PHYS 499 Interim Report

Improving APOGEE Survey Radial Velocity Precision Using Machine Learning Techniques

Yiwei Chai

advised by Prof. Cullen Blake and Prof. Bhuvnesh Jain

December 21, 2021

1 Introduction

The radial-velocity (RV) method of detecting exoplanets uses observations of Doppler shifts in the spectrum of a star to determine the potential existence of an orbiting exoplanet. RV measurements can also provide information about an exoplanet's minimum mass. This method is best suited to detecting hot Jupiters orbiting close to a star, due to the strength of the gravitational pull in these cases, and limitations in existing instrumentation. Currently, about 20.6% of exoplanets have been discovered via the RV method.

The Apache Point Observatory Galactic Evolution Experiment (APOGEE) survey uses near-infrared spectroscopy to acquire high-resolution spectra of stars across the entire Milky Way. The large-scale nature of the survey APOGEE means that it is able to gather spectra data for several hundreds of stars simultaneously. The spectra from each 'visit' is then reduced to provide an estimated RV. This means that APOGEE is well-suited to increasing the data points available for the RV detection of exoplanets, of which there are relatively few, particularly in comparison to what is available for transiting exoplanets. However, the challenge presented by APOGEE data is that RV thresholds, at 50-100 m/s, are currently lower than expected, limiting its usability in the search for exoplanets. We hypothesise that the worse than expected RV thresholds are caused by systematic errors. By using a principal component analysis (PCA) algorithm to identify and remove these errors, I hope to increase RV precision by at least a factor of 2, and show this is a viable method for identifying exoplanet candidates.

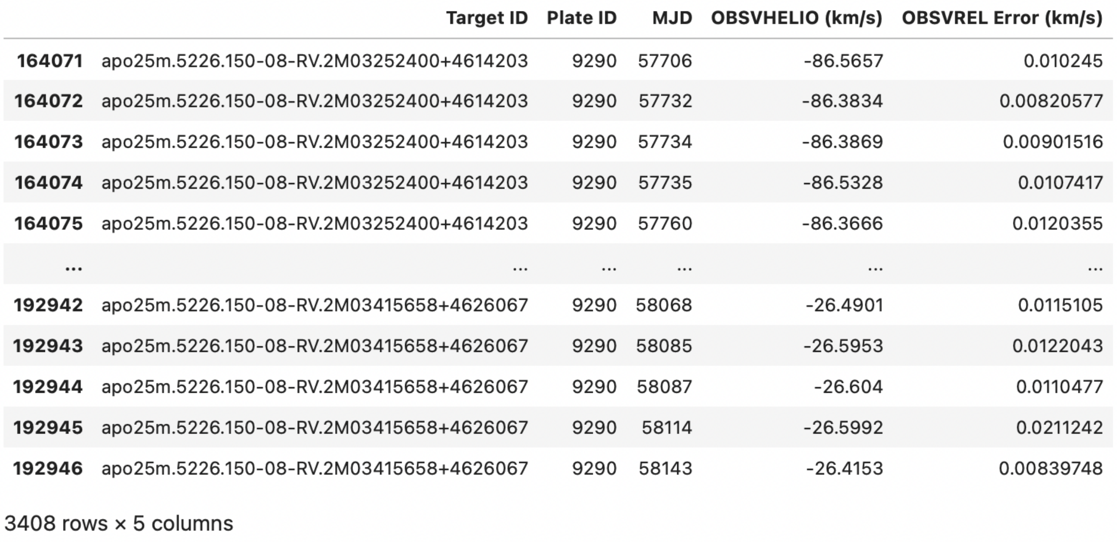
This semester, I focussed on data cleaning to produce a suitable dataset with which to perform PCA. I also conducted preliminary analysis of the cleaned data to determine if a pattern in the average RV scatter could be detected by eye.

2 Methods

2.1 Data Cleaning

APOGEE provides two different data models of RV measurements to work with: the allStar Catalogue, and the allVisit Catalogue. I decided to work with the allVisit Catalogue, which provides the raw data from each stellar 'visit', as opposed to the allStar Catalogue, which provides the averages of all the visits for each star and thus may have already eliminated the types of effects I am hoping to capture with the PCA analysis.

From the allVisit Catalogue, I sought to select a suitable sample of stars that have been observed on the same 'visit' (i.e. night), in order to minimise independent variables within the data. The allVisit Catalogue is stored in a FITS file, with the pertinent data labels being `OBSVHELIO` (heliocentric relative RV from observed spectrum template matching, in units of km/s), and `OBSVRELERR` (RV error of relative of RV from observed spectrum template matching, also in units of km/s). Using the `astropy.io.fits` package, I converted the FITS data into a `pandas` dataframe. I removed `OBSVHELIO` values of greater than $|300|$ km/s, the rotation velocity of the Milky Way, in order to eliminate erroneous measurements. Similarly, I removed all null values for `OBSVRELERR`. I then identified the mode number of visits per star on each plate, and selected for plates with a mode visit count of 10 or greater. From this, I found 7 plates fitting these conditions. I chose to work with plate 9290, which had the highest mode visit count per target, at 16 visits. This gave me a sample of 4153 total visits. I created a new dataframe for all data from plate 9290, and removed all data for targets which were visited less than 16 times. My final sample was 3408 visits, or 16 visits per 213 stars. I stored this in a CSV file.



	Target ID	Plate ID	MJD	OBSVHELIO (km/s)	OBSVREL Error (km/s)
164071	apo25m.5226.150-08-RV.2M03252400+4614203	9290	57706	-86.5657	0.010245
164072	apo25m.5226.150-08-RV.2M03252400+4614203	9290	57732	-86.3834	0.00820577
164073	apo25m.5226.150-08-RV.2M03252400+4614203	9290	57734	-86.3869	0.00901516
164074	apo25m.5226.150-08-RV.2M03252400+4614203	9290	57735	-86.5328	0.0107417
164075	apo25m.5226.150-08-RV.2M03252400+4614203	9290	57760	-86.3666	0.0120355
...	...	...	...	...	...
192942	apo25m.5226.150-08-RV.2M03415658+4626067	9290	58068	-26.4901	0.0115105
192943	apo25m.5226.150-08-RV.2M03415658+4626067	9290	58085	-26.5953	0.0122043
192944	apo25m.5226.150-08-RV.2M03415658+4626067	9290	58087	-26.604	0.0110477
192945	apo25m.5226.150-08-RV.2M03415658+4626067	9290	58114	-26.5992	0.0211242
192946	apo25m.5226.150-08-RV.2M03415658+4626067	9290	58143	-26.4153	0.00839748

3408 rows × 5 columns

Figure 1: `pandas` dataframe for final sample of plate 9290 data.

Code for this section can be found here.

2.2 Calculating Average RV Scatter

My next step was to plot the average RV scatter over time. To do this, I calculated the average RV scatter using the following equation:

$$\Delta RV_j = RV_j - \langle RV \rangle \tag{1}$$

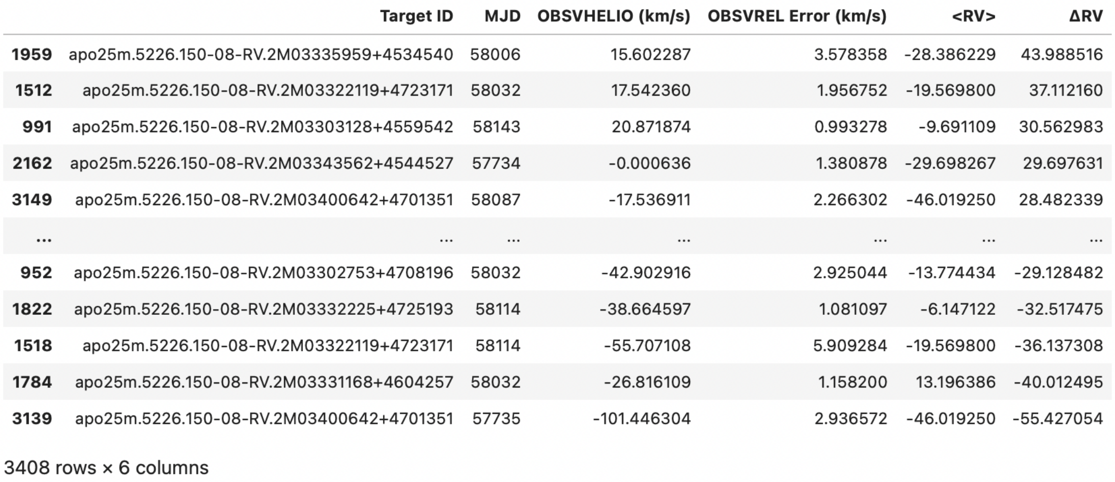
where,

$$\langle RV \rangle = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

and,

$$w_i = \frac{1}{\sigma_i^2} \tag{3}$$

Here, $_j$ denotes the epoch, ΔRV_j is the weighted average RV scatter, RV_j is the j^{th} value of `OBSVHELIO`, $\langle RV \rangle$ is the weighted average of all `OBSVHELIO`, w_i is the i^{th} value of the weight, x_i is the i^{th} value of ΔRV_j, and σ_i is the i^{th} value of `OBSVRELERR`.



	Target ID	MJD	OBSVHELIO (km/s)	OBSVREL Error (km/s)	<RV>	ΔRV
1959	apo25m.5226.150-08-RV.2M03335959+4534540	58006	15.602287	3.578358	-28.386229	43.988516
1512	apo25m.5226.150-08-RV.2M03322119+4723171	58032	17.542360	1.956752	-19.569800	37.112160
991	apo25m.5226.150-08-RV.2M03303128+4559542	58143	20.871874	0.993278	-9.691109	30.562983
2162	apo25m.5226.150-08-RV.2M03343562+4544527	57734	-0.000636	1.380878	-29.698267	29.697631
3149	apo25m.5226.150-08-RV.2M03400642+4701351	58087	-17.536911	2.266302	-46.019250	28.482339
...	...	...	...	...	...	...
952	apo25m.5226.150-08-RV.2M03302753+4708196	58032	-42.902916	2.925044	-13.774434	-29.128482
1822	apo25m.5226.150-08-RV.2M03332225+4725193	58114	-38.664597	1.081097	-6.147122	-32.517475
1518	apo25m.5226.150-08-RV.2M03322119+4723171	58114	-55.707108	5.909284	-19.569800	-36.137308
1784	apo25m.5226.150-08-RV.2M03331168+4604257	58032	-26.816109	1.158200	13.196386	-40.012495
3139	apo25m.5226.150-08-RV.2M03400642+4701351	57735	-101.446304	2.936572	-46.019250	-55.427054

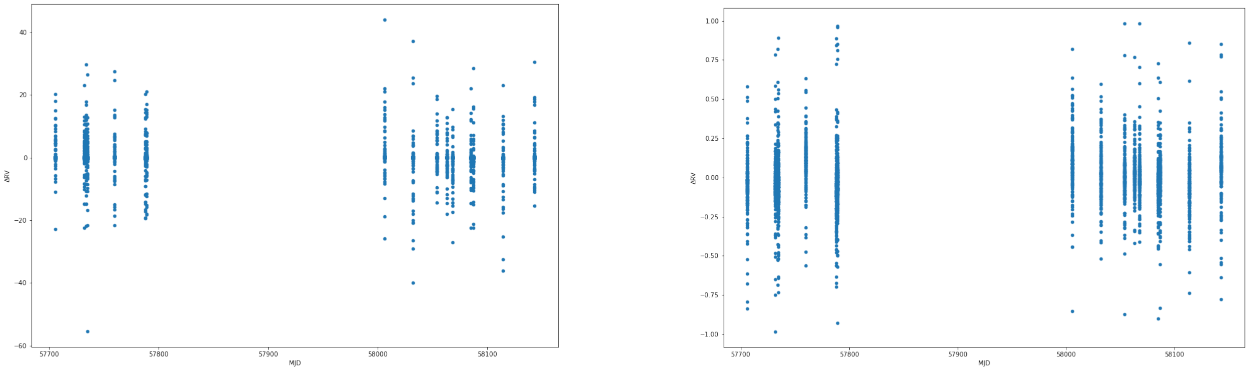
3408 rows × 6 columns

Figure 2: `pandas` dataframe for plate 9290 data, with added calculations for $\langle RV \rangle$ $[km/s]$ and ΔRV_j $[km/s]$. Sorted by descending order of ΔRV_j values.

Code for this section can be found here.

3 Results and Analysis

I plotted ΔRV_j $[km/s]$ versus MJD in order to determine if an overall trend in ΔRV_j could be identified by eye. From these initial plots, I found that the data points seemed to be more populous closer towards 0 ΔRV_j (see Figure 3). However, further refinement of the data was needed to identify the potential trend.



(a) Initial plot

(b) Plot constrained to $\Delta RV_j < |1|$ km/s

Figure 3: Plots of ΔRV_j $[km/s]$ versus MJD

In order to do this, I calculated the weighted average of all ΔRV_j data per each MJD, to obtain a single data point for each of the 16 nights of observation contained on plate 9290. I plotted this new weighted average against the MJD. This showed some clustering between $|50 - 100|$ m/s, which seems promising in terms of identifying an overall trend in average RV scatter (see Figure 4). However, I expected an average RV scatter on the order of 1-10 m/s, which is less one order of magnitude than my result.

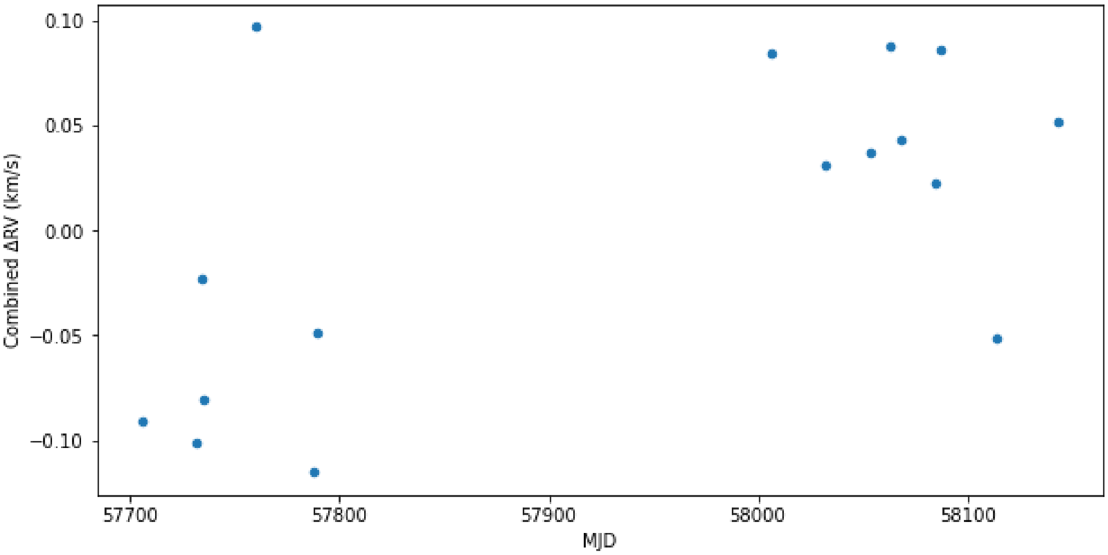


Figure 4: Plot of weighted average of ΔRV_j per MJD versus MJD.

4 Conclusion

My immediate next steps are to identify potential reasons for the larger than expected average RV scatter. They are as follows:

1. Calculate $\sigma_{\Delta RV_j}$, the standard error of ΔRV_j, as well as $\sigma_{\Delta RV_c}$, the standard error of the combined ΔRV_j
2. Plot statistical error bars for $\sigma_{\Delta RV_j}$ and $\sigma_{\Delta RV_c}$ on their respective graphs
3. Fit line to free parameters and check gradient for anything interesting
4. Check distribution of all data points used to create ΔRV_j; if not well-behaved, then try looking at flags listed on SDSS site to remove for other errors
5. Repeat this semester's process for all other plates; combining plots for all plates will fill in the gaps in MJD

After gaining a better picture of what is contributing to the higher than expected combined ΔRV_j, I can begin to consider additional systematic factors. For example, I can look at star distance from the centre of each plate, or sky elevation in relation to APOGEE's location. Currently, the dataset is 2-dimensional, rendering it less suitable for PCA. By increasing the number of dimensions in the data, PCA will be a more appropriate tool for identifying potential systematic errors.

References

Deshpande et al. 2013, The SDSS-II APOGEE Radial Velocity Survey of M Dwarfs I: Description of Survey and Science Goals

Price-Whelan et al. 2018, Binary companions of evolved stars in APOGEE DR14: Search method and catalog of 5,000 companions

Troup et al. 2016, Companions to APOGEE Stars I: A Milky Way-Spanning Catalog of Stellar and Substellar Companion Candidates and Their Diverse Hosts