Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **468,** 3775–3784 (2017)
Advance Access publication 2017 March 16

doi:10.1093/mnras/stx664

Searching for planetary signals in Doppler time series: a performance evaluation of tools for periodogram analysis

Matteo Pinamonti,[1,2]★ Alessandro Sozzetti,[3] Aldo S. Bonomo[3] and Mario Damasso[3]

[1]*Dipartimento di Fisica, Università degli Studi di Trieste, via G. B.Tiepolo 11, I-34143 Trieste, Italy*
[2]*INAF – Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[3]*INAF – Osservatorio Astrofisico di Torino, Via Osservatorio 20, I-10025 Pino Torinese, Italy*

Accepted 2017 March 14. Received 2017 March 13; in original form 2017 February 15

ABSTRACT

We carry out a comparative analysis of the performance of three algorithms widely used to identify significant periodicities in radial-velocity (RV) data sets: the generalized Lomb–Scargle (GLS) periodogram, its modified version based on Bayesian statistics (BGLS) and the multifrequency periodogram scheme called FREquency DEComposer (FREDEC). We apply the algorithms to a suite of numerical simulations of (single and multiple) low-amplitude Keplerian RV signals induced by low-mass companions around M-dwarf primaries. The global performance of the three period search approaches is quite similar in the limit of an idealized, best-case scenario (single planets, circular orbits, white noise). However, GLS, BGLS and FREDEC are not equivalent when it comes to the correct identification of more complex signals (including correlated noise of stellar origin, eccentric orbits, multiple planets), with variable degrees of efficiency loss as a function of system parameters and degradation in completeness and reliability levels. The largest discrepancy is recorded in the number of false detections: the standard approach of residual analyses adopted for GLS and BGLS translates in large fractions of false alarms (∼30 per cent) in the case of multiple systems, as opposed to ∼10 per cent for the FREDEC approach of simultaneous multifrequency search. Our results reinforce the need for the strengthening and further development of the most aggressive and effective *ab initio* strategies for the robust identification of low-amplitude planetary signals in RV data sets, particularly now that RV surveys are beginning to achieve sensitivity to potentially habitable Earth-mass planets around late-type stars.

Key words: methods: data analysis – methods: statistical – techniques: radial velocities – planetary systems.

1 INTRODUCTION

The growing evidence from transit (e.g. *Kepler*) and radial-velocity (RV, e.g. HARPS, HARPS-N) surveys points towards a high occurrence rate of low-mass ($\leq$30 $M_\oplus$), small-size ($\leq$3 $R_\oplus$) planets (e.g. Mayor et al. 2011; Howard 2013), with a large fraction of late-type M dwarfs hosting habitable-zone terrestrial-type companions (see e.g. Winn & Fabrycky 2015, and references therein). The combined statistical inferences from HARPS and *Kepler* indicate that planets in the range between Super Earths and Neptunes are not only very common but they are often found in multiple systems, tightly packed close to the central star and almost perfectly coplanar when seen in transit (e.g. Batalha et al. 2013; Fabrycky et al. 2014; Rowe et al. 2014). The observational evidence is posing a formidable challenge for planet formation and evolution models, but it is also inducing a fundamental change of perspective in RV observing strategy. The ubiquitousness of multiple systems with low-mass components requires a very significant investment of observing time for a proper modelling of the complex signals. Usually, multiyear campaigns with hundreds of RVs are presented in discovery announcements of Super Earths and Neptune-like planets (e.g. Bonfils et al. 2013b; Astudillo-Defru et al. 2015). In addition, the analysis of low-amplitude signals is often complicated by stellar activity that can induce false positive signals mimicking the RV signature of a low-mass planet and induce systematic effects comparable in magnitude to (and even exceeding) the amplitudes of the sought after Keplerian signals (e.g. Pepe et al. 2013).

In the search for low-mass planets with spectroscopic surveys, the first step in the investigation of unevenly spaced RV time series relies on the identification of statistically significant periodic signals via a variety of implementations of a periodogram analysis.

★ E-mail: m.pinamonti.astro@gmail.com

The Lomb–Scargle periodogram (LS; Lomb 1976; Scargle 1982), which performs a full sine-wave fit over a large grid of trial frequencies, has historically been the first tool adopted for the task. More recently, some authors have extended the LS formalism to include weights for the measurement errors and constant offsets for the data in the generalized Lomb–Scargle (GLS) periodogram (Zechmeister & Kürster 2009) and generalizations based on Bayesian probability theory in the Bayesian Lomb–Scargle and Bayesian generalized Lomb–Scargle periodograms (BLS, BGLS) (Bretthorst 2001; Mortier et al. 2015). Due to the high fraction of low-mass multiple-planet systems, and also to the presence of activity-related signals in the data, the correct identification of multiple, low-amplitude signals is of course a central issue in RV time series analysis as applied to exoplanet science. However, all the above algorithms fit only a single sine-wave, or Keplerian signal, and multiple signals must be detected via subsequent fits and residual analysis. To overcome some of the shortcomings of standard periodograms when dealing with data containing two or more periodicities, Baluev (2013) has developed the multifrequency periodogram FREquency DEComposer (FREDEC).

In this work, we expand on the study by Mortier et al. (2015) and carry out a set of detailed numerical experiments aimed at (1) gauging the relative effectiveness of the GLS, BGLS and FREDEC algorithms, including completeness and false positives, and (2) understanding their biases and limitations when applied to the systematic search of single and multiple low-amplitude periodic signals produced by low-mass companions, using M dwarfs as choice of reference for the central star. The performance evaluation in the presence of representative complex signals element constitutes a novel analysis that has not been undertaken before, to our knowledge. This comparative study should not be interpreted as a way of ranking the intrinsic effectiveness of a periodogram analysis method against another. Rather, it has to be seen as one of the steps that will help towards the definition and implementation of the most aggressive and effective strategies (e.g. Dumusque et al. 2017; Hara et al. 2017, and references therein) for a robust identification of terrestrial planetary systems with state-of-the-art instrumentation (e.g. HARPS, HARPS-N) that guarantees metre-per-second accuracy, as well as next-generation facilities for extreme precision RV measurements, such as ESPRESSO. In Section 2, we describe the numerical setup adopted in our study, while the main results of our suite of simulations are presented in Section 3. We provide a summary and discussion of our findings in Section 4.

2 SIMULATION SETUP

2.1 Assumptions and caveats

The suite of simulated catalogues of RV observations described below and utilized in the analysis has been produced using a set of working assumptions and simplifications. In particular:

(i) the comparative performance evaluation of GLS, BGLS and FREDEC is expressed in terms of the dependence of the efficiency of signal recovery (parametrized through the theoretical false alarm probability FAP) on the main orbital elements it is expected to depend upon, i.e. orbital period P, eccentricity e, RV semi-amplitude K and the 'signal-to-noise' ratio K/σ, where σ is the single-measurement RV error. The adoption of the theoretical FAP rather than its calculation via bootstrap methods was dictated by the need to keep processing time within reasonable boundaries given the computational resources at our disposal;

(ii) the RV measurements are affected by a random (Gaussian) noise component. In one experiment, a simple synthetic stellar activity signal was added to the RV data. This was done as a metric of comparison with recent literature works, while a full-scale study of the effect of correlated stellar noise is left for future developments. We also did not consider the presence of outer companions, stellar or planetary, that would introduce long-term RV drifts;

(iii) up to two low-mass planets where simulated. The growing evidence for the existence of compact multiple systems with a number of planets significantly exceeding two naturally calls for relaxation of this assumption. Our aim is to identify proxies for interpreting in a simple manner any differences in behaviour of the three algorithms that might arise in the case of two-planet systems that might be used in a future work for easing the understanding of the efficiency of periodogram analyses carried out with a variety of methods in cases of even more complex RV signals.

(iv) In the simulations, we included the elements of the window function appropriate for reproducing the gaps in the data due to the seasonality of the observations as well as the alternation between day and night. The number of RV measurements per season (a few tens) was that typical of current RV surveys, rather than that used in very intensive observational campaigns (with hundreds of data points) focused on few targets. No prescriptions were made for either the generation of gaps in the data due to long stretches of bad weather or the generation of RVs with large uncertainties as if obtained under not optimal weather conditions.

2.2 Synthetic catalogues

We created several catalogues of synthetic RV time series. Each time series consists of N RV measurements y_i distributed over a number N_s of observing seasons, their respective times t_i and the associated errors σ_i ($i = 1, \ldots, N$). The Keplerian RV signal induced by the jth planetary companion is evaluated through the standard formula:

$$y_j(t) = K_j[e_j \cos\omega_j + \cos(\nu_j(t) + \omega_j)] + \gamma, \tag{1}$$

with ω_j the longitude of periastron, $\nu_j(t)$ the true anomaly and γ a constant offset. One obtains $\nu_j(t)$ in terms of e_j and the eccentric anomaly $E_j(t)$ as

$$\tan\frac{\nu_j(t)}{2} = \sqrt{\frac{1+e_j}{1-e_j}}\tan\frac{E_j(t)}{2}, \tag{2}$$

with $E_j(t)$ determined via iterative solution of Kepler's equation:

$$E_j(t) - e_j \sin E_j(t) = M_j(t) = 2\pi\frac{t - T_{0,j}}{P_j}, \tag{3}$$

where $M_j(t)$ is the mean anomaly and $T_{0,j}$ the time of periastron passage. From the orbital parameters, we can recover the planets' minimum mass $M_{p,j}\sin i_j$ using the relation:

$$M_{p,j}\sin i_j \propto K_j P_j^{1/3} M_\star^{-2/3}(1-e_j^2)^{1/2}, \tag{4}$$

where $M_\star$ is the mass of the primary. The values of $M_\star$ and γ were kept constant to $M_\star = 0.5\ \mathrm{M}_\odot$ and $\gamma = 0.0$ m s^{-1}, respectively, throughout our study.

The instrumental noise was modelled as purely white, with the single-measurement error σ_i drawn from a Gaussian distribution with standard deviation of 1.5 m s^{-1}, which is representative of typical values of internal errors in Doppler time series of relatively bright M dwarfs. The generation of the synthetic systems and relative RV signals was carried out with a set of prescriptions detailed below.

2.2.1 Single-planet circular orbits catalogue

The first catalogue consists of 10 000 synthetic systems composed of a single companion on a circular orbit ($e = 0.0$). The orbital parameters and RV amplitudes were drawn from the following distributions:

P: log-uniformly distributed over the interval [10.0, 365.25] d;
K: uniformly distributed over [1.5, 5.0] m s^{-1};
T_0: uniformly distributed over the range: [0, P].

Given the range of K and the adopted value of $M_\star$, the corresponding interval of minimum planetary masses is between $\sim$3 M$_\oplus$ and 30 M$_\oplus$. All 10 000 RV time series were generated with $N = 60$ observations uniformly distributed over $N_s = 3$. The season duration was set close to 6 months, with a daily observing window of approximately 12 h.

2.2.2 Single-planet eccentric orbits catalogue

The second catalogue is composed of 10 000 synthetic eccentric systems and their relative time series. The probability distribution function adopted for e was the Beta distribution, following the recipe of Kipping (2013):

$$\mathcal{P}_\beta(e; a, b) = \frac{1}{B(a, b)} e^{a-1}(1 - e)^{b-1}, \quad (5)$$

with $a = 0.867$ and $b = 3.03$. The remainder of the simulation setup was identical to that described in Section 2.2.1.

2.2.3 Multiplanet circular orbits catalogue

The third catalogue is composed of 10 000 synthetic two-planet systems on circular orbits and their relative time series. To generate each pair of companions, we first use the same P distribution as in the first two catalogues and then assign the orbital period P' of the second planet following the distribution of period ratios observed for Kepler candidates by Steffen & Hwang (2015):

$$\mathcal{P}(\mathcal{R}) \propto \mathcal{R}^{-1.26}, \quad (6)$$

where $\mathcal{R} = P_o/P_i$, P_i and P_o being the periods of the inner and outer planet, respectively. The relation is valid for $\mathcal{R} \gtrsim 2$. We do not require $P = P_i$, so the probability density function for P' is

$$\mathcal{P}(P'; P) = \begin{cases} \left(\dfrac{P}{P'}\right)^{-1.26}, & \text{if P}' < \text{P}/2, \\ \left(\dfrac{P'}{P}\right)^{-1.26}, & \text{if P}' > \text{P}/2. \end{cases} \quad (7)$$

P' was also required to be in the interval [10.0, 365.25] d. All other parameters in the simulated catalogue were generated following the same prescriptions as in Section 2.2.1. The resulting period ratio distribution is shown in Fig. 1.

We denote the largest and smallest amplitudes K_M and K_m, respectively, and the corresponding periods P_M and P_m. The distribution function of amplitude ratios is shown in Fig. 2.

2.2.4 Multiplanet eccentric orbits catalogue

The last catalogue generated encompassed a set of 10 000 eccentric two-planet systems and their corresponding RV time series. As done in Section 2.2.2, the e values for both orbits were drawn from the Beta distribution (Kipping 2013). In order to avoid unrealistic configurations corresponding to clearly dynamically unstable orbits, the masses, orbital separations and eccentricities of a pair of synthetic planets were generated in order to fulfil the analytic Hill-stability criterion (Giuppone, Morais & Correia 2013, and references therein):

$$\left(\mu_1 + \mu_2 \frac{a_1}{a_2}\right)\left(\mu_1\gamma_1 + \mu_2\gamma_2\sqrt{\frac{a_2}{a_1}}\right)^2 > \alpha^3 + 3^{4/3}\mu_1\mu_2\alpha^{5/3}, \quad (8)$$

with $\mu_i = m_i/m_\star$, $\alpha = \mu_1 + \mu_2$, a_i the semimajor axis of planet i and $\gamma_i = \sqrt{1 - e_i^2}$. Systems violating this criterion were discarded.

Since the stability criterion penalizes highly eccentric orbits, in order to avoid a statistically insignificant sample of highly eccentric wide systems, we cut the eccentricities distribution at the $e = 0.5$ level, which includes roughly 90 per cent of the systems.

In order to study the sensitivity to the $P/2$ harmonics of eccentric orbits, we raised the period ratio lower limit in equation (7) to $\mathcal{R} = 2.5$ to avoid overlapping with signals from planets in 2 : 1 resonance.

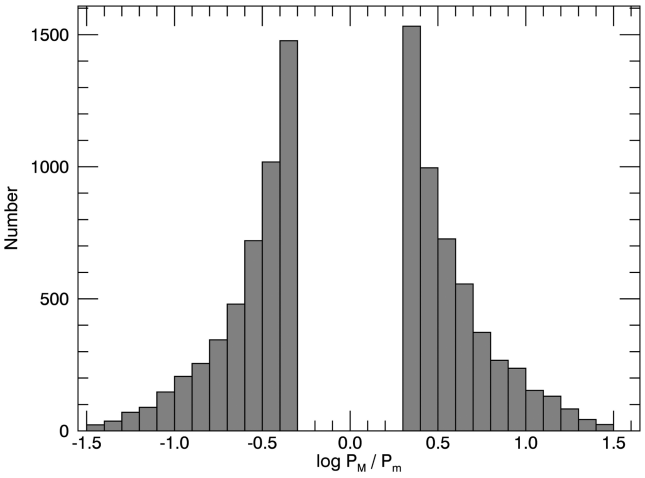

Figure 1. Period ratio distribution function, with P_M and P_m the period of the planet with the larger and smaller amplitudes, respectively.

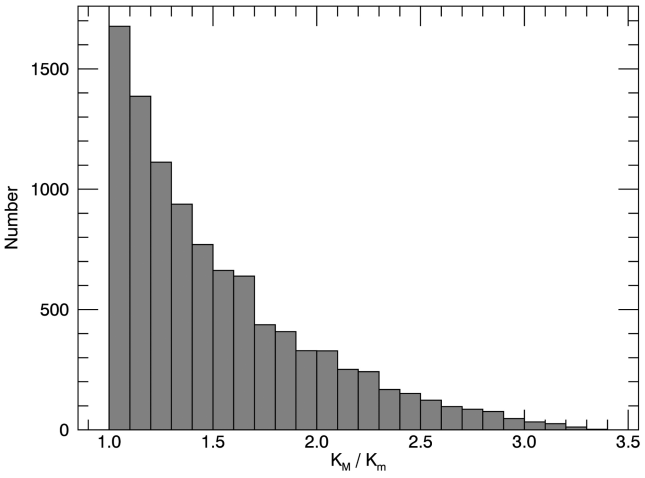

Figure 2. Amplitude ratio distribution function, with K_M and K_m the larger and smaller amplitudes, respectively.

3 RESULTS

The comparative study of the efficiency of the three period search algorithms presented here is carried out applying

sequentially GLS, BGLS and FREDEC to each of the four simulated data sets described in Section 2.2. Indeed, other studies in the past (e.g. Walker et al. 1995; Nelson & Angel 1998; Cumming, Marcy & Butler 1999, 2008; Eisner & Kulkarni 2001; Endl et al. 2002, 2006; Cumming 2004; Narayan, Cumming & Lin 2005; Bonfils et al. 2013a; Faria et al. 2016) have focused on gauging the sensitivity of RV planet searches to single-planet architectures utilizing periodogram analysis tools applied to synthetic as well as actual data sets in a variety of situations (large/small number of observations, periods shorter/longer than the duration of the observations, small and large companion masses). The systematic performance evaluation of GLS, BGLS and FREDEC in the single-planet case is useful in this context as it provides the opportunity to define and train on grounds that are better understood the comparison metrics to be used later for the comparative analysis of multiple circular and Keplerian signals, which has not been investigated in the past.

For the purpose of maximizing the homogeneity of the analysis, we have set the maximum value of FAP considered for evaluation of a signal at 10 per cent, driven by the inbuilt FAP < 0.1 limit in FREDEC (see Baluev 2013, Section 4.2). For GLS, the FAP has been calculated following equations (24) and (25) in Zechmeister & Kürster (2009). For BGLS, we followed Mortier et al. (2015) and adopted as FAP value the relative probability between the two highest peaks. In practice, statistically significant detections are considered only those with FAP below the threshold $\mathrm{FAP_{thr}} = 1 \times 10^{-3}$.

To further quantify the quality of the results of the different algorithms, we also calculated for each time series the true fractional error between the best output period P_{out} and the true simulated one P_{in}:

$$\Delta P = \frac{P_{\mathrm{in}} - P_{\mathrm{out}}}{P_{\mathrm{in}}}, \tag{9}$$

and considered a correct identification of a given period when $\Delta P < 0.1$. For FREDEC, we considered a planetary system as correctly identified if all the input periods were recovered in the output set with a fractional error lower than 10 per cent, even in the presence of additional output periodicities, as well as we considered as wrong solutions that did not contain the input periods, even if they contained some of their harmonics.

To compare the algorithms, we describe their performances by means of two global performance metrics: the completeness $C = N_{\mathrm{corr}}/N_{\mathrm{cat}}$ identifies the fraction of correctly identified planets signals N_{corr} with respect to the total simulated planets in the catalogue N_{cat}; the reliability $R = N_{\mathrm{corr}}/(N_{\mathrm{corr}} + N_{\mathrm{FP}})$ is the ratio of correct detections to the total of correct plus false alarms N_{FP}. Finally, we quantify dependences of the performance on the relevant parameters by using simple scaling relations expressing, for example, the detection efficiency as a function of the ratio K/σ between planetary signal amplitude and single-measurement uncertainty. All the analysis is carried out using $\mathrm{FAP} < \mathrm{FAP_{thr}}$.

3.1 Sanity check on white noise

The standard experiment to gauge the false alarm rate in the presence of pure white noise due to the statistical FAP threshold adopted for each algorithm should give expected results (e.g. 1 per cent of false positives for an FAP of 1 per cent). We have generated 10 000 time series with pure white noise, $N = 60$, and $N_s = 3$, and run the three algorithms sequentially. We show in Fig. 3 the fraction of false alarms as function of FAP threshold.

We can see that all three curves are systematically lower than the dashed line, corresponding to the ideal relation between FAP and number of false positives. All three algorithms appear robust against false positives, within the limits of the FAP definition for each method.

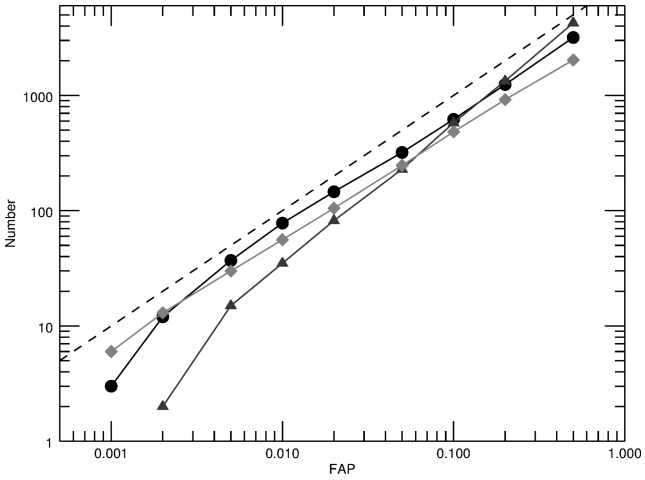

Figure 3. Number of false positives found in 10 000 white noise realizations as a function of the FAP threshold: black circles for the GLS, dark grey triangles for BGLS and light grey squares for FREDEC. The dashed line is the theoretical expectation.

Table 1. Circular orbits catalogue results.

	C (per cent)	FP fraction (per cent)	R (per cent)
GLS	94.0%	0.3%	99.6%
BGLS	87.9%	0.0%	100.0%
FREDEC	87.8%	0.4%	99.6%

3.2 Single-planet circular orbits catalogue

We applied GLS, BGLS and FREDEC on the circular orbits catalogue computing the periodograms at 10^3 logarithmically spaced periods over the interval $[1, 10^3]$ d.

In Table 1 are shown the overall C and R values for the three algorithms, along with the fraction of false positive signals found in the catalogue. All methods show very high C values, GLS performing slightly better (∼6 per cent) than BGLS and FREDEC. Reliability levels are virtually at 100 per cent for all methods, given the extremely low fraction of false positive signals. There is however a significant discrepancy in the level of concordance between the three methods, that is the fraction of detected systems that is common: only 80 per cent of all detected signals is in common between GLS, BGLS and FREDEC. These effects are best understood by looking at the structure of the dependence of the FAP on K/σ in the three cases.

As shown in Fig. 4 (upper two panels and bottom left panel), the FAP decreases approximately log-linearly with increasing K/σ, as expected, BGLS highlighting a steeper dependence and much larger spread in (statistically significant) FAP values in any given bin in K/σ. Furthermore, we note that for BGLS very high FAP values are obtained even for $K/\sigma \gtrsim 3$, which is not the case for GLS. FREDEC also highlights a systematically different behaviour with respect to GLS, stemming from its simultaneous multifrequency identification approach. In this case, the small fraction of high-FAP systems that is recorded, independently of K/σ, corresponds to systems in which

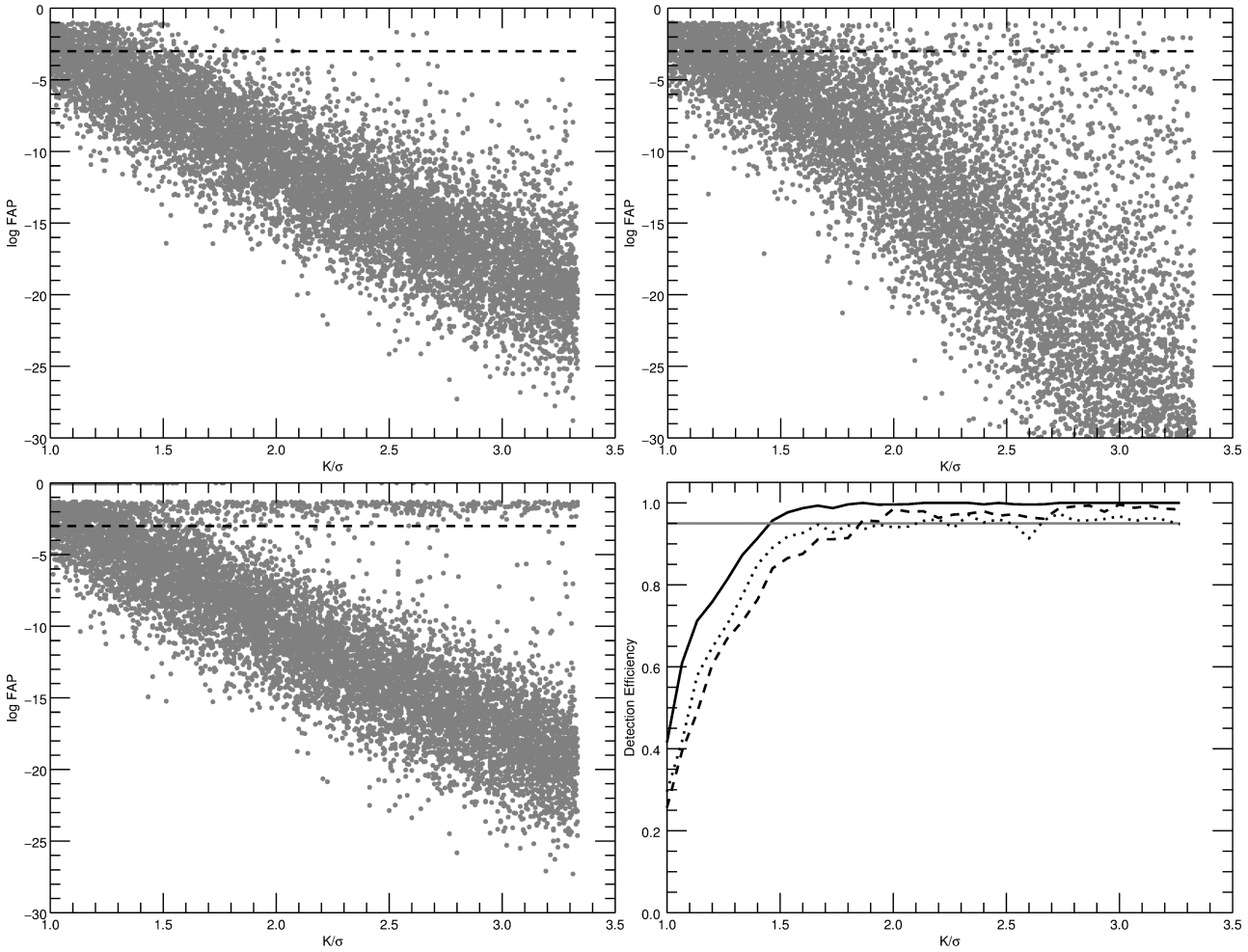

Figure 4. Dependence of the FAP on the K/σ for GLS (top left), BGLS (top right) and FREDEC (bottom left) applied to the circular orbits catalogue. The black dashed line represents the 10^{-3} FAP level. Bottom right: detection efficiency as a function of K/σ, for the circular orbits catalogue. The solid back line is for GLS, the dashed black line for BGLS and the dotted black line for FREDEC. The grey solid line indicates the 95 per cent level of detections with FAP $<$ FAP$_{\rm thr}$.

more than 1 signal is identified by FREDEC. No such cases are seen below the FAP$_{\rm thr}$ level.

The bottom right panel of Fig. 4 quantifies the dependence of detection efficiency on K/σ. For GLS, $K/\sigma \simeq 1.5$ is enough for correct recovery of the signals with >95 per cent efficiency, while this result is achieved by BGLS at $K/\sigma \simeq 2.0$. Unlike the other two methods, FREDEC never reaches close to the 100 per cent efficiency level, due to the systematic effect described above, that identifies ≈5 per cent of low-FAP systems, independently of K/σ. Overall, GLS appears ∼10 per cent more efficient than the other two algorithms, even in the limit of $K/\sigma \approx 1$. The results obtained here are in agreement with the findings of Cumming (2004) but highlight slight differences between the three algorithms.

We show in Fig. 5 the behaviour of FAP with P for GLS, BGLS and FREDEC. No clear dependence of the FAP on the period of the detected signals is derived. This confirms the behaviour found by Cumming (2004) using the LS periodogram coupled to a Keplerian fit, i.e. that the detection threshold is independent of P, for P shorter than the time span of the observations. However, a clear loss in sensitivity for BGLS is seen for periods around 180 d. This effect is related to the simulated length of the observing seasons and is observed neither in GLS nor in FREDEC. The feature in

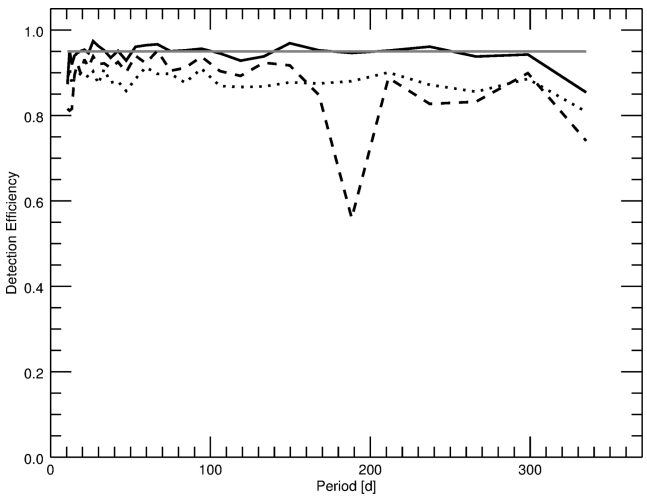

Figure 5. Dependence of the FAP on the orbital period for the three algorithms applied to the circular orbits catalogue. Line coding as in Fig. 4.

Table 2. Eccentric orbits catalogue results.

	C (per cent)	FP fraction (per cent)	R (per cent)
GLS	86.1%	0.9%	99.0%
BGLS	80.0%	0.2%	99.8%
FREDEC	76.0%	0.8%	98.9%

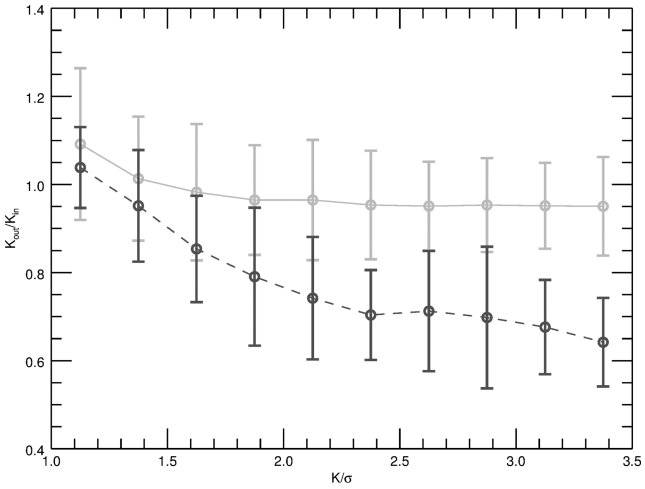

Figure 6. The ratio $K_{\rm out}/K$in as function of K/σ, for the $e < 0.5$ (light grey) and $e \geq 0.5$ (dark grey) samples.

correspondence of ∼180 d disappears from the BGLS analysis in the limit of higher sampling and unequal duration of each observing season (results not shown).

3.3 Single-planet eccentric orbits catalogue

All of the three algorithms fit pure sine waves.[1] We applied them (with the same boundaries in trial period as before) to a catalogue of eccentric signals, to gauge their different biases and limitations (such as spurious detections of harmonics produced by eccentric signals) in the correct identification of P and K as a function of the eccentricity. In the analysis, we distinguished between high and low eccentricity signals, the threshold being set to $e = 0.5$.

Also in this case, we find that GLS and BGLS are in excellent agreement on the output values of the first periodogram analysis when both their signals are significant. Table 2 shows again C, R and fraction of false positives of the different algorithms based on the analysis of the eccentric catalogue. We can see that both C and R are lower than for the circular orbit catalogue, while the fraction of false positives is higher. The behaviour of the individual algorithms is the same as before, with GLS being the most complete and BGLS the most reliable. As expected, most of the incorrect identifications come from time series in which no significant period is found and/or those with particularly high eccentricity. We next take a closer look at the results of the individual algorithms.

We show in Fig. 6 the ratio $K_{\rm out}/K_{\rm in}$ of the fitted amplitude to the input K value expressed as a function of K/σ for two regimes of eccentricity for GLS. The derived K is systematically underestimated for the high-e subsample. It is worth noticing that the result is opposite to that observed by Shen & Turner (2008) in their analysis of eccentric RV signals. In that work, a systematic overestimate of the fitted K values is a result of force-fitting Keplerian orbits with non-zero e even in the limit of $K/\sigma \simeq 1$, for which systematically large, and statistically not significant, eccentricities are obtained. The results for BGLS (not shown) are essentially identical.

[1] Zechmeister & Kürster (2009) also presented a fully Keplerian version of the GLS periodogram. The algorithm is significantly heavier computationally than its circular version, and it would have required applying Keplerian fits to the data analysed with BGLS and FREDEC as well in order to keep homogeneity, thus making this study impractical given the available computational resources.

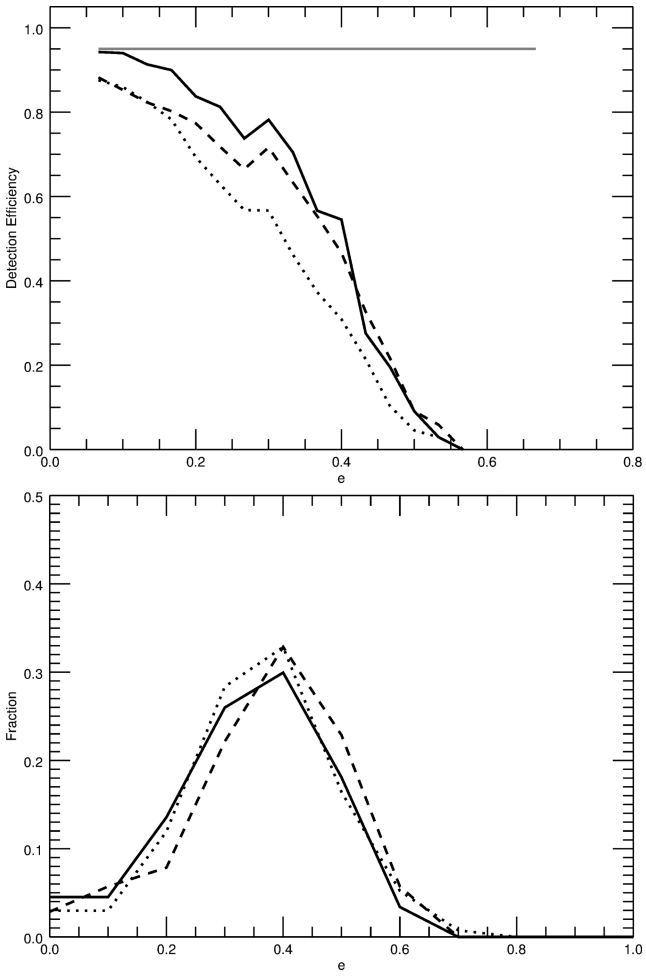

Figure 7. Top: detection efficiency above the 10^{-3} FAP threshold as function of eccentricity. The solid back line is for GLS, the dashed black line for BGLS and the dotted black line for FREDEC. The grey solid line indicates the 95 per cent level. Bottom: histogram of the fraction of significant periods identified in the residuals as function of eccentricity. Line coding is the same as in the upper panel.

Cumming (2004) observed a quick decrease in detection efficiency for systems with $e \gtrsim 0.6$, finding that for too high eccentricities it is impossible to reconstruct the planetary signals. We derive in Fig. 7 (top panel) a very similar result for all signal detection algorithms. For both GLS and BGLS, the detection efficiency drops to 50 per cent at $e \simeq 0.4$, and no signals are detected (even with the largest K/σ values) for $e \gtrsim 0.6$. As for FREDEC, the behaviour is also similar to that of GLS and BGLS, with its detection efficiency reaching zero for $e \approx 0.6$ (Fig. 7). However, an even steeper dependence of the algorithm on e is seen, with the efficiency already lower by a factor of 2 with respect to GLS and BGLS at $e \simeq 0.4$.

As force-fitting a full Keplerian orbit to a low-amplitude signal often results in badly constrained (and artificially high) e values, in practice signal subtraction is often carried out assuming a circular orbit. We carried out a GLS and BGLS analysis (with the same FAP thresholds as before) on the residuals to a circular-orbit fit to learn about the possible distortions in the time series induced by this approximation, particularly in the limit of high eccentricities for which residual power at first and higher order harmonics is expected.

From the results of the residual analysis, we note that the fraction of significant signals found increases with increasing e (Fig. 7, bottom panel), up to the eccentricity limit set by detection efficiency dropping to zero. For GLS, in 70 per cent of these systems the significant signal in the residuals is the first harmonic ($P/2$) of the input period. For BGLS, this happens in 55 per cent of the cases. As for FREDEC, twice as many multiple significant signals are identified with respect to the circular orbit case. In this sample, the first harmonic at $P/2$ is found in 49 per cent of the cases, with a mean eccentricity of $\langle e \rangle = 0.41$, which is significantly higher than the average on the subsample and on the whole catalogue.

Finally, for all algorithms, we tested whether increasing the length of the RV monitoring (up to five observing seasons) and/or doubling the number of observations per seasons (40 instead of 20) allowed us to (a) improve detection efficiency and/or (b) mitigate the underestimation of the K value. No statistically significant changes in the behaviour shown in Figs 6 and 7 were detected.

3.4 Additional experiment: correlated noise

As an additional experiment, we tested the performance of GLS and BGLS on a catalogue with a more realistic stellar noise model. We added a simple correlated stellar activity signal, modelled with the analytical recipe by Aigrain, Pont & Zucker (2012). Our model considered 200 stellar spots, a realistic value for an M dwarf (Barnes, Jeffers & Jones 2011) and a rotation period of 30 d; no differential rotation was included. We generated different spot distributions and sizes, in order to produce stellar activity signals with amplitudes $K_\star$ ranging between 1.5 m s^{-1} and 5 m s^{-1}. The planetary parameters were generated as in the circular orbits catalogue of Section 2.2.1.

We compared the results with an analogous catalogue with the same planetary signals but no stellar activity, in order to quantify the decrease in detection efficiency of the planetary signals present in each time series. For both algorithms, we used the same measure of relative detection efficiency utilized by Vanderburg et al. (2016) ($R_{\rm S/N}$, see their equation 1). For GLS, this is the square root of the ratio between the periodogram power measured with and without the stellar signal included, while for BGLS, the quantity is the ratio between two Bayesian probabilities. An analogous experiment was not carried out using FREDEC, as no direct output in terms of periodogram power can be obtained from the software in its release.

Vanderburg et al. (2016) found that the presence of correlated stellar noise produces a systematic degradation of $R_{\rm S/N}$ at all orbital periods investigated, with a stronger effect in the neighborhood of the stellar rotation period and its first two harmonics. As we can see in the bottom panel of Fig. 8, our analysis using GLS confirms the systematic effect. Furthermore, the simulations allow us to quantify the dependence of the loss of detection efficiency as a function of the amplitude ratio $K_\star/K_P$ (kept constant at $K_\star/K_P = 2$ by Vanderburg et al. 2016). The result is shown in the upper panel of Fig. 8, in which we plot the relative detection efficiency as function of $K_\star/K_P$. We can see that for $K_\star/K_P \simeq 2$ the detection efficiency integrated over all periods drops by about 30 per cent.

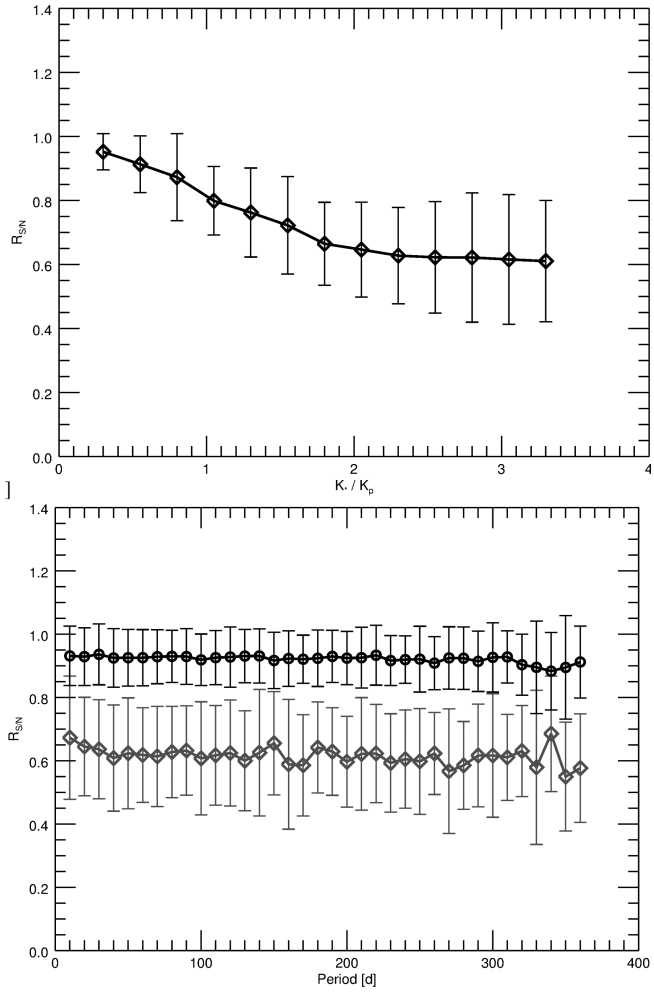

Figure 8. Top: relative detection efficiency as function of the amplitude ratio $K_\star/K_P$. The dots and error bars indicate the binned means and standard deviations. Bottom: relative detection efficiency as function of orbital period. The upper line is for the time series with $K_\star/K_P < 1$, the lower one for $K_\star/K_P > 2$.

The effect at the stellar rotation period and harmonics, discussed by Vanderburg et al. (2016), is not present in our analysis. The drop in $R_{S/N}$ observed by Vanderburg et al. (2016) was due to the subtraction of the fitted stellar activity signal from the RV data set, translating in additional dilution of the planetary signal. Instead, we did not use any mock activity indicators to correct the RVs for the stellar signals but simply studied the results of the periodogram analysis. It is also important to remember that the magnitude of the effect at $P_{\rm rot}$ and its harmonics depends on the stellar spot configuration, but investigation of these aspects is beyond the scope of this experiment.

The results with BGLS (not shown) follow similar trends, with the probability of the peak in presence of stellar activity being typically 10^2 and 10^3 times lower at $K_\star/K_P \simeq 2$ and $K_\star/K_P \simeq 3$, respectively.

3.5 Multiplanet circular orbits catalogue

For GLS and BGLS, analysis of the multiple-planet simulations in the case of circular orbits proceeded (adopting the same periodogram setup as before) up to the period search in the RV residuals after removal of the second planetary signal. For FREDEC, up to three significant peaks were recorded.

Table 3. Multiplanet circular orbits catalogue results.

	C (per cent)	FP fraction (per cent)	*R* (per cent)
GLS	73.1%	21.2%	77.5%
BGLS	61.0%	28.7%	68.0%
FREDEC	72.8%	8.5%	89.5%

We start by comparing directly the output periods of the GLS and BGLS algorithms. The first most significant period is identified by both GLS and BGLS in 100 per cent of the cases, thus both algorithms return the same results as in the single circular orbits catalogue (see Section 3.2). As expected, the same result is obtained in the analysis of the residuals after removal of the first and second significant periodicity, whenever the identified periods are the same for both algorithms (thus giving the same output structure of the post-fit residuals).

As we can see in Table 3, the levels of completeness and reliability for the correct detection of both injected planets are significantly lower than in the one planet case (see Table 1). Interestingly, BGLS shows the worst *C* value for this catalogue, thus proving its difficulties in dealing with multiple signals, as stated by Mortier et al. (2015). Both GLS and BGLS are prone to a large number of false positives, thus decreasing their *R* value. While completeness for FREDEC is similar to that of GLS, its *R* is significantly higher on the face of a much smaller number of false positives. This is likely due to the simultaneous multiple period search approach intrinsic to FREDEC.

The top panel Fig. 9 captures, for three methods, the effect on the global efficiency of detection of both signals on the ratio of amplitudes K_M/K_m. Efficiency never rises above ∼80 per cent for either of the three algorithms. This value is maximum at $K_M/K_m \approx 1$, the loss of ∼20 per cent being due to the sample of systems with similar amplitudes, both close to the single-measurement precision. At $K_M/K_m \approx 3$, efficiency is lower by a typical factor of 2–3, quantifying the difficulty in identifying correctly a second planet with $K_m \simeq \sigma$ in the presence of a larger amplitude signal, within the simulated observational scenario. Among the three methods, BGLS appears to suffer the most, performing typically a factor 1.3–2 with respect to GLS and FREDEC. We next turn to discuss some detailed features of the analysis carried out with each of the algorithms.

No significant signals are detected by GLS in 5.7 per cent of the systems. This occurs when both the input amplitudes are small, typically with $K/\sigma \lesssim 1.8$ in both cases, and with the amplitude ratio being typically close to unity. There is no clear dependence on the periods or their ratio. For 18.6 per cent of the systems, only one significant period is identified. The input periods of this subsample are usually both long (typically ∼150 d), and the ratio between the largest and the smallest amplitude is typically ∼2. In Fig. 10, we show the period distribution for the output and input for this subsample: the distribution is almost the same, except for a clear aliasing effect for a significant fraction of systems with the strongest signal at ∼1 yr, which are identified instead as being systems at 6 months of orbital period. The above results highlight some of the potential limitations for detection of these specific architectures of multiple-planet systems.

GLS finds two significant periodicities in 75.1 per cent of the time series. In the overwhelming majority of cases (96 per cent), two input signals are both identified correctly. In the remainder of the cases, incorrect identification of one or both periods is related to systems in which aliases created by the window function and its

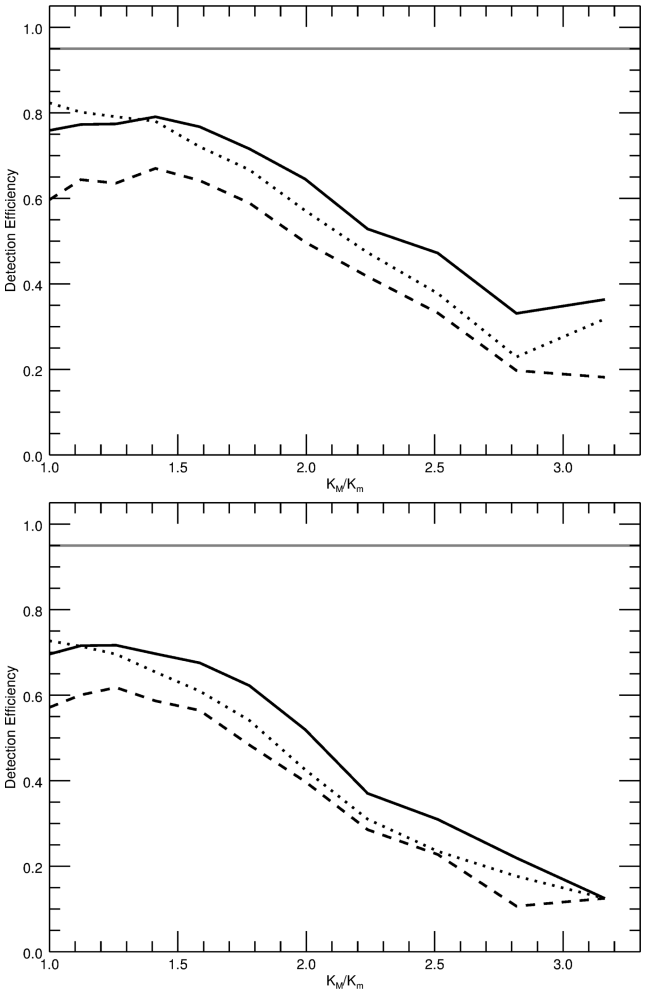

Figure 9. Top: detection efficiency above the 10^{-3} FAP in the multiple-planet, circular orbits case. The solid back line is for GLS, the dashed black line for BGLS and the dotted black line for FREDEC. The grey solid line indicates the 95 per cent level. Bottom: the same, for the multiple Keplerian orbits case. Line coding is the same as in the upper panel.

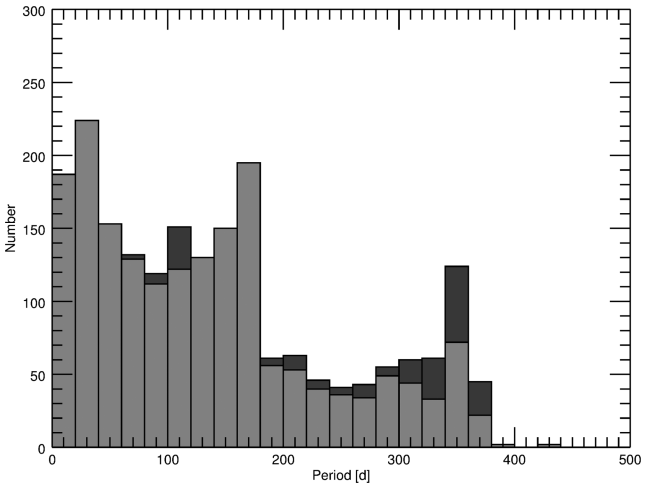

Figure 10. Period distribution for GLS, on the multiplanet circular orbits catalogue. The light grey area shows the output and the dark grey the strongest input signal, for the subsample with only one significant period identified.

Table 4. Multiplanet eccentric orbits catalogue results.

	C (per cent)	FP fraction (per cent)	R (per cent)
GLS	65.0%	26.3%	71.2%
BGLS	55.2%	33.1%	62.5%
FREDEC	62.2%	12.4%	83.4%

harmonics are detected. In only 0.65 per cent of the cases a third additional significant period is found after removal of the first two. This small sample is dominated by short-period aliases.

In the BGLS analysis, the fraction of 0, 1, 2 and 3 detected periods (with FAP $< 10^{-3}$) is 10.3 per cent, 27.5 per cent, 61.9 per cent and 0.3 per cent, respectively. The global features of the sub-samples in the four cases are essentially identical to those discussed for the GLS cases. It is worth noticing the significant increase in null detections and in detections of only one period, which explains the lower C value for BGLS in this experiment.

In the FREDEC analysis, the fraction of 0, 1, 2 and 3 detected periods (with FAP $< 10^{-3}$) is 18.7 per cent, 0.0 per cent, 80.8 per cent and 0.6 per cent, respectively. The distributions of amplitudes and periods in the cases of no detections are similar to those of GLS and BGLS, although with somewhat larger average ratio of amplitudes and $K/\sigma \simeq 1.5$ for the smaller of the two amplitudes in a system. The fraction of systems with two detected period is characterized by slightly longer periods and smaller amplitudes with respect to the GLS and BGLS cases, and a slightly lower fraction (90 per cent) of systems with both periods correctly identified is also recorded. Similarly to GLS and BGLS, incorrect identification of one or both periods is related to systems in which aliases are detected that are created by the window function and its harmonics. Contrary to GLS and BGLS, in the 0.6 per cent of cases with three significant periods detected, the sample is dominated by longer period aliases (e.g. 1 yr).

3.6 Multiplanet eccentric orbits catalogue

The analysis of the multiple eccentric orbits catalogue was performed as in the previous section. The completeness and reliability levels of the algorithms are listed in Table 4, and as we can see both are lower than for the previous catalogue. Again BGLS shows the lowest C value, and also in this case FREDEC and GLS have comparable C values but the former has higher R and half has many false positives. The lower completeness levels translate in larger values of null detections and detections of only one significant period.

The dependence of the number of detected systems on the main parameters (amplitude, period, eccentricity) generally follows the same behaviour observed in the previous experiments, for all methods. In particular, the mean amplitude increases with increasing number of signals found (as in Section 3.5) and the average eccentricity of both planets is lowest ($\sim$0.15) when all signals are correctly identified (as in Section 3.3). As in the multiple circular orbits case, the overall behaviour of detection efficiency for all three methods is mostly sensitive to the amplitude ratio, as demonstrated by the plot in the bottom panel of Fig. 9. The impact of eccentric orbits is quantified in an additional efficiency loss of 10–20 per cent, slightly increasing towards larger $K_{\rm M}/K_{\rm m}$ values.

There is however one difference: in the circular catalogue, the average period increased when more signals than in the input were found, while in this case it decreases. This is likely because the excess of signals recovered is due to poorly reconstructed orbits, which for the circular catalogue is mainly due to not optimal sampling (and thus long periods), while for the eccentric catalogue the extra signals found are also due to harmonics of the eccentric orbits, whose impact becomes more significant with better orbit sampling. As a matter of fact, the fraction of cases when in addition to both the input signals one or both the harmonics are detected is 72.2 per cent, 77.8 per cent and 86.3 per cent for the GLS, BGLS and FREDEC, respectively, thus dominating over spurious detections.

4 SUMMARY AND DISCUSSION

In this paper, we have carried out an extensive suite of numerical experiments aimed at a direct performance evaluation of three commonly adopted algorithms (GLS, BGLS and FREDEC) in the search of significant periodicities in RV data sets, indicative of the presence of planetary companions. Using simple scaling relations (detection efficiency) and global performance metrics (completeness, reliability, false positives fraction), we have gauged the strengths and weaknesses of the three period search algorithms when run on a variety of classes of Doppler signals (one and two planets, circular and fully Keplerian orbits) of low amplitude ($1 \lesssim K/\sigma \lesssim 3$), with representative realizations of observational strategies, different measurement noise prescriptions (simple Gaussian noise, stellar correlated noise) and adopting as reference an M dwarf primary. The main results can be summarized as follows.

(i) The degree of completeness and reliability are very high for GLS, BGLS and FREDEC in the single-planet, circular orbit case, with GLS being slightly more complete than the latter two methods. As a consequence, the fraction of false positives is very low. The overall detection efficiency is close to 100 per cent for all methods as long as $K/\sigma \gtrsim 2$, with a sharp decrease below 50 per cent in the limit $K/\sigma \simeq 1$. Also in this cases, GLS appears to be slightly (10–15 per cent) more efficient than BGLS and FREDEC in signal recovery when RV amplitudes get close to the single-measurement error.

(ii) The effect of eccentricity on correct signal identification by all methods is significant, as expected. A typical loss of 10 per cent in completeness is found, with GLS returning again the largest C value. Reliability of detections remains however close 100 per cent given the mild increase in false detections. The latter are a clear function of increasing e, as long as detection efficiency remains above $\sim$50 per cent. The loss in efficiency of period recovery is a steep decreasing function of e, dropping to zero for all algorithms for $e \gtrsim 0.6$. However, FREDEC shows a higher sensitivity to this parameter, with detection efficiency reduced by up to a factor of 2 in regime of intermediate e.

(iii) A preliminary investigation of the levels of degradation of detection efficiency in the presence of stellar correlated noise indicates efficiency losses of 20 per cent to 40 per cent in the range $1 \lesssim K_\star/K_P \lesssim 3$ for GLS and decrease of 2–3 orders of magnitude in the Bayesian probability of a detection for BGLS in the same $K_\star/K_P$ interval.

(iv) The difficulty in correctly identifying multiple planets is quantified through a typically reduced completeness level between 70 per cent (circular orbits) and 60 per cent (for Keplerian orbits), with BGLS performing slightly worse (10 per cent) with respect to the other two methods. Within the realm of the simulation scenario, and based on an analysis of the dependence of detection efficiency on the amplitude ratio $K_{\rm M}/K_{\rm m}$, the limitations induced by sub-optimal orbit sampling (particularly in the case of eccentric orbits) indicate as the most challenging architectures those

containing signals with very similar amplitudes and $K \lesssim 1.8$ m s^{-1}. In configurations containing two long-period companions with dissimilar amplitudes, the one with the lowest K value is not detected in a significant fraction of cases (particularly for $K_M/K_m \gtrsim 2$). Degradation in the degree of reliability is also clear on the face of large fractions (∼30 per cent) of false detections. In this respect, FREDEC appears more reliable than GLS and BGLS, with a false positive rate ∼10 per cent.

The results presented in this paper complement and extend the comparative analysis of period search tools for planet detection in RV data sets carried out by Mortier et al. (2015). Our study encompasses a wide range of single-planet architectures, it includes a preliminary assessment of the effects of increasing levels of stellar correlated noise, and it addresses for the first time some of the complications induced by multiple-planet architectures. The most important lessons learned are the following: (1) even under idealized, best-case conditions (one planet, circular orbits, white noise, well-sampled orbits) different period search algorithms do not perform in an exactly identical fashion, particularly when it comes to the regime of signal amplitudes close to the single-measurement error; (2) in the presence of more complex signals, the most conspicuous element to underline is the different behaviour in the identification of false alarms: the standard approach of successive signals removal and investigation of the residuals (using GLS and BGLS) appears to be prone to as much as three times the amount of false positives obtained by an approach in which all statistically significant signals are searched simultaneously (using FREDEC), even in the idealized case of perfectly circular orbits.

The analysis presented here is by no means exhaustive. Within the scope of this work, our results nevertheless underscore the urgent need for strengthening and further developing sophisticated analysis techniques for the simultaneous identification of low-amplitude planetary signals in the presence of stellar activity. This is a crucial topic in the case of low-mass M-type hosts, for which stellar noise is often coupled to complex planetary RV signals induced by small-mass multiple systems, as testified by the significant literature presenting disputes on the nature, interpretation and sometimes existence of multiple planets around some of our nearest low-mass neighbours (e.g. GJ 581, Kapteyn's star. See Anglada-Escudé et al. (2016b), and references therein). This is a particularly sensitive issue as M dwarf primaries constitute the fast track to the identification of potentially habitable terrestrial-type planets, whose abundance, albeit with large uncertainties, appears to be very high (e.g. Dressing & Charbonneau 2013; Kopparapu 2013; Bonfils et al. 2013a,b, Tuomi et al. 2014; Anglada-Escudé et al. 2016a).

It will be certainly necessary to use the largest possible set of observational constraints, including simultaneous photometric measurements for determining rotation periods and activity signals and spectroscopic indicators and/or RV measurements at different wavelengths for mitigating and (hopefully) removing activity signals (e.g. Vanderburg et al. 2016, and references therein). It will be equally important, however, to pursue aggressively advances in the path to the determination of the complete information content of RV data sets not only via techniques that shy away from the standard residual analysis and implement global model fitting approaches (e.g. Dumusque et al. 2017; Hara et al. 2017) but also through the application of improved methodologies for the simultaneous, robust identification of credible signals in time series (with very small fractions of false alarms), of which algorithms such as FREDEC constitute possible seeds. This necessity is expected to become pressing very soon, with facilities for ultra-high precision RV work such as ESPRESSO that will seek to find (multiple) planetary signals with amplitudes even orders of magnitude smaller than other sources (primarily stellar in nature) of correlated RV variations.

ACKNOWLEDGEMENTS

MP acknowledgesthe financial support of the 2014 PhD fellowship programme of INAF. MD acknowledges funding from INAF through the Progetti Premiali funding scheme of the Italian Ministry of Education, University, and Research. We also thank the anonymous referee for the swift and useful review.

REFERENCES

Aigrain S., Pont F., Zucker S., 2012, MNRAS, 419, 3147
Anglada-Escudé G. et al., 2016a, Nature, 536, 437
Anglada-Escudé G. et al., 2016b, ApJ, 830, 74
Astudillo-Defru N. et al., 2015, A&A, 575, A119
Baluev R. V., 2013, Astron. Comput., 3, 50
Barnes J. R., Jeffers S. V., Jones H. R. A., 2011, MNRAS, 412, 1599
Batalha N. M. et al., 2013, ApJS, 204, 24
Bonfils X. et al., 2013a, A&A, 549, A109
Bonfils X. et al., 2013b, A&A, 556, A110
Bretthorst G. L., 2001, in Mohammad-Djafari A., ed., AIP Conf. Ser. Vol. 568, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 241
Cumming A., 2004, MNRAS, 354, 1165
Cumming A., Marcy G. W., Butler R. P., 1999, ApJ, 526, 890
Cumming A., Butler R. P., Marcy G. W., Vogt S. S., Wright J. T., Fischer D. A., 2008, PASP, 120, 531
Dressing C. D., Charbonneau D., 2013, ApJ, 767, 95
Dumusque X. et al., 2017, A&A, 598, A133
Eisner J. A., Kulkarni S. R., 2001, ApJ, 550, 871
Endl M., Kürster M., Els S., Hatzes A. P., Cochran W. D., Dennerl K., Döbereiner S., 2002, A&A, 392, 671
Endl M., Cochran W. D., Kürster M., Paulson D. B., Wittenmyer R. A., MacQueen P. J., Tull R. G., 2006, ApJ, 649, 436
Fabrycky D. C. et al., 2014, ApJ, 790, 146
Faria J. P. et al., 2016, A&A, 589, A25
Giuppone C. A., Morais M. H. M., Correia A. C. M., 2013, MNRAS, 436, 3547
Hara N. C., Boué G., Laskar J., Correia A. C. M., 2017, MNRAS, 464, 1220
Howard A. W., 2013, Science, 340, 572
Kipping D. M., 2013, MNRAS, 434, L51
Kopparapu R. K., 2013, ApJ, 767, L8
Lomb N. R., 1976, Ap&SS, 39, 447
Mayor M. et al., 2011, preprint (arXiv:1109.2497)
Mortier A., Faria J. P., Correia C. M., Santerne A., Santos N. C., 2015, A&A, 573, A101
Narayan R., Cumming A., Lin D. N. C., 2005, ApJ, 620, 1002
Nelson A. F., Angel J. R. P., 1998, ApJ, 500, 940
Pepe F. et al., 2013, Nature, 503, 377
Rowe J. F. et al., 2014, ApJ, 784, 45
Scargle J. D., 1982, ApJ, 263, 835
Shen Y., Turner E. L., 2008, ApJ, 685, 553
Steffen J. H., Hwang J. A., 2015, MNRAS, 448, 1956
Tuomi M., Jones H. R. A., Barnes J. R., Anglada-Escudé G., Jenkins J. S., 2014, MNRAS, 441, 1545
Vanderburg A., Plavchan P., Johnson J. A., Ciardi D. R., Swift J., Kane S. R., 2016, MNRAS, 459, 3565
Walker G. A. H., Walker A. R., Irwin A. W., Larson A. M., Yang S. L. S., Richardson D. C., 1995, Icarus, 116, 359
Winn J. N., Fabrycky D. C., 2015, ARA&A, 53, 409
Zechmeister M., Kürster M., 2009, A&A, 496, 577

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