Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **505,** 1699–1717 (2021)
Advance Access publication 2021 May 12

https://doi.org/10.1093/mnras/stab1323

Separating planetary reflex Doppler shifts from stellar variability in the wavelength domain

A. Collier Cameron [1,2]★ E. B. Ford [2,3,4,5] S. Shahaf [6] S. Aigrain,[7] X. Dumusque,[8] R. D. Haywood,[9,10]† A. Mortier [11,12] D. F. Phillips,[9] L. Buchhave [13] M. Cecconi,[14] H. Cegla [8,15] R. Cosentino,[14] M. Crétignier [8] A. Ghedina,[14] M. González,[14] D. W. Latham,[9] M. Lodi,[14] M. López-Morales,[9] G. Micela,[16] E. Molinari,[17] F. Pepe,[8] G. Piotto,[18] E. Poretti [14] D. Queloz,[11] J. San Juan [14] D. Ségransan,[8] A. Sozzetti [19] A. Szentgyorgyi,[9] S. Thompson,[11] S. Udry[8] and C. Watson[20]

Affiliations are listed at the end of the paper

Accepted 2021 May 2. Received 2021 April 28; in original form 2020 October 13

ABSTRACT

Stellar magnetic activity produces time-varying distortions in the photospheric line profiles of solar-type stars. These lead to systematic errors in high-precision radial-velocity measurements, which limit efforts to discover and measure the masses of low-mass exoplanets with orbital periods of more than a few tens of days. We present a new data-driven method for separating Doppler shifts of dynamical origin from apparent velocity variations arising from variability-induced changes in the stellar spectrum. We show that the autocorrelation function (ACF) of the cross-correlation function used to measure radial velocities is effectively invariant to translation. By projecting the radial velocities on to a subspace labelled by the observation identifiers and spanned by the amplitude coefficients of the ACF's principal components, we can isolate and subtract velocity perturbations caused by stellar magnetic activity. We test the method on a 5-yr time sequence of 853 daily 15-min observations of the solar spectrum from the HARPS-N instrument and solar-telescope feed on the 3.58-m Telescopio Nazionale Galileo. After removal of the activity signals, the heliocentric solar velocity residuals are found to be Gaussian and nearly uncorrelated. We inject synthetic low-mass planet signals with amplitude $K = 40$ cm s^{-1} into the solar observations at a wide range of orbital periods. Projection into the orthogonal complement of the ACF subspace isolates these signals effectively from solar activity signals. Their semi-amplitudes are recovered with a precision of ~ 6.6 cm s^{-1}, opening the door to Doppler detection and characterization of terrestrial-mass planets around well-observed, bright main-sequence stars across a wide range of orbital periods.

Key words: methods: statistical – techniques: radial velocities – techniques: spectroscopic – Sun: photosphere – planets and satellites: general.

1 INTRODUCTION

For decades, Doppler spectroscopy has been one of the most productive methods to discover and characterize exoplanets. Improvements in the precision, wavelength calibration, and stability of high-resolution échelle spectrographs have allowed exoplanet surveys to probe planets with radial velocity (RV) semi-amplitudes of just $\sim$1 m s^{-1}. New generations of spectrographs such as CARMENES (Quirrenbach et al. 2014), ESPRESSO (Mégevand et al. 2014), EXPRES (Jurgenson et al. 2016), HARPS-3 (Thompson et al. 2016), HPF (Ninan et al. 2018), and NEID (Schwab et al. 2016) are being designed and commissioned with improved resolution, spectral coverage, wavelength calibration, and stabilization systems (Wright & Robertson 2017). Recently, ESPRESSO has achieved 30 cm s^{-1} precision per RV observation on Proxima Cen (Suárez Mascareño et al. 2020), and EXPRESS has achieved 58 cm s^{-1} precision per RV observation on HD 3651 (Brewer et al. 2020).

Even with present instruments, the ability of spectroscopic surveys to detect and characterize low-mass planets is often limited by stellar variability and the stability of the wavelength calibration, rather than photon noise or instrumental errors (e.g. Saar & Donahue 1997; Queloz et al. 2001; Haywood et al. 2014). The purpose of this study is to devise a practical new approach to measuring stellar RVs in a way that mitigates the errors due to line-shape changes caused by stellar variability. To achieve this, we make use of the fact that changes in the shape of spectral lines may influence the apparent RV, but changes in the range rate (the first derivative with respect to time of the distance from the star's centre to the Solar-system barycentre) induce only a shift and do not affect the line shape or depth. Related approaches exploiting profile-shape changes of even and odd character to disentangle shifts from activity have been published recently by Zhao & Tinney (2020) and Holzer et al. (2020) while de

★ E-mail: acc4@st-andrews.ac.uk
† NASA Sagan Fellow.

Beurs et al. (2020) have employed a neural-network machine learning to relate activity-related RV shifts to cross-correlation function (CCF) profile-shape changes in the same solar data set examined here.

One feature of the method we present is that it makes use of existing data products, i.e. the CCFs between the observed spectra and a digital mask (see Section 2), and the RVs derived from them. The CCF $C(v, t)$ is a function of both barycentric velocity v and time t. The temporal variability of the CCF includes both Doppler shifts and line shape changes. The derivatives of the CCF with respect to velocity [e.g. $C^{'}(v, t)$, $C^{''}(v, t)$] are also functions of velocity and time.

Our long-term goal is to improve the detection sensitivity and robustness of spectroscopic planet searches. As a specific objective towards that goal, we aim to devise efficient, shift-invariant metrics that can contribute to characterizing the detailed line-profile shape at each epoch. We describe the building blocks for our new method in Section 2. Then, in Section 3, we propose a new algorithm to compute such metrics, based on a novel combination of existing data products and techniques. We apply these metrics as predictors for the contribution of stellar variability to the apparent Doppler shift and infer a cleaned velocity time series. In Section 4, we verify and validate the method based on injection and recovery tests using solar observations. While our data-driven method makes no assumptions about the physical origin of the stellar variability, the tests in this paper focus on magnetic activity of the Sun. Finally, we discuss the implications of our work for future spectroscopic planet surveys in Section 5.

2 CROSS-CORRELATION FUNCTION AND ITS AUTOCORRELATION

The HARPS, HARPS-N, and ESPRESSO data reduction systems (DRS) derive RVs from stellar échelle spectra by computing the CCF of the spectrum with a digital line mask matched approximately to the spectral type of the target star (Baranne et al. 1996; Pepe et al. 2002).

The data presented in this paper were re-analysed with the ESPRESSO data-reduction pipeline (Pepe et al. 2021). The reduction procedure differs from that used for previous analyses of HARPS-N data (cf. Collier Cameron et al. 2019) in several important respects. These are described in detail by Dumusque et al. (2020). The wavelength scale is derived using a new line list tailored for the primary HARPS-N ThAr calibration lamp. A single master wavelength calibration is used for all observations. The wavelength of each pixel in the extracted spectrum is corrected for drift relative to the master calibration using the daily wavelength calibrations with the primary ThAr lamp, and the simultaneous reference source (either a secondary ThAr lamp or a Fabry–Perot). Prior to cross-correlation with a synthetic mask, the pixel wavelengths are transformed to the reference frame of the Solar system barycentre along the line of sight to the target. The CCFs are estimated on a common grid of pixels at uniformly spaced intervals of $h = 0.82$ km s^{-1} in velocity space, using a blaze-corrected, inverse-variance weighted cross-correlation with a mask of line wavelengths and weights appropriate to the target spectral type. The CCF sampling interval matches the velocity increment per physical CCD pixel in the instrument. The velocity scale of the CCF is in the reference frame of the Solar system barycentre, with the drift correction applied.

Being the cross-correlation of a stellar absorption spectrum with a positive line mask, the CCF computed by the DRS resembles a single stellar absorption line, with a pseudo-continuum level which is accurately and consistently normalized to unity. The resulting CCF profile is fitted with one minus a Gaussian function described by three parameters: central velocity (v, relative to the mask), full width at half-maximum depth (FWHM) and central line depth as a fraction of the pseudo-continuum. Similarly, the bisector inverse slope (BIS) of the profile (Queloz et al. 2001) is recorded as a measure of profile asymmetry.

Stellar activity compromises the fidelity of this method of RV measurement. The contrast between bright convective upflows in photospheric granules and cooler downflows in intergranular lanes imposes an inherent asymmetry on the line profile (Dravins, Lindegren & Nordlund 1981). Magnetic activity, the finite lifetime of the granulation pattern and P-mode oscillations all cause the already non-Gaussian shape of the observed CCF to vary with time. As the star rotates, dark star-spots produce line-absorption deficits which migrate across the profile from blue to red, introducing time-varying amounts of skew and kurtosis into the spectral-line shapes (Saar & Donahue 1997; Dumusque, Boisse & Santos 2014). These magnetic regions alter the local convective velocity and line-profile asymmetry, as well as the local brightness weighting of the stellar rotation profile (Meunier, Lagrange & Desort 2010). In faculae-dominated stars like the Sun, magnetic suppression of granular convection in faculae causes even stronger time-varying profile asymmetries than sunspots, combining rotational Doppler shifts with foreshortening-dependent changes in the radial-tangential velocity field (Meunier et al. 2010; Cegla et al. 2019).

Several previous studies (e.g. Aigrain, Pont & Zucker 2012; Dumusque et al. 2012; Rajpaul et al. 2015) have explored whether the estimated velocities could be improved by decorrelating with other measurements, such as the FWHM or BIS. While the FWHM and BIS contain information about the CCF shape, they are not sufficient to describe the detailed changes in the CCF. In order to make use of all information contained in the spectra, some studies have suggested analysing stellar variability by applying a principal-component analysis (PCA) to the observed spectroscopic time series (e.g. Davis et al. 2017; Jones et al. 2017). These methods are promising on simulated data sets, but applying this approach to actual observations is challenging due to details of the spectrograph and its calibration process. In this paper, we apply the PCA approach to the CCF instead of the raw spectrum. This approach leverages extensive investments in developing a robust pipeline to measure the CCF. Of course, analysing only the CCF does reduce the total information content of the spectrum. We offer suggestions for how the method could be generalized to extract more information in Section 5.

To illustrate the impact of stellar variability on the CCF and to provide a test data set, we created a sequence of 886 daily solar CCFs from the HARPS-N solar telescope (Cosentino et al. 2014; Dumusque et al. 2015; Phillips et al. 2016), spanning the period from 2015 July 29 to 2020 March 6. The resulting distortions of the CCF profile, described in detail by Collier Cameron et al. (2019), are seen most clearly in the residuals $R(v_i, t_j) = \mathbf{C}(v_i, t_j) - \langle \mathbf{C}(v_i) \rangle$ obtained by subtracting the time-averaged profile of the entire 5-yr sequence from each CCF in the sequence (Fig. 1, left-hand panels).

Full details of the HARPS-N solar observations are given by Collier Cameron et al. (2019), who used a Gaussian mixture model to assign a probability to each observation that it is unaffected by uneven transparency across the solar disc and corrected the velocities for differential extinction. Here we select only those observations with (1) probabilities greater than 99 per cent of being good and (2) with velocity corrections for differential extinction less than 10 cm s^{-1}. In summer these conditions are satisfied for up to 4 h d^{-1}, and in winter for up to 2 h d^{-1}.

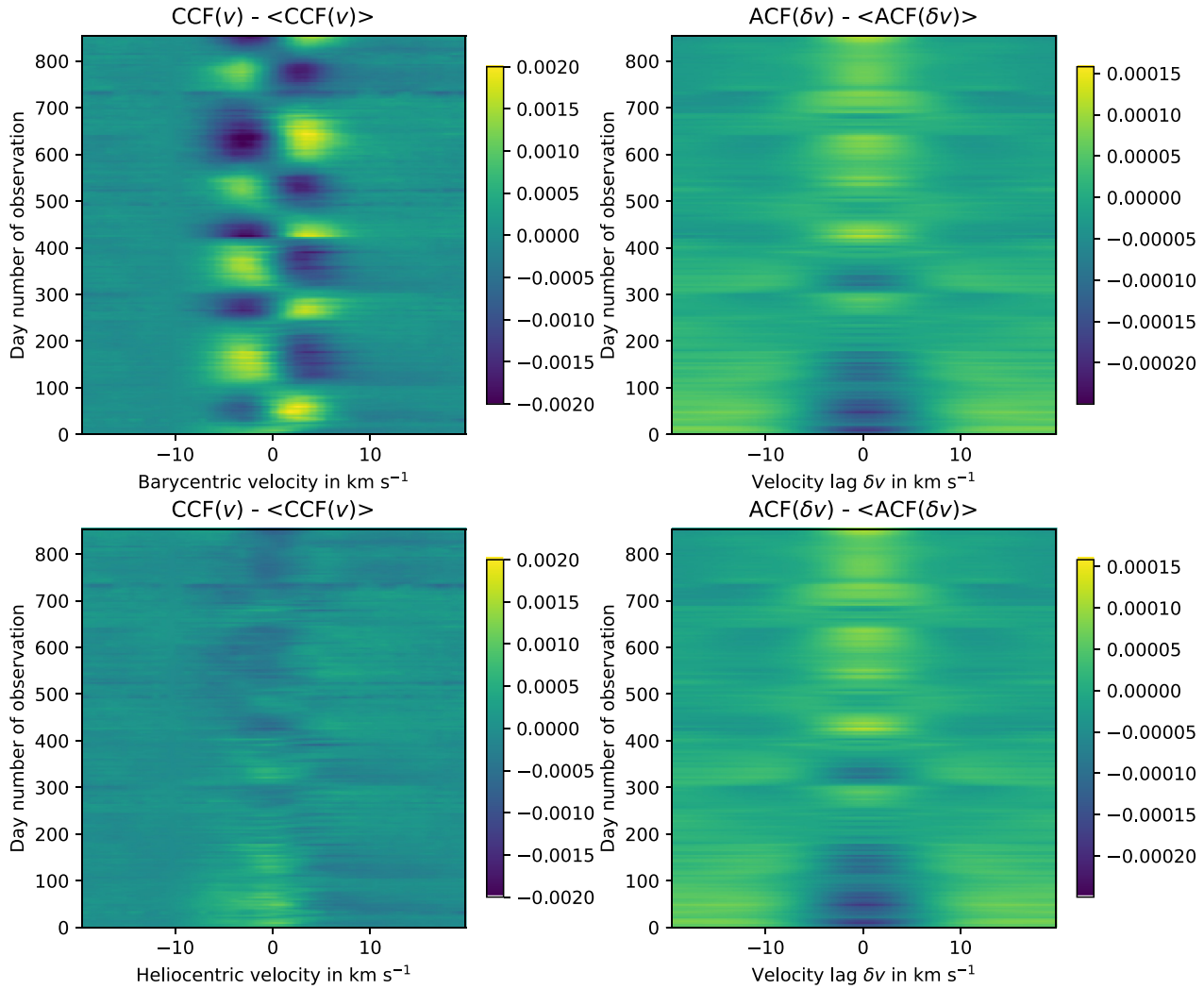

Figure 1. Time sequence of residual solar CCFs and the ACFs of the CCFs, spanning the period from 2015 July to 2020 March. Each row represents 1 d of observation. Days on which no data were obtained are not shown, so although time increases up the vertical axis, the time-scale is not linear. The barycentric residual CCF (upper left) shows the solar reflex motion about the Solar system barycentre. Dominated by Jupiter, it has a semi-amplitude of 12 m s^{-1} at Jupiter's synodic period of 398 d. The residual CCF in the heliocentric frame (lower left) shows a secular change in line depth, deepening as the solar activity level declines over the period of observation. The residual ACF (right panels) shows temporal variability that is correlated with that of the CCF (compare left and right panels), but is unchanged by Doppler shifts being applied to the CCF (compare upper right and lower right panels).

Our 5-min exposure time is dictated by the need to average out solar *p*-mode oscillations. Light from the Sun is gathered by a 76 mm objective lens of 200 mm focal length, scrambled in an integrating sphere and fed into the spectrograph via an optical-fibre feed to the calibration unit and a neutral-density filter which attenuates the throughput by a further factor 15 (Phillips et al. 2016). The overall throughput is comparable to night-time HARPS-N exposures for a star of magnitude 5.5, for which we use the same exposure time in good seeing without saturating the detector. This gives signal-to-noise ratio (SNR) $\simeq$ 350 or so in the continuum in échelle order 60, which translates to SNR $\simeq$ 5000 in the CCF. For stellar RV observations we use 15-min blocks of contiguous exposures to mitigate the effects of *p*-mode oscillations. Within the windows that satisfy our selection criteria each day, we select at random a set of three contiguous CCFs to mimic an RV observation of a $V = 5.5$ star, and form a weighted average CCF using the square of the mean SNR of the CCF as the weighting factor. Thus, we anticipate that our test data set is dominated by solar magnetic activity, granulation and/or instrumental issues. We encourage future studies to investigate how well the algorithm can mitigate spectral variability on shorter time-scales.

2.1 Translation to the heliocentric frame

The HARPS-N DRS was designed primarily for stellar RV measurement, computing the CCF in the reference frame of the Solar-system barycentre in the direction of the target, as described above. As a result, the instantaneous CCF is Doppler-shifted by the component of the Sun's barycentric motion in the observer's direction, as is apparent from the upper-left panel of Fig. 1. To convert CCFs derived from solar spectra to the heliocentric reference frame, the CCF profiles must be shifted by the line-of-sight component ϵ of the Sun's reflex motion about the barycentre.

The barycentric to heliocentric velocity corrections were computed using the JPL HORIZONS software of Giorgini et al. (1996).

We use a Taylor-series approximation to eliminate the solar barycentric motion from the CCF time series, adding scaled derivatives of the instantaneous profile shape at time t_j to the barycentric CCF:

$$\mathbf{C}(v_i+\epsilon,t_j)=\mathbf{C}(v_i,t)+\epsilon\mathbf{C}'(v_i,t_j)+\frac{\epsilon^2}{2}\mathbf{C}''(v_i,t_j)+\mathcal{O}\epsilon^3. \quad (1)$$

The derivatives are calculated numerically using equations (A1) and (A2) (see Appendix A). The differences between neighbouring CCF values are substantially less than unity. The barycentric to heliocentric velocity correction is never greater than ± 14.7 m s^{-1}, which is much less than the $h = 820$ m s^{-1} sampling interval between neighbouring CCF elements. The truncation error in equation (1) is therefore significantly less than $\epsilon^3/12h^3 \simeq 2.6 \times 10^{-8}$.

We validated the fidelity of the shift by calculating heliocentric velocities from the shifted profiles using the methodology of Appendix B. We computed barycentric velocities from the original un-shifted CCFs by the same method, then applied the barycentric to heliocentric velocity correction. The RMS scatter of the difference between the two resulting sets of heliocentric velocities was 0.008 m s^{-1}. The RMS scatter in the difference between the heliocentric velocities calculated from the shifted CCFs and the DRS velocities transformed to the heliocentric frame was 0.044 m s^{-1}.

The resulting CCF time series, shown at lower left in Fig. 1, is effectively that of a star with no planets. This image shows that the form of the solar CCF is far from static, with dramatic changes in profile shape taking place on all time-scales from days to years.

2.2 Autocorrelation of CCF and shift invariance

Since our goal is to separate the effects of genuine dynamical Doppler shifts from spurious shifts caused by line-shape changes, we aim to characterize changes in CCF profile shape in a way that is invariant to translation in velocity space. The autocorrelation function (ACF) of the CCF has the desired property that it is invariant to translation (Adler & Konheim 1962). The ACF $A(\delta v)$ is the expectation value of the vector cross-product of the CCF with itself at a sequence of lags δv:

$$A(\delta v)=\mathrm{E}(\mathrm{CCF}(v)\cdot\mathrm{CCF}(v+\delta v)). \quad (2)$$

For the sake of brevity, in this manuscript, we refer to the ACF of the CCF as simply 'the ACF'.

The CCF series has m rows representing individual observations and l columns representing individual velocity bins. We compute the ACF of every CCF in the time series. This is done by sequentially shifting the CCF by integer numbers of velocity steps, or CCF 'pixels', modulo the number l of elements in the CCF, and co-multiplying by the unshifted CCF:

$$A(v_i,t_j)=\sum_{i'=1}^{l}C(v_{i'},t_j)C(v_{\mathrm{Mod}(i'-i,l)},t_j). \quad (3)$$

This set of circular shifts and cross-products is repeated for every observation, to obtain a time sequence of ACFs. The m rows of the ACF time series have the same length l as the original CCFs and are normalized to a mean value of unity. There is sufficient pseudo-continuum to either side of the dip in the CCFs to ensure that this circular autocorrelation procedure is sensitive to long-range correlations while minimising edge effects.

The right-hand panels of Fig. 1 show that, despite the strong differences between the residual CCFs in the barycentric and heliocentric frames, their ACFs are very similar. The similarity is, however, only approximate. The autocorrelation domain is not infinite, and the circular shift method employed is vulnerable to edge effects if the CCFs are strongly shifted. In Fig. 2 we compare the column variances of the barycentric ACF time series with the column variances of the residuals obtained by subtracting the barycentric from the heliocentric ACF time series. We find that the temporal variance of the residual ACF is between 2.5 and 4.5 orders of magnitude smaller than the temporal variance of either the barycentric or heliocentric ACF, at every point in the profile. We conclude that for the purposes of this study, the ACF is *effectively* invariant to the solar reflex motion around the Solar-system barycentre.

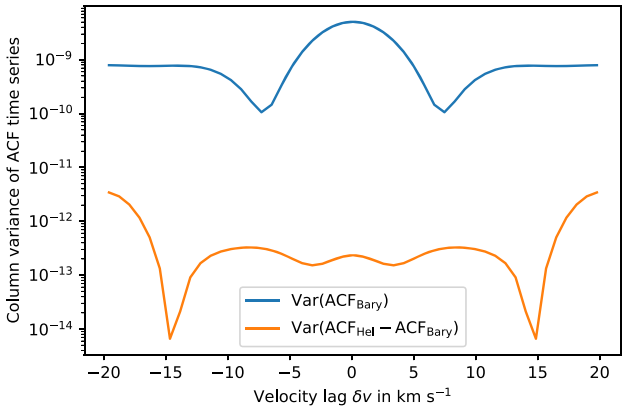

Figure 2. Column variances of the ACF time series derived from CCFs in the barycentric frame, compared to column variances of the difference between the heliocentric and barycentric ACFs. The column variance of the difference is 2.5–4.5 orders of magnitude smaller than the column variance of either ACF time series.

2.3 Principal-component analysis of stellar variability

The principal modes of variability in the CCF can be isolated by calculating the singular-value decomposition (SVD) of the ensemble of CCFs:

$$\mathbf{C}(v_i,t_j)=\langle\mathbf{C}(v_i)\rangle+\mathbf{U}_C(t_j)\cdot\mathrm{diag}(\mathbf{S}_C)\cdot\mathbf{P}_C(v_i). \quad (4)$$

The same method yields the principal components of the ensemble of ACFs:

$$\mathbf{A}(\delta v_i,t_j)=\langle\mathbf{A}(\delta v_i)\rangle+\mathbf{U}_A(t_j)\cdot\mathrm{diag}(\mathbf{S}_\mathrm{A})\cdot\mathbf{P}_\mathrm{A}(\delta v_i). \quad (5)$$

The diagonal matrices $\mathbf{S}_C$ and $\mathbf{S}_A$ list the singular values (eigenvalues) of the principal components in decreasing order. Fig. 3 (top) shows the eigenvectors $\mathbf{P}_{C,k}(v_i)$ and $\mathbf{P}_{A,k}(v_i)$ (also known as loadings) of the leading ($k = 1\cdots6$) principal components of the heliocentric CCF (left) and ACF (right) time series. They represent orthonormal modes of profile variability.

The columns of $\mathbf{U}_{C,k}(t_j)$ and $\mathbf{U}_{A,k}(t_j)$ define an orthonormal basis in the time domain. Each column comprises the coefficients (also known as scores) that define the temporal behaviour of the corresponding eigenvector. Fig. 3 (bottom) shows the scores of the leading 6 eigenvectors for all the individual observations in the time series ensemble, plotted against barycentric Julian date. The ACF is calculated in such a way that it is an even function, so its eigenvectors are also even functions. Those of the CCF display a mix of even and odd character. Nonetheless, there are strong similarities in the temporal behaviours of their scores.

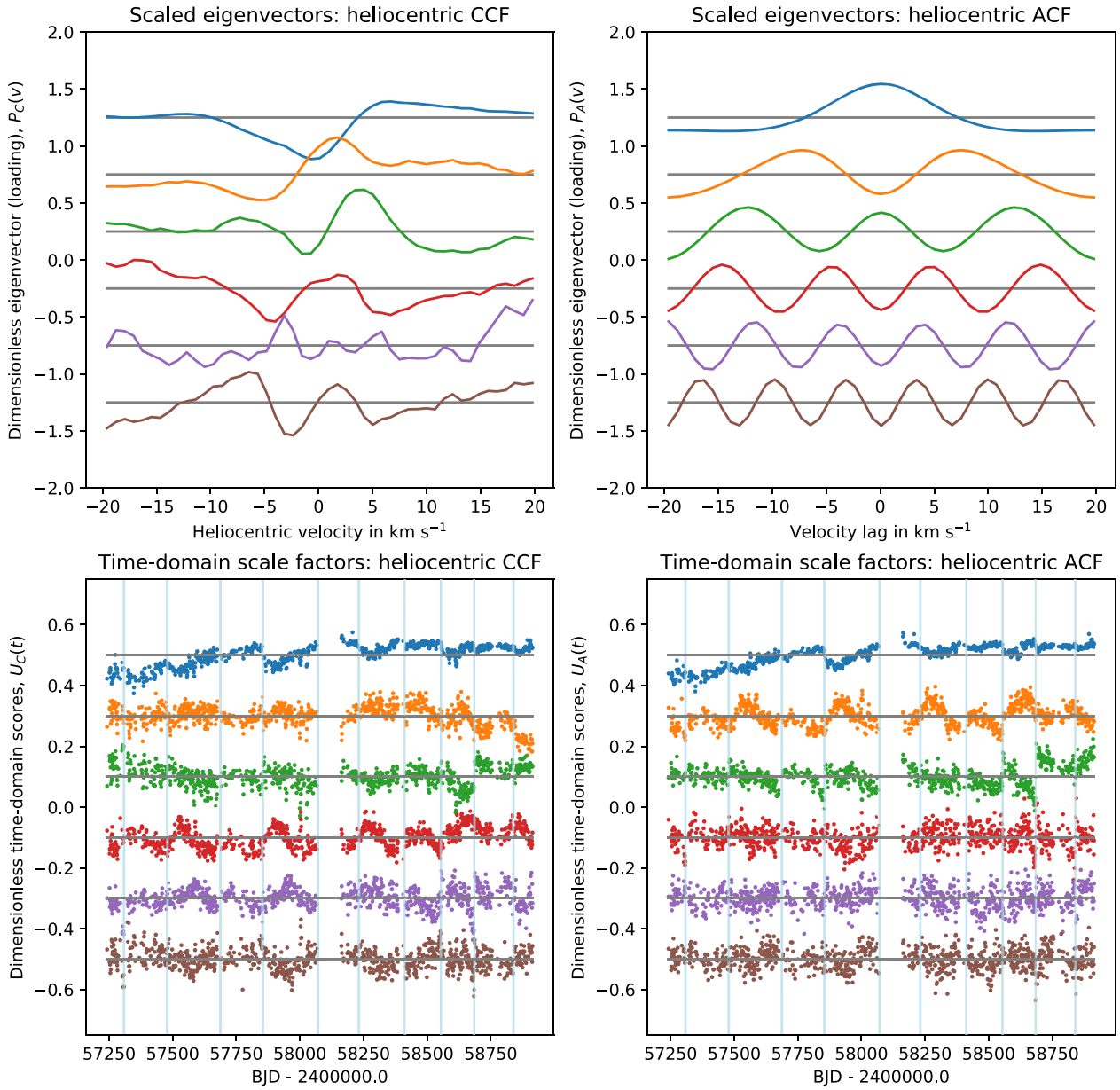

Figure 3. The first six basis vectors (loadings) of the singular-value decomposition of the CCF (upper left) and the ACF of the CCFs (upper right) of the heliocentric time series capture the highest-variance stellar and instrumental behaviours. In the lower panels, their time-domain coefficients (scores) are plotted against barycentric Julian date. Both the basis vectors and scores are normalized and have been arbitrarily shifted in the vertical direction for clarity. The colours of the scores in the lower panel match the corresponding loadings in the upper panel. The sign of the basis vector (prior to shifting) is arbitrary. Vertical light-blue lines in the lower panels denote the dates of cryostat warm-ups.

The scores of the first principal component of both the CCF and the ACF, plotted in blue at the top of all four panels, show a secular upward trend with a superposed signal of higher frequency. The form of the trend, and the shape of the corresponding CCF eigenvector, indicates that this mode of variation affects both the depth and asymmetry of the line profile. It bears a strong resemblance to the variability of the CCF area (i.e. the product of the FWHM and central line depth) noted by Collier Cameron et al. (2019). These authors attributed the trend in CCF area to a secular decline in solar network flux and the faster variations to passages of active-region faculae across the solar disc. Thus, one sees that the ACF is able to recover a very similar series of scores in a way that is insensitive to line shifts. We will exploit this property for separating true Doppler shifts from stellar variability in Section 3.

The time variations of the scores of the second principal component of the ACF (orange traces, second from top in right-hand panels of Fig. 3) and the fourth principal component of the CCF (red traces, fourth from top in left-hand panels) are also similar, though the CCF version appears noisier. Collier Cameron et al. (2019) noted the same pattern of variability in the FWHM of the Gaussian profile fitted to the CCF by the HARPS-N DRS, arising from seasonal changes in the apparent solar rotational broadening. The Earth's orbital eccentricity gives rise to an annual modulation in its orbital angular velocity, and hence the apparent solar rotation rate. The six-month oscillation in

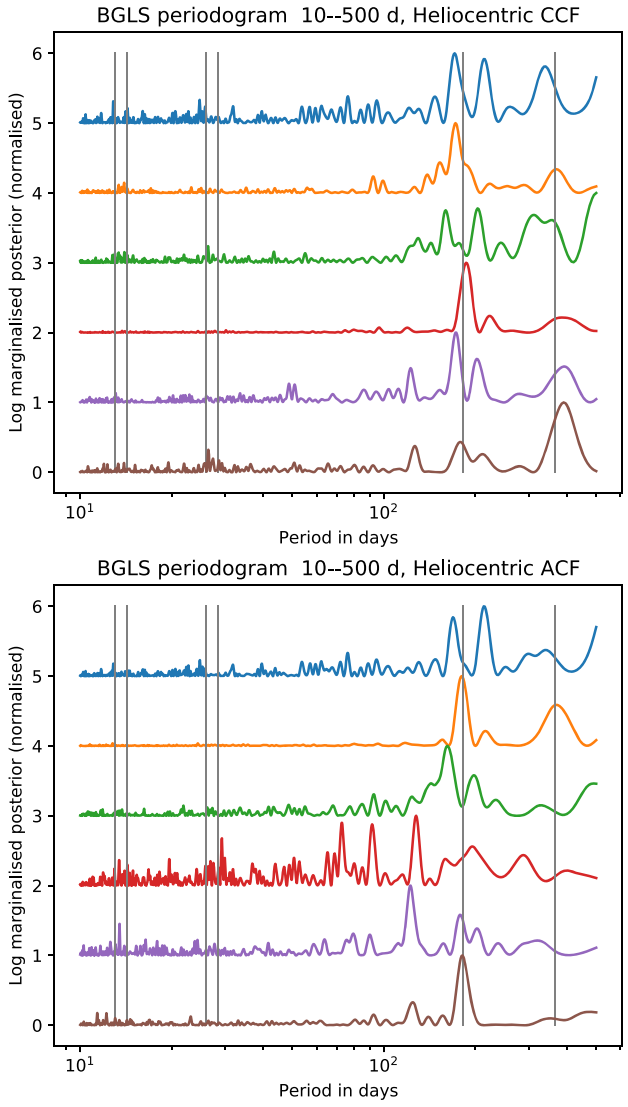

Figure 4. Bayesian Generalized Lomb–Scargle periodograms of the six leading principal components of the residual CCF (upper) and ACF (lower) of the heliocentric time series, in the same order as their counterparts in Fig. 3. The *y*-axis is the posterior probability density marginalized over the amplitude, phase, and zero-point of each time series, rescaled to a peak value of unity for display purposes. Several of the principal components of both time series exhibit power around the first harmonic of the solar synodic rotation period, denoted by vertical bars at $P = 13.0$ and 14.25 d. A similar pair of bars marks the solar rotation period. Other bars show periodicities of 6 months (solar obliquity) and 1 yr (Earth orbital eccentricity). Successive traces are offset by 1 unit for clarity.

the obliquity of the solar rotation axis to the Earth's orbital plane also affects the rotational broadening. The Bayesian Generalized Lomb–Scargle (BGLS; Mortier et al. 2015) periodogram of the second principal component of the ACF (Fig. 4) shows both periods clearly. The corresponding eigenvector for the CCF resembles the second derivative of the line profile, as expected for the CCF changing in width.

The third principal component of the ACF (green traces, third from top in right-hand panels of Fig. 3 and the CCF (green, third from top in left-hand panels) also resemble each other. They show apparently stochastic discontinuities followed by quasi-exponential decay with a time constant of order a few tens of days. These discontinuities are of instrumental origin. There is a very slow leak in the continuous-flow cryostat of the HARPS-N CCD. The cryostat has to be warmed up approximately twice per year to drive off the water that starts to obstruct the flow of liquid nitrogen. It has been observed that these warm-ups cause a sudden change in the asymmetry of the PSF, which takes a few weeks to decay (Dumusque et al. 2020). During the period of these observations, warm-ups were carried out at JD (24)57161.5, 57308.5, 57478.5, 57687.5, 57854.5, 58071.5, 58231.5, 58412.5, 58554.5, 58684.5, and 58839.5. These clearly coincide with the discontinuities in the third principal component of the ACF.

Examination of the BGLS periodograms (Fig. 4) of the leading principal components of the CCF reveals power at half the solar rotation period in the first, second, third, and fifth principal components. The ACF shows power at this period in the first, fourth, and fifth components. These components probably track profile-shape changes caused by sunspot groups and faculae traversing the visible solar hemisphere. There is surprisingly little power at the solar rotation period in the principal components of either the CCF or the ACF.

The CCF shows power at 6 months and/or 1 yr in its second and fourth components; the ACF shows power at these periods in the second and sixth components. Since the heliocentric time series by definition contains no solar reflex motion, these periodic shifts must also be associated with CCF profile-shape variability arising from Earth's orbital motion.

The second principal component of the CCF and the third component of the ACF in Fig. 4, which we have identified with cryostat warm-ups, shows no power at the solar rotation period or its harmonics in either the CCF or the ACF, but we see significant structure on time-scales upwards of 100 d. This indicates that the changes in profile shape caused by cryostat warm-ups are different in character from any form of rotationally modulated solar activity.

Overall, we see that in the CCF, a simple profile shift would have non-zero projection on to multiple eigenvectors, including those that primarily represent broadening, skew and kurtosis. The eigenvectors of the corresponding components of the CCF have a mix of even and odd characteristics, and their odd parts should therefore affect the measured RVs. However, most of the time variations of the CCF appear broadly similar to those of the shift-invariant profile-shape changes probed by the ACF. This raises the possibility that *the ACF can be used to deduce the contribution of profile shape changes to the measured RVs*. We conclude that, at least for the heliocentric solar time series, principal-component analysis of the ACF could provide an effective means of separating the effects of dynamical shift from those of stellar and instrumental profile variability.

3 THE RV RESPONSE TO ACF TIME-DOMAIN VARIATIONS

To achieve this, we treat the time series of scores for each principal component of the ACF as the coefficients of a set of unknown eigenvectors representing orthogonal modes of variability in the shape of the CCF. These unknown eigenvectors of the CCF will affect the measured RV to a greater or lesser degree depending on whether they are pre-dominantly of even or odd character.

3.1 Projection into the ACF time-domain subspace

The set of RV observations has m elements, and can be thought of as a vector $\mathbf{v}_{\rm obs}$ belonging to an m-dimensional space S. The l

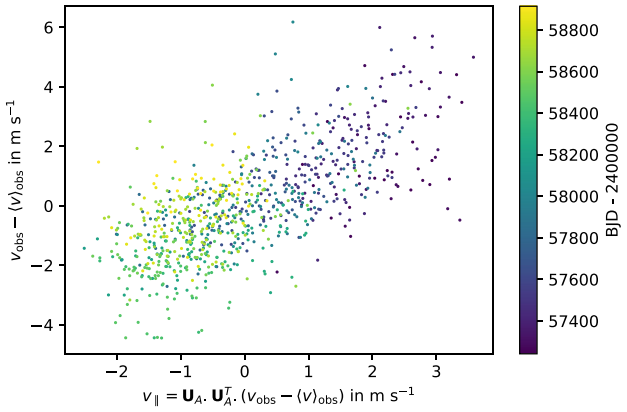

Figure 5. Observed velocities $\mathbf{v}_{obs}$ plotted against the shape-driven velocity component $\mathbf{v}_\parallel$ computed using SCALPELS projection. The individual points are colour-coded by date of observation.

orthonormal columns of $\mathbf{U}_A$ have the same dimension as S, and define an l-dimensional subspace $U \subset S$, centred on the origin of S.

We first subtract the inverse-variance weighted mean $\langle v \rangle_{obs}$ from the vector $\mathbf{v}_{obs}$ of RVs measured with the data-reduction pipeline, to ensure orthogonality. We project the difference $\mathbf{v}_{obs} - \langle v \rangle_{obs}$ on to the time-domain manifold spanned by the basis formed by the columns of the matrix $\mathbf{U}_A$. Each row of $\mathbf{U}_A$ corresponds to an individual observation identified with a unique time-stamp. For conciseness we refer to this observation-identifier space as the 'time domain'. The inner product of the kth column of $\mathbf{U}_A$ (i.e. the scores associated with kth basis vector for ACF decomposition) with the velocities yields the response $\alpha_k = \mathbf{U}_{A,k}^T \cdot (\mathbf{v}_{obs} - \langle v \rangle_{obs})$ of the RV to the time variation of $\mathbf{U}_{A,k}$. The vector of response factors is then

$$\hat{\alpha} = \mathbf{U}_A^T \cdot (\mathbf{v}_{obs} - \langle v \rangle_{obs}). \tag{6}$$

The sum of the scaled velocity contributions from all principal components of the ACF is then $\mathbf{v}_\parallel = \mathbf{U}_A \cdot \hat{\alpha}$. This velocity vector $\mathbf{v}_\parallel$ lies within the subspace U, and gives a complete model of the RV perturbations arising from the changes in profile shape to which the ACF is sensitive:

$$\mathbf{v}_\parallel = \mathbf{U}_A \cdot \mathbf{U}_A^T \cdot (\mathbf{v}_{obs} - \langle v \rangle_{obs}). \tag{7}$$

The product $P_\parallel = \mathbf{U}_A \cdot \mathbf{U}_A^T$ is a projection operator. The operator $P_\perp = (\mathbf{I} - \mathbf{U}_A \cdot \mathbf{U}_A^T)$ projects on to the subspace orthogonal to $\mathbf{U}_A$. The residual velocities

$$\mathbf{v}_\perp = P_\perp \cdot (\mathbf{v}_{obs} - \langle v \rangle_{obs}) = \mathbf{v}_{obs} - \langle v \rangle_{obs} - \mathbf{v}_\parallel \tag{8}$$

lie outside the subspace $\mathbf{U}_A$, and therefore $\mathbf{v}_\perp$ preserves the shifts to which the ACF is insensitive. Information is, however, lost in this process. The resulting velocities $\mathbf{v}_\perp$ are biased down by the projection of the velocities on to $\mathbf{U}_A$, as discussed in Section 4.

From here on, we refer to $\mathbf{v}_{obs}$ as the 'measured' or 'observed' velocities'; $\mathbf{v}_\parallel$ as the 'model' or 'shape-driven' velocities; and $\mathbf{v}_\perp$ as the 'shift-driven' velocities. Fig. 5 shows that the shape-driven velocity perturbations, $\mathbf{v}_\parallel$, are strongly correlated with the observed velocities, reproducing faithfully the long-term and short-term fluctuations dominated by stellar activity.

Given a set of measured RVs ($\mathbf{v}_{obs}$) and the corresponding array of CCFs from which they were derived, equations (3), (5), (7), and (8) constitute a simple linear projection method for deriving the shape-driven perturbations to the RV ($\mathbf{v}_\parallel$), so as to provide a substantially cleaned set of shift-driven RVs ($\mathbf{v}_\perp$).

3.2 Outlier clipping

The shape of the CCF is sensitive to more than just solar activity. Changes in spectrograph focus can affect the FWHM of the CCF, while cryostat warm-ups perturb the skewness of the profile. Noisy CCFs, saturated exposures, or undetected cloud obscuration of part of the solar disc, can also cause temporary profile distortions which may not correlate with any of the highest-variance principal components.

Such anomalous observations may indeed generate their own basis functions when SVD is applied to the ACF time series. Their coefficients are normally close to zero, except when an anomaly occurs. They then appear as outliers in the corresponding columns of $\mathbf{U}_A$. Such points can be masked as bad (0) if their absolute deviations lie further from the median value of the column than a specified number of median absolute deviations (MAD), and good (1) otherwise. If even one of the coefficients for an observation is an extreme outlier, it is likely that the entire observation is contaminated. We therefore create a one-dimensional rejection mask in the time domain from the product of the column masks. For the solar data we found that clipping at 6 times the MAD within each column of $\mathbf{U}_A$ provided a stable set of basis vectors at the cost of reducing the total number of usable days of observation from 886 to 853.

This clipping procedure ensures a clean set of basis vectors, but does not detect outliers caused by unwanted velocity shifts, such as might be caused by an anomalous drift measurement. If present, these must be identified and clipped separately.

3.3 Rank reduction and column re-ordering

Following outlier clipping and masking, the singular-value decomposition of both the CCF and the ACF is re-computed from the surviving observations. It should be noted that all figures in this paper from Fig. 1 onward are based on the masked data set only.

The subspace defined by $\mathbf{U}_A$ has as many dimensions as there are pixels in each row of the ACF array. The CCF and the ACF, however, only display a small number of modes of variability that are detectable above the noise level. Only the highest-variance components of $\mathbf{U}_A$ are needed to capture adequately the shape changes in the ACF. The remaining low-variance components serve only to fit noise. These low-variance components may show spurious correlations with the RV signal, leading to overfitting of $\mathbf{v}_\parallel$. It is therefore possible (and desirable) to model the activity signal adequately and avoid overfitting noise using a reduced number of dimensions.

To determine the optimal size of the null space, we used leave-one-out cross-validation (Celisse 2014). Holding out each row $\mathbf{A}_j$ of the ACF in turn, we decompose the remaining rows and compute the singular-value decomposition:

$$\mathbf{A}_{i \neq j} = \mathbf{U}_{i \neq j} \cdot \mathrm{Diag}(S_{i \neq j}) \cdot V_{i \neq j}. \tag{9}$$

We reconstruct an estimate $\hat{\mathbf{U}}_j$ of the missing jth row of $\mathbf{U}_A$ by fitting the eigenvectors and eigenvalues to the jth row of the ACF:

$$\hat{\mathbf{U}}_j = \frac{(\mathbf{A}_j \cdot V_{i \neq j}^T)}{(\mathrm{Diag}(S_{i \neq j}) \cdot V_{i \neq j}) \cdot V_{i \neq j}^T}. \tag{10}$$

After having repeated this procedure for all rows, we find that the reconstruction of the kth column $\hat{\mathbf{U}}_k^T$ reproduces $\mathbf{U}_{A,k}^T$ with good fidelity for $k < 25$ or so. The ratio of the median absolute deviation (MAD) of $\mathbf{U}_k^T - \hat{\mathbf{U}}_k^T$ to MAD($\hat{\mathbf{U}}_k^T$) rises to values close to unity for values of $k > k_{crit}$ for which the leave-one-out cross-validation indicates that the reconstruction is poor, as shown in Fig. 6.

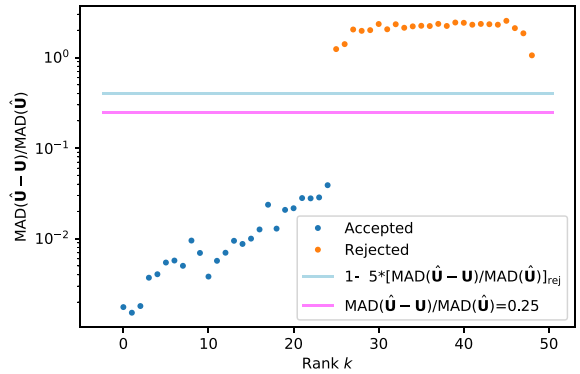

Figure 6. Rank reduction with leave-one-out cross-validation. The ratio of the median absolute deviation (MAD) of $\mathbf{U}_k^T - \hat{\mathbf{U}}_{A,k}^T$ to MAD($\hat{\mathbf{U}}_k^T$) rises sharply to values above unity for values of $k > k_{\rm crit}$. Here $k_{\rm crit} = 25$.

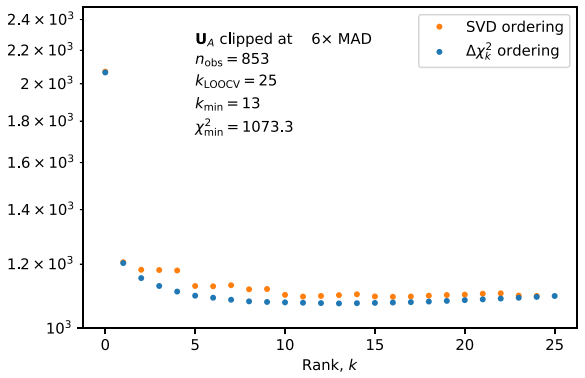

Figure 7. As more dimensions are added to the subspace defined by the time-domain coefficients $\hat{\mathbf{U}}$ of the principal components of the ACF, the χ^2 of $\mathbf{v}_\perp$ decreases to a minimum value, then increases gradually as overfitting degrades the solution. The minimum is reached more rapidly when the columns of U are sorted in order of decreasing $\delta\chi^2$ (blue) rather than in order of their corresponding singular values (orange). With $\delta\chi^2$ sorting, the optimal size of the null space is defined by the minimum at $k = 13$.

Following projection of the RV data into the reduced space defined by the surviving columns of $\hat{\mathbf{U}}$, we find that the quality of the fit between the RV data and the shape model $\mathbf{v}_\parallel$ improves rapidly at first, reaches a minimum then increases gradually as more velocity components are added to the model and overfitting starts to degrade the solution. In other words, we need even fewer principal components to model $\mathbf{v}_\parallel$ than we need to reproduce the ACF itself.

SVD orders the principal components of the ACF in descending order of their eigenvalues $\mathbf{S}_A$. This ordering does not take the velocity projection into account, so the ordering of principal components does not reflect accurately their contributions to the RV. Instead, we reorder the columns of $\hat{\mathbf{U}}$ into the sequence that gives the fastest decrease in χ^2, obtaining the optimal fit to the RV data with the smallest number of basis vectors, as shown in Fig. 7.

Davis et al. (2017) found that 4 or 5 principal components were sufficient to capture the temporal behaviour of synthetic spectra produced by a noise-free SOAP2.0 (Dumusque et al. 2014) simulation of star-spot activity and facular suppression of convective blueshift on a rotating stellar model. As we have seen, the HARPS-N solar data contain additional profile distortions arising from changes in the instrument and Earth's orbital motion, so more principal components are needed.

We find a good compromise between outlier clipping, number of surviving days of observation, and minimal number of basis vectors when we clip $\mathbf{U}_A$ at 6 times the MAD, as noted above. With $\delta\chi^2$ reordering, the χ^2 of $\mathbf{v}_\perp$is minimized at $k_{\rm min} = 13$. We therefore use the 13 leading principal components of $\hat{\mathbf{U}}_A$, ordered by $\delta\chi^2$, to define the null space. We note, however, that the results that follow are only very weakly sensitive to the number of principal components used over the range $6 < k < 13$ or so.

The projections of the observed velocities on to the components of $\mathbf{U}_A$ corresponding to the 6 largest values of $\hat{\alpha}$ are plotted against time in Fig. 8.

3.4 SCALPELS analysis of solar data

We refer to the projection of the observed velocities on to the orthogonal complement of the time-domain scores $\mathbf{U}_A$ of the ACF together with the outlier clipping (Section 3.2) and rank reduction (Section 3.3) algorithms collectively as *Self-Correlation Analysis of Line Profiles for Extracting Low-amplitude Shifts* (SCALPELS). The reader is referred to Appendix C1 for a concise listing of the main steps of the algorithm.

In a blind RV survey, planet-candidate detection is typically conducted using analysis of some form of periodogram such as Lomb–Scargle (e.g. Lomb 1976; Scargle 1982; Zechmeister & Kürster 2009) or marginalized posterior versus orbital period (e.g. Mortier et al. 2015) computed from the RVs. Periodogram peaks are identified and fitted with Keplerian orbit models. This method is, however, susceptible to confusion with rotationally modulated signals from the host star.

To assess the effectiveness of the SCALPELS algorithm for suppressing stellar noise, we apply it to the daily averaged solar RVs in the heliocentric frame. In the absence of any dynamical shifts or instrumental calibration drift, the measured RVs should show only variations caused by line-profile shape changes arising from solar activity, changes in the instrumental point-spread function, and changes in the apparent solar rotation rate arising from Earth's orbital eccentricity and the solar obliquity. Given that the heliocentric solar data set contains no planet signals, a frequency search for candidate periodic signals provides a means of establishing the planet-detection threshold for comparable data sets.

In Fig. 9 we show the observed heliocentric RVs minus their own mean, together with the shape-driven velocities obtained by SCALPELS projection and the shift-driven velocity difference between the two. The histograms of the observed velocities and the shift-driven velocities are also shown. The distribution of observed velocities is severely non-Gaussian, with a bimodal character arising from short-term (days–weeks) and long-term activity (years) variability.

After subtracting the SCALPELS-identified shape-driven velocity residual velocities $\mathbf{v}_\parallel$, the shift-driven velocities $\mathbf{v}_\perp$ are nearly constant with respect to time, with a local RMS scatter of 1.25 m s^{-1}. The Anderson–Darling 1-sample test (Scholz & Stephens 1987) indicates that the distribution of the shift-driven velocities is indistinguishable from a normal distribution (Fig. 9, lower panels), with standard deviation $\sigma = 1.25$ m s^{-1}.

Any stellar activity signature remaining in the shift-driven velocity time series is likely to show temporal correlations and departures from uncorrelated Gaussian noise. There appear to be weakly correlated residuals with amplitudes of a few tens of cm s^{-1} on a range of time-scales upward of about 200 d. The origin of these slow drifts is unclear. They could be a shift-like manifestation of

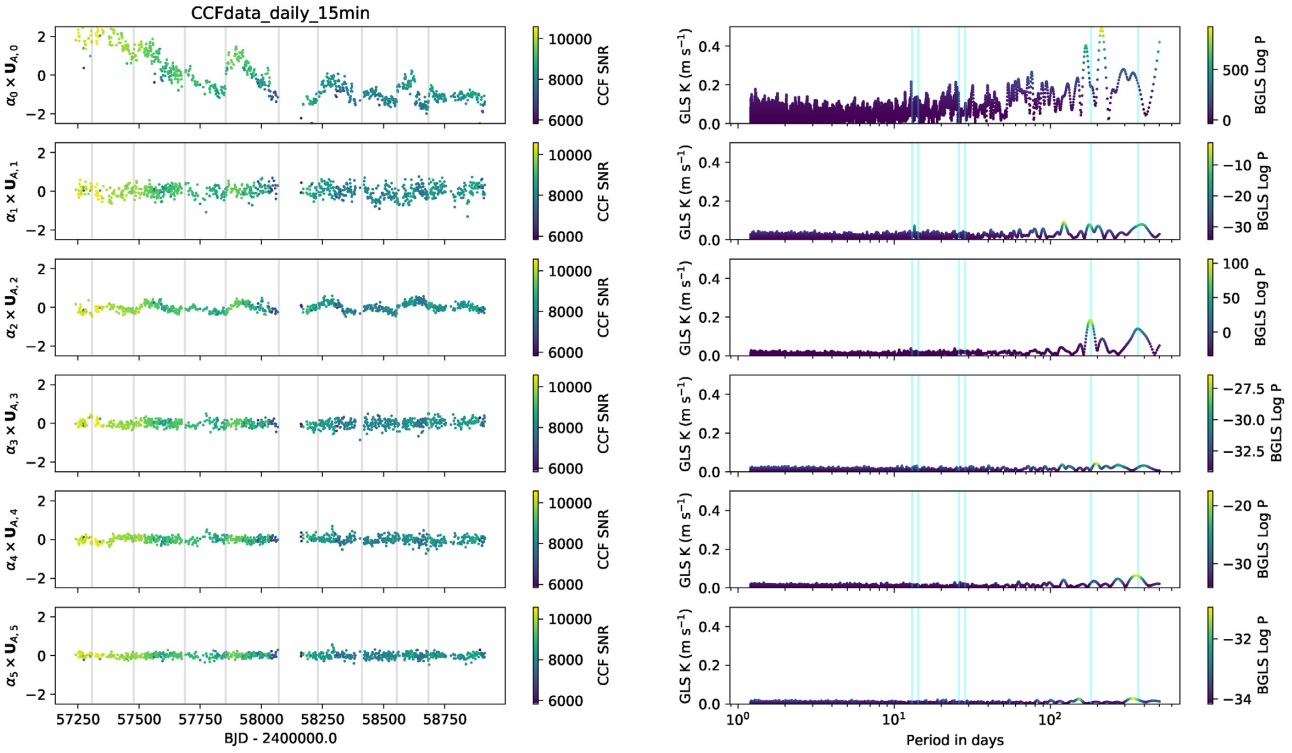

Figure 8. Left panel: The first six vector components of the projection of the solar heliocentric RVs (in m s^{-1}) into the rank-reduced ACF time-domain subspace, ordered by their projection coefficients $\hat{\alpha}$ and colour-coded by the SNR of the CCF. Vertical grey lines denote the dates of cryostat warm-ups. Right panel: Generalized Lomb–Scargle K amplitude periodograms of the same six time series, colour-coded by Bayesian GLS likelihood (Mortier et al. 2015). The vertical blue lines are at periods of 13.0, 14.25, 26.0, 28.5, 182.63, and 365.25 d. Power is seen near the solar rotation period in the first component. Weak power is seen at half the rotation period in the second component. The third component shows the annual and six-monthly modulation of the CCF FWHM caused by Earth's orbital eccentricity and the solar obliquity, respectively. Annual signals are also present in the second and fifth components.

solar activity. Secular drifts induced by the instrument are also a possibility.

The Ljung–Box Q test (Ljung & Box 1978) suggests that the shift-driven velocities remain weakly correlated at all autocorrelation lags up to at least 100 d of observation. Therefore, it is likely that some activity-driven velocity components remain in $\mathbf{v}_{\perp}$, but as the upper-right panel of Fig. 9 shows, they are substantially reduced relative to the original, observed velocity time series. The shape-driven signals are, as expected, strongly correlated. This offers improved detection prospects for small planetary-orbit signals at periods of tens to hundreds of d.

In Fig. 10 we show periodograms (in terms of the best-fitting semi-amplitude of a sinusoid as a function of its period) for RVs measured with the data-reduction system, transformed to the heliocentric reference frame. The periodogram of the raw velocities shows numerous candidate signals with semi-amplitudes of order 0.4 m s^{-1}, particularly between 13 and 26 d, close to the solar synodic rotation period and its first harmonic. The SCALPELS projection shows a very similar pattern of semi-amplitudes.

These peaks are strongly suppressed in the semi-amplitude periodogram of the shift-driven RVs (Fig. 10, bottom trace), which shows no strong frequency structure. This is important, since the background level of the periodogram peaks in the cleaned time series with no planets present, effectively sets the sensitivity for detecting planets after applying SCALPELS to clean the velocity measurements. Peaks with amplitudes greater than 0.30 m s^{-1} are seen in the shift-driven RVs at $P = 191.47$ and 30.45 d. Their amplitudes are reduced to 0.320 and 0.339 m s^{-1} respectively, nearly a factor of two less than those found in the observed velocities measured with the data-reduction system. We note that the excess of power at around 200 d is commensurate with the average interval between cryostat warm-ups.

4 ALGORITHM TESTS WITH PLANETS INJECTED INTO SOLAR OBSERVATIONS

We now turn to the problem of determining the impact of the SCALPELS signal separation on detection thresholds when weak planetary signals are present. We begin by injecting four periodic shift signals into the heliocentric CCF time series, using equation (1) to shift the rows by the small amounts required. The periods of these signals were well-spaced in log period, at non-integer periods of 7.142, 27.123, 101.543, and 213.593 d. The 27.1-d period was chosen deliberately to be close to the solar synodic rotation period. The injected signals are sinusoidal in form, with semi-amplitudes $K = 40$ cm s^{-1}. This is less than the amplitude of the strongest signals arising from solar variability in the upper trace of Fig. 10, but greater than the activity-driven signals remaining after subtracting the SCALPELS projection from the RV measurements. For a 1 $M_{\odot}$ star the corresponding planet masses are 1.2, 1.9, 2.9, and 3.7 $M_{\oplus}$.

Before proceeding, we must consider the methodology used to extract velocities from the shifted CCFs and to estimate their precision from the covariances in the rows and columns of the CCF. The reader is referred to Appendix B for the details of this methodology.

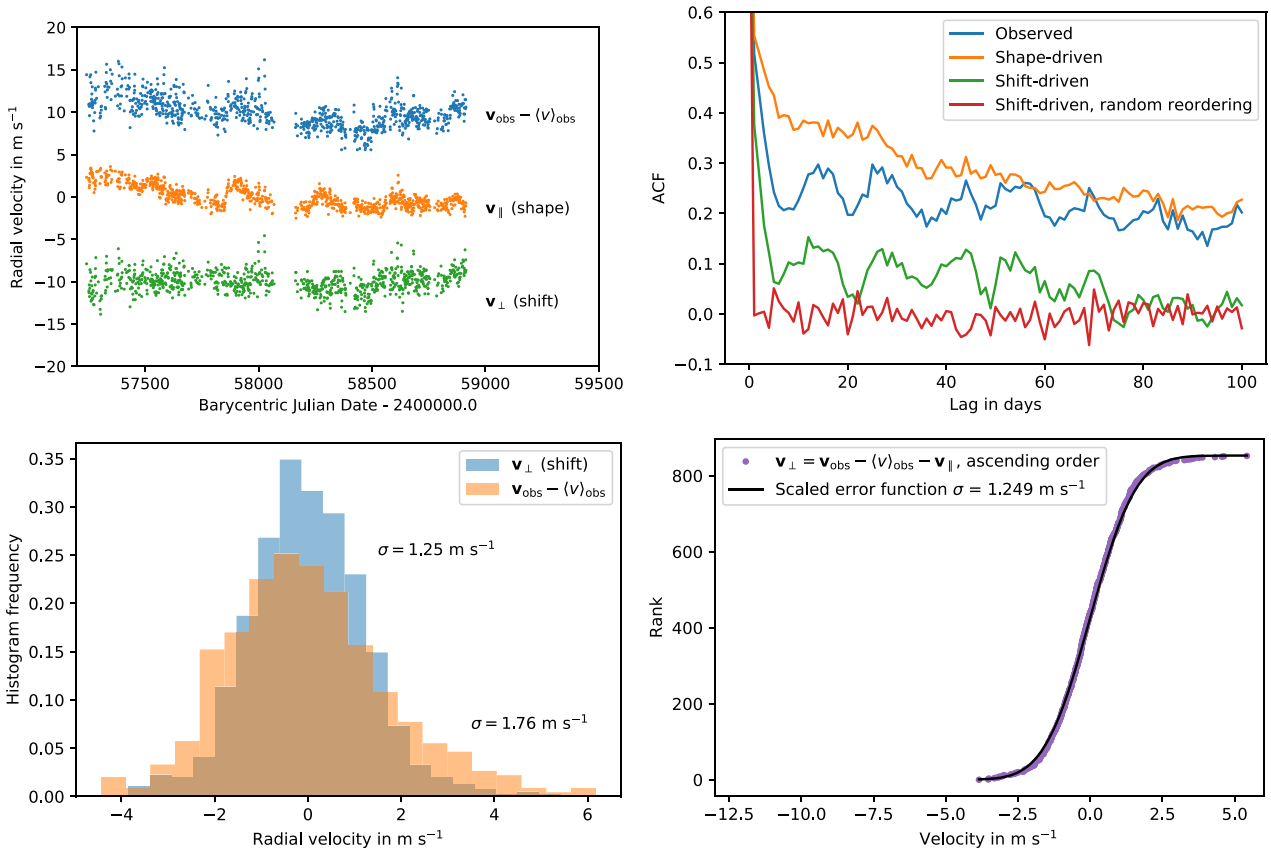

Figure 9. Upper left: observed RVs transformed to the heliocentric reference frame, together with their SCALPELS-separated shape-driven and shift-driven velocity components, offset by ± 10 m s^{-1} for clarity. Upper right: ACFs of the RVs, corrected for missing dates of observation and normalized to unity at zero lag. The original RVs (blue) are seen to be strongly correlated at all time lags, and the shape-driven velocities (orange) even more so. The autocorrelation of the shift-driven velocities decays substantially more rapidly than the observed or shape-driven velocities. Any correlation at lags longer than ∼70 d is negligible. The red curve is the ACF of the shift-driven velocities shuffled into random order, and is effectively uncorrelated. Lower left: Histograms show that the RMS scatter has been reduced from 1.76 m s^{-1} in the original data to 1.25 m s^{-1} in the shift-driven velocities. Lower right: shift-driven velocities sorted in ascending order and overplotted with the cumulative normal distribution with $\sigma = 1.25$ m s^{-1}.

4.1 Recovery of weak injected planet signals

Following signal injection, RVs were again measured from the CCF time series using equation (B1).

Fig. 11 shows the periodograms obtained from these velocities, from the SCALPELS projection of the shape-driven velocity component, and from the differences between them representing pure shifts.

The periodogram of the velocities measured from the shifted CCFs does not enable us to distinguish clearly between the injected signals and RV variability intrinsic to the Sun or the instrument. The injected signals at 7.1, 27, and 100 d are detected fairly unambiguously, but the 213-d signal is suppressed and there are also many false detections of amplitude comparable to the injected signals.

The periods, semi-amplitudes and uncertainties of the five strongest signals recovered from the periodogram of the shift-driven velocities after subtraction of the shape-driven model are listed in columns 3, 4, and 5 of Table 1. Among these, four of the five strongest signals are very close to the frequencies of the injected planet signals. The mean of their semi-amplitudes is 0.428 ± 0.01 m s^{-1}, somewhat above the expected sample uncertainty of the injected values. They dominate over all residual variability and zero-point jitter signals except for a spurious 0.435 m s^{-1} signal at $P = 185.1$ d. This latter period is so close to half a year that it is likely to arise from an as-yet unidentified effect of observing the Sun from the Earth, which would not be expected to affect exoplanet searches.

4.2 Simultaneous modelling of stellar variability and planetary motion

The amplitudes of the injected planet signals do not appear to be strongly attenuated in the upper trace of Fig. 11, but they are buried in a forest of activity-related peaks of similar amplitude. In the bottom trace they stand out above the suppressed activity signals. Their amplitudes could none the less be affected by activity if the planet signals themselves contaminate the SCALPELS projection. This could occur if the injected shift signals are not perfectly orthogonal to all elements of $\mathbf{U}_A$, and hence partly absorbed in the SCALPELS projection. The data set has a finite length, so irregularly sampled superpositions of Keplerian signals will not be perfectly orthogonal to any randomly chosen vector in the same space. Moreover, a periodogram fits only a single sinusoid per frequency sample, so that cross-talk between multiple signals can lead to incorrect amplitude estimates.

The orbital perturbations of any planet and the SCALPELS projection process must therefore be modelled self-consistently for the signal separation to recover their semi-amplitudes as reliably as

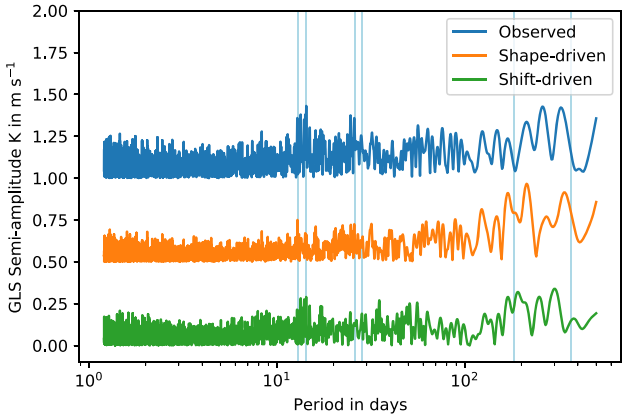

Figure 10. Semi-amplitude periodogram demonstrating the value of signal separation by projection of the observed RVs on to the principal time-domain components of the ACF. The top periodogram (blue) is for the measured velocities derived from the CCF. The second trace (orange) is for the shape-driven velocities $\mathbf{v}_{\parallel}$ produced by the SCALPELS projection. The third periodogram (green) is that of the shift-driven velocities $\mathbf{v}_{\perp}$ remaining after subtraction of the shape-driven velocities from the observations. Light-blue bars denote the approximate ranges of the solar rotation period and its first harmonic, and periods of 6 months and 1 yr.

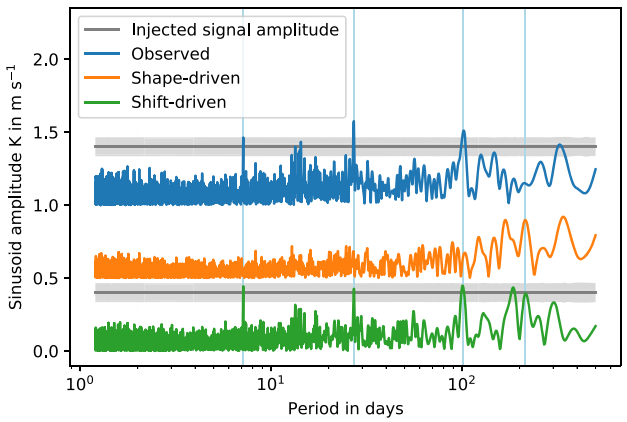

Figure 11. Periodograms of velocities derived from the heliocentric solar CCFs when four sinusoidal signals of 40 cm s^{-1} have been injected at the periods denoted by the vertical blue lines. The traces are as defined in the caption of Fig. 10. The uncertainty in the amplitude of the sinusoid is almost independent of period. Its 1σ limits are indicated by the shaded region around the horizontal grey lines showing the amplitude of the injected signal. The four dominant peaks in the lower trace indicate successful recovery of all four signals, at amplitudes that are consistent with the injected values, and whose scatter is consistent with the expected uncertainty.

possible. Once the periods of candidate signals have been determined – either through prior knowledge of transits or via periodogram search(es) – parameter estimation and signal separation can be achieved in a single linear calculation.

For a set of n planet signals, the net orbital velocity vector $\mathbf{v}_{\rm orb}$ can be modelled as the product of a set of coefficient pairs $\theta_{\rm orb} = \{A_1, B_1, \cdots, A_n, B_n\}$ with an array of time-domain function pairs $\mathbf{F} = \{\cos\omega_1 t_j, \sin\omega_1 t_j, \cdots, \cos\omega_n t_j, \sin\omega_n t_j\}$, ω_k being the orbital frequency of the kth planet:

$$\mathbf{v}_{\rm orb} = \mathbf{F}\cdot\theta_{\rm orb}. \tag{11}$$

For simplicity we assume circular orbits here, though eccentric orbits could in principle be fitted with periodic signals including additional Fourier components.

The complete model of the RV data is then the sum of the model orbital velocity variations and the shape-driven velocity variations. The only unknowns are the amplitudes and phases of the orbital basis functions $\mathbf{F}$. We can solve for these using the method of least-squares

The simultaneous solution involves computing the shape-driven variations from the difference between the observed RVs and the model velocities. In the projection-operator language of Section 3, we solve for the vector $\theta_{\rm orb}$ that minimizes $\chi^2 = (\mathbf{P}_\perp\cdot\delta\mathbf{v}^T)\cdot\Sigma^{-1}\cdot(\mathbf{P}_\perp\cdot\delta\mathbf{v})$, where $\delta\mathbf{v}\equiv\mathbf{v}_{\rm obs} - \langle v_{\rm obs}\rangle - \mathbf{F}\cdot\theta_{\rm orb}$.

Defining $\mathbf{v}_\perp = \mathbf{P}_\perp\cdot\mathbf{v}_{\rm obs}$ and $\mathbf{F}_\perp = \mathbf{P}_\perp\cdot\mathbf{F}$, the goodness of fit is quantified by

$$\chi^2 = (\mathbf{v}_\perp - \mathbf{F}_\perp\cdot\theta_{\rm orb})^T\cdot\Sigma^{-1}\cdot(\mathbf{v}_\perp - \mathbf{F}_\perp\cdot\theta_{\rm orb}), \tag{12}$$

which is minimized by solving for $\theta_{\rm orb}$:

$$(\mathbf{F}_\perp^T\cdot\Sigma^{-1}\cdot\mathbf{F}_\perp)\cdot\theta_{\rm orb} = \mathbf{F}_\perp^T\cdot\Sigma^{-1}\cdot\mathbf{v}_\perp. \tag{13}$$

The log likelihood of the data given the model is

$$\begin{aligned}\ln\mathcal{L} &= -\frac{1}{2}(\mathbf{v}_\perp - \mathbf{F}_\perp\cdot\theta_{\rm orb})^T\cdot\Sigma^{-1}\cdot(\mathbf{v}_\perp - \mathbf{F}_\perp\cdot\theta_{\rm orb})\\ &\quad -\frac{1}{2}\ln|\Sigma| - \frac{m}{2}\ln(2\pi)\\ &= -\frac{1}{2}(\mathbf{v}_\perp - \mathbf{F}_\perp\cdot\theta_{\rm orb})^T\cdot\Sigma^{-1}\cdot(\mathbf{v}_\perp - \mathbf{F}_\perp\cdot\theta_{\rm orb})\\ &\quad -\frac{1}{2}\left(\sum_{j=1}^{m}\ln\Sigma_{jj}\right) - \frac{m}{2}\ln(2\pi).\end{aligned} \tag{14}$$

Here we use the simplifying assumption that the RV measurements are uncorrelated. The covariance matrix is diagonal and its log determinant is $\sum_{j=1}^{l}\ln\Sigma_{jj}$. The diagonal elements $\Sigma_{jj} = {\rm Var}(\mathbf{v}(t_j))$ are calculated using equations (B2) and (B8). If the RV data are sufficiently densely sampled, a time-dependent covariance model with a kernel incorporating the stellar rotation period and active-region lifetime could also be included – see, e.g. Gilbertson et al. (2020). For convenience, we summarize the algorithm in Appendix C2.

In Table 1, columns 6–8 list the amplitudes and uncertainties of the sinusoidal signals recovered from the data at the known periods of the injected signals. The standard deviation of the four individual recovered semi-amplitudes is $\sigma = 0.010$ m s^{-1}. The sample mean and standard deviation ($\sigma/\sqrt{4}$) of the signal amplitudes are 0.433 ± 0.005 m s^{-1}, again somewhat above the injected value. The four individual signals deviate from the injected values by amounts that are consistent with their individual estimated 0.066 m s^{-1} semi-amplitude uncertainties.

The improvement in signal separation is also apparent from Fig. 12. The top two traces are almost the same as those in Fig. 11, but the balance of the signal separation is changed by the explicit modelling of the orbital motion at the known periods. The periodogram of the fitted orbital model illustrates the apparent signal attenuation that can occur when fitting multiple signals with a single sinusoid. The final residuals are very similar to those of Fig. 10.

The precision of the recovered semi-amplitudes is poorer when the velocities are left uncorrected for profile-shape variations, by setting the dimension $k_{\rm max}$ of the null space to zero. If sinusoids are fitted to the raw $\mathbf{v}_{\rm obs}$ at the same four periods, we obtain semi-amplitudes 0.466, 0.560, 0.481, and 0.129 ms $^{-1}$, whose sample mean and standard deviation are 0.409 ± 0.083 m s^{-1}. Fig. 13 shows clearly the improvement in fidelity of the recovered amplitudes when the optimal shape model is applied.

Table 1. Frequencies, periods, and semi-amplitudes of the strongest signals in the periodograms of raw and shape-corrected apparent velocities. The first two columns give the periods and semi-amplitudes of the four injected signals. Columns 3–5 give the periods, semi-amplitudes, and uncertainties of all peaks with amplitudes greater than 33 cm s^{-1} in the periodogram of residual velocities remaining after subtraction of the SCALPELS projection, as in a blind RV search. The final three columns give the same information, from simultaneous modelling of CCF shape changes and planetary motion, made with prior knowledge of the four injected periods, as described in (Section 4.2).

Injected		Velocities from residual CCF			Velocities from simultaneous fit		
P (d)	K (m s^{-1})	P (d)	K (m s^{-1})	σ_K (m s^{-1})	P (d)	K (m s^{-1})	σ_K (m s^{-1})
7.142	0.400	7.144	0.443	0.064	7.142	0.451	0.065
27.123	0.400	27.133	0.426	0.066	27.123	0.430	0.066
101.543	0.400	100.625	0.448	0.064	101.543	0.426	0.066
213.593	0.400	215.219	0.395	0.065	213.593	0.427	0.069
–	–	185.101	0.435	0.065	–	–	–

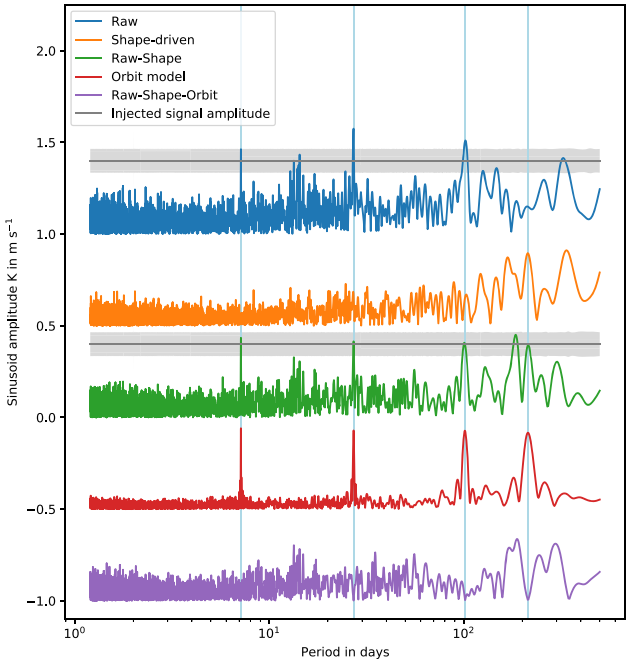

Figure 12. As for Fig. 11, but for the case where signal separation is performed simultaneously with orbit fitting given prior knowledge of the orbital periods. The middle (green) periodogram shows the difference between the observed and shape-driven velocities. The fourth (red) trace shows the periodogram of the fitted model of the four orbital signals. The bottom (magenta) trace shows the residuals after subtraction of both the shape-driven and orbital RV models. The scatter in the amplitudes of the four dominant peaks in the middle trace is consistent with the expected uncertainty, and their mean amplitude is unbiased relative to the injected values.

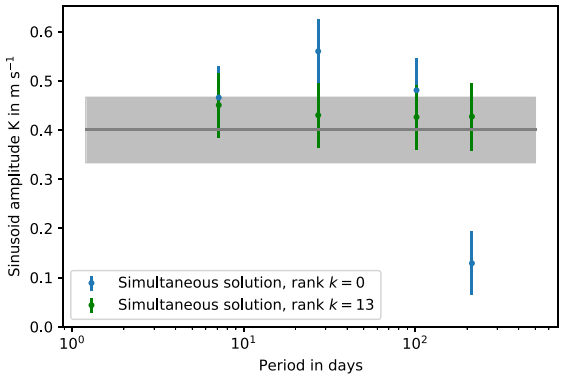

Figure 13. The recovered amplitudes of the injected signals are shown in green for the simultaneous SCALPELS fit with $k_{\max} = 13$, and in blue for simultaneous sinusoidal fits to the uncorrected original RV data ($k_{\max} = 0$) at the injected periods. The grey line and band show the original signal level and the formal 1σ uncertainty on the recovered amplitude. The result demonstrates clearly the improvement in consistency and fidelity in the recovered signal amplitudes when SCALPELS is used.

The formally propagated 0.066 m s^{-1} uncertainties in the semi-amplitudes are $\sim 1.6\sigma_v/\sqrt{N_{obs}}$ for $N_{obs} = 853$ if we adopt a single-measurement precision of $\sigma_v \simeq 1.25$ m s^{-1} based on the RMS scatter in the heliocentric solar velocities after removal of the shape perturbations (see Section 3.4).The RV amplitude precision appears to scale as expected for uncorrelated random variables to a level $\sim$20 times better than the single-measurement precision.

The average of the recovered semi-amplitudes shows no evidence of bias relative to the injected value when signal separation is performed with prior knowledge of the orbital period, as is the case with transiting planets. Thus, the RV semi-amplitudes inferred from the cleaned velocities can be significantly more reliable than RV semi-amplitudes inferred from original velocity measurements.

We conclude that the SCALPELS method succeeds in reducing correlations between apparent velocities due to stellar variability, based solely on line shape changes and without making use of time-domain information. This decoupling from time-domain information allows planet signals to be recovered with good fidelity even when they fall close to the stellar rotation period, as is the case with the 27-day signal injected here.

5 DISCUSSION

5.1 Summary

We have presented a new algorithm for extracting precise RV estimates from high-resolution spectroscopic planet surveys. The algorithm begins with a list of CCFs for each observation epoch, constructs a reduced-rank representation of stellar variability and reconstructs CCFs which have been cleaned of most stellar variability. We demonstrated the algorithm using observations of the solar spectrum from HARPS-N. We verified and validated that the

algorithm can accurately detect multiple simulated planets injected into solar observations, spanning a wide range of orbital periods.

5.2 Planets injected into solar observations

Our algorithm recovered the RV signals of injected planets with semi-amplitudes of just 40 cm s^{-1} with an SNR of 6 based on a time series of 853 CCF measurements.

The semi-amplitude uncertainty of 6.6 cm s^{-1} implies that it is possible for intensive Doppler spectroscopy campaigns to measure the masses of transiting planets (i.e. with well-measured orbital period and inclination) at **∼ 15 per cent** precision for RV semi-amplitudes as small as $K = 40$ cm s^{-1} for a solar twin, even if there are multiple planets spanning a wide range of orbital periods.

The precision with which we can measure velocity semi-amplitudes is insensitive to the velocity semi-amplitude of the planet (for small velocity semi-amplitudes). Thus, our results suggest that intensive Doppler spectroscopy campaigns could detect and measure the masses (times sine of inclination angle) to 6σ precision for planets with RV semi-amplitudes of ∼60 cm s^{-1}, if there were independent evidence for a planet at a given period (e.g. transit, direct imaging) orbiting a sufficiently bright ($V \simeq 6$) Sun-like star.

5.3 Areas for Further Research

The algorithm presented successfully mitigates the impact of solar variability. Nevertheless, there are multiple lines for additional research that are likely to lead to further improvements in the Doppler sensitivity.

5.3.1 Simulations using realistic observing cadence

For planet injection tests in Section 4, we used CCFs from HARPS-N solar observations taken on 853 of a total 886 d. The observations span 1681 d, and over 70 per cent of the observations are spaced by just 1 d. While there are seasonal shutdowns each year and gaps of up to two weeks, this observing cadence is more favourable than that of most targets of Doppler planet surveys. We recommend that future simulations of Doppler planet surveys explore how the planet detection sensitivity depends on observing cadence in the presence of intrinsic stellar variability and after incorporating advanced methods for mitigating stellar variability such as presented here. Hall et al. (2018) have carried out such a study for the HARPS-3 project, which will produce stellar data sets of comparable duration and quality to the solar data set examined here, albeit with seasonal gaps. A similar sensitivity study using HARPS-N solar data has been published recently by Langellier et al. (2020), who concluded that a decade or more of observations could be needed to achieve a 5σ detection of a 50 cm s^{-1} signal with a period of 225 d, given an instrumental white-noise uncertainty of 80 cm s^{-1} and a perfect model of activity-driven RV variability. Haywood et al. (2020) analysed an 8-yr sequence of synthetic solar RVs derived from SDO images, with a six-month duty cycle and decorrelation against hemispherically averaged magnetic field. They succeeded in recovering an injected signal with $K = 30$ cm s^{-1} and $P = 300$ d. Our results, using real data but an idealized observing pattern, give similar cause for optimism.

5.3.2 Granulation

The CCF time series used for verifying our algorithm was based on 15-min daily averages of the HARPS-N solar CCFs. Each exposure time is 300 s, yielding an SNR in the range $250 < \mathrm{SNR} < 400$. This is comparable to that of exoplanet observations for a host star of magnitude 5.5 using the same exposure time. Combining three contiguous such exposures in a 15-min block is reasonably effective at averaging-out the side effects of p-mode pulsations which occur on a ∼5 min time-scale (Chaplin et al. 2019). Modern exoplanet surveys typically choose an exposure time of at least 15 min, averaging subexposures if necessary to average out spectral variability due to pulsations and avoid saturation. Such exposures are repeated at intervals of 2–3 h to mitigate the effects of granulation. Our use of a single 15-min block per day is likely to be less effective at eliminating intrinsic stellar variability due to the granulation pattern, as studied in detail by Meunier et al. (2015). We tested this by averaging all data satisfying the data-selection constraints in Section 2 throughout each day of observations, rather than down-selecting to a single 15-min block. The longer daily baseline reduced the day-to-day scatter in $\mathbf{V}_\perp$ from 1.25 to 1.08 m s^{-1}. This is less than the improvement expected if photon counts were the limiting noise source, but consistent with the improvement expected through averaging over velocity fluctuations arising from photospheric granulation noise.

In principle, granulation might imprint on the stellar spectrum differently than variability on the time-scales of magnetic activity (Cegla 2019). It is likely that some of the 13 modes of RV variability found in our data may be attributable to granulation.

5.3.3 Searching for low-mass planets with outer giants

Some complexities of planetary-system architecture are beyond the scope of this initial study, but merit further investigation. For example, signals from terrestrial planets with sub-m s^{-1} amplitudes may be superposed on much larger signals of giant planetary companions at longer orbital periods. If the giant's orbit is well-characterized, it can be included directly in the model. Indeed the barycentric data considered here contain such a signal, with the synodic period of Jupiter. Rather than simply work in the heliocentric frame, we repeated the analysis of Section 4.2 by injecting the same four signals into the barycentric CCFs. To mimic the situation where the giant's orbit is incomplete or poorly determined, we performed a GP regression with a squared-exponential covariance kernel to smooth out variations on time-scales longer than 300 d. All four of the injected signals were recovered successfully at the correct amplitudes.

More complex cases might involve transiting systems in which a temperate Earth-sized object is accompanied by a few strongly interacting compact short-period companions. In this case, gravitational interactions might make the fully linear algebra approach of the method less effective. As shown in Fig. 9, however, signal separation and detection is reasonably effective even in a 'blind-search' scenario where simultaneous linear fitting is not possible. Furthermore, even a nonlinear parametric model of the orbital reflex motion can be included as a fitting function and its parameters optimized together with the coefficients of ACF basis vectors in an MCMC scheme, at the cost of some additional computational overheads.

5.3.4 Reconstructing full spectra

In this paper, we reconstructed the shape-driven velocity signal using a reduced-rank representation and scores derived from the ACF of the CCFs. In principle, the same approach could be used to reconstruct the CCFs themselves or even the full spectra. Then, the reconstructed spectra could become objects for further analysis, e.g.

performing line-by-line analysis (Dumusque 2018). We anticipate that a reconstruction of the spectra would likely lead to more significant deviations from the observed spectra than those we found when reconstructing the velocities alone from the ACF. This is because the CCF, from which the ACF is derived, combines information from a large list of spectral lines, even though different lines are known to respond differently to stellar variability (Dravins et al. 1981; Toner & Gray 1988). Furthermore, variability of the continuum contributes very little to the CCF. Future studies could analyse the residual spectra to gain insights into more subtle ways in which stellar variability manifests in the observed spectra (e.g. Davis et al. 2017; Thompson et al. 2017; Dumusque 2018; Wise et al. 2018; Cretignier et al. 2020).

5.3.5 Multiple CCF masks

The results of full-spectrum reconstruction could inform the development of additional stellar variability indicators and/or the design of CCF masks. As noted above, summarizing the spectrum with a single CCF averages over the different responses of lines which are more or less sensitive to stellar activity. For this study, we have used the CCF generated by the HARPS-N DRS. However, one could design multiple CCF masks to compute multiple CCFs at each observation epoch. The choice of mask could be based on astrophysical insights or machine-learning approaches. Either way, each mask would include spectral lines exhibiting common patterns of response to stellar variability, so as to provide greater sensitivity for recognizing line shape variations.

Once CCFs have been computed using multiple masks, the ACF can be computed for each CCF-mask pair. The underlying algorithm presented in Section 3 can be applied to compute time-domain subspaces either separately for each mask or simultaneously. The resulting velocities derived from each mask could be analysed separately for diagnostic purposes and then combined for inference (e.g. Zechmeister et al. 2018). The use of multiple masks naturally leads to multivariate time series for estimates of the velocity and stellar variability indices.

5.3.6 Time-domain modelling

Stellar variability generally distinguishes itself from planet-induced orbital motion both in the wavelength domain (e.g. relative depths of lines, line shapes) and in the time domain (e.g. deviations from a strictly Keplerian signal). We caution, however, that subtler effects such as activity-driven changes in the stellar gravitational redshift may manifest as pure shifts at amplitudes up to 10 cm s^{-1} (Cegla et al. 2012). In this paper, we have described an improved algorithm for measuring RVs in the presence of stellar variability that does not make use of time-domain information. That is, the computation of $\mathbf{v}_{\|}$ depends only on the ensemble of measured CCFs, but not on the times at which the observations were taken. For the purpose of demonstrating our algorithm, we have used traditional maximum likelihood estimation based solely on the cleaned velocities. Our algorithm naturally produces additional stellar variability indicators [i.e. $\mathbf{U}_A(t)$ and $\mathbf{U}_C(t)$] that could be modelled simultaneously with the velocities, following methods developed by Rajpaul et al. (2015) and generalized in Jones et al. (2017).

We recommend that future simulations explore the potential for further improvements in the sensitivity to small planets by combining algorithms for utilizing information in the wavelength and time domains.

5.3.7 Integrating into Doppler planet survey toolboxes

The algorithm developed in his paper can be implemented efficiently using a standard linear algebra toolbox. Thus, it can be readily integrated into existing or future software packages to analyse Doppler planet search observations. For example, one could inspect the posterior marginalized over all parameters except orbital period (i.e. a periodogram), similar to Mortier et al. (2015), but replacing the standard likelihood with equation (14). When considering observations of a star potentially hosting multiple planets, a periodogram of the reconstructed CCFs could be applied iteratively, removing one signal at a time. Alternatively, one could apply sparse regression techniques, i.e. fit for the semi-amplitudes of many periodic signals simultaneously, while applying regularization to the semi-amplitude, so as to find a family of maximum likelihood solutions for each plausible number of planets.

This could be implemented efficiently for an a priori unknown number of signals using either the alternating direction method of multipliers (ADMM) algorithm or the spectral projected-gradient algorithm (as in Hara et al. 2017). Once the putative orbital periods have been identified, then MCMC-based techniques can be used to perform parameter estimation (e.g. Ford 2006; Nelson, Ford & Payne 2014) replacing the standard likelihood with our equation (14). Then, one could compare the Bayes factor or ratio of marginalized likelihoods (i.e. 'evidences') for models with various numbers of planets. While there is still active research in finding efficient and robust algorithms for estimating Bayes factors for models with several planets, replacing the likelihood with equation (14) would be algorithmically straightforward and is expected to add only a modest additional cost.

5.3.8 Design of Doppler planet surveys

The observing strategy of Doppler planet surveys (e.g. number of observations per star, distribution of duration between observations) can have a significant impact on the sensitivity for detecting planets and its dependence on orbital period. Simulations of Doppler planet surveys have been used to inform survey design choices (e.g. Ford 2008; Burt et al. 2018; Cloutier et al. 2018; Hall et al. 2018).

Previous studies have typically ignored or adopted simplistic models for intrinsic stellar variability (e.g. assuming that a fixed fraction of spurious velocities due to stellar variability can be corrected). Our algorithm for simultaneously inferring the effects of stellar variability and planetary signals could be readily incorporated into more advanced survey simulations incorporating explicit models of the effects of stellar activity such as that published recently by Damasso et al. (2019).

Now that there is a concrete and computationally efficient strategy for mitigating stellar variability, we recommend that future studies conduct new survey simulations to compare candidate Doppler planet survey strategies. We anticipate that our method will significantly increase the value of observing the same star many times, since we substantially reduce the correlation of inferred velocities at different times. On the other hand, it may be that the improved ability to mitigate stellar activity reduces the importance of obtaining a dense set of observations over the rotation period or active region lifetime. Monte Carlo simulations are necessary to understand the interaction of these two effects and the implications for future planet surveys. In addition to informing survey strategies that do not make use of information from previous observations, our algorithms could be folded into adaptive scheduling algorithms that maximize a merit

function (or minimize a cost function), so as to maximize the efficiency of a Doppler survey (e.g. Ford 2008).

ACKNOWLEDGEMENTS

This article is a result of collaborative scholarly efforts from the residency of the Research Group on Big Data and Planets at the Israel Institute for Advanced Studies. We acknowledge valuable discussions with Tsevi Mazeh, Eric Feigelson, Shay Zucker, Lev Tal-Or, and Dovi Poznanski. We thank the anonymous referee for numerous insightful suggestions that led to major improvements, including the use of leave-one-out cross-validation and streamlining of the mathematical approach used in Section 4.2. ACC acknowledges support from the Science and Technology Facilities Council (STFC) consolidated grant number ST/R000824/1 and UKSA grant ST/R003203/1. EBF acknowledges support from NSF award #1616086, NASA grant #80NSSC18K0443, and Heising-Simons Foundation grant #2018-0851. This work was supported by a grant from the Simons Foundation/SFARI (675601, EBF). EBF acknowledges the support of the Ambrose Monell Foundation and the Institute for Advanced Study. EBF acknowledges support from the Penn State Eberly College of Science, Department of Astronomy & Astrophysics, Institute for Computational & Data Sciences, the Center for Exoplanets and Habitable Worlds, the Center for Astrostatistics. DFP acknowledges NASA award number NNX16AD42G. This work was performed partly under contract with the California Institute of Technology (Caltech)/Jet Propulsion Laboratory (JPL) funded by NASA through the Sagan Fellowship Program executed by the NASA Exoplanet Science Institute (RDH). AM acknowledges support from the senior Kavli Institute Fellowships, funded by the Kavli Foundation. XD is grateful to the Branco Weiss Fellowship – Society in Science for its financial support. HMC acknowledges financial support from the National Centre for Competence in Research (NCCR) PlanetS, supported by the Swiss National Science Foundation (SNSF), as well as a UK Research and Innovation Future Leaders Fellowship. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Framework Programme (grant agreement No. 851555). This work has been carried out in the framework of the National Centre for Competence in Research *PlanetS* supported by the Swiss National Science Foundation (SNSF). The authors gratefully acknowledge the support of the International Team 453 by the International Space Science Institute (Bern, Switzerland). Based on observations made with the Italian *Telescopio Nazionale Galileo* (TNG) operated by the *Fundación Galileo Galilei* (FGG) of the *Istituto Nazionale di Astrofisica* (INAF) at the *Observatorio del Roque de los Muchachos* (La Palma, Canary Islands, Spain). The HARPS-N project has been funded by the Prodex Program of the Swiss Space Office (SSO), the Harvard University Origins of Life Initiative (HUOLI), the Scottish Universities Physics Alliance (SUPA), the University of Geneva, the Smithsonian Astrophysical Observatory (SAO), the Italian National Astrophysical Institute (INAF), the University of St Andrews, Queen's University Belfast, and the University of Edinburgh. The citations in this paper have made use of NASA's Astrophysics Data System Bibliographic Services. This publication makes use of The Data & Analysis Center for Exoplanets (DACE), which is a facility based at the University of Geneva (CH) dedicated to extrasolar planets data visualisation, exchange and analysis. DACE is a platform of the Swiss National Centre of Competence in Research (NCCR) PlanetS, federating the Swiss expertise in Exoplanet research. The DACE platform is available at https://dace.unige.ch.

DATA AVAILABILITY

The HARPS-N solar data products for the first 3 yr of the period described in this paper are publicly available through the DACE platform (https://dace.unige.ch) developed in the frame of PlanetS. All other research data underpinning this publication (https://doi.org/10.17630/9ec51274-cbea-445f-b188-112f4734c6e9) and the PYTHON code and notebook used to prepare all diagrams in this paper (https://doi.org/10.17630/957d6320-1969-43c8-b6bb-a8ae53d1f657) will be made available through the University of St Andrews Research Portal.

REFERENCES

Adler R. L., Konheim A. G., 1962, Proc. Amer. Math. Soc., 13, 425
Aigrain S., Pont F., Zucker S., 2012, MNRAS, 419, 3147
Baranne A. et al., 1996, A&AS, 119, 373
Bouchy F., Pepe F., Queloz D., 2001, A&A, 374, 733
Brewer J. M. et al., 2020, AJ, 160, 67
Burt J., Holden B., Wolfgang A., Bouma L. G., 2018, AJ, 156, 255
Cegla H., 2019, Geosc, 9, 114
Cegla H. M. et al., 2012, MNRAS, 421, L54
Cegla H. M., Watson C. A., Shelyag S., Mathioudakis M., Moutari S., 2019, ApJ, 879, 55
Celisse A., 2014, Annals of Statistics, 42, 1879
Chaplin W. J., Cegla H. M., Watson C. A., Davies G. R., Ball W. H., 2019, AJ, 157, 163
Cloutier R., Doyon R., Bouchy F., Hébrard G., 2018, AJ, 156, 82
Collier Cameron A. et al., 2019, MNRAS, 487, 1082
Cosentino R. et al., 2014, SPIE, 91478C, SPIE.9147
Cretignier M., Dumusque X., Allart R., Pepe F., Lovis C., 2020, A&A, 633, A76
Damasso M., Pinamonti M., Scandariato G., Sozzetti A., 2019, MNRAS, 489, 2555
Davis A. B., Cisewski J., Dumusque X., Fischer D. A., Ford E. B., 2017, ApJ, 846, 59
de Beurs Z. L. et al., 2020, preprint (arXiv:2011.00003)
Dravins D., Lindegren L., Nordlund A., 1981, A&A, 96, 345
Dumusque X. et al., 2012, Nature, 491, 207
Dumusque X. et al., 2015, ApJ, 814, L21
Dumusque X., 2018, A&A, 620, A47
Dumusque X., Boisse I., Santos N. C., 2014, ApJ, 796, 132
Dumusque X. et al., 2021, A&A, 648, 10
Ford E. B., 2006, ApJ, 642, 505
Ford E. B., 2008, AJ, 135, 1008
Gilbertson C., Ford E. B., Jones D. E., Stenning D. C., 2020, ApJ, 905, 155
Giorgini J. D. et al., 1996, BAAS, 28, 1158
Hall R. D., Thompson S. J., Handley W., Queloz D., 2018, MNRAS, 479, 2968
Hara N. C., Boué G., Laskar J., Correia A. C. M., 2017, MNRAS, 464, 1220
Haywood R. D. et al., 2014, MNRAS, 443, 2517
Haywood R. D. et al., 2020, ApJ, preprint (arXiv:2005.13386)
Holzer P., Cisewski-Kehe J., Fischer D., Zhao L., 2020, preprint (arXiv:2005.14083)
Jones D. E., Stenning D. C., Ford E. B., Wolpert R. L., Loredo T. J., Dumusque X., 2017, preprint (arXiv:1711.01318)
Jurgenson C. et al., 2016, SPIE, 99086T, SPIE.9908
Langellier N. et al., 2020, ApJ, preprint (arXiv:2008.05970)
Ljung G. M., Box G. E. P., 1978, Biometrika, 65, 297
Lomb N. R., 1976, Ap&SS, 39, 447
Mégevand D. et al., 2014, SPIE, 91471H, SPIE.9147
Meunier N., Lagrange A.-M., Desort M., 2010, A&A, 519, A66
Meunier N., Lagrange A.-M., Borgniet S., Rieutord M., 2015, A&A, 583, A118
Mortier A., Faria J. P., Correia C. M., Santerne A., Santos N. C., 2015, A&A, 573, A101

Nelson B., Ford E. B., Payne M. J., 2014, ApJS, 210, 11
Ninan J. P. et al., 2018, SPIE, 107092U, SPIE10709
Pepe F. et al., 2002, A&A, 388, 632
Pepe F. et al., 2021, A&A, 645, A96
Phillips D. F. et al., 2016, SPIE, 99126Z, SPIE.9912
Queloz D. et al., 2001, A&A, 379, 279
Quirrenbach A. et al., 2014, SPIE, 9147, 91471F
Rajpaul V., Aigrain S., Osborne M. A., Reece S., Roberts S., 2015, MNRAS, 452, 2269
Saar S. H., Donahue R. A., 1997, ApJ, 485, 319
Scargle J. D., 1982, ApJ, 263, 835
Scholz F. W, Stephens M. A., 1987, Journal of the American Statistical Association, 82, 918
Schwab C. et al., 2016, SPIE, 99087H, SPIE.9908
Suárez Mascareño A. et al., 2020, A&A, 639, A77
Thompson S. J. et al., 2016, SPIE, 9908, 99086F
Thompson A. P. G., Watson C. A., de Mooij E. J. W., Jess D. B., 2017, MNRAS, 468, L16
Toner C. G., Gray D. F., 1988, ApJ, 334, 1008
Wise A. W., Dodson-Robinson S. E., Bevenour K., Provini A., 2018, AJ, 156, 180
Wright J. T., Robertson P., 2017, RNAAS, 1, 51
Zechmeister M. et al., 2018, A&A, 609, A12
Zechmeister M., Kürster M., 2009, A&A, 496, 577
Zhao J., Tinney C. G., 2020, MNRAS, 491, 4131

APPENDIX A: CALCULATION OF PROFILE DERIVATIVES

Derivatives of the rows of the CCF with respect to velocity are required for applying RV shifts to the rows of the CCF in both correction from the barycentric to the heliocentric reference frame (Section 2.1) and for injecting synthetic orbital reflex motion signals (Section 4).

The numerical derivatives are calculated with a simple differencing scheme:

$$\mathbf{C}'(v_i, t_j) = \frac{\mathbf{C}(v_{i+1}, t_j) - \mathbf{C}(v_{i-1}, t_j)}{2h} \tag{A1}$$

and

$$\mathbf{C}''(v_i, t_j) = \frac{\mathbf{C}(v_{i+1}, t_j) - 2\mathbf{C}(v_i, t_j) + \mathbf{C}(v_{i-1}, t_j)}{h^2}. \tag{A2}$$

Here h is the velocity increment per CCF sampling interval in velocity. The current HARPS-N DRS outputs CCFs on to a standard grid of $l = 161$ velocities with $h = 0.25$ km s^{-1}. Here $\mathbf{C}(v_i, t_j)$, $\mathbf{C}'(v_i, t_j)$, and $\mathbf{C}''(v_i, t_j)$ are estimates of the CCF and its derivatives, which we describe below.

Simple numerical derivatives computed directly from the data have the undesirable property that they amplify noise. If the profile shape was time-invariant, this problem could be overcome by fitting the Taylor-series model derived from the mean profile, to a sequence of CCFs shifted by orbital reflex motion. However, magnetically active regions rotating across the star's visible hemisphere distort the CCF profile (Meunier et al. 2010). This time variation in the shape of the profile alters the RV measured by fitting a fixed profile or parametric function to the CCF. Currently, it is unclear whether the cross-terms between such intrinsic stellar variability and small Doppler shifts caused by planets are sufficiently small to be neglected. Therefore, this paper attempts to estimate the derivative of the CCF at each observation time and focuses on solar data for which very high signal-to-noise data is available.

Rather than estimating the derivative from a constant mean spectrum, we estimate the derivative of the CCF at each epoch t_j based on a reduced-rank reconstruction of the CCF using the mean profile and the $k_{\max} = 10$ leading principal components obtained by singular-value decomposition (SVD) of the full ensemble of CCFs. The optimal choice of $k_{\max}$ that best separates the effects of profile shape changes from the effects of oversampled photon noise is calculated as described in Appendix B1.

APPENDIX B: TAYLOR-SERIES VELOCITY MEASUREMENT FROM CCF

The RVs derived by the HARPS-N DRS are obtained by fitting a Gaussian profile to the activity-distorted CCF. Pure Doppler shifts simply displace the distorted profile in velocity. Following the formulation of Bouchy, Pepe & Queloz (2001) for line shifts that are small compared to the line width, we measure Doppler shifts using a first-order Taylor-series approximation to the instantaneous CCF profile shape at the time t_j of the jth observation. For shifts much smaller than the 0.25 km s^{-1} velocity increment per CCF array element, the first derivative $\mathbf{C}'(v_i, t_j)$ of the instantaneous profile $\mathbf{C}$ suffices to measure the displacement of the mean-subtracted CCF $\mathbf{R}(v_i, t_j) = \mathbf{C}(v_i, t_j) - \langle \mathbf{C}(v_i) \rangle$ where $\langle \mathbf{C}(v_i) \rangle$ is the time-averaged profile:

$$\mathbf{v}(t_j) = \frac{\mathbf{R}(v, t_j)^T \cdot \Sigma^{-1}(t_j) \cdot \mathbf{C}'(v, t_j)}{\mathbf{C}'(v, t_j)^T \cdot \Sigma^{-1}(t_j) \cdot \mathbf{C}'(v, t_j)}. \tag{B1}$$

The derivatives $\mathbf{C}'(v_i, t_j)$ are calculated from a reduced-rank representation of $\mathbf{C}(v_i, t_j)$ as described in Appendix A. The covariance matrix Σ is calculated as described in Appendix B1 below. The corresponding variances of the estimated RVs are given by

$$\mathrm{Var}(\mathbf{v}(t_j)) = \frac{1}{\mathbf{C}'(v, t_j)^T \cdot \Sigma^{-1}(t_j) \cdot \mathbf{C}'(v, t_j)}. \tag{B2}$$

In order to ensure that the sampling of the CCF did not bias the scale of the derived velocities, we performed linear regression of both the DRS velocities and the Taylor-series velocities against the solar barycentric RV computed with JPL HORIZONS (Giorgini et al. 1996). The scale factors differed from unity by 0.8 and 2.4 per cent, respectively. The barycentric RVs were scaled by these factors for all subsequent calculations, to ensure that the velocities were transformed correctly to the heliocentric reference frame. The agreement between the two sets of corrected velocities is illustrated in Fig. B2.

B1 Covariance matrix for time-varying CCFs

For calculating instantaneous RVs and their precisions from equations (B1) and (B2), the covariance matrix Σ must contain information about the SNR of each observation and the correlations between neighbouring pixels. The covariances between different pixels in the time series of CCFs fall into two categories.

The first is systematic uncertainty arising from temporal variability of the profile shape. The **sample** covariance matrix $\Sigma = \mathrm{Cov}(\mathbf{R}) = \mathbf{R}^T \cdot \mathbf{R}/m$**estimates** the systematic covariances between the columns of $\mathbf{R}$, where each column has m elements, one per row of the CCF time series. These arise from correlations between different parts of the line profile.

The second is random measurement error arising from the finite SNR of the original spectrum. The CCF sampling interval is matched to the velocity increment per CCD pixel of the instrument. Although neighbouring pixel values in the original spectrum are statistically independent, interpolation during rebinning and calculation of the CCF introduces correlations between neighbouring CCF samples.

We therefore need to account for the correlations between CCF elements at neighbouring values of velocity.

The dispersion of HARPS-N gives a near-constant velocity increment per physical CCD pixel in the dispersion direction of 0.82 km s^{-1} (Dumusque et al. 2020). The DRS delivers a default CCF that is also sampled in velocity increments of 0.82 km s^{-1}, so uncorrelated photon-noise fluctuations in the extracted spectra are spread over about 2 CCF velocity increments after rebinning and interpolation.

The systematic covariances and the instantaneous correlated-noise pattern are unrelated to each other, and enter into the velocity and error calculation in different ways. To estimate the independent systematic uncertainty that affects every observation through profile shape changes, the systematic covariance matrix should be used with equation (B2), separately from the small-scale noise pattern.

The full covariance matrix can be computed via singular-value decomposition of **R**, as defined in equation (4):

$$\frac{1}{m}\mathbf{R}^T \cdot \mathbf{R} = \frac{1}{m}(\mathbf{U}_C \cdot \mathrm{diag}(\mathbf{S}_C) \cdot \mathbf{P}_C^T))^T \cdot (\mathbf{U}_C \cdot \mathrm{diag}(\mathbf{S}_C) \cdot \mathbf{P}_C^T)) = \frac{1}{m}(\mathbf{P}_C \cdot \mathrm{diag}(\mathbf{S}_C^2) \cdot \mathbf{P}_C^T). \quad \text{(B3)}$$

This approach gives results that are identical to a direct evaluation of the unweighted sample covariance matrix, with a maximum fractional deviation of 1 part in 1000. More importantly, it allows us to calculate a reduced-rank version of the covariance matrix, using only the leading $k_{\max}$ principal components. With an appropriate choice of $k_{\max}$, a reduced-rank reconstruction of the mean-subtracted CCF residuals

$$\mathbf{R}_k \equiv \sum_{k=1}^{k_{\max}} \mathbf{U}_{C,k} \cdot \mathrm{diag}(\mathbf{S}_{C,k}) \cdot \mathbf{P}_{C,k}^T, \quad \text{(B4)}$$

gives a representation of the large-scale temporal covariance pattern that can be written as

$$\frac{1}{m}\mathbf{R}_{k_{\max}}^T \cdot \mathbf{R}_{k_{\max}} = \frac{1}{m}(\mathbf{P}_{C,k_{\max}} \cdot \mathrm{diag}(\mathbf{S}_{C,k_{\max}}^2) \cdot \mathbf{P}_{C,k_{\max}}^T). \quad \text{(B5)}$$

This produces the low-pass filtered covariance matrix shown in the upper panel of Fig. B1.

When this reduced-rank version is subtracted from the full covariance matrix, we obtain the covariance matrix of the remaining high-frequency noise, as shown in the middle panel of Fig. B1. Its strongest feature is a ridge of covariance with an approximately triangular cross-section, whose amplitude depends on the average SNR of the original spectra, and whose width reflects the sampling of the CCF. The peak variance along its diagonal is about 100 times smaller than the peak amplitude of the large-scale covariance pattern.

The profile of the ridge is seen clearly when the values of the covariance matrix are averaged along diagonals parallel to the leading diagonal, and plotted against velocity lag relative to the leading diagonal, as shown in the bottom panel of Fig. B1. Using a triangular fit

$$T(v\,|A, v_0, \delta v) = A\ \max\left(1 - \frac{|v - v_0|}{\delta v}, 0\right) \quad \text{(B6)}$$

to this average profile we obtain an average base half-width $\delta v = 0.82\,\mathrm{kms}^{-1}$. The base width of the ridge perpendicular to the diagonal is therefore about 2 CCF velocity increments, as expected from matching of the CCF sampling interval to the spectrograph resolution element and linear interpolation in the calculation of the CCF.

We optimize $k_{\max}$ to give a clear separation between the large-scale column covariances and the covariances between neighbouring elements in each row arising from oversampling of photon noise. To achieve this we divide the peak value in the bottom panel of Fig. B1 by the range of all other diagonal means with lags differing by more than 1 km s^{-1} (slightly more than the CCF sampling interval) in velocity from the leading diagonal. This ratio shows a well-defined peak at $k_{\max} = 6$, where the separation between the two variance patterns is optimized.

The row variances σ_j^2 of $\mathbf{R} - \mathbf{R}_{k_{\max}}$ reflect the SNR of the individual observations, so we use them as the scale factors $A_j = \sigma_j^2$

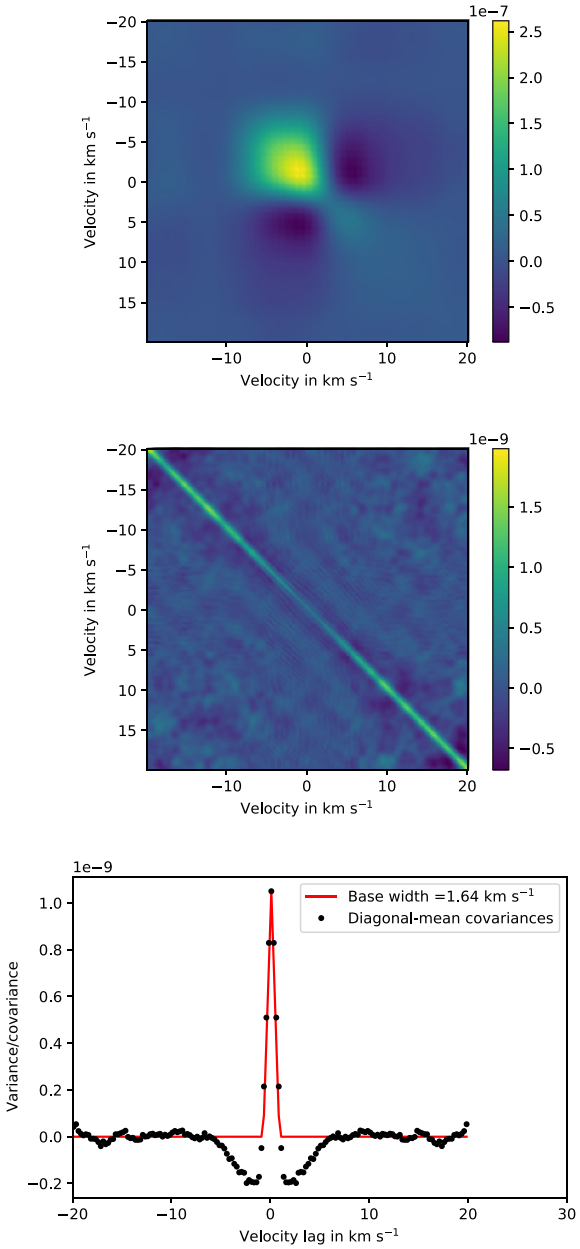

Figure B1. Upper panel: reduced-rank ($k_{\max} = 10$) covariance matrix of the sequence of CCFs residuals, showing the large-scale covariance due to profile variability. Middle panel: Residual covariance matrix obtained by subtracting the reduced-rank representation from the full covariance matrix. Lower panel: Mean residual covariance along diagonals as a function of horizontal velocity offset from the leading diagonal. The diagonal ridge is triangular in cross-section, with base width 1.64 km s^{-1}, which is equivalent to two physical pixels on the spectrograph CCD.

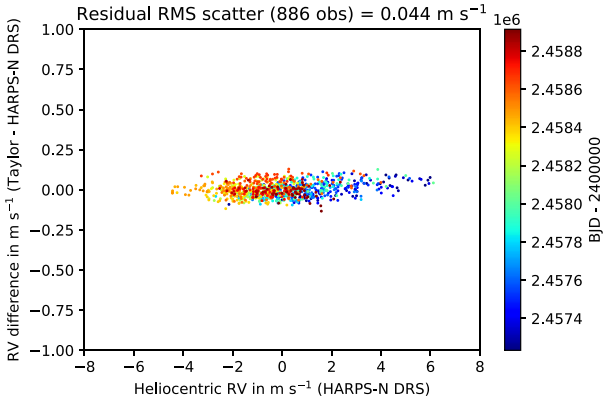

Figure B2. Apparent RVs derived from the CCF time series in the heliocentric frame via the Taylor approximation (equation B1) versus apparent RVs from the HARPS-N DRS after applying the barycentric to heliocentric correction. The mean values of both sets of velocities have been subtracted. The standard deviation of the Taylor-series velocities is 2.287 ms $^{-1}$. The two estimators of radial velocity are highly correlated and have slope near unity. The points are colour-coded with BJD-2400000.0. The RMS scatter in the differences between the two sets of measured velocities is 0.044 m s^{-1}.

of the triangular profile. This leads to the following model for the covariance matrix of the high-frequency noise of an observation made at time t_j:

$$\Sigma_{\xi,\eta}(t_j) \simeq \sigma_j^2 \times \max\left(1 - \frac{|v_\xi - v_\eta|}{\delta v}, 0\right). \quad \text{(B7)}$$

This form of the covariance matrix is suitable for calculating the velocities and their precision. The resulting velocities differ from those obtained assuming spatially uncorrelated noise only at the 15 cm s^{-1} level. Their precision also depends, however, on the scatter introduced by profile-shape changes. When fitting an orbit to the velocities, it is therefore more appropriate to use velocity variances calculated with equation (B2) with a covariance matrix that also includes the reduced-rank model of the covariances arising from the time-varying shape of the CCF:

$$\Sigma_{\xi,\eta}(t_j) \simeq \sigma_j^2 \times \max\left(1 - \frac{|v_\xi - v_\eta|}{\delta v}, 0\right) + \frac{1}{m}\mathbf{R}^T_{k_{\max}} \cdot \mathbf{R}_{k_{\max}}. \quad \text{(B8)}$$

B2 Comparison with DRS velocities

We use this covariance matrix with equations (B1) and (B2) to estimate the RVs $\delta\mathbf{v}(t_j)$ and their variances due to photon noise and profile-shape changes.

In Fig. B2 we show that RVs measured using this method are almost identical to those reported by the HARPS-N DRS. The overall RMS scatter in the differences between the two estimates of the velocity is 0.044 m s^{-1}.

The formal errors derived from equation (B2) are typically 0.11 m s^{-1}, though the data set includes a small number of days of sparse data yielding uncertainties greater than 0.2 m s^{-1}. Such small error estimates are not surprising. The formal uncertainties computed for individual observations by the DRS indicate an average of 0.24 m s^{-1} photon noise per exposure. We average many such exposures per day, so the photon noise is insignificant in comparison to the systematic uncertainties arising from profile-shape variations, zero-point errors and calibration drift.

The independent systematic variances arising from large-scale profile-shape changes are re-computed from equation (B2) using the reduced-rank covariance matrix $\Sigma = \mathrm{Cov}(\mathbf{R}_{k_{\max}})$ from equations (B3) and (B4). The resulting systematic error per observation is 0.82 ± 0.017 m s^{-1}, which is very close to the RMS amplitude of the scatter in Fig. 9.

APPENDIX C: SCALPELS ALGORITHM

C1 Simple blind search

Input: C (2D CCF time series), **x** (CCF array velocities), $\mathbf{v}_{\rm obs}$ (RV measurements), $\sigma(\mathbf{v}_{\rm obs})$ (RV error estimates).

(i) Compute ACF with equation (3) and normalize.
(ii) Compute SV decomposition of ACF with equation (5).
(iii) Perform MAD clip at appropriate threshold on columns of $\mathbf{U}_A$.
(iv) Recompute SVD of surviving ACF rows.
(v) Withhold each row of ACF in turn, perform SVD, reconstruct corresponding row of $\hat{\mathbf{U}}_A$ (equations 9, 10).
(vi) Keep columns of $\mathbf{U}_A$ s.t. $\mathrm{MAD}(\mathbf{U}_k^T - \hat{\mathbf{U}}_k^T)/\mathrm{MAD}(\hat{\mathbf{U}}_k^T) \ll 1$.
(vii) Compute $\hat{\alpha}$ (equation 6).
(viii) Rearrange elements of $\hat{\alpha}$ and columns of $\mathbf{U}_A$ in order of descending $|\delta\chi^2(\mathbf{v}_\perp)|$ (Sect 3.3).
(ix) Select number of basis vectors to keep (Sect 3.3).
(x) Compute $\mathbf{v}_\parallel$ (equation 7) in reduced-rank basis.
(xi) Compute $\mathbf{v}_\perp = \mathbf{v}_{\rm obs} - \mathbf{v}_\parallel$.

Return: $\mathbf{v}_\parallel$ (shape-driven RV), $\mathbf{v}_\perp$ (shift-driven RV).

C2 Simultaneous sinusoidal fit

Input: C (2D CCF time series), **x** (CCF array velocities), $\mathbf{v}_{\rm obs}$ (RV measurements), $\sigma(\mathbf{v}_{\rm obs})$ (RV error estimates), $\{\omega_1, \cdots, \omega_n\}$ (orbital frequencies for n planets).

Perform steps (i) - (vii) above and see Section 4.2.

(i) Compute $\mathbf{F} = \{\cos\omega_1 t_j, \sin\omega_1 t_j, \cdots, \cos\omega_n t_j, \sin\omega_n t_j\}$.
(ii) Compute reduced-rank covariance matrix (equation B5).
(iii) Compute row variances σ_j^2 of $\mathbf{R} - \mathbf{R}_{k_{\max}}$ (Appendix B).
(iv) Compute model of full covariance matrix (equation B8).
(v) Compute $\mathbf{C}'$ and $\mathrm{Var}(\mathbf{v}(t_j))$ (equation B2).
(vi) Construct $\Sigma = \mathrm{Diag}(\mathrm{Var}(\mathbf{v}(t_j)))$.
(vii) Construct $\mathbf{P}_\perp = (\mathbf{I} - \mathbf{U}_A \cdot \mathbf{U}_A^T)$ in reduced-rank basis.
(viii) Compute $\mathbf{v}_\perp = \mathbf{P}_\perp \cdot \mathbf{v}_{\rm obs}$.
(ix) Compute $\mathbf{F}_\perp = \mathbf{P}_\perp \cdot \mathbf{F}$.
(x) Solve equation (13) to obtain $\theta_{\rm orb}$.
(xi) Compute $\mathrm{Var}(\theta_{\rm orb}) = 1/\mathrm{Diag}(\mathbf{F}_\perp^T \cdot \Sigma^{-1} \cdot \mathbf{F}_\perp)$.
(xii) Compute $\mathbf{v}_{\rm orb} = \mathbf{F} \cdot \theta_{\rm orb}$.
(xiii) Compute $\mathbf{v}_{\rm resid} = \mathbf{v}_\perp - \mathbf{F}_\perp.\theta_{\rm orb})$.
(xiv) Compute $\mathbf{v}_\parallel = \mathbf{v}_{\rm obs} - \mathbf{v}_\perp$.

Return: RV amplitudes and variances, $\mathbf{v}_\parallel$, $\mathbf{v}_\perp$.

[1]*SUPA School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews KY16 9SS, UK*
[2]*Israel Institute for Advanced Studies, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, Israel*
[3]*Center for Exoplanets and Habitable Worlds, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA 16803, USA*
[4]*Department of Astronomy and Astrophysics, 525 Davey Laboratory, The Pennsylvania State University, University Park, PA 16803, USA*
[5]*Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16803, USA*

[6]*School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel*
[7]*Department of Physics, University of Oxford, Keble Road Oxford, OX1 3RH, UK*
[8]*Observatoire Astronomique de l'Université de Génève, 51 Chemin des Maillettes, Ch-1290 Sauverny, Switzerland*
[9]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 01238, USA*
[10]*Astrophysics Group, University of Exeter, Exeter EX4 2QL, UK*
[11]*Astrophysics Group, Cavendish Laboratory, University of Cambridge, J.J. Thomson Avenue, Cambridge CB3 0HE, UK*
[12]*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[13]*DTU Space, National Space Institute, Technical University of Denmark, Elektrovej 328, DK-2800 Kgs. Lyngby, Denmark*
[14]*INAF – Fundación Galileo Galilei, Rambla José Ana Fernandez Pérez 7, E-38712 Breña Baja, Tenerife, Spain*
[15]*Physics Department, University of Warwick, Coventry CV4 7AL, UK*
[16]*INAF – Osservatorio Astronomico di Palermo, Piazza del Parlamento 1, I-90134 Palermo, Italy*
[17]*INAF - Osservatorio Astronomico di Cagliari, via della Scienza 5, I-09047 Selargius, Italy*
[18]*Dip. di Fisica e Astronomia Galileo Galilei – Università di Padova, Vicolo dell'Osservatorio 2, I-35122 Padova, Italy*
[19]*INAF – Osservatorio Astrofisico di Torino, via Osservatorio 20, I-10025 Pino Torinese, Italy*
[20]*Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, University Road, Belfast BT7 1NN, UK*

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