

Applications of Data Science to Exoplanets

Coursework Submission

A REPORT PRESENTED
BY
STEVEN DILLMANN

Departments
Department of Applied Mathematics and Theoretical Physics
Department of Physics (Cavendish Laboratory)
Institute of Astronomy

Degree
MPhil Data Intensive Science

Module
Applictions of Data Science to Exoplanets

Supervision
Dr. Vinesh Maguire Rajpaul



ST JOHN'S COLLEGE
UNIVERSITY OF CAMBRIDGE
21ST JUNE 2024

List of Figures

1	Hubble Space Telescope high-resolution image of the target system containing three stars. Component A is the host of the planet candidates. Their apparent magnitudes, converted to the TESS bandpass are: A 8.88 ± 0.02 and BC 8.80 ± 0.02 (combined).	1
2	<i>Top Panel:</i> TESS lightcurve of the target system from more than 31,000 epochs. <i>Middle Panel:</i> Zoom in on the first part of the lightcurve. <i>Bottom Panel:</i> Zoom in on the last part of the lightcurve. The gaps in TESS lightcurves are often due to specific mission operations or observational constraints.	2
3	Lomb-Scargle periodogram of the lightcurve (black) and period corresponding to the maximum power (magenta square).	2
4	<i>Top Panel:</i> Full TESS lightcurve (black) with $> 3\sigma$ outliers (magenta) removed. <i>Middle Panel:</i> Zoom in on the first part of the lightcurve. <i>Bottom Panel:</i> Zoom in on the last part of the lightcurve.	3
5	<i>Top Panel:</i> Last part of the lightcurve after outlier removal (black) including the first 5 strongest sine model fits (green, orange, magenta, turquoise, purple). <i>Bottom Panels:</i> Evolution of the last part of lightcurve (black) during the first 5 sine fitting iterations (green, orange, magenta, turquoise, purple).	4
6	<i>Top Panel:</i> Full TESS lightcurve (black) with the 50 strongest sinusoidal trends and systematic variations (red) removed. <i>Middle Panel:</i> Zoom in on the first part of the lightcurve. <i>Bottom Panel:</i> Zoom in on the last part of the lightcurve.	5
7	The TLS power spectrum per period of the lightcurve (black) and period corresponding to the maximum power (red square) for two different period ranges.	6
8	<i>Top Panel:</i> Full TESS lightcurve (black) with the best-fit transit model (red). <i>Middle Panel:</i> Zoom in on the first part of the lightcurve. <i>Bottom Panel:</i> Zoom in on the last part of the lightcurve.	6
9	<i>Top Panel:</i> Phase-folded lightcurve (black) and transit model (red), where the phase is normalised between 0 and 1 such that the mid-transit time is placed at 0.5. <i>Bottom Panel:</i> Zoom in on the transit. The phase-folded model curve is not an exact representations of the model used during the search - they are intended for visualization purposes only as explained in Heller et al. (2024).	7
10	<i>Top Panel:</i> Full TESS lightcurve after masking the data (black) with the transit mask (red). <i>Middle Panel:</i> Zoom in on the first part of the lightcurve. <i>Bottom Panel:</i> Zoom in on the last part of the lightcurve.	8
11	The TLS power spectrum per period of the masked lightcurve (black) and period corresponding to the maximum power (blue square) for two different period ranges.	9
12	<i>Top Panel:</i> Full TESS lightcurve (black) with the first (red) and second (blue) TLS model. <i>Middle Panel:</i> Zoom in on the first part of the lightcurve. <i>Bottom Panel:</i> Zoom in on the last part of the lightcurve.	9
13	<i>Top Panel:</i> Phase-folded lightcurve (black) and second transit model (blue). <i>Bottom Panel:</i> Zoom in on the transit. The phase-folded model curve is for visualization purposes only (Heller et al., 2024).	10
14	Doppler time-series data of CB 01223. The data points on the left half of the plot are from Instrument 1 and the data points on the right half of the plot are from Instruments 2, 3 and 4. <i>Top Panel:</i> RV time-series (red). <i>Middle Panel:</i> FWHM time-series (blue). <i>Bottom Panel:</i> BS time-series (green).	13
15	Lomb-Scargle periodograms with the power corresponding to a 1.0% false alarm probability level highlighted with a horizontal line. The strongest period is highlighted with a circle. <i>Top Panel:</i> Periodogram for the RV data (gold-red scheme). <i>Middle Panel:</i> Periodogram for the FWHM data (silver-blue scheme). <i>Bottom Panel:</i> Periodogram for the BS data (bronze-green scheme).	14

16	Scatter plots of the RV against stellar activity indicators and a linear model fit highlighted with a linear line. The residuals are displayed below the scatter plots. <i>Left Panel:</i> Scatter plots and fits for RV-FWHM (silver-blue scheme). <i>Right Panel:</i> Scatter plots and fits for RV-BS (bronze-green scheme).	15
17	Correlation matrix between RV, FWHM and BS.	15
18	Exploratory long-term trend fitting of the RV, FWHM and BS data with polynomials.	16
19	Trace plot of the nested sampling results for the stellar activity only model. The purple lines are the medians of the unweighted sample posterior distributions. The golden lines represent the 95% CI.	20
20	Corner plot of the posterior parameter distributions for the stellar activity only model. The purple lines are the medians of the weighted sample posterior distributions. The parameter errors represent the 68% CI.	21
21	Multidimensional GP regression output with the best-fit parameters from the stellar activity only model.	22
22	Trace plot of the nested sampling results for the 1 planet model (circular orbit).	23
23	Corner plot of the posterior parameter distributions for the 1 planet model (circular orbit).	24
24	Multidimensional GP regression and simplified Keplerian model output with the best-fit parameters from the 1 planet model (circular orbit).	25
25	Trace plot of the nested sampling results for the 1 planet model (eccentric orbit).	26
26	Corner plot of the posterior parameter distributions for the 1 planet model (eccentric orbit).	27
27	Multidimensional GP regression and full Keplerian model output with the best-fit parameters from the 1 planet model (eccentric orbit).	28
28	Trace plot of the nested sampling results for the 2 planet model (circular orbits).	29
29	Corner plot of the posterior parameter distributions for the 2 planet model (circular orbits).	30
30	Multidimensional GP regression and full Keplerian model output with the best-fit parameters from the 2 planet model (circular orbits).	31
31	Trace plot of the nested sampling results for the 2 planet model (eccentric orbits).	32
32	Corner plot of the posterior parameter distributions for the 2 planet model (eccentric orbits).	33
33	Multidimensional GP regression and full Keplerian model output with the best-fit parameters from the 2 planet model (eccentric orbits).	34

List of Tables

1	Known parameters of star component A derived from K-band photometry, with Gaia parallax.	11
2	Physical and orbital properties of the detected exoplanets.	12
3	Strongest periods in the periodograms and their associated FAPs.	14
4	Linear model fits and correlation of RV-FWHM and RV-BS.	15
5	Prior probability distributions for the stellar activity model parameters.	18
6	Prior probability distributions for the planetary signal model parameters.	19
7	Parameter estimation results for Model I (0 planets, stellar activity GP only).	35
8	Parameter estimation results for Models II and III (1 planet).	35
9	Parameter estimation results for Models IV and V (2 planets).	35
10	Bayesian evidences for the tested models.	36
11	Physical and orbital properties of the detected exoplanet.	36

List of Listings

Contents

List of Figures	iii
List of Tables	iv
List of Listings	v
1 Transit Method	1
1.1 Analyse the lightcurve to identify the strongest period	1
1.2 Filter the lightcurve to search for any planet candidate(s)	1
1.2.1 Outlier Removal	2
1.2.2 Iterative Sine Fitting Removal	3
1.2.3 Detrending	4
1.2.4 Planet Candidate 1	5
1.2.5 Planet Candidate 2	8
1.3 Derive estimates of the properties of the planetary systems	11
1.3.1 Property Estimation with the Transit Method	11
1.3.2 Results & Summary	12
2 Radial Velocity Method	13
2.1 Q1 & Q2: Periodic Variability and Long-term Trends	13
2.1.1 Data	13
2.1.2 Exploratory Analysis	13
2.2 Q3 & Q4: Exoplanet Detection and Characterisation	17
2.2.1 Stellar Activity Modelling	17
2.2.2 Planetary Signal Modelling	18
2.2.3 Model I: No Planets, Stellar Activity GP only	19
2.2.4 Models II & III: One Planet	19
2.2.5 Models IV & V: Two Planets	20
2.2.6 Posterior Summaries	35
2.2.7 Model Comparison	36
2.2.8 Property Estimation with the Radial Velocity Method	36
2.2.9 Results & Summary	36
A Use of auto-generation tools	39
A.1 Co-Pilot	39
A.2 ChatGPT	39
A.2.1 Prompt 1	39

Total L^AT_EX Word Count: 2995/3000 (excluding Figures, Tables, Bibliography and Appendix)

1 Transit Method

In this chapter, we analyse the lightcurve of a TESS target and use the transit method to identify potential exoplanet candidates. The target system is shown in Figure 1.

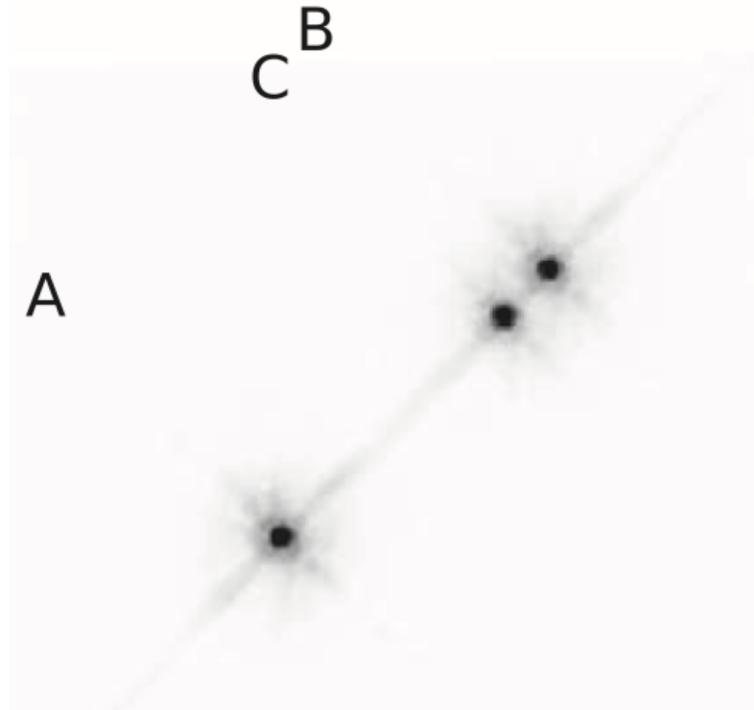


Figure 1: Hubble Space Telescope high-resolution image of the target system containing three stars. Component A is the host of the planet candidates. Their apparent magnitudes, converted to the TESS bandpass are: A 8.88 ± 0.02 and BC 8.80 ± 0.02 (combined).

1.1 Analyse the lightcurve to identify the strongest period

Figure 2 shows the raw lightcurve of the target, created with the `lightkurve` package (Cardoso et al., 2018). Stellar lightcurves often exhibit oscillations in their brightness over time associated with stellar activity. Magnetic activity on stellar surfaces causes starspots (Strassmeier, 2009), which are cooler areas and therefore darker than the surrounding surface. As the star rotates, these spots move in and out of view. This effect is called rotation modulation (Berdyugina, 2005) and causes periodic dips in the star's brightness.

Lomb-Scargle periodogram

The periodic oscillations visible in the lightcurve are more effectively analysed in the frequency domain, using a Lomb-Scargle periodograms (Lomb, 1976; Scargle, 1982). A periodogram estimates the discrete Fourier transform of the time-domain lightcurve (VanderPlas, 2018). Figure 3 shows the Lomb-Scargle periodogram of the lightcurve, created with the `astropy` package (Robitaille et al., 2013). By identifying the maximum power in the periodogram, we obtain a strongest rotation period estimate of $P_{max} = 1.39837$ days. This period is consistent with the apparent sinusoidal trend in the lighcurve.

1.2 Filter the lightcurve to search for any planet candidate(s)

To identify potential exoplanet candidates within the target system, a series of data processing steps need to be employed.

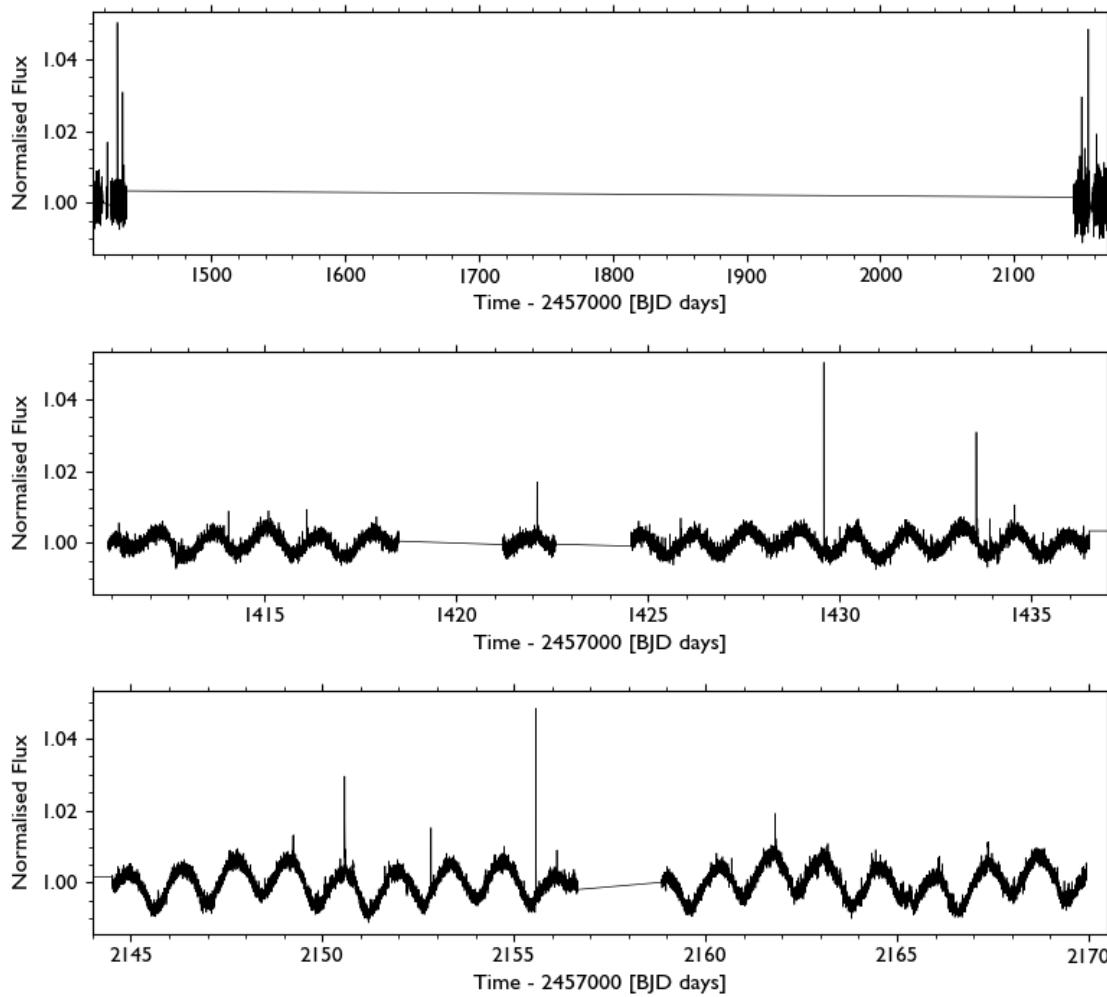


Figure 2: *Top Panel:* TESS lightcurve of the target system from more than 31,000 epochs. *Middle Panel:* Zoom in on the first part of the lightcurve. *Bottom Panel:* Zoom in on the last part of the lightcurve. The gaps in TESS lightcurves are often due to specific mission operations or observational constraints.

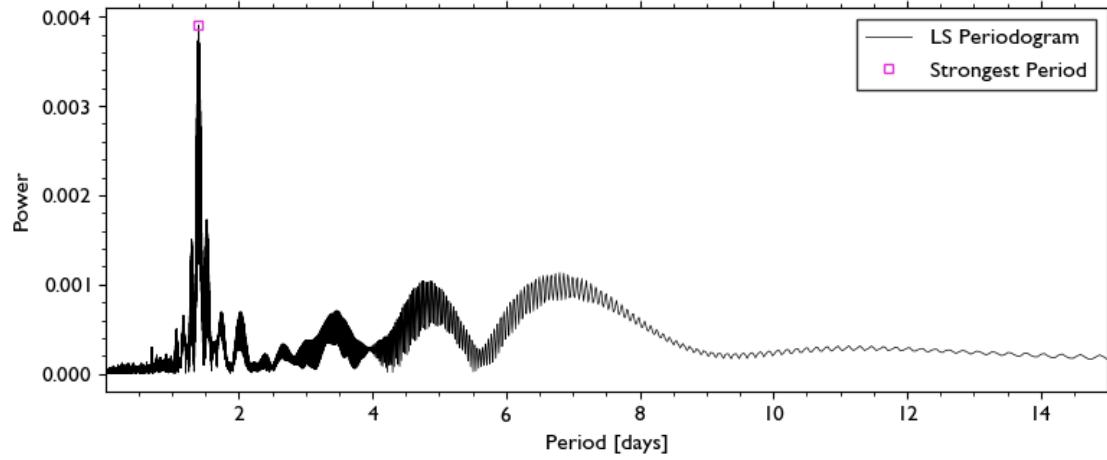


Figure 3: Lomb-Scargle periodogram of the lightcurve (black) and period corresponding to the maximum power (magenta square).

1.2.1 Outlier Removal

In addition to rotational modulation with starspots, the lightcurve also contains stellar flares (Pettersen, 1989) that appear as outliers in Figure 2. We handle outliers in the lighcurve by removing

data points that deviate more than 3 standard deviations from the mean of the data. Figure 4 shows the removed outliers and the cleaned lightcurve.

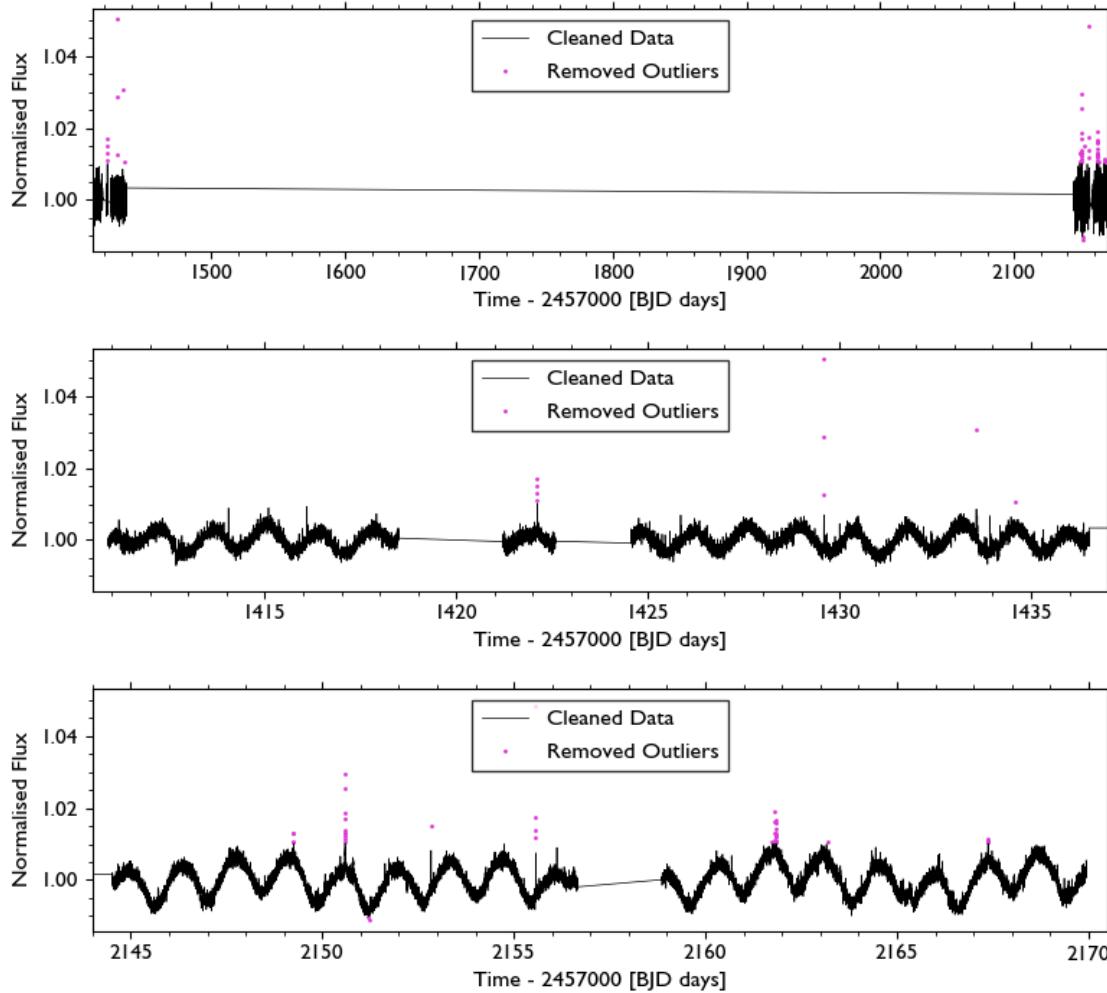


Figure 4: *Top Panel:* Full TESS lightcurve (black) with $>3\sigma$ outliers (magenta) removed. *Middle Panel:* Zoom in on the first part of the lightcurve. *Bottom Panel:* Zoom in on the last part of the lightcurve.

1.2.2 Iterative Sine Fitting Removal

The effects of rotational modulation in a spotty star often complicate the detection of exoplanets with the transit method because their periodic signals can obscure signals from planetary transits. Other sources of variability are asteroseismic oscillations (Brown and Gilliland, 1994) and magnetic cycles (Wilson, 1978). Here, we use a technique called iterative sine fitting to model and remove sinusoidal rotation signals. This involves iteratively fitting sinusoidal models to the lightcurve and removing the most significant periodic components in the following way:

1. Create a Lomb-Scargle periodogram of the lightcurve and find the period of maximum power.
2. Define a Lomb-Scargle model with the strongest period.
3. Divide the lightcurve by the model to remove the signal with the strongest period.
4. Repeat this process with the new lightcurve for a given number of iterations.

We run the process for 50 iterations. The number of iterations was adjusted iteratively by maximising and retaining as much of the signal-to-noise ratio (SNR) of the first and second planetary signals in Section 1.2.4 and Section 1.2.5. Figure 5 visualises the first 5 iterations of the process.

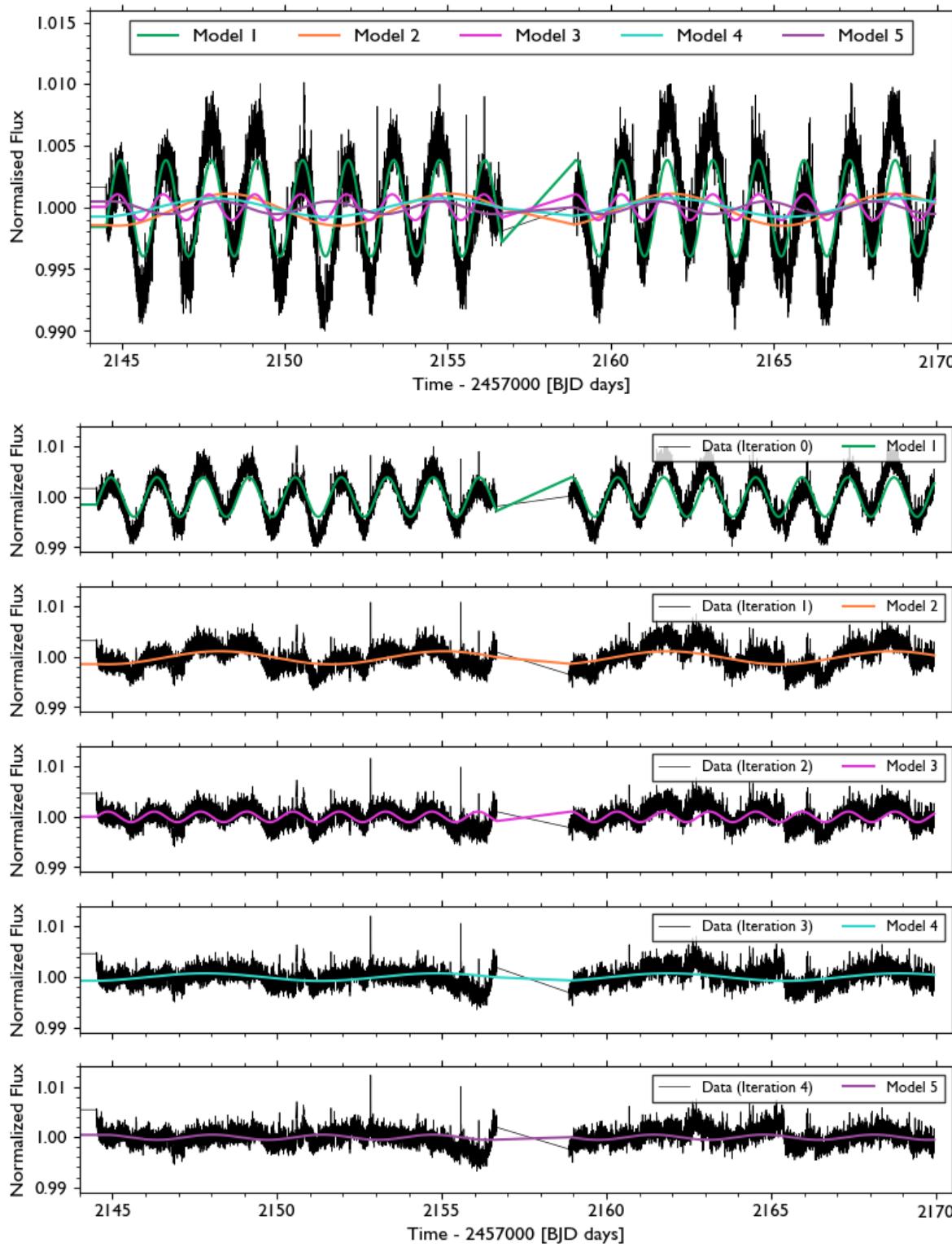


Figure 5: *Top Panel:* Last part of the lightcurve after outlier removal (black) including the first 5 strongest sine model fits (green, orange, magenta, turquoise, purple). *Bottom Panels:* Evolution of the last part of lightcurve (black) during the first 5 sine fitting iterations (green, orange, magenta, turquoise, purple).

1.2.3 Detrending

Long-term systematic trends and instrumental noise can also complicate the analysis. We use flattening with a window length of 500 days to detrend our data. This value was chosen as it provides strong SNRs for the planetary signals. Figure 6 shows the residual data after the detrending and iterative sine fitting process. The residuals shows no or minimal sinusoidal trends or outliers.

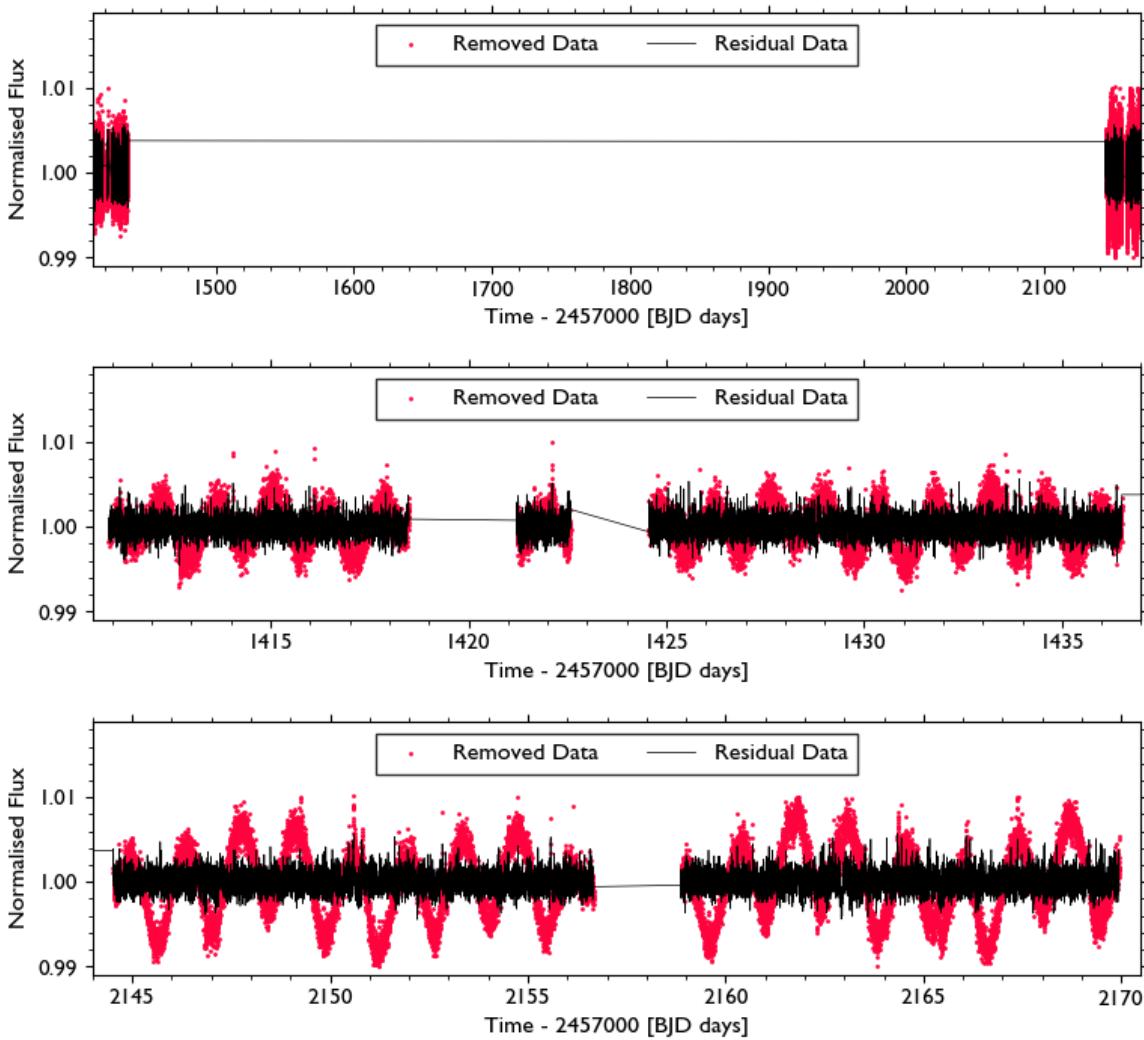


Figure 6: *Top Panel:* Full TESS lightcurve (black) with the 50 strongest sinusoidal trends and systematic variations (red) removed. *Middle Panel:* Zoom in on the first part of the lightcurve. *Bottom Panel:* Zoom in on the last part of the lightcurve.

1.2.4 Planet Candidate 1

We now use the filtered dataset to search for planetary candidates with the transit method.

Transit Least Squares

We use the Transit Least Squares (TLS) algorithm in the `transitleastsquares` package ([Hippke and Heller, 2019a,b](#)) to search for transit candidates. The advantage of TLS over the commonly used Box Least Squares (BLS) algorithm ([Kovács et al., 2002](#)) is that TLS takes into account stellar limb darkening and planetary ingress and egress. TLS has a higher detection efficiency and a similar false alarm rate compared to BLS ([Heller et al., 2019](#)). Figure 7 shows the median-smoothed power spectrum per period as defined in [Hippke and Heller \(2019a\)](#). The best-fit period of the first planetary candidate is found to be $P_1 = 5.35871 \pm 0.00028$ days and the best-fit transit depth is $\delta_1 = 1946 \pm 232$ ppm with a SNR of 6.5. The mid-transit time of the first transit is 1412.713 days (BJD – 2457000) and we estimate the transit duration to be $\tau_1 = 0.05339$ days. Further results are discussed in Section 1.3. Figure 8 shows the lightcurve and the best-fit transit model.

Phase Folding

To visually confirm the planetary transit, we use phase-folding. With the expected period obtained from the TLS fit, we can fold the lightcurve at this period and make the transit signal more pronounced. Figure 9 shows the folded lightcurve and the transit model with a clear dip.

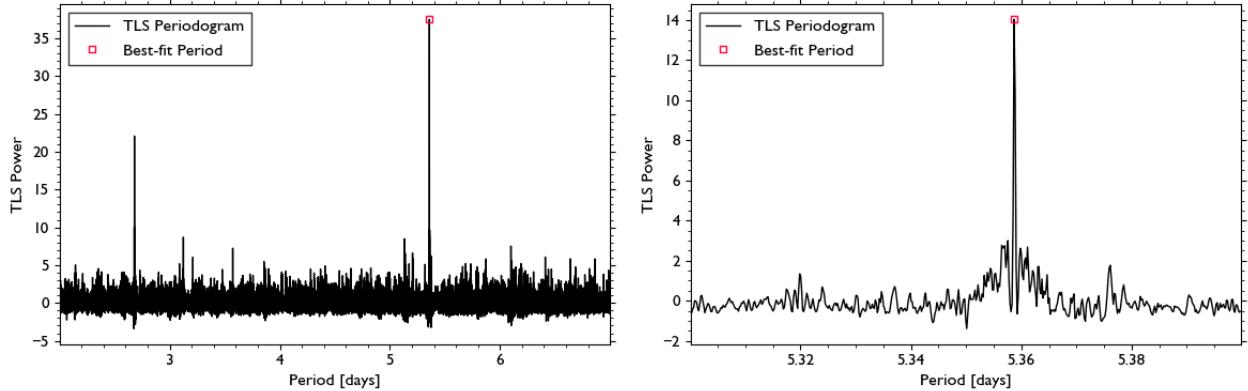


Figure 7: The TLS power spectrum per period of the lightcurve (black) and period corresponding to the maximum power (red square) for two different period ranges.

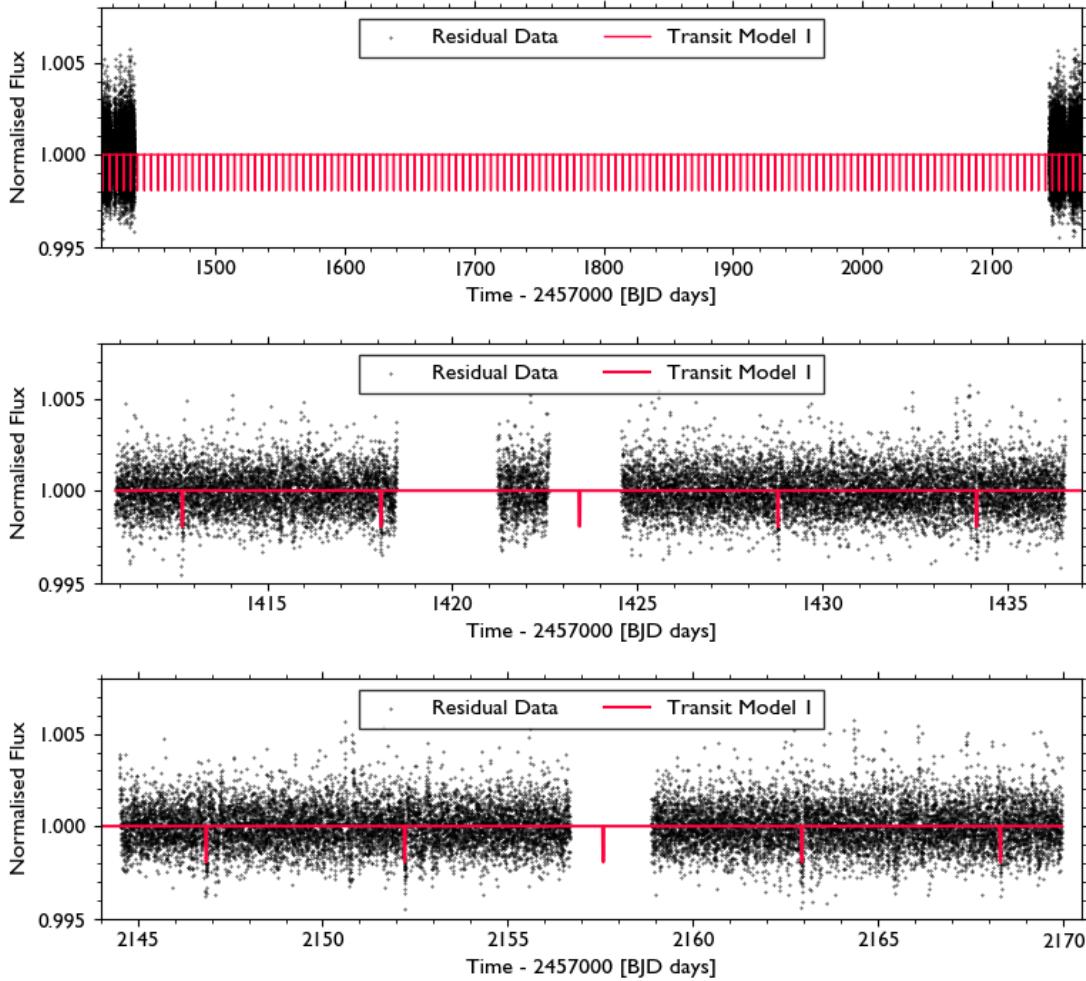


Figure 8: *Top Panel:* Full TESS lightcurve (black) with the best-fit transit model (red). *Middle Panel:* Zoom in on the first part of the lightcurve. *Bottom Panel:* Zoom in on the last part of the lightcurve.

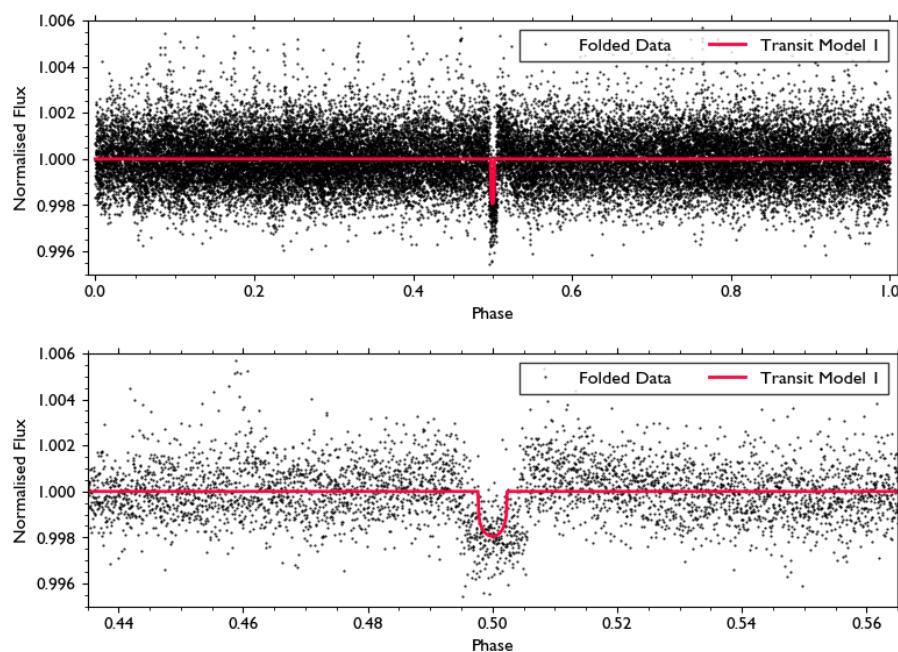


Figure 9: *Top Panel:* Phase-folded lightcurve (black) and transit model (red), where the phase is normalised between 0 and 1 such that the mid-transit time is placed at 0.5. *Bottom Panel:* Zoom in on the transit. The phase-folded model curve is not an exact representation of the model used during the search - they are intended for visualization purposes only as explained in Heller et al. (2024).

1.2.5 Planet Candidate 2

Now that we have identified the first planetary candidate, we can search for additional transits.

Transit Mask

In order to more effectively search for a second transit, we can remove the previously identified signal by creating a transit mask. Masking the data associated with the dominant transit period removes its dominance in the TLS power spectrum, allowing us to uncover additional, less prominent transits. Figure 10 shows the masked data and transit mask.

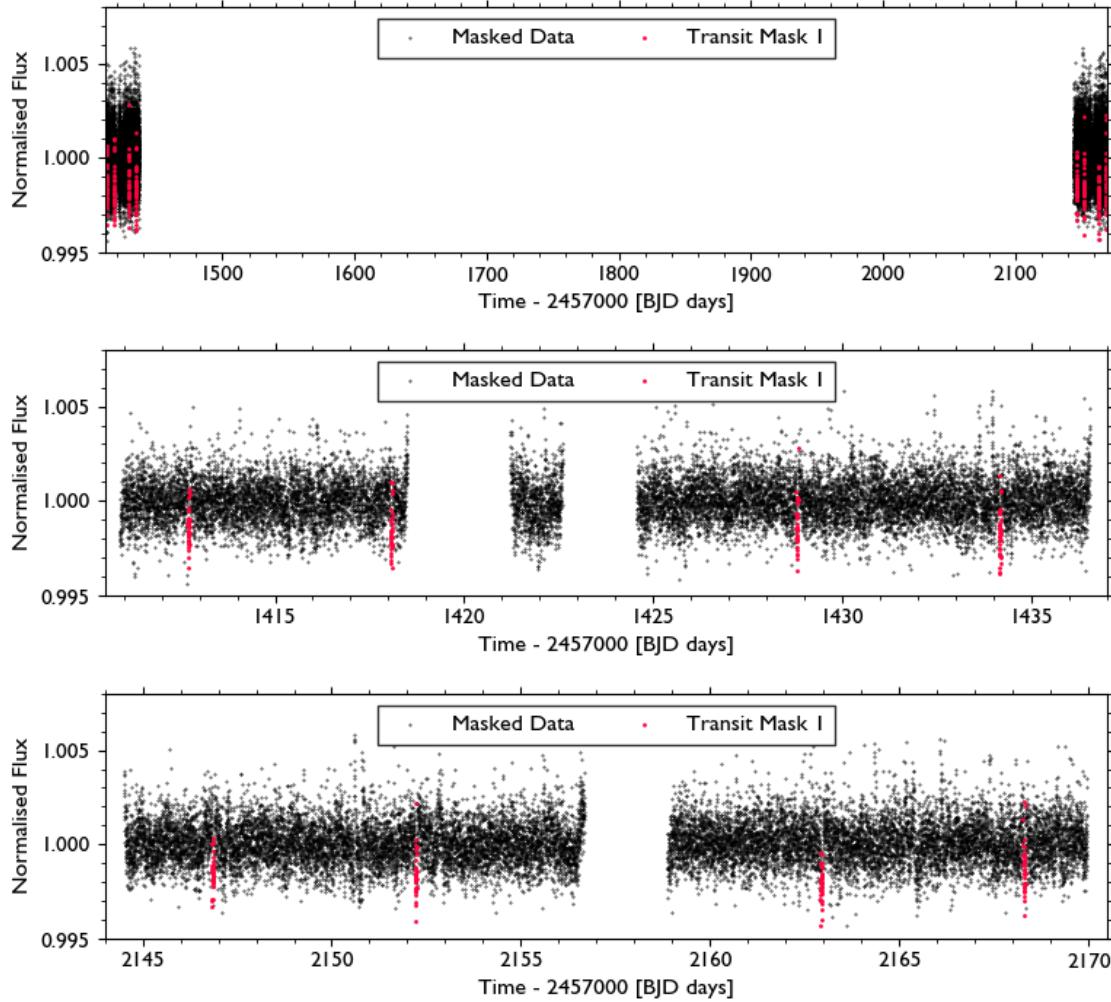


Figure 10: *Top Panel:* Full TESS lightcurve after masking the data (black) with the transit mask (red). *Middle Panel:* Zoom in on the first part of the lightcurve. *Bottom Panel:* Zoom in on the last part of the lightcurve.

Transit Least Squares

We repeat the same steps as for the first transit candidate and apply the TLS algorithm to search for a second planetary candidate in the lightcurve. Figure 11 shows the TLS power spectrum Hippke and Heller (2019a) for the masked lightcurve. We can see that the strong period from the first candidate is not present anymore, effectively unveiling an additional candidate. The best-fit period of the second planetary candidate is found to be $P_2 = 3.12392 \pm 0.00009$ days and the best-fit transit depth is $\delta_2 = 983 \pm 342$ ppm with a SNR of 3.8. The mid-transit time of the first transit is 1412.580 days (BJD – 2457000) and we estimate the transit duration to be $\tau_1 = 0.01971$ days. Further results are discussed in Section 1.3. Figure 12 shows the lightcurves and the best-fit models for both planetary candidates.

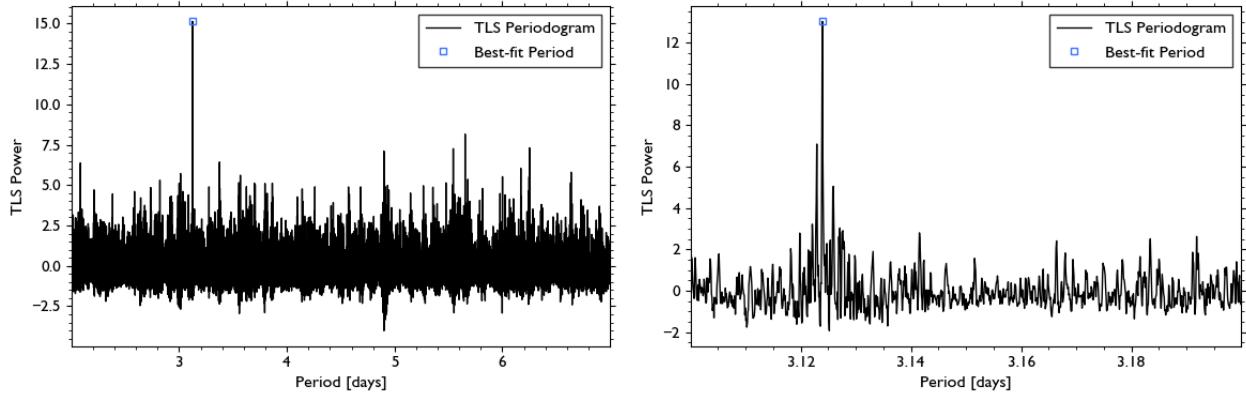


Figure 11: The TLS power spectrum per period of the masked lightcurve (black) and period corresponding to the maximum power (blue square) for two different period ranges.

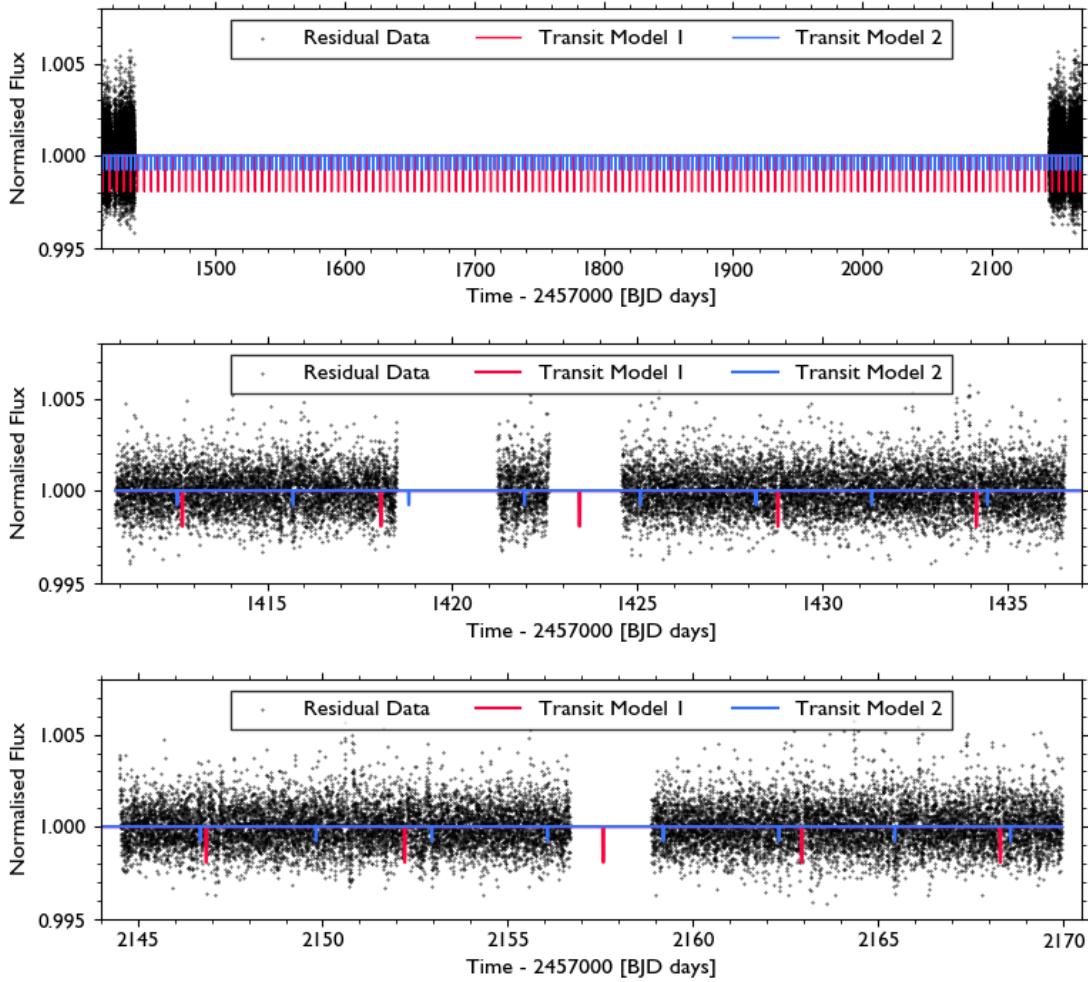


Figure 12: *Top Panel:* Full TESS lightcurve (black) with the first (red) and second (blue) TLS model. *Middle Panel:* Zoom in on the first part of the lightcurve. *Bottom Panel:* Zoom in on the last part of the lightcurve.

Phase Folding

We visually confirm the second planetary transit by phase-folding the lightcurve at the best-fit period from the obtained from the second TLS model. Figure 13 shows the folded lightcurve and the associated transit model with a less pronounced dip as compared to the first transit.

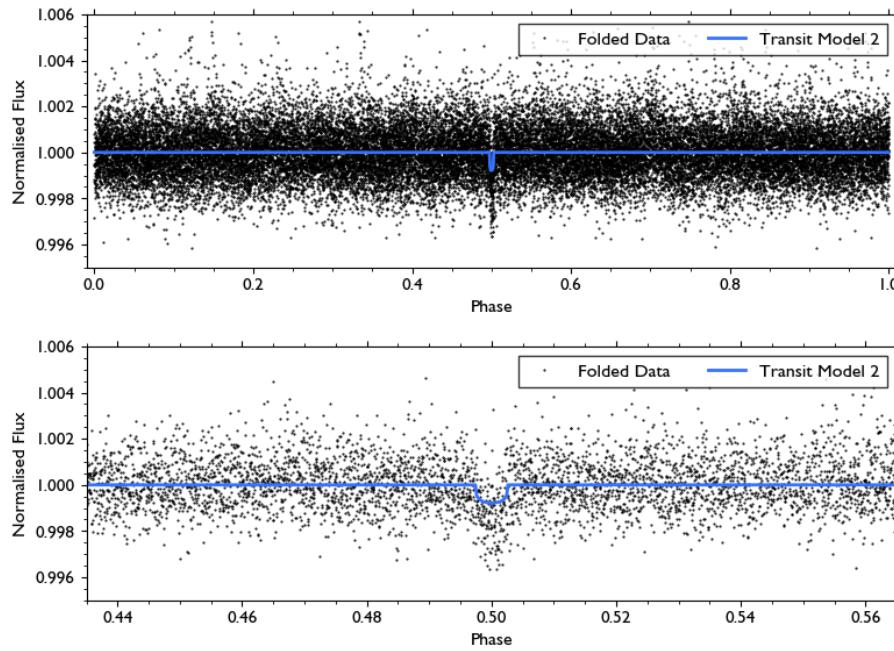


Figure 13: *Top Panel:* Phase-folded lightcurve (black) and second transit model (blue). *Bottom Panel:* Zoom in on the transit. The phase-folded model curve is for visualization purposes only ([Heller et al., 2024](#)).

1.3 Derive estimates of the properties of the planetary systems

Table 1 lists the known stellar parameters of component A in the target system, which is the host of any exoplanets.

Table 1: Known parameters of star component A derived from K-band photometry, with Gaia parallax.

Parameter	Value
Mass M_s	$0.257 \pm 0.014 M_{\odot}$
Radius R_s	$0.268 \pm 0.027 R_{\odot}$

1.3.1 Property Estimation with the Transit Method

We can estimate a number of properties of the planetary systems with the transit method.

Planetary radius

The transit depth $\delta = \Delta F/F$ is the fractional reduction of the stellar flux F due to the planetary transits. Assuming the star is a homogeneous disk with radius R_s and the planet is a black disk transiting in front of the star edge-on or nearly edge-on, the planetary radius R_p is found with:

$$\delta \approx \left(\frac{R_p}{R_s} \right)^2 \Rightarrow R_p = \sqrt{\delta} \cdot R_s \quad (1)$$

Semi-major axis

Using Kepler's Third Law and assuming that the planetary mass M_p is negligible compared to the stellar mass M_s , i.e. $M_p \ll M_s$, the orbital semi-major axis a is found with:

$$P^2 = \frac{4\pi^2 a^3}{G(M_s + M_p)} \approx \frac{4\pi^2 a^3}{GM_s} \Rightarrow a = \left(\frac{GM_s P^2}{4\pi^2} \right)^{1/3}, \quad (2)$$

where G is the gravitational constant and P is the orbital period of the exoplanet.

Inclination

For $a \gg R_s$, $\cos i \ll 1$ and using $\arcsin x \approx x$ for small x , the transit duration τ can be approximated as (Odunlade, 2010):

$$\tau = \frac{P}{\pi} \sin^{-1} \left(\frac{R_s \sqrt{(1 + R_p/R_s)^2 - (a/R_s \cos i)^2}}{\sin i} \right) \approx \frac{PR_s}{\pi a} \sqrt{\left(1 + \frac{R_p}{R_s}\right)^2 - \left(\frac{a}{R_s} \cos i\right)^2}, \quad (3)$$

where i is the orbital inclination. Defining the projected distance between the planet and star centers as the impact parameter $b = \frac{a}{R_s} \cos i$, we can find b with:

$$\tau \approx \frac{PR_s}{\pi a} \sqrt{\left(1 + \frac{R_p}{R_s}\right)^2 - b^2} \Rightarrow b = \sqrt{\left(1 + \frac{R_p}{R_s}\right)^2 - \left(\frac{\tau \pi a}{PR_s}\right)^2} \quad (4)$$

Finally, this allows us to find the orbital inclination i with:

$$b = \frac{a}{R_s} \cos i \Rightarrow i = \arccos \frac{bR_s}{a} \quad (5)$$

1.3.2 Results & Summary

The property estimates for both planetary candidates are shown in Table 2.

Table 2: Physical and orbital properties of the detected exoplanets.

Parameter	Planet 1	Planet 2
Period P [days]	5.35871 ± 0.00028	3.12392 ± 0.00014
Transit Depth δ [fraction]	0.00195 ± 0.00023	0.00098 ± 0.00034
Transit Duration τ [days]	0.05339	0.01971
Radius R_p [R_{\oplus}]	1.291 ± 0.151	0.918 ± 0.185
Semi-major axis a [AU]	0.03810 ± 0.00069	0.02659 ± 0.00082
Impact Parameter b	0.417 ± 0.001	0.941 ± 0.001
Inclination i [deg]	89.22 ± 0.08	87.47 ± 0.25

Discussion

The estimated properties are physically plausible. However, the transit depths and consequently the planetary radii may be underestimated, as the lightcurve also includes light from the other two stars in the host system. To account for this, one could incorporate a dilution factor D of the form:

$$D = \frac{f_A}{f_A + f_B + f_C}, \quad (6)$$

where the individual fluxes can be obtained from the given magnitudes m_A , m_B and m_C with:

$$f_X = f_0 \cdot 10^{-0.4m_X}, \quad (7)$$

where f_0 is some reference value.

Conclusion

We successfully identified two exoplanets in the host system using the lightcurve. The first planet has a period of 5.36 days and a radius of $1.29 R_{\oplus}$. The second planet has a period of 3.12 days and a radius of $0.92 R_{\oplus}$. To confirm the detection and improve the parameter estimates, one could leverage data from other facilities, combine the transit method with the RV method, and use more sophisticated parameter fitting techniques.

2 Radial Velocity Method

In this chapter, we analyse high-resolution Doppler spectrographic measurements of the K-dwarf star CB 01223 with the radial velocity (RV) method.

2.1 Q1 & Q2: Periodic Variability and Long-term Trends

2.1.1 Data

Figure 14 shows the time series for the RV data and two stellar activity indicators: (i) the Full Width at Half Maximum (FWHM) of the Cross-Correlation Function (CCF) and (ii) the bisector span (BS) of the CCF. Visual inspection of the data shows signs of periodic variability and a long-term rising trend of the RV data and stellar activity when comparing earlier measurements (Instrument 1) with more recent measurements (Instruments 2, 3 and 4).

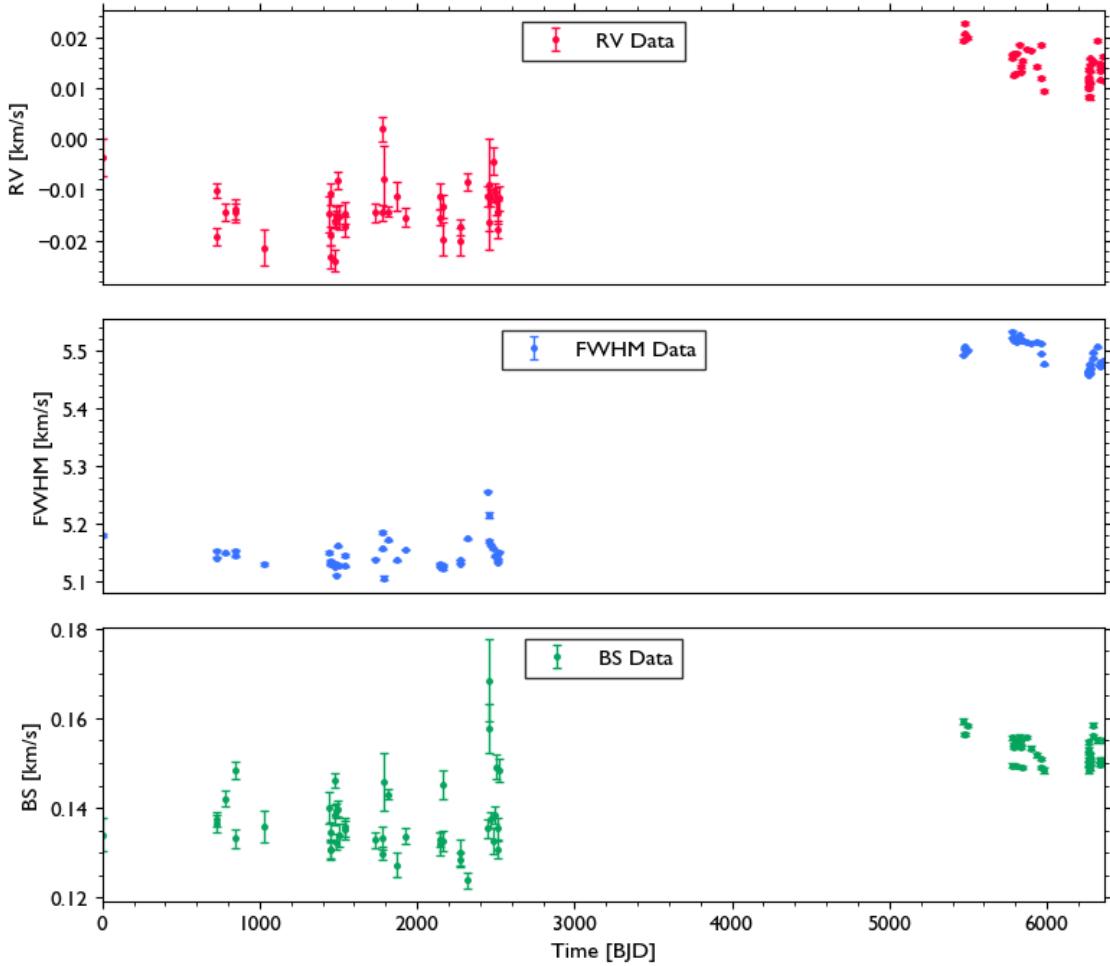


Figure 14: Doppler time-series data of CB 01223. The data points on the left half of the plot are from Instrument 1 and the data points on the right half of the plot are from Instruments 2, 3 and 4. *Top Panel:* RV time-series (red). *Middle Panel:* FWHM time-series (blue). *Bottom Panel:* BS time-series (green).

2.1.2 Exploratory Analysis

Monitoring the stellar activity indicators in addition to the RV measurements can help distinguish between RV variations due to stellar activity and RV variations due to an orbiting exoplanet. Note that any of the analysis in this section is purely exploratory and are not meant to substitute rigorous modelling of the data.

Periodic Variability

Changes in the FWHM and BS can suggest changes in the stellar activity affecting the RV measurements. The previously introduced Lomb-Scargle periodograms (Lomb, 1976; Scargle, 1982) can help detect periodic signals in this data. Figure 15 shows the periodograms for the RV, FWHM and BS data, created with astropy (Robitaille et al., 2013). We see a number of noisy peaks but also a few very significant peaks. The known stellar rotation period of $P_{rot} = 31 \pm 10$ days is clearly visible in the BS periodogram. We quantify the significance of the peaks in terms of a false alarm probability (FAP), which encodes the probability of measuring a peak of a given power (or higher) under the assumed condition that the data consists of Gaussian noise only. There are two clear peaks with $FAP < 1.0\%$ in the all periodograms simultaneously. The strongest period in all periodograms is found to be ≈ 0.9974 days and the corresponding FAPs listed in Table 3 suggest this signal to be highly significant.

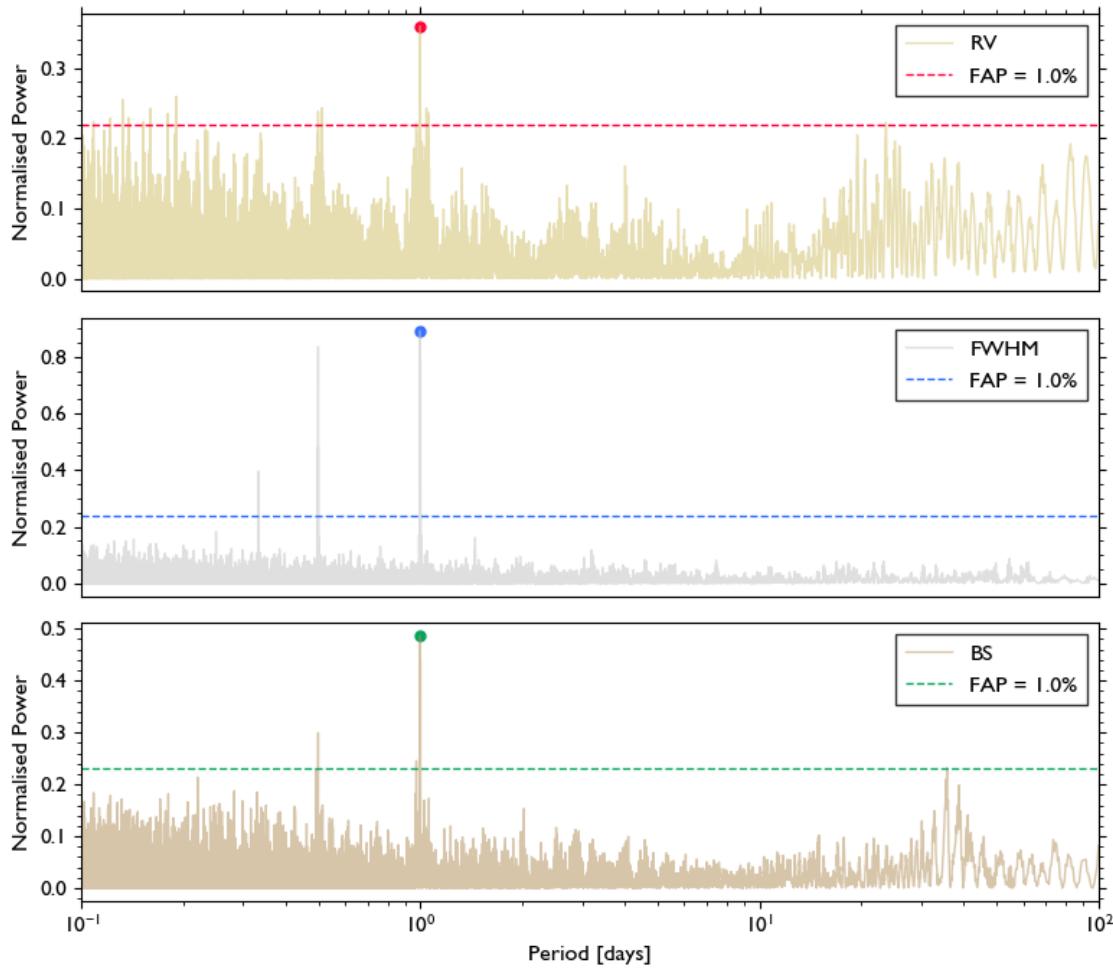


Figure 15: Lomb-Scargle periodograms with the power corresponding to a 1.0% false alarm probability level highlighted with a horizontal line. The strongest period is highlighted with a circle. *Top Panel:* Periodogram for the RV data (gold-red scheme). *Middle Panel:* Periodogram for the FWHM data (silver-blue scheme). *Bottom Panel:* Periodogram for the BS data (bronze-green scheme).

Table 3: Strongest periods in the periodograms and their associated FAPs.

	Period at max. power	FAP at max. power
RV	0.9974	$2.0 \cdot 10^{-5}$
FWHM	0.9974	$5.1 \cdot 10^{-35}$
BS	0.9974	$7.4 \cdot 10^{-10}$

We visualise the relation between the RV and stellar activity indicators with scatter plots, linear

model fits and the corresponding residuals in Figure 16. Moreover, we plot the correlation matrix between RV, FWHM and BS in Figure 17. Table 4 summarises the Pearson correlation coefficients (Pearson, 1895) and their corresponding p-values, as well as the linear model fit results. The correlation for RV-FWHM is 0.97, and 0.77 for RV-BS. The linear model fit is better for RV-FWHM based on the distribution of the residuals and performance metrics like the R^2 value and the F-statistic.

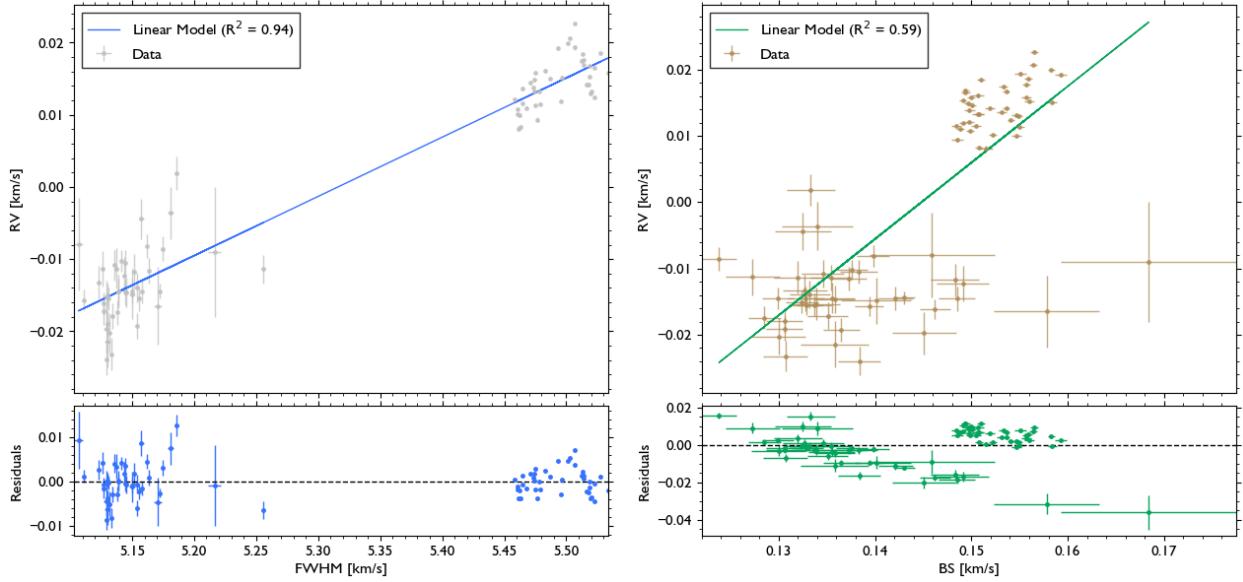


Figure 16: Scatter plots of the RV against stellar activity indicators and a linear model fit highlighted with a linear line. The residuals are displayed below the scatter plots. *Left Panel:* Scatter plots and fits for RV-FWHM (silver-blue scheme). *Right Panel:* Scatter plots and fits for RV-BS (bronze-green scheme).

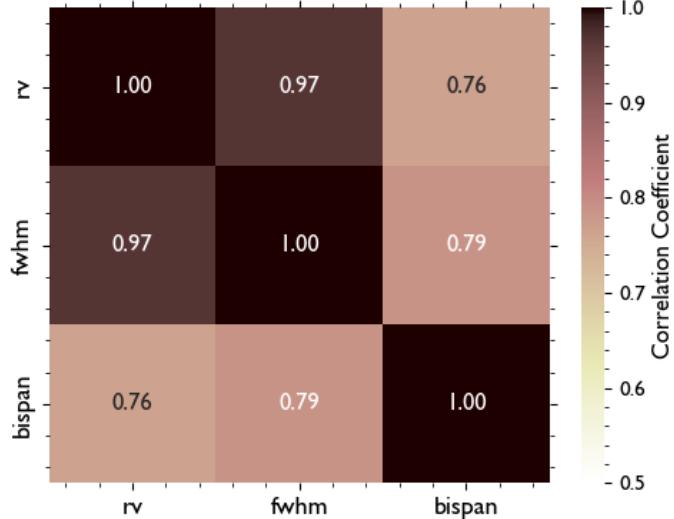


Figure 17: Correlation matrix between RV, FWHM and BS.

Table 4: Linear model fits and correlation of RV-FWHM and RV-BS.

	Linear Model Slope	R^2	F-statistic	Correlation (p-value)
RV-FWHM	0.08	0.94	1199	$0.97 (1.0 \cdot 10^{-50})$
RV-BS	1.15	0.59	116	$0.77 (2.5 \cdot 10^{-17})$

There is a relatively strong, positive and significant correlation between both stellar activity indicators and the RV measurements. And simple linear models show a clear positive relation ship between them. All of these exploratory findings suggest that the RV data is highly affected by

(quasi-)periodic variability due to stellar activity. Stellar (quasi-)variability may be due to (i) oscillation and granulation (Michel et al., 2008; Mathur et al., 2011) with periods on the scale of minutes and hours, (ii) rotationally modulated activity (Dumusque et al., 2011; Boisse et al., 2012) with periods on the scale of days to weeks, and (iii) magnetic cycles (Da Silva et al., 2012) on the scale of years. Additional possible origins of (quasi-)periodic variability are due to sampling effects (e.g. inadequate sampling rate, sparse sampling etc.) and instrumental effects (e.g. instrumental drift (Pepe et al., 2014)). However, the RV periodogram may also contain peaks that are associated with planetary signals that are hidden within stronger signals from stellar activity. The question whether or not the RV data exhibits evidence of planetary signals is addressed in Section 2.2.

Long-term Trends

We assess the apparent long-term trend by fitting the RV and stellar activity indicator time series with polynomial models, as shown in Figure 18. The quality of the fits is measured with the adjusted R^2 value to account for the different complexities of the model. We achieve better fits with models that exhibit long-term cyclic features in the RV and stellar activity indicator data. This may suggest that the long-term trend in the data may be due to long magnetic cycles.

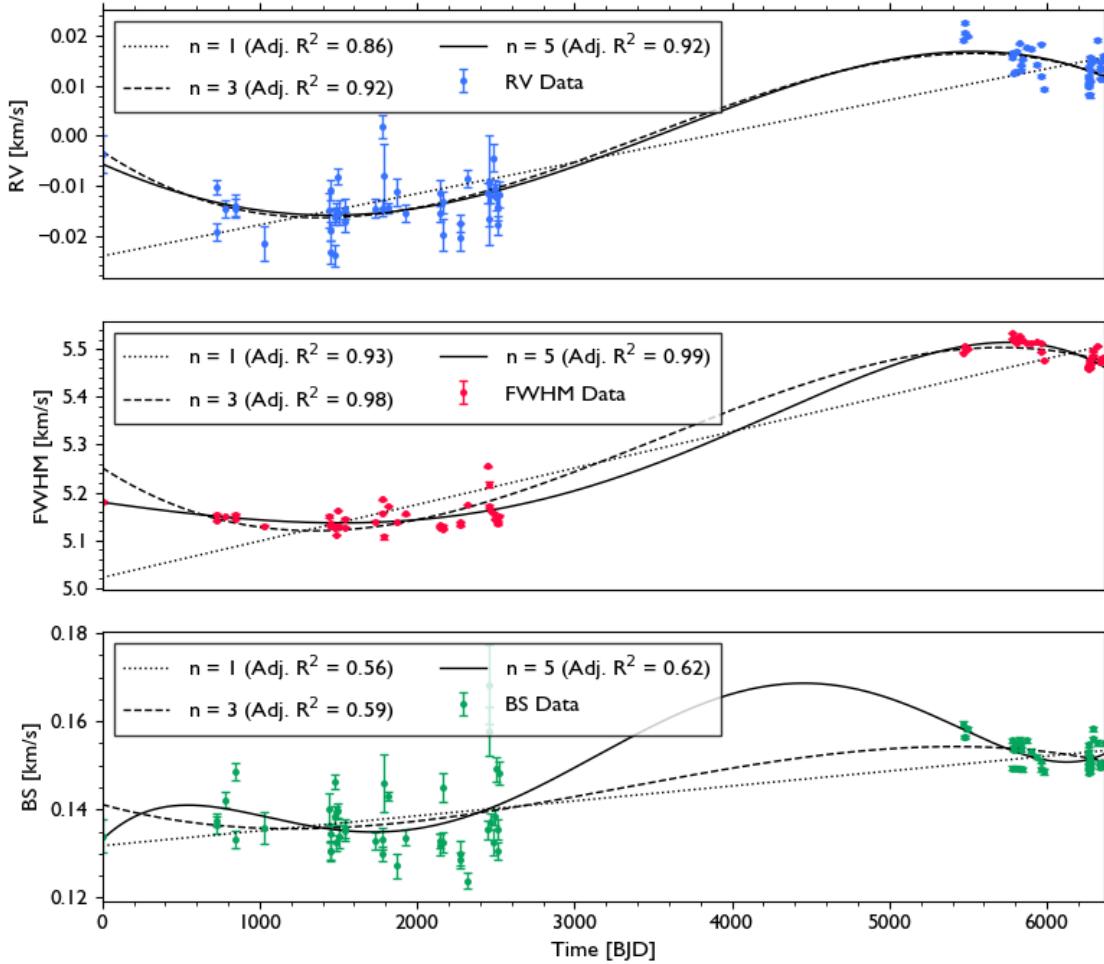


Figure 18: Exploratory long-term trend fitting of the RV, FWHM and BS data with polynomials.

2.2 Q3 & Q4: Exoplanet Detection and Characterisation

In this section, we search for planetary signals in the RV data and account for stellar variability with a joint model of RV, FWHM and BS. For faster convergence of our parameter estimation, we only use data from Instruments 2, 3 and 4.

2.2.1 Stellar Activity Modelling

We model the stellar activity in our data using Gaussian process (GP) regression ([Williams and Rasmussen, 1995](#)). GPs are powerful non-parametric models, which can model correlated noise.

Joint RV, FWHM and BS Model

We present a multidimensional GP approach to model the RV, FWHM and BS data jointly, in order to leverage correlations between the stellar activity indicators and the actual RV data. This approach helps to model stellar activity better than a simple unidimensional RV model and provides a more robust parameter estimation framework.

Quasi-periodic Kernel

The core of a GP is the covariance function, also known as the kernel. It defines how points in the input space are correlated ([Roberts et al., 2013](#)). Starspots and rotational modulation introduce period signals. However, they also evolve over time, making the resulting stellar activity a quasi-periodic signal. We therefore consider a quasi-periodic kernel k_{QP} for our covariance function, given by the following equation ([Rajpaul et al., 2015](#)):

$$k_{QP}(x, x') = k_{SE}(x, x') \cdot k_{ESS}(x, x') = h^2 \cdot \exp\left(-\frac{[x - x']^2}{2\lambda_e^2}\right) \cdot \exp\left(-\frac{\sin^2\left[\frac{\pi|x-x'|}{P_{rot}}\right]}{2\lambda_p^2}\right), \quad (8)$$

where h is the amplitude parameter, λ_e corresponds to the evolution time scale, λ_p determines the roughness and structure per period and P_{rot} is the characteristic period. The quasi-periodic kernel k_{QP} is a product of the squared-exponential kernel k_{SE} and the exponentiated, squared, sinusoidal kernel function k_{ESS} ([Angus et al., 2018](#)). The k_{ESS} kernel can capture periodicity due to stellar rotation, while the k_{SE} can capture slow evolutions of e.g. starspots over time. Combining both into k_{QP} offers the flexibility to model the complex nature of stellar activity. Note that we do also use an additive small jitter term to the kernels for more stable matrix inversion.

Shared Parameters

The GP model uses quasi-periodic kernels for each time-series with most of the hyperparameters shared as in [Faria et al. \(2022\)](#). The only parameter that is not shared is the amplitude parameter h , for which we introduce the individual parameters h_{RV} , h_{FWHM} and h_{BS} . Sharing the length scales and periodicity parameters leverages the underlying physical correlations between these measurements and effectively models common sources of variability.

Priors

Because we adopt a Bayesian approach for parameter estimation, we need to specify parameter priors. Because CB 01223 has previously been observed photometrically by TESS and a rotation period of $P_{rot} = 31 \pm 10$ days has been derived, we propose a Gaussian prior for the period P_{rot} centered around 31 days with a standard deviation of 10 days, i.e. $P_{rot} \sim \mathcal{N}(31, 10)$. For the remaining parameters, due to our lack of prior knowledge, we adopt uninformative priors as in [Damasso and Del Sordo \(2017\)](#), with bounds refined through iterative experimentation and convergence analysis. The bounds of these uniform priors are given in Table 5.

Table 5: Prior probability distributions for the stellar activity model parameters.

Parameter	Prior	Lower Bound	Upper Bound
P_{rot} [days]	$\mathcal{N}(31, 10)$	-	-
λ_p	Uniform	0	20
λ_e [days]	Uniform	0	2500
h_{RV} [km/s]	Uniform	0	1
h_{FWHM} [km/s]	Uniform	0	10
h_{BS} [km/s]	Uniform	0	1

2.2.2 Planetary Signal Modelling

We model planetary signals using Keplerians.

Keplerian Model

The radial velocity $RV(t)$ of a star due to an orbiting planet is given by:

$$RV(t) = K [\cos(\omega + \nu(t)) + e \cos(\omega)], \quad (9)$$

where ω is the argument of periastron, $\nu(t)$ is the true anomaly at time t , e is the orbital eccentricity, and K is the radial velocity semi-amplitude given by:

$$K = \left(\frac{2\pi G}{P} \right)^{1/3} \frac{M_p \sin i}{(M_s + M_p)^{2/3}} \frac{1}{\sqrt{1 - e^2}}, \quad (10)$$

where G is the gravitational constant, P is the orbital period of the planet, M_p is the planetary mass, M_s is the stellar mass, i is the orbital inclination, and e is the orbital eccentricity. The true anomaly is found by solving the following equations:

$$\nu(t) = 2 \arctan \left(\sqrt{\frac{1+e}{1-e}} \tan \left(\frac{E(t)}{2} \right) \right) \quad \text{and} \quad E(t) - eE(t) = \frac{2\pi(t - t_c)}{P}, \quad (11)$$

where T_c is the time of periastron passage and $E(t)$ is the eccentric anomaly. The orbital period P is related to the semi-major axis a via Kepler's third law:

$$P^2 = \frac{4\pi^2 a^3}{G(M_s + M_p)} \quad (12)$$

Simplified Keplerian Model

In the case of a circular orbit, we define the simplified Keplerian model as:

$$RV(t) = K \sin \left(\frac{2\pi(t - T_c)}{P} \right) \quad (13)$$

Prior for RV Semi-amplitude K

The observed scatter in the RV data is < 9 m/s, based on the standard deviation of the RVs from Instruments 2, 3 and 4. This suggests that any detected planets likely have RV semi-amplitudes < 9 m/s (Rajpaul, personal communication, June 18, 2024). We thus also do not expect K to vary over several orders of magnitude and thus use a uniform prior $K \sim \mathcal{U}(K_{min}, K_{max})$ and set the upper bound to $K_{max} = 9$ m/s. Note that while deriving a prior from the data itself is not considered best practice, we faced convergence issues throughout the project and thus use short-cuts for demonstration and time-saving purposes.

Prior for Orbital Period P

As we only consider RV data from Instruments 2, 3 and 4 with the associated duration of the data being 895 days, we cannot reliably detect planetary candidates with periods > 895 days (Rajpaul, personal communication, June 18, 2024). Therefore, we use $P \sim \mathcal{U}(P_{min}, P_{max})$ with $P_{max} \approx 895$ as our period prior.

Prior for Orbital Eccentricity e

As for the eccentricity e , we use a physically-motivated prior with the Beta distribution $\mathcal{B}(e_{min}, e_{max})$ given by:

$$\mathcal{B}(e | e_{min}, e_{max}) = \frac{\Gamma(e_{min} + e_{max})}{\Gamma(e_{min})\Gamma(e_{max})} e^{e_{min}-1} (1-e)^{e_{max}-1}, \quad (14)$$

where the limits are set to $e_{min} = 0.867$ and $e_{max} = 3.03$ as in [Kipping \(2013\)](#). This distribution represents the underlying distribution of exoplanet eccentricities particularly well.

Other Priors

Due to a lack of prior knowledge, we adopt uninformative priors for the other parameters with the bounds shown in Table 6.

Table 6: Prior probability distributions for the planetary signal model parameters.

Parameter	Prior	Lower Bound	Upper Bound
K [m/s]	Uniform	0	9
P [days]	Uniform	0	895
T_c [days]	Uniform	5471.8	6366.7
e	$\mathcal{B}(0.867, 3.03)$	-	-
ω [rad]	Uniform	0	2π

2.2.3 Model I: No Planets, Stellar Activity GP only

First, we fit the data with the multidimensional GP model only. We obtain the best-fit parameters using nested sampling with the `dynesty` package ([Speagle, 2020](#)) due to the high-dimensional parameter space, multi-modal posteriors and robust model comparison. Figure 19 shows the trace plot of the sampling results and Figure 20 shows the corner plot of the posterior distributions. Figure 21 shows the multidimensional GP regression output with the best-fit parameters.

2.2.4 Models II & III: One Planet

We model the case of 1 planetary signal with a full and a simplified Keplerian. In practice, this involves subtracting the Keplerian signal from the GP mean for the RV time series. The residuals are then used to compute the log-likelihood of the GP model.

Model II: Simplified Keplerian (circular orbit)

Using the simplified Keplerian to model planetary signals involves 3 additional parameters: K , P and T_c . Figure 22 shows the trace plot of the sampling results and Figure 23 shows the corner plot of the posterior distributions. Figure 24 shows the GP regression including the Keplerian model with the best-fit parameters.

Model III: Full Keplerian (eccentric orbit)

Using a full Keplerian involved 2 further parameters: e and ω . Figure 25 shows the trace plot of the sampling results and Figure 26 shows the corner plot of the posterior distributions. Figure 27 shows the GP regression including the Keplerian model with the best-fit parameters.

2.2.5 Models IV & V: Two Planets

For the case of 2 planets, the parameter estimation results are shown in Figures 28 and 29 for the circular orbit model (Model IV), and in Figures 31 and 32 for the eccentric orbit (Model V).

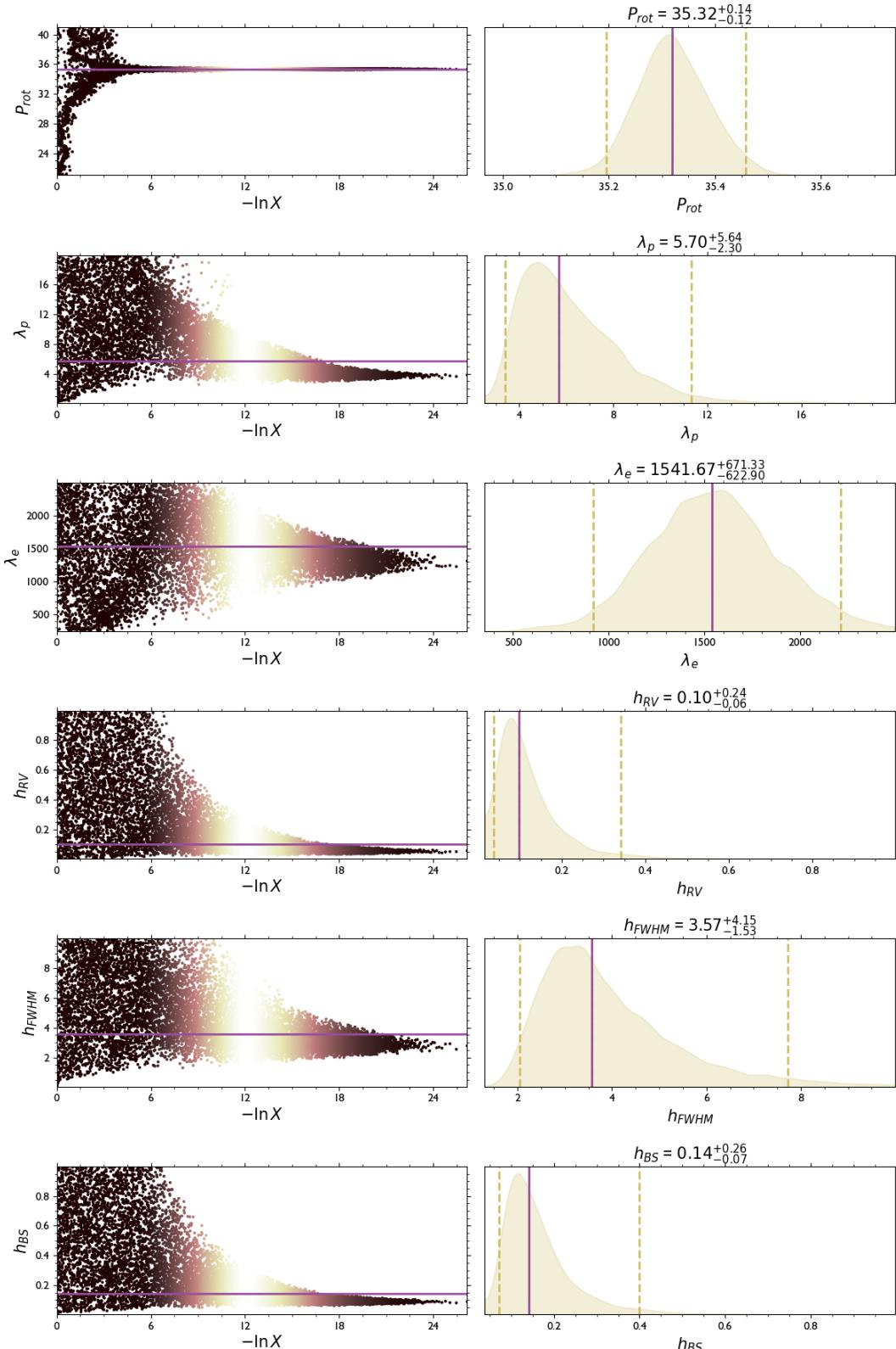


Figure 19: Trace plot of the nested sampling results for the stellar activity only model. The purple lines are the medians of the unweighted sample posterior distributions. The golden lines represent the 95% CI.

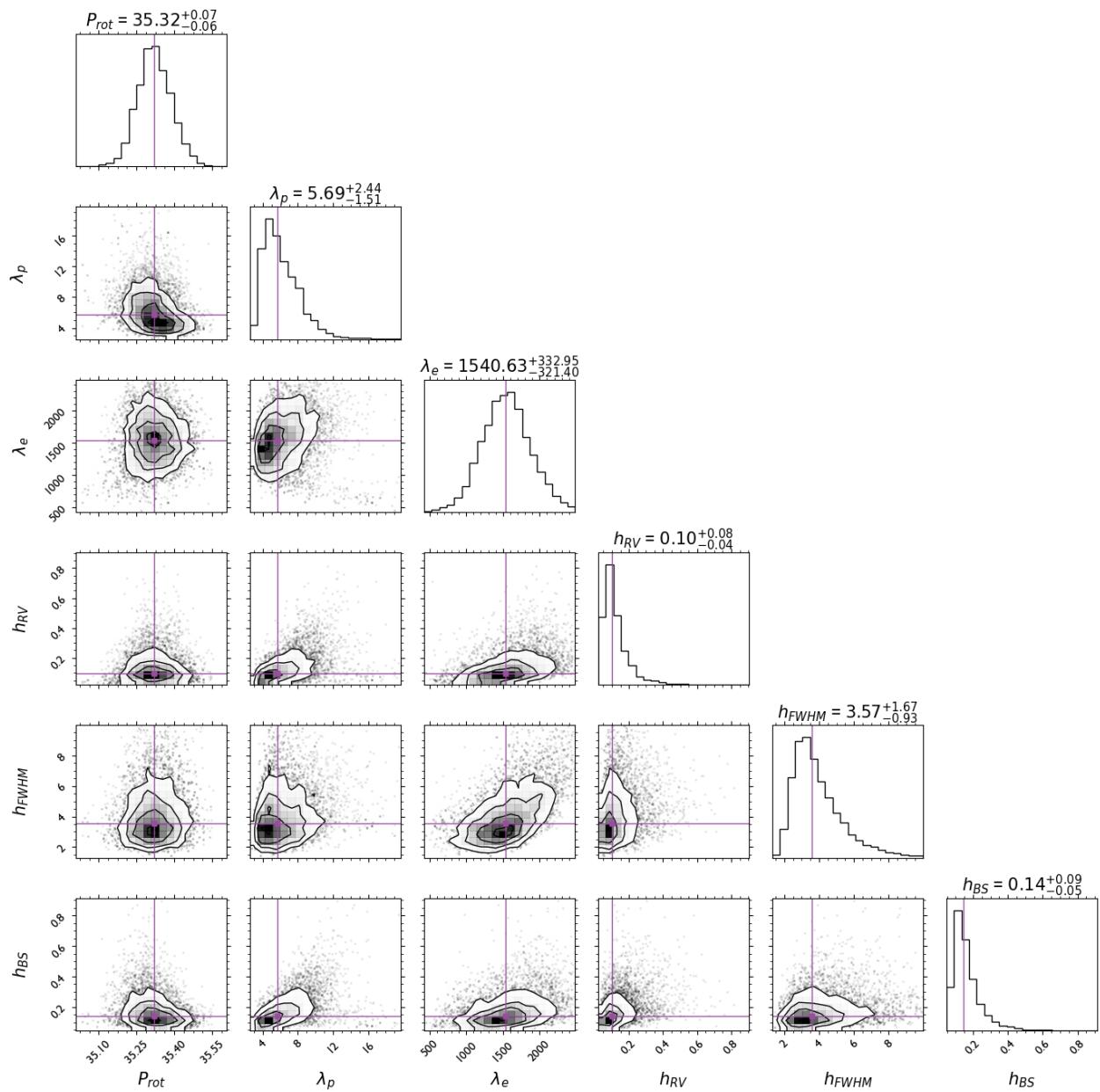


Figure 20: Corner plot of the posterior parameter distributions for the stellar activity only model. The purple lines are the medians of the weighted sample posterior distributions. The parameter errors represent the 68% CI.

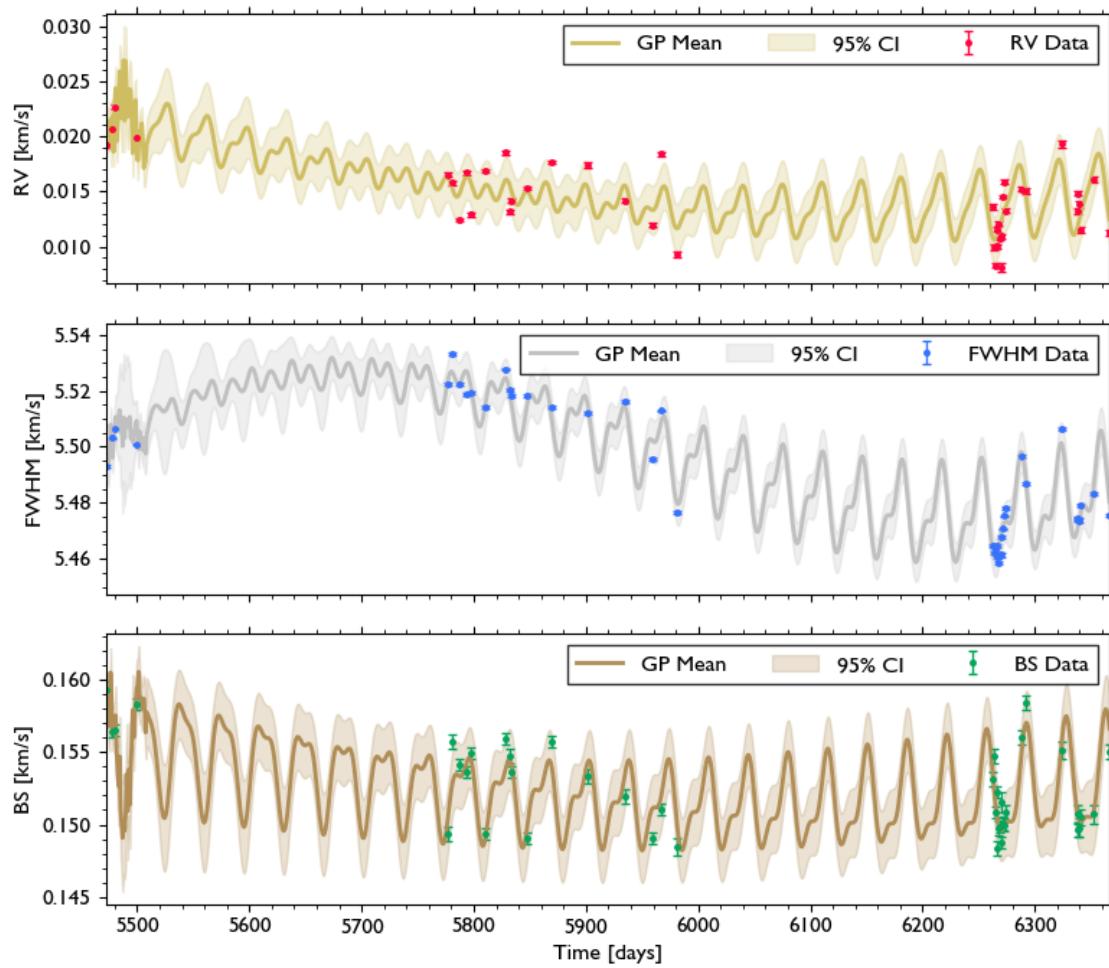


Figure 21: Multidimensional GP regression output with the best-fit parameters from the stellar activity only model.

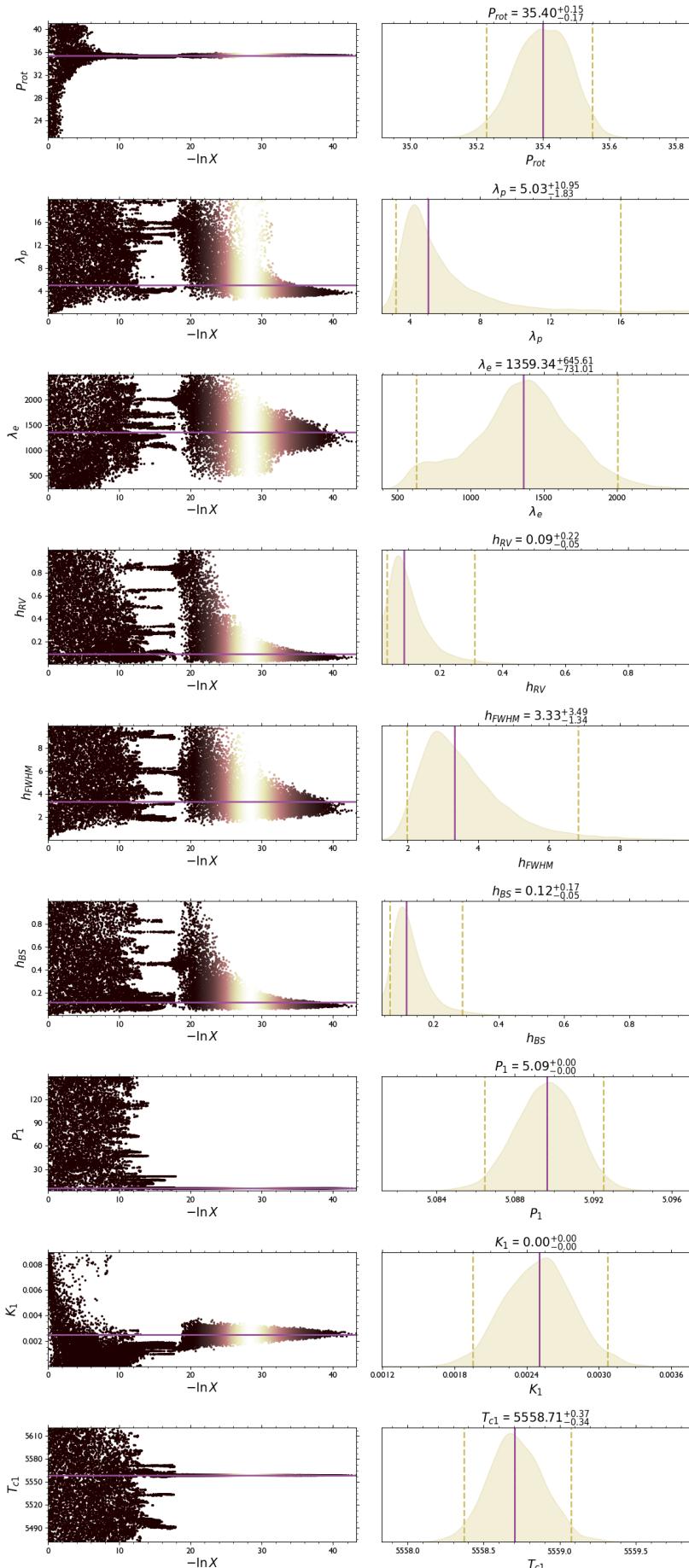


Figure 22: Trace plot of the nested sampling results for the 1 planet model (circular orbit).

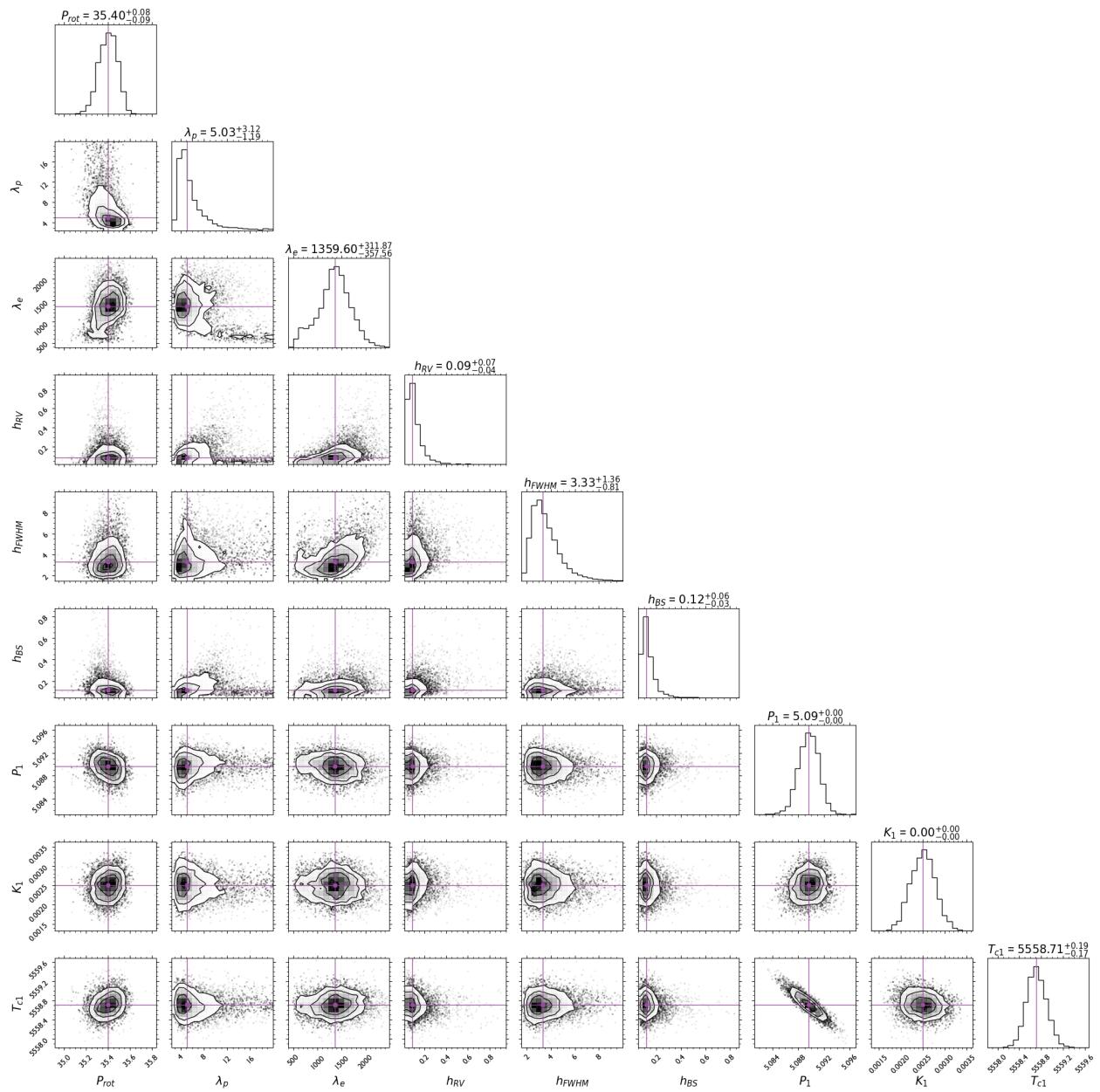


Figure 23: Corner plot of the posterior parameter distributions for the 1 planet model (circular orbit).

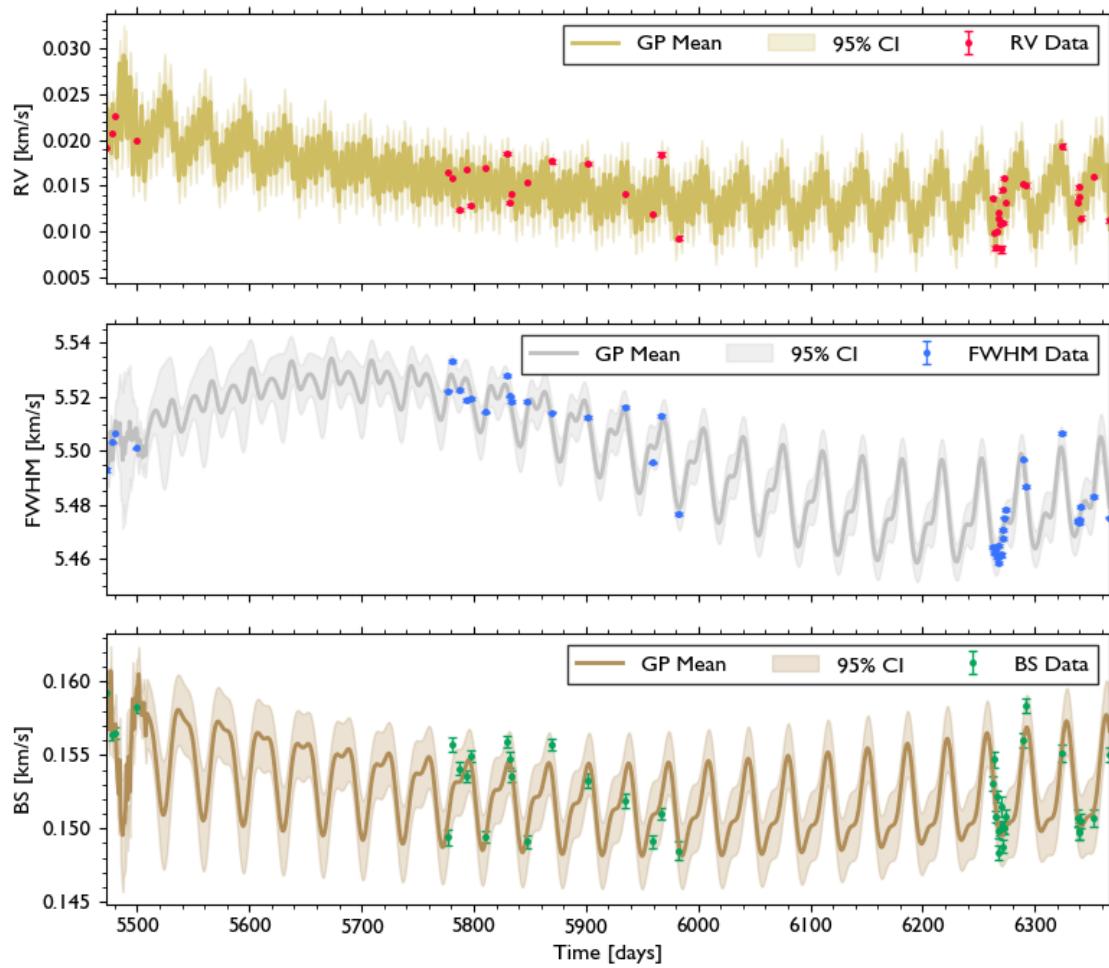


Figure 24: Multidimensional GP regression and simplified Keplerian model output with the best-fit parameters from the 1 planet model (circular orbit).

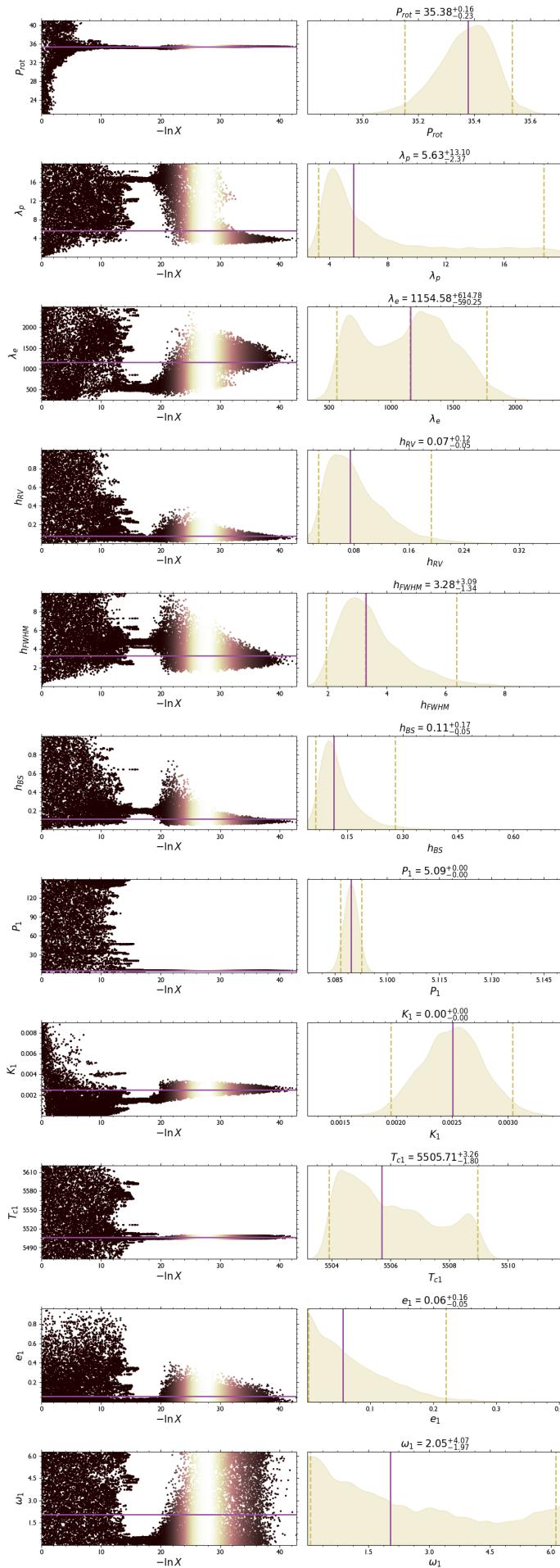


Figure 25: Trace plot of the nested sampling results for the 1 planet model (eccentric orbit).

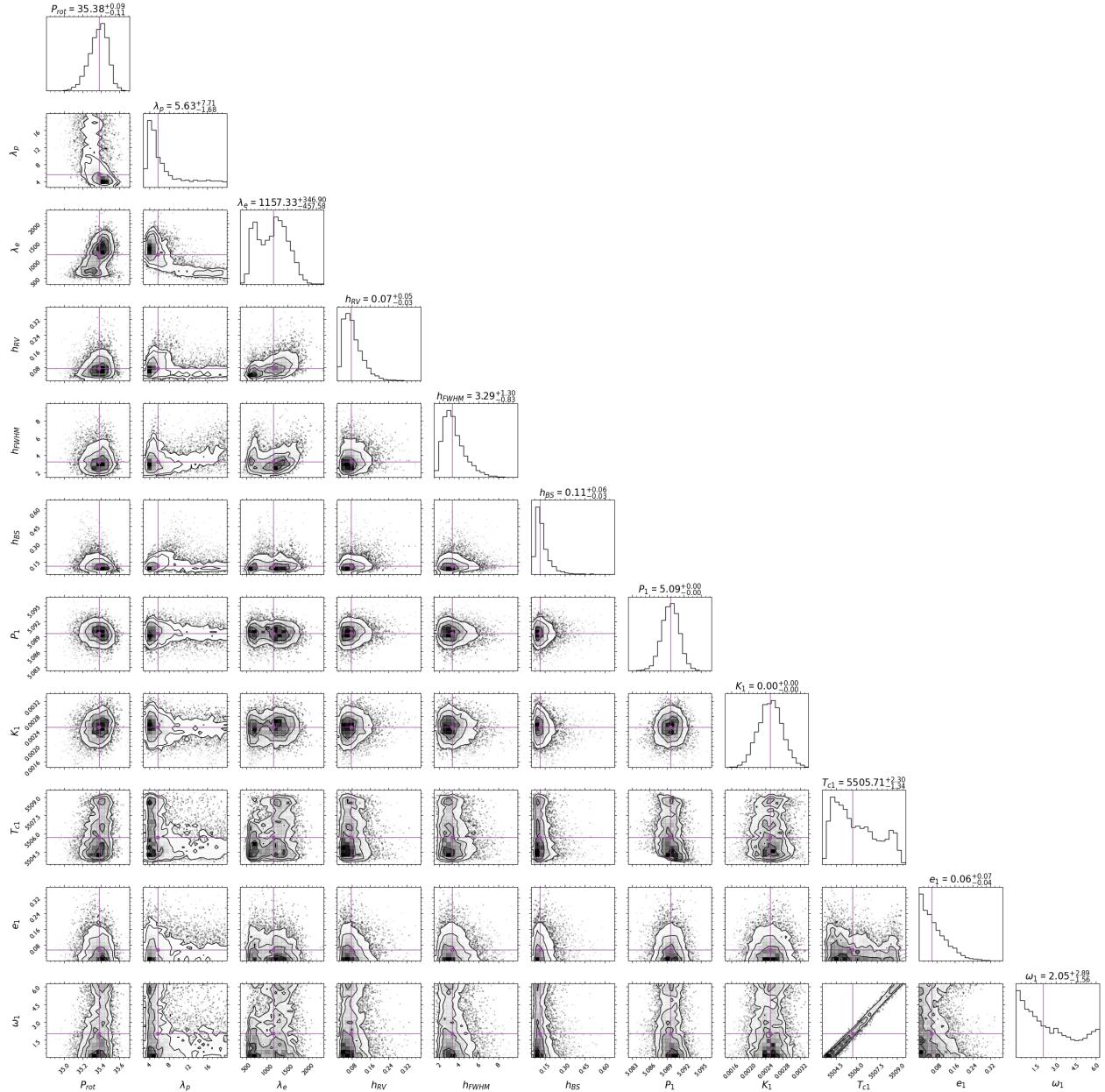


Figure 26: Corner plot of the posterior parameter distributions for the 1 planet model (eccentric orbit).

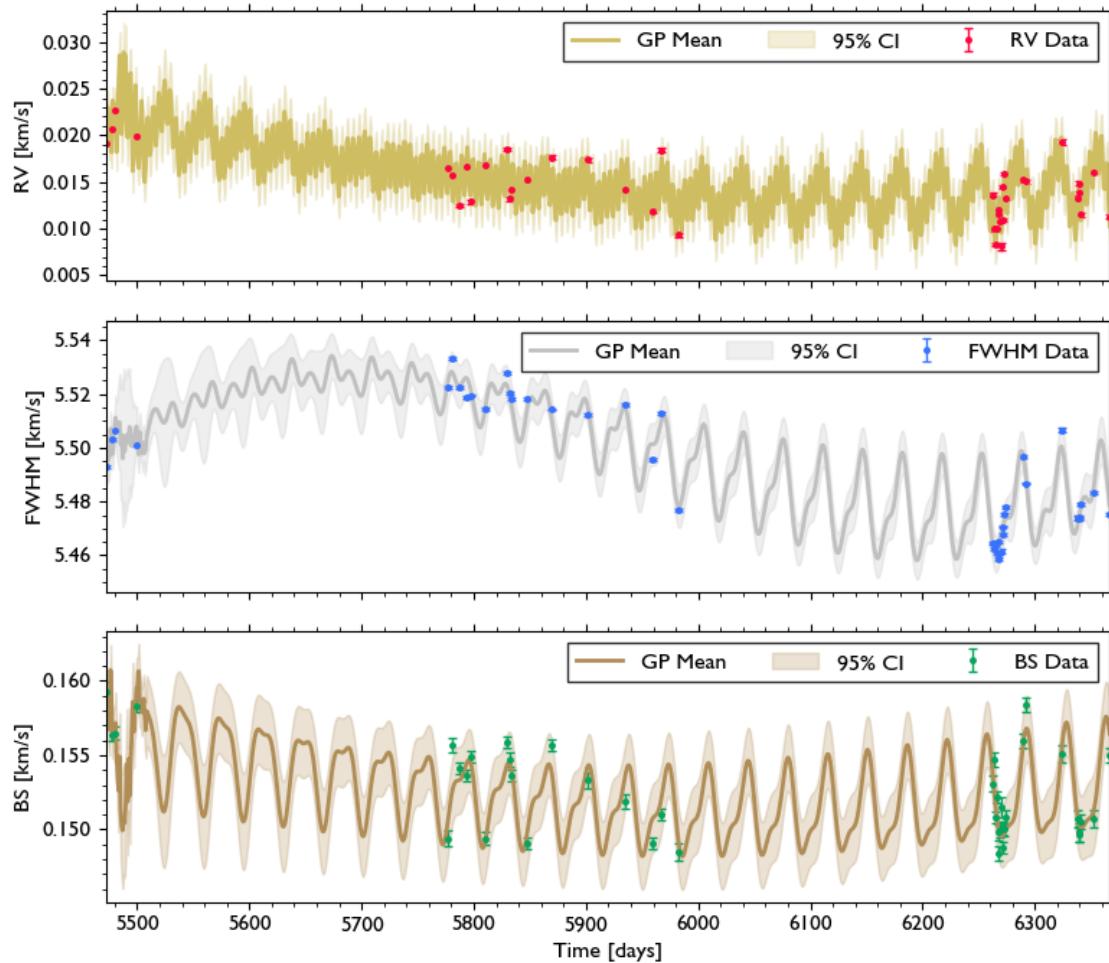


Figure 27: Multidimensional GP regression and full Keplerian model output with the best-fit parameters from the 1 planet model (eccentric orbit).

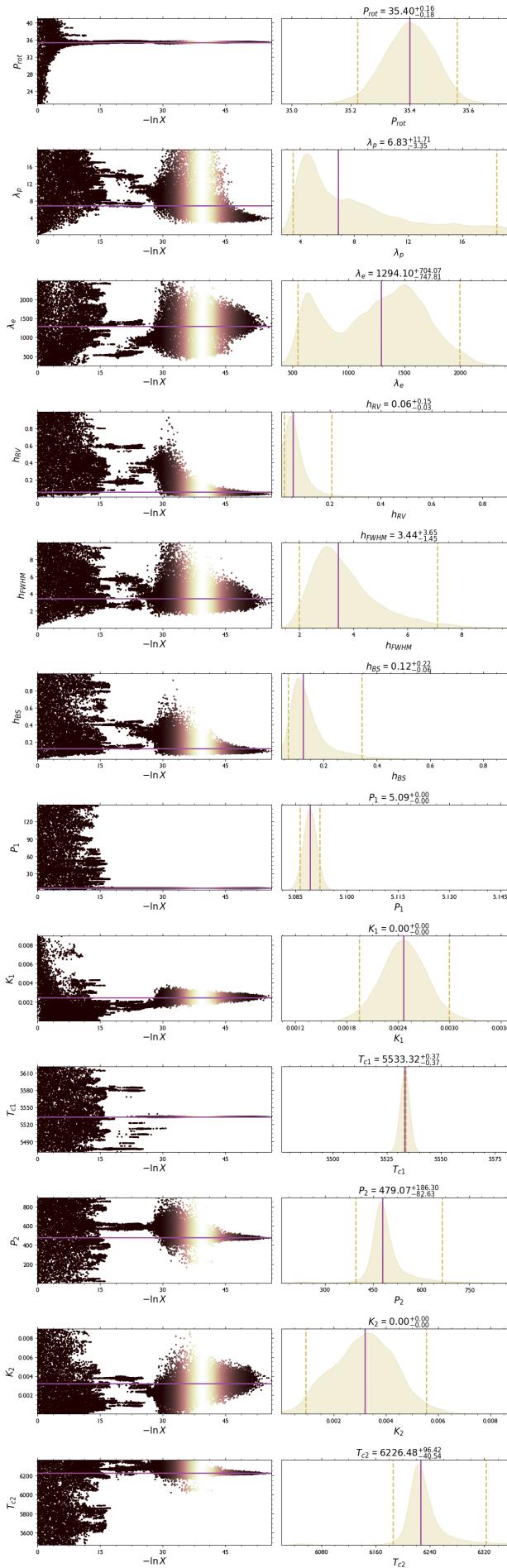


Figure 28: Trace plot of the nested sampling results for the 2 planet model (circular orbits).

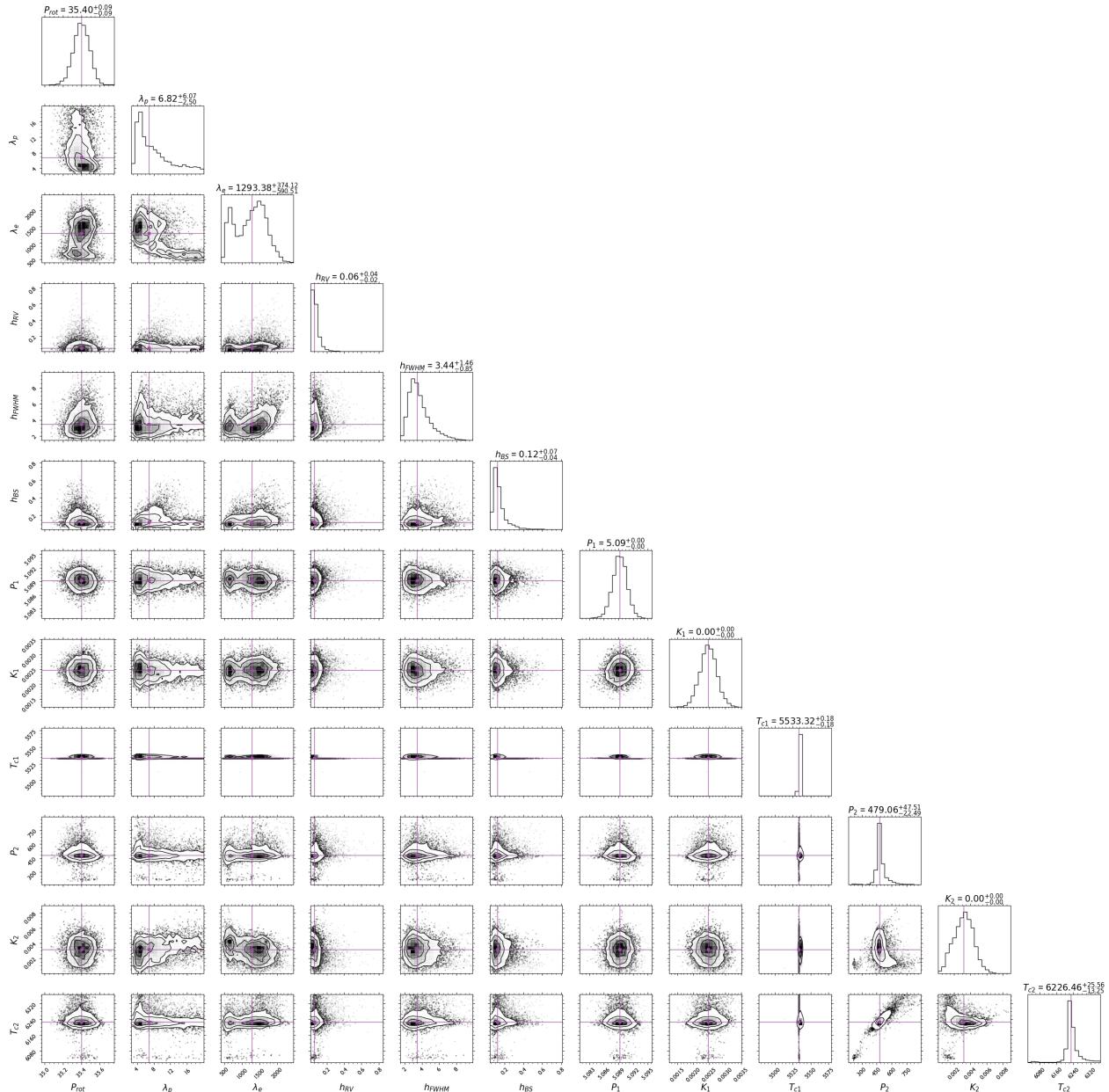


Figure 29: Corner plot of the posterior parameter distributions for the 2 planet model (circular orbits).

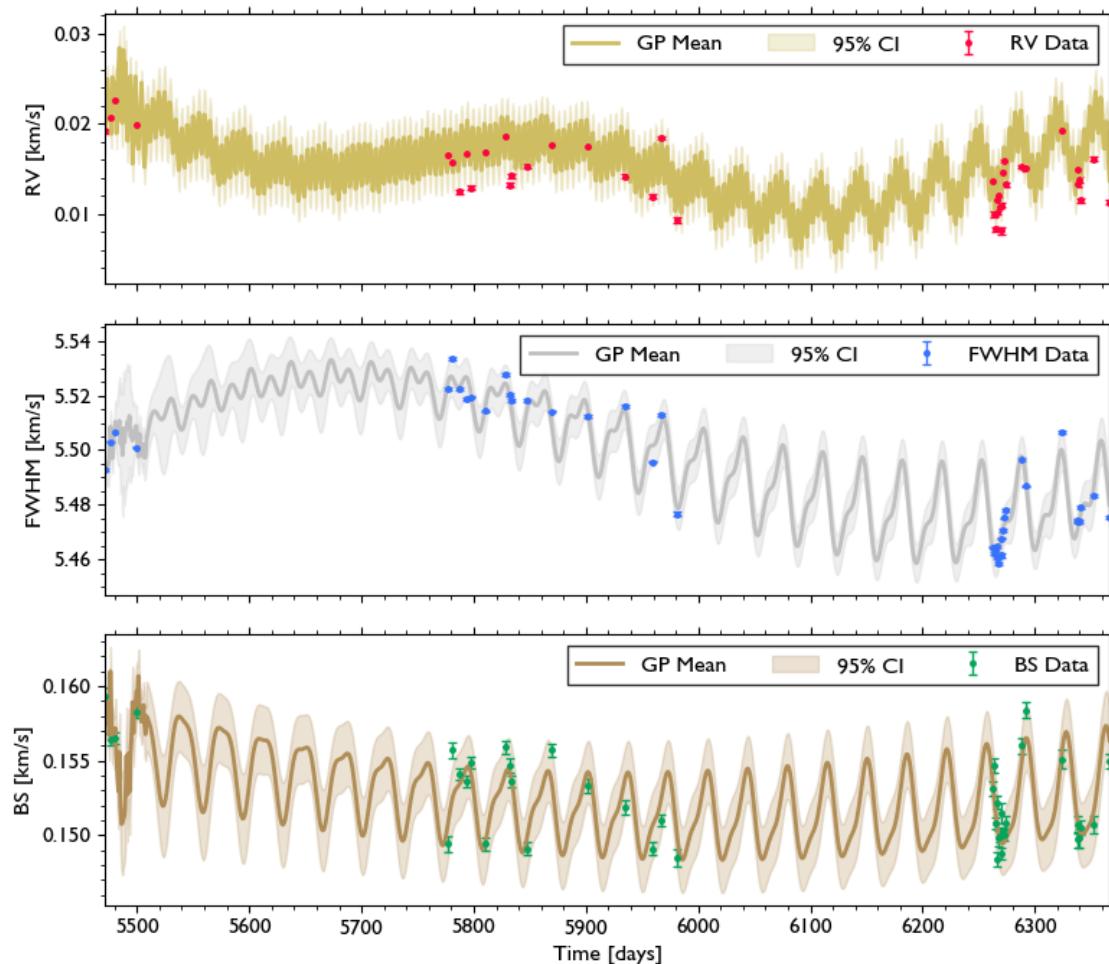


Figure 30: Multidimensional GP regression and full Keplerian model output with the best-fit parameters from the 2 planet model (circular orbits).

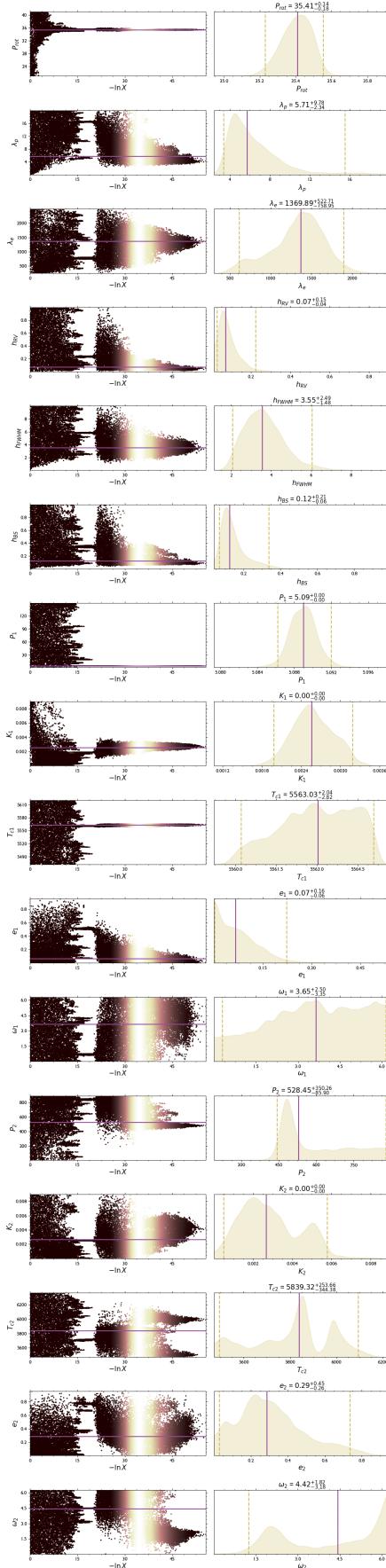


Figure 31: Trace plot of the nested sampling results for the 2 planet model (eccentric orbits).

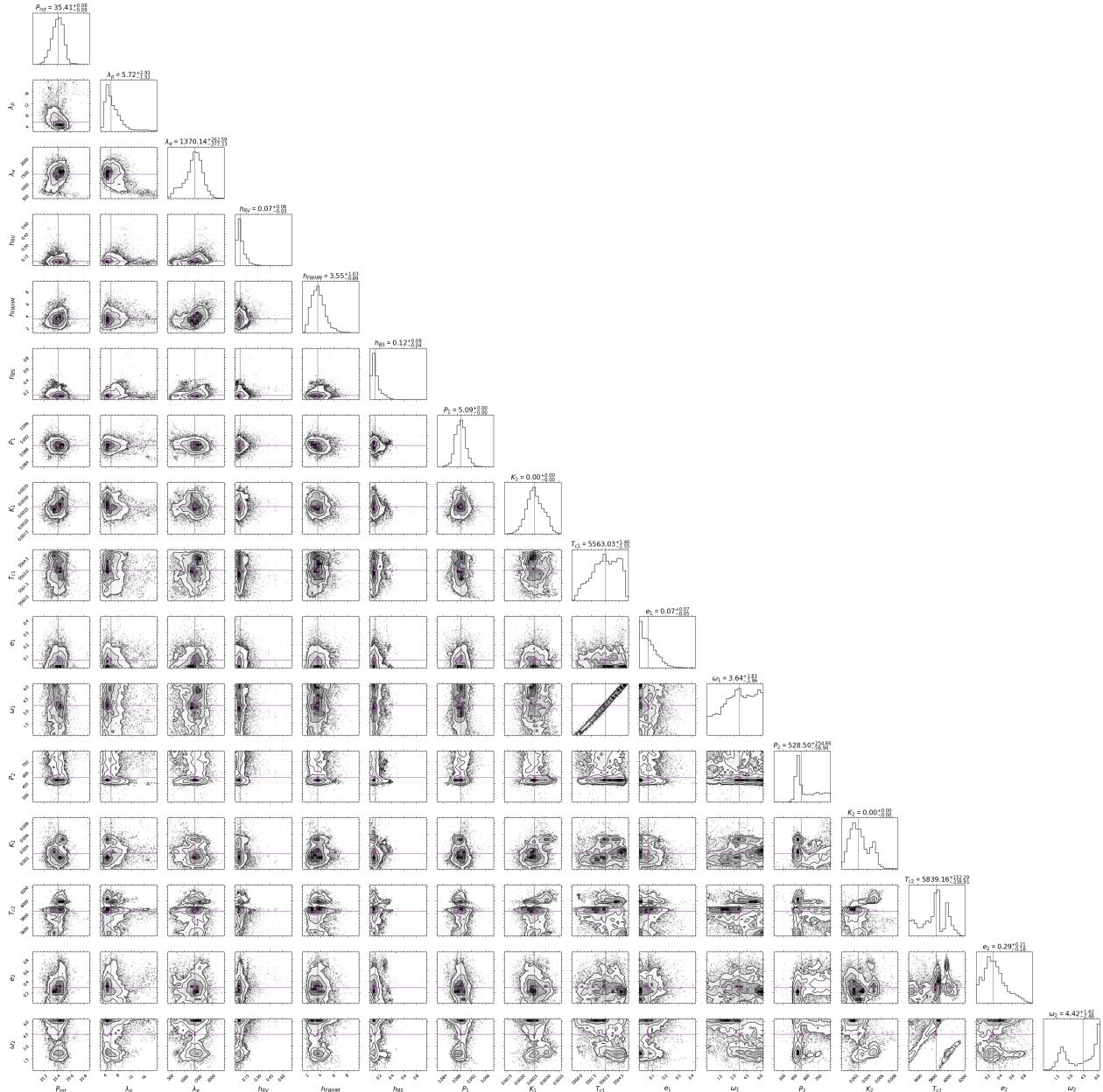


Figure 32: Corner plot of the posterior parameter distributions for the 2 planet model (circular orbits).

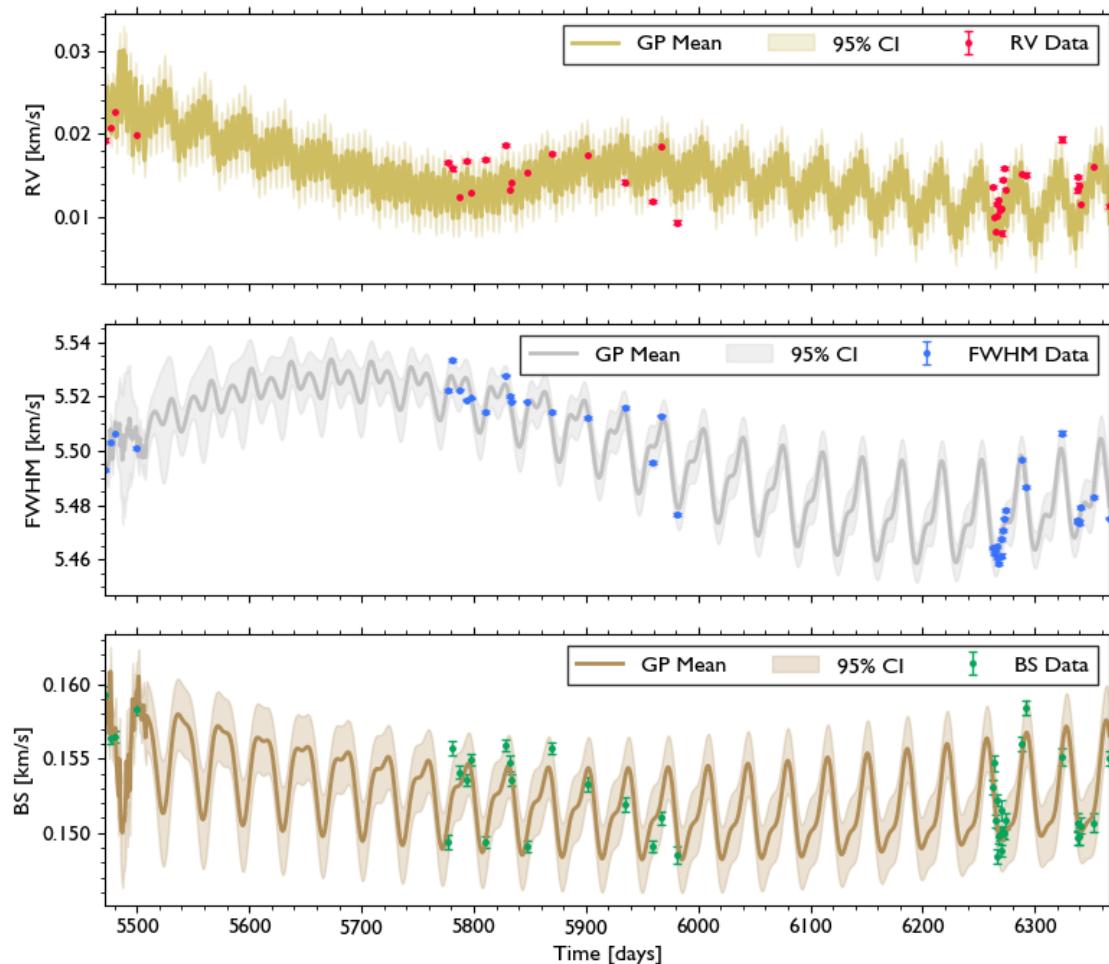


Figure 33: Multidimensional GP regression and full Keplerian model output with the best-fit parameters from the 2 planet model (eccentric orbits).

2.2.6 Posterior Summaries

Tables 7, 8 and 9 summarise the parameter estimation results for all models

Table 7: Parameter estimation results for Model I (0 planets, stellar activity GP only).

Parameter	Estimate
P_{rot} [days]	$35.32^{+0.07}_{-0.06}$
λ_p	$5.69^{+2.44}_{-1.51}$
λ_e [days]	1541^{+333}_{-321}
h_{RV} [km/s]	$0.10^{+0.08}_{-0.04}$
h_{FWHM} [km/s]	$3.57^{+1.67}_{-0.93}$
h_{BS} [km/s]	$0.14^{+0.09}_{-0.05}$

Table 8: Parameter estimation results for Models II and III (1 planet).

Parameter	Circular Orbit Estimate	Eccentric Orbit Estimate
P_{rot} [days]	$35.40^{+0.08}_{-0.09}$	$35.38^{+0.09}_{-0.11}$
λ_p	$5.03^{+3.12}_{-1.19}$	$5.63^{+7.71}_{-1.68}$
λ_e [days]	1360^{+312}_{-358}	1157^{+347}_{-458}
h_{RV} [km/s]	$0.09^{+0.07}_{-0.04}$	$0.07^{+0.05}_{-0.03}$
h_{FWHM} [km/s]	$3.33^{+1.36}_{-0.81}$	$3.29^{+1.30}_{-0.83}$
h_{BS} [km/s]	$0.12^{+0.06}_{-0.03}$	$0.11^{+0.06}_{-0.03}$
P_1 [days]	$5.090^{+0.001}_{-0.002}$	$5.090^{+0.001}_{-0.002}$
K_1 [m/s]	$2.50^{+0.27}_{-0.29}$	$2.51^{+0.27}_{-0.29}$
T_{c_1} [days]	5558.7 ± 0.2	$5505.7^{+2.3}_{-1.3}$
e_1	-	$0.057^{+0.073}_{-0.043}$
ω_1 [rad]	-	$2.05^{+2.89}_{-1.56}$

Table 9: Parameter estimation results for Models IV and V (2 planets).

Parameter	Circular Orbit Estimate	Eccentric Orbit Estimate
P_{rot} [days]	35.40 ± 0.09	$35.41^{+0.08}_{-0.09}$
λ_p	$6.82^{+6.07}_{-2.50}$	$5.72^{+2.93}_{-1.52}$
λ_e [days]	1293^{+374}_{-591}	1370^{+263}_{-377}
h_{RV} [km/s]	$0.06^{+0.04}_{-0.02}$	$0.07^{+0.06}_{-0.03}$
h_{FWHM} [km/s]	$3.44^{+1.46}_{-0.85}$	$3.55^{+1.03}_{-0.89}$
h_{BS} [km/s]	$0.12^{+0.07}_{-0.04}$	$0.12^{+0.08}_{-0.04}$
P_1 [days]	5.089 ± 0.001	5.089 ± 0.001
K_1 [m/s]	2.47 ± 0.26	$2.56^{+0.37}_{-0.29}$
T_{c_1} [days]	5533.3 ± 0.2	$5563.7^{+1.5}_{-1.6}$
e_1	-	$0.066^{+0.074}_{-0.051}$
ω_1 [rad]	-	$3.64^{+1.83}_{-1.96}$
P_2 [days]	$479.1^{+47.5}_{-22.5}$	$528.5^{+254.9}_{-56.9}$
K_2 [m/s]	$3.20^{+1.16}_{-1.32}$	$2.65^{+2.13}_{-1.29}$
T_{c_2} [days]	$6226.5^{+25.6}_{-13.3}$	$5839.2^{+152.3}_{-238.9}$
e_2	-	$0.287^{+0.207}_{-0.156}$
ω_2 [rad]	-	$4.42^{+1.62}_{-2.46}$

2.2.7 Model Comparison

Nested sampling provides a robust estimate of the evidence $\ln \mathcal{Z}$, which quantifies how well a model explains the data. We use the Bayes factor $\Delta \ln \mathcal{Z}$ as a powerful tool for model comparison. There is strong evidence in favour of models with 1 planet as compared models with 0 or 2 planets. The most favourable model is Model III. However, the MAP of the posterior distribution for e is very close to 0, which is effectively to a circular orbit. This explains why the evidence for Model II is very similar. Given that the Bayes factor of these models is only $\Delta \ln \mathcal{Z}$ (II - III) = -0.42 , we consider both models to estimate the planetary properties of the detected planet.

Table 10: Bayesian evidences for the tested models.

Model	Description	$\ln \mathcal{Z}$
I	0 planets (stellar activity only)	434.02 ± 0.15
II	1 planet (circular)	461.70 ± 0.24
III	1 planet (free eccentricity)	462.12 ± 0.23
IV	2 planets (circular)	456.78 ± 0.28
V	2 planets (free eccentricity)	459.51 ± 0.26
$\Delta \ln \mathcal{Z}$ (I - II)	0 minus 1 planet (circular)	-27.6
$\Delta \ln \mathcal{Z}$ (I - II)	0 minus 1 planet (free eccentricity)	28.10
$\Delta \ln \mathcal{Z}$ (II - III)	1 planet (circular minus free eccentricity)	-0.42
$\Delta \ln \mathcal{Z}$ (IV - V)	2 planets (circular minus free eccentricity)	-2.73
$\Delta \ln \mathcal{Z}$ (II - IV)	1 minus 2 planets (circular)	4.92
$\Delta \ln \mathcal{Z}$ (III - V)	1 minus 2 planets (free eccentricity)	2.61

2.2.8 Property Estimation with the Radial Velocity Method

Assuming $M_s + M_p \approx M_s$, we can estimate the minimum planetary mass $M_p \sin i$ and the orbital semi-major axis a with:

$$M_p \sin i \approx K \left(\frac{P}{2\pi G} \right)^{1/3} M_s^{2/3} \cdot \sqrt{1 - e^2}, \quad (15)$$

$$a \approx \left(\frac{M_s G P^2}{4\pi^2} \right)^{1/3}, \quad (16)$$

where M_s is the stellar mass.

2.2.9 Results & Summary

The estimated properties of the detected exoplanet are given in Table 11. As expected, the results are very similar because of the very low e estimate in the eccentric orbit.

Table 11: Physical and orbital properties of the detected exoplanet.

Parameter	Circular Orbit Estimate	Eccentric Orbit Estimate
$M_p \sin i [M_\oplus]$	$5.25^{+0.57}_{-0.61}$	$5.24^{+0.57}_{-0.61}$
$a [\text{AU}]$	0.0512 ± 0.0003	0.0512 ± 0.0003

In summary, we find an exoplanet around the star CB 01223 with a period of 5.1 days and a minimum mass of $5.25 M_\oplus$. To confirm the detection and improve the parameter estimates, one could leverage data from other facilities and build more sophisticated models.

Bibliography

- Angus, R., Morton, T., Aigrain, S., Foreman-Mackey, D., and Rajpaul, V. (2018). Inferring probabilistic stellar rotation periods using gaussian processes. *Monthly Notices of the Royal Astronomical Society*, 474(2):2094–2108.
- Berdyugina, S. V. (2005). Starspots: a key to the stellar dynamo. *Living Reviews in Solar Physics*, 2(1):8.
- Bosse, I., Bonfils, X., and Santos, N. (2012). Soap-a tool for the fast computation of photometry and radial velocity induced by stellar spots. *Astronomy & Astrophysics*, 545:A109.
- Brown, T. M. and Gilliland, R. L. (1994). Asteroseismology. *Annual Review of Astronomy and Astrophysics*, 32(1):37–82.
- Cardoso, J. V. d. M., Hedges, C., Gully-Santiago, M., Saunders, N., Cody, A. M., Barclay, T., Hall, O., Sagear, S., Turtelboom, E., Zhang, J., et al. (2018). Lightkurve: Kepler and tess time series analysis in python. *Astrophysics Source Code Library*, pages ascl–1812.
- Da Silva, J. G., Santos, N., Bonfils, X., Delfosse, X., Forveille, T., Udry, S., Dumusque, X., and Lovis, C. (2012). Long-term magnetic activity of a sample of m-dwarf stars from the harps program-ii. activity and radial velocity. *Astronomy & Astrophysics*, 541:A9.
- Damasso, M. and Del Sordo, F. (2017). Proxima centauri reloaded: Unravelling the stellar noise in radial velocities. *Astronomy & Astrophysics*, 599:A126.
- Dumusque, X., Santos, N., Udry, S., Lovis, C., and Bonfils, X. (2011). Planetary detection limits taking into account stellar noise-ii. effect of stellar spot groups on radial-velocities. *Astronomy & Astrophysics*, 527:A82.
- Faria, J., Mascareño, A. S., Figueira, P., Silva, A., Damasso, M., Demangeon, O., Pepe, F., Santos, N., Rebolo, R., Cristiani, S., et al. (2022). A candidate short-period sub-earth orbiting proxima centauri. *Astronomy & Astrophysics*, 658:A115.
- Heller, R., Rodenbeck, K., and Hippke, M. (2019). Transit least-squares survey-i. discovery and validation of an earth-sized planet in the four-planet system k2-32 near the 1: 2: 5: 7 resonance. *Astronomy & Astrophysics*, 625:A31.
- Heller, R., Rodenbeck, K., and Hippke, M. (2024). Python interface - tls. <https://transitleastsquares.readthedocs.io/en/latest/Python%20interface.html>.
- Hippke, M. and Heller, R. (2019a). Optimized transit detection algorithm to search for periodic transits of small planets. *Astronomy & Astrophysics*, 623:A39.
- Hippke, M. and Heller, R. (2019b). Tls: Transit least squares. *Astrophysics Source Code Library*, pages ascl–1910.
- Kipping, D. M. (2013). Parametrizing the exoplanet eccentricity distribution with the beta distribution. *Monthly Notices of the Royal Astronomical Society: Letters*, 434(1):L51–L55.
- Kovács, G., Zucker, S., and Mazeh, T. (2002). A box-fitting algorithm in the search for periodic transits. *Astronomy & Astrophysics*, 391(1):369–377.
- Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39:447–462.
- Mathur, S., Hekker, S., Trampedach, R., Ballot, J., Kallinger, T., Buzasi, D., Garcia, R. A., Huber, D., Jiménez, A., Mosser, B., et al. (2011). Granulation in red giants: observations by the kepler mission and three-dimensional convection simulations. *The Astrophysical Journal*, 741(2):119.

- Michel, E., Baglin, A., Auvergne, M., Catala, C., Samadi, R., Baudin, F., Appourchaux, T., Barban, C., Weiss, W. W., Berthomieu, G., et al. (2008). Corot measures solar-like oscillations and granulation in stars hotter than the sun. *Science*, 322(5901):558–560.
- Odunlade, A. A. (2010). *Overview of Exoplanet Research*. PhD thesis. http://www.astro.ex.ac.uk/people/alapini/Publications/PhD_chap1.pdf.
- Pearson, K. (1895). Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.
- Pepe, F., Ehrenreich, D., and Meyer, M. R. (2014). Instrumentation for the detection and characterization of exoplanets. *Nature*, 513(7518):358–366.
- Pettersen, B. (1989). A review of stellar flares and their characteristics. In *International Astronomical Union Colloquium*, volume 104, pages 299–312. Cambridge University Press.
- Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., and Roberts, S. (2015). A gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3):2269–2291.
- Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550.
- Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A. M., Kerzendorf, W. E., et al. (2013). Astropy: A community python package for astronomy. *Astronomy & Astrophysics*, 558:A33.
- Scargle, J. D. (1982). Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal, Part 1*, vol. 263, Dec. 15, 1982, p. 835–853., 263:835–853.
- Speagle, J. S. (2020). dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, 493(3):3132–3158.
- Strassmeier, K. G. (2009). Starspots. *The Astronomy and Astrophysics Review*, 17:251–308.
- VanderPlas, J. T. (2018). Understanding the lomb-scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16.
- Williams, C. and Rasmussen, C. (1995). Gaussian processes for regression. *Advances in neural information processing systems*, 8.
- Wilson, O. (1978). Chromospheric variations in main-sequence stars. *Astrophysical Journal, Part 1*, vol. 226, Dec. 1, 1978, p. 379-396., 226:379–396.

A Use of auto-generation tools

A copy of the README.md file is attached to this document and can be found on the next pages.

A.1 Co-Pilot

GitHub Co-Pilot assisted the author in producing all function docstrings present in the project repository and a number of descriptive comments in the code. No specific commands have been given, instead auto-completion suggestions have occasionally been accepted.

A.2 ChatGPT

A.2.1 Prompt 1

Prompt:

How do I create a custom colour map in Python?

Output:

To create a custom basic colormap in Python, you can use the `LinearSegmentedColormap` method. Here's a concise example:

```
1 import matplotlib.pyplot as plt
2 import matplotlib.colors as mcolors
3
4 # Define the colors for your colormap
5 colors = ["red", "green", "blue"]
6
7 # Create the colormap
8 cmap = mcolors.LinearSegmentedColormap.from_list("my_custom_cmap", colors)
9
10 # Use your colormap
11 plt.imshow([[0,1],[2,3]], cmap=cmap)
12 plt.colorbar()
13 plt.show()
```

Modification:

The output was modified by using custom colours to define custom colour maps.

```
1 sd2022_gold = '#cfbd62'
2 sd2022_silver = '#C0C0C0'
3 sd2022_bronze = '#b08d57'
4 cmap_gold = mcolors.LinearSegmentedColormap.from_list('my_colormap',
5     ['black',sd2022_gold, 'white'], N=256)
6 cmap_silver = mcolors.LinearSegmentedColormap.from_list('my_colormap',
7     ['black',sd2022_silver, 'white'], N=256)
8 cmap_bronze = mcolors.LinearSegmentedColormap.from_list('my_colormap',
9     ['black',sd2022_bronze, 'white'], N=256)
10 cmap_bronze_r = mcolors.LinearSegmentedColormap.from_list('my_colormap',
11     ['white',sd2022_bronze, 'black'], N=256)
```

Exoplanets - Coursework Submission (sd2022)

License MIT

Description

This project is associated with the submission of the coursework for the Exoplanets Module as part of the MPhil in Data Intensive Science at the University of Cambridge. The coursework assignment can be found here: [Exoplanets - Coursework Assignment](#). The associated report can be found here: [Exoplanets - Coursework Report](#).

Table of Contents

- [Installation and Usage](#)
- [Support](#)
- [License](#)
- [Project Status](#)
- [Authors and Acknowledgment](#)

Installation and Usage

To get started with the code associated with the coursework submission, follow these steps:

Requirements

- Python 3.9 or higher installed on your system.
- Conda installed (for managing the Python environment).
- Docker (if using containerisation for deployment).

Data

The data used for this project is found in the `sd2022/data` directory. The datasets are the following:

- [ex1_stars_image.png](#) (relevant for Exercise 1)
- [ex1_tess_lc.txt](#) (relevant for Exercise 1)
- [ex2_RVs.txt](#) (relevant for Exercise 2)

Please make sure to include these data files in the `sd2022/data` directory if you want to run the main notebooks [ex1_transit.ipynb](#), [ex2_rv.ipynb](#), [ex2_rv2.ipynb](#) and [ex2_rv3.ipynb](#).

Steps

You can either run the code locally using a `conda` environment or with a container using Docker. The Jupyter Notebooks associated with the different parts are located in the `sd2022/src` directory:

- TESS Lightcurve Planet Search: [ex1_lc.ipynb](#)
- Doppler Radial Velocity Planet Search: [ex2_rv_search.ipynb](#) (0 planets, stellar activity GP only), [ex2_rv2.ipynb](#) (1-2 planets, circular orbit) and [ex2_rv3.ipynb](#) (1-2 planets, eccentric orbit).

The Jupyter Notebooks will run faster locally on a high-spec computer (recommended).

Local Setup (Using Conda) [RECOMMENDED]

1. Clone the Repository:

Clone the repository to your local machine with the following command:

```
$ git clone https://gitlab.developers.cam.ac.uk/phy/data-intensive-scienc...
```

or simply download it from [Exoplanets - Coursework Repository \(sd2022\)](#).

2. Navigate to the Project Directory:

On your local machine, navigate to the project directory with the following command:

```
$ cd /full/path/to/sd2022
```

and replace `/full/path/to/` with the directory on your local machine where the repository lives in.

3. Setting up the Environment:

Set up and activate the `conda` environment with the following command:

```
$ conda env create -f environment.yml  
$ conda activate exo
```

4. Install ipykernel:

To run the notebook cells with `exo`, install the ipykernel package with the following command:

```
python -m ipykernel install --user --name exo --display-name "Python (e:
```

5. Open and Run the Notebook:

Open the `sd2022` directory with an integrated development environment (IDE), e.g. VSCode or PyCharm, select the kernel associated with the `exo` environment and run the Jupyter Notebooks (located in the `sd2022/src` directory).

Containerised Setup (Using Docker)

1. Clone the Repository:

Clone the repository to your local machine with the following command:

```
$ git clone https://gitlab.developers.cam.ac.uk/phy/data-intensive-scienc
```

or simply download it from [Exoplanets - Coursework Repository \(sd2022\)](#).

2. Navigate to the Project Directory:

On your local machine, navigate to the project directory with the following command:

```
$ cd /full/path/to/sd2022
```

and replace `/full/path/to/` with the directory on your local machine where the repository lives in.

3. Install and Run Docker:

You can install Docker from the official webpage under [Docker Download](#). Once installed,

make sure to run the Docker application.

4. Build the Docker Image:

You can build a Docker image with the following command:

```
$ docker build -t [image] .
```

and replace `[image]` with the name of the image you want to build.

5. Run a Container from the Image:

Once the image is built, you can run a container based on this image:

```
$ docker run -p 8888:8888 [image]
```

This command starts a container from the `[image]` image and maps port `8888` of the container to port `8888` on your local machine. The Jupyter Notebook server within the container will be accessible on JupyterLab at <http://localhost:8888>.

6. Access and Run the Notebook:

After running the container, you'll see logs in the terminal containing a URL with a token. It will look similar to this:

```
http://127.0.0.1:8888/lab?token=XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Navigate to <http://localhost:8888> and enter the token

`XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX`. Once you accessed JupyterLab, run the Jupyter Notebooks(located in the `sd2022/src` directory) with an `ipykernel` (Python 3).

Note: Make sure that no other Jupyter Notebook Servers are running. Otherwise, you might encounter 'Invalid credentials' issues when entering the token. Close any running Jupyter Notebook Servers. To stop a running server, use `Ctrl + C` in the terminal where you launched JupyterLab. Also make sure port `8888` is not occupied.

Support

For any questions, feedback, or assistance, please feel free to reach out via email at sd2022@cam.ac.uk.

License

This project is licensed under the [MIT License](#) - see the [LICENSE](#) file for details.

Project Status

The project is in a state ready for submission. All essential features have been implemented, and the codebase is stable. Future updates may focus on minor improvements, bug fixes, or optimisations.

Use of auto-generation tools

GitHub Co-Pilot assisted the author in producing all function docstrings present in the project repository. No specific commands have been given, instead auto-completion suggestions have occasionally been accepted. Other uses of auto-generation tools:

ChatGPT Prompt 1: How do I create a custom colour map in Python?

- ChatGPT output:

```
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors

# Define the colors for your colormap
colors = ["red", "green", "blue"]

# Create the colormap
cmap = mcolors.LinearSegmentedColormap.from_list("my_custom_cmap", colors)

# Use your colormap
plt.imshow([[0,1],[2,3]], cmap=cmap)
plt.colorbar()
plt.show()
```

- Modification of the output:

```
sd2022_gold = '#cfbd62'
sd2022_silver = '#C0C0C0'
sd2022_bronze = '#b08d57'
cmap_gold = mcolors.LinearSegmentedColormap.from_list('my_colormap', ['black', '#1f77b4', '#cfbd62'])
cmap_silver = mcolors.LinearSegmentedColormap.from_list('my_colormap', ['black', '#e0e0e0', '#C0C0C0'])
cmap_bronze = mcolors.LinearSegmentedColormap.from_list('my_colormap', ['black', '#b08d57', '#f0c2ab'])
cmap_bronze_r = mcolors.LinearSegmentedColormap.from_list('my_colormap', ['white', '#f0c2ab', '#b08d57'])
```

Authors and Acknowledgment

This project is maintained by [Steven Dillmann](#) at the University of Cambridge.

21st June 2024