Detecting expolanets with the Doppler method using the Lomb-Scargle periodogram.

Simon C. Debes
University of Copenhagen - Project outside course scope
(Dated: October 24, 2022)

In this report, features from spectrograph measurements of the light of the star HD34411 are found using the Lomb-Scargle periodogram, and then clustered using DBSCAN, in order to determine which features come from the star (signal), and which do not (noise). Another clustering on the cleaned data reveals that we can cluster together signals with different origins of the star. Both clusterings allow us to more accurately determine the radial velocity of the star. The radial velocity of the source of each feature was determined using the Doppler method. The variables used in the clustering is the mean, median, and standard deviation of calculated radial velocities of each of the 366 found features, their uncertainties, and the 5 strongest frequencies of the signal, and their corresponding intensity found using the Lomb-Scargle periodogram. Finally the radial velocity of the star is determined to be -0.123 ± 0.004 m/s, on the basis of which we cannot ascertain the existence of an orbiting exoplanet.

INTRODUCTION

If Earth was the only planet in our solar system, the Sun would have a radial velocity of around 8.95 cm/s, which is not a lot compared to the 30 m/s target radial velocity (RV) precision of the EXPRES spectrograph. Therefore, if we want to discover Earthlike planets, we need to increase the precision.

DOPPLER METHOD

Stars are easy to detect because they emit a lot of light, where in comparison, planets are quite dim. Luckily, how we observe the light emitted by stars, depend on the orbiting planets and their properties. In such cases, one can use the Doppler method (also known as the radial velocity method), where only the star the planet is orbiting is observed. If the observed star is orbited by one or more exoplanets, the star and the planets will all orbit the center of mass of the system. Usually the star will contain most of the mass of the system, and the corresponding radius of orbit will have a small radius, but it will often be non-negligible. This orbit gives the light source a radial velocity, causing blue and red shifting of the light when the star moves towards and away from Earth, respectively. By measuring the change in wavelength shift over time, we can infer the radial velocity of the star.

One drawback of the Doppler method is that one only measures the radial velocity in the plane of observation. Theoretically, it is possible for a star to have a high radial velocity due to a large orbiting planet, but only orbit in a plane perpendicular to the observation. In this case the distance from the star to the observer is constant, and thus no Doppler shift occurs, and the star appears stationary. Thus the Doppler method only produces a lower bound of the total radial velocity.

Data

The data used in this report is the Extreme Precision Spectrograph (EXPRES) data set, which is found by using a spectrograph on Earth to take a measurement of the light emitted from the star HD34411 over a period of time. More specifically 58 observations taken on different days, with an irregular sample rate over approximately 400 days (October 8th 2019 - November 27th 2020). The observations are taken somewhat regularly, except for one large gap midway as seen in figure 1.

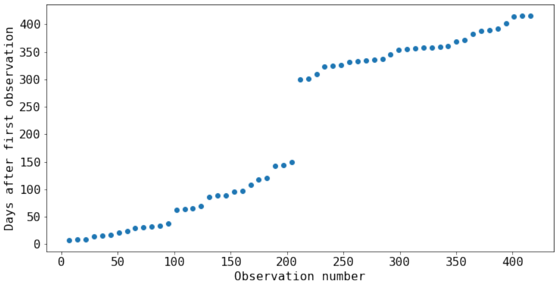

Figure 1: A plot showing how many days after the first observation, each observation was taken.

Initially only the intensity of light is recorded as a function of pixel placement. The wavelength of the light is found by comparing the signal to that of a thorium-argon (ThAr) lamp, where the wavelength of its emission lines are well known [1].

We are interested in finding the wavelength shift of the wobble of the stars orbit. This wobble is very small compared to the shift caused by the movement of Earth, which is by far the largest error. Thus a barycentric correction of the data is needed to remove this effect. In figure 3, the radial velocity is calculated without corrections, and the resulting period of the orbit is around 368 days, showing the dominating effect of Earth's orbit.

Other needed corrections are the scattering of light in Earth's atmosphere (tellurics), and the scattering of light within the detector (blaze). Once these corrections are made, the dominating signal should be light from the star in question.

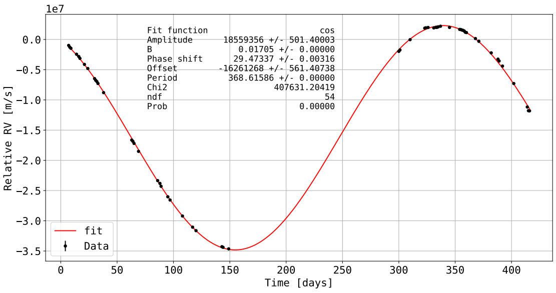

Figure 2: The relative radial velocity of the star without a barycentric correction as a function of time. The points have been fitted with a cosine function.

METHOD

Finding features

Each observation in the raw data is a grid of pixels. Each row in this grid is referred to as an (Echelle) order, and in this report, I have not considered every pixel, but only used one specific order in all observations.

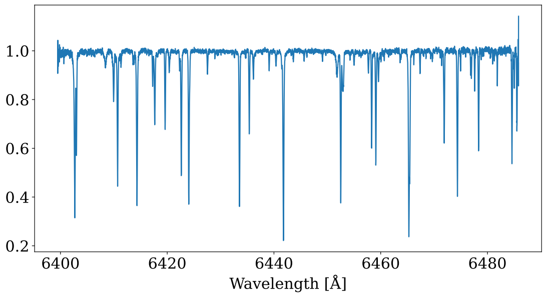

Figure 3: A plot showing one random order of the raw data.

`scipy.signal.find_peaks` was used to identify the location of the peaks (e.g. the ones shown in figure 3). This was done for all 58 observations. Once all features were located, I attempted to track each feature across all 58 observations. The problem here was that there was a lot of features that only appeared in one or a few measurements, and probably were not of interest to us, meaning that the nth peak in observation i, was not necessarily the nth peak in observation $i + 1$. To solve this problem, I made a histogram ranging over the entire wavelength spectrum, and bins containing at least 50 of the 58 points, would be counted as one feature. This method gave reasonable results, as there were some bins with a few features, many with 55-58 features, and none with more than 58. Now each feature had 50 to 58 measurements, and for each feature, a collection of summary variables were produced.

The variables were the mean, median, std and the error on these three, along with the 5 most prominent frequencies and their corresponding strength, found using the Lomb-Scargle Periodogram. Now each feature could be represented as a point in this 16-dimensional parameter space, and could be clustered.

Lomb-Scargle Periodogram

We are looking at a mesh of signals, each with a distinct wavelength/frequency, where only a few distinct wavelengths are of interest, and the rest are noise. This makes this problem well suited for Fourier analysis. Since the Fast Fourier Transform (FFT) makes an assumption of constant sample rate, the lesser known least squares spectral analysis method of Lomb-Scargle periodogram (LSP) is more suitable for this problem. The hope was that light from a certain source would emit a unique wavelength of light, such that all signals originating from this source, would share similar frequencies and intensities, and would thus more easily be clustered together.

Many of the strongest frequencies correspond to periods of less than one day. As there is no more than one observation per day, it would have been difficult for the algorithm to pick up on these frequencies, and they may just be noise. This effect could be negated by taking more measurements, and with shorter intervals.

Determining the radial velocity

The red/blueshifting shift all wavelengths equally, so two wavelength spectra recorded some time apart, would contain the same features, with the same spacing, but with the same constant offset. Once all the features are identified, it is possible to overlay the spectra from different observations, and slide them back and forth until the peaks were matched up. How much each observation has to be moved to line up, is then exactly how much the line has been shifted in wavelength between observations.
When the offset in wavelength for all the correct features in each observation had been identified, a 58x58 matrix could be constructed where the (i,j)'th entry was the difference in wavelength between observation i and observation j for each feature k, creating a $i \times j \times k$ tensor. Given a fixed k, we now have a matrix of the difference in radial velocity between observation i and j, $\Delta V_{RV}^{i,j}$.

Then I created an 58-dimensional zero vector $\bar{v}_{58}$, and minimised the χ^2 of the difference $\Delta V_{RV}^{i,j}$ and $\bar{v}_{58}$, to determine the radial velocity of the star for each of the 58 observations. The n'th entry in v_{58} was then the calculated radial velocity of the star for the n'th observation.

RESULTS

Principal Component Analysis (PCA) was applied to the aforementioned 16-dimensional data, allowing us to plot the data in two constructed dimensions that maximise the variance of the points. Under the expectation that some of the features were noise, the clustering algorithm DBSCAN was applied, to divide the data set in two clusters.

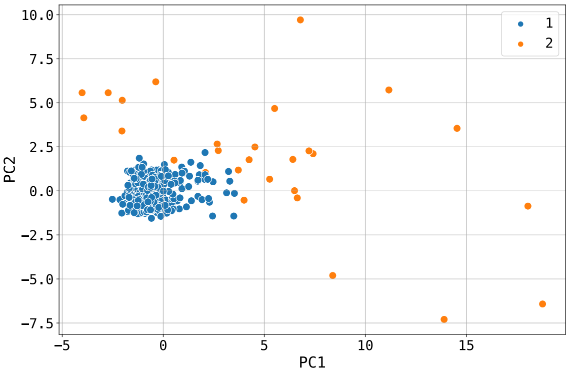

Figure 4: DBSCAN clustering on all features, plotted on the two principal components.

Figure 4 shows a tight blue cluster around the origin, and some orange points scattered a longer distance from the origin. As we expect the movement of the star to produce slight shifts between observations, we expect the signal to be similar in a lot of ways, and large outliers may be attributed to noise or measurement errors. To avoid ambiguities later on, I will therefore refer to the blue cluster (labeled cluster 1 in figure 4 and figure 5) as the good cluster containing the good points, and refer to the orange cluster (labeled cluster 2 in figure 4 and figure 5) as the bad cluster containing the bad points.

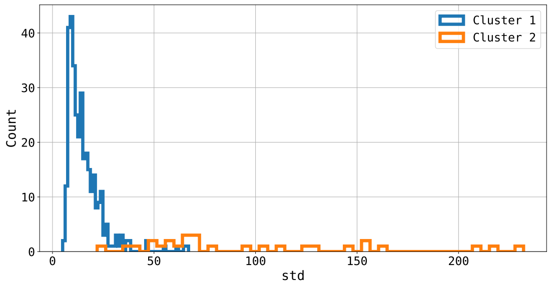

Figure 5: A histogram of the standard deviation of the clusters.

Figure 5 shows that the features in one cluster have much lower standard deviations. The points that were more tightly clustered around the origin, are also the points with the lower standard deviation. We also expected actual signal to have a lower standard deviation, as they would only vary a little bit from day to day, where as features that varied a lot would be presumed to be noise.

	Good cluster	Bad cluster
Counts	339	27
Mean RV $[m/s]$	-1.90	22.63
RMS $[m/s]$	9.94	171.31
Mean std $[m/s]$	15.33	101.52

Table I: Summary variables comparing the two clusters.

Table I shows that the bad cluster has a much higher standard deviation, and the RV velocity on average is larger. This indicates that these features change a lot from from measurement to measurement. From figure 6 we can also see that features in the bad cluster generally vary a lot more drastically in radial velocity than those in the good cluster.

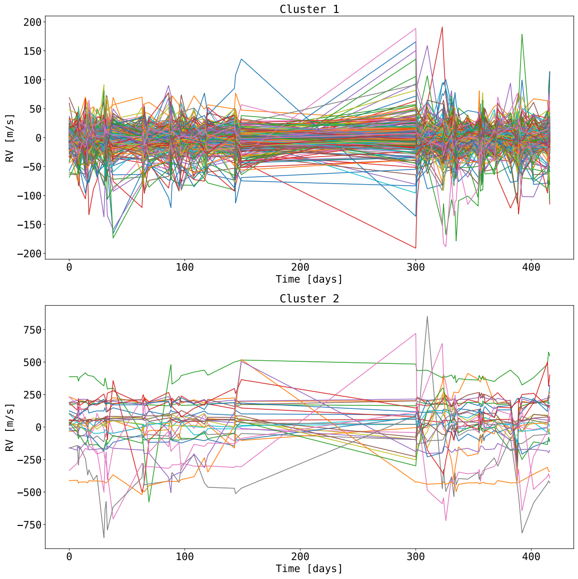

Figure 6: The top plot shows the radial velocity of each feature on a given day, for the features in cluster 1, and the bottom plot shows the corresponding plot for cluster 2. Note that the lines are only there to make it easier to follow a specific feature, and have no meaning in themselves.

Further clustering

A hierarchical clustering algorithm (`scipy.cluster.hierarchy.fcluster`) was used to further cluster the data. The expectation was that signal would originate from different parts of the star, making them inherently different, and that our clustering algorithm would be able to recognize this. This hierarchical clustering algorithm takes `n_clusters` as a parameter, and as there's no way of knowing the right number of clusters, 10 was deemed reasonable. To highlight the benefit of excluding noise, we'll see the 10-fold hierarchical clustering on both the cleaned and uncleaned data. These 10 clusters will from now on be referred to as HA-clusters, to distinguish them from the other sets of clusters found.

Clustering on all points

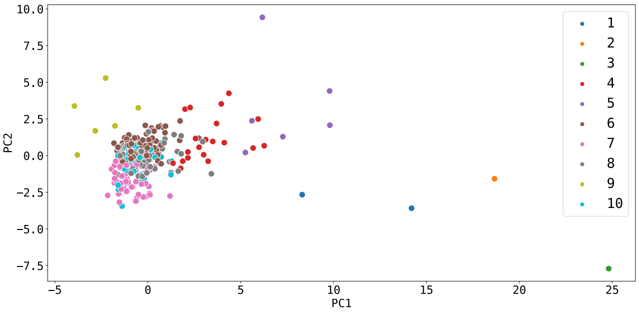

Figure 7: Hierarchical clustering of all features.

Figure 7 shows the clustering done on all features. There is clearly noise present, as a few points are very far away from the others, and some clusters only contain one or two points.

Cluster	$\mathbf{N_{good}}$	$\mathbf{N_{bad}}$	**Mean** $\mathbf{v_r}$	**RMS**	**std**	$\mathbf{F_{max}}$	$\mathbf{I_{max}}$
Good	339	0	-1.90	9.94	15.33	1.50	0.45
Bad	0	27	22.63	171.31	101.52	1.15	5.02
All	339	27	-0.09	47.50	21.69	1.50	5.02
HA1	0	2	-328.08	331.19	220.10	1.02	0.65
HA2	0	1	-170.82	170.82	130.32	1.00	1.40
HA3	0	1	-384.94	384.94	156.16	1.00	5.02
HA4	9	11	5.33	48.66	63.66	1.04	0.56
HA5	0	6	90.34	157.73	98.35	1.15	0.69
HA6	133	0	-0.1238	6.17	12.64	1.5	0.31
HA7	59	0	-2.24	11.99	20.69	1.5	0.38
HA8	86	0	-2.85	11.26	15.42	1.21	0.42
HA9	0	6	182.10	184.54	72.02	1.12	0.27
HA10	52	0	-1.79	8.52	13.26	1.5	0.32

Table II: Summary variables comparing the various clusters. N_{good} and N_{bad} show how many features that have been classified as good or bad, respectively, occur in the given cluster. F_{max} and I_{max} show the maximum frequency found in features in the given cluster and I_{max} shows the corresponding intensity of that frequency.

Table II shows that when doing the 10-fold clustering, there is a clear separation of the good and the bad features in the clusters. Only one of the ten clusters contain features from both. As expected, the clusters with bad features, have a much higher standard deviation and root mean square. Most of the bad features have an I_{max} larger than even that largest I_{max} of all the good points, meaning that there seems to be an upper bound of intensity of the frequencies of actual signal, and if it surpassed, it is more probable that it is noise.

Clustering on good points

The same clustering algorithm was also applied to the data, after all features in the bad cluster had been removed.

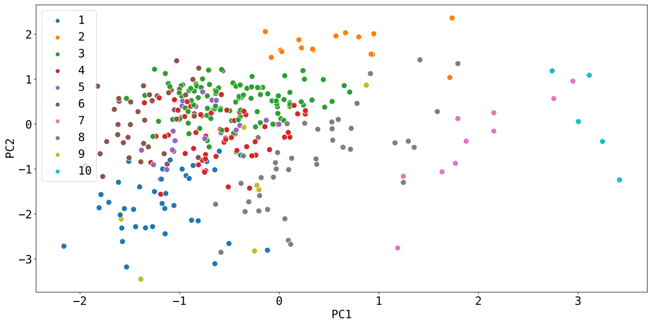

Figure 8: Hierarchical clustering of the features in the good cluster.

As expected, we see a similar structure from figure 7 in figure 8, except the solitary points far away from the origin have been removed, and the points exist on a much smaller range on the principal components, as they are more similar.

Cluster	N	Mean v_r	RMS	std	F_{max}	I_{max}
Good	339	-1.90	9.94	15.33	1.50	0.45
Bad	27	22.63	171.31	101.52	1.15	5.02
All	366	-0.09	47.50	21.69	1.50	5.02
H1	38	-1.72	7.47	21.79	1.50	0.29
H2	16	-3.05	7.36	20.73	1.45	0.30
H3	100	0.33	4.33	10.69	1.50	0.31
H4	57	0.27	4.79	12.16	1.50	0.30
H5	21	-0.47	3.53	11.51	1.13	0.23
H6	43	-1.72	5.33	11.99	1.50	0.23
H7	10	2.87	14.53	24.82	1.01	0.45
H8	42	-0.50	6.96	17.88	1.46	0.34
H9	7	-43.28	43.87	33.37	1.24	0.32
H10	5	-39.46	40.61	57.17	1.03	0.42
Jakob	366+	?	1.67	?	?	?
Lily 1	366+	?	1.78	?	?	?

Table III: Summary variables comparing the various clusters after bad features have been excluded.

Now the features within a given cluster are much more homogeneous, and the RMS and std are much lower across the clusters. Even H5 with only 21 points now has a lower std than HA7 and HA8, which contained 59 and 86 of the good points, respectively. This supports the claim that the clustering has been successful.
Table III compares my results to those of two people who have done the same calculations before me. Jakob and Lily had a lower RMS, but they also used data with 188 observations, where I had 58. Cluster H5 got as low as 3.53, which is not extremely far off, given the smaller data sample. This analysis would have to be performed on a data set of equal size to compare fairly.

The radial velocity of the star was calculated by taking a weighted average of the mean radial velocity of the 58 measurements of each individual feature. The radial velocity of HD34411 was calculated to be -0.123 ± 0.004 m/s. The uncertainty is found using error propagation on the errors from the original data, although it does seem very small.

DISCUSSION

The calculated radial velocity of the star is only a lower bound, due to the Doppler method, as mentioned earlier. Therefore we cannot conclude that there is no exoplanet, just that the findings of this study does not confirm the presence of an exoplanet.

Since I found no evidence of an exoplanet, I focused mostly on the method, and thus there were places where one could be more meticulous, to obtain more accurate results, such as obtaining and using more measurements.

Some of the largest frequencies found using the Lomb-Scargle periodogram were around 1/day, and since all measurements were taken on different days, shortest period between two measurements are around 1 day (assuming the measurements were taken around the same time of the day). The low sample rate and high frequency can easily hide any periodicity in the data. For example, if you measure the angle of the suns position at noon everyday, you could conclude that it never moves. This effect would be negated by using more (and more frequent) data.

CONCLUSION

It was possible to separate signal and noise efficiently, with reasonable separation. We see that most of the features clustered as noise also are the features with the highest standard deviation, which holds up against our expectation. The std and RMS of the features decrease after excluding the noise.
The further clustering of the signal produced clusters that looked as expected, and some that still might have contained noise, as they contained a few points that deviated greatly (namely H9 & H10). Most clusters did appear to be "correct", and the std and RMS was reduced even more.

After excluding the noise from the data, it was possible to calculate the redial velocity of the star to be -0.123 ± 0.004 m/s, which is not significant enough for us to ascertain the existence of an orbiting planet.

OUTLOOK

The data from EXPRES, used in this report comes as FITS files, with certain filters, such that you can access the raw measurements, or e.g. the barycentric corrected data to save you the trouble of doing it yourself. The work done in this report has the potential to become another filter in the EXPRES data, such that each feature can have a label saying if its noise or not, and which cluster it belongs in. This would allow for easy noise cleaning, resulting in more precise calculations.

REFERENCES

[1] Fei Zhao et al., 2021, *Statistical modeling of an astro-comb for high precision radial velocity observation*

[2] EXPRES Stellar-Signals Project - Yale Astronomy.
http://exoplanets.astro.yale.edu/science/activity.php
#dataESPcd%20
Visited on 16/05-22.

[3] An Extreme Precision Radial Velocity Pipeline: First Radial Velocities from EXPRES.
https://arxiv.org/abs/2003.08851