Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 866–883 (2022)
Advance Access publication 2021 October 7

https://doi.org/10.1093/mnras/stab2889

PYANETI – II. A multidimensional Gaussian process approach to analysing spectroscopic time-series

Oscar Barragán [1]★ Suzanne Aigrain,[1] Vinesh M. Rajpaul [2] and Norbert Zicher[1]

[1]*Sub-department of Astrophysics, Department of Physics, University of Oxford, Oxford OX1 3RH, UK*
[2]*Astrophysics Group, Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0HE, UK*

Accepted 2021 October 4. Received 2021 September 28; in original form 2021 August 17

ABSTRACT

The two most successful methods for exoplanet detection rely on the detection of planetary signals in photometric and radial velocity time-series. This depends on numerical techniques that exploit the synergy between data and theory to estimate planetary, orbital, and/or stellar parameters. In this work, we present a new version of the exoplanet modelling code `pyaneti`. This new release has a special emphasis on the modelling of stellar signals in radial velocity time-series. The code has a built-in multidimensional Gaussian process approach to modelling radial velocity and activity indicator time-series with different underlying covariance functions. This new version of the code also allows multiband and single transit modelling; it runs on `Python 3`, and features overall improvements in performance. We describe the new implementation and provide tests to validate the new routines that have direct application to exoplanet detection and characterization. We have made the code public and freely available at https://github.com/oscaribv/pyaneti. We also present the codes `citlalicue` and `citlalatonac` that allow one to create synthetic photometric and spectroscopic time-series, respectively, with planetary and stellar-like signals.

Key words: methods: numerical – techniques: photometry – techniques: spectroscopy – planets and satellites: general.

1 INTRODUCTION

We are now living in a fascinating era in which we know that 'other Suns' are the centres of concentric systems of many worlds (cf. Bruno 1584). Over the past few decades, astronomers have developed techniques that allow us to 'translate' photons into worlds (see e.g. Struve 1952; Bozza, Mancini & Sozzetti 2016). More than 4800[1] exoplanets discovered to date have shown that worlds are abundant and diverse. Interestingly, most of these exoplanets have been discovered indirectly by observing planet-induced variations in their host stars. The transit (Henry et al. 1999; Charbonneau et al. 2000) and radial velocity (RV; Mayor & Queloz 1995) methods are the most successful techniques today in terms of number of discovered exoplanets. However, the current number of discovered and characterized exoplanets is dwarfed by the estimated number of exoplanets in our Galaxy (e.g. Petigura, Howard & Marcy 2013; Batalha 2014). Present and future exoplanet search and characterization instruments, such as *TESS* (Ricker et al. 2015), *CHEOPS* (Broeg et al. 2013), *PLATO* (Rauer et al. 2014), *ESPRESSO* (Pepe et al. 2010), and *SPIRou* (Donati et al. 2018), will provide us with a plethora of photometric and spectroscopic data with exoplanet induced signals waiting to be discovered.

Light curves and RVs are compared with models to infer planetary, orbital, and/or stellar parameters. These analyses are generally performed numerically using diverse models combined with computational techniques. Fortunately, the wealth of exoplanetary data has spurred the development of a variety of exoplanet numerical tools. To our knowledge, the codes that allow one to analyse jointly RV and transit data are: `allesfitter` (Günther & Daylan 2020), `EXOFAST` (Eastman, Gaudi & Agol 2013; Eastman et al. 2019), `exoplanet` (Foreman-Mackey et al. 2021), `ExoStricker` (Trifonov 2019), `juliet` (Espinoza, Kossakowski & Brahm 2019), `MCMCI` (Bonfanti & Gillon 2020), `PASTIS` (Díaz et al. 2014; Santerne et al. 2015), `PlanetPAck` (Baluev 2013), `TLCM` (Csizmadia 2020), and `pyaneti` (Barragán, Gandolfi & Antoniciello 2019). These software packages cover a wide range of programming languages and models, and have been used extensively in the literature.

One of the current challenges in exoplanet detection is related to stellar signals in our data. Particularly, RV variations caused by stellar activity jeopardize our ability to detect planet-induced Doppler signals (e.g. Queloz et al. 2001; Rajpaul, Aigrain & Roberts 2016). Some methods try to remove stellar activity during the RV extraction to produce activity-free RV times-series (e.g. Collier Cameron et al. 2020; Cretignier et al. 2020; Rajpaul, Aigrain & Buchhave 2020). Others attempt to filter the activity induced signals in the RV time-series (e.g. Hatzes et al. 2010, 2011; Pepe et al. 2013; Barragán et al. 2018). Gaussian Processes (GPs) have become a widely used tool to model activity induced RVs given their ability to describe stochastic variations (e.g. Haywood et al. 2014; Grunblatt, Howard & Haywood 2015; Fulton et al. 2018). However, the flexibility that GPs offer may also be their major drawback if not used carefully. Rajpaul et al. (2015) proposed a method of using spectroscopic activity-indicators together with RVs in order to constrain the activity induced signal in the RV time-series. This can be done by extending the GP approach to a multidimensional GP that exploits the correlations between the

★ E-mail: oscaribv@gmail.com, oscar.barragan@physics.ox.ac.uk
[1]As of Aug 11, 2021, http://exoplanet.eu.

different time-series. This method has proven useful in disentangling planetary and stellar induced signals in RV data with different levels of activity (e.g. Barragán et al. 2019; Mayo et al. 2019).

In this work, we present a new version of the multiplanet modelling code pyaneti[2] (Barragán et al. 2019). This updated version of the code focuses on multidimensional GP regression in order to model planetary and activity induced signals in spectroscopic times-series. This new version also allows one to model multiband transits and single transit events, and features various performance improvements. This new version of pyaneti has already been used in recent exoplanet characterization works (e.g. Carleo et al. 2020; Eisner et al. 2021, 2020; Georgieva et al. 2021).

This manuscript is part of a series of papers under the project *GPRV: Overcoming stellar activity in radial velocity planet searches* funded by the European Research Council (ERC, P.I. S. Aigrain). The paper is organized as follows: for the sake of self-completeness, we provide a short recap on Gaussian processes in Section 2. Section 3 describes the multidimensional GPs, with a special emphasis on connecting the activity indicators and the RV time-series. We describe the new implementation of pyaneti in Section 4. Section 5 describes the tests used to validate the code and we conclude in Section 6.

2 A BRIEF OVERVIEW OF GAUSSIAN PROCESSES

In this manuscript we do not provide a detailed description of Gaussian processes. Instead, we will provide the basics of GPs needed in order to apply them in data analysis, specifically, in the context of RV and light curve modelling. For further details, we advise the reader to consult specialist literature (e.g. Rasmussen & Williams 2006; Roberts et al. 2013).

A *stochastic process* is a system which evolves in continuous space (in our case time) while undergoing fluctuations. It is possible to describe the system as a finite set of random variables that are related by a given mathematical entity (Coleman 1974). If the mathematical object that describes the relation between the random variables is a multivariate normal distribution, then the stochastic process is a *Gaussian process*. Following Tracey & Wolpert (2018), a GP assumes that the marginal joint distribution of function values at any finite set of input locations, $\boldsymbol{t} = t_{i,(i=1,\cdots,N)}$, is given by a multivariate Gaussian distribution

$$P(\boldsymbol{t}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{K}|}} \exp\left[-\frac{1}{2}(\boldsymbol{t}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{K}^{-1} (\boldsymbol{t}-\boldsymbol{\mu})\right], \quad (1)$$

where $\boldsymbol{\mu}$ is a vector containing mean values, and $\boldsymbol{K}$ is a matrix containing the information about the correlation between the variables. The only condition about the matrix $\boldsymbol{K}$ is that it has to be symmetric and positive semidefinite. We note that for a given $\boldsymbol{\mu}$ and $\boldsymbol{K}$ we can draw an infinite number of curves as random samples of equation (1). As these curves are not characterized by explicit sets of parameters, GPs are referred as non-parametric functions (see e.g. Roberts et al. 2013, for more details).

The important aspect is then to find a way to compute our $\boldsymbol{\mu}$ and $\boldsymbol{K}$ entities that describe a particular GP. One advantage of GPs being defined over a continuous space is that the mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{K}$, can be computed from evaluations of continuous parametric functions at the positions $\boldsymbol{t}$. Equation (1) depends only on $\boldsymbol{\mu}$ and $\boldsymbol{K}$; therefore, a GP can be fully described with a mean and a covariance kernel function (see Rasmussen & Williams 2006, for more details).

[2]From the Italian word *pianeti*, which means *planets*.

2.1 Mean and covariance functions

Mean functions are the deterministic part of a GP. It can be any function $\mu(t;\boldsymbol{\phi})$ that depends on a set of parameters, $\boldsymbol{\phi}$, and the variable describing the continuous space, t. For example, a mean function can be a straight line, a sinusoid, a Keplerian, or a transit model.

A covariance (also called kernel) function $\gamma(t_i, t_j; \boldsymbol{\Phi})$ describes how two locations, t_i and t_j, are related according to some parameters $\boldsymbol{\Phi}$. Such kernel functions can be tuned in order to describe physical/instrumental signals, such as noise, periodicity, long-term evolution, etc. We describe below some examples of covariance kernel functions widely used in astronomical literature.

One of the simplest covariance matrix is computed with the white noise kernel

$$\gamma_{\mathrm{WN}}(t_i, t_j) = \sigma_i^2 \delta_{ij}, \quad (2)$$

where σ_i is the error associated with the datum i and δ_{ij} is the Kronecker delta. This kernel creates a diagonal covariance matrix, and is used to take into account uncertainties in data. Another widely used kernel is the squared exponential

$$\gamma_{\mathrm{SE}}(t_i, t_j) = A^2 \exp\left(-\frac{|t_i - t_j|^2}{2\lambda^2}\right), \quad (3)$$

where A is an amplitude that works as a scale factor that determines the typical deviation from the mean function, and λ is the length-scale, which can be interpreted as the characteristic distance for which two points are strongly correlated. This kernel generates smooth functions with a typical length-scale λ. Fig. 1 shows some examples of functions drawn using the γ_{SE} kernel and different mean functions.

Also widely used are the Matérn family of kernels. They are based on the standard Gamma function and the modified Bessel function of second order (see e.g. Rasmussen & Williams 2006, for more details). Two examples of the Matérn kernels are the Matérn 3/2 Kernel

$$\gamma_{\mathrm{M32}}(t_i, t_j) = A^2 \left(1 + t_{3/2}\right) \exp\left(-t_{3/2}\right), \quad (4)$$

with $t_{3/2} \equiv \sqrt{3}|t_i - t_j|\lambda^{-1}$, and the Matérn 5/2 Kernel

$$\gamma_{\mathrm{M52}}(t_i, t_j) = A^2 \left(1 + t_{5/2} + \frac{t_{5/2}^2}{3}\right) \exp\left(-t_{5/2}\right), \quad (5)$$

with $t_{5/2} \equiv \sqrt{5}|t_i - t_j|\lambda^{-1}$. The parameters A and λ have the same role as for the Squared Exponential kernel, but in these cases the resulting functions are less smooth. Fig. 1 shows GP samples drawn using the γ_{M32} kernel.

A widely used kernel in astronomy, especially in exoplanet research, is the Quasi-Periodic (QP) kernel (as defined by Roberts et al. 2013)

$$\gamma_{\mathrm{QP}}(t_i, t_j) = A^2 \exp\left\{-\frac{\sin^2\left[\pi\left(t_i - t_j\right)/P_{\mathrm{GP}}\right]}{2\lambda_{\mathrm{p}}^2} - \frac{\left(t_i - t_j\right)^2}{2\lambda_{\mathrm{e}}^2}\right\}, \quad (6)$$

where A has the same meaning as for the Squared Exponential kernel, P_{GP} is the characteristic period of the GP, λ_{p} the inverse of the harmonic complexity (how complex variations are inside each period), and λ_{e} is the long-term evolution time-scale (similar to the λ for the squared exponential kernel). We show some examples of functions created using the γ_{QP} kernel in Fig. 1.

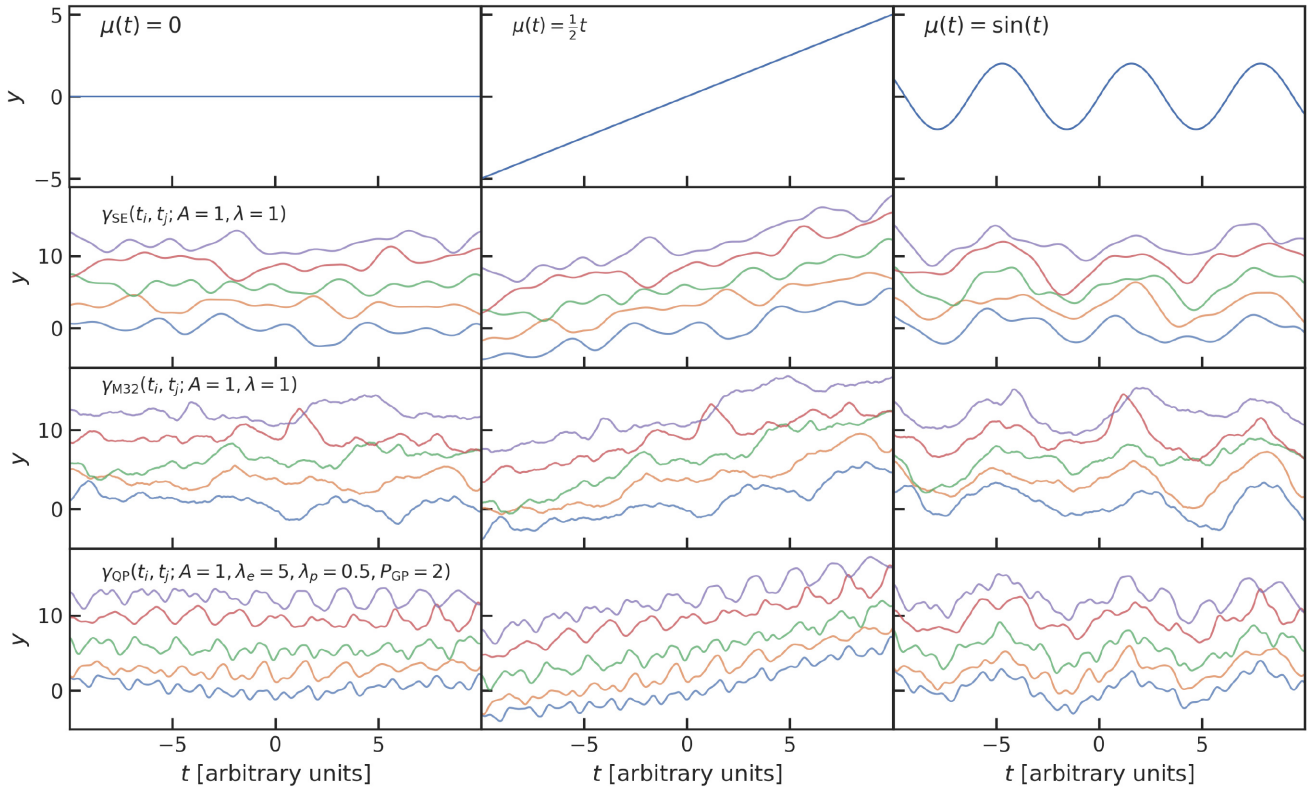

Figure 1. Example of functions generated by GPs with different mean and covariance functions. Left-hand, middle, and right-hand panel corresponds to GPs with mean functions $\mu = 0$, $\mu = 1/2\,t$, and $\mu = \sin(t)$, respectively. Top panel shows a plot with the respective mean function. From top to bottom, the second, third, and fourth panels show five GPs samples from square exponential (with $A = 1$, $\lambda = 1$), Matérn 3/2 (with $A = 1$, $\lambda = 1$), and Quasi-Periodic kernels (with $A = 1$, $\lambda_e = 5$, $\lambda_p = 0.5$, $P_{GP} = 2$), respectively.

Because the QP kernel generates stochastic periodic signal, this choice of covariance function is widely used to model stellar activity signals in both photometry and RVs (Haywood et al. 2014; Rajpaul et al. 2015). In a general context the GP period, P_{GP}, can be interpreted as the stellar rotation period, the long-term evolution time-scale, λ_e, can be associated with the active region lifetime on the stellar surface; and the inverse harmonic complexity, λ_p, can be associated with the activity regions distribution on the stellar surface (see e.g. Aigrain et al. 2015).

We note that for the cases in which the GP has a relatively small evolution time-scale ($\lambda_e \lesssim P_{GP}$), the periodicity of the GP is practically irrelevant (as pointed out by Rajpaul et al. 2015). Therefore, special care has to be taken when dealing with signals in which the evolution time scale is smaller than the expected periodicity. It is better to use a QP kernel only in cases when $P_{GP} < \lambda_e$. If that is not case, the QP kernel might not be appropriate and some other kernels should be considered. In this work we ensure that when we use the QP kernel, $\lambda_e > P_{GP}$ is satisfied.

Fig. 1 also shows an example of the non-parametric behaviour of the GPs (see Rasmussen & Williams 2006; Roberts et al. 2013, for more details). While in a parametric deterministic model a given set of parameters will give always the same curve, in the non-parametric case, a given set of parameters can give different curves with the condition that the random variables satisfy their intrinsic correlation. Given that the parameters of GPs do not have the same interpretation as for parametric functions, they are often called *hyper -parameters*.

2.2 Gaussian process regression

We can perform regression using GPs if we assume that our data (a finite set of variables $\boldsymbol{y}$, taken at times $\boldsymbol{t}$) are samples of a GP. The mean function μ can be a physically motivated parametric model (e.g. Keplerian or transit curves), and the covariance kernel function, $\gamma(t_i, t_j)$, can encompass any intrinsic correlation in our data set (e.g. stellar activity and/or instrumental systematics).

Given that a finite set of variables drawn from a GP is described by a multivariate normal distribution, we can use this property to write a logarithmic Gaussian likelihood to marginalize over variables as

$$\ln \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\Phi}) = -\frac{1}{2}\left(N_{\rm obs} \ln 2\pi + \ln |\boldsymbol{K}| + \boldsymbol{r}^{\intercal} \boldsymbol{K}^{-1} \boldsymbol{r}\right), \tag{7}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\Phi}$ are the mean and covariance functions parameters, respectively; $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{\mu}$ is the vector of residuals of the data $\boldsymbol{y}$ and the mean function $\boldsymbol{\mu} = \mu(\boldsymbol{t})$ evaluated at the times $\boldsymbol{t}$, $\boldsymbol{K}$ is the covariance matrix for the observations $\boldsymbol{t}$ given a kernel function $\gamma(t_i, t_j)$, and $N_{\rm obs}$ is the number of observations. Equation (7) can be optimized or sampled with different numerical techniques in order to infer the parameters of the mean and covariance functions.

GP regression is the cornerstone of the updated version of `pyaneti`. Mean functions can be constructed easily with Keplerian and transit models, while any other intrinsic correlation can be absorbed by the covariance function. If we use a physically motivated kernel function, we can also learn about the underlying mechanism that gives rise to the correlation (e.g. stellar activity). Section 4 describes how GP regression is included inside `pyaneti`.

3 MULTIDIMENSIONAL GAUSSIAN PROCESSES

We have described how we can use a GP to describe non-parametric functions over a continuous space on $\mathbb{R}$ (in our case time). It is possible to extend this idea and to assume that the same GP can describe correlations in multiple continuous spaces that may be related to each other.

Multidimensional (also called multivariate) GPs provide a solid and unified framework to make the prediction of GPs on $\mathbb{R}^N$, where N is the number of dimensions (for more details about the mathematical formalism of multi-dimensional GPs see e.g. Alvarez, Rosasco & Lawrence 2011; Chen, Fan & Wang 2020). An useful application of multi-dimensional GP is regression.

The general idea of multidimensional GP regression is to transform the multidimensional problem into a 'big' 1D GP regression. This is done by vectorizing the set of observations $[\boldsymbol{t}, \boldsymbol{y}]_i$ (where $i = 1, \cdots, N$) as a big vector in $\mathbb{R}$. The residual vector $\boldsymbol{r}$ is created as a concatenation of residuals vectors $\boldsymbol{r}_i$, for each dimension i. The covariance matrix $\boldsymbol{K}$ is constructed of small sub-matrices $\boldsymbol{k}^{l,m}$ that describe the correlations between the different dimensions l and m. This allows one to reformulate the multidimensional GP regression as a conventional GP regression (see Section 2.2). In the remainder of this section we will describe how we can use multidimensional GPs to model activity induced signals in spectroscopic time-series.

3.1 Multidimensional GPs for spectroscopic time-series

Rajpaul et al. (2015) proposed a framework to model stellar activity in RV time-series simultaneously with the activity indicators using a multidimensional GP approach. This approach assumes that the stellar induced signals in all observables can be described by the same latent GP and its time derivative. Jones et al. (2017) and Gilbertson et al. (2020) expanded the work of Rajpaul et al. (2015) by adding higher derivatives and generalizing the work to a generic set of activity indicators. In this work we also generalize the work of Rajpaul et al. (2015) to a generic set of activity indicators, but we maintain the approach of using only the first GP derivative.

3.1.1 *Physical motivation*

It has been shown that there is an intrinsic relation between the area covered by active regions on the stellar surface and the stellar-induced RV variations (e.g. Boisse et al. 2009; Aigrain, Pont & Zucker 2012). Following this idea, we can assume that we can describe a function, $G(t)$, which is a latent unobserved variable that represents the projected area of the visible stellar disc that is covered by active regions as function of time. Such active regions affect photometric and spectroscopic observed parameters (e.g. RVs or activity indicators) in different ways. Some of them are only affected by the projected area that is covered by active regions, i.e. they can be described as $G(t)$ with some scale factor; others are also affected by how these regions evolve in time on the stellar surface, i.e. they are described by $G(t)$ and its time derivatives ($\dot{G}(t)$, $\ddot{G}(t)$, etc.).

RVs are affected by the position of the active regions on the stellar surface and how these regions evolve in time (see e.g. Dumusque, Boisse & Santos 2014). Therefore, in our assumption that $G(t)$ describes the area covered by active regions of the visible stellar disc, activity-induced RV data can be described by $G(t)$ (e.g. to account for the convective blueshift) and its time derivatives (to account for the evolution of the spots on the stellar surface). Some activity indicators, such as $\log R'_{\rm HK}$(flux of the Calcium II H & K lines relative to the bolometric flux) or $S_{\rm HK}$ (flux of the Calcium II H & K lines relative to the local continuum), are only affected by the fraction of the area that is covered by the stellar surface (see e.g. Isaacson & Fischer 2010; Thompson et al. 2017), i.e. they can be described only by $G(t)$. Some other activity indicators, such as the bisector inverse slope (BIS), are also affected by how the active regions evolve on the stellar surface (see e.g. Dumusque et al. 2014), requiring higher time derivatives of $G(t)$ in order to describe them.

A set of contemporaneous time-series that contain activity-induced signals of a given star (RVs, $\log R'_{\rm HK}$, etc.) can then be modelled simultaneously assuming they are described by the same underlying function $G(t)$ and its derivatives. We can assume that our function $G(t)$ is generated by a GP, given its flexibility to model stochastic signals and the property that any affine operator (including linear and/or derivative) applied to a GP yields another GP (see Rajpaul et al. 2015, and references therein). We also note that, if we assume that our underlying function comes from a GP, we can use a multidimensional GP approach to exploit the correlation between observations in the different time-series, assuming each one is a different continuous space described by the same underlying GP-drawn function. The advantage of this approach is that RV time-series contain stellar activity and planet-induced variations, while activity indicators are only sensitive to activity. Therefore, activity indicators can help to constrain the activity induced signal in RV time-series, thus allowing one to disentangle the planetary signals from the activity signals (Rajpaul et al. 2015).

3.1.2 *Theoretical approach*

We follow the approach described by Rajpaul et al. (2015) and we will assume that we have a set of N time-series, $\mathcal{A}_{i=1,\cdots,N}$, each one with M points. Although this is not necessary for the multidimensional GP approach, we make this assumption given that for spectroscopic time-series the RVs and activity indicators are computed from the same spectra. A set of N time-series that are characterized by the same GP-drawn function $G(t)$ and its derivative $\dot{G}(t)$ can be described as

$$\begin{aligned} \mathcal{A}_1 &= A_1 G(t) + B_1 \dot{G}(t) \\ &\vdots \\ \mathcal{A}_N &= A_N G(t) + B_N \dot{G}(t), \end{aligned} \tag{8}$$

where the variables A_1, B_1, $\cdots$, A_N, B_N, are free parameters which relate the individual time-series to $G(t)$ and $\dot{G}(t)$. We note that each $\mathcal{A}_i$ is a GP by itself, given the property that any derivative operator applied to a GP and the sum of two GPs is also a GP. We can assume that the GP has zero mean because in the GP regression (see Section 2.2) we use the residuals vector to evaluate the likelihood, i.e. we remove the mean function from the observations.

To create the covariance matrix, we need to define how points are correlated between all time-series $\mathcal{A}_i$ and $\mathcal{A}_j$. Following Rajpaul et al. (2015), the covariance between two observations at times t_i and t_j between the time-series $\mathcal{A}_l$ and $\mathcal{A}_m$ is given by

$$\begin{aligned} k^{l,m}(i,j) = A_l A_m \gamma^{G,G}(i,j) + B_l B_m \gamma^{dG,dG}(i,j) \\ + A_l B_m \gamma^{G,dG}(i,j) + A_m B_l \gamma^{dG,G}(i,j), \end{aligned} \tag{9}$$

where $\gamma^{G,G}(i,j)$ denotes the covariance between (non-derivative) observations of G at times t_i and t_j; $\gamma^{G,dG}(i,j)$ refers to the covariance between an observation of G at time t_i and an observation of $\dot{G}$ at time t_j; $\gamma^{dG,G}(i,j)$ refers to the covariance between an observation of $\dot{G}$ at time t_i and an observation of G at time t_j; and $\gamma^{dG,dG}(i,j)$ denotes the covariance between two observations of $\dot{G}$ at times t_i and

t_j. In Appendix A, we show the gamma terms ($\gamma^{G,G}$, $\gamma^{dG,G}$, $\gamma^{G,dG}$, and $\gamma^{dG,dG}$) for the squared exponential, Matérn 5/2, and QP kernels.

The 'big' covariance matrix $\boldsymbol{K}_{\rm big}$ that describes the covariance between all the N time-series is

$$\boldsymbol{K}_{\rm big} = \begin{pmatrix} \boldsymbol{k}^{1,1} & \boldsymbol{k}^{1,2} & \dots & \boldsymbol{k}^{1,N} \\ \boldsymbol{k}^{2,1} & \boldsymbol{k}^{2,2} & \dots & \boldsymbol{k}^{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{k}^{N,1} & \boldsymbol{k}^{N,2} & \dots & \boldsymbol{k}^{N,N} \end{pmatrix} \tag{10}$$

where each $\boldsymbol{k}^{l,m}$ is computed using equation (9) for any covariance function γ. If γ is a valid kernel function, the matrix $\boldsymbol{K}$ is a valid covariance matrix that we can use in GP regression. This matrix also has the property of being symmetric and positive definite therefore we need to compute only the upper triangle part of the matrix, while the lower panel matrices can be computed as $\boldsymbol{k}^{j,i} = (\boldsymbol{k}^{i,j})^{\intercal}$.

At this point we have all the mathematical framework needed to perform multidimensional GP regression for RVs and activity indicators using the GP regression described in Section 2.2. We can create a residual vector $\boldsymbol{r}_i$ for each dimension i, e.g. the residuals corresponding to the RVs can be computed by subtracting Keplerian models, while the residuals of an activity indicator can be computed by subtracting constant offsets. We then create the residual vector $\boldsymbol{r}$ by concatenating the $\boldsymbol{r}_i$ for each dimension. We can create the 'big' covariance matrix $\boldsymbol{K}_{\rm big}$ using equation (10) with any valid kernel (e.g. the ones in Section 2.1). Once we have $\boldsymbol{r}$ and $\boldsymbol{K}_{\rm big}$ we can use the likelihood given by equation (7) to infer the parameters of our multidimensional GP using our preferred numerical method. Section 4 describes how this framework is included inside `pyaneti`.

3.2 Comparison between multidimensional GP and other approaches

A common approach in the literature to describe stellar signals in RV time-series consists of modelling ancillary time-series (these can be light curves or activity indicators) with a GP to infer the hyper-parameters for a given kernel. The hyper-parameters inferred from one or more ancillary time series are then used to inform hyper-parameters priors for the GP modelling of the RV data using the same kernel (see e.g. Haywood et al. 2014; Grunblatt et al. 2015). The process of retrieving kernel hyper-parameters for a GP modelling is known as *training* a GP, and we therefore refer to this approach as the 'training-GP' approach. Training the GP in this way ensures that the stellar activity signal in the RVs is modelled with the same characteristic features – such as periodicity, degree of smoothness, characteristic evolution time-scale – as the ancillary time-series. In this approach, the functions describing the ancillary time-series and the RVs share the similar covariance properties, but they are otherwise independent of each other, and their shapes are entirely unrelated.

A slight variation on the training-GP approach involves modelling the ancillary time-series and the RVs simultaneously, using independent GPs sharing the same covariance function (see e.g. Suárez Mascareño et al. 2020; e.g. Osborn et al. 2021). In this case, both the RVs and the ancillary time-series are used to constrain the GP hyper-parameters, and the functions describing them share exactly the same covariance properties, but they are still independent of each other and have unrelated shapes.

These approaches make weaker assumptions about the relationship between the ancillary time-series and the RVs than those made in the multidimensional model developed by Rajpaul et al. (2015) and implemented in `pyaneti`. They result in more flexible models for the RV time-series, with the associated risk of overfitting (where potential planetary signals can be absorbed or modified by the activity model). On the other hand, our assumptions are based on a fairly simplistic toy model, whose limitations will no doubt become apparent once the framework is applied to a large enough sample of high-precision data-sets. Such failures should however be easy to diagnose, as they would lead to a poor fit to the data. In Section 5.2, we present a comparison of planetary signal recovering between the 'training-GP' method described in this section and the multi-GP approach.

It is also important to note that, in our multidimensional GP model, the functions used to describe the activity signal in the RVs have different covariance properties from those used to model the activity indicators. The fact that the RVs and activity indicators are modelled as different linear combinations of the underlying GP and its time-derivative results in markedly different harmonic complexity, for example (see Section 3.3 for a more detailed discussion). While we are not aware of examples in the literature, the 'training-GP' approach could be generalized to take into account the derivatives of a GP to describe time-series. In Section 5.2.2, we show an example on how to train a GP to use the hyper-parameters to model RVs taking into account the derivative of the chosen kernel. A more complete quantitative comparison between different methods to model stellar signals is given by Ahrer et al. (2021).

3.3 On the usefulness of the GP derivatives

In this section, we describe the importance of taking into account the derivatives of the GP to model RV data when assuming that our GP generates a function that describes the surface covered by active regions on the stellar surface. We base our discussion on the QP kernel, but the conclusions can be extended to other kernels.

Let us suppose we have a 2D GP to describe two time-series, S_1 and S_2, that behave as

$$\begin{aligned} S_1 &= G(t), \\ S_2 &= \dot{G}(t). \end{aligned} \tag{11}$$

Equation (11) was computed from equation (8) with $A_1 = B_2 = 1$, and $A_2 = B_1 = 0$. Fig. 2 shows some samples of S_1 and S_2 time-series that were created using a QP kernel with $P_{\rm GP} = 1$, $\lambda_{\rm e} = 10$, and different values of $\lambda_{\rm p}$.

From the examples in the top panel of Fig. 2 we can see that if the S_1 signal has a high harmonic complexity, then the contemporaneous S_2 would have an apparently higher harmonic complexity. We can see that this behaviour is expected from the derivatives of the QP kernel (see Appendix A). From equations (A13) and (A14), we can see that when $\lambda_{\rm p} \lesssim 1$ there are some 'τ terms' (terms that include the τ parameter) that add extra 'wiggles' to the behaviour of the S_2 curve. This has a direct implication when training GPs with ancillary observations (that may behave as S_1) to model RVs (that may behave as S_2). Setting a prior on $\lambda_{\rm p}$ for the S_2 signal based on our S_1 signal may lead to biased results.

The previous discussion has special implications when training GPs to model RVs using light curves. In general, light curves and RV data are not taken simultaneously, so the active regions on the stellar disc might not be the same between the two data sets (see e.g. Aigrain et al. 2012; Barragán et al. 2021b). But even if they were, the harmonic complexity extracted from a light curve may not be the same as for the activity induced RV signal if modelled only with $G(t)$, i.e. without accounting for the GP derivatives. There are examples of the importance of using the GP derivative when modelling RVs with high harmonic complexity (e.g. Barragán et al. 2019, 2021a) .

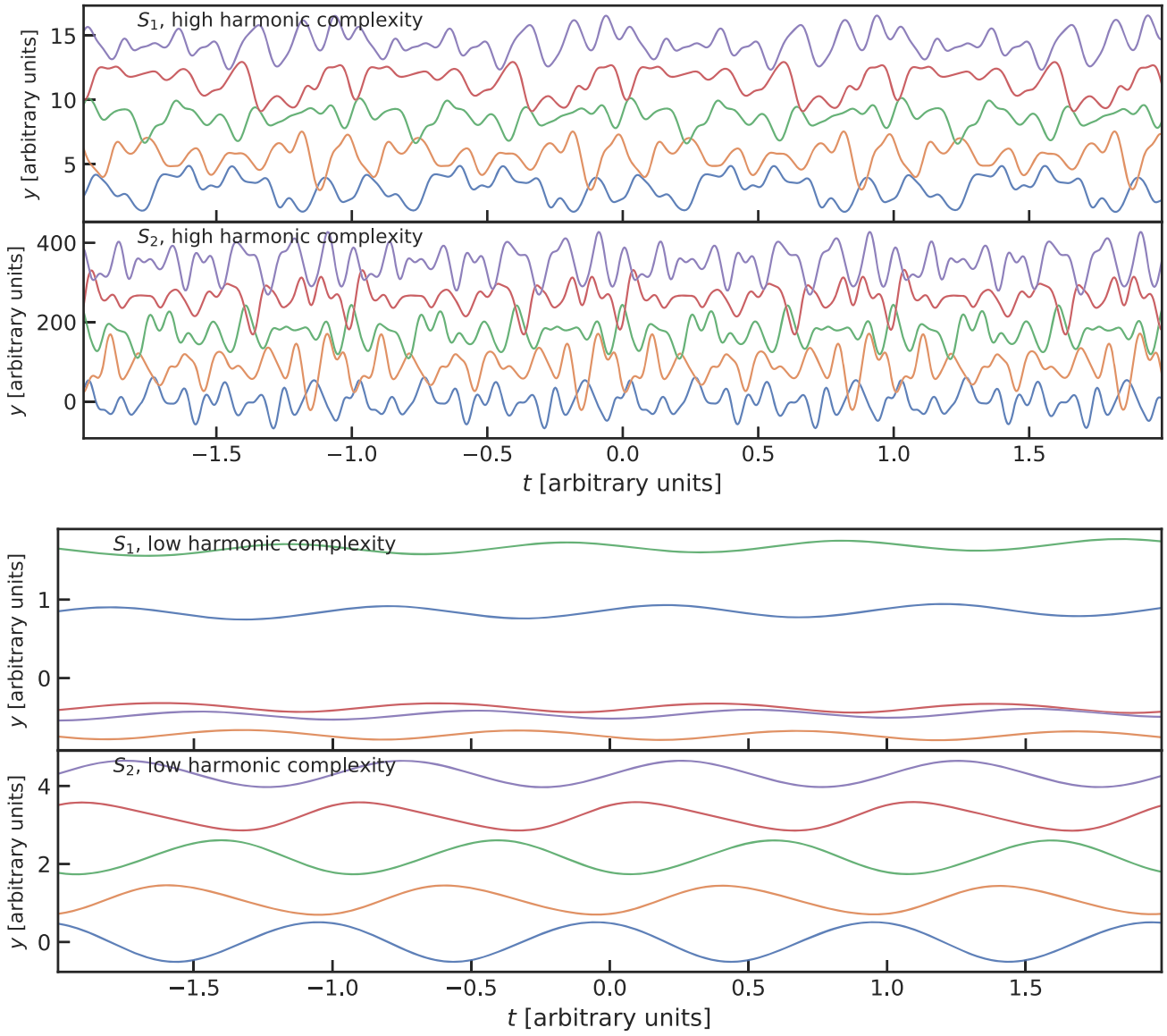

Figure 2. Example of samples from a multidimensional GP (two dimensions in this case) created with a QP kernel with $P_{GP} = 1$, $\lambda_e = 10$ and high (λ_p=0.1, top panel) and low λ_p=10, lower panel) levels of harmonic complexity. Each panel shows two subplots with samples of S_1 and S_2 time-series (See equation 11). Corresponding S_1 and S_2 draws from the same sample are shown with the same colour.

In the other extreme of low harmonic complexity ($\lambda_p \gg 1$) we see that S_1 and S_2 behave as quasi-sinusoidal signals (lower panel of Fig. 2). If we consult equations (A13) and (A14), we can see that in the limit when $\lambda_p \gg 1$, all τ terms are irrelevant and the behaviour of the GP and derivatives is quasi-sinusoidal with a long-term evolution regulated by λ_e. In this case using the derivative of the GP to model the RVs may not be crucial for modelling the activity induced signal. There are some examples in the literature of activity induced RV signals that behave similar to the activity indicators in the low harmonic complexity regime (e.g. Serrano et al., submitted).

We also note that the fact that RVs depend on how active regions evolve in the stellar surface may cause an apparent asynchrony with the activity indicators. Thus phenomenon of RV and activity indicators being out of phase has been reported in the literature (e.g. Collier Cameron et al. 2019). If the stellar signal induced in the RVs depends on a combination of the kind $AG(t) + B\dot{G}$, this can generate curves that may seem similar to the one of the activity indicator, that may vary as $AG(t)$, but with an apparent phase shift. This may be most applicable to the low harmonic complexity case in which signals tend to look more sinusoidal. For example, in Georgieva et al. (2021) the stellar signal in the RV and activity indicators time-series is similar, but the RV curve seems to have a different phase (seemingly ahead) of the contemporary activity indicator time-series. In the multidimensional GP approach the time derivatives of $G(t)$ allow this behaviour to be accounted for.

We have discussed the importance of the GP derivatives based on the QP kernel. However, any other kernel that has strong variations on short time-scales may suffer from the same effect in which the derivatives of the GP are relevant to model RV time-series. We note that this topic needs to be explored further (e.g. Nicholson et al., in preparation), but a detailed description of this issue is beyond the scope of this paper. We mention it, however, to apprise the reader of

the usefulness of including the derivatives of the GP when modelling RV time-series with pyaneti.

3.4 Limitations of the multidimensional GP approach

We warn the reader that the multidimensional GP approach should not be taken as a magic recipe that will always solve our stellar induced RV variation problems.

The first limitation is that this framework assumes a relation between all time-series via a function $G(t)$ and its time derivatives. This may work in many cases, but we warn that this is still only a first-order approach to the problem. Stellar activity signals may be more complicated than the underlying model assumed for this framework. Therefore the multidimensional GP approach may lead to activity signals being fit imperfectly.

Problems may also arise if the data sampling is not optimal. The main idea behind GP regression is to exploit the correlation between our observations. This implies that if observations have time separations larger than the GP time-scales that we want to characterize, then the GP model may be poorly constrained. Therefore, special care has to be taken when using GPs to modelling data with sparse sampling. In Section 4.4.2, we show and describe citlalatonac, a code that simulates spectroscopic-like time-series that can be helpful to plan RV observation campaigns for active stars.

It is also worth mentioning that a downside of GP regression is the computational cost of matrix inversion. This is particularly relevant to the multidimensional GP approach where the dimension of the matrix to invert increases with the number of time-series to model. However, in the case of spectroscopic time-series, there are rarely more than several hundred observations per star. Therefore, the multidimensional GP regression to model RVs and activity indicators remains computationally treatable.

4 THE NEW pyaneti

Barragán et al. (2019) describe the data analysis approach taken by pyaneti. Briefly, pyaneti uses a implementation of a Markov chain Monte Carlo (MCMC) sampler based on EMCEE (Foreman-Mackey et al. 2013) to create marginalized posterior distributions for exoplanet RV and transit parameters. The code uses a built-in Gaussian likelihood together with user-input priors for each sampled parameter. The demanding computational routines are written in FORTRAN and the subroutines are wrapped into PYTHON using the f2py tool included in the numpy (Harris et al. 2020) package.

The new version of pyaneti is an extension of the original package presented in Barragán et al. (2019). The biggest update is the generalization of the Gaussian likelihood that includes correlation using a GPR following equation (7). The code can also perform multiband transit and single transit modelling. The rest of this section describes in more detail the new additions to the code.

4.1 Gaussian processes

Barragán et al. (2019) describe the functions used to model multi-planet signals in both RV and transit data sets. These equations are used as mean functions of the GP to compute the residual vector $\boldsymbol{r}$, allowing pyaneti to perform multiplanet fits with RV and/or transit data together with GP regression using equation (7) and user input priors. The elements of the covariance matrix inside pyaneti are created as

$$K(t_i, t_j) = \gamma\left(t_i, t_j\right) + \left(\sigma_i^2 + \sigma_{\rm jitter}^2\right)\delta_{ij}, \tag{12}$$

where $\gamma(t_i, t_j)$ is any valid kernel function, δ_{ij} is the Kronecker delta, σ_i the white noise associated with the datum i, and $\sigma_{\rm jitter}$ is a jitter term. We note that the same jitter term can be shared by a collection of points with the same underlying systematics, e.g. a jitter term associated with a common instrument. We note that in the case where no correlation is assumed in the data, i.e. $\gamma(t_i, t_j) = 0$, equation (7) reduces to the white noise Gaussian likelihood implemented in the previous version of pyaneti (Barragán et al. 2019). We have implemented in pyaneti all kernels described in Section 2.1, but more can be added easily if needed.

We also incorporated the multidimensional GP approach described in Section 3 into pyaneti within the RV modelling routines. The new version of pyaneti can reproduce the original approach given by Rajpaul et al. (2015), but also allows one to combine arbitrarily many time-series to use multiple activity indicators. The residual vector for the RV data is computed subtracting Keplerian signals, and for the activity indicators with a constant offset. We note that pyaneti can deal with different instrumental offsets for the RV and activity indicators.

To create the 'big' covariance matrix, equation (10), we define the sub-matrices $k_{\rm wn}^{l,m}(i, j)$ to account for white noise as

$$k_{\rm wn}^{l,m}(i, j) = k^{l,m}(i, j) + \left(\sigma_{i,l}^2 + \sigma_{{\rm jitter},l}^2\right)\delta_{ij}\delta_{lm}, \tag{13}$$

where $k^{l,m}(i,j)$ is computed using equation (9) and a valid covariance function, $\sigma_{i,l}$ and $\sigma_{{\rm jitter},l}$ are the white noise and jitter term associated to the dimension l; and δ_{ij} and δ_{lm} are Kronecker deltas. The δ_{lm} between two different dimensions, l and m, ensures that the nominal errors are only added in the diagonal of the "big" matrix. pyaneti creates the $\boldsymbol{K}_{\rm big}$ covariance matrix using the sub-matrices computed with equation (13). The code has built-in routines to perform multi-dimensional GP regression using the squared exponential (equation 3), Matérn 5/2 (equation 4), and QP (equation 6) Kernels. Appendix A show the derivatives included into pyaneti to compute equation (9) for these covariance functions.

4.1.1 *Dimensionality problem*

We note that pyaneti suffers from the high computational cost of GP regression, which in general entails matrix inversion. The computational time of a matrix inversion scales as $\mathcal{O}(N^3)$, where N is the number of data points. This is generally not a problem for RV data sets, which include relative few observations (usually less than 1000). However, it becomes a problem when modelling light curve time-series with thousands of observations.

Fortunately, in the case of light curves, the time-scales of the planetary transits are small compared with those associated with stellar variability. This makes it relatively easy to remove such trends from light curves. This *detrending* is a common approach in the literature when one is only interested on modelling transit signals in flattened light curves (see e.g. Hippke et al. 2019). A reason for fitting GPs to full light curves with transits would be to argue that the light curve might provide information on GP hyper-parameters that can also be used simultaneously with RV data. Yet this may not be optimal because usually light curves and RVs are not observed simultaneously, and there is evidence that light curves and RVs may not be constrained by the same time-scales, as discussed in Section 3.3. Since the GP implementation of pyaneti is mainly focused on the RV analysis, we did not explore GP computation acceleration for light-curve analyses. We note that pyaneti could still be used to model binned light curves in order to try to estimate

hyper-parameters of the light curve, but a GP modelling of the light curve including transits is not advisable with pyaneti.

We note that progress has been made in fast matrix inversion for GP regression. We considered implementing the GP regression operations included in pyaneti using the george package (Ambikasaran et al. 2015); this would require explicitly coding the derivative kernels we use to model each time-series in george, a feasible but non-trivial endeavour. It has not been necessary to do it so far, given the typical size of the data sets we are modelling, but it is something to be considered for a future implementation of pyaneti. Another, even faster option for GP regression on large data sets is the celerite package (Foreman-Mackey 2018), but that is not suitable for this work as it is restricted to 1D data sets.

4.2 Multiband fit

Multiband photometric follow-up of transiting planets has become common. This is because the number of ground and space-based instruments has increased, and more transiting planets are found around relatively bright stars. For this reason, we have added multiband transit modelling into pyaneti. The code solves for the same orbital parameters for all bands, but independently samples the wavelength-dependent parameters, i.e. limb darkening coefficients, cadence and integration time, and scaled planet radius.

As in the previous version, the code uses the Mandel & Agol (2002) equations to model transits by assuming the star limb darkening can be modelled as a quadratic law. The code samples for two limb darkening coefficients for each band following the q_1 and q_2 parametrization described in Kipping (2013). The code also allows for a different cadence and integration time for each band (see Kipping 2010). Finally, the code also allows modelling for the same scaled planet radius $R_p/R_\star$ for all bands, as well as an independent $R_{p,i}/R_\star$ for each band i. The latter can be useful to test false-positive scenarios (e.g. Parviainen et al. 2019), or to fit transit depths at different wavelengths as used in transmission spectroscopy (e.g. Charbonneau et al. 2002).

4.3 Single transit fit

Single transit events can be caused by transiting planets with periods longer than the observational window. Fortunately, they can be detected by methods that do not rely on periodicity of transit-like events (see e.g. Osborn et al. 2016; Eisner et al. 2021).

The problem when dealing with mono-transits is that the period and the semimajor axis cannot be determined. These parameters are important because they determine the velocity at which the planet moves during the transit. In order to solve this, in the new version of pyaneti we fix a period to a dummy value larger than our observing window, and we sample for a dummy scaled semimajor axis $a_{\rm dummy}$. While $a_{\rm dummy}$ does not have a physical sense, but it ensures that the transit shape is sampled.

Therefore, in order to model a single transit, pyaneti samples the time of mid-transit T_0, impact parameter b, scaled planet radius R_p, $a_{\rm dummy}$, and limb darkening coefficients q_1 and q_2 (Kipping 2013). With these parameters we can estimate the transit duration, and if we assume that the planetary orbit is circular, we can estimate the orbital period albeit with relatively large uncertainty (see Osborn et al. 2016).

4.4 citlalicue and citlalatonac

We have created codes to create synthetic stellar photometric and spectroscopic time-series, called citlalicue and citlalatonac, respectively.[3]

4.4.1 citlalicue: the light-curves creator

citlalicue is a Python module that allows one to create synthetic light curves. It is totally independent of pyaneti. It can be easily installed using pip install citlalicue. The module has a class called citlali that contains all the attributes and methods needed to create a synthetic stellar light curves with transits, periodic modulation, and white noise. The current version of the code uses a QP kernel (equation 6) to simulate stellar variability, and it allows one to create transits for any number of planets using pytransit (Parviainen 2015). An example of how to create a light curve using citlalicue is given here https://github.com/oscaribv/citlalicue/blob/master/example_light_curves.ipynb.

citlalicue also has a class called detrend that allows one to detrend light curves using GPs. detrend takes a plain-text file with light-curve data containing time and flux (and errors as input if available). The code allows one to mask out the transits from the data as well as to fit simultaneously for the transits and GP. The former is recommended given that it is faster. The code uses george (Ambikasaran et al. 2015) to perform a fast GP regression that enables the modelling of the variability in the light curve. The code allows for an iterative optimization with a sigma clipping algorithm, where the threshold can be tuned by the user (the default is 5). Once the optimal model is found by the code, the inferred trend is removed from the light curve, creating a flattened signal with transits. An example of how to detrend a light curve using citlalicue is available here https://github.com/oscaribv/citlalicue/blob/master/example_detrending.ipynb. Examples of the detrending capability of citlalicue can be found in e.g. Barragán et al. (2021b) and Georgieva et al. (2021).

4.4.2 citlalatonac: the spectroscopic time-series creator

The Python package citlalatonac uses pyaneti in order to create synthetic spectroscopic (RVs and activity indicators-like) time-series. This package comes together with pyaneti when the latter is cloned directly from its GitHub repository. The code creates samples of a multidimensional GP following equation (8). This simulates spectroscopic-like signals (RVs and activity indicators) of an active star, assuming they all are generated by the same underlying GP.

The main class of the package is named citlali. When citlali is called in Python, the user needs to specify the time range in which the synthetic data will be created (e.g. this range can be an observing season), the number of time-series to create, the amplitudes of the signals, following equation (8), the kernel to use, and the kernel parameters. By default the first time-series is called rv and it is always treated as RV-like, i.e. the planet-induced signals are added to this time-series only. The class includes methods that allow one to add as many planets as needed, as well as white and red noise.

[3]In Aztec mythology, Citlalicue (goddess) and Citlalatonac (god) are the creators of the stars. The words root, *Citlali*, is the Nahuatl word for star.

The package also includes the `create_real_times` function that allows one to create realistic sampling of targets at a given observatory using `astropy` (Astropy Collaboration 2013, 2018). This utility can be useful for estimating the number of data points needed in order to measure the Doppler semi-amplitude of a given target, even if the star is active. Such numbers can be valuable while writing a telescope proposal. A practical example of how to use `citlalatonac` to create synthetic time-series of a target observed at a given observatory can be found here https://github.com/oscaribv/pyaneti/blob/master/pyaneti_extras/synthetic_k2100.ipynb.

5 TESTS

5.1 Recovering multi-GP hyper-parameters

We created a set of synthetic spectroscopic-like time-series using `citlalatonac` (see Appendix 4.4) in order to test the ability of `pyaneti` to recover parameters using a multidimensional GP. We assume that we have three time-series that are described by the same underlying function, $G(t)$, as

$$\begin{aligned} S_1 &= A_1 G(t) + B_1 \dot{G}(t), \\ S_2 &= A_2 G(t), \\ S_3 &= A_3 G(t) + B_3 \dot{G}(t). \end{aligned} \tag{14}$$

We compute equation (14) using equation (8) with 3 time-series. We set the values for the amplitudes $A_1 = 0.005\,\mathrm{km\,s^{-1}}$, $B_1 = 0.05\,\mathrm{km\,s^{-1}\,d}$, $A_2 = 0.02\,\mathrm{km\,s^{-1}}$, $B_2 = 0\,\mathrm{km\,s^{-1}\,d}$, $A_3 = 0.02\,\mathrm{km\,s^{-1}}$, $B_3 = -0.05\,\mathrm{km\,s^{-1}\,d}$. We assume a mean function of zero for all three time series, and we use a QP covariance function with hyper-parameters λ_e=30 d, λ_p=0.3, and $P_{GP} = 5$ d. We created 50 simultaneous observations taken randomly in a window of 50 d. We added white noise with standard deviation of $0.001\,\mathrm{km\,s^{-1}}$ for S_1, $0.005\,\mathrm{km\,s^{-1}}$ for S_2, and $0.010\,\mathrm{km\,s^{-1}}$ for S_3. The synthetic time-series data are shown in Fig. 3. The `Jupyter` notebook used to create the synthetic data is available here https://github.com/oscaribv/pyaneti/blob/master/inpy/example_toyp1/toy_model1.ipynb

We performed multidimensional GP modelling of the data set using `pyaneti`. Priors and parameters used are defined in Table 1. We perform an MCMC analysis with 100 Markov chains. We use the last 5000 iterations of converged chains, with a thin factor of 10, to create the posterior distributions. We assume chains have converged when their Gelman et al. (2004) criterion $\boldsymbol{R}$ is smaller than 1.02 for all the sampled parameters (for more details see Gelman et al. 2004; Barragán et al. 2019). The inferred parameters are shown in Table 1, and the inferred models, together with the data, are shown in Fig. 3. We have made available the data and input file in `pyaneti` for this example; it can be run as `./pyaneti.py example_toyp1` from the main `pyaneti` directory.

From Table 1 we can see that the code is able to recover the injected amplitudes and kernel parameters within the error bars. Something to note is that the code is able to recover the value of $B_2 = 0\,\mathrm{km\,s^{-1}\,d}$. This is important because it means that the analysis is able to differentiate between the pure $G(t)$ curves from those that depends on $\dot{G}(t)$. This has a practical application to understand the behaviour of the activity indicators that we use in our modelling.

5.2 Recovering Keplerian signals with multi-instrument data

We perform another test similar to the one described in Section 5.1, but this time we added some more challenges to the test. In this case, we assume that we have contemporaneous observations of two time-series that behave as

$$\begin{aligned} S_1 &= A_1 G(t) + B_1 \dot{G}(t), \\ S_2 &= A_2 G(t), \end{aligned} \tag{15}$$

with $A_1 = 0.005\,\mathrm{km\,s^{-1}}$, $B_1 = 0.05\,\mathrm{km\,s^{-1}\,d}$, $A_2 = 0.02\,\mathrm{km\,s^{-1}}$, and $B_2 = 0\,\mathrm{km\,s^{-1}\,d}$. We use a QP covariance function with hyper-parameters $\lambda_e = 20$ d, $\lambda_p = 0.5$, and $P_{GP} = 5$ d. We assume that the spectroscopic data comes from two different instruments, I_1 and I_2, with an offset of zero for each time-series. For instrument I_1 we created 20 random observations between 0 and 60 d, each datum with an error bar of $0.003\,\mathrm{km\,s^{-1}}$. For instrument I_2 we created 30 random observations in the same range, each one with an error bar of $0.005\,\mathrm{km\,s^{-1}}$. We included two Keplerian signals in the RV-like time-series (S_1). One signal is associated with a circular orbit ($\sqrt{e}\sin\omega = 0$, and $\sqrt{e}\cos\omega = 0$, following Anderson et al. 2011, parametrization) and the other one with an eccentricity of 0.3 and angle of periastron of $\pi/3$ ($\sqrt{e}\sin\omega = 0.47$, and $\sqrt{e}\cos\omega = 0.27$). The amplitude of the Keplerian signals is significantly smaller than the amplitudes of the activity-like signal. The parameters used to create both signals are listed in Table 2. The `Jupyter` notebook used to create the synthetic data is available here https://github.com/oscaribv/pyaneti/blob/master/inpy/example_toyp2/toy_model2.ipynb. We show the synthetic time-series in Fig. 4. We perform four different analyses to the data using different techniques. For all the cases we assume that we know the ephemeris of the Keplerian signals and set Gaussian priors on the time of minimum conjunction, T_0, and period, P, for the two signals.

5.2.1 *Run 1*

We first perform a 1D GP modelling with `pyaneti` modelling only the RVs with a QP kernel. We assume that the only information that we have of the GP hyper-parameters are the ranges where the true values lie. Table 2 shows the sampled parameters and priors we use for this run that we name as *Run 1*. We perform an MCMC sampling with 100 Markov chains. We create the posterior distributions with the last 5000 iterations of converged chains with a thin factor of 10. This generates distributions with 50 000 independent points per each sampled parameter.

The inferred parameters are shown in Table 2. The first thing we note is that for this case the recovered λ_e and P_{GP} are consistent with the true values, but the value of λ_p is smaller than the true vale used to create the time-series. This is expected given that we are modelling the data without taking into account the GP derivative (see discussion in Section 3.3). However, the parameter true values used to create the Keplerian signals are recovered within the confidence interval.

5.2.2 *Run 2*

We then perform a *Run 2* in which we train our GP based on our activity-indicator-like signal (S_2). As we mentioned in Section 3.2, this is a common approach in the literature. We first do a 1D GP modelling of the S_2 signal using a QP kernel in order to obtain posterior distributions for the hyper-parameters λ_e, λ_p, and P_{GP}. We then model the RV data following the normal approach in the literature (e.g. Grunblatt et al. 2015), i.e. we use a QP kernel with Gaussian priors on λ_e, λ_p, and P_{GP} based on our S_2 analysis. The MCMC details are identical to the ones described in Section 5.2.1. Table 2 shows priors and inferred values for all the sampled parameters.

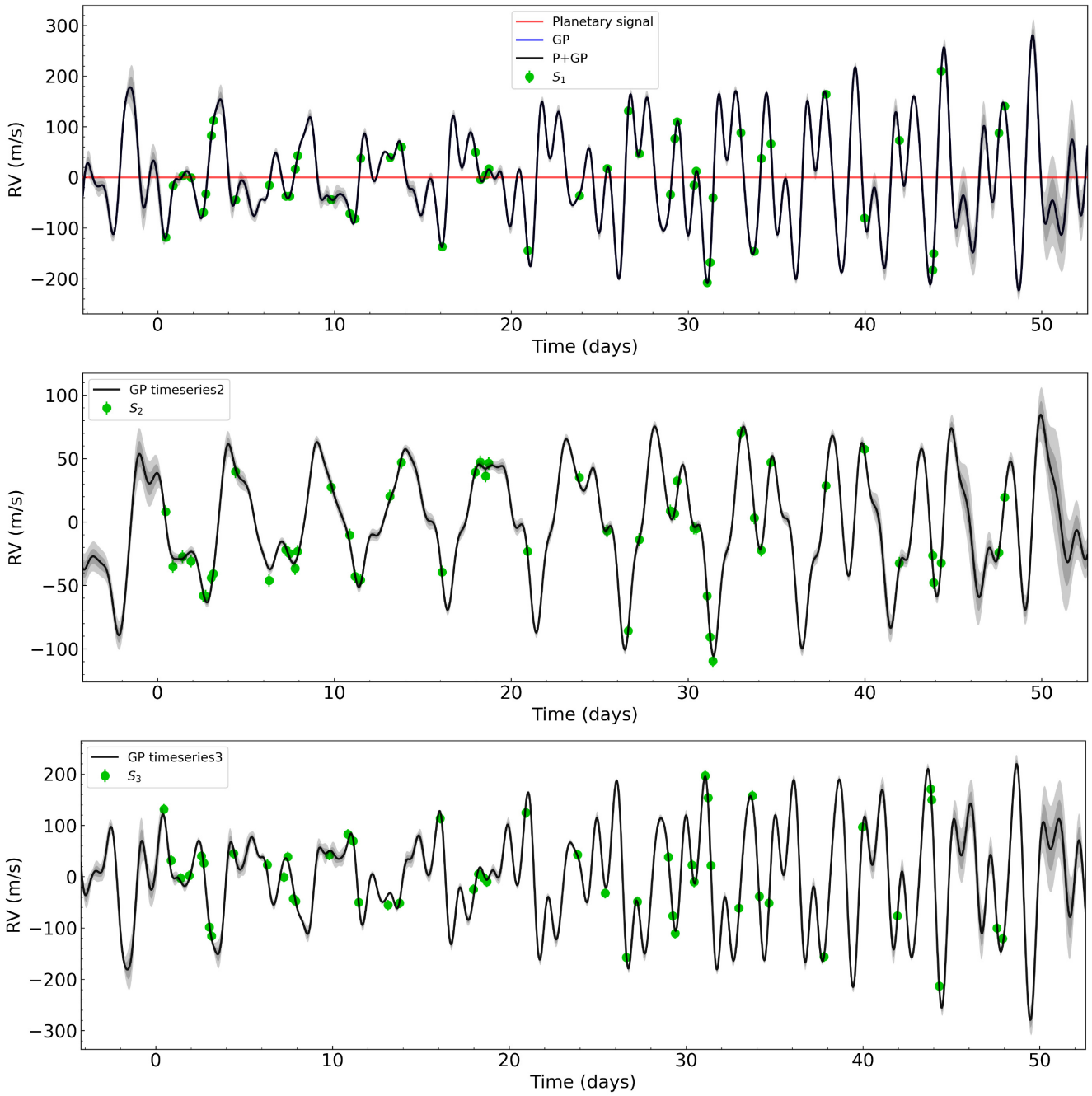

Figure 3. S_1, S_2, and S_3 time-series described in Section 5.1. The green markers in each panel represent the synthetic observations with inferred offsets extracted. The solid dark lines shows the inferred mean of the predictive distribution of our multidimensional GP, with dark and light shaded areas showing the 1σ and 2σ credible intervals of the corresponding GP model, respectively. These plots are shown as they are provided by `pyaneti`. The code assumes that the first time-series is RV-like, and it may contain Keplerian signals that are represented with a red curve, in this case no Keplerian signals are included and the curve appears as a horizontal red line at 0.

From Table 2 we can see that the results for this run are similar to the ones from the *Run 1*. We note that despite the Gaussian prior that we set on λ_p the recovered value for this parameter is significantly smaller than the Gaussian prior mean. This is again expected given that we are modelling the RV data only with a QP kernel, without accounting for the time derivative (See Section 3.3). None the less, the recovered values of the Doppler semi-amplitudes of the of the coherent signals are recovered with a significance similar to the ones in *Run 1*.

5.2.3 *Run 3*

We explored another possibility on the modelling of the RV time-series training the GP. But this time for the RV modelling we are including the first time derivative of the GP. The GP training comes from the same analysis of the S_2 time-series described in Section 5.2.2. But, for the RV 1D GP regression we construct our covariance matrix with the kernel

$$\gamma(t_i, t_j) = A_1^2 \gamma_{\rm QP}^{G,G} + B_1^2 \gamma_{\rm QP}^{dG,dG}, \tag{16}$$

Table 1. System parameters for toy model.

Parameter	Real value	Prior[a]	Inferred value
A_1 (km s^{-1})	0.005	$\mathcal{U}[0, 0.5]$	$0.0103^{+0.0053}_{-0.0042}$
B_1 (km s^{-1} d)	0.05	$\mathcal{U}[0, 0.5]$	$0.063^{+0.014}_{-0.011}$
A_2 (km s^{-1})	0.05	$\mathcal{U}[-0.5, 0.5]$	$0.064^{+0.014}_{-0.011}$
B_2 (km s^{-1} d)	0	$\mathcal{U}[-0.5, 0.5]$	$-0.00041^{+0.00055}_{-0.00062}$
A_3 (d)	0.005	$\mathcal{U}[-0.5, 0.5]$	$0.0031^{+0.0043}_{-0.0047}$
B_3 (km s^{-1} d)	−0.05	$\mathcal{U}[-0.5, 0.5]$	$-0.063^{+0.011}_{-0.014}$
λ_e (d)	20	$\mathcal{U}[1, 80]$	$20.01^{+1.34}_{-1.34}$
λ_p	0.3	$\mathcal{U}[0.01, 5]$	$0.331^{+0.02}_{-0.021}$
P_{GP} (d)	5	$\mathcal{U}[4, 6]$	$4.9949^{+0.0059}_{-0.0054}$
Offset S_1	0	$\mathcal{U}[-0.7, 0.7]$	$0.0007^{+0.0047}_{-0.0043}$
Offset S_2	0	$\mathcal{U}[-0.7, 0.7]$	$-0.007^{+0.027}_{-0.022}$
Offset S_3	0	$\mathcal{U}[-0.7, 0.7]$	$-0.0016^{+0.0029}_{-0.003}$

[a]$\mathcal{U}[a, b]$ refers to uniform priors between a and b, $\mathcal{N}[a, b]$ to Gaussian priors with mean a and standard deviation b.
[b]Inferred parameters and errors are defined as the median and 68.3 per cent credible interval of the posterior distribution.

that includes the derivative term of the QP kernel. Appendix 3.3 shows the full form of the $\gamma_{QP}^{G,G}$ and $\gamma_{QP}^{dG,dG}$ terms. We perform a 1D GP modelling of the RV data following the priors described in Table 2 as Run 3. The MCMC configuration follows the same parameters as in Section 5.2.1.

Table 2 show the recovered parameters for this run. We note that for this case, the recovered value for λ_p is consistent with the true value, as expected now that we are including the derivative of the GP. However, we note that even if we include the derivative, the recovered values of the Doppler semi-amplitudes are not more precise that the values recovered in Runs 1 and 2. This can be explained because even if we are using a better model (that we know agrees with the model used to create the synthetic data), this approach still does not constrain better the shape of the underlying function describing the stellar signal in the RV data.

5.2.4 Run 4

We then perform a 2D GP modelling with `pyaneti` following the approach described in Section 3.1. Table 2 shows the sampled parameters, priors we use, and derived parameters. Again, the MCMC configuration is the same as the one described in Section 5.2.1. This example can be reproduced by running `./pyaneti.py example_toyp2` in the main `pyaneti` directory.

Fig. 4 shows the derived time-series and phase-folded models for this case. We can see in Table 2 that all derived parameters agree with the true values within the error bars. Specially, the value of λ_p agrees with the true value as in the *Run 3*, as expected given that we are using the GP derivative in this case. We note that in this case, the derived semi-amplitudes for both Keplerian signals are recovered with a relative higher precision than in the previous runs.

5.2.5 Comparison between the different Runs

Table 2 shows the parameter value for all the four Runs described in this Section, and Fig. 5 shows the recovered posterior distribution for the Keplerian signals b and c for each case. From Table 2 we can see that all Runs are able to recover the planetary-like signals within the error bars. This is relevant since we created this data set with activity-like amplitudes significantly larger than the Keplerian ones, with the intention to show that the code is able to recover coherent signals in the RV-like time-series. This provides some tentative evidence that if RV observations are planned with suitable observing campaigns and with the right instruments, we may be able to find planetary signals even in cases with extreme activity.

From Table 2 and Fig. 5, we can also see that run that provides better precision on the detected Doppler semi-amplitudes is *Run 4*. We can argue that *Runs* 1 and 2 are not optimal to analyse this problem, because of the way we create the synthetic RV time-series. But we may think that Runs 3 and 4 are equivalent: the model of the S_2 data is created with a function draw using a QP kernel, while the S_1 time-series with a function created with a QP kernel and its first time derivative. And from Table 2, we see that the QP kernel hyper-parameters are fully consistent within the error bars for both runs. However, we obtain better detection of the Keplerian signals on *Run 4* where we use the multidimensional GP approach. This is explained by the discussion in Section 3.2, where we mentioned that the multidimensional GP approach ensures that the underlying function $G(t)$ is the same for all time-series. This result shows the advantage of using the activity indicators to model stellar signals within a multidimensional GP framework.

These tests also illustrate `pyaneti`'s ability to handle different instruments. However, we caution that this capability should be used with care. In particular, multiple instruments do not only have different offsets between them, but they may also observe in different wavelength ranges. This means that the stellar signal that each instrument observes may be different. Therefore, they cannot necessarily be treated as the same underlying signal that the multi-GP approach described in this manuscript assumes.

5.3 Multiband transit modelling

We create a synthetic light curve using `citlalicue` (see Section 4.4) in order to test the ability of `pyaneti` to model multiband transit modelling. We created data assuming we have a flattened light curve of a system with two transiting planets observed with two different instruments. For the first instrument, named 'B1', the data ranges from 0 to 15 d, with one data point every 5 min, with a precision of 500 ppm per datum. The limb darkening coefficients for this fictitious observed star with instrument B1 are $u_1 = 0.25$ and $u_2 = 0$ ($q1 = 0.06$ and $q2 = 0.50$, following Kipping 2013), following the quadratic law of Mandel & Agol (2002). For the second instrument, named 'B2', data go from 20 to 30 d, with one observations every 5 min with a precision of 100 ppm. The limb darkening coefficients for this fictitious star with instrument B2 are $u_1 = 0.50$ and $u_2 = 0.25$ ($q1 = 0.56$ and $q2 = 0.33$, following Kipping 2013). We injected two transiting planets with circular orbits into the light curves. We also assume that both planets are transiting a star with a density of 1.4 g cm^{-3}. The time of transit T_0, orbital period P, impact parameter b, and scaled planet radius r_p for each planet are given in Table 3. The `Jupyter` notebook used to create this synthetic data set is provided here https://github.com/oscaribv/pyaneti/blob/master/inpy/example_multiband/multiband_transits.ipynb, and Fig. 6 shows the synthetic light curves for both bands.

We perform a multiband and multiplanet modelling with `pyaneti`. Table 3 shows the sampled parameters and priors we use. We note that we sample for a different limb darkening coefficients for each band. We also sample for the stellar density, and recover the scaled semimajor axis for each planet using Kepler's third

Table 2. System parameters for toy model in Section 5.2.

	Real	Run 1		Run 2		Run 3		Run 4	
Parameter	Value	Prior[a]	Value[b]	Prior[a]	Value[b]	Prior[a]	Value[b]	Prior[a]	Value[b]
$T_{0,b}$ (d)	1	$\mathcal{N}[1, 10^{-3}]$	$1. \pm 0.001$	$\mathcal{N}[1, 10^{-3}]$	$1. \pm 0.001$	$\mathcal{N}[1, 10^{-3}]$	$1. \pm 0.001$	$\mathcal{N}[1, 10^{-3}]$	$1. \pm 0.001$
P_b (d)	3	$\mathcal{N}[3, 10^{-3}]$	$3. \pm 0.001$	$\mathcal{N}[3, 10^{-3}]$	$3. \pm 0.001$	$\mathcal{N}[3, 10^{-3}]$	$3. \pm 0.001$	$\mathcal{N}[3, 10^{-3}]$	$3. \pm 0.001$
$\sqrt{e}_b \sin\omega_b$	0	$\mathcal{U}[-1, 1]$	$0.19^{+0.39}_{-0.48}$	$\mathcal{U}[-1, 1]$	$0.16^{+0.4}_{-0.48}$	$\mathcal{U}[-1, 1]$	$0.05^{+0.43}_{-0.46}$	$\mathcal{U}[-1, 1]$	$0.08^{+0.27}_{-0.31}$
$\sqrt{e}_b \cos\omega_b$	0	$\mathcal{U}[-1, 1]$	$-0.1^{+0.36}_{-0.3}$	$\mathcal{U}[-1, 1]$	$-0.14^{+0.35}_{-0.26}$	$\mathcal{U}[-1, 1]$	$-0.15^{+0.38}_{-0.25}$	$\mathcal{U}[-1, 1]$	$0.15^{+0.17}_{-0.23}$
K_b (m s^{-1})	5	$\mathcal{U}[0, 500]$	$7.41^{+3.0}_{-2.6}$	$\mathcal{U}[0, 500]$	$7.27^{+3.02}_{-2.66}$	$\mathcal{U}[0, 500]$	$6.82^{+2.87}_{-2.7}$	$\mathcal{U}[0, 500]$	$4.89^{+1.01}_{-0.88}$
$T_{0,c}$ (d)	2	$\mathcal{N}[2, 10^{-3}]$	$2. \pm 0.001$	$\mathcal{N}[2, 10^{-3}]$	$2. \pm 0.001$	$\mathcal{N}[2, 10^{-3}]$	$2. \pm 0.001$	$\mathcal{N}[2, 10^{-3}]$	$2. \pm 0.001$
P_c (d)	10	$\mathcal{N}[10, 10^{-3}]$	$10. \pm 0.001$	$\mathcal{N}[10, 10^{-3}]$	$10. \pm 0.001$	$\mathcal{N}[10, 10^{-3}]$	$10. \pm 0.001$	$\mathcal{N}[10, 10^{-3}]$	$10. \pm 0.001$
$\sqrt{e}_c \sin\omega_c$	0.47	$\mathcal{U}[-1, 1]$	$0.19^{+0.39}_{-0.48}$	$\mathcal{U}[-1, 1]$	$0.16^{+0.4}_{-0.48}$	$\mathcal{U}[-1, 1]$	$0.05^{+0.43}_{-0.46}$	$\mathcal{U}[-1, 1]$	$0.08^{+0.27}_{-0.31}$
$\sqrt{e}_c \cos\omega_c$	0.27	$\mathcal{U}[-1, 1]$	$-0.1^{+0.36}_{-0.3}$	$\mathcal{U}[-1, 1]$	$-0.14^{+0.35}_{-0.26}$	$\mathcal{U}[-1, 1]$	$-0.15^{+0.38}_{-0.25}$	$\mathcal{U}[-1, 1]$	$0.15^{+0.17}_{-0.23}$
K_c (m s^{-1})	10	$\mathcal{U}[0, 500]$	$11.49^{+3.67}_{-2.87}$	$\mathcal{U}[0, 500]$	$11.47^{+3.26}_{-2.77}$	$\mathcal{U}[0, 500]$	$11.61^{+3.06}_{-2.59}$	$\mathcal{U}[0, 500]$	$9.97^{+1.0}_{-0.89}$
A_1 (km s^{-1})	0.005	$\mathcal{U}[0, 0.5]$	$0.075^{+0.026}_{-0.015}$	$\mathcal{U}[0, 0.5]$	$0.081^{+0.022}_{-0.016}$	$\mathcal{U}[0, 0.5]$	$0.0109^{+0.0209}_{-0.0083}$	$\mathcal{U}[0, 0.5]$	$0.006^{+0.0024}_{-0.0018}$
B_1 (km s^{-1}d)	0.05	...	...	...	...	$\mathcal{U}[0, 0.5]$	$0.051^{+0.018}_{-0.013}$	$\mathcal{U}[0, 0.5]$	$0.056^{+0.015}_{-0.011}$
A_2 (km s^{-1})	0.05	...	...	...	...	...	...	$\mathcal{U}[-0.5, 0.5]$	$0.054^{+0.015}_{-0.011}$
B_2 (km s^{-1}d)	0	...	...	...	...	...	...	$\mathcal{U}[-0.5, 0.5]$	$0.15^{+0.49}_{-0.47}10^{-3}$
λ_e(d)	20	$\mathcal{U}[1, 80]$	$26.59^{+5.94}_{-4.84}$	$\mathcal{N}[21, 5]$	$24.16^{+3.62}_{-3.4}$	$\mathcal{N}[21, 5]$	$22.64^{+3.8}_{-3.46}$	$\mathcal{U}[1, 80]$	$20.56^{+1.6}_{-1.53}$
λ_p	0.5	$\mathcal{U}[0.01, 5]$	$0.273^{+0.037}_{-0.035}$	$\mathcal{N}[0.59, 0.10]$	$0.308^{+0.045}_{-0.037}$	$\mathcal{N}[0.59, 0.10]$	$0.48^{+0.06}_{-0.055}$	$\mathcal{U}[0.01, 5]$	$0.487^{+0.041}_{-0.033}$
P_{GP} (d)	5	$\mathcal{U}[4, 6]$	$5.032^{+0.012}_{-0.014}$	$\mathcal{N}[5.03, 0.03]$	$5.031^{+0.012}_{-0.013}$	$\mathcal{N}[5.03, 0.03]$	$5.03^{+0.014}_{-0.015}$	$\mathcal{U}[4, 6]$	$5.012^{+0.01}_{-0.011}$
I_1 offset S_1	0	$\mathcal{U}[-0.5, 0.5]$	$0.005^{+0.031}_{-0.029}$	$\mathcal{U}[-0.5, 0.5]$	$0.003^{+0.033}_{-0.03}$	$\mathcal{U}[-0.5, 0.5]$	$-0.0019^{+0.0088}_{-0.0063}$	$\mathcal{U}[-0.5, 0.5]$	$-0.0006^{+0.0027}_{-0.0028}$
I_2 offset S_1	0	$\mathcal{U}[-0.5, 0.5]$	$0.001^{+0.031}_{-0.029}$	$\mathcal{U}[-0.5, 0.5]$	$-0.001^{+0.033}_{-0.03}$	$\mathcal{U}[-0.5, 0.5]$	$-0.0052^{+0.009}_{-0.0066}$	$\mathcal{U}[-0.5, 0.5]$	$-0.0013^{+0.0028}_{-0.0027}$
I_1 offset S_2	0	...	...	...	...	...	...	$\mathcal{U}[-0.5, 0.5]$	$-0.008^{+0.025}_{-0.024}$
I_2 offset S_2	0	...	...	...	...	...	...	$\mathcal{U}[-0.5, 0.5]$	$-0.008^{+0.025}_{-0.024}$

[a] and [b] defined as in Table 1.

law. We sample the parameter space with 100 independent chains. We create the posterior distributions for each sampled parameter with the last 5000 iterations of converged steps using a thin factor of 10. This example can be reproduced in `pyaneti` by running `./pyaneti.py example_multiband` in the main `pyaneti` directory.

Fig. 6 shows the phase-folded light curves for each transiting planet. We show the inferred parameters in Table 3. We can see that the code is able to recover the orbital and planet parameters for both transiting signals. We also note that `pyaneti` can recover the band-dependent parameters for this example (limb darkening coefficients). In Section 5.5, we describe some real stellar systems where the multiband capabilities of the code have been used.

5.4 Single transit event

We create a single transit event light curves using `citlalicue` (see Appendix 4.4) in order to test the ability of `pyaneti` to model single transits. We create the data assuming we have a flattened light curve of a system with one planetary transit. The data range from 9 to 11 d, with one data point every 5 min, with a precision of 100 ppm per datum. The limb darkening coefficients for this fictitious star are $u_1 = 0.25$ and $u_2 = 0$ ($q1 = 0.06$ and $q2 = 0.50$, following Kipping 2013), following the quadratic law of Mandel & Agol (2002). We injected a transiting planet assuming a circular orbit around a star with a Sun-like density. The time of transit is $T_0 = 10$ d, orbital period $P = 30$ d, scaled semimajor axis $a/R_\star = 40.6$, impact parameter $b = 0.5$ and scaled planet radius $r_p = 0.25$. The code needed to create this synthetic data set is provided in this link https://github.com/oscaribv/pyaneti/blob/master/inpy/example_single/example_single.ipynb.

We perform a single transit modelling with `pyaneti` by indicating `is_single_transit = True` in the input file for this system. Table 4 shows the sampled parameters and priors we use. We sample the parameter space with 100 independent chains and create the posterior distributions with the last 5000 iterations of converged chains with a thin factor of 10. This example can be reproduced in `pyaneti` by running `./pyaneti.py example_single` within the main `pyaneti` directory.

Fig. 7 shows the inferred model for the single transit and Table 4 shows the inferred parameters. `pyaneti` is able to recover the injected values of T_0, b, r_p, q_1, and q_2 within 1σ error bars. We note that for single transit fits, the scaled semimajor axis and periods given by the code do not have the same meaning as for normal runs fitting multiple transits. The scaled semimajor axis is a dummy value sampled to take into account the transit shape, while the orbital period is a derived parameter, i.e. we do not sample for it directly, we compute it with the other sampled parameters assuming the orbit is circular. This capability of the code has been used before to estimate periods for transit signals detected by the *TESS* mission (e.g. Eisner et al. 2021).

5.5 Real planetary systems

We have shown how `pyaneti` is able to recover the injected planetary and orbital parameters in specific examples of synthetic spectroscopic-like and photometry-like time-series. This demonstrates that if we believe that our RV and transit data behave as the models described in this paper, `pyaneti` will be able to provide reliable parameter estimates. Fortunately, the new implementations of the code have already been applied to real data in peer-reviewed literature. In this section, we describe how `pyaneti` has been used in these analyses. We describe scenarios in which all the new additions of the code have been used combined. It is not our intention to reproduce the analyses or plots published in the aforesaid manuscripts. We only describe how `pyaneti` was used

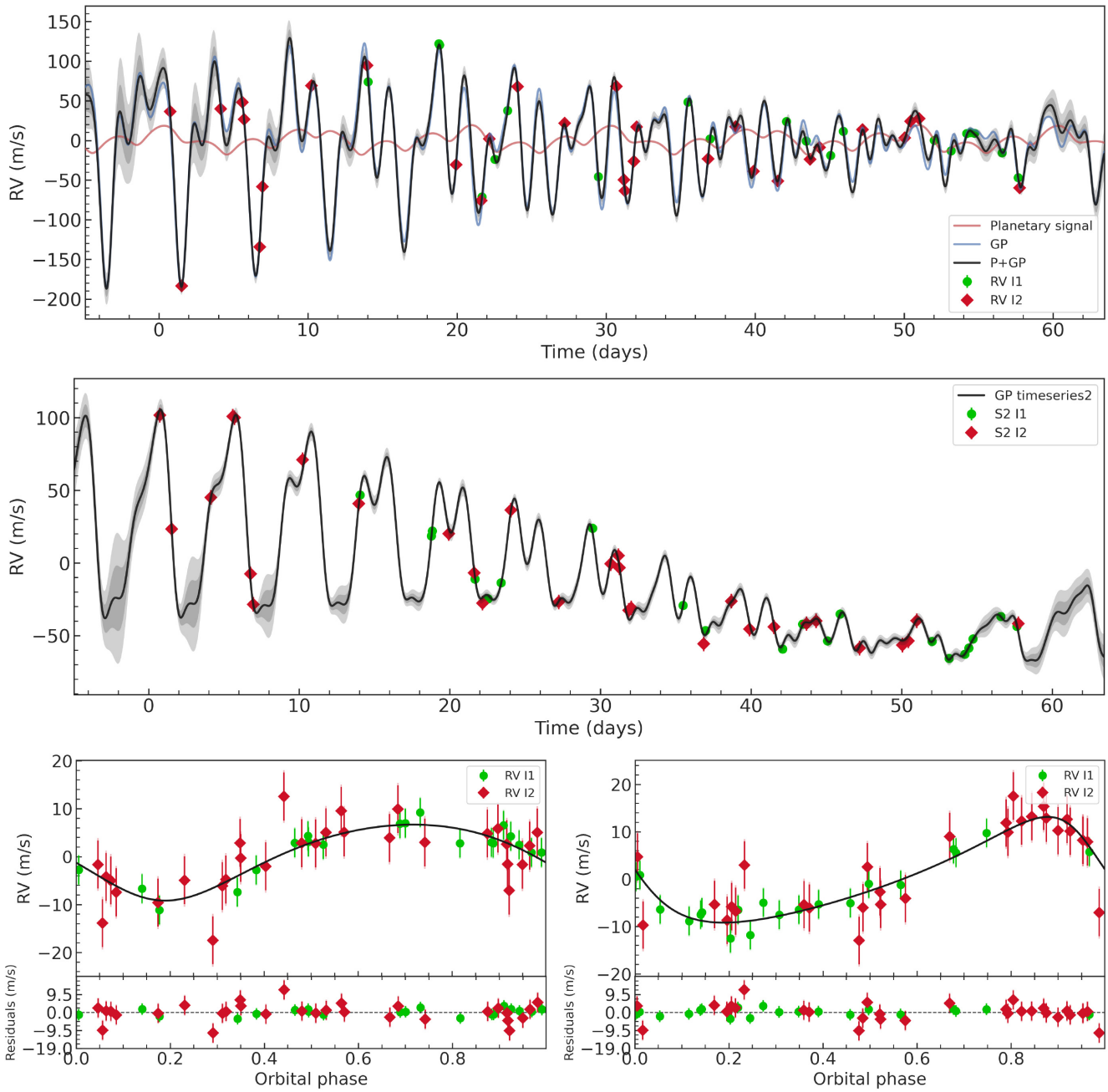

Figure 4. S_1, and S_2 time-series described in Section 5.2. Top panel: The green (instrument I_1) and red markers (instrument I_2) in each panel represent the synthetic measurements with inferred offsets extracted. Solid dark line and shadow regions are as in Fig. 3. The RV-like time-series also included the RV model for the two planets (red line). Bottom panel: Synthetic RV-like data folded on the orbital period of each injected planet following the subtraction of the systemic velocities, GP signal, and the other planet. The plots also show the inferred RV model for each planet (solid black line). These plots were generated automatically by `pyaneti`.

in the relevant paper and we also provide examples to reproduce the analyses in such papers.

5.5.1 K2-100

Barragán et al. (2019) published the RV detection of K2-100 b, a transiting exoplanet orbiting a young active star in the Praesepe cluster. In that manuscript we used the ability of the code to fit multiband transit photometry simultaneously with a 3D GP approach for spectroscopic time-series including a Keplerian component.

The data were modelled using a multidimensional GP approach with a QP kernel, for which the harmonic complexity detected in the spectroscopic time-series was relatively high ($\lambda_p = 0.6$). Therefore, the use of the GP derivatives for the detection of the planetary signal was crucial in this case (see discussion in Section 3.3). Fig. 8 shows a 20 d subset of the RV and log R'_{HK} time-series of fig. 2 in Barragán et al. (2019). This figure shows that the log R'_{HK} time-series behaves as the S_1 curve in the example in Section 3.3, and the RV data as the S_2 curve, i.e. the RV curve behaves as the time derivative of the log R'_{HK} curve. We note that K2-100 is a fast rotating and spot-dominated

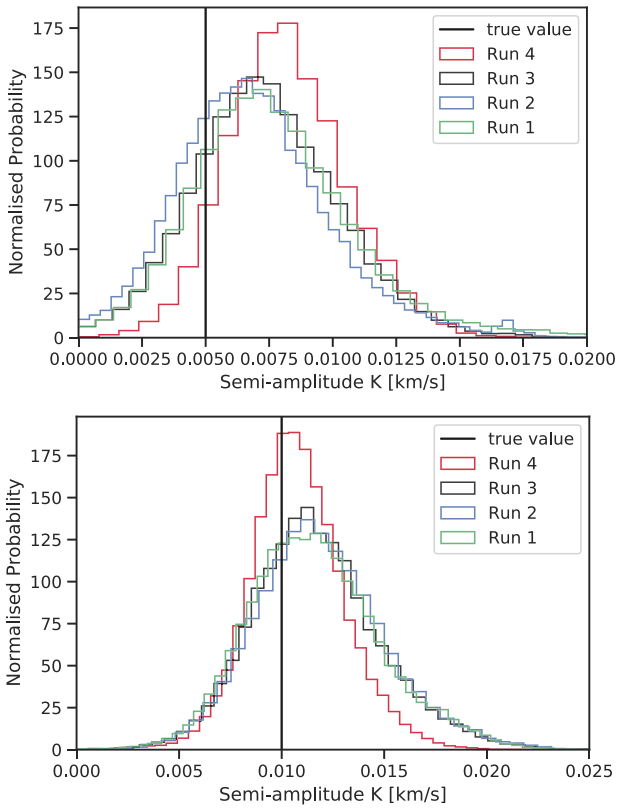

Figure 5. Probability distributions for the Doppler semi-amplitudes for the signals b (upper panel) and c (lower panel) as described in Section 5.2. Each colour corresponds to the different Runs described in the text. The true value of the parameter is shown with a vertical line in each case.

Table 3. System parameters for multiband toy model.

Parameter	Real value	Prior[(a)]	Inferred value[(b)]
$T_{0,\,b}$ (d)	4	$\mathcal{U}[3.95, 4.05]$	$3.9998^{+0.0012}_{-0.0014}$
P_b (d)	3	$\mathcal{U}[2.95, 3.05]$	$3.0001^{+0.0002}_{-0.00018}$
b_b	0.25	$\mathcal{U}[0, 1]$	$0.26^{+0.17}_{-0.17}$
$r_{p,b}/R_\star$	0.025	$\mathcal{U}[0, 0.1]$	$0.025^{+0.00039}_{-0.00027}$
$T_{0,\,c}$ (d)	3	$\mathcal{U}[2.95, 3.05]$	$3.0008^{+0.0011}_{-0.0011}$
P_c (d)	10	$\mathcal{U}[9.95, 10.05]$	$9.9995^{+0.00061}_{-0.00056}$
b_c	0.7	$\mathcal{U}[0, 1]$	$0.652^{+0.027}_{-0.04}$
$r_{p,c}/R_\star$	0.05	$\mathcal{U}[0, 0.1]$	$0.04898^{+0.00051}_{-0.00059}$
$\rho_\star$ (g cm^{-3})	1.4	$\mathcal{U}[0.01, 5]$	$1.41^{+0.55}_{-0.37}$
$q_{1,b1}$	0.06	$\mathcal{U}[0, 1]$	$0.04^{+0.044}_{-0.025}$
$q_{2,b1}$	0.50	$\mathcal{U}[0, 1]$	$0.49^{+0.33}_{-0.34}$
$q_{1,b2}$	0.56	$\mathcal{U}[0, 1]$	$0.666^{+0.111}_{-0.091}$
$q_{2,b2}$	0.33	$\mathcal{U}[0, 1]$	$0.231^{+0.093}_{-0.087}$

Note. – Same Note as Table 1.

young active star, and we expect that for young stars with similar characteristics, the GP derivative will be crucial to model the activity induced signals in the RV data.

We made available different setups to reproduce the analysis of the K2-100 system from the `pyaneti` main directory. To reproduce the multiband analysis of the system, the relevant command is `./pyaneti.py example_multiband_k2100`. To model the RV together with activity indicators, the command is `./pyaneti.py example_timeseries_k2100`. And to reproduce the full modelling including multiband transits and RV together with activity indicators, type `./pyaneti.py example_full_k2100`.

5.5.2 *TOI-1260*

Georgieva et al. (2021) report the detection and characterization of two mini-Neptunes transiting TOI-1260. They use `pyaneti` to perform joint transit and RV modelling using a multiplanet, multiband, and multidimensional GP configuration.

As for K2-100, the time-series were modelled using a multidimensional GP approach with a QP kernel. However, for this case, the harmonic complexity was moderate ($\lambda_p = 1.4^{+1.0}_{-0.5}$). Even in this case of moderate harmonic complexity, the multidimensional GP approach improved the planetary detection when compared with other common approaches (see Georgieva et al. 2021, for more details).

5.6 Execution time

All examples presented in Section 5 were run on a personal laptop with 8 × 1.90 GHz Intel® Core™ i7-8650U CPUs with Ubuntu 18.04. For comparison, we also ran all the examples with the same setup in a cluster with 12 × 3.40 GHz Intel® Xeon® CPUs. We compiled the code with `GFortran` and ran all the examples in parallel. Table 5 show the execution time for the different setups presented in Section 5.

We note that `pyaneti` is able to produce ready-to-publish results in a relative short time even on a personal laptop. However, the execution time can be longer, depending on several parameters, such as the MCMC configuration, number of planets being modelled, number of data, number of CPUs used, etc. For example, the setup `example_full_k2100` is a complex problem that combines multiband analysis and multidimensional GP regression. The total model samples 38 parameters simultaneously. Therefore, to reproduce the results by Barragán et al. (2019) requires a relatively high computational power.

6 CONCLUSIONS

In this manuscript we present a major update to the `pyaneti` code. This new version of `pyaneti` allows one to perform GP regression, as well as multiband and single transit fits. The biggest advantage of this new version of `pyaneti` with respect to other codes is the multidimensional GP regression that is included within the RV modelling routines. We present some tests to show how `pyaneti` can recover the model parameters in different data set configurations. We will continue updating `pyaneti` in the future according to the needs of the exoplanet community.

We described how we expanded the likelihood to account for data correlation using a multidimensional GP. In this manuscript, we use the bulit-in MCMC described in Barragán et al. (2019) to sample the parameter space. However, we note that this new likelihood can be used with other sampling techniques. In the near future we plan to add different sampling techniques in `pyaneti` with other MCMC (e.g. Foreman-Mackey et al. 2013; Karamanis, Beutler & Peacock 2021) sampling algorithms, as well as incorporate nested sampling (e.g. Feroz, Hobson & Bridges 2009; Handley, Hobson & Lasenby 2015; Speagle 2020) methods.

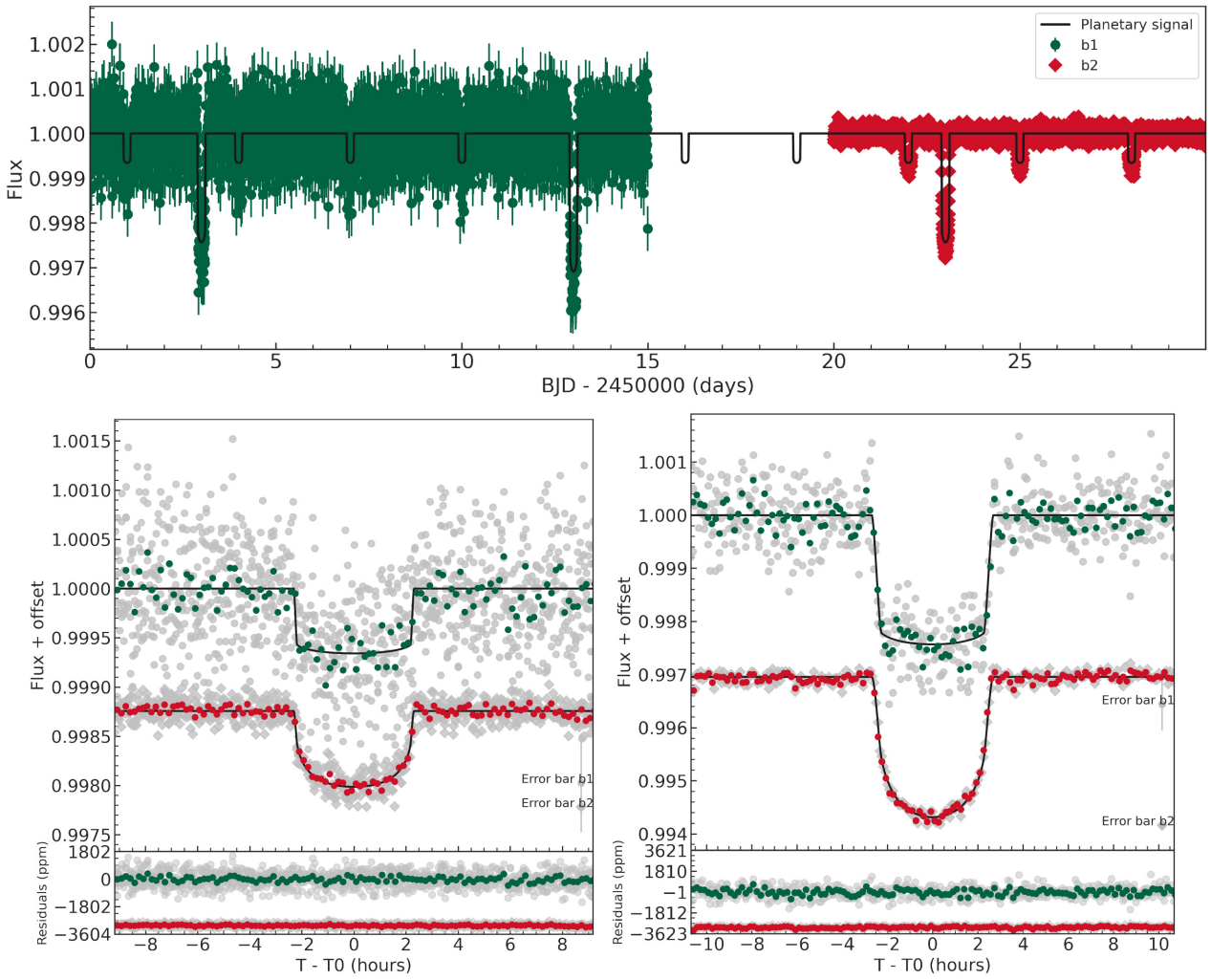

Figure 6. Top panel: Synthetic light curve created with the fictitious B1 (green circles) and B2 (red squares) instruments. The light-curve model with the two transiting signals is shown with a black thick line. Bottom panel: Phase-folded light curves for injected planet signal *b* (left) and *c* (right). Each plot shows the data for instrument B1 (grey circles) and B2 (grey squares) separated by an offset. The plots also show the data in 10-min bins for each instrument (B1 as green and B2 as red circles) together with the inferred model (black line) for each case. These plots are generated automatically by `pyaneti`.

Table 4. System parameters for single transit toy model.

Parameter	Real value	Prior[(a)]	Inferred value[(b)]
$T_{0,\,b}$ (d)	10	$\mathcal{U}[9.5, 10.5]$	$9.99999^{+0.00092}_{-0.00081}$
b_b	0.50	$\mathcal{U}[0, 1]$	$0.52^{+0.25}_{-0.37}$
$r_{p,\,b}/R_\star$	0.025	$\mathcal{U}[0, 0.1]$	$0.02485^{+0.0009}_{-0.00061}$
q_1	0.15	$\mathcal{U}[0, 1]$	$0.146^{+0.142}_{-0.083}$
q_2	0.50	$\mathcal{U}[0, 1]$	$0.3^{+0.39}_{-0.22}$
$a_{\rm dummy}$ [(c)]	···	$\mathcal{U}[1.1, 1000]$	$26.41^{+4.06}_{-6.8}$
Derived parameters			
$P_{b,\,\rm circ}$ (d)[(d)]	30	···	$34.1^{+47.6}_{-14.6}$

Note. – (*a*) and (*b*) same as Table 1. [(c)] This is a dummy scaled semimajor axis that `pyaneti` needs to sample to deal with the transit shape, it does not have a physical sense. [(d)] Note that the period is a derived parameter.

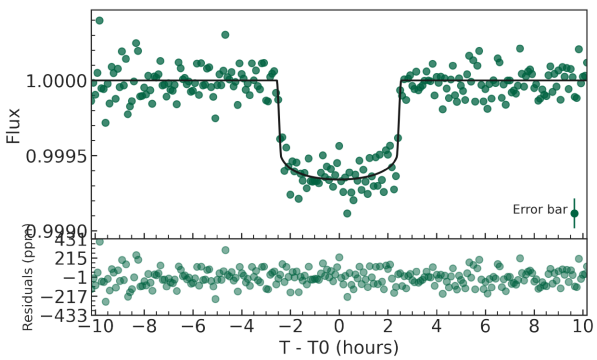

Figure 7. Model for the single transit modelling test. Synthetic data are shown as green circles. The inferred model is shown with a black line. This plot is shown as is provided automatically by `pyaneti`.

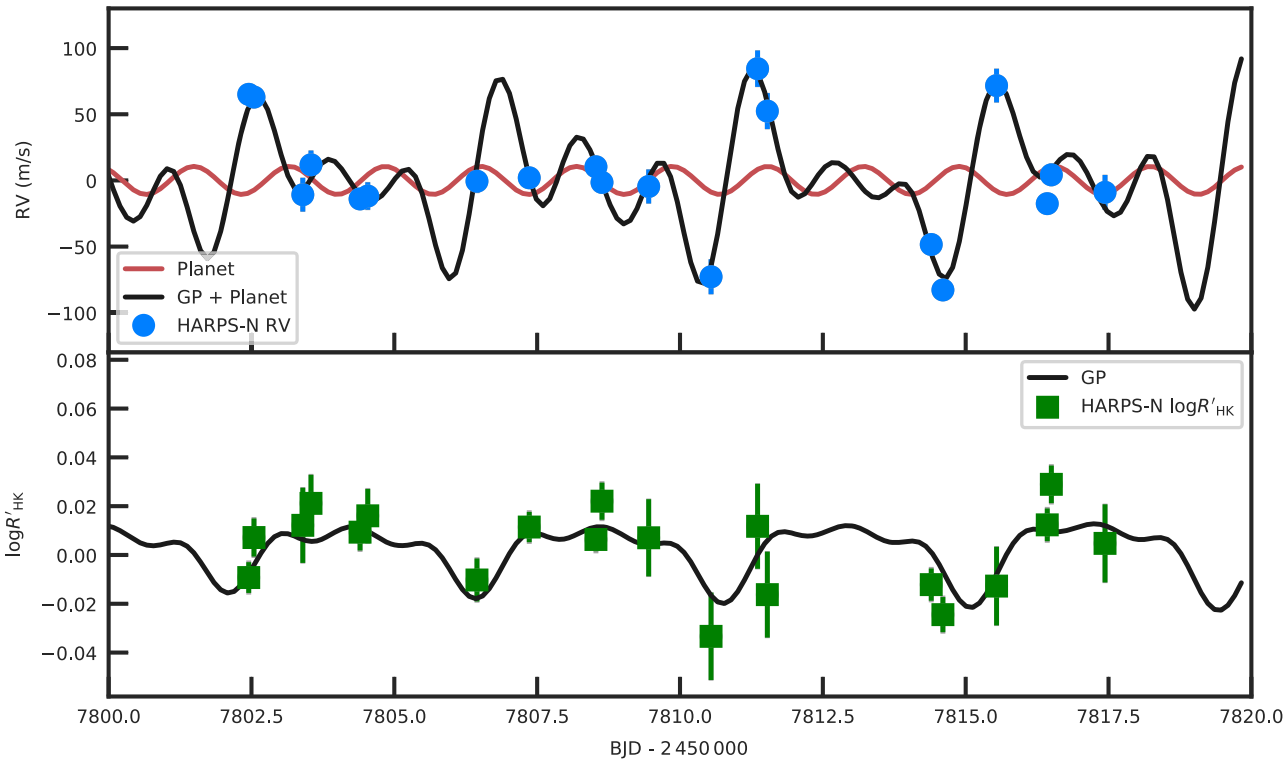

Figure 8. A 20 d subset of RV (upper panel) and log $R'_{\rm HK}$ (lower panel) time-series for K2-100. Both time-series have had the inferred offsets in Barragán et al. (2019) subtracted. Measurements are shown as filled symbols with error bars. For the RV time-series, we show the planet-induced signal alone (red line) and also the subtracted activity signal (black line). For the log $R'_{\rm HK}$ time-series, we show the activity model as a black line.

Table 5. Execution time for test runs in Section 5.

Setup	Time in laptop	Time in cluster
`example_toyp1`	$29^{\rm m}47^{\rm s}$	$11^{\rm m}21^{\rm s}$
`example_toyp2`	$23^{\rm m}10^{\rm s}$	$8^{\rm m}35^{\rm s}$
`example_multiband`	$17^{\rm m}59^{\rm s}$	$6^{\rm m}1^{\rm s}$
`example_single`	$1^{\rm m}03^{\rm s}$	$35^{\rm s}$
`example_timeseries_k2100`	$38^{\rm m}34^{\rm s}$	$18^{\rm m}51^{\rm s}$
`example_multiband_k2100`	$31^{\rm m}43^{\rm s}$	$12^{\rm m}20^{\rm s}$
`example_full_k2100`	$8^{\rm h}40^{\rm m}$	$3^{\rm h}16^{\rm m}$

Together with the code, we also provided some discussion on the use of GPs to model RV time-series. We pointed out how training GPs with activity indicators or light curves may not be the best approach for all cases of stellar activity. Even if the discussion is open, these aspects should be taken into account when modelling RV time-series using GPs.

We also presented `citlalicue` and `citlalatonac`. These numerical tools allow one to create synthetic photometric and spectroscopic time-series that may be useful to plan observing campaigns, especially in the cases in which the stellar signal is significantly larger than the planetary ones.

ACKNOWLEDGEMENTS

We thank the anonymous referee for their helpful suggestions that improved the quality of this paper. This publication is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 865624). NZ and SA acknowledge support from the UK Science and Technology Facilities Council (STFC) under Grant Code ST/N504233/1, studentship no. 1947725.

DATA AVAILABILITY

The data and software underlying this article are available as online supplementary material accessible following the links provided in the online version.

REFERENCES

Ahrer E. et al., 2021, MNRAS, 503, 1248
Aigrain S. et al., 2015, MNRAS, 450, 3211
Aigrain S., Pont F., Zucker S., 2012, MNRAS, 419, 3147
Alvarez M. A., Rosasco L., Lawrence N. D., 2011, preprint(arXiv:1106.6251)
Ambikasaran S., Foreman-Mackey D., Greengard L., Hogg D. W., O'Neil M., 2015, IEEE Trans. Pattern Anal. Mach. Intell., 38, 252
Anderson D. R. et al., 2011, ApJ, 726, L19
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Baluev R. V., 2013, Astron. Comput., 2, 18
Barragán O. et al. 2021a, preprint (arXiv:2120.13069)
Barragán O. et al., 2018, A&A, 612, A95
Barragán O. et al., 2019, MNRAS, 490, 698
Barragán O., Aigrain S., Gillen E., Gutiérrez-Canales F., 2021b, Res. Notes Am. Astron. Soc., 5, 51
Barragán O., Gandolfi D., Antoniciello G., 2019, MNRAS, 482, 1017
Batalha N. M., 2014, Proc. Natl. Acad. Sci., 111, 12647
Boisse I. et al., 2009, A&A, 495, 959
Bonfanti A., Gillon M., 2020, A&A, 635, A6
Bozza V., Mancini L., Sozzetti A., 2016, Methods of Detecting Exoplanets, Vol. 428, Springer, Heidelberg, Germany

Broeg C. et al., 2013, European Physical Journal Web of Conferences. p. 03005
Bruno G., 1584, De l'infinito, universo e mondi
Carleo I. et al., 2020, AJ, 160, 114
Charbonneau D., Brown T. M., Latham D. W., Mayor M., 2000, ApJ, 529, L45
Charbonneau D., Brown T. M., Noyes R. W., Gilliland R. L., 2002, ApJ, 568, 377
Chen Z., Fan J., Wang K., 2020, preprint(arXiv:2010.09830)
Coleman R., 1974, What is a Stochastic Process?. Springer, Netherlands, Dordrecht, p. 1
Collier Cameron A. et al., 2019, MNRAS, 487, 1082
Collier Cameron A. et al., 2021, MNRAS, 505, 1699
Cretignier M., Dumusque X., Allart R., Pepe F., Lovis C., 2020, A&A, 633, A76
Csizmadia S., 2020, MNRAS, 496, 4442
Díaz R. F., Almenara J. M., Santerne A., Moutou C., Lethuillier A., Deleuil M., 2014, MNRAS, 441, 983
Donati J.-F. et al., 2018, SPIRou: A NIR Spectropolarimeter/High-Precision Velocimeter for the CFHT. p. 107
Dumusque X., Boisse I., Santos N. C., 2014, ApJ, 796, 132
Eastman J. D. et al., 2019, preprint(arXiv:1907.09480)
Eastman J., Gaudi B. S., Agol E., 2013, PASP, 125, 83
Eisner N. L. et al., 2020, MNRAS, 494, 750
Eisner N. L. et al., 2021, MNRAS, 501, 4669
Espinoza N., Kossakowski D., Brahm R., 2019, MNRAS, 490, 2262
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Foreman-Mackey D. et al., 2021, JOSS, 6, 3285
Foreman-Mackey D., 2018, Res. Notes Am. . Society, 2, 31
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Fulton B. J., Petigura E. A., Blunt S., Sinukoff E., 2018, PASP, 130, 044504
Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 2004, Bayesian Data Analysis, 2nd edn. Chapman and Hall/CRC
Georgieva I. Y. et al., 2021, MNRAS, 505, 4684
Gilbertson C., Ford E. B., Jones D. E., Stenning D. C., 2020, ApJ, 905, 155
Grunblatt S. K., Howard A. W., Haywood R. D., 2015, ApJ, 808, 127
Günther M. N., Daylan T., 2021, ApJS, 254, 13
Handley W. J., Hobson M. P., Lasenby A. N., 2015, MNRAS, 450, L61
Harris C. R. et al., 2020, Nature, 585, 357
Hatzes A. P. et al., 2010, A&A, 520, A93
Hatzes A. P. et al., 2011, ApJ, 743, 75
Haywood R. D. et al., 2014, MNRAS, 443, 2517
Henry G. W., Marcy G., Butler R. P., Vogt S. S., 1999, IAU Circ. 7307
Hippke M., David T. J., Mulders G. D., Heller R., 2019, AJ, 158, 143
Isaacson H., Fischer D., 2010, ApJ, 725, 875
Jones D. E., Stenning D. C., Ford E. B., Wolpert R. L., Loredo T. J., Gilbertson C., Dumusque X., 2017, preprint(arXiv:1711.01318)
Karamanis M., Beutler F., Peacock J. A., 2021, MNRAS, 508, 3589
Kipping D. M., 2010, MNRAS, 408, 1758
Kipping D. M., 2013, MNRAS, 435, 2152
Mandel K., Agol E., 2002, ApJ, 580, L171
Mayo A. W. et al., 2019, AJ, 158, 165
Mayor M., Queloz D., 1995, Nature, 378, 355
Osborn H. P. et al., 2016, MNRAS, 457, 2273
Osborn H. P. et al., 2021, MNRAS, 502, 4842
Parviainen H. et al., 2019, A&A, 630, A89
Parviainen H., 2015, MNRAS, 450, 3233
Pepe F. A. et al., 2010, in McLean I. S., Ramsay S. K., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III. SPIE, Bellingham, p. 77350F
Pepe F. et al., 2013, Nature, 503, 377
Petigura E. A., Howard A. W., Marcy G. W., 2013, Proc. Natl. Acad. Sci., 110, 19273
Queloz D. et al., 2001, A&A, 379, 279
Rajpaul V. M., Aigrain S., Buchhave L. A., 2020, MNRAS, 492, 3960
Rajpaul V., Aigrain S., Osborne M. A., Reece S., Roberts S., 2015, MNRAS, 452, 2269
Rajpaul V., Aigrain S., Roberts S., 2016, MNRAS, 456, L6
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning, MIT Press
Rauer H. et al., 2014, Exp. Astron., 38, 249
Ricker G. R. et al., 2015, J. Astron. Telesc. Instrum. Syst., 1, 014003
Roberts S., Osborne M., Ebden M., Reece S., Gibson N., Aigrain S., 2013, Phil. Trans. R. Soc. A: Math. Phys. Eng. Sci., 371, 20110550
Santerne A. et al., 2015, MNRAS, 451, 2337
Speagle J. S., 2020, MNRAS, 493, 3132
Struve O., 1952, The Observatory, 72, 199
Suárez Mascareño A. et al., 2020, A&A, 639, A77
Thompson A. P. G., Watson C. A., de Mooij E. J. W., Jess D. B., 2017, MNRAS, 468, L16
Tracey B. D., Wolpert D. H., 2018, preprint(arXiv:1801.06147)
Trifonov T., 2019, Astrophysics Source Code Library, record ascl:1906.004

APPENDIX A: COVARIANCE FUNCTIONS AND THEIR DERIVATIVES

In this appendix, we show the form of the $\gamma_{i,j}^{G,G} = \gamma^{G,G}(t_i, t_j)$,

$$\gamma_{i,j}^{G,dG} = \gamma^{G,dG}(t_i, t_j) = \frac{\partial}{\partial t}\gamma^{G,G}(t_i, t)\Big|_{t=t_j}, \tag{A1}$$

$$\gamma_{i,j}^{dG,G} = \gamma^{dG,G}(t_i, t_j) = \frac{\partial}{\partial t}\gamma^{G,G}(t, t_j)\Big|_{t=t_i}, \tag{A2}$$

and

$$\gamma_{i,j}^{dG,dG} = \gamma^{dG,dG}(t_i, t_j) = \frac{\partial}{\partial t'}\left(\frac{\partial}{\partial t}\gamma^{G,G}(t', t)\Big|_{t=t_i}\right)\Big|_{t'=t_j}, \tag{A3}$$

for the squared exponential, Matérn 5/2, and Quasi-periodic kernels. These quantities can be used to compute the matrices given in equation (9) needed to compute the big covariance matrix given in equation (10) for the multidimensional GP regression.

A1 Square exponential kernel

The squared exponential covariance function is written as

$$\gamma_{\mathrm{SE},i,j}^{G,G} = \exp\left[-\frac{(t_i - t_j)^2}{2\lambda^2}\right], \tag{A4}$$

where we have omitted the amplitude term shown in equation (3). The covariance between the derivative observation i and the non-derivative observation j is

$$\gamma_{\mathrm{SE},i,j}^{G,dG} = -\gamma_{\mathrm{SE},i,j}^{dG,G} = \frac{t_i - t_j}{\lambda^2}\gamma_{\mathrm{SE},i,j}^{G,G}, \tag{A5}$$

and the covariance between the derivative observation i and the derivative observation j is

$$\gamma_{\mathrm{SE},i,j}^{dG,dG} = \left[\frac{1}{\lambda^2} - \frac{(t_i - t_j)^2}{\lambda^4}\right]\gamma_{\mathrm{SE},i,j}^{G,G}. \tag{A6}$$

A2 Matérn 5/2 Kernel

If we define

$$t_{5/2} \equiv \frac{\sqrt{5}\,|t_i - t_j|}{\lambda}, \tag{A7}$$

we can write the Matérn 3/2 Kernel as

$$\gamma_{\mathrm{M52},i,j}^{G,G} = \left(1 + t_{5/2} + \frac{t_{5/2}^2}{3}\right)\exp\left(-t_{5/2}\right), \tag{A8}$$

where we have omitted the amplitude term shown in equation (4). The covariance between the derivative of the observations i and the non-derivative observations of j is then

$$\gamma^{G,dG}_{\mathrm{M52},i,j} = -\gamma^{dG,G}_{\mathrm{M52},i,j} = \frac{\sqrt{5}}{3\lambda} t_{5/2} \left(1 + t_{5/2}\right) \exp\left(-t_{5/2}\right) \mathrm{sgn}(t_i - t_j), \tag{A9}$$

where sgn is the sign function. The covariance between the derivative of the observations i and the derivative observations of j is then

$$\gamma^{dG,dG}_{\mathrm{M52},i,j} = -\frac{5}{3\lambda^2} \left(t^2_{5/2} - t_{5/2} - 1\right) \exp\left(-t_{5/2} t\right). \tag{A10}$$

A3 Quasi-Periodic Kernel

In order to use the QP kernel in the multidimensional GP approach, we need to use equation (6) without amplitude term as

$$\gamma^{G,G}_{\mathrm{QP},i,j} = \exp\left\{ -\frac{\sin^2\left[\pi \left(t_i - t_j\right)/P_{\mathrm{GP}}\right]}{2\lambda_{\mathrm{p}}^2} - \frac{\left(t_i - t_j\right)^2}{2\lambda_{\mathrm{e}}^2} \right\}, \tag{A11}$$

If we define

$$\tau \equiv \frac{2\pi(t_i - t_j)}{P_{\mathrm{GP}}}, \tag{A12}$$

the covariance between the derivative observation i and the non-derivative observation j for the QP kernel given by equation (A11) is

$$\gamma^{G,dG}_{\mathrm{QP},i,j} = -\gamma^{dG,G}_{\mathrm{QP},i,j} = -\gamma^{G,G}_{\mathrm{QP},i,j} \left(\frac{\pi \sin\tau}{2P_{\mathrm{GP}}\lambda_{\mathrm{p}}^2} + \frac{t_i - t_j}{\lambda_{\mathrm{e}}^2} \right). \tag{A13}$$

Finally, the covariance between the derivative observation i and the derivative observation j is then

$$\begin{aligned} \gamma^{dG,dG}_{\mathrm{QP},i,j} = \gamma^{G,G}_{\mathrm{QP},i,j} \Bigg[& -\left(\frac{\pi \sin\tau}{2P_{\mathrm{GP}}\lambda_{\mathrm{p}}^2} \right)^2 - \frac{\tau \sin\tau}{2\lambda_{\mathrm{p}}^2\lambda_{\mathrm{e}}^2} + \frac{\pi^2 \cos\tau}{P_{\mathrm{GP}}^2\lambda_{\mathrm{p}}^2} \\ & - \left(\frac{t_i - t_j}{\lambda_{\mathrm{e}}^2} \right)^2 + \frac{1}{\lambda_{\mathrm{e}}^2} \Bigg]. \end{aligned} \tag{A14}$$

This paper has been typeset from a TEX/LATEX file prepared by the author.