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

David Jones
SAMSI

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

April 12, 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

April 12, 2017
Exoplanets in the News: Trappist-1

[Image of the TRAPPIST-1 System with planets labeled b to h]

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

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 = 

\[
\begin{align*}
\text{Radial Velocity (m/s)} & \quad 0 \quad 10 \quad 20 \quad 30 \quad 40 \\
\text{Days} & \quad -1 \quad 0 \quad 1 \quad 2
\end{align*}
\]

\[
\begin{align*}
\text{Stellar Activity RV Signal (m/s)} & \quad 0 \quad 10 \quad 20 \quad 30 \quad 40 \\
\text{Days} & \quad -5 \quad 0 \quad 5
\end{align*}
\]

\[
\begin{align*}
\text{Corrupted RV (m/s)} & \quad 0 \quad 10 \quad 20 \quad 30 \quad 40 \\
\text{Days} & \quad -10 \quad -5 \quad 0 \quad 5
\end{align*}
\]

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 do we get the (corrupted) RV times series?

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

Single observation:

Data matrix =

- Data reduction to (corrupted) RV time series:
How to stop the corruption!

- **Statistical opportunity:** use other information from the spectrum to recover the corruption and subtract it out.
- There is far more information in the spectrum changes than a single times series:

![Comparison quiet photosphere and spot spectra](image-url)
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 / proxy 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
Finding proxies using GPCA: “Generalized” PCA

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

Raw data =

$$\begin{pmatrix}
\text{Spectrum at } t_1 \\
\text{Spectrum at } t_2 \\
\vdots \\
\text{Spectrum at } t_{25}
\end{pmatrix}$$

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:
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
Rules for stellar activity model
Rules for stellar activity model

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

- **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 \( t_1, \ldots, t_m \in T \), the vector \( (X(t_1), \ldots, X(t_m)) \) 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 \( \Sigma \)
e.g.

\[
\text{Cov}(X(t), X(s)) = \beta^2 \exp \left( -\frac{(t - s)^2}{\lambda^2} \right)
\]
Model from Rajpaul et al. 2015

Figure credit: Rajpaul et al. 2015

Dependent Gaussian processes:

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

\[ \log R'_{HK}(t) = a_{21} X(t) + \sigma_2 \epsilon_2(t) \]

\[ \text{BIS}(t) = a_{31} X(t) + a_{32} \dot{X}(t) + \sigma_3 \epsilon_3(t) \]

Stellar activity proxies \[
\left\{
\begin{align*}
\Delta RV(t) &= a_{11} X(t) + a_{12} \dot{X}(t) + \sigma_1 \epsilon_1(t) \\
\log R'_{HK}(t) &= a_{21} X(t) + \sigma_2 \epsilon_2(t) \\
\text{BIS}(t) &= a_{31} X(t) + a_{32} \dot{X}(t) + \sigma_3 \epsilon_3(t)
\end{align*}
\right.
\]

Covariance function for \(X(t)\):

\[
\text{Cov}(X(t), X(s)) = K(t, s) = \exp \left( -\frac{\sin^2(\pi(t - s)/\tau)}{2\lambda_p^2} - \frac{(t - s)^2}{2\lambda_e^2} \right)
\]
Constructing the covariance matrix

\[
\begin{pmatrix}
\Delta RV(t) \\
\log R'_{HK}(t) \\
BIS(t)
\end{pmatrix} \sim N(0, \Sigma)
\]

\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\
\Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\
\Sigma_{31} & \Sigma_{32} & \Sigma_{33}
\end{pmatrix}
\]

▶ **Example:** \(\Sigma^{(1,2)}\) gives the covariance between observations of \(\Delta RV(t)\) and \(\log R'_{HK}(t)\)

▶ **Calculation:** we use the fact that

\[
\text{Cov}(X(t), \dot{X}(s)) = \frac{\partial K(t, s)}{\partial s}
\]

\[
\text{Cov}(\dot{X}(t), \dot{X}(s)) = \frac{\partial^2 K(t, s)}{\partial t \partial s}
\]
They weight the measurement errors to get a better fit to the first component (RV)
Limitations of Rajpaul et al. approach

1. Ad-hoc inflation of measurement errors to ‘improve RV fit’
2. No reason it should work for out proxies – fails for DM proxies
3. Highly constrained covariance matrix, leading to bizarre predictions
I tried a number of things ...

What worked well:
- Adding in $\dot{X}(t)$
- Adding an independent GP to GPCA1 / GPCA2

What didn’t work well:
- Inflating the measurement errors of GPCA1 (and GPCA2)
- Nugget terms
- Other covariance functions: periodic, sum of two squared exponential kernels, geometric, cosine
- Priors (did help in some cases)
- Allow GPCA1 to use $\ddot{X}(t)$
General class of models we consider

RV.corruption\((t_i) = a_{11}X(t_i) + a_{12}\dot{X}(t_i) + a_{13}\ddot{X}(t_i) + a_{14}Y_1(t_i) + \sigma_{i1}\epsilon_1(t_i)\)

Proxy1\((t_i) = a_{21}X(t_i) + a_{22}\dot{X}(t_i) + a_{23}\ddot{X}(t_i) + a_{24}Y_2(t_i) + \sigma_{i2}\epsilon_2(t_i)\)

Proxy2\((t_i) = a_{31}X(t_i) + a_{32}\dot{X}(t_i) + a_{33}\ddot{X}(t_i) + a_{34}Y_3(t_i) + \sigma_{i3}\epsilon_3(t_i)\)

\[\vdots\]

- 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 Akaike information criterion (AIC), Bayesian Information Criterion (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

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC.rank</th>
<th>BIC.rank</th>
<th>no.paras</th>
<th>dev</th>
<th>AIC</th>
<th>BIC</th>
<th>CV.rank</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rajpaul</td>
<td>2313</td>
<td>2242</td>
<td>8</td>
<td>133</td>
<td>-573</td>
<td>-558</td>
<td>337</td>
<td>-39</td>
</tr>
<tr>
<td>GPCA1+GP</td>
<td>424</td>
<td>372</td>
<td>12</td>
<td>20</td>
<td>-678</td>
<td>-655</td>
<td>2262</td>
<td>18397</td>
</tr>
<tr>
<td>min.AIC</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>-695</td>
<td>-680</td>
<td>19</td>
<td>-45</td>
</tr>
<tr>
<td>min.BIC</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>-695</td>
<td>-680</td>
<td>19</td>
<td>-45</td>
</tr>
<tr>
<td>min.CV</td>
<td>116</td>
<td>47</td>
<td>12</td>
<td>9</td>
<td>-689</td>
<td>-666</td>
<td>1</td>
<td>-46</td>
</tr>
</tbody>
</table>
Typical AIC / BIC optimal model fit

<table>
<thead>
<tr>
<th></th>
<th>log.period</th>
<th>log $\lambda_p$</th>
<th>log $\lambda_e$</th>
<th>$\chi$ coeff</th>
<th>$\dot{\chi}$ coeff</th>
<th>$\ddot{\chi}$ coeff</th>
<th>$Y$ coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV.corruption</td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.21</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>GPCA1</td>
<td></td>
<td></td>
<td></td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPCA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.18</td>
</tr>
<tr>
<td>Joint</td>
<td>2.30</td>
<td>-1.08</td>
<td>21.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 \( \approx 7.5 \) m/s for stellar activity)
- \( \tau \)=planet orbital period (compared with 10 days for stellar period)
Null distribution for AIC / BIC optimal model (GPCA)

- 350 simulated datasets without a planet
- BIC: $m \ln n - 2 \ln L(\hat{\theta})$
- $\Delta \text{BIC} = \text{null model BIC} - \text{null model plus planet model BIC}$
Looking for Planets

- 50 simulations for each planet setting (not complete)
- Semi-amplitude: $K = 0.1, 0.25, 0.5, 1, 2 \text{ m/s}$
  (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)

<table>
<thead>
<tr>
<th>$K$</th>
<th>$\tau = 5$</th>
<th>$\tau = 6$</th>
<th>$\tau = 7$</th>
<th>$\tau = 8$</th>
<th>$\tau = 9$</th>
<th>Avg. power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 m/s (1.3%)</td>
<td>6.84</td>
<td>1.30</td>
<td>-7.55</td>
<td>3.30</td>
<td>-4.55</td>
<td>0.02</td>
</tr>
<tr>
<td>0.25 m/s (3.3%)</td>
<td>8.63</td>
<td>12.19</td>
<td>5.14</td>
<td>5.96</td>
<td>3.73</td>
<td>0.12</td>
</tr>
<tr>
<td>0.5 m/s (6.7%)</td>
<td>44.72</td>
<td>75.08</td>
<td>87.22</td>
<td>63.76</td>
<td>39.99</td>
<td>0.79</td>
</tr>
<tr>
<td>1 m/s (13.4%)</td>
<td>150.53</td>
<td>267.30</td>
<td>286.25</td>
<td>273.08</td>
<td>153.20</td>
<td>0.96</td>
</tr>
<tr>
<td>2 m/s (26.8%)</td>
<td>213.79</td>
<td>353.26</td>
<td>521.04</td>
<td>442.55</td>
<td>362.91</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Closer look for period = 7 days
Power for period = 7 days

Planet with 7 day orbit

Detection power

GPCA proxies
Period estimation for period = 7 days
Amplitude estimation for period = 7 days
DM is running . . .
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:
- DM proxies and other proxies
- Test for a variety of inclinations and spot latitudes
- Test on evolving spots and other stellar activity phenomena
- Real data challenges e.g. calibrating null, finding periods etc.
Fit to naively evolving spot data

<table>
<thead>
<tr>
<th></th>
<th>log.period</th>
<th>log $\lambda_p$</th>
<th>log $\lambda_e$</th>
<th>$X$ coeff</th>
<th>$\dot{X}$ coeff</th>
<th>$\ddot{X}$ coeff</th>
<th>$Y$ coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCA1</td>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPCA2</td>
<td></td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>GPCA3</td>
<td></td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>Joint</td>
<td></td>
<td>2.30</td>
<td>-0.90</td>
<td>3.20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


