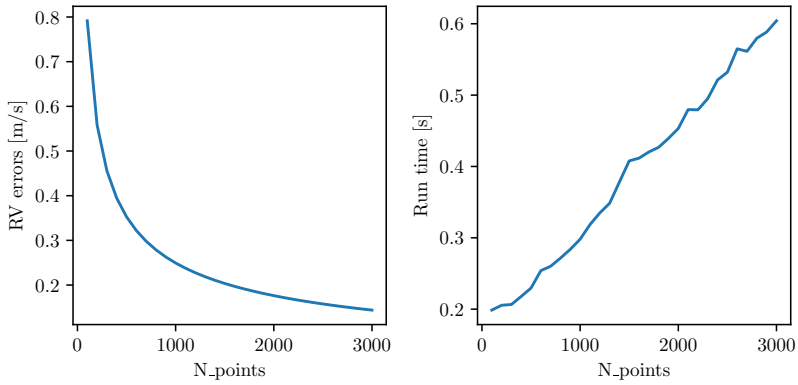


**Figure 13:** Number of data points in each LFC peak fit, determined by the average distance between peaks in each order.

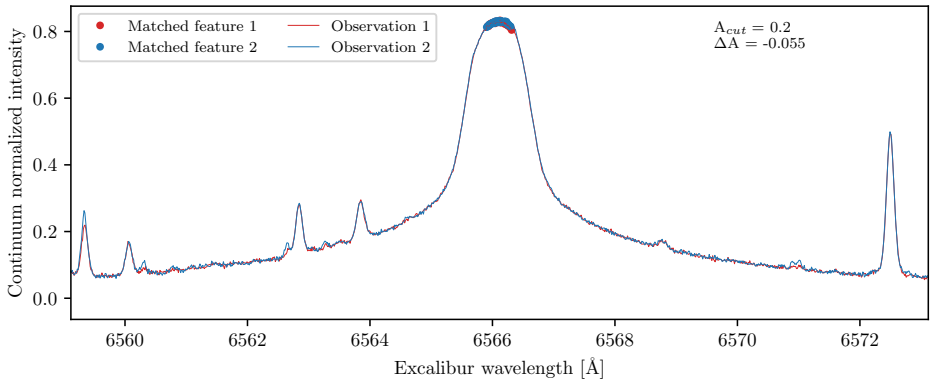


**Figure 14:** Example of a bad match that made it through the filters. This match is actually correct, but yields a bad result because it is too broad for the standard data slice size.

## Computational run times

**todo: add number of function calls for the fitting**

**todo: add number of peak analyzed and so on**

**todo: add run times**

# Appendix B: Additional results

RV extracts from three more stars: HD 101501, HD 10700 and HD 26965. In each figure the top shows my final extracted RV shifts using excalibur calibrated, barycentric-corrected data (data column `bary_excalibur`), the center shows Lily Zhao et al's results using a chunk-by-chunk (CBC) analysis technique [7], and the bottom shows residuals from subtracting Lily's results from mine.

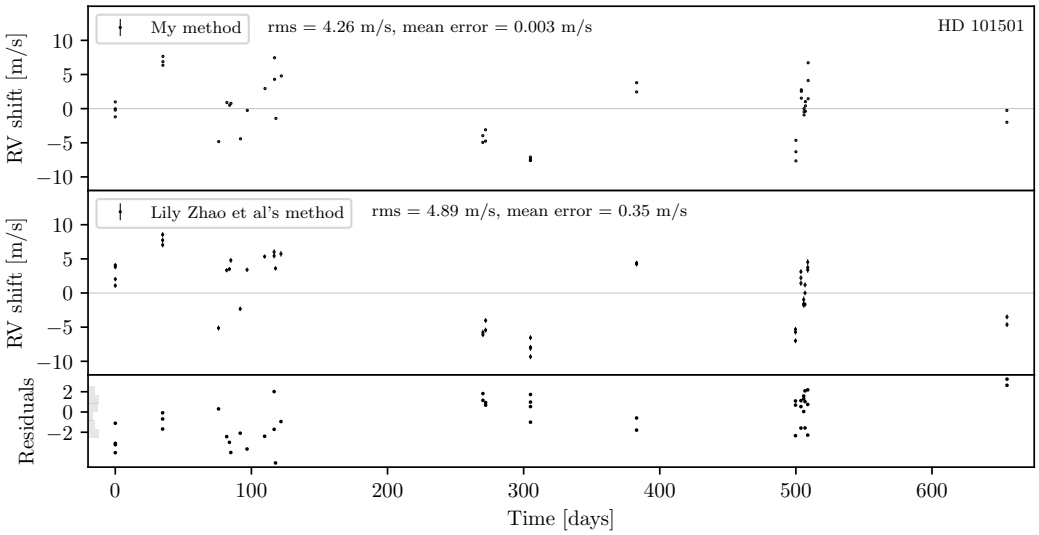


**Figure 15:** HD 101501
45 observations made between 2019-2-10 and 2020-11-26.

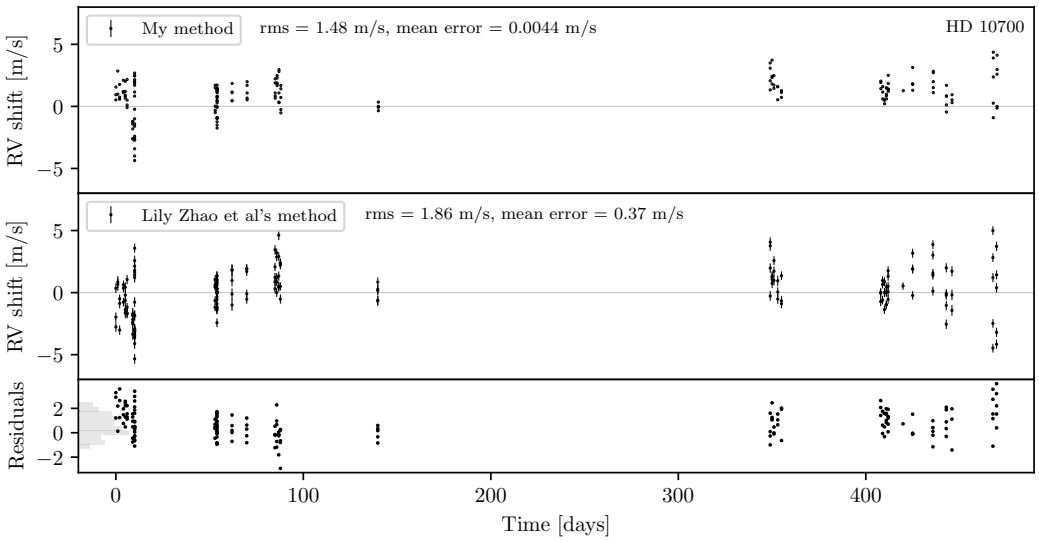


**Figure 16:** HD 10700
174 observations made between 2019-8-15 and 2020-11-27.

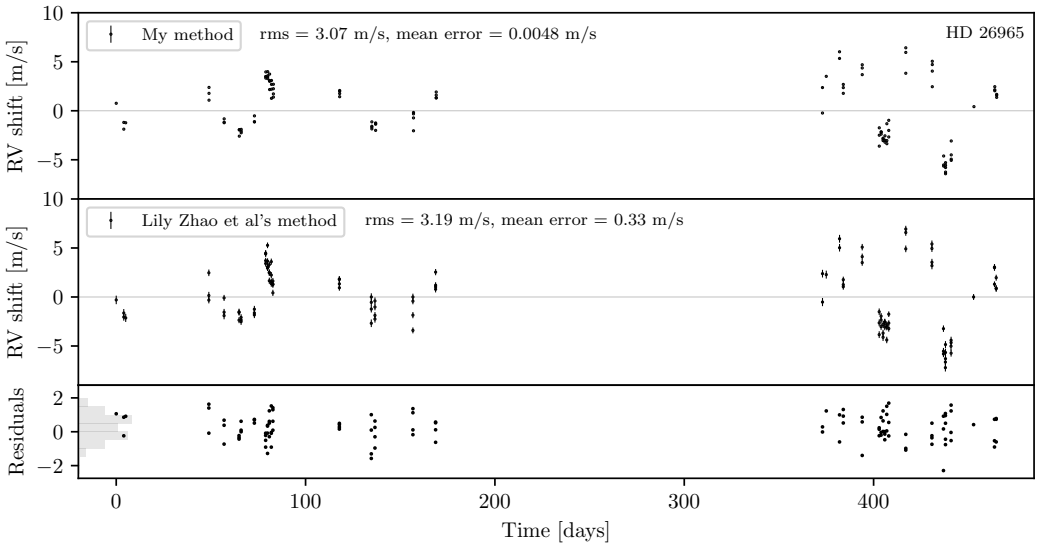


**Figure 17:** HD 26965
114 observations made between 2019-8-20 and 2020-11-27.