November 13, 2020: ESSPN Check-In 2

At this meeting, all teams were allowed to present their methods with accompanying slides. Questions were asked through the Zoom chat. The Zoom chat and accompanying slides have been uploaded to the Yale Box folder and shared on the slack.

Anyone may leave comments on this document on Box by highlighting text in the preview window; the option to leave a comment will then pop up.

Andrews

The Andrews team is implementing **SCALPLES** (Self-Correlation Analysis of Line Profiles for Extraction of Low-amplitude Shifts), which was developed along with Penn State team members. SCALPLES works by taking the residual time series from EXPRES using CCFs. CCFs generated by Eric Ford were used for the results submitted.

The autocorrelation function of any spectrum or CCF is invariant to translation. This means this autocorrelation function is sensitive to line shape changes without interference from shifts due to planets. A time series of autocorrelation functions (of the CCF) is used and put through single-value decomposition that describes the time series as a set of orthonormal eigenvectors. These eigenvectors can be thought of as modes of line-shape changes guided by time-domain coefficients.

Using the orthonormal basis, you can recreate a model of the RV shift due to shape changes. The results with EXPRES data of HD 101501 show that the model and the RVs are pretty strongly correlated, suggesting that most of the RV variation is shape driven.

CCA

The CCA team is developing a method centered on the idea of using **residuals from wobble** templates to isolate line shape variations. They then determine how correlated a set of meta data (i.e. CBC RVs or BIS) is with the residuals to wobble templates; i.e. to changes in line shape, by regressing the wobble RVs against different activity indicators.

The vector of these regressions can be used to derive the amount of the RV shift that is attributed to line shape variations, and therefore presumably to activity (see slide 8). The amount of deviation that can be derived will change with the meta data that the wobble residuals are regressed against, and the relative information content of this meta data.

Geneva

Geneva used a series of published code on the EXPRES spectra. First, the spectra were merged to join all orders. **RASSINE** (Cretignier+ 2020) was used to normalize the spectra. **Kit-Cat** (Cretignier+ 2020) was used to carefully select lines. **YARARA** (Cretignier+ in prep) was used to form a good template and remove tellurics.

The cleaned spectra were used to derive **LBL RVs** (as per the Dumusque paper). A PCA was then run on the LBL RVs. The slides show the first three components and their periodograms. The second component, for example, shows a significant period at 17 days, the stellar rotation period, and so likely traces back to activity.

OxBridGen

The first method OxBridGen implemented was **pairwise spectral GP modeling**. As a qualitative sketch, first all spectra are modeled with GPs. Each GP model is then compared with every other model, including an RV shift accompanying each model. This gives the pairwise shifts, which allows for an easy calculation of total differential RV shifts for all exposures.

The pairwise model makes the process more computationally convenient. The GP only has to deal with two exposures' worth of data points at a time and there are fewer free parameters that need to be sampled at once. This framework also avoids the need to form a template that can smear out features and remove time coherent symbols.

To deal with activity, the spectra get divided into small chunks. You can then consider time or wavelength dependent properties of each chunk. Chunks with high error or strong correlation with activity indicators may be excluded. There are lots of ways to identify correlation with chunks. Next steps would be adding extra covariance terms to account for changes rather than simply throwing away chunks.

The second method OxBridGen implemented was a **multi-dimensional Gaussian processes** method. This method assumes that all activity indicators can be explained by the same Gaussian process. Slide 17 show an animation of a GP model for RVs (upper panel) and an indicator (lower panel). This animation demonstrates that the biggest changes in both panels happen at the same time.

The photometry was used to get a prior on the rotation period (due to sparse sampling of the spectral data). A multi-dimensional GP was then run for different combinations of RV + one activity indicator using a quasi-periodic kernel. The results submitted were using H-alpha emission, which returned an RMS of 0.33 m/s. Other indicators gave residuals of 40-50 cm/s.

Penn State

The Penn State team focused on wavelength domain methods. Firstly, Penn State implemented a slightly different take on **SCALPLES**. They used their own RVs, rather than the provided ones, and incorporated less basis vectors in the final model.

They also implemented a **doppler-constrained PCA (DCPCA)** method (Jones+ 2017/2020). This returned unreasonably precise residuals, and so a linear regression to the basis vectors was used instead. They will move back to using GPs on data sets with more data points.

Each method was tried using three different types of CCF masks. The first type of mask, as a point of reference, is the classic ESPRESSO G9 mask trimmed for tellurics and order boundaries. Masks were also made using the spectrum itself and the VALD line data base. Lines with inconsistent FWHM from observation to observation were flagged. Line blends according to the VALD line database (even if they couldn't be seen) were flagged. The two other masks were built up in this way with different filtering parameters. More filtering means more bias (since you are using less lines) but less variance (since the lines you are using are cleaner). The ESPRESSO mask has a factor of 2 to 3 more lines that those constructed from the spectrum itself + VALD.

Moving forward, they will also be trying a **multivariate GP** model using scores from the basis vectors for both SCALPES and the DCPCA method in addition to trying the different CCF masks. They will also try to

incorporate a telluric model to allow correcting for tellurics rather than just throwing out those regions of the spectra. The **Fiesta** Method, a Fourier decomposition of the CCF will also be tried.

Porto

Porto is implementing a **Gaussian process regressive network (GPRN)** framework. The RVs are fitted along with any other time series that can give you information on activity with exactly the same time stamps as the RV information (which describes most indicators, but not the photometry as given). The framework is similar to that of a neural network. Nodes can be any form of GP kernel. These nodes are connected by weights, that are also GP guided.

The main difference between GPRN and a standard GP is that with a standard GP with say a semi-quasi periodic kernel, the amplitude is constant. The GPRN framework allows the amplitude to vary as it is described by a GP itself. The amplitudes are therefore non-stationary, which works better for analyzing non-stationary signals such as activity.

There are no activation functions in this framework despite being like a neural network.

Sidera

Sidera methods focus on exploring the data in the frequency domain. They have constructed metrics of coherence and phase lag as well as de-biased periodograms. Phase wasn't discussed at the meeting but the other two were.

With a classic Lomb-Scargle periodogram, there is no improvement if you have more observations. Additionally, with a small number of data points, there will be a lot of frequency leak, i.e. one frequency will affect other frequencies. This is particularly an issue in the presence of red noise. Their **de-biased periodograms** include a background model that links consecutive observations. The significance at each frequency is decided by an MCMC, constructing a different FAP at each frequency.

Coherence measurements are a mathematical answer to the question "to what extent does an activity indicator predict RV coherence". This coherence metric is given by the Fourier transform of the cross correlation between RV and a given activity indicator squared, which is then normalized by the Fourier transform of the RV and indicator individually squared (see slide 29 for the equation). The indicator varies between 0 and 1, where a value of 1 means the RV signal is entirely due to activity.

The coherence metric revealed that indicators that are sensitive to line shapes better trace the RVs. The Bi-Gaussian indicator did particularly well. Coherence can be used to rank indicators in terms of usefulness as well as reveal which indicators are coherent and therefore present repeated information.

A simple activity correction was done using the indicators identified as most relevant via the coherence metric. Activity was decorrelated from RVs using a linear fit (since the phase metric didn't reveal any phase lag between the RVs and the indicators). Decorrelation was done interactively on the most coherent indicator until there were no more significant coherences, suggesting that most of the variance had been modeled out.

YaleWI

The **Hermite-Gaussian based Radial Velocity (HGRV)** method formulates the estimation of the RV from a stellar spectrum in terms of simple linear regression. It takes account of the (approximate) Gaussian shape of many absorption features and the ability to accurately estimate a star's template spectrum.

This formulation is able to account for the heteroskedastic nature of the noise, does not require interpolation, and easily allows for unbiased estimation of the RV's standard error. It is also shown to have higher RV-estimation accuracy than the EXPRES team's implementation of a traditional cross-correlation function.

The **Stellar Activity F-Statistic for Exoplanet surveys (SAFE)** extends the HGRV model by including higher-order Hermite-Gaussian functions in the linear model. These additional terms are tuned to approximately resemble signals of stellar activity based on SOAP 2.0. The SAFE is the F-statistic calculated for testing the hypothesis that the all the coefficients of these additional terms are simultaneously zero (i.e., that there is no stellar activity signal present). Through simulation, the SAFE is shown to follow its theoretical null distribution when only a Doppler-shift is present. The SAFE is also found to have higher statistical power (i.e., probability of detection) than many classical stellar activity indicators. Currently, the primary purpose of the SAFE is to detect the presence of stellar activity, but may be extended to help correct apparent RV's.

LienhardMortier

The LienhardMortier team is making use of the Haywood+ 2020 result that showed how well stellar RVs matched with the unsigned magnetic flux of the Sun, making it a very promising indicator. Their **Zeeman Splitting Least Squares Deconvolution (ZLSD)** method models the effects of Zeeman splitting in the spectrum, and so can isolate this indicator from the spectral data.

Slide 36 demonstrates how the Zeeman effect is expected to manifest in spectral lines, mainly broadening. One of the parameters also reveals the geometry of the magnetic fields on the surface of the star. Line information (wavelength, depth, and landé factor) is read in from VALD. RVs can be estimated from where there is broadening and from where there isn't, allowing for the separation of where there is expected to be changes due to activity.

Warwick

The Warwick team is working on improving **CCF masks** by using the observations themselves. Slide 39 shows an example of the process using data from CARMENES. Each line is characterized using a fitting function. The fit parameters (FWHM, depth, contrast) are then synthesized into a statement of how good that line is, which is used to select lines.

A mask has been made for the EXPRES data and the code works, but a bit more vetting is needed. The team is happy to share their CCFs with any participants as well as the activity indicators derived from their CCFs.

Austin(-ish)

Zoe De Beurs, Andrew Vanderburg, and Christopher Shallue will be joining as a new team. Zoe, a fourth-year undergraduate student at UT Austin (looking at graduate programs!) will be implementing a **neural network** to detect activity, as described in her recent paper on arxiv.

AmazonML

A team led by Christopher Leet (who used to be an undergraduate at Yale and now is also looking into graduate programs!) is working with a couple of people from Amazon's Auto Machine Learning Project. The focus is on **explainable machine learning** (XML), which provides a human understandable model and at best can be translated into an analytical model. This grants insight into physical models and allows analytic errors to be quantified.

They have only just joined, but intend to try a range of methods.

Initial Results

Initial results are shown on slides 44 and 45.

The first slide shows a grid of plots. Each row represents a different method tried by a team. The first column plots the cleaned (activity-less) RVs submitted for each method (blue). The second column plots the modeled noise/activity RVs (orange). The third column shows the periodogram for both sets of RVs. In all plots, the original EXPRES data is depicted in grayish-black. For the periodograms, the periodogram representing the observation's window function is also plotted in green.

The second slide shows a box plot for each data point composed of the results from all methods (including several variants of the Penn State methods, which therefore may be slightly overweighed). The x-axis is now observation number (and so while corresponds to time, does not have the same scale). The top plot shows box plots for the cleaned RVs; the second plot shows box plots for the modeled noise/activity RVs. These give a sense of how much the methods agree for each data point.

Note: the bottom plot shown during the meeting was wrong, but has been fixed in the uploaded version.

Some Data Comments

We originally planned to release data for HD 217014 (51 Peg), but have now decided instead to release data of HD 26965. There is more data for HD 26965, it is of higher quality (taken longer after commissioning), and is more scientifically interesting. Different publications have characterized signals found in HD 26965 data as being due to either planets or activity. One of the keystone results of this project will be the ability to settle this question using literally a world of experience.

The next three data sets to be released will have a lot more data taken over a spread of more nights. The EXPRES team is currently working on some pipeline changes that are expected to be implemented over the weekend/early next week. Updates will be sent out when we know when the next round of data, extracted with the newest pipeline, will be released.

The CCFs from July 1 for HD 101501 (and a few other observations) do exhibit greater variation than other CCFs, mostly in the baseline of the CCF (i.e. not in the actually main dip of the CCF). We believe this is due to an ADC failure during the night. We acknowledge that the EXPRES CCFs are not optimized, as we have stopped actively developing this code. Several teams, such as Penn State and Warwick, will likely be able to offer better CCFs soon and have kindly volunteered this addition.

There is far less throughput in the blue for EXPRES. Many methods noted better performance after making sharp order cutoffs or keeping to the region informed by the LFC. However, we have seen that some of the densest RV information areas is immediately blue of the LFC region, and we know that there are more lines and less tellurics in the blue. Pick your poison!

The spectral data have NaNs on the edges where there wasn't enough information to get a good extracted value. The telluric model also reports values of -1 where good values could not be derived. Participants should be careful to expect and mask these values.

Standardizing Results

Many different groups are using similar inputs, but there is some tension between teams using their own derived values vs. the provided values, which may give more standardized results. It was proposed that methods should be required to give results using the provided data, for more apples to apples results. Teams would then also be allowed to submit results using their best-looking values.

The submission guidelines will be updated with any requests in this vein.

Standard Training Set

It would be helpful to have a standard set of data with a known truth for everyone to test their methods on. This would ideally have a known planet signal as well as a known activity signal. Such a data set will also help with the reproducibility of the different methods, especially those with many tunable parameters.

The idea of a known truth is, of course, inherently fraught as we lose realism either in the known properties of the star/planets with actual observations, or in the true noise properties with simulated data. We will work on injecting planet signals (see below), but participants are also encouraged to use the HARPS-N stellar data, which has a well understood signal but is taken with a different instrument. EXPRES, led by Joe Llama, does have an operating solar telescope as of August 2020, and we are continuing to build up this data set.

Injected Planet Signal

Similarly, there should be a data set with a known injected planet signal. An essential metric of how well a method for mitigating stellar activity performs is assuring that the method will not also erase planet signals. A method that does so is obviously no good, and so we must be able to test this.

Injecting a planet signal will have to be done carefully. First and foremost, we must only slide the wavelengths of the stellar lines. The telluric lines should remain in the same place. Otherwise, methods that implicitly deal with tellurics may respond differently to these injected signals than methods that do not or simply mask out tellurics.

Optimism

There exist techniques that truly do reduce the scatter in the RVs in a significant way. We should all feel optimistic that these methods and this project is moving towards disentangling stellar signals on the level needed!

We have also produced perhaps the world's first set of ensemble RVs, i.e. RV measurements reached through a community effort. These ensemble results themselves are valuable and can be used as a metric of success. Even just these initial results show a lot of agreement between teams; very encouraging!