Exoplanet detection: some statistical challenges

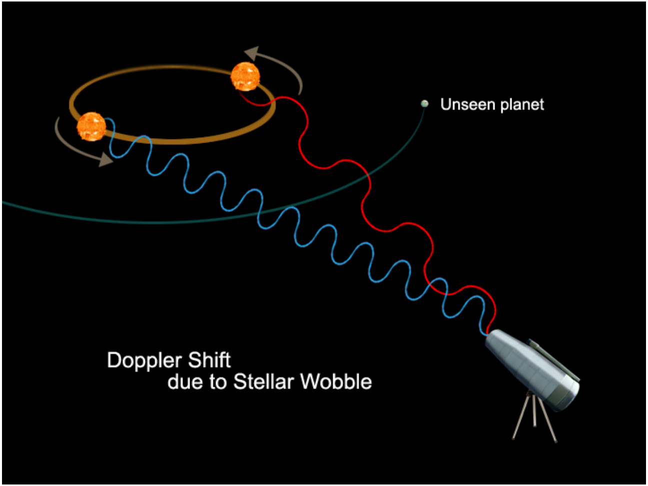
David Jones

Texas A&M University

Based on work with David Stenning, Eric Ford, Robert Wolpert, Thomas Loredo, and Xavier Dumusque

November 13, 2018

Radial velocity (RV) method

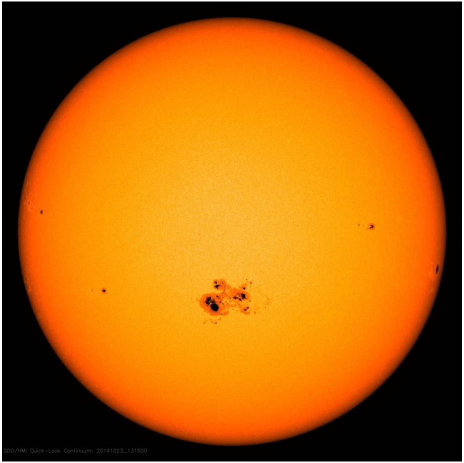


NASA, `https://www.nasa.gov/`

Radial velocity (RV) method

NASA, `https://exoplanets.nasa.gov/interactable/11/`

Stellar activity e.g. spots

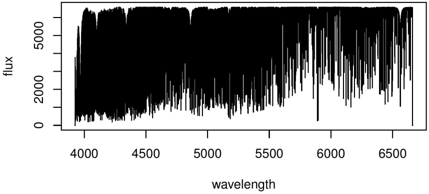


NASA, `https://www.nasa.gov/`

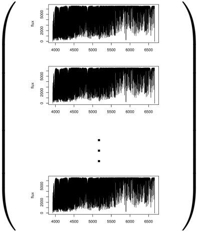
How do we get the RV times series?

- Observation times: $t_1, t_2, \ldots, t_n$

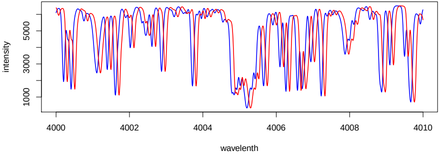
Single observation – a vector of dimension p:

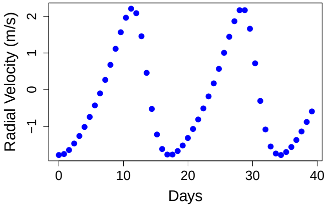


Data matrix $Y_{n \times p} =$



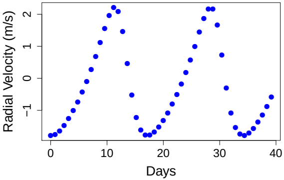
- Astronomers typically reduce the data to RV time series:



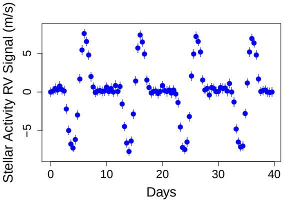


RV corruption 

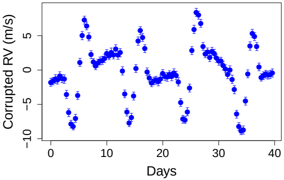
Corrupted RV $=$



$+$



$=$



Keplerian model for RV due to a planet

Keplerian model e.g. Danby (1988)

$$M(t) = \frac{2\pi t}{\tau} + M_0$$
$$E(t) - e\sin E(t) = M(t)$$
$$\tan\frac{\phi(t)}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E(t)}{2}$$

RV due to planet: $v(t) = K(e\cos\omega + \cos(\omega + \phi(t))) + \gamma$

Parameters:

K = velocity semi-amplitude

τ = planet orbital period

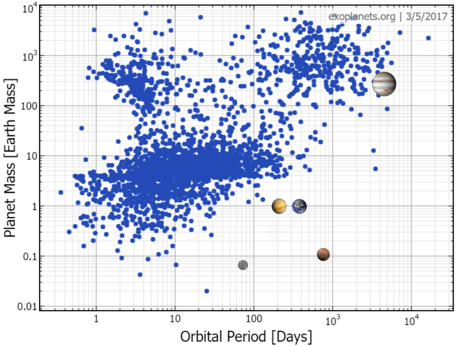
M_0 = mean anomaly at $t = 0$

e = eccentricity

γ = systematic velocity parameter

ω = argument of periapsis

So is it difficult to find a real planet?

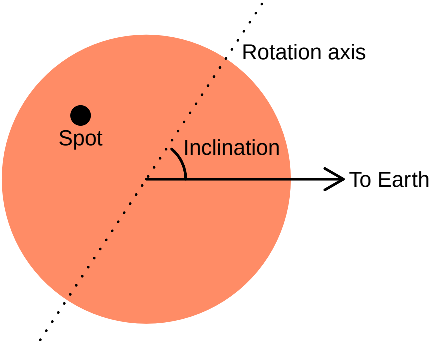


http://exoplanets.org

- There are many planets, and large planets and planets with short orbital periods can be easy to find, but **Earth-like planets** are hard to find
- Some notable detections have turned out to be **false positives**:
  - e.g. *Ghost in the time series: no planet for Alpha Cen B*, by Rajpaul, Aigrain, & Roberts (2015)
- In other cases, the **strength of evidence** for a planet may be (very!) inaccurately quantified – coming next!

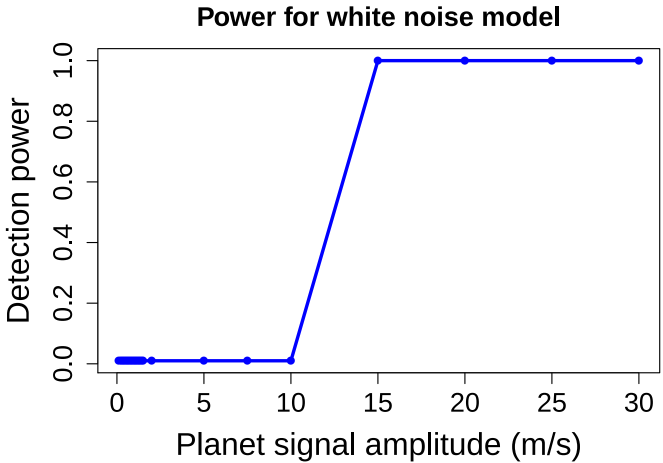
Key tool: SOAP 2.0 simulation software

Dumusque et al 2014: Spot Oscillation And Planet (SOAP) 2.0 radial velocity simulation software.



White noise model

White noise stellar activity model: $v_i = v_{\mathrm{pred}}(t_i|\theta) + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$



Five challenges

1. Assessing evidence / Bayes factor estimation
2. Constructing stellar activity proxies
3. RV and stellar activity proxy modeling
4. Activity model selection / evaluation
5. Analyzing multiple stars jointly

Challenge I: Assessing evidence / Bayes factor estimation

Basic correlated RV noise model

RV observations: $v_i = v_{\mathrm{pred}}(t_i|\theta) + \epsilon_i$

Correlated noise: $\epsilon \sim \mathrm{Normal}(\mathbf{0}, \mathbf{\Sigma})$, where

$$\Sigma_{i,j} = K_{i,j} + \delta_{i,j}\left(\sigma_i^2 + \sigma_J^2\right)$$

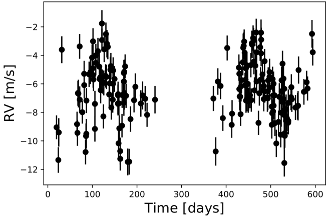
$$K_{i,j} = \alpha^2 \exp\left[-\frac{1}{2}\left\{\frac{\sin^2[\pi(t_i - t_j)/\tau]}{\lambda_p^2} + \frac{(t_i - t_j)^2}{\lambda_e^2}\right\}\right],$$

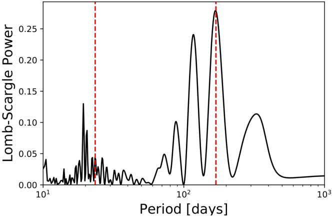
Likelihood:

$$\log \mathcal{L}(\theta) = -\frac{1}{2}(\mathbf{v} - \mathbf{v}_{\mathrm{pred}}(\theta))^T \Sigma^{-1} (\mathbf{v} - \mathbf{v}_{\mathrm{pred}}(\theta)) - \frac{1}{2}\log|\mathrm{det}\mathbf{\Sigma}| - \frac{n_{\mathrm{obs}}}{2}\log(2\pi)$$

Multi-modal posteriors (plus other challenges)

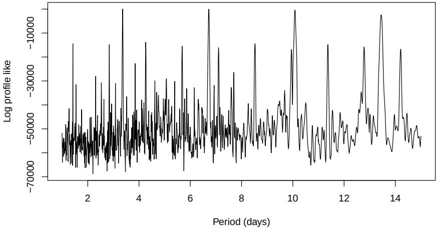
Lomb-Scargle periodogram: essentially looks at the deviance between a sinusodal model and a constant model, e.g., see VanderPlas (2018)



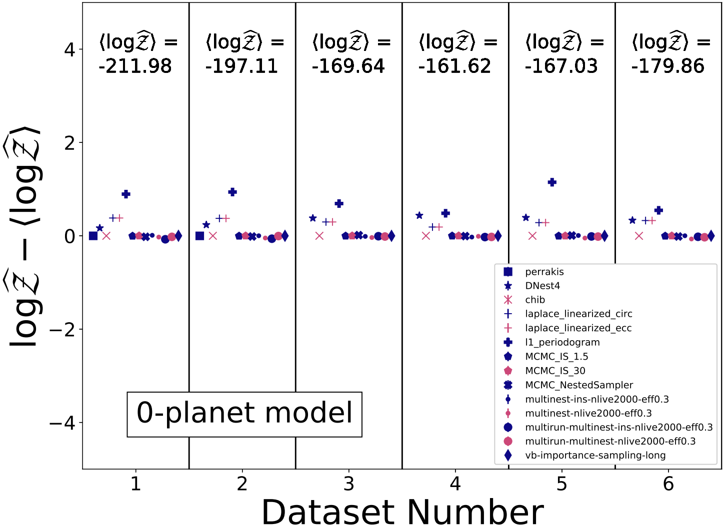


Nelson et al. (2018)

`https://arxiv.org/abs/1806.04683`



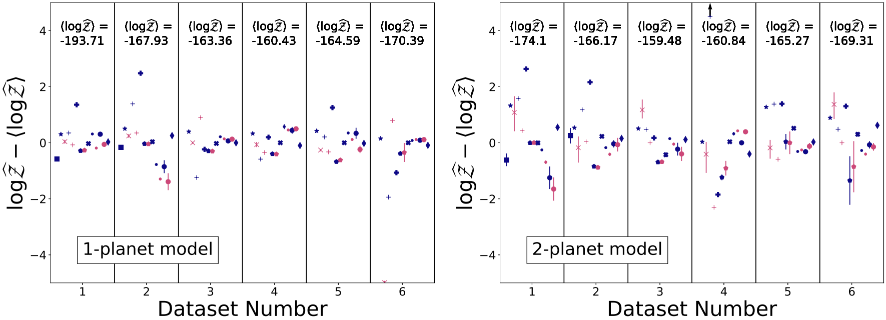
Estimated Bayes factors: EPRV III data challenge



Nelson et al. (2018)

`https://arxiv.org/abs/1806.04683`

Estimated Bayes factors: EPRV III data challenge



Nelson et al. (2018)

`https://arxiv.org/abs/1806.04683`

Equi-energy samplers:

- Equi-energy sampler: Kou, Zhou, & Wong (2006)
- Generalized Wang-Landau algorithm: Liang (2005), Liang, Liu, & Carroll (2012), Bornn et al. (2013)
- Additional bridge sampling step: Wang, Jones, & Meng (2018+)

Period finding:

- Lomb-Scargle periodogram, Lomb (1976), Scargle (1982)
- Supersmoother, Friedman (1984)
- Conditional entropy, Graham et al. (2013)
- Multi-band case e.g. VanderPlas & Ivezic (2015)

Yang Chen & David Jones have done some preliminary investigations in search of an approach that does not involve an exhaustive search

Challenge II: constructing stellar activity proxies

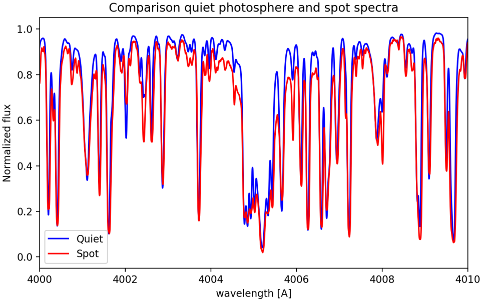
Physically motivated proxies

Motivation:

- If we can determine the level of activity, maybe we can work out if the RV signal is due to a planet or not

Examples:

- Normalized flux
- BIS
- $\log R'_{HK}$



Physically motivated proxies

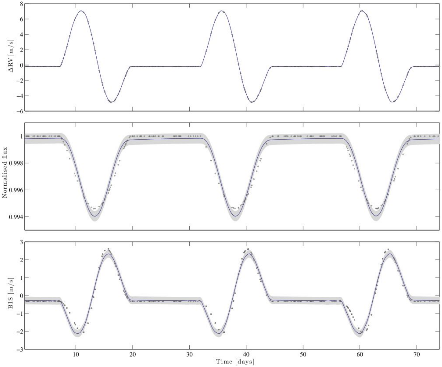


Figure credit: Rajpaul et al. 2015

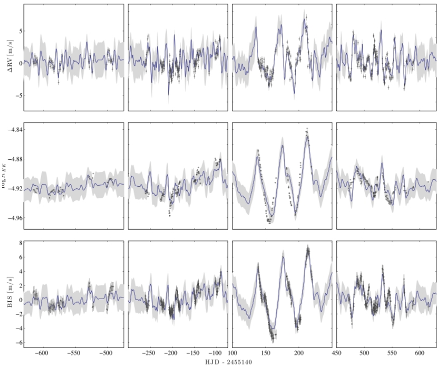


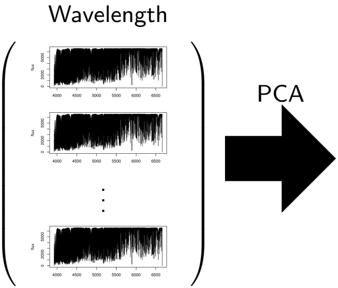
Figure credit: Rajpaul et al. 2015

Motivation for an automatic approach:

- Not clear that two or three proxies is enough
- For different stars / types of stars it may be best to use different proxies

Davis et al. (2017) investigate the use of PCA coefficients as activity proxies

$Y_{n\times p} =$ Time



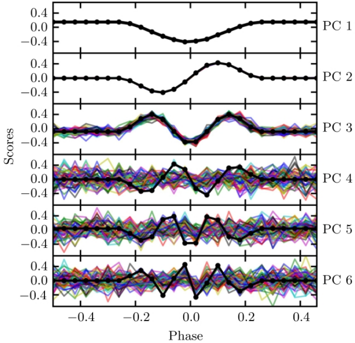


Figure credit: Davis et al. (2017)

Simple insight: we cannot get a pure planet RV signal, but we can get pure stellar activity . . . which can potentially help us find a planet in the corrupted RV signal

Automated Discovery of Activity Proxies

Simple insight: we cannot get a pure planet RV signal, but we can get pure stellar activity . . . which can potentially help us find a planet in the corrupted RV signal

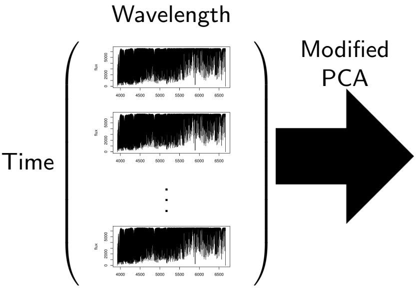
Our modified PCA:

1. Extract RV: compute the apparent RV component, w, and remove it from Y

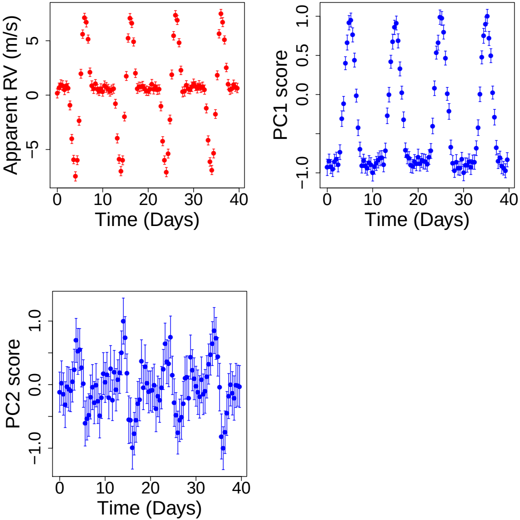
$$\tilde{Y} = Y - \frac{Yww^T}{\sum_i |w_i|^2}$$

2. Find remaining structure: apply a dimension reduction technique (e.g. PCA) and use the new coordinates as proxies

Automated Discovery of Activity Proxies

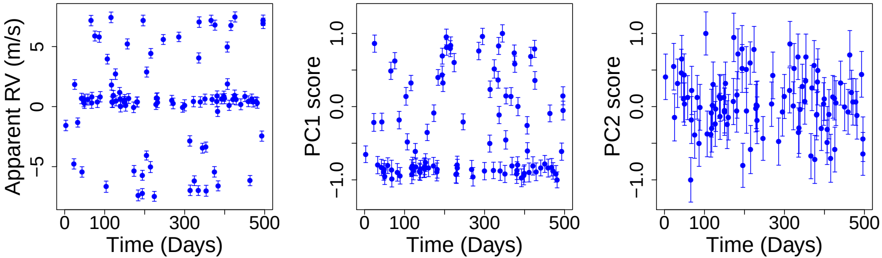


RV corruption and 2 PCA scores:

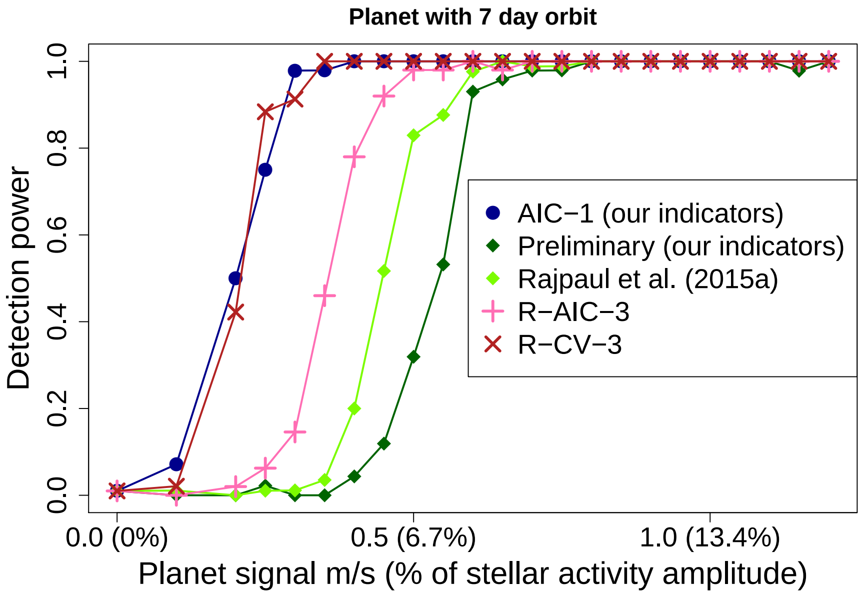


- **Key:** a planet will have no effect on the stellar activity proxies (blue signals)

The data we use looks more like this



Comparison to Rajpaul et al. (2015)



Further dimension reduction approaches

For more complex forms of stellar activity, other techniques may extract more of the relevant information:

- Independence component analysis (ICA)
- Diffusion maps

Challenge III: RV and stellar activity proxy modeling (in the case of a single spot)

- **Def:** a Gaussian process is a stochastic process $X(t)$, $t \in T$ s.t. for any $t_1, \ldots, t_m \in T$, the vector $(X(t_1), \ldots, X(t_m))$ has a multivariate Normal distribution.
- e.g. apparent RV time series $\sim N(0, \Sigma)$
- Quasi-periodic covariance function

$$\mathrm{Cov}(X(t), X(s)) = \exp\left(-\underbrace{\frac{\sin^2(\pi(t-s)/\tau)}{2\lambda_p^2}}_{\text{periodic}} - \underbrace{\frac{(t-s)^2}{2\lambda_e^2}}_{\text{local}} \right)$$

Model from Rajpaul et al. 2015

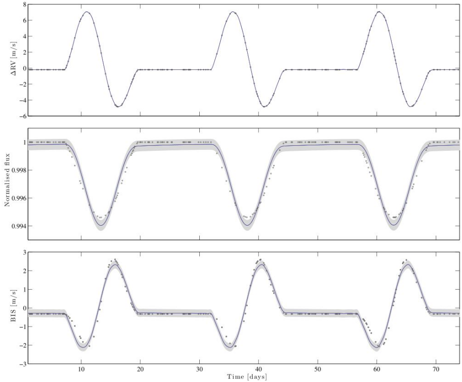


Figure credit: Rajpaul et al. 2015

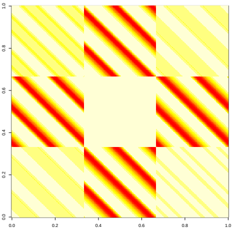
Dependent Gaussian processes:

Stellar activity proxies

$$\begin{aligned}
\Delta\mathrm{RV}(t) &= a_{11}X(t) + a_{12}\dot{X}(t) + \sigma_1\epsilon_1(t) \\
\log R'_{HK}(t) &= a_{21}X(t) + \sigma_2\epsilon_2(t) \\
\mathrm{BIS}(t) &= a_{31}X(t) + a_{32}\dot{X}(t) + \sigma_3\epsilon_3(t)
\end{aligned}$$

Constructing the covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma^{(1,2)} & \Sigma^{(1,2)} & \Sigma^{(1,3)} \\ \Sigma^{(2,1)} & \Sigma^{(2,2)} & \Sigma^{(2,3)} \\ \Sigma^{(3,1)} & \Sigma^{(3,2)} & \Sigma^{(3,3)} \end{pmatrix}$$



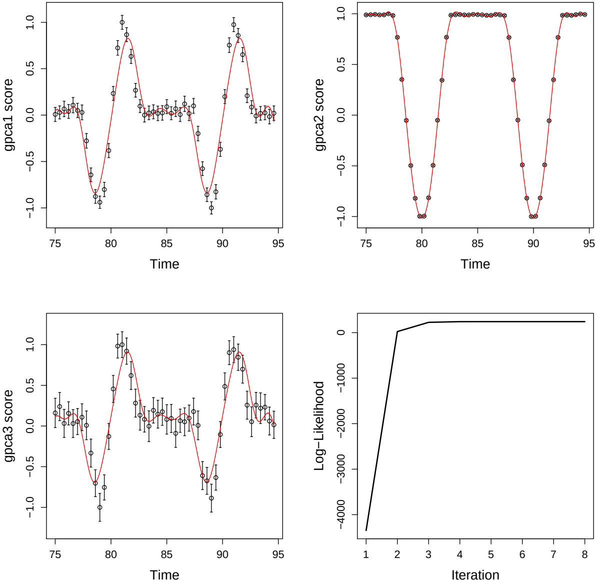
- **Example:** $\Sigma^{(1,2)}$ gives the covariance between observations of $\Delta\mathrm{RV}(t)$ and $\log R'_{HK}(t)$
- **Calculation:** we use the fact that

$$\mathrm{Cov}(X(t), \dot{X}(s)) = \frac{\partial K(t,s)}{\partial s}$$

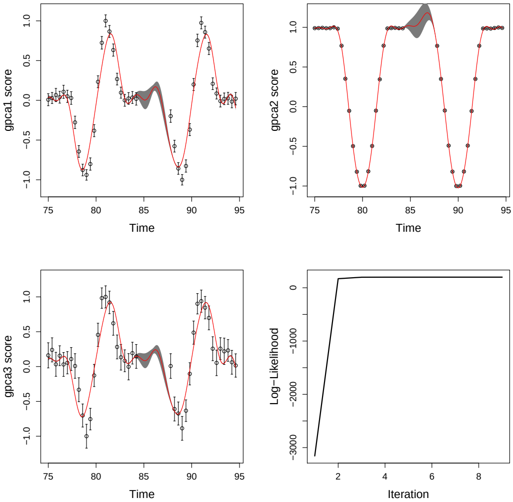
$$\mathrm{Cov}(\dot{X}(t), \dot{X}(s)) = \frac{\partial^2 K(t,s)}{\partial t \partial s}$$

See Theorem 2.2.2 in Adler (2010)

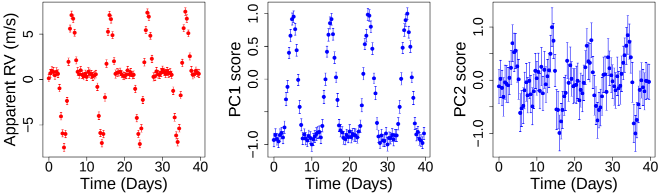
Rajpaul et al. model applied to GPCA scores: MLE fit



Overly constrained, causing strange behaviour



General class of GP models we consider



$$\text{apparent.RV}(t_i) = a_{11}X(t_i) + a_{12}\dot{X}(t_i) + a_{13}\ddot{X}(t_i) + a_{14}Y_1(t_i) + \sigma_{i1}\epsilon_1(t_i)$$
$$\text{Proxy1}(t_i) = a_{21}X(t_i) + a_{22}\dot{X}(t_i) + a_{23}\ddot{X}(t_i) + a_{24}Y_2(t_i) + \sigma_{i2}\epsilon_2(t_i)$$
$$\text{Proxy2}(t_i) = a_{31}X(t_i) + a_{32}\dot{X}(t_i) + a_{33}\ddot{X}(t_i) + a_{34}Y_3(t_i) + \sigma_{i3}\epsilon_3(t_i)$$
$$\dots$$

- Green shows model proposed by Rajpaul et al. (2015)
- In our approach some of the a_{ij}'s are set to zero

Note: adaptation of *Linear Model of Co-regionalization* (LMC) e.g. see Journel and Huijbregts (1978), Osborne et al. (2008), and Alvarez and Lawrence (2011)

Better modeling approaches?

Thoughts / comments:

- **Taylor:** indefinitely extending the Taylor series approach doesn't seem like a good idea
- **Quasi-periodic:** in practice, spots will change at least every couple of stellar rotations, so periodic behaviour will constantly be changing
- **Mean function:** if the mean function is very structured then it may be best to model this more explicitly, rather than using a zero mean GP
- **Kernel learning:** e.g. spectral density modeled by Gaussian mixture (Wilson & Adams, 2013), a Bayesian version (Olivia et al. 2016), transform input (time) before applying standard kernel (Wilson et al., 2016)
- **Non-stationarity?** as spots come and go, stationarity may not be a good assumption

Impossible challenge? learn dependence structure between time series, but also allow the dependence to develop over time.

Challenge IV: model selection / evaluation

Stage 1: Preliminary model selection

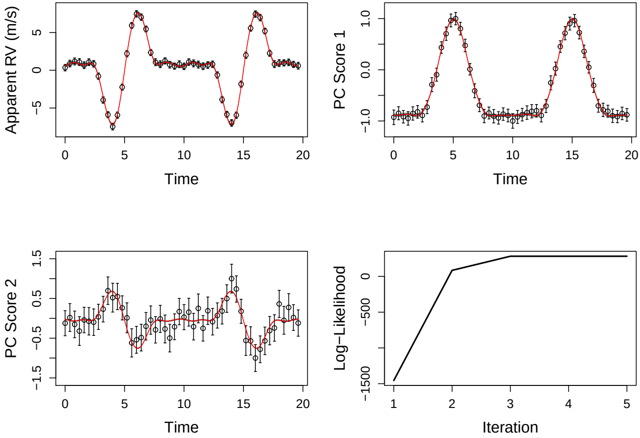
Number of models $= 3375$

Goal: short-list adequate stellar activity models for second stage

Criteria for short-listing models:

1. AIC
2. BIC
3. CV criterion

Typical AIC / BIC 1st ranked model fit



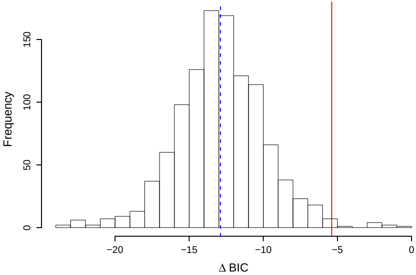
	X coeff	$\dot{X}$ coeff	$\ddot{X}$ coeff	Y coeff
apparent.RV	Y	Y		
PC1	Y		Y	
PC2		Y		

Stage 2: Hypothesis Testing

How much **power** does the LRT have?

- H_0: no planet
- H_A: planet

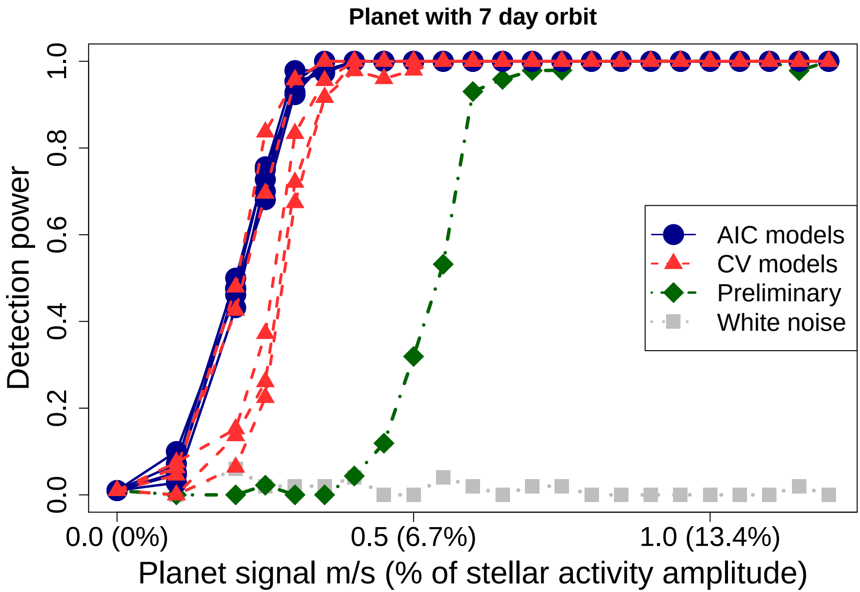
Power computation: null distribution generated via SOAP 2.0 simulations for Sun-like stars with a single spot



Question: How to generate null distribution in general?

- Unknown and time varying activity
- Different types of star

Detection Power: orbital period = 7 days

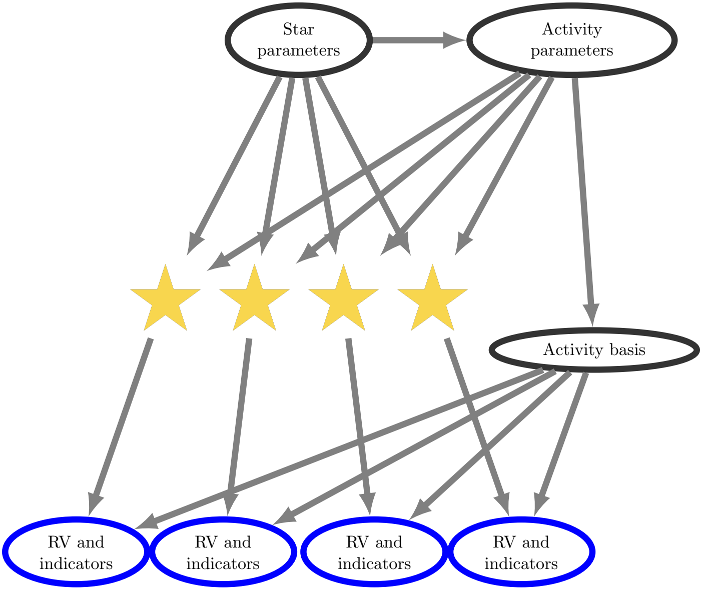


Challenge V: analyzing multiple stars jointly

Questions:

- If we have multiple "similar" stars, all with their own activity, can we gain from pooling information across stars?
- E.g. can we learn basis vectors to capture activity for this type of star
- Since in practice, we won't know the exact form of activity, we want a way to learn likely forms of activity, so we can integrate over these rather than integrating with respect to our prior on the type of activity

Possible hierarchical structure



Five challenges

1. Assessing evidence / Bayes factor estimation
2. Constructing stellar activity proxies
3. RV and stellar activity proxy modeling
4. Activity model selection / evaluation
5. Analyzing multiple stars jointly

Thanks! Questions?