The radial velocity method is one of the most successful techniques for detecting exoplanets. It works by detecting the velocity of a host star induced by the gravitational effect of an orbiting planet, specifically the velocity along our line of sight, which is called the radial velocity of the star. As astronomical instrumentation has improved, radial velocity surveys have become sensitive to low-mass planets that cause their host star to move with radial velocities of 1 m/s or less. While analysis of a time series of stellar spectra can in theory reveal such small radial velocities, in practice intrinsic stellar variability (e.g., star spots, convective motion, pulsations) affects the spectra and often mimics a radial velocity signal. This signal contamination makes it difficult to reliably detect low mass planets and planets orbiting magnetically active stars. A principled approach to recovering planet radial velocity signals in the presence of stellar activity was proposed by Rajpaul et al. (2015) and involves the use of a multivariate Gaussian process model to jointly capture time series of the apparent radial velocity and multiple indicators of stellar activity. We build on this work in two ways: (i) we propose using dimension reduction techniques to construct more informative stellar activity indicators that make use of a larger portion of the stellar spectrum; (ii) we extend the Rajpaul et al. (2015) model to a larger class of models and use a model comparison procedure to select the best model.
for the particular stellar activity indicators at hand. A novel aspect of the Rajpaul et al. (2015) model and our larger class of models is that they use both a Gaussian process and its derivatives, which imposes scientifically motivated structure. By combining our high-information stellar activity indicators, Gaussian process models, and model selection procedure, we achieve substantially improved planet detection power compared to previous state-of-the-art approaches.

1. Motivation. In this paper, we present a statistical framework to improve the sensitivity of astronomical surveys for detecting exoplanets, i.e., planets orbiting stars other than the Sun. As a planet orbits a star, the planet’s gravitational force causes the star to orbit around the center of mass of the system. Consequently, the starlight appears to be alternately shifted to longer (redder) and shorter (bluer) wavelengths, as the star moves away from and towards the observer, respectively, due to the Doppler effect caused by this motion.

While the Doppler effect due to a planet is very small, modern astronomical instrumentation can detect these Doppler shifts by carefully analyzing the spectrum (i.e., the intensity of starlight as a function of wavelength) of the host star (e.g., Butler et al., 1996; Baranne et al., 1996; Mayor et al., 2003). An example spectrum is shown in the left panel of Figure 1. The spectrum contains many tens of thousands of spectral “lines” (i.e., dips in the stellar spectrum) due to absorption of light of specific wavelengths determined by the quantum mechanical energy levels of atoms in the upper layers of the star’s atmosphere. The spectral lines have non-zero width due to several factors including the rotation of the star, which causes some blurring across wavelengths. Figure 2 (discussed in Section 3.2) zooms in on one small portion of the spectrum, so individual spectral lines can be distinguished. The Doppler effect is detected by measuring how the observed wavelengths of the spectral lines shift from one observation to another. In practice this shift is extremely small, but astronomers are able to obtain the high precision needed by analyzing the spectrum over a broad range of wavelengths containing many thousands of spectral lines.

The velocity of a star projected onto the observer’s line of sight is known as the radial velocity (RV) of the star. Currently, radial velocities can be determined up to ∼1 m/s through the Doppler shift analysis described above. To hunt for planets, astronomers choose candidate stars and for each collect a time series of radial velocity observations spanning many days up to years (e.g., Mayor et al., 2011; Pepe et al., 2011; Fischer, Marcy and Spronck, 2013; Butler et al., 2017). The shape of the RV signal expected due to a planet is well understood based on Newton’s laws of motion and gravity.
Fig 1: Example stellar spectrum (left) and an illustration of the scientific model for a radial velocity signal due to a planet (right). There are many spectral lines in the left panel and most of them cannot be separated by eye.

The right panel of Figure 1 shows an example RV signal produced by the scientific model for a planet with a 17 day orbital period (further details of the scientific model are given in Section 3.1). The magnitude of the periodic changes in the RV signal are related to the mass of the planet, the mean separation between the star and planet, and the viewing geometry. Hunting for periodic RV signals such as that in Figure 1 is known as the radial velocity method and is one of the most successful methods for detecting exoplanets.

A key challenge for the radial velocity method, and the focus of this paper, is that intrinsic stellar variability, such as magnetic activity on the surface of the star, can mimic a Doppler shift signals. In the past, such perturbations have been modeled as an additional noise term, referred to as RV “jitter” (e.g., Ford, 2006). However, spurious RV signals arise due to physical processes that vary with stellar type. For main sequence stars similar to or slightly cooler than the Sun, there is an envelope of convective gas at the stellar surface (technically, the photosphere). Stellar magnetic fields affect the convective motion, leading to groups of small, dark and relatively cool star spots, often surrounded by brighter regions, known as faculae. As a result, the morphology of the spectral lines changes over time due to a combination of the convective motions in the vicinity of the star spots and the changing radial velocity of the spot regions as the star rotates.

To help visualize the way that spots (or faculae) affect the stellar spectrum and can induce an apparent RV signal, consider a star with rotational axis in the plane of the sky. As the star rotates, the light from one side is red shifted and the light from the other side is blue shifted. For a star with uniform (or
axisymmetric) surface brightness, these effects broaden each spectral line, but do not result in a net shift of the spectral lines. Next, consider adding a single, non-evolving spot on the surface of the star. When the spot is on the hemisphere of the star hidden from the observer, the unperturbed stellar spectrum is observed. Once the spot rotates onto the hemisphere facing the observer, the stellar spectrum is affected by the size, magnitude and location of the spot. When the spot is visible and on the side of the star rotating towards (away from) the observer, the amount of blue-shifted (red-shifted) light is decreased, causing a distortion in the spectral line shapes. At the spectral resolution of a typical planet-hunting spectrograph, the distortion will not be resolved. When comparing spectra with and without the spot, the spectral line will appear to be shifted or skewed, creating an apparent Doppler shift. Thus, as a spot (or faculae) rotates across the stellar disk it leads to a spurious signal in the RV time series. Because the strength and location of star spot (faculae) groups typically evolve on timescales somewhat longer than the rotational period, these apparent Doppler shifts in the RV time series can be quasi-periodic, where the period is close to that of the stellar rotation. Because stellar rotation periods are often similar to plausible orbital periods for planets, distinguishing spots and faculae from planets can be particularly problematic (e.g., see discussion of Dumusque et al., 2012 in Rajpaul et al., 2015 and Rajpaul, Aigrain and Roberts, 2015). A further spurious RV signal is caused by the strong magnetic fields inside the spot (and facula) regions which reduce convection and therefore blueshift (see Dravins, Lindegren and Nordlund (1981), Cavallini, Ceppatelli and Righini (1985), Meunier, Desort and Lagrange (2010), and Dumusque, Boisse and Santos (2014)). However, such signals are not necessarily on the same timescales as planetary orbital periods.

In addition to magnetically active regions, namely spots and faculae, other forms of stellar variability affect the stellar spectrum. For example, for some stars, stellar variability can be dominated by stellar pulsations or granulation, the spatially correlated patterns of convective motions near the star surface. While also important, these forms of variability typically cause apparent Doppler shifts that change on timescales less than two days (see Del Moro et al., 2004; Del Moro, 2004; Arentoft et al., 2008). Since typical planetary orbital periods are substantially longer than two days, these additional forms of variability can be mitigated using alternative strategies to those required to deal with magnetically active regions. Therefore, this paper instead focuses on the effects of localized magnetically active regions.

---

1In astronomy, the term *stellar activity* tends to refer to magnetically active regions (the focus here), whereas *stellar variability* is more general.
particularly stellar spots. Nevertheless, we anticipate that our statistical framework may prove useful for additional forms of stellar variability. In particular, all forms of stellar variability share some common features that can be used to distinguish them from true Doppler shifts due to planets. First, a true Doppler shift affects the entire spectrum in the same way, i.e., all the spectral lines are shifted. In contrast, intrinsic stellar variability affects different lines differently, depending on the properties of each spectral line and the height in the star’s photosphere at which the light absorption takes place (and the nature of the variability). Second, unlike a pure Doppler shift, intrinsic stellar variability often changes the shapes of spectral lines, with typical effects being increased skewsness, broadening, and depth changes. Finally, while a single planet induces a strictly periodic Doppler shift, apparent Doppler shifts due to stellar variability will not be strictly periodic. For instance, the size, strength and location of spots and faculae evolve, resulting in quasi-periodic perturbations. For further details on the physical nature of stellar variability, the reader is referred to the relevant astronomy literature, e.g., Jenkins (2013), Dumusque, Boisse and Santos (2014), Borgniert, Meunier and Lagrange (2015), and Haywood et al. (2016).

Careful study and modeling of stellar activity is crucial because it is currently one of the leading challenges for exoplanet detection efforts. This is especially true in the hunt for Earth-like planets because these typically produce only small RV signals that are easily drowned out by spurious signals due to spots and other phenomena. Given the three key differences in how planets and stellar activity affect the observed spectrum described above, it is in principle possible to distinguish stellar activity from true Doppler shifts. However, progress towards developing statistical frameworks to achieve this separation in practice has so far been limited. Recently, Davis et al. (2017) demonstrated that stellar activity imparts a detectable signal on the stellar spectrum for the resolution and signal-to-noise (SNR) of upcoming astronomical spectrographs. However, that study did not provide a means of making use of the detailed information in spectroscopic time series. A sophisticated Gaussian process framework to model spectroscopic information about stellar activity has been proposed by Rajpaul et al. (2015), and is reviewed in Section 2.1. However, the summaries of the stellar activity that they model are likely less informative than those constructed by Davis et al. (2017), and they do not explain how their model can be adapted to make use of other summaries. Furthermore, the impact of the stellar activity modeling that Rajpaul et al. (2015) propose on planet detection power is unclear.

This paper addresses the two-fold challenge posed by stellar activity for exoplanet detection: (i) “How can one construct spectroscopic activity indi-
cators that quantify the different effects of activity on an individual spectrum?”, and (ii) “How can a statistical model integrate information contained in the temporal evolution of these activity indicators, so as to further constrain the spurious RV signals produced by the activity and thereby separate them from the true RV perturbations induced by planets?”. We propose a general approach to tackling these questions and demonstrate its potential using realistic stellar spectra generated with the Spot Oscillation and Planet (SOAP) 2.0 simulation software developed by Dumusque, Boisse and Santos (2014). In particular, rather than introducing a single model for all stellar activity indicators, we propose a general class of multivariate Gaussian process models and develop a model selection procedure to find the best model for the specific activity indicators at hand. Secondly, instead of specifying stellar activity in advance, we introduce a way to adapt dimension reduction techniques and thereby construct customized high-information stellar activity indicators from the data. The construction of these indicators incorporates a simple pre-processing step that separates them from RV signals and thereby preserves interpretability and facilitates modeling. We illustrate this indicator construction procedure for SOAP 2.0 data using principal component analysis (PCA) as the base dimension reduction algorithm. For both these new activity indicators and some existing ones, we show that applying our model selection procedure leads to substantially higher planet detection power than simply using an existing model from the literature.

Our paper is organized as follows. Section 2 reviews the state of the art in stellar activity modeling and stellar activity indicators. Section 3 discusses how we generate simulated data with which we test our methodology, including both planetary radial velocity signals and spectral perturbations due to stellar activity. Section 4 presents our general class of statistical models for jointly capturing stellar activity indicators and the RV time series, and then introduces our two stage model selection procedure. Section 5 explains and illustrates our method for constructing new high-information stellar activity indicators. Section 6 applies our model selection procedure in the case of our new indicators and compares the resulting planet detection power to that for an existing model in the literature. We discuss the results, the implications for future RV planet surveys, and areas for future research in Section 7. Our code and simulated spectra are available on GitHub at https://github.com/djones2013/improving-planet-detection-power.

2. Overview of State-of-art in Doppler Planet Surveys.

2.1. State-of-the-art stellar activity modeling. Recently, Rajpaul et al. (2015) (hereafter R15) proposed a flexible stellar activity model that jointly
captures changes in the apparent RV time series due to stellar activity and times series of two stellar activity indicators (sometimes called stellar activity proxies). When a planet is suspected an astronomer can assess how well the R15 stellar activity model fits with and without a component explaining a planet signal. If the planet signal component is found to be needed then this suggests that there is a planet present, e.g., R15 use the Bayesian information criterion (BIC) to assess the evidence for a planet. We consider the R15 model to be representative of the state-of-the-art of stellar activity modeling because it is one of a few existing models in the literature that meets the key criteria for good performance identified in a recent comparison of methods by Dumusque et al. (2017). The criteria include modeling both apparent RV signals and stellar activity indicators, using a red-noise model for apparent RV signals, and performing hypothesis tests for planets in a statistically justified way. In addition to satisfying these criteria, the R15 model is also physically well-motivated and uses relatively robust numerical methods. Furthermore, the R15 model (and slight variants) has proved useful for both detecting and characterizing low-mass exoplanets and for recognizing spurious claims of planets that were in fact due to stellar activity, see R15 and Rajpaul, Aigrain and Roberts (2015). Therefore, the R15 model will serve as a point of reference for our study.

Let \( u(t) \) and \( q_j(t) \), for \( j = 1, 2 \), denote the values of the apparent RV and two stellar activity indicators, respectively, at time \( t \). Given observation times \( t_1, \ldots, t_n \), the R15 model is

\[
\begin{align*}
\text{(2.1)} & \quad u(t_i) = m_0 + a_{01}X(t_i) + a_{02}\dot{X}(t_i) + \epsilon_{0i} \\
\text{(2.2)} & \quad q_1(t_i) = m_1 + a_{11}X(t_i) + \epsilon_{1i} \\
\text{(2.3)} & \quad q_2(t_i) = m_2 + a_{21}X(t_i) + a_{22}\dot{X}(t_i) + \epsilon_{2i}
\end{align*}
\]

where \( X(t) \) and its derivative \( \dot{X}(t) \) are unknown functions and the \( \epsilon_{ji} \)'s are independent random errors with

\[
\text{(2.4)} \quad \epsilon_{ji} \sim N(0, \sigma_j^2),
\]

for \( j = 0, 1, 2 \) and \( i = 1, \ldots, n \). The \( m_j \)'s, \( \sigma_j \)'s, and \( a_{jk} \)'s are unknown parameters to be inferred from the data, for \( j = 0, 1, 2 \) and \( k = 1, 2 \) (note, \( a_{12} = 0 \)). In words, the three time series are each modeled as a linear combination of some function \( X(t) \) and its derivative \( \dot{X}(t) \), plus some random noise. Since the function \( X(t) \) (and \( \dot{X}(t) \)) is unknown it is modeled as a zero mean Gaussian process. Gaussian processes play an important part in our modeling and will be reviewed briefly in Section 4.1. The reader is referred to Rasmussen and Williams (2006) for a comprehensive introduction.
to Gaussian processes and to R15 for full details of their model. In the atmospheric sciences literature, Hewer et al. (2017) also make use of a model incorporating derivatives of a Gaussian process.

The model (2.1)-(2.4) is for a specific choice of the stellar activity indicators $q_1$ and $q_2$ (discussed in Section 2.2). Its form was motivated by physical arguments in Aigrain, Pont and Zucker (2012) which demonstrate that under a simple spot model the apparent RV can be expressed as a function of the area of the spot projected onto our line of sight and the derivative of this area. When the activity indicators available are similar to those of R15, then this model can be applied unaltered. However, for more general indicators or other forms of stellar activity, it is not immediately clear exactly how the model should be adapted without developing new physical justifications. Our approach detailed in Section 4 is to define a larger class of models and select the best model for the activity indicators at hand.

2.2. Stellar activity indicators. Stellar activity indicators are functionals of individual spectra that summarize the level and nature of stellar activity at the time of the observation. Their purpose is to reduce the notion of stellar activity from complex changes in high-dimensional stellar spectra to a few simple time series. A number of photometric and spectroscopic stellar activity indicators have been proposed, the latter usually being more promising (because spectroscopic observations are much higher resolution than photometric observations). The spectroscopic indicators used by R15 are $\log R'_{\text{HK}}$ and BIS, which measure the relative size of two specific spectral lines (compared with a reference) and the asymmetry of spectral lines, respectively. That is, in R15 the indicators $\log R'_{\text{HK}}$ and BIS take the place of the generic indicators $q_1$ and $q_2$ in equations (2.2) and (2.3), respectively. R15 also use normalized flux (light intensity) as a photometric stellar activity indicator. Specifically, they replace $\log R'_{\text{HK}}$ by normalized flux in the case of SOAP 2.0 data (see Section 3.2 below), which does not include $\log R'_{\text{HK}}$.

Some stellar activity measures are summaries of very specific parts of the stellar spectrum, as in the case of $\log R'_{\text{HK}}$. More generally a stellar activity indicator can be any functional $g$ of the stellar spectra observed. This raises the question of how to choose the functional $g$ in order to capture as much information as possible. Indeed, the measures $\log R'_{\text{HK}}$ and BIS were designed for purposes other than planet detection, and therefore in the current context are somewhat arbitrary. There is no reason to expect that such indicators

\[ \text{In technical terms, } \log R'_{\text{HK}} \text{ is a measure of the chromospheric emission induced by the magnetic fields present inside faculae and spots. It is therefore an indirect measure of the filling factor of spots and faculae. BIS is the inverse slope of the bisector of the cross-correlation function.} \]
capture all of the relevant information about stellar activity contained in the spectra. Given the subtle effects of stellar activity on the stellar spectrum, the most powerful activity indicators must combine information from many spectra lines. Therefore, we suggest an approach to address the challenge of choosing the best $g$ in Section 5.

3. Physical Models for Generating Input Data.

3.1. Keplerian model for planetary RV signals. The RV signal due to a single planet orbiting a star is well understood and can be described precisely using a Keplerian model, see for example Danby (1988). Specifically, the RV induced by a single planet system is given by

$$v(t) = K(e \cos \omega + \cos(\omega + \phi(t))) + \gamma$$

where $e$ is the eccentricity of the orbit, $\omega$ is an orbital orientation angle known as the argument of periapsis, and $\gamma$ and $K$ are velocity offset and amplitude parameters, respectively. The offset $\gamma$ corresponds to the average motion of the planetary system’s center of mass relative to the center of mass of the Solar System and any instrument velocity offset. The angle $\phi(t)$ is called the true anomaly and indicates the phase of the star in its elliptical orbit of the center of mass. This angle is determined by a system of three equations which are given in Appendix A and which depend on the planet orbital period $\tau_p$ (as well as $e$ and another physical parameter $M_0$ called the mean anomaly at $t = 0$). The model (3.1) accurately captures the RV signal due to a planet and is therefore appropriate for producing realistic simulations on which to test our methodology. See the right of Figure 1 for an example planet RV signal.

As described in Section 1, the planet signal $v(t)$ is often corrupted by an apparent RV signal due to stellar activity which we must model in order to reliably detect the planet and infer its basic properties. In practice there are additional challenges including highly irregular observations, seasonal observation gaps, and multiple planet or star systems. Our models are flexible and can in principle incorporate such complications but here we focus our attention on stellar activity because currently this is the main obstacle limiting our ability to detect low-mass exoplanets.

3.2. SOAP 2.0 model for stellar spectra. The Spot Oscillation and Planet (SOAP) 2.0 software simulates spectra for a star with spots or faculae. Each spectrum is a vector giving the expected value of the observed intensity of the light from the star at many different wavelengths. From these expected
Fig 2: Zoomed view of the spectrum of a star with a single spot (generated by SOAP 2.0). The blue line shows the spectrum when the spot is behind the star (and therefore not effecting the spectrum), and the red line shows the spectrum when the spot is on the visible hemisphere of the star and the spot center projects onto the star’s rotational axis.

intensities, observed intensities can be simulated using a Poisson model. SOAP 2.0 generates simulated spectra by computing linear combinations of real, spatially resolved spectra of quiet and active regions of the Sun, accounting for both rotational Doppler shifts and limb darkening (an effect causing the surface brightness at the center of the star’s disk to be greater than regions near the edge of disk). Thus, SOAP 2.0 simulations are highly realistic for the Sun and likely useful for gaining insights about the stellar activity of many Sun-like stars. SOAP and SOAP 2.0 were developed by Boisse, Bonfils and Santos (2012) and Dumusque, Boisse and Santos (2014), respectively.

Using SOAP 2.0 we simulate \( n = 125 \) stellar spectra at equally spaced stellar rotation phases in the case of a single spot that covers 1% of one hemisphere of the star surface (and no planet). Specifically, the stellar rotation phases at which spectra are generated are \(-0.5, -0.496, -0.492, \ldots, 0.496\). Since SOAP 2.0 outputs spectra at a much higher spectral resolution (\( \sim 10^6 \)) than can be expected for current and near-future instruments, we reduce the resolution to the more realistic value of \( 1.5 \times 10^5 \). (In particular, we convolve the higher resolution spectra with a Gaussian line spread function and resample using cubic splines.) This reduction in resolution corresponds to a reduction in the number of wavelengths recorded from 523,732 in the SOAP 2.0 output to \( p = 237,944 \) in our datasets. We set the key configuration parameters, namely the inclination of the stellar rotation axis relative to the line of sight and the spot latitude from the stellar equator, to 90° and 40°,
respectively. We write the SOAP 2.0 data as an $n \times p$ matrix $Y$, where the rows and columns of $Y$ correspond to the $n$ phases and $p$ wavelengths, respectively. In Section 5, we use this ideal data to construct stellar activity indicators and as a starting point for generating realistic final datasets for the analysis and model selection in Section 6, i.e., datasets incorporating noise and with observation times spread over many stellar rotations.

A time series of the apparent RV signal due to the spot can straightforwardly be computed by applying existing methodology to the spectra $Y$. However, the spot has complex effects on the spectrum in addition to a Doppler shift. This can be seen in Figure 2 which shows part of a simulated spectrum when the spot is behind the star (blue) and again when the spot is closest to us (red). The Doppler shift is very small and is not visible, but many other changes can be seen. The effects of the spot on the spectrum are not easy to summarize but Section 6 demonstrates that they often allow us to separate the RV corruption from a planet signal, at least for SOAP 2.0 simulated data.

4. Stellar activity modeling framework.

4.1. General class of stellar activity models. The R15 Gaussian process based model in (2.1)-(2.4) is designed for the specific stellar activity indicators $\log R'_{\text{HK}}$ (or normalized flux) and BIS. We now extend their model to a class of models that can capture a greater variety of behaviors of stellar activity indicators. We begin with a brief review of Gaussian processes.

A real-valued continuous time stochastic process $\{X(t)\}$ is a Gaussian process (GP) if for every finite set of times $t_1, \ldots, t_n$ the vector $(X(t_1), \ldots, X(t_n))$ has a multivariate Gaussian distribution. Spatial processes and other kinds of processes can also be Gaussian processes, but here all Gaussian processes that we refer to will be temporal. A Gaussian process is specified via a mean function and a covariance function which determine what the mean vector and covariance matrix should be for the vector $(X(t_1), \ldots, X(t_n))$ given any set of times $t_1, \ldots, t_n$. Here we make the usual assumptions that for any $t$ the mean of $X(t)$ is zero and the process is stationary in time so that the covariance between two observations of the process $X(t)$ and $X(t')$ only depends on the value of $|t - t'|$. Specifically, we adopt the quasi-periodic covariance function

$$K(t, t') = \exp \left( \frac{-\sin^2(\pi(t - t')/\tau_s)}{2\lambda_p^2} - \frac{(t - t')^2}{2\lambda_e^2} \right),$$

(4.1)

where $\tau_s$ is the stellar period and $\lambda_p$ and $\lambda_e$ are parameters governing the relative importance of periodic and local correlations and the time-scale of
local correlations, respectively. The fact that this covariance function leads
to a positive definite covariance matrix follows from the fact that the prod-
uct of two valid covariance functions yields a valid covariance function, see
Rasmussen and Williams (2006) page 95. In the current case the two under-
lying covariance functions are the standard periodic and squared exponen-
tial covariance functions, which correspond to the first and second terms of
the exponent in (4.1), respectively. The covariance function (4.1) was also
adopted by R15. A quasi-periodic covariance function makes sense in the
current context because for a fixed spot the stellar activity signal should
be similar for each rotation of the star but will change over longer intervals
due to evolution of the spot or other phenomena not explicitly modeled. We
denote the parameters of the covariance function by \( \phi = (\tau_s, \lambda_p, \lambda_e) \).

The R15 model given by (2.1)-(2.4) also makes use of the derivative of
\( X(t) \). The derivative of a Gaussian process with covariance function
\( K(t, t') \) (if it exists as is the case for (4.1)) is also a Gaussian process and has
covariance function

\[
\frac{\partial^2}{\partial t \partial t'} K(t, t').
\]

Furthermore, \( \partial K(t, t')/\partial t' \) gives the covariance between \( X(t) \) and \( \dot{X}(t') \). The
covariance function for higher order derivatives of \( X(t) \) can be obtained in
an analogous way. These results follow from Theorem 2.2.2 in Adler (2010).

R15 mention in their discussion section that the second derivative of \( X(t) \)
could be useful for modeling the BIS time series, and also that it could be
helpful to include an independent “GP component to model correlated in-
strumental noise.” The purpose behind the latter ingredient is to capture
deviations from the model in (2.1)-(2.4) which makes the strong assump-
tion that only a single function \( X(t) \) and its derivative are needed to model
both stellar activity and instrumental effects on all three time series. We
investigated fitting a number of stellar activity indicators (constructed as
described in Section 5) and concluded that the above two extensions offer
substantially improved flexibility regarding the indicators that can be mod-
eled. The scientific motivation that R15 give for including GP derivatives to
capture the relationships between the apparent RV, \( \log R'_{\text{HK}} \), and BIS does
not necessarily directly carry over to other stellar activity indicators. Nev-
evertheless, it seems plausible that indicators capturing similar information to
\( \log R'_{\text{HK}} \) and BIS would have qualitatively similar relationships to each other
and the apparent RV, and we have found this to be the case in practice.

Other model improvement ideas such as changing the covariance function
for \( X(t) \) are likely to be important for fine tuning to specific indicators but
were found to be less essential for expanding the range of indicators that can be modeled adequately.

We propose the following general class of models for jointly capturing the apparent RV signal due to stellar activity and \( l \) stellar activity indicators:

\[
\begin{align*}
   u(t) &= m_0 + a_{01}X(t) + a_{02}\dot{X}(t) + a_{03}\ddot{X}(t) + a_{04}Z_0(t) + \epsilon_{0i} \\
   q_1(t) &= m_1 + a_{11}X(t) + a_{12}\dot{X}(t) + a_{13}\ddot{X}(t) + a_{14}Z_1(t) + \epsilon_{1i} \\
   & \vdots \\
   q_l(t) &= m_l + a_{l1}X(t) + a_{l2}\dot{X}(t) + a_{l3}\ddot{X}(t) + a_{l4}Z_l(t) + \epsilon_{li},
\end{align*}
\]

for \( i = 1, \ldots, n \), where the \( \epsilon_{ji} \) are independent with

\[
\epsilon_{ji} \sim N(0, \sigma_{ji}).
\]

Here the \( \sigma_{ji} \)'s are known measurement uncertainties, \( q_1, \ldots, q_l \) are arbitrary stellar activity indicators, and \( Z_0, \ldots, Z_l \) are independent zero mean GPs with the covariance function (4.1). The parameters for the covariance functions of \( Z_0, \ldots, Z_l \) are assumed to be the same and are denoted \( \phi_Z \) (they are allowed to be different to \( \phi \), the covariance parameters for \( X \)). It is standard practice in astronomy to provide observation uncertainties, i.e., in practice, the \( \sigma_{ji} \)'s would be derived from the spectra based on measurement uncertainties for each pixel as provided by a standard astronomical image reduction pipeline. We denote by \( \Sigma \) the \((l+1)n \times (l+1)n \) covariance matrix implied by the model (4.2)-(4.5), and specify its form in Appendix B.

The inclusion of the second derivative of \( X(t) \) in our class of models explicitly permits a third signal shape that is not just a linear combination of \( X(t) \) and \( \dot{X}(t) \). The independent GP components \( Z_0, \ldots, Z_l \) permit deviations from the signals allowed by the model assuming each output is a linear combination of \( X(t), \dot{X}(t), \) and \( \ddot{X}(t) \). In any specific case, not all the coefficients \( a_{jk} \), for \( j = 0, \ldots, l \), and \( k = 1, \ldots, 4 \), in (4.4) will be non-zero. Indeed, such a model would likely be too flexible and absorb any RV signal from a planet that is mixed in with \( u \). The goal of the model selection procedure introduced in Section 4.2 is to decide which coefficients should be non-zero.

For conciseness we write the observation times as \( t = (t_1, \ldots, t_n)^T \) and the time series data as \( s = (u, q_1, \ldots, q_l) \), where \( u = (u(t_1), \ldots, u(t_n))^T \) and \( q_j = (q_{j1}(t_1), \ldots, q_{j}(t_n))^T \), for \( j = 1, \ldots, l \). Denoting the parameters for the model (4.2)-(4.4) by \( \theta_{\text{act}} = (m_0, \ldots, m_l, a_{01}, \ldots, a_{04}, \ldots, a_{l1}, \ldots, a_{l4}, \phi, \phi_Z) \), the log-likelihood is given by

\[
\begin{equation}
   l_{\text{act}}(\theta_{\text{act}}|t, s) = -\frac{(l+1)n}{2} \log(2\pi) - \frac{1}{2} |\Sigma| - \frac{1}{2} (s - m)^T \Sigma^{-1} (s - m),
\end{equation}
\]
where \( m = (m_0^T, m_1^T, \ldots, m_l^T) \) and \( 1_n \) denotes a column vector of \( n \) ones.

Our model capturing both a planet and stellar activity is simply (4.2)-(4.4) except that \( u(t) \) in (4.2) is replaced by

\[
(4.7) \quad u_p(t) = u(t) + v(t) - \gamma,
\]

where \( v(t) \) is given by the planet RV signal model in (3.1) and the offset \( \gamma \) is subtracted out because \( m_0 \) (in (4.2)) already provides an offset. We refer to this model incorporating a planet as the full model and denote the corresponding log-likelihood by \( l_{\text{full}} \). We write the parameters of the full model as \( \theta_{\text{full}} = (\theta_{\text{act}}, \alpha) \), where \( \alpha = (K, M_0, \tau_p, \omega, e) \) are the parameters describing the planet and its orbit.

The model (4.2)-(4.5) and the R15 model (2.1)-(2.4) are adaptations of the linear model of co-regionalization (LMC), see for example Journel and Huijbregts (1978), Osborne et al. (2008), and Alvarez and Lawrence (2011). The LMC approach captures dependence between multiple outputs by modeling each output as a linear combination of independent GPs with at least some of the GPs appearing in more than one of the linear combinations. Our model and that of R15 allow more structured dependences between outputs than LMC by the inclusion of GP derivatives. Our approach also differs from both that of LMC and R15 in that our model selection procedure (see Section 4.2 below) allows us to learn the relationships between the outputs instead of fixing them by fitting a pre-specified model.

4.2. Two stage model selection procedure. Model selection criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are often used to choose between models. In the current context these criteria can be applied to identify models in the class (4.2)-(4.5) that capture the RV corruption and stellar activity indicators at hand. However, there is a trade-off between selecting a model that captures the stellar activity well and making sure that the chosen model is not so flexible that it absorbs true planetary signals. Therefore the best model for exoplanet detection purposes cannot be determined from AIC, BIC, and similar criteria alone (even if we ignore the usual limitations of these criteria). Instead we now propose a two stage model selection procedure that focuses on our ultimate goal of planet detection.

Given a stellar activity model, we would like to quantify the evidence for a planet by computing the Bayes factor for the model with only stellar activity relative to the full model with both stellar activity and a planet (or multiple planets). In practice, computing the marginalized likelihood is
computationally expensive. For the purposes of this paper, we will instead test for a planet by performing a likelihood ratio test (LRT) where the null model is the stellar activity model and the alternative model is the corresponding full model. Thus, for our purposes, the best stellar activity model in the class (4.2)-(4.5) is that which yields the highest power for detecting planets under the LRT. Since there are many possible models in the class (4.2)-(4.5), and many choices for the planet parameter vector $\alpha$, it is not possible to determine the detection power for all possibilities through simulation. Furthermore, the typical number of spectral observations available for each star (on the order of dozens to a few hundred) is insufficient to apply asymptotic arguments to determine detection power. Instead, we propose a two stage model selection procedure where the first stage eliminates many of the models in the class (4.2)-(4.5) through an assessment of their adequacy for capturing the stellar activity indicators at hand, and the second stage chooses between the remaining models using a simulation study that evaluates the planet detection power they offer. The planet configurations we use in the stage two simulation study correspond to typical realistic systems that yield RV signals close to or below the current detection threshold of $\sim 1$ m/s. If no activity model dominates the others in terms of detection power for the simulated planet configurations, we will choose the model that produces the highest detection power for the planets of most interest or the planets that are deemed most probable to accompany the star in question based on additional scientific considerations.

The performance of a two stage procedure of the form above can in general be improved by lowering the threshold for allowing a model to pass the first stage and expanding the range of simulations used to assess power in the second stage. However, these improvements will come at a computational cost which must be taken into consideration. We postpone theoretical investigations along these lines to future work and here focus on illustrating the method in the exoplanet detection context.

In principle our procedure can be applied with any number of activity indicators, but for concreteness we fix $l = 2$. We can immediately exclude models in which $a_{j1} = a_{j2} = a_{j3} = a_{j4} = 0$ for any $j \in \{0, 1, 2\}$ because a non-trivial model is required for each time series. We also decide to fix $a_{04} = 0$ because including an independent GP for modeling the apparent RV signal will intuitively lead to a stellar activity model that can also capture RV signals from planets. Such a model would have low power because the benefit of using the full model in the presence of a planet would be small. This leaves $(2^3 - 1)(2^4 - 1)^2 = 1575$ possible models which we choose between in stage one of our procedure using AIC, BIC, and cross validation (CV).
Stone (1977) showed that AIC is asymptotically equivalent to leave-one-out CV, but here it is helpful to use both AIC and CV because asymptotic arguments do not apply, and because we found that it is better to leave out multiple observations for CV. The specifics of our CV procedure are given in Appendix C. For the sake of illustration we specify that the best five models under each criterion will be evaluated at stage two of our procedure.

Given the models selected in stage one of our procedure, we move to stage two where the planet detection power of the models is evaluated. The test statistic $\Delta$ for our LRT is simply the difference in maximized log likelihood for the current stellar activity model and the corresponding full model, which includes the planet model (3.1). That is,

$$
\Delta = l_{\text{act}}(\hat{\theta}_{\text{act}}) - l_{\text{full}}(\hat{\theta}_{\text{full}})
$$

where $\hat{\theta}_{\text{act}}$ and $\hat{\theta}_{\text{full}}$ are maximum likelihood estimates for $\theta_{\text{act}}$ and $\theta_{\text{full}}$, respectively. Optimization details are given in Appendix D. To construct the null distribution of $\Delta$ we use SOAP 2.0 to simulate 1000 datasets of the form $\{u(t_i), q_1(t_i), q_2(t_i)\}_{i=1,...,n}$ and compute $\Delta$ for each dataset.

Next, we simulate datasets with injected planet signals. In practice, the specific simulations that should be performed will depend on the instrument in question, our understanding of what signals can realistically be detected, and the type of planets we are hunting. Often it will be of particular interest to determine the minimum amplitude of a planet RV signal needed to reach a detection power of 0.5, which we call the detection threshold. In this paper, we will choose between the candidate stellar activity at stage two of our model selection procedure by computing the detection threshold for planets with an orbital period of 7 days. This period is favorable in the sense that it does not coincide with the stellar rotational period in our SOAP 2.0 simulations (10 days) or its harmonics. To generate datasets with planets, we first generate (new) datasets of the form $\{u(t_i), q_1(t_i), q_2(t_i)\}_{i=1,...,n}$ and then add the planet signal $v(t_i; \alpha)$ (given by (3.1)) to $u(t_i)$, for $i = 1, \ldots, n$. The parameter vectors of the injected signals are given by $\alpha = (K, M_0 = 1.5, \tau_p = 7, \omega = 1, e = 0.2, \gamma = 0)$ for amplitudes $K = 0.1, 0.2, 0.25, \ldots, 1.45, 1.5, 2$ m/s, and 50 simulations are generated under each setting. In practice, the parameters $M_0, \omega, e,$ and $\gamma$ have less impact on the planet signal than $K$ and $\tau_p$ and the fixed values used here are reasonably typical. (Note that all the planet parameters are fitted in our procedure.) Finally, for each dataset, we compute $\Delta$ for the current stellar activity model being considered and reject the null of no planet if $\Delta$ is greater than the 0.99 quantile of the relevant null distribution, i.e., that constructed using the current stellar activity model. The number of rejections across the 50 simulations at a given value of $K$
provides an estimate of the detection power at that planet signal amplitude. The approximate detection threshold is then determined by finding the $K$ value for which the detection power is approximately 0.5. The final model we select is the one with the smallest detection threshold.

5. Constructing new stellar activity indicators.

5.1. Motivation for Doppler-constrained dimension reduction approaches. Traditional activity indicators such as log $R'_{\text{HK}}$ and BIS are unlikely to be the most suitable stellar activity indicators for predicting the RV perturbation induced by stellar activity. It is desirable to have a data-driven method for identifying stellar activity indicators that are informative for our goals of detecting and characterizing exoplanets. Recently, Davis et al. (2017) proposed a principled way of constructing stellar activity indicators and determining the proportion of the available information captured. They use principal component analysis (PCA) to decompose stellar spectra into orthogonal vectors along which variation in the spectra is observed across time. In particular, their principal component (PC) vectors are the eigenvectors of the matrix $Y^T Y$, where $Y$ is the same as the $n \times p$ SOAP 2.0 data matrix introduced in Section 3.2, except that they use $n = 25$. (They also use a slightly different process for reducing the SOAP 2.0 high resolution spectra to a realistic resolution.) For SOAP 2.0 data simulated with a single spot and no planet, one to four principal components are typically sufficient to explain nearly all of the spectral variation observed across different stellar rotation phases, even at spectral resolution and SNR significantly greater than current and next-generation spectrographs. For current planet surveys, only one or two principal component scores can be measured with significant SNR. PCA has also been used for tackling related challenges in astronomy, for example González et al. (2008) use PCA to denoise spectra.

It is not immediately apparent how the PCA based stellar activity indicators proposed by Davis et al. (2017) can be used in practice because each PC score may be composed of components from both stellar activity and the true Doppler shift. Ideally, some PC scores would be affected only by stellar activity, allowing them to be used to infer the impact of the activity on the apparent RV. Similarly, we do not know how much a true Doppler shift would project onto each of the PC vectors developed to reconstruct spectra of a star with spots. Indeed, with traditional PCA we may need more than one principal component, or even all the principal components, in order to capture an RV signal. To address this difficulty we now propose an adaptation of PCA in order to represent the apparent RV signal due to stellar activity as a single component.
5.2. Doppler-constrained PCA method. The idea is to specify a $p \times 1$ component $\mathbf{w}$ that corresponds to the apparent RV, then to compute

$$
\tilde{Y} = Y - \frac{Y \mathbf{w} \mathbf{w}^T}{\sum_i |w_i|^2},
$$

and finally to apply PCA to $\tilde{Y}$. To proceed with this approach we first identify $\mathbf{w}$. Let the star light intensity at wavelength $\lambda$ be denoted by $f(\lambda)$, so that the function $f(\cdot)$ gives the stellar spectrum (at some fixed time $t$). For a source moving with a radial velocity of $v = cz$ the Doppler shift is $z$, where $c$ is the speed of light, i.e., the observed intensity at wavelength $\lambda$ is given by $f((1+z)\lambda)$. Thus, a Taylor expansion tells us that the new spectrum at wavelength $\lambda$ will approximately be given by $f(\lambda) + zf'(\lambda)$, where $f' = df/d\log \lambda$. Since the relevant Doppler shifts are typically very small ($z \approx 10^{-8}$) this approximation is very accurate and we can therefore represent Doppler shifts as scalar multiples of the vector $\mathbf{w} = (f'(\lambda_1), \ldots, f'(\lambda_p))^T$, where $\lambda_i$, for $i = 1, \ldots, p$, are the recorded wavelengths. In the case of actual astronomical observations (as opposed to SOAP 2.0 simulations), it is necessary to make various corrections to the Doppler shift before using the Taylor expansion, see Appendix E.

Another complication is that the stellar spectrum changes due to the presence and location of stellar activity. Therefore, the function $f'(\cdot)$ varies with stellar rotation phase. Fortunately, these differences can be regarded as second-order, so it is reasonable to treat $\mathbf{w}$ as fixed across time. Here, we compute the mean spectrum across the different phases observed, denoted $\bar{f}(\cdot)$, and use $\bar{f}'(\cdot)$ in place of $f'(\cdot)$ for all phases.

Once $\mathbf{w}$ is computed, the vector of apparent radial velocity observations $(u(t_1), \ldots, u(t_n))$ is obtained by $cY \mathbf{w}/\sum_i |w_i|^2$, where the scaling $c$ is the speed of light in m/s. Then we apply PCA to $\bar{Y}$ and use $q_{PC_j}(t_i)$, for $j = 1, 2, \ldots, \min\{n, p\} - 1$, to denote the values of our resulting stellar activity indicators at time $t_i$, for $i = 1, \ldots, n$.

The main benefit of our approach is that the apparent Doppler component $\mathbf{w}$ and the stellar activity indicators are orthogonal. Thus, any planet RV signal will only appear in the $\mathbf{w}$ component scores, and we can use the orthogonal indicators to attempt to recover and remove any apparent RV signal due to the spot. In particular, we can use the class of models introduced in Section 4.1. Intuitively, we expect our indicators to be highly informative because PCA guarantees that the proportion of the variance explained by PC1 is maximized (after the removal of the $\mathbf{w}$ component from the data, see 5.1), then of the remaining variance the proportion explained by PC2 is maximized, and so on. Of course, high variance explained does not neces-
sarily mean that the PC scores are related to the apparent RV corruption due to a spot, as explained in Davis et al. (2017). Furthermore, higher order PCs may be important for distinguishing different types of stellar activity. In practice, for the case of a spot, we find that the RV corruption is related to at least the first three PC score time series for spectra with sufficiently high resolution and high SNR. However, since the PC1 scores are least sensitive to noise they are usually the most informative for recovering the RV corruption.

5.3. Generating realistic Doppler-constrained PCA activity indicator observations using SOAP 2.0. We now describe the final stellar activity simulations to be used in our model selection procedure of Section 4.2. Figure 3 summarizes our simulation of realistic activity signals of the form \( \{u(t_i), q_{PC1}(t_i), q_{PC2}(t_i)\}_{i=1,...,n} \) from the SOAP 2.0 output \( Y \), for the case of a star with a single spot. First, we apply our Doppler-constrained PCA
approach to the noiseless spectra contained in $Y$ to obtain ideal basis vectors. The top row of Figure 3 shows the corresponding PC scores (blue dots) for 100 randomly selected phases, i.e., the ideal $u, q_{PC1}$, and $q_{PC2}$ signals with a few phase gaps. Since we want our final simulations to include noise we also plot two standard deviation error bars. These are computed pointwise by applying our Doppler-constrained PCA approach to 200 noisy realizations of the spectra (at the 100 selected phases). The noise was generated for each wavelength using a Poisson distribution, which realistically reproduces random variations in the number of photons detected. Note, that although these latter spectra are noisy, the scores are still computed by projecting onto the previously obtained ideal basis vectors. SOAP 2.0 is based on real observations of the Sun so the use of the ideal basis vectors makes sense for Sun-like stars, but an alternative approach will likely be needed for stars that differ significantly from the Sun. We address this issue further in Section 7.

The bottom row of Figure 3 shows an example final dataset used in our analyses of Section 6. We have injected noise based on the two standard deviation error bars. The phases of the observations are the same as in the top row, but we have now randomly selected a rotation of the star from 1 to 50 for each observation to be recorded. Since the rotation period of the star is 10 days, the total observation time is 500 days, which is representative of an intensive observing program spanning two calender years.

6. Application of model selection procedure.

6.1. Preliminary model. The R15 model (2.1)-(2.4) was not designed to capture our PCA based stellar activity indicators both because our indicators differ to $\log R'_{\text{HK}}$ and BIS and because the SOAP 2.0 settings R15 used were different. However, the plot of $\{u(t_i), \log R'_{\text{HK}}(t_i), \text{BIS}(t_i)\}_{i=1,...,m}$ in Figure 3 of R15 qualitatively resembles the first three panels of Figure 3 here which show $\{u(t_i), q_{PC1}(t_i), q_{PC2}(t_i)\}_{i=1,...,n}$. We therefore plot the maximum likelihood fit of the R15 model (very slightly modified) to the data $\{u(t_i), q_{PC1}(t_i), q_{PC2}(t_i)\}_{i=1,...,n}$ in the left half of Figure 4. The only difference with the model given by (2.1)-(2.4) and that fitted is that the standard deviations $\sigma^2_{ij}$ are replaced by the known observation specific standard deviations $\sigma^2_{ji}$, for $j = 0, 1, 2$, and $i = 1, \ldots, n$. We will refer to this model as the preliminary model since it is our best initial guess of a suitable model for our indicators based on the existing literature.

The red lines in the top row of Figure 4 show inferred values of the functions $h_1(t) = m_0 + a_{01}X(t) + a_{02}\dot{X}(t)$, $h_2(t) = m_1 + a_{11}X(t)$, and $h_3(t) = m_2 + a_{21}X(t) + a_{22}\dot{X}(t)$, i.e., $\hat{m}_j + (\text{Cov}(\hat{h}_j(t), s_1), \ldots, \text{Cov}(\hat{h}_j(t), s_{3n}))^T \Sigma^{-1} (s - \hat{m})$, for $j = 0, 1, 2$, and $t \in [0, 40]$. The hat notation indicates that the pa-
Fig 4: Preliminary model fit to \( \{u(t_i), q_{PC1}(t_i), q_{PC2}(t_i)\}_{i=1,...,n} \) (top row) and AIC-1 model fit (middle row). Note that during fitting all the signals were normalized for numerical stability but \( u(t) \) is plotted on the original m/s scale for interpretability. The bottom row shows the residuals for the preliminary model fit.

The parameter vector \( \theta \) has been replaced by its maximum likelihood estimate \( \hat{\theta} \). A bright green region representing a 95% confidence interval for the functions \( h_j(t) \), for \( j = 0, 1, 2 \), is plotted but is mostly covered by the red line due to very small uncertainties. The reason for the small uncertainties is that \( X(t) \) (and hence \( \hat{X}(t) \)) can be precisely inferred by combining information across the three time series.

Despite the high precision, the preliminary model fit is unsatisfactory because in the case of \( q_{PC2} \) the inferred function \( h_1 \) shows systematic deviations from the observations. This can be seen from the bottom row of Figure 4, which shows the residuals for the preliminary model fit. In particular, the middle panel of the bottom row gives the \( q_{PC2} \) residuals and we can see that for phases near 0 and 1 the residuals tend to be positive, and for phases near 0.3 and 0.7 they tend to be negative. Finally, the fit overshoots the
signal peak and so again many of the residuals are positive near phase 0.5. In further investigations we found that introducing large gaps between the observations produced implausible predictions for the intervening times due to the highly constrained negative correlations imposed by the use of both $X(t)$ and its derivative. R15 identified similar difficulties and attempted to alleviate them by imposing upper bounds on their noise parameters (the $\sigma_k$’s in (2.4)) and thereby weighting the influence of the RV time series more heavily when optimizing the GP parameters. Their logic was that only the RV time series ultimately needs to be captured. However, the success of this approach in the absence of a planet only suggests that by fitting the stellar activity induced RV corruption with a GP and its derivative, some stellar activity indicators are also approximately recovered. In practice, we are of course interested in the reverse problem: by fitting a model to stellar activity indicators, we want to recover the RV corruption. In the case of a star with a planet, by inflating the influence of the RV time series in fitting the stellar activity model, one risks absorbing some of the planet signal that could have been extracted by more careful modeling of the stellar activity indicators and their relationship to RV corruption. Therefore, we do not alter the $\sigma_k$’s to affect the balance of residuals among the RV and stellar activity indicator time series.

6.2. Initial stellar activity model selection. To improve upon the preliminary model we apply the two stage model selection procedure introduced in Section 4.2 to identify the best model among the class (4.2)-(4.5). Table 1 gives a partial summary of the results of stage one of our selection procedure, which identifies the top models for capturing stellar activity based on AIC, BIC, and CV. The second and third columns list the number of parameters in the model and the deviance, and columns four through six list the ranking of the models by the three criteria. Columns seven through nine give the relative criteria values for each model, e.g., $\Delta\text{AIC}$ is the AIC value for the model in question after the minimum AIC value found across all the models has been subtracted. Smaller values are preferable for all quantities listed in the table. The top five ranked models are the same under the AIC and BIC criteria so we have only listed them once as AIC models.

The top five ranked models under AIC (and BIC) are also found to be good models under the CV criterion, with the exception of AIC-1. Similarly the top five CV models have good AIC and BIC ranks, though the performance of CV-4 and CV-5 is somewhat worse than the performance of the top three CV models in terms of $\Delta\text{AIC}$ and $\Delta\text{BIC}$. The preliminary model performs poorly under all three of the criteria in Table 1, as might
Table 1
Summary of results for stage one of our model selection procedure presented in Section 4.2. The preliminary model and the top five models under each of the AIC and CV criteria are listed. The top five models under BIC are the same as those under AIC.

| Model  | no. paras | dev. | AIC rank | BIC rank | CV rank | ∆AIC | ∆BIC | ∆CV |
|--------|-----------|------|----------|----------|---------|------|------|------|
| Preliminary | 11 58 | 595 429 841 | 55 55 3.28 |
| AIC-1 | 11 2 | 1 1 774 | 0 0 2.47 |
| AIC-2 | 12 1 | 2 2 1 | 1 3 0.00 |
| AIC-3 | 12 1 | 3 3 6 | 1 4 0.09 |
| AIC-4 | 12 2 | 4 4 15 | 2 5 0.16 |
| AIC-5 | 12 2 | 5 5 7 | 2 5 0.09 |
| CV-1 | 12 1 | 2 2 1 | 1 3 0.00 |
| CV-2 | 13 1 | 9 9 2 | 3 8 0.05 |
| CV-3 | 13 1 | 7 7 3 | 2 8 0.09 |
| CV-4 | 15 2 | 15 16 4 | 8 18 0.09 |
| CV-5 | 16 3 | 31 25 5 | 11 24 0.09 |

Table 2
Maximum likelihood estimates of the AIC-1 stellar activity model coefficients. Blank entries mean the coefficients are set to zero. All the outputs were normalized, but for interpretability the \( u(t) \) coefficient estimates in m/s are given in parentheses.

| \( u(t) \) (m/s) | \( X \) coeff \( (a_{j1}) \) | \( \dot{X} \) coeff \( (a_{j2}) \) | \( \ddot{X} \) coeff \( (a_{j3}) \) | \( Z_j \) coeff \( (a_{j4}) \) |
|------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| \( q_{PC1}(t) \) | -0.03 (-0.22) | -0.29 (-2.20) | 0.45 | 0.07 |
| \( q_{PC2}(t) \) | 0.24 | | | |

be expected from the discussion in Section 6.1 and the fit shown in the top of Figure 4. For comparison, the fit of the AIC-1 model to the same data is shown in the middle row of Figure 4. Close inspection reveals that the AIC-1 fit for \( q_{PC1} \) does not have the systematic deviations from the data that we observed for the preliminary model fit. Table 2 gives the maximum likelihood estimates (MLEs) of the model coefficients for AIC-1. The MLEs of the kernel parameters are \( \log \hat{\tau}_s = 2.30 \), \( \log \hat{\lambda}_p = -0.92 \), and \( \log \hat{\lambda}_e = 9.98 \). The interpretation of these values is that the stellar period \( \tau_s \) is approximately 10 days and that given the relatively small value of \( \lambda_p \) and large value of \( \lambda_e \), periodic correlations are strong and local correlations are weak. Thus, the kernel parameter values are indeed consistent with a star with a single spot producing stellar activity signals that are periodic over a 10 day period (the star rotation period). The AIC-1 model turns out to be contained in all nine of the models listed below it in Table 1, i.e., the coefficients \( a_{01}, a_{02}, a_{11}, a_{13}, \) and \( a_{22} \), are again non-zero in these nine models. This can be seen from the right panel of Figure 5, which shows the relative frequencies of the model coefficients across the five AIC and five CV models in Table 1. Further checks revealed that in fact the top 49 AIC ranked models and the top 44 CV ranked models contain the AIC-1 model.
Given the above discussion, it might be surprising that the AIC-1 model has a poor CV score. The reason for the poor score is that the model only has 11 parameters and is too restrictive in the sense that it tends to under-represent prediction uncertainty, particularly for $q_{PC1}$. Since our CV score is the conditional log-likelihood of the test data, the AIC-1 model is strongly penalized for this unreliable prediction uncertainty quantification. Nonetheless, in Section 6.3 we find that the strong constraints imposed by the AIC-1 model are helpful (or at least not harmful) for the purpose of planet detection because in most cases the model cannot mistakenly fit a planet signal. For more complex forms of stellar activity that behave less periodically such a constrained model may be less useful for planet detection.

The models in Table 1 do not make use of the independent GP components $Z_1$ and $Z_2$, with the exception of CV-4 and CV-5 which are found to perform relatively poorly in Section 6.3 below. However, in Section 6.4 (below), we find that the flexibility offered by $Z_1$ and $Z_2$ does help to improve planet detection power in the case of modeling the stellar activity indicators used by R15. This is because the R15 indicators do not closely follow the structure imposed by using linear combinations of $X(t)$ and its derivatives. In our ongoing investigations we have also found the extra GP components useful in the case where we use diffusion maps instead of PCA to construct Doppler-constrained activity indicators. Further discussion of diffusion map based indicators can be found in Section 7.

6.3. Optimal models for planet detection. Next, we proceed to stage two of our model selection procedure, as described in Section 4.2. In order to choose between the stellar activity models in Table 1, we assesses the planet detection power under the different candidate models. As explained in Section 4.2, the detection powers are computed for a planet with an orbital period of $\tau_p = 7$ days, which avoids aliasing with the stellar rotation period of 10 days. That is, $\tau_p$ is distinct from the stellar rotation period and its harmonics and we therefore do not need to deal with the problem of separating two periodic signals that could be realistically modeled by one periodic signal.

Figure 5 shows the power for detecting planets under each model at a range of planet signal amplitudes and $\tau_p = 7$ days. In the plot, we see that the top five AIC models (circles of different shades of blue) have a detection threshold of approximately 0.2 m/s (i.e., this is the amplitude at which 0.5 detection power is achieved), and have detection power close to 1 for planet signal amplitudes equal to or greater than 0.4 m/s. Turning to the top five CV models (triangles of different shades of red), the performance of CV-1
and CV-2 is very similar to that of the top five AIC models, but CV-3, CV-4, and CV-5 all yield lower detection power for small amplitude planet signals. The problem with the CV-3, CV-4, and CV-5 stellar activity models is that they are too flexible and therefore sometimes absorb part of the planet signal making it harder to detect. In the case of CV-4 and CV-5 the source of the extra flexibility is mainly due to the inclusion of the independent GP $Z_1$, and in the case of CV-3 it is due to the coefficients $a_{12}$ and $a_{21}$ being non-zero (the latter is also non-zero in the preliminary model).

The preliminary model (dark green diamonds) is seen to have substantially worse performance than all ten of the AIC and CV models presented. Indeed, the detection threshold for the preliminary model is about 0.55 m/s. This demonstrates the advantage of selecting a model appropriate for the stellar activity indicators at hand, and that it is possible to do this selection automatically even in the case of indicators such as ours which have no immediate physical interpretation. The white noise model (grey squares) considers only the RV signal and treats any RV corruption from stellar activity as independent Gaussian realizations with a fixed standard deviation (plus measurement error). This approach can be valuable for analyzing legacy RV datasets that typically have sparsely spaced observations and no stellar activity information (see Ford, 2006). However, each of the models that makes use of stellar activity indicators is significantly more powerful for detecting low-mass planets, including the preliminary model.

The relatively poor performance of CV-3, CV-4, and CV-5 shown in Figure 5 demonstrates the importance of including the second stage of our
model selection procedure. Indeed, these models are ranked among the best for capturing the stellar activity signals according to CV, and also have reasonably good AIC and BIC ranks (see Table 1). It is only stage two of our model selection procedure that allows us to identify the top five AIC models, CV-1, and CV-2 as better models for planet detection (at least under the scenario captured by our simulations). The fact that AIC-1, BIC-1, and CV-1 are all good models for planet detection is something of a coincidence, and in general we do not recommend relying on the AIC, BIC, and CV selection criteria alone; we only suggest using them to short-list candidate models.

In addition to detecting planets it is of interest to infer their properties, and in Appendix F we summarize our orbital period ($\tau_p$) and signal amplitude ($K$) estimation results for the AIC-1 model. Another important consideration is how planet detection power varies as a function of period, and we discuss this in Appendix G, again for the AIC-1 model. Our main finding is that detection power is lower for planets with orbital periods that coincide with the stellar rotation period or its harmonics.

6.4. Comparison to Rajpaul et al. (2015a). The poor performance of the preliminary model seen in Figure 5 is not particularly surprising because the model was originally designed by R15 to capture the evolution of the apparent RV, normalized flux (or log $R'_{HK}$), and BIS, rather than the evolution of our stellar activity indicators. Therefore, we repeat our detection power analysis for the R15 model, but using the R15 activity indicators in order to compare their utility for planet detection to that of our PCA based activity indicators. Both normalized flux and BIS are provided in SOAP 2.0 output. To mimic exoplanet RV observations, we must add noise at scales representative of RV data. For normalized flux, we compute the noise levels directly using the Poisson uncertainties for the wavelength level flux measurements. For BIS, we assume a constant standard deviation of 0.038 m/s, where the specific value is motivated by scaling actual BIS error measurements in Dumasque et al. (2012) to account for the increased spectral resolution and SNR of our input data.

The light green line with diamond points in the left panel of Figure 6 shows the detection power under the R15 model and using the normalized flux and BIS stellar activity indicators described above (for a planet with a 7 day orbital period). We note that again there is a slight difference between the original R15 model and the model we fit because we have used the observation specific standard deviations, rather than estimating a single standard deviation for each times series (under an informative prior), see equation (2.4). For comparison, the detection power under the AIC-1
Fig 6: The left panel shows the detection power under the R15 model (light green diamonds) using their normalized flux and BIS indicators. For comparison some results from Figure 5 are plotted again: the detection power using our PCA based indicators under the AIC-1 model (blue circles) and preliminary (or R15) model (dark green diamonds). The R-AIC-3 (pink “+” symbols) and the R-CV-3 (dark red “x” symbols) models are the best models short-listed using AIC and CV, respectively, for the R15 indicators. The right panel compares the \( q_{PC2} \) and BIS indicators computed from noiseless spectra outputted from SOAP 2.0. The indicators are standardized (centered and normalized) for the purpose of the comparison.

(blue circles) and preliminary (dark green diamonds) models applied with our PCA based indicators are also shown (these lines are the same as in Figure 5). The R15 model has a detection threshold of about 0.45 m/s and therefore performs better than the preliminary model, but still substantially worse than the AIC-1 model.

It is not clear if the superiority of the AIC-1 model results are due to the use of our PCA based indicators or our model selection procedure or both. To investigate this we also applied our model selection procedure in the case of the normalized flux and BIS indicators. We again found the five best AIC, BIC, and CV ranked models, though we only plot a selection in Figure 6. We write, for example, “R-AIC-3” to mean the model ranked third by AIC in the case of the normalized flux and BIS indicators used by R15. The R-CV-3 model (dark red “x” symbols) is the top performing model, though several other models have very similar performance. The model performs substantially better than the R15 model and indeed comparably to the AIC-1 model for our stellar activity indicators (blue circles). The R-CV-3 model and the maximum likelihood estimates of its coefficients are given in Table 3. Interestingly, the model uses the extra GP components \( Z_1 \) and \( Z_2 \).
Table 3

Maximum likelihood estimates of the R-CV-3 stellar activity model coefficients. Blank entries mean the coefficients are set to zero. All the outputs were normalized, but for interpretability the \( u(t) \) coefficient estimates in m/s are given in parentheses.

| \( u(t) \) (m/s) | \( X \) coeff \( (a_{j1}) \) | \( \dot{X} \) coeff \( (a_{j2}) \) | \( \ddot{X} \) coeff \( (a_{j3}) \) | \( Z_j \) coeff \( (a_{j4}) \) |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( q_{PC1}(t) \) | 0.00 (0.02)     | 0.10 (0.76)     | 0.58 (4.37)     | 0.17            |
| \( q_{PC2}(t) \) | 0.03            | 0.56            | 0.55            | 0.11            |

Investigation showed that \( Z_1 \) and \( Z_2 \) are needed in order to capture the R15 indicators, which do not have the same level of symmetry as our PCA based indicators (in particular, the BIS minima are not as low as the maxima are high, see the right panel of Figure 6). The left of Figure 6 also shows the detection power for the R-AIC-3 model (pink “+” symbols), which is the best performing R-AIC model identified. This model performs poorly compared with the R-CV-3 model, but still considerably better than the R15 model. The R-AIC-3 model again makes use of the extra GP component \( Z_2 \). The models that only use \( X(t) \) and its first two derivatives fit poorly to stellar activity data, including the R15 model (see their Figure 3).

Our conclusion is that for this simple case of a single constant spot it is our model selection procedure rather than our PCA based stellar activity indicators that offers improved detection power. Therefore, for the time being, astronomers may prefer to keep using normalized flux (or \( \log R'_{HK} \)) and BIS as stellar activity indicators, but the model should be updated to the R-CV-3 model or a similar model. In particular, although the R15 model has the appealing feature of being physically motivated, it seems that planet detection power can be improved by using a more flexible model that can better capture the stellar activity time series jointly, at least in case of a constant spot. For cases where the stellar spectra are available to the investigators, it may be preferable to use our PCA based indicators for computational and robustness reasons, because the AIC-1 model (for our indicators) requires six fewer parameters than the R-CV-3 model (for the R15 indicators). Alternatively, investigators may wish to transform the BIS indicator so that the blueshifts (troughs) and redshifts (peaks) are more symmetric because this will likely make it possible to use a simpler GP model. For example, such a transformation could be based on transforming the BIS indicator to look more like our \( q_{PC2} \) indicator. The right panel of Figure 6 compares \( -q_{PC2} \) and BIS for noiseless SOAP 2.0 data (for a star with a single constant spot), and suggests that scaling blueshifts and redshifts differently (after centering) could be a good choice of transformation.

With a longer term perspective, we emphasize that our indicators were
constructed automatically without the need for physical derivations. This is important because it means that our approach can likely be easily generalized to more complex stellar activity phenomena, such as evolving spots, where different or additional indicators may be needed. The PCA approach ensures that when additional indicators are used they contain different information to those already included, whereas there is no such guarantee for expert-identified indicators chosen for their individual interpretations. Similarly, our indicator construction method could also allow custom indicators to be used for each star or type of star.

7. Discussion. We have considered the problem of using high-resolution Doppler spectroscopy to detect low-mass planets with RV amplitudes significantly less than the apparent RV induced by intrinsic stellar variability. We propose a constrained dimension reduction procedure to construct high-information stellar activity indicators. We then model the temporal evolution of the RV and stellar activity indicators with a multivariate GP model. Specifically, we consider a family of GP models and use a model selection procedure to identify a model that can jointly capture the activity indicators and RV corruption due to a stellar spot. By incorporating planet detection power into our model selection procedure, we ensure that the models selected are effective for detecting low-mass planets. We have demonstrated that our model selection procedure leads to a substantial improvement in planet detection power compared to that obtained using the state-of-the-art approach of R15. Furthermore, we have found that our automatically derived stellar activity indicators are at least comparable to the physically motivated indicators used by R15, and suspect they may offer detection power improvements for more complex stellar activity.

Our general procedures can be applied more broadly than to the specific stellar activity indicators used in this initial study. Firstly, our approach for constructing Doppler-constrained stellar activity indicators can make use of any dimension reduction technique. Our Doppler-constrained PCA method is simply one example. We are currently exploring the use of diffusion maps as an alternative to PCA because diffusion maps do not have the constraint of projecting onto a linear subspace as does PCA. Thus, diffusion maps might capture non-linear information that PCA cannot preserve. This non-linear information could particularly be helpful for complex situations such as when the stellar spot evolves over time. Our initial investigations also suggest that capturing non-linear structure helps to improve planet detection power when there is aliasing between the stellar rotation period and the planet orbital period, see the discussion of Table 4 in Appendix G.
Secondly, by using a model selection procedure we avoid having to rely on a single model and therefore can model a variety of stellar variability indicators. Since our procedure assesses the planet detection power this can be used to compare the relative performance of various stellar variability indicators, as well as different models. Lastly, although we have assumed a single spot of constant size, our investigations suggest that our class of models is able to capture evolving spots. More generally, the extent to which it is reasonable to use the specific class of models (4.2)-(4.5) for modeling complex stellar activity is a topic of our ongoing work. If a more general class of models is ultimately found to be needed in the case of multiple evolving stellar variability phenomena, one option is to use the Gaussian process regression networks framework introduced by Wilson, Ghahramani and Knowles (2012). In our context this would involve using GPs to model temporal changes in the coefficients of our model (4.2)-(4.5), which would permit the relationships between outputs to evolve over time.

Despite the promise of our approach, there are a number of limitations in the astrophysical scope of our work that we intend to address in future research. Firstly, the SOAP 2.0 spectra that we use to construct our PCA basis vectors do not include noise, whereas real spectra do. We have incorporated noise for each individual simulated dataset, but the basis vectors that we project onto remain the same across datasets. This makes sense for stars like the Sun because the SOAP 2.0 output used to create the basis vectors is based on real observations of the Sun, and these are of far higher precision than we could hope for in the case of any other star. However, for stars unlike the Sun, the basis vectors will instead need to be inferred from the data at hand rather than using SOAP 2.0. Consequently, there will be additional noise in the time series observations which might reduce the usefulness of higher order PCs and ultimately our ability to detect planets. Preliminary investigation suggests that this will not be a fundamental limitation.

A second area for expanding the scope of our work is to generalize our two stage model selection procedure to stars that differ significantly from the Sun. Our power calculations rely on simulated datasets from a star with the same spot configuration as the observed dataset, but in practice the precise properties of the star and spot are unknown. In future work we hope to address both of the above issues by marginalizing over a wide range of astrophysically plausible spot geometries. By incorporating our existing model into a hierarchical Bayesian model, one could jointly model stellar activity from multiple stars with similar spectroscopic properties. This would allow us to infer a population level stellar activity model that can be used for constructing basis vectors and generating replicate datasets.
Acknowledgements. The work reported here began as part of the Program on Statistical, Mathematical and Computational Methods for Astronomy hosted by the Statistical and Applied Mathematical Sciences Institute (SAMSI).

References.

Adler, R. J. (2010). *The geometry of random fields*. Classics in Applied Mathematics. SIAM.

Aigrain, S., Pont, F. and Zucker, S. (2012). A simple method to estimate radial velocity variations due to stellar activity using photometry. *Monthly Notices of the Royal Astronomical Society* 419 3147–3158.

Alvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research* 12 1459–1500.

Arentoft, T., Kjeldsen, H., Bedding, T. R., Bazot, M., Christensen-Dalsgaard, J., Dall, T. H., Karoff, C., Carrier, F., Eggenberger, P., Sosnowska, D. et al. (2008). A multisite campaign to measure solar-like oscillations in procyon. I. Observations, data reduction, and slow variations. *The Astrophysical Journal* 687 1180.

Baranne, A., Queloz, D., Mayor, M., Adriaensky, G., Knispel, G., Kohler, D., Lacroix, D., Meunier, J.-P., Rimbaud, G. and Vin, A. (1996). ELODIE: A spectrograph for accurate radial velocity measurements. *Astronomy & Astrophysics Supplement Series* 119 373–390.

Boisse, I., Bonfils, X. and Santos, N. (2012). SOAP: A tool for the fast computation of photometry and radial velocity induced by stellar spots. *Astronomy & Astrophysics* 545 A109.

Borgniet, S., Meunier, N. and Lagrange, A.-M. (2015). Using the Sun to estimate Earth-like planets detection capabilities-V. Parameterizing the impact of solar activity components on radial velocities. *Astronomy & Astrophysics* 581 A133.

Butler, R. P., Marcy, G. W., Williams, E., McCarthy, C., Dosanjh, P. and Vogt, S. S. (1996). Attaining Doppler precision of 3 M s-1. *Publications of the Astronomical Society of the Pacific* 108 500.

Butler, R. P., Vogt, S. S., Laughlin, G., Burt, J. A., Rivera, E. J., Tuomi, M., Teske, J., Arriagada, P., Diaz, M., Holden, B. and Keiser, S. (2017). The LCES HIRES/Keck Precision Radial Velocity Exoplanet Survey. *The Astronomical Journal* 153 208.

Cavallini, F., Ceppatelli, G. and Righini, A. (1985). Asymmetry and shift of three Fe I photospheric lines in solar active regions. *Astronomy & Astrophysics* 143 116–121.

Danby, J. (1988). Fundamentals of celestial mechanics. Willmann-Bell, 1988. 2nd ed., rev. & enl.

Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. and Ford, E. B. (2017). Insights on the spectral signatures of RV jitter from PCA. In American Astronomical Society Meeting Abstracts 229.

Del Moro, D. (2004). Solar granulation properties derived from three different time series. *Astronomy & Astrophysics* 428 1007–1015.

Del Moro, D., Berrilli, F., Duval, T. and Kosovichev, A. (2004). Dynamics and structure of supergranulation. *Solar Physics* 221 23–32.

Dravins, D., Lindegreen, L. and Nordlund, Å. (1981). Solar granulation-Influence of convection on spectral line asymmetries and wavelength shifts. *Astronomy & Astrophysics* 96 345–364.
DUMUSQUE, X., BOISSE, I. and SANTOS, N. (2014). SOAP 2.0: A tool to estimate the photometric and radial velocity variations induced by stellar spots and plages. The Astrophysical Journal 796 132.

DUMUSQUE, X., PEPE, F., LOVIS, C., SÉGRANSAN, D., SAHLMANN, J., BENZ, W., BOUCHY, F., MAYOR, M., QUELOZ, D., SANTOS, N. and UDRI, S. (2012). An Earth-mass planet orbiting α Centauri B. Nature 491 207–211.

DUMUSQUE, X., BORSA, F., DAMASSO, M., DIAZ, R. F., GREGORY, P., HARA, N., HATZES, A., RAJPAL, V., TUOMI, M., AIGRAIN, S. et al. (2017). Radial-velocity fitting challenge-II. First results of the analysis of the data set. Astronomy & Astrophysics 598 A133.

FISCHER, D. A., MARCY, G. W. and SPRONCK, J. F. (2013). The twenty-five year Lick planet search. The Astrophysical Journal Supplement Series 210 5.

FORD, E. B. (2006). Improving the Efficiency of Markov Chain Monte Carlo for Analyzing the Orbits of Extrasolar Planets. The Astrophysical Journal 642 505-522.

GONZÁLEZ, M. M., RAMOS, A. A., CARROLL, T., KOPF, M., VÉLEZ, J. R. and SEMEL, M. (2008). PCA detection and denoising of Zeeman signatures in polarised stellar spectra. Astronomy & Astrophysics 486 637–646.

HAYWOOD, R. D., COLLIER CAMERON, A., UNRUH, Y., LOVIS, C., LANZA, A., LLAMA, J., DELELUI, M., FARES, R., GILLON, M., MOUTOU, C., PEPE, F., POLLACCO, D., QUELOZ, D. and SÉGRANSAN, D. (2016). The Sun as a planet-host star; proxies from SDO images for HARPS radial-velocity variations. Monthly Notices of the Royal Astronomical Society 457 3637–3651.

HEWER, R., FRIEDERICHS, P., HENSE, A. and SCHLATHER, M. (2017). A Matérn based multivariate Gaussian random process for a consistent model of the horizontal wind components and related variables. Journal of the Atmospheric Sciences 2017.

JENKINS, J. L. (2013). Observing the Sun: A Pocket Field Guide. Springer Science & Business Media.

JOURNEL, A. G. and HULIBRECTS, C. J. (1978). Mining geostatistics. Academic press.

LOREDO, T. J., BERGER, J. O., CHERNOFF, D. F., CLYDE, M. A. and LIU, B. (2012). Bayesian methods for analysis and adaptive scheduling of exoplanet observations. Statistical Methodology 9 101–114.

MAYOR, M., PEPE, F., QUELOZ, D., BOUCHY, F., RUPPRECHT, G., LO CURTO, G., AVILA, G., BENZ, W., BERTAUX, J. L., BONFILS, X., DALL, T., DEKKER, H., DELABRE, B., ECKERT, W., FLEURY, M., GILLOTTE, A., GOJAK, D., GUZMAN, J. C., KOLHER, D., LIZON, J. L., LONGINOTTI, A., LOVIS, C., MEGEVAND, D., PASQUINI, L., REYES, J., SIVAN, J. P., SOSNOWSKA, D., SOTO, R., UDRI, S., VAN KESTEREN, A., WEBER, L. and WEILENMAN, U. (2003). Setting New Standards with HARPS. The Messenger 114 20-24.

MAYOR, M., MARMIER, M., LOVIS, C., UDRI, S., SÉGRANSAN, D., PEPE, F., BENZ, W., BERTAUX, J.-L., BOUCHY, F., DUMUSQUE, X., LO CURTO, G., MORDASINI, C., QUELOZ, D. and SANTOS, N. (2011). The HARPS search for southern extra-solar planets XXXIV. Occurrence, mass distribution and orbital properties of super-Earths and Neptune-mass planets. arXiv preprint arXiv:1109.2497.

MEUNIER, N., DESORT, M. and LAGRANGE, A.-M. (2010). Using the Sun to estimate Earth-like planets detection capabilities-II. Impact of plages. Astronomy & Astrophysics 512 A39.

OSBORNE, M. A., ROBERTS, S. J., ROGERS, A., RAMCHURN, S. D. and JENNINGS, N. R. (2008). Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. In Proceedings of the 7th international conference on Information processing in sensor networks 109–120. IEEE
Pepe, F.,Lovis, C.,Segransan, D.,Benz, W.,Bouchy, F.,Dumusque, X.,Mayor, M.,Queloz, D.,Santos, N. and Udry, S. (2011). The HARPS search for Earth-like planets in the habitable zone-I. Very low-mass planets around HD 20794, HD 85512, and HD 192310. Astronomy & Astrophysics 534 A58.

Rajpaul, V.,Aigrain, S. and Roberts, S. (2015). Ghost in the time series: no planet for Alpha Cen B. Monthly Notices of the Royal Astronomical Society: Letters 456 L6–L10.

Rajpaul, V.,Aigrain, S.,Osborne, M. A.,Reece, S. and Roberts, S. (2015). A Gaussian process framework for modelling stellar activity signals in radial velocity data. Monthly Notices of the Royal Astronomical Society 452 2269–2291.

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning 1. MIT press Cambridge.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion. Journal of the Royal Statistical Society. Series B (Methodological) 44–47.

Wilson, A.,Ghahramani, Z. and Knowles, D. A. (2012). Gaussian Process Regression Networks. In Proceedings of the 29th International Conference on Machine Learning (ICML-12) 599–606.

Wright, J. and Eastman, J. (2014). Barycentric Corrections at 1 cm s⁻¹ for Precise Doppler Velocities. Publications of the Astronomical Society of the Pacific 126 838.
Appendices

A. Kepler planet model details. The true anomaly function \( \phi(t) \) in (3.1) is given by solving the following system of equations

\[
\tan \frac{\phi(t)}{2} = \left( \frac{1 + e}{1 - e} \right) \tan \frac{E(t)}{2}
\]

(A.1)

\[ E(t) - e \sin E(t) = M(t) \]

(A.2)

\[ M(t) = \frac{2\pi t}{\tau_p} + M_0, \]

(A.3)

where \( \tau_p \) is the orbital period of the planet, \( M_0 \) is known as the mean anomaly at \( t = 0 \), and \( e \) is the orbital eccentricity.

B. Covariance matrix calculation. Here we specify the covariance matrix \( \Sigma \) implied by the model (4.2)-(4.4) and the covariance function (4.1). Let \( A^{(a,b)} \) be the matrix whose \( (i, i') \) entry is \( \text{Cov} \left( \frac{\partial}{\partial t_i} X(t_i), \frac{\partial}{\partial t_{i'}} X(t_{i'}) \right) \), for \( a, b = 0, 1, 2 \). Then, for model parameters \( a_{jk} \), for \( j = 0, 1, 2 \) and \( k = 1, 2, 3, 4 \), the \( n \times n \) diagonal block of \( \Sigma \) corresponding to output \( j \) (i.e., the square block with rows and columns \( jn + 1, \ldots, (j + 1)n \)) is given by

\[
\sum_{k_1=1}^{3} \sum_{k_2=1}^{3} a_{jk_1} a_{jk_2} A^{(k_1-1, k_2-1)} + a_{4}^2 B^{(0,0)},
\]

(B.1)

for \( j = 0, 1, 2 \), where the \( (i, i') \) entry of \( B^{(0,0)} \) is \( \text{Cov}(Z_0(t_i), Z_0(t_{i'})) \) (recall \( Z_0, Z_1, \) and \( Z_2 \) all have the same covariance function parameters \( \phi_Z \)). Similarly, the off-diagonal \( n \times n \) block corresponding to the covariance between outputs \( j_1 \) and \( j_2 \) (i.e., the square block with rows \( j_1 n + 1, \ldots, (j_1 + 1)n \) and columns \( j_2 n + 1, \ldots, (j_2 + 1)n \)) is given by

\[
\sum_{k_1=1}^{3} \sum_{k_2=1}^{3} a_{j_1 k_1} a_{j_2 k_2} A^{(k_1-1, k_2-1)}.
\]

(B.2)

Thus, all that remains is to specify \( A^{(a,b)}_{ii'} \), for \( a, b = 0, 1, 2 \). The term \( A^{(0,0)}_{ii'} \) is (4.1) with \( t = t_i \) and \( t' = t_{i'} \), and the remaining terms are given by

\[
A^{(0,1)}_{ii'} = -A^{(1,0)}_{ii'} = T_1 A^{(0,0)}_{ii'}
\]

(B.3)

\[
A^{(1,1)}_{ii'} = -T_1 A^{(0,1)}_{ii'} + T_2 A^{(0,0)}_{ii'}
\]

(B.4)

\[
A^{(0,2)}_{ii'} = A^{(2,0)}_{ii'} = -A^{(1,1)}_{ii'}
\]

(B.5)
(B.6) \[ A_{ii'}^{(1,2)} = -A_{ii'}^{(2,1)} = T_1 A_{ii'}^{(1,1)} + 2T_2 A_{ii'}^{(0,1)} + 2T_3 A_{ii'}^{(0,0)} \]

(B.7) \[ A_{ii'}^{(2,2)} = -T_1 A_{ii'}^{(1,2)} + 3T_2 A_{ii'}^{(0,2)} - 6T_3 A_{ii'}^{(0,1)} + 4T_4 A_{ii'}^{(0,0)} \]

where, writing \( \lambda_{ij} = 2\pi(t_i - t_j)/\tau \),

(B.8) \[ T_1 = \frac{\pi \sin(\lambda_{ij})}{2\tau_s \lambda_p^2} + \frac{t_i - t_j}{\lambda_p^2} \]

(B.9) \[ T_2 = \frac{\pi^2 \cos(\lambda_{ij})}{\tau_s^2 \lambda_p^2} + \frac{1}{\lambda_p^2} \]

(B.10) \[ T_3 = \frac{\pi^3 \sin(\lambda_{ij})}{\tau_s^3 \lambda_p^2} \]

(B.11) \[ T_4 = \frac{\pi^4 \cos(\lambda_{ij})}{\tau_s^4 \lambda_p^2}. \]

C. Cross validation details. Here we describe the CV approach used in stage one of our procedure, see Section 4.2. As mentioned in Section 4.2, we found that the models favored by leave-one-out cross validation tend to be overly complex and have low power for planet detection. This is unsurprising because leave-one-out prediction is substantially easier than identifying the component of a corrupted RV signal that is due to stellar activity. Indeed, the latter case is more similar to predicting all the apparent RV observations. Therefore, to somewhat better approximate the problem at hand, we instead leave out blocks of observations.

In a single repetition of our CV procedure we randomly select a test block of \( b = 5 \) consecutive observation times to hold back. Then we find the maximum likelihood parameter estimates \( \hat{\theta}_T \) based on the observations at the remaining \( n - b \) times, which we refer to as the training data. Let the subscripts \( B \) and \( T \) attached to vectors or matrices denote elements corresponding to the test block and training data, respectively. In the case of matrices the first subscript refers to the rows and the second refers to the columns. Dropping constants, our cross validation score is

\[ \frac{1}{2} \left( (s_B - \hat{\mu})^T \hat{V}^{-1} (s_B - \hat{\mu}) - \log |\hat{V}| \right) \]

where \( \hat{\mu} = \hat{m}_B - \hat{\Sigma}_{BT} \hat{\Sigma}_{TT}^{-1} (s_T - \hat{m}_T) \) and \( \hat{V} = \hat{\Sigma}_{BB} - \hat{\Sigma}_{BT} \hat{\Sigma}_{TT}^{-1} \hat{\Sigma}_{TB} \). Here \( \hat{\Sigma} \) is the estimated covariance matrix constructed using all the observation times \( t \) and the estimated model parameters \( \hat{\theta}_T \). Similarly, \( \hat{m} \) is the maximum likelihood estimate of \( m \) based on the training data. Thus, the cross validation score is the negative log conditional likelihood of the data held back under
the parameters $\hat{\theta}_T$ and conditional on the training data. It makes sense to use the log conditional likelihood rather than the log likelihood because our Gaussian process model is non-parametric, which means the unconditional likelihood of the test block observations is not very informative about the predictive power of the model. We repeat the above cross validation procedure $r = 10$ times for each model. The final cross validation score for a given model is

$$CV = \frac{1}{r} \sum_{k=1}^{r} CV_k$$

where $CV_k$ denotes the value of (C.1) for the $k$th repetition of the procedure, for $k = 1, \ldots, r$. To ensure a fairer comparison of the models, we re-use the same 10 test blocks for all models.

D. Details of optimization procedure. Optimization of the parameters of the models in the class (4.2)-(4.4) and additional planet signal parameters was mostly straightforward, but there were three aspects of our approach that were specific to the context, and we detail them here. Firstly, the parameters were divided into four blocks: (i) the model coefficients $a_{jk}$, for $j = 0, 1, 2, k = 1, 2, 3, 4$, (ii) the covariance function parameters $\phi$ for $X$, (iii) the covariance function parameters $\phi_Z$ for $Z_j$, for $j = 0, 1, 2$, and (iv) the mean function parameters including $m_j$, for $j = 0, 1, 2$ and, in the case of the full model, the planet parameter vector $\alpha$. Block (i) was optimized, followed by block (ii), and so on. We iterated through the blocks until we had completed at least 10 cycles and the log-likelihood (4.6) converged.

Secondly, rather than directly optimizing the full log-likelihood, we first optimized the log-likelihood for stellar indicator $q_1$ (again using parameter blocks as described above), and then the log-likelihood for $q_1$ and $q_2$, and finally the full log-likelihood. At each stage the optimized parameters were input into the next stage as initial values. This approach proved more successful than direct optimization of the full log-likelihood because the stellar activity indicators are unaffected by potential planet signals making it easier to find the global mode of their log-likelihood, and because in our case the indicators have a natural information ordering (for parameter estimation) in that the measurement errors of the $q_{PC1}$ observations are smaller than those of $q_{PC2}$. We repeated the above procedure for 10 initializations of the parameters and chose the run that resulted in the highest log-likelihood.

Thirdly, although most of the optimization was done using standard functions in the R software package, period and angle parameters required more care. The period parameter $\tau_s$ in the covariance function (4.1) was optimized using a fine grid search. To optimize the planet parameters we first
performed a fine 2D grid search on \( \tau_p \) and \( M_0 \), where for each candidate pair \((\tau_p, M_0)\) we used regression to quickly optimize the other planet parameters. In particular, following Loredo et al. (2012), we re-wrote (3.1) as

\[ v(t) = \beta_0 + \beta_1 (e + \cos(\phi(t))) + \beta_2 \sin(\phi(t)), \]

where \( \beta_0 = \gamma = m_0 \), \( \beta_1 = K \cos(\omega) \), and \( \beta_2 = -K \sin(\omega) \). The linear parameters \( \beta_0 \), \( \beta_1 \), and \( \beta_3 \) were then inferred by regressing the residuals for the radial velocity observations under the current stellar activity model fit against \( e + \cos(\phi(t)) \) and \( \sin(\phi(t)) \). For this step \( e \) was fixed at a typical value (in practice there is little information to infer \( e \)). After the initial grid search, the parameters values found were used to initialize a joint gradient ascent optimization of the planet parameters.

E. Doppler shift corrections for actual astronomical observations. While the Doppler shift due to a planet orbiting other stars is small, the Doppler shift due to the motion of the observing instrument is much larger. That is we must consider the Earth’s motion around Sun, and the Earth’s rotation, which combined give a Doppler shift of around 30 km/s (or \( z \approx 10^{-4} \)). Therefore, when analyzing actual astronomical observations, one must first compute the Doppler shift due to the motion of the observatory relative to the Solar System barycenter and transform the wavelength scale so that each spectrum is observed in the Solar System barycenter frame. This is now a standard astronomical procedure (e.g., Wright and Eastman, 2014). Only then could one use the linear Taylor expansion discussion in Section 5.2 to approximate any remaining Doppler shifts (which will be very small). However, in our study we use SOAP 2.0 simulated observations which are already in a stable reference frame, so no adjustments are necessary and we directly make use of the Taylor expansion.

F. Estimation of Keplerian planet model parameters. In addition to detecting planets it is of interest to infer their properties. Figure 7 summarizes the performance of the MLEs of the velocity amplitude \( K \) and orbital period \( \tau_p \) under the AIC-1 stellar activity model for the simulations used in Figure 5 (note, no simulations were performed for \( K = 0.15 \) m/s). In Figure 7 each boxplot shows the relative errors in estimating \( K \) (left plot) or the estimates of \( \tau_p = 7 \) (right plot) for a given simulation value of \( K \). For values of \( K \) greater than about 0.4 m/s, we can see that both \( K \) and \( \tau_p \) are reliably well inferred. Since the detection threshold for the AIC-1 is about 0.2 m/s, this means that the model can detect some planets even before it is able to reliably infer properties accurately. This is to be expected because planet detection requires less information than inferring planet properties. The relative errors in estimating \( K \) seem to indicate a small negative bias. This is due to the stellar activity model attempting to fit parts of the planet
Fig 7: Difference between the MLE and true value for: the planet’s RV amplitude MLE relative to the true amplitude (left) and the orbital period in days (right).

G. Detection power as a function of orbital period. In practice, we would like to be able to detect planets with a wide range of orbital properties. Based on experience analyzing RV observations, the detection sensitivity is most sensitive to RV amplitude and orbital period. Therefore, we proceed to investigate detection power under the AIC-1 model as a function of orbital period. The planet signals injected were given by $\alpha = (K, M_0 = 1.5, \tau_p, \omega = 1, e = 0.2, \gamma = 0)$ for a two-way grid of the amplitude $K$ and the period $\tau_p$, where the values used were $K = 0.1, 0.25, 0.5, 1, 2$ m/s and $\tau_p = 5, 6, \ldots, 10$ days. The other simulation and LRT details are the same as in Section 6.3.

Table 4 shows the detection power found under each setting. The numbers in parentheses in the left column give the amplitude of the planet signal as a percentage of the amplitude of the apparent RV caused by the spot $K_{\text{spot}}$, which is about 7.5 m/s for our SOAP 2.0 settings. Consistent with the findings in Section 6.3, the detection threshold seems to be between 0.1 m/s and 0.25 m/s, except for orbital periods of 5 and 10 days. For orbital periods of 5 and 10 days, the detection power is lower because these orbital
periods are difficult to separate from the 10 day stellar rotation period (and its harmonics) that appears in the stellar activity signal. This is known as the aliasing problem in astronomy. Our ongoing investigations suggest that detection power for aliased planets can be increased above the values in Table 4 by modifying the activity indicator construction method of Section 5.2 to use dimension reduction techniques that, unlike PCA, are able to find non-linear structure.

Table 4
Detection power for a range of orbital periods ($\tau_p$) and planet signal amplitudes ($K$) under the AIC-1 stellar activity model.

| K          | $\tau_p = 5$ | $\tau_p = 6$ | $\tau_p = 7$ | $\tau_p = 8$ | $\tau_p = 9$ | $\tau_p = 10$ |
|------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 0.1m/s (1.3%) | 0.00          | 0.00          | 0.12          | 0.06          | 0.10          | 0.00          |
| 0.25m/s (3.3%) | 0.00          | 0.84          | 0.88          | 0.80          | 0.84          | 0.00          |
| 0.5m/s (6.7%)  | 0.06          | 1.00          | 1.00          | 1.00          | 1.00          | 0.06          |
| 1m/s (13.4%)   | 0.56          | 1.00          | 1.00          | 1.00          | 1.00          | 0.26          |
| 2m/s (26.8%)   | 1.00          | 1.00          | 1.00          | 1.00          | 1.00          | 0.32          |