Bachelor Project in Physics

Jakob Harteg, wmc573

June 2022, University of Copenhagen

todo: standardize figure sizes

1 Introduction

To discover Earth analogs around other stars, next generation spectrographs must measure radial velocity (RV) with 10 cm/s precision.

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 spectrographs such as EXPRES, we are slowly approaching the precision necessary for the discovery of Earth-sized planets around Sun-like stars. To achieve such precision however, it is necessary to understand and mitigate many effects. This includes the movement of the Earth relative to the center of mass of the solar system (bary-centric corrections), light scattering in the Earth's atmosphere (tellurics), light scattering inside the spectrograph (blaze). Furthermore a general wavelength calibration of the spectrograph data is needed, the quality of which of course directly influences the precision of the final radial velocities that can be obtained. To perform this calibration on the EXPRES spectrograph a laser frequency comb (LFC) is used. The full procedure from raw data to results, also spoken of as the pipeline, is extensive and described in more detail in [6]. This project mainly revolves around performing the calibration and subsequent computation of the radial velocities, and will ignore remaining parts of the pipeline by using already corrected data. The aim of this project is method exploration and development, and not exoplanet detection.

todo: Write that the EXPRES pipeline is very long and complicated and in this paper I will describe a few basic things in the pibeline along with the few things that I have actually worked on.

todo: Write that this paper is intented as an introduction to other students with similar backgrounds to mine interested in working with LFC calib and RV extractions, giving an overview of the data, the problem, and the specific things that I have tried, in hope that it can kickstart their understanding of the project.

2 Theory

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 larger this movement will be. For now, it is not possible to directly observe exoplanets. One way of indirect detection however is thus measuring the movement of the star. We can measure the relative movement of a star 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. The potential indirect signal of a planet will thus consist of a periodic doppler shift in the spectrum of the host star. For this method in general, several years of data collection is necessary. One should at least have one full orbit of observational data, and, to decrease statistical uncertainties, several orbits is preferential.

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]). In the hunt for detecting Earth-sized planets around Sun-like stars, a new generation of spectrographs are appearing called extreme-precision radial velocity (EPRV) spectrographs. The EXtreme PREcision Spectrograph (EXPRES) is one such spectrograph, data from which this project is based on. **todo: move this to the introduction**

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 λ is the observed wavelength, λ_0 is the emited wavelength, Φ 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. **todo:** I don't quite know how to check analytically if this will have an influence on the errors on A given by imniuit. Runing the program it doesn't... So, we can neglect the denominator completely.

If the unit vector $\mathbf{k}$ 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 $\mathbf{k}$ is treat the resulting $\mathbf{v}$ 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 emited wavelength scaled by a factor $(1 + v/c)$. This formular allows us to compute the minimum relative velocity shift, v, between two observations, λ and λ_0.

2.2 Description of the instrument

The EXtreme PREcision Spectrograph or EXPRES is situated at the Lowell Observatory's 4.3m Lowell Discovery Telescope (LDT) near Flagstaff, Arizona, U.S.A. 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 $3800-7800$Å. When a photon hits a pixel it is converted into a charge, and each pixel can thus suply 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. **todo: maybe add illustration**

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.

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.

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)

The observations for HD34411 are plotted in figure 1. Most days have 3-4 observations, and there are significant gaps in the data as well.
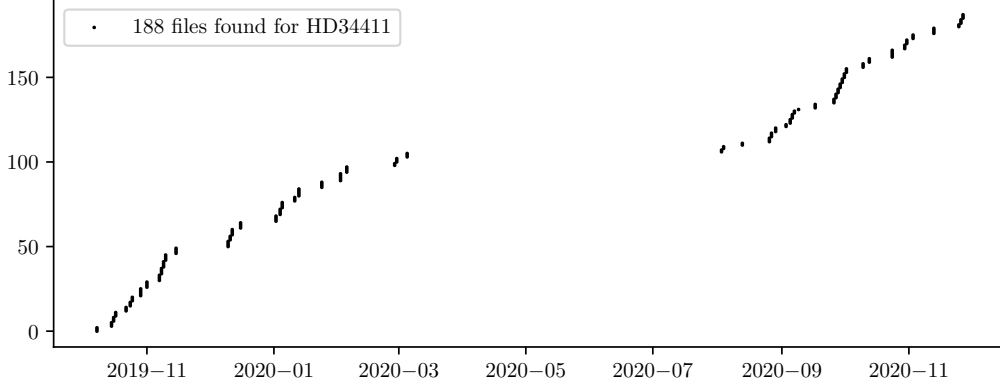

Figure 1: EXPRES observations of HD34411

todo: For development of a LFC calibration method data from what was used

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 refered 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 86×7920, 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 unuseful. The data comes with a mask which shows which pixels are should be used.

To perform the calibration the following variables are used: `spectrum`, `uncertainty`, `wavelength` and `continuum`. To perform the RV extraction on the pre-calibrated data the following variables are used: `bary_excalbur`, `excalibur_mask`, `spectrum`, `uncertainty` and `continuum`.

2.3.2 Noise

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 devitation is then the square root of photon counts. Read noise is calculated emperically, details of which will not be discussed, but is assumed to be consistent throughout each night of observation.

2.3.3 Correction: Scatter / blaze

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. It is modelled and subtracted by measureing the photon count in between orders.

2.3.4 Correction: Tellurics

Tellurics (general definition being originating from the Earth) in this context refers to the contamination that ground based spectrographs have to deal with, which occurs as the light passes through the Earth's atmosphere, encountering molecules such as oxygen and water vapor. On EXPRES the technique used is called SELENITE [https://arxiv.org/pdf/1903.08350.pdf].

2.3.5 Correction: Bary centric corrections

todo: not sure what to write here, if anything

3 Data Analysis

3.1 Calibration

The calibration is needed to map each pixel on the CCD to a specific wavelength. Such a map is refered to as a wavelength solution. To do this, we need a light source with known frequencies, preferably many discret peaks. EXPRES uses a Thorium Argon lamp for an initial trial wavelength solution and then a laser frequency comb (LFC) for an more precise solution.

The Thorium Argon lamp produces 4,000 lines across 82 orders, which can be identified and mapped to a wavelength through a *line atlas*. An intial wavelength solution for all pixels is then produced by linear interpolation. (I this project I have not done this calibration).
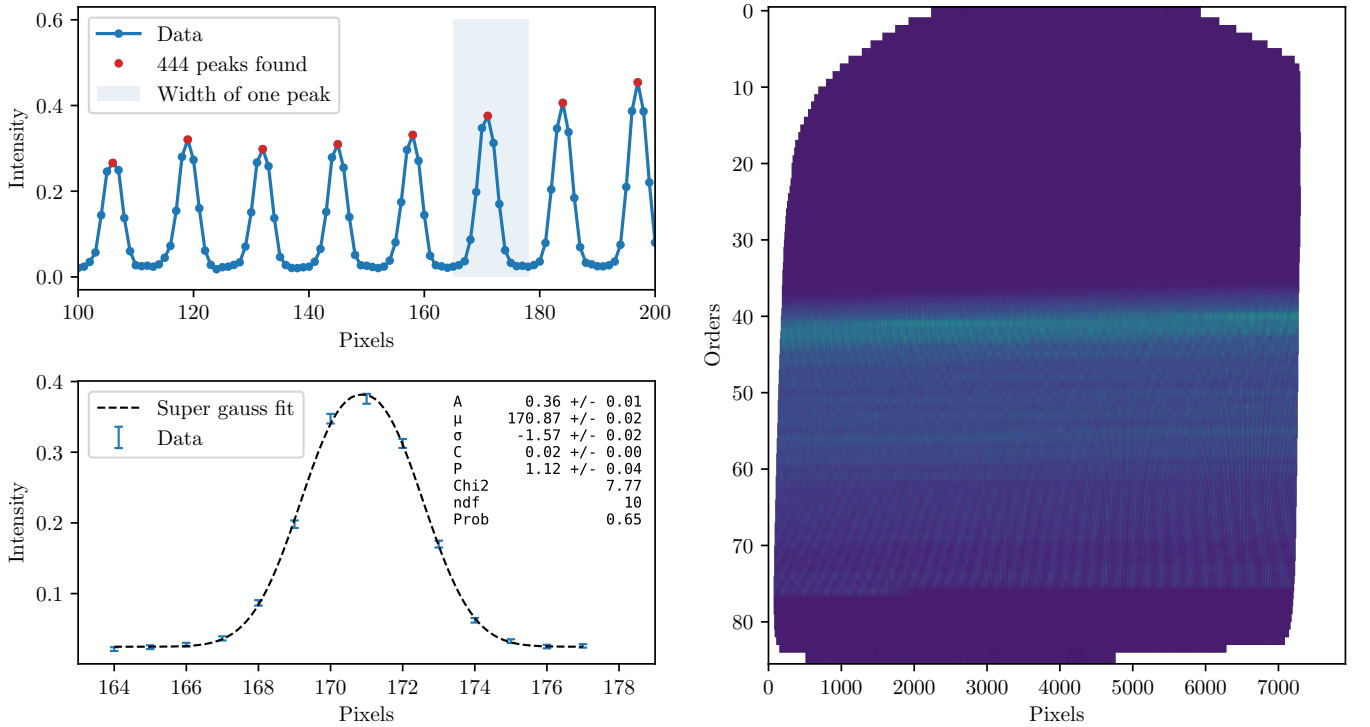

Figure 2: Right: Measured intensities for the LFC across the CCD (unitless). Upper left: illustration of a few LFC peaks in order 65. Peaks are identified with scipy peak finder. Lower left: each peak is fitted with a super gauss to find the exact top of the peak with uncertainties.

The LFC generates a series of equadistant (evenly spaced) spectral lines, typically 20,000 lines across 50 orders. The range of the LFC is thus shorter, and for this reason the ThAr exposures can also be used for a rough calibration outside the LFC range. The frequencies of the LFC peaks are given by the relation

$$v_n = v_{\text{rep}} \times n + v_{\text{offset}} \qquad (3)$$

5

for integers n. The repetition rate v_{rep} and offset frequency v_{offset} are referenced against a GPS-disciplined quartz oscillator, providing calibration stability corresponding to a fractional uncertainty of less than 8×10^{-12} for integration times greater than 1 s. (p. 8, [6]). The values I have used in the calibration, $v_{\text{rep}} = 14e9$ and $v_{\text{offset}} = 6.19e9$, were provided by Lars Buchhave, but may be outdated. See figure 2 right side for a plot of the intensities measured across the CCD.

The following procedure is followed to determine the location of the LFC peaks on the CCD: 1) Find peaks using scipy peak finding algorithm 2) make data slices around each peak with the size of the average distance between peaks, 3) using iminuit do a χ^2 minimisation fit to each peak with a super-gauss plus a linear background. See figure 2 left side.

A super-gauss, defined in eq. (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.

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

The fit then is a minimisation of

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

Where N is the number of data points, x is pixel-space, y_i and σ_i is the measured photon count and uncertainty respectively. The fit returns the values and uncertainties for the parameters A, B, C, P, μ, σ when the value of χ^2 is minimized.

We are most interested in μ, which gives the position of the LFC peak on the CCD (in pixel-space). With the intial rough wavelength solution dervied from the ThAr lamp (pre-calculated in the data set that I've used) 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.

Of course we need to have a wavelength solution for all points on the CCD and to do that I have explored two approaches: cubic interpolation and polynomial fitting. **todo now: why poly-fit? show plot** I can evaluate the quality of the interpolation calibration by choosing to omit every second peak from the interpolation and then computing the residuals between the omitted peaks and the resulting interpolation function. For the polynomial, I can compute residuals simply by subtracting the location peaks from the fit function. Residuals from the two methods are compared in figure 3.
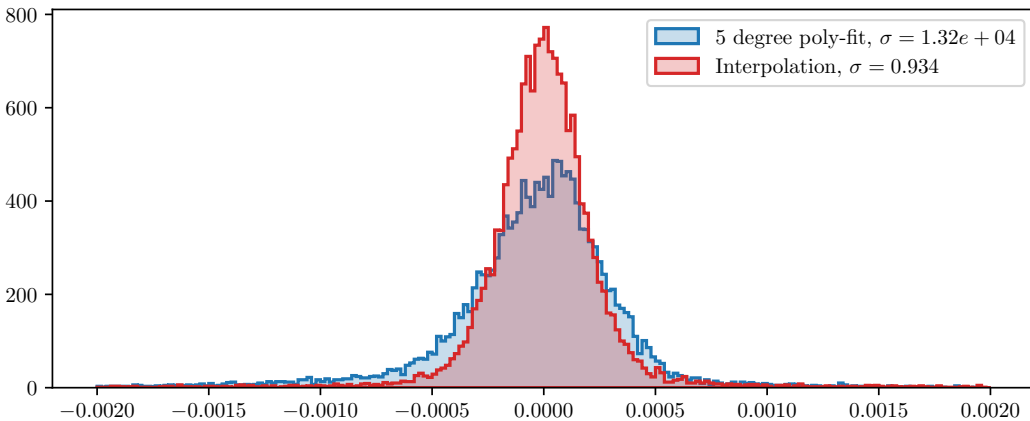

Figure 3: Residuals from calibrations performed through poly-fit and interpolation. Both results contain approximately the same amount of points, but the poly-fit has a much larger spread and therefore appears smaller. **todo: polyfit has 18373 lines, interp has 16888. Why?** Might be because I only use order 40-76 in the interp and all orders with more than 10 peaks for the polyfit. **todo now: Do new polyfit calib** **todo: find out x units**

The standard deviation of the residuals from the interpolation come out much smaller than that of the polyfit (values specified in figure 3), in this example, suggesting that the interpolation method is superior. It is also worth noting that because the interpolation was done on only half the data points, it will be even better when performed on all data points. A similar comparison was done to determine that the polynomial fit got better with increasing degrees until 5th. todo: specify what file?

todo: add graph comparing residuals using gauss vs super gauss

todo: perhaps add plot of changes in parameters across the CCD

todo: specify run times for calibration using poly fit and interp

3.1.1 Errors in the calibration data

The LFC fits files come with an uncertatiny on the photon count (spectrum). It appears however that this uncertatiny might be a bit underestimated. We can see this by plotting the χ^2- and P-values for the LFC peak super-gauss fits, as done in figure 4. The χ^2 value should be roughly equal to the number of degrees of freedom in the fit, which is:

$$N_{\text{dof}} = N_{\text{data-points}} - N_{\text{fit-parameters}} = 13 - 5 = 8. \tag{6}$$

The number of data points per fit I set to be the rounded average distance bewteen peaks in a given order. Although the LFC should generate equadistant peaks, this does vary bewteen 13 and 18 points as we go through orders, most fits having 13 data points (see figure 7 in appendix A). We should therefore see a spike in the χ^2 values roughly around 8, falling of at a slower pace to the right side. Looking again at figure 4, this is not the case for the errors as provided (scale-factor 1). But multiplying by $\sqrt{3}$ we get much closer. $\sqrt{10}$ is overdoing it, however. I looked at several other scaling factors bewteen $\sqrt{3}$ and $\sqrt{10}$, see figure 8 in appendix A for details, and $\sqrt{3}$ comes closest to giving a peak at 8.

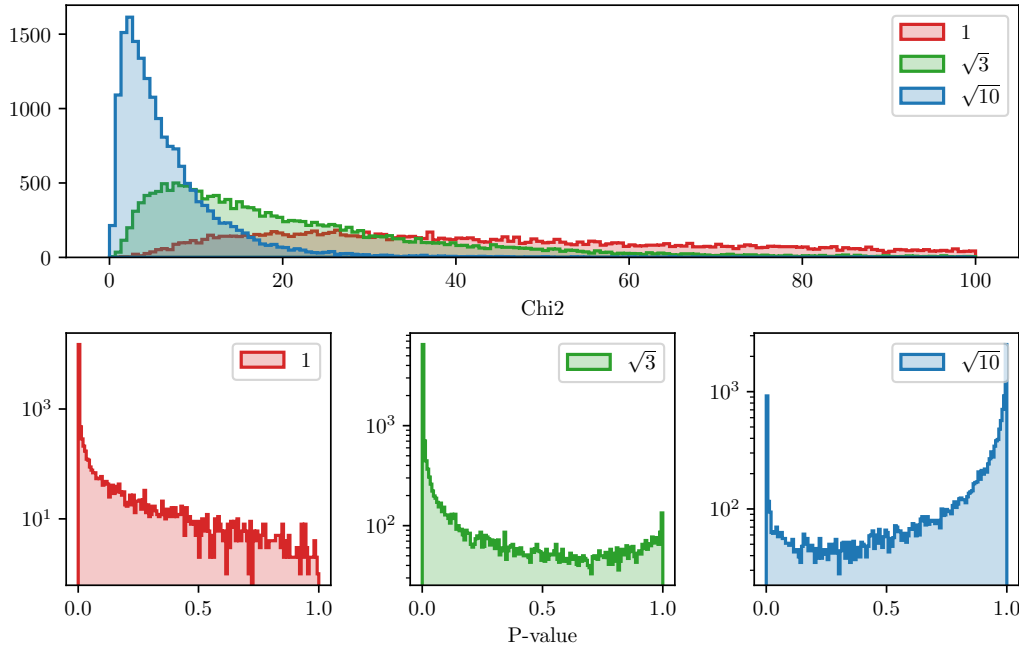Another check is that the probability distribution should be roughly flat. todo: explain why, help .

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 datails.

todo: explain why we use factors of square-root: Something to do with putting into the chi2 which is a square sum, so the square-root goes away.

Effects on calibration:
The errors used during the production of the calibration residuals shown in figure 3 I have already multiplied by $\sqrt{3}$. This gaves us a $\sigma = 0.934$, without this correction I got $\sigma = 2.34$.

> **todo: Computing the sigma for different error scaling factors gave very jumpy results.. why?**

3.2 RV extraction

> **todo: mention that so far I am using pre-calibrated data, because we are missing the right data**

> **todo: include plot of spectrum and continuum normalized spectrum**

In signal processing, cross-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other.

To extract radial velocity we need to measure the doppler shift between spectras from different days of observation. The most straight-forward way to do this is to compute the cross-correlation, since, in signal processing in general, the cross-correlation is exactly a measure of the similarity of two data series as a function of the displacement of one relative to the other.

To perform a smooth minimization of a shift parameter, it is however necessary to have continuous data, so the first step is to perform a cubic interpolation of the spectra data. In practice I do this inside my chi2 minimization function. Taking the spectra data and wavelength solution for from two different files, but adding a shift parameter to the wavelength solution of one of the observations before performing the interpolation. The interpolation functions are then evaluated in the range of wavelength solutions common to both observations, using $N = 1000$ steps. **todo: Inset plot of different run times vs. errors** Plot in `rv_fixes.ipynb` but unsure about the units.

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(x; A)}{\sigma_i} \right]^2 \tag{7}$$

where y_i are the unshifted interpolated photon counts for one file and $f(x; A)$ returns interpolated photon counts for the same wavelenth solution as y_i but with an added wavelength shift, A. The errors, σ_i, are also computed through interpolation.[1]

I then compute the cross-correlation and obtain the wavelength shift as a minimization of eq (7) using iminuit. When it comes to choosing data for fitting there is some freedom. I've explored 2 options: 1) fitting order by order and 2) fiting feature by feature, i.e. locating and individual peaks in the spectra, fitting features obvisouly being the more complicated approach.

Finding and matching features across observations

> **todo:**

Extracting relative shifts from over constrained system - matrix reduction:

The end desired end result is a plot of the relative radial velocities as a function of time or observation. If we were to simply compare each observation with the following one, our results would become correlated in the sense that the difference between day 1 and day 10, let's say, will depend on the quality of all observations in between. To circumvent this, we can compute the relative shift between all observations, yielding an $N \times N$ upper triangular matrix, where each cell is the shift bewteen observations i and j, and thus with a diagonal of 0. I will call this matrix ΔV_r^{ij}, see figure **todo:** The following chi2 minimization with the fit parameters V_r^i (an array of length N) will then find a list of values that best describe the whole matrix. In other words, it finds the relative shift for each observation with all other observations.

$$\chi^2 = \sum_{i,j=0}^{N} \left[\frac{\Delta V_r^{ij} - (V_r^i - V_r^j)}{\sigma(\Delta V_r^{ij})} \right]^2 \quad : \quad i < j. \tag{8}$$

---

[1] The errors, σ_i, which of course also need to be continuous, are computed by: 1) computing another two cubic interpolations for the non-shifted observation, one with photon counts *plus* photon count errors and another one with photon counts *minus* photon count errors. Both interpolations are then evaluted on the same common wavelength range with 1000 steps. The error for each data point is then computed as the mean difference between the measured value and the upper and lower error.
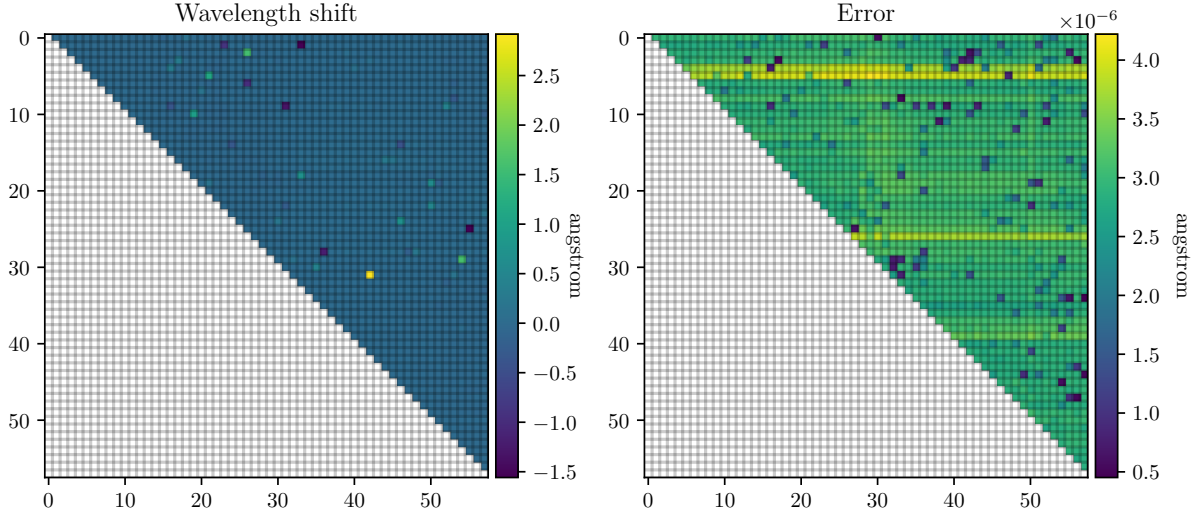
Figure 5: Each cell shows the average wavelength shift computed for features between all observations

This will give us a list of relative wavelengh shifts in angstroms. In figure 6 is plotted the final result for HD34411, but for un-calibrated and un-barycentric-corrected data. This is purely for illustration of the method, since using this data, we can get a nice signal of the Earth's movement around the bary-center of the solar system.

The fit, in red, with parameters on the right, is not doing very well (considering the chi2- and p-values), the errors are pressumably much too small, but it does yield a period (wavelength) close to one earth year.
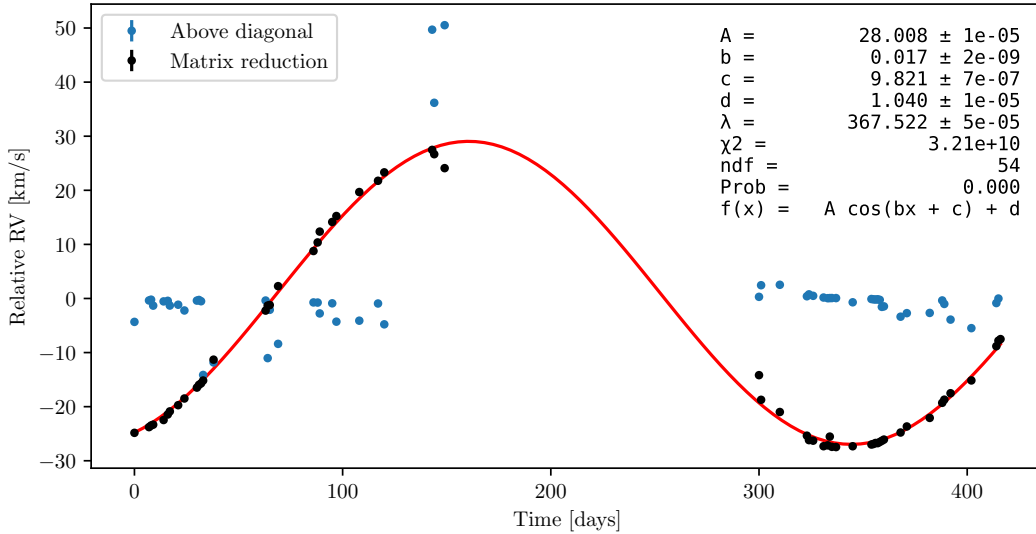

Figure 6: Computed relative wavelength shifts for HD34411 using non-calibrated, non-barycentric-corrected data (column = "wavelength"). Blue: the shifts from day to day. Black: values computed through the chi2 minimization.

But of course we would like to have radial velocities, which allow us to determine the minimum mass of the planet. Converting to radial velocity turns out to be more complicated than expected, as there is no direct conversion factor, but it depends on the specific wavelenght. In general a (non-relativistic) shift in velocity from a shift in wavelength is given by

$$\Delta v = \frac{\lambda_1 - \lambda_0}{\lambda_0} \times c. \tag{9}$$

But as mentioned for this we not just the shift but also the actual position of each of

9

the peaks.

> **todo:** So I need to do gauss fits (?) for the features and find the exact location. Perhaps compute the RMS per line, like they do in eq 5 of excalibur paper:

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

> **todo: add run times**

4 (Results)

5 Discusion

Future ideas:

- Auto encoder

6 Conclusion

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.
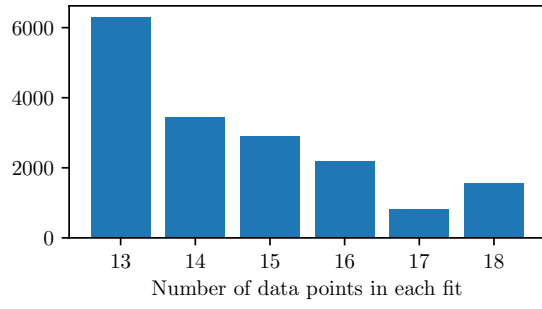
A   LFC photon count errors

Figure 7: Number of data points in each LFC peak fit, determined by the average distance between peaks in each order.
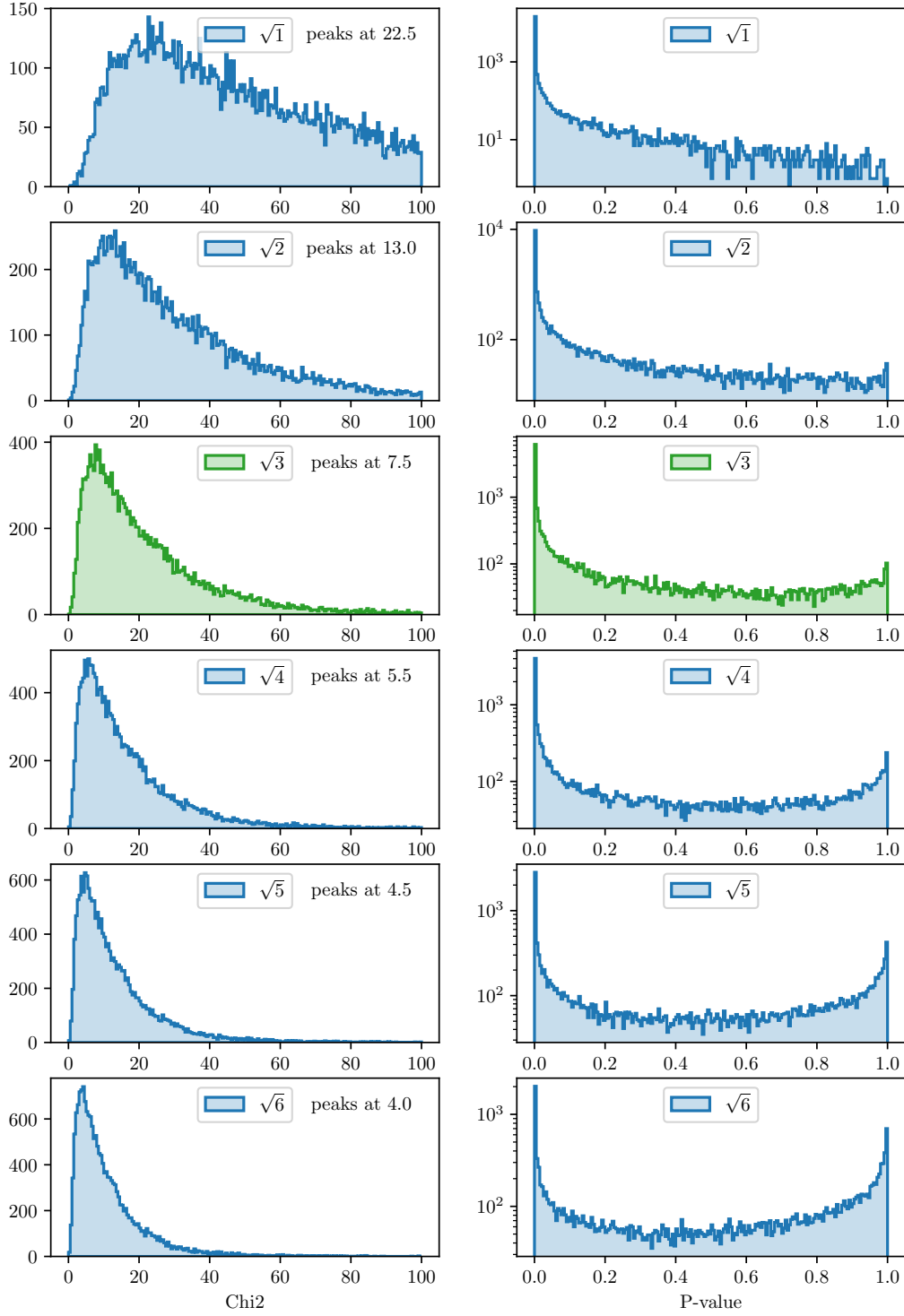
Figure 8: Chi2-values and P-values from individual LFC peak super-gauss fits with photon count (spectrum) errors multiplied by different scale-factors.