# IMPROVING EXOPLANET DETECTION POWER: MULTIVARIATE GAUSSIAN PROCESS MODELS FOR STELLAR ACTIVITY

By David E. Jones[1,a], David C. Stenning[2,b], Eric B. Ford[3,4,5,6,c],
Robert L. Wolpert[7,e], Thomas J. Loredo[8,f], Christian Gilbertson[3,d] and
Xavier Dumusque[9,g]

[1]*Department of Statistics, Texas A&M University,* [a]*david.jones@tamu.edu*

[2]*Department of Statistics and Actuarial Science, Simon Fraser University,* [b]*dstennin@sfu.ca*

[3]*Department of Astronomy and Astrophysics, Pennsylvania State University,* [c]*ebf11@psu.edu,* [d]*cjg66@psu.edu*

[4]*Center for Exoplanets & Habitable Worlds, Pennsylvania State University*

[5]*Center for Astrostatistics, Pennsylvania State University*

[6]*Institute for Computational & Data Sciences, Pennsylvania State University*

[7]*Department of Statistical Science, Duke University,* [e]*rlw@duke.edu*

[8]*Cornell Center for Astrophysics and Planetary Science, Cornell University,* [f]*loredo@astro.cornell.edu*

[9]*Observatoire Astronomique de l'Université de Genève,* [g]*Xavier.Dumusque@unige.ch*

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. Low-mass planets typically cause their host star to move with radial velocities of 1 m/s or less. By analyzing a time series of stellar spectra from a host star, modern astronomical instruments can, in theory, detect such planets. However, 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. A principled approach to recovering planet radial velocity signals in the presence of stellar activity was proposed by Rajpaul et al. (*Mon. Not. R. Astron. Soc.* **452** (2015) 2269–2291). It uses 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 new high-information stellar activity indicators; and (ii) we extend the Rajpaul et al. (*Mon. Not. R. Astron. Soc.* **452** (2015) 2269–2291) model to a larger class of models and use a power-based model comparison procedure to select the best model. Despite significant interest in exoplanets, previous efforts have not performed large-scale stellar activity model selection or attempted to evaluate models based on planet detection power. In the case of main sequence G2V stars, we find that our method substantially improves planet detection power, compared to previous state-of-the-art approaches.

## 1. Introduction.

1.1. *Motivation.* In this paper we present a statistical framework to improve the sensitivity of astronomical surveys for detecting exoplanets, that is, 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
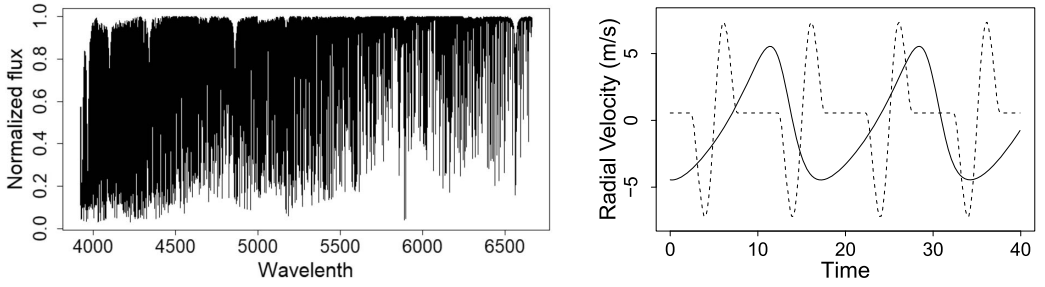
FIG. 1. *Example stellar spectrum (*left*) and an illustration of the scientific model for a RV signal due to a planet (*right*, solid line) and the RV corruption due to a spot (*right*, dashed line).*

toward the observer, respectively. That is, the observed starlight contains a periodic signal, due to the Doppler effect, resulting from the star's motion.

In radial velocity exoplanet surveys, the raw data are a time series of high-dimensional spectra, and it is from these that the magnitude of the Doppler effect is measured, as we now explain. A spectrum is the intensity of starlight as a function of wavelength, and 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) because the star's atmosphere absorbs light at specific wavelengths. The Doppler effect is detected by measuring how the observed wavelengths of the spectral lines are shifted over time. In particular, if $f$ denotes the unperturbed stellar spectrum and $z(t)$ denotes the Doppler shift at time $t$, then the observed spectrum at time $t$ and wavelength $\lambda$ is given by

$$(1) \qquad f_t(\lambda) = f\left(\frac{\lambda}{1 + z(t)}\right) + \epsilon_{t\lambda},$$

where $\epsilon_{t\lambda}$ denotes noise. While the Doppler shifts $z(t)$ due to a planet are very small, modern astronomical instrumentation can detect them by jointly analyzing many spectral lines, for example, Butler et al. (1996), Baranne et al. (1996), Mayor et al. (2003).

In practice, the computed Doppler shifts are typically converted to a more physically interpretable quantity, namely, the velocity of the star projected onto the observer's line of sight which is known as the *radial velocity* (RV) of the star. The radial velocity is denoted $v(t)$ and is given by $v(t) = cz(t)$, where $c$ is the speed of light. Currently, radial velocities can be determined down to $\sim$1 m/s (Fischer et al. (2016)), and upcoming instruments will likely be even more sensitive. To hunt for planets, astronomers choose candidate stars and for each collect a time series of RV observations spanning many days up to years, for example, 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. The right panel of Figure 1 shows an example RV signal produced by the scientific model for a large planet with a 17 day orbital period (solid line). Hunting for periodic RV signals such as this is known as the *radial velocity method* and is one of the most successful methods for detecting exoplanets. However, astronomers would like to reach below the current detection threshold of 1 m/s because many low-mass planets induce RV signal amplitudes smaller than this threshold, for example, Earth induces an RV signal of about nine cm/s.

A key challenge for the radial velocity method in practice, and the focus of this paper, is that planet RV signals can be mimicked by intrinsic stellar variability, that is, phenomena occurring in the outer part of the observed star. For instance, stellar magnetic fields can lead to groups of small, dark and relatively cool *star spots*, often surrounded by brighter regions known as *faculae*. As a monitored star rotates, any spots (or faculae) will move across the visible stellar disk and lead to a spurious signal in the observed RV time series. Stellar rotation

periods are often similar to plausible planetary orbital periods and, therefore, distinguishing spots and faculae from planets can be particularly problematic, for example, the signal discovered by Dumusque et al. (2012) was later found to be spurious by Rajpaul et al. (2015) and Rajpaul, Aigrain and Roberts (2015). An example of the RV corruption due to a single, nonevolving star spot is shown by the dashed line in the right panel of Figure 1. This dashed line and the solid line of the example planet RV signal are easily distinguished, but, in practice, separation is much more difficult due to sparse observations, noise, unknown periods, the temporal evolution of active regions and the potential for multiple active regions at once.

Despite these difficulties, it is often, in principle, possible to distinguish stellar activity from a true planet signal by using more of the raw data, rather than only computing the RV time series. Specifically, we can exploit the fact that a true Doppler shift affects the entire spectrum in the same way (i.e., *all* the spectral lines are shifted; see (1)), whereas the effect of intrinsic stellar activity is different for each spectral line. However, progress toward developing statistical frameworks that make use of this fact has so far been limited. Recently, Davis et al. (2017) used principal component analysis (PCA) to construct multiple univariate measures of stellar activity, known as stellar activity indicators, based on low-noise spectra. They showed that the evolution of these indicators over time is different for stellar activity and Doppler shifts. On the other hand, they did not provide a full statistical method for making use of their stellar activity indicators to detect planets. Meanwhile, Rajpaul et al. (2015) proposed a detection approach based on a sophisticated Gaussian process model that jointly captures an RV time series and some conventional stellar activity indicators. However, the impact of their modeling framework on planet detection power is unclear.

This paper unifies and extends the work of Davis et al. (2017) and Rajpaul et al. (2015) to propose a powerful framework for selecting stellar activity models and developing new high-information stellar activity indicators. Currently, most exoplanet detection methods either do not use statistical models for stellar activity at all, or use a single convenient or physically motivated model, without any empirical or theoretical statistical validation of the detection performance. To address this limitation, this paper provides a framework, which can evaluate and compare a large number of stellar activity models in terms of statistical power for detecting exoplanets, thereby facilitating automated and substantially improved model selection. This automated approach also makes it easy to use novel stellar activity indicators, such as those of Davis et al. (2017), whereas previous approaches have generally been restricted to specific indicators. Our second contribution is to propose a modified version of the Davis et al. (2017) indicator construction technique. Specifically, we introduce a simple preprocessing step to separate the indicators from RV signals (real or corrupted) and, thereby, preserve interpretability and facilitate modeling. For both these new activity indicators and some conventional ones, we show that applying our model selection procedure identifies models that have high power for detecting small exoplanets in the presence of both small and large spots, whereas existing approaches either do little to address stellar activity or have substantially lower power in the case of large spots.

Exoplanet detection is a developing area of research, and our work does not address all the scientific and statistical challenges posed by the analysis of exoplanet survey data. In this paper we restrict our focus to spot-related activity on main sequence G2V (MS-G2V) stars, that is, stars similar to the Sun. An initial focus on spots is reasonable because, although other forms of stellar variability[1] also affect the stellar spectrum, they typically cause

---

[1]We will mainly use the term *stellar activity* because in astronomy this tends to refer to magnetically active regions (e.g., spots), whereas *stellar variability* is more general. For further details on the nature of stellar variability, the reader is referred to the relevant astronomy literature, for example, Jenkins (2013), Dumusque, Boisse and Santos (2014), Borgniet, Meunier and Lagrange (2015), Haywood et al. (2016).

spurious RV signals that change on relatively short timescales; for example, see Del Moro et al. (2004), Del Moro (2004), Arentoft et al. (2008). Second, to carefully study exoplanet detection power for MS-G2V stars with spot activity we need a ground truth dataset, which, unfortunately, is not available due to the inherent challenges in astronomy, i.e., there is no way to know for sure what type of activity a distant star has or whether a planet is orbiting it. To overcome this challenge, we construct a realistic dataset of 1800 MS-G2V stars with various stellar spot configurations by synthesizing spectra of the Sun, using the Spot Oscillation and Planet (SOAP) 2.0 software which was developed by Boisse, Bonfils and Santos (2012) and Dumusque, Boisse and Santos (2014). This approach has the added advantage of avoiding the unique complications associated with each individual instrument that likely differ to the complications of future instruments, which our study can help to guide. Use of SOAP 2.0 data is a well-established approach for developing methods to mitigate stellar activity, for example, Rajpaul et al. (2015), Dumusque (2016), Davis et al. (2017), though previous SOAP 2.0 based investigations have not specifically studied detection power. The recent work of Damasso et al. (2019) did consider detection performance (and various biases) associated with different activity models, using simulations of stellar activity from a quasi-periodic Gaussian process. However, our SOAP 2.0 simulations are, in some ways, more realistic, being based on real spectra, and we compare a much larger number of models which are also more sophisticated. For example, we consider the Rajpaul et al. (2015) model and extend it, whereas Damasso et al. (2019) do not.

This article is organized as follows. Section 2 overviews the paper and provides pseudocode for the key algorithms. Section 3 introduces the raw input data and the SOAP 2.0 package. Section 4 presents our approach for defining high-information stellar activity indicators. Section 5 presents our general class of statistical models for jointly capturing stellar activity indicators and the RV time series. Section 6 then introduces our model selection procedure. Section 7 applies our model selection procedure and compares the planet detection power under the top performing models to that obtained using the Rajpaul et al. (2015) model. Section 8 discusses the results, implications for future RV planet surveys and areas for further research. Our code and dataset are provided in the Supplementary Material (Jones et al. (2022)) and is also available on GitHub at https://github.com/david-jones-statistics/improving-planet-detection-power.

## 2. Overview and key algorithms.
Our paper contains five main components:

*Part* 1. Stellar activity data and preprocessing.
*Part* 2. Defining new stellar activity indicators.
*Part* 3. Model class for stellar activity indicators.
*Part* 4. Ideal power-based model selection.
*Part* 5. Practical power-based model selection.

We now overview each of these parts and explain how they fit together. Details will be given in Sections 3–7.

2.1. *Stellar activity dataset synthesis and preprocessing.* The raw data to be used in simulating spectra consists of highly precise solar spectra from both active and quiet subregions of the Sun, a main sequence G2V star, and was collected by Wallace, Hinkle and Livingston (1998) and Wallace, Hinkle and Livingston (2005). The raw data can be considered ground-truth stellar activity because the effects of planets in the Solar System are well known and were accounted for in the data collection process.

With this raw data as input, we apply the Spot Oscillation and Planet (SOAP) 2.0 software developed by Boisse, Bonfils and Santos (2012) and Dumusque, Boisse and Santos (2014) to synthesize the observations into a time series of $n$ spectra of the full visible disk of the Sun.

The power of SOAP 2.0 is that it is able to synthesize the spectra so as to accurately represent observations of an MS-G2V star under different stellar activity and viewing configurations, for example, different spot sizes and locations. This reconstruction of different situations is possible because the raw input spectra are spatially localized and precise, and the underlying geometry is relatively simple.

To achieve our two goals of defining new stellar activity indicators and performing power-based model selection, we, in fact, use SOAP 2.0 multiple times. First, we run SOAP 2.0 with a single large spot configuration to obtain a low-noise stellar activity dataset $Y_s$, an $n \times p$ matrix, where $n = 125$ and $p = 237{,}944$. Each matrix row is a $p$-dimensional spectrum of the star at one of $n$ regularly spaced observation times. In Section 2.2 below, we use $Y_s$ to *define* new high-information stellar activity indicators.

Next, we use SOAP 2.0 to obtain many additional datasets for empirically computing exoplanet detection power, using the procedure described in Sections 2.3 and 2.4. In particular, we run SOAP 2.0 $M = 1800$ more times with $M$ different stellar activity settings, plus adjustments to take account of the lower sign-to-noise (SNR) and irregular observation cadences typical in exoplanet surveys. The result is a collection, $Y_A$, of $M$ observation time vectors together with their corresponding data matrices, that is, $Y_A = \{(t^{(m)}, Y^{(m)}) : m = 1, \ldots, M\}$, where each value of $m$ corresponds to a different spot size and latitude setting. Further details regarding SOAP 2.0 and the datasets $Y_s$ and $Y_A$ are given in Section 3.

2.2. *Defining new stellar activity indicators.*   Since the Sun can be observed precisely, the dataset $Y_s$ is almost noiseless, and we use it to *define* new stellar activity indicators. In spectroscopic surveys a Doppler shift of the spectra of a host star is the only observable impact of an exoplanet. We, therefore, seek to separate this effect from the effects of stellar activity. Since stellar activity can also cause an apparent Doppler shift, it is not possible to immediately separate exoplanet signals from all stellar activity. However, stellar activity *can* be isolated by considering variations in the stellar spectra that are orthogonal to a Doppler shift. To summarize such variations, we apply a modified principal component analysis (PCA) to the data matrix $Y_s$, as detailed in Algorithm 1 below.

ALGORITHM 1 (Doppler-constrained PCA).

1. Input: $n \times p$ data matrix $Y_s$.
2. Based on $Y_s$, approximate the component $w$ corresponding to a Doppler shift of a spectrum; see Appendix A.
3. Subtract the Doppler component from the observed spectra,

$$(2) \qquad\qquad \widetilde{Y}_s = Y_s - \frac{Y_s w w^T}{\sum_{i=1}^{p} |w_i|^2}.$$

4. Apply PCA to $\widetilde{Y}_s$ to obtain the basis vectors $v_k$, for $k = 1, \ldots, n-1$.

The details of Algorithm 1 are discussed in Section 4. The collection $\mathcal{I}_{\text{MS-G2V}} = \{w, v_1, \ldots, v_L\}$ defines $L$ stellar activity indicators plus the RV signal, where $L \leq n-1$ must be chosen. Given a dataset, the stellar activity indicators (and RV signal) are obtained by projecting onto the vectors in $\mathcal{I}_{\text{MS-G2V}}$, that is, the indicators are the scores, which we denote by $u(t_i)$ and $q_{\text{PC}j}(t_i)$, for $j = 1, \ldots, L$ (here, $u$ corresponds to $w$). The indicators for $Y_s$ are plotted in the top row of Figure 2. The scores $q_{\text{PC}j}(t_i)$, for $j = 1, \ldots, L$, will always summarize only stellar activity. The scores $u(t_i)$ summarize either: (a) stellar activity *and* a planet signal (if there is a planet) or (b) only stellar activity (if there is not a planet). In the case of Figure 2, there is in fact no planet. For convenience, we write $s = (u, q_{\text{PC}1}, \ldots, q_{\text{PC}L})$, where $u = (u(t_1), \ldots, u(t_n))^T$ and $q_{\text{PC}j} = (q_{\text{PC}j}(t_1), \ldots, q_{\text{PC}j}(t_n))^T$, for $j = 1, \ldots, L$, and refer to $s$ as the *stellar activity indicators*. We also write $t = (t_1, \ldots, t_n)^T$.

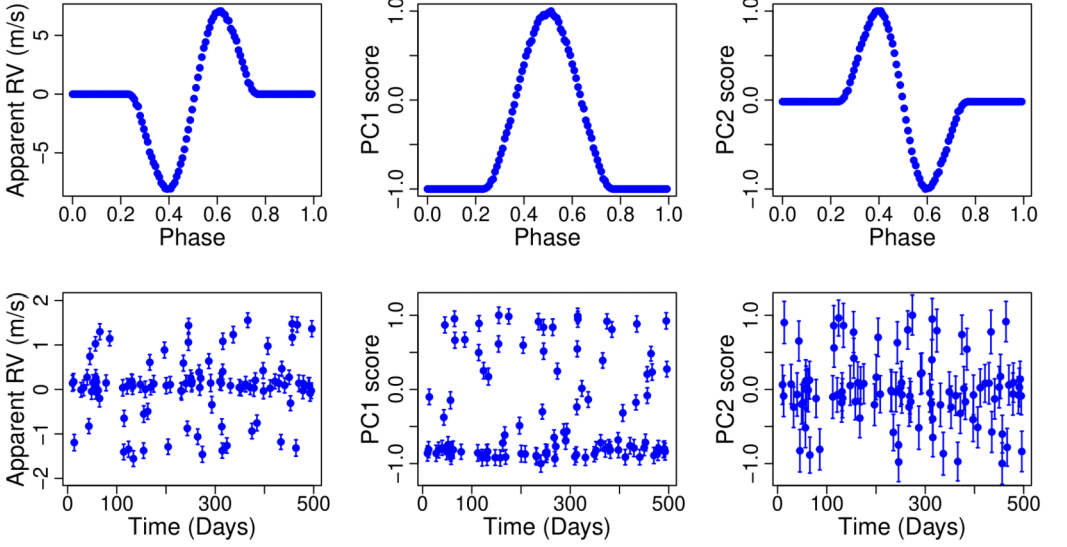FIG. 2. *Top row*: *Noiseless realizations of* $u$, $q_{PC1}$ *and* $q_{PC2}$ *obtained by projecting* $Y_s$ *onto the components of* $\mathcal{I}_{MS\text{-}G2V}$. *Bottom row*: *Realizations of* $u$, $q_{PC1}$, $q_{PC2}$ *for one of the noisy datasets contained in* $Y_A$, *plotted over time as opposed to in phase.*

The indicators, defined by $\mathcal{I}_{MS\text{-}G2V}$, were derived using the specific spot size and latitude appearing in $Y_s$, but we found them to be almost identical for other configurations, suggesting that these indicators are generally useful, at least in the case of MS-G2V stars which tend to have similar patterns of stellar activity. In other words, for a new dataset from an MS-G2V it is reasonable to expect that its stellar activity will again be well captured by the stellar activity indicator components $\mathcal{I}_{MS\text{-}G2V}$. Conventional stellar activity indicators are similarly defined, based on features of a low-noise or ideal spectrum, but they are typically not constructed in a way that necessarily extracts maximal information or even ensures that the indicators capture different information to each other.

2.3. *Ideal power-based model selection.* Stellar activity is highly complex, and it is generally challenging to jointly model multiple stellar activity indicators. We, therefore, propose a flexible class of models in Section 5.2 and then use a model selection approach to find the best models within the class. For now, we focus on our model selection approach and postpone specification of the model class until Section 5.

Heuristically, in the context of exoplanet detection the point of stellar activity modeling is to infer how much of the apparent RV signal shown in the top left panel of Figure 2 is due to stellar activity by making use of the two indicator time series plotted in the top middle and top right panels. In other words, simply modeling the stellar activity data is not the primary objective, our goal is to *detect exoplanets via a hypothesis test*. The null hypothesis of the test is that there is no planet and a stellar activity model can sufficiently capture the data; the alternative hypothesis is that there is a planet and, therefore, an additional model component is needed to capture the planet. Given our goal of planet detection, the correct way to select the model is to evaluate the planet detection power under each candidate stellar activity model and then select the model yielding the highest power. This is what Algorithm 2 below does. For convenience, we split the collection $Y_A$ into two parts $Y_N$ ($M_N = 1000$ datasets) and $Y_P$ ($M_P = 800$ datasets), which are used for generating a null distribution and generating datasets with planet signals, respectively.

ALGORITHM 2 (Power-based model selection).

1. Input: Set $\mathcal{L}$ of candidate stellar activity models, SOAP 2.0 dataset collection $Y_A = (Y_N, Y_P)$, vector $(K_1, \ldots, K_R)$ of planet signal magnitudes (in m/s), Type I error probability $\alpha$.

2. Compute null distribution and critical value:

   (i) Compute activity indicators $s_N^{(m)}$ from $(t_N^{(m)}, Y_N^{(m)})$, for $m = 1, \ldots, M_N$.
   
   (ii) For each $l \in \mathcal{L}$, compute the likelihood ratio test (LRT) statistic

$$(3) \qquad \Delta(t_N^{(m)}, s_N^{(m)}, l) = l_{\text{act}}(\hat{\theta}_{\text{act}}^{(m)}; t_N^{(m)}, s_N^{(m)}, l) - l_{\text{full}}(\hat{\theta}_{\text{full}}^{(m)}; t_N^{(m)}, s_N^{(m)}, l),$$

   for $m = 1, \ldots, M_N$, and let $\mathcal{F}_l$ denote the resulting distribution. Here, $l_{\text{act}}$ and $l_{\text{full}}$ are the likelihood functions under the *stellar activity only model* and the *full model*, which, additionally, includes the planet component given by (9) in Section 5.1.

   (iii) For each $l \in \mathcal{L}$, set $c_l$ to be the $1 - \alpha$ quantile of $\mathcal{F}_l$.

3. Compute power:

   (i) Compute activity indicators $s_P^{(m)} = (u^{(m)}, q_{\text{PC1}}^{(m)}, \ldots, q_{\text{PCL}}^{(m)})$ from $(t_P^{(m)}, Y_P^{(m)})$, for $m = 1, \ldots, M_P$.

   (ii) For $r \in \{1, \ldots, R\}$, inject planet RV signal of magnitude $K_r$ m/s into $u^{(m)}$, for $m = (r - 1)n_p + 1, \ldots, rn_P$, where $n_P = M_P / R$.

   (iii) For each $l \in \mathcal{L}$, compute the empirical power for each value of $r$

$$(4) \qquad p(l, r) = \frac{1}{n_P} \sum_{m=(r-1)n_p+1}^{rn_P} 1_{\{\Delta(t_P^{(m)}, s_P^{(m)}, l) > c_l\}}.$$

4. Select $l \in \mathcal{L}$ that has the smallest value of $r$ satisfying $p(l, r) \geq 0.5$.

Note that, since each dataset has a different spot size and latitude, the power computed in (4) marginalizes over these configuration settings. This is important because the configuration is unknown in practice.

2.4. *Practical power-based model selection.* Unfortunately, Algorithm 2 cannot be applied directly in practice, because the number of candidate models is very large. Instead, in Algorithm 3 below, we first shortlist some stellar activity models using three standard model selection criteria: the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and a cross-validation score (CV). The CV score we use is based on leaving out multiple observations, and the details are given in Appendix E. To construct the short list we make use of $M_R = 10$ datasets $(t_R^{(m)}, Y_R^{(m)})$, $m = 1, \ldots, M_R$, that well represent the distribution of spot sizes and latitudes. These can either be selected from $Y_A$ or simulated from SOAP 2.0 in the same manner. Here, we use the latter approach and choose the representative latitudes and sizes, using a maximin Latin hypercube design; see Appendix B for details.

ALGORITHM 3 (Power-based model selection in practice).

1. Input: Set $\mathcal{L}$ of candidate stellar activity models, datasets $(t_R^{(m)}, Y_R^{(m)})$, for $m = 1, \ldots, M_R$.

2. Compute activity indicators $s_R^{(m)}$ from $(t_R^{(m)}, Y_R^{(m)})$, for $m = 1, \ldots, M_R$.

3. For each $l \in \mathcal{L}$, compute

$$(5) \qquad \overline{\Delta \text{AIC}}(l) = \frac{1}{M_R} \sum_{m=1}^{M_R} \left\{ \text{AIC}(l, t_R^{(m)}, s_R^{(m)}) - \min_{k \in \mathcal{L}} \text{AIC}(k, t_R^{(m)}, s_R^{(m)}) \right\},$$

and, similarly, compute $\overline{\Delta \text{BIC}}(l)$ and $\overline{\Delta \text{CV}}(l)$.

4. Shortlist the top five models under each of the three selection criteria.
5. Apply Algorithm 2 to the short list of models.

In (5), AIC($l, \boldsymbol{t}_R^{(m)}, \boldsymbol{s}_R^{(m)}$) denotes the AIC for model $l$ and dataset ($\boldsymbol{t}_R^{(m)}, \boldsymbol{s}_R^{(m)}$). The results of applying Steps 1–4 of Algorithm 3 are presented in Table 1 in Section 7.2. The final model selected is denoted $\mathcal{M}_{\text{MS-G2V}}$. To increase the chance that the model selected has the highest detection power, one can increase the number of models shortlisted in Step 4 of Algorithm 3 at the expense of more computation.

For a new dataset from an MS-G2V star, it is reasonable to expect that its stellar activity will again be well captured by the indicators, defined by $\mathcal{I}_{\text{MS-G2V}}$, and the corresponding model $\mathcal{M}_{\text{MS-G2V}}$, identified by Algorithm 3. Thus, in principle, no new indicator and model development is needed when applying our method to a new dataset (from an MS-G2V star): we simply compute the stellar activity indicators (using $\mathcal{I}_{\text{MS-G2V}}$) and perform a likelihood ratio test. The null distribution used by the test is, again, that previously found in Step 2 of Algorithm 2. Similarly, since the model is fixed, the Type I error rate $\alpha$ input into Algorithm 2 and the power for detecting an exoplanet with signal amplitude $K_r$ will be that computed in Step 3 of Algorithm 2.

Nonetheless, given the possibility of more complex variations in the form of activity across MS-G2V stars, a case for recomputing the indicator components $\mathcal{I}_{\text{MS-G2V}}$ for each new dataset could be made, and this approach is illustrated in Section 8.2. However, except in the presence of very large spots, we found that principal components computed from noisy datasets were generally noisy themselves and less useful for exoplanet detection than $\mathcal{I}_{\text{MS-G2V}}$. Furthermore, recomputing the PCA basis for each dataset adds substantial computation to Algorithm 3. Therefore, in the main part of this paper we use the fixed indicators, given by $\mathcal{I}_{\text{MS-G2V}}$ and the model $\mathcal{M}_{\text{MS-G2V}}$. In this respect our method is similar to conventional approaches which pre-define indicators and models based on physical knowledge. The difference is that we derive the indicators empirically (from low-noise data) and select a model from a large class based on power for exoplanet detection.

## 3. Spectroscopic time series datasets.

3.1. *Dataset $Y_s$ used to define stellar activity indicators.* The low-noise dataset $Y_s$ is an $n \times p$ matrix containing a times series of $n = 125$ spectra each of dimension $p = 237{,}944$; that is, each of the $n$ spectra consist of the measured light intensity at the same 237,944 wavelengths. The dataset captures stellar activity in the form of a single stellar spot and is used as an ideal dataset with respect to which we define new stellar activity indicators by running Algorithm 1, introduced in Section 2.2; see Section 4.2 for full details. To obtain $Y_s$, we ran SOAP 2.0 a single time to synthesize the raw input spectra, described in Section 2.1, into a time series of spectra of the full visible disk of the Sun. The observation times are equally spaced and have solar rotation phases $p_i = i/125$, for $i = 0, \ldots, 124$. For our SOAP 2.0 settings, the activity takes the form of a single large spot that covers 1% of one hemisphere of the solar surface, the spot latitude from the Solar equator is 40° and the inclination of the solar rotation axis relative to the line of sight is 90°.

Unfortunately, it is not practical to observe stars more distant than the Sun with the same resolution and signal-to-noise as our raw input data. Thus, we modify SOAP 2.0 to reduce the output resolution to a practically achievable level. The result is that the number of wavelengths recorded is reduced from 523,732 (in the raw input spectra) to $p = 237{,}944$.[2] This

---

[2]More precisely, the resolution is reduced from $\sim 10^4$ to $1.5 \times 10^5$ by convolving the higher-resolution spectra with a Gaussian line spread function and resampling using cubic splines.

modification is necessary because if we were to define stellar activity indicators based on unrealistically high resolution spectra, then the indicators may primarily capture spectral features that would not be observable in practice.

3.2. *Dataset $Y_A$ used to perform detection power-based model selection.* The dataset $Y_A$ is a collection of observations from $M = 1800$ MS-G2V stars, generated by SOAP 2.0. For each star, $Y_A$ contains a vector of $n_A = 100$ observation times and a corresponding data matrix containing $n_A$ spectra, that is, $Y_A = \{(t^{(m)}, Y^{(m)}_{n_A \times p}) : m = 1, \ldots, M\}$. The dataset is representative of exoplanet survey observations of 1800 MS-G2V stars with varying degrees of stellar activity, and Algorithm 3 uses it to perform model selection by empirically evaluating planet detection power under different stellar activity models; see Section 6. To obtain $Y_A$, we ran SOAP 2.0 $M$ times, where for each run the spot size and latitude were drawn from the distribution of configurations expected for the Sun (or similar stars). Specifically, the distributions are

$$(6) \qquad S_{\text{lat}} \sim \pi \mathcal{N}_{ab}(\mu, \sigma^2) + (1 - \pi)\mathcal{N}_{ab}(-\mu, \sigma^2),$$

$$(7) \qquad S_{\text{size}} \sim \rho \mathcal{LN}_d(\theta_1, \gamma_1^2) + (1 - \rho)\mathcal{LN}_d(\theta_2, \gamma_2^2),$$

where $\mathcal{N}_{ab}$ denotes a truncated Gaussian distribution with $a = -90°$ and $b = 90°$ being the lower lower and upper bound, respectively, and $\mathcal{LN}_d$ denotes a lower truncated log-Gaussian distribution with lower bound $d = 10$ micro solar hemispheres (MSH). To make (6) and (7) consistent with the distributions given in Baumann and Solanki (2005) and Mandal, Karak and Banerjee (2017), respectively, we set $\pi = 0.5$, $\mu = 15.1$, $\sigma = 7.3$, and $\rho = 0.4$, $\theta_1 = \log(46.51) + \log(2.14)$, $\theta_2 = \log(90.24) + \log(2.49)$, $\gamma_1 = \sqrt{\log(2.14)}$, $\gamma_2 = \sqrt{\log(2.49)}$ (where log denotes the natural logarithm). Although, in practice, there may be multiple spots, they typically form in localized groups, which can be approximated as a single spot (as is done here), and the distribution (7) is consistent with this approximation.

To ensure that the detection powers computed by Algorithm 3 are relevant, we set the observation cadence and signal-to-noise (SNR) to be reasonable for exoplanet surveys. Specifically, for each star (i.e., each value of $m$) we set the total observation window to be 500 days. Next, we randomly select $n_A = 100$ phases $p^{(m)}_1, \ldots, p^{(m)}_{100}$, from among $p_i = i/125$, $i = 0, \ldots, 124$ (without replacement). Finally, we randomly generate a rotation cycle of the star $r^{(m)}_i \in \{0, \ldots, 49\}$ and set the observation time to be $t^{(m)}_i = (r^{(m)}_i + p^{(m)}_i)\tau^{(m)}_s$, for $i = 1, \ldots, n_A$, where $\tau_s$ is the stellar rotation period. In this work we set $\tau^{(m)}_s = \tau_s = 10$ days. To generate spectra with a given SNR, we add noise $\epsilon_{t\lambda}$ to the spectrum at each wavelength. Specifically,

$$(8) \qquad \epsilon_{t\lambda} \sim \mathcal{N}(0, \beta_t f_t(\lambda)/\text{SNR}),$$

where $\beta_t = \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} f_t(\lambda) \, d\lambda$. This noise distribution is based on the fact that a Poisson distribution accurately models the uncertainty involved in the physical process of photons being detected by a telescope. A Gaussian is used in (8) because $f_t(\lambda)$ is continuous, due to observation stage processing steps. We set $\text{SNR} = 500$, unless otherwise specified. This is higher than typical of existing observations but still feasible for exoplanet surveys. Investigating this setting is valuable because an important goal of exoplanet astronomy is establishing what planets could be detected with improved instruments, before the instruments are constructed (because it is very expensive and time-consuming to design and build astronomical instruments).

## 4. Defining new stellar activity indicators.

4.1. *Spectroscopic stellar activity indicators.* It is prohibitively challenging to directly model the temporal evolution of the high-dimensional spectra stored in the rows of the data matrix $Y_s$ (or $Y_A^{(m)}$, for $m = 1, \ldots, M$). Instead, astronomers typically model time series of a small number of stellar activity indicators, which are functionals of individual spectra, summarizing the level and nature of stellar activity. The indicators used by Rajpaul et al. (2015) (hereafter, R15) are $\log R'_{HK}$ and BIS, which measure the emission in the core of the Calcium II H & K spectral lines and the asymmetry of an "averaged" version of the spectral lines, respectively. It is nontrivial to compute $\log R'_{HK}$, and R15 suggest replacing it by normalized flux when $\log R'_{HK}$ is not available.

Some stellar activity indicators are summaries of very specific parts of the stellar spectrum, for example, $\log R'_{HK}$. More generally, a stellar activity indicator can be any functional $g$ of the stellar spectra observed. This raises the question of how to choose $g$ in order to capture as much information as possible. Indeed, the measures $\log R'_{HK}$ and BIS were designed for purposes other than planet detection, and, therefore, in the current context are somewhat arbitrary. It is thus desirable to have an empirical method for identifying stellar activity indicators that are informative for our goals of detecting and characterizing exoplanets.

4.2. *Defining new stellar activity indicators via Doppler-constrained PCA.* Recently, Davis et al. (2017) proposed using principal component analysis (PCA) to construct stellar activity indicators in an empirical way. However, it is not immediately apparent how their stellar activity indicators can be used in practice because it is unclear if any given principal component corresponds to stellar activity, a planet RV signal or both.

Algorithm 1 in Section 2.2 specifies our simple adaptation of PCA which separates the apparent RV signal from the PCA scores and thereby addresses the difficulties posed by the Davis et al. (2017) indicators. In summary, PCA is applied to $\widetilde{Y}_s = Y_s - Y_s \boldsymbol{w}\boldsymbol{w}^T / \sum_{i=1}^p |\boldsymbol{w}_i|^2$, where the $p \times 1$ component $\boldsymbol{w}$ corresponds to a Doppler shift and is specified below. We denote the PCA components by $\boldsymbol{v}_j$, for $j = 1, \ldots, n-1$. The collection of basis vectors $\mathcal{I}_{MS\text{-}G2V} = \{\boldsymbol{w}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_l\}$ define the apparent RV signal and our stellar activity indicators for MS-G2V stars. Based on the findings of Davis et al. (2017) and our own experience, $L$ does not need to be larger than 2 or 3, so we choose $L = 2$.

Since Doppler shifts of the spectral wavelengths are very small, they closely correspond to shifts in the direction of the spectrum derivative scaled by the wavelength (as can be seen from a Taylor expansion of (1)). We, therefore, choose $\boldsymbol{w}$ in Algorithm 1 to be the derivative of the average spectrum in $Y_s$, that is, $\boldsymbol{w} = (\lambda_1 \bar{f}'(\lambda_1), \ldots, \lambda_p \bar{f}'(\lambda_p))^T$; see Appendix A for full details. We emphasize that $\boldsymbol{w}$ is simply the direction along which a Doppler shift occurs, not the magnitude or sign of an observed Doppler shift. Therefore, even if there was a planet in the dataset used to define $\boldsymbol{w}$ (which for $Y_s$ there is not), then this would not change $\boldsymbol{w}$ significantly.

The bottom row of Figure 2 shows our stellar activity indicators for one of the stars in the cadence and SNR adjusted dataset $Y_A$. Specifically, we have plotted $u(t_i)$ and $q_{PCj}(t_i)$, for $j = 1, 2$ and $i = 1, \ldots, 100$. Two standard deviation error bars are plotted to indicate the uncertainty induced by the spectrum level noise, as described by (8).

## 5. Framework for modeling planet signals and stellar activity.

5.1. *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

$$(9) \qquad v(t) = K\left(e\cos\omega + \cos(\omega + \phi(t))\right) + \gamma,$$

where $e$ is the eccentricity of the planetary 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 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 C and which depend on the planet orbital period $\tau_p$, in addition to other parameters. The right panel of Figure 1 shows an example planet RV signal (solid line).

5.2. *General class of stellar activity models.* We now propose a class of Gaussian process models to jointly capture changes in the RV corruption due to stellar activity and times series of multiple stellar activity indicators, for example, those in the bottom row of Figure 2. The indicators can be general and need not be the Doppler-constrained PCA scores discussed in Section 4.2. Our approach is an extension of the model proposed by R15. More generally, our model class can be viewed as an adaptation of the linear model of coregionalization (LMC); see, for example, Journel and Huijbregts (1978), Osborne et al. (2008) and Álvarez and Lawrence (2011).

A real-valued continuous time stochastic process $\{X(t)\}$ is a Gaussian process (GP) if for every finite set of times $t_1, \ldots, t_\eta$ the vector $(X(t_1), \ldots, X(t_\eta))$ has a multivariate Gaussian distribution. 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

$$(10) \qquad \kappa(t, t') = \exp\left(-\frac{\sin^2(\pi(t - t')/\tau_s)}{2\lambda_p^2} - \frac{(t - t')^2}{2\lambda_e^2}\right),$$

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. That this covariance function leads to a positive definite covariance matrix follows from the fact that the product of two valid covariance functions yields a valid covariance function; see Rasmussen and Williams ((2006), page 95). The same covariance function was also adopted by R15. The motivation for a quasi-periodic covariance structure is that, for a given 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 reader is referred to Rasmussen and Williams (2006) for a comprehensive introduction to Gaussian processes.

We also make use of the first two derivatives of $X(t)$ in our model class (12)–(15) below, where we have denoted them by $\dot{X}(t)$ and $\ddot{X}(t)$. The inclusion of $\dot{X}(t)$ is motivated by physical arguments in Aigrain, Pont and Zucker (2012), which demonstrate that the RV corruption, due to a spot, can be approximated by a function of the area of the spot projected onto our line of sight and the derivative of this area. We additionally include $\ddot{X}(t)$ to capture second order effects of the spot area, as suggested by R15 (though, for simplicity, they did not use this second derivative). The derivative of a Gaussian process with covariance function $\kappa(t, t')$ (if it exists, as is the case for (10)) is also a Gaussian process and has covariance function

$$(11) \qquad \frac{\partial^2}{\partial t \partial t'} \kappa(t, t').$$

Furthermore, $\partial \kappa(t, t')/\partial t'$ 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).

Let $u(t)$ and $q_j(t)$, for $j = 1, \ldots, L$, denote the values of the RV corruption and $l$ generic stellar activity indicators, respectively, at time $t$. Given observation times $t_1, \ldots, t_{n_A}$, we propose the following class of models:

$$(12) \qquad u(t_i) = m_0 + a_{01}X(t_i) + a_{02}\dot{X}(t_i) + a_{03}\ddot{X}(t_i) + a_{04}Z_0(t_i) + \xi_{0i},$$

$$(13) \qquad q_1(t_i) = m_1 + a_{11}X(t_i) + a_{12}\dot{X}(t_i) + a_{13}\ddot{X}(t_i) + a_{14}Z_1(t_i) + \xi_{1i},$$

$$\vdots$$

$$(14) \qquad q_L(t_i) = m_L + a_{L1}X(t_i) + a_{L2}\dot{X}(t_i) + a_{L3}\ddot{X}(t_i) + a_{L4}Z_L(t_i) + \xi_{Li},$$

for $i = 1, \ldots, n_A$, where the $\xi_{ji}$ are independent with

$$(15) \qquad \xi_{ji} \sim \mathcal{N}(0, \sigma_{ji}^2),$$

for $j = 0, \ldots, L$. Here, the $a_{jk}$ are unknown parameters to be inferred from the data, and $Z_0, \ldots, Z_L$ are independent zero mean GPs with the covariance function (10). The $\sigma_{ji}$ are treated as given measurement uncertainties because it is standard practice in astronomy for uncertainties to be supplied with all observations. In particular, the uncertainties associated with the underlying spectra are provided by the observation team (or, in the case of our simulations, they are determined by (8)) and the resulting $\sigma_{ji}$ in (15) are then computed using Monte Carlo simulation. 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 differ from $\phi$, the covariance parameters for $X$). We denote by $\Sigma$ the $(L+1)n_A \times (L+1)n_A$ covariance matrix implied by the model (12)–(15) and specify its form in Appendix D.

The purpose of the independent GP components $Z_0, \ldots, Z_L$ is to permit structured deviations from linear combinations of the physically-motivated terms $X(t)$, $\dot{X}(t)$ and $\ddot{X}(t)$. R15 suggested that their approach could perhaps be improved by adding in these independent GP components, and, indeed, when we tried modeling a variety of stellar activity indicators using (12)–(15), we found that this additional flexibility is sometimes helpful (despite our inclusion of $\ddot{X}(t)$).

We now specify the log-likelihood function of the model (12)–(15). For conciseness, we write the observation times as $\boldsymbol{t} = (t_1, \ldots, t_{n_A})^T$ and the time series data as $\boldsymbol{s} = (\boldsymbol{u}, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_L)$, where $\boldsymbol{u} = (u(t_1), \ldots, u(t_{n_A}))^T$ and $\boldsymbol{q}_j = (q_j(t_1), \ldots, q_j(t_{n_A}))^T$, for $j = 1, \ldots, L$. Denoting the parameters for the model (12)–(14) 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

$$(16) \qquad l_{\text{act}}(\theta_{\text{act}}; \boldsymbol{t}, \boldsymbol{s}) = -\frac{(L+1)n_A \log(2\pi) + |\Sigma| + (\boldsymbol{s} - \boldsymbol{m})^T \Sigma^{-1}(\boldsymbol{s} - \boldsymbol{m})}{2},$$

where $\boldsymbol{m} = (m_0 \mathbf{1}_n^T, m_1 \mathbf{1}_n^T, \ldots, m_L \mathbf{1}_n^T)$ and $\mathbf{1}_n$ denotes a vector of $n$ ones. Equation (16) gives the general form of the likelihood, and the model index $l$, which appears in (3), simply corresponds to different choices of $\Sigma$.

Our model capturing both a planet and stellar activity is simply (12)–(14), except that $u(t)$ in (12) is replaced by

$$(17) \qquad u_p(t) = u(t) + v(t) - \gamma,$$

where $v(t)$ is given by (9) and the offset $\gamma$ is subtracted out because $m_0$ (in (12)) 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 $\theta_{\text{full}} =$

$(\theta_{\text{act}}, \Gamma)$, where $\Gamma = (K, M_0, \tau_p, \omega, e)$ are the parameters describing the planet and its orbit. In Algorithms 2 and 3 we use maximum likelihood to fit all the parameters in the stellar activity only and full models.

5.3. *Rajpaul et al.* (2015) *model as a special case.* The model proposed by R15 for the RV corruption and the stellar activity indicators $\log R'_{\text{HK}}$ and BIS is a special case contained within our model class (12)–(15). Setting $q_1$ and $q_2$ to be the indicators $\log R'_{\text{HK}}$ and BIS, respectively, their model is

$$u(t_i) = m_0 + a_{01} X(t_i) + a_{02} \dot{X}(t_i) + \xi_{0i}, \tag{18}$$

$$q_1(t_i) = m_1 + a_{11} X(t_i) + \xi_{1i}, \tag{19}$$

$$q_2(t_i) = m_2 + a_{21} X(t_i) + a_{22} \dot{X}(t_i) + \xi_{2i}. \tag{20}$$

A minor difference from the model they specify is that we assume $\xi_{ji} \sim N(0, \sigma_{ij}^2)$, for $j = 0, 1, 2$ and $i = 1, \ldots, n_A$, where the $\sigma_{ij}$ are the individual observation uncertainties recorded. In contrast, R15 set $\sigma_{ij} = \sigma_j$ and estimate $\sigma_j$, for $j = 0, 1, 2$. We use the recorded uncertainties $\sigma_{ij}$ for all model fits in this paper (providing measurement uncertainties is standard practice in astronomy).

We consider the approach by Rajpaul et al. (2015) to be representative of the current 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 stellar activity modeling performance identified in a recent comparison of methods by Dumusque et al. (2017). 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, we will use the R15 model as a point of reference in evaluating our results in Section 7.

**6. Power-based model selection.** If we set the stellar activity model to be (12)–(15) allowing all the coefficients to be nonzero, then the power of the resulting likelihood ratio test (LRT) for detecting a planet is suboptimal for realistic planets, because the stellar activity model is too flexible and absorbs planet signals. That is, the test statistic $\Delta = l_{\text{act}}(\hat{\theta}_{\text{act}}) - l_{\text{full}}(\hat{\theta}_{\text{full}})$ tends to be small even when a planet is present. Thus, we cannot allow all the coefficients in (12)–(14) to be nonzero but must try to find the best submodel for the specific stellar activity indicators at hand.

The ideal way to choose between the models in the class (12)–(15) is to apply Algorithm 2 in Section 2.3 which selects the best model by comparing the power for planet detection under each candidate model. Unfortunately, since each coefficient $a_{jk}$ in (12)–(14) can be included or not, there are $2^{12} = 4096$ models, and the computational cost of evaluating the power for all of them is prohibitively high. Some of the models can be ruled out straightaway, but there are still too many to apply Algorithm 2 directly. Instead, we apply Algorithm 3 in Section 2.4, which first creates a short-list of models using the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and cross validation (CV), and then applies Algorithm 2 to this short-list.

We now give additional details regarding Algorithm 3. 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 nontrivial model is required for each time series. We also fix $a_{04} = 0$ because including an independent GP for modeling the RV corruption will intuitively lead to low power; the model would be able to capture RV signals from planets without the need for a planet model. This leaves $(2^3 - 1)(2^4 - 1)^2 = 1575$ possible stellar activity models from

which we create a short list based on the ability of the models to capture the stellar activity indicators $u(t_i)$ and $q_{\text{PC}j}(t_i)$, for $j = 1, 2$ and $i = 1, \ldots, n$, plotted in the bottom row of Figure 2.

In Steps 3 and 4 of Algorithm 3, we screen the models and create the short list by computing the criteria $\overline{\Delta\text{AIC}}$, $\overline{\Delta\text{BIC}}$ and $\overline{\Delta\text{CV}}$ which average across 10 screening datasets and are specified by (5). The specifics of our CV procedure are given in Appendix E. For the sake of illustration, we include the best five models under each criterion in the short list.

Step 5 of Algorithm 3 applies Algorithm 2 to assess the planet detection power under the short-listed models. For each model, Step 2 of Algorithm 2 approximates the null distribution of the LRT statistic $\Delta$ by computing it for the 1000 replicate datasets stored in $Y_N$ (recall that $Y_A = (Y_N, Y_P)$). Details of the optimization approach used are given in Appendix F. Step 3(ii) of Algorithm 2 injects planet signals into the 800 replicate datasets stored in $Y_P$. In practice, the specific null and planet-injected datasets that should be used in the power approximations will depend on the type of activity expected, the instrument in question, our understanding of what signals can realistically be detected and the type of planets we are hunting. In this paper we inject planets with an orbital period of $\tau_p = 7$ days and RV signal amplitudes close to or below the current detection threshold of $\sim 1$ m/s. Note that it is principally the relative values of $\tau_s$ and $\tau_p$ that impact detection power, rather than their absolute values. Indeed, we found that, for most values of $(\tau_s, \tau_p)$, the detection power is similar to that under our setting $(\tau_s, \tau_p) = (10, 7)$, with the exception being $\tau_s \approx \tau_p$ (or cases where major harmonics of the periods overlap), which leads to lower power. This insensitivity to $(\tau_s, \tau_p)$ is why it suffices to use only one setting in the model selection procedure, even though the values for any specific dataset will be different. To inject a planet signal, we simply add $v(t_i; \Gamma)$ (given by (9)) to $u(t_i)$, for $i = 1, \ldots, n_A$, and remove the superfluous offset term; see (17). The parameter vectors of the injected signals are given by $\Gamma = (K, M_0 = 1.5, \tau_p = 7, \omega = 1, e = 0.2, \gamma = 0)$ for amplitudes $K = 0.01, 0.02, \ldots, 0.16$ m/s. Planet signals corresponding to each setting of $\Gamma$ are injected into 50 replicate datasets. 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.

In Step 3(iii) of Algorithm 2, for each model $l \in \mathcal{L}$, we compute $\Delta$ for each of the 800 planet-injected datasets. For each dataset we reject the null of no planet if $\Delta$ is greater than the 0.99 quantile (denoted $c_l$ in Algorithm 2) of the relevant null distribution, that is, that constructed using the stellar activity model $l$. The number of rejections across the 50 simulations at a given value of $K$ provides an estimate $p(l, K)$ of the detection power at that planet signal amplitude under model $l$; see (4).

Step 4 of Algorithm 2 selects the final model $\mathcal{M}_{\text{MS-G2V}}$ to be that with the smallest *detection threshold*, which we define to be the minimum amplitude of a planet RV signal $K$ such that $p(l, K) \geq 0.5$. If no activity model dominated the others in terms of detection power for the simulated planet configurations, we would choose the final model according to additional considerations, for example, the best overall criterion rank (i.e., the minimum value of the sum of the ranks under the three selection criteria), the highest detection power for the planets of most interest or the planets that are deemed most probable to accompany the type of star in question.

## 7. Application of model selection procedure.

### 7.1. *Preliminary model.*
The R15 model (18)–(20) was not designed to capture our PCA based stellar activity indicators because our indicators differ to $\log R'_{\text{HK}}$ and BIS. However, the plot of $\{u(t_i), \log R'_{\text{HK}}(t_i), \text{BIS}(t_i)\}_{i=1,\ldots,n}$ in Figure 3 of R15 qualitatively resembles the top row of Figure 2 here, which shows $\{u(t_i), q_{\text{PC}1}(t_i), q_{\text{PC}2}(t_i)\}_{i=1,\ldots,n}$ for the noiseless
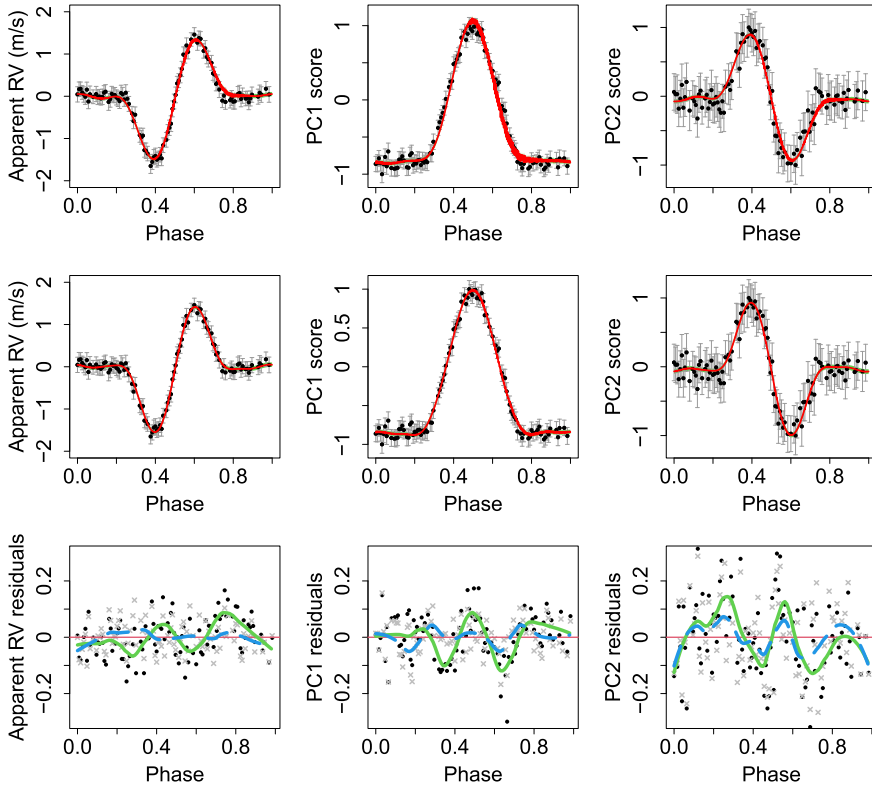
FIG. 3. *Preliminary model fit to $\{u(t_i), q_{PC1}(t_i), q_{PC2}(t_i)\}_{i=1,...,n}$ (top row) and AIC-1 model fit (middle row). Note that during fitting all the signals were normalized for numerical stability but $u(t)$ is plotted on the original m/s scale for interpretability. The bottom row shows the residuals for the preliminary model fit (points), the solid green curve is a LOESS (local polynomial regression) fit to the plotted residuals and the dashed blue line is a LOESS fit to the AIC-1 model residuals (grey crosses).*

dataset $Y_s$. To investigate further, in the top row of Figure 3 we plot the maximum likelihood fit of the R15 model to our PCA-based stellar activity indicators shown in the bottom row of Figure 2. For visualization purposes the $x$-axis is stellar rotation phase rather than time, though the stellar rotation period was treated as unknown when fitting the model. A lighter region representing a 95% confidence interval for the underlying functions is plotted but is mostly covered by the fitted function (bold line), due to very small uncertainties. The reason for the small uncertainties is that $X(t)$ (and hence $\dot{X}(t)$) can be precisely inferred by combining information across the three time series. We will refer to the R15 model as the *preliminary model* since it is our best initial guess of a suitable model for our indicators based on the existing literature.

Despite the high precision the preliminary model fit is unsatisfactory because the inferred underlying function shows systematic deviations from the observations. This can be seen from the bottom row of Figure 3, which shows the residuals for the preliminary model fit, with the solid green lines being LOESS (local linear regression) fits to the preliminary model residuals (black dots). For instance, consider the middle panel of the bottom row which gives the $q_{PC1}$ residuals: we can see oscillations, with the residuals being positive near phase 0.5, and negative either side of 0.5. Furthermore, the root mean squared residuals for the three panels are 0.067, 0.081, 0.141, respectively. These are all larger than the corresponding values under the AIC-1 model, identified in Section 7.2 below, which are 0.053, 0.062, 0.124, respectively. The dashed blue lines in the bottom row show LOESS fits to the AIC-1 model residuals (grey crosses), and the middle row shows the fit of the AIC-1 model to the data.

TABLE 1
*Summary of results for stage one of our model selection procedure presented in Section 6. The preliminary model and the top five models under the $\overline{\Delta AIC}$, $\overline{\Delta BIC}$, and $\overline{\Delta CV}$ criteria are listed. Repeated models are omitted*

| Model | No. paras | $\overline{dev.}$ | $\overline{\Delta AIC}$ rank | $\overline{\Delta BIC}$ rank | $\overline{\Delta CV}$ rank | $\overline{\Delta AIC}$ | $\overline{\Delta BIC}$ | $\overline{\Delta CV}$ |
|---|---|---|---|---|---|---|---|---|
| Saturated | 21 | 0 | 48 | 217 | 137 | 9.2 | 29.5 | 0.46 |
| Preliminary | 11 | 70 | 1008 | 933 | 788 | 58.9 | 59.6 | 1.44 |
| AIC-1 | 11 | 11 | 1 | 2 | 1 | 0.0 | 0.7 | 0.00 |
| AIC-2 | 12 | 10 | 2 | 5 | 2 | 1.3 | 4.5 | 0.06 |
| AIC-3 | 11 | 13 | 3 | 3 | 5 | 1.7 | 2.4 | 0.11 |
| AIC-4 | 10 | 15 | 4 | 1 | 3 | 1.9 | 0.0 | 0.07 |
| AIC-5 | 13 | 10 | 5 | 8 | 7 | 2.5 | 8.4 | 0.13 |
| BIC-4 | 11 | 15 | 6 | 4 | 8 | 3.3 | 4.0 | 0.13 |
| CV-4 | 15 | 12 | 37 | 54 | 4 | 8.3 | 19.4 | 0.11 |

7.2. *Results of applying Algorithm* 3. Table 1 summarizes the short list of models obtained in Step 4 of Algorithm 3 and also the saturated and preliminary models. The saturated model is that where all the coefficients $a_{jk}$ in (12)–(14) are allowed to be nonzero. The second and third columns list the number of parameters in the shortlisted model and the average deviance across the 10 screening datasets input into Algorithm 3, respectively. Columns four through six list the ranking of the models by the three selection criteria. Columns seven through nine give the criteria values, $\overline{\Delta AIC}$, $\overline{\Delta BIC}$ and $\overline{\Delta CV}$, which are given by (5). Smaller values are preferable for all quantities listed in the table. Most of the top five $\overline{\Delta BIC}$ and $\overline{\Delta CV}$ ranked models are the same as the top $\overline{\Delta AIC}$ models, and we have only listed each unique model once. For brevity, we denoted the top $\overline{\Delta AIC}$ ranked model by AIC-1 and use similar notation for other ranked models.

The preliminary model performs poorly under all three of the criteria in Table 1, as might be expected from the discussion in Section 7.1 and the residuals shown in the bottom row of Figure 3. Table 2 gives the maximum likelihood estimates (MLEs) of the model coefficients for AIC-1. The MLEs of the kernel parameters are $\log \hat{\tau}_s = 2.30$, $\log \hat{\lambda}_p = -0.95$ and $\log \hat{\lambda}_e = 9.07$. These kernel parameter values are consistent with a star with a single spot and a 10-day rotation period. The BIC-1 model (AIC-4 and CV-3) has nonzero coefficients $a_{02}$, $a_{11}$, $a_{13}$ and $a_{22}$ and turns out to be contained in all seven models listed in the bottom section of Table 1; that is, these coefficients are nonzero in these seven models. This can be seen from the left panel of Figure 4 which shows the relative frequencies of the model coefficients across the models listed in the bottom section of Table 1. Indeed, the left panel of Figure 4 makes it clear that the short-listed models differ from the BIC-1 model only by the addition of one or two nonzero coefficients.

TABLE 2
*Maximum likelihood estimates of the AIC-1 stellar activity model coefficients for the screening dataset with the largest spot. 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*

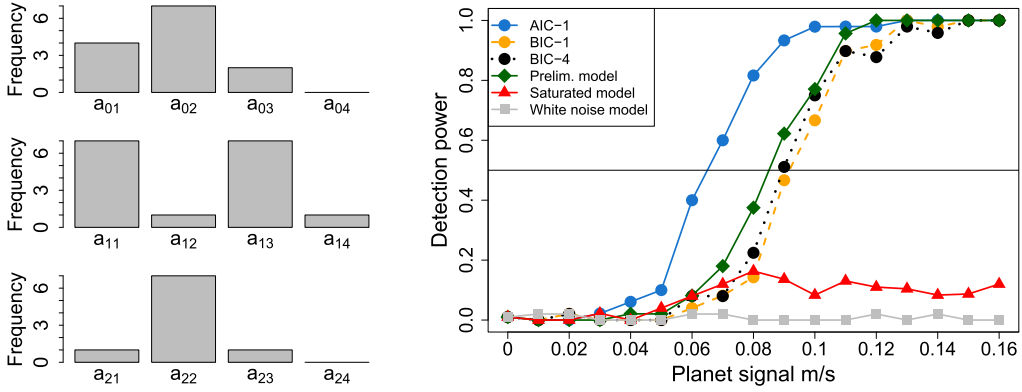| | $X$ coeff ($a_{j1}$) | $\dot{X}$ coeff ($a_{j2}$) | $\ddot{X}$ coeff ($a_{j3}$) | $Z_j$ coeff ($a_{j4}$) |
|---|---|---|---|---|
| $u(t)$ (m/s) | $-0.01$ ($-0.02$) | $-0.31$ ($-0.48$) | | |
| $q_{PC1}(t)$ | 0.50 | | 0.08 | |
| $q_{PC2}(t)$ | | 0.31 | | |

FIG. 4. (*Left*) *parameter frequencies across the top ranked* $\overline{\Delta\text{AIC}}$ / $\overline{\Delta\text{BIC}}$ / $\overline{\Delta\text{CV}}$ *models given in Table* 1. (*Right*) *planet detection power for a selection of the models summarized in Table* 1 (*left*). *The other models in Table* 1 *perform similarly to the AIC-1 model.*

Next, the right panel of Figure 4 shows a selection of the results of Step 5 of Algorithm 3, that is, it shows the power for detecting planets under different models in Table 1 at a range of planet signal amplitudes. We set the Type I error rate of the LRT to 0.01, and thus, for planet RV signals of 0 m/s, the detection power is 0.01. The steep slopes of the power curves are a typical feature in planet detection studies because planets and stellar activity affect observed spectra quite differently, and, consequently, small increases in the planet RV signal amplitude can make it significantly easier to detect the planet. In the plot we see the AIC-1 model (blue circles, solid line) has a detection threshold of between 0.06 m/s and 0.07 m/s (i.e., the amplitude at which 0.5 detection power is achieved) and has detection power close to 1 for planet signal amplitudes $\geq 0.1$ m/s. Four of the other top ranked models in Table 1 had almost identical performance and are not plotted. The BIC-1 (AIC-4 and CV-3) model has lower power (orange circles, dashed line) because it is too simple, and, in particular, it does not include the coefficient $a_{01}$, which is small when fitted, but still impacts the deviance and detection power. The BIC-4 model (black circles, dotted line) is the same, except it includes the coefficient $a_{21}$, which seems to be superfluous in that it does not lower the average deviance noticeably, and, indeed, BIC-4 performs similarly to BIC-1. In summary, several models perform almost identically well, but we set $\mathcal{M}_{\text{MS-G2V}}$ to be the AIC-1 model because it has the best overall rank (the sum of the three selection criteria ranks). Also, note that the AIC-1 model is among the top five models shortlisted under each of the three individual criteria considered in Table 1 (see the AIC-1 row) which suggests that our approach is somewhat robust to the specific screening criterion used.

The AIC-1 model selected is quite similar to the preliminary model, since it only adds $a_{13}$ and drops $a_{21}$. This similarity is unsurprising because the preliminary or R15 model (for similar proxies) was well motivated by physical insights, given by Aigrain, Pont and Zucker (2012), based on the area of a spot and its derivative; see Section 5.2. Nonetheless, the right panel of Figure 4 shows that the preliminary model (green diamonds) performs substantially worse than the AIC-1 model. This is likely because the coefficient $a_{13}$ is needed to correct the systematic derivations discussed in Section 7.1, while the flexibility offered by $a_{21}$ is not essential and makes separating stellar activity and exoplanet signals harder. On the other hand, the preliminary model performs slightly better than the BIC-1 and BIC-4 models. The saturated model (red triangles) is overly flexible and, therefore, absorbs planet signals which results in low power across the range of planet signal amplitudes considered. The white noise model (solid 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, it results in almost no power for the small planet signal amplitudes considered in Figure 4.

In addition to detecting planets, it is of interest to infer their properties, and Appendix G summarizes the 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 in preliminary studies we found that the power is lower for planets with orbital periods that coincide with the stellar rotation period or its harmonics.

7.3. *Comparison to Rajpaul et al. (2015).* That the preliminary model performs poorly compared to the AIC-1 model (see Figure 4) is unsurprising because the model was originally designed by R15 to capture the evolution of the RV corruption, normalized flux (or $\log R'_{HK}$), and BIS, rather than the evolution of our indicators. Therefore, we repeat our detection power analysis for the R15 model but use 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 were computed during the SOAP 2.0 runs used to create the $Y_A$ datasets, and we add noise consistent with that of the $Y_A$ datasets. Specifically, for normalized flux we computed the noise levels based on a Poisson model for variations in the number of photons detected. For BIS we assumed a constant standard deviation of 0.0198 m/s, where the specific value was motivated by scaling actual BIS error measurements in Dumusque et al. (2012) to account for the increased spectral resolution and SNR of our input data.

The red "+" symbols in the left panel of Figure 5 show the detection power under the R15 model when using the normalized flux and BIS activity indicators described above (for a planet with a seven-day orbital period). For comparison, the detection power under the AIC-1 (blue circles) is replotted from Figure 4). The R15 model has a detection threshold of about 0.07 m/s and, therefore, performs almost as well as the AIC-1 model and better than the preliminary model (see Figure 4). However, the dark red "×" symbols show that the detection power of the R15 model is substantially reduced when detecting planets in the presence of an 866 micro hemispheres (MSH) stellar spot which is the 95% quantile of the spot size distribution (7) used in our main simulations. In other words, in the case of particularly active MS-G2V stars, the R15 model has reduced performance. In contrast, the dark blue triangles
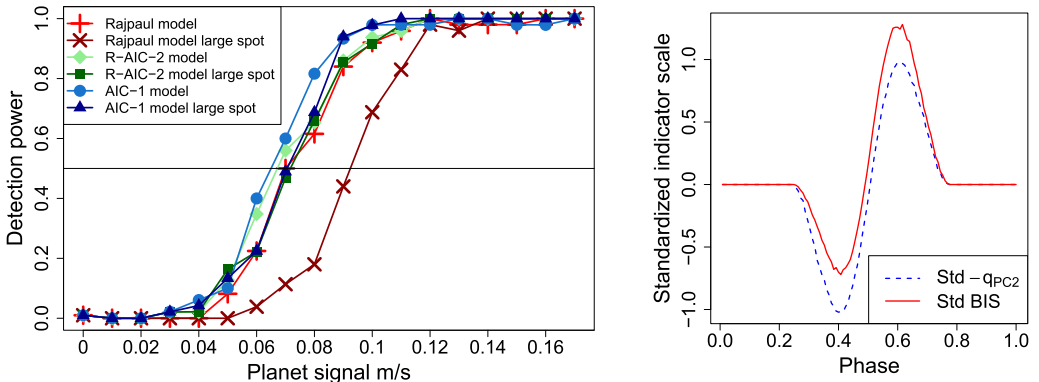


FIG. 5. *The left panel shows the detection power under the R15 model, using their normalized flux and BIS indicators (red "+" symbols), and again in the case where the datasets with planets all have a large spot (dark red "×" symbols). Also shown is the detection power for the R-AIC-2 model (green diamonds) and again in the large spot case (dark green squares). For comparison, the AIC-1 model results (with our proxies) is reproduced from Figure 4 (blue circles), and the AIC-1 model detection power in the case of a large spot is plotted (dark blue triangles). The right panel compares the $q_{PC2}$ and BIS indicators computed from the noiseless data $Y_s$. The indicators are standardized (centered and normalized) for the purpose of the comparison.*

show that the detection power of the AIC-1 model is almost unaffected by the presence of such a large spot.

It is unclear if the superiority of the AIC-1 model in the case of large spots is due to our PCA based indicators or our model selection procedure or both. To investigate this, we applied our model selection procedure in the case of the normalized flux and BIS indicators. We again found the five best $\overline{\Delta AIC}$, $\overline{\Delta BIC}$ and $\overline{\Delta CV}$ ranked models. The top performing model in terms of detection threshold was that ranked second by $\overline{\Delta AIC}$, which we denote by R-AIC-2. Figure 4 shows the detection power of the R-AIC-2 model (green diamonds) and its detection power in the case of a large spot (dark green squares). The R-AIC-2 model performs similarly to the AIC-1 model in both cases and substantially better than the R15 model in the case of large spots. The R-AIC-2 model and the maximum likelihood estimates of its coefficients are given in Table 3 in Appendix H.

In conclusion, for the simple case of a single constant spot it is principally our model selection procedure rather than our PCA-based stellar activity indicators that offers improved detection power, and this improvement is mainly obtained in the case of large spots, as might be expected. Therefore, for the time being, astronomers may prefer to keep using normalized flux (or $\log R'_{HK}$) and BIS as indicators, but the model should be updated to the R-AIC-2 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 the 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 five fewer parameters than the R-AIC-2 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. The right panel of Figure 5 compares $-q_{PC2}$ and BIS (after centering) for the noiseless data $Y$. The asymmetry suggests that scaling blueshifts and redshifts differently 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 potentially also allow custom indicators to be used for each star or type of star, and this is discussed further in Section 8.2.

## 8. Discussion.

8.1. *Scope of our framework.*    Our general procedures apply broadly and are not limited to the specific indicator and model choices discussed here, as we now explain. First, in Algorithm 1, PCA can be straightforwardly replaced by another dimension reduction technique. We are currently exploring the use of diffusion maps (e.g., Coifman et al. (2005), Coifman and Lafon (2006)) which do not have the constraint of projecting onto a linear subspace as does PCA. Capturing nonlinearity could be particularly helpful for more complex situations, such as when stellar spots evolve over time. Our initial investigations also suggest that capturing nonlinear structure may help to improve planet detection power when there is aliasing between the stellar rotation period and the planet orbital period.

Second, our automated model selection procedure allows investigators to easily find the best model for new stellar variability indicators and to compare the relative performance

of competing indicators. Lastly, although we have assumed a single spot of constant (but unknown) size, our ongoing work suggests 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 (12)–(15) for modeling complex stellar activity is a topic of future work.

8.2. *Generalizing to other star types.*   One possible solution is to apply Algorithm 1 to every dataset and, thereby, obtain star-specific indicators from the noisy spectra observed. We now briefly investigate the potential of this approach.

The thin blue line in the upper part of the left panel of Figure 6 shows a section of the principal component corresponding to $q_{PC1}$ when Algorithm 1 is applied directly to a noisy dataset containing a 661 MSH spot (fairly large under the spot size distribution (7)). For comparison, the think black line in the upper part of the left panel shows the corresponding principal component contained in $\mathcal{I}_{MS\text{-}G2V}$ (computed from noiseless data). The bottom part of the plot shows the corresponding section of the spectrum (green). It is clear that, in the case of the noisy dataset, the principal component is dominated by noise, and, indeed, we found it almost useless for planet detection. Here, we instead focus on the case of a very large 10k MSH spot, that is, in both null and test simulations in Algorithm 2 we fix the spot size at 10k MSH, rather than simulating it from (7). The thin orange line in the upper part of the left panel of Figure 6 shows that, in this case, the first indicator in $\mathcal{I}_{MS\text{-}G2V}$ can be reasonably approximated from one of the noisy datasets. The right panel shows the resulting detection power under the AIC-1 model (see Table 2), for this 10k MSH spot case, when recomputing the indicator vectors for each dataset by applying Algorithm 1 (red diamonds). In this large spot scenario, although the AIC-1 model was found to still be the best model if recomputing the indicator vectors for each dataset, this was not the case when using the fixed indicators defined by $\mathcal{I}_{MS\text{-}G2V}$. Thus, for comparison with the fixed indicator case, we plot the new top ranked $\overline{\Delta AIC}$ model, denoted 10k-AIC-1 (blue circles) and given in Appendix I.[3] In
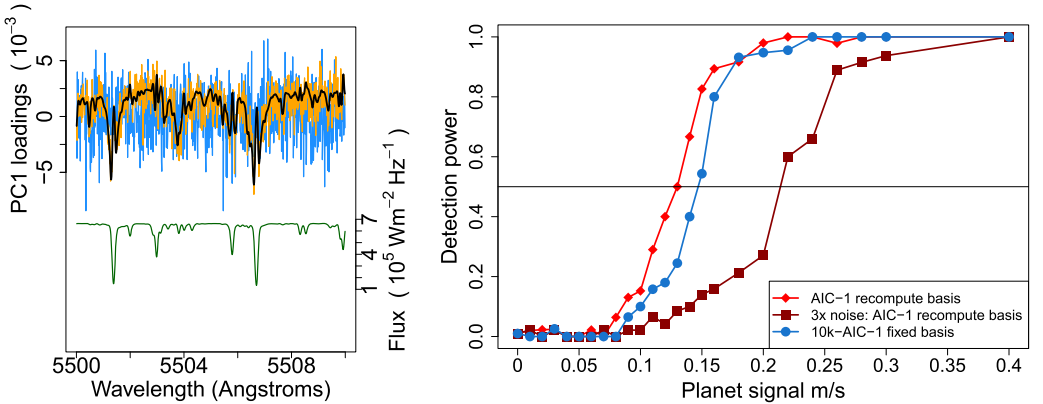


FIG. 6.   (*Left*) *PC*1 *loadings* (*for a small portion of the input wavelength range*) *obtained by applying Algorithm* 1 *to a noisy dataset with a* 661 *MSH spot* (*blue line, upper half*) *and a* 10k *MSH spot* (*orange green line, upper half*). *For comparison, the corresponding loadings from* $\mathcal{I}_{MS\text{-}G2V}$ *are also shown* (*thick black line, upper half*). *In the bottom half of the panel, a small segment of the quiet spectrum is plotted* (*green*). (*Right*) *planet detection power under the AIC-1 model when the activity indicator vectors are recomputed for each noisy dataset* (*red diamonds*), *in the case where all null and test simulations use a* 10k *MSH spot. For comparison, we also plot the detection power when the noise standard deviation is inflated by a factor of three* (*dark red squares*), *that is,* SNR = 167, *and when we use the fixed indicators* $\mathcal{I}_{MS\text{-}G2V}$ *with the* 10k-AIC-1 *model* (*blue circles*).

[3]Interestingly, in our preliminary studies with lower SNR the AIC-1 model was still the best model in the large spot case, suggesting that SNR is an important factor in the model selection.

conclusion, there appears to be no loss of power due to recomputing the indicators vectors, in the case of a 10k MSH spot, and, in fact, there is a small gain in power. Of course, even for large spots, if the spectra are sufficiently noisy, then we expect that the detection power will decrease, especially when recomputing the indicators for each dataset. Confirming this, the dark red squares show the detection power when the noise is inflated by a factor of three (i.e., SNR = 167), and we recompute the indicators for each dataset.

The left panel of Figure 6 suggests that it may be fruitful to apply a penalization or smoothing approach when deriving stellar activity indicators from noisy datasets. It also suggests that observing stars during high-activity periods could be useful for deriving reliable indicators, whereas, in the past, observing currently highly active stars has been intentionally avoided (in exoplanet surveys).

The second key challenge is that, for star types other than MS-G2V, the distribution of spot related activity is less well understood and may not be well approximated by the distribution for the Sun (as is done in Algorithm 2). A potential solution is to use a hierarchical Bayesian structure to jointly model stellar activity from multiple stars with similar spectroscopic properties. This would allow us to infer population-level stellar activity parameters and a corresponding model that could be used for generating replicate datasets from such stars. The population inference could also be used to infer indicator vectors (e.g., $\mathcal{I}_{\text{MS-G2V}}$), thereby avoiding application of Doppler-constrained PCA directly to the noisy spectra of individual stars.

## APPENDIX A: DOPPLER COMPONENT FOR ALGORITHM 1

Let the star light intensity at rest frame wavelength $\lambda^{(0)}$, when normalized by the total light intensity across wavelengths, be denoted by $f(\lambda^{(0)})$ so that the function $f(\cdot)$ gives the rest frame normalized stellar spectrum (at some fixed time). For a source moving with a radial velocity of $v = cz$, the Doppler shift is $z$, where $c$ is the speed of light in m/s, that is, the observed intensity at observed wavelength $\lambda$ is given by $f(\lambda/(1+z))$. Thus, a Taylor expansion tells us that the observed intensity at observed wavelength $\lambda$ will approximately be given by $f(\lambda) - z\lambda f'(\lambda)$, where $f' = df/d\lambda$. Since the relevant Doppler shifts are typically very small ($z \approx 10^{-8}$), this Taylor approximation is very accurate, and we can, therefore, represent Doppler shifts as scalar multiples of the vector $(\lambda_1 f'(\lambda_1), \ldots, \lambda_p f'(\lambda_p))^T$, where $\lambda_i$ is the recorded wavelength, for $i = 1, \ldots, p$.

A complication is that $f'(\cdot)$ is not known exactly, and the derivative of observed stellar spectrum $f_t'(\cdot)$ changes over time, due to the presence of stellar activity, any planets and noise; see (1). However, fortunately, the temporal changes to $f_t'(\cdot)$ can be regarded as second-order, so it is reasonable to treat $\boldsymbol{w}$ as fixed across time. Here, we compute the mean observed spectrum across the different observation times, denoted $\bar{f}(\cdot)$, and set $\boldsymbol{w} = (\lambda_1 \bar{f}'(\lambda_1), \ldots, \lambda_p \bar{f}'(\lambda_p))^T$. We approximate the derivative $\bar{f}'$ by finite differencing which is sufficiently accurate because the $p = 237{,}944$ wavelengths are recorded on a very fine grid.

## APPENDIX B: DETAILS OF MAXIMIN LATIN HYPERCUBE DESIGN USED IN ALGORITHM 3

We computed a minimax Latin hypercube design on the square $[0, 1]^2$ using the `maximinESE_LHS` function in the R package `DiceDesign` by Dupuy et al. (2015). A minimax Latin hypercube design (Morris and Mitchell (1995)) maximizes the minimum distance between the design points while preserving the Latin hypercube property that none of the design points have any co-ordinates in common. Given the design on $[0, 1]^2$, we applied the quantile function of the spot latitude and size distributions, given by (6) and (7)

to the $x$ and $y$ coordinates of the design points, respectively, and thereby obtained a design representative of the two-dimensional spot latitude-size distribution. This minimax Latin hypercube based design has a computational advantage over random sampling from (6) and (7) because the latter approach would require more datasets (since some of them would have similar sizes and latitudes).

## APPENDIX C: KEPLER PLANET MODEL DETAILS

The true anomaly function $\phi(t)$ in (9) is given by solving the following system of equations:

$$\tan \frac{\phi(t)}{2} = \left(\frac{1+e}{1-e}\right)^{\frac{1}{2}} \tan \frac{E(t)}{2},$$
(21)

$$E(t) - e \sin E(t) = M(t),$$
(22)

$$M(t) = \frac{2\pi t}{\tau_p} + M_0,$$
(23)

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.

## APPENDIX D: COVARIANCE MATRIX CALCULATION

Here, we specify the covariance matrix $\Sigma$, implied by the model (12)–(14) and the covariance function (10). Let $A^{(a,b)}$ be the matrix whose $(i, i')$ entry is $\text{Cov}(\frac{d^a}{dt_i^a} X(t_i), \frac{d^b}{dt_{i'}^b} X(t_{i'}))$, for $a, b = 0, 1, 2$. Then, for model parameters $a_{jk}$, for $j = 0, \ldots, L$ and $k = 1, 2, 3, 4$, the $n_A \times n_A$ diagonal block of $\Sigma$ corresponding to output $j$ (i.e., the square block with rows and columns $jn_A + 1, \ldots, (j+1)n_A$) is given by

$$\sum_{k_1=1}^{3} \sum_{k_2=1}^{3} a_{jk_1} a_{jk_2} A^{(k_1-1,k_2-1)} + a_4^2 B^{(0,0)} + D_j,$$
(24)

for $j = 0, \ldots, L$, where $D_j$ is a diagonal matrix with diagonal elements are $\sigma_{j1}, \ldots, \sigma_{jn_A}$, and the $(i, i')$ entry of $B^{(0,0)}$ is $\text{Cov}(Z_0(t_i), Z_0(t_{i'}))$ (recall $Z_0, \ldots, Z_L$ all have the same covariance function parameters $\phi_Z$). Similarly, the off-diagonal $n_A \times n_A$ block corresponding to the covariance between outputs $j_1$ and $j_2$ (i.e., the square block with rows $j_1 n_A + 1, \ldots, (j_1 + 1)n_A$ and columns $j_2 n_A + 1, \ldots, (j_2 + 1)n_A$) is given by

$$\sum_{k_1=1}^{3} \sum_{k_2=1}^{3} a_{j_1 k_1} a_{j_2 k_2} A^{(k_1-1,k_2-1)}.$$
(25)

Thus, all that remains is to specify $A_{ii'}^{(a,b)}$, for $a, b = 0, 1, 2$. The term $A_{ii'}^{(0,0)}$ is (10) with $t = t_i$ and $t' = t_{i'}$, and the remaining terms are given by

$$A_{ii'}^{(0,1)} = -A_{ii'}^{(1,0)} = T_1 A_{ii'}^{(0,0)},$$
(26)

$$A_{ii'}^{(1,1)} = -T_1 A_{ii'}^{(0,1)} + T_2 A_{ii'}^{(0,0)},$$
(27)

$$A_{ii'}^{(0,2)} = A_{ii'}^{(2,0)} = -A_{ii'}^{(1,1)},$$
(28)

$$A_{ii'}^{(1,2)} = -A_{ii'}^{(2,1)} = T_1 A_{ii'}^{(1,1)} + 2T_2 A_{ii'}^{(0,1)} + 2T_3 A_{ii'}^{(0,0)},$$
(29)

$$A_{ii'}^{(2,2)} = -T_1 A_{ii'}^{(1,2)} + 3T_2 A_{ii'}^{(0,2)} - 6T_3 A_{ii'}^{(0,1)} + 4T_4 A_{ii'}^{(0,0)},$$
(30)

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

$$(31) \qquad T_1 = \frac{\pi \sin(\lambda_{ij})}{2\tau_s \lambda_p^2} + \frac{t_i - t_j}{\lambda_e^2},$$

$$(32) \qquad T_2 = \frac{\pi^2 \cos(\lambda_{ij})}{\tau_s^2 \lambda_p^2} + \frac{1}{\lambda_e^2},$$

$$(33) \qquad T_3 = \frac{\pi^3 \sin(\lambda_{ij})}{\tau_s^3 \lambda_p^2},$$

$$(34) \qquad T_4 = \frac{\pi^4 \cos(\lambda_{ij})}{\tau_s^4 \lambda_p^2}.$$

## APPENDIX E: CROSS-VALIDATION DETAILS

Here, we describe the CV approach used in Algorithm 3; see Section 2.4 and Section 6. We first investigated standard leave-one-out cross-validation but found that the models favored by this approach 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 an apparent RV signal that is due to stellar activity. Indeed, the latter case is more similar to predicting the RV corruption for *all* the 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

$$(35) \qquad \frac{1}{2}\big((s_B - \hat{\mu})^T \hat{V}^{-1}(s_B - \hat{\mu}) - \log|\hat{V}|\big),$$

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 nonparametric 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 $B = 10$ times for each model. The final cross-validation score for a given model is

$$(36) \qquad \text{CV} = \frac{1}{B}\sum_{k=1}^{B} \text{CV}_k,$$

where $\text{CV}_k$ denotes the value of (35) for the $k$th repetition of the procedure, for $k = 1, \ldots, B$. To ensure a fairer comparison of the models, we reuse the same 10 test blocks for all models.

## APPENDIX F: DETAILS OF OPTIMIZATION PROCEDURE

Optimization of the parameters of the models in the class (12)–(14) 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. First, 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 $\Gamma$. 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 (16) converged.

Second, rather than directly optimizing the full log-likelihood, we first optimized the log-likelihood for stellar indicator $q_{PC1}$ (again using parameter blocks, as described above), and then the log-likelihood for $q_{PC1}$ and $q_{PC2}$ 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.

Third, 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 (10) 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 rewrote (9) 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.

## APPENDIX G: 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 orbital period $\tau_p$ and velocity amplitude $K$ under the AIC-1 stellar activity model for the simulations used in Figure 4. In Figure 7 each boxplot shows the estimates of $\tau_p = 7$ (right panel) or the relative errors in estimating $K$ (left panel) for a given simulation value of $K$. For larger values of $K$, we can see that both $\tau_p$ and $K$ are reasonably well inferred. Since the detection threshold for the AIC-1 model is about 0.065 m/s, the model may detect some planets even when it is unable to accurately infer their properties. This is to be expected because planet detection requires less information than inferring planet properties.

## APPENDIX H: PARAMETERS ESTIMATES FOR THE R-AIC-2 MODEL

In Section 7.3 we considered modeling the R15 activity indicators, as opposed to our PCA based indicators, and applied Algorithm 3. The R-AIC-2 model was the best model identified, and its nonzero coefficients and their maximum likelihood estimates are given in Table 3. A number of other models were found to yield very similar planet detection power.
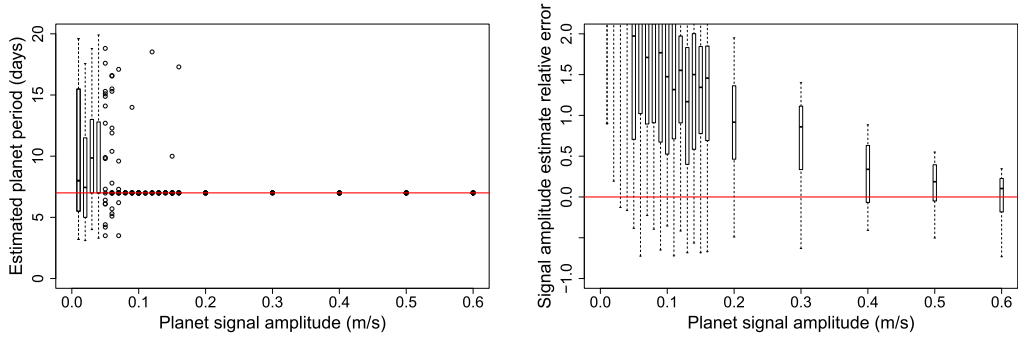
FIG. 7. *The orbital period MLE in days against the true planet RV amplitude (left); and the relative error of the MLE of the planet RV amplitude K against the true planet RV amplitude (right). In the left panel the red line shows the true orbital period of seven days, and in the right panel it indicates zero relative error.*

TABLE 3
*Maximum likelihood estimates of the R-AIC-2 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*

|  | $X$ coeff ($a_{j1}$) | $\dot{X}$ coeff ($a_{j2}$) | $\ddot{X}$ coeff ($a_{j3}$) | $Z_j$ coeff ($a_{j4}$) |
|---|---|---|---|---|
| $u(t)$ (m/s) |  |  | 0.37 (0.96) |  |
| Norm. flux | 0.05 | −0.29 |  | 0.14 |
| BIS |  | 0.12 | −0.32 | 0.17 |

Interestingly, the model uses the extra GP components $Z_1$ and $Z_2$. Investigation showed that, at least, $Z_2$ is needed in order to properly 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 5). Models that only use $X(t)$ and its first two derivatives fit the R15 indicators comparatively poorly, including the R15 model (see their Figure 3).

## APPENDIX I: PARAMETERS ESTIMATES FOR THE 10K-AIC-1 MODEL

Figure 6 in Section 8.2 plots the detection power for the 10k-AIC-1 model, in the case where all null and test datasets have a 10k MSH spot. This model was obtained by again running our model selection procedure, Algorithm 3, but where all the datasets had a 10k MSH spot. The MLEs of the model coefficients are given in Table 4.

TABLE 4
*Maximum likelihood estimates of the 10k-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*

|  | $X$ coeff ($a_{j1}$) | $\dot{X}$ coeff ($a_{j2}$) | $\ddot{X}$ coeff ($a_{j3}$) | $Z_j$ coeff ($a_{j4}$) |
|---|---|---|---|---|
| $u(t)$ (m/s) |  | −0.03 (−0.46) | −0.37 (−5.70) |  |
| $q_{\mathrm{PC1}}(t)$ |  | 0.33 |  | −0.15 |
| $q_{\mathrm{PC1}}(t)$ | 0.01 |  | 0.38 |  |

## SUPPLEMENTARY MATERIAL

**Simulated data and power analysis code** (DOI: 10.1214/21-AOAS1471SUPP; .zip). The file contains raw spectra simulated using SOAP 2.0 and R code to perform the Doppler-constrained PCA and detection power analysis discussed in Sections 4.2 and 6, respectively.

## REFERENCES

ADLER, R. J. (2010). *The Geometry of Random Fields*. *Classics in Applied Mathematics* **62**. SIAM, Philadelphia, PA. MR3396215 https://doi.org/10.1137/1.9780898718980.ch1

AIGRAIN, S., PONT, F. and ZUCKER, S. (2012). A simple method to estimate radial velocity variations due to stellar activity using photometry. *Mon. Not. R. Astron. Soc.* **419** 3147–3158.

ÁLVAREZ, M. A. and LAWRENCE, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *J. Mach. Learn. Res.* **12** 1459–1500. MR2813145

ARENTOFT, T., KJELDSEN, H., BEDDING, T. R., BAZOT, M., CHRISTENSEN-DALSGAARD, J., DALL, T. H., KAROFF, C., CARRIER, F., EGGENBERGER, P. et al. (2008). A multisite campaign to measure solar-like oscillations in procyon. I. Observations, data reduction, and slow variations. *Astrophys. J.* **687** 1180.

BARANNE, A., QUELOZ, D., MAYOR, M., ADRIANZYK, G., KNISPEL, G., KOHLER, D., LACROIX, D., MEUNIER, J.-P., RIMBAUD, G. et al. (1996). ELODIE: A spectrograph for accurate radial velocity measurements. *Astron. Astrophys. Suppl. Ser.* **119** 373–390.

BAUMANN, I. and SOLANKI, S. K. (2005). On the size distribution of sunspot groups in the Greenwich sunspot record 1874–1976. *Astron. Astrophys.* **443** 1061–1066. https://doi.org/10.1051/0004-6361:20053415

BOISSE, I., BONFILS, X. and SANTOS, N. (2012). SOAP: A tool for the fast computation of photometry and radial velocity induced by stellar spots. *Astron. Astrophys.* **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. *Astron. Astrophys.* **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. *Publ. Astron. Soc. Pac.* **108** 500.

BUTLER, R. P., VOGT, S. S., LAUGHLIN, G., BURT, J. A., RIVERA, E. J., TUOMI, M., TESKE, J., ARRIA-
    GADA, P., DIAZ, M. et al. (2017). The LCES HIRES/Keck precision radial velocity exoplanet survey. *Astron.
    J.* **153** 208.
COIFMAN, R. R. and LAFON, S. (2006). Diffusion maps. *Appl. Comput. Harmon. Anal.* **21** 5–30. MR2238665
    https://doi.org/10.1016/j.acha.2006.04.006
COIFMAN, R. R., LAFON, S., LEE, A. B., MAGGIONI, M., NADLER, B., WARNER, F. and ZUCKER, S. W.
    (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps.
    *Proc. Natl. Acad. Sci. USA* **102** 7426–7431.
DAMASSO, M., PINAMONTI, M., SCANDARIATO, G. and SOZZETTI, A. (2019). Biases in retrieving planetary
    signals in the presence of quasi-periodic stellar activity. *Mon. Not. R. Astron. Soc.* **489** 2555–2571.
DANBY, J. M. A. (1988). *Fundamentals of Celestial Mechanics*, 2nd ed. Willmann-Bell, Inc., Richmond, VA.
    MR0981360
DAVIS, A. B., CISEWSKI, J., DUMUSQUE, X., FISCHER, D. A. and FORD, E. B. (2017). Insights on the spectral
    signatures of stellar activity and planets from PCA. *Astrophys. J.* **846** 59.
DEL MORO, D. (2004). Solar granulation properties derived from three different time series. *Astron. Astrophys.*
    **428** 1007–1015.
DEL MORO, D., BERRILLI, F., DUVALL, T. and KOSOVICHEV, A. (2004). Dynamics and structure of super-
    granulation. *Sol. Phys.* **221** 23–32.
DUMUSQUE, X. (2016). Radial velocity fitting challenge—I. Simulating the data set including realistic stellar
    radial-velocity signals. *Astron. Astrophys.* **593** A5.
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. *Astrophys. J.* **796** 132.
DUMUSQUE, X., PEPE, F., LOVIS, C., SÉGRANSAN, D., SAHLMANN, J., BENZ, W., BOUCHY, F., MAYOR, M.,
    QUELOZ, D. et al. (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., RAJPAUL, V.,
    TUOMI, M. et al. (2017). Radial-velocity fitting challenge—II. First results of the analysis of the data set.
    *Astron. Astrophys.* **598** A133.
DUPUY, D., HELBERT, C., FRANCO, J. et al. (2015). DiceDesign and DiceEval: Two R packages for design and
    analysis of computer experiments. *J. Stat. Softw.* **65** 1–38.
FISCHER, D. A., MARCY, G. W. and SPRONCK, J. F. (2013). The twenty-five year Lick planet search. *Astrophys.
    J., Suppl. Ser.* **210** 5.
FISCHER, D. A., ANGLADA-ESCUDE, G., ARRIAGADA, P., BALUEV, R. V., BEAN, J. L., BOUCHY, F., BUCH-
    HAVE, L. A., CARROLL, T., CHAKRABORTY, A. et al. (2016). State of the field: Extreme precision radial
    velocities. *Publ. Astron. Soc. Pac.* **128** 066001.
FORD, E. B. (2006). Improving the efficiency of Markov chain Monte Carlo for analyzing the orbits of extrasolar
    planets. *Astrophys. J.* **642** 505–522. https://doi.org/10.1086/500802
HAYWOOD, R. D., COLLIER CAMERON, A., UNRUH, Y., LOVIS, C., LANZA, A., LLAMA, J., DELEUIL, M.,
    FARES, R., GILLON, M. et al. (2016). The Sun as a planet-host star: Proxies from SDO images for HARPS
    radial-velocity variations. *Mon. Not. R. Astron. Soc.* **457** 3637–3651.
JENKINS, J. L. (2013). *Observing the Sun: A Pocket Field Guide*. Springer, New York.
JONES, D., STENNING, D., EB, F., WOLPERT, R., LOREDO, T., GILBERTSON, C. and DUMUSQUE, X. (2022).
    Supplement to "Improving exoplanet detection power: Multivariate Gaussian process models for stellar activ-
    ity." https://doi.org/10.1214/21-AOAS1471SUPP
JOURNEL, A. G. and HUIJBREGTS, C. J. (1978). *Mining Geostatistics*. Academic Press, San Diego.
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. *Stat. Methodol.* **9** 101–114. MR2863601
    https://doi.org/10.1016/j.stamet.2011.07.005
MANDAL, S., KARAK, B. B. and BANERJEE, D. (2017). Latitude distribution of sunspots: Analysis using
    sunspot data and a dynamo model. *Astrophys. J.* **851** 70. https://doi.org/10.3847/1538-4357/aa97dc
MAYOR, M., PEPE, F., QUELOZ, D., BOUCHY, F., RUPPRECHT, G., LO CURTO, G., AVILA, G., BENZ, W.,
    BERTAUX, J. L. et al. (2003). Setting new standards with HARPS. *Messenger* **114** 20–24.
MAYOR, M., MARMIER, M., LOVIS, C., UDRY, S., SÉGRANSAN, D., PEPE, F., BENZ, W., BERTAUX, J.-
    L., BOUCHY, F. et al. (2011). The HARPS search for southern extra-solar planets XXXIV. Occurrence,
    mass distribution and orbital properties of super-Earths and Neptune-mass planets. Preprint. Available at
    arXiv:1109.2497.
MORRIS, M. D. and MITCHELL, T. J. (1995). Exploratory designs for computational experiments. *J. Statist.
    Plann. Inference* **43** 381–402. https://doi.org/10.1016/0378-3758(94)00035-T
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* 7*th International Conference on Information Processing in Sensor Networks* 109–120. IEEE Computer Society, Los Alamitos.

PEPE, F., LOVIS, C., SEGRANSAN, D., BENZ, W., BOUCHY, F., DUMUSQUE, X., MAYOR, M., QUELOZ, D., SANTOS, N. et al. (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. *Astron. Astrophys.* **534** A58.

RAJPAUL, V., AIGRAIN, S. and ROBERTS, S. (2015). Ghost in the time series: No planet for Alpha Cen B. *Mon. Not. R. Astron. Soc. Lett.* **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. *Mon. Not. R. Astron. Soc.* **452** 2269–2291.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning.* MIT Press, Cambridge, MA. MR2514435

WALLACE, L., HINKLE, K. and LIVINGSTON, W. (1998). An atlas of the spectrum of the solar photosphere from 13,500 to 28,000 cm$^{-1}$ (3570 to 7405 Å). *Astrophys. Space Sci. Proc.*

WALLACE, L., HINKLE, K. and LIVINGSTON, W. (2005). An atlas of sunspot umbral spectra in the visible from 15,000 to 25,500 cm$^{-1}$ (3920 to 6664 Å).