The Astronomical Journal, 159:73 (26pp), 2020 February https://doi.org/10.3847/1538-3881/ab5190

Quantifying the Bayesian Evidence for a Planet in Radial Velocity Data

Benjamin E. Nelson[1,2], Eric B. Ford[3,4,5,6], Johannes Buchner[7,8,9,10], Ryan Cloutier[11,12,13], Rodrigo F. Díaz[14,15], João P. Faria[16,17], Nathan C. Hara[18,21], Vinesh M. Rajpaul[19], and Surangkhana Rukdee[20]

[1] Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), Department of Physics and Astronomy, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA
[2] Northwestern Institute for Complex Systems, 600 Foster Street, Evanston, IL 60208, USA
[3] Center for Exoplanets and Habitable Worlds, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA, 16802, USA
[4] Department of Astronomy & Astrophysics, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802, USA
[5] Institute for CyberScience, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802, USA
[6] Center for Astrostatistics, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802, USA
[7] Millennium Institute of Astrophysics, Vicuña MacKenna 4860, 7820436 Macul, Santiago, Chile
[8] Pontificia Universidad Católica de Chile, Instituto de Astrofísica, Casilla 306, Santiago 22, Chile
[9] Excellence Cluster Universe, Boltzmannstr. 2, D-85748, Garching, Germany
[10] Max Planck Institute for Extraterrestrial Physics, Gießenbachstraße 1, D-85748 Garching bei München, Germany
[11] Department of Astronomy & Astrophysics, University of Toronto, 50 St. George Street, Toronto, Ontario, M5S 3H4, Canada
[12] Centre for Planetary Sciences, Department of Physical & Environmental Sciences, University of Toronto Scarborough, 1265 Military Trail, Toronto, Ontario, M1C 1A4, Canada
[13] Institut de recherche sur les exoplanètes, Département de physique, Université de Montréal, 2900 boul. Édouard-Montpetit, Montréal, Quebec, H3T 1J4, Canada
[14] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales. Buenos Aires C1428, Argentina
[15] CONICET—Universidad de Buenos Aires, Instituto de Astronomía y Física del Espacio (IAFE). Buenos Aires C1428, Argentina
[16] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
[17] Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Ruado Campo Alegre, 4169-007 Porto, Portugal
[18] Observatoire de Genève, Université de Genève, 51 ch. des Maillettes, 1290 Versoix, Switzerland
[19] University of Cambridge, Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue, Cambridge CB3 0HE, UK
[20] Pontificia Universidad Católica de Chile, Center of Astro-engineering UC-AIUC, Avda. Vicuña Mackenna 4860, Macul, Santiago, Chile
Received 2018 June 28; revised 2019 October 23; accepted 2019 October 26; published 2020 January 28

Abstract

We present results from a data challenge posed to the radial velocity (RV) community: namely, to quantify the Bayesian "evidence" for $n = \{0, 1, 2, 3\}$ planets in a set of synthetically generated RV data sets containing a range of planet signals. Participating teams were provided the same likelihood function and set of priors to use in their analysis. They applied a variety of methods to estimate $\widehat{\mathcal{Z}}$, the marginal likelihood for each n-planet model, including cross-validation, the Laplace approximation, importance sampling, and nested sampling. We found the dispersion in $\widehat{\mathcal{Z}}$ across different methods grew with increasing n-planet models: $\sim$3 for zero planets, $\sim$10 for one planet, $\sim 10^2$–10^3 for two planets, and $>10^4$ for three planets. Most internal estimates of uncertainty in $\widehat{\mathcal{Z}}$ for individual methods significantly underestimated the observed dispersion across all methods. Methods that adopted a Monte Carlo approach by comparing estimates from multiple runs yielded plausible uncertainties. Finally, two classes of numerical algorithms (those based on importance and nested samplers) arrived at similar conclusions regarding the ratio of $\widehat{\mathcal{Z}}$s for n- and $(n + 1)$-planet models. One analytic method (the Laplace approximation) demonstrated comparable performance. We express both optimism and caution: we demonstrate that it is practical to perform rigorous Bayesian model comparison for models of $\leqslant$3 planets, yet robust planet discoveries require researchers to better understand the uncertainty in $\widehat{\mathcal{Z}}$ and its connections to model selection.

Unified Astronomy Thesaurus concepts: Exoplanet detection methods (489); Radial velocity (1332); Astrostatistics techniques (1886); Model selection (1912); Time series analysis (1916); Algorithms (1883); Bayes factor (1919); Nested sampling (1894); Importance sampling (1892)

1. Introduction

Early Doppler surveys of nearby solar-like stars provided the first census of exoplanet systems. Relatively massive and short orbital period planets with strong radial velocity (RV) signals made up most of this sample, but instrumental upgrades and extended monitoring facilitated the detection of lower-mass and longer-period planets. State-of-the-art RV instruments can reach precisions better than $1\,\mathrm{m\,s^{-1}}$, and continued improvements in spectrograph technologies and stellar modeling (see review by Fischer et al. 2016) hope to achieve a precision sufficient to detect an exo-Earth, an Earth-mass planet orbiting at habitable zone distances from their host stars. This is roughly $10\,\mathrm{cm\,s^{-1}}$ for a solar-mass star.

The journey to this milestone has been fraught with methodological and astrophysical hurdles. One of the most notable are new stellar processes that emerged at the $\sim 1\,\mathrm{m\,s^{-1}}$ level, including but not limited to starspots rotating in and out of view, plages, granulation, stellar oscillations, and long-term stellar activity cycles (Bastien et al. 2014; Cegla et al. 2014; Haywood et al. 2014). Some of these nuisance signals have been previously mistaken as low-mass and/or long-period planets, until follow-up photometric or spectroscopic activity measurements could explain the observed periodicities otherwise (e.g., Robertson & Mahadevan 2014; Kane et al. 2016). In some cases, false-positive detections can arise from aliases in the RV time series itself (e.g., Dawson & Fabrycky 2010; Rajpaul et al. 2016).

In light of these challenges, the RV community needs to improve their analysis of RV data. Dumusque (2016) and

[21] NCCR PlanetS CHEOPS Fellow, Switzerland.

Dumusque et al. (2017) posed a data challenge to the RV community, in which teams had to disentangle planetary signals from other nuisance signals using a set of synthetically generated RV data and activity indicators (bisector span, FWHM of the cross-correlation function, the calcium activity index $\log R'\{hk\}$) and whatever methods they deemed appropriate. Methods that performed best took into account activity indicators, incorporated correlated noise models, and imposed some kind of Bayesian framework. In the longer term, many groups have strayed from a traditional frequentist framework, which attempts to reject the null hypothesis of a no-planet model being compatible with the RV data, and experimented with various algorithms to compute a quantitative evidence for n versus $n + 1$ planets. The Bayesian "evidence" refers to the fully marginalized likelihood, i.e.,

$$\mathcal{Z} \equiv p(\boldsymbol{d}|\mathcal{M}) = \int p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}, \quad (1)$$

where $\boldsymbol{d}$ is a set of real velocity data, $\mathcal{M}$ is the underlying physical and noise model, and $\boldsymbol{\theta}$ is the set of model parameters that describe $\mathcal{M}$. For two models $\mathcal{M}_1$ and $\mathcal{M}_2$, one can update $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ (the ratio of model prior beliefs) with $p(\boldsymbol{d}|\mathcal{M}_1)/p(\boldsymbol{d}|\mathcal{M}_2)$ (the Bayes factor) to calculate $p(\mathcal{M}_1|\boldsymbol{d})/p(\mathcal{M}_2|\boldsymbol{d})$ (the posterior odds ratio, POR).

The art of exoplanet detection ultimately comes down to a decision on whether or not the data support the existence of a planet. The Bayes factor can be interpreted against empirical scales (Kass & Raftery 1995; Jeffreys 1998), see for example Gregory (2007). However, to make decisions with known false-positive and false-negative rates, thresholds on B (or correspondingly POR) need to be calibrated with extensive simulations.

In this work, we focus on the preliminary step toward such Bayesian exoplanet inference: the numerically reliable computation of $\mathcal{Z}$. In particular, we would like to know if different methods converge to similar conclusions about the evidence for n planets, given the exact same data sets and assuming the exact same noise model and prior beliefs. Some examples in RV of methods for computing the Bayesian evidence include thermodynamic integration (Gregory 2007), nested sampling (NS; Feroz & Hobson 2014), geometric path Monte Carlo (Hou et al. 2014), transdimensional Markov Chain Monte Carlo (MCMC; Brewer & Donovan 2015), and importance sampling (Nelson et al. 2016; Jenkins et al. 2017). The above studies were applied to real RV data for systems with suspect planets. The methods were not developed in the same context: each study considered a different RV data set, noise model, and set of n-planet hypotheses, so the relative strengths of these model comparison algorithms are largely unknown. Ford & Gregory (2007) compared several methods for zero- and one-planet models, and Guo (2012) applied some promising methods to multiplanet systems.

Inspired by these previous studies, we designed a data challenge for the RV community to compare different algorithms and implementations for performing model comparison. Participants were given six synthetic RV data sets and a set of n-planet models, where $n = \{0, 1, 2, 3\}$. They were asked to compute quantitative estimates for $\mathcal{Z}$ ($\widehat{\mathcal{Z}}$ henceforth) for each model and their respective uncertainties using whatever computational methods and simplifying assumptions that they choose. This challenge took place in association with a breakout session at The Third Workshop on Extremely Precise Radial Velocities at The Pennsylvania State University in 2017 August 14–17 (EPRV3 henceforth). Some teams participated remotely, while others exchanged ideas during the breakout sessions.

Table 1
Common Variable Names Used Throughout the Manuscript

Variable	Description
$\mathcal{M}_n$	The RV model with n planets
$\boldsymbol{d}$	The RV data
$\boldsymbol{t}$	Times
$\boldsymbol{v}$	RVs
$\boldsymbol{\sigma}$	RV uncertainties
$\boldsymbol{\theta}$	The model parameters
P_i	Orbital period for ith planet
K_i	RV semiamplitude for ith planet
e_i	Eccentricity for ith planet
ω_i	Argument of pericenter for ith planet
M_i	Mean anomaly for ith planet at a fixed epoch
C	RV zero-point offset
σ_J	RV jitter
α	Amplitude of κ
λ_e	Scale length of the exponential component of κ
λ_p	Scale length of the periodic component of κ
τ	Period of periodic component of κ
n	Number of planets
	Statistical Parameters
$\mathcal{Z}$	The fully marginalized likelihood
$\mathcal{L}(\boldsymbol{\theta})$, $p(\boldsymbol{d}\|\boldsymbol{\theta})$	The likelihood function
$p(\boldsymbol{\theta})$	The joint prior probability distribution
Σ	Covariance matrix in likelihood function
κ	Quasi-periodic kernel defined by α, λ_e, λ_p, τ
	Meta-analysis Parameters
$\widehat{\mathcal{Z}}$	Estimate of the fully marginalized likelihood
$\sigma_{\widehat{\mathcal{Z}}}$, $\sigma_{\log\widehat{\mathcal{Z}}}$	Uncertainty in each $\widehat{\mathcal{Z}}$ and $\log\widehat{\mathcal{Z}}$, respectively
$D_{\widehat{\mathcal{Z}}}$, $D_{\log\widehat{\mathcal{Z}}}$	Dispersion in $\widehat{\mathcal{Z}}$ and $\log\widehat{\mathcal{Z}}$, respectively

There are four questions we hoped to answer in the EPRV3 Evidence Challenge:

1. What is the dispersion in reported $\widehat{\mathcal{Z}}$s (i.e., $D_{\widehat{\mathcal{Z}}}$), and how does it change with increasing model complexity (i.e., number of planets)?
2. Does each method's reported uncertainty in $\widehat{\mathcal{Z}}$ (i.e., $\sigma_{\widehat{\mathcal{Z}}}$) accurately reflect the observed dispersion?
3. How does $D_{\widehat{\mathcal{Z}}}$ and $\sigma_{\widehat{\mathcal{Z}}}$ affect our ability to favor n- versus $(n + 1)$-planet models for different data sets?
4. Within the context of this study, which methods should be recommended, avoided, and/or further developed?

This paper summarizes the results of the data challenge. In Section 2, we present the assumed observational and statistical models. In Section 3, we present brief summaries of the different methods that teams employed. In Section 4, we compare everyone's results across many parameters of interest. Finally, in Section 5, we discuss the relative strengths of these methods in the context of the challenge. We reserve a set of variable names to be used throughout the paper, described in Table 1.

Table 2
Simulated Planet Properties

Data Set Number	Detectability	P (days)	K (m s^{-1})	e (unitless)	ω (rad)	M (rad)	C (m s^{-1})	σ_J (m s^{-1})
1	easy	12.1	1.86	0.08	0.0	0.87	1.46	0.6
	easy	42.4	2.44	0.04	2.0	2.99		
2	easy	15.96	2.12	0.05	0.1	0.18	6.33	0.6
	difficult	120.5	1.36	0.31	1.3	0.82		
3	difficult	40.4	1.25	0.1	3.0	4.16	−8.28	0.6
	difficult	91.9	1.19	0.1	0.3	0.33		
4	easy	169.1	1.58	0.22	2.1	0.06	−6.23	0.6
	impractical	23.45	0.74	0.04	6.5	4.37		
5	difficult	31.1	0.75	0.04	0.2	3.31	−4.55	0.6
	impractical	10.9	0.67	0.02	6.2	4.14		
6	difficult	40.4	1.25	0.1	3.0	4.16	−10.7	0.6
	difficult	91.9	1.19	0.1	0.3	0.33		

Note. Each data set contains two planets with a variety of orbits and masses, which we also summarize with their supposed level of detectability (to be referenced again in Figures 2 and 5). Note that data sets 3 and 6 have the same injected planets.

2. Observational and Statistical Models

Each participating team used a standardized set of assumptions for the physical and statistical models. Here, we describe the process used to generate the data sets in detail.

We provided six simulated data sets. The data sets were generated with a set of consistent properties: (1) each data set was an RV time series, including the times of observations ($\boldsymbol{t}$), the "measured" RVs ($\boldsymbol{v}$), and the measurement uncertainties ($\boldsymbol{\sigma}$); (2) the number of observations was fixed at 200; (3) the data were drawn over an observing baseline of 600 days; and (4) each data set included a single velocity offset and correlated Gaussian noise to model stellar activity. We also injected two planets into each data set with a wide range of orbital and mass properties to be described in Section 2.1.

2.1. Physical Model

In each data set, the RV of the star was computed via n-body integrations using Newtonian gravity, one star and two planets. While the full model formally included mutual planetary interactions, we fully expect that it would be well described by the linear superposition of two Keplerian orbits plus a constant velocity offset and a noise term. We estimate the difference between these two assumptions to be less than a couple of cm s^{-1} across all data sets.

The simulation returned a set of line-of-sight velocities of the star $\boldsymbol{v}_{\rm pred}(\boldsymbol{t}|\boldsymbol{\theta})$ for a set of input times $\boldsymbol{t}$ and mass/orbital parameters $\boldsymbol{\theta}$. For the sake of computational efficiency, we restricted the range of injected planet orbital periods to between 10 and 2400 days. Table 2 describes the orbital and mass properties of each pair of planets, along with each data set's input zero-point offset and jitter. Note that data sets 3 and 6 have the exact same injected planets, but the zero-point offset, time series, and noise realizations are different. The generated data sets are shown in Figure 1.

We designed these six data sets with a range of planet detectability in mind. Some planetary signals were relatively easy to identify ($K/\sigma > 1$), which may facilitate efficient computation of $\widehat{\mathcal{Z}}$. Some were relatively difficult ($K/\sigma \sim 1$) or nearly impractical ($K/\sigma < 1$) to find, which could lead to challenging $\widehat{\mathcal{Z}}$ calculations. To reiterate, the main purpose of this challenge is to determine how accurately different algorithms can compute the evidence of n planets in RV data, not their ability to disentangle real planets from astrophysical noise. However, we are interested in how the variation in teams' calculations of $\widehat{\mathcal{Z}}$ depends on the strength of a supposed planetary signal.

2.2. Statistical Model

A likelihood function ($\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M})$) and prior probability distribution on the model parameters ($p(\boldsymbol{\theta})$) are needed to compute the integral in Equation (1). Below, we specify both of these distributions.

2.2.1. Likelihood

Each simulated data point was generated according to

$$v_i = v_{\rm pred}(t_i|\boldsymbol{\theta}) + \epsilon_i, \tag{2}$$

where v_i is a component of $\boldsymbol{v}$, t_i is a component of $\boldsymbol{t}$, and ϵ_i is the perturbation to the measurement due to noise. The noise vector was drawn from a multivariate normal distribution with covariance matrix Σ, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma)$. Therefore, the appropriate likelihood is a multivariate normal distribution, centered on the predictions of the model (parameterized by $\boldsymbol{\theta}$),

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{v} - \boldsymbol{v}_{\rm pred}(\boldsymbol{\theta}))^T \Sigma^{-1} (\boldsymbol{v} - \boldsymbol{v}_{\rm pred}(\boldsymbol{\theta})) - \frac{1}{2}\log|\det \Sigma| - \frac{n_{\rm obs}}{2}\log(2\pi). \tag{3}$$

The Gaussian noise is correlated from one observation to the next. Σ is given by

$$\Sigma_{i,j} = \kappa_{i,j} + \delta_{i,j}(\sigma_i^2 + \sigma_J^2), \tag{4}$$

where $\kappa_{i,j}$ is a quasi-periodic kernel, $\delta_{i,j}$ is the Kronecker delta, and σ_J^2 is the amplitude of an additional unknown noise term (often casually referred to as RV "jitter").

As argued by Haywood et al. (2014) and Rajpaul et al. (2015), we expect some degree of periodicity in stellar activity, modulated by the rotation of the star, which motivates our

choice of a quasi-periodic kernel. It is defined by

$$\kappa_{i,j} = \alpha^2 \exp\left[-\frac{1}{2}\left\{\frac{\sin^2[\pi(t_i - t_j)/\tau]}{\lambda_p^2} + \frac{(t_i - t_j)^2}{\lambda_e^2}\right\}\right], \quad (5)$$

where the hyperparameters are fixed at the following values: $\alpha = \sqrt{3}$ m s^{-1}, $\lambda_e = 50.0$ days, $\lambda_p = 0.5$ (unitless), and $\tau = 20.0$ (days). These values were given for the Evidence Challenge, so teams did not need to marginalize over these hyperparameters.

2.2.2. Priors

In the Bayesian framework, the prior probability density function specifies the state of information prior to taking the observation. It could thus vary from system to system, or as additional information becomes available (for example, from transits). To enable direct comparisons of results across teams, we asked that they adopt a common set of priors, described below. We use a prior that is plausible and convenient to implement, albeit not necessarily informed by orbital mechanics or the latest exoplanet statistics.

We assumed a prior that factorizes in terms of each planet's orbital period (P_i), RV semiamplitude (K_i), eccentricity (e_i), argument of pericenter (ω_i), and mean anomaly at epoch (M_i), as well as the RV offset (C) and the white-noise term (σ_J). Note that for the purpose of computing evidences, teams adopted an orbital period prior ranging from 1.25 to 10^4 days.

1. For each planet's orbital period, we assumed a truncated Jeffreys prior, $p(P)\,dP = \frac{dP}{P} \times \frac{1}{\log(P_{\max}/P_{\min})}$ for $P_{\min} \leqslant P \leqslant P_{\max}$. For the primary analysis, we assumed $P_{\min} = 1.25$ days and $P_{\max} = 10^4$ days for each of the planets. For an alternative analysis, we provided specific values of $P_{\min,i}$ and $P_{\max,i}$ for each planet and data set to be described in Section 2.2.3.
2. For each planet's RV semiamplitude, we assumed a truncated modified Jeffreys prior, $p(K)\,dK = \frac{dK}{K_0(1 + K/K_0)} \times \frac{1}{\log(1 + K_{\max}/K_0)}$ for $0 < K \leqslant K_{\max}$, where $K_0 = 1$ m s^{-1} and $K_{\max} = 999$ m s^{-1}.
3. For each planet's eccentricity, we assumed a truncated Rayleigh distribution, $p(e)\,de = \frac{e\,de}{\sigma_e^2}\exp\left(-\frac{e^2}{2\sigma_e^2}\right)\Big/\left[1 - \exp\left(-\frac{e_{\max}^2}{2\sigma_e^2}\right)\right]$ from $0 \leqslant e < e_{\max} = 1$ and zero for $e \geqslant e_{\max} = 1$, where $\sigma_e = 0.2$.
4. For each planet's argument of pericenter, we assumed a uniform distribution, $p(\omega)\,d\omega = d\omega/2\pi$ from $0 \leqslant \omega < 2\pi$ radians.
5. For each planet's mean anomaly, we assumed a uniform distribution, $p(M)\,dM = dM/2\pi$ from $0 \leqslant M < 2\pi$ radians.
6. For the additional white-noise term, we assumed a truncated modified Jeffreys prior, $p(\sigma_J)\,d\sigma_J = \frac{d\sigma_J}{\sigma_{J,0}(1 + \sigma_J/\sigma_{J,0})} \times \frac{1}{\log(1 + \sigma_{J,\max}/\sigma_{J,0})}$ for $0 < \sigma_{J,0} \leqslant \sigma_{J,\max}$, where $\sigma_{J,0} = 1$ m s^{-1} and $\sigma_{J,\max} = 99$ m s^{-1}.
7. For the RV velocity offset, we assumed a uniform distribution, $p(C)\,dC = dC/2C_{\max}$ from $-C_{\max} \leqslant C \leqslant C_{\max}$, where $C_{\max} = 1000$ m s^{-1}.

Here, the log refers to the natural logarithm. The combined prior for a given n-planet model is

$$p(\{P_i, K_i, e_i, \omega_i, M_i\}_{i=1..n}, \sigma_J, C) = p(\sigma_J)p(C)\prod_{i=1}^{n} p(P_i)p(K_i)p(e_i)p(\omega_i)p(M_i). \quad (6)$$

2.2.3. Two Sets of Priors for Orbital Periods

We previously described a prior where $P_{\min} = 1.25$ days and $P_{\max} = 10^4$ days for each of the planets (the broad prior, henceforth). Note that even for a very well-behaved data set (i.e., one dominant posterior mode if we assume $P_1 < P_2 < P_3$), the posterior would have $n!$ modes corresponding to the number of permutations for ordering n planets. If a team only explores one mode, they would have to renormalize their orbital period prior by a factor of $n!$. However, for the challenge, we imposed an order restriction so teams will ignore this degeneracy when computing $\widehat{\mathcal{Z}}$.

Based on preliminary results reported at the EPRV3 breakout sessions, we noticed that different groups sometimes focused their exploration of parameter space on different regions, particularly in terms of the orbital periods. This made it difficult to directly compare methods. We decided to impose a second choice of priors for orbital period that force all groups to explore the same regions of parameter space in orbital period (the narrow prior, henceforth). That is, we specified different values of $P_{\min,i}$ and $P_{\max,i}$ for each planet and each data set. The values (in days) are as follows for each data set:

1. Data set 1: $P_{\min,1} = 39.8107$, $P_{\max,1} = 44.6684$, $P_{\min,2} = 11.4815$, $P_{\max,2} = 12.8825$, $P_{\min,3} = 10.0$, $P_{\max,3} = 10.7152$.
2. Data set 2: $P_{\min,1} = 15.4882$, $P_{\max,1} = 16.2181$, $P_{\min,2} = 14.7911$, $P_{\max,2} = 17.0608$, $P_{\min,3} = 158.489$, $P_{\max,3} = 251.189$.
3. Data set 3: $P_{\min,1} = 81.2831$, $P_{\max,1} = 107.152$, $P_{\min,2} = 38.0189$, $P_{\max,2} = 42.658$, $P_{\min,3} = 16.5959$, $P_{\max,3} = 17.5792$.
4. Data set 4: $P_{\min,1} = 138.038$, $P_{\max,1} = 204.174$, $P_{\min,2} = 15.1356$, $P_{\max,2} = 16.5959$, $P_{\min,3} = 398.107$, $P_{\max,3} = 1000.0$.
5. Data set 5: $P_{\min,1} = 29.5121$, $P_{\max,1} = 32.3594$, $P_{\min,2} = 10.7152$, $P_{\max,2} = 11.4815$, $P_{\min,3} = 18.197$, $P_{\max,3} = 19.9526$.
6. Data set 6: $P_{\min,1} = 79.4328$, $P_{\max,1} = 141.254$, $P_{\min,2} = 31.6228$, $P_{\max,2} = 50.1187$, $P_{\min,3} = 316.228$, $P_{\max,3} = 398.107$.

These $P_{\min,i}$ and $P_{\max,i}$ values do not necessarily bound true orbital parameters used to generate the data sets. These merely represent a set of reasonable period ranges for each data set to facilitate more direct comparison of different methods. They were chosen without knowledge of the true planet parameters.

2.2.4. Prior over Models

Participants submitted their $\mathcal{Z}$ estimates for the evidence for each $\mathcal{M}_n$, assuming that is the correct n-planet model. In case some participants performed a non-Bayesian analysis, it would be useful to have something that can be compared between Bayesian and non-Bayesian estimates. For those analyses that could not report the marginalized likelihood, we compared the POR to whatever they provide that they think is analogous to a

Table 3
Evidence Challenge Teams and Methods

Method Class	Team Name	Method Name
Computationally cheap	Feng	Bayesian Information Criterion
	Feng	Chib's approximation
	Ford	Laplace approximation
	Hara	ℓ_1 periodogram + Laplace approximation
Importance samplers	Díaz	Perrakis Estimator
	Nelson	Ratio Estimator (MCMC+Importance Sampling)
	Team PUC	Variational Bayes with Importance Sampling
Nested samplers	Rajpaul	MCMC Nested Sampling
	Team PUC	MULTINEST (Nested Sampling)
	Team PUC	MULTINEST(Importance Nested Sampling)
	Team PUC	Multirun-MULTINEST (Nested Sampling)
	Team PUC	Multirun-MULTINEST (Importance Nested Sampling)
	Faria	Diffusive Nested Sampling
Prediction based	Cloutier	Leave-One-Out Cross-validation
	Cloutier	Time-series Cross-validation

POR. To estimate PORs, we must define a prior over $\mathcal{M}_n$,

$$p(\mathcal{M}_n) = \begin{cases} \beta^n & \text{for } n = 1, 2, 3 \\ 1 - \sum_{i=1}^{3} \beta^i & \text{for } n = 0 \end{cases}, \quad (7)$$

and set $\beta = 1/3$. Any participants submitting non-Bayesian estimates were instructed to take this into consideration, so that they could calibrate their estimates appropriately.

3. Methods for Calculating the Marginal Likelihoods

In this section, we will briefly list and describe each method used in the EPRV3 Evidence Challenge. They are described in greater detail in the Appendix. Table 3 provides a list of the teams and methods they employed. Most of the submissions used a unique sampling technique, but some were simply different tunings for the same sampling algorithm. For example, Team PUC submitted MULTINEST results using NS and importance nested sampling (INS) approaches. For each of those algorithms, they also submitted a variety of different MULTINEST tunings (i.e., adjusting the number of live points or the efficiency parameter). When describing each method, we specifically refer to the particular choice of algorithm as opposed to every algorithm and tuning combination.

3.1. Bayesian Computationally Cheap Methods

1. Bayesian Information Criterion (Appendix A.1): The BIC is defined as $-2\log\mathcal{L}_{\max} + k\log N$, where $\mathcal{L}_{\max}$ is the value of the maximum likelihood, k is the number of free parameters, and N is the number of data points. Smaller BIC values suggest higher model probability. Two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ can be compared with $\exp[-(\mathrm{BIC}_{\mathcal{M}_2} - \mathrm{BIC}_{\mathcal{M}_1})/2]$, similar to a Bayes factor. The BIC is derived under very strong simplifying assumptions. Under infinite data, $N \to \infty$, the evidence integral is assumed to become a single, infinitely narrow peak, independent of any prior. In realistic data sets, the posterior has finite width, so the BIC is at best a poor approximation of a Bayesian evidence into question.
2. Chib's Approximation: Chib's approximation is based on the fact that the evidence is the normalization constant of the posterior density at a given point in the parameter space. To estimate the evidence, we choose a point with high posterior probability and calculate the evidence using the one-block sampling of parameter space (Equations (9) and (10) in Chib & Jeliazkov 2001). We divide the MCMC chain into 100 subsamples and calculate the distribution of the evidence.
3. Laplace Approximation (Appendix A.3): The Laplace approximation computes the required integral analytically by approximating the target distribution as a Gaussian. For this challenge, we numerically integrate over the orbital period (grid search) and jitter parameter (Gauss–Legendre quadrature) and apply the Laplace approximation to approximate the remaining model parameters. For this challenge, we used either a circular or epicyclic approximation for the planetary motion to facilitate rapid computation.
4. ℓ_1 periodogram (Appendix A.4): This method relies on the basis pursuit de-noising algorithm (Chen et al. 1998) and is detailed in Hara et al. (2017). It is an alternative to the Lomb–Scargle periodogram or its generalizations, and can be read similarly, but mitigates the problem of aliasing. We here use two ways to assess the significance of its peaks: the false-alarm probabilities (FAPs) as provided by Baluev (2008) and a Laplace approximation of the evidence of the model given by its n tallest peaks.

3.2. Bayesian Importance Samplers

Importance sampling is a integration technique that draws from a simple, normalized distribution that approximates the target distribution, the posterior. If the two distributions are close matches, the integral estimator is accurate and efficient.

1. Perrakis estimator (Appendix A.6): In the Perrakis estimator (Perrakis et al. 2014), the importance sampling function is constructed from the product of marginal posterior densities. Samples are drawn by shuffling the vector elements of joint posterior samples (e.g., from a previous MCMC run) across samples. Additionally, the estimator requires an estimation of the marginal posterior densities of each parameter, which are approximated from a normalized histogram of the marginal samples.
2. Ratio estimator (MCMC + importance sampling; Appendix A.5): This importance sampling technique adopts for the sampling distribution a truncated Gaussian with mean and covariance estimated from a previous MCMC run. For each model and data set, we perform 20 separate MCMC runs, apply this algorithm for each case, and calculate $\widehat{\mathcal{Z}}$ using the median and standard deviation based on the 20 different estimates.
3. Variational Bayes with importance sampling (Appendix A.7): A mixture of Gaussians is used for the importance sampling proposal distribution. For the initial guess of the mixture, multiple global maxima searches are performed. Variational Bayes is an iterative

procedure that optimally updates the Gaussians to match the target distribution better. It samples from the mixture proposal distribution, evaluates the target distribution, and adjusts the parameters of the Gaussians. As with the above techniques, importance sampling estimates the integral.

3.3. Bayesian Nested Samplers

NS is an efficient technique for estimating Bayesian evidence integrals (and numerical quadrature more generally). It computes the geometric size at various likelihood $\mathcal{L}$ thresholds. That threshold is continuously increased, such that the volume decreases exponentially. The gradual increase overcomes the difficulty to handle multimodal posterior distributions (compared to, e.g., MCMC). NS allows both parameter estimation and model comparison. $\mathcal{Z}$ is the integral over likelihood and volume at each likelihood threshold.

Internally, however, NS requires a method for drawing a new random point from the prior with the condition that its likelihood is higher than the current likelihood threshold.

1. MCMC nested sampling (Appendix A.8): Rajpaul's implementation used a semiadaptive MCMC scheme for this purpose; this was chosen as a foil to MULTINEST (below), which instead makes use of a more sophisticated ellipsoidal rejection scheme and clustering algorithm for drawing new points.
2. MULTINEST (Appendix A.9): A robust NS technique, which draws a new uniformly random point with higher likelihood through an ellipsoidal rejection sampling scheme (Shaw et al. 2007; Feroz et al. 2009). Existing live points are clustered into multiple ellipsoids, from which points are drawn. Studying the algorithm parameters, we vary the number of live points (`nlive` = 400–2000) and the target efficiency (inverse of the ellipsoid expansion factor) from 0.3 to 0.01.
3. MULTINEST using INS: An alternative summation of MULTINEST draws that interprets the ellipsoid draws as an importance sampling process (Cameron & Pettitt 2014; Feroz et al. 2019). While the standard NS technique may reject many drawn points failing the likelihood constraint ($\mathcal{L} > \mathcal{L}_i$), INS uses all the points drawn to improve the estimation. The uncertainty on $\log \widehat{\mathcal{Z}}$ can become very small, with up to an order of magnitude higher accuracy than a typical NS (Feroz et al. 2019). However, applying INS in this exoplanet problem, we found that the INS estimator leads to overly small uncertainties. This is shown in the Appendix, Figures 7 and 8.
4. Multirun-MULTINEST (with NS and INS): Examining MULTINEST $\log \widehat{\mathcal{Z}}$ estimates, we find scatter far exceeding the reported uncertainties (in both NS and INS, to be discussed in detail in Section 4.2 and Appendix A.9.2). To obtain robust estimates with realistic uncertainties, we define quantities over multiple runs. We define the multirun evidence estimate as the median $\log \widehat{\mathcal{Z}}$ across runs. For an estimate of the uncertainty on $\log \widehat{\mathcal{Z}}$, we add in quadrature the median absolute deviations (scatter) and the median reported uncertainty. The multirun results are also shown in Figures 7 and 8.
5. Diffusive NS (Appendix A.10): The Diffusive Nested Sampling algorithm (DNS; Brewer et al. 2011) is a Monte Carlo method based on NS. Unlike classic NS, which samples from the prior subject to a hard likelihood constraint, DNS explores a mixture of successively nested distributions, each occupying about e^{-1} times the enclosed prior mass of the previous one. Using a mixture of distributions allows DNS to "go back" to a lower likelihood threshold. After an initial phase where these distributions are created, DNS starts sampling from the complete mixture with uniform weights, which means that the prior is also included in the target distribution, improving the sampling efficiency in multimodal posteriors.

3.4. Prediction-based Methods

1. Leave-One-Out Cross-validation (Appendix A.2.1): In general, cross-validation techniques are commonly used in the field of machine learning to evaluate model performance and inform model selection as an alternative to calculating the fully marginalized likelihood. Cross-validation techniques are used to evaluate the predictive power of a model by splitting the input data set into N training and testing sets. Competing models are then fit to each training set with an objective function (i.e., Equation (3)) being evaluated on the testing set with the optimized model; the score. This formalism helps avoid overfitting of data as models that appear to provide excellent fits to training data will exhibit poor scores on previously unseen testing data if they are actually overfitting. The relative scores between competing models are used for model selection. Leave-one-out cross-validation refers to a particular strategy for train/test splitting wherein N unique splits of the RV time series $\bar{v}$ are made. Each training set contains $N - 1$ of the RV measurements, with the remaining measurement being used for testing.
2. Time-series Cross-validation (Appendix A.2.2): The principle behind time series cross-validation is equivalent to that of leave-one-out cross-validation but differs in the method of train/test splitting. As is the case with RV time series featuring temporally correlated signals—from planets or possibly from stellar activity—removing a single random measurement fails to remove all of signal associated with that measurement. Time series cross-validation works to alleviate this bias by constructing training sets from subsets of the sequential measurements containing at least $N_{\rm min} = 20$ measurements. Each unique training set will then contain $N_{\rm min} + i$ measurements for $i = 0,\ldots,N - N_{\rm min} - 1$. In the single-step forecasting method used here, the corresponding testing sets are the next sequential measurement, i.e., $N_{\rm min} + i + 1$.

4. Results

The four main goals associated with the Evidence Challenge are (1) to better understand the dispersion of estimates of the marginal likelihood ($D_{\widehat{\mathcal{Z}}}$) and how much this varies with the number of planets in the model, (2) to see if the reported uncertainty of $\log \widehat{\mathcal{Z}}$ ($\sigma_{\widehat{\mathcal{Z}}}$) accurately reflects the empirical $D_{\widehat{\mathcal{Z}}}$, (3) to understand how $D_{\widehat{\mathcal{Z}}}$ and $\sigma_{\widehat{\mathcal{Z}}}$ affect our ability to compare the evidence for n- versus $(n + 1)$-planet models, and (4) to

identify promising methods for use and refinement in future studies. In this section, we will address the first three questions and leave the fourth for Section 5.

The methods used to estimate $\mathcal{Z}$ are labeled in the figures based on their directory names in the Evidence Challenge's Github repository.[22]

First, we compare $\log\widehat{\mathcal{Z}}$ (always in base 10) from Bayesian methods that compute it, i.e., without the prediction-based methods in Table 3. We are most interested in the differences and dispersion in the $\log\widehat{\mathcal{Z}}$, not necessarily their absolute values, so we plot each method's $\log\widehat{\mathcal{Z}}$–$\langle\log\widehat{\mathcal{Z}}\rangle$, where $\langle\log\widehat{\mathcal{Z}}\rangle$ is the median $\log\widehat{\mathcal{Z}}$ among the methods being considered.

Note that Team PUC submitted roughly half of the total analyses considered. Most of these were different variations on MULTINEST, in which they varied algorithm settings (number of live points [nlive] and efficiency [eff]) and sampling techniques (NS versus INS, a single run versus multiple runs). This study focuses on comparing methods for estimating $\log\widehat{\mathcal{Z}}$, rather than the choice of algorithm settings for any one method. Therefore, in this section, we include results provided by one set of MULTINEST runs (those with nlive = 2000 and eff = 0.3) which appears to perform well. By including MULTINEST results based on a single set of settings when calculating the median $\log\widehat{\mathcal{Z}}$, we prevent the results from appearing heavily biased toward the MULTINEST results in the figures that follow. An analysis of all MULTINEST results is presented in Appendix A.9.2. All results submitted to the Evidence Challenge are available for further analysis at the Github repository.

4.1. Dispersion in $\log\widehat{\mathcal{Z}}$ ($D_{\log\widehat{\mathcal{Z}}}$)

Figure 2 summarizes the Bayesian results submitted to the Evidence Challenge. Each pixel corresponds to one estimate of $\log\widehat{\mathcal{Z}}$ based on a particular method, orbital period prior, data set, and number of planets included in the model. The color is $\log\widehat{\mathcal{Z}}$–$\langle\log\widehat{\mathcal{Z}}\rangle$, and the color scale spans 10 orders of magnitude in $\widehat{\mathcal{Z}}$. Black pixels are unreported values. We grouped methods into three different classes based on the sample of methods submitted: "computationally cheap," "importance samplers," and "nested samplers." In essence, paler colors correspond to $\log\widehat{\mathcal{Z}}$ values closer to $\langle\log\widehat{\mathcal{Z}}\rangle$, and more saturated colors stray farther from the median. Purple colors are biased toward larger $\log\widehat{\mathcal{Z}}$ with respect to $\langle\log\widehat{\mathcal{Z}}\rangle$, and orange colors are biased toward smaller values. We do not consider reported uncertainties ($\sigma_{\widehat{\mathcal{Z}}}$) here but present that information in Figures 3 and 4.

In most cases, we do not know the true value of $\log\widehat{\mathcal{Z}}$. Thus, it is difficult to quickly evaluate the accuracy of each estimate. For the zero-planet model ($\mathcal{M}_0$, two parameters), multiple teams performed brute force calculations via a very fine grid or large number of Monte Carlo samples to provide a comparison point. However, brute force was not practical for models with $\geqslant$1 planet (7+ parameters). Therefore, we focus our attention on $\log\widehat{\mathcal{Z}}$ estimates relative to $\langle\log\widehat{\mathcal{Z}}\rangle$ and $D_{\log\widehat{\mathcal{Z}}}$, emphasizing that $\langle\log\widehat{\mathcal{Z}}\rangle$ should not be regarded as the "true" $\log\widehat{\mathcal{Z}}$. The dispersion in results across methods can be seen by comparing the color of pixels across rows in Figure 2. All Bayesian methods provided very similar estimates for $\log\widehat{\mathcal{Z}}$ for $\mathcal{M}_0$, with less than a factor of $D_{\log\widehat{\mathcal{Z}}} \sim 0.5$ in variation or $D_{\widehat{\mathcal{Z}}} \sim 3$. However, $D_{\log\widehat{\mathcal{Z}}}$ grows to $\sim$1 for $\mathcal{M}_1$, $\sim$2–3 for $\mathcal{M}_2$, and >3 for $\mathcal{M}_3$.

We also observe differences among the classes of algorithms. Computationally cheap methods have the greatest variability and appear to estimate systematically higher $\log\widehat{\mathcal{Z}}$ values than the results provided by the importance and nested samplers. In practice, this would imply that the computationally cheap methods are typically more confident in the evidence for additional planets. Overall, the importance samplers seem slightly biased to smaller $\log\widehat{\mathcal{Z}}$ relative to the nested samplers, which tend to report larger values of $\log\widehat{\mathcal{Z}}$. In consideration of this, we reanalyzed the $\log\widehat{\mathcal{Z}}$ results excluding the computationally cheap methods, recalculated the $\langle\log\widehat{\mathcal{Z}}\rangle$, and found that the patterns in $\log\widehat{\mathcal{Z}}$–$\langle\log\widehat{\mathcal{Z}}\rangle$ did not significantly change.

Different teams could be computing the evidence for planets at different orbital periods, which may contribute to a substantial fraction of the dispersion seen here. However, we see similar dispersion when teams were instructed to use the narrow period prior. Interestingly, some methods seem to have greater dispersion for the narrow priors, denoted by the more saturated pixels in the left column of Figure 2. We found that some teams renormalized their orbital period prior when they imposed this narrower range while others did not. We corrected for this as noted in the _renormalized.txt files in the Evidence Challenge repository, but significant dispersion remained. In particular, Chib's approximation and the Laplace approximation for circular orbits calculate a $\widehat{\mathcal{Z}}$ for models with $\geqslant$1 planet that can be over five orders of magnitude different from the other methods.

4.2. Uncertainty in $\log\widehat{\mathcal{Z}}$ ($\sigma_{\log\widehat{\mathcal{Z}}}$)

Figure 3 displays the $\log\widehat{\mathcal{Z}}$ results assuming the broad priors and includes the uncertainties in $\log\widehat{\mathcal{Z}}$ ($\sigma_{\log\widehat{\mathcal{Z}}}$). Every panel corresponds to a different n-planet model, and each panel is divided into six subpanels for the six different data sets. Each subpanel plots every method's $\log\widehat{\mathcal{Z}}$–$\langle\log\widehat{\mathcal{Z}}\rangle$, and we display $\langle\log\widehat{\mathcal{Z}}\rangle$ for that data set and model near the top. Figure 4 is in the same format as Figure 3 but displays the results for the narrow period prior. These figures are designed to emphasize $D_{\log\widehat{\mathcal{Z}}}$ across all data sets and how it compares to each reported $\sigma_{\log\widehat{\mathcal{Z}}}$.

For both priors, we find that most methods claim a high degree of precision in $\log\widehat{\mathcal{Z}}$ that does not reflect the observed scatter in estimates of $\log\widehat{\mathcal{Z}}(D_{\log\widehat{\mathcal{Z}}})$. In other words, the estimates are mutually exclusive to an extreme degree. Analytic methods like the Laplace approximation did not report estimates for the uncertainty $\sigma_{\log\widehat{\mathcal{Z}}}$. However, a handful of methods appear to report reasonable $\sigma_{\log\widehat{\mathcal{Z}}}$: the MCMC + importance sampling ratio estimator and variations of multirun-MULTINEST. One common feature among these methods is that $\sigma_{\log\widehat{\mathcal{Z}}}$ was based on comparing the estimates of $\log\widehat{\mathcal{Z}}$ from multiple runs of the same method, rather than an internal estimate of uncertainty based upon a single run. Despite being more computationally expensive, this Monte Carlo approach seems to provide more plausible uncertainty estimates. The MCMC + importance sampling ratio estimator shows particularly large error bars for some data sets in Figure 4. This is likely due to many MCMC runs not converging for those models, thus providing a poor importance sampling density for

[22] https://github.com/EPRV3EvidenceChallenge

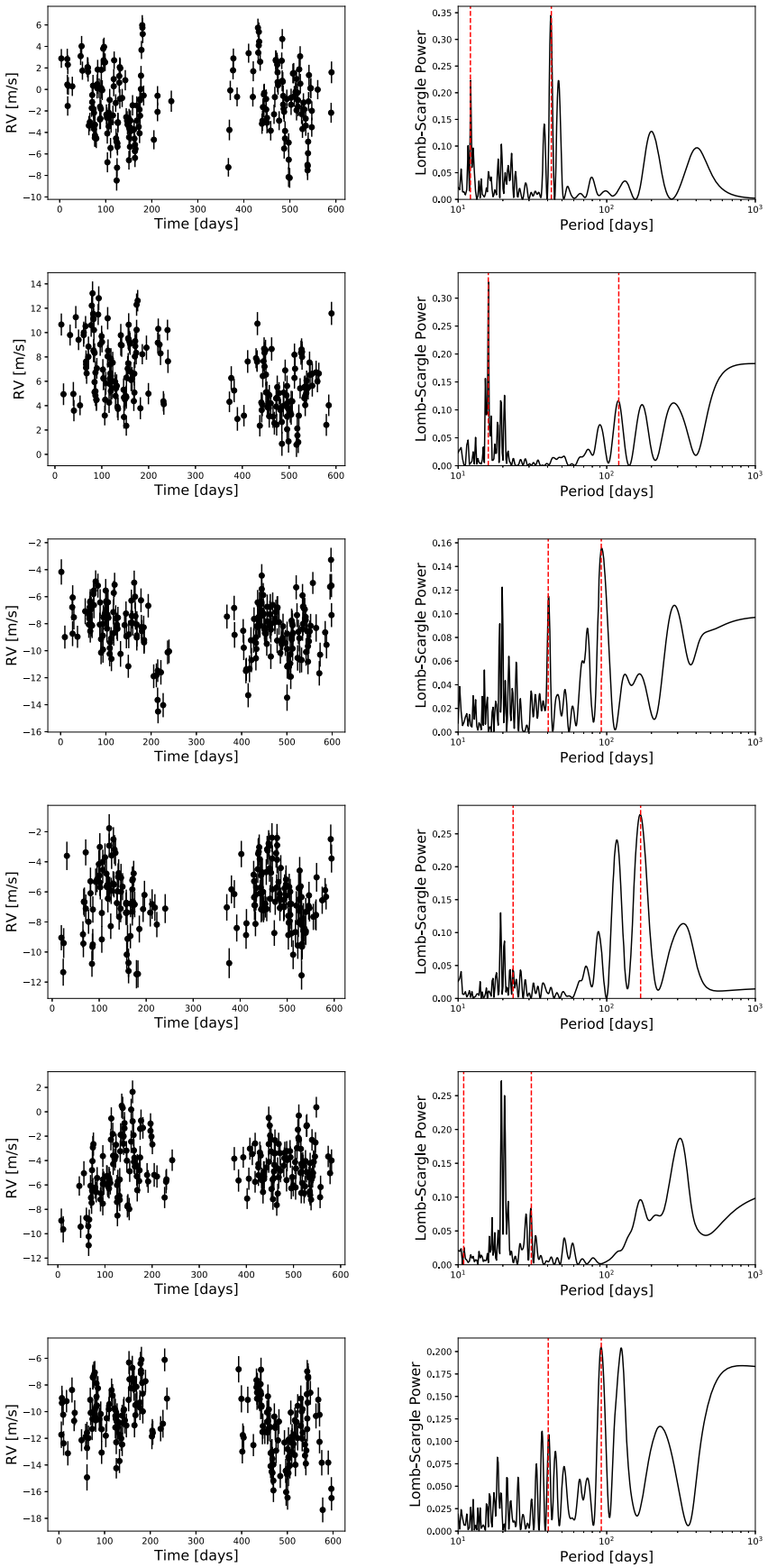

Figure 1. For the Evidence Challenge, we generate six radial velocity data sets (left). The Lomb–Scargle periodograms (right) show the relative strengths of periodic signals in the data sets, with the orbital periods of injected planets indicated (vertical red dashed lines).

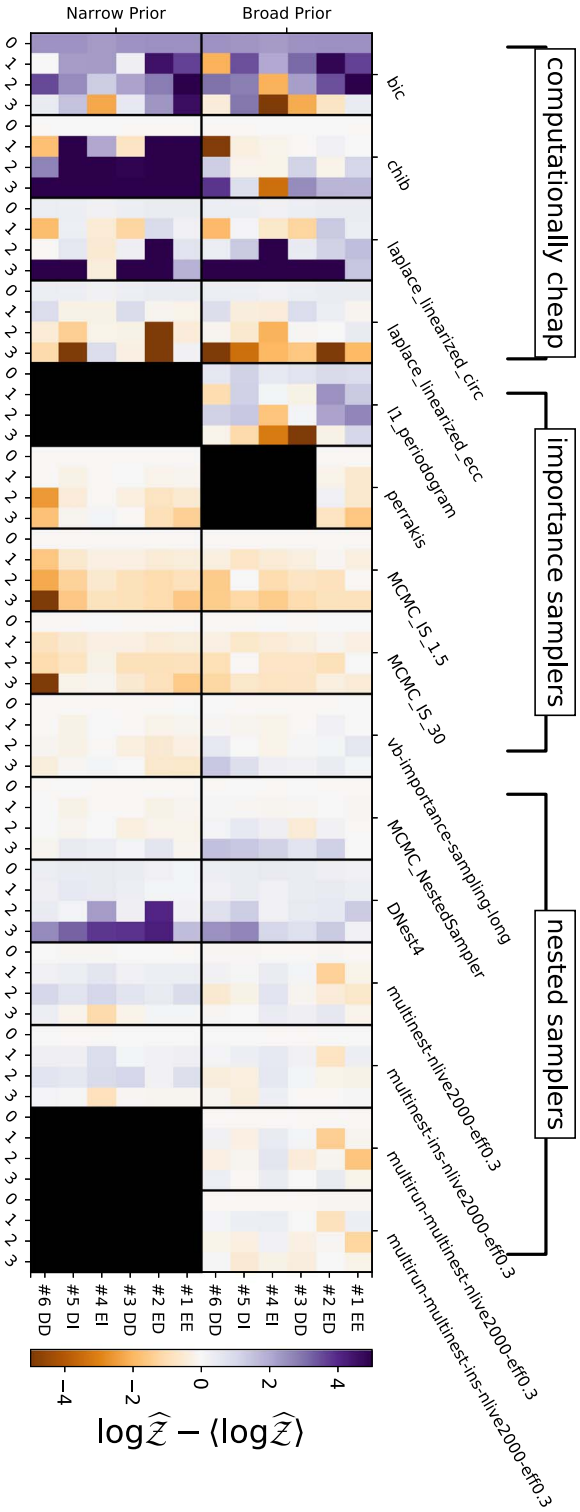

Figure 2. Summary of $\log\widehat{\mathcal{Z}}$ results across all data sets and models. A row of pixels corresponds to an n-planet model, where $n = \{0, 1, 2, 3\}$. Columns correspond to one of the six data sets, each simulated with two planets of varying levels of detectability ("easy" = "E," "difficult" = "D," impractical = "I"). Rows of pixels are grouped with black outlines by method. The left (right) grouped columns correspond to the model with narrow (broad) period priors. The color of each pixel shows $\log\widehat{\mathcal{Z}}$ with respect to the median $\log\widehat{\mathcal{Z}}$ ($\langle\log\widehat{\mathcal{Z}}\rangle$) for that particular data set and model, in order to emphasize the level of scatter seen in all computed $\log\widehat{\mathcal{Z}}$. Any $|\log\widehat{\mathcal{Z}} - \langle\log\widehat{\mathcal{Z}}\rangle|$ greater than 5 is set to a color at the end of the color scale. Black pixels are unreported values.

the estimator. Team PUC directly compared $\sigma_{\log\widehat{\mathcal{Z}}}$ across multiple MULTINEST runs in Appendix A.9.2.

4.3. How $D_{\log\widehat{\mathcal{Z}}}$ Affects Odds Ratios

We see significant dispersion in $\log\widehat{\mathcal{Z}}$ across methods even when assuming the same statistical model. How does this affect our interpretation of n- versus $(n+1)$-planet models? In practice, the evidence is rarely used by itself. Instead, we compare $\log\widehat{\mathcal{Z}}$ for different models by taking ratios of their respective $\widehat{\mathcal{Z}}$ to compute a Bayes factor or POR for assessing the evidence of the nth planet. Methods that initially appear to generate biased estimates of $\log\widehat{\mathcal{Z}}$ could provide an accurate odds ratio if the apparent bias cancels out.

Figure 5 shows the POR results for each method and data set in a format very similar to that of Figure 2. However, instead of results for each individual n-planet model, each pixel corresponds to the POR for a particular pair of models to be compared (for a given method, prior, and data set). For instance, a pixel corresponding to the one-planet versus zero-planet model comparison is denoted as simply "1v0." The color of each pixel is log 10 of the POR, and the color scale spans 10 orders of magnitude in POR. In essence, the bluer pixels favor the $(n+1)$-planet model, redder pixels favor the n-planet model, and pale pixels find roughly similar evidence for the n- and $(n+1)$-planet models. Black pixels are unreported values.

In addition to the Bayesian methods shown in the previous figures, Figure 5 also includes two results based on prediction-based methods: Leave-One-Out Cross-validation, and Time-series Cross-validation. In each case, the team was asked to report a quantity that would be as analogous to a POR as practical given their method.

We discuss several trends in the computed odds ratios across data sets, priors, and method class. After results were submitted, we revealed that each data set contained two planets with different levels of detectability (see Section 2.1). Note that there was an error in the evidence calculations of the ℓ_1 periodogram, and these were revised after the true answers were revealed.

4.3.1. Initial Observations for POR Estimates

Nearly all methods found odds ratios favoring $\mathcal{M}_1$ over $\mathcal{M}_0$. There was more variability across methods when comparing the evidence for $\mathcal{M}_2$ and $\mathcal{M}_1$. Aside from a few exceptions, methods generally did not find odds ratios favoring $\mathcal{M}_3$, across all data sets.

4.3.2. Results for POR by Method Class

We previously identified four classes of algorithms based on everyone's submissions: Bayesian computationally cheap methods, Bayesian importance samplers, and Bayesian nested samplers, and prediction-based methods. The latter two classes of methods require large numbers of model evaluations ($>10^3$) to compute $\mathcal{Z}$. The former two are comprised of (semi)analytic methods or methods that require relatively fewer model evaluations.

We find the numerical Bayesian methods qualitatively agree on the strength of the evidence for n versus $(n+1)$ planets for nearly all data sets and model comparison permutations

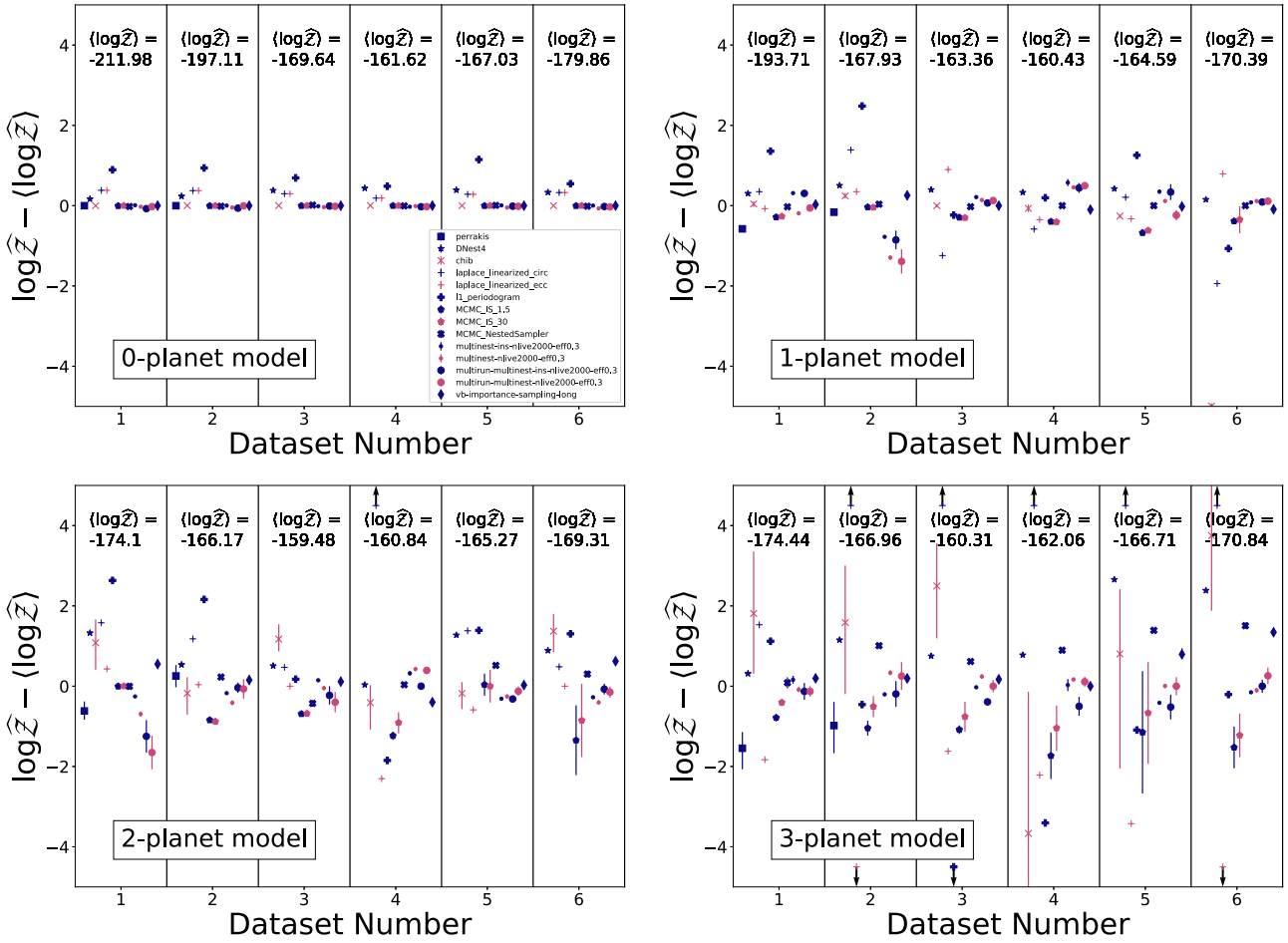

Figure 3. $\log \widehat{\mathcal{Z}}$ estimates for $\mathcal{M}_0$ (upper left), $\mathcal{M}_1$ (upper right), $\mathcal{M}_2$ (lower left), and $\mathcal{M}_3$ (lower right) models assuming broad orbital period priors. All figures show $\log \widehat{\mathcal{Z}}$ with respect to the median value for each data set and model; $\langle \log \widehat{\mathcal{Z}} \rangle$ is displayed at the top of each figure. The symbols correspond to different methods, and colors correspond to different implementations (e.g., input parameters or assumptions) of the same method. Error bars show 1σ equivalent uncertainties in $\log \widehat{\mathcal{Z}}$, some of which are too small to resolve. Methods reporting $|\log \widehat{\mathcal{Z}} - \langle \log \widehat{\mathcal{Z}} \rangle| > 5$ are denoted with arrows pointing outside of the figure bounds.

considered. Even when they do favor detecting an additional planet, these numerical methods tend to report less extreme PORs than the computationally cheap methods, as denoted by the paler pixels for the 2v1 and 3v2 comparisons. However, the computationally cheap methods and prediction-based ones often do not agree on the sign or strength of the evidence for finding an additional planet. Furthermore, they tend to have a much stronger interpretation for either n- or $(n+1)$-planet models, as denoted by the more saturated pixels.

Of the computationally cheaper methods, the Laplace approximation using a linear approximation for eccentric orbits also displayed qualitative agreement with the more computationally expensive methods. We address this importance in Section 5.5.

4.3.3. Results for POR by Data Set and Priors

Here, we assess the reported odds ratios in light of the expected difficulty in detecting the planets in each data set. Data set 1 contained two easily detectable planets. Data set 2 contained an easily detectable planet and two planets that we expected would be difficult to detect. Data sets 3 and 6 contained two planets that we expected would be difficult to detect. These two data sets used the same planet masses and orbits, but different zero-point offsets, observation times, and realization of measurement noise. Data sets 4 and 5 had "easy-impractical" and "difficult-impractical" planets, respectively.

For the broad prior, most methods found decisive evidence for at least one planet in data sets 1, 2, 3, and 6. The notable expectations were the prediction-based methods, which disagreed on the evidence for one planet in data sets 2, 5, and 6. In particular, Leave-One-Out Cross-validation found marginal evidence for a planet in data set 2 and favored no planets in data set 5. All of the remaining methods reported qualitatively similar results for the 1v0 case. For the narrow prior, we see the cross-validation methods had similar disagreements for the 1v0 case in the same data sets. Moreover, Chib's approximation had a much stronger 1v0 interpretation for data sets 4 and 5 than other methods.

For both priors, there is more interesting variability in the POR for the 2v1 and 3v2 cases. There are only two planets in each data set, so the "correct" result is unlikely to have a POR strongly favoring $\mathcal{M}_3$, but could have a POR either near unity or strongly favoring $\mathcal{M}_1$ or $\mathcal{M}_2$.

For data set 1, all methods found strong evidence for at least two planets. Overall, this matches well with the planets' expected level of detectability. The only exception was Chib's approximation, which found strong evidence for three planets when the narrow prior was imposed. For data set 2 and the broad prior, all methods found strong evidence for one planet,

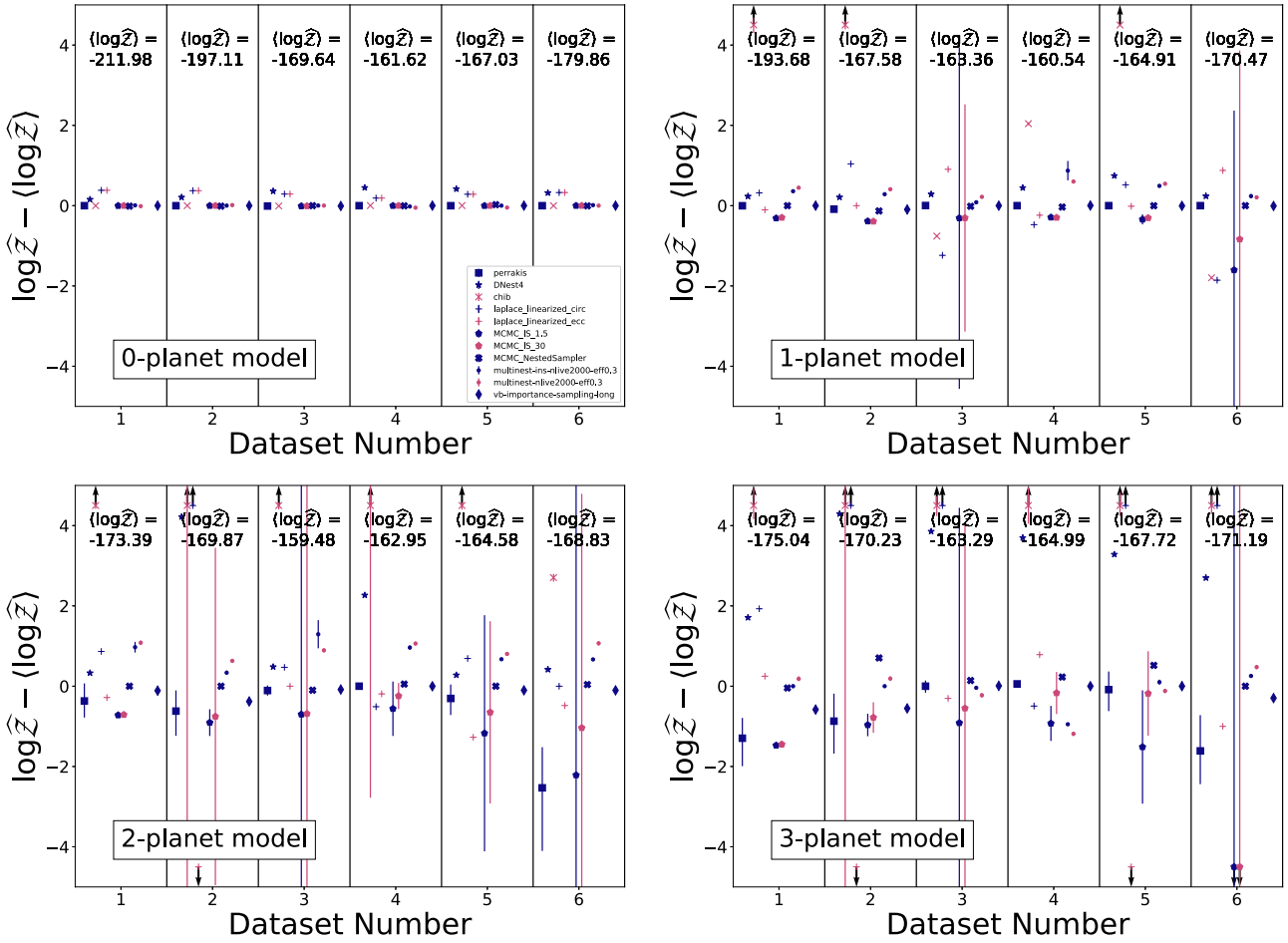

Figure 4. Following the same format as Figure 3, $\log \widehat{\mathcal{Z}}$ estimates of each n-planet model but assuming narrow orbital period priors.

and most found weak to marginal evidence for two planets. For the narrow prior, most methods did not find evidence for a second planet, but the narrow prior interval did not bracket the true orbital period for the second planet (120.5 days). For data set 3, all of the Bayesian methods found evidence for both planets using either set of priors. The narrow prior bracketed the true orbital period values (40.4 and 91.9 days). For data set 4, methods typically found weak evidence for one planet and no evidence for more planets. The supposedly easy-to-detect planet had $P = 169.1$ days, $K = 1.58\,\mathrm{m\,s^{-1}}$, and $e = 0.22$. Perhaps having a P near half Earth's orbital period and this particular noise realization made it more difficult to detect than expected. For data set 5, methods typically found weak evidence for one planet and comparable to no evidence for a second planet, similar to data set 4. In this case, the narrow prior did bracket the true orbital period values (31.1 and 10.9 days). For data set 6, methods found strong evidence for at least one planet and mostly weak evidence for two planets. These conclusions are moderately different from those for data set 3, which contained the exact same planets.

Comparing results for the narrow and broad priors, most methods reported less decisive evidence against three planets when they were allowed to choose a planet at any orbital period (i.e., paler red pixels in the right grouped column than the left). When the narrow prior was imposed, methods typically found evidence for fewer planets.

Note that these odds ratios calculations are based on a physical model that assumes Keplerian orbits. In some cases, the separation between two of the three planets was small (e.g., as imposed by the narrow priors for data sets 1 and 2). We suspect that these scenarios would likely break the Keplerian assumption, and if teams had been instructed to apply an n-body model, then evidence calculations might be affected.

5. Discussion

The Evidence Challenge was envisioned as an opportunity to empirically characterize the accuracy, precision, and robustness of various methods for computing the marginal likelihood of realistic RV data sets.

5.1. Scatter in Estimates

Upon characterizing the dispersion in $\log \widehat{\mathcal{Z}}$, we find reasons for both caution and optimism.

On one hand, estimates for $\log \widehat{\mathcal{Z}}$ often differed by one to two orders of magnitude for the test cases considered. This dispersion is seen across different classes of methods and even within some individual methods. Furthermore, the internal estimates of uncertainty in $\log \widehat{\mathcal{Z}}$ often significantly underestimated the observed dispersion of estimates. For the methods that estimated the uncertainty in $\log \widehat{\mathcal{Z}}$ based on multiple runs, the Monte Carlo uncertainties sometimes spanned >1 order of magnitude, particularly for multiplanet models. Therefore, we

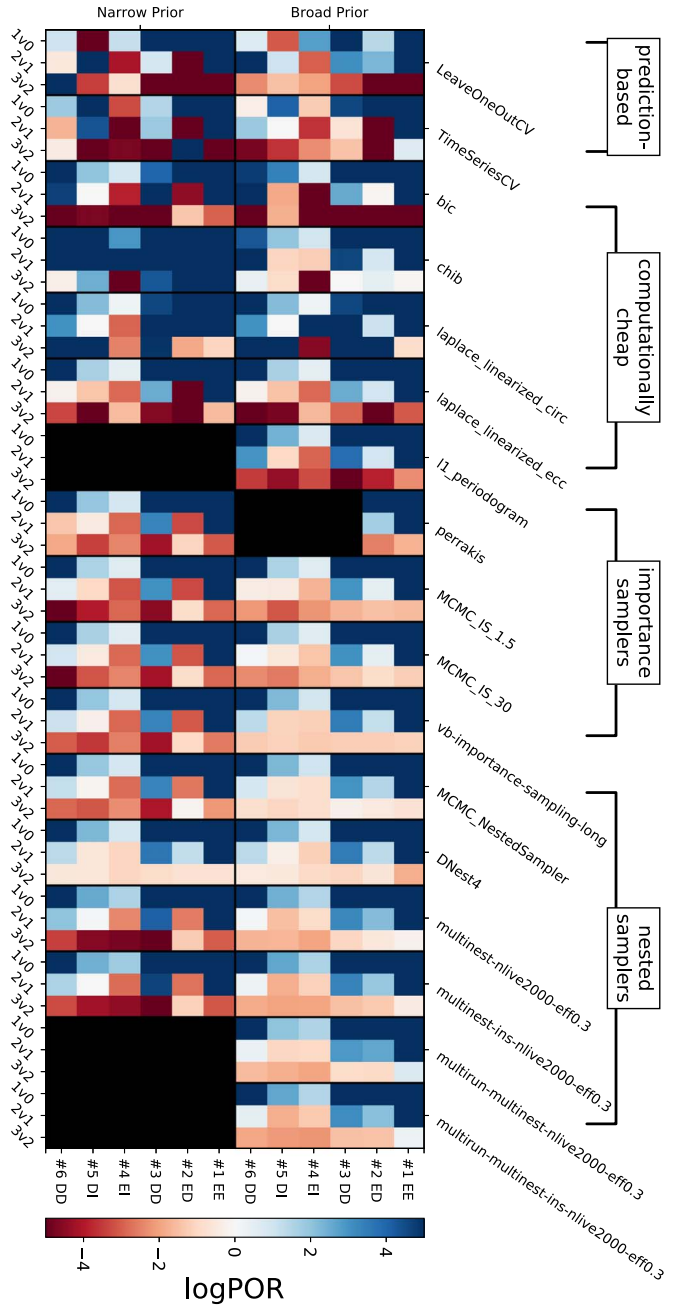

Figure 5. Summary of logPOR results across all data sets and models. A row of pixels corresponds to an odds ratio of an n- vs. $(n+1)$-planet model comparison (i.e., 1v0, 2v1, 3v2). Pixel columns correspond to one of the six data sets, and we also denote the detectability of the two injected planets (easy="E," difficult="D," impractical="I"). Rows of pixels are grouped by method with black outlines. The left (right) grouped column corresponds to the model with narrow (broad) period priors. Pixel colors indicate the logPOR value for that particular data set and model pair to be compared: blue pixels favor the $(n+1)$-planet model, red pixels favor the n-planet model. Any $|\log \mathrm{POR}|$ value greater than 5 is set to a color at the end of the color scale. Black pixels are unreported values.

recommend caution when claiming strong evidence for multiple planets based on an estimated POR within a few orders of magnitude of unity.

On the other hand, it is reassuring to find that the computationally intensive Bayesian methods provided PORs that would lead to similar qualitative conclusions (i.e., favoring an n-planet or $(n+1)$-planet model by at least $>10^4$, or too close to call). For data sets with many high-precision observations (such as those considered here), the POR is likely to deviate from unity by many orders of magnitude, allowing for robust conclusions despite the limitations of existing methods for estimating marginal likelihoods. However, we caution that the POR is more likely to be within a few orders of magnitude of unity for smaller data sets and/or data sets with reduced measurement precision. Additionally, the observed dispersion in marginalized likelihoods increases with the number of planets in the model. Therefore, we caution that even greater estimated PORs are likely necessary to support strong claims for the evidence of more than three planets in a given system, if they are derived with different methods.

Conventional wisdom would suggest that computationally cheap methods are not as robust at estimating $\log \widehat{\mathcal{Z}}$ or logPOR as the more computationally intensive methods. Indeed, most of the computationally cheap methods often disagreed with the computationally intensive methods, especially for cases where the latter found an odds ratio within two orders of magnitude of unity. Furthermore, the likelihood shows complex, multimodal shapes in some data sets, which are missed when only characterizing the best-fit location.

5.2. Non-Bayesian Methods

Here we discuss some alternatives to the Bayesian evidence for deciding the detection of a planet.

Among the submitted results, the prediction-based methods often resulted in a different qualitative conclusion about the evidence for a second or third planet. This is not surprising, since these methods are not estimating the PORs. Future improvements to these methods might reduce the number of false positives and false negatives, including via calibration of the algorithm.

Information criteria, rooted in information theory, quantify if the additional complexity of models is worth storing. One example is the Akaike Information Criterion (AIC; Akaike 1974, see also Watanabe 2013). For our sample size ($N = 200$), the AIC punishes complex models more severely than the BIC (the $2 \times k$ term is replaced by $7.39 \times k$). Considering the results of the BIC, this would introduce several false negatives (the white pixels in the BIC results of Figure 5).

A frequentist approach to distinguish models would be to identify the maximum likelihood $\mathcal{L}_{\max}$ and investigate whether this statistic is substantially higher in the more complex model than in a simpler one ($\mathrm{LR} = \mathcal{L}_{i+1,\max}/\mathcal{L}_{i,\max}$). Because the simpler model is embedded in the parameter edge of the more complex model ($K(i+1) = 0$), analytic formulas do not hold to judge the LR. Instead, the significance (p-value) of the LR improvement has to be found by generating random data sets assuming the best-fit parameters of the simpler model, fitting both the simpler and more complex model (parametric bootstrap). This is however computationally expensive, and even more so when $\mathcal{Z}$ would be considered as the statistic.

5.3. Caveats and Limitations

This Evidence Challenge considered only six data sets, which is not enough to represent the full diversity seen in real RV data sets (e.g., number of observations, observing baselines, planet signal-to-noise ratios (S/Ns), time series, etc.). Therefore, it is unclear how robust our conclusions are to a wider range of RV data quality. These data sets were

designed considering the expected future of the RV field: prioritizing low-mass planets (low RV S/N) with hundreds of observations over multiple observing seasons. On the one hand, these specific concerns about the accuracy and precision of marginal likelihood estimates demonstrated here are not necessarily problematic for the vast majority of previously RV-discovered planets, because most of these planets are relatively more massive (i.e., higher RV S/N) and often had complementary follow-up observations. Furthermore, this analysis was based on RV observations alone with no other forms of supporting ancillary, activity-sensitive data (e.g., transits, activity indicators).

The Evidence Challenge provided an idealized scenario where each team was provided a standardized model, set of priors, and the precise noise model that was used to generate the RV data. When analyzing real data, different teams might reasonably choose to impose different sets of priors. In such cases, if teams explicitly state their statistical model and provide posterior samples, other researchers could reweigh the results using another set of priors (assuming there is sufficient overlap between the posteriors under the two priors). Unfortunately, the exact noise model that generates real data will not be available. Therefore, conclusions about the strength of the evidence for an nth planet must be tempered by uncertainty in the noise model. In the spirit of starting simple, each team was provided the exact values of the other hyperparameters in Equation (5) (e.g., stellar rotation period, correlation lengths) and instructed to hold these parameters fixed. These would need to be estimated or marginalized over for real data (e.g., Faria et al. 2016; Millholland et al. 2018), ideally at the same time as the planetary parameters. Marginalizing over additional hyperparameters would have made it more challenging to estimate evidence accurately, due to increased dimensionality and the potential for multimodal posteriors (Dumusque et al. 2017). In addition to these numerical difficulties, there is an additional challenge of model misspecification, as realistic astrophysical noise is likely more complex than a simple mathematical model.

With recent improvements in the precision, accuracy, and stability of spectrographs, the limitations of current and next-generation RV surveys will often come from stellar astrophysics, rather than random measurement noise. Astronomers are actively seeking new methods of characterizing intrinsic spectroscopic variability of the target stars due to a wide variety of effects (e.g., starspots, granulation, convection, pulsations). In principle, one could estimate the evidence for a model which includes a likelihood on $\boldsymbol{d}$ including both apparent RV measurements and various stellar activity indicators (e.g., $\log R'\{hk\}$). Multivariate Gaussian process noise models seem a particularly promising approach (e.g., Rajpaul et al. 2016; Jones et al. 2017). However, performing the computations necessary for rigorous statistical inference with such models will be even more challenging than for the simple noise model considered in this Evidence Challenge. As astronomers develop more powerful statistical models for analyzing spectroscopic time series, it will likely be useful to perform additional data challenges with such models.

In principle, it is possible that the observed $D_{\log \widehat{\mathcal{Z}}}$ overestimates the dispersion if each method were ideally implemented and tuned. Teams analyzed these data sets independently using a wide variety of codes and tools on platforms with different compilers, libraries, operating systems, and hardware. We cannot eliminate the possibility that some teams may have reported results based on a buggy implementation of an algorithm or chose algorithm settings that resulted in less than ideal performance of the algorithm. In any case, the observed $D_{\log \widehat{\mathcal{Z}}}$ reflects a combination of random and systematic errors intrinsic to each method, finite-precision numerical calculations, and perhaps human errors, similar to those that would arise if these teams had been analyzing real astronomical data sets.

Finally, the evidence estimates submitted do not fully represent the array of statistical methods available to perform quantitative model comparison (e.g., Ford & Gregory 2007). In particular, no results were submitted based on methods using thermodynamic integration. It would also be useful to investigate other computationally cheap methods such as AIC, DIC, or WAIC (Gelman et al. 2014). Other researchers are encouraged to develop and apply alternative methods to the same data sets available in the Evidence Challenge Github repository, as they evaluate methods and implementations.

Table 4
Number of Likelihood Evaluations ($n_{\mathcal{L}}$) Reported to Calculate $\log \widehat{\mathcal{Z}}$ for Data Set 2 and $\mathcal{M}_2$, Assuming Broad Period Priors

Method (Directory Name)	$n_{\mathcal{L}}$	$\log \widehat{\mathcal{Z}} - \langle \log \widehat{\mathcal{Z}} \rangle$
`chib`	1000000	−0.342
`laplace_linearized_circ`	264	1.012
`laplace_linearized_ecc`	319	−0.128
`vb-importance-sampling`	261979	−0.449
`vb-importance-sampling-long`	2883983	−0.012
`MCMC_NestedSampler`	8814939	0.062
`MULTINEST-nlive400-eff0.3`	173460	−0.516
`MULTINEST-nlive400-eff0.01`	768668	0.551
`MULTINEST-nlive2000-eff0.3`	1017587	−0.578
`MULTINEST-ins-nlive400-eff0.3`	173460	0.018
`MULTINEST-ins-nlive400-eff0.01`	768668	0.984
`MULTINEST-ins-nlive2000-eff0.3`	1017587	−0.34
`multirun-MULTINEST-nlive400-eff0.3`	1164856	0.012
`multirun-MULTINEST-nlive400-eff0.01`	5093831	0.588
`multirun-MULTINEST-nlive2000-eff0.3`	5132502	−0.234
`multirun-MULTINEST-ins-nlive400-eff0.3`	1164856	0.107
`multirun-MULTINEST-ins-nlive400-eff0.01`	5093831	1.106
`multirun-MULTINEST-ins-nlive2000-eff0.3`	5132502	−0.204

Note. Similar methods with different tuning parameters or simplifying assumptions are grouped together. The median $\log \widehat{\mathcal{Z}}$ for this set of methods is −166.005.

5.4. Computational Costs

On top of the reported evidence values, roughly half of the teams also provided benchmarking results for their methods, detailing the number of likelihood evaluations, wall-clock time, and/or number of cores required for the evidence calculation. This gives a useful, yet incomplete, picture of the efficiency of these methods. We will take a qualitative look at these results, focusing on the total number of likelihood evaluations ($n_{\mathcal{L}}$, henceforth) of one particular problem: data set 2 and $\mathcal{M}_2$, assuming broad priors. Table 4 shows $n_{\mathcal{L}}$ and the evidence estimate $\log \widehat{\mathcal{Z}}$ relative to the median.

Focusing first on computationally cheap methods (first three rows in Table 4), the Laplace approximation required the fewest $n_{\mathcal{L}}$. These were mainly used in the grid search for the $(n+1)$th planet, as the integral calculation itself was analytic. For the other data sets, Ford reported a wide range of $n_{\mathcal{L}}$, from $n_{\mathcal{L}}=1$ for $\mathcal{M}_0$ up to $n_{\mathcal{L}}\sim 10^5$ for $\mathcal{M}_3$. In general, $\widehat{\mathcal{Z}}$ computed via the circular approximation deviates from other methods by one to several orders of magnitude. For Chib's approximation, Feng used a constant $n_{\mathcal{L}}=10^6$ across all data sets and models.

The remaining methods listed in Table 4 are computationally expensive and include variational Bayes with importance sampling, MCMC-based NS, and variations of MULTINEST. For the MCMC nested sampler, Rajpaul used the largest $n_{\mathcal{L}}$ for this particular case. A future study could investigate whether it is possible for this algorithm to achieve similarly accurate results with fewer $n_{\mathcal{L}}$. For other data sets and models, the number of model evaluations spanned a large range ($n_{\mathcal{L}}\sim 10^6$–10^7) with no clear pattern across different models or data sets. For MULTINEST, $n_{\mathcal{L}}$ increases for larger `nlive` and smaller `eff`. However, interpreting the number of likelihood evaluations also requires understanding the robustness of the results. The $\log\widehat{\mathcal{Z}}$ differences of MULTINEST variations are analyzed in detail in Appendix A.9.2. Briefly, low-efficiency runs (i.e., the `-eff0.01` suffix) show consistent estimates, while `-eff0.3` is unstable. This could suggest that the true $\log\widehat{\mathcal{Z}}$ is above the median (+0.5 or +1.0). In all variants, multiple runs increased the $\log\widehat{\mathcal{Z}}$ estimate, indicating that substantial parts of the integral are often missed. This is also seen in the importance sampling technique increasing the estimate when run longer. With this in mind, $n_{\mathcal{L}}>10^6$ with low efficiency and/or multiple runs seems to be required.

The same trends also hold for the importance NS estimator, which use the same run. However additionally, enabling importance NS requires substantially more memory. Unexplained systematic differences between the INS and classic INS remain (also seen in Table 4). These indicate that the MULTINEST integrations is encountering some difficulties.

Some methods like Chib's approximation and the MCMC + importance sampling ratio estimator rely on a set of posterior samples to estimate $\mathcal{Z}$. If reliable posterior samples were already available (via a database or published along with an RV data analysis), then this would substantially reduce the number of additional likelihood evaluations needed.

5.5. Promising Methods for Future Studies

With the aforementioned results and caveats in mind, we now address the fourth question of the Evidence Challenge: which methods should be recommended, avoided, and/or further developed? In practice, it is difficult to reliably estimate the true value for the odds ratio of high-dimensional (12+ parameter) models. However, we consider the numerical Bayesian methods (i.e., MCMC+importance sampling, variational Bayes+Importance sampling, the Perrakis estimator, MCMC+Nested Sampler, DNEST4, and multirun-MULTINEST) to be more reliable because they provided a consistent set of conclusions. Among this set of evidence estimators, DNEST4 demonstrated the widest deviations from the consensus of the other methods. To reiterate, we found that it is important to estimate uncertainties in the evidence based on multiple independent runs of Monte Carlo algorithms, rather than trusting internal uncertainty estimates based on a single run or posterior sample.

We also identify one computationally cheap method that was consistent with the numerical methods: the Laplace approximation with a linearized eccentric model. This is important because this suggests a (semi)analytic method has comparable performance to methods that often require orders of magnitude more model evaluations. Other than the grid search to find plausible planets, the most computationally expensive part of the Laplace approximation is a single log determinant calculation of the Hessian matrix described in Appendix A.3. For this study, the Laplace approximation demonstrates a nice balance between efficiency and robustness, which would be particularly appealing for analyzing a large number of data sets or data sets with expensive model evaluations. Because this model adopted a linear expansion of the Keplerian motion, it would not be appropriate for application to systems with "high" eccentricity planets. For planets near the threshold of detection, the linear approximation can be much more precise than measurement precision even for sizable eccentricities (e.g., 0.3), because the error term is of order $\sim Ke^2$. We also note that the BIC results generally shared the same sign, but sometimes claimed much more extreme odds ratios in cases where other methods found more marginal ratios.

5.6. Areas for Future Research

Recently, Butler et al. (2017) released RVs for 1642 stars and identified/classified significant signals for each case. Having demonstrated the viability of multiple methods for computing evidence for one, two, and three planet models, one could apply these methods to perform a systematic analysis of these systems. Due to the varied number and precision of RV observations, one should estimate the uncertainty for evidence of each combination of model and data set. When interpreting the results of such an analysis, one should also consider the robustness of conclusions to the choice of likelihood function and potential for model misspecification.

Previous studies that have compared methods for computing marginal likelihoods for RV data were limited to relatively few data sets. Our study was also limited to six RV data sets and four n-planet models, partially because some methods would not scale well to thousands of synthetic data sets. Regardless, this first step at identifying efficient methods will help drive next-generation RV analyses.

Our results illustrate a few of the challenges in the responsible analysis of RV data sets. In order to support current and upcoming RV planet surveys, we recommend much broader evidence challenges that would involve analyzing large number of simulated data sets, so as to understand the rate at which different methods favor nonexistent planets. Such studies could (1) test the robustness and false-discovery rates of the algorithms that performed well over a wider range of RV baselines, cadences, and planet S/Ns by analyzing thousands of simulated RV data sets; (2) compare estimates of the evidence for more sophisticated noise models or more sophisticated physical models (i.e., some that impose stability criterion for multiplanet systems); and (3) compare estimates of the evidence for heterogeneous data sets (i.e., RVs + activity indicators). Interpreting results from the current and next generation of RV surveys will be increasingly complex (e.g., combining a large number of observations, correlated noise models, stellar activity indicators). Therefore, studies such as those recommended above will be critical to establishing the robustness of RV detections and mass measurements.

B.E.N. acknowledges support from CIERA and the Data Science Initiative at Northwestern University. B.E.N. also acknowledges support from the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. E.B.F. acknowledges support from the Institute for CyberScience and the Center for Exoplanets and Habitable Worlds, which is supported by The Pennsylvania State University, the Eberly College of Science, and the Pennsylvania Space Grant Consortium. E.B.F. also acknowledges support from NSF award 1616086, NASA Exoplanets Research Program grant #NNX15AE21G, and supporting collaborations within NASA's Nexus for Exoplanet System Science (NExSS). J.B. acknowledges support from the CONICYT-Chile grants Basal-CATA PFB-06/2007, FONDECYT Postdoctorados 3160439, and the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC120009, awarded to The Millennium Institute of Astrophysics, MAS. This research was supported by the DFG cluster of excellence "Origin and Structure of the Universe." R.C. thanks the Canadian Institute for Theoretical Astrophysics for use of the Sunnyvale computing cluster throughout this work. R.C. is partially supported in this work by the National Science and Engineering Research Council of Canada. J.P.F. acknowledges support from the fellowship with reference SFRH/BD/93848/2013 funded by Fundação para a Ciência e Tecnologia (FCT, Portugal) and POPH/FSE (EC), and also from FCT through national funds and FEDER through COMPETE2020, in the form of grants UID/FIS/04434/2013 & POCI-01-0145-FEDER-007672 and PTDC/FIS-AST/1526/2014 and POCI-01-0145-FEDER-016886. R.F.D. and J.P.F. acknowledge financial support from the organizers to attend the EPRV3 meeting. N.C.H. acknowledges the financial support of the National Centre for Competence in Research PlanetS of the Swiss National Science Foundation (SNSF).

Software: emcee (Foreman-Mackey et al. 2013), george (Ambikasaran et al. 2015), matplotlib (Hunter 2007), MULTINEST (Feroz et al. 2009, 2019), PyMULTINEST (Buchner et al. 2014), pypmc (Jahn et al. 2018).

Appendix

In this section, we will describe the methods presented in Section 3 in greater detail. As mentioned previously, some variables in the following subsections may share common notation with other variables seen in the main manuscript. For such conflicts, we recommend the reader treat these variables as "locally defined" within that method's subsection.

A.1. Feng, BIC

The BIC measures the plausibility of a model through the Laplace approximation of a Gaussian likelihood distribution. It further assumes that under $N \rightarrow \infty$, the posterior becomes dominated by an infinitely narrow peak, which is insensitive to the prior and linear data terms (Konishi & Kitagawa 2008). Despite these strong simplifications, the BIC is frequently used because the posterior density for many inference problems is dominated by a single Gaussian-like distribution and is not sensitive to prior distribution. To compare with other evidence estimators, we follow Kass & Raftery (1995) to approximate the evidence by using $E = e^{-\mathrm{BIC}/2}$, where $BIC = -2\ln\mathcal{L}_{\max} + k\ln N$, $\mathcal{L}_{\max}$ is the maximum likelihood, k is the number of free parameters, and N is the number of data points. Considering such approximation, we use the evidence ratio to assess the performance of the BIC.

The maximum likelihood is calculated through MCMC posterior sampling using DRAM, an adaptive Metropolis algorithm (Haario et al. 2006). The Gelman–Rubin criteria is used to judge whether a chain approximately converges to a stationary distribution (Gelman & Rubin 1992). We draw one million posterior samples using DRAM, drop the first half of the chain as the burn-in part, divide the rest sample into 100 subsamples, and calculate the distribution of $\mathcal{L}_{\max}$ and BIC from these subsamples.

A.2. Cloutier, Cross-validation

In general, cross-validation (CV) is a technique used to evaluate the predictive power of a particular model on an input data set. CV is commonly used to assess model overfitting as overly complex models can often be fine-tuned to produce a high data likelihood while not necessarily generalizing to unseen data (e.g., future observations) and thus demonstrating poor predictive power.

In CV, the first step is splitting the input data set in a training and a test data set. The model parameters are optimized with the training data set. The predictive power of this model with the best-fit parameters is then evaluated on the (previously unseen) test data. The "score" is a user-defined objective function that measures the quality of the prediction. This procedure is often repeated for multiple possible splits of the data. To summarize the model's predictive power, the scores are averaged over the split to give a single score. To select a model, competing models can be compared by their scores. Generally, overly complex models overfit the training data and poorly predict the test data, giving low scores. Overly simple models produce low scores in general, because they poorly fit both training and test data. Good models generalize well from the training data to the test data and have the highest scores.

We note that numerous flavors of CV exist, and the exact nature of the train/test splitting can vary depending on the flavor of CV used. A general summary of the various CV techniques can be found in Arlot & Celisse (2010).

A.2.1. Leave-one-out CV

Leave-one-out CV (LOOCV) represents a common form of train/test splitting in CV. When considering the set of N RV observations $\boldsymbol{v}$, LOOCV first splits the data into N training/testing sets. In each split, one observation is left out as the test data and the training set contains the other $N-1$ observations. For each split, the best-fit parameters $\bar{\theta}$ for each model $\mathcal{M}_n$ under consideration are optimized using a user-defined technique such as least-squares minimization or gradient descent methods. As such, the resulting best-fit parameters may be the maximum a posteriori point estimate, the maximum likelihood parameters, or similar depending on the employed objective function. We have adopted to identify the maximum a posteriori model parameters via MCMC ensemble sampling (Foreman-Mackey et al. 2013) to search for global maxima in

the posterior parameter space. For each model, we employed 100 chains that are run until the chain lengths are $\gtrsim$10 times the average autocorrelation time among the chains. Each P_i is initialized to the period value of a significant peak in the Lomb–Scargle periodogram of the RVs (in descending order of power) while all other orbital parameters are assigned random initial values drawn from their respective priors. In subsequent MCMC simulations on the same data set but under a unique planet model, all parameters for planets featured in both models were initialized to their MAP values from the previous MCMC.

The predictive power of the model is then calculated as the value of the objective function of the testing set under the optimized model. Here, this is the likelihood $\mathcal{L}$ evaluated only on the left-out data point. The final score for each model's predictive power is obtained from the median score among the N splits. We quantify the score dispersion with the median absolute deviation.

We now discuss how to quantify the preference for one model over another within the framework of this challenge. The median score describing a model's predictive power clearly is calculated from the median $\ln \mathcal{L}(\bar{\theta})$ of a single data point. This differs from the Bayesian evidence, which integrates the value $\mathcal{L}(\bar{\theta})$ computed on the full input data set. Therefore, these values cannot be compared directly. However, a useful analogy may be made between the score ratio obtained from LOOCV and the evidence odds ratio obtained from Bayesian techniques. Recall that the training set in each LOOCV split is a single measurement. Therefore, in order to compare scores to Bayesian evidences, one must account for the difference in scale between individual observations and the full N data set. An applicable correction is applied by multiplying the score per observation—obtained in each split from LOOCV—by N. The ratio of median scaled scores can then be used to compute the odds ratio from LOOCV. It is worth re-emphasizing that odds ratios derived in this way are not the same as Bayesian odds ratios.

A.2.2. Time-series CV

In general, LOOCV (see Appendix A.2.1) is only applicable when the measurements within the input data set are independent. In the case when the input data set features correlated observations, standard CV techniques such as LOOCV need to be modified as removing a single random observation fails to remove all associated information due to temporal correlations with the remaining observations. RV time series are often highly correlated in time due to the presence of periodic planetary signals and correlated noise arising from stellar activity (e.g., Astudillo-Defru et al. 2017; Cloutier et al. 2017; Bonfils et al. 2018). The latter signal is present in all of the simulated time series used throughout this study and consequently warrants an alternative form of CV.

One such form of CV used when treating temporally correlated data sets is known as time-series CV (TSCV). TSCV is a variant of LOOCV that measures the predictive power of competing models on a set of observations that are known to be correlated in time. The procedure follows almost identically to LOOCV but differs in the method of train/test splitting. In TSCV, training sets are constructed from a chronologically ordered input data set $\bar{v} = v_1,\ldots,v_N$. The training set t ($t \in [N_{\rm min}, N-1]$) contains the data v_1, …, v_t, and the corresponding testing set is v_{t+1}, the chronologically next observation. For each train/test split, the value of the index t is increased from a minimum training set size $N_{\rm min}$, which we fix to 20, to the full size of the input data set minus one (i.e., $N-1$). Therefore, just like in LOOCV, the testing set in each split is always a single observation, and the scale of each split's calculated score is consistent with the values obtained from LOOCV. TSCV features only $N - N_{\rm min} - 1$ splits, compared to the N splits computed in LOOCV. Quantifying each model's predictive power proceeds identically to LOOCV via the median score and its median absolute deviation over the $N - N_{\rm min} - 1$ splits. The odds ratio comparing competing models is again computed after scaling each model's score per observation by N, before computing the score ratios for each pair of competing models.

A.3. Ford, Laplace Approximation

The Laplace approximation can provide a fast and accurate method for approximating the integral of a function with a single dominant mode that is well separated from the boundary of the integration domain. In particular, consider the integral $\int dx\, \exp f(x)$ and insert the second-order Taylor series for $f(x)$, expanding about x_o the location of the global mode. Then,

$$f(x) \simeq f(x_o) + \frac{1}{2}\sum_{a,b}\frac{\partial^2 f}{\partial x_a \partial x_b}(x - x_o)^2, \tag{8}$$

and the first term can be brought outside the integral. The remaining integral can be approximated analytically if one extends the limits of integration to infinity. Then,

$$\int dx\, \exp f(x) \simeq \left[\frac{(2\pi)^2}{|\det H(x_o)|}\right]^{1/2} \exp f(x_o), \tag{9}$$

where $H(x_o)$ is the Hessian matrix, $\frac{\partial^2 f}{\partial x_a \partial x_b}$, evaluated at x_o. The Laplace approximation can be understood as proportional to the maximum value of $\exp f(x)$ times the width of the global mode. The maximum a posteriori value, the AIC, and the "Bayesian" Information Criterion which are sometimes used as heuristics for model comparison include the maximum posterior value, but do not properly account for the width of the posterior mode. In comparison to the BIC, the Laplace approximation here exploits information about both the priors and the posterior width. Thus, it is expected to be more reliable when the number of observations is finite, and particularly for RV data sets where the number of observations is not very large.

The accuracy of the Laplace approximation depends on the posterior density. For the application to RV survey data, formally the posterior for models with $n \geqslant 1$ planets is highly multimodal, particularly in terms of the orbital period. Fortunately, the posterior for RV data sets is often dominated by a single posterior mode. Indeed, one could adopt a criterion for "detecting" a planet based on the posterior probability distribution being dominated by a single mode. Therefore, we anticipate that the Laplace approximation is likely to be accurate for a data set with n planets if all n planets have been strongly detected, but is likely to be inaccurate for calculating the marginal likelihood to a model with $n+1$ planets.

In practice, the most difficult part of approximating the marginal likelihood via the Laplace approximation is identifying the dominant posterior mode. This is nontrivial for a full Keplerian model. Further, it is possible for the formal posterior

mode to occur at a very high eccentricity and to correspond to a such a narrow spike that the marginal likelihood is actually dominated by the integral around another mode. While it is possible for the marginalized likelihood to strongly favor an n-planet model even if the posterior has multiple significant modes, this implies that there is significant uncertainty in the orbit of the object. This has occurred in the literature for actual exoplanet data sets when aliasing issues cause there to be significant uncertainty in the orbital period of planet (e.g., 55 Cnc e, Dawson & Fabrycky 2010). In principle, one could apply the Laplace approximation around multiple posterior modes to estimate the marginal likelihood. For this study, we instead apply the Laplace approximation to a simplified model, so as to avoid this difficulty. In particular, we construct one of two linearized models for the RV perturbation due to each planet. In the first model, we assume that each planet follows a circular orbit and induces a stellar RV of $v_{\rm pred}(t|A, B, P) = A\cos(2\pi t/P) + B\sin(2\pi t/P)$. In the second model, we adopt an epicycle approximation to each planet's orbit, in which case the RV perturbation can be written as $v_{\rm pred}(t|A, B, P) = A_1\cos(2\pi t/P) + B_1 * \sin(2\pi t/P) + A_2\cos(4\pi t/P) + B_2\sin(4\pi t/P)$. If the orbital period and the covariance matrix are fixed, then there is a single global mode and one can find the values of A and B that maximize the likelihood via linear algebra. Once the posterior mode (conditioned on orbital period and parameters to the covariance matrix) is identified, one can rapidly evaluate the model and the Hessian at the posterior mode.

To find the orbital periods corresponding to the posterior mode, we adopt an iterative approach adding one planet at a time. When evaluating the marginal likelihood for the n-planet model, we perform a brute force grid search over the period of the nth planet, while holding the orbital period of planets 1 through $n-1$ fixed at the values which maximized the posterior probability under the $(n-1)$-planet model. The grid is uniformly spaced in orbital frequency with a density proportional to the frequency range being searched, the time span of observations, and the root mean square of the velocity residuals under the best-fit $(n-1)$-planet model. To avoid local maxima due to aliases with previous planets, we exclude orbital periods periods within 20% of the orbital period of one of the first $n-1$ planets identified. We apply the Laplace approximation with either the circular or epicycle model to compute the posterior probability marginalized over all model parameters other than the orbital periods and the parameters in the covariance matrix.

For each set of orbital periods, we compute the posterior probability given the orbital period and marginalized over the covariance matrix (i.e., σ_J) using 40-point Gauss–Legendre quadrature, as provided by the Julia FastGaussQuadrature.jl package.[23] Initially, we attempted to perform integration over σ_J via the Laplace approximation, but found that this often introduced a nontrivial error due to the cubic term in the expansion about the modal σ_J. This approach is conceptually similar to the Integrated Nested Laplace Approximations technique for latent Gaussian models (Rue et al. 2017).

Finally, we integrate the posterior probability over the orbital period of the nth planet via the Laplace approximation to arrive at the marginalized posterior probability given an nth planet model, where orbits are approximated as circular or epicycles. The orbital period of the nth planet that maximizes the marginalized posterior probability is adopted for future calculations involving $n+1$ planets.

The Laplace approximation combined with the circular model can be interpreted as a Bayesian periodogram, i.e., a brute force search/integration over orbital period combined with a fast approximate model conditional on the orbital periods. This method has the advantage of performing a global search of parameter space for each planet. We anticipate that the Laplace approximation will underestimate the marginal likelihood for models that include more planets that are justified by the data. In these cases, multiple small posterior modes would contribute significantly to the marginalize probability, but our particular implementation only includes one mode. In principle, this could be addressed by summing over multiple posterior modes, but such generalizations are beyond the scope of this study. In practice, this is not a serious limitation, because there is relatively little scientific value in precisely calculating the marginal probability for a model which is not dominated by a single mode (i.e., there are qualitative uncertainties in the orbit of at least one planet).

We anticipate that our Laplace approximation method will be accurate for planetary systems with weak to modest detections, as the posterior would be dominated by a single model and the RV amplitude is small enough that the deviations from circular orbit are small compared to the measurement precision. In order to address this limitation, we performed a similar calculation using the epicycle approximation, so the physical model error is reduced from $O(Ke)$ to $O(Ke^2)$. We anticipate that this will improve the Laplace approximation for planets with strong detections and modest eccentricities. Unfortunately, this also comes with the risk of the model finding spurious posterior modes at high or even unphysical eccentricities. We address the issue of unphysical eccentricities (i.e., $e \geqslant 1$) when using the epicycle model by drawing 100 samples for the inferred A and B coefficients given the modal values of orbital periods and σ_J, and computing what fraction of those samples correspond to an eccentricity less than unity. We multiplied the marginal posterior probability for that set of orbital periods by the fraction of accepted samples. While this eliminated totally unphysical models, it does not make the physical model accurate in the high-eccentricity regime. For systems with high-eccentricity planets, our linearized models will introduce a nonrandom error term. Curiously, it is also possible that the high computational efficiency of this method may result in it finding a narrow posterior mode that other methods may have overlooked due to the difficulty of performing a global search with a nonlinear model. Therefore, when there are significant differences between the marginal likelihood computed via the Laplace approximation and other methods using a Keplerian model, it may not be obvious whether the differences are primarily due to the limitations of the Laplace approximation, the use of an approximate physical model, or the more comprehensive search of parameter space possible with the Laplace approximation.

A.4. Hara, ℓ_1 Periodogram

A.4.1. Overview

In the present work, most of the presented techniques aim at approximating closely the evidence of a model with a given number of planets, in order to perform model comparison. The method presented in this section is similar in that it aims at finding how many planets are orbiting a given star, but differs in that its initial goal is not to compute evidences. Its aim is to

[23] https://github.com/ajt60gaibb/FastGaussQuadrature.jl

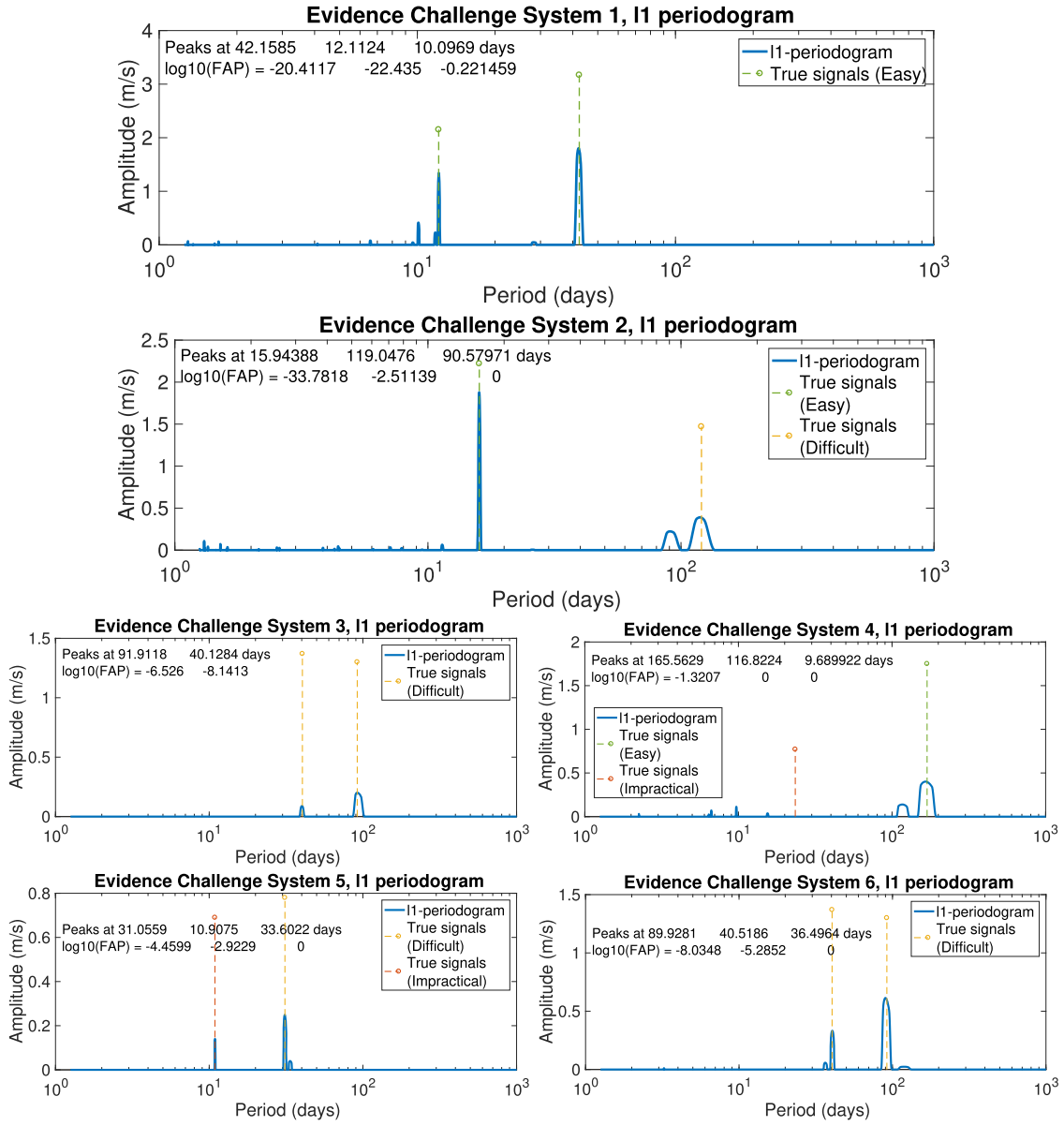

Figure 6. ℓ_1 periodograms of the evidence challenge systems (in blue). The period and semiamplitude of the injected signals are represented by the stems, whose color gives their detection difficulty as defined in Section 2. The legend "Peaks at..." indicates the location of the two or three tallest peaks of the ℓ_1 periodogram in order of decreasing amplitude. The legend "log10(FAP)..." gives the logarithm in base 10 of the false-alarm probability of the signals at the periodicity given above. These figures were obtained before the true location of the periods was known. A version of them without the stems indicating the true signals is available on the GitHub page EPRV3EvidenceChallenge/Inputs/Hara/l1_periodogram.

provide a quick and reliable search for periodicities in RV data while avoiding some caveats of the Lomb–Scargle periodogram (Lomb 1976; Scargle 1982) or its generalizations.

Indeed, it is well known that if several sources of periodicity are present in the signal, due to alias combinations, the maximum of the periodogram might be attained at a period that does not correspond to any signal actually in the data (Dawson & Fabrycky 2010). One solution to that problem is to search for several periods at once, which might be computationally costly.

The alternative we suggest is not to search for best-fitting models with one or a few periodicities, but directly for a Fourier spectrum of the true RV signal. This seemingly more complicated problem will be greatly simplified by an assumption: there are not many planets in the signal. In other words, the signal is sparse in the frequency domain.

The result of our method is an estimate of the Fourier spectrum that we call the ℓ_1 periodogram. Its plot can be read similarly to a classical periodogram, with a significance attached to each peak, but has much fewer peaks due to aliasing. Figure 6 shows the ℓ_1 periodograms we obtain for the six systems of the evidence challenge (in blue). The periods and semiamplitudes of the true planets are given by the stems, with the level of difficulty of their detection in color code as defined in Section 2. For instance, on system 1, the three main peaks are at 42.1, 12.1, and 10.01 days and have respective FAPs $10^{-20.4}$, $10^{-22.4}$, and $10^{-0.22}$; the true signals were two "easy" planets at 42.4 and 12.1 days. The method is fast, that is, it takes typically 5–10 s to run on each data set of this challenge, 20–30 s including the statistical significance assessment on an i7, 2.5 GHz laptop processor. After the challenge,

some further work enabled us to bring these computation times on the evidence challenge data sets down to an average of 1.5 s for the ℓ_1 periodogram calculation and 4.6 s including statistical significance assessment. Note that more conservative values of the FAPs were obtained later, but we chose to plot figures that were publicly available before the results were unveiled.

How the plot is obtained and how the significance is computed are discussed respectively in Appendices A.4.2 and A.4.3. We discuss how our method fits in the present challenge in Appendix A.4.4.

A.4.2. Basis Pursuit De-noising

Let us denote by $\boldsymbol{d}_t$ the data we would have obtained without noise, so that $\boldsymbol{d} = \boldsymbol{d}_t + \boldsymbol{n}$, $\boldsymbol{n}$ being the noise. The variable we wish to estimate is the Fourier spectrum x of $\boldsymbol{d}_t$. To obtain a finite-sized variable, we approximate x by its discretization on a fine grid of equally spaced frequency: $\boldsymbol{x} = (x(\omega_k))_{k=1\,..\,N}$, where $(\omega_k)_{k=1\,..\,N}$ span between 0 and an Ω to be determined. The data then admit the following representation: $\boldsymbol{d}_t = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is an $N_{\rm obs} \times 2N$ matrix whose entries are $A_{kl} = \cos \omega_k t_l$ for $l = 1\,..\,N$ and $A_{kl} = \sin \omega_k t_l$ for $l = N + 1...N$, $l = 1...N_{\rm obs}$.

Obviously, $\boldsymbol{d}_t$ is unknown; we want to find an $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x}$ is close to $\boldsymbol{d}$. For instance, in the sense of the usual Euclidean norm, we can impose $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{d}\|_{\ell_2} < \epsilon$ for some $\epsilon > 0$, where $\|\boldsymbol{z}\|_{\ell_2} = \sqrt{\sum_{k=1}^{2N} z_k^2}$ for any $\boldsymbol{z} \in \mathbb{R}^{N_{\rm obs}}$. As said above, we know that the true signal contains only a few nonzero frequencies (a few planets). It seems reasonable to search for an $\boldsymbol{x}$ that satisfies the inequality and has as few nonzero components as possible. Unfortunately, trying to minimize the number of nonzero components subject to a quadratic constraint is NP-hard (Ge et al. 2011).

We use a proxy of the number of nonzero components of $\boldsymbol{x}$, which is the sum of the absolute values of the coefficients, $\sum_{k=1}^{2N} |x(\omega_k)| =: \|\boldsymbol{x}\|_{\ell_1}$, also termed the ℓ_1 norm of $\boldsymbol{x}$. So, we solve

$$\text{argmin} \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to.} \quad \|\boldsymbol{W}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{d})\|_{\ell_2} \leqslant \epsilon, \tag{10}$$

with $\boldsymbol{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$, $\boldsymbol{\Sigma}$ being the covariance matrix of the noise. The quantity ϵ sets the trade-off between sparsity and agreement with observations. The minimization problem (10) is known as Basis Pursuit De-Noising in the signal processing literature (Chen et al. 1998). Other formulations of ℓ_1 penalties are possible; for instance, Bourguignon et al. (2007) used the Lagrange multiplier version of Equation (10) for spectral estimation. Unlike the number of nonzero components, the ℓ_1 norm is a convex penalty function. Because the constraint $\|\boldsymbol{W}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{d})\|_{\ell_2} \leqslant \epsilon$ defines a convex set, the problem (10) has only one local minimum and is fast to solve.

There have been several algorithms written to solve Equation (10). We selected SPGL1 (Van Den Berg & Friedlander 2008). Several parameters of the algorithms have to be tuned, such as the frequency grid width and spacing, and the tolerance ϵ. In Hara et al. (2017), we provide a method to tune the algorithm parameters; we introduce the $\boldsymbol{W}$ matrix to take into account correlated noise and additional processing steps. We then obtain a quantity $(x^\sharp(\omega_k))_{k=1\,..\,N}$ that can be plotted versus the frequency grid and gives an estimate of the Fourier spectrum, just like in Figure 6 (in blue). Note that the ℓ_1 periodogram is used to find periodic candidates. This is not a good estimator of semiamplitudes, which are underestimated due to the ℓ_1 penalization in Equation (10).

Because several periodicities are searched at the same time, one can expect that the problem of aliases adding up together to give a spurious tallest peak is mitigated. Indeed, the number of misleading peaks is drastically reduced (Hara et al. 2017). However, as in the case of the classical periodogram, significances of the peaks are to be determined.

A.4.3. Significance

We used two methods to evaluate the significance of the peaks; their common feature is to test the improvement made by fitting a periodic signal at the $(n + 1)$th tallest peak of the ℓ_1 periodogram compared to fitting only the n first. For instance, on the system 1 of the Evidence Challenge (see Figure 6, top) we compare the models with a sinusoidal signal at 42.16 days (maximum peak) to nothing, then a model with two sines at 42.16 and 12.11 days to one signal at 42.16 day and so on.

The first way to proceed, as described in Hara et al. (2017), is to compute the significance as if the period of the peaks had been found by a residual periodogram (Baluev 2008). These periodograms generalize the Lomb–Scargle one, and consist in comparing the likelihood of a model that constitutes the null hypothesis H_0 to a model with the H_0 model plus a sine function at a frequency ω. Here, we use the null hypothesis "the signal contains k planets at periods P_1 ...P_k," and the significance for an additional planet is tested. The value of the periodogram at frequency ω is

$$P(\omega) = \alpha \frac{\chi^2_{H_0,\omega} - \chi^2_{H_0}}{\chi^2_{H_0}}, \tag{11}$$

where $\chi^2_{H_0}$ and $\chi^2_{H_0,\omega}$ are, respectively, the χ^2 of the null hypothesis model and of the model with the null hypothesis plus a sinusoidal model at frequency ω, and α is a positive constant. To assess whether an additional periodic signal must be included in the model, one can compute the probability that the random variable "maximum of the periodogram," $P_{\rm max}$, exceeds the maximum value of the periodogram of the data under the null hypothesis, that is, the p-value

$$p = \Pr\{P_{\rm max} \geqslant \max_{\omega} P(\omega) | H_0\}. \tag{12}$$

The assessment of the statistical significance of an ℓ_1 periodogram peak can be done sequentially by using as the null hypothesis a model with sines at the n tallest peaks. Denoting by ω_{n+1} the location of the $(n+1)$th tallest peak, we then use $P(\omega_{n\,+\,1})$ in place of $\max_\omega P(\omega)$ in Equation (12). The values reported in Figure 6 are the p-values computed with formula (5) of Baluev (2008).

The second significance testing method we used for this challenge is a Laplace approximation of the evidence at the period found, as in Appendix A.3. We approximate the evidence of the n-planet model as in formula (5) of Kass & Raftery (1995),

$$\log \mathcal{Z}_n \approx \log \mathcal{L}(\boldsymbol{d}|\widehat{\boldsymbol{\theta}_n}) + \log p_n(\widehat{\boldsymbol{\theta}_n}) + \frac{1}{2}(-\log|\widehat{\boldsymbol{I}_n}| + d_n \log 2\pi), \tag{13}$$

where d_n is the number of parameters of the model, p_n is the prior on the parameters of an n sines model, and $\widehat{\boldsymbol{I}_n}$ is the information matrix evaluated at $\widehat{\boldsymbol{\theta}_n}$. The parameters $\widehat{\boldsymbol{\theta}_n}$ are fitted

with a nonlinear sinusoidal fit initialized at the periods of the n tallest peaks of the periodogram. Note that the fit includes an error term in quadrature of the nominal errors in the maximum likelihood estimation. The Laplace approximation is here computed with an analytical formula we derived. The log of the odds ratio is then approximated by $\log B = \log \mathcal{Z}_{n+1} - \log \mathcal{Z}_n$. The approximated evidences and odds ratio are reported, respectively, in Figures 2 and 5.

A.4.4. Discussion

Residual periodograms are robust tools with a well-founded theory, but they do not necessarily indicate correctly the period of the variation in the data. The ℓ_1 periodogram is thought to be an alternative to residual periodograms and has approximately the same computational workload but mitigates the aliasing problem (for details see Hara et al. 2017).

Significance tests on basis pursuit solutions are a notoriously difficult problem. The present challenge constitutes a good test of applying FAPs or odds ratio, developed in other contexts, to test significance in our case. It seems reasonable because if there are planets, they will appear in general on the ℓ_1 periodogram tallest peaks, and the remaining peaks will be noise. Significance tests such as FAPs or odds ratio should validate the signals until a peak due to noise is selected. The results of the challenge we obtain are consistent with this scenario.

A.5. Nelson, Ratio Estimator (MCMC Importance Sampling)

Importance sampling is essentially a more general form of Monte Carlo integration to estimate $\mathcal{Z}$. We multiply the numerator and denominator of the integrand in Equation (1) by $g(\boldsymbol{\theta})$, a distribution with a known normalization:

$$\mathcal{Z} = \int \frac{\mathcal{L}(\boldsymbol{\theta})p(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{14}$$

This does not change the value of $\mathcal{Z}$, but Equation (14) is in a convenient form such that $\mathcal{Z}$ can be estimated numerically by drawing N samples from $g(\boldsymbol{\theta})$,

$$\widehat{\mathcal{Z}} \approx \frac{1}{N} \sum_{\boldsymbol{\theta}_i \sim g(\boldsymbol{\theta})} \frac{\mathcal{L}(\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}. \tag{15}$$

The efficiency of importance sampling depends strongly on the chosen $g(\boldsymbol{\theta})$. Assuming our parameter space contains one dominant posterior mode, we choose a multivariate normal with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$ for $g(\boldsymbol{\theta})$. For each model considered, we estimate $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ from a set of posterior samples obtained via MCMC.

One good strategy with importance sampling is to pick a $g(\boldsymbol{\theta})$ that is heavier in the tails than $\mathcal{L}(\boldsymbol{\theta})p(\boldsymbol{\theta})$. This makes it easier to sample from low-probability parts of the posterior distribution and prevents any samples from resulting in extremely large weights. However, the chance of sampling from the posterior mode will decrease with increasing dimensionality, which could lead to an inefficient and inaccurate estimate of $\widehat{\mathcal{Z}}$ (see a discussion of the "typical set" in Betancourt 2017). One way around this is to sample from $g(\boldsymbol{\theta})$ within some truncated subspace, $\mathcal{T}$. This new distribution $g_{\mathcal{T}}(\boldsymbol{\theta})$ is proportional to $g(\boldsymbol{\theta})$ inside $\mathcal{T}$ and renormalized to be a proper probability density. Thus, Equation (15) can be rewritten as

$$f \times \widehat{\mathcal{Z}} \approx \frac{1}{N} \sum_{\boldsymbol{\theta}_i \sim g_{\mathcal{T}}(\boldsymbol{\theta})} \frac{\mathcal{L}(\boldsymbol{\theta})p(\boldsymbol{\theta}_i)}{g_{\mathcal{T}}(\boldsymbol{\theta}_i)}, \tag{16}$$

where f is a factor that specifies what fraction of $\mathcal{L}(\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)$ lies within $\mathcal{T}$. We can estimate f with the previously mentioned MCMC sample. By counting what fraction of our posterior samples fell within $\mathcal{T}$, $f_{\rm MCMC}$, we can rearrange Equation (16) to give us $\widehat{\mathcal{Z}}$,

$$\widehat{\mathcal{Z}} \approx \frac{1}{N \times f_{\rm MCMC}} \sum_{\boldsymbol{\theta}_i \sim g_{\mathcal{T}}(\boldsymbol{\theta})} \frac{\mathcal{L}(\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)}{g_{\mathcal{T}}(\boldsymbol{\theta}_i)}. \tag{17}$$

There are two competing effects when choosing the size of our subspace $\mathcal{T}$. If $\mathcal{T}$ is large (i.e., occupies nearly all of the posterior distribution), then $f_{\rm MCMC}$ approaches 1, and we return to a basic importance sampling algorithm that may not be efficient in high dimensions. If $\mathcal{T}$ occupies a much smaller region, then we are more likely to sample from near the posterior mode, but $f_{\rm MCMC}$ approaches 0, making it difficult to accurately estimate $\widehat{\mathcal{Z}}$. This necessitates carefully choosing an appropriate $\mathcal{T}$ that will provide a robust estimate for $\widehat{\mathcal{Z}}$. Guo (2012) and Nelson et al. (2016) provide more detailed prescriptions and investigations of this method.

Here, we compute $\widehat{\mathcal{Z}}$ for all models using small (1.5) and large (30) truncated subspaces. Our parameterization for $g(\boldsymbol{\theta})$ is P, K, $\sqrt{e}\sin\omega$, $\sqrt{e}\cos\omega$, and $\omega + M$ for each planet, and C and σ_J for the zero-point offset and jitter, respectively. We run 20 independent MCMCs per model per data set and compute a $\widehat{\mathcal{Z}}$ value based on every MCMC. We report the median and standard deviation for each set of 20 $\widehat{\mathcal{Z}}$ values.

A.6. Díaz, Perrakis

The Perrakis estimator is an importance sampling estimator described in detail in Perrakis et al. (2014). The importance sampling density used is the product of the marginal posterior distributions of parameter blocks. In our case, we chose one-dimensional blocks, so that the importance sampling function is

$$g(\boldsymbol{\theta}) = \prod_{i=0}^{D} p(\theta_i|\boldsymbol{d}) \ ,$$

so that the samples are drawn from the marginal posterior distributions,

$$\theta_i^{(n)} \sim p(\theta_i|\boldsymbol{d}) \ \text{ for } \ i = 1, 2, ..., D \ ,$$

for a D-dimensional model. This produces the estimator

$$\widehat{\mathcal{Z}} = N^{-1} \sum_{i=0}^{N} \frac{p(\boldsymbol{d}|\theta_1^{(n)}, \theta_2^{(n)}, ..., \theta_D^{(n)}) p(\theta_1^{(n)}, \theta_2^{(n)}, ..., \theta_D^{(n)})}{\prod_{j=0}^{D} p(\theta_j^{(n)}|\boldsymbol{d})} \ . \tag{18}$$

The estimate can be computed based on joint posterior samples drawn using, for example, an MCMC algorithm, but requires two additional elements: draws from the marginal posterior distributions of the parameter blocks and an estimate of the marginal densities that appear in the denominator of Equation (18). The former is promptly obtained by shuffling the elements of the parameter vector across MCMC samples. This breaks the correlation between parameters and leads to samples that are drawn from the product of (independent)

marginal posteriors. More details and a discussion on this are given in Perrakis et al. (2014).

As we used one-dimensional parameter blocks, the marginal posterior densities are approximated by the corresponding normalized histogram. Of course, to obtain a precise estimate, a large posterior sample and small bin sizes are required. However, we checked that the result does not change significantly with bin size. This estimation could be improved by modeling the marginal distributions or using a kernel density estimation.

The resulting estimate, which we named the Perrakis estimator, was previously employed in the analysis of exoplanet data in a number of articles (e.g., Díaz et al. 2016a, 2016b; Bonfils et al. 2018). Here, we obtained a sample of size 5000 from the importance sampling function to perform the computation of the evidence estimate, $\widehat{\mathcal{Z}}$. The uncertainty was computed by repeating the computation 600 times. At each time, a new sample is considered: a new subsample of the joint posterior distribution is taken, and a new shuffling is performed.

The joint posterior sample was obtained using the affine-invariant ensemble sampler by Goodman & Weare (2010) implemented by Foreman-Mackey et al. (2013). For each data set and model, we ran 300 walkers for 30,000 iterations.

A.7. Team PUC, Variational Bayes with Importance Sampling

Johannes Buchner used an integration algorithm based on variational Bayes (VB) and Importance Sampling. The method is very similar to the one described in Beaujean & Caldwell (2013) and uses their PYPMC package (Jahn et al. 2018).

The method proceeds as follows:

1. Identify likelihood maxima to guess an initial mixture. The original technique used points from several MCMC chains. Here, a single MULTINEST run (see Appendix A.9) is used to obtain initial posterior points. This just serves to identify an initial mixture density and does not rely on MULTINEST sampling correctly. The posterior points are divided into groups based on their likelihood value and clustered further into subgroups. This is analogous to multiple MCMC chains split into segments in Beaujean & Caldwell (2013).
2. Generate an initial Gaussian mixture density from the above groups. The intent is to develop a mixture that closely describes the posterior well so that importance sampling is efficient.
3. Run Variational Bayes to optimize the proposal mixture density against the posterior points.
4. Define an Importance Sampler based on the optimized mixture. Set N to 1000 times the number of model parameters.
5. Loop:
   (a) Draw N importance samples from the mixture and evaluate their likelihood.
   (b) If the importance sampling integral uncertainty is below the threshold $\sigma_{\widehat{\mathcal{Z}}} < 0.3$ and the effective sampling size is above 100, terminate.
   (c) Otherwise: increase N by a factor of 1.4. This implies that the total number of samples drawn increases exponentially.
   (d) Update the proposal mixture density with Variational Bayes.
   (e) In every third loop, the previous step is not done. Instead, the proposal mixture density is recreated from scratch (as above), but with one more point group. That group is created by starting a simple MCMC chain from the point with the highest weight, after a simple likelihood optimization.

Iteratively optimizing with Variational Bayes is effective in making the importance sampler efficient and improves the integration uncertainty. However, a limitation is that the number of mixture components cannot increase. If importance sampling discovers a new small peak, VB typically does not place a component there. To solve this, step 5e recreates the mixture from scratch (with up to 10 components). The local MCMC run helps in identifying the size of the potential new component. In the subsequent iteration, all previous samples are used to optimize the mixture, and the number of components can shrink again (often drastically).

We also include a long run from this algorithm, where we initialize from the combination of 10 MULTINEST preruns (to mitigate the problems named in the MULTINEST section), higher number of importance samples, and integrate to a higher effective sampling size (20,000) before terminating. At the cost of many likelihood evaluations, this should be safer. For some data sets, this stringent termination criterion was never reached, and the runs were terminated manually.

A.8. Rajpaul, MCMC Nested Sampler

NS is a technique developed by Skilling (2004) and Skilling (2006) for Bayesian model comparison via estimation of Bayesian evidence integrals. As NS produces samples from the posterior PDFs of model parameters as a trivial byproduct of the evidence integral estimation, it may be thought of as a reversal of the usual approach to Bayesian inference. Although Skilling's original formulation was designed with Bayesian inference in mind, NS is in fact a general method for numerical integration that may be applied to any continuous integrals.

NS proceeds by exploring the volume above a given likelihood threshold. That threshold is continuously increased, such that the volume decreases by a constant factor (exponential shrinkage). This allows NS to keep track of the volume and likelihood value for making a Lebesgue integral. At a late point, the volume is small and the likelihood flat, so that the remainder does not contribute to the integral, and the algorithm terminates.

The shrinkage of NS is achieved by having e.g., 100 live points sampling the prior space uniformly and then removing one. This reduces the represented volume by $\sim 1/100$. Next, the algorithm samples a new point with a likelihood higher than the removed point. The number of live points therefore determines the speed of the shrinkage and how coarsely the space is sampled.

The error of the integral estimate is given in Skilling (2004). The usual implementation assumes that the bulk of the integral can be found around some (rather than multiple) shrinkage; in practice, this is a sufficient approximation.

Internally, NS requires an algorithm for drawing a new, random point from the prior with the condition that its likelihood is higher than the current likelihood threshold. Several general solutions for these constrained drawing algorithm exist, including those relying on local steps (e.g., MCMC, Galilean Monte Carlo, HMC, POLYCHORD—and

those reconstructing the volume enclosed by the likelihood contour, e.g., MULTINEST, RADFRIENDS). See Buchner (2016) for a more detailed discussion.

Here, Rajpaul implemented the MCMC sampler from Veitch & Vecchio (2010) to generate samples within a standard NS routine. This implementation is different from MULTINEST (see below) in that it replaces the clustering algorithm or ellipsoidal rejection schemes with a semiadaptive MCMC exploration of the prior range. In particular, the present implementation used a mixture of the following proposal schemes to draw new samples: a Student-t distribution (with $\nu = 2$ degrees of freedom) based on the Cholesky-decomposed covariance matrix of the live points, differential evolution using two randomly selected points from the current live points, and affine-invariant walk and stretch moves (see Goodman & Weare 2010).

The algorithm as presented by Veitch & Vecchio has two main parameters that can be adjusted: N, the number of live points; and M, the number of MCMC iterations. By tuning N and M, any desired level of evidence accuracy can (in principle) be achieved, albeit at the expense of increasing computational burden, with the total number of likelihood evaluations scaling linearly with both N and M. Based on recommendations given by Veitch & Vecchio, and to strike a balance between a reasonable computation time and (ostensible) accuracy, Rajpaul fixed $N = 1000$ and $M = 1000$, such that estimation of a given model's evidence would require of order 10^6 likelihood evaluations.

Rajpaul noted a priori that his own experience was that MULTINEST was typically faster and better-suited to higher-dimensional (>10-dimensional) problems than the above algorithm from Veitch & Vecchio. Nevertheless, the MCMC sampler from Veitch & Vecchio was implemented for this evidence challenge to provide a foil to the more popular MULTINEST NS algorithm, discussed below.

A.9. Team PUC, *MULTINEST*

Team PUC (Johannes Buchner and Surangkhana Rukdee from Pontificia Universidad Católica de Chile) employed NS with the constrained drawing algorithm MULTINEST. MULTINEST's multimodal ellipsoidal sampling (Shaw et al. 2007; Feroz et al. 2009) encloses the existing random points into best-fitting ellipsoids. These are enlarged by a certain factor (inverse of the efficiency parameter). New points are drawn from the enlarged ellipsoids and rejected if below the likelihood threshold. Therefore, the ellipsoids reduce the space to be sampled, making MULTINEST fast (in terms of number of likelihood evaluations needed). However, if the ellipsoids accidentally cut away parameter space regions, e.g., because the enlargement is too small or the contours do not look similar to ellipsoids, the estimate can be biased.

A.9.1. Algorithm Variations

MULTINEST has two parameters, the number of live points `nlive` and the target efficiency `eff` (inverse of the ellipsoid enlargement). We chose a standard configuration (`MULTINEST-nlive400-eff0.3`) and two variations, increasing either the number of live points (`MULTINEST-nlive2000-eff0.3`) or the enlargement (`MULTINEST-nlive400-eff0.01`).

Importance NS is a modification of NS where the rejected points can improve the estimate (Cameron & Pettitt 2014; Feroz et al. 2019). To some degree, this also mitigates the above-mentioned issues of imperfect ellipsoid sampling. MULTINEST computes both the standard NS estimator and the importance NS estimator. The results are named correspondingly (MULTINEST-ins-nlive400-eff0.3, MULTINEST-ins-nlive400-eff0.01, MULTINEST-ins-nlive2000-eff0.3).

A.9.2. Scatter between *MULTINEST* Runs

We observe that there are substantial scatter and outliers in the evidences between MULTINEST runs. Figures 7 and 8 shows the scatter and assigned errors for repeated runs of data set 1 and data set 4, respectively. Panel columns represent the three MULTINEST configurations and panel rows show different numbers of modeled planets. Each panel shows the comparison between the NS estimator and INS estimator for six runs. In most cases, the INS estimator gives a smaller error bar to compare to the NS estimator. However, it sometimes shows outliers; for example, in data set 4 (Figure 8) Run 3 with one planet, increasing `nlive` from 400 (left column) to 2000 (middle column) yields smaller errors. Decreasing the efficiency from 0.3 to 0.01 (right column) gives systematic offsets between NS and INS estimators.

Throughout, the quoted uncertainties of MULTINEST are smaller than the scatter between runs. Low outliers can come from undiscovered solutions, but increasing the number of live points did not eradicate this completely. Imperfect ellipsoids can also lead to scatter in the estimate. Indeed, decreasing the efficiency also decreases the scatter, but at great computational cost. Using the INS estimator instead of the standard NS generally leads to overly small uncertainties. One conclusion is that running MULTINEST just once gives unreliable uncertainty estimates, which cannot be completely eradicated by decreasing the efficiency or increasing the number of live points.

To represent this additional uncertainty in MULTINEST, we define a multirun estimator. We ran MULTINEST six times and combine the evidence estimate as the median of individual estimates:

$$\log \widehat{\mathcal{Z}} = \mathrm{median}(\log \widehat{\mathcal{Z}}_i).$$

The multirun error is defined as the median of the absolute deviations and the median individual error estimates added in quadrature:

$$\sigma^2_{\widehat{\mathcal{Z}}} = \mathrm{median}(\sigma_i)^2 + \mathrm{median}(|\log \widehat{\mathcal{Z}}_i - \log \widehat{\mathcal{Z}}|)^2 .$$

This gives appropriate errors when MULTINEST is having trouble and shows substantial scatter, yet is robust against individual outliers. The bottom of each panel of Figures 7 and 8 shows our MULTINEST multirun estimators.

A.10. Faria, Diffusive Nested Sampling

One of the main challenges with the NS algorithm is to generate new particles from the likelihood-constrained prior. As described above, a number of methods have been proposed for this (and used in the current work). However, some of those methods, and NS in general, tend to suffer from the curse of dimensionality, with sampling efficiency decreasing rapidly with the dimension of the parameter space. This is particularly problematic if the posterior distribution is multimodal or highly correlated. Brewer et al. (2011) introduced a new algorithm, which they called Diffusive Nested Sampling (DNS), designed to be as flexible and general as a more standard MCMC, but

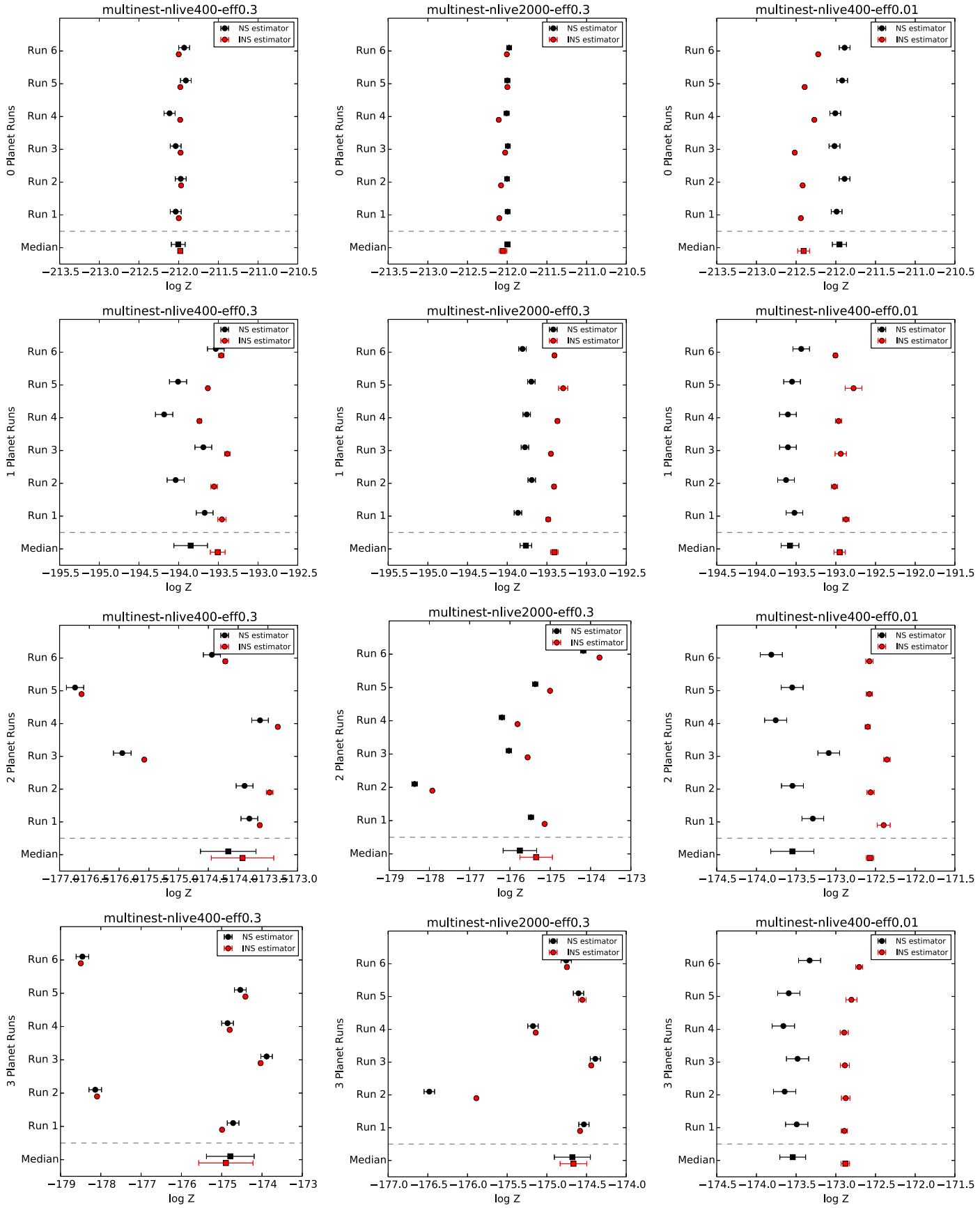

Figure 7. Scattering of MULTINEST $\log \widehat{\mathcal{Z}}$ estimates from runs against data set 1. Panels show our three MULTINEST configurations (columns) and number of planets used (rows). The Nested Sampling (NS) and Importance Nested Sampling (INS) estimates are shown in black and red, respectively. Scattering between estimates is often larger than the quoted uncertainties. Also, there are outliers. The multirun estimator (median) is shown at the bottom of each plot.

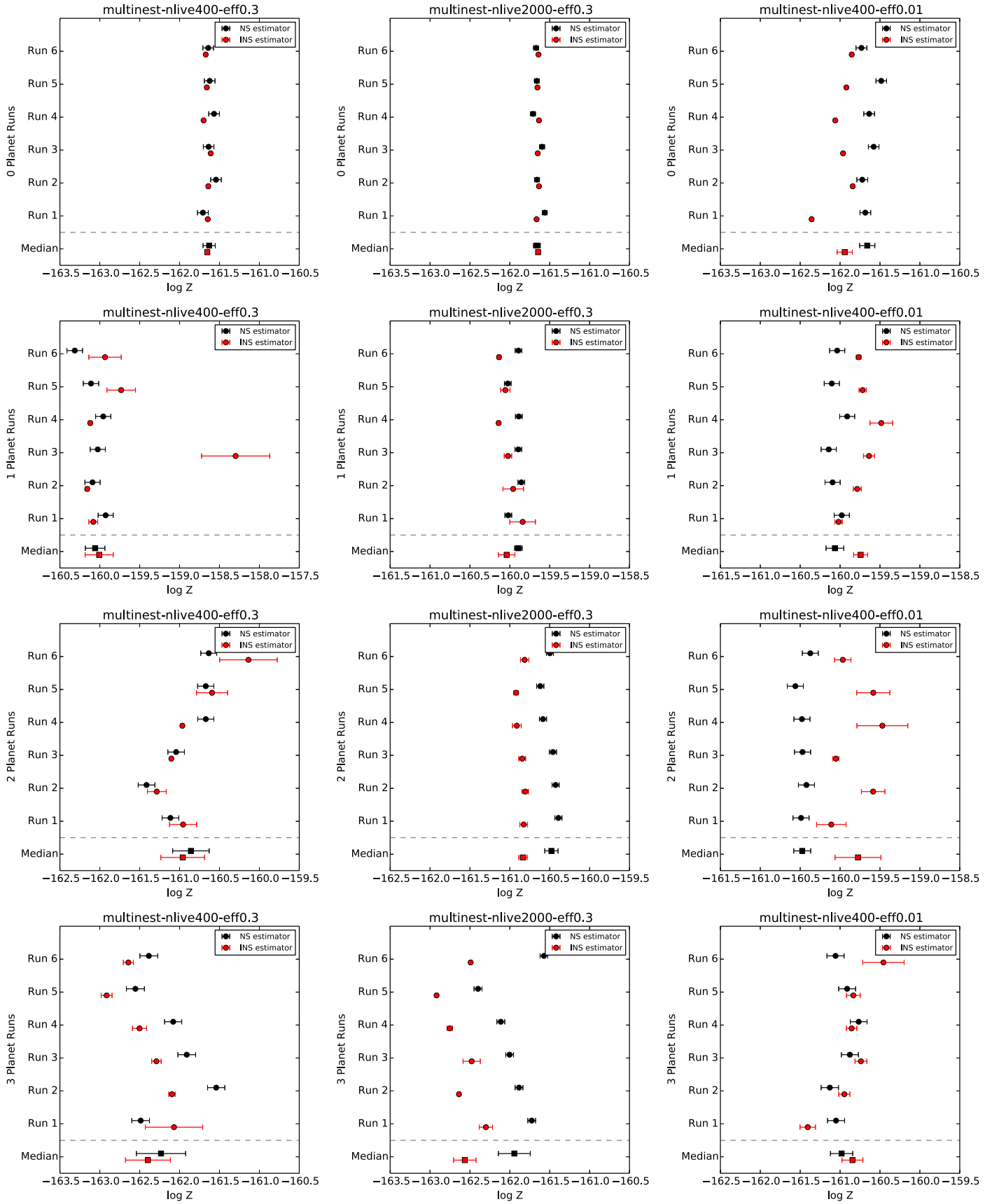

Figure 8. Same as Figure 7, but for data set 4.

also capable of efficiently exploring difficult constrained distributions. The algorithm introduces a slight but important improvement to the classic NS approach, in that it attempts to sample from a mixture of successively constrained distributions, instead of using one single hard constraint at each step.

DNS starts by generating a particle from the prior (call this distribution $p_{\mathcal{L}_0}$) and evolving it with an MCMC, storing all the intermediate likelihood values. After a given number of iterations, it finds the $1 - e^{-1} \sim 63\%$ quantile of all the likelihood values and records it as $\mathcal{L}_1$; this creates a new level occupying about e^{-1} times the mass of $p_{\mathcal{L}_0}$. All of the likelihood values lower than $\mathcal{L}_1$ are then discarded. At this point, (classic) NS would continue sampling from the prior constrained to $\mathcal{L}_1$ (call it $p_{\mathcal{L}_1}$). In contrast, DNS attempts to sample from a weighted sum of the two distributions $p_{\mathcal{L}_0}$ and $p_{\mathcal{L}_1}$. An MCMC is used to evolve the particle with this mixture of distributions as the target, and once enough samples have been obtained from $p_{\mathcal{L}_1}$, we again find the $1 - e^{-1}$ quantile of all the likelihood values and record it as $\mathcal{L}_2$. Likelihood values smaller than $\mathcal{L}_2$ are removed. The particle then explores a mixture of $p_{\mathcal{L}_0}$, $p_{\mathcal{L}_1}$, and $p_{\mathcal{L}_2}$, and this process continues until a maximum number of levels is created.

Once all the levels have been obtained, the particle simply continues to explore the mixture of all the levels until the algorithm is terminated. In order to create the mixture of distributions, we need to provide a weighting scheme for each component. Simple uniform weights for all distributions would work, albeit inefficiently. Brewer et al. (2011) proposed exponentially decaying weights with a scale length Λ, which describes how far (down in likelihood) the particle is able to go in order to explore more freely. When the desired number of levels has been created, the weights can be changed to uniform, and further samples are drawn from all the component distributions. The algorithm can then continue to sample for as long as required, with the evidence and posterior samples converging to their true values. Each time a new level is created, its constrained distribution covers about e^{-1} times as much prior mass as the last distribution. Therefore, the X-value of the kth level can be estimated as $\exp(-k)$. However, as the levels are being created, their actual X values can be modified from this theoretical expectation. This means that the weight of each distribution is actually different, and the exploration is thus not completely correct. The X values can nevertheless be corrected. At a given level k, the values of the likelihood will be higher than the upper level's likelihood cutoff a fraction X_{k+1}/X_k of the time. Thus, we can use the actual fraction of samples in which this happens as an estimate of the true ratio of the X values for consecutive levels.

In summary, the DNS algorithm is essentially an application of the Metropolis–Hastings algorithm to a distribution other than the posterior. Changing the target distribution improves upon other MCMC algorithms by providing the value of the evidence in one single run and being less sensitive to the presence of complicated features in the posterior. Classic NS also shares these advantages, but DNS improves upon the classic algorithm by alleviating the problem of sampling from the likelihood-constrained prior. Because the target distribution used by DNS always includes the prior distribution as one of the components of the mixture, sampling from posteriors with substantial multimodality is still possible and even efficient.

A.10.1. Details

In this work, Faria used the DNS algorithm implemented in the DNEST4 package (Brewer & Foreman-Mackey 2018). The specific application of DNEST4 to the exoplanet problem is implemented in a new open-source package called `kima` (J. P. Faria et al. 2019, in preparation). The code allows the posterior distribution for the orbital parameters and the value of the evidence for a model $\mathcal{M}_n$ with n planets to be calculated.

The DNS algorithm has a few options, which need to be set for each run. We set the scale length Λ to 25 and require 500 samples from the consecutively constrained distributions before creating a new level. The maximum number of levels is determined automatically by DNEST4 (see Brewer et al. 2011). For all the simulated data sets, we obtained 100,000 samples from the DNS target distribution. This corresponds to different numbers of posterior samples for each data set and for each model.

In the DNS algorithm, there is no explicit global search step as the algorithm is always free to explore the full prior volume. This means that once the settings mentioned above are fixed, the results were computed automatically for all data sets, without any data-set-dependent input.

For the analysis with constrained priors for the orbital period, the prior probability density function was set to 0 outside of the provided period bounds. Inside the bounds, the prior is still a Jeffreys between 1.25 and 10^4 days.

The error we report for the evidence value is calculated from one single run, by the probabilistic reassignment of X values to the samples, as in standard NS (see Brewer et al. 2011). These errors are likely to be overly optimistic.

ORCID iDs

Benjamin E. Nelson https://orcid.org/0000-0003-3010-2334
Eric B. Ford https://orcid.org/0000-0001-6545-639X
Johannes Buchner https://orcid.org/0000-0003-0426-6634
Ryan Cloutier https://orcid.org/0000-0001-5383-9393
Rodrigo F. Díaz https://orcid.org/0000-0001-9289-5160
João P. Faria https://orcid.org/0000-0002-6728-244X
Vinesh M. Rajpaul https://orcid.org/0000-0001-7576-6703
Surangkhana Rukdee https://orcid.org/0000-0001-5423-1005

References

Akaike, H. 1974, ITAC, 19, 716
Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2015, ITPAM, 38, 252
Arlot, S., & Celisse, A. 2010, Stat. Surv., 4, 40
Astudillo-Defru, N., Díaz, R. F., Bonfils, X., et al. 2017, A&A, 605, L11
Baluev, R. V. 2008, MNRAS, 385, 1279
Bastien, F. A., Stassun, K. G., Pepper, J., et al. 2014, AJ, 147, 29
Beaujean, F., & Caldwell, A. 2013, arXiv:1304.7808
Betancourt, M. 2017, arXiv:1701.02434
Bonfils, X., Astudillo-Defru, N., Díaz, R., et al. 2018, A&A, 613, A25
Bourguignon, S., Carfantan, H., & Böhm, T. 2007, A&A, 462, 379
Brewer, B. J., & Donovan, C. P. 2015, MNRAS, 448, 3206
Brewer, B. J., & Foreman-Mackey, D. 2018, J. Stat. Softw., 86, 1
Brewer, B. J., Pártay, L. B., & Csányi, G. 2011, S&C, 21, 649
Buchner, J. 2016, S&C, 26, 383
Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125
Butler, R. P., Vogt, S. S., Laughlin, G., et al. 2017, AJ, 153, 208
Cameron, E., & Pettitt, A. 2014, Stat. Sci., 29, 397
Cegla, H. M., Stassun, K. G., Watson, C. A., Bastien, F. A., & Pepper, J. 2014, ApJ, 780, 104

Chen, S. S., Donoho, D. L., & Saunders, M. A. 1998, SIAM J. Sci. Comput., 20, 33
Chib, S., & Jeliazkov, I. 2001, J. Am. Stat. Assoc., 96, 270
Cloutier, R., Astudillo-Defru, N., Doyon, R., et al. 2017, A&A, 608, A35
Dawson, R. I., & Fabrycky, D. C. 2010, ApJ, 722, 937
Díaz, R. F., Rey, J., Demangeon, O., et al. 2016a, A&A, 591, A146
Díaz, R. F., Ségransan, D., Udry, S., et al. 2016b, A&A, 585, A134
Dumusque, X. 2016, A&A, 593, A5
Dumusque, X., Borsa, F., Damasso, M., et al. 2017, A&A, 598, A133
Faria, J. P., Haywood, R. D., Brewer, B. J., et al. 2016, A&A, 588, A31
Feroz, F., & Hobson, M. P. 2014, MNRAS, 437, 3540
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, OJAp, 2, 10
Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, PASP, 128, 066001
Ford, E. B., & Gregory, P. C. 2007, in ASP Conf. Ser. 371, Statistical Challenges in Modern Astronomy IV, ed. G. J. Babu & E. D. Feigelson (San Francisco, CA: ASP), 189
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Ge, D., Jiang, X., & Ye, Y. 2011, Math. Program., 129, 285
Gelman, A., Hwang, J., & Vehtari, A. 2014, Statistics and Computing, 24, 997
Gelman, A., & Rubin, D. B. 1992, StaSc, 7, 457
Goodman, J., & Weare, J. 2010, CAMCS, 5, 65
Gregory, P. C. 2007, MNRAS, 381, 1607
Guo, P.-C. 2012, PhD thesis, Univ. Florida
Haario, H., Laine, M., Mira, A., & Saksman, E. 2006, Statistics and Computing, 16, 339
Hara, N. C., Boué, G., Laskar, J., & Correia, A. C. M. 2017, MNRAS, 464, 1220
Haywood, R. D., Cameron, A. C., Queloz, D., et al. 2014, IJAsB, 13, 155
Hou, F., Goodman, J., & Hogg, D. W. 2014, arXiv:1401.6128
Hunter, J. D. 2007, CSE, 9, 90
Jahn, S., Beaujean, F., & Straub, D. 2018, fredRos/pypmc: v1.1.2 Better build support, Zenodo, doi:10.5281/zenodo.1158068
Jeffreys, H. 1998, International Series of Monographs on Physics (3rd ed.; Oxford: Oxford Univ. Press)
Jenkins, J. S., Jones, H. R. A., Tuomi, M., et al. 2017, MNRAS, 466, 443
Jones, D. E., Stenning, D. C., Ford, E. B., et al. 2017, arXiv:1711.01318
Kane, S. R., Thirumalachari, B., Henry, G. W., et al. 2016, ApJL, 820, L5
Kass, R. E., & Raftery, A. E. 1995, J. Am. Stat. Assoc., 90, 773
Konishi, S., & Kitagawa, G. 2008, Information Criteria and Statistical Modeling (New York: Springer)
Lomb, N. R. 1976, Ap&SS, 39, 447
Millholland, S., Laughlin, G., Teske, J., et al. 2018, AJ, 155, 106
Nelson, B. E., Robertson, P. M., Payne, M. J., et al. 2016, MNRAS, 455, 2484
Perrakis, K., Ntzoufras, I., & Tsionas, E. G. 2014, Computational Statistics Data Analysis, 77, 54
Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, MNRAS, 452, 2269
Rajpaul, V., Aigrain, S., & Roberts, S. 2016, MNRAS, 456, L6
Robertson, P., & Mahadevan, S. 2014, ApJL, 793, L24
Rue, H., Riebler, A., Sørbye, S. H., et al. 2017, AnRSA, 4, 395
Scargle, J. D. 1982, ApJ, 263, 835
Shaw, J. R., Bridges, M., & Hobson, M. P. 2007, MNRAS, 378, 1365
Skilling, J. 2004, in AIP Conf. Proc. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. R. Fischer, R. Preuss, & U. von Toussaint (Melville, NY: AIP), 395
Skilling, J. 2006, BayAn, 1, 833
Van Den Berg, E., & Friedlander, M. P. 2008, SIAM Journal on Scientific Computing, 31, 890
Veitch, J., & Vecchio, A. 2010, PhRvD, 81, 062003
Watanabe, S. 2013, Journal of Machine Learning Research, 14, 867