Detecting planets: jointly modeling radial velocity and stellar activity time series

David Jones

SAMSI

Collaborators: David Stenning, Eric Ford, Robert Wolpert, Tom Loredo

March 7, 2017

Detecting planets: jointly modeling radial velocity and stellar activity time series

Or . . . using GPs to find EPs

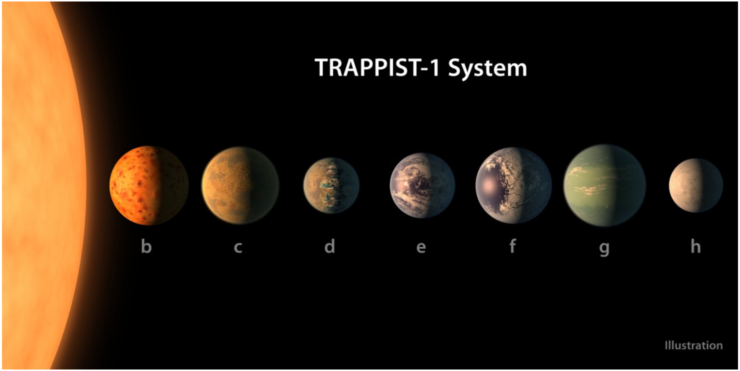
David Jones

SAMSI

Collaborators: David Stenning, Eric Ford, Robert Wolpert, Tom Loredo

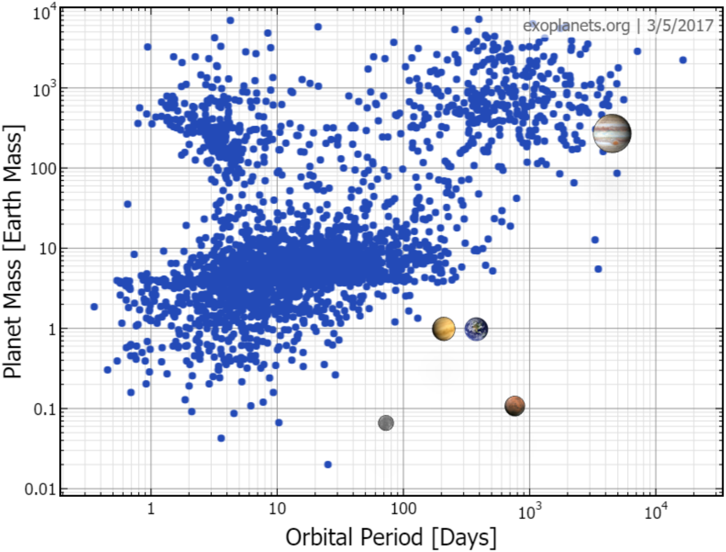
March 7, 2017

Exoplanets in the News: Trappist-1



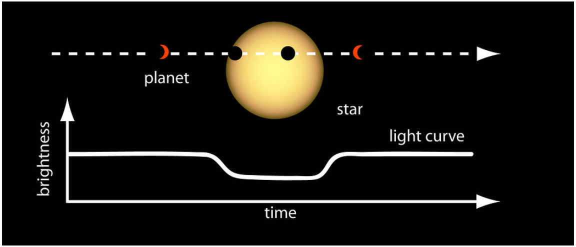
https://www.eso.org

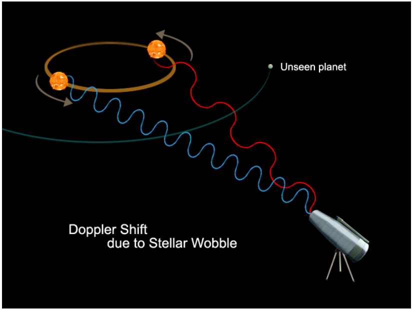
So why keep looking for planets?



http://exoplanets.org

Transit and radial velocity methods





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

Radial velocity method

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

Radial velocity signal

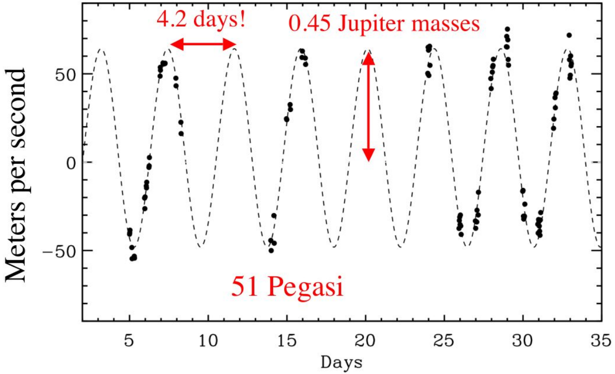


Figure credit: John Asher Johnson, Harvard

- Usually the radial velocity signal is smaller and is corrupted by stellar activity

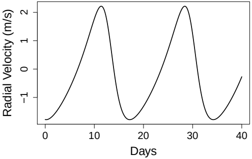
Stellar activity

- Corrupted RV $=$ RV $+$ stellar activity $+$ noise

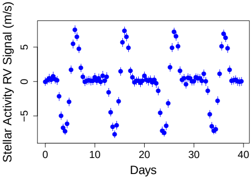
SOHO,
`https://sohowww.nascom.nasa.gov/bestofsoho/Movies/sunspots.html`

RV corruption 

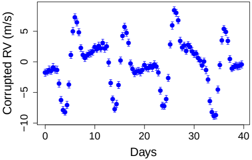
Corrupted RV =



+



=



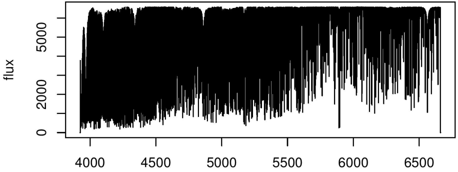
Challenges:

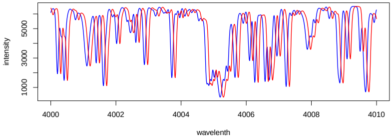
- Earth like planets usually give $< 1ms^{-1}$ signal ...slower than walking speed!
- Multiple and evolving stellar activity phenomena
- Highly irregular observations and lower SNR

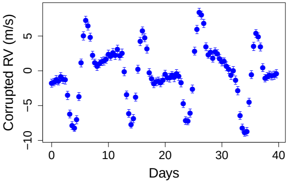
How to stop the corruption!

Statistical opportunity: use information from the spectrum to recover the corruption and subtract it out

- Observation times: $t_1, t_2, \ldots, t_n$
- Raw data is spectrum at each time point e.g.







- Much more information than a single univariate time series is available

Recent approach: Rajpaul et al. 2015

- Rajpaul et al. 2015 jointly model the corrupted RV time series and stellar activity proxies using dependent Gaussian processes
- Spot only (no planet) example from Rajpaul et al. 2015:

RV corruption =

Proxy 1 =

Proxy 2 =

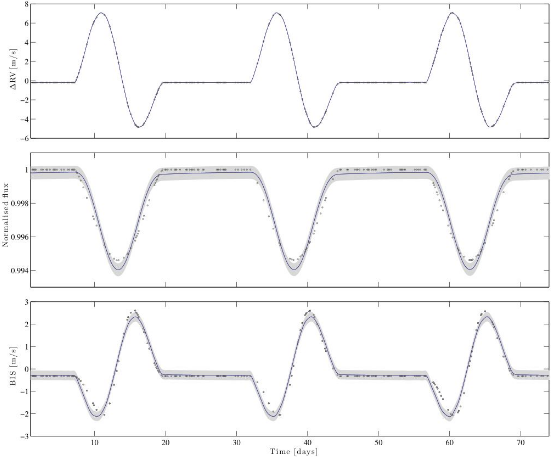


Figure credit: Rajpaul et al. 2015

Real data looks like this . . .

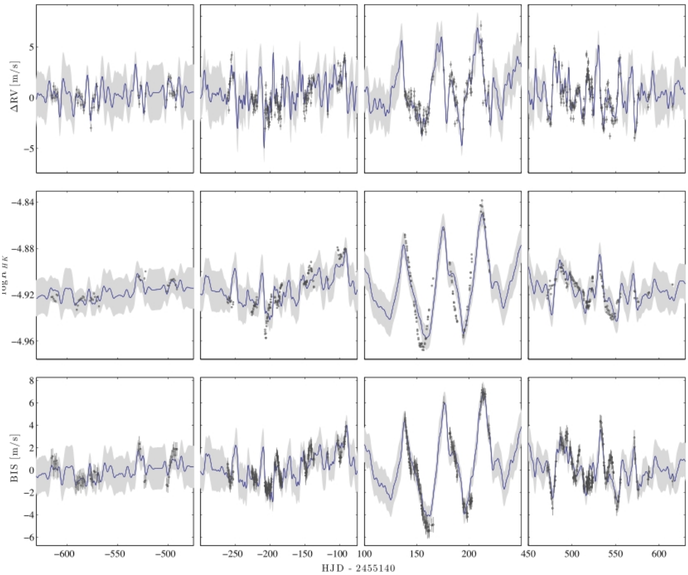


Figure credit: Rajpaul et al. 2015

Our goals

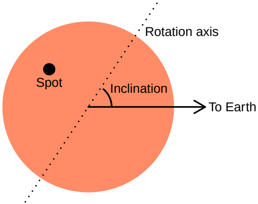
1) More informative proxies - GPCA and diffusion maps (David Stenning)
2) Identify more flexible models to capture new proxies and address existing limitations
3) Model comparison procedure

Goal 1: new stellar activity proxies

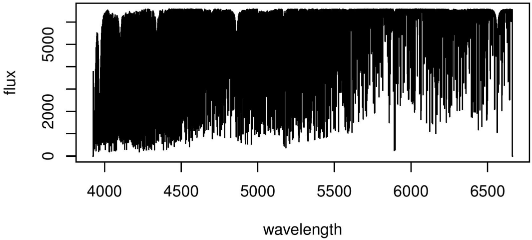
Simulated Stellar Activity Data: NO PLANET YET!

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

- Settings: one spot, stellar inclination 90 degs, spot latitude 40 degs



- Simulated 25 spectra per stellar rotation with 237,944 wavelengths per spectra



Spot Effects

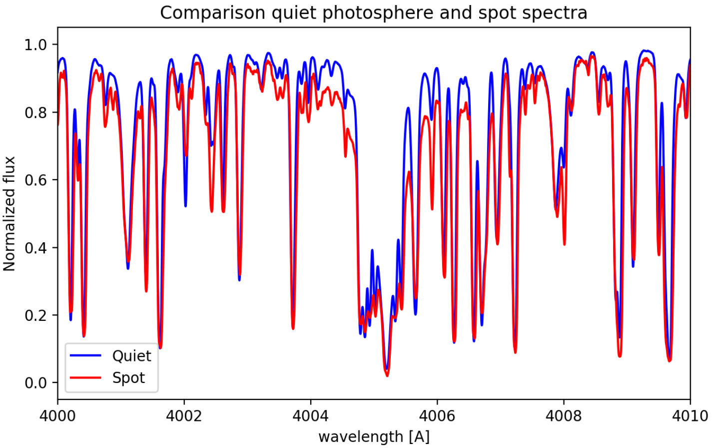


Figure credit: David Stenning

Finding proxies using GPCA: "Generalized" PCA

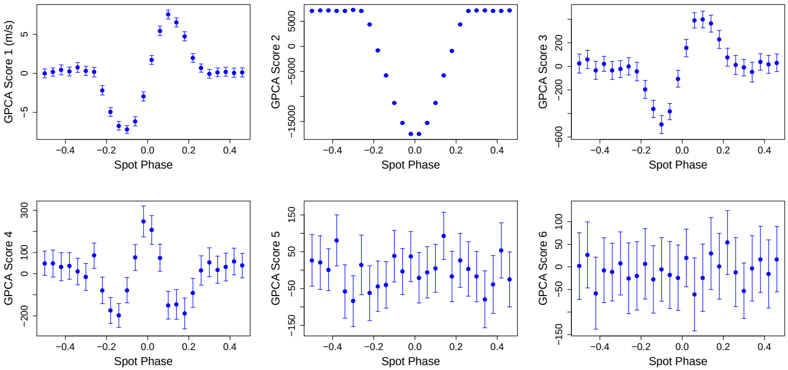
Observation times: $t_1, t_2, \ldots, t_n$

$$\text{Raw data} = \underbrace{\begin{pmatrix} \text{Spectrum at } t_1 \\ \text{Spectrum at } t_2 \\ \vdots \\ \text{Spectrum at } t_{25} \end{pmatrix}}_{\text{237,944 wavelengths}}$$

- Davis et al. (2017) investigate the use of PCA coefficients as activity proxies
- We use the following **GPCA:**
  1. First basis vector is chosen to correspond to the radial velocity
  2. Subsequent orthogonal vectors are chosen to maximize the variation explained as in PCA

RV corruption and GPCA proxies: SOAP data

RV corruption and 5 PCA scores for SOAP 2.0 simulated data:



Diffusion maps

- David Stenning's focus
- Removes linear subspace restriction
- Illustration example:

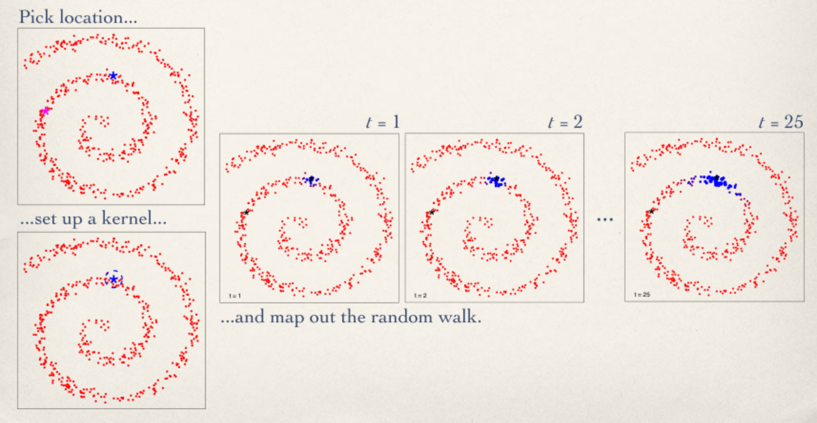


Figure credit: Peter Freeman, CMU, `https://hea-www.harvard.edu/astrostat/CAS2010/pfreeman_CAS2010aug24.pdf`

Diffusion maps

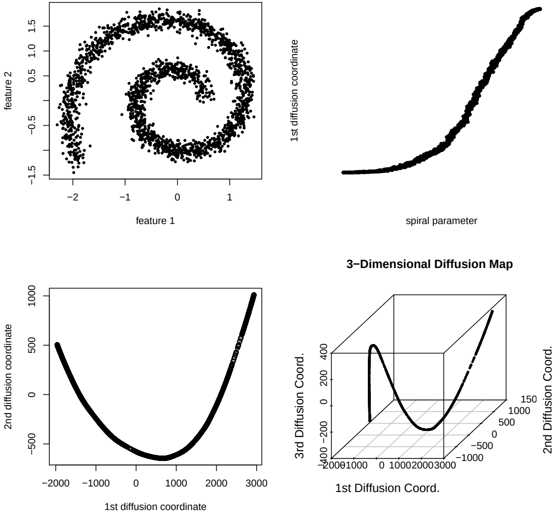
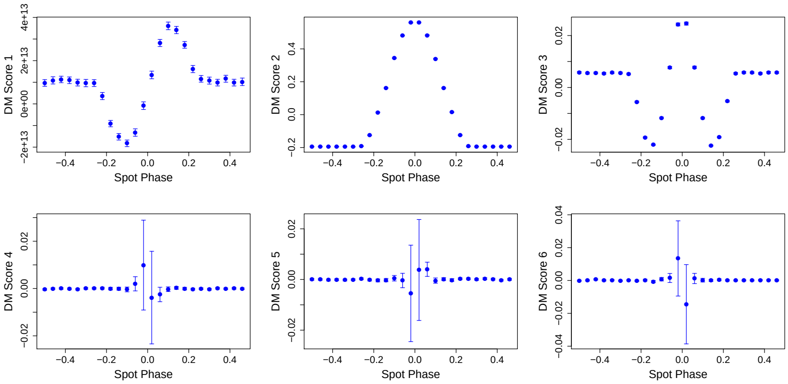


Figure credit: David Stenning

RV corruption and DM proxies: SOAP data

RV corruption and 5 DM scores for SOAP 2.0 simulated data:



Goal 2: identify more flexible models

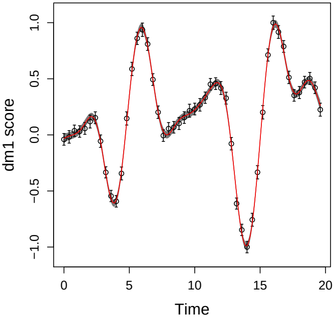
Model rules

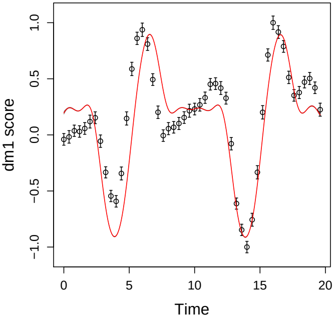
Model rules

- **Be sufficiently flexible:** stellar activity proxies must be well jointly modeled so that the component corrupting the RV signal can be efficiently removed

Model rules

- **Be sufficiently flexible:** stellar activity proxies must be well jointly modeled so that the component corrupting the RV signal can be efficiently removed
- **Don't eat the planet**





Gaussian processes

- **Def:** a Gaussian process is a stochastic process $X(t)$, $t \in T$ s.t. for any $t_1, \dots, t_m \in T$, the vector $(X(t_1), \dots, X(t_m))$ has a multivariate Normal distribution.
- e.g. centred radial velocity time series $\sim N(0, \Sigma)$
- Typically a parametric form is assumed for the covariance matrix Σ e.g.

$$\mathrm{Cov}(X(t), X(s)) = \beta^2 \exp\left(-\frac{(t-s)^2}{\lambda^2}\right)$$

Model from Rajpaul et al. 2015

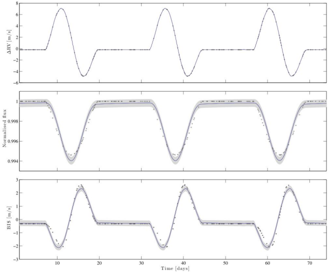


Figure credit: Rajpaul et al. 2015

Dependent Gaussian processes:

Stellar activity proxies

$$\Delta\mathrm{RV}(t) = a_{11}X(t) + a_{12}\dot{X}(t) + \sigma_1\epsilon_1(t)$$

$$\left\{ \begin{aligned} \log R'_{HK}(t) &= a_{21}X(t) + \sigma_2\epsilon_2(t) \\ \mathrm{BIS}(t) &= a_{31}X(t) + a_{32}\dot{X}(t) + \sigma_3\epsilon_3(t) \end{aligned} \right.$$

Covariance function for $X(t)$:

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

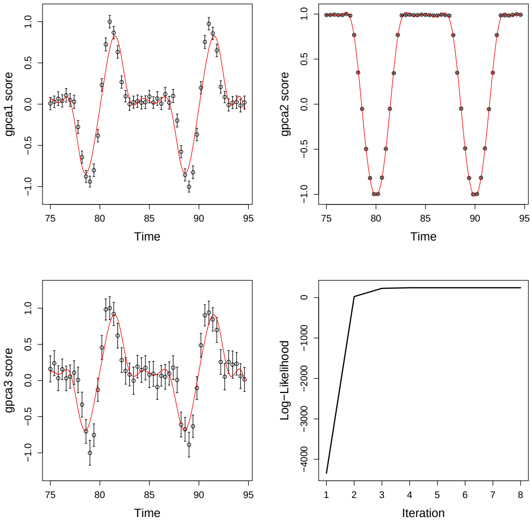
Constructing the covariance matrix

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

- **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

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

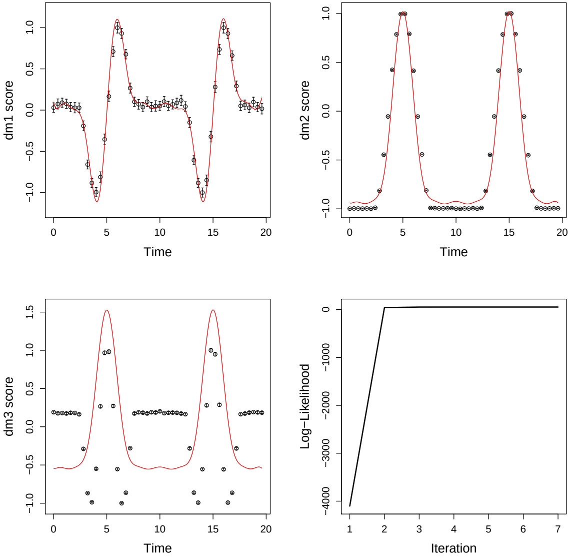
Rajpaul et al. model applied to GPCA scores: MLE fit



- They weight the measurement errors to get a better fit to the first component (RV)

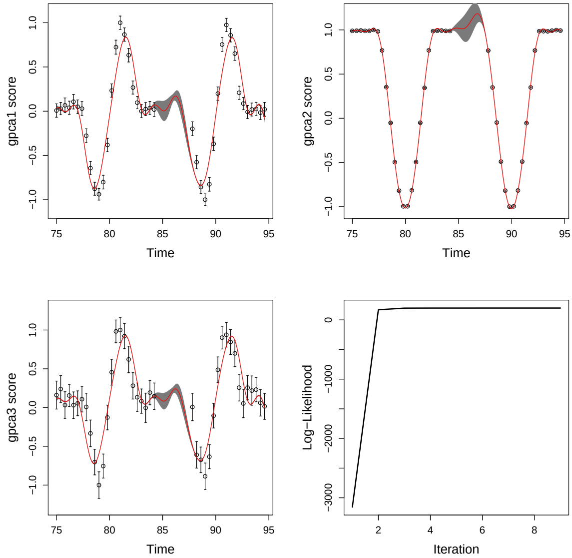
Additional limitations of Rajpaul et al. model

1. Can't capture DM scores with only $X(t)$ and $\dot{X}(t)$



Additional limitations of Rajpaul et al. model

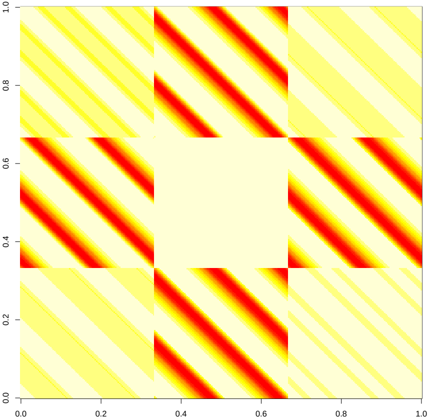
2. Overly constrained, causing strange behaviour



Additional limitations of Rajpaul et al. model

$$
\begin{aligned}
\mathrm{GPCA1}(t_i) &= a_{11}X(t_i) + a_{12}\dot{X}(t_i) + \sigma_{1i}\epsilon_1(t_i) \\
\mathrm{GPCA2}(t_i) &= a_{21}X(t_i) \qquad\qquad\qquad + \sigma_{2i}\epsilon_2(t_i) \\
\mathrm{GPCA3}(t_i) &= a_{31}X(t_i) + a_{32}\dot{X}(t_i) + \sigma_{3i}\epsilon_3(t_i)
\end{aligned}
$$

Negative entries of covariance matrix:



I tried a number of things ...

What worked well:

- Adding in $\ddot{X}(t)$
- Adding an independent GP to GPCA2 / GPCA3

What didn't work well:

- Inflating the measurement errors of GPCA2 (and GPCA3)
- Nugget terms
- Other covariance functions: periodic, sum of two squared exponential kernels, geometric, cosine
- Priors (did help in some cases)
- Allow GPCA2 to use $\dot{X}(t)$

General class of models we consider

$$\text{Output1}(t_i) = a_{11}X(t_i) + a_{12}\dot{X}(t_i) + a_{13}\ddot{X}(t_i) + a_{14}Y_1(t_i) + \sigma_{i1}\epsilon_1(t_i)$$
$$\text{Output2}(t_i) = a_{21}X(t_i) + a_{22}\dot{X}(t_i) + a_{23}\ddot{X}(t_i) + a_{24}Y_2(t_i) + \sigma_{i2}\epsilon_2(t_i)$$
$$\text{Output3}(t_i) = a_{31}X(t_i) + a_{32}\dot{X}(t_i) + a_{33}\ddot{X}(t_i) + a_{34}Y_3(t_i) + \sigma_{i3}\epsilon_3(t_i)$$
$$\ldots$$

- Some of the a_{ij}'s will be set to zero
- $Y_1(t), Y_2(t), Y_3(t), \ldots$ are independent GPs
  BUT: $Y_1(t), Y_2(t), Y_3(t), \ldots$ have the same covariance parameters (different to $X(t)$)

Covariance function:

$$K(t, s) = \exp\left(-\frac{\sin^2(\pi(t-s)/\tau)}{2\lambda_p^2} - \frac{(t-s)^2}{2\lambda_e^2}\right)$$

Goal 3: model selection

Three stages

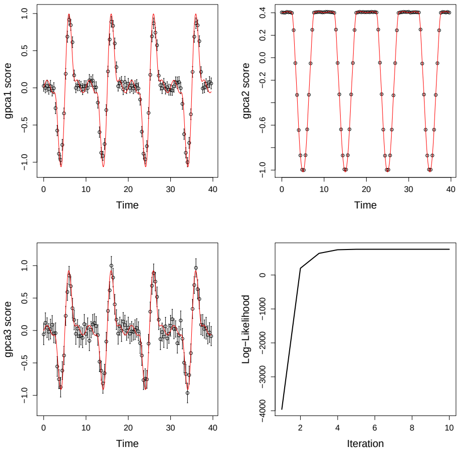
1. **Preliminary stellar activity model search** using AIC, BIC, and cross validation
2. **Simulation study** to assess planet finding power for few top model choices (BIC based)
3. Choose best model and use proper Bayes factor / better approximation to calibrate test and perform search

Preliminary GPCA model selection summary

- BIC: $m \ln n - 2 \ln L(\hat{\theta})$
- CV criterion: $-$log-like for 20% randomly missing data
- Number of models = 3375

Model	AIC.rank	BIC.rank	no.paras	dev	AIC	BIC	CV.rank	CV
Rajpaul	2313	2242	8	133	-573	-558	337	-39
GPCA2+GP	424	372	12	20	-678	-655	2262	18397
min.AIC	1	1	8	10	-695	-680	19	-45
min.BIC	1	1	8	10	-695	-680	19	-45
min.CV	116	47	12	9	-689	-666	1	-46

Typical AIC / BIC optimal model fit



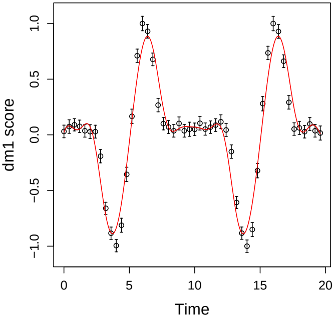
	log.period	log λ_p	log λ_e	X coeff	$\dot{X}$ coeff	$\ddot{X}$ coeff	Y coeff
GPCA1				0.01	0.21		
GPCA2				0.27		0.05	
GPCA3					0.18		
Joint	2.30	-1.08	21.16				

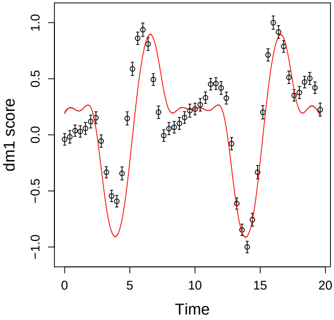
Hypothesis Testing

Question: does the stellar activity model help us find planets?

How much power does the following test have?

- H_0: no planet - stellar activity model is sufficient
- H_A: planet - need additional model for RV signal due to a planet





Adding in a planet: Keplerian model

Taken from Loredo et al. 2012:

$$M(t) = \frac{2\pi t}{\tau} + M_0$$

$$E(t) - e \sin E(t) = M(t)$$

$$\tan \frac{\phi(t)}{2} = \left(\frac{1+e}{1-e} \right) \tan \frac{E(t)}{2}$$

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

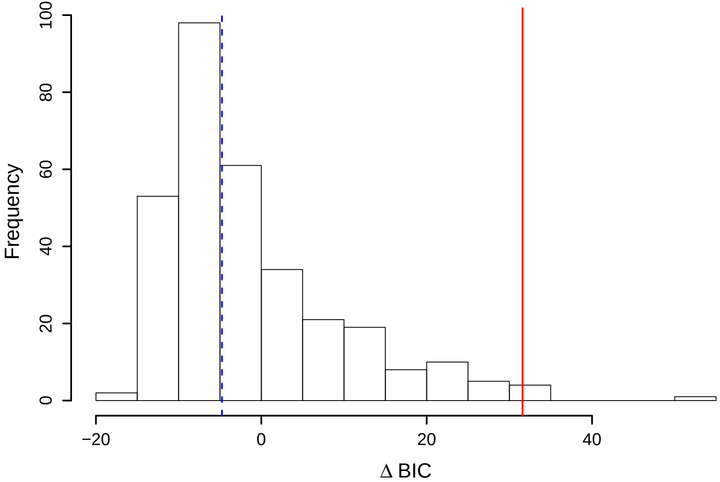
Parameters varied:

K=velocity semi-amplitude (compared with ≈ 7.5 m/s for stellar activity)

τ=planet orbital period (compared with 10 days for stellar period)

Null distribution for AIC / BIC optimal model

- 350 simulated datasets without a planet
- BIC: $m \ln L(\hat{\theta}) - 2 \ln L(\hat{\theta})$
- ΔBIC $=$ null model BIC $-$ null model plus planet model BIC

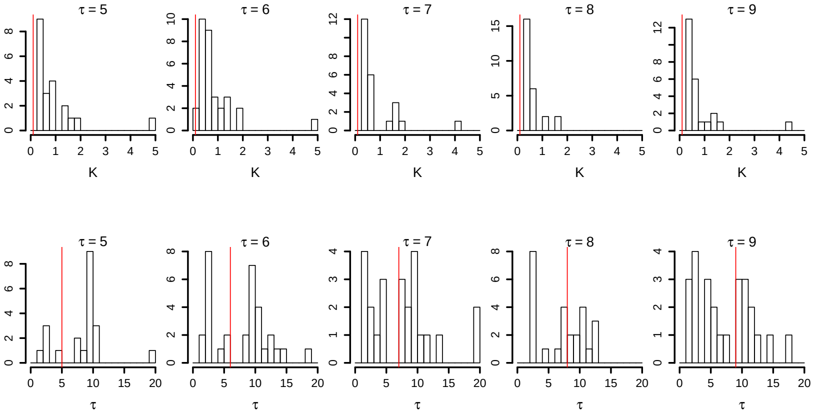


Looking for Planets

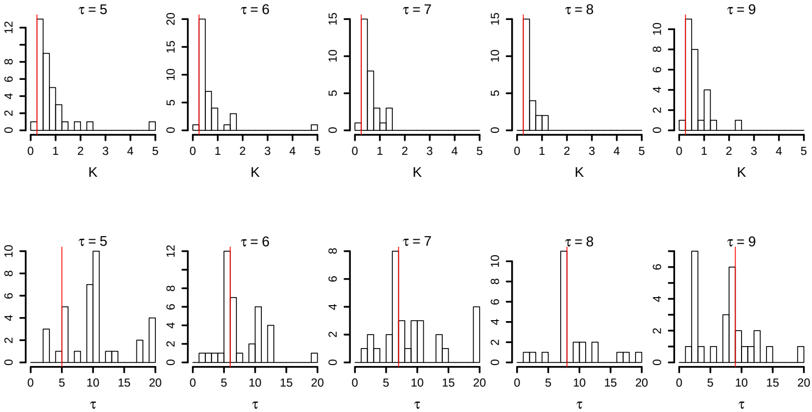
- 50 simulations for each planet setting (not complete)
- Semi-amplitude: $K = 0.1, 0.25, 0.5, 1, 2$ m/s (corresponds to 1.3%, 3.3%, 6.7%, 13.4%, 26.8% of stellar activity amplitude)
- Period: $\tau = 5, 6, \ldots, 9$ (compared with 10 for stellar rotation)

	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	Avg. power
K=0.1 m/s (1.3%)	6.84	1.30	-3.08	3.30	-4.55	0.02
K=0.25 m/s (3.3%)	8.63	12.19	5.21	5.96	3.73	0.12
K=0.5 m/s (6.7%)	44.72	75.08	71.46	63.76	39.99	0.79
K=1 m/s (13.4%)	150.53	267.30	250.70	273.08	153.20	0.96
K=2 m/s (26.8%)	213.79	353.26	396.91	442.55	362.91	1.00

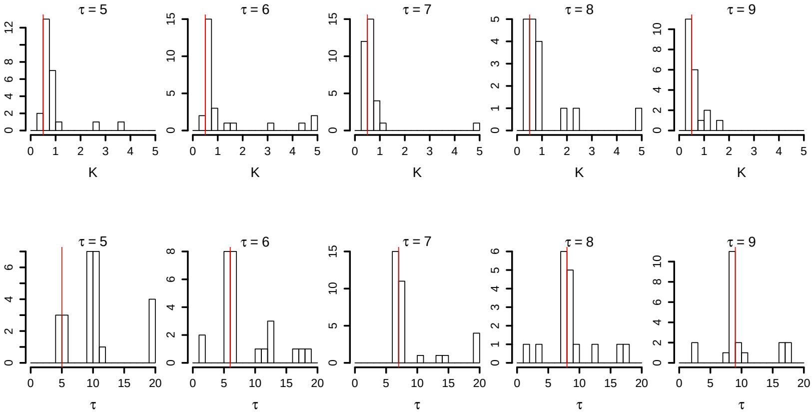
$K = 0.1$m/s (1.3% of SA)



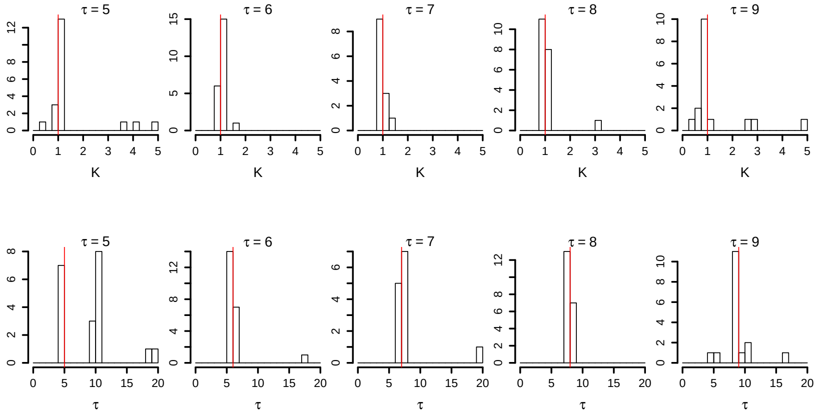
$K = 0.25\text{m/s}$ (3.3% of SA)



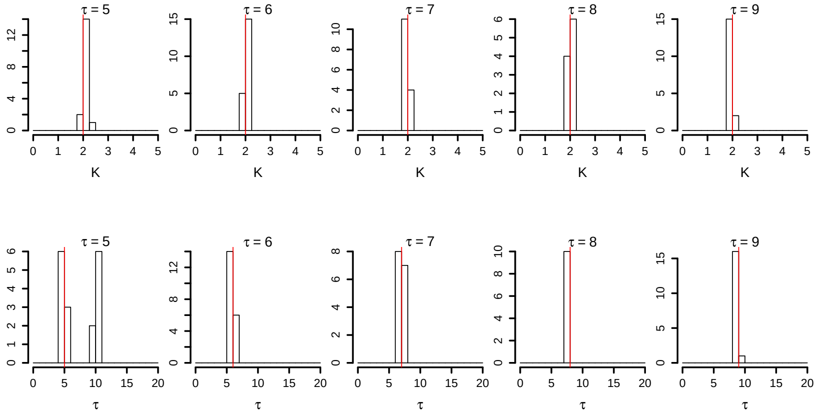
$K = 0.5\text{m/s}$ (6.7% of SA)



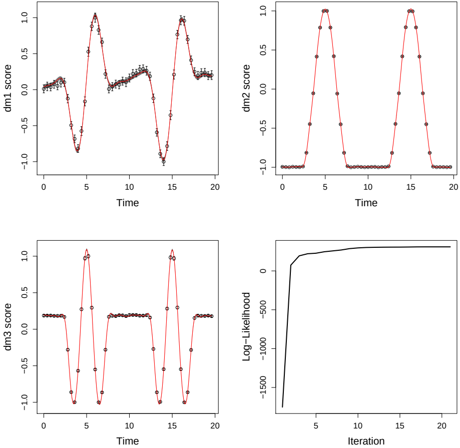
$K = 1\text{m/s}$ (13.4% of SA)



$K = 2\text{m/s}$ (26.8% of SA)

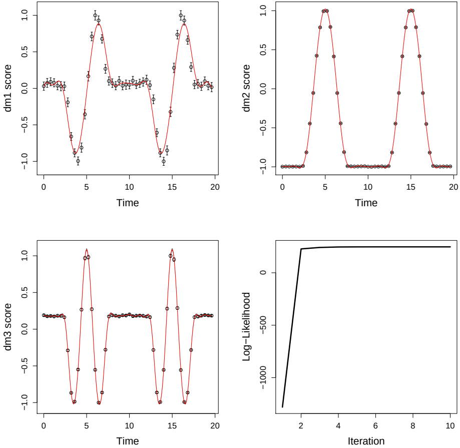


DM BIC-optimal model - eats the planet!



	log.period	log λ_p	log λ_e	X coeff	$\dot{X}$ coeff	$\ddot{X}$ coeff	Y coeff
DM1				0.00	-0.5		
DM2	2.30	-1.40	10.00	0.02		-0.03	0.27
DM3	2.30	-1.40	10.00	-0.09		-0.15	-0.35
Joint	2.50	10.00	0.35				

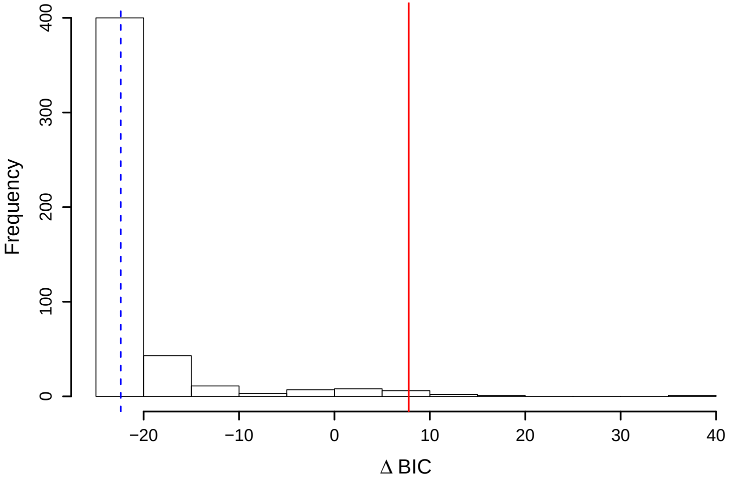
Current best DM model



	log.period	log λ_p	log λ_e	X coeff	$\dot{X}$ coeff	$\ddot{X}$ coeff	Y coeff
DM1				-0.05	-0.58		
DM2				0.77			
DM3	2.30	-0.51	1.23			-0.39	0.34
Joint	2.17	-0.33	1.38				

Null distribution for selected model

- 500 simulated datasets without a planet
- BIC: $m \ln n - 2 \ln L(\hat{\theta})$
- ΔBIC = null model BIC − null model plus planet model BIC



Avg. power results - as of 1pm!

	$\tau = 5$
K=0.1 m/s (1.3%)	0.33
K=0.25 m/s (3.3%)	0.35
K=0.5 m/s (6.7%)	0.82
K=1 m/s (13.4%)	
K=2 m/s (26.8%)	

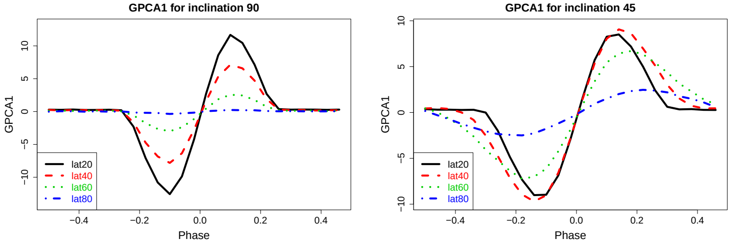
Summary and next steps

Summary:

1) Identify informative stellar activity proxies
2) Propose a flexible class of models
3) Select the optimal model for the purpose of planet detection

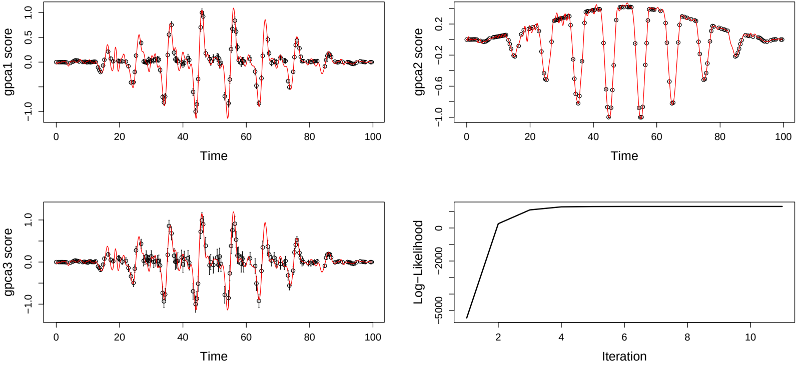
Next steps and future directions:

- Test for a variety of inclinations and spot latitudes



- Test on evolving spots and other stellar activity phenomena
- Real data challenges e.g. finding periods with erratic sampling
- Other proxies
- Scheduling observations

Fit to naively evolving spot data



	log.period	log λ_p	log λ_e	X coeff	$\dot{X}$ coeff	$\ddot{X}$ coeff	Y coeff
GPCA1				0.01	0.15		
GPCA2				0.18		0.04	
GPCA3					0.16		
Joint	2.30	-0.90	3.20				

References

1. Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. (2015). A Gaussian process framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3), 2269-2291.
2. Dumusque, X., Boisse, I., & Santos, N. C. (2014). SOAP 2.0: A tool to estimate the photometric and radial velocity variations induced by stellar sports and plages. *The Astrophysical Journal*, 796(2), 132.
3. Davis, A. B., Cisewski, J., Dumusque, X., Fischer, D., & Ford, E. B. (2017). Insights on the spectral signatures of RV jitter from PCA. In *American Astronomical Society Meeting Abstracts*, 229.
4. Loredo, T. J., Berger, J. O., Chernoff, D. F., Clyde, M. A., & Liu, B. (2012). Bayesian methods for analysis and adaptive scheduling of exoplanet observations. *Statistical Methodology*, 9(1), 101-114.
5. Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning (2006). The MIT Press.