Ivo Vladislavov Petrov

A5: Exoplanets

Coursework Report

Data-Intensive Science

March 28, 2024

Word count: 2938 (using Overleaf)

Contents

List of Figures

Chapter 1

Introduction

In recent years, the search for exoplanets—planets orbiting stars beyond our solar system—has evolved dramatically through developments in surveying telescopes and machine learning approaches. Among all detection techniques, the transit and radial velocity (RV) methods have succeeded in finding the most exoplanets. In particular, previous work has even managed to find exoplanets in stars with radii small as $0.3R_\odot$ [1]. In this work, we present our approach to detecting a planet candidate orbiting the brightest star of a 3-star system using data from the *Transiting Exoplanet Survey Satellite* (TESS) [2] using a combination of *gaussian processes* (GP) and nested sampling. We further demonstrate our methodology for detecting planetary systems using radial velocities on a synthetic dataset, while accounting for any stellar activity signals.

Chapter 2

Transiting planet in a 3-star system

In this section we describe how we evaluated the discovery of an exoplanet candidate around a multistellar system. We first filter the data from any stellar activity signals using a Gaussian process regression. Following that, we fit a transit model to the residuals using nested sampling and extract the relevant system and planetary properties. Below we explain the pipeline in detail and present the rationale behind the decisions.

2.1 Initial observations

First of all, we state some of our key assumptions. We assume the 3 stars shown in Sec. 2.1 are gravitationally connected. It has been given that the transiting exoplanet candidate is orbiting around star A, which seems to be more distant from the BC pair. Thus, we further assume a Keplerian orbit for the exoplanet due to the fact that any gravitational influence from B or C will be negligible.

We notice a variability in flux that seems to have a periodic nature, as can be directly observed in the light curve (Sec. 2.1). We will assume it is caused by stellar activity of the host star A. This is a more plausible alternative than other explored alternatives, such as an effect from a W UMa-type eclipsing binaries [3]. The main argument against the latter is that the observed flux changes seem to have a small amplitude (less than 1%), as well as the irregular nature of the signal. Therefore, in our interpretation of the GP parameter fit, we will assume the effects come from the activity of star A, as it is considerably brighter than both of the other stars ($m_{B+C} \approx m_A$).
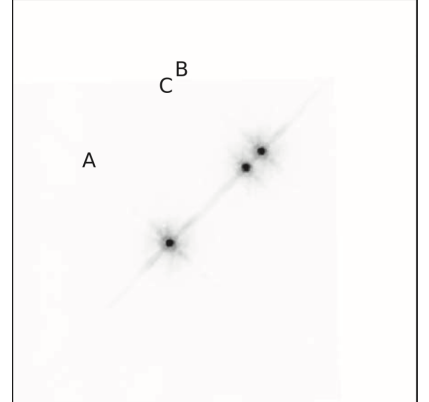
Figure 2.1: Stellar system high-fidelity image.

2.2 Background

Gaussian Processes Gaussian processes are supervised learning method that represents the data as a multivariate-normally distributed collection of points. They are parameterized through a kernel function $k(x_1, x_2)$, which determines the covariance of a pair of points. To this end, we utilize a GP with constant mean and the *Quasi-periodic kernel*, introduced in Nicholson and Aigrain [4]. Nicholson and Aigrain [4] also demonstrate strong correlations between the parameters of the GP and physical properties of the system, which we will explore in Sec. 2.4.

Figure 2.2: The light curve observed in the first epoch of the TESS survey (top). We show the filtered signal, as well as the fitted transits (bottom).

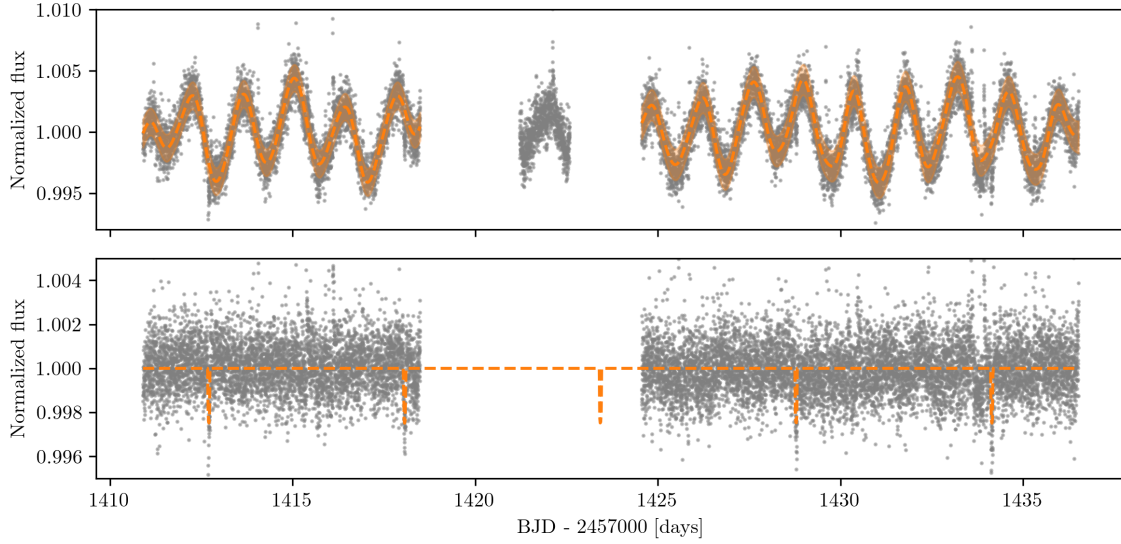**Definition 2.2.1** The *quasi-periodic* kernel can be parameterized by an amplitude A, period P, periodic length scale Γ and exponential length scale l, as shown below:

$$k(t_1, t_2) = k(\tau) = A^2 e^{-\frac{\tau^2}{2l^2}} e^{-\Gamma \sin^2(\frac{\pi\tau}{P})} \tag{2.1}$$

,where $\tau = |t_1 - t_2|$.

Light curve fitting   Light curve modelling techniques vary heavily in terms of complexity in detail - ranging from a simple box plot to capturing effects of the planet's atmosphere through transmission spectroscopy [5]. Considering our signal isn't sufficient for any complex modelling, we list the parameters we are interested in below. In particular, we employ the implementation provided by the *batman* package [6]. It provides utilities for modelling properties of the orbit and transit, assuming a quadratic limb darkening model for more precise fits.

- T_0 - the time of inferior conjunction.
- P - the transit period
- R_P - the radius of the planet (measured in $R_\odot$)
- a_P - the semi-major axis of the orbit (measured in $R_\odot$)
- i - the orbital inclination.
- e - the orbital eccentricity.
- w - the argument of periapse.
- u_1, u_2 - quadratic limb darkening coefficients.

Usually, a scale parameter is added to the light curve, however we noticed that fitting it only adds to the complexity of the model with no detectable benefit.

2.3 Methodology

We now present our methodology that utilises the aforementioned techniques

Activity filtering We first investigate the periodic nature of the stellar activity. To this end, we utilize a simple Lomb-Scargle periodogram [7] and report a periodic relationship with a period of $T = 1.39$ days. We notice that this matches the activity oscillations, emphasizing the need for their filtration. The relevant part of the periodogram can be found in Fig. 2.3. We now proceed to fit a zero-mean GP using the *scikit-learn* package. The fit also includes a hyperparameter optimization which maximizes the marginal log-likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \frac{1}{2}\boldsymbol{y}^T(\boldsymbol{K}+\sigma_y\boldsymbol{I})\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}+\sigma_y\boldsymbol{I}| - \frac{n}{2}\log 2\pi$$

Here we denote the covariance matrix as $\boldsymbol{K}$, assume a constant standard deviation $\sigma_y = \max_i \sigma_{y_i}$ and a number of datapoints n. We further denote the prediction of the fitted GP as $\boldsymbol{m}(\boldsymbol{t})$.

> **Remark:** A convenient property of GP kernels is that they are multiplicative, i.e for any valid kernels k_1, k_2, $k(t_1, t_2) = k_1(t_1, t_2) \times k_2(t_1, t_2)$ is also a valid kernel. Therefore, we implement the Quasi-periodic kernel as the product of the *constant kernel*, *exponential squared kernel* and the *periodic kernel*. We also add a trainable jitter term for numerical stability.
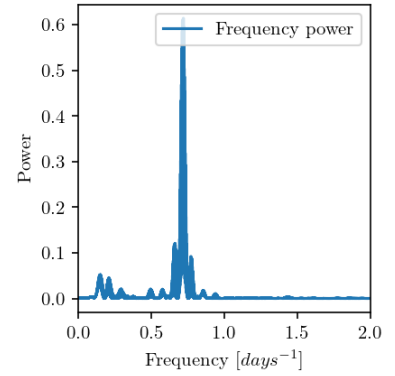

Figure 2.3: Lomb-Scargle periodogram for the light curve of the system.

We obtain a zero-mean flux by simply removing the mean of the entire sequence from the data. We further *mask out any outliers and estimated transit spots* for a robust fit estimation. Finally, we remove the GP fit from the original signal $\boldsymbol{y} \Leftarrow \boldsymbol{y} - \boldsymbol{m}(\boldsymbol{t})$, and also keep track of the GP errors associated with each point $\boldsymbol{\sigma}_{\mathrm{GP}} >> \boldsymbol{\sigma}_y$. We show the result of the fit in Sec. 2.1. This figure encapsulates the first epoch spanning between BJD+2458410 and BJD+2458435. We show the second epoch that covers dates between BJD+2459140 and BJD+2459170 in App. .1 .

2.4 Transit modelling

Having filtered the activity signal, we proceed to fitting the transit using `batman`. We perform the fit in a Bayesian framework, defining priors on each parameter $\pi(\boldsymbol{\theta})$, as well as a likelihood $\mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta})$. Using the nested sampling algorithm [8], implemented in the *dynesty* package [9], we obtain samples from the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{y}) \sim \mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Pre-processing Because the data is spread in a range of > 900 days, it is difficult to ensure a precise determination of the transit period and the mid-transit time using weakly informed priors. To this end, we first fit box plots in different sections and obtain a T_0 and P for each transit. We can therefore infer the epoch for each transit: epoch $= |\frac{T_0 - T_{0,\mathrm{first}}}{P}|$. We then fit the obtained T_0 versus the inferred epochs to obtain a more accurate measure for T_0 and P. In particular, we obtain $T_0 = 1412.7076 \pm 0.0037 BJD + 2457000$ days and $P = 5.358737 \pm 0.000038$
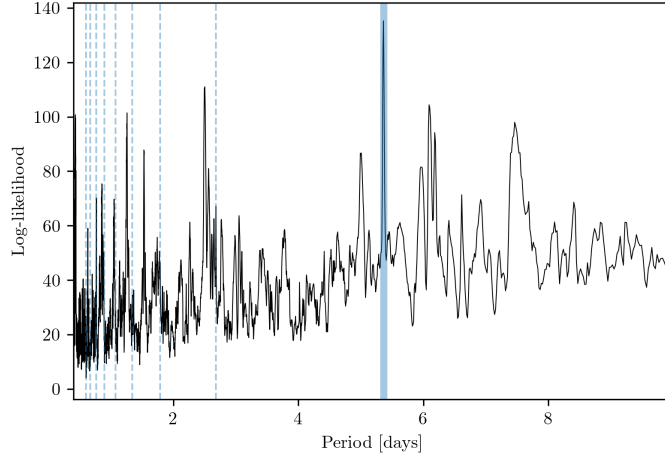
Figure 2.4: A periodogram for the observed transits determined by the box least squares fit.

days. An example periodogram for one such fit can be seen in Fig. 2.4. Considering these parameters are quite well defined with very small uncertainties, it is preferable to optimize the rest of parameters. To achieve that, we fold the light curve using the found T_0 and P assuming a linear relationship to merge the transits into a single light curve. The result of this operation, alongside the best fit can be found in Fig. 2.5.
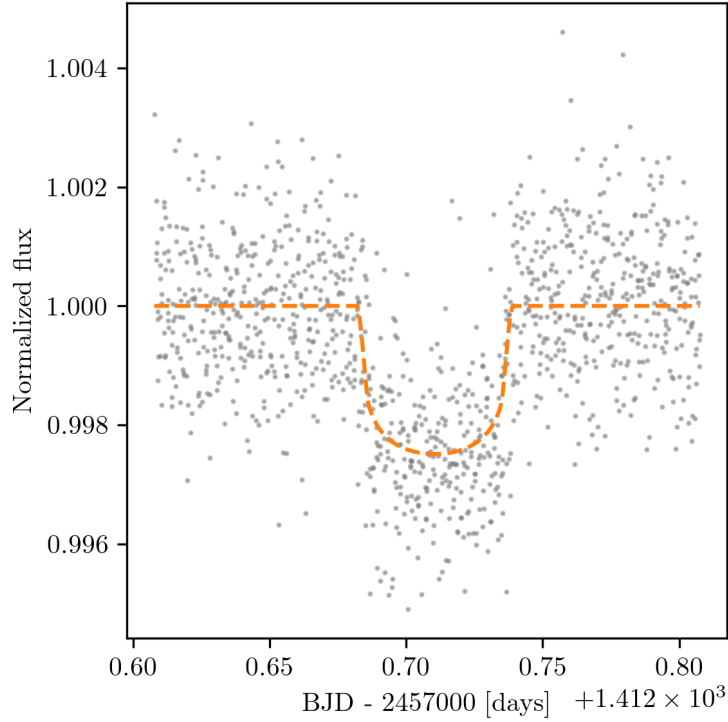
Figure 2.5: The fitted curve for the folded transit.

Nested Sampling We now proceed to fitting the light curve using Bayesian inference. We detail the priors for all parameters described in Sec. 2.2 in Table 2.1. We imbue a heavily informed prior on T_0 due to the high certainty we have obtained it with above, and do not fit for P due to having a single combined transit. We use weakly informed priors for the rest of the parameters. It is notable that we bound i to not go below 75°, as observed transits usually have $i \approx 90°$ (otherwise the signal would be weaker). Following [10], we apply an

5

eccentricity (e) prior using a Beta distribution which biases the parameter to take values close to 0 due to the unlikely nature of a highly elliptical orbit, as well as to ensure numerical stability. The bounds on a and R_p were determined through informed intuition for the parameters (a planet should not be as large as the star and must orbit in relatively close proximity to it).

Furthermore, we define a likelihood function, which assumes a Gaussian distribution around each point with a standard deviation:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_{\text{GP},i}} e^{-(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i(\boldsymbol{\theta}))^2/2\boldsymbol{\sigma}_{\text{GP},i}^2}$$

Where $\hat{\boldsymbol{y}}_i(\boldsymbol{\theta})$ is the prediction of the model with parameters $\boldsymbol{\theta}$. We utilize the errors from the GP, as they are higher than the measurement errors by more than an order of magnitude. We then iterate until convergence and report the inferred mean and 68% confidence interval for each parameter in Table 2.2. The fit using the mean parameters is visualized in Fig. 2.5. Diagnostic plots, such as the corner plot or trace plot are deferred to App. .1 , where we can see that the relevant parameters behave well.

Table 2.1: Prior distributions on relevant hyperparameters in the batman model.

Parameter	Prior	Values/bounds
T_0 [days][1]	Gaussian	$\mu = 0, \sigma = 4 \times 10^{-3}$
i [°]	Uniform	(75, 90)
a [$R_\odot$]	Uniform	(2.0, 100.0)
R_p [$R_\odot$]	Jeffrey's	(0.01, 0.3)
u_1	Uniform	(0, 1)
u_2	Uniform	(0, 1)
e	Beta	a=0.867, b=3.03
w [°]	Uniform	(-180, 180)

Table 2.2: Posterior mean and 68% confidence interval for relevant parameters. We also list any useful alternative units, incorporating accurate error propagation.

Parameter	Posterior	Alternative units
T_0 [BJD]	$2458412.71060 \pm 0.00054$	N/A
i [°]	$88.38^{+1.28}_{-1.17}$	N/A
a [$R_\odot$]	$23.793^{+6.02}_{-7.07}$	$0.030^{+0.008}_{-0.009}$ AU
R_p [$R_\odot$]	$0.0476^{+0.0025}_{-0.0027}$	$1.40 \pm 0.16 R_\oplus$
u_1	0.27 ± 0.18	N/A
u_2	$0.40^{+0.37}_{-0.3}$	N/A
e	$0.25^{+0.23}_{-0.19}$	N/A
w [°]	$-16.5^{+151.6}_{-133.6}$	N/A
P [days]	5.358737 ± 0.000038	N/A

2.5   Results

In this section, we briefly discuss the implications of the discovered parameter values.

The maximum likelihood estimation for the GP parameters . Nicholson and Aigrain [4] relate the period P to the stellar rotation period P_{rot}, attributing the deviation in flux to the rotation of dark spots. They further claim that the spot evolution time τ is highly correlated with the length scale l. Additional observations are made with regards to relating Γ to other stellar parameters, however it is difficult to make any conclusions given the reported uncertainty.

Therefore, we can claim that the star likely exhibits a rotation period $P_{rot} = \approx 1.43 \pm 0.06$ days, as well as a spot evolution time $\tau \approx 1.77 \pm 0.12$ days. We notice the former agrees well with the Lomb-Scargle periodogram described in Sec. 2.3, showing our findings are likely stable and consistent.

We find that the results are consistent with a planet that is slightly larger than Earth ($R_p = 1.4 R_\oplus$), which orbits A in very close proximity ($a = 0.03$ AU) with a high frequency ($P = 5.4$ days). We further notice that we have large margin of errors for parameters u_1, u_2, e, w. This likely due to the model being overdetermined for our specific dataset, meaning we can only poorly model the limb darkening and eccentricity effects with the current precision. We unfortunately cannot infer any properties belonging to B or C, as the signal received from them is too weak.

Chapter 3

Planetary system discovery using radial velocities

In this part, we demonstrate a simple approach for detecting exoplanets, while incorporating information about the stellar activity.

3.1 Background

Radial Velocities The radial velocity technique, also known as the Doppler spectroscopy method, is a powerful and widely used method for detecting exoplanets orbiting distant stars. This technique relies on observing the changes in the velocity of a star as it moves towards or away from Earth. When a planet orbits a star, conservation of angular momentum causes the star to oscillate slightly. This oscillation induces a periodic shift in the star's spectral lines due to the Doppler effect, from which properties of the planet can be inferred.

For modelling, we utilize the *radvel* package [11]. Multiple planets can be modelled at the same time, each having a different impact on the host star's radial velocity. In particular, we consider systems with 0, 1 or 2 planets.

Stellar Variability It has been shown that stellar activity can significantly change the measurement of radial velocities through sunspots, granulation or long-term magnetic cycles [12, 13]. Unfortunately, it is difficult to model the high-frequency granulation due to the sparse nature of the data.

We nonetheless attempt to model the other two effects through 2 spectrally-derived activity indicators of the cross-correlation function – the Full Width at Half-Maximum (FWHM), and the bisector inverse slope (BIS) [14]. To this end, we utilize Gaussian processes that can model the hidden activity, which are described in the next paragraph.

Gaussian Processes We utilize the same technique to model non-exoplanet related variability in the radial velocities, as described in Sec. 2.2, up to using the same Quasi-Periodic kernel. There are a few studies that incorporate this method in RV studies to achieve more stable results, even without the need of activity indicators [15, 16].

Nested Sampling We model our data in a Bayesian framework, obtaining a posterior distribution for the parameter set, from which we sample from. To this end, nested sampling, as implemented in *dynesty* [9], can be both used to obtain said samples, as well as the *evidence*

for the underlying model. The latter is key for model selection, where we can determine the relative validity of our models through the Bayes factor [17]:

> **Definition 3.1.1** The Bayes factor K for two models M_1, M_2 over a dataset $\mathcal{D}$ can be given as:
> $$K = \frac{P(\mathcal{D}|M_1)}{P(\mathcal{D}|M_2)} \tag{3.1}$$
> Here $P(\mathcal{D}|M_1)$ can be interpreted as the *evidence* for the data under the model assumptions.

3.2 Methodology

3.2.1 Trends and periodicity

Before proceeding to the radial velocity analysis. We observe a trend of increase in all variables (RV, FWHM, BIS), as shown in Fig. 3.1.
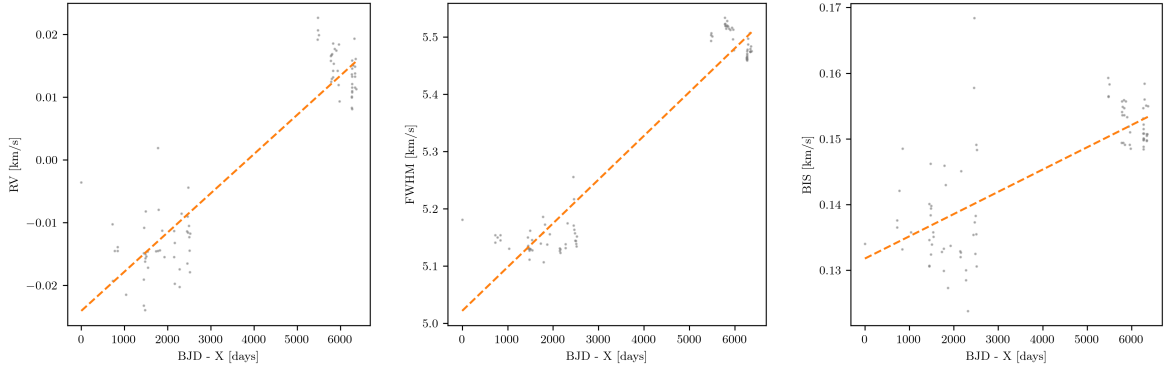

Figure 3.1: Trends for the radial velocity, FWHM and BIS.

Considering the activity indicators are also affected, we can assume this change is a long-term effect, as K-dwarf stars can exhibit magnetic cycles of more than 20 years [18]. While some periodicity can be observed which cannot be solely explained through acitivity or planetary effects, we attempt to discover the nature of said periodicity. We detrend the data with respect to time using a second degree polynomial and apply a Lomb-Scargle periodogram on the data, as seen in Fig. 3.2.

Unfortunately, no significant periodicity can be captured using this model, but we observe a strong peak around $P = 35.5$ days, which is within the expected rotation period of the star $P_{rot} = 31 \pm 10$ days. This is likely caused by the rotation of colder stellar spots. On the other hand, we can observe significant variability in nearby data points, likely showing deviations caused by planetary motion.

3.2.2 Modelling

We defined the model in a basis of $\log(P), T_c, \sqrt{e}\cos\omega, \sqrt{e}\cos\omega, k$. This basis ensures that the period is positive, and that eccentricity-related behaviour is well-defined. We utilize the same priors for multiple planets, which are listed in Table 3.1. We derived these priors by following the radvel
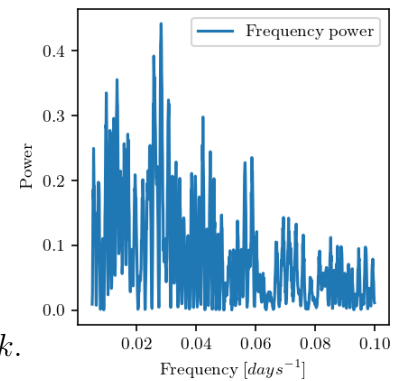

Figure 3.2: Lomb-Scargle periodogram for the BIS of the system.

Table 3.1: Prior and posterior distributions on relevant hyperparameters in the model.

Parameter	Prior	Values/bounds	Posterior
$A\ [\frac{km}{s}]$	Jeffrey's	$(0.001, 1)$	0.003 ± 0.001
$l\ [days]$	Gaussian	$\mu = 31\sqrt{2}, \sigma = 10\sqrt{2}$	$35.1^{+15.2}_{-28.2}$
$\tau\ [days]$	Gaussian	$\mu = 31, \sigma = 10$	32.3 ± 9.8
$l_{per}\ [days]$	Gaussian	$\mu = \sqrt{50}, \sigma = 1$	7.9 ± 0.9
$\log(P)\ [\log days]$	Gaussian	$\mu = 1, \sigma = 1$	$1.63^{+0.00}_{-0.93}$
$T_c\ [BJD]$	Gaussian	$\mu = 6000, \sigma = 3$	$5998.9^{+1.15}_{-0.9}$
$k\ [\frac{km}{s}]$	Uniform	$(10^{-5}, 0.1)$	$2.4^{+0.6}_{-1.3} \times 10^{-3}$
$\sigma_{INST1}\ [\frac{km}{s}]$	Jeffrey's	$(10^{-7}, 1)$	$0.003^{+0.002}_{-0.001}$
$\sigma_{INST2-4}\ [\frac{km}{s}]$	Jeffrey's	$(10^{-7}, 1)$	$0.0019^{+0.0011}_{-0.0014}$
$\sqrt{e}\cos\omega$	Uniform	$(0, 0.2)$	0.08 ± 0.07
$\sqrt{e}\sin\omega$	Uniform	$(0, 0.2)$	$0.12^{+0.05}_{-0.08}$

guidelines when fitting Gaussian processes. The Gaussian process model described in Sec. 3.1 is parameterized by an amplitude A, period τ, exponential length l and periodic length l_{per}. By definition l_{per} is equivalent to Γ, as described in Sec. 2.2, but the value is inversed. Additional terms $\gamma_{INST_i}, \sigma_{INST_i}$ for modelling the different spectrograph mean and jitter were added for better inference. The priors for these parameters are given in Table 3.1. We fit a model, and compute the corresponding evidence - obtaining $\log(Z_0) = 331.067 \pm 0.076$, $\log(Z_1) = 334.074 \pm 0.129$, and $\log(Z_2) = 334.566 \pm 0.105$ for 0, 1 and 2 planets respectively. This results in a Bayes factor of $K_{1,0} = \frac{Z_1}{Z_0} = e^{\log(Z_1) - \log(Z_0)} = 20.22$ and $K_{2,1} = 1.63$. These results imply a high likelihood for the existence of a planet, but not for a second one.

3.3 Results

Having confirmed the likely presence of a single planet, we now infer the properties of said planet using the native *radvel* Markov Chain Monte Carlo method. The obtained posteriors are shown in Table 3.1. From this we can infer that the orbiting planet has a semi-amplitude of $a = 0.0512^{+0.0003}_{-0.0055} AU$ and a mass of $M_p = 5.8^{+3.1}_{-4.2} M_J$, showing the impressive capabilities of RV-based methods. This planet is very likely to be a hot Jupiter, which is to be expected considering the short orbital period.

Bibliography

[1] JM Almenara, X Bonfils, EM Bryant, A Jordán, G Hébrard, E Martioli, ACM Correia, N Astudillo-Defru, C Cadieux, L Arnold, et al. Toi-4860 b, a short-period giant planet transiting an m3. 5 dwarf. *Astronomy & Astrophysics*, 683:A166, 2024.

[2] George R Ricker, Joshua N Winn, Roland Vanderspek, David W Latham, Gáspár Á Bakos, Jacob L Bean, Zachory K Berta-Thompson, Timothy M Brown, Lars Buchhave, Nathaniel R Butler, et al. Transiting exoplanet survey satellite. *Journal of Astronomical Telescopes, Instruments, and Systems*, 1(1):014003–014003, 2015.

[3] L Binnendijk. The orbital elements of w ursae majoris systems. *vistas in Astronomy*, 12:217–256, 1970.

[4] Belinda A Nicholson and Suzanne Aigrain. Quasi-periodic gaussian processes for stellar activity: From physical to kernel parameters. *Monthly Notices of the Royal Astronomical Society*, 515(4):5251–5266, 2022.

[5] Lisa Kaltenegger, Zifan Lin, and Sarah Rugheimer. Finding signs of life on transiting earthlike planets: high-resolution transmission spectra of earth through time around fgkm host stars. *The Astrophysical Journal*, 904(1):10, 2020.

[6] Laura Kreidberg. batman: basic transit model calculation in python. *Publications of the Astronomical Society of the Pacific*, 127(957):1161, 2015.

[7] Jacob T VanderPlas. Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplement Series*, 236(1):16, 2018.

[8] John Skilling. Nested sampling. *Bayesian inference and maximum entropy methods in science and engineering*, 735:395–405, 2004.

[9] Joshua S Speagle. dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, 493(3):3132–3158, 2020.

[10] David M Kipping. Parametrizing the exoplanet eccentricity distribution with the beta distribution. *Monthly Notices of the Royal Astronomical Society: Letters*, 434(1):L51–L55, 2013.

[11] Benjamin J Fulton, Erik A Petigura, Sarah Blunt, and Evan Sinukoff. Radvel: the radial velocity modeling toolkit. *Publications of the Astronomical Society of the Pacific*, 130(986):044504, 2018.

[12] Jean C Costes, Christopher A Watson, Ernst de Mooij, Steven H Saar, Xavier Dumusque, Collier Cameron, David F Phillips, Maximilian N Günther, James S Jenkins, Annelies Mortier, et al. Long-term stellar activity variations and their effect on radial-velocity measurements. *Monthly Notices of the Royal Astronomical Society*, 505(1):830–850, 2021.

[13] Nadège Meunier. Stellar variability in radial velocity. *arXiv preprint arXiv:2104.06072*, 2021.

[14] Nuno C Santos, Michel Mayor, Dominique Naef, Didier Queloz, and Stephane Udry. Bisector analysis as a diagnostic of intrinsic radial-velocity variations. In *Symposium-International Astronomical Union*, volume 202, pages 121–123. Cambridge University Press, 2004.

[15] Vinesh Rajpaul, Suzanne Aigrain, Michael A Osborne, Steven Reece, and S Roberts. A gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3):2269–2291, 2015.

[16] Samuel K Grunblatt, Andrew W Howard, and Rapha"elle D Haywood. Determining the mass of kepler-78b with nonparametric gaussian process estimation. *The Astrophysical Journal*, 808(2):127, 2015.

[17] Zoltan Dienes. Using bayes to get the most out of non-significant results. *Frontiers in psychology*, 5:85883, 2014.

[18] S Boro Saikia, Sandra V Jeffers, Julien Morin, P Petit, CP Folsom, SC Marsden, J-F Donati, R Cameron, JC Hall, V Perdelwitz, et al. A solar-like magnetic cycle on the mature k-dwarf 61 cygni a (hd 201091). *Astronomy & Astrophysics*, 594:A29, 2016.

Appendix

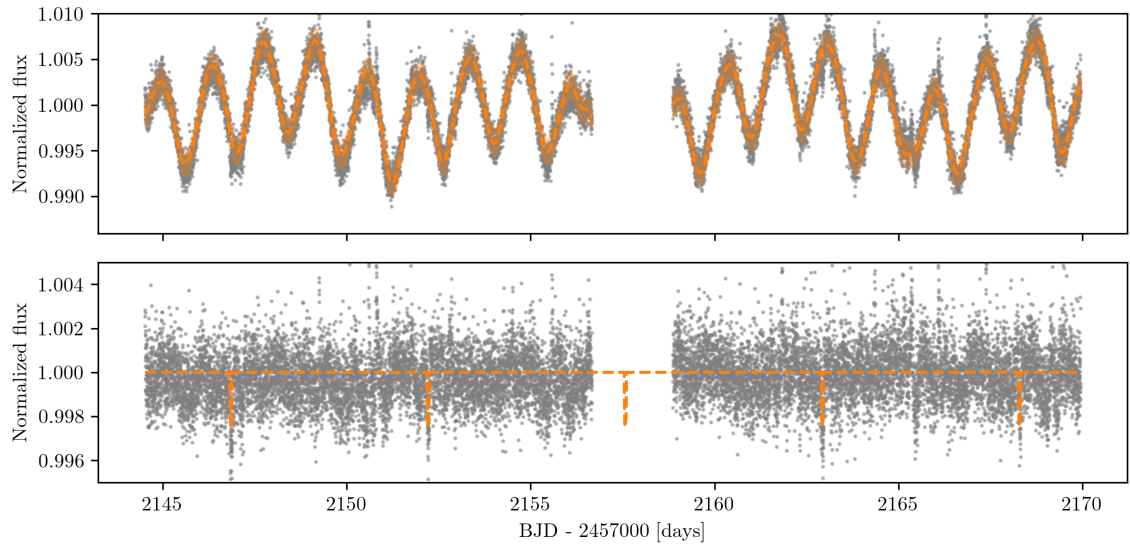## .1 Deferred figures (Transit)

Figure 3: The light curve observed in the second epoch of the TESS survey (top). We show the filtered signal, as well as the fitted transits (bottom).
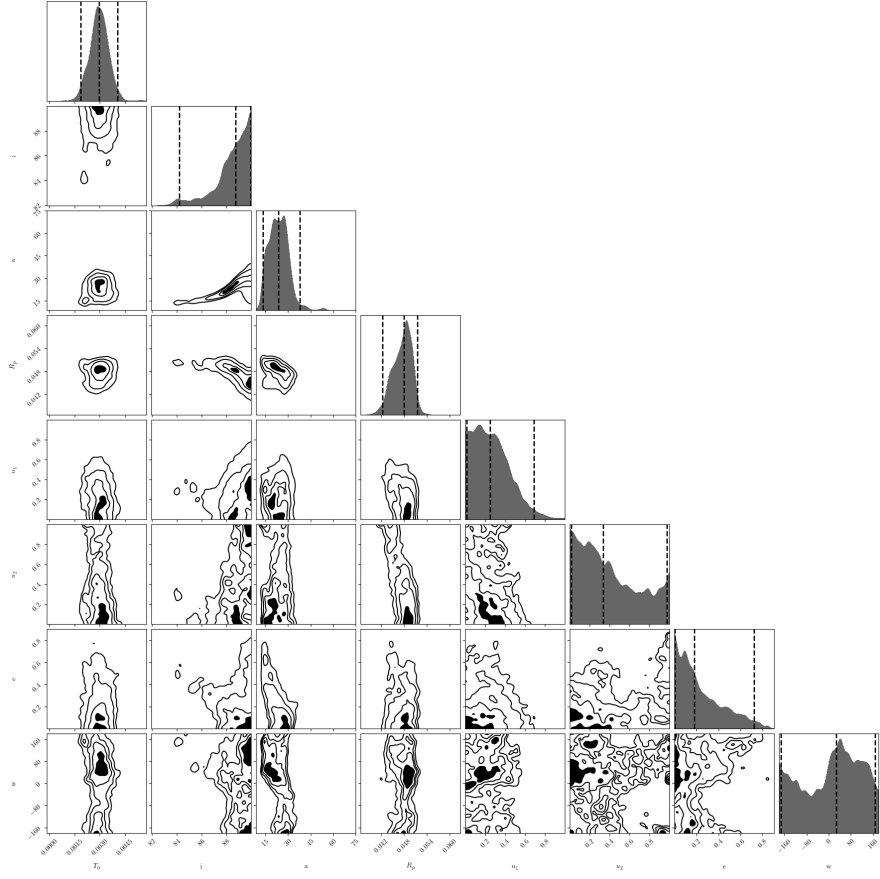
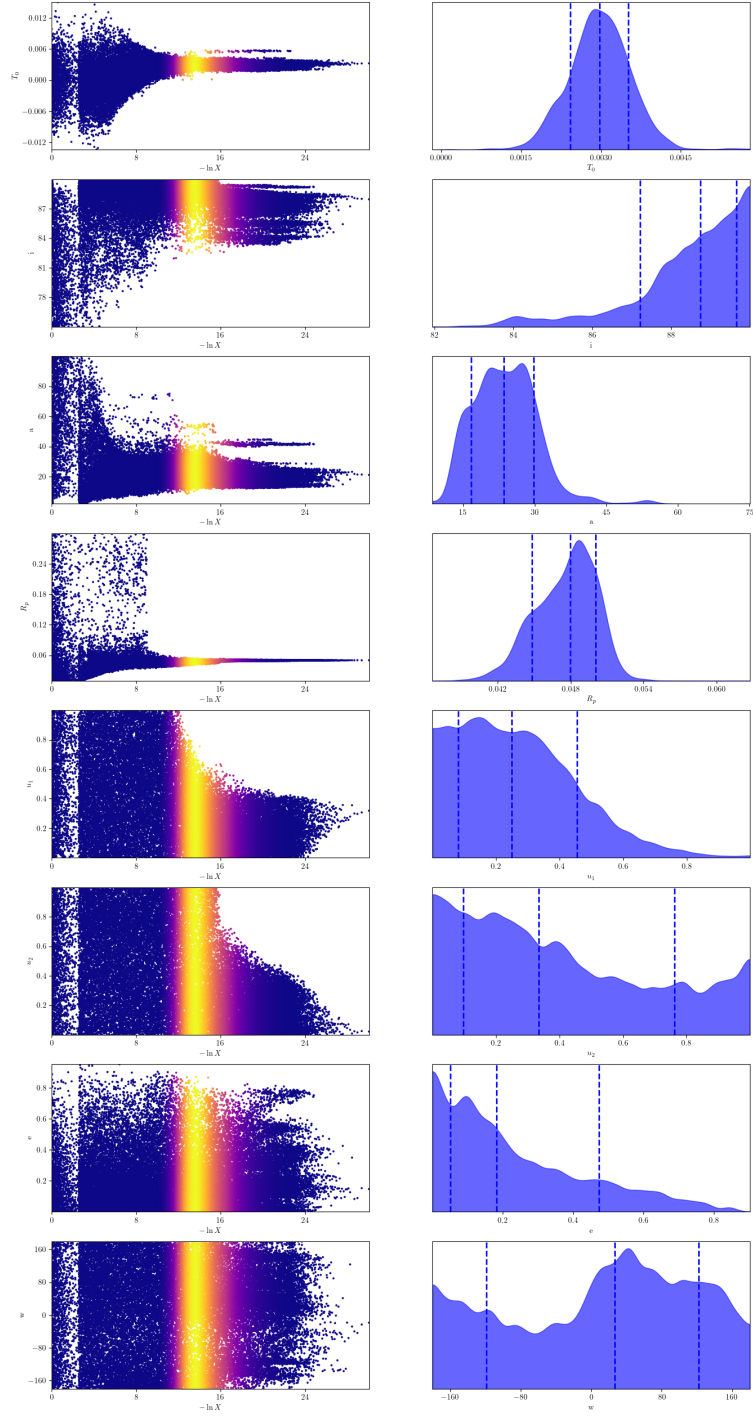Figure 4: Corner plot for the sampled posterior distribution.

Figure 5: Traceplot for the sampled posterior distribution.