Astronomy & Astrophysics manuscript no. RV_challenge_paper_II_v4 ©ESO 2016
September 14, 2016

Radial-Velocity Fitting Challenge ⋆

II. First results of the analysis of the data set

X. Dumusque[1,2] ⋆⋆, F. Borsa[3], M. Damasso[4], R. Díaz[1], P. C. Gregory[5], N.C. Hara[6], A. Hatzes[7], V. Rajpaul[8], M. Tuomi[9], S. Aigrain[8], G. Anglada-Escudé[9,10], A.S. Bonomo[4], G. Boué[6], F. Dauvergne[6], G. Frustagli[3], P. Giacobbe[4], R. D. Haywood[2], H. R. A. Jones[9], M. Pinamonti[11,12], E. Poretti[3], M. Rainer[3], D. Ségransan[1], A. Sozzetti[4], and S. Udry[1]

[1] Observatoire de Genève, Université de Genève, 51 ch. des Maillettes, CH-1290 Versoix, Switzerland e-mail: xavier.dumusque@unige.ch
[2] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA
[3] INAF – Osservatorio Astronomico di Brera, Via E. Bianchi 46, 23807 Merate (LC), Italy
[4] INAF – Osservatorio Astrofisico di Torino, via Osservatorio 20, 10025, Pino Torinese, Italy
[5] Physics and Astronomy Department, University of British Columbia, 6224 Agricultural Rd., Vancouver, BC V6T 1Z1, Canada
[6] ASD, IMCCE-CNRS UMR8028, Observatoire de Paris, UPMC, 77 Av. Denfert-Rochereau, 75014 Paris, France
[7] Thüringer Landessternwarte Tautenburg, Sternwarte 5, 07778 Tautenburg, Germany
[8] Sub-department of Astrophysics, Department of Physics, University of Oxford, Oxford OX1 3RH, UK
[9] University of Hertfordshire, Centre for Astrophysics Research, Science and Technology Research Institute, College Lane, AL10 9AB, Hatfield, UK
[10] School of Physics and Astronomy, Queen Mary University of London, 327 Mile End Rd., E1 4NS, London, UK
[11] Dipartimento di Fisica, Universita degli Studi di Trieste, via G. B.Tiepolo 11, I-34143 Trieste, Italy
[12] INAF – Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143, Trieste, Italy

Received XXX; accepted XXX

ABSTRACT

Context. Radial-velocity (RV) signals arising from stellar photospheric phenomena are the main limitation for precise RV measurements Those signals induce RV variations an order of magnitude larger than the signal created by the orbit of Earth-twins, thus preventing their detection.

Aims. Different methods have been developed to mitigate the impact of stellar RV signals. The goal of this paper is to compare the efficiency of these different methods to recover extremely low-mass planets despite stellar RV signals. However, because observed RV variations at the meter-per-second precision level or below is a combination of signals induced by unresolved orbiting planets, by the star, and by the instrument, performing such a comparison using real data is extremely challenging.

Methods. To circumvent this problem, we generated simulated RV measurements including realistic stellar and planetary signals. Different teams analyzed blindly those simulated RV measurements, using their own method to recover planetary signals despite stellar RV signals. By comparing the results obtained by the different teams with the planetary and stellar parameters used to generate the simulated RVs, it is therefore possible to compare the efficiency of these different methods.

Results. The most efficient methods to recover planetary signals **take into account the different activity indicators,** use red-noise models to account for stellar RV signals and a Bayesian framework to provide model comparison in a robust statistical approach. Using the most efficient methodology, planets can be found down to $K/N = K_{\rm pl}/{\rm RV}_{\rm rms} \times \sqrt{N_{\rm obs}} = 5$ with a threshold of $K/N = 7.5$ at the level of 80-90% recovery rate found for a number of methods. These recovery rates drop dramatically for K/N smaller than this threshold. In addition, for the best teams, no false positives with $K/N > 7.5$ were detected, while a non-negligible fraction of them appear for smaller K/N. A limit of $K/N = 7.5$ seems therefore a safe threshold to attest the veracity of planetary signals for RV measurements with similar properties to those of the different RV fitting challenge systems.

Key words. techniques: radial velocities – planetary systems – stars: oscillations – stars: activity – methods: data analysis

1. Introduction

The radial-velocity (RV) technique is an indirect method that measures with Doppler spectroscopy the stellar wobble induced by a planet orbiting its host star. The technique is sensitive not only to possible companions, but also to signals induced by the host star. Now that the $\rm m\,s^{-1}$ precision level has been reached by the best spectrographs, it is clear that solar-like stars introduce signals at a similar level. Those stellar signals, often referred to as *stellar jitter*, currently prevent the RV technique from detecting and measuring the mass of Earth-twins orbiting solar-type stars, i.e., Earth analogues orbiting in the habitable zone of GK dwarfs, because such planets induce signals an order of magnitude smaller. It is therefore extremely important to investigate new approaches to mitigate the impact of stellar signals if we want the RV technique to be efficient at characterizing the Earth-twins that will be found by TESS (Ricker et al. 2014) and PLATO (Rauer et al. 2014).

⋆ Based on observations collected at the La Silla Parana Observatory, ESO (Chile), with the HARPS spectrograph at the 3.6-m telescope.

⋆⋆ Society in Science – Branco Weiss Fellow (url: http://www.society-in-science.org)

At the $\mathrm{m\,s^{-1}}$ precision level, RV measurements are affected by stellar signals, that depend on the spectral type of the observed star (Dumusque et al. 2011c; Isaacson & Fischer 2010; Wright 2005). For GK dwarfs, those stellar signals can be decomposed, to our current knowledge, in four different components:

- solar-type oscillations (Dumusque et al. 2011c; Arentoft et al. 2008; O'Toole et al. 2008; Kjeldsen et al. 2005),
- granulation phenomena (Dumusque et al. 2011c; Del Moro et al. 2004; Del Moro 2004; Lindegren & Dravins 2003; Dravins 1982),
- short-term activity signals on the stellar rotation period timescale (Haywood et al. 2016; Borgniet et al. 2015; Robertson et al. 2015, 2014; Dumusque et al. 2014a; Boisse et al. 2012; Saar 2009; Meunier et al. 2010; Saar & Donahue 1997),
- and long-term activity signals on the magnetic cycle period timescale (Lanza et al. 2016; Díaz et al. 2016; Meunier & Lagrange 2013; Lovis et al. 2011; Dumusque et al. 2011a; Makarov 2010).

For more details about these signals and their origins, readers are referred to Section 2 in Dumusque (2016) and references therein.

Stellar signals creates RV variations that are larger than the signal induced by small-mass exoplanets, such as Earth-twins. There is several examples in the literature, where by analyzing the same RV measurements different teams detected different planetary configurations. This is the case of the famous planetary system GJ581, for which the number of planet detected is ranging between 3 and 6 (Hatzes 2016; Anglada-Escudé & Tuomi 2015; Robertson et al. 2014; Baluev 2013; Vogt et al. 2012; Gregory 2011; Vogt et al. 2010; Mayor et al. 2009), of HD40307, for which 4 to 6 planets have been announced (Díaz et al. 2016; Tuomi et al. 2013), and GJ667C, for which 3 to 7 planets have been detected (Feroz & Hobson 2014; Anglada-Escudé et al. 2012; Gregory 2012). All those systems are affected by stellar signals, and therefore depending on the model used to analyze the data, different teams arrives to different conclusions. This shows that optimal models do not exist at the moment to analyze RV measurements affected by stellar signals and this pushes the community towards finding an optimal solution. The RV fitting challenge is one of the efforts pursued today in this direction. The development of the HARPS-N solar telescope (Dumusque et al. 2015) is another one that should deliver the optimal data set for characterizing and understanding stellar signals in detail.

In principle, the nature of RV stellar and planetary signals is different. RV signal induced by a planet is periodic over time, while stellar signals are in the best case semi-periodic. In addition, a planet induces a pure Doppler shift of the observed stellar spectrum, while stellar signals change the shape of the spectral lines. Therefore, it should be possible to find techniques to differentiate between planetary and stellar signals.

Stellar oscillations are often averaged out in RV surveys by fixing an exposure time to 15 minutes. To obtain the best RV precision, it is also possible to observe the same star several times per night, with measurements spread out during the night, to sample better the signature of granulation and supergranulation (Dumusque et al. 2011c). It has been shown that this simple approach reduces the observed daily RV rms of measurements, however it does not fully average out this signal (Meunier et al. 2015; Dumusque et al. 2011c), and more optimal techniques need to be investigated. For short-term activity, which is by far the most difficult stellar signal to deal with due to the non-periodic, stochastic, long-term signals arising from the evolution and decay of active regions, several correction techniques have been investigated:

- fitting sine waves at the rotation period of the star and harmonics (Boisse et al. 2011),
- using red-noise models to fit the data (e.g. Feroz & Hobson 2014; Gregory 2011; Tuomi et al. 2013),
- using the FF′ method if contemporaneous photometry exists (Dumusque et al. 2015; Haywood et al. 2014; Aigrain et al. 2012),
- modeling activity-induced signals in RVs with Gaussian process regression, whose covariance properties are shared either with the star's photometric variations (Haywood et al. 2014; Grunblatt et al. 2015) or a combination of several spectroscopic indicators (Rajpaul et al. 2015), or determined from the RVs themselves (Faria et al. 2016),
- using linear correlations between the different observables, i.e., RV, bisector span (BIS SPAN) and full width at half maximum (FWHM) of the cross correlation function (CCF, Baranne et al. 1996; Pepe et al. 2002), photometry (Robertson et al. 2015, 2014; Boisse et al. 2009; Queloz et al. 2001), and magnetic field strength (Hébrard et al. 2014),
- checking for season per season phase incoherence of signals (Santos et al. 2014; Dumusque et al. 2014b, 2012),
- avoiding the impact of activity by using wavelength dependence criteria for RV signal (e.g. in HD40307 and HD69830, Tuomi et al. 2013; Anglada-Escudé & Butler 2012).

Finally, long-term activity seems to correlate well with the calcium chromospheric activity index, which provides a promising approach to mitigation of this source of stellar RV noise (Lanza et al. 2016; Díaz et al. 2016; Meunier & Lagrange 2013; Dumusque et al. 2012).

The goal of this paper is to test the efficiency of different approaches to retrieve low eccentricity planetary signals despite stellar signals. To do so, we present the results of a RV fitting challenge, where several teams analyzed blindly the same set of real and simulated RV measurements affected by planetary and stellar signals. Each team used their own method to recover planetary signals despite stellar signals. At the $\mathrm{m\,s^{-1}}$ precision level reached by the best spectrographs, RV measurements are affected by unresolved planets, but also stellar and instrumental signals. Without knowing which part of the RV variations is due to planets and which is due to the star or the instrument, it is extremely difficult to test which method is the most efficient at finding low-mass planets despite stellar signals. For such an exercise, it is crucial to use simulated RV measurements so that a comparison can be performed between the results of the different analysis and what was initially injected into the data. The set of simulated and real RV measurements used for this RV fitting challenge is described in detailed in Dumusque (2016). As said in this paper, most of the planets injected in the data have very low eccentricities, which is common is observed multi-planetary systems. Those RVs correspond to typical quiet solar-like stars targeted by high-precision RV surveys. Therefore, the conclusions of this paper are relevant for most high-precision RV surveys.

In Sections 2 and 3, we describe the methods used by the different teams to recover planetary signals despite stellar signals; Section 2 focuses on methods relying on a Bayesian framework, while Section 3 on other methods. For those sections, the number assigned to each team does not have any particular meaning. In Section 4, we discuss the results of the different teams and compare the efficiency of their method to recover low-mass planetary signals despite stellar signals. We conclude in Section 5.

Table 1. Techniques to deal with stellar signals used by the different teams, as well as planetary and stellar parameters reported. P$_{rot}$ corresponds to the stellar rotation period, P, K T_0, ecc and ω to the planetary period, semi-amplitude, transit time, eccentricity and argument of periastron, respectively.

	Team	Techniques	P_{rot}	P	K	T_0	ecc	ω
1	Torino	Bayesian framework with Gaussian process to account for red noise	Yes	Yes	Yes	Yes	Sometimes	Sometimes
2	Oxford	Bayesian framework with Gaussian process to account for red noise	No	Yes	Yes	Yes	Yes	Yes
3	M. Tuomi	Bayesian framework with Moving Average to account for red noise	Yes	Yes	Yes	No	No	No
4	P. Gregory	Bayesian framework with apodized Keplerians to account for red noise	No	Yes	Yes	Yes	Yes	Yes
5	Geneva	Bayesian framework with white noise	No	Yes	Yes	No	No	No
6	A. Hatzes	Pre-whitening	Yes	Yes	Yes	No	No	No
7	Brera	Filtering in frequency space	No	Yes	Yes	Yes	Yes	Yes
8	IMCCE	Compressed sensing and filtering in frequency space (preliminary results)	Yes	Yes	Yes	No	No	No

2. Methods to deal with stellar signals using a Bayesian Framework

In total eight different teams have analyzed the RV fitting challenge data set, using different approaches. This section is dedicated to the description of these different methods. The first five teams used a Bayesian framework and model comparison to find the most favorable solution for each system. Teams 1 through 4 used red-noise models to account for stellar signals, while team 5 used a white noise model. Team 6, 7 and 8 used *pre-whitening*,compressed sensing and/or filtering in the frequency domain. Table 1 summarizes the different techniques used to deal with stellar signals, and the different stellar and orbital parameters reported by each team.

2.1. Team 1: Torino team - Bayesian framework with Gaussian process regression to account for stellar signals

The Torino team is composed of, in order of contribution, M. Damasso, A. Sozzetti, R. D. Haywood, A.S. Bonomo, M. Pinamonti, and P. Giacobbe. Their activities are in the framework of the *Global Architecture of Planetary Systems* (GAPS) project (e.g., Poretti et al. 2015). This team analyzed the 14 valid systems of the RV fitting challenge[1]. For each system, team 1 reported the period, semi-amplitude, time of periastron passage and sometimes eccentricity and argument of periastron of the detected planets, as well as their best estimate of the stellar rotation period. For most of the planetary system, team 1 analyzed only the most significant signals with small p-values, as the team had not enough time and computational power to explore more complex solutions obtained by adding smaller significance signals. For the same reason, as a first approach team 1 favored circular over more complex eccentric orbits.

2.1.1. General framework

The Torino team used Gaussian processes (GP, Rasmussen 2006) to model, in a non-parametric way, stellar activity effects in RV data. GPs are a powerful tool for mitigating the contribution of stellar activity in RV measurements, especially when contemporaneous stellar activity indicators are available (e.g. light curves, spectroscopic activity indexes).

When simulating short-term activity signals for the RV fitting challenge data set, Dumusque (2016) only considered slow rotators, i.e., $v \sin i < 4\,\mathrm{km\,s^{-1}}$. In this case, plages are expected to be responsible for the majority of the short-term activity RV variation (see introduction, and for more details Haywood et al. 2016; Dumusque et al. 2014a; Meunier et al. 2010). The calcium activity index log(R$'_{HK}$), which is a measure of the emission in the core of the Ca II H&K spectral lines, was provided within the RV fitting challenge data set, and appeared to be the best proxy to trace out short-term activity contribution to the RVs.

The real strength of a GP regression approach is that it is non-parametric and does not assume any physical model about active regions or the physical processes at play. Here, the only assumption made by Team 1 is that short-term activity variations in RV and log(R$'_{HK}$) share the same covariance properties. This assumption is reasonable as the short-term activity signal in RV and log(R$'_{HK}$) is induced by stellar rotation and active region evolution.

Team 1 first trained a GP on the time series of the activity index log(R$'_{HK}$), and then injected the resultant covariance function into another GP that is part of a RV model. This GP absorbs correlated noise due to the stellar activity through a global fit (Keplerian signals + GP). This approach is inspired in particular by the works of Haywood et al. (2014) and Grunblatt et al. (2015). For the Kepler-78 system discussed in this last paper, the same global fit was able to recover a periodicity in the RV consistent with the stellar rotation period found via photometry.

For the sake of a homogeneous analysis, all the systems were processed generally using the same recipe, following a sequence of few steps, as described in the appendix (see Section B.1). In all cases, the team used only one covariance function and analyzed the full data sets, i.e., without rejecting outliers, without binning the data to get a single point per night, and without dividing the analysis into sub-sets. The analysis did not use the information provided by the bisector span (BIS SPAN, Queloz et al. 2001) of the CCF.

Because an optimal way of correctly identifying the number of planetary signals, and their orbital properties, based on the available RV data is through the evaluation of the Bayesian statistical evidence $\mathcal{Z}$ for each model, the Torino group carried out this task by testing a limited number of models and calculating $\mathcal{Z}$, which is notoriously complicated to assess using only one analytical procedure among those proposed in literature.

The details about the method used by team 1 to analyze the data of the RV fitting challenge can be found in the appendix of the paper (Section B.1). In the next subsection we illustrate the method using as example system 2.

[1] as explained in Dumusque (2016), system 6 was not considered due to a problem when generating the RV measurements.

2.1.2. Example for system 2

In Fig. 1, we show the results obtained by team 1 for system 2. The original RVs data show a very significant correlation with log(R$'_{HK}$) (Spearman's rank correlation coefficient ρ = 0.88), thus they were first corrected using a linear regression with log(R$'_{HK}$). The GLS periodogram of the RV residuals shows a peak at ~ 2727 days, that corresponds to a peak value observed in the GLS periodogram of the log(R$'_{HK}$) time series at exactly the same frequency, probably due to a long-term stellar activity cycle. We removed this signal from the RV residuals and log(R$'_{HK}$) time series by fitting a sinusoid with the same periodicity. Team 1 correctly recovered the rotation period of the star ($P_{\rm rot}$=25 days) through a GP analysis of the detrended log(R$'_{HK}$) time series. They recovered two out of the 5 planets injected in system 2. The GLS periodogram of the corrected RVs shows a series of significant frequencies with p-value<0.1%, the most prominent ones corresponding to the orbital period of the two recovered planets and to the first harmonic of the stellar rotation period. Despite the dominant frequency related to the stellar rotation appearing at the first harmonic $P_{\rm rot}/2$, the GP noise model converged towards the stellar rotation period. There is also a couple of peaks with p-value<0.1%, one of them clearly corresponding to the ~20 days period planet that Team 1 did not report. To prevent announcing false positives, Team 1 assumed this peak to be produced by differential rotation of the star, because of a period close to stellar rotation.

2.2. Team 2: Oxford team - Bayesian framework with GP regression to account for stellar signals

Team 2 is composed of, in order of contribution, V. Rajpaul and S. Aigrain. This team analyzed the first 5 systems of the RV fitting challenge. For each system, team 2 reported the period, semi-amplitude, time of periastron passage, eccentricity and argument of periastron of the detected planets. Team 2 did not report stellar rotation periods.

2.2.1. General framework

Team 2 analyzed the data of the RV fitting challenge using, like team 1, a GP regression to account for red noise induced by stellar signals. However their approach is slightly different and is described in details in Rajpaul et al. (2015). Rather than using only the calcium activity index log(R$'_{HK}$) as a proxy for activity and first training the GP on the activity indicator and then using the best estimate of the GP hyper-parameter to fit the RVs, team 2 modeled all time series (log(R$'_{HK}$), FWHM, BIS SPAN, RV) simultaneously. Team 2 treated each time series as a linear combination of a single unobserved GP and its derivative, adding a polynomial function to fit long-term trends. In addition, and only for the RV data, team 2 added one or several Keplerians. This approach is statistically more robust as it does not require an iterative fitting process, however, it is more computationally demanding.

At the outset, team 2 wanted to marginalize fully over all the hyper-parameters and parameters of their model, using an MCMC sampler. However, because this increased computational times by orders of magnitude, team 2 found it was not able to produce reasonable results within the time allocated for the RV fitting challenge. Team 2 therefore found the maximum *a posteriori* (MAP) values for all model hyper-parameters and parameters. Then, with the GP covariance hyper-parameters fixed at their MAP values (type-II maximum likelihood approximation), team 2 used a nested sampling algorithm to explore the parameter space of interest, i.e., the planet parameters. Note that team 2 used uniform priors for all parameters, except for eccentricity for which a log-uniform prior was used. As a result of the nested sampler, team 2 obtained posterior distributions for all parameters, as well as model evidences (log $\mathcal{Z}$). Team 2 compared models with up to N = 5 planets and added planets until the evidence for a model with an additional planet was smaller than the model without this extra complexity, i.e., log $\mathcal{Z} \mid (N+1) <$ log $\mathcal{Z} \mid (N)$.

Team 2 notes that given the type-II maximum likelihood approximation, computed evidence values are perhaps not reliable. Team 2 also note that the code used to pre-process the data before GP regression included a polynomial-subtraction component that removed long-term trends from the time series. In retrospect, this ended up removing all planetary signals with periods longer than a couple of months – a simple though costly error.

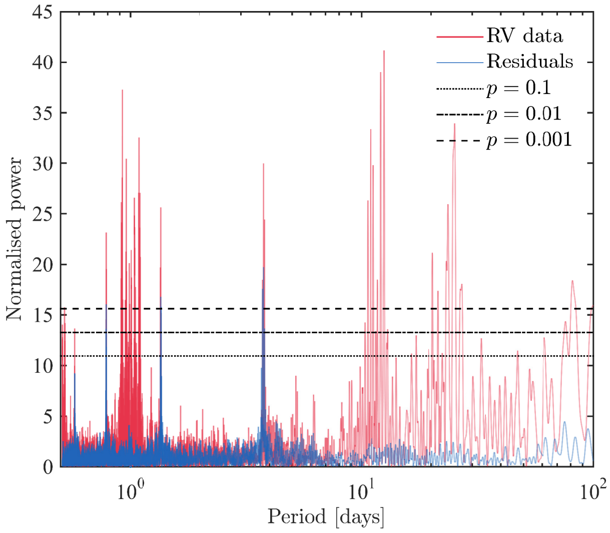


Fig. 4. Periodograms of the raw RVs (red) and of the residual RVs (blue) after removing the best GP plus 2-planet model to account for stellar signals and the 3.77 and 10.64-day planets present in the time series. The horizontal lines show from top to bottom the 0.1, 1 and 10% p-values used as a first guess to estimate signal significance. The non-white residuals suggest an imperfect model.

2.2.2. Example for system 2

In Figs. 2 and 3, we show the best-fit obtained by team 2 on system 2 of the RV fitting challenge. In Fig. 4 we compare the periodogram of the raw RVs with the residual RVs after removing the same best-fit. We see that the GP fitted simultaneously to the RV, log(R$'_{HK}$) and BIS SPAN allows to strongly mitigate the variations induced by stellar signals. For more examples on the technique used by team 2, readers are referred to Rajpaul et al. (2015).

2.3. Team 3: M. Tuomi and G. Anglada-Escudé - Bayesian framework with first order Moving Average to account for stellar signals

Team 3 is composed of, in order of contribution, M. Tuomi, G. Anglada-Escudé and H. R. A. Jones. This team analyzed the 14 systems of the RV fitting challenge. For each system, team 3 reported the period and semi-amplitude of the detected planets, as well as their best estimate of the stellar rotation period.

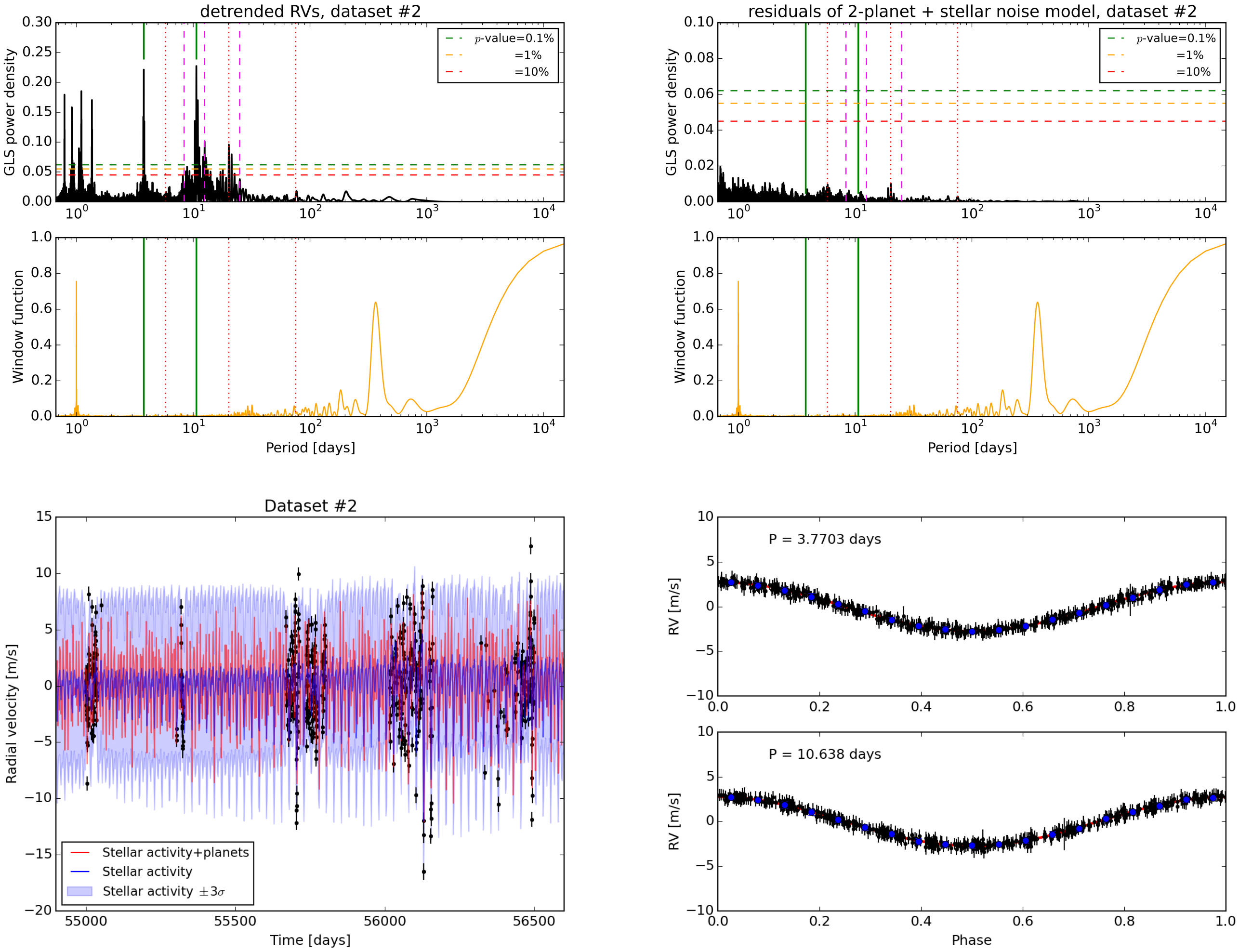


Fig. 1. *Top left panel:* GLS periodogram of the challenge RV dataset number 2 analysed by the team 1, and the corresponding window function. The original RVs data were first corrected through a linear regression with log(R'_{HK})and then through a sinusoidal fit to remove a 2727-day periodicity likely related to a long-term stellar activity cycle. Vertical lines indicate *i)* the orbital frequencies of the planetary candidates identified by the team (solid green lines), which turned out to be real; *ii)* the missed planets (dotted red lines); *iii)* the simulated stellar rotation period and its first two harmonics (dashed magenta lines). *Top right panel:* GLS periodogram of the RV residuals after the best-fit global model including the two planets detected was removed from the detrended dataset. A bootstrap analysis performed on these RV residuals shows no evidence of significant signals left. *Bottom left panel:* RVs detrended with the best-fit global model superposed (solid red line) and the mean contribution due to the stellar activity predicted by the GP (blue solid line). The shaded area spans the 3σ region centered around the best-fit noise model. *Bottom right panel:* RV curves folded at the periods of the two candidate Keplerian solutions found by the team, with the superposed red continuous line representing the best-fit orbital model. Data in each plot refers to a single planet, and are obtained from the original RVs by subtracting the offset, the GP stellar activity noise model, and the Keplerian of the other planet.

2.3.1. General framework

Team 3 also analyzed the data of the RV fitting challenge using a correlated noise model to account for stellar activity. However, in their case, team 3 used a first order moving average component with exponential smoothing. Team 1 used a GP and trained it on the log(R'_{HK}) data to obtain the best hyper-parameters that are then used to fit the RVs. This implies therefore an iterative fitting process, which can be dangerous. Team 2 overcomed this problem by fitting simultaneously all the time series with a GP. However a strong assumption is made during the process: the covariance of short-term activity should be the same in the RVs than in the activity observables (log(R'_{HK}), BIS SPAN and FWHM). Using a first order moving average, team 3 avoided this assumption, which could imply significantly different results if this assumption turns out not to be valid. In addition, team 3 also considered in their RV model linear correlations with the different activity observables, therefore fitting everything at once implying a robust statistical approach.

The details about the method used by team 3 to analyze the data of the RV fitting challenge can be found in the appendix of the paper (Section B.2). In the next subsection we illustrate the method using as example system 2.

2.3.2. Example for system 2

Here we present the results of the analysis of system 2 performed by team 3.

Analysis of the activity indicators of system 2

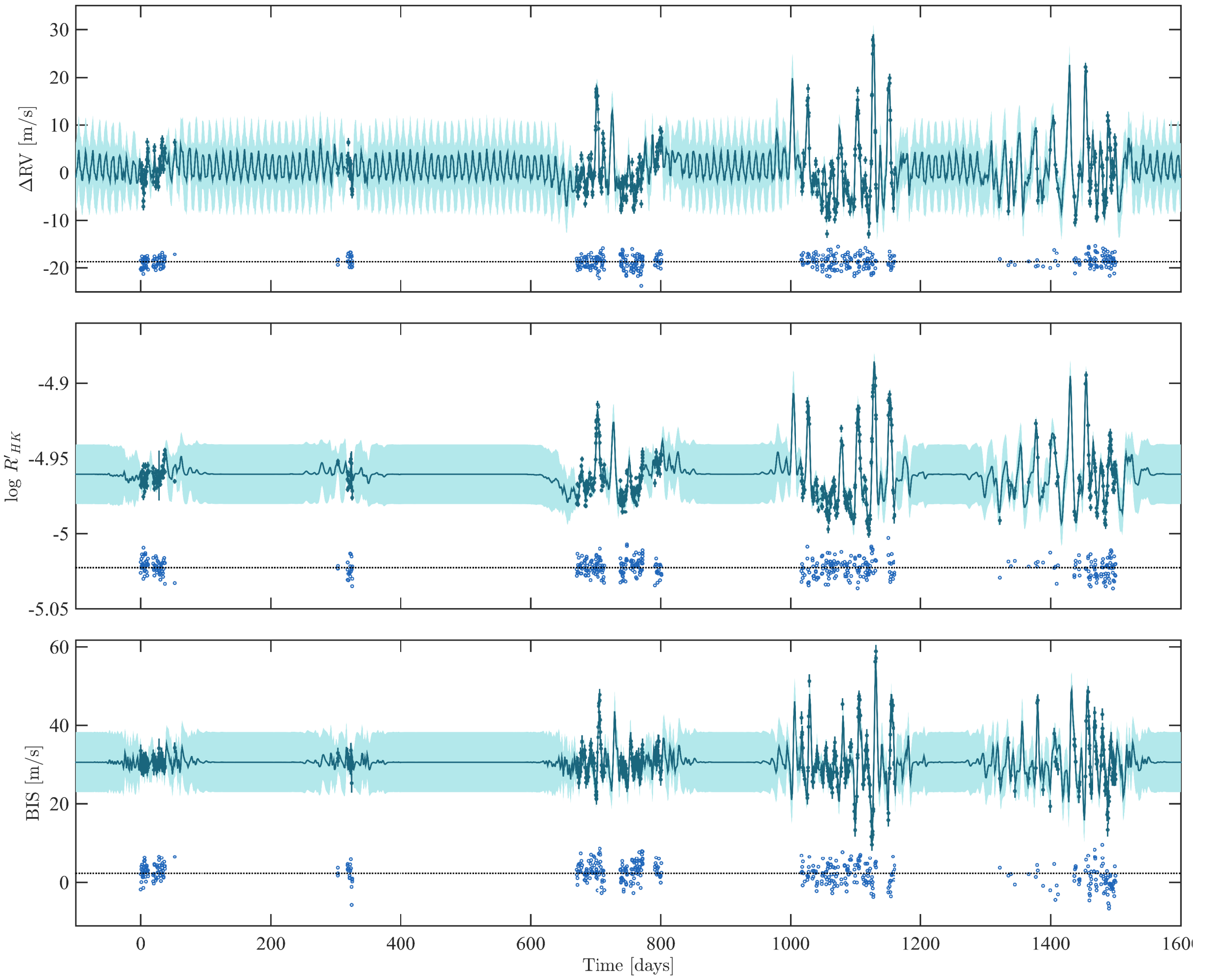


Fig. 2. GP model MAP fit to system 2 for which team 2 could recover accurately two of the four injected planetary signals. As in other datasets, the crude pre-processing pipeline used by team 2 wrongly removed signals longer than a couple of months, therefore removing the signal of the 75-day planet present in the data. All plotted time series were fitted simultaneously, using a single set of GP hyperparameters. The green dots indicate the raw time series; the solid lines are model posterior means, and the shaded regions denote $\pm\sigma$ posterior uncertainty. The blue dots indicate model residuals. A zoom on the fourth epoch of observation can be seen in Fig. 3.

The team first analyzed the activity indicators by calculating the likelihood-ratio periodograms (see Fig. 5). This analysis indicates a strong signal at a period of 12.5 days in the BIS SPAN time series. The time-series for both FWHM and log(**R**$'_{HK}$) **activity indices show a very strong signal at 24.9 days, twice the period found in the BIS SPAN value, suggesting that this is the rotation period. The fact that BIS SPAN shows a signal at half the rotation is the expected period for a spots showing only one half of the rotation (Dumusque et al. 2014a). Signals in the activity indices were also searched when adding a first order moving average component to the model but they didn't change the periods found much in this case. Fig. 5 shows that these signals (12.4 and 24.9 days) are detected well below the usual 1% (even 0.1%) p-value thresholds (shown as horizontal lines).**

Analysis of the RVs of system 2

Team 3 analyzed the RVs with a statistical model including a linear trend overtime, linear correlations with the three activity indicators given, and correlated noise according to a first order moving average model. The likelihood-ratio periodogram and signal searches without including correlation terms would lead to a very different answer from the correct one. A model without including correlations would show a rather strong signal at the same period as the BIS SPAN and subsequent inclusion of Keplerians requires fitting several sinusoids (all spuriously generated by rotation and activity) before finally spotting the first real planet candidate. The difference between the raw RVs and the RVs corrected from activity signal using linear correlations is highlighted in Fig. 6. A model including the linear correlation terms directly spots three unambiguous Keplerian signals. Fig. 7 shows the likelihood periodograms with and without correlation terms for the first signal search, and the subsequent likelihood periodograms obtained when adjusting the signals under investigation together with all the model free parameters. Likelihood periodograms are used to obtain a quick look at the solution landscape when one new planet is added, but the actual search and verification is then done using tempered delayed rejection and adaptive Metropolis (DRAM) samplings (Haario et al. 2001; Haario 2006) of the posterior density. This DRAM samplings allow to explore all the new periods while allowing re-adjusting all the previously included Keplerian signals. The posterior contours resemble the likeli-

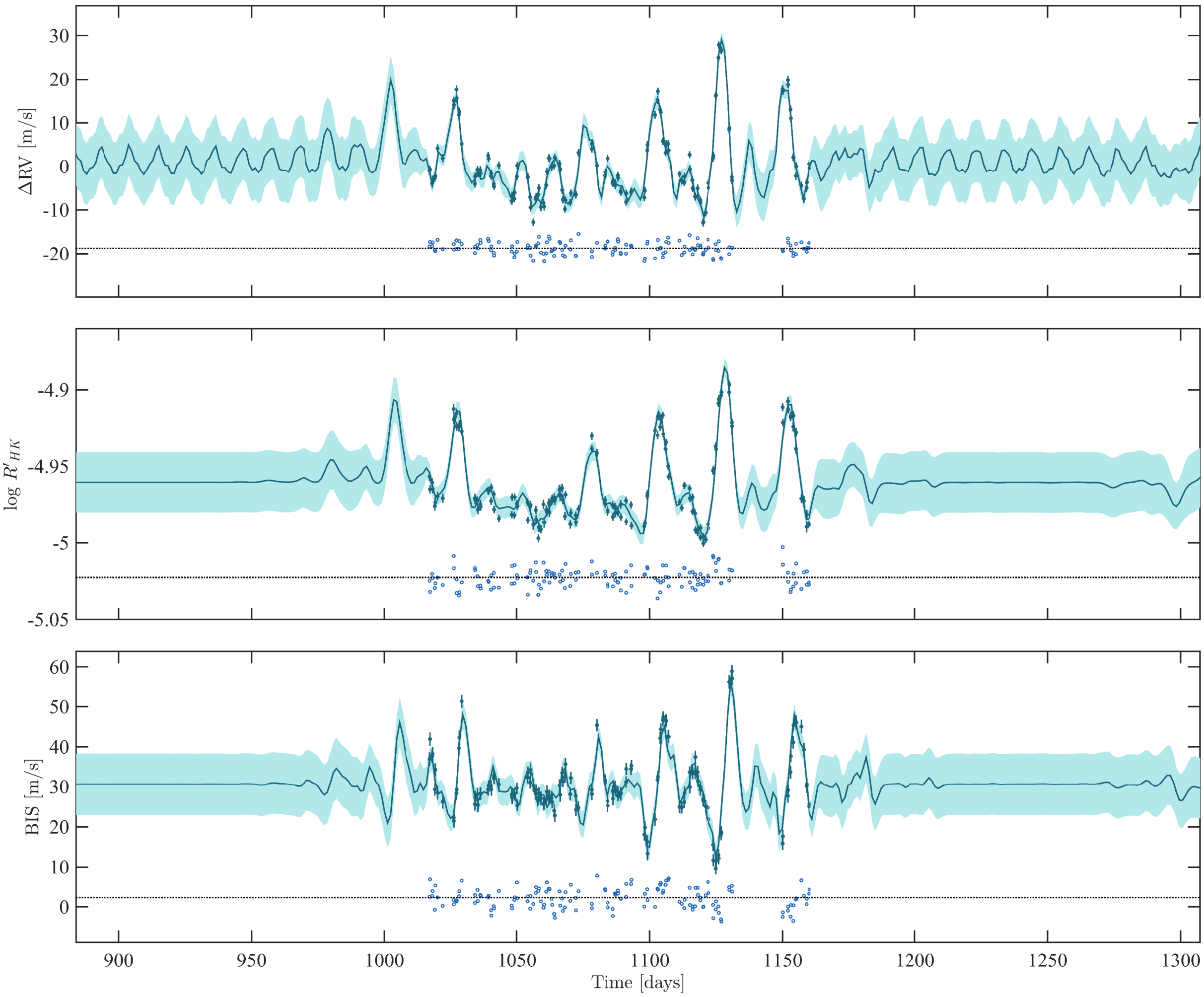


Fig. 3. Same legend as Fig. 2, just zoomed in on the fourth epoch of observation

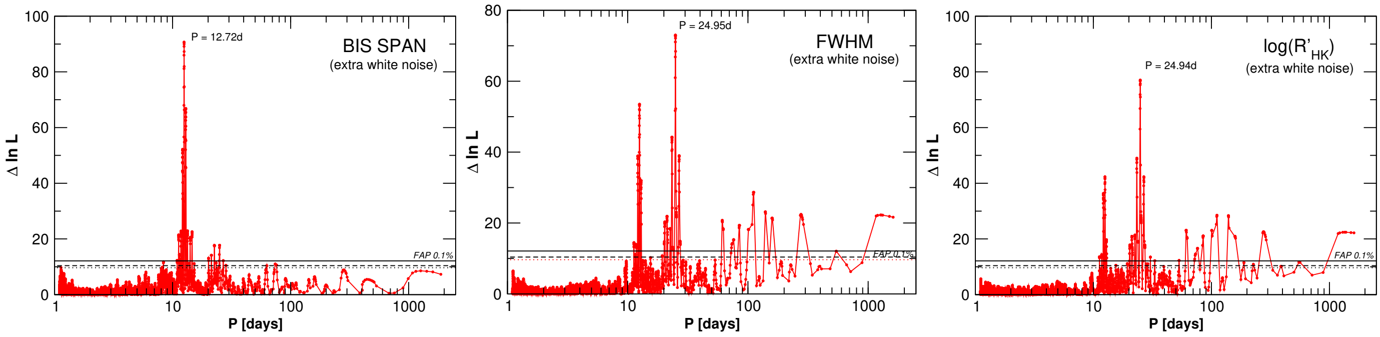


Fig. 5. Likelihood periodogram searches of the strongest periodic signal in the time series of the three activity indices provided (BIS SPAN, FWHM and log(R'_{HK})). Both the rotation period (most prominent in FWHM and log(R'_{HK}) index) and its first harmonic (most prominent feature in BIS SPAN) are clearly seen in the activity time-series. In the three periodograms, the tested model contains sinusoid (ciruclar orbit), an offset, a linear trend and a extra white-noise jitter component.

hood periodograms shown in Fig. 7. A more detailed description of the methodology used by team 3 is given in the appendices of the paper.

For a signal to be tagged as a planet candidate several conditions must be met:

- the period of the signal has to be well-constrained from above and below,
- the amplitude of the signal has to be statistically significantly different from zero (the zero value must be excluded from the 99% credibility interval),
- all other local maxima (peaks in posterior or likelihood periodogram) must be 100 times smaller than the preferred solution (uniqueness condition), and
- a model with this extra signal is statistically more significant than a model without it when computing a Bayes

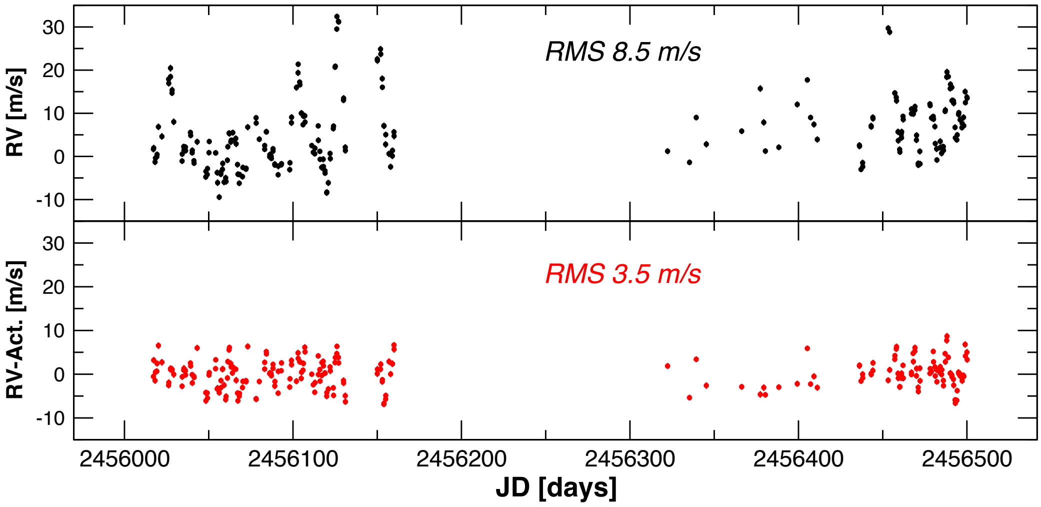


Fig. 6. Original Doppler measurements (top, black) and residual time-series (bottom, red) after adjusting a model with no periodic signal but with all activity correlation terms included. Even without subtracting any Keplerian signal, the reduction in the scatter in the RV time-series is apparent to the eye.

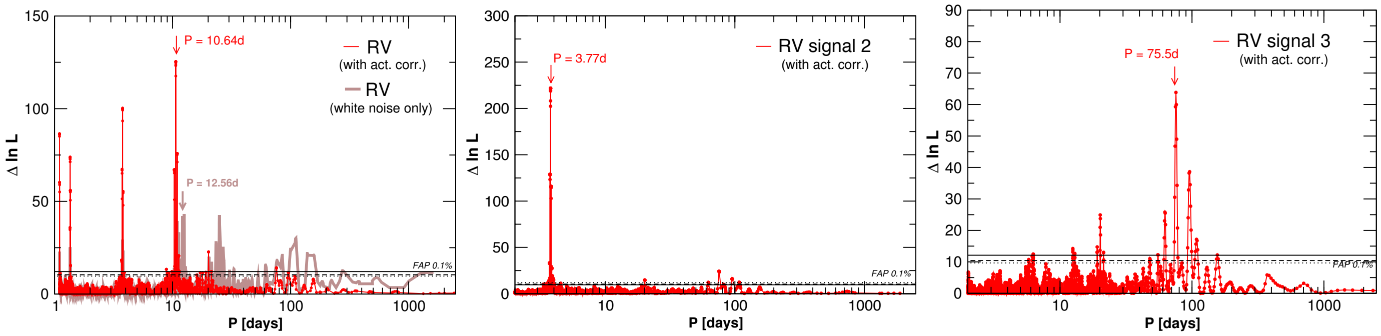


Fig. 7. Likelihood periodogram searches for the first, second and third signals in the time-series when adjusting the full noise model (first order moving average term plus linear correlations) at the same time as the Keplerian signals. In the first panel, we show the likelihood periodogram of the original RV time-series with and without accounting for correlations. In that case we would incorrectly conclude that the rotation period is a planet candidate. The correlations also reduce the impact of unstructured noise (eg. caused by high-frequency, sub-day jitter) which is made obvious by the much higher significance of the first planet candidate against the detection of the rotation period in the model without activity correlations.

factor using the mixture of posterior and prior densities (Newton & Raftery 1994).

As a threshold, it is often said that the more complex model must have a Bayes factor 150 times higher than the simpler one (Kass 1995), but team 3 uses a threshold of 10^4 in Doppler time series to acknowledge that the space of model is likely incomplete (eg. sinusoids fitting instrumental and activity features can still improve the model without implying the presence of extra Keplerian signals). This threshold was considered sufficient and adopted after examination and combination of data from the UVES and HARPS spectrographs, that is, signals producing improvements on the model in the UVES data below 10^4 were often not confirmed when combining the measurements with available HARPS data (Tuomi et al. 2014).

2.4. *Team 4: P. Gregory - Bayesian framework with apodized Keplerians to account for red noise*

Team 4 is composed of P. Gregory. He analyzed the first 5 systems of the RV fitting challenge. For each system, he reported the period, semi-amplitude, time of periastron passage, eccentricity and argument of periastron of the detected planets. He did not report stellar rotation periods.

2.4.1. General framework

P. Gregory analyzed the RV fitting challenge data set with a novel approach using apodized Keplerians. Stellar short-term activity creates semi-periodic signals due to stellar rotation and active region evolution, unlike the periodic signal induced by planets. He therefore decided to fit every significant signal in the RVs using Keplerians that could change their semi-amplitude as a function of time using a Gaussian apodization function. The two parameters that characterize each Gaussian apodization function, are fitted as free parameters. If the timescale appears to be much shorter than the time span of the RV measurements, it implies a non-stationary signal as a function of time, and therefore this signal is flagged as being induced by stellar activity. In addition, as team 3, P. Gregory also includes in its RV model a correlation with log(R'_{HK}) to account for the RV effect of magnetic cycles in a statistically robust approach.

A detailed step-by-step approach can be found in Gregory (2016). We however give a small summary of the approach in the appendix of the paper (Section B.3). In the following subsection, we give an example of how the method works for system 2.

2.4.2. Example for system 2

Fig. 8 shows an example of the Bayesian Fusion MCMC results for system 2, for which apodized Keplerians were used to characterize both planetary and stellar activity signals.

Fig. 9 shows a comparison of the GLS periodograms of RV and FWHM data, each corrected by the removal of the best-fit log(R'_{HK}) correlation model. There are three traces in each of the 6 panels. The black trace is the modified RV periodogram. The blue trace is the negative of the modified FWHM periodogram and the red trace shows the difference, i.e., the black trace plus the blue traces. One can clearly see blue trace counterparts to the black trace around periods of 12.5 and 20.2 days, indicating they are likely stellar activity signals. In contrast there is no blue trace counterpart to the peaks near 3.77 and 10.6 days.

2.5. Team 5: Geneva team - Bayesian framework with white noise

Team 5 is composed of, in order of contribution, R. Díaz, D. Ségransan and S. Udry. Team 5 analyzed the two first systems of the RV fitting challenge. For each system, team 5 reported the period, semi-amplitude and eccentricity of the detected planets. Team 5 did not report stellar rotation periods.

2.5.1. General framework

Team 5 considered models including several Keplerians to represent planetary companions and velocity variations related to stellar rotation period. To account for the correlation between RV and log(R'_{HK}) induced by magnetic activity cycles, team 5 fitted log(R'_{HK}) using a third order polynomial. The posterior distributions of this fit are then used as priors for a third order polynomial added to the model used to fit the RVs. In other words, the low-frequency structure of the log(R'_{HK}) is included in the model of the RVs. A source of additional white noise whose amplitude was set to scale linearly with activity (measured by the log(R'_{HK}) index) was added to the model to account for the known correlation between stellar activity level and velocity jitter. The model is fully described in Díaz et al. (2016).

After fitting several models with different numbers of Keplerians, team 5 used two estimation of the Bayesian evidence log $\mathcal{Z}$ to compare between models: the Chib & Jeliazkov (2001) and the Perrakis (2014) estimators. The best model is the one that exhibits a reasonable instrumental white noise component and not too-low evidence using the two estimators.

3. Methods to deal with stellar signals without using a Bayesian Framework

3.1. Team 6: A. Hatzes - Pre-whitening

Team 6 is composed of A. Hatzes. He analyzed the 14 systems of the RV fitting challenge. For each system, he reported the period and semi-amplitude of the detected planets, as well as the best estimate of the stellar rotation period.

3.1.1. General framework

A. Hatzes used the so-called *pre-whitening* procedure. One first computes the Discrete Fourier Transform (DFT) to find the dominant peak in the Fourier amplitude spectrum. A least squares sine fit to the data is made using this frequency and the resulting sine fit is subtracted from the data. One then performs a DFT on the residual data to find the next dominant peak. In finding a subsequent signal in the data, a simultaneous fit is made using all the previously found sine functions. The process stops when the final peak amplitude is less than about four times the mean amplitude of the surrounding noise peaks. Kuschnig et al. (1997) established that this corresponds to a *p*-value of about 1%. Examples of this process performed on RV data can be found for CoRoT-7 (Hatzes et al. 2010) and GL 581 (Hatzes 2013). The DFT analysis was only performed out to the nominal Nyquist frequency of 0.5 d^{-1}. This means that periods shorter than 2 days were not actively searched for even if they were in the data.

Given the large number of time series, the program Period04 (Lenz & Breger 2005) was used to perform *pre-whitening*. This program provides a convenient environment for computing DFTs, selecting peaks in the amplitude spectrum, fitting those, and searching for additional signals in the residual data. The program also provides an option for computing the signal-to-noise ratio (amplitude of a peak divided by the computed mean noise level).

The *pre-whitening* procedure was performed on all time series. Significant peaks found in the RV data were compared to those found in the activity indicators (log(R'_{HK}), BIS SPAN, and FWHM). If a significant peak found in the RV did not have a corresponding peak in the activity indicators it was identified as a planet. Signals found in the RVs and in the activity indicators were attributed to activity, with the dominant peak chosen as the rotation period.

Note that in this case, only fitting sine waves is dangerous, because not removing the correct solution for a planet or stellar signals can then perturb the residuals and lead to the detection of false positives.

3.2. Team 7: Brera team - Filtering in frequency space

Team 7 is composed of, in order of contribution, F. Borsa, G. Frustagli, E. Poretti and M. Rainer, from INAF - Brera Astronomical Observatory. Their activities are in the framework of the *Global Architecture of Planetary Systems* (GAPS) project (e.g. **?**). Team 7 analyzed the 14 systems of the RV fitting challenge. For each system, team 7 reported the period, semi-amplitude, time of periastron passage, eccentricity and argument of periastron of the detected planets. Team 7 did not report stellar rotation periods.

3.2.1. General framework

Team 7 decided to try an approach that is not model dependant for stellar signals. This approach is based on filtering signals in the frequency domain of the RVs, using the frequency information found in the different activity indicators. The details about the method used by team 7 to analyze the data of the RV fitting challenge can be found in the appendix of the paper (Section B.4). In the next subsection we illustrate the method using as example system 2.

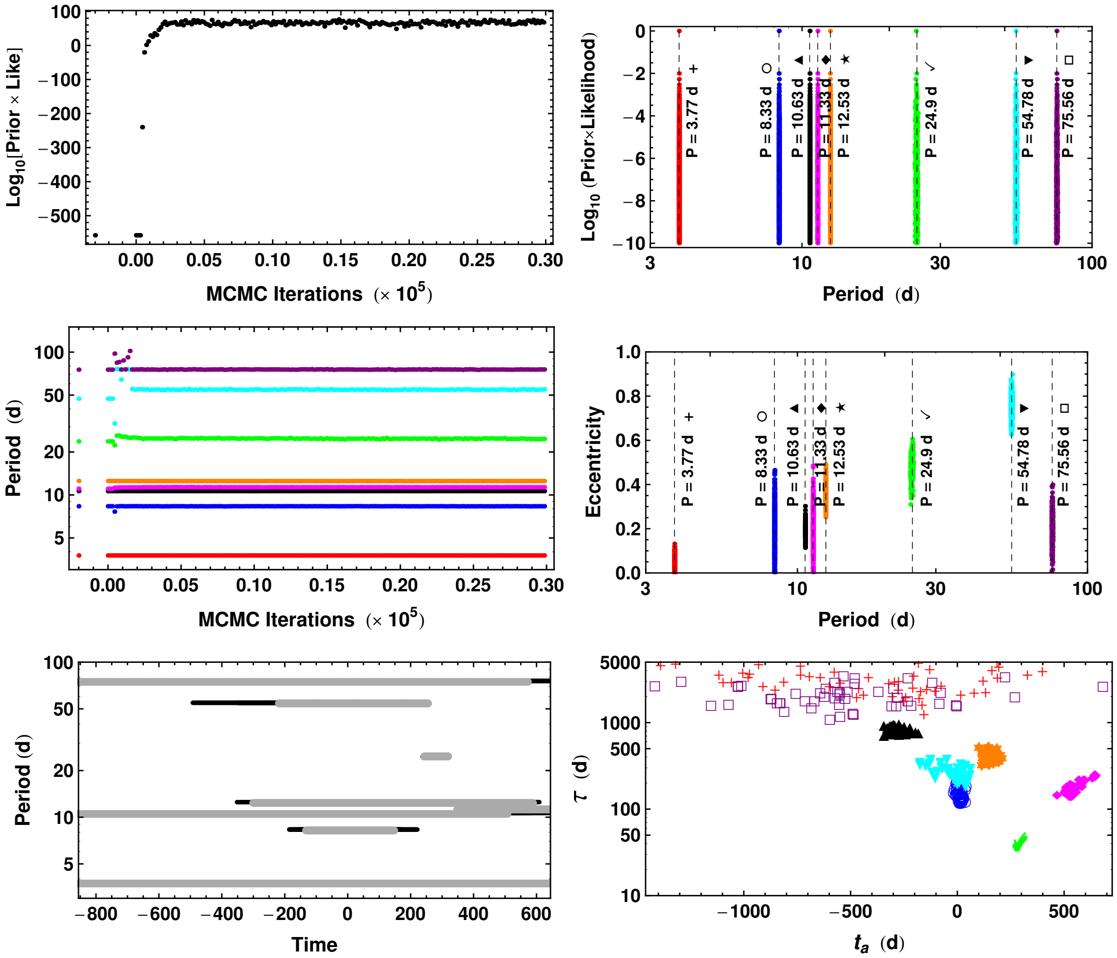


Fig. 8. *Upper left:* Log_{10}[Prior × Likelihood] versus iterations for the 8 signal apodized Keplerian model used to fit system 2. *Upper right:* Log_{10}[Prior × Likelihood] versus period showing the 8 periods detected. *Middle left:* Values of the 8 unknown period parameters versus iteration number. *Middle right:* Eccentricity parameters versus period parameters. *Lower left:* Apodization window for each signal (gray trace for MAP values of the apodization time constant τ and the apodization window center time t_a, black for a representative set of samples which is mainly hidden below the gray). *Lower right:* Apodization time constant versus apodization window center time for each signal (Credit: Gregory 2016).

3.2.2. Example for system 2

As an example, Fig. 10 shows the different steps used by team 7 to detect the 3.77-day planetary signal present in System 2.

3.3. Team 8: IMCCE team - Compressed sensing and frequency filtering

Team 8 is composed of, in order of contribution, N. Hara, F. Dauvergne and G. Boué, from the Institut de Mécanique Céleste et de Calcul des Éphémérides in Paris (IMCCE). Team 8 analyzed the 14 data sets of the RV fitting challenge. For each system, team 8 reported the period and semi-amplitude of the detected planets, as well as the best estimate of the stellar rotation period.

3.3.1. General Framework

The IMCCE team used an approach based on Compressed Sensing (or Compressive Sampling, see Donoho 2006; Candès et al. 2006) and frequency filtering. This method was devised to avoid fitting the planets one by one. Indeed, after removing a certain number of signals, the tallest peak of the periodogram of the residuals might not correspond to a real planet. One might even face the case where the maximum of the periodogram of the raw data is significant but spurious. The usual way to circumvent this issue is to fit a complete model accounting for several planets and sometimes noise parameters. In that case, one uses MCMC methods or genetic algorithms to explore the whole parameter space and avoid being trapped in a suboptimal local minimum. The compressed sensing framework allows in a certain sense to search all the planets at once while considering an objective function which has only one minimum. As the minimization problem is convex, one can design fast algorithms.

The key is to use an *a priori* information: the signal is supposed to be "simple" in a certain sense. Here, it means that there exists a set of vectors, termed the dictionary, such that a linear combination of a few of its entries reproduces the signal. For instance, dictionaries made of wavelets are appropriate to represent most images. This feature is exploited for the JPEG2000 format (Taubman & Marcellin 2002). The image is stored via its significant wavelet coefficients, which are a few compared to the total number of pixels. The MP3 and AAC audio formats rely on the same principles.

In our case, the movement of a star due to its planets is quasi-periodic. In other words, it is a linear combination of a few sine functions $\exp^{-i\omega t}$ and $\exp^{i\omega t}$. However, we do not measure the

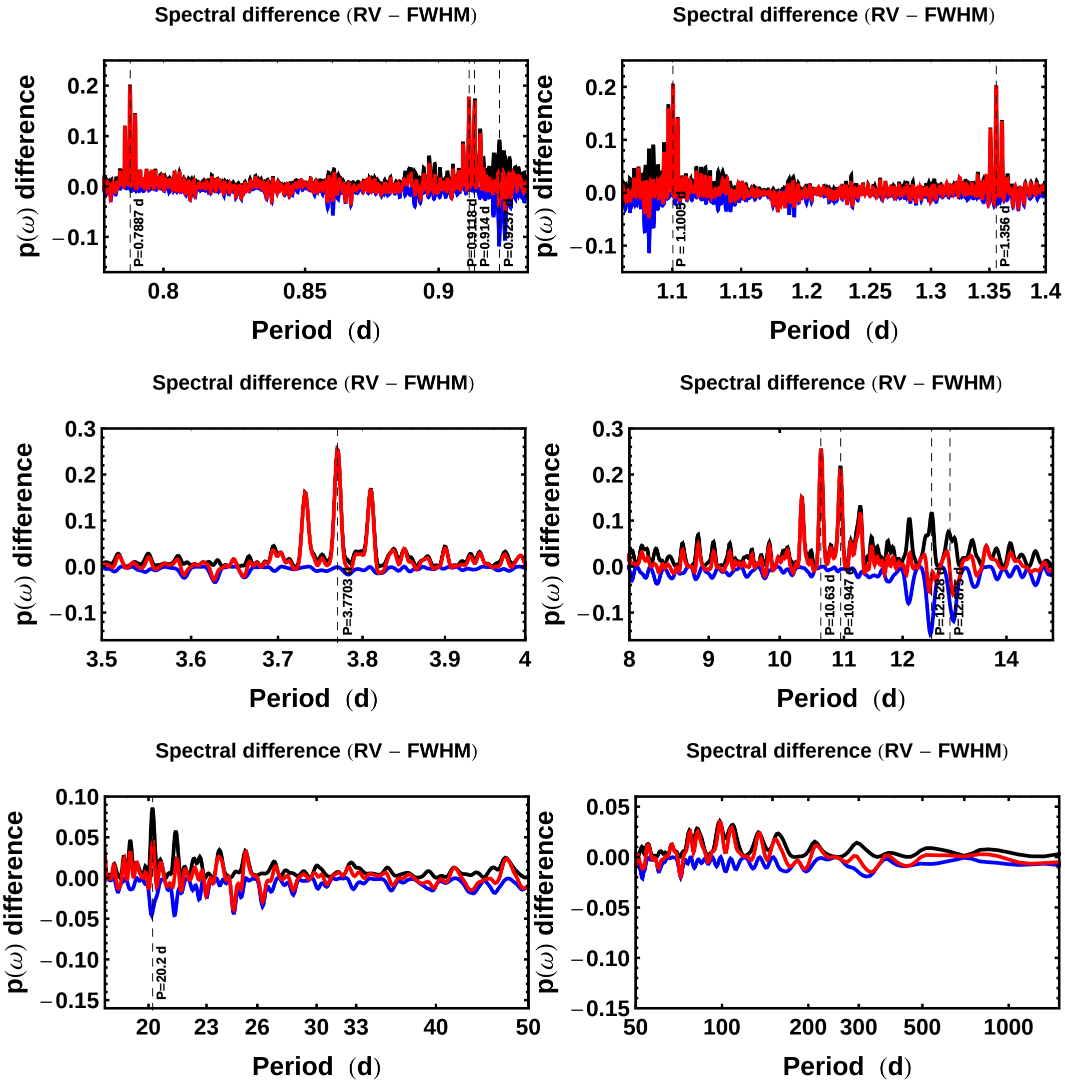


Fig. 9. Comparison between GLS periodograms of RV and FWHM data for system 2, each modified by the removal of the best-fit $\log(R'_{HK})$ correlation model (hereafter called modified RV and FWHM). There are three traces in each of the 6 panels. The black trace is the modified RV periodogram, the blue trace is the negative of the modified FWHM periodogram, and the red trace is the difference, i.e., the black trace plus the blue traces. Each plot show an interesting portion of the periodogram. As explained in the text, interesting signals that can be seen in both the modified RV (black) and the modified FWHM (blue) are not fitted as they probably are the result of stellar activity, while signals only present in the modified RVs are considered (Credit: Gregory 2016).

motion of the star only. The signal is also contaminated - and in the present challenge, dominated - by the stellar activity. To account for this effect, in addition to the sine functions, frequency-filtered FWHM, bisector span and $\log R'_{HK}$ were incorporated into the dictionary.

Once the dictionary is defined, one searches for a combination of a few of its element which is close to the observations. The output of that procedure are the coefficients of the linear combination of dictionary elements reproducing the data within a certain tolerance. This vector is plotted versus the frequency (or the period), just like a GLS periodogram. To avoid the confusion with this one, the figure obtained is termed ℓ_1-periodogram (see Fig. 11).

The procedure applied to the RV fitting Challenge data was at an intermediate stage of development and is outlined in subsection 3.3.2 on an example. This preliminary method was not very robust, and was greatly improved afterwards. For a precise description of the most recent version, see (Hara et al. 2016), where system 2 of the RV Fitting Challenge is treated in detail along with real radial velocity signals.

3.3.2. Example for system 2

In this section the results of the method applied to the second system of the RV Fitting Challenge are presented. The system contains five planets, whose periods and true amplitudes are represented in red in Fig. 11. The blue curve on the top plot of Fig. 11 is the GLS periodogram of the raw RV data (Zechmeister & Kürster 2009), displayed for comparison.

The dictionary is made of sine functions $\exp^{-i\omega_k t}$ and $\exp^{i\omega_k t}$ for $n = 3.10^5$ frequencies, $\omega_k = 3k\pi/n$ radian per day, $k = 0..n-1$. To obtain a representation of the activity, each of the FWHM, bisector span and $\log(R'_{HK})$ signals are bandpass filtered by projection onto five families of orthonormal polynomials. The family j, $j = 1..5$ is made of $D_j - d_j + 1$ polynomials of degrees

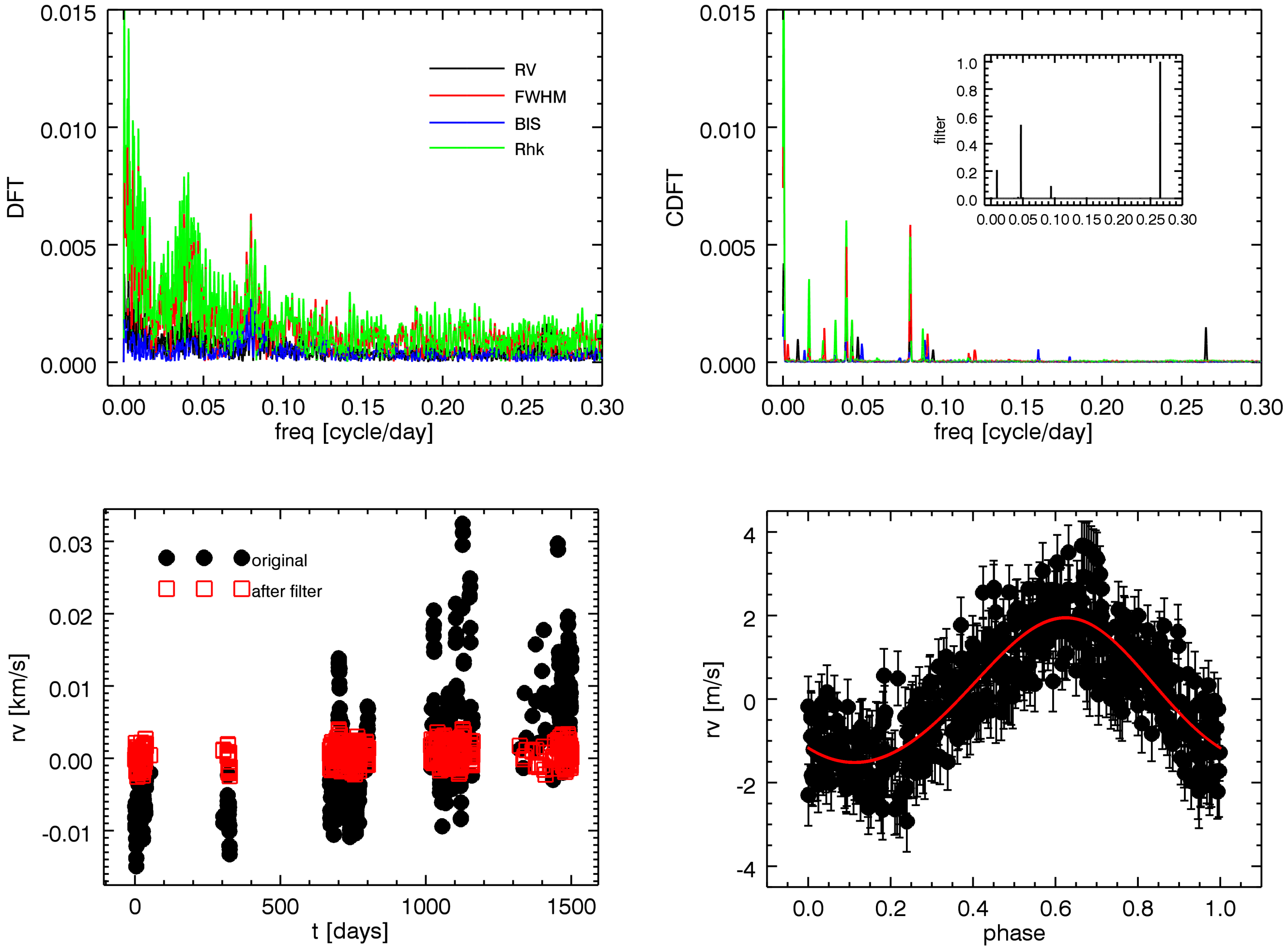


Fig. 10. Different steps used by team 7 to detect the 3.77-day planetary signal present in System 2. *Upper left panel:* DFT of the RVs, FWHM, BIS SPAN and log(R'_{HK}). *Upper right panel:* Cleaned DFT (CDFT) obtained after applying the CLEAN algorithm to the DFTs of all observables. In the inset, the resulting pass-planet filter in the frequency domain, that is later applied to the DFT of the RVs to mitigate the effect of stellar signals. *Lower left panel:* The RVs before (black circles) and after (red squares) applying the pass-planet filter. *Lower right panel:* The phase-folded result of the Keplerian fit to the 3.77-day planetary signal found in the RVs.

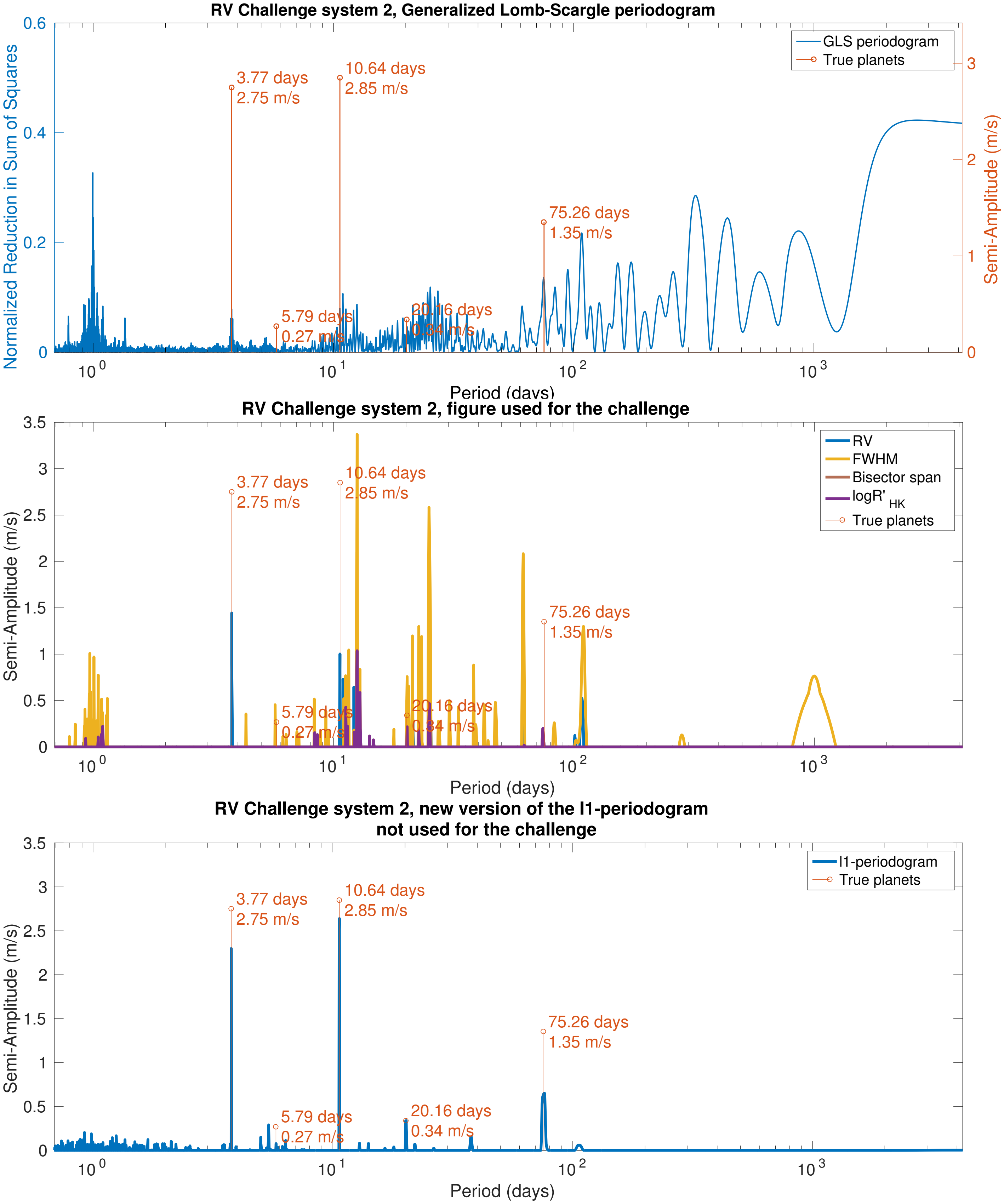
d_j to D_j. Here $D_j = d_{j+1} - 1$, and $d_1 = 0$, $d_2 = 15$, $d_3 = 60$, $d_4 = 160$, $d_5 = 300$.

The ℓ_1-periodograms - in a version used for the RV Fitting Challenge - of the RV, FWHM and bisector span are represented in the middle plot of Fig.11. We then selected planetary signals following this principle: If a "high" peak of the ℓ_1-periodogram of the radial velocity data is sufficiently far from peaks of the ℓ_1-periodogram of the FWHM and bisector span, it is retained. If this peak is too close to at least one of the peaks of the FWHM or bisector span ℓ_1-periodogram, it is discarded. Here, the three smallest peaks do not appear. The 10.64 days periodicity does show up, but was discarded due to its proximity to features of the other signals. Finally, we see clearly the 3.77 days periodicity, which was indeed selected.

As said above, the IMCCE team kept on working on the method. If one subtracts the estimated activity of the star before performing the ℓ_1 minimization and with further improvements, one obtains the bottom plot in Fig. 11. In this case, the four strongest signals appear without ambiguity. There are also two signals close to 5.4 and 37 days, which are signatures of the first harmonics of the eccentric orbits at 10.64 and 75.26 days. We however could not see clearly the 5.79 days periodicity, which seems to be buried in the noise. A further study shows that the five planets plus the harmonic of the 10.64 days orbit are statistically significant in some sense. This system is treated in detail in (Hara et al. 2016).

4. Results

In this section, we analyze the results of the different teams, in term of stellar rotation periods found, planetary signals detected and false positives announced. We also discuss the accuracy of orbital parameters recovered, as well as the realism of the simulated systems generated for the purpose of the RV fitting challenge. Because a wrong estimate of the stellar rotation period can lead to the detection of false positives, this is the first point we discuss.



RV Challenge system 2, Generalized Lomb-Scargle periodogram

RV Challenge system 2, figure used for the challenge

RV Challenge system 2, new version of the l1-periodogram not used for the challenge

Fig. 11. *Top:* GLS of the RV fitting challenge system 2 (raw time series). The red horizontal lines correspond to the true planetary signals injected into the data. *Middle:* Figure used for the challenge, blue: ℓ_1-periodogram of the RVs, yellow: ℓ_1-periodogram of FWHM, purple: ℓ_1-periodogram of log(R$'_{HK}$). Only the 3.77-day signal was detected by team 8. *Bottom:* New version of the ℓ_1-periodogram (Hara et al. 2016).

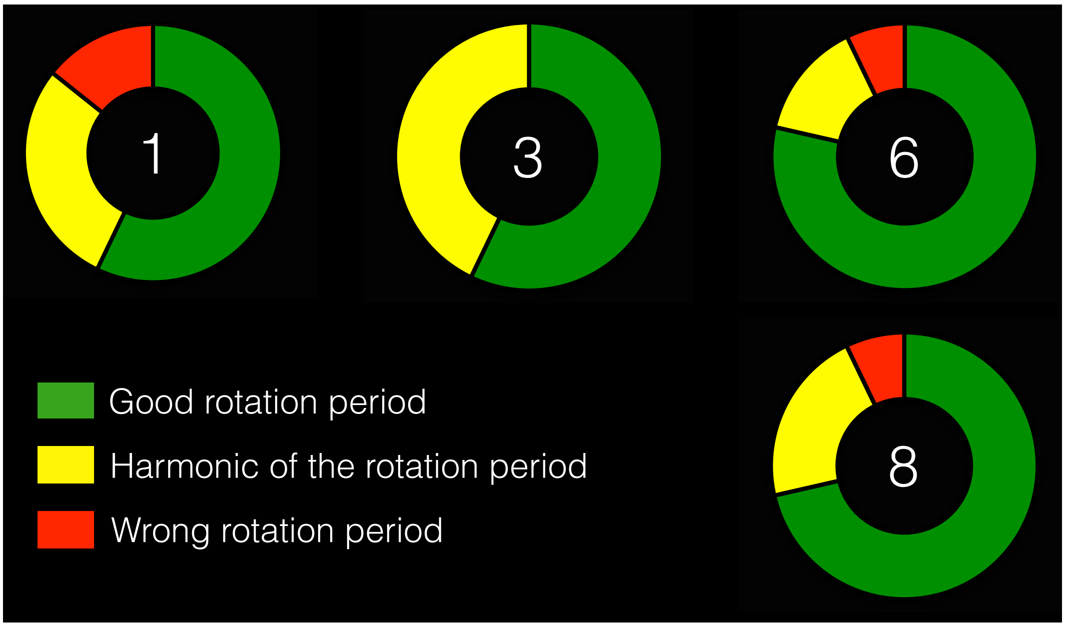


Fig. 12. Stellar rotation periods detected by teams 1, 3, 6 and 8 for the 14 systems of the RV fitting challenge.

4.1. Detection of stellar rotation periods

As described in detail in Dumusque (2016), the data of the RV fitting challenge include planetary signals, but also stellar signals, i.e., oscillations, granulation, short-term activity and long-term activity signals. Among all these stellar signals, the most difficult to deal with is short-term activity, induced by active regions, i.e., spots and plages, rotating with the stellar surface. Because several active regions rotating with the star are present simultaneously on the stellar surface, the observed RV signal induced by short-term activity is characterized by signals at the stellar rotation period $P_{\rm rot}$, and its harmonics ($P_{\rm rot}/2$, $P_{\rm rot}/3$, . . . , Boisse et al. 2011). Therefore, one of the very important aspects to differentiate between planetary signal and short-term activity signal is the detection of the stellar rotation period. If a signal is found in the RVs with a periodicity similar to $P_{\rm rot}$ or its harmonics, it is very likely that this signal is induced by active regions. For teams that used a GP regression to model short-term activity, it is essential for them to have a good guess of the stellar rotation period, otherwise the flexibility of a GP applied to the RVs could model planetary signals.

Teams 1, 3, 6 and 8 reported stellar rotation periods for all the RV fitting challenge systems, while the other teams did not explicitly derive such an estimate. All the teams that performed this analysis compared the signals found in the RVs with the ones found in the other observables sensitive to activity, i.e., the calcium activity index log(R'_{HK}), the BIS SPAN and the FWHM. A clear detection at the same period in the RVs and any other observables was assigned to a non-planetary component, because certainly due to short-term activity. Team 1 used a GLS periodogram to have a first estimate of the stellar rotation period, and then fitted a GP using a MCMC starting at this first estimate (see Equation B.1). Team 3 smooths the time series of the different activity observables using a moving average, and then looked for the stellar rotation period using a GLS periodogram. Team 6 looked at significant peaks in the DFT of the RVs and different activity observables, and finally team 8 at significant peaks in the GLS and ℓ_1-periodograms of the RVs, the BIS SPAN and the FWHM.

In Figs. A.1 to A.6, we show for each team and system, the stellar rotation period found. In Fig. 12, we summarize those results by classifying them as:

- correct rotation period detected in green,
- detection of an harmonic of the true rotation period in yellow,
- and wrong rotation period announced in red).

There are only a small number of wrong periods detected, which is positive. We will see their impact in Section 4.2 when analyzing the detection of planetary signals. However, we can see that in 20 to 45% of the cases, the different teams detected a harmonic of the stellar rotation period, and not the true period used to model the data. This can be a problem as we expect short-term activity to induce signals at $P_{\rm rot}$, $P_{\rm rot}/2$, $P_{\rm rot}/3$ and so on, but not at $2P_{\rm rot}$ and $3P_{\rm rot}$. Therefore, if the detected stellar rotation period is in fact $P_{\rm rot}/2$, it is possible to confuse an activity signal found at $P_{\rm rot}$ with a planet. We will see further that this case happened when team 1 analyzed systems 5, 9, 10, 11 and 12, team 3 analyzed system 13 and team 7 analyzed system 5.

We know that depending on the active region configuration on the stellar surface, and depending on the sampling of the data, the first harmonic of the rotation period ($P_{\rm rot}/2$) can have more power that the fundamental ($P_{\rm rot}$, Boisse et al. 2011). To prevent confounding a signal due to short-term activity with a planetary signal when the detected stellar rotation period is a harmonic of the real period, an easy solution is simply to reject signals at $2P_{\rm rot}$ and $3P_{\rm rot}$. In addition, we can use the average activity level of a star and its spectral type to guess its rotation period. First demonstrated by Noyes et al. (1984), and then updated by Mamajek & Hillenbrand (2008), a relation exists between the average $\log({\rm R}'_{HK})$ level, the spectral type and the stellar rotation period, with a few day error. Therefore, when analyzing the RVs of an old star, for which the rotation period is longer than 20 days, using such a relation can tell us if the period detected is the true stellar rotation period, or a harmonic of it. The spectral type of the stars were not given for the RV fitting challenge, therefore the different teams could not use this relation to estimate rotation periods. This was done on purpose because, as explain in detail in Dumusque (2016), only the variation of the $\log({\rm R}'_{HK})$ was properly simulated for the RV fitting challenge, and not the absolute value of it. Therefore using the average $\log({\rm R}'_{HK})$ level given in the RV fitting challenge dataset to calculate rotation periods would give wrong rotational period estimates. If this would have been possible, several yellow detections in Fig. 12 would turn green, therefore only leaving a few mistakes. This is something that should be taken into account for any further RV fitting challenges.

Because the different teams did not make mistakes on the same systems, it is difficult to conclude on the origin of these mistakes. However, among all the teams, team 3 performed the best as it reported no mistakes. Therefore, the technique used by this team to estimate stellar rotation periods from the activity observables ($\log({\rm R}'_{HK})$, FWHM and BIS SPAN), consisting on first modeling correlated noise using a moving average and then analyzing the residuals to find the stellar rotation period, seems to be the most robust (see Section 2.3).

4.2. Detection of planetary signals

In this section, we analyze the results of the different teams in terms of planetary detection. In total, 14 planetary systems were given, including a total of 45 simulated planetary signals and 6 published planetary signals probably present in real datasets 9,10,11 and 14 (α Centauri Bb and Corot-7b, c and d). As we can see in Dumusque (2016), and in Figs. A.1 to A.6, the semi-amplitude of the planetary signals was ranging between 0.16 and 5.85 $\rm m\,s^{-1}$, with rather low eccentricities. The RV fitting challenge time series for systems 9, 10, 11, and 14 are real observations obtained with HARPS, while the time series of all the other systems were simulated using the modelization of stellar signals described in Dumusque (2016). Note that no planetary signal were present in simulated systems 4, 8 and 13. In addition, no planetary signal was injected in real system 9, however the time series used for this system are the published HARPS measurements of α Centauri B that led to the discovery of an Earth-mass planet (Dumusque et al. 2012), therefore a 0.5 $\rm m\,s^{-1}$planetary signal might be recovered when analyzing this system.

4.2.1. Comparing the results obtained by the different teams

To compare the results between the different team, we define the K/N ratio as:

$$K/N = \frac{K_{\rm pl}}{{\rm RV}_{\rm rms}}\sqrt{N_{\rm obs}}, \qquad (1)$$

where $K_{\rm pl}$ is the semi-amplitude of each planetary signal, N_{obs} is the number of observation in each system, and ${\rm RV}_{\rm rms}$ is the RV rms of each system once the best-fit of a model consisting of a linear correlation with $\log({\rm R}'_{HK})$ plus a second order polynomial as a function of time was removed. This model allows removal of the effect of magnetic cycles (Meunier & Lagrange 2013; Dumusque et al. 2011b) and any long-term drift in the RVs due to binary companions.

In Fig. 13, we summarize the results of the different teams when analyzing only the first five systems and all the systems. For each signal detected, we assign a different color flag depending on the true signals present in the data. The different possibilities are:

- dark green: the team recovered a planetary signal that exists in the data and would have published the result.
- light green: the team recovered a planetary signal that exists in the data but is not confident enough in its detection for publication.
- yellow: the team recovered a planetary signal that exists in the data and would have published the result, however the semi-amplitude or period is wrong compared to the truth or an alias of the true signal was detected.
- grey: the team recovered a planetary signal that exists in the data but is not confident enough in its detection for publication. The semi-amplitude or period is wrong compared to the truth or an alias of the true signal was detected.
- white: non-detected planetary signal for which $K/N > 7.5$.
- cyan: non-detected planetary signal for which $K/N \leq 7.5$.
- orange: the team recovered a planetary signal that does not exist in the data but is not confident enough in its detection for publication.
- red: false positive or false negative, i.e., the team recovered a planetary signal that does not exist in the data and would have published the result, or the team rejected with confidence the detection of a true signal, respectively.

To study planetary population, we believe that the most important criteria are publishable planets with correct parameters (dark green flag), false positives or false negatives (red flag) and non-detection of planetary signals (white flag for $K/N > 7.5$, cyan flag for $K/N \leq 7.5$). The selection of the threshold $K/N =$

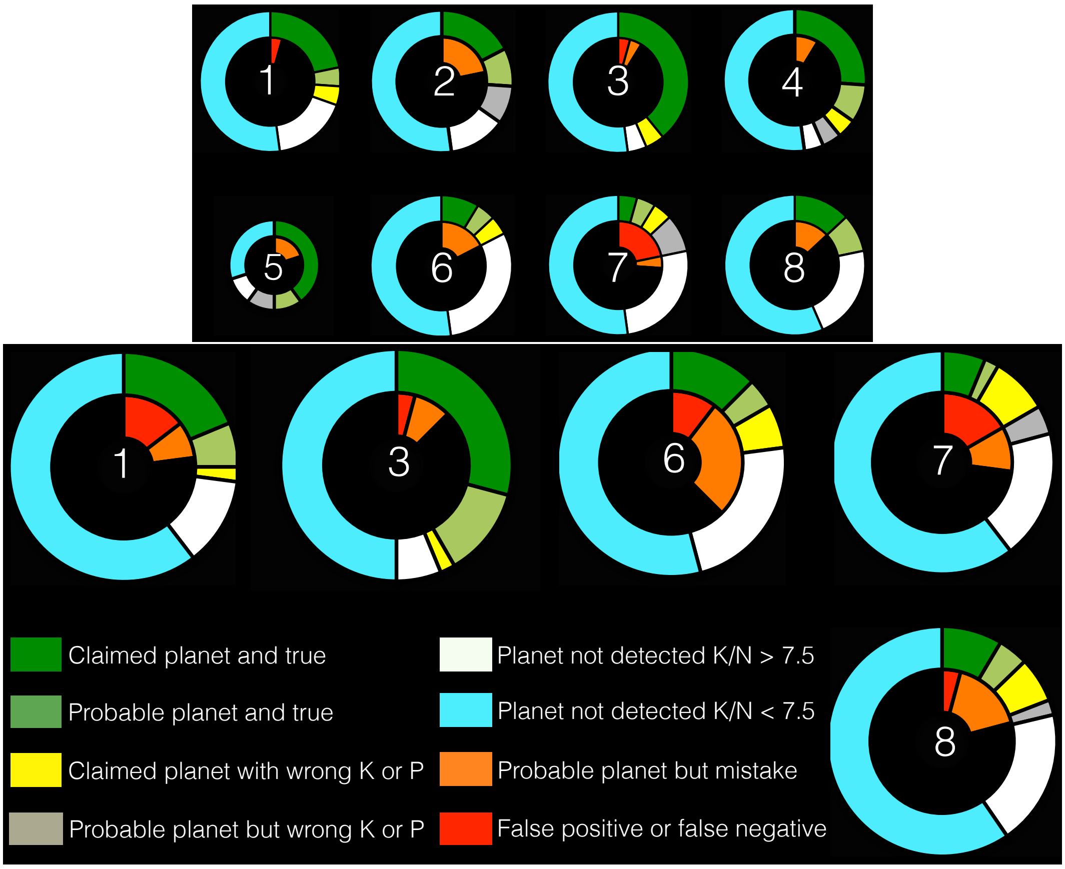


Fig. 13. *Top:* Summary of the signals detected in the first 5 systems of the RV fitting challenge data set. All the teams have analyzed those system, expect team 5 that only looked at system 1 and 2. The different color flags are defined in the legend and in more details in the second paragraph of Section 4.2. For each team, the outer circle diagram represents true planetary signals that were present in the data, and show how well the different teams could recover those. The inner circle diagram represents signals announced by the teams, but that were not present in the provided RV measurements. *Bottom:* Same but considering all the systems in the RV fitting challenge. Only teams 1, 3, 6, 7 and 8 performed this analysis. Size of the circle diagrams represents the number of systems analyzed: large size for all 14 systems, medium size for the first 5 systems and small size for the first 2 systems.

7.5 will be discussed in the next paragraph. Publishable signals that are slightly wrong (yellow flag) represent only a small fraction of the detections and therefore should not strongly bias planetary population statistics. All the other signals flagged as light green, grey, and orange would not have been published. Therefore, in Fig. 13, the most successful teams in terms of planet detection should have a large dark green region, while having small red, white and cyan regions. Given those criteria, we can separate the teams in two different groups: teams 1 to 5, and teams 6 to 8. This delimitation separates teams that used a Bayesian framework with red-noise models, which allows to compare between different solutions and model stellar signals, from teams that used other frameworks (see Table 1 and Sections 2 and 3 for more details about the different techniques used).

In Figs. 14 and 15, we plot the different planets that have been detected by the different teams as a function of the K/N ratio. We also highlight the false positives and false negatives. Note that only one false negative was announced: the true planetary signal in system 11 with a K/N ratio of 6 rejected by team 8 (see Fig. A.5 and the lowest red dot for team 8 in Fig. 15). Except for this signal, all the other mistakes correspond to false positives therefore we will only discuss false positive in the rest of the section.

Looking at the results of the different teams for the first five systems (Figs. 14), we see that teams 1, 3, 6 and team 4 were able to detect confidently (i.e. color flags dark green and yellow) planetary signals with a K/N ratio as low as 6 and 7.5, respectively. Excluding false positives found at the stellar rotation period when the correct rotation period was detected a priori, because those mistakes could have been avoided (hatched red dots), those teams did not detect any false positives above a K/N level of 5. For the other teams, i.e. 2, 5, 7 and 8, it was more difficult to detect confidently planetary signals with a small K/N ratio, and the threshold between detecting and not detecting a planetary signal is closer to K/N = 10. We note also that team 7 detected a lot of false positives, therefore the filtering technique in frequency space they used does not seem optimal to prevent false positives.

On the first 5 systems of the RV fitting challenge, there was not many planetary signals with a K/N ratio between 5 and 10. It is therefore worth analyzing the results of teams 1,3, 6, 7 and 8 with the entire data set to get a better idea at which threshold in K/N planets start to be detected. In Fig. 15, all the teams that analyzed the entire data set were able to confidently detect planetary signals with a K/N level above 7.5. However teams 1 and 3 detected most of the planetary signals above this threshold, which is not the case for teams 6, 7 and 8. We therefore see here a significant difference between techniques using a Bayesian framework with model comparison in addition to red noise models and technique that do not.

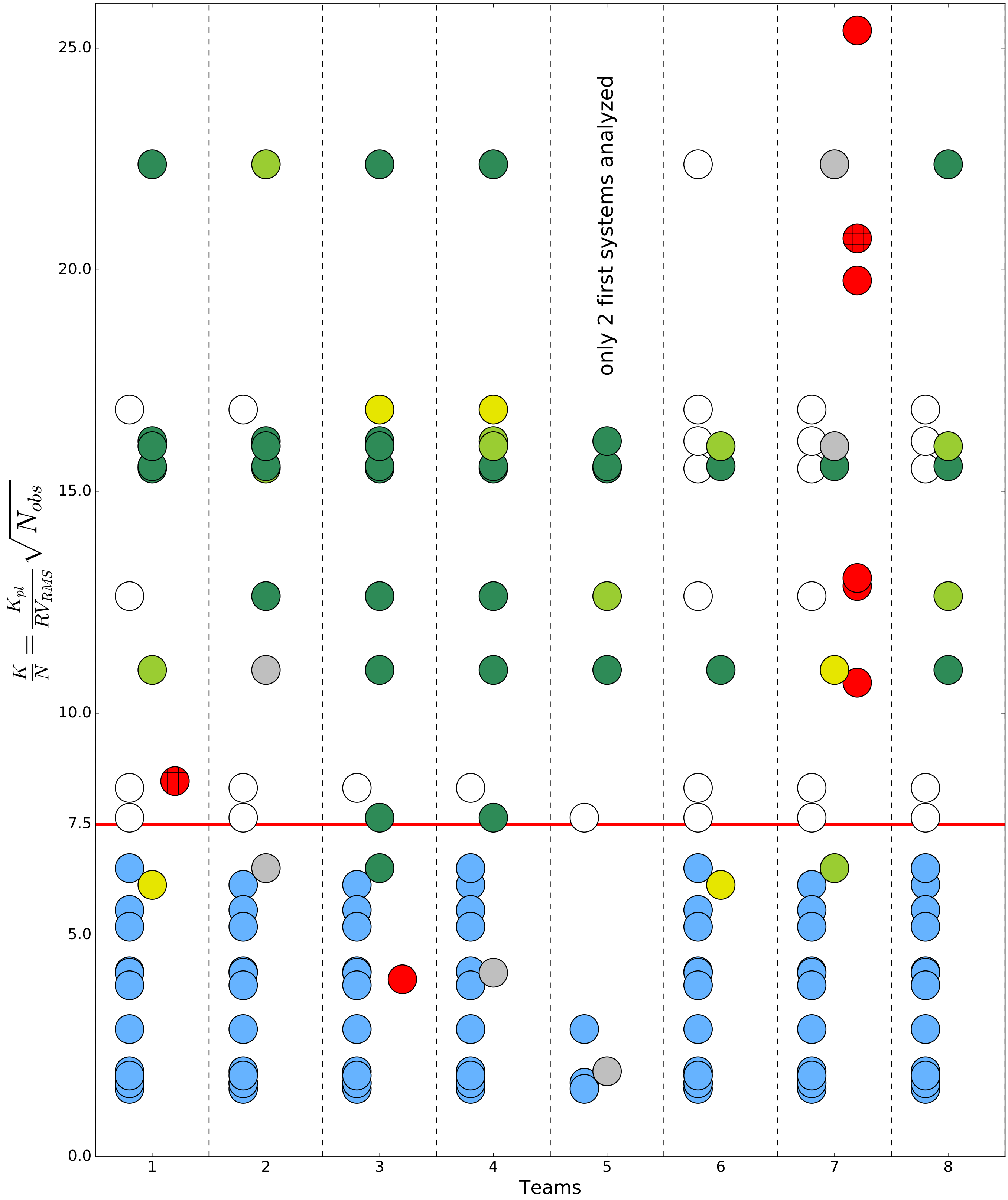


Fig. 14. K/N ratio for all the planets present in the 5 first systems of the RV fitting challenge, in addition to the false positives and the false negatives announced for those 5 first systems. All the teams analyzed those 5 first systems, except team 5 that only looked at the first two systems, which explain why there is less dots corresponding to planetary signals. The different color flags are defined in the legend of Fig. 13 and in more details in the second paragraph of Section 4.2. We separate the false positives or false negatives appearing in red in two categories. Either they cannot be explained easily (plain red dots), or the activity signal at the stellar rotation period has been confused with a planetary signal despite the fact that the correct stellar rotation period was found a priori (hatched red dots). The red horizontal line corresponds to a K/N ratio of 7.5. Note that the RV rms used to calculate K/N is the rms of the raw RVs once the best-fit of a model consisting of a linear correlation with log(R'_{HK}) plus a second order polynomial as a function of time was removed. This model allows removing the effect of magnetic cycles and any long-term drift in the RVs. We removed from this plot the 2 planets in system 11 and 12 that have an orbital period longer than 3000 days, much longer than the timespan of the data, which explain why they were not detected by any team despite their large K/N values (see Section 4.2.4).

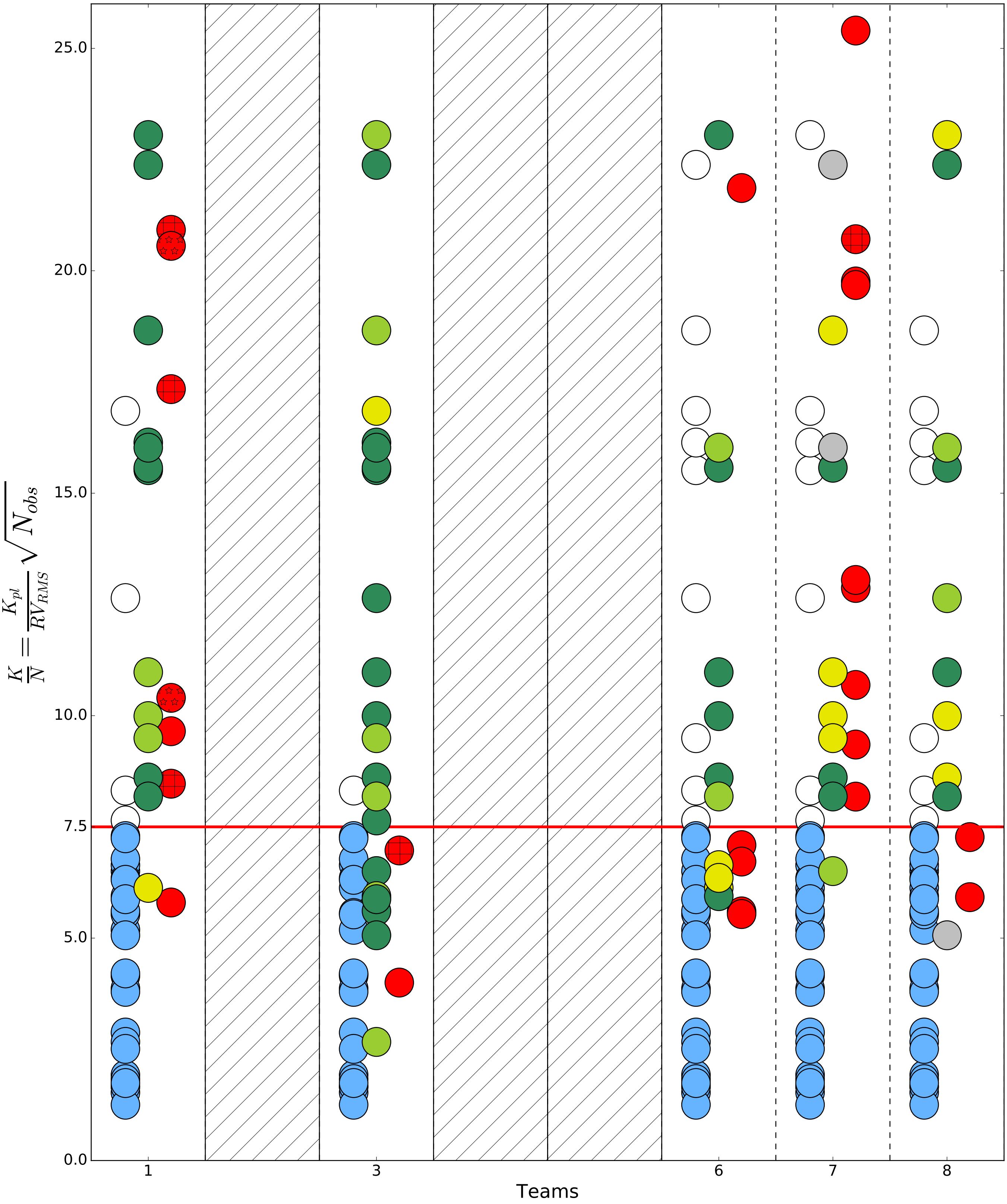


Fig. 15. Same as Fig. 14 but for all the signals announced by the different teams in the entire dataset of the RV fitting challenge. Only team 1, 3, 6, 7 and 8 performed this full analysis. We separate the false positives or false negatives appearing in red in three categories. Either they cannot be explained easily (plain red dots), or the activity signal at the stellar rotation period has been confused with a planetary signal despite the fact that the correct stellar rotation period was found a priori (hatched red dots) or the activity signal at the stellar rotation period has been confused with a planetary signal when a wrong stellar rotation period was found a priori (red dots with stars).

In Figs. 15, we distinguish between three types of false positives: those that cannot be explained easily (plain red dots), those that correspond to the stellar rotation period when the correct rotation period was detected a priori (hatched red dots), and those that correspond to the stellar rotation period when a wrong rotation period was detected a priori (red dots with stars). The second type of false positive could have been avoided assuming that all signals close to the rotation period should be excluded, and the third type could have been avoided using a better algorithm to estimate the stellar rotation period[2]. We therefore decided to exclude the second and third-type false positives discussed jsut above from the following discussion. When doing so, we see that team 1 detected 2 false positives at K/N ratios of 6 and 9.7, therefore the confidence level to detect a planetary signal without risking a false positive is close to $K/N = 10$. For team 3 this level is closer to $K/N = 5$, and for team 6 and 7, closer to $K/N = 7.5$. Finally, team 7 have detected too many false positives, up to $K/N = 25$, making it impossible to estimate a threshold between confidently detecting a planetary signal without risking a false positive. As a general conclusion, the limit between confident and non-confident detections is somewhere close to $K/N = 7.5$. Only the method used by team 3 allows to detect a few candidates with a K/N ratio between 5 and 7.5, without risking of announcing a false positive.

In Table 2, we report for each team the recovery rate of planetary signals detected and publishable with K/N ratios above and below 7.5. Planetary signals detected correspond to color flags dark green, light green, yellow and grey, and publishable planets to color flag dark green and yellow only (see definition of color flags in the second paragraph of Section 4.2). We show the results when studying only the first 5 systems, analyzed by all the teams except team 5 that only worked on system 1 and 2, and when studying all the RV fitting challenge system, analyzed by teams 1, 3, 6, 7, and 8.

When looking at detected planetary signals with $K/N > 7.5$, it is clear that teams 1 to 5, which used a Bayesian framework with model comparison in addition to red noise models, were more successful at finding those type of planetary signals. There is however a difference between detecting planetary signals, and being confident in those to publish them. Only publishable results will be used for planetary statistics, therefore those are the most important. When looking at publishable planetary signals with $K/N > 7.5$, we arrive to a similar ranking in performance. However, except in the special case of team 3 when analyzing the 5 first models, 20 to 30% of planetary detections with $K/N > 7.5$ will not be good enough to lead to publications. Those signals are however detected, which is a valuable argument to get more data for a system and thus publish it at a later stage.

As we can see in Fig. 15, compared to team 1 and 3, team 6 to 8 have a significantly larger proportion of detections flagged as yellow, i.e. planetary signals for which the teams are confident in the detection, however the period or semi-amplitude differs from the true solution, or the detection corresponds to an alias of the true signal. Including those solutions would bias any statistical analysis on planetary semi-amplitude and period distributions. Therefore, when searching for planetary signal for which $K/N > 7.5$, techniques using Bayesian model selection with red-noise models allow to recover more signals and give better estimates of the orbital period and semi-amplitude (see Section 4.2.2 for more details).

When looking at detected planetary signals with $K/N \leq 7.5$, we have very few number of detections and even less of publishable planetary signals. It is therefore difficult to draw out strong conclusions. Only team 3 and 6 were able to find a significant number of candidates, 6 and 4 respectively. However, 3 planetary signals out of the 4 found by team 6 have incorrect period or semi-amplitude (yellow color flag), in addition to 3 false positives announced. The results found by team 3 are therefore more robust.

The K/N ratio is used here as a measure of the detectability of planetary signals that were present in the RV fitting challenge dataset. In the case of the RV fitting challenge, systems 1 to 13 were very similar, with:

- a large number of measurements, between 433 and 527,
- planetary signals much shorter than the timespan of the data,
- planetary signals with a good phase coverage,
- and stellar signals similar to what is observed for the Sun, with a RV rms ranging from 1.8 to 5 $\mathrm{m\,s^{-1}}$once the best-fit of a model consisting of a linear correlation with $\log(\mathrm{R}'_{HK})$ plus a second order polynomial as a function of time was removed to the raw RVs.

For those systems, we show that for most of the teams it was possible to confidently detect planetary signals above a threshold in K/N of 7.5, without announcing false positives. However teams that used a Bayesian framework with model comparison in addition to red noise models where able to detect nearly all the planetary signals above this threshold, which was not the case for the other teams.

Systems 14 and 15 exhibit a higher level of stellar signals, with a RV rms of 8.9 and 7.6 $\mathrm{m\,s^{-1}}$once the best-fit of a model consisting of a linear correlation with $\log(\mathrm{R}'_{HK})$plus a second order polynomial as a function of time was removed to the raw RVs. In addition, those two systems presented only 170 measurements. However, in these very different cases compared to system 1 to 13, most of the team were able to detect the signals of Corot-7c and Corot-7d with a K/N ranging from 8 to 10, while only team 1 was able to detect the $K/N = 5$ signal of Corot-7b (see Fig. A.6 in the appendix). Therefore, it seems that this K/N threshold of 7.5, and probably 5 for team 3, can be applied to quite different set of data. We however only have a few systems in the RV fitting challenge to test this hypothesis and a detailed study of the behavior of this threshold as a function of number of measurements, ratio of the planet period to the timespan of the measurements, phase coverage of the signal and level of stellar signals would be something extremely useful to explore.

This threshold $K/N = 7.5$, or 5 for team 3, is the best that can currently be done by the different teams when analyzing the RV fitting challenge data set. We expect that this level goes down with ongoing progress in the different methods used to detect planetary signals in the presence of stellar signals.

4.2.2. Accuracy of estimated planetary period and semi-amplitude

Detecting a true planetary signal and being confident in its veracity is a difficult task, even more when $K/N \leq 7.5$. We discussed in the previous sections that some techniques to deal with stellar signals are performing better. In this section, we look at the parameters found for each planet detected by the different teams, and compare them to the true parameters that were used to generate those planetary signals. When dealing with real data including planetary signals like in system 14, we compared with the latest published parameters.

[2] For the two cases of third-type false positive, only detected by team 1, all the other teams were able to find the correct stellar rotation period.

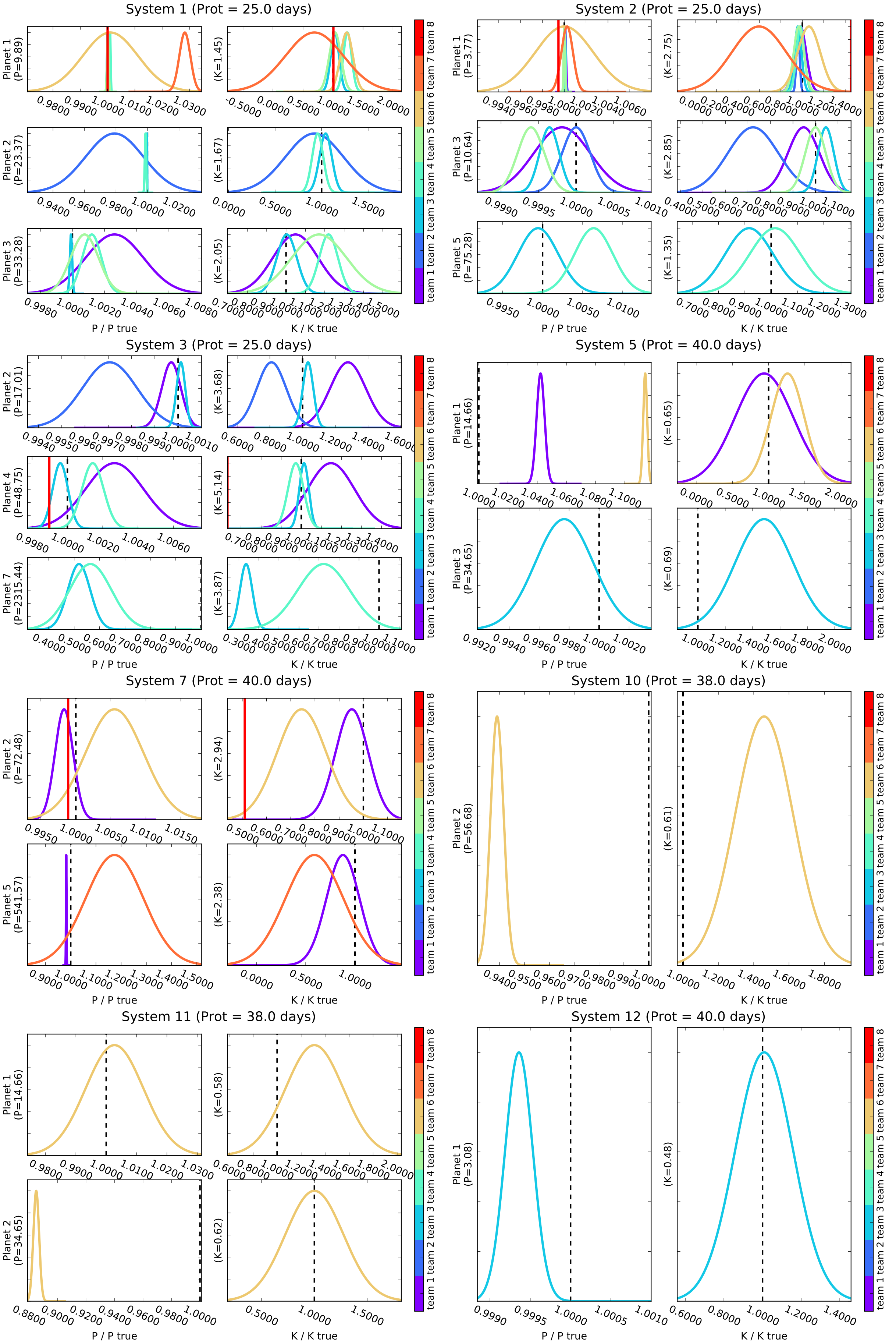


Fig. 16. Periods and semi-amplitudes reported by each team for the planets detected in systems 1, 2, 3, 5, 7, 10, 11 and 12. We divided those parameters by the values of the true signals, so that a perfect estimate would fall on one. Team 8 was the only one not reporting error bars on their parameters, therefore we just show their best estimates as red vertical lines.

Table 2. Recovery rate of planetary signals detected (dark green, light green, yellow and grey color flags), of publishable planets with correct orbital parameters (dark green and yellow color flags) and of false positives and false negatives (red color flag) for each team. Recovery rates between 0 and 33, 33 and 66, and 66 and 100% are highligthed in red, yellow and green, respectively.

	Bayesian framework + red-noise models					Other techniques		
	1: Torino	2: Oxford	3: Tuomi	4: Gregory	5: Geneva	6: Hatzes	7: Brera	8: IMCCE
Detected planetary signals $K/N > 7.5$								
5 first systems (total 10)	80% (8)	70% (7)	90% (9)	90% (9)	83% (5/6)	30% (3)	40% (4)	50% (5)
all systems (total 18)	68% (12)	-	83% (15)	-	-	39% (7)	50% (9)	50% (9)
Publishable planetary signals $K/N > 7.5$								
5 first systems (total 10)	50% (5)	40% (4)	90% (9)	70% (7)	67% (4/6)	20% (2)	20% (2)	30% (3)
all systems (total 18)	50% (9)	-	61% (11)	-	-	28% (5)	39% (7)	39% (7)
Detected planetary signals $K/N \leq 7.5$								
5 first systems (total 13)	8% (1)	8% (1)	8% (1)	8% (1)	25% (1/4)	8% (1)	15% (2)	0%
all systems (total 30)	3% (1)	-	20% (6)	-	-	13% (4)	7% (2)	3% (1)
Publishable planetary signals $K/N \leq 7.5$								
5 first systems (total 13)	0%	0%	8% (1)	0%	0%	8% (1)	8% (1)	0%
all systems (total 30)	3% (1)	-	13% (4)	-	-	13% (4)	3% (1)	0%

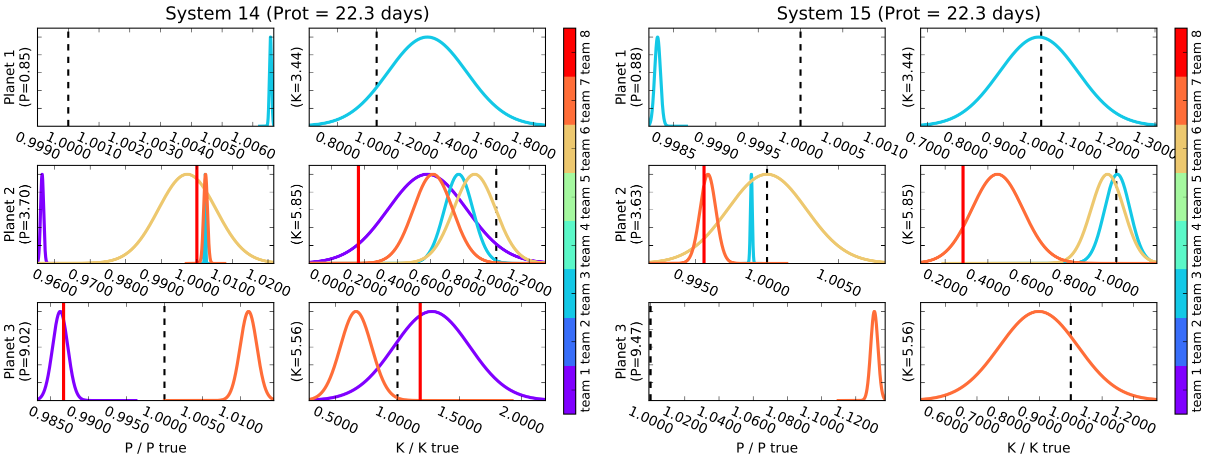


Fig. 17. Same as Fig. 16 for systems 14 and 15.

In Figs. 16 and 17, we show for each system and each planet detected, the period and semi-amplitude parameters found by each team. We divided those parameters by the values of the true signals, so that a perfect estimate would fall on one. Team 8 was the only team that did not report error bars on their measurements, therefore we just show their best estimates as red vertical lines in Figs. 16 and 17. Note that we included in those figures all the signals for which the teams were confident in (dark green and yellow color flags).

It is difficult to conclude which team have recovered the best period and semi-amplitude parameters for the planetary signals present in the RV fitting challenge data set, as some teams discovered more signals than others. However, if we look at system 1, 2, 3, 14 and 15, for which many teams detected a lot of signals, team 3 found the best estimate for the different parameters. Therefore using a moving average model to account for stellar signals seems the best approach to deal with the correlated noise induced by RV stellar signals. This does not mean that team 3 always found the best parameters, and as a general conclusion, the errors on the period and semi-amplitude parameters are often underestimated, certainly due to the fact that the models used to account for stellar signals are not perfect.

4.2.3. Comparing the results obtained by teams using a GP regression and teams using other red-noise modelings

When looking in Fig. 14 at the first 5 systems, we notice that GP regression techniques (team 1 and 2) could not confidently recover 3 planetary signals with $K/N > 7.5$ compared to teams that used other red-noise models (team 3 and 4). This is probably due to the fact that stellar signals do not have the same covariance in RVs than in the activity observables. Just as an example, an equatorial spot on a star seen equator-on will induce a sinusoidal variation with a period of $P_{\rm rot}/2$ when the spot will pass on the visible hemisphere (e.g., Dumusque et al. 2014a). Then no signal is observed when the spot is behind the star. For the signal in log(R$'_{HK}$), projection effect comes into play; log(R$'_{HK}$) increases when a spot moves from the limb to the stellar disc center, and symmetrical decreases when a spot moves from the disc center to the opposite side of the limb. Therefore, the observed signal is half of a sine wave with the stellar rotation period. Then, like for the RVs, no signal is seen when the spot is behind the star. Although the signals in RVs and log(R$'_{HK}$) seems to have different periods, they do not as after a full stellar rotation period, the same signal will appear again in RVs and log(R$'_{HK}$) if there is no spot evolution. However, because many spots are present on the stellar surface at the same time, in addition to their evolution, the RVs of stars spot-dominated will tend to have a significant signal at period $P_{\rm rot}/2$, while the $\log(\mathrm{R}'_{HK})$ will present a significant signal at $P_{\rm rot}$. This example shows that the RVs and the activity observables can have a different covariance and therefore could explain why teams 1 and 2 were confident in fewer true planetary signals than team 3 and 4 when analyzing the first five systems. We believe that further investigations should be done to confirm or reject this argument.

When looking at the 14 systems of the RV fitting challenge, team 1 announced 6 false positives with $K/N > 7.5$ (see Fig. 15). Out of those 6 false positives, one cannot be explained easily (plain red dots), three are due to a confusion with the stellar rotation period, thus stellar activity, despite the fact that the correct stellar rotation period was found a priori (hatched red dots), and the two last one due to a confusion with the stellar rotation period knowing that a wrong stellar rotation period was found a priori (red dots with stars). Although it is difficult to draw any conclusion on the first and the two last false positives, for the three other ones, it seems that the GP regression used by team 1 was not able to fully model stellar activity and that an extra Keplerian with a period close to stellar rotaion was needed to better explain the observed RV variations. This is therefore something that should be explored in detail as we do not want GP regression to create some false-positives. Team 2 used GP regression with a different formalism, unfortunately it is not possible to compare the results of team 1 and 2 because team 2 only analyzed the first five systems, for which only one false positive was announced by team 1.

4.2.4. Detection of long period planets

In the different systems of the RV fitting challenge, we injected long-period signals to see if they could be recovered despite the RV stellar signal induced by magnetic cycles. In total 6 planetary signals with periods longer than 500 days were present in the data:

- 596 and 2315 days for system 3, with K/N = 8.3 and 16.9, respectively (K=1.91 and 3.87 $\mathrm{m\,s^{-1}}$),
- 616 days for system 5, with K/N = 5.2 (K=0.55 $\mathrm{m\,s^{-1}}$),
- 542 days for system 7, with K/N = 18.7 (K=2.38 $\mathrm{m\,s^{-1}}$),
- 3245 days for system 11, with K/N = 15.8 (K=1.54 $\mathrm{m\,s^{-1}}$),
- 3407 days for system 12, with K/N = 19.2 (K=1.64 $\mathrm{m\,s^{-1}}$).

Regarding our previous discussion about signal smaller than K/N = 7.5 (see Section 4.2.1), it is clear that the 596-day period signal in system 3, and the long-period signal injected in system 5 are difficult to find. However, all the other signals have $K/N > 15$ and could have been discovered by the different teams, mainly those using a Bayesian framework with red-noise models to mitigate the impact of stellar signals.

The long-period signals in systems 3, 11 and 12 are close to 6, 9 and 9 years, much longer than the 4-year time span of the RVs for these systems. Therefore, the RVs do not cover an entire phase of those signals, which makes it very difficult to characterize them, as in general orbital periods need closure for correct parameter estimation (e.g. Black & Scargle 1982). Team 3 and 4 reported a signal at 1202 and 1306 days for system 3, which is half of the period of the real signal. Because the RVs do not cover an entire phase, the power of the signal is transferred to its first harmonic, i.e., half of its period. The two other planetary signals were not detected, and this can be explained by the fact that the different teams added in their RV model a polynomial up to the second order to account for any drift in the data, which absorbs any long-period planetary signals if the time span of the data is much shorter than the orbital period of the planets.

The 542-day signal in system 7 has a shorter period than the time span of the data and was thus confidently announced by team 1 and 7, and detected but not confidently by team 3.

4.2.5. Detection of short period planets

In opposition to long-period planetary signals, 6 signals in the RV fitting challenge have periods shorter than 5 days:

- 3.77 days for system 2 with K/N = 15.6 (K=2.75 $\mathrm{m\,s^{-1}}$),
- 1.12 days for system 3 with K/N = 4.2 (K=0.96 $\mathrm{m\,s^{-1}}$),
- 0.82 day for system 10 with K/N = 7.3 (K=0.67 $\mathrm{m\,s^{-1}}$),
- 3.08 days for system 12 with K/N = 5.6 (K=0.48 $\mathrm{m\,s^{-1}}$),
- 0.85 day for system 14 with K/N = 5.1 (K=3.44 $\mathrm{m\,s^{-1}}$),
- and 0.88 days for system 15 K/N = 5.9 (K=3.44 $\mathrm{m\,s^{-1}}$).

Out of these 6 signals, all the teams recovered the one in system 2, which can be explained by the high K/N ratio. The 5 other signals all have $K/N \leq 7.5$ and therefore, following our discussion in Sections 4.2.1, they were difficult to find. Three of them, in systems 12, 14 and 15, were confidently recovered only by team 3. This shows that the technique used by team 3 seems more efficient at finding short-period planets. A probable explanation is that the moving average model used by team 3 includes correlation between points on short-period timescales, which therefore reduces the effect of granulation (timescale up to 2 days) and of stellar short-term activity over a few day timescale that can be important for active stars (systems 14 and 15). Short-period planets are therefore easier to find.

The short-period planets in systems 14 and 15 are the true and simulated version of Corot-7b, the first Earth-radius planet ever detected. This planet was first found by photometry (Léger et al. 2009), and then confirmed with the RV measurements of system 14 (Queloz et al. 2009). It is however interesting to see that only team 3 was able to recover this planet without imposing as priors the period and time of transit derived from photometry.

The planets recovered by team 3 in system 14 and 15 have a very similar periods and a smaller K/N ratio than the planet not detected in system 10. Therefore the K/N ratio is not the only criterion to separate detections from non-detection. The detection of the planets in system 14 and 15 and not in system 10 can be explained by two effects here: i) for system 14 and 15, the signals from short-term activity is dominating the other sources of stellar signal; short-term activity is better characterized and therefore easier to model with the moving average, and ii) the 0.82-day planetary signal has an amplitude much larger than the expected perturbations induced by the other sources of stellar signal, i.e., granulation and stellar oscillations.

When comparing the K/N ratio of the planets in systems 10 and 12, we would guess that the one in system 10 is easier to recover. However, team 3 could not recover it but could confidently detect the other. This can be explained by the longer period of the planet in system 12, 3 days, compared to close to 1 day for the planet in system 10. Indeed, this difference in period implies a better phase coverage of the planet with a longer orbital period. Nine measurements per orbit can be obtained for a 3-day period planet when observing with a strategy of 3 measurements per night (similar to the sampling of the different systems in the RV fitting challenge) compared to only 3 for a 1-day period planet. In addition, planetary signals close to 1 day are more affected by granulation signal that affects RV measurement on a timescale smaller than 2 days, which makes them harder to find. Team 3 was the only team to recover the 3-day signal, and this is probably because it is the only team that considered, with its moving average model, correlation between points on short-period timescales. The moving average reduces the impact of granulation on RV measurements, and therefore increases the significance of planetary signals with similar or shorter periods.

Speaking about planetary signals with periods shorter than 5 days, we need to discuss system 9,10 and 11, for which the RVs have been extracted from the HARPS measurements that led to the detection of α Centauri Bb. The RVs and the activity observables for system 9 are the raw data published in Dumusque et al. (2012). We only reversed time, added a Gaussian noise of 0.05 $\mathrm{m\,s^{-1}}$, and changed the gamma velocity of the star, so that the time series for this system could not be recognized (Dumusque 2016). These modifications should not perturb the 0.5 $\mathrm{m\,s^{-1}}$ planetary signal of α Centauri Bb present in the data. This planetary signal should also be present in system 10 and 11, however for those systems we added extra planets, which can perturb the detection of this small semi-amplitude signal. The K/N ratio for α Centauri Bb in system 9, 10 and 11 would be 5.7, 5.4 and 5.1 respectively, implying a very challenging detection according to the discussion in Section 4.2.1. None of the teams were able to recover the signal of α Centauri Bb. However, team 3 was able to confidently recover the simulated planetary signal of α Centauri Bb in simulated system 12. Based on this result, team 3 should have been able to detect the signal of α Centauri Bb in systems 9,10 and 11. It is therefore possible that the signal of α Centauri Bb announced in Dumusque et al. (2012) is in fact a spurious one induced by a combination of the sampling of the data and of the model used to fit stellar activity, as questioned by Rajpaul et al. (2016). The signal of α Centauri Bb is however at the limit of what can be done with current methods to deal with stellar signals and more RV measurements are needed to really conclude on the existence or not of α Centauri Bb.

4.3. Results for real RV data compared to simulated ones

Testing the efficiency of different techniques on simulated data can be useless, if those simulated data are not realistic.

To be able to test the realism of simulated data, Dumusque (2016) included in the data set of the RV fitting challenge some real observations done with HARPS, and then simulated RVs as close as possible to those real data. Thus systems 6 and 7, 9 and 13, 11 and 12, and 14 and 15, including real and simulated data, can be compared. Unfortunately, as discussed in Dumusque (2016), system 6 cannot be used. The comparison of the other systems is described below and each time we refer to the RV rms, this one is calculated on the raw RVs once the best-fit of a model consisting of a linear correlation with log(R'_{HK}) plus a second order polynomial as a function of time was removed:

System 9 and 13 The two systems have a similar RV rms, 1.82 and 2.06 $\mathrm{m\,s^{-1}}$, respectively, therefore the level of stellar signal present in the simulated data seems realistic. For real system 9, only team 1 could not find the correct stellar rotation period, and team 1 and 6 announced 3 false positives. For system 13, only team 3 could find the correct stellar rotation period while the other teams found the first harmonic. Team 3 and 8 announced 2 false positives. It is difficult to conclude as no similar mistake was done on the two systems, however it seems that it was as difficult to analyze the real and the simulated data.

System 11 and 12 The simulated data seems to have a realistic level of stellar signal as the two systems have a similar RV rms, 2.04 and 1.78 $\mathrm{m\,s^{-1}}$, respectively. Regarding stellar rotation period, all the teams could recover the correct value, except team 1 for simulated system 12. By making this mistake, team 1 announced a false positive at $P_{\mathrm{rot}}/2$. Four mistakes were done on system 12, while only two were done on system 11. In addition, team 3, 6 and 8 could recover the $K/N = 5.95$ signal at 15 days orbiting system 11 (note however that team 8 announced it as false-negative), while no one could recover the same signal in system 12. From the comparison of these two systems, it seems that is was more difficult for the different teams to find planetary signals in the simulated data, and easier to make some mistakes.

System 14 and 15 These two systems exhibit similar RV rms, 8.86 and 7.64 $\mathrm{m\,s^{-1}}$, respectively, therefore implying at first order a correct modelization of stellar signals. From the comparison of those two systems presenting the real and simulated data of Corot-7, we find that it was easier to detect the correct rotation period in the real time series. Regarding false positives, none was announced for system 14, while 2 were detected in system 15. Except team 3 that found exactly the same solution, the different teams were more confident in the signals found in system 14 than in system 15.

In general, it was slightly more difficult for the different teams to analyze simulated data. However, we note that team 3, that performed the best at the exercise of the RV fitting challenge, found very similar solutions when analyzing real and simulated data. We therefore believe that even if not perfect, the simulated data are realistic enough to be used to test the efficiency of techniques to recover planetary signature despite stellar signals.

5. Conclusion

In total, 8 different teams participated in the analysis of the RV fitting challenge data set. They all used different techniques to find the low eccentricity planets that were hidden inside stellar signals. Except system 14 and 15, that present the real and simulated RVs of the active star Corot-7, all the other systems present a typical level of stellar signal for inactive G-K dwarfs. Those stars are the typical targets of most high-precision RV surveys searching for low-mass planets, and therefore the conclusions made here can be applied to most of the RV measurements gathered up to now.

With 14 different systems, 48 planets with semi-amplitude ranging between 0.16 and 5.85 $\mathrm{m\,s^{-1}}$, and different modelizations of stellar signals, the number of parameter is huge, and it is difficult to draw some strong conclusions with the analysis of only 8 different teams. In addition, the data set of the RV fitting challenge was given to the different teams 8 months before the deadline. Techniques used by team 1 to 5, based on a Bayesian framework with red-noise models, required significantly more computational time than the other techniques used by teams 6 to 8. As a result, team 2 and 4 could only analyze the first five systems out of 14, team 5 only the first two, and teams 1 to 5 used statistical shortcuts to find planetary signals in the data, or could not test all possible models, taking the risk of biasing their final results. Readers should therefore be aware that the results presented in this paper are preliminary, and depends on (1) how much time each team was able to invest in the challenge, (2) how mature their analytical methods were, and (3) how experienced the team members were with such analyses. Looking at the results presented in this paper, it seems that some techniques work better at recovering planets despite stellar signals, however further investigation need to be performed to be confident in the conclusions presented here. Note that the best techniques all require intensive computational efforts.

A first important step before finding planets is the detection of the stellar rotation period. For team 1 and 2, this period is used in their model that accounts for short-term activity, for the other teams, this period and its harmonics defines regions in period space were planetary signal should be excluded because likely due to short-term activity. Finding the correct stellar rotation period is therefore crucial to reduce the number of false positives in the end, and team 3, using its moving average model to account for stellar signals, performed the best at this exercise. Among all the teams that reported explicitly a stellar rotation period, we notice that only a small number of mistakes were done. However in many cases, a harmonic of the stellar rotation period was found, which can be dangerous because then a signal at the true rotation period can be confounded with a planet. To distinguish between the true stellar rotation period and a harmonic of it, an activity level-rotation calibration as the one developed by Mamajek & Hillenbrand (2008) can be used. This was however not possible here due to lack of information in the RV fitting challenge data set, but this is something that people analyzing RV data should strongly consider to prevent false positives (see Section 4.1).

When looking at the recovery rate of planetary signals for each team, teams can be separated in two groups. Teams 1, 2, 3, 4 and 5 that used a Bayesian framework with red-noise models and teams 6, 7 and 8 that used *pre-whitening*, compressed sensing and/or filtering techniques in the frequency domain to deal with stellar signals. The first group discovered more true planetary signals than the second one, and also made fewer mistakes. In addition, when asked if those detections are significant enough to lead to publications, the first group of teams was also more confident in announcing a planetary signal. The planets for which the K/N ratio (see Eq. 1) was above 7.5 were nearly all recovered by the best teams. Below this threshold, the detection rate drops to 20% at best. Note that team 3 was able to find the smallest K/N ratio true planetary signals, with K/N ratios between 5 and 7.5, without announcing false positives. Below $K/N = 5$, no planetary signals were confidently recovered, it is therefore a lower limit for planetary detections using data with similar properties as those of the RV fitting challenge (see Section 4.2.1).

Regarding accuracy when estimating the best orbital parameters for planetary signals qualified as publishable most of the teams recovered the correct orbital parameters within $3-\sigma$ from the truth. A few signals were however out of the $3-\sigma$ limit, which is probably due to the fact that the models used in this paper to account for stellar signals are not perfect. This is not surprising as models to account for stellar activity are not perfect, however those are the best we have so far (see Section 4.2.2).

Besides recovering real planetary signal in the data and giving correct orbital parameters, it is very important that the false positive rate stays low. Above a threshold of 7.5 in K/N ratio, team 7 announced nine false positives, team 1 six, team 6 one, and the other teams none. The technique use by team 7 is therefore prone to false positive and cannot be used to reliably detect planets. Team 1 also announced several false positives, however a few of them correspond to the stellar rotation period, despite the fact that the correct rotation period was found a priori. Therefore, although their GP regression has the correct stellar rotation period, it seems that the GP regression cannot fully model stellar signals and that an extra sinusoidal signal is needed. Further investigation on GP modeling needs therefore to be performed to be sure that GP regression does not create false-positives. For the time being, signals close to the stellar rotation period or its harmonics should always be associated to stellar activity to prevent false positives (see Section 4.2.3).

For planetary signals with periods longer than 500 days, several effects make their detection difficult. It is common that drifts in the data are observed due to magnetic cycle effects and long-period binaries. To remove such long-period signals, the different teams corrected the RVs from magnetic cycle effects by using the observed long-term correlation between the RVs and the different activity observables (log(R'_{HK}), BIS SPAN, FWHM), and removed the effect of binaries by fitting polynomials as a function of time. People analyzing RVs data should be aware that such a model can absorb the signal of planets that have orbital periods similar or longer than the time span of the data, and that orbits need closure before inferring planet parameters (see Section 4.2.4).

When analyzing the recovery of planets with periods shorter than 5 days, teams 3 found 4 out of 6 planets, including 3 with $K/N \leq 7.5$, while all the other teams found only the planet for which $K/N > 7.5$. It seems therefore that the moving average model used by team 3 is more sensitive to short-period planets because such a model consider measurement correlation on short-period timescales, which therefore mitigate the effect of granulation on quiet stars, and the strong short-timescale effect of short-term activity on active stars like Corot-7. We would therefore encourage people using GP modeling, or apodized Keplerians, to add on top of their model a correlation between measurements on short-period timescales, as this seems critical to detect short-period planetary signals with small K/N ratios (see Section 4.2.5).

The RV rms of real and simulated systems was similar, going in the direction that the different sources of stellar signals were realistically taken into account. Team 3, that performed the best at the exercise of the RV fitting challenge found very similar solutions between real and simulated data. However, it was slightly more difficult for the other teams to analyze simulated data. We therefore believe that even if not perfect, the simulated data are realistic enough to be used to test the efficient of techniques to recover planetary signature despite stellar signals (see Section 4.3

With more time, each technique can be improved, and the different teams are making progress (see Gregory 2016; Hara et al. 2016). The Oxford team also made some important progresses (priv. comm.). Now they are able to perform a full Bayesian marginalisation over all parameters (planets + GP), which give them much more reliable Bayesian model evidences. Following a private communication with N.C Hara from team 8, It seems that their method is now delivering similar performances in terms of planetary detection as Bayesian framework techniques using red-noise models and with a much shorter computational time (see bottom plot in Fig. 11 and Hara et al. 2016). However, following the first results of the RV fitting challenge presented here, techniques using a Bayesian framework and red-noise models seem the most efficient at modeling the effect of stellar signals, and therefore detecting true planetary signals while limiting the number of false positives. Moving average, GP regression and apodized Keplerian modelizations should be investigated further, to see the sensitivity of these models to planets at short and long-periods, to planets with a similar period than stellar rotation, to planet with high and low K/N ratios, to multi-planet systems.

The goal of the RV fitting challenge was to test the efficiency of the different techniques to recover planets in RV data given the presence of stellar signals, while limiting the number of false positives. As we can see in the different discussions above, the

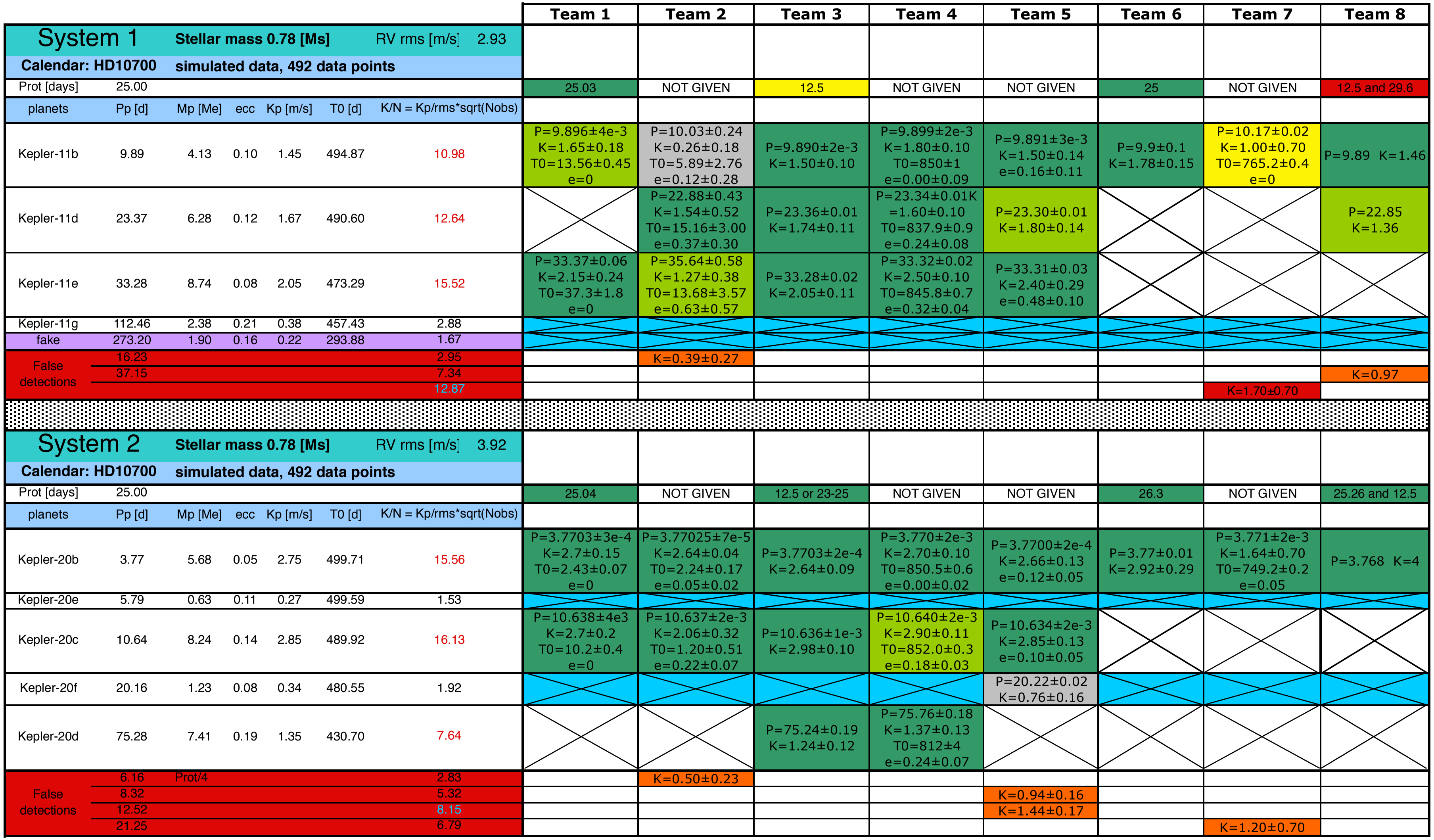
Bayesian framework and moving average model used by team 3 performed the best. Then, in second position comes the Bayesian framework and apodized Keplerian model used by team 4, followed by the Bayesian framework and GP model used by team 1 in third position. Although team 1 performed well in analyzing system 6 to 15 in terms of true planetary signals detected, they announced a lot of false-positives at the stellar rotation period. Further investigation need to be performed to test if those false positives originate from the GP regression they used, or from another part of their method.

Team 3 was able to confidently discover a few planetary signals with K/N ratios between 5 and 7.5 without announcing false positives, and nearly all the planetary signal with $K/N > 7.5$. Team 4 and 1 detected confidently most of the signals for which $K/N > 7.5$, and none below this threshold. In conclusion, for RV measurement similar to those of the RV fitting challenge, a ratio $K/N = 7.5$ seems to be a threshold separating confident detection from non-detection of planetary signals. Note however that the method used by team 3 could confidently detect ~20% of the planetary signals with K/N ratios as low as 5, without announcing false positives.

Acknowledgements. We thanks the two referees of this paper for their valuable comments that greatly improved the paper and made the results more impactful and significant. We are extremely grateful to A. Correia, D. A. Fischer, E. B. Ford, J. S. Jenkins, A. Santerne, S. H. Saar, and G. Scandariato for interesting discussions about the design of the RV fitting challenge data set and the best way to organize it. X. Dumusque is grateful to the Society in Science−The Branco Weiss Fellowship for its financial support. S. Aigrain acknowledges support from the Leverhulme Trust and the UK Science and Technology Facilities Council. The Brera team acknowledges financial support from INAF through the "Progetti Premiali" funding scheme of the Italian Ministry of Education, University, and Research. M. Damasso acknowledges funding from Progetto Premiale INAF "Way to Other Worlds" (WoW), decreto 35/2014. A. S. Bonomo is grateful to the European Union Seventh Framework Programme (FP7/2007-2013) for funding under Grant agreement number 313014 (ETAEARTH). R. D. Haywood gratefully acknowledges a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. M. Tuomi, G. Anglada-Escudé and H. R. A. Jones acknowledge the Leverhulme Trust funding scheme for grants RPG-2014-281 and STFC ST/M001008/1.

References

Aigrain, S., Pont, F., & Zucker, S. 2012, MNRAS, 419, 3147
Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2014, ArXiv e-prints [arXiv:1403.6015]
Anglada-Escudé, G., Arriagada, P., Vogt, S. S., et al. 2012, ApJ, 751, L16
Anglada-Escudé, G. & Butler, R. P. 2012, ApJS, 200, 15
Anglada-Escudé, G. & Tuomi, M. 2015, Science, 347, 1080
Arentoft, T., Kjeldsen, H., Bedding, T. R., et al. 2008, ApJ, 687, 1180
Baluev, R. V. 2013, MNRAS, 429, 2052
Baranne, A., Queloz, D., Mayor, M., et al. 1996, A&AS, 119, 373
Black, D. C. & Scargle, J. D. 1982, ApJ, 263, 854
Boisse, I., Bonfils, X., & Santos, N. C. 2012, A&A, 545, 109
Boisse, I., Bouchy, F., Hébrard, G., et al. 2011, A&A, 528, A4
Boisse, I., Moutou, C., Vidal-Madjar, A., et al. 2009, A&A, 495, 959
Borgniet, S., Meunier, N., & Lagrange, A.-M. 2015, A&A, 581, A133
Candès, E., J., Romberg, J., K., & Tao, T. 2006, Communications on Pure and Applied Mathematics, 59, 1207
Chib, S. & Jeliazkov, I. 2001, Journal of the American Statistical Association, 96, 279
Del Moro, D. 2004, A&A, 428, 1007
Del Moro, D., Berrilli, F., Duvall, Jr., T. L., & Kosovichev, A. G. 2004, Sol. Phys., 221, 23
Díaz, R. F., Ségransan, D., Udry, S., et al. 2016, A&A, 585, A134
Donoho, D. 2006, Information Theory, IEEE Transactions on, 52, 1289
Dravins, D. 1982, ARA&A, 20, 61
Dumusque, X. 2016, A&A, 593, A5
Dumusque, X., Boisse, I., & Santos, N. C. 2014a, ApJ, 796, 132
Dumusque, X., Bonomo, A. S., Haywood, R. D., et al. 2014b, ApJ, 789, 154
Dumusque, X., Glenday, A., Phillips, D. F., et al. 2015, ApJ, 814, L21
Dumusque, X., Lovis, C., Ségransan, D., et al. 2011a, A&A, 535, A55
Dumusque, X., Lovis, C., Udry, S., & Santos, N. C. 2011b, in IAU Symposium, Vol. 276, IAU Symposium, ed. A. Sozzetti, M. G. Lattanzi, & A. P. Boss, 530–532
Dumusque, X., Pepe, F., Lovis, C., et al. 2012, Nature, 491, 207
Dumusque, X., Udry, S., Lovis, C., Santos, N. C., & Monteiro, M. J. P. F. G. 2011c, A&A, 525, A140
Faria, J. P., Haywood, R. D., Brewer, B. J., et al. 2016, A&A, 588, A31
Feroz, F. & Hobson, M. P. 2014, MNRAS, 437, 3540
Ford, E. B. 2006, ApJ, 642, 505
Foreman-Mackey, D. 2015, George: Gaussian Process regression, Astrophysics Source Code Library
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Gregory, P. C. 2011, MNRAS, 415, 2523
Gregory, P. C. 2012, ArXiv e-prints [arXiv:1212.4058]
Gregory, P. C. 2013 (Springer)
Gregory, P. C. 2016, MNRAS, 458, 2604
Gregory, P. C. & Fischer, D. A. 2010, MNRAS, 403, 731
Grunblatt, S. K., Howard, A. W., & Haywood, R. D. 2015, ApJ, 808, 127
Haario, H., Saksman, E., & Tamminen, J. 2001, Bernoulli, 7, 223
Haario, H., L. M. M. A. S. E. 2006, Statistics and Computing, 16, 339
Hara, N. C., Boué, G., Laskar, J., & Correia, A. C. M. 2016, MNRAS[arXiv:1609.01519]
Hatzes, A. P. 2013, Astronomische Nachrichten, 334, 616
Hatzes, A. P. 2016, A&A, 585, A144
Hatzes, A. P., Dvorak, R., Wuchterl, G., et al. 2010, A&A, 520, A93
Haywood, R. D., Collier Cameron, A., Queloz, D., et al. 2014, MNRAS, 443, 2517
Haywood, R. D., Collier Cameron, A., Unruh, Y. C., et al. 2016, MNRAS, 457, 3637
Hébrard, É. M., Donati, J.-F., Delfosse, X., et al. 2014, MNRAS, 443, 2599
Isaacson, H. & Fischer, D. 2010, ApJ, 725, 875
Kass, R. E., R. A. E. 1995, Journal of the American Statistical Association, 90, 773
Kjeldsen, H., Bedding, T. R., Butler, R. P., et al. 2005, ApJ, 635, 1281
Kuschnig, R., Weiss, W. W., Gruber, R., Bely, P. Y., & Jenkner, H. 1997, A&A, 328, 544
Lanza, A. F., Molaro, P., Monaco, L., & Haywood, R. D. 2016, A&A, 587, A103
Léger, A., Rouan, D., Schneider, J., et al. 2009, A&A, 506, 287
Lenz, P. & Breger, M. 2005, Communications in Asteroseismology, 146, 53
Lindegren, L. & Dravins, D. 2003, A&A, 401, 1185
Lovis, C., Dumusque, X., Santos, N. C., et al. 2011, ArXiv e-prints [arXiv:1107.5325]
Makarov, V. V. 2010, ApJ, 715, 500
Mamajek, E. E. & Hillenbrand, L. A. 2008, ApJ, 687, 1264
Mayor, M., Bonfils, X., Forveille, T., et al. 2009, A&A, 507, 487
Meunier, N., Desort, M., & Lagrange, A.-M. 2010, A&A, 512, A39
Meunier, N. & Lagrange, A.-M. 2013, A&A, 551, A101
Meunier, N., Lagrange, A.-M., Borgniet, S., & Rieutord, M. 2015, A&A, 583, A118
Newton, M. & Raftery, A. 1994, Journal of the Royal Statistical Society, 56, 3
Noyes, R. W., Hartmann, L. W., Baliunas, S. L., Duncan, D. K., & Vaughan, A. H. 1984, ApJ, 279, 763
O'Toole, S. J., Tinney, C. G., & Jones, H. R. A. 2008, MNRAS, 386, 516
Pepe, F., Mayor, M., Galland, F., et al. 2002, A&A, 388, 632
Perrakis, K., N. I. T. E. G. 2014, Computational Statistics and Data Analysis, 77, 54
Pont, F., Sing, D. K., Gibson, N. P., et al. 2013, MNRAS, 432, 2917
Poretti, E., Boccato, C., Claudi, R., et al. 2016, Mem. Soc. Astron. Italiana, 87, 141
Queloz, D., Bouchy, F., Moutou, C., et al. 2009, A&A, 506, 303
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
Rasmussen, C. E.; Williams, C. K. I. 2006, Gaussian Processes for Machine Learning, 2nd edn., ed. M. Press (MIT Press)
Rauer, H., Catala, C., Aerts, C., et al. 2014, Experimental Astronomy, 38, 249
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9143, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 20
Roberts, D. H., Lehar, J., & Dreher, J. W. 1987, AJ, 93, 968
Robertson, P., Endl, M., Henry, G. W., et al. 2015, ApJ, 801, 79
Robertson, P., Mahadevan, S., Endl, M., & Roy, A. 2014, Science, 345, 440
Saar, S. H. 2009, in American Institute of Physics Conference Series, Vol. 1094, 15th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun, ed. E. Stempels, 152–161
Saar, S. H. & Donahue, R. A. 1997, ApJ, 485, 319
Santos, N. C., Mortier, A., Faria, J. P., et al. 2014, A&A, 566, A35
Taubman, D. S. & Marcellin, M. W. 2002, Proceedings of the IEEE, 90, 1336
Tuomi, M., Anglada-Escudé, G., Gerlach, E., et al. 2013, A&A, 549, A48
Tuomi, M. & Jones, H. R. A. 2012, A&A, 544, A116
Tuomi, M., Jones, H. R. A., Barnes, J. R., Anglada-Escudé, G., & Jenkins, J. S. 2014, MNRAS, 441, 1545
Vogt, S. S., Butler, R. P., & Haghighipour, N. 2012, Astronomische Nachrichten, 333, 561
Vogt, S. S., Butler, R. P., Rivera, E. J., et al. 2010, ApJ, 723, 954
Wright, J. T. 2005, PASP, 117, 657
Wright, J. T. & Howard, A. W. 2009, ApJS, 182, 205
Zechmeister, M. & Kürster, M. 2009, A&A, 496, 577



System 1	Stellar mass 0.78 [Ms]				RV rms [m/s]	2.93	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
Calendar: HD10700	simulated data, 492 data points													
Prot [days]	25.00						25.03	NOT GIVEN	12.5	NOT GIVEN	NOT GIVEN	25	NOT GIVEN	12.5 and 29.6
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
Kepler-11b	9.89	4.13	0.10	1.45	494.87	10.98	P=9.896±4e-3 K=1.65±0.18 T0=13.56±0.45 e=0	P=10.03±0.24 K=0.26±0.18 T0=5.89±2.76 e=0.12±0.28	P=9.890±2e-3 K=1.50±0.10	P=9.899±2e-3 K=1.80±0.10 T0=850±1 e=0.00±0.09	P=9.891±3e-3 K=1.50±0.14 e=0.16±0.11	P=9.9±0.1 K=1.78±0.15	P=10.17±0.02 K=1.00±0.70 T0=765.2±0.4 e=0	P=9.89 K=1.46
Kepler-11d	23.37	6.28	0.12	1.67	490.60	12.64		P=22.88±0.43 K=1.54±0.52 T0=15.16±3.00 e=0.37±0.30	P=23.36±0.01 K=1.74±0.11	P=23.34±0.01K =1.60±0.10 T0=837.9±0.9 e=0.24±0.08	P=23.30±0.01 K=1.80±0.14			P=22.85 K=1.36
Kepler-11e	33.28	8.74	0.08	2.05	473.29	15.52	P=33.37±0.06 K=2.15±0.24 T0=37.3±1.8 e=0	P=35.64±0.58 K=1.27±0.38 T0=13.68±3.57 e=0.63±0.57	P=33.28±0.02 K=2.05±0.11	P=33.32±0.02 K=2.50±0.10 T0=845.8±0.7 e=0.32±0.04	P=33.31±0.03 K=2.40±0.29 e=0.48±0.10			
Kepler-11g	112.46	2.38	0.21	0.38	457.43	2.88								
fake	273.20	1.90	0.16	0.22	293.88	1.67								
False detections	16.23					2.95		K=0.39±0.27						
	37.15					7.34								K=0.97
						12.87							K=1.70±0.70	
System 2	Stellar mass 0.78 [Ms]				RV rms [m/s]	3.92								
Calendar: HD10700	simulated data, 492 data points													
Prot [days]	25.00						25.04	NOT GIVEN	12.5 or 23-25	NOT GIVEN	NOT GIVEN	26.3	NOT GIVEN	25.26 and 12.5
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
Kepler-20b	3.77	5.68	0.05	2.75	499.71	15.56	P=3.7703±3e-4 K=2.7±0.15 T0=2.43±0.07 e=0	P=3.77025±7e-5 K=2.64±0.04 T0=2.24±0.17 e=0.05±0.02	P=3.7703±2e-4 K=2.64±0.09	P=3.770±2e-3 K=2.70±0.10 T0=850.5±0.6 e=0.00±0.02	P=3.7700±2e-4 K=2.66±0.13 e=0.12±0.05	P=3.77±0.01 K=2.92±0.29	P=3.771±2e-3 K=1.64±0.70 T0=749.2±0.2 e=0.05	P=3.768 K=4
Kepler-20e	5.79	0.63	0.11	0.27	499.59	1.53								
Kepler-20c	10.64	8.24	0.14	2.85	489.92	16.13	P=10.638±4e3 K=2.7±0.2 T0=10.2±0.4 e=0	P=10.637±2e-3 K=2.06±0.32 T0=1.20±0.51 e=0.22±0.07	P=10.636±1e-3 K=2.98±0.10	P=10.640±2e-3 K=2.90±0.11 T0=852.0±0.3 e=0.18±0.03	P=10.634±2e-3 K=2.85±0.13 e=0.10±0.05			
Kepler-20f	20.16	1.23	0.08	0.34	480.55	1.92					P=20.22±0.02 K=0.76±0.16			
Kepler-20d	75.28	7.41	0.19	1.35	430.70	7.64			P=75.24±0.19 K=1.24±0.12	P=75.76±0.18 K=1.37±0.13 T0=812±4 e=0.24±0.07				
False detections	6.16	Prot/4				2.83		K=0.50±0.23						
	8.32					5.32					K=0.94±0.16			
	12.52					8.15					K=1.44±0.17			
	21.25					6.79							K=1.20±0.70	

Fig. A.1. Summary of signal detection for RV fitting challenge systems 1 and 2 reported by the different teams. Color flags are defined in the legend of Fig. 13 and in more details in the second paragraph of Section 4.2. Note that the RV rms shown here is the one obtained from the raw RVs once the best-fit of a model consisting of a linear correlation with log(R'_{HK}) plus a second order polynomial as a function of time was removed.

							Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
System 3	Stellar mass 0.78 [Ms]				RV rms [m/s]	5.09					NOT PERFORMED			
Calendar: HD10700	simulated data, 492 data points													
Prot [days]	25.00						25.04	NOT GIVEN	12.5	NOT GIVEN		26.7	NOT GIVEN	24.96
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
HD10180b	1.12	1.32	0.00	0.96	498.92	4.18								
HD10180d	17.01	12.42	0.15	3.68	488.49	16.04	P=17.005±3e-3 K=4.7±0.4 T0=11.19±0.89 e=0.12±0.06	P=16.96±0.02 K=2.98±0.33 T0=12.51±2.56 e=0.08±0.10	P=17.012±3e-3 K=3.80±0.12	P=16.992±3e-3 K=3.98±0.14 T0=845±3 e=0.00±0.04		P=16.97±0.02 K=5.05±0.28	P=17.77±0.05 K=1.5±0.7 T0=756.5±0.6 e=0.03	P=16.9 K=2.2
fake	26.30	1.50	0.08	0.38	484.04	1.66								
HD10180e	48.75	24.89	0.06	5.14	484.22	22.40	P=48.88±0.08 K=5.9±0.6 T0=45.68±0.19 e=0.08±0.04	P=48.73±0.08 K=4.17±0.35 T0=28.54±13.16 e=0.02±0.20	P=48.73±0.02 K=5.21±0.12	P=48.82±0.03 K=5.00±0.20 T0=822±2 e=0.09±0.03			P=48.63±0.38 K=2.3±0.7 T0=967.0±1.2 e=0.8	P=48.7 K=3.25
fake	201.50	3.20	0.20	0.42	423.41	1.83								
HD10180g	595.98	21.19	0.13	1.91	122.54	8.32								
HD10180h	2315.44	67.26	0.16	3.87	-140.26	16.86			P=1202±102 K=1.30±0.12	P=1306±191 K=2.80±0.50 T0=145±201 e=0.00±0.61				
False detections	6.26	Prot/4				6.28						K=1.44±0.20		
	9.18					4.01			K=0.92±0.11					
	156.39					13.07							K=3.0±0.7	
System 4	Stellar mass 0.78 [Ms]				RV rms [m/s]	3.31					NOT PERFORMED			
Calendar: HD10700	simulated data, 492 data points													
Prot [days]	25.00						25.04	NOT GIVEN	12.5	NOT GIVEN		27.03	NOT GIVEN	27.3 and 12.54
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
NONE							NONE							
False detections	0.94					4.49				K=0.67				
	6.25	Prot/4				< 11 for all		K=1.41±0.05	K=1.02±0.14,			K=1.62±0.50		K=1.24
	10.99					12.53						K=1.87±0.35		
	11.74					10.12				K=1.51				
	13.18					13.87		K=2.07±020						
	14.56					10.72							K=1.6±0.7	
	37.70					20.77							K=3.1±0.7	

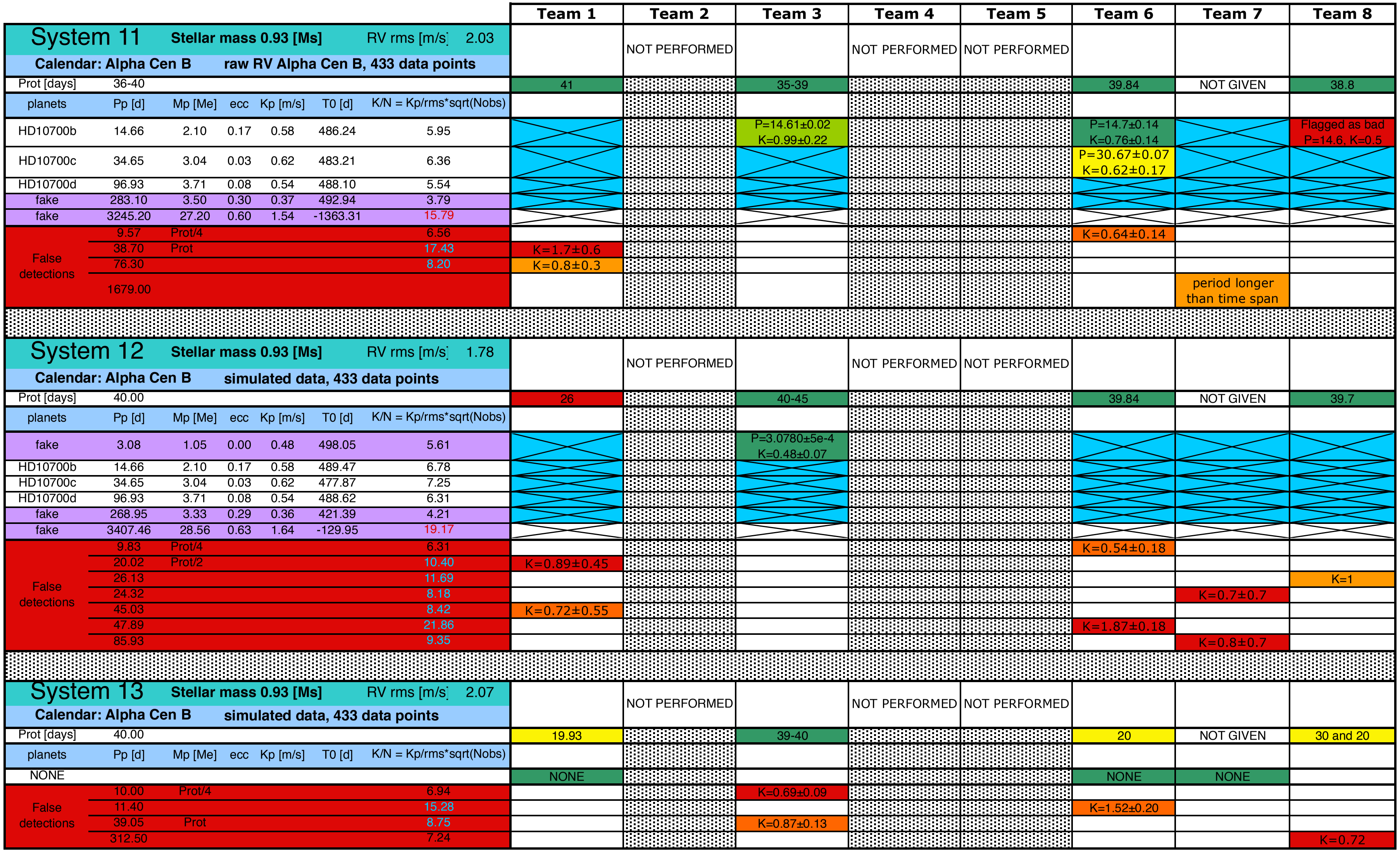
Fig. A.2. Same as Fig. A.1 but for RV fitting challenge systems 3 and 4.

							Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
System 5	Stellar mass 0.80 [Ms]				RV rms [m/s]	2.43					NOT PERFORMED			
Calendar: HD192310	simulated data, 527 data points													
Prot [days]	40.00						19.99	NOT GIVEN	19-21	NOT GIVEN	▨	40	NOT GIVEN	40.2, 19.9 and 44
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)					▨			
HD10700b	14.66	2.10	0.17	0.65	486.82	6.14	P=15.28±0.04 K=0.61±0.26 T0=99.2±1.4 e=0	☒	☒	☒	▨	P=16.33±0.02 K=0.82±0.15	☒	☒
fake	26.20	1.70	0.25	0.44	481.41	4.16	☒	☒	☒	P=0.96 K=0.73, siderial alias of 26.42d	▨	☒	☒	☒
HD10700c	34.65	3.04	0.03	0.69	467.37	6.52	☒	P=34.98±0.02 K=1.65±0.25 T0=24.59±1.60 e=0.93±0.31	P=34.57±0.07 K=1.02±0.15	☒	▨	☒	P=34.32±0.19 K=1.4±0.7 T0=812.3±1.1 e=0.01	☒
HD10700e	173.16	4.43	0.05	0.59	421.08	5.57	☒	☒	☒	☒	▨	☒	☒	☒
fake	283.10	3.50	0.30	0.41	462.39	3.87	☒	☒	☒	☒	▨	☒	☒	☒
HD10700f	616.32	6.34	0.03	0.55	414.69	5.20	☒	☒	☒	☒	▨	☒	☒	☒
False detections	9.80	Prot/4				9.22 and 11.01		K=0.98±0.06			▨	K=1.17±0.15		
	19.92	Prot/2				19.84					▨		K=2.1±0.7	
	39.70	Prot				8.50	K=0.9±0.3				▨			
	379.73					25.51					▨		K=2.7±0.7	
	1270.00					17.95					▨			K=1.9
System 7	Stellar mass 0.80 [Ms]				RV rms [m/s]	2.93		NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: HD192310	simulated data, 527 data points													
Prot [days]	40.00						20.03	▨	20	▨	▨	55.2	NOT GIVEN	20.04 and 33.5
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)		▨		▨	▨			
fake	38.32	1.55	0.02	0.34	464.01	2.66	☒	▨	P=38.18±0.07 K=0.54±0.08	▨	▨	☒	☒	☒
HD192310b	72.48	16.39	0.13	2.94	467.62	23.03	P=72.36±0.09 K=2.8±0.2 T0=57.24±0.77 e=0	▨	P=72.43±0.05 K=2.87±0.07	▨	▨	P=72.88±0.30 K=2.19±0.3	☒	P=72.4 K=1.5
fake	100.99	1.92	0.24	0.32	412.95	2.51	☒	▨	☒	▨	▨	☒	☒	☒
fake	303.80	1.47	0.12	0.16	263.19	1.25	☒	▨	☒	▨	▨	☒	☒	☒
HD192310c	541.57	24.72	0.33	2.38	251.45	18.65	P=532+0.6 K=2.09±0.4 T0=8.72±0.54 e=0	▨	P=534±4 K=2.47±0.08	▨	▨	☒	P=635.55±62.2 K=1.4±0.7 T0=322.3±16 e=0.09	☒
False detections	1622.00						perhaps, orbit not complete	▨		▨	▨			

Fig. A.3. Same as Fig. A.1 but for RV fitting challenge systems 5 and 7.

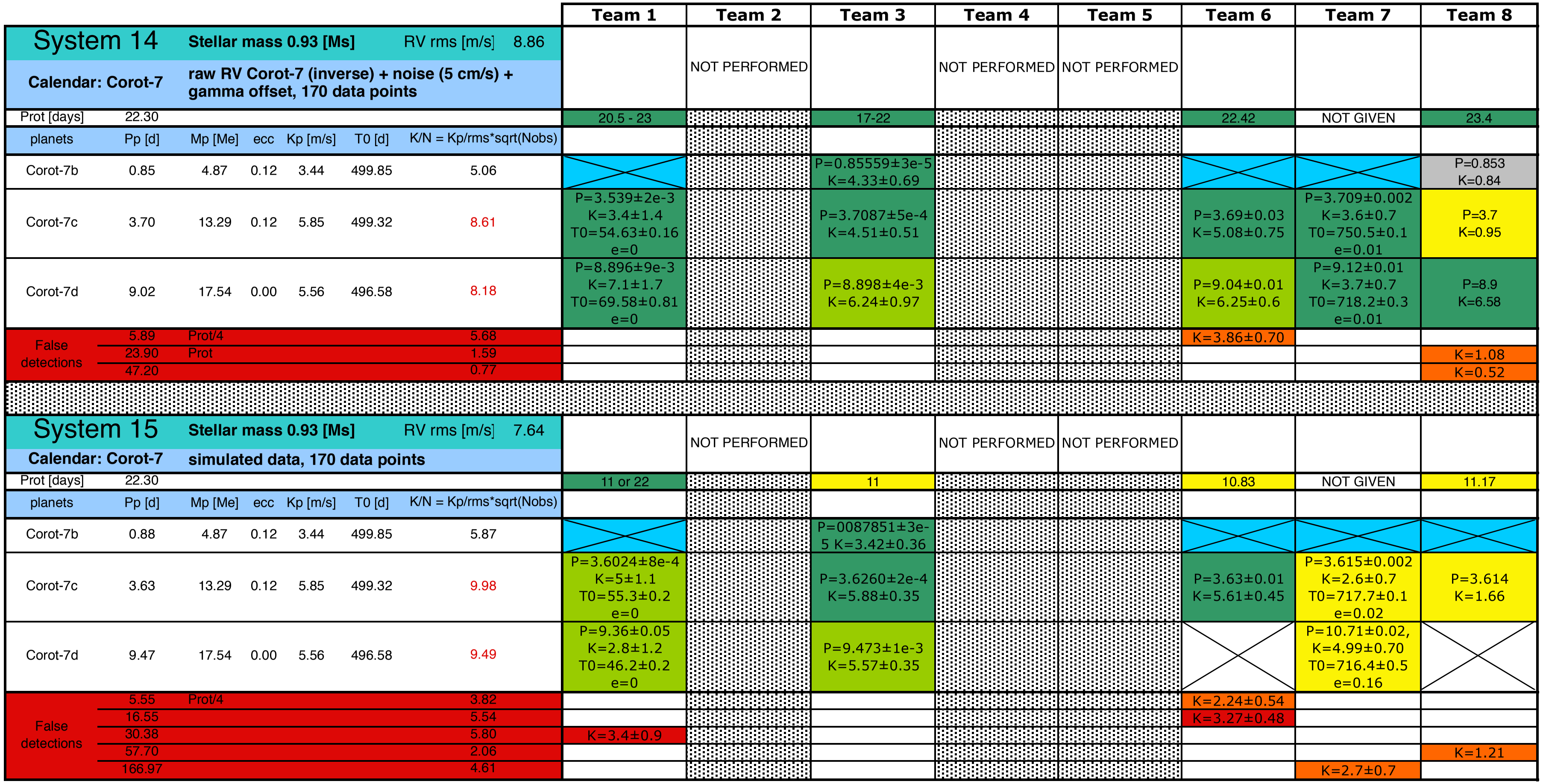
							Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
System 8	Stellar mass 0.80 [Ms]			RV rms [m/s]	2.10			NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: HD192310	simulated data, 527 data points													
Prot [days]	40.00						19.99		35-39			36.74	NOT GIVEN	39.7, 20.1 and 34
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
NONE							NONE							
False detections	10.00	Prot/4				6.89			K=0.63±0.10					
	10.30	Prot/4				10.39						K=0.95±0.15		
	15.40					19.68						K=1.8±0.15		
	54.18					19.68							K=1.8±0.7	
System 9	Stellar mass 0.93 [Ms]			RV rms [m/s]	1.82			NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Alpha Cen B	raw RV Alpha Cen B (inverse) + noise (5 cm/s) + gamma offset, 433 data points													
Prot [days]	36-40						46		35-39			39.52	NOT GIVEN	38.9
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
NONE														
False detections	9.56	Prot/4				7.43						K=0.65±0.10		
	14.49					5.60						K=0.49±0.1		
	20.92	Prot/2				13.72			K=1.20±0.21					
	24.87					7.09						K=0.62±0.1		
	38.74	Prot				20.58	K=1.8±0.3							
	74.40					6.17	K=0.54±0.29							
	165.00					2.29								K=0.2
	1529.00												period longer than time span	
System 10	Stellar mass 0.93 [Ms]			RV rms [m/s]	1.91			NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Alpha Cen B	raw RV Alpha Cen B, 433 data points													
Prot [days]	36-40						41		35-39			39.52	NOT GIVEN	38.9
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
fake	0.82	0.93	0.05	0.67	499.89	7.30	☒		☒			☒	☒	☒
HD85512b	56.68	3.49	0.11	0.61	498.78	6.65	☒		☒			P=53.22±0.15 K=0.89±0.10	☒	☒
fake	296.30	1.62	0.05	0.16	212.66	1.74	☒		☒			☒	☒	☒
False detections	9.57	Prot/4				8.28						K=0.76±0.11		
	15.04					6.75						K=0.62±0.13		
	38.72	Prot				21.03	K=1.93±0.38							
	84.14					9.70	K=0.89±0.39							
	1498.00												period longer than time span	

Fig. A.4. Same as Fig. A.1 but for RV fitting challenge systems 8, 9 and 10.



							Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
System 11	Stellar mass 0.93 [Ms]				RV rms [m/s]	2.03		NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Alpha Cen B	raw RV Alpha Cen B, 433 data points													
Prot [days]	36-40						41		35-39			39.84	NOT GIVEN	38.8
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
HD10700b	14.66	2.10	0.17	0.58	486.24	5.95	☒		P=14.61±0.02 K=0.99±0.22			P=14.7±0.14 K=0.76±0.14	☒	Flagged as bad P=14.6, K=0.5
HD10700c	34.65	3.04	0.03	0.62	483.21	6.36	☒		☒			P=30.67±0.07 K=0.62±0.17	☒	☒
HD10700d	96.93	3.71	0.08	0.54	488.10	5.54	☒		☒			☒	☒	☒
fake	283.10	3.50	0.30	0.37	492.94	3.79	☒		☒			☒	☒	☒
fake	3245.20	27.20	0.60	1.54	-1363.31	15.79	☒		☒			☒	☒	☒
False detections	9.57	Prot/4				6.56						K=0.64±0.14		
	38.70	Prot				17.43	K=1.7±0.6							
	76.30					8.20	K=0.8±0.3							
	1679.00												period longer than time span	
System 12	Stellar mass 0.93 [Ms]				RV rms [m/s]	1.78		NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Alpha Cen B	simulated data, 433 data points													
Prot [days]	40.00						26		40-45			39.84	NOT GIVEN	39.7
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
fake	3.08	1.05	0.00	0.48	498.05	5.61	☒		P=3.0780±5e-4 K=0.48±0.07			☒	☒	☒
HD10700b	14.66	2.10	0.17	0.58	489.47	6.78	☒		☒			☒	☒	☒
HD10700c	34.65	3.04	0.03	0.62	477.87	7.25	☒		☒			☒	☒	☒
HD10700d	96.93	3.71	0.08	0.54	488.62	6.31	☒		☒			☒	☒	☒
fake	268.95	3.33	0.29	0.36	421.39	4.21	☒		☒			☒	☒	☒
fake	3407.46	28.56	0.63	1.64	-129.95	19.17	☒		☒			☒	☒	☒
False detections	9.83	Prot/4				6.31						K=0.54±0.18		
	20.02	Prot/2				10.40	K=0.89±0.45							
	26.13					11.69								K=1
	24.32					8.18							K=0.7±0.7	
	45.03					8.42	K=0.72±0.55							
	47.89					21.86						K=1.87±0.18		
	85.93					9.35							K=0.8±0.7	
System 13	Stellar mass 0.93 [Ms]				RV rms [m/s]	2.07		NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Alpha Cen B	simulated data, 433 data points													
Prot [days]	40.00						19.93		39-40			20	NOT GIVEN	30 and 20
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
NONE							NONE					NONE	NONE	
False detections	10.00	Prot/4				6.94			K=0.69±0.09					
	11.40					15.28						K=1.52±0.20		
	39.05	Prot				8.75			K=0.87±0.13					
	312.50					7.24								K=0.72

Fig. A.5. Same as Fig. A.1 but for RV fitting challenge systems 11, 12 and 13.



							Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
System 14	Stellar mass 0.93 [Ms]			RV rms [m/s]		8.86		NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Corot-7	raw RV Corot-7 (inverse) + noise (5 cm/s) + gamma offset, 170 data points													
Prot [days]	22.30						20.5 - 23		17-22			22.42	NOT GIVEN	23.4
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
Corot-7b	0.85	4.87	0.12	3.44	499.85	5.06	☒		P=0.85559±3e-5 K=4.33±0.69			☒	☒	P=0.853 K=0.84
Corot-7c	3.70	13.29	0.12	5.85	499.32	8.61	P=3.539±2e-3 K=3.4±1.4 T0=54.63±0.16 e=0		P=3.7087±5e-4 K=4.51±0.51			P=3.69±0.03 K=5.08±0.75	P=3.709±0.002 K=3.6±0.7 T0=750.5±0.1 e=0.01	P=3.7 K=0.95
Corot-7d	9.02	17.54	0.00	5.56	496.58	8.18	P=8.896±9e-3 K=7.1±1.7 T0=69.58±0.81 e=0		P=8.898±4e-3 K=6.24±0.97			P=9.04±0.01 K=6.25±0.6	P=9.12±0.01 K=3.7±0.7 T0=718.2±0.3 e=0.01	P=8.9 K=6.58
False detections	5.89	Prot/4				5.68						K=3.86±0.70		
	23.90	Prot				1.59								K=1.08
	47.20					0.77								K=0.52
System 15	Stellar mass 0.93 [Ms]			RV rms [m/s]		7.64		NOT PERFORMED		NOT PERFORMED	NOT PERFORMED			
Calendar: Corot-7	simulated data, 170 data points													
Prot [days]	22.30						11 or 22		11			10.83	NOT GIVEN	11.17
planets	Pp [d]	Mp [Me]	ecc	Kp [m/s]	T0 [d]	K/N = Kp/rms*sqrt(Nobs)								
Corot-7b	0.88	4.87	0.12	3.44	499.85	5.87	☒		P=0087851±3e-5 K=3.42±0.36			☒	☒	☒
Corot-7c	3.63	13.29	0.12	5.85	499.32	9.98	P=3.6024±8e-4 K=5±1.1 T0=55.3±0.2 e=0		P=3.6260±2e-4 K=5.88±0.35			P=3.63±0.01 K=5.61±0.45	P=3.615±0.002 K=2.6±0.7 T0=717.7±0.1 e=0.02	P=3.614 K=1.66
Corot-7d	9.47	17.54	0.00	5.56	496.58	9.49	P=9.36±0.05 K=2.8±1.2 T0=46.2±0.2 e=0		P=9.473±1e-3 K=5.57±0.35			☒	P=10.71±0.02, K=4.99±0.70 T0=716.4±0.5 e=0.16	☒
False detections	5.55	Prot/4				3.82						K=2.24±0.54		
	16.55					5.54						K=3.27±0.48		
	30.38					5.80	K=3.4±0.9							
	57.70					2.06								K=1.21
	166.97					4.61							K=2.7±0.7	

Fig. A.6. Same as Fig. A.1 but for RV fitting challenge systems 14 and 15.

Appendix A: Summary of the planet detection results obtained by the different teams

Figs. A.1, A.2, A.3, A.4, A.5 and A.6 show a summary of the planet detection results obtained by the different teams.

Appendix B: Details about the different algorithmes used by the different teams to analyze the data of the RV fitting challenge

Appendix B.1: Team 1

Appendix B.1.1: The step-by-step approach

Here is a detailed summary of the different steps performed by team 1 to analyze the data of the RV fitting challenge.

1. *Pre-treatment phase: removing long-term trends of stellar origin:* When generating the data of the RV fitting challenge, Dumusque (2016) considered magnetic cycles and their effect on the different observables, i.e., log(R$'_{HK}$), BIS SPAN and full width at half maximum (FWHM) of the CCF. In this case, strong correlation between log(R$'_{HK}$) and FWHM, and between log(R$'_{HK}$) and RV are expected (Lovis et al. 2011; Dumusque et al. 2011b; Lindegren & Dravins 2003). To test those correlations, the Torino team calculated Spearman's rank correlation coefficients and found, in most cases, a very strong correlation between these observables, i.e., $\rho > 0.9$. In the case of significant correlation, i.e., $\rho > 0.5$, team 1 detrended the RV and the log(R$'_{HK}$) using linear fits between log(R$'_{HK}$) and RV, and log(R$'_{HK}$) and FWHM, respectively (Meunier & Lagrange 2013). Detrending the RV and the log(R$'_{HK}$) allows suppressing almost entirely the long-term activity effect induced by magnetic cycles, and therefore leaves only the short-term activity effect, that team 1 further modeled using a GP. In the case of systems 9, 10, and 11, RV were detrended with a linear fit as a function of time, as a significant long-term signal, probably due to a binary, was still visible after correcting for the magnetic cycle effect.

2. *GP regression of the activity index* log(*R$'_{HK}$*)*:* To model log(R$'_{HK}$) with a GP, team 1 used the combination of a *rational quadratic* (RQ) and a *quasi-periodic* (QP) covariance function (Pont et al. 2013; Rasmussen 2006):

$$k_{RQ,QP}(t,t') = A^2 \exp\left(-\frac{sin^2[\pi(t-t')/\theta]}{2L^2}\right) \times \left(1+\frac{(t-t')^2}{2\alpha l^2}\right)^{-\alpha} + \sigma_t^2\delta_{tt'}, \quad (B.1)$$

where t and t' represent epochs of observations, θ the stellar rotation period, σ_t is the uncertainty of the measurement at time t, and $\delta_{tt'}$ is the Kronecker's delta. When there is no a suitable guess about the timescale over which the data are varying, the RQ kernel can be assumed as a reasonable choice because it is intended to model the data by accounting for many different timescales. In fact, it is equivalent to an infinite sum of squared exponential (SE) kernels:

$$k_{SE}(t,t') = h^2 \exp\left[-\frac{(t-t')^2}{2l^2}\right], \quad (B.2)$$

with different length-scales l (Rasmussen 2006), with the inverse squared timescales l^{-2} distributed according to a Gamma distribution with parameters α and $\beta = l^{-2}$. When $\alpha \rightarrow \infty$ the RQ kernel converges to the SE kernel. The function $k_{RQ,QP}(t,t')$ describes the degree of correlation between each pair of measurements at times t and t', reducing to uncorrelated noise, i.e., white noise, when t=t'. This form of covariance function is suitable for data sets spanning a few years. For example, for the long-term photometry data set of HD189733, Pont et al. (2013) discussed the choice of a $k_{RQ,QP}(t,t')$ instead of a simpler exponential decay covariance function to model the observed signal due to stellar activity.

The best-fit values of the covariance function hyper-parameters were obtained using an MCMC analysis. Initial guess for hyper-parameter θ was derived by performing a periodogram analysis with the Generalized Lomb-Scargle algorithm (GLS, Zechmeister & Kürster 2009). After a burn-in phase, typically consisting of 1500 steps per chain, team 1 maximized the following log-likelihood function:

$$\ln \mathcal{L} = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(det\,\mathbf{K}) - \frac{1}{2}\underline{r}^T \cdot \mathbf{K}^{-1} \cdot \underline{r}, \quad (B.3)$$

where $\mathbf{K}$ is the covariance matrix built from the covariance function in Equation B.1, and $\underline{r}$ is the detrended log(R$'_{HK}$). The best-fit estimates of the hyper-parameters, inferred from their posterior distributions, were used as guess values for the subsequent modeling of the RVs, as explained below. Team 1 derived stellar rotation periods from the posterior distribution of θ.

3. *First identification of significant signals; GLS analysis of the RV time series:* The Torino team applied the GLS algorithm to search for significant signals in the original RVs. Team 1 explored the frequency space below the Nyquist frequency and estimated peak significance using p-values determined through a bootstrap with replacement analysis consisting of 10'000 random shuffles of the data by keeping the time stamps fixed. Team 1 selected for further considerations only peaks with p-values<10^{-3} (0.1%), except for systems 14 and 15, because of the lower number of data points.

Team 1 iteratively removed sinusoidal fits from the data, with periodicity corresponding to the periodogram peaks, and obtained guess values for the orbital period of the candidate Keplerian signals. Team 1 looked at the window function to discard aliases.

As a general rule, only significant RV signals with period shorter than the data time span were considered, except for system 7, where a signal with a longer period than the data time span was modeled with a Keplerian in the global fit, despite the inability of characterizing reliably the potential orbit. Moreover, the approach followed by the team was conservative, i.e. aimed at avoiding as much false positives as possible, favoring the analysis of signals with the highest semi-amplitudes.

4. *RV model and MCMC analysis:* After the analysis of the GLS periodogram, and the identification of significant signals that could be due to planetary candidates, the Torino team performed a global fit of the RVs with a model consisting of Keplerian orbits and correlated noise, to account for short-term stellar activity signals. This correlated noise is modeled using the GP covariance function seen in Equation B.1. The training of the GP on the log(R$'_{HK}$) gives initial guess for the GP hyper-parameters used when fitting the RVs. Doing so, team 1 assumes that short-term activity signals seen in RV and log(R$'_{HK}$) have a similar covariance.

The **general** Keplerian model fitted to the RVs is described by:

$$\begin{aligned}\Delta RV_{Kep}(t_i) &= \sum_{j=1}^{n_{planet}} \Delta RV_{Kep,j}(t_i) + \gamma \\ &= \sum_{j=1}^{n_{planet}} K_i \left[\cos(\nu(t_i, T_{0j,peri.}, P_j) + \omega_j) + e_j \cos(\omega_j)\right] \\ &+ \gamma. \qquad \text{(B.4)}\end{aligned}$$

Instead of fitting e_j and ω_j separately, team 1 introduced:

$$C_i = \sqrt{e_i} \cdot \cos\omega_i \qquad S_i = \sqrt{e_i} \cdot \sin\omega_i, \qquad \text{(B.5)}$$

to uniformly sample the eccentricity parameter space (Ford 2006). Short-term stellar activity is fitted simultaneously by the GP applied to the RV residuals obtained by subtracting the Keplerian model from the raw RV data. The best-fit is found by maximizing the log-likelihood seen in Equation B.3. Note however that in this case the array $\underline{r}$ represent the RV residuals.
The MCMC analysis used a number of random walkers, typically in the range 50-150, and was characterized by a burn-in phase, in general consisting of 1500 steps. For each fitted parameter the team adopted non-informative, uniform priors. The hyper-parameters of the covariance function were constrained within a range with reasonable finite lower and upper limits comprising the best-fit estimates found with the analysis of log(R$'_{HK}$), except for the semi-amplitude term A of the covariance function, which for the RVs is necessarily different from that of log(R$'_{HK}$) and was only imposed to be positive. No upper limits were fixed for $T_{0j,peri.}$ and semi-amplitude K, while the orbital periods were constrained over ranges of reasonable semi-amplitude centered on the guessed values obtained from the GLS periodogram analysis. To test the convergence of the different chains, team 1 used the Gelman-Rubin statistics as described in Ford (2006). The best estimate of each parameter is derived using the median of its posterior distribution, with their asymmetric uncertainties derived from the 16^{th} and 84^{th} percentile ($1-\sigma$ uncertainty).

5. *Model selection:* The GP analysis requires a significant computational effort. Due to the relatively short timescale of the RV fitting challenge, team 1 could only test a limited number of different models for each system. Team 1 performed a Bayesian selection based on the truncated posterior mixture (TPM) method described in Tuomi & Jones (2012). In some cases, team 1 tested models with an equal number of planets, but fixing or not the eccentricities to zero. In few other cases, when signals could be of planetary or stellar nature, team 1 compared models with a different number of planets, limiting the analysis to circular orbits. Finally, when the Bayesian analysis showed to be inconclusive, team 1 selected the model with fewest parameters, following the "Occam razor" principle. Note however this was not the case for system 15, because the three candidate signals appear to be well modeled by a sinusoid, even if the true nature of one Keplerian was flagged as doubtful.

Appendix B.1.2: Algorithms and Tools

Here is a list of the different tools that team 1 used to perform the analysis:

- Spearman's rank correlation coefficients were evaluated with the R_CORRELATE function, which is part of the IDL library.
- Linear fits (log(R$'_{HK}$) vs. FWHM and log(R$'_{HK}$) (or time) vs. RV), and estimation of the GP hyper-parameters and Keplerian parameters were performed using the publicly available EMCEE Affine Invariant Markov Chain Monte Carlo Ensemble sampler, developed by Foreman-Mackey et al. (2013) (see also http://dan.iel.fm/emcee/current/).
- The GP regression analysis was performed with the George Python library developed by Foreman-Mackey (2015) and Ambikasaran et al. (2014), and publicly available at http://dan.iel.fm/george/current/
- The search for sinusoidal modulations in the log(R$'_{HK}$) and RV data were performed with the Generalized Lomb-Scargle (GLS) algorithm developed by Zechmeister & Kürster (2009).

Appendix B.2: Team 3

Appendix B.2.1: The step-by-step approach

Here is a detailed summary of the different steps performed by team 3 to analyze the data of the RV fitting challenge.

1. *First identification of significant signals* Team 3 first analyzed all time series of the RV fitting challenge using a likelihood-ratio periodogram including a first order moving average, i.e. a correlation dependence of each data point with their preceding neighbor (see last term of Equation B.6). Significant signals in activity observables, i.e. log(R$'_{HK}$), BIS SPAN and FWHM, were associated to stellar activity effect, and significant signal in the RVs were associated to potential planetary candidates if a similar signal was not seen in the activity observables.
2. *RV model and MCMC analysis:*
   To fit the RVs, team 3 used a model composed of:
   - one or several Keplerians,
   - a polynomial function up to the 2nd order to fit any long-term trend due to distant companions,
   - linear correlation with activity observables, to account for the effect of magnetic cycles,
   - a Gaussian white noise ϵ_i with zero mean and variance $\sigma_i^2 + \sigma^2$, where σ_i is given by the data and σ is a free parameter to account for additional instrumental white noise,
   - and a first order moving average component with exponential smoothing accounting for the intrinsic correlations in the RVs.

   In this case, the RV model that team 3 used can be described as:

$$\begin{aligned}\Delta RV_{tot}(t_i) &= \Delta RV_{Kep}(t_i) + \epsilon_i + Cte + \alpha\, t_i + \beta\, t_i^2 + \\ &+ c_{01}\,\text{BIS SPAN} + c_{02}\,\text{FWHM} + c_{03}\,\log(R'_{HK}) \\ &+ \phi\left[\Delta RV_{tot}(t_{i-1}) - \Delta RV_{Kep}(t_{i-1})\right] exp^{\frac{t_{i-1}-t_i}{\tau}} \text{(B.6)}\end{aligned}$$

   where Cte, α, β, and c_{01}, c_{02}, c_{03} are the free parameters of the polynomial fit and to account for correlation with activity observables, respectively. The parameter ϕ measures the strength of the correlation between consecutive measurements, τ is the correlation timescale, and ΔRV_{Kep} is the Keplerian model described in Equation B.4. Team 3 analyzed the data of the RV fitting challenge using adaptive-Metropolis Markov Chain Monte Carlo samplings (Haario et al. 2001), maximizing the following log-likelihood:

$$\ln \mathcal{L} = -\frac{1}{2}\ln(2\pi(\sigma_i^2+\sigma^2)) - \left[\frac{(RV(t_i)-\Delta RV_{tot}(t_i))^2}{2(\sigma_i^2+\sigma^2)}\right]. \quad (B.7)$$

Note that in their analysis, team 3 fixed the correlation timescale τ to 4 days, as this value seemed to give good results on previous analysis of HARPS high-cadence data.

3. *Model selection:* Team 3 used the MCMC samplings to calculate the integrated likelihoods and obtain Bayesian estimates for model probabilities when assuming equal prior probabilities. Team 3 applied the method based on the mixture of posterior and prior densities, described in Newton & Raftery (1994), to compare between models.
   When detecting a Keplerian signal, team 3 interpreted it as existing if:
   (a) including the signal in the model increased the model probability by a factor of 1000;
   (b) the corresponding signal was unique in the period space such that there were no other periods with posterior density (i.e. local maxima) in excess of 0.1% of the global maximum;
   (c) and the period and the semi-amplitude of the signal were well constrained from above and below, and the semi-amplitude, in particular, statistically significantly different from zero.
   In addition to these criteria (Tuomi et al. 2014), team 3 interpreted the signal to be related to activity-induced variations if any of the activity indices showed a significant signal at the same period.

Appendix B.2.2: Algorithms and Tools

Here is a list of the different tools that team 3 used to perform the analysis:

- Moving average (Baluev 2013; Tuomi et al. 2013)
- Adaptive Metropolis MCMC algorithm (Haario et al. 2001)
- Model selection using posterior and prior mixture (Newton & Raftery 1994)

Appendix B.3: Team 4

Appendix B.3.1: The step-by-step approach

Here is a detailed summary of the different steps performed by team 4 to analyze the data of the RV fitting challenge.

1. *First identification of significant signals* To look for significant signals in the RVs, P. Gregory corrected the RVs from the effect of the magnetic cycle using the log(R$'_{HK}$)-RV correlation, and considered any signal in a GLS periodogram with p-values smaller than 0.01. Note that contrary to team 1 and 3, P. Gregory estimated p-values analytically.

2. *RV model and MCMC analysis:* Once the first peak is detected, P. Gregory runs a Bayesian Fusion MCMC (Gregory 2013) analysis to find the best parameters for the signal, and look for extra signals in the residuals using a GLS periodogram. To fit the RV data, P. Gregory used the following model:

$$\begin{aligned}\Delta RV_{tot}(t_i) &= \sum_{j=1}^{n_{signals}} \Delta RV_{Kep,j}(t_i) \times exp\left[-\frac{(t_i-t_{a,j})^2}{2\tau_j^2}\right] \\ &+ \gamma + \epsilon_i + a\,\log(R'_{HK}), \qquad (B.8)\end{aligned}$$

   where $n_{signals}$ is the number of significant signals in the data independent of their nature, i.e. planetary or stellar activity, $t_{a,j}$ and τ_j are the center and timescale of the apodized window of signal j, and a is a free parameter to account for a possible correlation between RV and log(R$'_{HK}$).

3. *Distinguishing planetary from stellar short-term activity signals*: If the signal j is induced by a planet, the apodized term $exp\left[-\frac{(t_i-t_{a,j})^2}{2\tau_j^2}\right]$ will essentially be constant over the duration of the data because the semi-amplitude of the signal is constant. In this case τ will be greater than the time span of the data. On the other hand, the apodized term will strongly vary as a function of time in the case of stellar activity, due to appearance and disappearance of active regions on the stellar surface. In this case τ will be smaller than the time span of the data.
   To help distinguish between planetary and stellar activity signals, P. Gregory used a second approach based on the FWHM. The FWHM is first corrected for the effect of the magnetic cycle using the log(R$'_{HK}$)-FWHM correlation. Then any significant signal found in either the initial corrected RVs or the later stage RV fit residuals, that coincides with a significant signal in the corrected FWHM, is associated with stellar activity.
   At each stage in the RV analysis, the number of Keplerian signals was extended to include the period with the highest peak in the periodogram of the residuals from the previous model as a starting point for a Bayesian Fusion MCMC exploration in parameter space.

4. *Model selection:*
   Model comparison was based on Bayes factors computed using the Nested Restricted Monte Carlo (NRMC) estimator (Section 2.2 in Gregory 2016, Gregory 2013, Gregory & Fischer 2010, and in more details in Section 1.6 of the *Supplement to Bayesian Logical Data Analysis for the Physical Sciences* available in the resources section of the Cambridge University Press website for P. Gregory's Textbook *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support*).

Appendix B.3.2: Algorithms and Tools

Here is a list of the different tools that P. Gregory used to perform the analysis:

- GLS periodogram (Zechmeister & Kürster 2009) to look for significant signals,
- Bayesian Fusion MCMC (Gregory 2013) to explore parameter space,
- and Nested Restricted Monte Carlo estimator (Gregory 2013; Gregory & Fischer 2010) to compare between different models.

Appendix B.4: Team 7

Appendix B.4.1: The step-by-step approach

1. *DFT of all observables and cleaning from the spectral window:* To move from the time-domain to the frequency domain for the RV, BIS SPAN, FWHM and log(R$'_{HK}$), team 7 used, like team 6, a DFT. In the frequency domain, any uneven data as a function of time will be affected by the sampling of the signal. To reduce the effect of sampling, team 7 used the CLEAN algorithm (Roberts et al. 1987). The team

got as a result a cleaned DFT (CDFT) for all time series (see upper-left panel of Fig. 10). Particular attention is made to carefully select the *gain* parameter, in order to remove as much spurious frequency peaks as possible without removing significant signals.

2. *Removing stellar signals:* To remove stellar signals, team 7 subtracted the CDFT of all the activity observables (BIS SPAN, FWHM, log(R'_{HK})) from the CDFT of the RVs. The obtained CDFT, exempt of stellar signals, is used to create a *pass-planet* filter in the frequency domain (see upper-right panel of Fig. 10). By applying this filter to the DFT of the RVs, team 7 obtained RVs in the frequency domain that are cleaned from any stellar signals. At this stage, any significant signal in those filtered RVs should be due to planets. Team 7 selected the highest peak and recorded its period, semi-amplitude and phase.
3. *Fitting planets:* To fit the planetary signal found at the previous step with a Keplerian, team 7 first transformed the filtered RVs back into the time-domain using an inverse DFT, and then used the RVLIN package (Wright & Howard 2009) to fit the planetary signal, fixing the initial parameters to what was previously found.
4. *Iterative process:* Once team 7 found the best-fit for the planet inducing the strongest RV signal, it removed the signal from the raw RVs. Team 7 applied the CLEAN algorithm on the residual RVs and restarted the whole process from the beginning. To be conservative and prevent the detection of false positives, team 7 stopped when the semi-amplitude of the signal found in the filtered residual RVs was smaller than the average uncertainty of the RV measurements.

This method presents the advantage of being independent of any model to account for stellar signals, and it is computationally very fast compared to Bayesian methods. However this technique presents the disadvantage that planetary signals are only fitted one by one, thus it is difficult to constrain orbital parameters such as eccentricity and argument of periastron. In addition, the imperfect removal of a signal causes the introduction of spurious frequencies that can lead to false detections. In addition, lack of statistics forced team 7 to stop at a S/N level of 1 (S/N once the data have been filtered from stellar signals), preventing the detection of small S/N planetary signals. Finally significant signal around one day were not considered, to avoid strong residuals of the spectral window not fully cancelled by the cleaning process.

Appendix B.4.2: Algorithms and Tools

Here is a list of the different tools that team 7 used to perform the analysis:

- Discrete Fourier Transform (as described in Roberts et al. 1987)
- CLEAN algorithm to remove sampling effects (Roberts et al. 1987)
- RVLIN package to fit Keplerians (Wright & Howard 2009)