Semi-supervised standardized detection of extrasolar planets

S. Sulis, D. Mary, L. Bigot, M. Deleuil

HAL Id: insu-03863229
https://insu.hal.science/insu-03863229

Submitted on 21 Nov 2022

Semi-supervised standardized detection of extrasolar planets

S. Sulis[1] , D. Mary[2], L. Bigot[2], and M. Deleuil[1]

[1] Université Aix-Marseille, CNRS, CNES, LAM, 13007 Marseille, France
e-mail: sophia.sulis@lam.fr
[2] Université Côte-d'Azur, Observatoire de la Côte d'Azur, CNRS, Lagrange UMR 7293, CS 34229, 06304, Nice Cedex 4, France

ABSTRACT

Context. The detection of small exoplanets with the radial velocity (RV) technique is limited by various poorly known noise sources of instrumental and stellar origin. As a consequence, current detection techniques often fail to provide reliable estimates of the significance levels of detection tests in terms of false-alarm rates or p-values.

Aims. We designed an RV detection procedure that provides reliable p-value estimates while accounting for the various noise sources typically affecting RV data. The method is able to incorporate ancillary information about the noise (e.g., stellar activity indicators) and specific data- or context-driven data (e.g. instrumental measurements, magnetohydrodynamical simulations of stellar convection, and simulations of meridional flows or magnetic flux emergence).

Methods. The detection part of the procedure uses a detection test that is applied to a standardized periodogram. Standardization allows an autocalibration of the noise sources with partially unknown statistics (algorithm 1). The estimation of the p-value of the test output is based on dedicated Monte Carlo simulations that allow handling unknown parameters (algorithm 2). The procedure is versatile in the sense that the specific pair (periodogram and test) is chosen by the user. Ancillary or context-driven data can be used if available.

Results. We demonstrate by extensive numerical experiments on synthetic and real RV data from the Sun and αCenB that the proposed method reliably allows estimating the p-values. The method also provides a way to evaluate the dependence of the estimated p-values that are attributed to a reported detection on modeling errors. It is a critical point for RV planet detection at low signal-to-noise ratio to evaluate this dependence. The python algorithms developed in this work are available on GitHub.

Conclusions. Accurate estimation of p-values when unknown parameters are involved in the detection process is an important but only recently addressed question in the field of RV detection. Although this work presents a method to do this, the statistical literature discussed in this paper may trigger the development of other strategies.

Key words. techniques: radial velocities – stars: activity – planets and satellites: detection – methods: statistical

1. Introduction

When radial velocity (RV) time series are analysed, exoplanet detection tests aim at deciding whether the time series contains one or several planetary signatures plus noise (the alternative hypothesis, noted $\mathcal{H}_1$), or only noise (the null hypothesis, noted $\mathcal{H}_0$). With the recent advent of very stable and high-precision spectrographs (Pepe et al. 2010; Jurgenson et al. 2016), instrumental noise has reached levels that are so low (Fischer et al. 2016) that the noise generated by stellar activity has become the limiting factor for the detection of Earth-like planets. Stellar noise is frequency dependent (colored) because stellar variability results from several phenomena that contribute simultaneously to the observations and evolve at different timescales. The main sources of stellar noise that have been identified so far include magnetic cycles, starspots, faculae, inhibition of the convective blueshift by plages, meridional circulation, convective motions of the stellar atmosphere (granulation and supergranulation), and acoustic oscillations. The overall stellar noise affects the frequency range in which planets can be found and may dominate the ~10 cm s^{-1} amplitude level that is typical of an Earth-like exoplanet orbiting a Sun-like star at 1 AU.

The development of techniques to mitigate stellar activity for the study of exoplanets is an active field of research. Several methods and tools have been developed to account for these different sources in the RV time series (see Meunier 2021 for

a recent review). First, some specific observation strategies are commonly used to reduce the contribution of short-timescale activity components (Hatzes et al. 2011; Dumusque et al. 2011). Second, to correct for variability that evolves at the stellar rotation period (spots and plages), the most basic technique consists of fitting a model to the RV dataset (e.g., sinusoidal functions at the stellar rotation period (Boisse et al. 2011), or quasi-periodic Gaussian processes (see, e.g., Aigrain et al. 2012; Baluev 2011, 2013b; Rajpaul et al. 2015, 2021; Jones et al. 2022)). To constrain the parameters of these models, activity indicators (e.g., the S-index, Wilson 1968; log R'_{HK}, Noyes et al. 1984; or the bisector span, Queloz et al. 2001) are commonly used if correlations with the RV dataset are found. Some relations with photometry (Aigrain et al. 2012; Yu et al. 2018) or the study of the chromatic dependence of the stellar activity (Meunier et al. 2017; Tal-Or et al. 2018; Dumusque 2018) can also be used. Recent advances in RV extraction processes from stellar spectra in particular are very promising (Davis et al. 2017; Collier Cameron et al. 2021; Cretignier et al. 2021; de Beurs et al. 2022).

However, despite the diversity of existing techniques to reduce the stellar noise, residual noise sources at unknown levels translate into uncertainties in the interpretation of the RV detection tests. In practice, it is difficult to provide a reliable estimation of the significance level of a reported detection. For this reason, examples of a debated detection are numerous (see as examples the cases of HD 41248 b and c, Jenkins & Tuomi 2014;

Santos et al. 2014; HD 73256 b, Udry et al. 2003; Ment et al. 2018; AD Leo b, Tuomi et al. 2018; Carleo et al. 2020; Aledebaran c, Hatzes et al. 2015; Reichert et al. 2019; or Kapteyn b and c, Anglada-Escude et al. 2014; Bortle et al. 2021). We discuss the particular case of αCenBb at the end of this paper as an example (Dumusque et al. 2012; Hatzes 2013; Rajpaul et al. 2016; Toulis & Bean 2021).

The main question we address here is the evaluation of reliable significance levels of detection tests in the presence of unknown colored noise. Significance levels are often interpreted through analyses of false-alarm probabilities (FAP) or of p-values. The former are fixed before the tests are conducted, while the latter measure a degree of surprise of the observed test statistic under the null hypothesis. In this work, we consider p-values. The estimation of these values has been amply discussed in the literature for the case when the noise model contains several unknown parameters (Bayarri & Berger 2000). Connections of our approach with the literature are discussed in Sect. 3.3. An accurate estimation of the p-values is critical for designing reliable detection methods of small-planet RV signatures.

Because the sampling grid is irregular, an RV detection is most often performed through the Lomb–Scargle periodogram (LSP; Scargle 1982) or variants of the periodograms (see, e.g., Ferraz-Mello 1981; Cumming et al. 1999; Cumming 2004; Reegen 2007; Bourguignon et al. 2007; Zechmeister & Kürster 2009; Baluev 2013a; Jenkins et al. 2014; Tuomi et al. 2014, Gregory 2016, Hara et al. 2017, 2022). While the marginal distribution of the periodogram components at each frequency is often known for white Gaussian noise (WGN) of known variance (e.g., a χ^2 distribution for the LSP), the joint distribution of the periodogram components is much more difficult to characterize because the components exhibit dependences that are dictated by the sampling grid (or equivalently, by the spectral window).

In practice, specific noise models are applied to correct RV data for the activity, and the p-values of the test statistics are evaluated on the RV residuals assuming they follow the statistics of a WGN, or that they have a known covariance matrix. When the noise is considered WGN, some methods are based on approximate analytical expressions (Baluev 2008). Other methods compute the FAP using the formula that would be obtained for independent components, where an ad hoc number of independent frequencies is plugged in (see Horne & Baliunas 1986; Schwarzenberg-Czerny 1998; Cumming 2004, as well as Süveges 2014 and Süveges et al. 2015 for a critical analysis of this approach). Finally, another family of approaches is based on Monte Carlo or bootstrap techniques (Paltani 2004; Schwarzenberg-Czerny 2012), sometimes aided by results from extreme-value distributions (Süveges 2014; Süveges et al. 2015).

In the large majority of cases, however, the noise in RV residuals is not white, but colored and with an unknown power spectral density (PSD). Providing reliable FAP estimates in this situation remains, to our knowledge, an open problem. The assumption that the noise PSD is flat leads logically to increased false alarms in the PSD bumps, and to a loss of detection power in the PSD valleys (e.g., see Fig. 2 in Sulis et al. 2016). This problem is only partially solved in practical approaches where a covariance model is fitted to the data and used to produce residuals that are assumed to be white. For instance, Delisle et al. (2020) recently presented an analytical formulation of the p-value resulting from the generalization of Baluev (2008) to the case of colored noise when the covariance matrix is known. Their results indicate that the FAP estimates are highly dependent on the considered covariance matrix, with variations

directly related to the mismatch between the true and assumed covariance matrix. Similar effects were analyzed in Sulis et al. (2016) and in Sect. 3.5 of Sulis (2017). To our knowledge, all methods aimed at evaluating the FAP or p-values associated with periodogram tests assume that the noise is eventually white, either directly or because the model used for the covariance is considered exact, so that the signal can be whitened. Since estimation errors in the noise statistics lead to fluctuations in the p-value estimates, our incomplete knowledge of the noise sources and their statistics represents a fundamental limitation to how much we can trust these estimates. The present work deals precisely with this problem. Another critical point is the consistency of the p-value estimate when different noise models are used in the detection process (this point was very clearly raised by Rajpaul et al. 2016, for instance). The controversy about a low-mass planet detection often resides in the different noise models that are assumed for the data. Because i) noise models can never be exact and ii) the estimated p-values are model dependent, incorporating model errors in the estimation of these values appears to be a relevant (but new) approach, and we follow this path below.

In this paper, we propose a new RV planet detection procedure that can generate reliable estimates of the significance levels of detection tests by means of p-values. The approach is built on a previous study (Sulis et al. 2020), which was limited to the case of regularly sampled time series and primarily considered the effect of granulation noise. Both of these limitations are removed in the present paper.

The approach is flexible and can deal with the main characteristics that are encountered in practice with RV detection: irregular sampling and various noise sources with partially unknown noise statistics. The basic principle of the procedure is to propagate errors on the noise model parameters to derive accurate p-values of detection tests by means of intensive Monte Carlo simulations. The choice of the noise models involved in the procedure is entirely left to the user, as is the choice of the periodogram and detection test. We note that without loss of generality, the procedure can be applied to an RV dataset resulting from any extraction process (including, e.g., the extraction techniques recently described in Collier Cameron et al. 2021 or Cretignier et al. 2021). In addition, the procedure can exploit training time series of stellar or instrumental noise, if these data are available, for instance, through auxiliary observations or realistic simulations.

The core of the new detection approach is presented in Sect. 2 (algorithm 1). The numerical procedure for estimating the p-values of the test statistics produced by algorithm 1 is presented in Sect. 3 (algorithm 2). Validations of the p-value estimation procedure under different configurations are given in Sects. 4 and 5. We conclude with the application of the detection procedure to the RV data of αCenB, where an Earth-mass planet detection has been debated (Sect. 6). In this last section, we show that the proposed detection procedure can be used to evaluate the robustness of a planet detection with respect to specific model errors. Examples of how to use the algorithms we developed in this work are available on GitHub[1], and practical details are given in Appendix A.

2. Semi-supervised standardized detection

Our notations are as follows: Bold small (capital) letters without subscript indicate column vectors (matrices). Vector (matrix)

---

[1] https://github.com/ssulis/3SD

components are denoted by subscripts. For instance, we write vector $\mathbf{x}$ as $\mathbf{x} = [\mathbf{x}_1, \ldots \mathbf{x}_N]^\top$. We denote by $\widehat{z}$ the estimate of a quantity z. Throughout the paper, the notation | means "conditional on" (and it is not necessarily used for probability densities). Appendix B summarizes the main notations of this work.

2.1. Hypothesis testing

We consider an RV time series under test, $\mathbf{x}$, irregularly sampled at N time instants $\mathbf{t} = [t_1, t_2, \ldots, t_N]^\top$. Our detection problem can be cast as a binary hypothesis problem of the form

$$\begin{cases} \mathcal{H}_0 : \mathbf{x} = \mathbf{d}|\mathcal{M}_{\mathbf{d}}(\boldsymbol{\theta}_{\mathbf{d}}) + \mathbf{n}, \\ \mathcal{H}_1 : \mathbf{x} = \mathbf{s}|\mathcal{M}_{\mathbf{s}}(\boldsymbol{\theta}_{\mathbf{s}}) + \mathbf{d}|\mathcal{M}_{\mathbf{d}}(\boldsymbol{\theta}_{\mathbf{d}}) + \mathbf{n}. \end{cases} \quad (1)$$

The signal $\mathbf{d}$ accounts for various effects that typically affect RV data. A nonexhaustive list of such effects includes long-term instrumental trends, stellar magnetic activity, or RV trends due to an additional companion in the system. This nuisance signal, which is the unknown mean under $\mathcal{H}_0$, cannot be estimated by numerical simulations. While $\mathbf{d}$ can be of random origin, it is customary to model it as a deterministic signature (see, e.g., Dumusque et al. 2012). This is our approach in the following. We refer to $\mathbf{d}$ as a nuisance signal and write generically that $\mathbf{d}$ is generated from some model $\mathcal{M}_{\mathbf{d}}$ with parameters $\boldsymbol{\theta}_{\mathbf{d}}$.

The noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a zero-mean stochastic noise of unknown covariance matrix $\boldsymbol{\Sigma}$ that we decompose into

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{c}} + \boldsymbol{\Sigma}_w, \quad (2)$$

with $\boldsymbol{\Sigma}_w$ a diagonal covariance matrix whose entries account for possibly different uncertainties on the RV measurement. If they are equal, $\boldsymbol{\Sigma}_w = \sigma_{\mathbf{w}}^2 \boldsymbol{I}$, with $\sigma_{\mathbf{w}}^2$ the variance of a WGN (noted $\mathbf{w}$ below), and $\boldsymbol{\Sigma}_{\mathrm{c}}$ the unknown covariance matrix of a colored noise We note that, in this paper, $\boldsymbol{\Sigma}_{\mathrm{c}}$ is considered fixed. Temporal variations of $\boldsymbol{\Sigma}_{\mathrm{c}}$ such as those considered by Baluev (2015) might in principle be considered in our numerical procedure by sampling according to a time-dependent covariance structure, but this is beyond the scope of this paper. A nonexhaustive list of effects that can be included in $\mathbf{n}$ entails meridional circulation, magnetic flux emergence, stellar granulation, supergranulation, and instrumental noise. The choice of which type of RV noise component fits in $\mathbf{d}$ or $\mathbf{n}$ depends on the assumptions made by the user about the RV time series under test.

Under the alternative hypothesis (one or more planets are present), time series $\mathbf{x}$ contains the same noise components $\mathbf{d}$ and $\mathbf{n}$, plus a deterministic signal $\mathbf{s}$ corresponding to the Keplerian signatures of planets (model $\mathcal{M}_{\mathbf{s}}$). The number of planets and their Keplerian parameters $\boldsymbol{\theta}_{\mathbf{s}}$ are unknown.

Throughout this paper, we consider the optional case in which context-driven data of the stochastic colored noise $\mathbf{n}$ in Eq. (1) can be made available through realistic simulations, for instance. We call these data the null training sample (NTS). By construction, this NTS is free of any possible contamination by the planetary signal.

For example, focusing on the stellar granulation noise, Sulis et al. (2020) showed that 3D magnetohydrodynamical (MHD) simulations can generate realistic synthetic time series of this RV noise source (see also Appendix C). This was also demonstrated in Palumbo et al. (2022) with synthetic stellar spectra

simulators. This suggests that using simulations of stellar granulation in the detection process could provide an alternative to the observational strategy of binning spectra over a night (Dumusque et al. 2011). Meunier et al. (2015) and Meunier & Lagrange (2019) showed that while convection (granulation and supergranulation) dominates at high frequencies, where it acts as a power-law noise ($\sim 1/f$), convection has energy at all frequencies, so that averaging on short timescales is not efficient enough to detect RV signatures of Earth-like planets that are hidden in the convection noise.

Other examples in which NTS can be obtained are 3D MHD simulations of stellar supergranulation (e.g., Rincon & Rieutord 2018), empirical simulations of stellar meridional flows (Meunier & Lagrange 2020), or specific data-driven observations representative of the instrumental noise (e.g., microtelluric lines; Cunha et al. 2014 Seifahrt et al. 2010) affecting the RV data. In general, all realistic synthetic time series of the stochastic colored noise sources affecting the RV data can be included in the NTS.

This setting poses the general question how this NTS might be leveraged to improve the detection process, both in terms of detection power and of the accuracy of the computed p-value. This approach was proposed and analyzed in Sulis et al. (2017a), but with $\mathbf{d} = \mathbf{0}$ and a regular sampling for $\mathbf{x}$ in model (1). In this approach, the detection test was automatically calibrated with the NTS through a standardization of the periodogram. While this preliminary work has demonstrated improved performances in terms of power and control of the p-value, it presents two important limitations for RV detection purposes. First, the approach accounts only for noise that can be generated through realistic simulations. Second, the scope was limited to regular sampling. In contrast, the detection procedure proposed below can cope with all components of noise, including instrumental and stellar noises of different origins ($\mathbf{d} \neq \mathbf{0}$) and any type of time sampling. This general procedure is semi-supervised in the sense that it incorporates ancillary sources of noise that affect the data. Because it is also based on a periodogram standardization, we call this procedure semi-supervised standardized detection (3SD for short). The 3SD procedure contains two algorithms: one for the detection, and one for evaluating the resulting significance levels.

2.2. Periodogram standardization

2.2.1. An NTS is available

We assume that the NTS, if available, takes the form of L sequences of synthetic time series of noise $\mathbf{n}$ in Eq. (1),

$$\mathcal{T}_{\mathrm{L}} := \{\mathbf{n}^{(i)}\}, i = 1, \ldots, L, \quad \mathbf{n}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (3)$$

An efficient way to use this side information in the detection process is to compute standardized periodograms (see Sulis et al. 2020). Let Ω denote the set of frequencies at which the periodogram, noted $\mathbf{p}$, is computed (Ω and the type of periodogram P are defined by the user). We note ν_k these frequencies and N_Ω their number. The periodograms of each of the L series of the NTS, $\mathbf{p}_\ell(\nu_k)$, with $\ell = 1, \ldots, L$ can be used to form an averaged periodogram, $\overline{\mathbf{p}}_{\mathrm{L}}(\nu_k)$, as the sample mean of these L periodograms,

$$\overline{\mathbf{p}}_{\mathrm{L}}(\nu_k) := \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{p}_\ell(\nu_k). \quad (4)$$

The standardized periodogram is then defined as

$$\widetilde{\mathbf{p}}(\nu_k) := \frac{\mathbf{p}(\nu_k)}{\overline{\mathbf{p}}_{\mathrm{L}}(\nu_k)}. \qquad (5)$$

In the following, $\mathbf{p}$ and $\overline{\mathbf{p}}_{\mathrm{L}}$ denote the vectors concatenating the periodogram ordinates at all frequencies in Ω, and the standardized periodogram is noted $\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_{\mathrm{L}}$ or simply $\widetilde{\mathbf{p}}$.

In the case of regular sampling and when $\mathbf{d} = \mathbf{0}$, Sulis et al. (2020) demonstrated that standardization avoids both a loss of detection power in valleys of the noise PSD and an increase in false alarm rate in its peaks (see Sulis et al. 2016, and Sect.3.5.3 of Sulis 2017). An important point of the present paper is to show that these desirable features also hold in the case of irregular sampling for different models of $\mathbf{d}$ and $\mathbf{n}$ (see Sects. 4 and 5).

2.2.2. No NTS is available

Simulations (e.g., 3D MHD simulations of stellar granulation) are computationally demanding, and generating very long time series (months or years) may therefore appear beyond reach for standard RV planet searches. It is also worth mentioning that helioseismology has revealed that 3D MHD simulations are unable to reproduce the power of large-scale flows on the Sun (supergranulation), which means that improvements are still needed in this field of research (Hanasoge et al. 2012). For these reasons, we also consider the case in which no NTS of $\mathbf{n}$ in Eq. (1) is available. In this case, the periodogram standardization relies on a parametric model $\mathcal{M}_{\mathbf{n}}$ whose parameters are estimated from the RV data. We note that, when an NTS is available, model $\mathcal{M}_{\mathbf{n}}$ is not necessary for the detection process in algorithm 1. Therefore, it was not made explicit in the notation used in model (1). In this approach, synthetic time series of $\mathbf{n}$ are generated and used to compute the averaged periodogram used in Eq. (4). Numerous examples of stochastic parametric models exist, and a (nonexhaustive) relevant list for RV data includes Harvey-like functions (Harvey 1985), autoregressive (AR) models (Brockwell & Davis 1991; Elorrieta et al. 2019), or Gaussian processes (Rasmussen & Williams 2005).

A particular noise configuration happens when noise $\mathbf{n}$ is white ($\mathbf{\Sigma}_{\mathbf{c}} = 0$, leading to $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$ in Eq. (2)). In this case, the standardization of the periodogram in Eq. (5) is simply performed by the estimated variance of the RV time series.

2.3. Considered detection procedure

The detection procedure proposed in this work is described in Fig. 1. We refer to this procedure as algorithm 1 in the following.

The first step (rows 1–4) aims at computing the periodogram $\mathbf{p}$ involved in Eq. (5). For this purpose, if the nuisance component $\mathbf{d}$ is present in Eq. (1), the algorithm starts by removing it from the time series $\mathbf{x}$ by fitting model $\mathcal{M}_{\mathbf{d}}$ (blue block). The residuals are used to compute $\mathbf{p}$ (row 4). If $\mathbf{d} = \mathbf{0}$ is not present in Eq. (1), then the blue block is ignored, and we directly compute $\mathbf{p}$ with the data under test ($\mathbf{x}$).

In this first step, the algorithm removes the nuisance signals, and detection is performed on the residuals. A more general detection approach such as the generalized likelihood ratio (Kay 1998) would require jointly estimating all unknown parameters under both hypotheses. This approach would lead to different nuisance parameters under $\mathcal{H}_0$ and $\mathcal{H}_1$ because the parameters of the nuisance signals are consecutively estimated with or without those of the signal (sinusoidal or Keplerian parameters). This procedure can be computationally very expensive
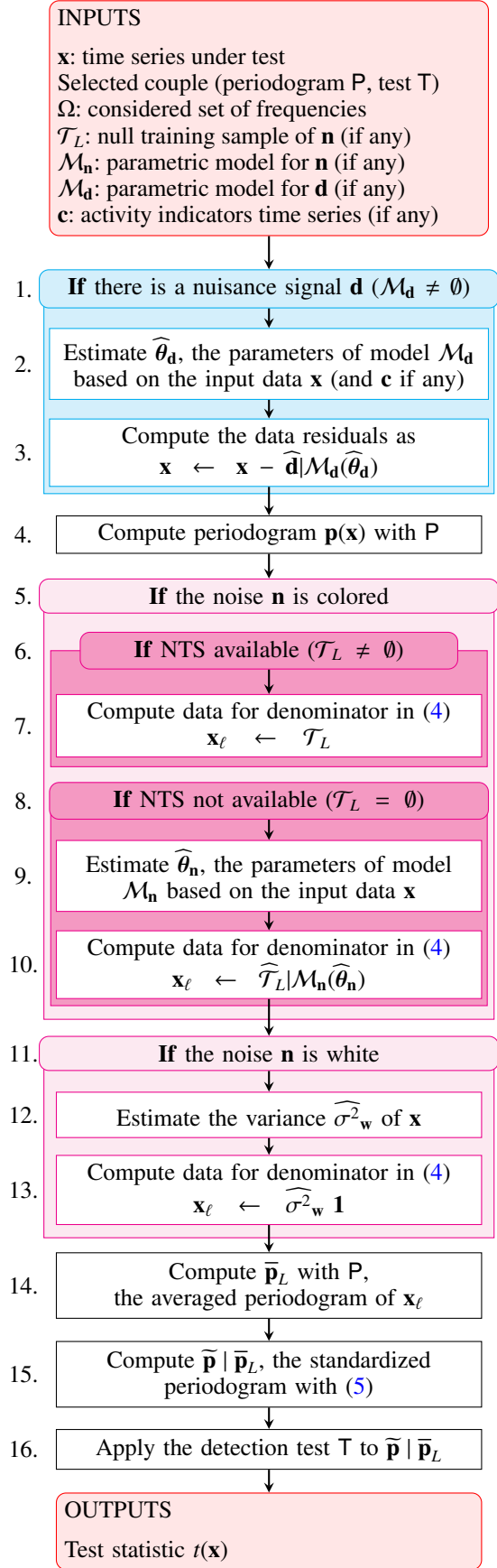
INPUTS

$\mathbf{x}$: time series under test
Selected couple (periodogram P, test T)
Ω: considered set of frequencies
$\mathcal{T}_L$: null training sample of $\mathbf{n}$ (if any)
$\mathcal{M}_{\mathbf{n}}$: parametric model for $\mathbf{n}$ (if any)
$\mathcal{M}_{\mathbf{d}}$: parametric model for $\mathbf{d}$ (if any)
$\mathbf{c}$: activity indicators time series (if any)

1. **If** there is a nuisance signal $\mathbf{d}$ ($\mathcal{M}_{\mathbf{d}} \neq \emptyset$)

2. Estimate $\widehat{\theta_{\mathbf{d}}}$, the parameters of model $\mathcal{M}_{\mathbf{d}}$ based on the input data $\mathbf{x}$ (and $\mathbf{c}$ if any)

3. Compute the data residuals as
$\mathbf{x} \;\leftarrow\; \mathbf{x} - \widehat{\mathbf{d}} | \mathcal{M}_{\mathbf{d}}(\widehat{\theta_{\mathbf{d}}})$

4. Compute periodogram $\mathbf{p}(\mathbf{x})$ with P

5. **If** the noise $\mathbf{n}$ is colored

6. **If** NTS available ($\mathcal{T}_L \neq \emptyset$)

7. Compute data for denominator in (4)
$\mathbf{x}_\ell \;\leftarrow\; \mathcal{T}_L$

8. **If** NTS not available ($\mathcal{T}_L = \emptyset$)

9. Estimate $\widehat{\theta_{\mathbf{n}}}$, the parameters of model $\mathcal{M}_{\mathbf{n}}$ based on the input data $\mathbf{x}$

10. Compute data for denominator in (4)
$\mathbf{x}_\ell \;\leftarrow\; \widehat{\mathcal{T}_L} | \mathcal{M}_{\mathbf{n}}(\widehat{\theta_{\mathbf{n}}})$

11. **If** the noise $\mathbf{n}$ is white

12. Estimate the variance $\widehat{\sigma^2_{\mathbf{w}}}$ of $\mathbf{x}$

13. Compute data for denominator in (4)
$\mathbf{x}_\ell \;\leftarrow\; \widehat{\sigma^2_{\mathbf{w}}} \mathbf{1}$

14. Compute $\overline{\mathbf{p}}_L$ with P, the averaged periodogram of $\mathbf{x}_\ell$

15. Compute $\widetilde{\mathbf{p}} | \overline{\mathbf{p}}_L$, the standardized periodogram with (5)

16. Apply the detection test T to $\widetilde{\mathbf{p}} | \overline{\mathbf{p}}_L$

OUTPUTS

Test statistic $t(\mathbf{x})$

Fig. 1. Detection part of the 3SD procedure (algorithm 1).

Table 1. Input parameters used in this work for algorithms 1 and 2, and a nonexhaustive list of additional examples that can be used.

Input parameter	Examples selected in this work for illustration	Other examples
Periodogram P	Lomb–Scargle periodogram (Sects. 4–6), Generalized Lomb–Scargle (Sect. 5.5)	Other periodograms' variants: e.g., Ferraz-Mello (1981), Cumming et al. (1999), Cumming (2004), Reegen (2007); Bourguignon et al. (2007), Baluev (2013a,b), Jenkins et al. (2014); Tuomi et al. (2014), Gregory (2016), Hara et al. (2017; 2022)
Detection test T	Max test T_M (Eq. (6)), Test of the N_C^{th} largest value, T_C (Eq. (9))	Any detection test (see e.g., in Sulis et al. 2017a,b): e.g., Higher Criticism (Donoho & Jin 2004), Berk-Jones (Berk & Jones 1979)
Noise models $\mathcal{M}_{\mathbf{d}}$	Gaussian process (Eq. (10)), Linear activity proxy with polynomials (Eq. (14)), Linear activity proxy (Eq. (16))	Any relevant model for $\mathbf{d}$: e.g., models for long term instrumental trends, stellar magnetic activity, binary trends
Noise models $\mathcal{M}_{\mathbf{n}}$	AR process (Eq. (8)) Harvey function (Eq. (13))	Any relevant model for $\mathbf{n}$: e.g., models for meridional circulation, stellar granulation, supergranulation, or instrumental noise
Ancillary time series c	log R'_{HK} (Sects. 5 and 6)	Any activity proxy for $\mathbf{d}$: e.g., S-index (Wilson 1968), FF' (Aigrain et al. 2012), bisector (Queloz et al. 2001)

or even intractable, however, when a maximum likelihood estimation is required over a large number of parameters. It is therefore discarded in our framework, where nuisance parameters are estimated only under $\mathcal{H}_0$. For weak planetary signatures, the perturbation of these signatures in the nuisance parameters is weak, so that the estimation of these parameters does not change much under both hypotheses.

The second step (rows 5–10) aims at computing the averaged periodogram $\overline{\mathbf{p}}_L$ involved in Eq. (5). For this purpose, if $\mathbf{n}$ in Eq. (1) is a colored noise for which an NTS is available (rows 6–7), we directly take the L null training series to compute $\overline{\mathbf{p}}_L$ in row 14. If no NTS is available (rows 8–10), we fit model $\mathcal{M}_{\mathbf{n}}$ to the time series under test $\mathbf{x}$ and compute a synthetic NTS (called $\widehat{\mathcal{T}}_L$) with the estimated model parameters $\widehat{\theta}_{\mathbf{n}}$. In this case, $\overline{\mathbf{p}}_L$ in row 14 is evaluated after generating a sufficiently large number of synthetic noise time series from $\mathcal{M}_{\mathbf{n}}$. If noise $\mathbf{n}$ is white (rows 11–13), the averaged periodogram in row 14 simply becomes the product of the estimated variance of $\mathbf{x}$ ($\widehat{\sigma_{\mathbf{w}}^2}$) by vector $\mathbf{1} := [1 \ldots 1]^\top$.

The final steps of the procedure consist of computing $\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L$ (row 15) and applying the detection test T (row 16). The procedure returns the test statistics $t(\mathbf{x})$.

2.4. Periodogram and test used in this work

In order to illustrate our detection framework, we mainly consider the Lomb–Scargle periodogram here (Scargle 1982). The numerator and the denominator in Eq. (5) are therefore LSP. For the detection test, we consider the test whose statistics is the largest periodogram peak. This test is referred to as Max test for short. The Max test statistic writes

$$T_{\mathrm{M}}(\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L) := \max_k \widetilde{\mathbf{p}}(\nu_k). \tag{6}$$

We emphasize that the 3SD detection procedure is not tied to the particular choice of the pair (periodogram and test) selected above. It can indeed be applied as such to other types

of periodograms and test statistics that exploit the periodogram ordinates differently (see Sulis et al. 2016, 2017a for examples of other detection tests). Table 1 summarizes examples of input parameters that can be used in algorithm 1.

When algorithm 1 has computed the test statistics of $\mathbf{x}$, the important question is to evaluate its p-value. This is the purpose of Sect. 3.

3. Numerical estimation of the *p*-values

For an observed test statistics t, the p-value is denoted by $p_{\mathrm{v}}(t)$ and defined as

$$p_{\mathrm{v}}(t) := \Pr(T > t \mid \mathcal{H}_0).$$

For example, when all parameters are known, the p-value of a test statistics t obtained by the Max test (6) is defined as

$$p_{\mathrm{v}}(t) := \Pr(T_M(\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L) > t \mid \mathcal{H}_0) = 1 - \Phi_{T \mid \mathcal{H}_0}(t), \tag{7}$$

with $\Phi_{T \mid \mathcal{H}_0}$ the cumulative distribution function (CDF) of the test statistics under $\mathcal{H}_0$.

A p-value estimation procedure in which all parameters of the statistical model are known under $\mathcal{H}_0$ is called an oracle in the statistical literature (see Donoho & Johnstone 1994; Roquain & Verzelen 2022; Mary & Roquain 2021 and references therein). We describe such an oracle procedure in Sect. 3.1.

When unknown nuisance parameters are present under $\mathcal{H}_0$, the computation of p-values is not straightforward. A common approach that allows accounting for uncertainties in the model parameters considers the unknown parameters under $\mathcal{H}_0$ as random and computes the expectation of p-values over the distribution of a prior parameter. This is the approach of the Monte Carlo procedure proposed in algorithm 2 (see Sect. 3.2). Connections with existing approaches are discussed in Sect. 3.3.

In practice, a desirable feature of the procedure is also that it provides a high-probability interval (e.g., 90%) for the

location of the true (i.e., oracle) p-value. The Monte Carlo sampling approach of algorithm 2 naturally provides estimates of such intervals, called 90% intervals below. As a complementary p-value estimate, the highest p-value (upper envelope of the sampled p-values) obtained from the simulation can also be used as a (often overly) conservative estimate (Bayarri & Berger 2000). In practice, a large difference between the two estimates is the sign of a strong dependence of the estimation results on the considered models and parameters.

3.1. Oracle

When all parameters under the null hypothesis are known, which is an ideal situation that is not met in practice, a straightforward way for evaluating the oracle p-values numerically consists of performing algorithm 1 many times (e.g., b) to obtain a set $\{t(\mathbf{x})\}_{j=1,\dots,b}$ of realizations of the test statistics t. From this set, we can compute an empirical estimate of $\Phi_{T|\mathcal{H}_0}$, called $\widehat{\Phi}_T$ for short, and an estimated p-value by plugging $\widehat{\Phi}_T$ into Eq. (7). The estimation of the p-value can be made arbitrarily accurate because $\widehat{\Phi}_T$ converges to $\Phi_{T|\mathcal{H}_0}$ as $b \to \infty$.

3.2. Algorithm for estimating the p-values

The generic procedure (algorithm 2) is presented in Fig. 2. The final expected p-value ($\overline{\widehat{p_v}}$) is obtained by sample expectation (row 19) over B p-values $\widehat{p}_v^{\{i\}}$ obtained in a major loop (green, rows 1–18). Each such p-value is obtained through a minor loop (rows 5–16) that measures an empirical CDF ($\widehat{\Phi}_T^{\{i\}}$). This loop (and this estimated CDF) includes an averaging over the priors $\pi_\mathbf{d}$ and $\pi_\mathbf{w}$.

The major loop samples B values of the parameters for $\mathbf{n}$ ($\widetilde{\boldsymbol{\theta}}_\mathbf{n}^{\{i\}}$), which are obtained by parametric bootstrap: a sample NTS is generated according to parameters $\widehat{\boldsymbol{\theta}}_\mathbf{n}$ estimated from the data (rows 3 and 4). In the nested minor loop (blue), the algorithm samples a number b of test statistics t produced by algorithm 1 conditionally on $\widetilde{\boldsymbol{\theta}}_\mathbf{n}^{\{i\}}$ (if $\mathbf{n}$ is colored), $\widetilde{\widehat{\sigma^2}}_\mathbf{w}^{\{i,j\}}$ (if $\mathbf{n}$ is white) and on $\widetilde{\boldsymbol{\theta}}_\mathbf{d}^{\{i,j\}}$ (if $\mathbf{d} \neq \mathbf{0}$ in Eq. (1)). These parameter values are obtained by randomly perturbing the corresponding input values $\widehat{\sigma^2}_\mathbf{w}$ and $\widehat{\boldsymbol{\theta}}_\mathbf{d}|\mathcal{M}_\mathbf{d}$ within reasonable uncertainties according to some prior distributions, called $\pi_\mathbf{w}$ (row 10) and $\pi_\mathbf{d}$ (row 13). If model $\mathcal{M}_\mathbf{d}$ involves the use of an ancillary series c, a synthetic series $\widehat{c}^{\{i,j\}}$ is generated in row 14 using the same model parameters $\widetilde{\boldsymbol{\theta}}_\mathbf{d}^{\{i,j\}}$ plus a WGN.

Each test statistics is then obtained by applying algorithm 1 to simulated time series $\mathbf{x}^{\{i,j\}}$ (row 16). If $\mathbf{n}$ is colored, algorithm 1 takes as input the L simulated NTS $\mathcal{T}_L^{\{i,j\}}$ (computed in row 7) for periodogram standardization. If $\mathbf{n}$ is white, this vector is sent as empty in algorithm 1.

A particular case of this algorithm arises when the noise is considered colored with a parametric model $\mathcal{M}_\mathbf{n}$ and no NTS is available. This case impacts the setting of L in rows 3 and 7 and is explained in Sect. 5.5.

For $\pi_\mathbf{w}$ and $\pi_\mathbf{d}$, we consider Gaussian and uniform priors (with independent components if multivariate) in the simulations below. The component distributions are centered on the parameter values estimated from the data. The scale parameters (widths of the intervals for uniform priors or variances for the Gaussian prior) are estimated from the estimation error bars.

3.3. Connection with the literature

In the statistical literature, several approaches exist for estimating p-values when unknown parameters are involved in the detection process (see Bayarri & Berger 2000 for a review). In these approaches, prior predictive p-values are obtained by averaging over a prior law of the parameters (Box 1980). In contrast, posterior predictive p-values require averaging over the posterior distribution (Guttman 1967; Rubin 1984). Sampling from the posterior would require a computationally involved procedure (e.g., with a Markov chain Monte Carlo method). For this reason, we do not follow the posterior predictive approach here and aim at keeping the sampling process simple. In our approach, this process is different depending on the parameters we consider. We list the different cases below.

- For the parameters of $\mathcal{M}_n$ (row 4 of algorithm 2), which correspond to a parametric model of a random process, a natural way of sampling parameters over a distribution that is consistent with our model is to sample the random process with the estimate at hand and to re-estimate the parameters. This is what we do here. This approach also means that we do not have to choose an explicit prior for $\boldsymbol{\theta}_n$, the parameters of which would need to be set or estimated.
- For the other unknown parameters $\sigma_\mathbf{w}$ and $\boldsymbol{\theta}_\mathbf{d}$ in rows 10 and 13 of algorithm 2, we did not use the same strategy for computational reasons. Here, we found it more efficient to sample these parameters from some priors (Gaussian and uniform), whose parameters were estimated from the estimated error bars in an empirical Bayes manner (Efron & Hastie 2016).

The performances of the generic algorithm 2 are evaluated on real and synthetic data in the next three sections.

4. Noise is granulation and WGN

4.1. Objectives

We first consider the simplest case in which the RV time series is only affected by the stochastic colored noise $\mathbf{n}$ in Eq. (1), of which a training set $\mathcal{T}_L$ is assumed to be available (i.e., we set $\mathbf{d} = \mathbf{0}$ in Eq. (1), and $\mathcal{M}_\mathbf{d} = \emptyset$ in algorithms 1 and 2). While this configuration is not realistic for most long-period planet RV searches, where magnetic activity phenomena (defined in $\mathbf{d}$) always play an important role in the observed RV time series, it can be useful for detecting low-mass ultra-short period planets (USP, defined with periods < 1 day). This has been discussed in Sulis et al. (2020). In addition to this very specific case, this section has two goals. The first goal is pedagogical because we illustrate the principle of the detection procedure with the simplest case. The second goal is to validate the proposed p-value estimation procedure when the NTS is composed with MHD simulations of stellar granulation and for irregular sampling. So far, this validation has only been performed for a regular sampling, in which case the FAP and p-values can be computed with analytical expressions (Sulis et al. 2020).

4.2. Validation of the 3SD procedure on solar data

4.2.1. Data, MHD simulations, and sampling grids

The validation of the procedure on real data poses two difficulties. First, establishing a ground-truth reference for the p-value that would be obtained with an NTS composed of genuine stellar

Fig. 2. Part of the 3SD procedure that estimates the *p*-value of algorithm 1 (algorithm 2).

convection noise requires implementing the oracle procedure (see Sect. 3.1), which in turn requires a very large number of genuine stellar noise time series. Fortunately, the GOLF/SoHO spectrophotometer has made 25-yr-long time series of solar RV observations available that can be used (e.g., see Appourchaux et al. 2018). Second, in this section, we focus on granulation and instrumental noise. We therefore need to disregard the low-frequency part of the noise PSD that is dominated by the magnetic activity. For this purpose, the detection test is applied

to frequencies $\nu > 56\,\mu$Hz alone because at frequencies below $56\,\mu$Hz, the PSDs of solar data and MHD simulations start diverging (see Sulis et al. 2020 and Appendix C).

Because GOLF observations cover a very long period of time and because the timescales characteristic for granulation are shorter than a day, we divided the total time series into smaller series and focused on this short-timescale solar variability alone. We selected $b = 1420$ two-day-long series from the ≈ 3000 available time series taken during the first 17 yr of GOLF

observations (the last years are particularly affected by detector aging).

These time series are regularly sampled (N_{reg} = 2880 data points) at a sampling rate Δt = 60s. As detailed in Sulis et al. (2020), we filtered out the high-frequency p-modes on each two-day-long series (with periods < 15 min because they are not reproduced by the MHD simulations), and we added a WGN with variance $\widehat{\sigma}_{\mathbf{w}}^2$ estimated from the high-frequency plateau visible in the periodogram of the GOLF data (see Appendix C). The resulting dataset constitutes a large set of noise time series from which the oracle reference results can be computed.

For the implementation of algorithm 2, the available NTS is composed with a 53-day-long noise time series obtained from 3D MHD simulations of solar granulation (Bigot et al., in prep.). After splitting them into two-day-long time series, 26 granulation noise time series are available, of which we randomly selected L = 5 for the standardization. Before running algorithm 2, we applied the same corrections as for the GOLF data to the NTS obtained from MHD simulations ($\mathcal{T}_5$) (i.e., filtering of the stellar acoustic modes and addition of an instrumental WGN). The NTS was finally sampled as the observations under test (see the three sampling grids detailed below).

Algorithm 2 also relies on parameters that set the size of the Monte Carlo simulations, B and b (taken here as B = 100 and $b = 10^4$) and on model $\mathcal{M}_{\mathbf{n}}$ with associated parameters $\widehat{\theta}_{\mathbf{n}}$. We chose an AR model because it suits our MHD NTS time series well and its parameters are fast to estimate. An AR process is defined as (Brockwell & Davis 1991)

$$\mathbf{n}(t_j) = \sum_{j=1}^{o_{AR}} \alpha_{AR,j}\, \mathbf{n}(t_j - j\Delta t) + \mathbf{w}(t_j), \qquad j = 1, \ldots, N, \qquad (8)$$

with $\mathbf{n}(t_j)$ the sample of $\mathbf{n}$ at time instant t_j, $\alpha_{AR,j} \in\,]-1; 1[$ the AR coefficients filter, o_{AR} the AR order, Δt the sampling time step, and $\mathbf{w}(t_j)$ the sample at time t_j of a WGN series $\mathbf{w}$ of variance $\sigma_{\mathbf{w}}^2$. The model parameters $\widehat{\theta}_{\mathbf{n}}$ are the AR order and the filter coefficients. The order was selected using Akaike's final prediction error criterion (Akaike 1969) and led to a value of 10, which remained fixed for the remaining procedure. The filter coefficients (estimated $B \times b + 1$ times) were obtained through the standard Yule-Walker equations (Brockwell & Davis 1991).

For this study, we considered three temporal sampling grids. First, we considered a regular sampling grid (Δt = 10 min), where we keep only 10% of the initial GOLF series to make the number of data points N that are comparable to ordinary RV time series. Second, we considered a regular sampling grid (Δt = 60 s) affected by a large central gap: two 2.4 h continuous observations separated by a 43.2 h gap. Third, we considered a random sampling grid, in which 10% of the data points are randomly selected in the initial regular grid of the GOLF series ($\Delta t \in [60, 2880]$ s). In each case, the sampling grid contains N = 288 data points. For the regular (irregular) sampling grids, the standardized periodogram was computed using the classical (Lomb–Scargle) periodogram. They are the same for regular sampling.

4.2.2. Validation of algorithm 2 on solar data

The p-value estimation procedure provided by algorithm 2 results in B = 100 estimates $\widehat{p}_v^{\{i\}}$ shown in gray in the top panels of Fig. 3. The average p-value $\overline{p}_v$ is shown in black and the p-value of the oracle procedure based on real stellar noise in green. To compute this oracle, we ran algorithm 1 with

1420 GOLF time series that we further permuted ten times by blocks. This led to an effective number of b = 14200 series. To compute each realization of the test, we randomly selected one series to compute the periodogram in row 4 of algorithm 1 and L = 5 other series to compute the averaged periodogram in row 14.

The beam of the gray curves represents particular estimates of the unknown relation $p_v(t)$ of algorithm 1, with parameters that are consistent with the NTS at hand. The true p-value of this procedure is unknown (for irregular sampling) because evaluating it exactly would require an infinite number of MHD simulations. If the simulations performed in algorithm 2 sample the space of parameters correctly, the truth is likely to lie somewhere in these curves, perhaps close to the mean where the bunch of curves is located.
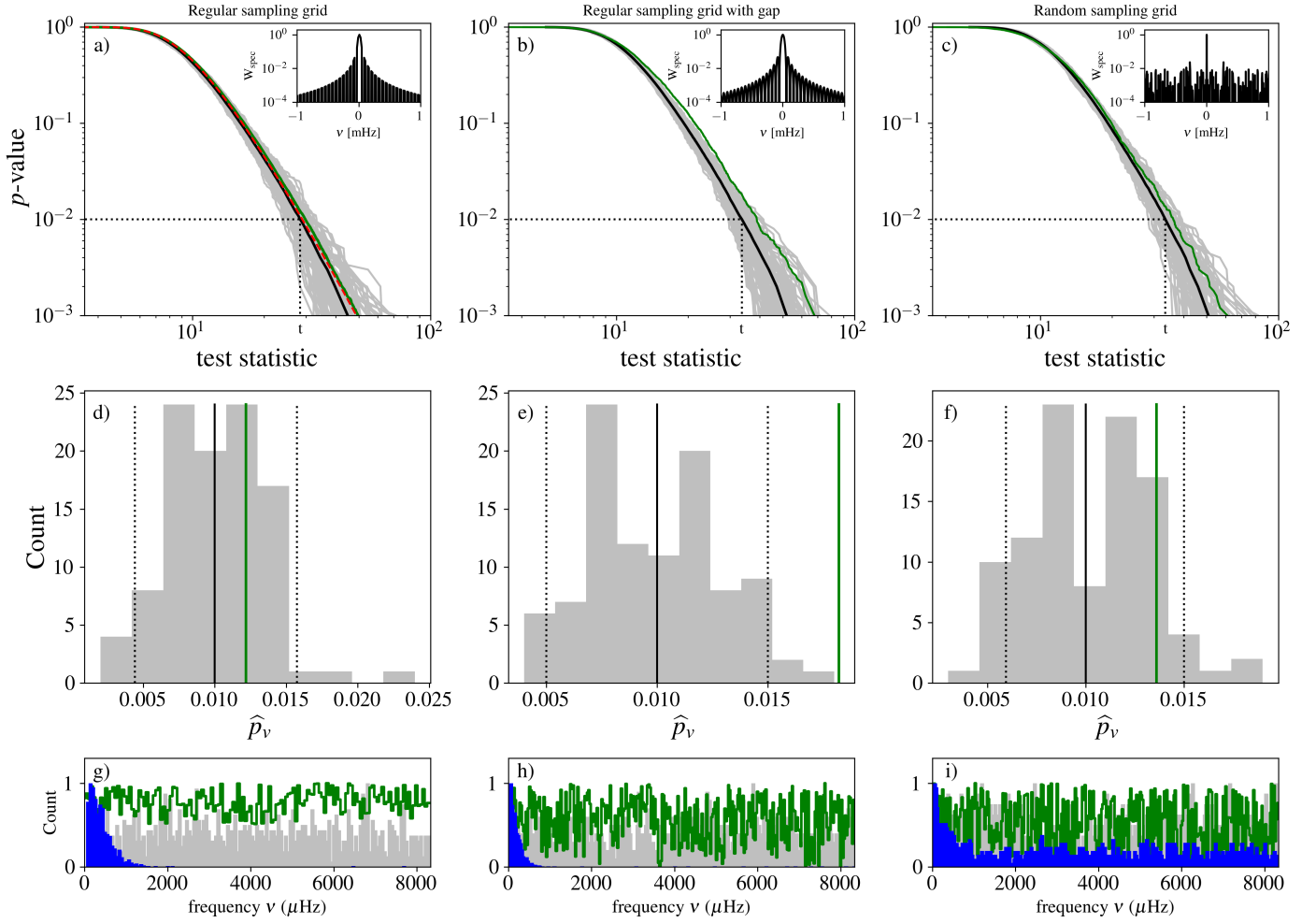
A comparison of the black and green curves shows that for the three considered sampling grids the approach based on MHD simulations is equivalent to using true stellar noise on average.

This shows that while algorithm 2 only has a reduced number (L = 5) of synthetic time series for the NTS at hand, it can generate a beam of situations whose mean succeeds in evaluating the true p-value of a standardized detection procedure supervised by genuine stellar noise.

We note a slight mismatch in the middle and right panel that probably arises because our ground-truth (GOLF) data contain some effects that were not accounted for in the p-value estimation procedure. Here, the highest p-value (upper envelope of the gray curves) provides a more conservative and more reliable estimate than the mean. The test statistics t corresponding to a mean p-value of 1% are 28.3, 33.6, and 33.5 for the three samplings. For these values, the simulations show that the p-value has an empirical probability of 90% to lie in the range [0.4%, 1.6%], [0.5%, 1.5%], and [0.6%, 1.5%] (panels in the middle row of Fig. 3). At these t values, the oracle p-values correspond to 1.2%, 1.8%, and 1.3%.

Because MHD simulations are made to represent stellar granulation noise with high accuracy, the good match observed in each case between the mean p-value obtained by algorithm 2 and the oracle based on true stellar noise suggests that the true p-value is close to the sample average. We also note that in the case of regular sampling, for which the p-value expression is known and given in Eq. (13) of Sulis et al. (2020), theory, the averaged estimate from algorithm 2, and the oracle with true stellar noise (panel a) match well.

Finally, the bottom panels show the histograms of the distribution of the frequency index in which the maximum of the standardized periodogram was found during the iterations of algorithm 2 of the 3SD procedure (gray) and oracle (green). We show the same distribution, but for the (nonstandardized) LSP in blue, that is, a procedure that would apply the Max test to $\mathbf{p}$ instead of $\widetilde{\mathbf{p}}$ in algorithm 1. Because the noise is colored, the LSP tends to peak at lower frequencies, where the PSD has higher values (see Fig. C.1). Colored noise inexorably causes an increased rate of false alarms at these frequencies, but this effect cannot be quantified or even suspected by the user when the noise PSD is unknown. In contrast, the standardization used in Eq. (5) leads to an autocalibration of the periodogram ordinates, which leads to a more reasonable (uniform) distribution of the frequency index of the maximum peak. These results show that the 3SD procedure is well calibrated and algorithm 2 succeeds in estimating the resulting p-value correctly. Because the distribution of the test statistics is very similar for data that are standardized with true noise and for data that are standardized by means of MHD noise time series, these results also demonstrate that the MHD

Fig. 3. Validation of the 3SD approach for a noise **n** composed of granulation plus WGN and **d** = 0 in Eq. (1). We investigate three temporal sampling grids, *from left to right*: regular, regular with a central gap, and random. *Top panels*: relation $p_v(t)$ for the detection procedure using a genuine NTS based on GOLF solar data (green; oracle). The $B = 100$ estimates $\widehat{p}_v^{\{i\}}$ produced by the 3SD procedure using the MHD-based NTS (gray). Sample mean value $\overline{\widehat{p}}_v$ of the gray curves (black). The spectral window of each sampling grid is shown in the inset panels. In *panel (a)*, the red line represents the analytical expression of p_v (see Eq. (13) of Sulis et al. 2020). The dotted lines indicate the test statistics for which $\overline{\widehat{p}}_v$ equals a mean *p*-value of 1%. *Middle panels*: empirical distribution of $\widehat{p}_v^{\{i\}}$ at t and 90% intervals (dotted lines) around the mean (solid black lines). *P*-values obtained with the oracle are shown with green vertical lines. *Bottom panels*: for one of the B loops of algorithm 2, we show an example of the empirical distribution of the $b = 10^4$ random values of the frequency index where the largest periodogram peak was found (gray). Empirical distribution of the Max test related to the oracle procedure (green; only contours are shown for clarity). The same distribution, but for a nonstandardized LSP is shown in blue (i.e., the Max test is applied to **p** instead of $\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L$).

simulations can efficiently be exploited in place of a real noise NTS for a periodogram standardization.

Finally, we carried out a similar validation of the output *p*-values, but for another pair (periodogram and test). We selected the GLS periodogram (Zechmeister & Kürster 2009) and the test statistics of the N_C^{th} maximum periodogram value, defined as

$$T_C(\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L) := \widetilde{p}_{(N-N_C+1)}, \tag{9}$$

with $N_C \in \mathbb{N}^+$ a parameter related to the number of periodic components, and $\widetilde{p}_{(k)}$ the *k*th-order statistic of $\widetilde{\mathbf{p}}$. We note that test T_C may be more discriminative against the null than the Max test $\max_k \widetilde{p}(v_k) = \widetilde{p}_{(N)}$ in Eq. (6) for multiple periodic components, as discussed by Chiu (1989) and Shimshoni (1971). The *p*-values resulting from algorithm 2 computed for $N_C = 10$ in Eq. (9) and for the random sampling grid defined in Sect. 4.2.1
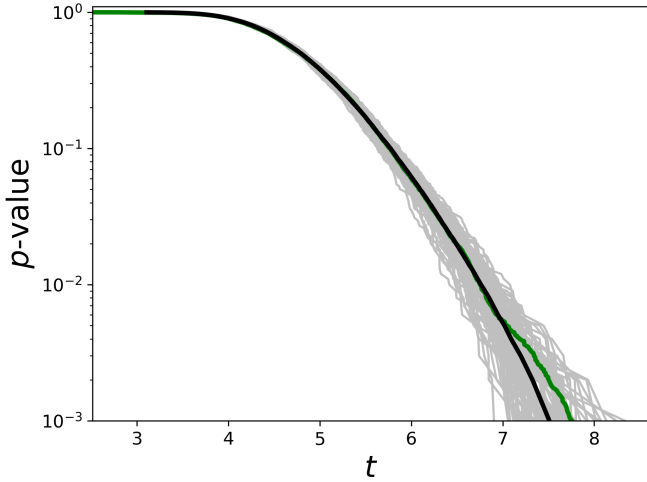
are shown in Fig. 4. These results illustrate that the reliability of the procedure is not tied to the pair (LSP and Max test).

5. Noise is magnetic activity, long term trends, granulation and WGN

5.1. Setting and objectives

In practice, many stellar activity processes affect the RV observations (Meunier 2021), and an NTS is not available for all of them. This is the case, for example, of stellar magnetic activity such as starspots and plages. We present a more general application of the 3SD procedure, in which classical techniques that are used to correct for the stellar variability of magnetic sources are coupled with the use of an NTS. We also investigate the case in which no NTS is available at the end of the section.

In this section, we consider as an illustration a nuisance signal **d** in Eq. (1) that encapsulates a magnetic signature that is

Fig. 4. Validation of algorithm 2 using the GLS periodogram and test T_C in Eq. (9). Oracle is shown in green, p-value estimates from algorithm 2 are shown in gray, and their mean value is plotted in black.

modulated with the stellar rotation period plus long-term trends that are modulated with the stellar cycle. Numerous techniques have been developed to estimate the contribution of magnetic activity to RV data (e.g., see Ahrer et al. 2021). We assume that an estimate of the noise activity signature has been obtained by one of these methods in the form of an ancillary time series for **d**. This ancillary time series is assumed to be the log R'_{HK} activity indicator in the following (but other indicators can be used as well, see Table 1).

The noise component **n** in Eq. (1) is composed of correlated granulation noise plus WGN. An NTS of **n** composed with L simulated RV time series obtained from MHD simulations is assumed to be available, except in the last part of this section, in which we relax this assumption.

The remaining section contains four parts. The first two parts describe the synthetic dataset with the computation of oracle p-values, and the validation of the p-value estimation algorithm with the impact of the priors on the computed p-values. In the third part, we study improvements that can be achieved by using generalized extreme values (GEV) to speed up the Monte Carlo simulations in algorithm 2. In the last part, we consider the particular (but practical) case in which no NTS is available for **n** in Eq. (1).

5.2. Synthetic RV time series and oracle p-values

We first created synthetic RV time series used to (i) generate an RV test time series that was later used in the input of algorithm 1, and (ii) compute the oracle p-values that were later used to evaluate the accuracy of the p-value estimates derived from algorithm 2. These time series entail magnetic, granulation, and instrumental noises generated as described below.

Component **d** was obtained as the realization of a Gaussian process (GP; Rasmussen & Williams 2005) of kernel

$$k(t_i, t_j) = \alpha_{GP}\, k_E(t_i, t_j) + k_M(t_i, t_j), \qquad (10)$$

with k_E and exponential sine square kernel defined as

$$k_E(t_i, t_j) = \exp\left\{-\Gamma_{GP} \sin^2\left(\pi|t_i - t_j|P_{GP}^{-1}\right)\right\}, \qquad (11)$$

and k_M a Matérn-3/2 kernel defined as

$$k_M(t_i, t_j) = \left(1 + \sqrt{3}\,|t_i - t_j|\,\lambda_{GP}^{-1}\right)\exp\left\{-\sqrt{3}\,|t_i - t_j|\,\lambda_{GP}^{-1}\right\}, \qquad (12)$$

with $\{t_i, t_j\}$ the time coordinates at indexes $i, j = 1, \ldots, N$. The hyperparameters of kernel k in Eq. (10) were set to $\alpha_{GP} = 2$, $\Gamma_{GP} = 22$ is the scale parameter, $P_{GP} = 25$ days is the rotation period, and $\lambda_{GP} = 1.609$ is the characteristic length scale. This GP was implemented using the GEORGE Python package[2] (see Ambikasaran et al. 2015 and online documentation[3] to find Eqs. (11) and (12)). Similar equations for kernels (11) and (12) can also be found in Eq. (8) of Espinoza et al. (2019) and Eq. (4) of Barragán et al. (2022), for instance.

This GP model with the same parameters was also used to compute the synthetic ancillary time series $\boldsymbol{c} := [c(t_1), \ldots, c(t_N)]^\top$, to which we added a WGN.

For component **n**, we created a stochastic colored noise time series with a Harvey function and a WGN. The power spectral density of this noise model has the form (Harvey 1985; Meunier et al. 2015)

$$S_H(\nu_k^+; \boldsymbol{\theta_n}) := \left(\frac{\frac{2\sqrt{2}}{\pi}\frac{a_H^2}{b_H}}{1 + \left(\frac{\nu_k^+}{b_H}\right)^{c_H}}\right) + \sigma_{\mathbf{w}}^2, \qquad (13)$$

where the vector $\boldsymbol{\theta_H} = [a_H, b_H, c_H, \sigma_{\mathbf{w}}^2]^\top$ collects the amplitude, characteristic frequency, and power of the Harvey function, and $\sigma_{\mathbf{w}}^2$ is the variance of the WGN. Notation ν_k^+ means that positive frequencies are considered. We calibrated the parameter vector $\boldsymbol{\theta_H}$ using a fit of the available MHD RV time series (see Sect. 4). In this case and in the following, we used the lmfit Python package (Newville et al. 2014), which makes nonlinear optimizations for curve-fitting problems. This choice was made for simplicity and because the algorithm is fast. For practical RV analyses, the user can optimize the fitting procedure to use a proper Markov chain Monte Carlo estimation procedure, for example.

Synthetic time series with a PSD that is exactly given by Eq. (13) can be simulated with the Fourier transform (FT). To compute a synthetic time series from Eq. (13), we generated a WGN $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, computed the fast FT of $\mathbf{w}$, performed a point-wise multiplication with $\sqrt{\mathbf{S}_H(\nu_k)}$, and finally took the real part of the inverse FT. An example of a synthetic ten-day-long time series obtained using models (10) and (13) is shown in Fig. 5 (black). From this time series, we created a large central gap between day 2 and day 8 in the time-sampling grid, and we randomly selected $N = 288$ data points in the two remaining intervals (gray). This synthetic RV time series is the series that we test in the next three subsections.
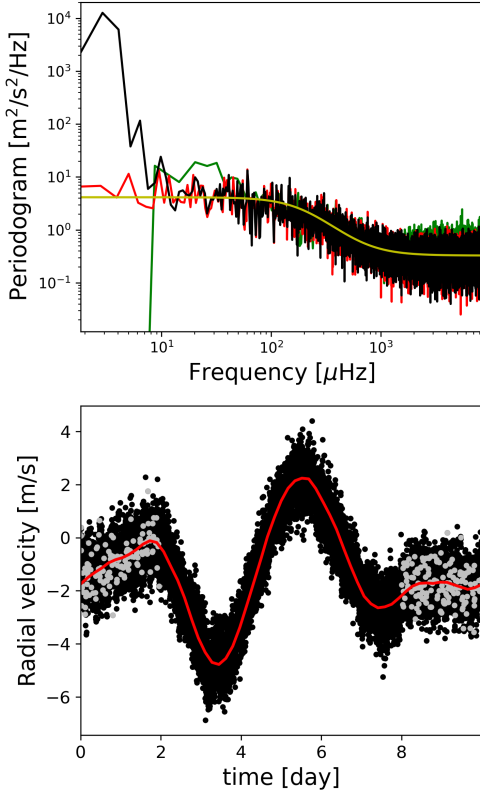
To compute the oracle procedure (which has knowledge of the exact parameters that were used to generate the synthetic time series, i.e., $\boldsymbol{\theta_H}$ in Eq. (13) and $[\alpha_{GP}, \Gamma_{GP}, P_{GP}, \lambda_{GP}]$ in Eq. (10)), we generated 10^4 synthetic noise time series as described above, and we ran algorithm 1 for each of the series. The resulting oracle ground truth is shown in green in Figs. 6–8 (all panels).

5.3. Validation of algorithm 2

We applied the 3SD procedure on one of the synthetic time series. For this purpose, we assumed a realistic situation in which

---

[2] https://github.com/dfm/george
[3] https://george.readthedocs.io/en/latest/user/kernels/

Fig. 5. Example of the synthetic time series generated to validate the 3SD approach when noise equals magnetic activity plus long-term trends, granulation, and WGN. *Top*: averaged periodograms of the two-day GOLF solar observations used in Sect. 2.4 (green) and the associated MHD training data (red), Harvey model fit to the MHD PSD (yellow), and averaged periodogram of the final synthetic data (GP+Harvey, black). *Bottom*: example of synthetic RV data generated with a regular sampling (black) based on a GP noise modeling for the magnetic activity (red). For the study, the dataset was also sampled irregularly (gray).

the noise models $\mathcal{M}_\mathbf{d}$ and $\mathcal{M}_\mathbf{n}$ that are used as inputs of algorithms 1 and 2 are different from the models that are used to generate the synthetic RV dataset in Sect. 5.2.

Following an example given in Dumusque et al. (2012) and Ahrer et al. (2021), we used for $\mathcal{M}_\mathbf{d}$ a simple parametric model[4] composed of a linear combination of $\boldsymbol{c}$ and three low-order polynomials,

$$\mathbf{d}(t_j \mid \boldsymbol{\theta}_\mathbf{d}) = \beta_\mathrm{d} \cdot \mathbf{c}(t_j) + \epsilon_\mathrm{d} \cdot t_j^2 + \zeta_\mathrm{d} \cdot t_j + \eta_\mathrm{d}, \quad j = 1, \ldots, N, \quad (14)$$

with $\boldsymbol{\theta}_\mathbf{d} = [\beta_\mathrm{d}, \epsilon_\mathrm{d}, \zeta_\mathrm{d}, \eta_\mathrm{d}]^\top$ the corresponding parameter vector.

For the available NTS $\mathcal{T}_\mathrm{L}$, we selected the $L = 5$ MHD time series described in Sect. 4.2.1. Using the LSP for P, the Max test for T, $\mathcal{T}_\mathrm{L}$, the model (14) for $\mathcal{M}_\mathbf{d}$, no model $\mathcal{M}_\mathbf{n}$, and the synthetic ancillary series $\boldsymbol{c}$, we then computed algorithm 1 and obtained a test value t.

To compute algorithm 2, the inputs parameters (P, T), $\mathcal{M}_\mathbf{d}$, and $\boldsymbol{c}$ were the same as for algorithm 1. For the additional inputs, we took the result from row 2 of algorithm 1 to obtain $\widehat{\boldsymbol{\theta}_\mathbf{d}}$, the vector of (four) parameters estimates, along with its scale parameter $\widehat{\boldsymbol{\delta}_\mathbf{d}}$. We selected a uniform prior $\pi_\mathbf{d}$ for the noise parameters

---

[4] This model is currently implemented in the code (see Appendix A). Other models for $\mathcal{M}_\mathbf{d}$, such as Gaussian process, are not implemented yet mainly for reasons of computational cost.

with interval widths set to twice the estimated standard error. We selected an AR model (8) for model $\mathcal{M}_\mathbf{n}$ (as in Sect. 4.1), and we fit this model to the NTS to obtain the estimated parameters $\widehat{\boldsymbol{\theta}_\mathbf{n}}$. We finally set up the size of the Monte Carlo simulations to $B = 100$ and $b = 1000$. The p-value estimates resulting from algorithm 2 are shown in the left panel of Fig. 6. When they are compared with the oracle p-values, algorithm 2 is shown to provide reliable p-value estimates at any t for colored noise and activity.

As a final validation step of algorithm 2, we finally examined the impact of the priors $\pi_\mathbf{d}$ (involved in row 13) on the derived p-values. The results are shown in Fig. 7. This example shows a weak impact, with an accurate estimation for any t in almost all cases. At the extreme of $10\,\sigma$ for the perturbation interval width, the p-value estimate becomes more conservative: at a given value of t, the estimated p-value is biased upward (blue curve). In practice, the influence of the chosen priors on the estimated p-values can always be confirmed by a simulation study of this type.

5.4. Generalized extreme values

Algorithm 2 does not rely on any model for the estimated CDF $\widehat{\Phi}_{T|\mathcal{H}_0}$ of the test statistics in Eq. (6). For the estimated mean p-value to be accurate, it therefore requires the parameter b involved in algorithm 2 to be large (typically $b \geq 1000$ or more). As $B \times b$ Monte Carlo simulations are involved (with B typically ≥ 100), algorithm 2 is computationally expensive. Fortunately, univariate extreme-value theory shows that the maximum of a set of identically distributed random variables follows a GEV distribution (Coles 2001). This suggests that GEV distributions can be used as a parametric model for the CDF of Eq. (6) to improve the estimation accuracy. This method was used by Süveges (2014) when only white noise is involved in model (1) (i.e., no colored noise or nuisance signal $\mathbf{d}$). Sulis et al. (2017b) extended the work of Süveges (2014) to the case of colored noise. Here, we evaluate the gains that this parameterization can contribute to the bootstrap procedure for the general case of colored noise and magnetic activity signals.
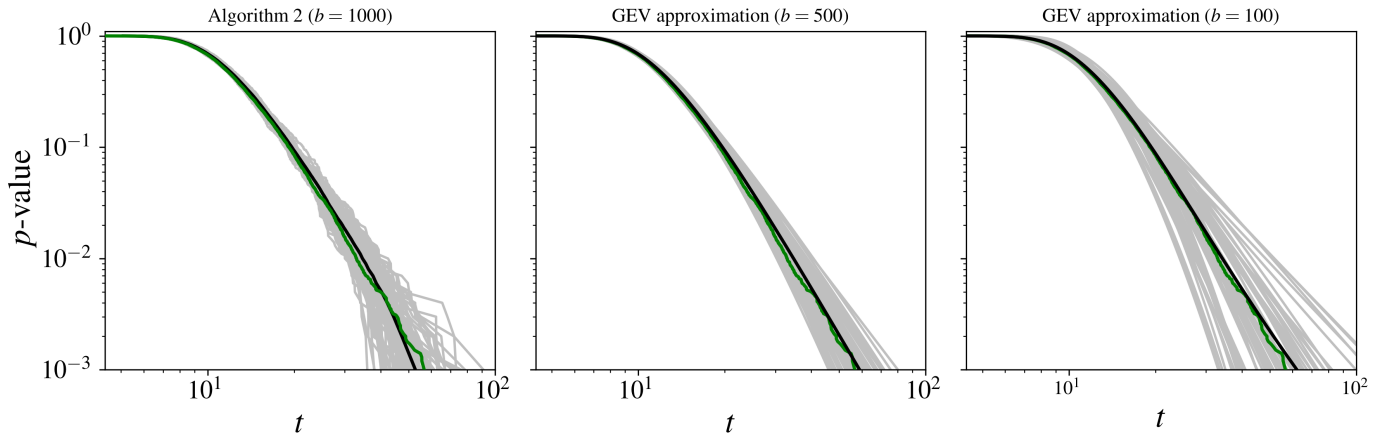
The GEV CDF depends on three parameters: the location $\mu \in \mathbb{R}$, the scale $\sigma \in \mathbb{R}^+$, and the shape $\xi \in \mathbb{R}$,

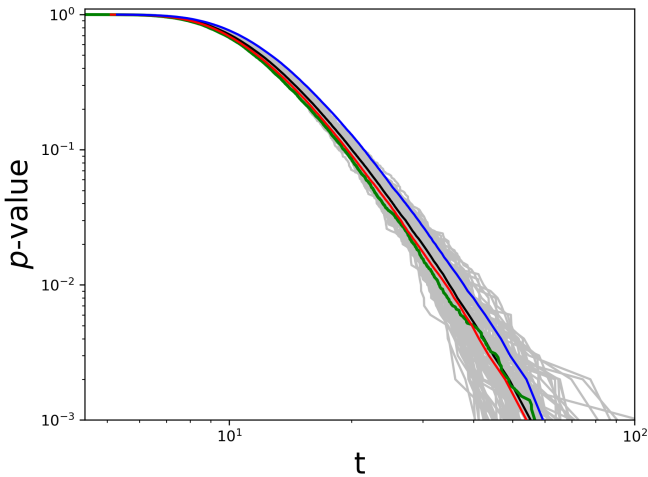$$G(t) = \mathrm{e}^{-\left[1 + \xi\left(\frac{t-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}}, \quad (15)$$

with $[x]_+ := \max(0, x)$. When $\xi \neq 0$, $G(t)$ in Eq. (15) is only defined for $t = 1 + \xi\frac{t-\mu}{\sigma} > 0$. When $\xi = 0$, $G(t)$ is derived by taking the limit at $t = 0$.

In algorithm 2, the unknown parameters $\boldsymbol{\theta}_{GEV} := [\xi, \mu, \sigma]^\top$ can be estimated using the $j = 1, \ldots, b$ test statistics $t(\mathbf{x}^{\{i,j\}})$, for instance, by maximum likelihood. In this case, the maximization can be obtained by an iterative method (Coles 2001). This leads to $i = 1, \ldots, B$ GEV parameters $\widehat{\boldsymbol{\theta}}_{GEV}^{\{i\}} := [\widehat{\xi}^{\{i\}}, \widehat{\mu}^{\{i\}}, \widehat{\sigma}^{\{i\}}]^\top$, from which the p-value can be estimated by $\widehat{p}_v^{\{i\}}(t) := 1 - G(t|\widehat{\boldsymbol{\theta}}_{GEV}^{\{i\}})$.

The results of the GEV parameterization are shown in the middle and right panels of Fig. 6 (for $b = 100$ and $b = 500$). The p-value estimates resulting from the GEV approximation are still given in a fairly tight 90% interval around the oracle. Moreover, when only half of the Monte Carlo realizations are used compared to the initial b value involved in algorithm 2 ($b = 500$ instead of 1000), a similar precision of the p-value estimates is obtained. Hence, the GEV approximation allows us to halve the computational time of algorithm 2 for the same precision

Fig. 6. *P*-value estimates resulting from algorithm 1 when noise equals magnetic activity plus long-term trends, granulation, and WGN, and where an NTS is available for granulation plus WGN. The mean *p*-value is shown in black. The oracle, which is the same for all three panels, is shown in green. *From left to right*: algorithm 2 computed without the GEV, and with the GEV approximation with *b* = 500 and *b* = 100.

Fig. 7. *P*-value estimates resulting for different choices of the priors $\pi_{\mathbf{d}}$ in row 13 of algorithm 2. Gray and black show the results for a Gaussian prior centered on the estimated parameter values with the scale parameters taken as 3 σ, with σ the estimated standard deviation of the parameter estimates. Gray shows examples of *p*-values $\widehat{p_{\mathrm{v}}}(t)$, and black shows the mean *p*-value. Red and blue show results for uniform priors with intervals centered on the estimated parameter values and widths taken as 1 σ (red) and 10 σ (blue). The true (oracle) *p*-value is shown in green.

(case *b* = 500). Reducing the computational time by ten (*b* = 100) leads to larger 90% intervals at constant *p*-value (e.g., the interval is ~40% larger at a *p*-value of 1% in the right panel of Fig. 6).

5.5. Algorithm 2 without NTS (practical case)

We consider now the common situation in which no NTS is available. In this case, the parameters of noise **n** can be estimated from the data (maybe along with the magnetic activity signal **d**) through a model $\mathcal{M}_{\mathbf{n}}$. Standardization is then performed by generating synthetic time series according to $\mathcal{M}_{\mathbf{n}}$ in algorithm 1.

When no NTS is available, two steps in algorithm 2 deserve particular attention. First, in rows 3 and 4, we set *L* = 1 because the noise parameters estimate $\widehat{\theta_{\mathbf{n}}}$ of model $\mathcal{M}_{\mathbf{n}}$ is based on the

unique time series under test **x**. Next, in row 7, as many synthetic time series can be generated that follow model $\mathcal{M}_{\mathbf{n}}(\theta_{\mathbf{n}})$ as necessary. Hence, in principle, $L \to \infty$. In practice, however, the choice of *L* is a compromise between estimation error and computation time. For the purpose of making massive Monte Carlo simulations, we used the low value *L* = 5.

As in Sect. 5.3, we applied the 3SD procedure to one of the synthetic time series described in Sect. 5.2. We first ran algorithm 1 with model $\mathcal{M}_{\mathbf{d}}$ such that

$$\mathbf{d}(t_j \mid \theta_{\mathbf{d}}) = \beta_{\mathbf{d}} \cdot \mathbf{c}(t_j), \qquad j = 1, \dots, N. \tag{16}$$
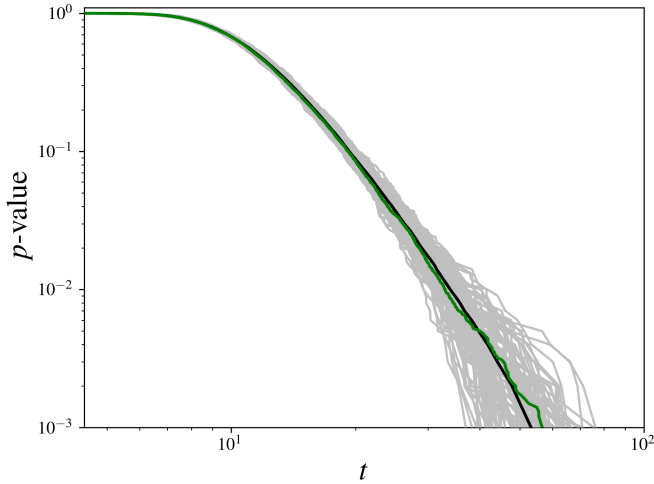
For model $\mathcal{M}_{\mathbf{n}}$, a Harvey function (13) was used. There is no NTS and the pair (P and T) was taken as the LSP periodogram and the Max test.

To compute algorithm 2, the input parameters (P, T), $\mathcal{M}_{\mathbf{d}}$, $\mathcal{M}_{\mathbf{n}}$, and **c** are the same as for algorithm 1. For the additional inputs, we obtain $\widehat{\theta_{\mathbf{d}}}$ and $\widehat{\delta_{\mathbf{d}}}$ from row 2 of algorithm 1 and set $\pi_{\mathbf{d}}$ as in Sect. 5.3. Parameters $\widehat{\theta_{\mathbf{n}}}$ were obtained from row 9 of algorithm 1 based on a fit of the LSP of **x** (as in Dumusque et al. 2011); this implicitly assumes that the LSP provides an acceptable approximation of the true PSD). We finally set up the size of the Monte Carlo simulations as *B* = 100 and *b* = 1000. The *p*-value estimates resulting from algorithm 2 are shown in Fig. 8. The estimated mean *p*-values and true *p*-values match very well, similarly to the case with an NTS (Sect. 5.3, see left panel of Fig. 6). This shows that the *p*-value can be robustly estimated even when no NTS is available.

We anticipate, however, that the performances of detection procedures based on estimates of the covariance structure from the data will probably depend on the RV data (noise sources, data sampling, etc.). The comparison of specific cases is left to a future paper.

6. Application to RV exoplanet detection: The debated detection of αCenBb

Alpha Centauri Bb is a debated ~1.13 Earth-mass (minimum mass) planet whose detection was reported around our solar neighborhood star αCenB with an orbital period of 3.2 days (Dumusque et al. 2012). It is challenging to achieve the detection of such a small planet because αCen is a triple star and the planetary RV semi-amplitude (~0.51 m s^{-1}) was at the limit of the

Fig. 8. Same as Fig. 6, but for a practical case without NTS for **n** in Eq. (1). In this case, the covariance structure of the stochastic noise source is estimated from the data with model $\mathcal{M}_\mathbf{n}$ given in Eq. (13).
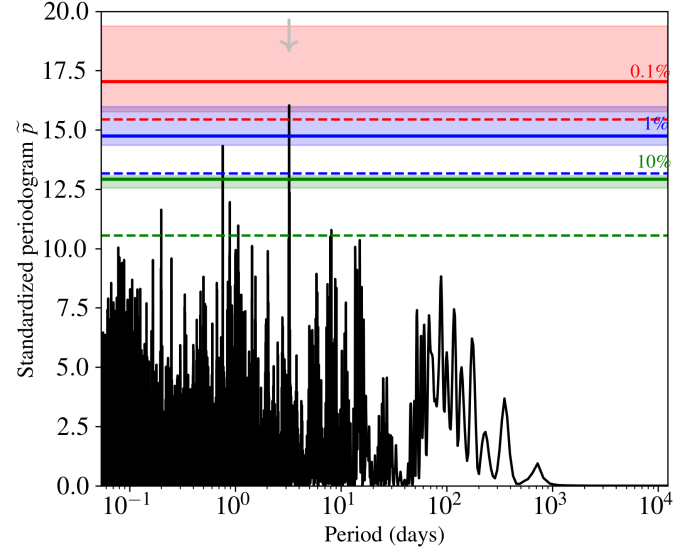

Fig. 9. Standardized Lomb–Scargle periodogram of the α CenB RV residuals. The noise model $\mathcal{M}_d$ used to create these residuals is taken from Dumusque et al. (2012). *p*-values estimated by a classical bootstrap procedure (see text) are shown with the dashed lines. Mean *p*-values derived with algorithm 2 are shown with solid lines, and their credibility interval with the shaded regions. The peak at 3.2 days is shown by the gray vertical arrow.

instrumental precision of HARPS ($\sim$1 m s^{-1}). The observed RV campaign contains $N = 459$ points, irregularly sampled during a time span of 4 yr between 2008 and 2011. The dataset contains large gaps between the observation years as a result of the visibility of the star, and small gaps that are due to the available nights.

The planet detection has been debated in a lively way: using analysis tools different from the detection paper in Hatzes (2013), using different stellar noise modeling in Rajpaul et al. (2016), and recently using a different technique of randomization inference to measure the *p*-value of the periodogram peak at 3.2 days by Toulis & Bean (2021).

In this section, we analyze this RV dataset as a case study for the 3SD procedure. The objective of the second part (Sect. 6.2) is also to demonstrate that algorithm 2 offers the possibility of investigating the effects of model errors on the frequency analysis and on the estimated *p*-values. The purpose is not to argue for or against a chosen noise model, but to study the effects of model mismatch. We first present the results of the 3SD procedure computed with the noise models involved in the detection paper (Dumusque et al. 2012). We then analyze the robustness of the results using a different noise model.

6.1. 3SD procedure with models from the detection paper

In short, the initial model $\mathcal{M}_\mathbf{d}$ described in the detection paper (Dumusque et al. 2012) contains three noise components. First, it contains the binary contribution ($\mathbf{m}_\mathrm{B}$) of αCen A, modeled with a second-order polynomial (three free parameters). Second, it contains the magnetic cycle, modeled with the linear relation of Eq. (16) and logR$'_\mathrm{HK}$ (for which periods < 40 days have been filtered out) taken as ancillary series ($\mathbf{m}_\mathrm{C}$, one free parameter). Third, it contains the magnetic (rotation modulated) activity ($\mathbf{m}_\mathrm{M}$), modeled as the sum of sine and cosine waves fitted at the stellar rotation period ($\sim$37 days) and its first four harmonics. Because of the stellar differential rotation, the authors fitted each of the 4 yr of observations individually to allow the stellar rotation period estimate to vary slightly from year to year. They also used a different number of harmonics for each of the four RV subsets. We refer to the supplementary material' of Dumusque et al. (2012) for the global model that is fit to each of the four

RV subsets (see their Eqs. p.7). In total, $\mathbf{m}_\mathrm{M}$ contains 19 free parameters.

The sum of components $\mathbf{m}_\mathrm{B}$, $\mathbf{m}_\mathrm{C}$, and $\mathbf{m}_\mathrm{M}$ forms the noise component **d** in model (1). The authors assumed that component **n** in (1) is uncorrelated and there is no NTS.
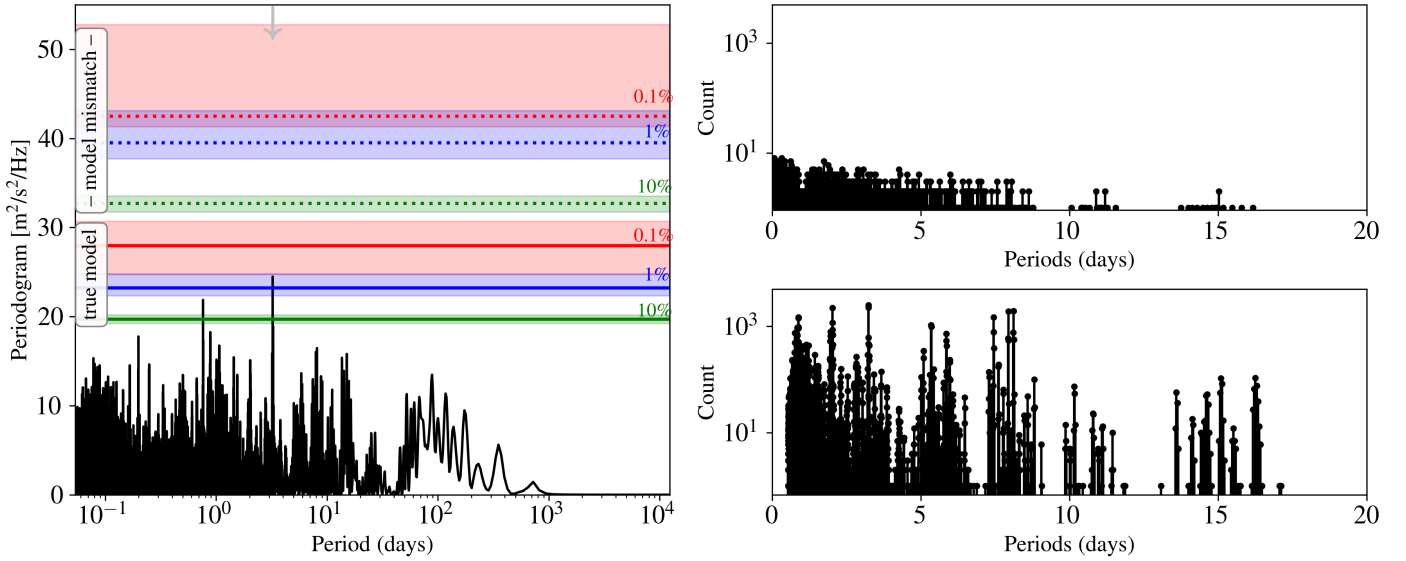
We applied the 3SD procedure to the RV data of αCenB with the following setting for algorithm 1:
- the Lomb–Scargle periodogram as P,
- the Max test as T,
- the set of linearly sampled frequencies as $\Omega := [9.4 \times 10^{-4}, 215.8]$ μHz (with $\Delta\nu = 1.88 \times 10^{-3}$ μHz, and $N_\Omega = 114750$ frequencies in total),
- no training time series or parametric model for the stochastic noise component **n** ($\mathcal{T}_\mathrm{L} = \emptyset$; $\mathcal{M}_\mathbf{n} = \emptyset$),
- the model defined in Dumusque et al. (2012) for component **d**, recalled above as $\mathcal{M}_\mathbf{d}$,
- the ancillary time series **c** as the 40-days-filtered logR$'_\mathrm{HK}$ time series available from the data release of Dumusque et al. (2012).

In row 11 of algorithm 1, we obtained a standard deviation $\widehat{\sigma}_\mathbf{w}$ of the data residuals of 1.23 m s^{-1}[5], in agreement with the value given in Dumusque et al. (2012). The standardized periodogram computed in row 14 is shown in Fig. 9. The algorithm computed a test value $t = 16.03$ (peak at 3.2 days).

We then ran algorithm 2 to evaluate the *p*-value of t with same inputs as algorithm 1 that are listed above. In addition, we plugged into algorithm 2 a number of Monte Carlo simulations of $B = 100$ and $b = 1000$, a Gaussian prior for $\pi_\mathbf{w}$ with a standard deviation of 0.04 m s^{-1} (1σ estimated error bar), and a uniform prior $\pi_\mathbf{d}$ with a 1σ interval width on each of the 23 free parameters of model $\mathcal{M}_\mathbf{d}$.

---

[5] For these data, different error bar estimates $\widehat{\sigma}_i$ are provided and were considered in the parameter estimation. For simplicity, algorithm 2 used a common uniform prior with an interval equal to $\widehat{\sigma}_\mathbf{w}$, but independent priors matched to $\widehat{\sigma}_i$ can be implemented here.

Fig. 10. Impact of noise model errors on the significance levels. *Left*: Lomb–Scargle periodogram of the α CenB RV residuals. The peak at 3.2 days is shown by the gray vertical arrow. Solid lines are the mean p-values estimated without a model error. The p-value at $t := \mathbf{p}(3.2\ d) = 24.48$ is within the 90% interval $[0.1\%, 0.7\%]$. Dotted lines are the mean p-values estimated with a model error (see text). The p-value at $t = 24.48$ this time is within the 90% interval $[59.5\%, 64.7\%]$. *Right*: distribution of the period index where the maximum periodogram value was found in row 15 of algorithm 2 (zoom at periods <20 days). The standard version of the 3SD procedure (simulating the situation without a model error) is shown *in the top panel*, and the hybrid version (detection process run with a model error) is shown *in the bottom panel*. A bias toward specific periods is visible *in the bottom panel*: the model error leaves residuals in which periodicities are imprinted.

The t values corresponding to mean p-values of 10, 1, and 0.1% are shown with solid green, blue, and red curves, respectively, in Fig. 9. The 90% intervals are shown with the corresponding colored shaded regions. The value $t := \widetilde{\mathbf{p}}(3.2\ d) = 16.03$ corresponds to an estimated mean p-value of 0.47% with a 90% interval of $[0.1\%, 0.7\%]$. We confirmed that these results are reasonably stable with respect to the prior model $\pi_{\mathbf{d}}$ for the $\mathcal{M}_{\mathbf{d}}$ model parameters in row 13 of algorithm 2 (e.g., a Gaussian instead of a uniform prior with interval widths set to 1σ standard error leads to $\overline{\widetilde{p}_v}(16.03) = 0.43\%$ with $[0.1\%, 0.7\%]$ 90% interval).

Dumusque et al. (2012) assumed that the data residuals (without a planet) are white. While the p-value evaluation procedure used in this paper was not specified, we assume that they used the classical bootstrap technique in which the p-values are evaluated by shuffling the dataset and keeping the observing time dates constant. The t values corresponding to the p-values of 10, 1, and 0.1% computed with this classical method are shown by dashed lines for comparison. The largest periodogram peak at 3.2 days lies in the region of p-values below 0.1%, which is consistent with the result of the publication paper (a p-value of 0.02%).

We conclude from this analysis that if the model assumed by Dumusque et al. (2012) is accurate, the 3.2 days peak is associated with a mean p-value of 0.47%, which is small but nevertheless about 23 times larger than the 0.02% p-value estimate given in the detection paper. In general, the p-value derived without taking the errors on the noise model parameters into account is underestimated.

6.2. Systematic study of noise model errors

We considered another noise model $\mathcal{M}_{\mathbf{d}}$ for $\mathbf{d}$. Following Rajpaul et al. (2016; see their Sect. 4), we still considered a nuisance

signal of the form $\mathbf{d} = \mathbf{m}_B + \mathbf{m}_C + \mathbf{m}_M$, but with some modifications. The estimation of the noise model parameters on the RV dataset was now performed in two steps. The binary and long-term trend components $\mathbf{m}_B$ and $\mathbf{m}_C$ were fit first. The magnetic activity component $\mathbf{m}_M$ was fitted in a second step on the data residuals as a Gaussian process with a quasi-periodic covariance kernel of the form of Eq. (10), with k_M fixed to 0 and parameters $[\alpha_{\mathrm{GP}}, \Gamma_{\mathrm{GP}}, P_{\mathrm{GP}}]$ free with P_{GP} in intervals of 30 and 42 days.

When we ran our detection algorithm 1 with this new noise model, we obtained final data residuals of $0.78\,\mathrm{m\,s^{-1}}$ standard deviation, and the peak at 3.2 days disappeared. Similarly to Rajpaul et al. (2016), we performed an injection test with a planetary signal at 3.2 days with an amplitude of $0.51\,\mathrm{m\,s^{-1}}$ to verify that the GP noise model does not overfit the planet signal. We found the periodogram peak of the data residuals to be the highest peak, confirming the observations of Rajpaul et al. (2016) that the detection outcome can be very sensitive to the choice of the noise model plugged in the detection algorithm.

Model errors are unavoidable in practice. For the same dataset, different experts may assume different noise models, and the magnitude of the absolute error resulting from these models can probably be a lower bound of the relative difference between the models. These observations suggest that automatic studies of the robustness of the detection process to model errors should be incorporated in the process of estimating p-values. For this purpose, a hybrid version of the 3SD procedure can be used.

This hybrid version simply consists of two modifications of the 3SD procedure. First, we allow in algorithm 2 that the procedure generates the noise-training series in rows 12 to 15 with a given model $\mathcal{M}_{\mathbf{d}}$ (the Gaussian process model), and to fit it in row 16 with another model, for instance, $\mathcal{M}'_{\mathbf{d}}$ (the model of Dumusque et al. 2012). Second, we do not use the standardization step in algorithm 1 (rows 5 to 13) because data residuals resulting from different models will have different dispersions, with a larger dispersion in presence of model error. Keeping

standardization (i.e., division by the estimated variance) would then lead to relatively smaller $\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L$ components for the model with the highest residuals and artificially hide the effects of the mismatch we seek to detect and show. The test values corresponding to p-values of 10, 1, and 0.1% computed in this hybrid configuration are shown with horizontal dotted lines in Fig. 10 (left panel). For comparison, the levels corresponding to the no model error configuration are shown with solid lines. The 90% intervals are again shown with shaded regions. We find that if the detection process indeed used the right noise model, the largest data peak at 3.2 days has a p-value of 0.62%. If, however, the detection process committed a slight mismatch in the model error, the actual p-value of the same peak increases to 62.02%. This means that peaks as large as the largest data peaks occur quite often, even without a planet, when the detection process is run under a model mismatch such as the one simulated here. Model mismatch is a likely situation in practice. A robust detection would be a detection whose p-value remains low under realistic model mismatches.

Interestingly, the hybrid version of the 3SD procedure further allows an investigation of the distribution of the period index where the $B \times b$ maximum values of $\mathbf{p}$ are found (see in Sect. 4). This allows indicating periodicities that are not caused by planets and may be used as a warning in the detection process when it finds a planet at these frequencies. This distribution is shown in Fig. 10 (bottom right panel). In this case, the distribution of the frequency index of the largest periodogram peaks significantly moves away from the uniform distribution that is seen when the noise model is correct (top right panel). When the noise model is correct, each frequency has been randomly hit a few times (the holes are caused by frequencies that are removed in the noise model). In contrast, when the noise model is not correct, the maximum peak was found 2504 times at the 3.2 -day period (compared with two times under no model error), or 1917 times at the 7.94 -day period (compared with one time without a model error).

In conclusion, when 1) the estimation errors on the noise parameters (assuming no model error) are taken into account and 2) the effects of possible realistic errors on the noise model are investigated, the degree of surprise brought by the peak at 3.2 days decreases substantially. This study of the α CenB dataset supports the conclusion that was reached in previous works (Hatzes 2013; Rajpaul et al. 2016; Toulis & Bean 2021), namely a likely false detection at 3.2 days, but with a different analysis. Beyond the particular case of the α CenB dataset, accounting for possible noise model errors appears to be a critical point, particularly when the amplitude of the planet signal is small compared to the nuisance signals.

7. Discussion and conclusions

Assessing the significance level of detection tests is critical for an exoplanet RV detection in the regime of a low signal-on-noise ratio and in presence of colored noise of unknown statistics. Impacts of errors on the parameter estimation and of model errors on the p-values have been little studied in the exoplanet community so far.

To overcome this limitation, we have presented a new method for computing p-values of detection tests for RV planet detection in presence of unknown colored noise. The developed Bayesian procedure is based on the principles of statistical standardization and allows using some training samples (NTS), if available (Sect. 2). It involves Monte Carlo simulations (Sect. 3) and

allows evaluating the robustness of the derived significance levels against specific model errors (Sect. 6.2). Our procedure builds on and extends our previous study (Sulis et al. 2020), which was limited to the case of regular sampling and was not able to handle magnetic activity signals. This procedure allows deriving in algorithm 2 the p-value corresponding to a test statistic produced in detection algorithm 1 while accounting for uncertainties in the noise model parameters.

Periodogram standardization can be implemented using a (possibly simulated) NTS. Examples of simulated NTS can come from MHD simulations of granulation, supergranulation, or any other stellar simulations that are shown to be reliable. This opens future collaborations between stellar modelers and RV exoplanet scientists (see Sect. 4). The NTS can also come from instrumental measurements, independent of the studied RV dataset. In addition, the procedure can take advantage of ancillary data to deal with nuisance signals such as the stellar magnetic activity (e.g., $\log R'_{HK}$ time series).

Overall, the proposed detection procedure is versatile in i) the periodograms (or other frequency analysis tools), ii) the detection tests (classical or adaptive), and iii) the noise models that can be plugged in. Examples of input parameters for the three items above are listed in Table 1. Additionally, Python codes are released on GitHub (see Appendix. A). Although not emphasized in this paper, the 3SD procedure also opens the possibility of using adaptive tests such as the higher criticsm or the Berk-Jones tests (Donoho & Jin 2004; Berk & Jones 1979). These tests are more powerful than classical detection tests (as the Max test presented in Eq. (6)) in the case of exotic planet signatures of small amplitudes (e.g., planets of high eccentricity or multiplanet systems) when the data sampling is regular (see Sulis et al. 2017a). How these adaptive tests may be incorporated in the proposed 3SD procedure to evaluate their performance in the context of irregular time sampling will be investigated in a further study.

We have illustrated how the procedure can be implemented in the presence of instrumental, granulation, and magnetic activity, of long-term drifts of stellar origin (magnetic cycle), and binary noises in model (1) (Sects. 4–6). Other RV noise sources can similarly be considered in the procedure when they are identified in the RV data.

Implementing the 3SD procedure may present some issues, but these can at least partially be solved. First, the procedure can be very time consuming owing to several factors (generation of NTS simulations, large number of Monte Carlo simulations, and the complexity scales with the number of parameters of the noise models). We have shown that the GEV framework allows us to reduce this computational cost (Sect. 5.4). Second, the computed p-values remain dependent on the choice of the input noise models. While this is not specific to our approach, we showed that a hybrid version of the procedure allows us to quantify the robustness of p-value estimates to specific model changes (Sect. 6.2). Third, the computed p-values are more conservative than the p-values obtained by random shuffling of the best-fit residuals, as is often done in RV data analysis (Sect. 6.1). This is because they take the main errors into account that are generated during the detection process (estimation error on the parameters and/or specific model error). Last, when the detection process makes some error in the assumed noise model, estimating and removing a nuisance signal from the RV dataset may lead to a nonuniform distribution of the frequency index where the largest periodogram peak is found (this is visible in the bottom right panel of Fig. 10). The reason is that imperfect subtraction of nuisance signals (caused by a model error) leads to residual

long-period signals in which the periodicities of the sampling grid are imprinted. Model errors always remain in practice, and the proposed algorithm allows diagnosing the effects of some of them.

Finally, throughout this paper, we have illustrated the proposed method on solar data, synthetic datasets, and HARPS data of αCenB. The next step is to analyze some still-to-be confirmed low-amplitude planet detection.

References

Ahrer, E., Queloz, D., Rajpaul, V. M., et al. 2021, MNRAS, 503, 1248
Aigrain, S., Pont, F., & Zucker, S. 2012, MNRAS, 419, 3147
Akaike, H. 1969, Ann. Inst. Stat. Math., 21, 243
Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2015, IEEE Trans. Pattern Anal. Mach. Intell., 38, 252
Anglada-Escude, G., Arriagada, P., Tuomi, M., et al. 2014, MNRAS, 443, L89
Appourchaux, T., Boumier, P., Leibacher, J. W., & Corbard, T. 2018, A&A, 617, A108
Baluev, R. V. 2008, MNRAS, 385, 1279
Baluev, R. V. 2011, Celest. Mech. Dyn. Astron., 111, 235
Baluev, R. V. 2013a, Astron. Comput., 2, 18
Baluev, R. V. 2013b, MNRAS, 429, 2052
Baluev, R. V. 2015, MNRAS, 446, 1493
Barragán, O., Aigrain, S., Rajpaul, V. M., & Zicher, N. 2022, MNRAS, 509, 866
Bayarri, M. J., & Berger, J. O. 2000, J. Am. Stat. Assoc., 95, 1127
Berk, R., & Jones, D. 1979, Z. Wahrscheinlichkeit, 47, 47
Boisse, I., Bouchy, F., Hébrard, G., et al. 2011, A&A, 528, A4
Bortle, A., Fausey, H., Ji, J., et al. 2021, AJ, 161, 230
Bourguignon, S., Carfantan, H., & Böhm, T. 2007, A&A, 462, 379
Box, G. E. P. 1980, J. Roy. Stat. Soc. A (General), 143, 383
Brockwell, P., & Davis, R. 1991, Time Series: Theory and Methods (Springer)
Carleo, I., Malavolta, L., Lanza, A. F., et al. 2020, A&A, 638, A5
Chiu, S.-T. 1989, J. Roy. Stat. Soc. B (Methodological), 51, 249
Coles, S. G. 2001, An Introduction to Statistical Modelling of Extreme Values (Springer-Verlag)
Collier Cameron, A., Mortier, A., Phillips, D., et al. 2019, MNRAS, 487, 1082
Collier Cameron, A., Ford, E. B., Shahaf, S., et al. 2021, MNRAS, 505, 1699
Cretignier, M., Dumusque, X., Hara, N. C., & Pepe, F. 2021, A&A, 653, A43
Cumming, A. 2004, MNRAS, 354, 1165
Cumming, A., Marcy, G. W., & Butler, R. P. 1999, ApJ, 526, 890
Cunha, D., Santos, N. C., Figueira, P., et al. 2014, A&A, 568, A35
Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D. A., & Ford, E. B. 2017, ApJ, 846, 59
de Beurs, Z. L., Vanderburg, A., Shallue, C. J., et al. 2022, AJ, 164, 49
Delisle, J. B., Hara, N., & Ségransan, D. 2020, A&A, 635, A83
Donoho, D., & Jin, J. 2004, Ann. Stat., 32, 962
Donoho, D. L., & Johnstone, I. M. 1994, Biometrika, 81, 425
Dumusque, X. 2018, A&A, 620, A47
Dumusque, X., Udry, S., Lovis, C., Santos, N. C., & Monteiro, M. J. P. F. G. 2011, A&A, 525, A140
Dumusque, X., Pepe, F., Lovis, C., et al. 2012, Nature, 491, 207
Dumusque, X., Cretignier, M., Sosnowska, D., et al. 2021, A&A, 648, A103
Efron, B., & Hastie, T. 2016, Computer Age Statistical Inference: Algorithms, Evidence, and Data Science, Institute of Mathematical Statistics Monographs (Cambridge University Press)
Elorrieta, F., Eyheramendy, S., & Palma, W. 2019, A&A, 627, A120
Espinoza, N., Kossakowski, D., & Brahm, R. 2019, MNRAS, 490, 2262
Ferraz-Mello, S. 1981, AJ, 86, 619

Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, PASP, 128, 066001
Gregory, P. C. 2016, MNRAS, 458, 2604
Guttman, I. 1967, J. Roy. Stat. Soc. B (Methodological), 29, 83
Hanasoge, S. M., Duvall, T. L., & Sreenivasan, K. R. 2012, Proc. Natl. Acad. Sci. U.S.A., 109, 11928
Hara, N. C., Boué, G., Laskar, J., & Correia, A. C. M. 2017, MNRAS, 464, 1220
Hara, N. C., Delisle, J.-B., Unger, N., & Dumusque, X. 2022, A&A, 658, A177
Harvey, J. 1985, in ESA Special Publication, Future Missions in Solar, Heliospheric & Space Plasma Physics, eds. E. Rolfe, & B. Battrick, 235, 199
Hatzes, A. P. 2013, ApJ, 770, 133
Hatzes, A. P., Fridlund, M., Nachmani, G., et al. 2011, ApJ, 743, 75
Hatzes, A. P., Cochran, W. D., Endl, M., et al. 2015, A&A, 580, A31
Horne, J. H., & Baliunas, S. L. 1986, ApJ, 302, 757
Jenkins, J. S., & Tuomi, M. 2014, ApJ, 794, 110
Jenkins, J. S., Yoma, N. B., Rojo, P., Mahu, R., & Wuth, J. 2014, MNRAS, 441, 2253
Jones, D. E., Stenning, D. C., Ford, E. B., et al. 2022, Ann. Appl. Stat., 16, 652
Jurgenson, C. et al. 2016, Proc. SPIE, 9908 99086T
Kay, S. M. 1998, Fundamentals of Statistical Signal Processing: Detection Theory, 1st edn., 2 (Prentice-Hall PTR)
Mary, D., & Roquain, E. 2021, ArXiv e-prints, [arXiv:2106.13501]
Ment, K., Fischer, D. A., Bakos, G., Howard, A. W., & Isaacson, H. 2018, AJ, 156, 213
Meunier, N. 2021, in Proceedings of the Evry Schatzman School 2019 "Interactions star-planet", eds. L. Bigot, J. Bouvier, Y. Lebreton, & A. Chiavassa
Meunier, N., & Lagrange, A. M. 2019, A&A, 625, A6
Meunier, N., & Lagrange, A. M. 2020, A&A, 638, A54
Meunier, N., Lagrange, A. M., Borgniet, S., & Rieutord, M. 2015, A&A, 583, A118
Meunier, N., Lagrange, A. M., & Borgniet, S. 2017, A&A, 607, A6
Newville, M., Stensitzki, T., Allen, D. B., & Ingargiola, A. 2014, LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python
Noyes, R. W., Hartmann, L. W., Baliunas, S. L., Duncan, D. K., & Vaughan, A. H. 1984, ApJ, 279, 763
Paltani, S. 2004, A&A, 420, 789
Palumbo, Michael L., I., Ford, E. B., Wright, J. T., et al. 2022, ApJ, 163, 11
Pepe, F. A. et al. 2010, in Proc. SPIE, Ground-based and Airborne Instrumentation for Astronomy III, 7735
Queloz, D., Henry, G. W., Sivan, J. P., et al. 2001, A&A, 379, 279
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
Rajpaul, V. M., Buchhave, L. A., Lacedelli, G., et al. 2021, MNRAS, 507, 1847
Rasmussen, C. E., & Williams, C. K. I. 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning) (The MIT Press)
Reegen, P. 2007, A&A, 467, 1353
Reichert, K., Reffert, S., Stock, S., Trifonov, T., & Quirrenbach, A. 2019, A&A, 625, A22
Rincon, F., & Rieutord, M. 2018, Living Rev. Solar Phys., 15, 6
Roquain, E., & Verzelen, N. 2022, Ann. Stat., 50, 1095
Rubin, D. B. 1984, Ann. Stat., 1151
Santos, N. C., Mortier, A., Faria, J. P., et al. 2014, A&A, 566, A35
Scargle, J. D. 1982, ApJ, 263, 835
Schwarzenberg-Czerny, A. 1998, MNRAS, 301, 831
Schwarzenberg-Czerny, A. 2012, in New Horizons in Time Domain Astronomy, eds. E. Griffin, R. Hanisch, & R. Seaman, 285, 81
Seifahrt, A., Käufl, H. U., Zängl, G., et al. 2010, A&A, 524, A11
Shimshoni, M. 1971, Geophys. J., 23, 373
Sulis, S. 2017, Ph.D. Thesis, Université Côte-d'Azur, Nice, France
Sulis, S., Mary, D., & Bigot, L. 2016, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4428
Sulis, S., Mary, D., & Bigot, L. 2017a, IEEE TSP, 65, 2136
Sulis, S., Mary, D., & Bigot, L. 2017b, in 2017 25th European Signal Processing Conference (EUSIPCO), 1095
Sulis, S., Mary, D., & Bigot, L. 2020, A&A, 635, A146
Süveges, M. 2014, MNRAS, 440, 2099
Süveges, M., Guy, L. P., Eyer, L., et al. 2015, MNRAS, 450, 2052
Tal-Or, L., Zechmeister, M., Reiners, A., et al. 2018, A&A, 614, A122
Toulis, P., & Bean, J. 2021, ArXiv e-prints, [arXiv:2105.14222]
Tuomi, M., Jones, H. R. A., Barnes, J. R., Anglada-Escudé, G., & Jenkins, J. S. 2014, MNRAS, 441, 1545
Tuomi, M., Jones, H. R. A., Barnes, J. R., et al. 2018, AJ, 155, 192
Udry, S., Mayor, M., Clausen, J. V., et al. 2003, A&A, 407, 679
Wilson, O. C. 1968, ApJ, 153, 221
Yu, J., Huber, D., Bedding, T. R., & Stello, D. 2018, MNRAS, 480, L48
Zechmeister, M., & Kürster, M. 2009, A&A, 496, 577

Appendix A: Algorithms available on GitHub

We do not describe the algorithms here (see Sect. 2.3 and 3.2), but rather the contents of the directory that is available online[6]. This directory contains algorithms written in Python 3.

In the directory "*Func/*", the user can find the implementation of algorithms 1 and 2. He/she can also find one function to evaluate the periodograms $\mathbf{p}$ and $\overline{\mathbf{p}}_L$ and one to evaluate the detection tests involved in algorithm 1. We also implemented several functions in a file to model a set of nuisance signal models $\mathcal{M}_d$, estimate the parameters θ_d of these models, and generate synthetic data from these models. Similarly, another file has been written for the set of nuisance signal models $\mathcal{M}_n$. Finally, the reader can find a file containing functions that save and read the results of the outputs of algorithm 2.

The current release implements the different setting (e.g., periodograms, considered set of frequencies, detection tests, and estimation of the noise model parameters) that were used in this work. Further settings will be added in the future (see Table. 1), but the users can already implement their own tools, such as new periodograms, detection tests, nuisance component models, parameter estimation procedure, prior laws, or stochastic component models. Note that the files containing algorithm 1, algorithm 2, and the read/save functiondo not need to be modified by the user.

As an illustration of how to use these released functions, we provide our liberally commented codes that were used to generate the numerical examples presented in Sect. 5.3 and 5.5 of the present paper (see "*Examples 1*" and "*Example 2*", respectively). The first example implements the case in which a null training sample (NTS) of the stochastic noise $\mathbf{n}$ is available. The second example implements the case without an NTS of the stochastic noise $\mathbf{n}$. In this case, the NTS is estimated from the RV series under test. We plan to improve this current release with new settings, a parallelization scheme, and practical examples of implementation for the most recent RV data analyses in the near future.
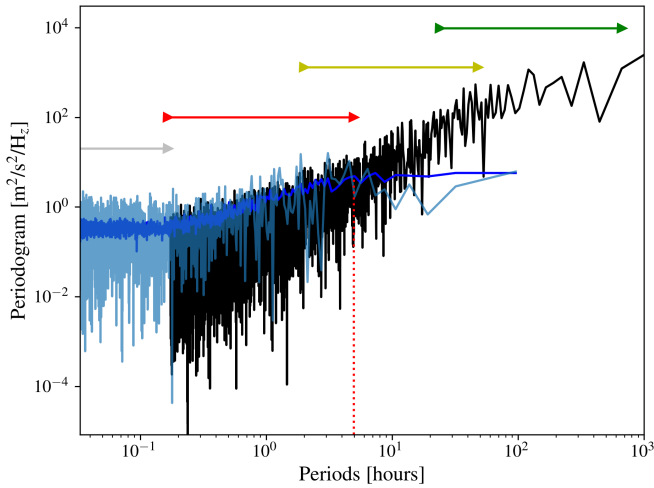
Appendix B: Main notations

Table B.1. Table of the main notations used in the paper.

Z, z	Z : Scalar random variable, z : one realization of Z
$\mathbf{z}$	Vector $[z_1, \ldots, z_N]^\top$
$\mathcal{H}_0, \mathcal{H}_1$	Null and alternative hypotheses
$\mathbf{x}$	Time series under test
$\mathbf{s}$	RV planet signal under $\mathcal{H}_1$
$\mathbf{n}$	Stochastic colored noise of covariance matrix Σ
$\mathbf{w}$	White Gaussian noise of variance $\sigma_{\mathbf{w}}^2$
$\mathbf{d}$	Nuisance signal
N	Number of data points
$\mathcal{T}_L$	Null training sample
L	Number of time series available in the NTS
$\mathbf{c}$	Ancillary time series that can be used to model $\mathbf{d}$
$\mathcal{M}_{\mathbf{s}}(\theta_{\mathbf{s}})$	Model of $\mathbf{s}$ depending on parameters vector $\theta_{\mathbf{s}}$
$\mathcal{M}_{\mathbf{n}}(\theta_{\mathbf{n}})$	Model of $\mathbf{n}$ depending on parameters vector $\theta_{\mathbf{n}}$
$\mathcal{M}_{\mathbf{d}}(\theta_{\mathbf{d}})$	Model of $\mathbf{d}$ depending on parameters vector $\theta_{\mathbf{d}}$
$\widehat{\theta}, \widehat{\delta}$	Estimate of parameter vector θ, and perturbation interval of this parameter
π	Parameters' prior distribution for $\widehat{\theta}$ with scale parameter $\widehat{\delta}$
ν_k	Frequencies
Ω	Indices set of considered frequencies
$\mathbf{p}$	Periodogram
$\overline{\mathbf{p}}_L$	Periodogram averaged with L training datasets
$\widetilde{\mathbf{p}} \mid \overline{\mathbf{p}}_L$	Periodogram standardized by $\overline{\mathbf{p}}_L$
T_M	Test of the highest periodogram value (Max test)
T_C	Test of the N_C^{th} highest periodogram value
(P, T)	Couple {periodogram, test}
Φ_{T}	CDF of the considered test statistic T
t	Test statistic
$\{B, b\}$	Number of Monte Carlo simulations in algorithm 2
p_v	p-value

---
[6] https://github.com/ssulis/3SD

Appendix C: Validation of MHD simulations with a ground-based RV dataset

The realism of the RVs extracted from 3D MHD simulations has been demonstrated in Sulis et al. (2020) by comparing them with spatial RV data taken from the GOLF/SoHO spectrophotometer (see Sect. 2.2.4 and Fig. 2 of Sulis et al. 2020). In this appendix, we extend this demonstration to solar ground-based RV data taken with HARPS-N (Collier Cameron et al. 2019; Dumusque et al. 2021). We selected the $L = 26$ synthetic MHD time series of solar granulation that are described in Sect. 4.2.1 and the 430-day solar time series of HARPS-N . Their estimated power spectral densities are shown in Fig. C.1. The match (amplitudes and frequency dependence of the noise) of the PSDs in the period regime where granulation dominates (red horizontal arrow) is very good. This example shows that MHD simulations, tuned for the stellar type of a target host star, could be used to provide realistic time series of the stellar granulation noise of this star.

Fig. C.1. PSD comparison of solar RVs taken with the HARPS-N telescope (black, 430 days) and synthetic RV extracted from 3D MHD simulations of solar granulation (light blue, 2 days). The averaged periodogram of the 26 MHD two-day time series is shown in dark blue. Period ranges dominates by high-frequency noise (instrumental and oscillations), granulation, supergranulation, and magnetic activity are indicated with gray, red, yellow, and green arrows. The dotted red vertical line indicates the period limit (∼5 hours) within which the PSD is no longer dominated by the granulation noise.