Bayesian Inference in exoplanet searches

XXVI Winter School in Astrophysics
Tenerife Canarias Spain

Phil Gregory
Physics and Astronomy
University of British Columbia

Nov. 2014

Bayesian Inference in the Exoplanet Search
Lectures

1) Exoplanet update, Bayesian primer & Periodograms

2) Fusion Markov chain Monte Carlo (FMCMC)
as a multi-planet Kepler periodogram

3) FMCMC analysis of HD 208487 and Gliese 581 RV data
equations and priors

4) Demonstration of FMCMC in Mathematica

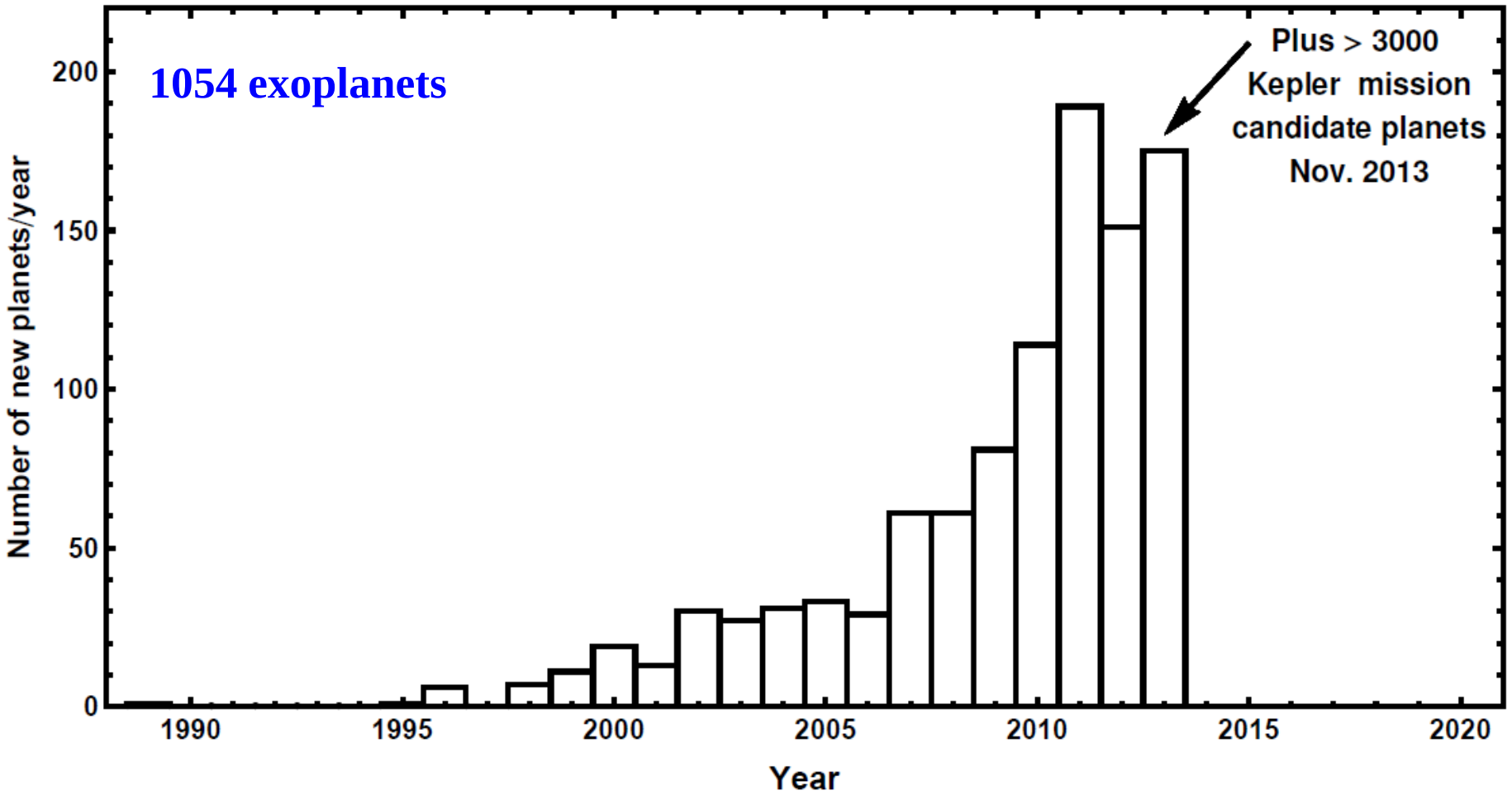
5) Bayesian model selection for exoplanets

6) Distinguishing stellar activity induced RV signals
- more Bayesian periodograms

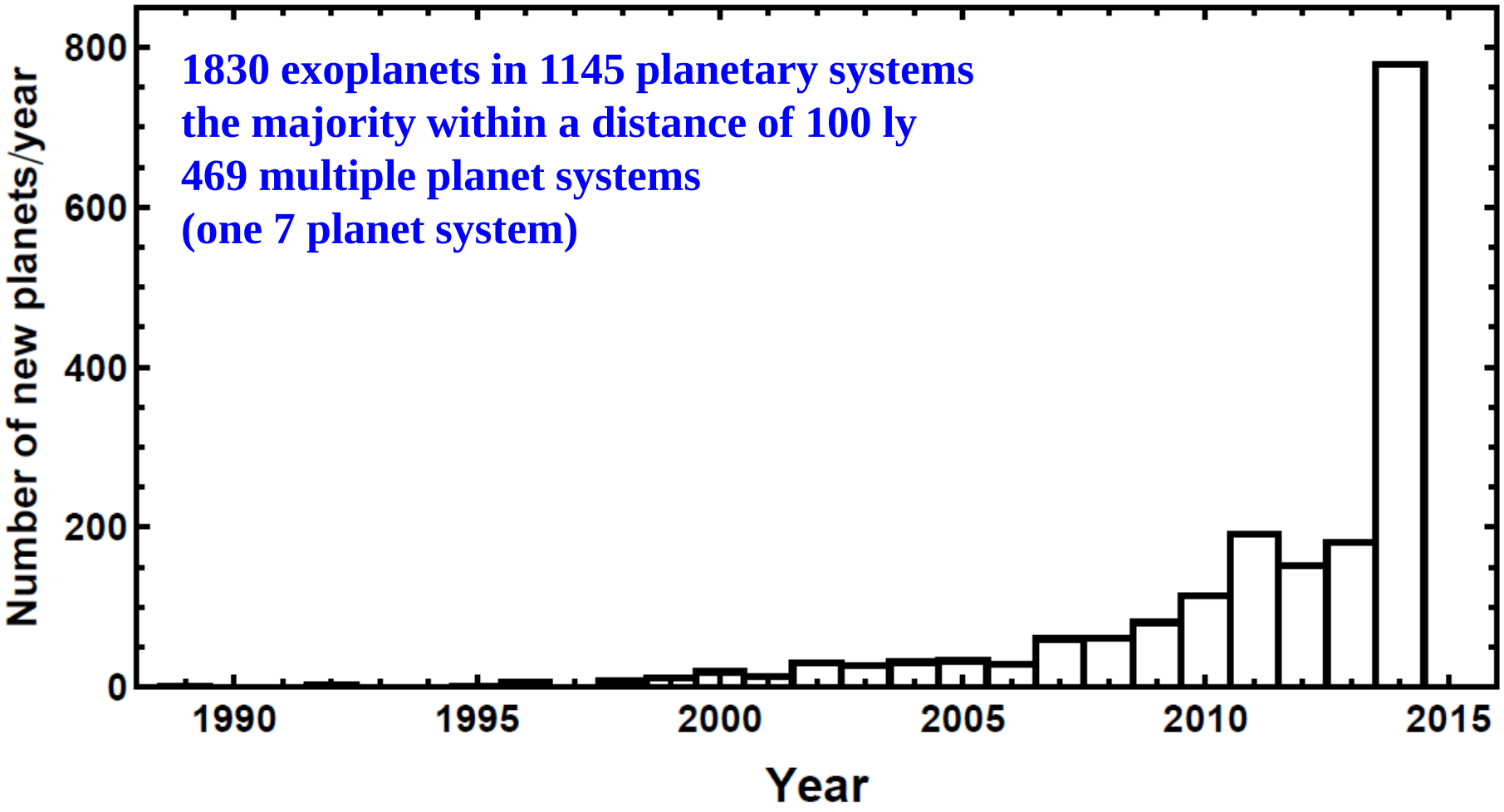
Index of topics

Exoplanet Update

Pace of Extra-solar Planet Discoveries Nov. 2013



Pace of Extra-solar Planet Discoveries Oct. 2014



22± 8 % of G & K dwarfs harbour a planet in HZ with radius = 1 to 2 r_e

(Petigura et al. 2013)

Average number of planets with r < 1.4 r_e in HZ of M dwarfs is = 0.53

(Dressing & Charboneau 2013: Kopparapu 2013; Giados 2013)

Revolutionary ALMA Radio Image Reveals Planetary Genesis

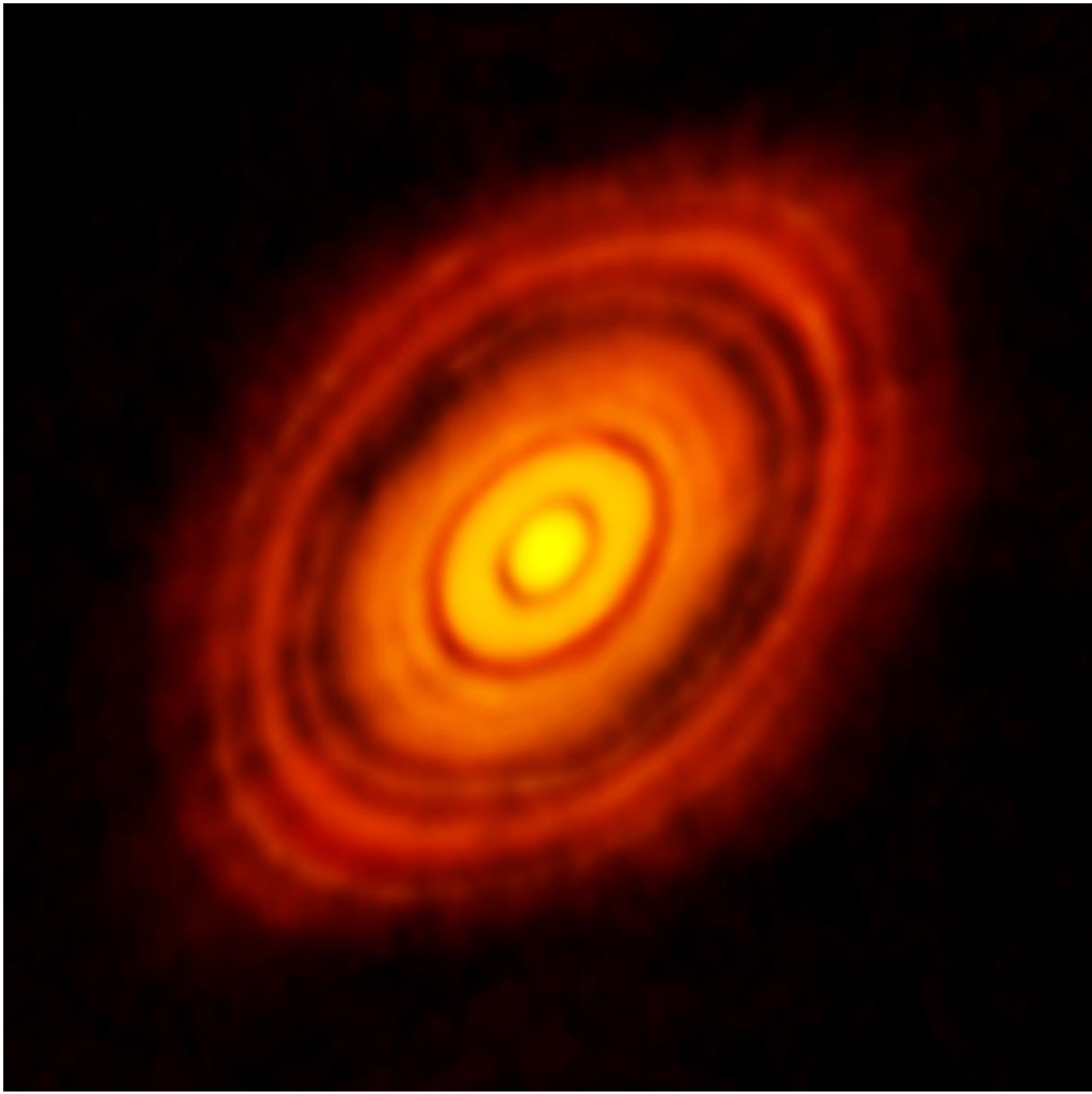
Nov. 5, 2014

HL Tau, no more than a million years old, yet already its disc appears to be full of forming planets.

d ~ 450 light-years

Image resolution 35 milliarcsec

Credit: ALMA (ESO/NAOJ/NRAO)



Breakdown of exoplanets detections by method

1830 exoplanets found in 1145 planetary systems

Detection method	No. of planets
Radial velocity (RV)	577
Transits	1151
Microlensing	32
Imaging	51
Timing (includes 5 pulsar planets)	16
Astrometry	2

The Extrasolar Planets Encyclopaedia
Jean Schneider
(CNRS-LUTH, Paris Observatory)

My own analysis is primarily based on radial velocity data sets.

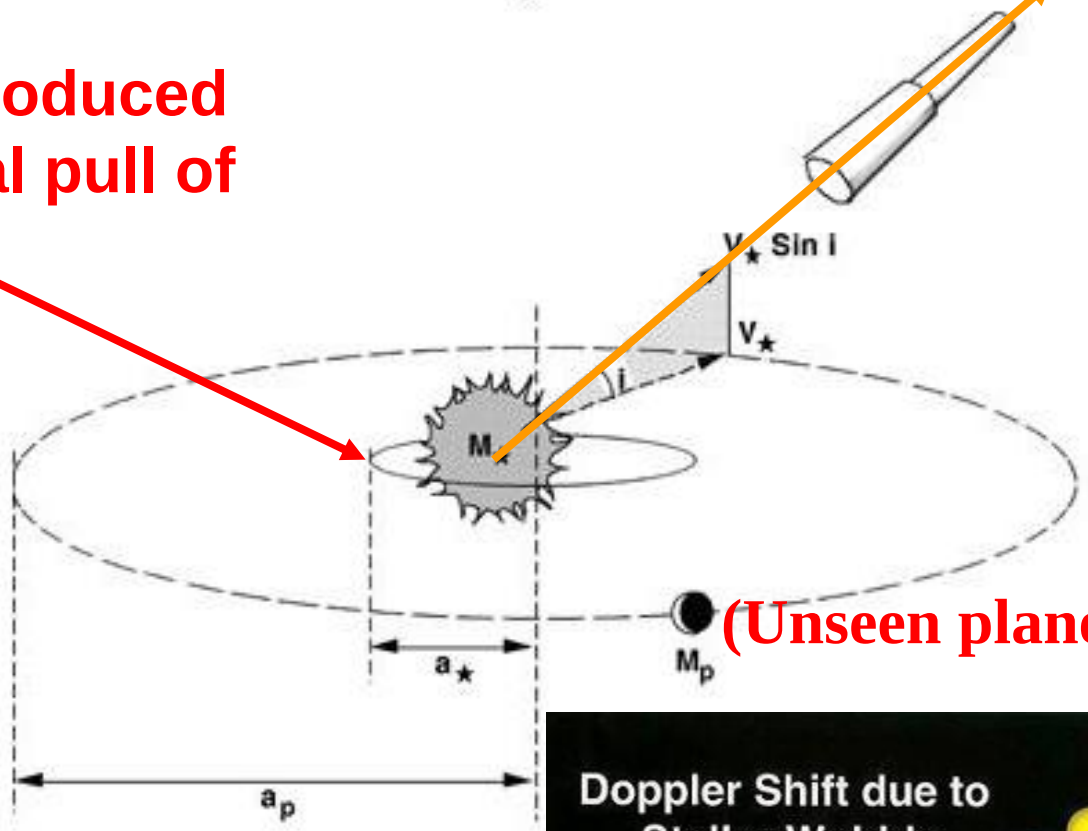
Currently best spectrographs are European HARPS (high accuracy radial velocity planet searcher) and HARPS North capable of long term sub m/s velocity precision. New instruments underway include European EXPRESSO and Yale EXPRESS capable of 10 cm/s precision.

The majority of the planets have been discovered with either the RV or transit method which give rise to time series with embedded periodic signals Which motivates the interest in periodograms.

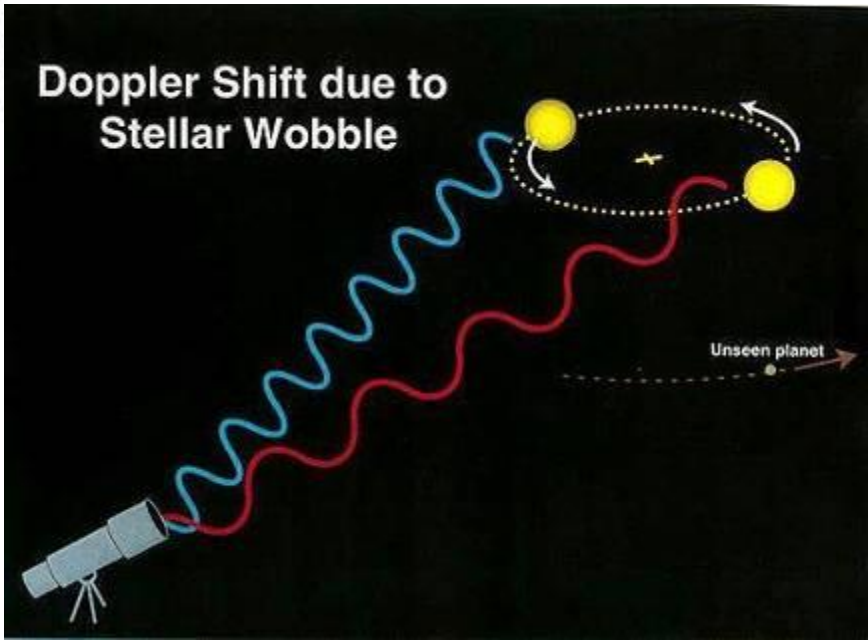
Radial Velocity Method

Earth

Orbit of star produced by gravitational pull of unseen planet.



(Unseen planet)

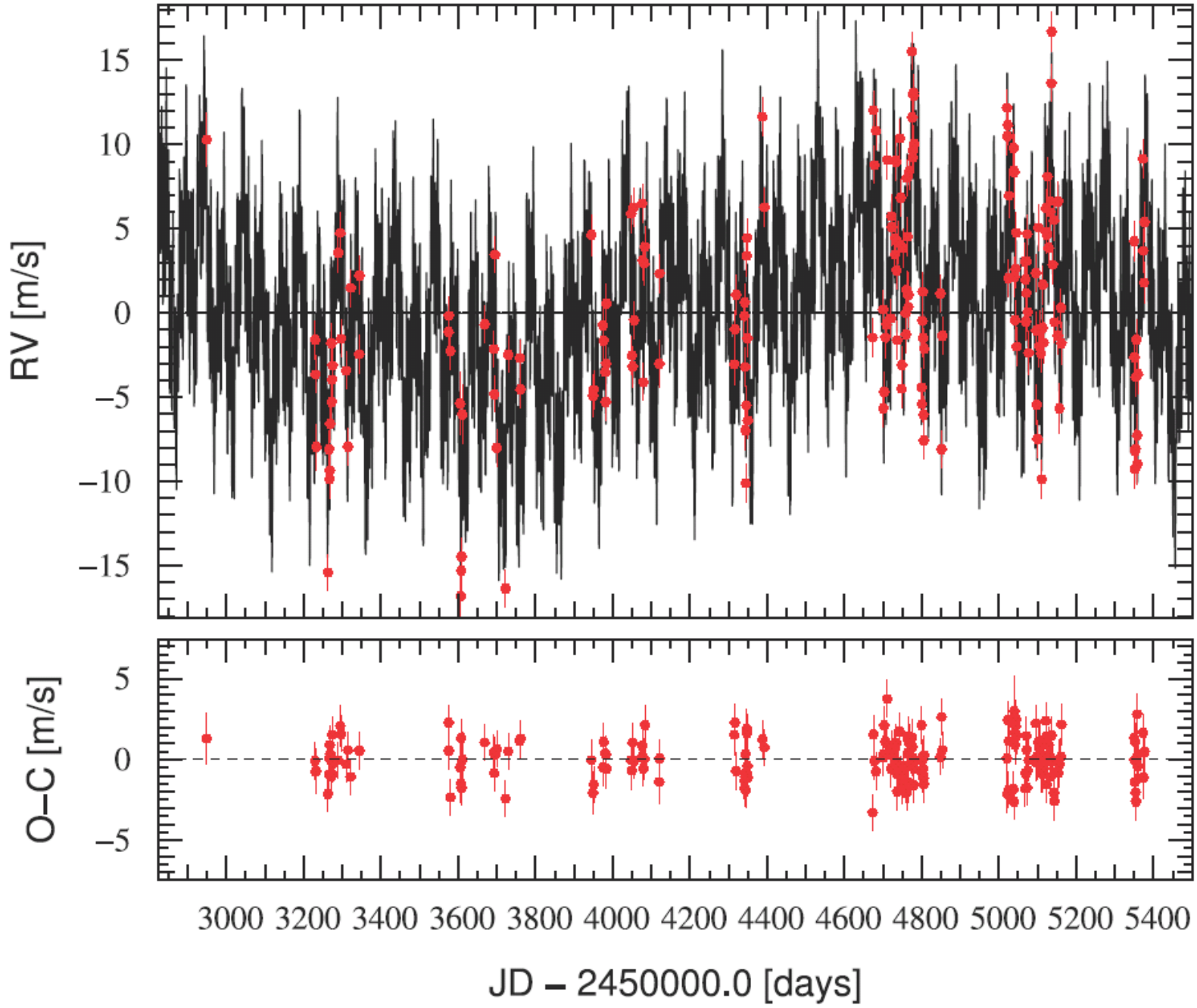


Because the star is a billion times brighter than a planet, only detect a small number of exoplanets directly.

Instead look for the reflex motion of the star due to gravitational tug from the planet.

7 Planet fit to HD10180

C. Lovis, et al., 2011, A&A, 528, 112



Bayesian Primer

What is Bayesian Probability Theory? (BPT)

BPT = a theory of extended logic

Deductive logic is based on Axiomatic knowledge.

In science we never know any theory of nature is true because our reasoning is based on incomplete information.

Our conclusions are at best probabilities.

Any extension of logic to deal with situations of incomplete information (realm of inductive logic) requires a theory of probability.

A new perception of probability has arisen in recognition that the mathematical rules of probability are not merely rules for manipulating random variables.

They are now recognized as valid principles of logic for conducting inference about any hypothesis of interest.

This view of, ``Probability Theory as Logic'', was championed in the late 20th century by E. T. Jaynes.

"Probability Theory: The Logic of Science"
Cambridge University Press 2003

It is also commonly referred to as Bayesian Probability Theory in recognition of the work of the 18th century English clergyman and Mathematician Thomas Bayes.

Logic is concerned with the truth of propositions.

A proposition asserts that something is true.

Examples of propositions:

$A \equiv$ "The newly discovered radio astronomy object is a galaxy."

$B \equiv$ "The measured redshift of the object is 0.150 ± 0.005 ."

$A \equiv$ "Theory X is correct."

$\overline{A} \equiv$ "Theory X is not correct."

$A \equiv$ "The frequency of the signal is between f and $f + df$."

We will need to consider compound propositions like

A,B which asserts that propositions ***A*** and ***B*** are true

A,B|C asserts that propositions ***A*** and ***B*** are true given that proposition **C** is true

Rules for manipulating probabilities

$$\text{Sum rule}: \quad p(A \mid C) + p(\overline{A} \mid C) = 1$$

$$\text{Product rule}: \quad p(A, B \mid C) = p(A \mid C)\, p(B \mid A, C)$$

$$= \underbrace{p(B \mid C)\, p(A \mid B, C)}$$

Re-arrange the two RH sides of product rule gives

Bayes theorem :

$$p(A \mid B, C) = \frac{p(A \mid C)\, p(B \mid A, C)}{p(B \mid C)}$$

Another useful version of the sum rule can be derived from the sum and product rules called the extended sum rule

$$p(A+B|C) = p(A|C) + p(B|C) - P(A,B|C)$$

where *A+B* ≡ proposition *A* is true or *B* is true or both are true

In science we are often reasoning about mutually exclusive propositions for which

$$p(A+B|C) = p(A|C) + p(B|C)$$

Note: *A+B* also commonly written as *A v B*

How to proceed in a Bayesian data analysis?

Identify the terms in Bayes' theorem and solve

Solution often requires repeated use of the product and sum rules

Prior probability **Likelihood**

$$p(H_i \mid D, I) = \frac{p(H_i \mid I) \times p(D \mid H_i, I)}{p(D \mid I)}$$

Posterior probability that H_i is true, given the new data D and prior information I

Normalizing constant

Every item to the right of the vertical bar | is assumed to be true

The likelihood $p(D \mid H_i, I)$, also written as $\mathcal{L}(H_i)$, stands for the probability that we would have gotten the data D that we did, if H_i and I are true.

As a theory of extended logic BPT can be used to find optimal answers to well posed scientific questions for a given state of knowledge, in contrast to a numerical recipe approach.

Two basic problems

1. Model selection (discrete hypothesis space)

"Which one of 2 or more models (hypotheses) is most probable given our current state of knowledge?"

e.g.

- Hypothesis or model M_0 asserts that the star has no planets.
- Hypothesis M_1 asserts that the star has 1 planet.
- Hypothesis M_i asserts that the star has *i* planets.

2. Parameter estimation (continuous hypothesis)

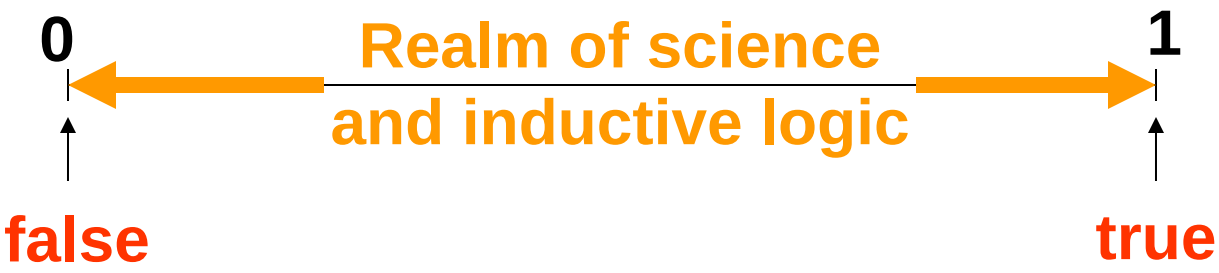
"Assuming the truth of M_1, solve for the probability density distribution for each of the model parameters based on our current state of knowledge."

e.g.

- Hypothesis H asserts that the orbital period is between P and P+dP.

Significance of this development

Probabilities are commonly quantified by a real number between 0 and 1.



The end-points, corresponding to absolutely false and absolutely true, are simply the extreme limits of this infinity of real numbers.

Bayesian probability theory spans the whole range.

Deductive logic is just a special case of Bayesian probability theory in the idealized limit of complete information.

Calculation of a simple Likelihood $p(D_i \mid M,X,I)$

Let d_i represent the i^{th} measured data value . We model d_i by,

$$d_i = f_i(X) + e_i$$

Model prediction for i^{th} data value
for current choice of parameters X

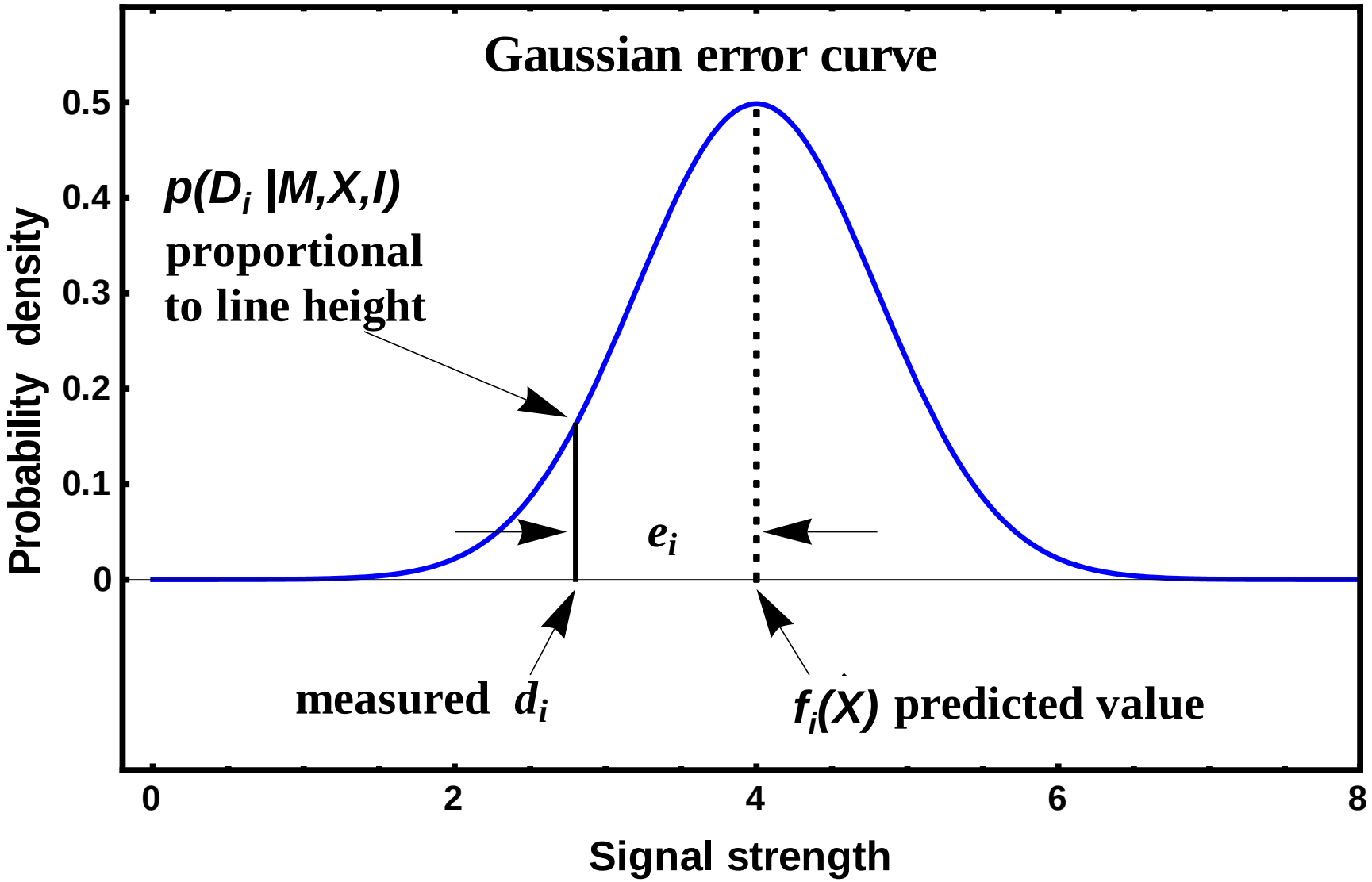
where e_i represents the error component in the measurement.

Since M, X assumed to be true, if it were not for the error e_i, d_i would equal the model prediction f_i .

Now suppose prior information I indicates that e_i has a Gaussian probability distribution. Then

$$p(D_i|M,X,I) = \frac{1}{\sigma_i \sqrt{2\pi}} Exp[-\frac{e_i^2}{2\sigma_i^2}]$$

$$= \frac{1}{\sigma_i \sqrt{2\pi}} Exp\left[-\frac{(d_i - f_i(X))^2}{2\sigma_i^2}\right]$$



Probability of getting a data value d_i a distance e_i away from the predicted value f_i is proportional to the height of the Gaussian error curve at that location.

Calculation of a simple Likelihood $p(D_i \mid M,X,I)$

For independent data the likelihood for the entire data set $D=(D_1,D_2,\ldots,D_N)$ is the product of N Gaussians.

$$p(D|M,X,I) = (2\pi)^{N/2} \left\{ \prod_{i=1}^{N} \sigma_i^{-1} \right\} Exp\left[-0.5 \sum_{i=1}^{N} \frac{(d_i - f_i(X))^2}{\sigma_i^2} \right]$$

The familiar χ^2 statistic used in least-squares

Maximizing the likelihood corresponds to minimizing χ^2

Recall: Bayesian posterior $\propto$ prior $\times$ likelihood

Thus, only for a uniform prior will a least-squares analysis yield the same solution as the Bayesian posterior.

Simple example of when not to use a uniform prior

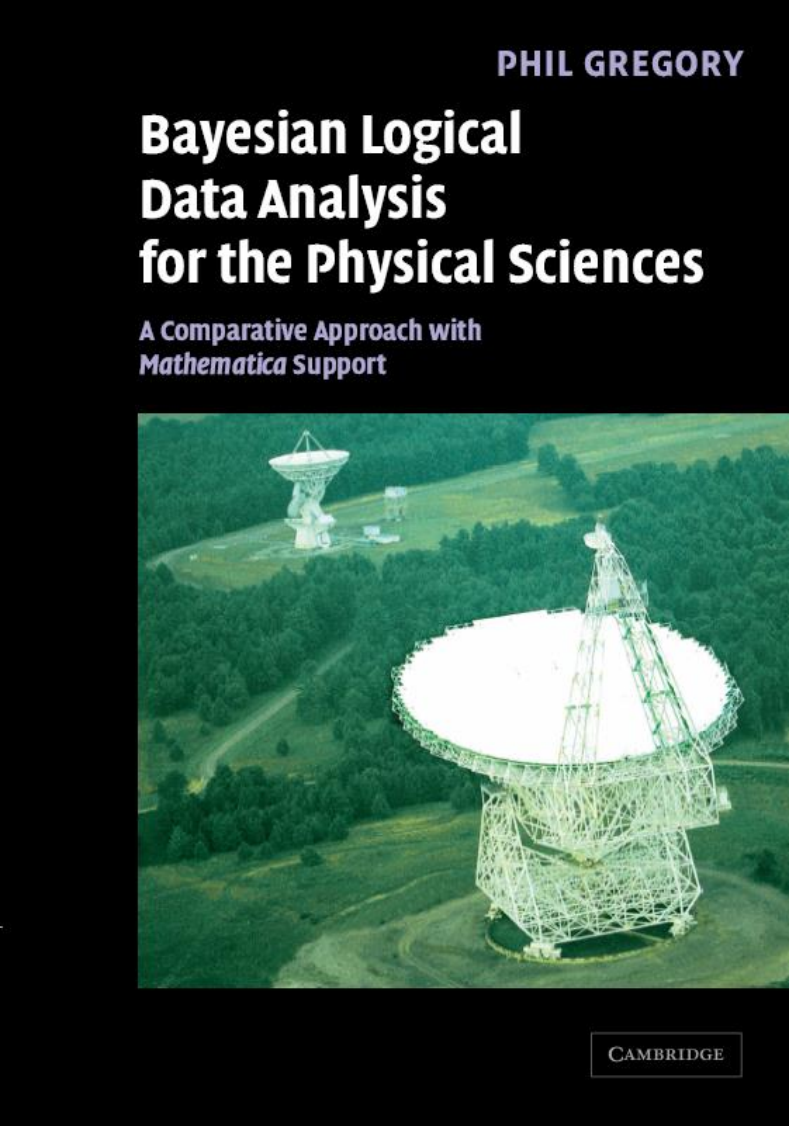
For extra-solar planet detection the prior range for the unknown orbital period *P* is very large from ~1 day to 1000 yr (upper limit set by perturbations from neighboring stars).

Suppose we assume a uniform prior probability density for the *P* parameter. This would imply that we believed that it was ~ 10^4 times more probable that the true period was in the upper decade (10^4 to 10^5 d) of the prior range than in the lowest decade from 1 to 10 d.

$$\frac{\int_{10^4}^{10^5} p(P|M,I)dP}{\int_1^{10} p(P|M,I)dP} = 10^4$$

Usually, expressing great uncertainty in some quantity corresponds more closely to a statement of scale invariance or equal probability per decade. A scale invariant prior has this property.

$$p(\ln P|M,I)\, d\ln P = \frac{d\,\ln P}{\ln(P_{max}/P_{min})}$$



Chapters

1. Role of probability theory in science
2. Probability theory as extended logic
3. The how-to of Bayesian inference
4. Assigning probabilities
5. Frequentist statistical inference
6. What is a statistic?
7. Frequentist hypothesis testing
8. Maximum entropy probabilities
9. Bayesian inference (Gaussian errors)
10. Linear model fitting (Gaussian errors)
11. Nonlinear model fitting
12. Markov chain Monte Carlo
13. Bayesian spectral analysis
14. Bayesian inference (Poisson sampling)

Includes 55 worked examples and many problem sets.

Resources and solutions
Includes free Mathematica based support software available from book website

2014: two supplementary chapters to be added to book website on:
1. Fusion Markov chain Monte Carlo
2. Intro. to hierarchical/multilevel Bayes

Bayesian and classical periodograms of interest for exoplanets

Introduction

Science is concerned with identifying and understanding structures or patterns in nature.

Periodic patterns have proven especially important. This is particular evident in the field of astronomy where the study of periodic phenomena yield:

- **Fundamental properties like mass and distance**
- **Interior structure of stars (stellar seismology)**
- **Extra solar planets**
- **Exotic states of matter (neutron stars & BH)**
- **Fundamental tests of physics**

Any significant advance in our ability to detect periodic phenomena will profoundly affect our capability of unlocking nature's secrets.

A Bayesian Revolution in Spectral Analysis

1) Fourier Power Spectrum (Schuster periodogram 1905)

The use of the Discrete Fourier Transform (DFT) is ubiquitous in spectral analysis as a result of the FFT introduced by Cooley and Tukey in 1965.

$$\text{periodogram} = C(f_n) = \frac{1}{N}\left|\sum_{k=1}^{N} d_k e^{-i2\pi\, n\Delta f\, k\Delta t}\right|^2$$

$$= \frac{1}{N}\left|\text{FFT}\right|^2$$

2) New Insights on the periodogram from Bayesian Probability Theory (BPT)

In 1987 E. T. Jaynes derived the DFT and periodogram directly from the principles of BPT and showed that the periodogram is an optimum statistic for the detection of a single stationary sinusoidal signal in the presence of independent Gaussian noise of variance σ^2.

He showed that the probability of the frequency of a periodic signal is given to a very good approximation by,

$$p(f_n \mid D, I) \propto \exp\left\{\frac{C(f_n)}{\sigma^2}\right\}$$

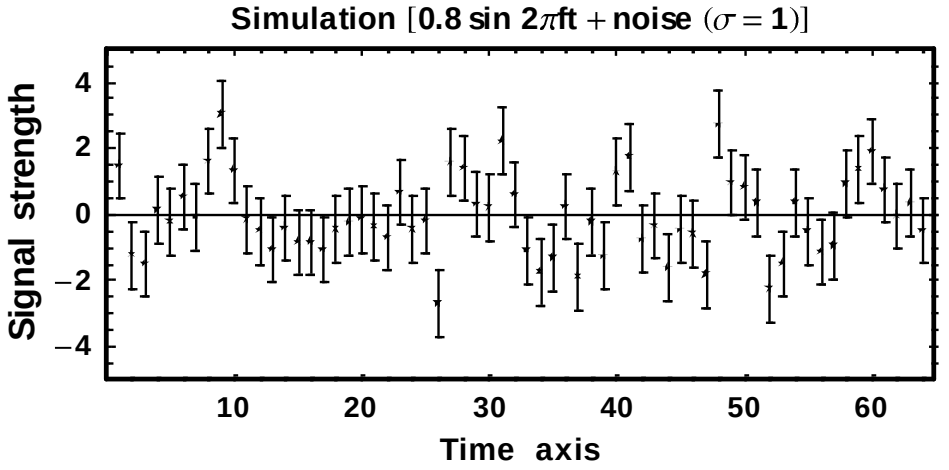
$$p(f_n \mid D, I) \propto \exp\left\{\frac{C(f_n)}{\sigma^2}\right\}$$

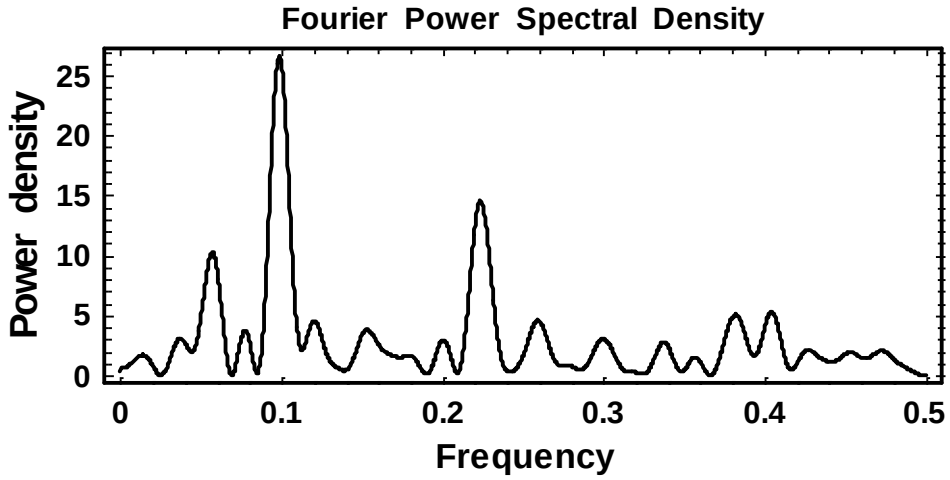
Thus $C(f_n)$ is indeed fundamental to spectral analysis but not because it is itself a satisfactory spectrum estimator.

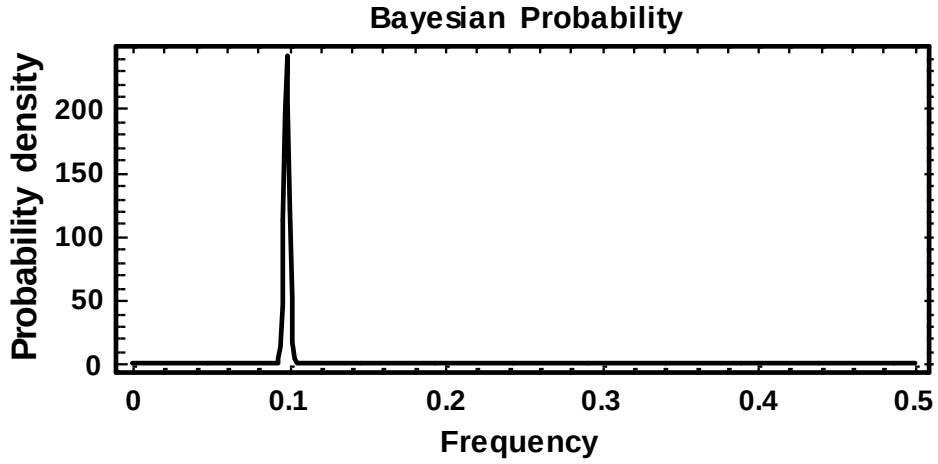
The proper algorithm to convert $C(f_n)$ to $p(f_n|D,I)$ involves first dividing $C(f_n)$ by the noise variance and then exponentiation.

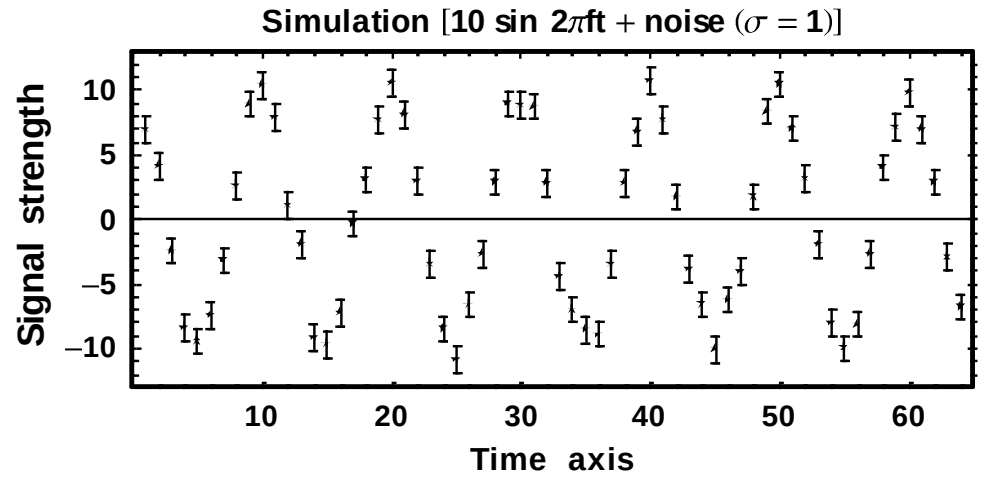
This naturally suppresses spurious ripples at the base of the periodogram as well as linear smoothing; but does it by attenuation rather than smearing, and therefore does not lose any resolution.

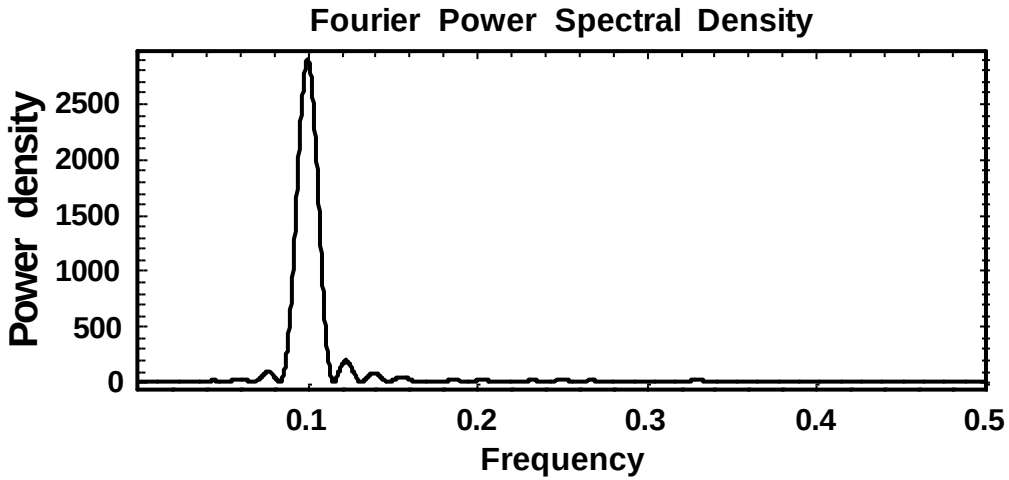
The Bayesian nonlinear processing of $C(f_n)$ also yields, when the data give evidence for them, arbitrarily sharp spectral peaks.

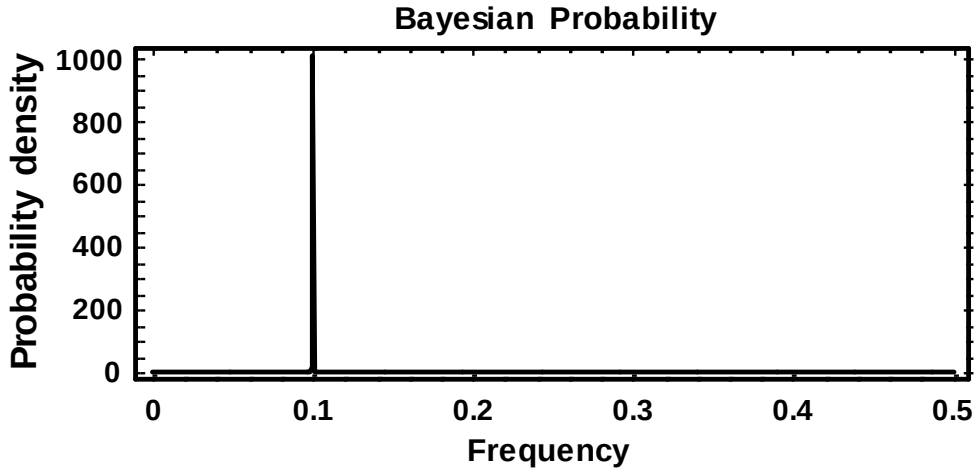

Simulation [0.8 sin 2πft + noise (σ = 1)]
Signal strength
4
2
0
−2
−4
10
20
30
40
50
60
Time axis


Fourier Power Spectral Density
Power density
25
20
15
10
5
0
0
0.1
0.2
0.3
0.4
0.5
Frequency


Bayesian Probability
Probability density
200
150
100
50
0
0
0.1
0.2
0.3
0.4
0.5
Frequency


Simulation [10 sin 2πft + noise (σ = 1)]
Signal strength
10
5
0
−5
−10
10
20
30
40
50
60
Time axis


Fourier Power Spectral Density
Power density
2500
2000
1500
1000
500
0
0.1
0.2
0.3
0.4
0.5
Frequency


Bayesian Probability
Probability density
1000
800
600
400
200
0
0
0.1
0.2
0.3
0.4
0.5
Frequency

What if σ is unknown

$$p(f_n \mid D, I) \propto \exp\left\{\frac{C(f_n)}{\sigma^2}\right\}$$

This equation assumes that the noise variance is a known quantity.

In some situations, the noise is not well understood, i.e., our state of knowledge is less certain. Even if the measurement apparatus noise is well understood, the data may contain a greater complexity of phenomena than the current signal model incorporates.

Again, Bayesian inference can readily handle this situation by treating the noise variance as a nuisance parameter with a prior distribution reflecting our uncertainty in this parameter. We need to integrate over this parameter to compute *p(f$_n$|D,I)*.

The resulting posterior can be expressed in the form of a Student's *t* distribution. The corresponding result for estimating the frequency of single sinusoidal signal (Bretthorst 1988, Bayesian Spectrum Analysis and Parameter Estimation, Springer) is given approximately by

$$p(f_n \mid D, I) \propto \left[1 - \frac{2C(f_n)}{N\overline{d^2}}\right]^{\frac{2-N}{2}}$$

where

$$\overline{d^2} = \frac{1}{N}\sum_j d_i^2$$

Bayesian Spectrum Analysis with Strong Prior Information of the Signal Model

Larry Bretthorst (Jaynes' last PhD student) extended Jaynes' work to more complex signal models with additive Gaussian noise and revolutionized the analysis of Nuclear Magnetic Resonance (NMR) signals.

Here one is dealing with multiple damped sinusoids.

See http://bayes.wustl.edu/ for a copy of Larry Bretthorst's papers and book.

Varian Corporation now offer an expert analysis package with their new NMR machines based on Bretthorst's Bayesian algorithm.

What if σ is unknown

$$p(f_n \mid D, I) \propto \exp\left\{\frac{C(f_n)}{\sigma^2}\right\}$$

This equation assumes that the noise variance is a known quantity.

In some situations, the noise is not well understood, i.e., our state of knowledge is less certain. Even if the measurement apparatus noise is well understood, the data may contain a greater complexity of phenomena than the current signal model incorporates.

Again, Bayesian inference can readily handle this situation by treating the noise variance as a nuisance parameter with a prior distribution reflecting our uncertainty in this parameter. We need to integrate over this parameter to compute *p(f$_n$|D,I)*.

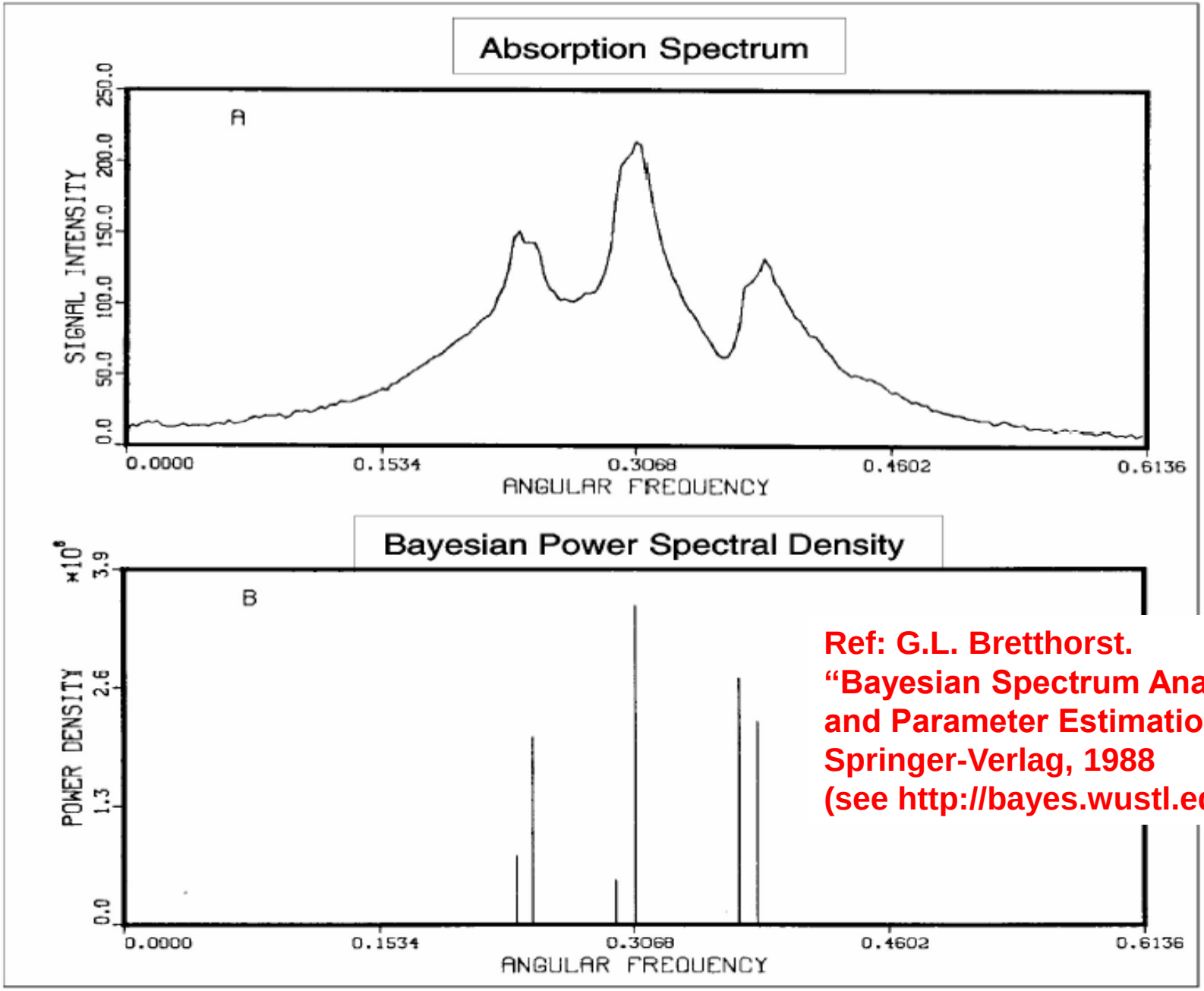
The resulting posterior can be expressed in the form of a Student's *t* distribution. The corresponding result for estimating the frequency of single sinusoidal signal (Bretthorst 1988, Bayesian Spectrum Analysis and Parameter Estimation, Springer) is given approximately by

$$p(f_n \mid D, I) \propto \left[1 - \frac{2C(f_n)}{N\overline{d^2}}\right]^{\frac{2-N}{2}}$$

where

$$\overline{d^2} = \frac{1}{N}\sum_j d_i^2$$

Analysis of Nuclear Magnetic Resonance Free Induction Decay Data



Ref: G.L. Bretthorst. "Bayesian Spectrum Analysis and Parameter Estimation", Springer-Verlag, 1988 (see http://bayes.wustl.edu)

The Bretthorst periodogram:

Bretthorst generalized Jaynes' insights to a broader range of single-frequency and multi-frequency estimation problems and sampling conditions.

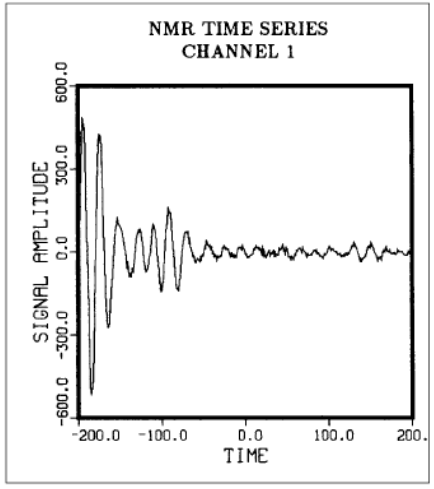
In the single-frequency case which we examine here, he established a connection between the Bayesian results and an existing frequentist statistic known as the Lomb-Scargle periodigram, which is a widely used replacement for the Schuster periodogram in the case of non-uniform sampling.

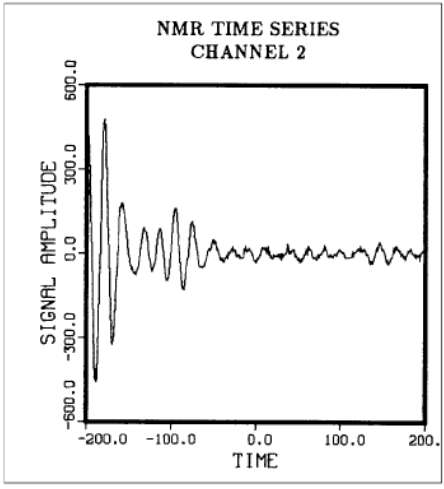
Bretthorst, G.L. (2001), American Institute of Physics Conference Proceedings, 568, pp. 241.

Bretthorst's analysis allows for the following complications:

1. Either real or quadrature data sampling. Quadrature data involves measurements of the real and imaginary components of a complex signal.

The figure show an example of quadrature signals occurring in NMR.





The Bretthorst periodogram

2. Allows for uniform or non-uniform sampling and for quadrature data with non-simultaneous sampling.

The analysis does not require the real and imaginary data samples to be simultaneous and successive samples can be unequally spaced in time.

3. Allows for non-stationary single sinusoid model of the form

(real channel)

$$d_R(t_i) = A\cos(2\pi f t_i - \theta)\,Z(t_i) + B\sin(2\pi f t_i - \theta)\,Z(t_i) + e_R(t_i),$$

(imaginary channel)

$$d_I(t_j') = A\cos(2\pi f t_j' - \theta)\,Z(t_j') + B\sin(2\pi f t_j' - \theta)\,Z(t_j') + e_I(t_j'),$$

The function $Z(t_i)$ describes an arbitrary modulation of the amplitude, e.g., exponential decay as exhibited in NMR signals.

In this analysis $Z(t_i)$ is assumed to be known, but in other analysis he allows it to have unknown parameters.

The Bretthorst periodogram

The angle θ is defined in such a way as to make the cosine and sine functions orthogonal on the discretely sampled times. In general, θ is frequency dependent.

Note: if the data are simultaneously sampled, $t_i = t'_j$, then the orthogonal condition is automatically satisfied so $\theta = 0$.

4. The noise terms $e_R(t_i)$ and $e_I(t'_j)$ are assumed to be IID Gaussian with an unknown σ. Thus, is a nuisance parameter, which is assumed to have a Jeffreys prior. By marginalizing over σ, any variability in the data that is not described by the model is assumed to be noise.

The Bretthorst periodogram

In this problem the main parameter of interest is the frequency ***f***. To compute ***p (f |D,I)*** we need to marginalize over the two amplitude parameters ***A***, ***B***, and ***σ***.

$$p(f|D,I) = \int dA\, dB\, d\sigma\, p(f,A,B,\sigma|D_R,D_I,I)$$

The RH side of this equation can be factored using Bayes' theorem and the product rule to yield

$$p(f|D,I) \propto \int dA\, dB\, d\sigma\; p(f|I)p(A|I)p(B|I)p(\sigma|I) \times$$

$$p(D_R|f,A,B,\sigma,I)\; p(D_I|f,A,B,\sigma,I)$$

Bretthorst assigns uniform priors for f, A & B and a Jeffreys prior for σ.

The Bretthorst periodogram

It turns out that the triple integral can be performed analytically using simple changes in the variables.

The final Bayesian expression for *p* (*f* |*D*,*I*), after marginalizing over amplitudes *A*, *B* & *σ* is given by

$$p(f|D,I) \propto \frac{1}{\sqrt{C(f)S(f)}}\left[N\overline{d^2} - \overline{h^2}\right]^{\frac{2-N}{2}},$$

where $$\overline{h^2} = \frac{R(f)^2}{C(f)} + \frac{I(f)^2}{S(f)},$$

The Bretthorst periodogram

where

$$R(f) \equiv \sum_{i=1}^{N_R} d_R(t_i)\cos(2\pi f t_i - \theta)Z(t_i) - \sum_{j=1}^{N_I} d_I(t'_j)\sin(2\pi f t'_j - \theta)Z(t'_j),$$

$$I(f) \equiv \sum_{i=1}^{N_R} d_R(t_i)\sin(2\pi f t_i - \theta)Z(t_i) + \sum_{j=1}^{N_I} d_I(t'_j)\cos(2\pi f t'_j - \theta)Z(t'_j),$$

$$C(f) \equiv \sum_{i=1}^{N_R} \cos^2(2\pi f t_i - \theta)Z(t_i)^2 + \sum_{j=1}^{N_I} \sin^2(2\pi f t'_j - \theta)Z(t'_j)^2$$

and

$$S(f) \equiv \sum_{i=1}^{N_R} \sin^2(2\pi f t_i - \theta)Z(t_i)^2 + \sum_{j=1}^{N_I} \cos^2(2\pi f t'_j - \theta)Z(t'_j)^2.$$

$$\theta = \frac{1}{2}\tan^{-1}\left[\frac{\sum_{i=1}^{N_R}\sin(4\pi f t_i)Z(t_i)^2 - \sum_{j=1}^{N_I}\sin(4\pi f t'_j)Z(t'_j)^2}{\sum_{i=1}^{N_R}\cos(4\pi f t_i)Z(t_i)^2 - \sum_{j=1}^{N_I}\cos(4\pi f t'_j)Z(t'_j)^2}\right]$$

Result

$$P(f|DI) \propto \frac{1}{\sqrt{C(f)S(f)}}\left[N\overline{d^2} - \overline{h^2}\right]^{\frac{2-N}{2}}$$

where the sufficient statistic $\overline{h^2}$ is given by $\overline{h^2} = \frac{R(f)^2}{C(f)} + \frac{I(f)^2}{S(f)}$

Simplifications

1. **When the data are real and the sinusoid is stationary, the sufficient statistic for single frequency estimation is the Lomb-Scargle periodogram; not the Schuster periodogram (power spectrum). However, the Schuster periodogram is often an excellent approximation.**

2. **When the data are real, but *Z(t)* is not constant, then $\overline{h^2}$ generalizes the Lomb-Scargle periodogram in a very straightforward manner to account for the decay of the signal.**

3. **For uniformly sampled quadrature data when the sinusoid is stationary, $\overline{h^2}$ reduces to a Schuster periodogram of the data.**

$$p(f_n \mid D, I) \propto \left[1 - \frac{2C(f_n)}{N\overline{d^2}}\right]^{\frac{2-N}{2}}$$

Strong prior information: signals obey Kepler's laws

The radial velocity equation, a nonlinear model

model prediction $f(t_i) = V + K[\cos\{\theta(t_i + \chi P) + \omega\} + e\cos\omega]$

V = a constant velocity.

K = velocity semi-amplitude = $\frac{2\pi\ a\ \sin i}{P\ \sqrt{1-e^2}}$,

where a = semi-major axis and i = inclination.

P = the orbital period.

e = the orbital eccentricity.

ω = the longitude of periastron.

χ = the fraction of an orbit, prior to the start of data taking, that periastron occurred at. Thus, χP = the number of days prior to $t_i = 0$ that the star was at periastron, for an orbital period of P days.

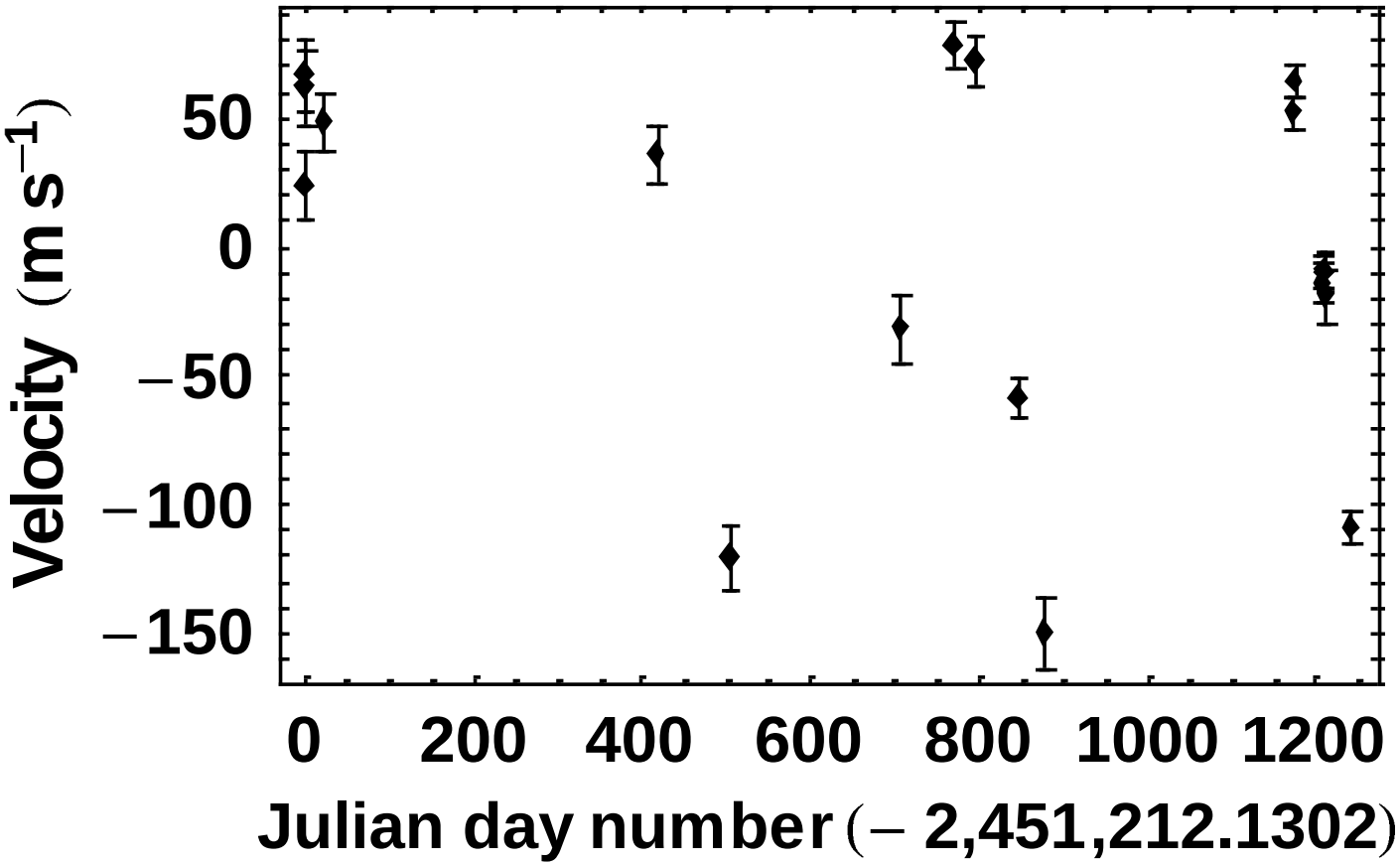
$\theta(t_i+\chi P)$ = the true anomaly, the angle of the star in its orbit relative to periastron at time t_i.

$$\frac{d\theta}{dt} - \frac{2\pi[1 + e\cos\theta(t_i + \chi\ P)]^2}{P(1-e^2)^{3/2}} = 0$$

Application to extra-solar planet data

Sparse radial velocity measurements for HD 73526

(ref. Tinney, G. C. 2003, Astrophsical Journal, 587, p. 423)

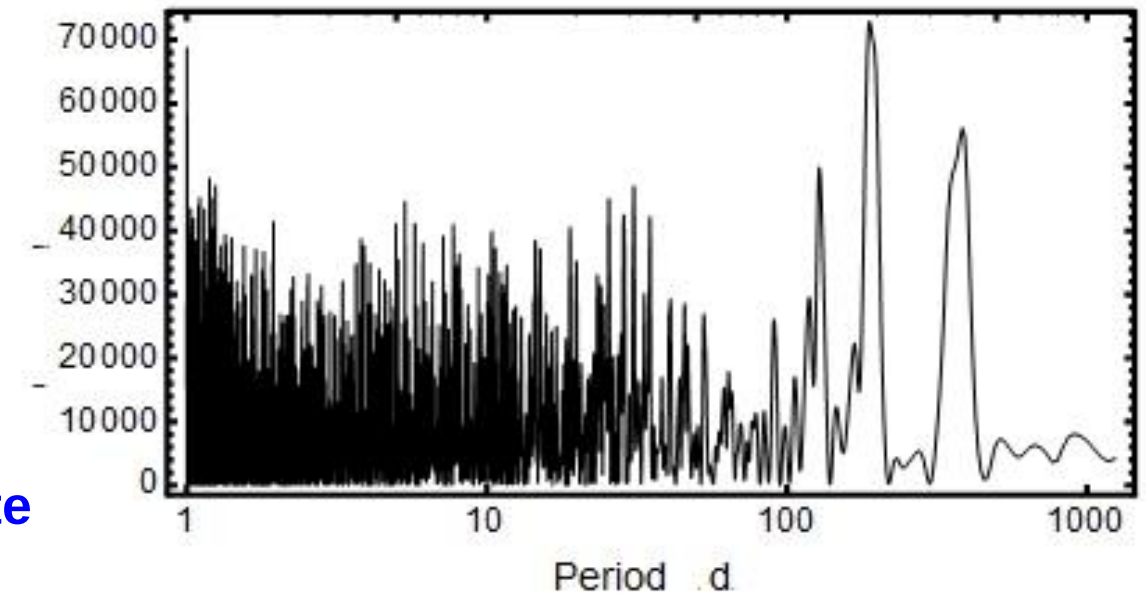


Conventional Nonlinear least-squares analysis requires a good initial guess at the parameter values.

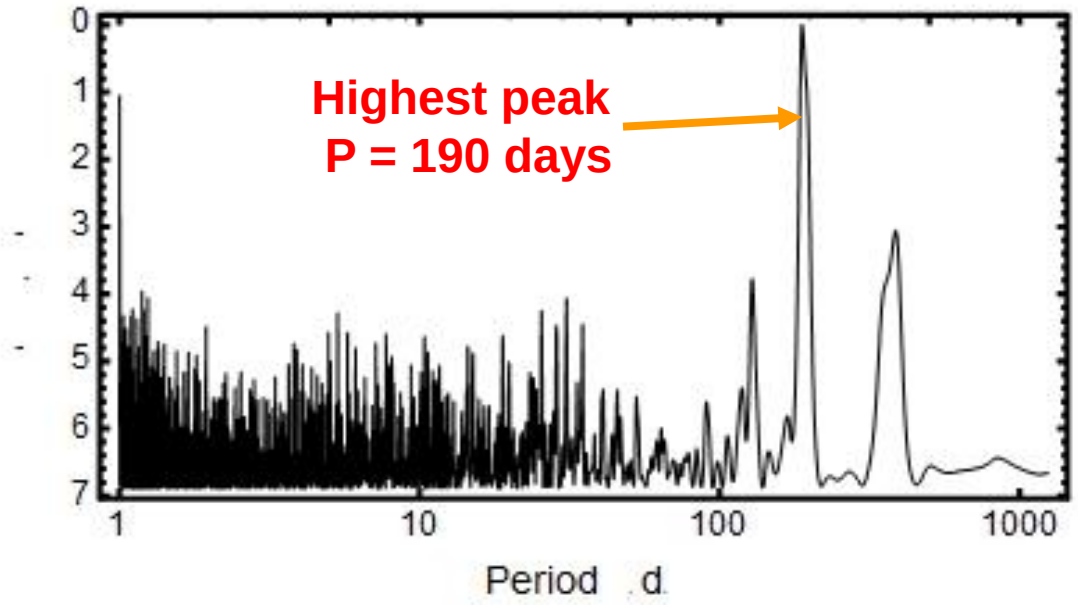
Need to use some form of periodogram to estimate the orbital period.

Here is a comparison of the Lomb-Scargle and Bretthorst's Bayesian generalization.

Lomb-Scargle Periodogram



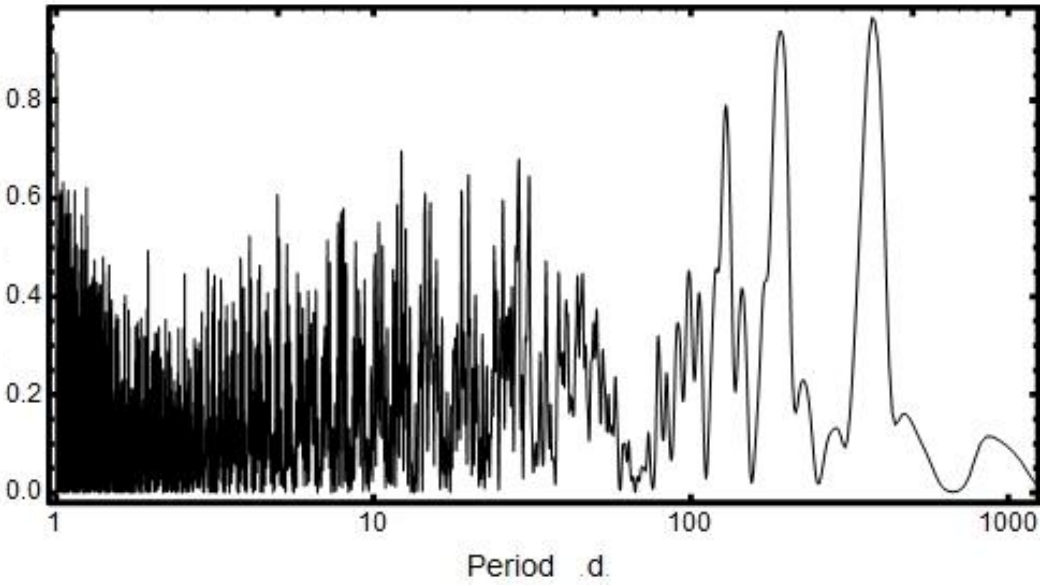
Bretthorst's Bayesian periodogram



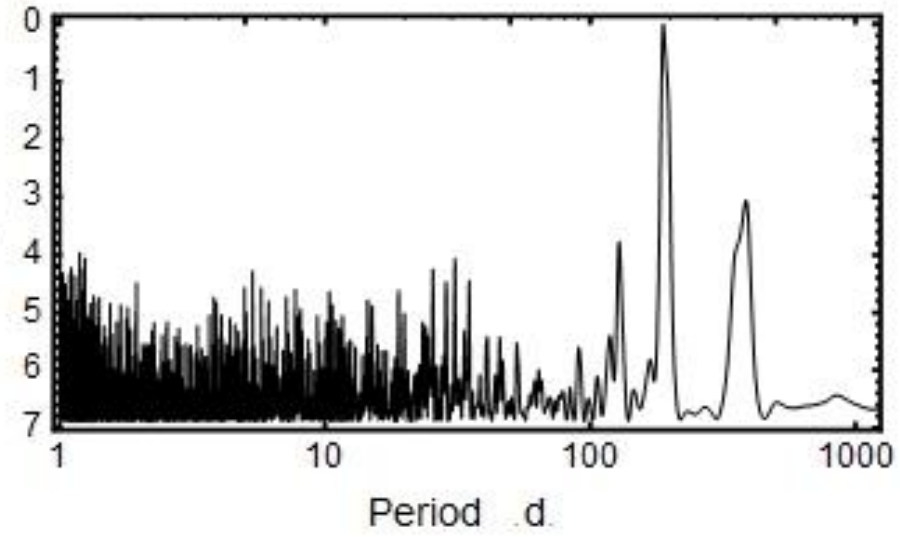
In 2009 (Zechmeister & Kurster, A&A, 496, 577, 2009) introduced a generalized Lomb-Scargle (GLS) periodogram that allows for a floating offset and weights.

What is missing now is a Bayesian version of the GLS.

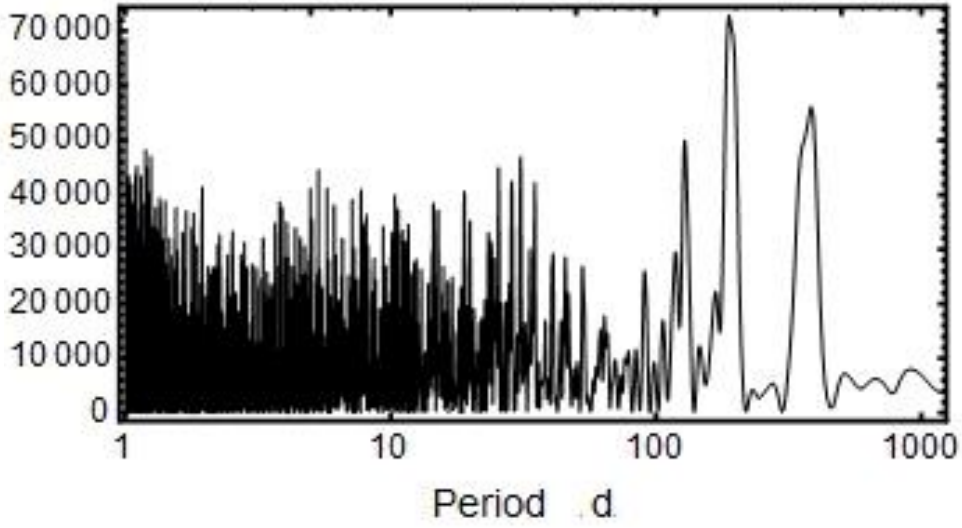
Generalized Lomb-Scargle Periodogram



Bretthorst's Bayesian periodogram



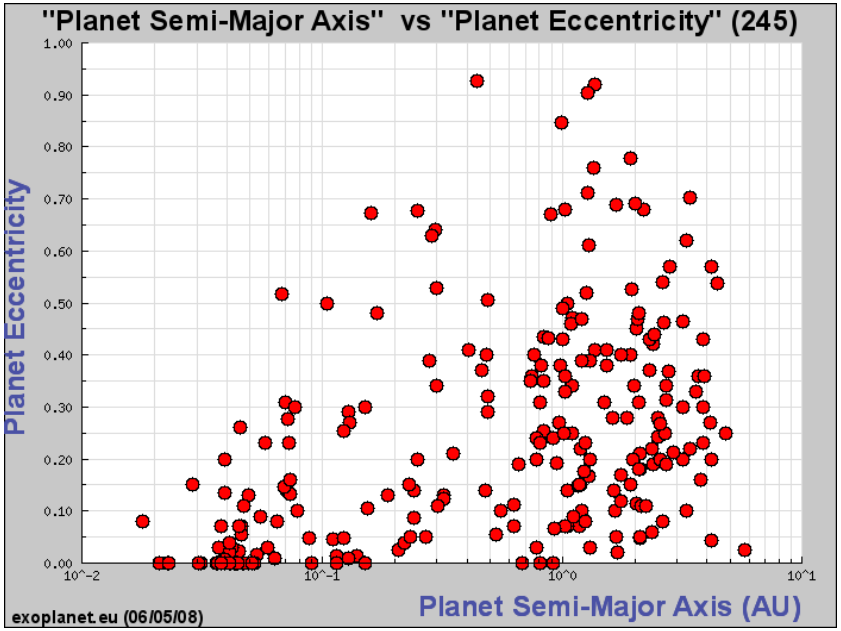
Lomb-Scargle Periodogram



Problem

The Lomb-Scargle periodogram, and Bretthorst's Bayesian generalization, assume a sinusoidal signal which is only optimum for circular orbits.

Why not develop a Bayesian Kepler periodogram designed for all Kepler orbits?



In 2005, Eric Ford and I independently published Kepler Periodograms based on a Markov chain Monte Carlo (MCMC) approach.

This will be the subject of the next lecture.

Markov chain Monte Carlo (MCMC)

Strong prior information: signals obey Kepler's laws

The radial velocity equation, a nonlinear model

model prediction $f(t_i) = V + K[\cos\{\theta(t_i + \chi P) + \omega\} + e\cos\omega]$

V = a constant velocity.

K = velocity semi-amplitude = $\frac{2\pi\ a\ \sin i}{P\ \sqrt{1-e^2}}$,

where a = semi-major axis and i = inclination.

P = the orbital period.

e = the orbital eccentricity.

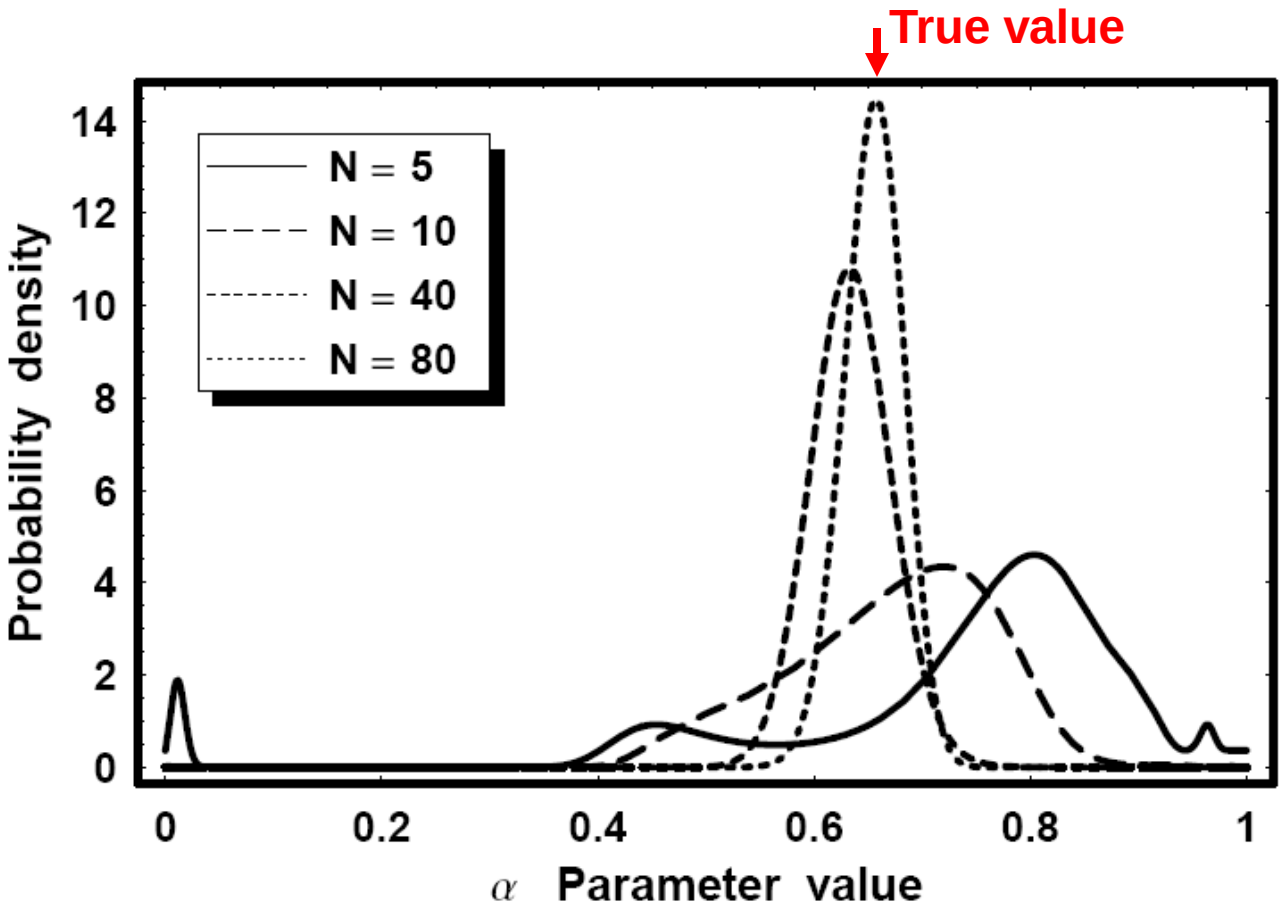
ω = the longitude of periastron.

χ = the fraction of an orbit, prior to the start of data taking, that periastron occurred at. Thus, χP = the number of days prior to $t_i = 0$ that the star was at periastron, for an orbital period of P days.

$\theta(t_i+\chi P)$ = the true anomaly, the angle of the star in its orbit relative to periastron at time t_i.

$$\frac{d\theta}{dt} - \frac{2\pi[1 + e\cos\theta(t_i + \chi\ P)]^2}{P(1-e^2)^{3/2}} = 0$$

The challenge of nonlinear models, multiple peaks



The Bayesian posterior density for a nonlinear model with a single parameter, α, for 4 simulated data sets of different size ranging from *N* = 5 to *N* = 80. The *N* = 5 case has the broadest distribution and exhibits 4 maxima.

Asymptotic theory says that the maximum likelihood estimator becomes more unbiased, more normally distributed and of smaller variance as the sample size becomes larger.

return

Bayesian parameter estimation

To find the marginal posterior probability density function (PDF) for the orbital period *P*, we need to integrate the joint posterior over all the other parameters.

$$p(P|D,M_1,I) = \int dK\, dV\, d\chi\, de\, d\omega\, ds\, p(P,K,V,\chi,e,\omega,s|D,M_1,I)$$

Marginal PDF for P

An 8 planet model has 42 parameters

Joint posterior probability density function (PDF) for the parameters

Markov chain Monte Carlo (MCMC) algorithms provide a powerful means for efficiently computing integrals in many dimensions to within a constant factor. This factor is not required for parameter estimation.

After an initial burn-in period, the MCMC produces an equilibrium distribution of samples in parameter space such that the density of samples is proportional to the target posterior PDF.

It is very efficient because, unlike straight Mont Carlo integration, it doesn't waste time exploring regions where the joint posterior is very small.

Starting point: Metropolis-Hastings MCMC algorithm

P(X|D,M,I) = target posterior probability distribution
(X represents the set of model parameters)

1. Choose X_0 an initial location in the parameter space . Set $t = 0$.

2. Repeat {

– Obtain a new sample Y from a proposal distribution $q(Y \mid X_t)$ that is easy to evaluate . $q(Y \mid X_t)$ can have almost any form.

I use a Gaussian proposal distribution. i.e., Normal distribution $N(X_t, \sigma)$

– Sample a Uniform (0, 1) random variable U.

–If $U \leq \frac{p(Y \mid D, I)}{p(X_t \mid D, I)} \times \frac{q(X_t \mid Y)}{q(Y \mid X_t)}$ then set $X_{t+1} = Y$

otherwise set $X_{t+1} = X_t$

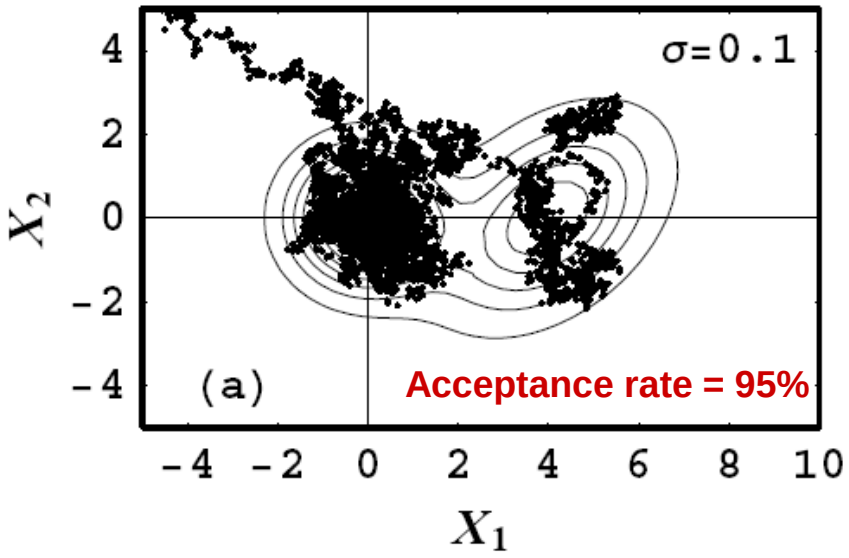
This factor =1
for a symmetric proposal
distribution like a Gaussian

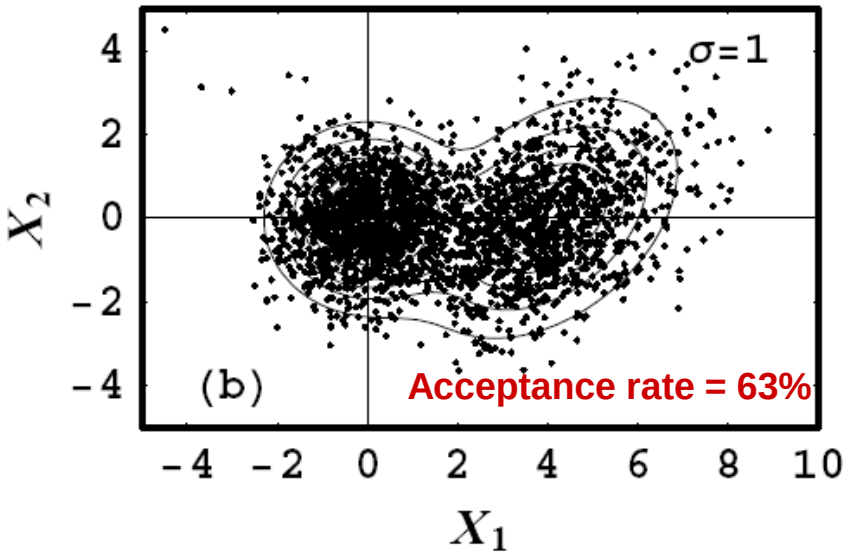
– Increment t }

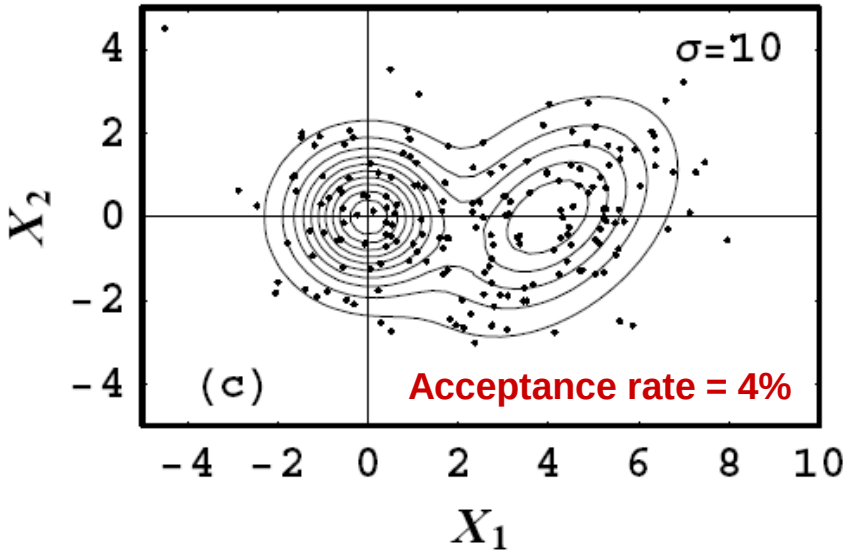
conditions

return

Toy MCMC simulations: the efficiency depends on tuning proposal distribution σ's. Can be a very difficult challenge for many parameters.







In this example the posterior probability distribution consists of two 2 dimensional Gaussians indicated by the contours

Parallel tempering MCMC

The simple Metropolis-Hastings MCMC algorithm can run into difficulties if the probability distribution is multi-modal with widely separated peaks. It can fail to fully explore all peaks which contain significant probability, especially if some of the peaks are very narrow.

One solution is to run multiple Metropolis-Hastings simulations in parallel, employing probability distributions of the kind

$$\pi(X \mid D, M, \beta, I) = p(X \mid M, I)\ p(D \mid X, M, I)^{\beta} \quad (0 < \beta \leqslant 1)$$

Typical set of β values = 0.09,0.15,0.22,0.35,0.48,0.61,0.78,1.0

β = 1 corresponds to our desired target distribution. The others correspond to progressively flatter probability distributions.

At intervals, a pair of adjacent simulations are chosen at random and a proposal made to swap their parameter states. The swap allows for an exchange of information across the ladder of simulations.

In the low β simulations, radically different configurations can arise, whereas at higher β, a configuration is given the chance to refine itself.

Final results are based on samples from the β = 1 simulation. Samples from the other simulations provide one way to evaluate the marginal likelihood for model selection problems.

Fusion Markov chain Monte Carlo (FMCMC)

Fusion MCMC

All of the studies reported here are implemented in a Bayesian framework using Fusion MCMC, a very general nonlinear model fitting method applicable to a wide range of problems.

In the exoplanet problem, the combination of nonlinear model, sparse sampling, multiple planets and huge prior period range of 0.5 d to 1000 yr, yields a highly multi-modal target distribution which is a problem for a straight Metropolis algorithm.

I have developed a new Markov chain Monte Carlo algorithm called Fusion MCMC. It combines Metropolis with:

a) parallel tempering*,
b) genetic crossover,
c) simulated annealing .

Each of these features facilitate the detection of a global minimum in chi-squared in a multi-modal environment. By combining all three, the algorithm greatly increases the probability of realizing this goal.

*** Also known as *Exchange Monte Carlo* (Hukushima & Nemoto 1996)**

Controlled Statistical Fusion

This statistical fusion approach has been achieved through the development of a unique multi-stage control system, hence the term controlled statistical fusion.

The control system also automates the tuning of MCMC proposal distributions for efficient exploration of the model parameter space even when the parameters are highly correlated.

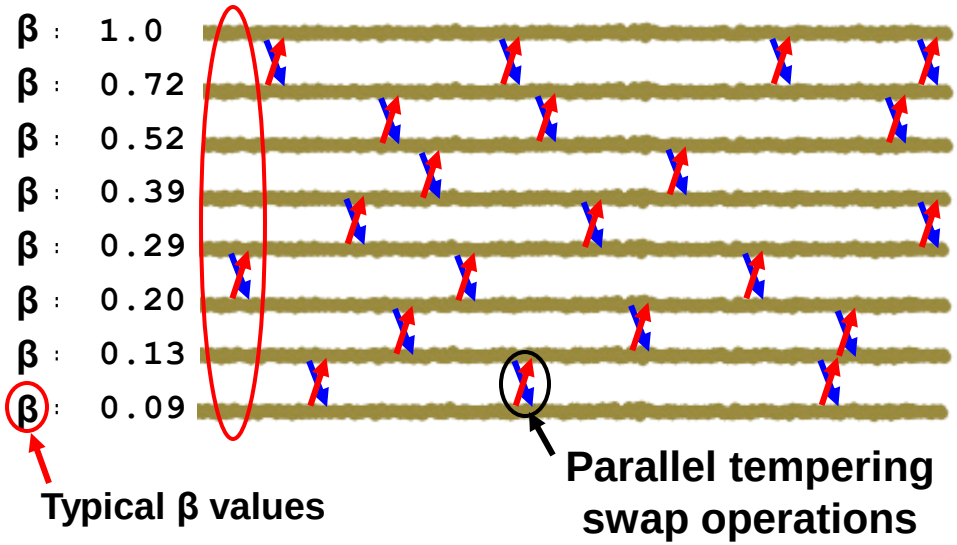
The fusion MCMC algorithm is implemented in *Mathematica* using parallelized code which utilizes all computer cores available.

Early discussion on Fusion MCMC: Gregory, P. C., Chapter 7 in Astrostatistical Challenges for the New Astronomy, Springer Series in Astrostatistics, Hilbe, J.M (ed), 2012, New York:Springer, pp. 121-148

Details

Fusion MCMC

8 parallel tempering Metropolis chains



Output at each iteration

parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike

8 parallel chains employed to avoid becoming trapped in a local probability maximum. Each samples a distributions of the form

$$\pi(\vec{X}|D, M_i, I, \beta) \propto p(\vec{X}|M_i, I) \times p(D|\vec{X}, M_i, I)^{\beta}$$

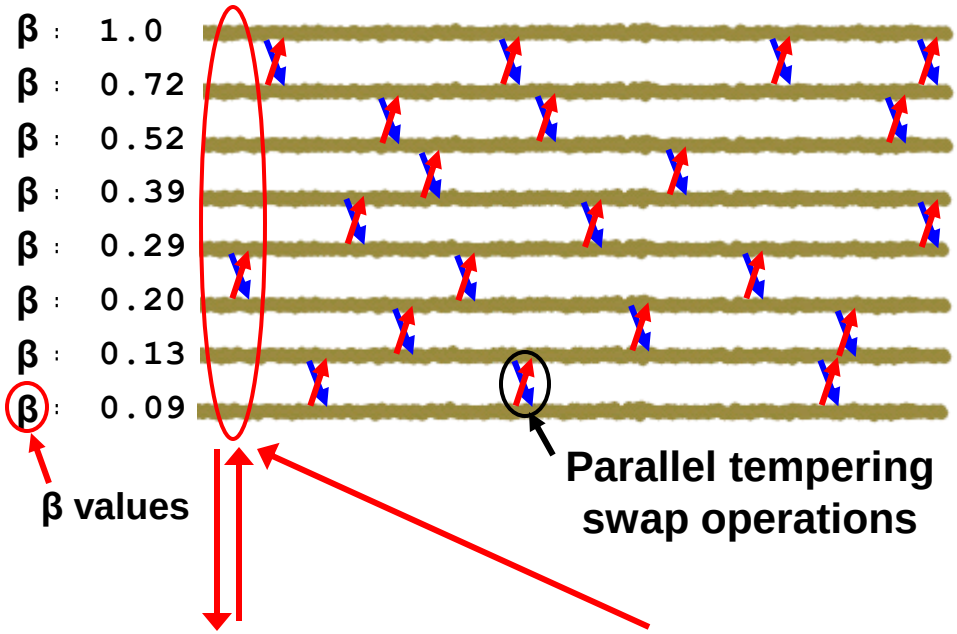
Range of β = 1 to 0

β = 1 corresponds to our desired target probability distribution. The others correspond to progressively flatter distributions.

At intervals, a pair of adjacent chains are chosen at random and a proposal made to swap their parameter states. The swap allows for an exchange of information across the ladder of chains.

Fusion MCMC

8 parallel tempering Metropolis chains



Output at each iteration

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

Anneal Gaussian proposal σ's → Refine & update Gaussian proposal σ's

2 stage proposal σ control system

error signal =
(actual joint acceptance rate – 0.25)

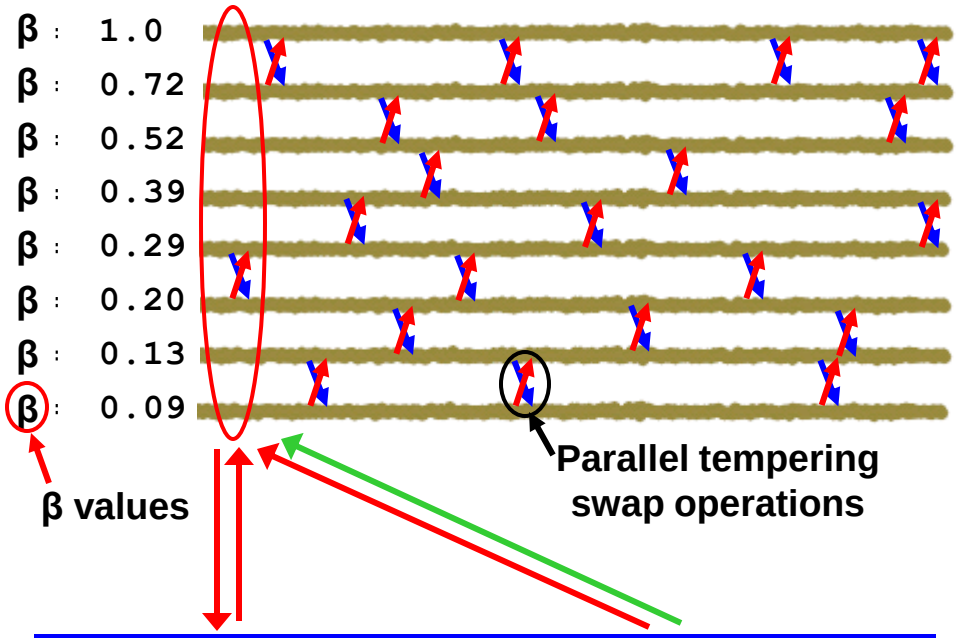
Effectively defines burn-in interval

Portion of Control System that automates the selection of an efficient set of σ values for the independent Gaussian proposal distributions ('I' proposals).

Return

Fusion MCMC

8 parallel tempering Metropolis chains



Output at each iteration

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

parameters, logprior + loglike, logprior + loglike

Monitor for parameters with peak probability

Peak parameter set:
If (logprior + loglike) > previous best by a threshold then update and reset burn-in

Anneal Gaussian proposal σ's

Refine & update Gaussian proposal σ's

2 stage proposal σ control system

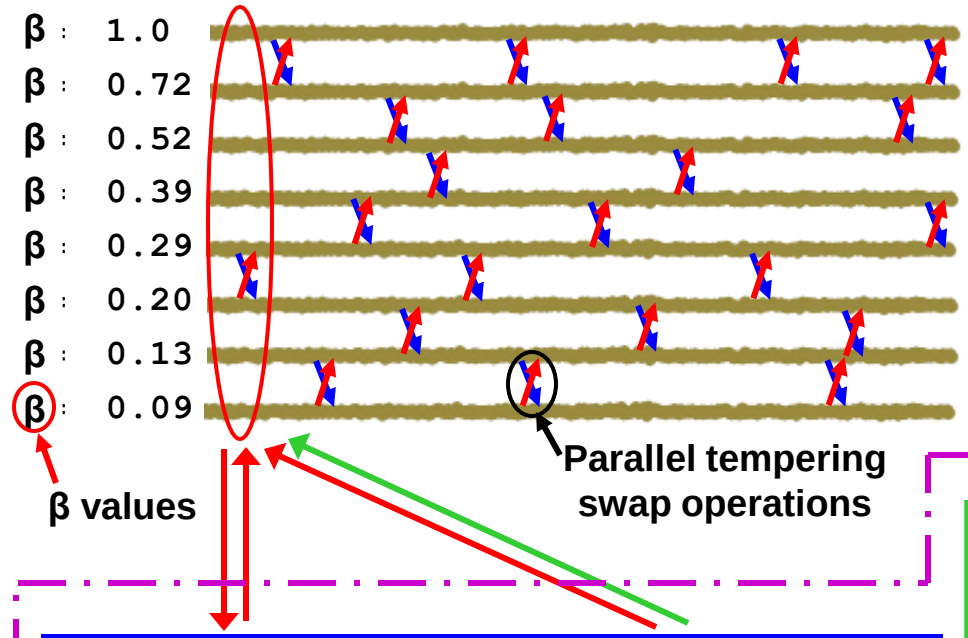
error signal =
(actual joint acceptance rate – 0.25)

Effectively defines burn-in interval

Part of control system that allows the MCMC to adaptively restart if it detects a significantly more probable peak in any chain.

Fusion MCMC

8 parallel tempering Metropolis chains



Output at each iteration

parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike
parameters, logprior + loglike, logprior + loglike

Peak parameter set:
If (logprior + loglike) > previous best by a threshold then update and reset burn-in

Monitor for parameters with peak probability

Anneal Gaussian proposal σ's

Refine & update Gaussian proposal σ's

2 stage proposal σ control system

error signal =
(actual joint acceptance rate – 0.25)

Effectively defines burn-in interval

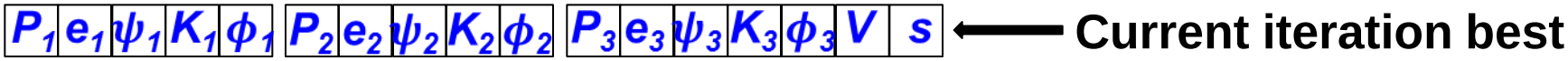
Genetic algorithm

Every 40th iteration perform gene crossover operation to breed a more probable parameter set.

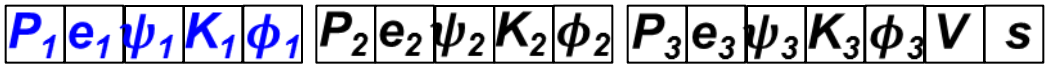
MCMC adaptive control system

Genetic breeding component

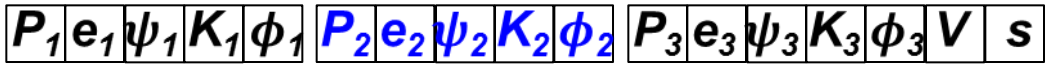
The 3 planet parameter string is divided into 3 genes as shown.



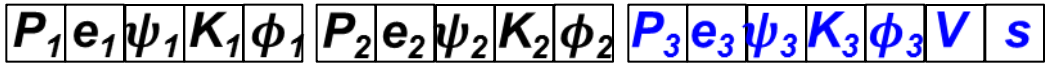
Parameters *V* and *s* are included as part of the last gene.



The genes from current iteration best are substituted for the corresponding gene in the best to date one at a time.



This form of breeding leads to transitions to higher Log[prior x Likelihood] values ~ 1.7 times MCMC steps.

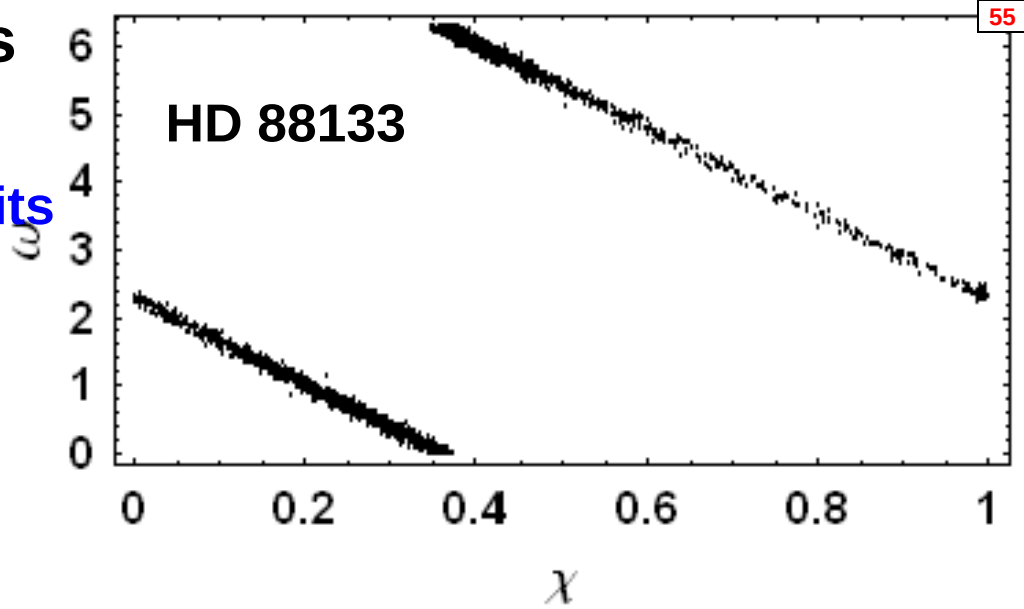


Substitutions from "Current iteration 2nd best" are also included and found to be ~ 70% as effective as "Current iteration best" substitutions. Overall, genetic breeding 2.8 times as effective as MCMC iterations.

Reverse substitutions from "Best to date" into "Current iteration best" were found to approximately 17 times less effective and not included.

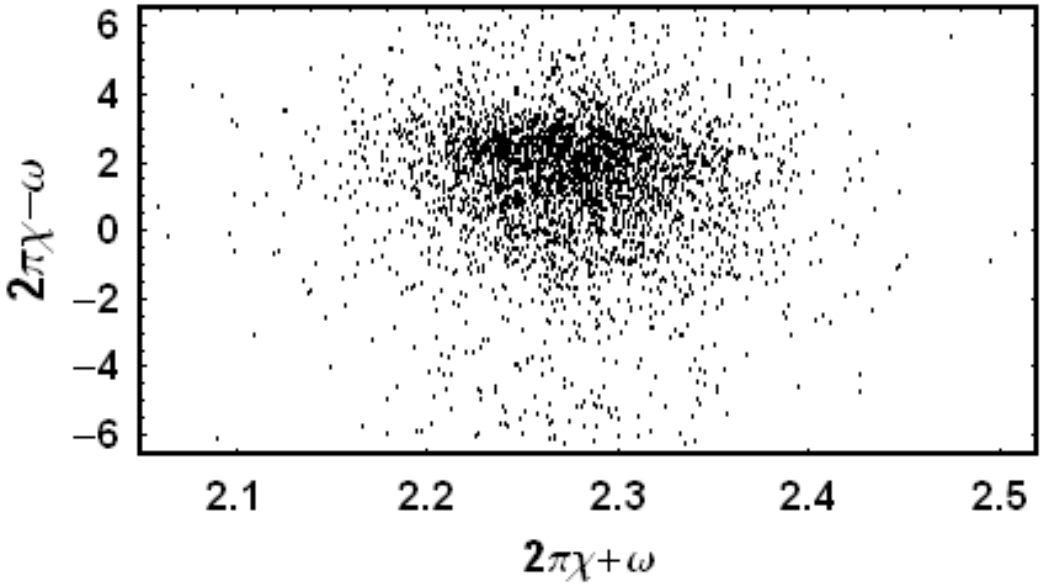
Highly correlated parameters

Top figure shows an exoplanet example. For low eccentricity orbits the parameters ω and χ are not separately well determined. This shows up as a strong correlation between ω and χ.



One option re-parameterization

The combination $2\pi\chi+\omega$ is well determined for all eccentricities. Although $2\pi\chi-\omega$ is not well determined for low eccentricities, it is at least orthogonal to $2\pi\chi+\omega$ as shown.



Another option

Algorithm learns about the parameter correlations during the burn-in and generates proposals with these statistical correlations.

How to deal with highly correlated parameters

Using only independent Gaussian proposals the ('l' scheme) the σ's need to be very small for any proposal to be accepted and consequently convergence is very slow.

Learn about parameter correlations during burn-in

The accepted 'l' proposals will generally cluster along the correlation path so every 2nd accepted 'l' proposal is appended to a correlated sample buffer (separate buffer for each tempering level).

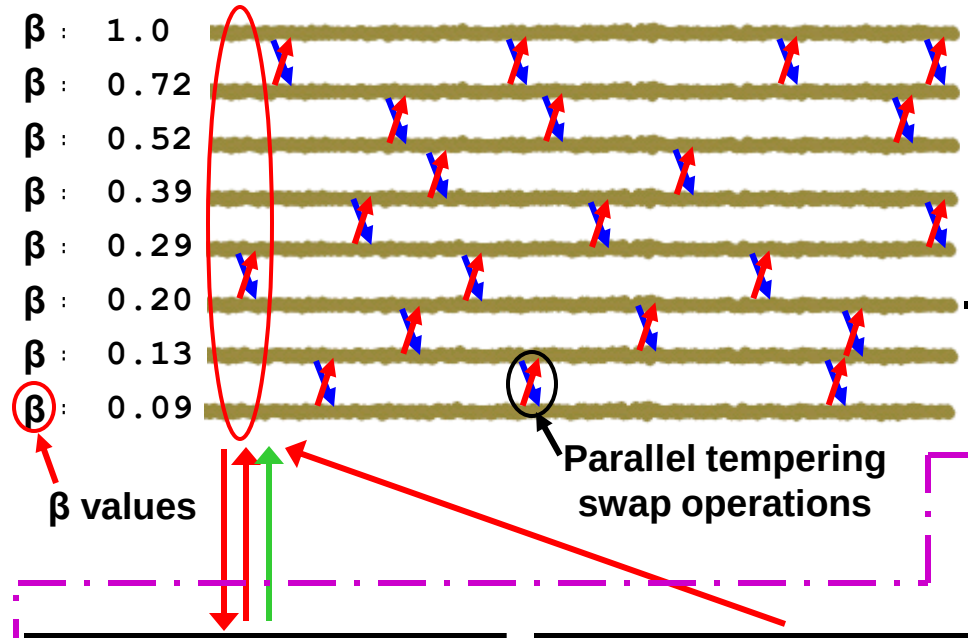
Only the 300 most recent additions to the buffer are retained.

A 'C' proposal is generated using the difference between a pair of randomly selected samples drawn from the correlated sample buffer (for that tempering level), after multiplication by a constant.

Value of constant is computed automatically by another control system module which ensures that the 'C' proposal acceptance rate is close to 25%.

Fusion MCMC

8 parallel tempering Metropolis chains



Automatic proposal scheme that learns about parameter correlations during burn-in (for each chain)

new parameter value

from proposal C

from proposal I

repeat

Add every 2nd to a buffer

latest 300 values

difference of random pairs

multiply by constant <1

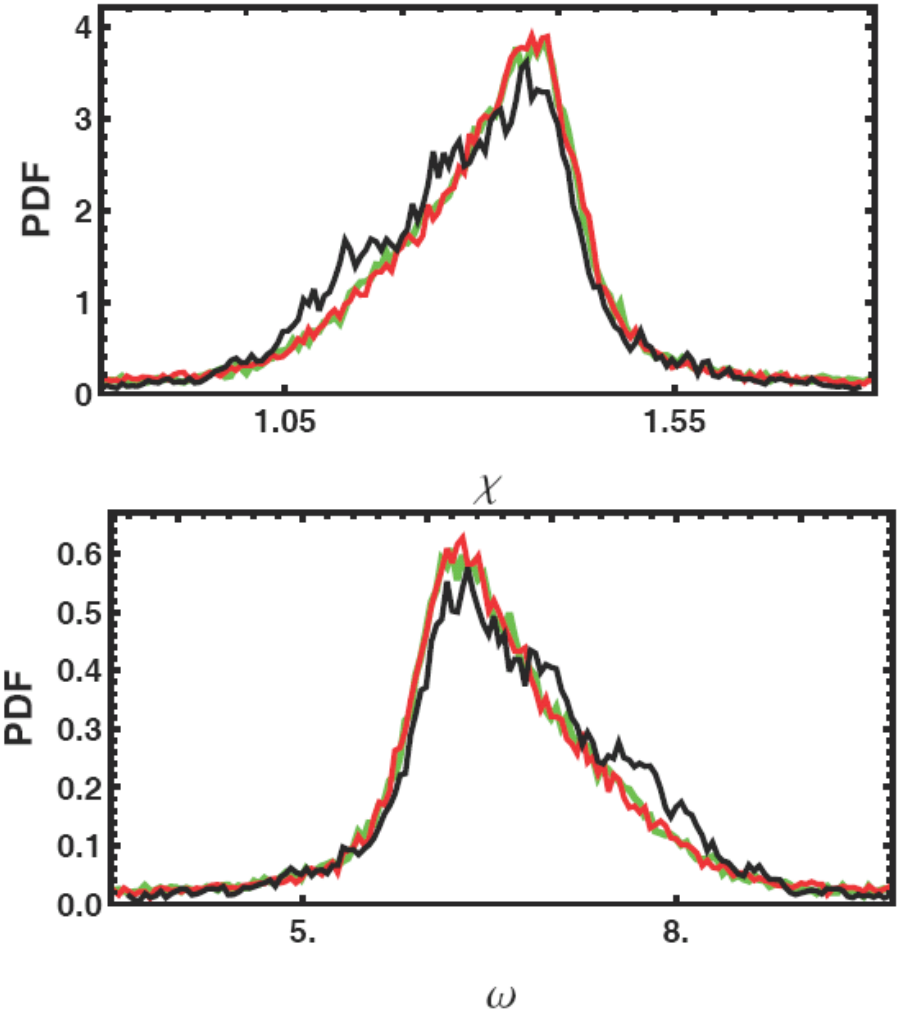
'I' proposals Independent Gaussian proposal scheme employed 50% of the time

'C' proposals Proposal distribution with built in parameter correlations used 50% of the time

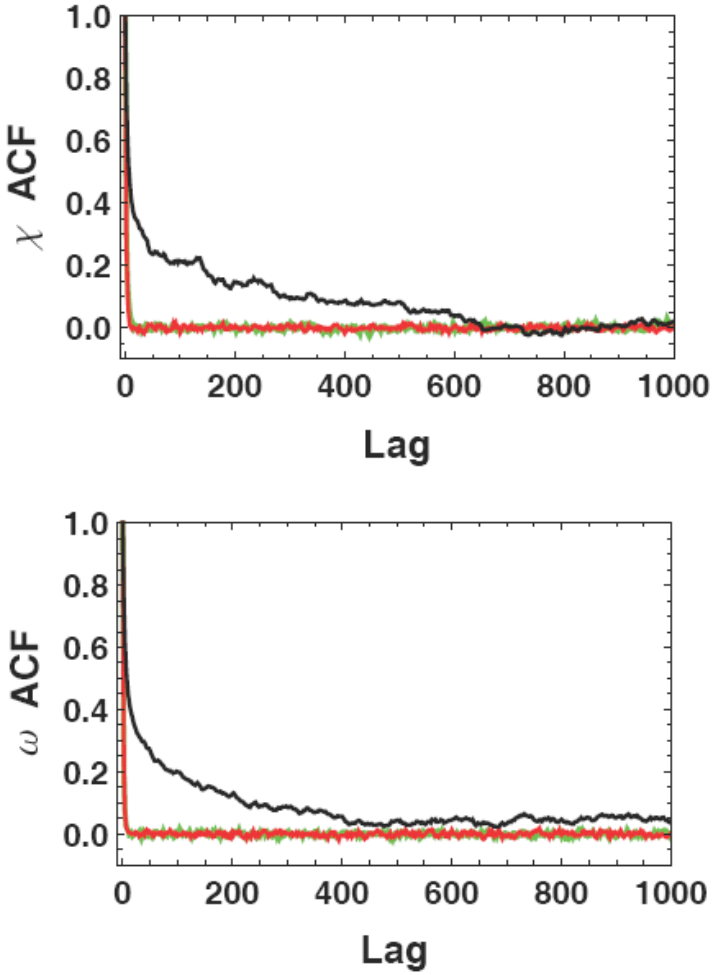
During burn-in control system adjusts constant so acceptance rate from C proposals = 25 %

MCMC adaptive control system

Testing ‘C’ proposal scheme



Autocorrelation function



Left panels show the MCMC marginal probability distributions for parameters χ and ω. Right panels show their MCMC autocorrelation functions.

Black trace = search in χ and ω using only ‘I’ proposals.

Red trace = search using both ‘I’ and ‘C’ proposals.

Green trace = ‘I’ search using transformed orthogonal coordinates.

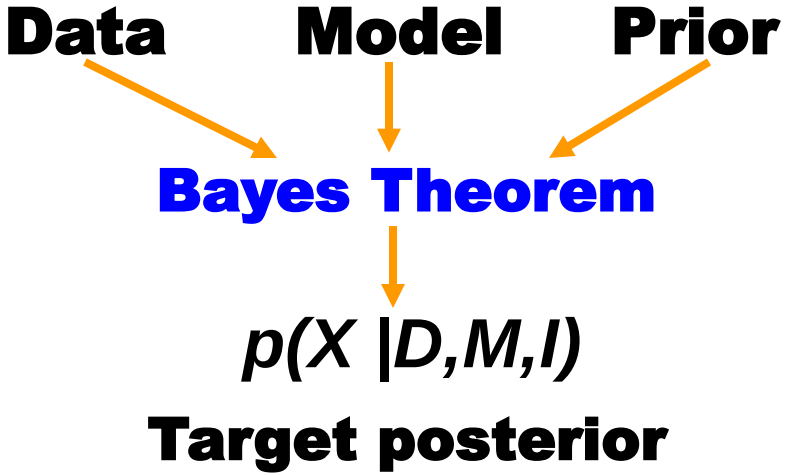
Summary of automatic ‘C’ proposal features

With very little computational overhead, the ‘C’ proposals provide the scale and direction for efficient jumps in a correlated parameter space with no additional chains.

The final proposal distribution is a random selection of ‘I’ and ‘C’ proposals. Each is employed 50% of the time.

The combination ensures that the whole parameter space can be reached and that the FMCMC chain is aperiodic.

The parallel tempering feature operates as before to avoid becoming trapped in a local probability maximum.



If you input a Kepler model the fusion MCMC becomes

A Kepler periodogram

Optimum for finding Kepler orbits and evaluating their probabilities.

Capable of simultaneously fitting multiple planet models.

A multi-planet Kepler periodogram

Multiple Planets complications: for a star being perturbed by multiple planets, there is no analytic expression for the exact radial velocity perturbation. However, in many cases the star's RV can be modeled well enough by the sum of multiple independent Keplerian orbits.

Model space considered

Symbol	Model	# of parameters
M_0	Constant velocity V + extra noise term s	2
M_1	V + elliptical orbit +extra noise term s	7
M_2	V + 2 elliptical orbits+extra noise term s	12
M_3	V + 3 elliptical orbits+extra noise term s	17
M_j	V + j elliptical orbits+extra noise term s	5j+2

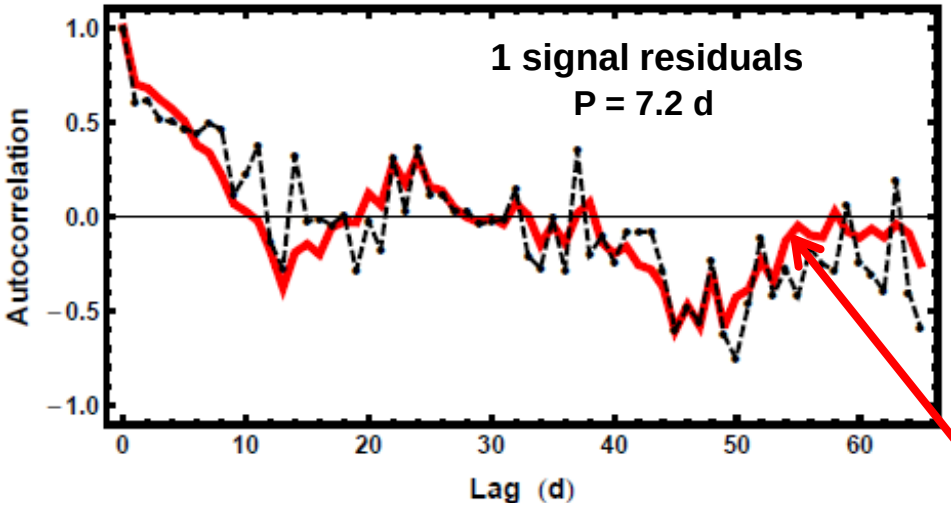
An extra noise term allows for an intrinsic stellar variability ("jitter") that we model as an additional source of uncorrelated Gaussian noise with variance s^2 and add to the measurement uncertainties in quadrature. s becomes an additional parameter to marginalize over.

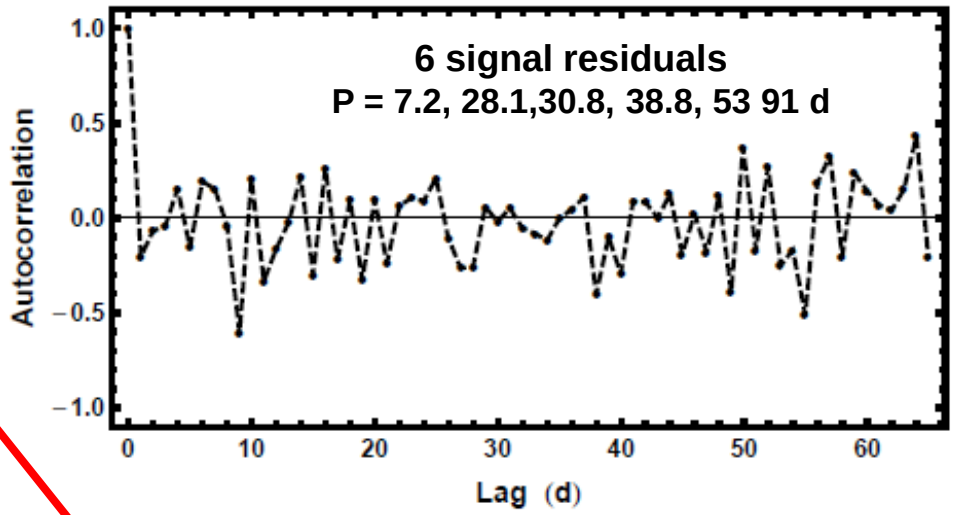
Note: some forms of stellar jitter (e.g., star spots) can produce Keplerian-like radial velocity variations.

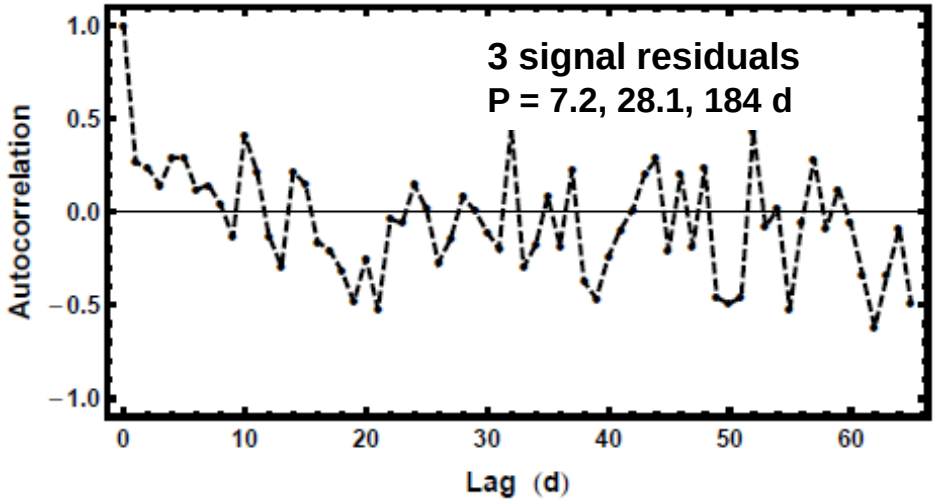
Autocorrelation function of residuals for Gliese 667C

$$\rho(j) = \frac{\sum_{\text{overlap}}[(x_i - \overline{x})\ (x_{i+j} - \overline{x})]}{\sqrt{\sum_{\text{overlap}}(x_i - \overline{x})^2} \times \sqrt{\sum_{\text{overlap}}(x_{i+j} - \overline{x})^2}}$$

where x_i is the i^{th} residual, j is the lag and x is the mean of the samples in the overlap region. Because the data are not uniformly sampled, for each lag all sample pairs that differed in time by this lag ± 0.1 d were utilized.







The solid red curve in the 1signal residuals is the average autocorrelation generated from 400 simulated data sets of a 5 signal model (28.1, 30.8, 38.8, 53.2, & 91 d periods) together with the quoted measurement errors.

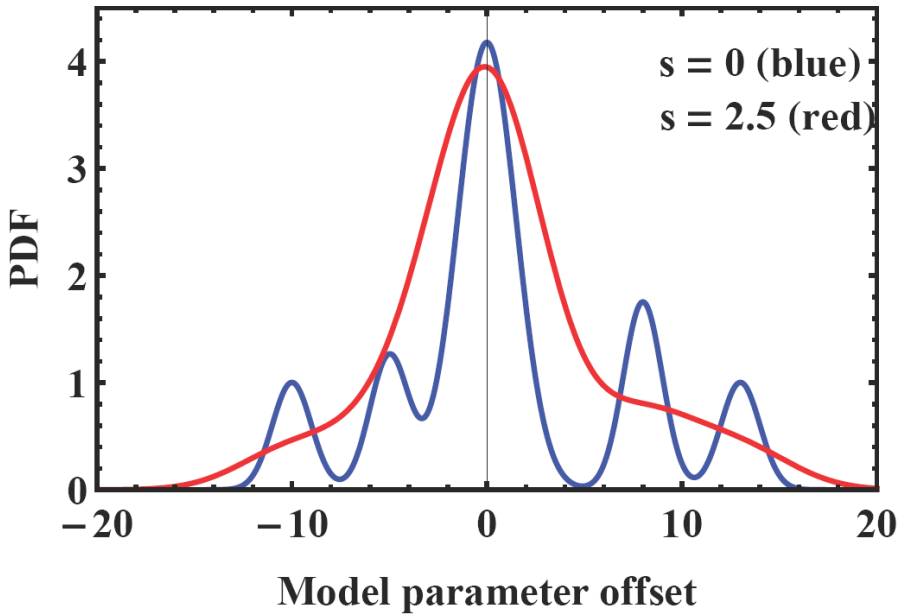
Annealing due to extra noise term, s

Inclusion of an extra noise term of unknown magnitude also gives rise to an annealing operation when the Markov chain is started far from the best-fit values.

If only known observational errors are included, the posterior probability distribution is often very "rough" with many local maxima throughout parameter space.

When *s* is included, Bayesian Markov chain automatically inflates *s* to include anything in the data that cannot be accounted for by the model with the current set of parameters and the known measurement errors.

This results in a smoothing out of the posterior surface and allows the Markov chain to explore the parameter space more quickly. The chain begins to decrease the value of *s* as it settles in near the best-fit parameters. This behavior is similar to simulated annealing, but does not require choosing a cooling scheme.



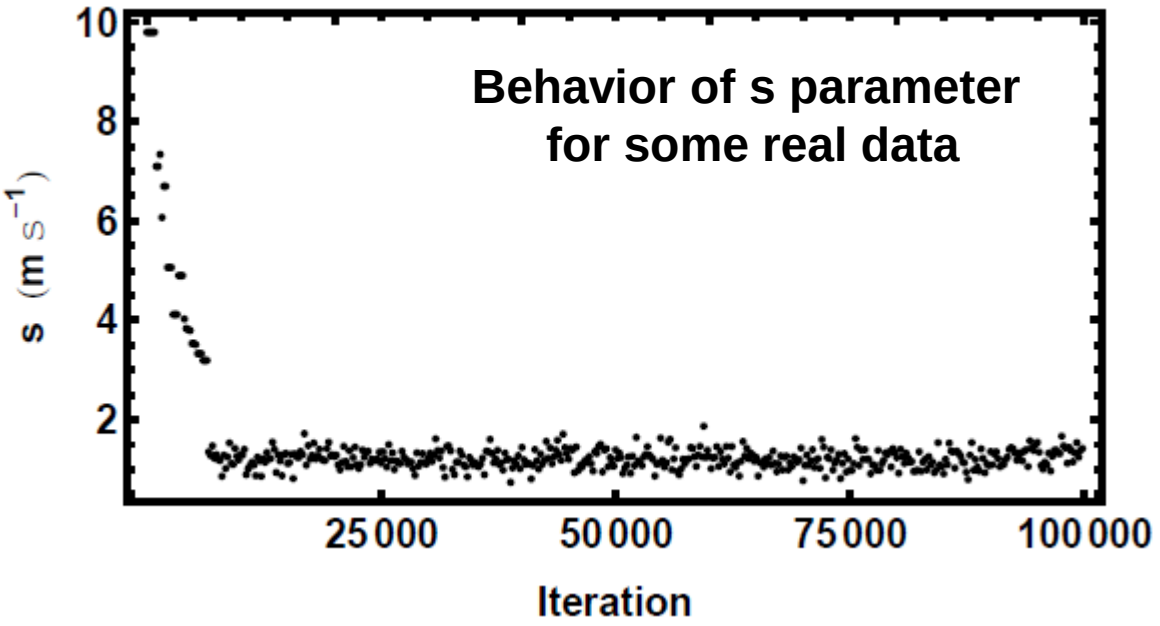
Annealing due to extra noise term, s

Inclusion of an extra noise term of unknown magnitude also gives rise to an annealing operation when the Markov chain is started far from the best-fit values.

If only known observational errors are included, the posterior probability distribution is often very "rough" with many local maxima throughout parameter space.

When *s* is included, Bayesian Markov chain automatically inflates *s* to include anything in the data that cannot be accounted for by the model with the current set of parameters and the known measurement errors.

This results in a smoothing out of the posterior surface and allows the Markov chain to explore the parameter space more quickly. The chain begins to decrease the value of *s* as it settles in near the best-fit parameters. This behavior is similar to simulated annealing, but does not require choosing a cooling scheme.



Example 1: HD208487

HD 208487

History

1) 2005, report of a P = 130d companion

C. G. Tinney et al. ApJ, 623, 1171

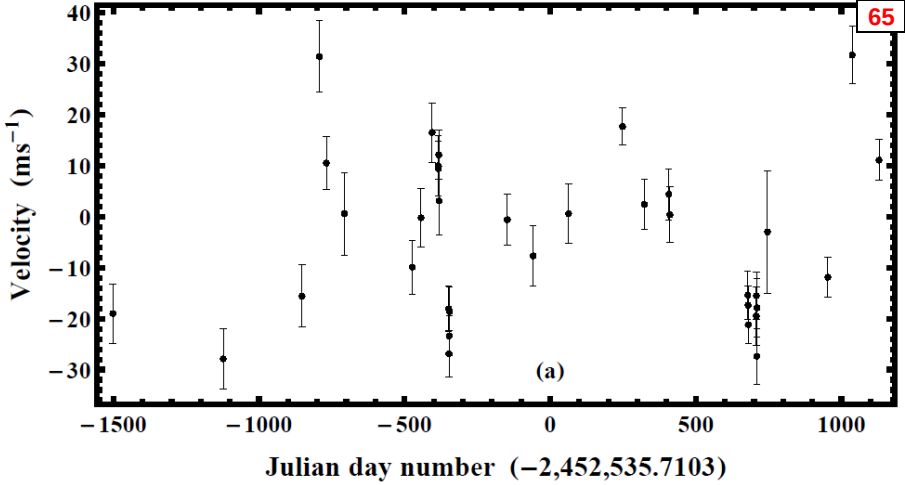
2) 02/2007, report of 2nd companion
P = 909± 90d

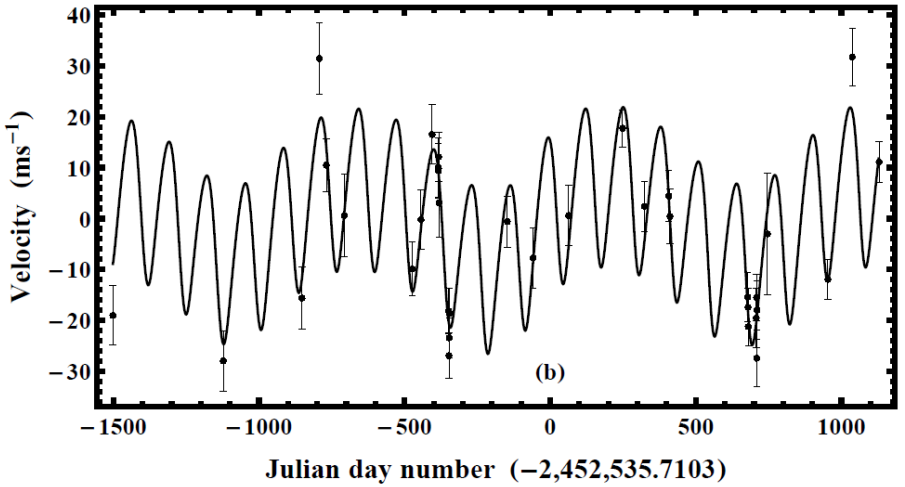
P. C. Gregory, MNRAS, 403, 731

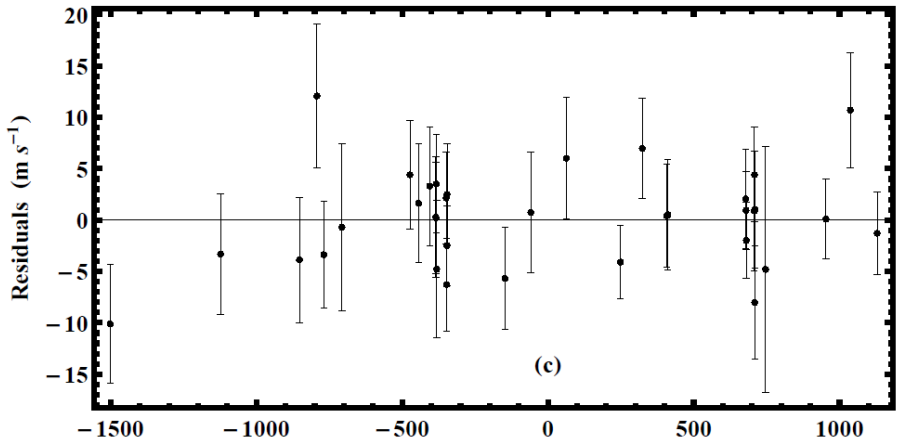
3) 03/2007, report of a 2nd companion,
P ~ 27 or 1000 days

J. T. Wright et al., ApJ, 657, 533

Will consider which is the real signal when I discuss aliases.







Parameter	prior	Lower bound	Upper bound
Orbital frequency	$p(\ln f_1, \ln f_2, \cdots \ln f_n \mid M_n, I) = \frac{n!}{[\ln(f_H/f_L)]^n}$ (n =number of planets)	1/0.5 d	1/1000 yr
Velocity K_i (m s^{-1})	Modified scale invariant[a] $\frac{(K+K_0)^{-1}}{\ln\left[1+\frac{K_{\max}}{K_0}\left(\frac{P_{\min}}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}\right]}$	0 ($K_0 = 1$)	$K_{\max}\left(\frac{P_{\min}}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}$ $K_{\max} = 2129$ K_{max} corresponds to a max. planet-star mass ratio = 0.01
V (m s^{-1})	Uniform	$-K_{\max}$	$K_{\max}$
e Eccentricity	$3.1(1-e)^{2.1}$	0	0.99
χ orbit fraction	Uniform	0	1
ω Longitude of periastron	Uniform	0	2π
s Extra noise (m s^{-1})	$\frac{(s+s_0)^{-1}}{\ln\left(1+\frac{s_{\max}}{s_0}\right)}$	0 (s_0 = 1 to 10)	$K_{\max}$

[a] Since the prior lower limits for K and s include zero, we used a modified scale invariant prior of the form

$$p(X|M,I) = \frac{1}{X+X_0}\,\frac{1}{\ln\left(1+\frac{X_{\max}}{X_0}\right)}$$

For $X \ll X_0$, $p(X|M,I)$ behaves like a uniform prior and for $X \gg X_0$ it behaves like a scale invariant prior. The $\ln\left(1+\frac{X_{\max}}{X_0}\right)$ term in the denominator ensures that the prior is normalized in the interval 0 to $X_{\max}$.

Search in frequency instead of period

The width of a spectral peak in a probability density plot, which reflects the accuracy of the frequency estimate, is determined by the duration of the data, the signal-to-noise (S/N) ratio and the number of data points.

More precisely, for a sinusoidal signal model, the standard deviation of the spectral peak, δf, for a S/N > 1, is given by

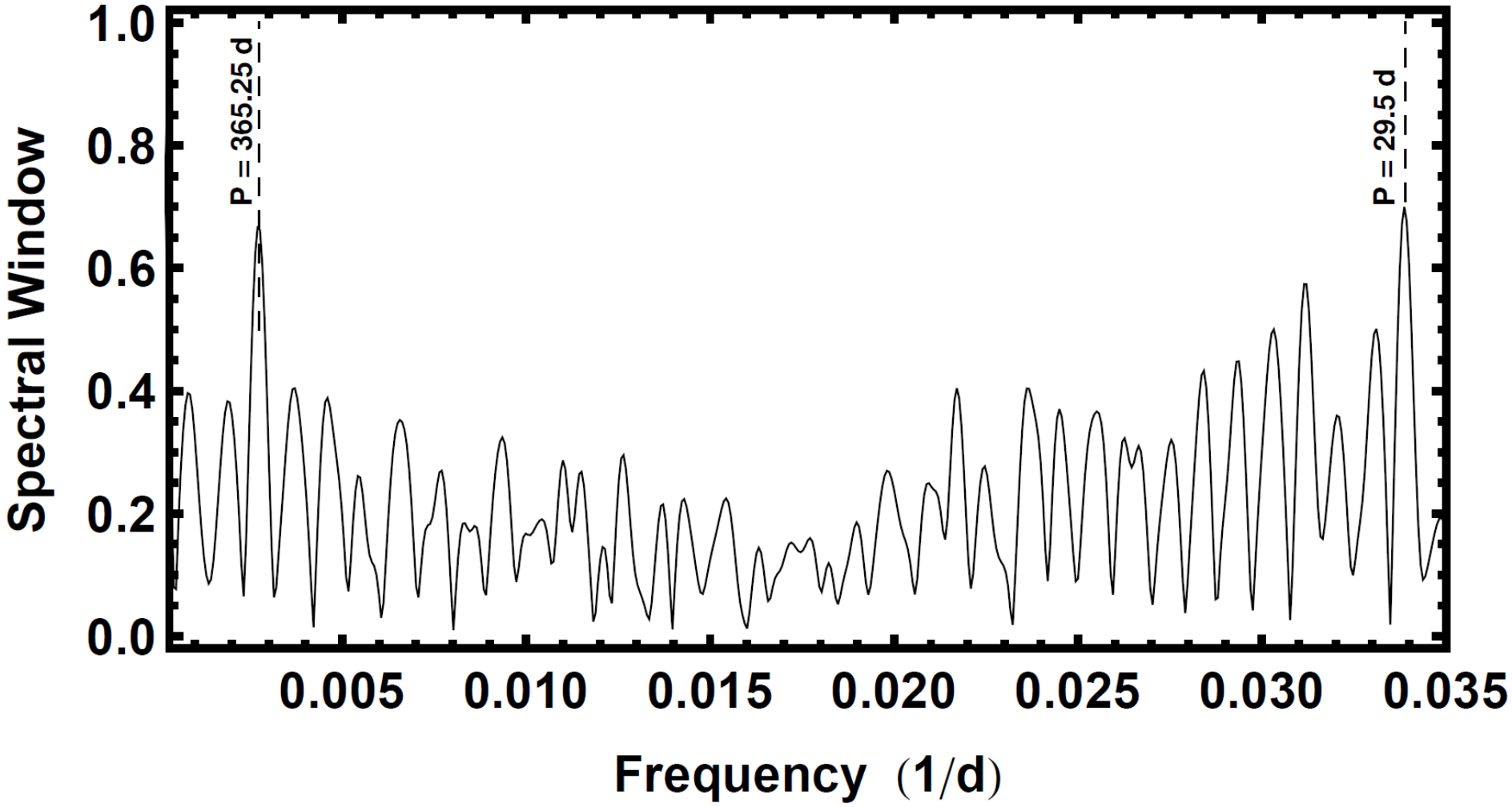
$$\delta f \approx \left(1.6 \frac{\mathrm{S}}{\mathrm{N}} \mathcal{T} \sqrt{N}\right)^{-1} \quad \mathrm{Hz}$$

where $\mathcal{T}$ = the data duration in s, and *N* = the # of data points in $\mathcal{T}$.

The thing to notice is that the width of any peak is independent of the frequency of the peak. Thus the same frequency proposal distribution will be efficient for all frequency peaks. This is not the case for a period search where the width of a spectral peak is $\propto P^2$.

Not only is the width of the peak independent of *f*, but the spacing of potential peaks is constant in frequency (roughly $f \sim 1/\mathcal{T}$), which is a another motivation for searching in frequency space.

Search in frequency instead of period



A portion of the spectral window function of the radial velocity data for HD 208487 demonstrating the uniform spacing of peaks in frequency. The 29.5 d peak corresponds to the synodic month.

Multiple orbital frequency prior

For a single frequency (period) model we use a scale invariant prior which can be written in two equivalent forms.

$$p(\ln f|M_1, I)\, d\ln f = \frac{d\ln f}{\ln(f_H/f_L)}$$

$$p(f|M, I)\, df = \frac{df}{f\ \ln(f_H/f_L)}$$

What form of frequency prior should we use for a multiple planet model? If we constrain the frequencies in an n planet search such that

$$(f_L \le f_1 \le f_2 \cdots \le f_n \le f_H).$$

From the product rule of probability theory and the above frequency constraints we can show that the multiple frequency prior is

$$p(\ln f_1, \ln f_2, \cdots \ln f_n|M_n, I) = \frac{n!}{[\ln(f_H/f_L)]^n}$$

Bretthorst (2003) derived a similar result involving n! in the numerator in connection of uniform frequency priors.

Orbital frequency search strategy

Two different approaches to searching in the frequency parameters were tried.

(a) an upper bound on $f_1 \leq f_2$ was utilized to maintain the identity of the two frequencies.

(b) both f_1 and f_2 were allowed to roam over the entire frequency range and the parameters re-labeled afterwards. In this second approach nothing constrains f_1 to always be below f_2 so that degenerate parameter peaks can occur.

Approach (b) was found to be more successful because in repeated blind period searches it regularly converged on the highest posterior peak, in spite of the huge period search range.

Approach (a) proved to be unsuccessful in finding the highest peak in many trials and in those cases where it did find the peak it required more iterations. Restricting $f_1 \leq f_2$ introduces an additional hurdle that appears to slow the MCMC period search

return

Orbital frequency search strategy

Provide the parameters are re-labeled after the MCMC run, such that the parameters associated with the lower frequency are always identified with planet 1 and vice versa, the two cases are equivalent and both require the same frequency prior.

return

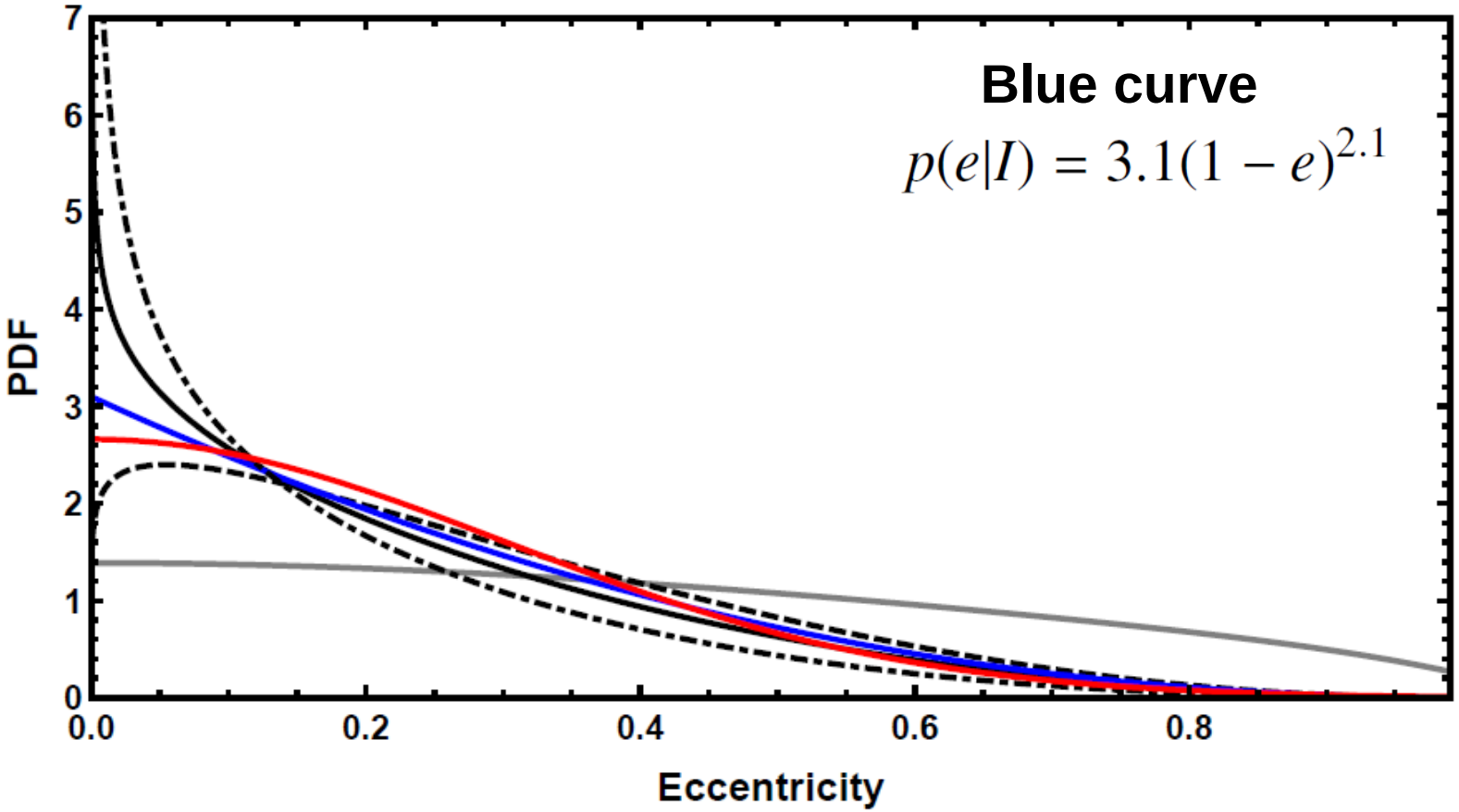
Parameter	prior	Lower bound	Upper bound
Orbital frequency	$p(\ln f_1, \ln f_2, \cdots \ln f_n \mid M_n, I) = \frac{n!}{[\ln(f_H/f_L)]^n}$ (n =number of planets)	1/0.5 d	1/1000 yr
Velocity K_i (m s^{-1})	Modified scale invariant[a] $\frac{(K+K_0)^{-1}}{\ln\left[1+\frac{K_{\max}}{K_0}\left(\frac{P_{\min}}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}\right]}$	0 ($K_0 = 1$)	$K_{\max}\left(\frac{P_{\min}}{P}\right)^{1/3}\frac{1}{\sqrt{1-e^2}}$ $K_{\max} = 2129$ K_{max} corresponds to a max. planet-star mass ratio = 0.01
V (m s^{-1})	Uniform	$-K_{\max}$	$K_{\max}$
e Eccentricity	$3.1(1-e)^{2.1}$	0	0.99
χ orbit fraction	Uniform	0	1
ω Longitude of periastron	Uniform	0	2π
s Extra noise (m s^{-1})	$\frac{(s+s_0)^{-1}}{\ln\left(1+\frac{s_{\max}}{s_0}\right)}$	0 (s_0 = 1 to 10)	$K_{\max}$

[a] Since the prior lower limits for K and s include zero, we used a modified scale invariant prior of the form

$$p(X|M,I) = \frac{1}{X+X_0}\,\frac{1}{\ln\left(1+\frac{X_{\max}}{X_0}\right)}$$

For $X \ll X_0$, $p(X|M,I)$ behaves like a uniform prior and for $X \gg X_0$ it behaves like a scale invariant prior. The $\ln\left(1+\frac{X_{\max}}{X_0}\right)$ term in the denominator ensures that the prior is normalized in the interval 0 to $X_{\max}$.

Eccentricity priors



Solid black line is best fit Beta distribution (Kipping 2013) to the eccentricity data of 396 high S/N exoplanets. The dashed and dot-dashed black lines are Kipping's Beta distribution fits to the subsets with periods > 382.3 d (median) and < 382.3 d, respectively. The red line is the Gaussian eccentricity prior adopted by Tuomi et al. (2012). The gray line is my earlier prior which attempted a modest correction for noise induced eccentricity bias. The blue line is prior employed in this work.

Prior for K

K_{max} corresponds to the velocity of a planet with a mass = 0.01 $M_\odot$ in a circular orbit with our shortest period of one day.

Since K_i depends on two of the other parameters, P_i , e_i , according to

$$K = \frac{m \sin i}{M_*} \left(\frac{2\pi G M_*}{P} \right)^{1/3} \left(1 + \frac{m}{M_*} \right)^{-2/3} \frac{1}{\sqrt{1 - e^2}},$$

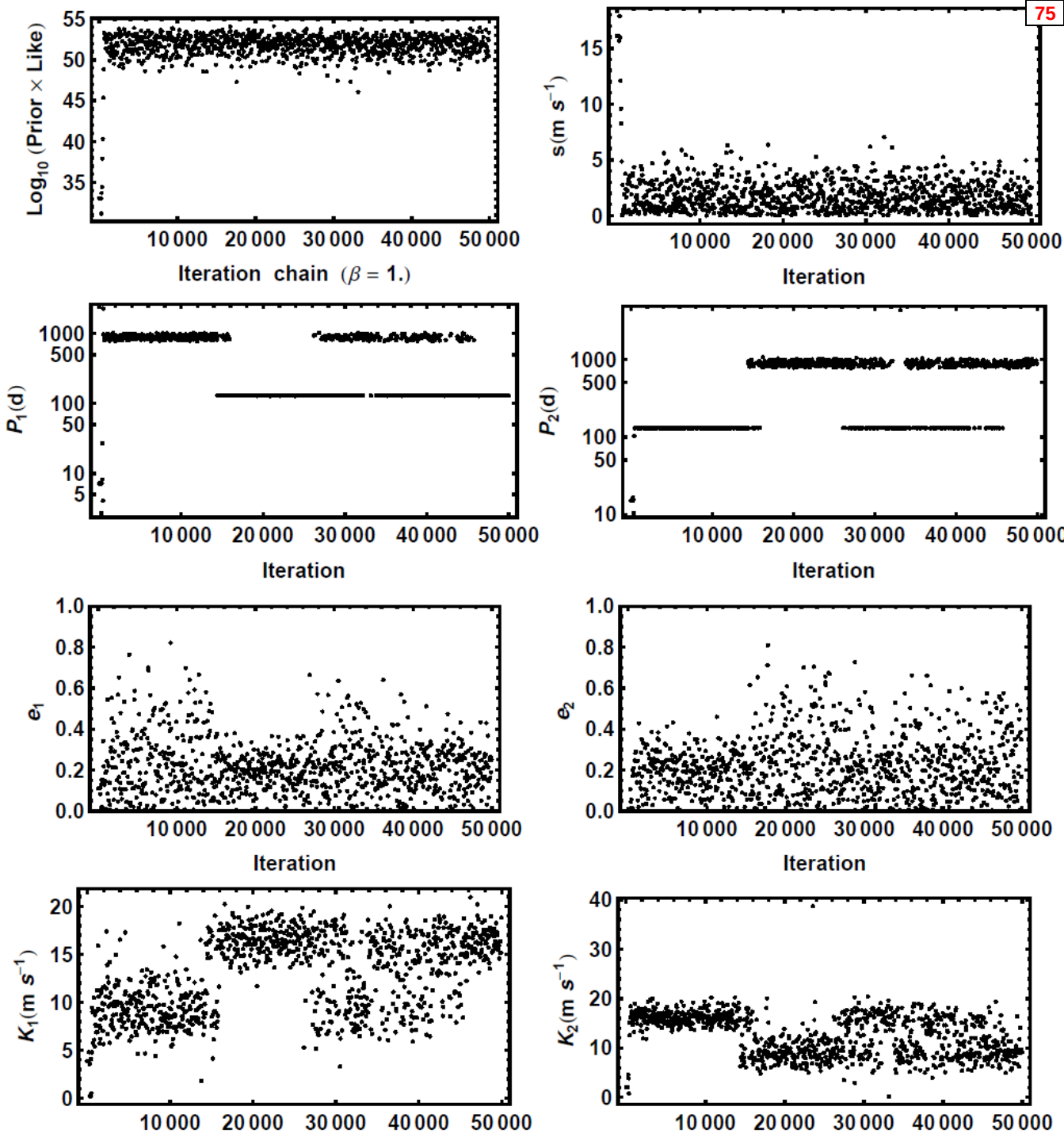
I used an upper bound for K_i given by

$$K_i = K_{\max} \left(\frac{P_{\min}}{P_i} \right)^{1/3} \frac{1}{\sqrt{1 - e_i^2}}$$

which allows the upper limit on K_i to depend on the proposed period and eccentricity. In my implementation of Metropolis, I make new proposals of all parameters at each iteration.

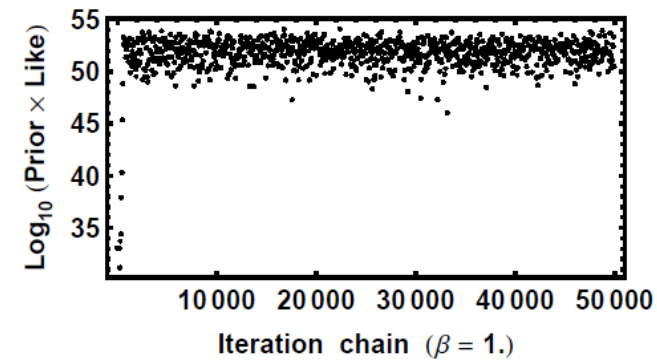
HD 208487

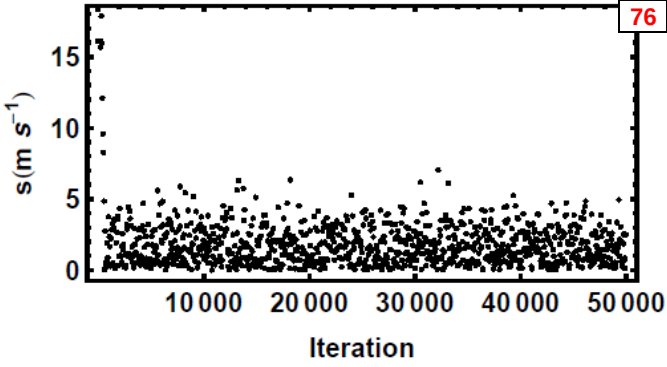
FMCMC results before relabeling parameters

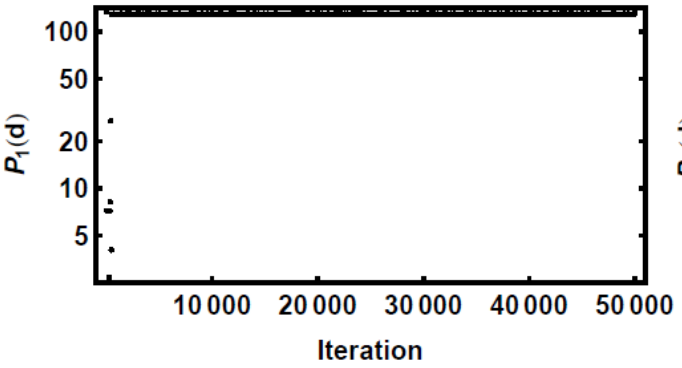


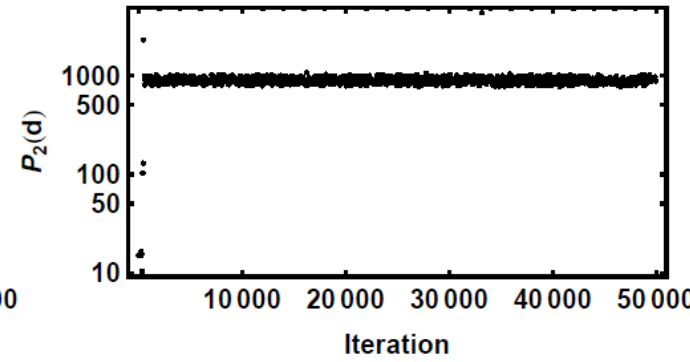
HD 208487

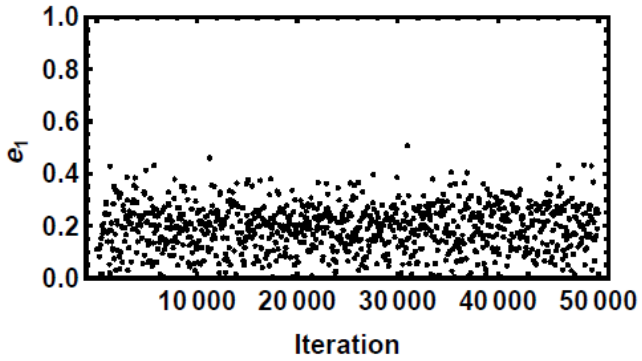
FMCMC results after relabeling parameters

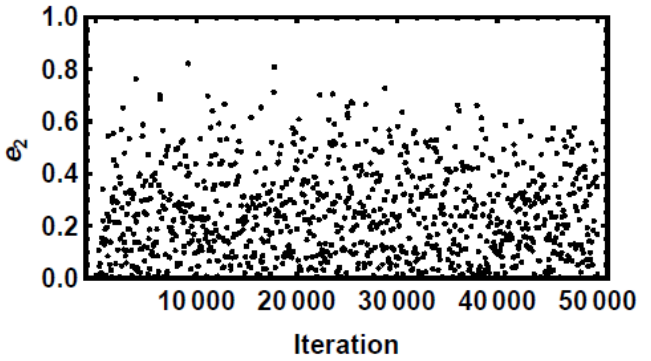


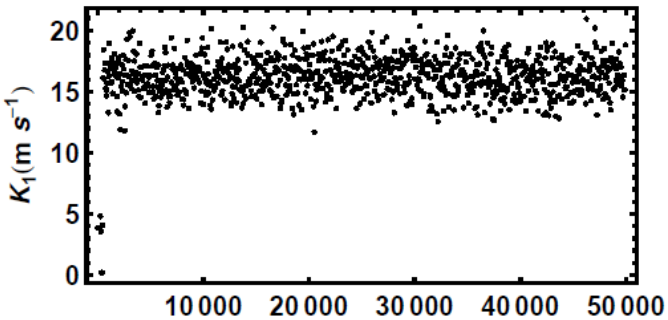


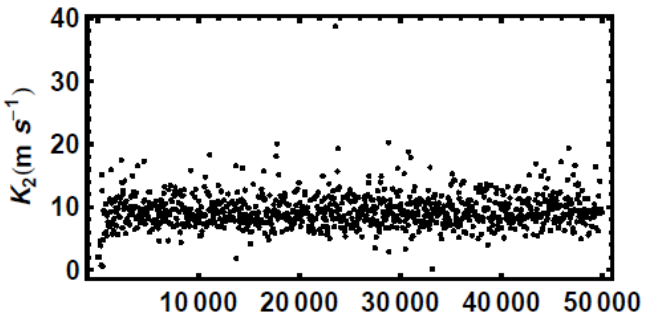






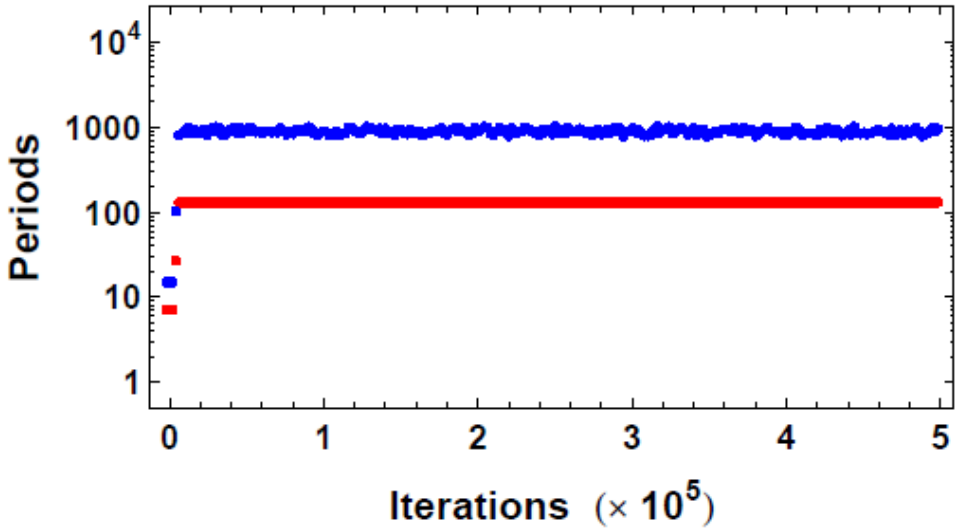


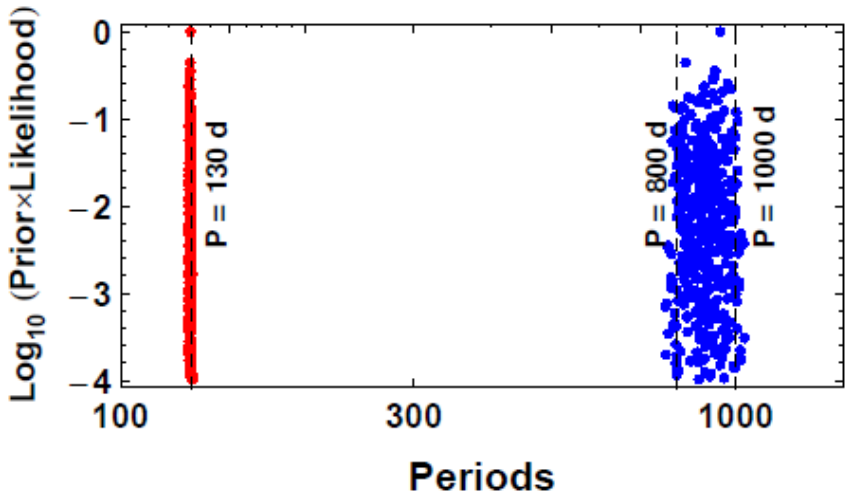


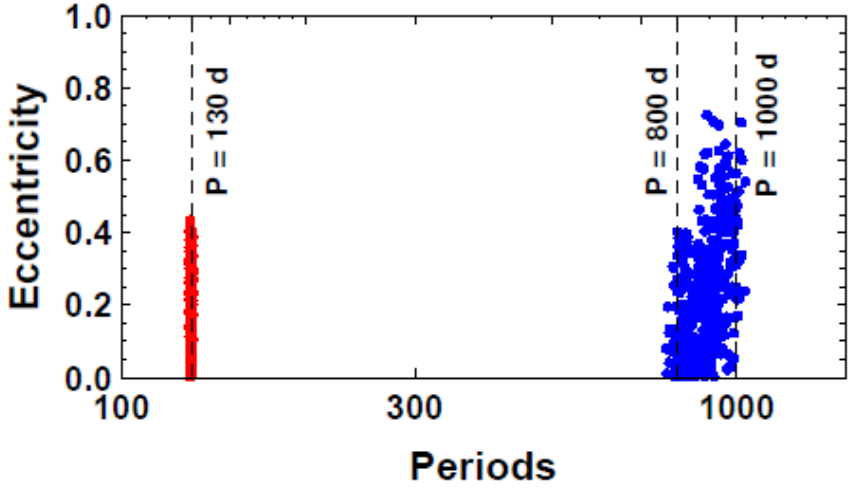


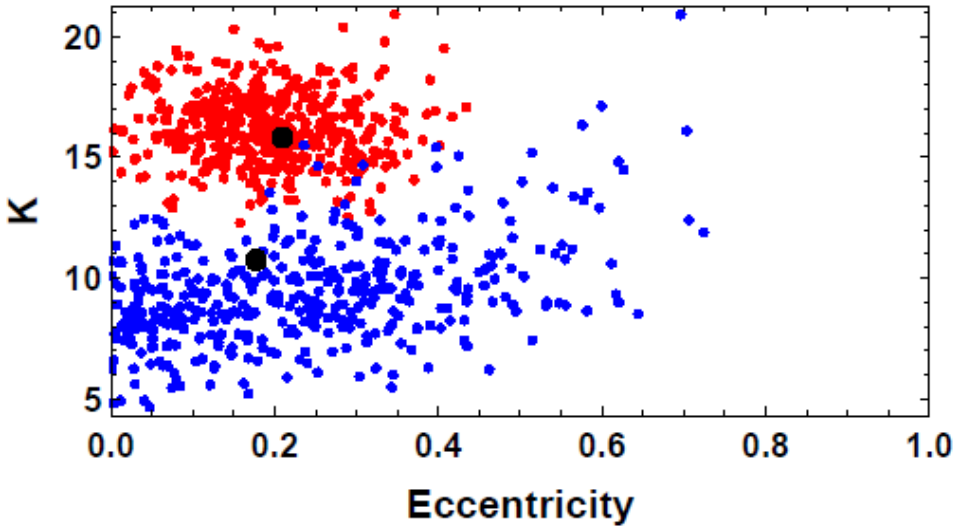
HD 208487 periodogram plots (2 planet model)

$p(f|M,I) \propto 1/f$

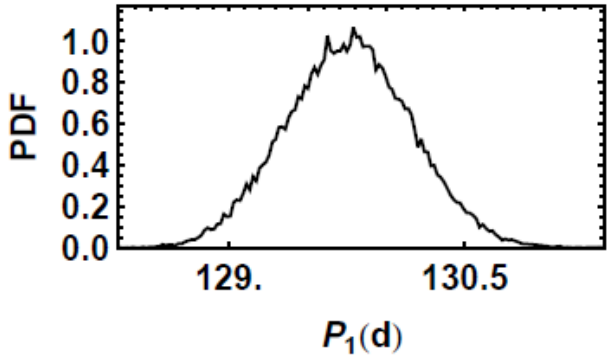


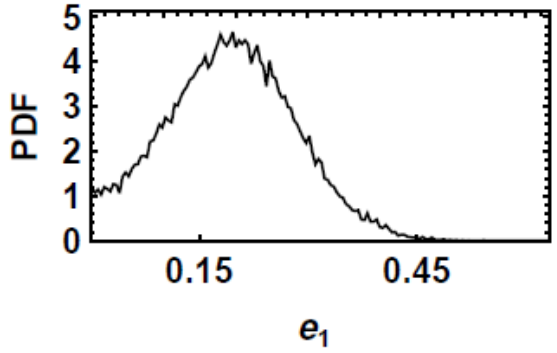


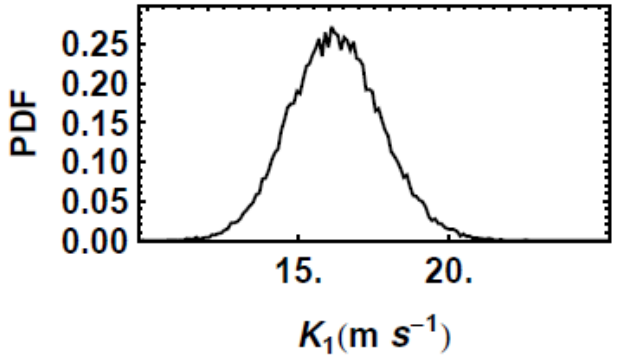


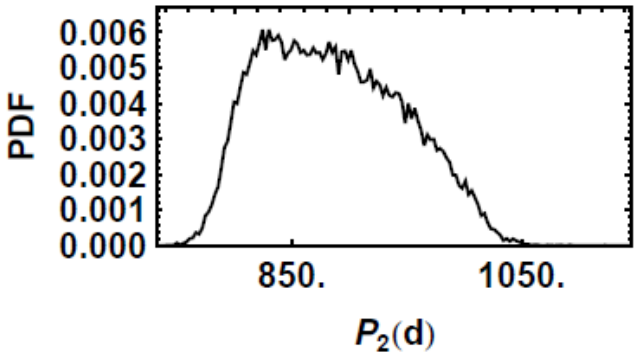


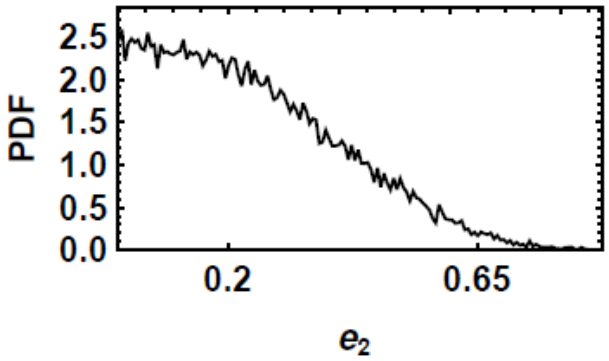
HD 208487 marginal distributions (2 planet model)

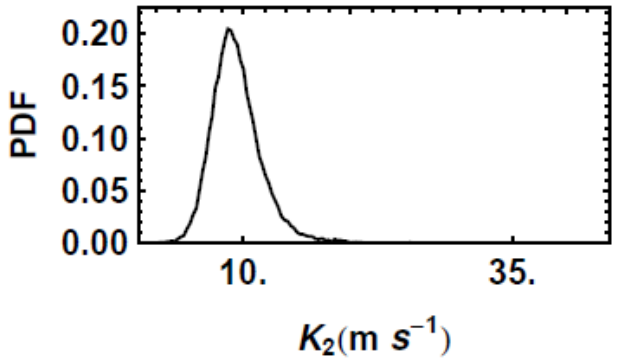


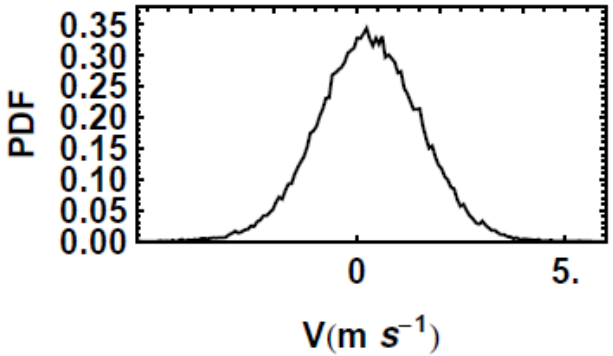


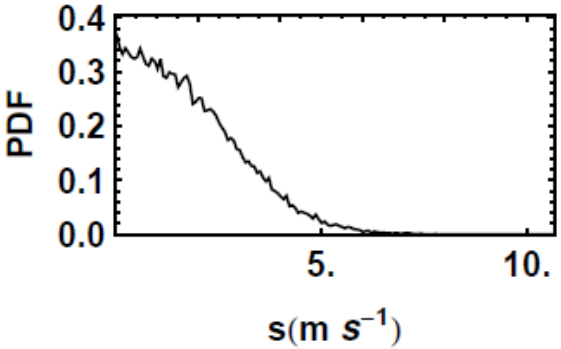






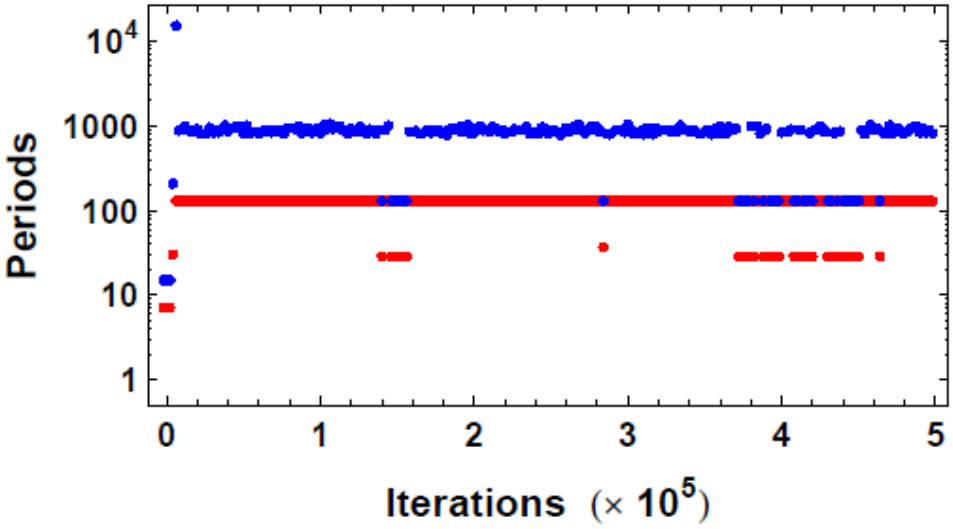


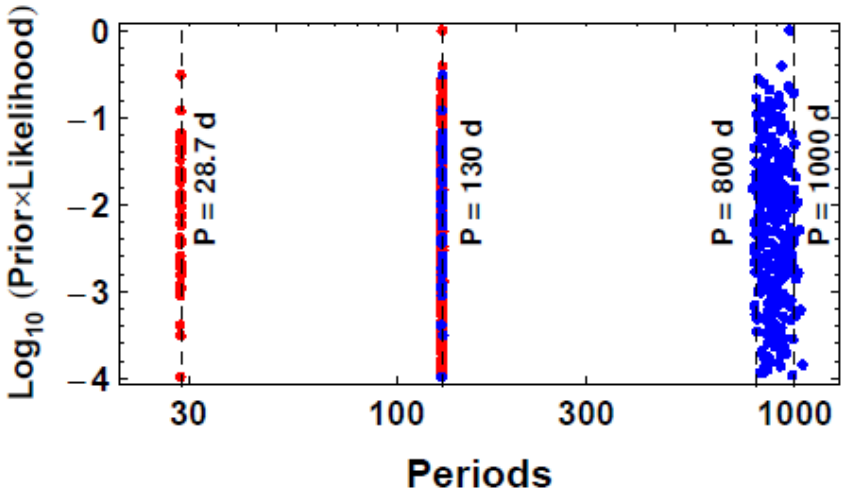


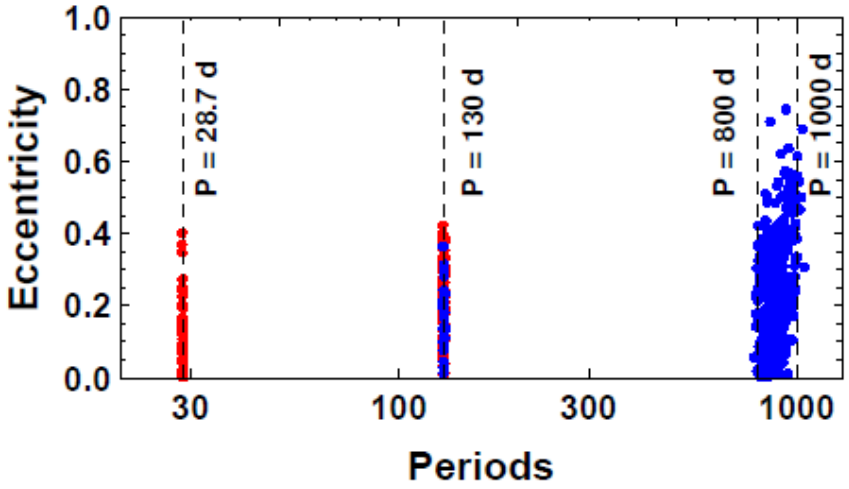


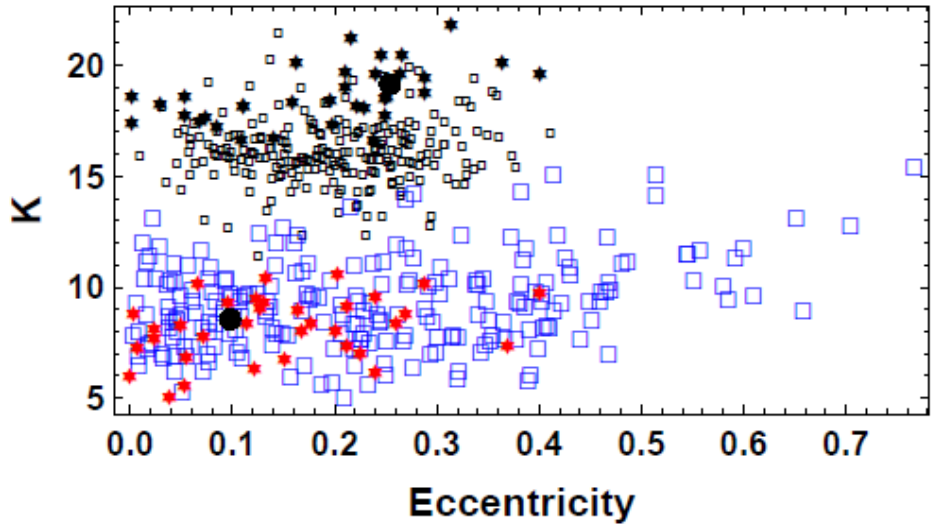
HD 208487 periodogram plots (2 planet model)

$p(f|M,I) \propto 1/\sqrt{f}$









Dynamical stability

Work in the 1970s and 80s showed that the motions of a system of a star with two planets (not in a low-order mean motion resonance) would be bounded in some situations. Two dominant definitions of stability emerged
(Barnes and Greenberg 2006)

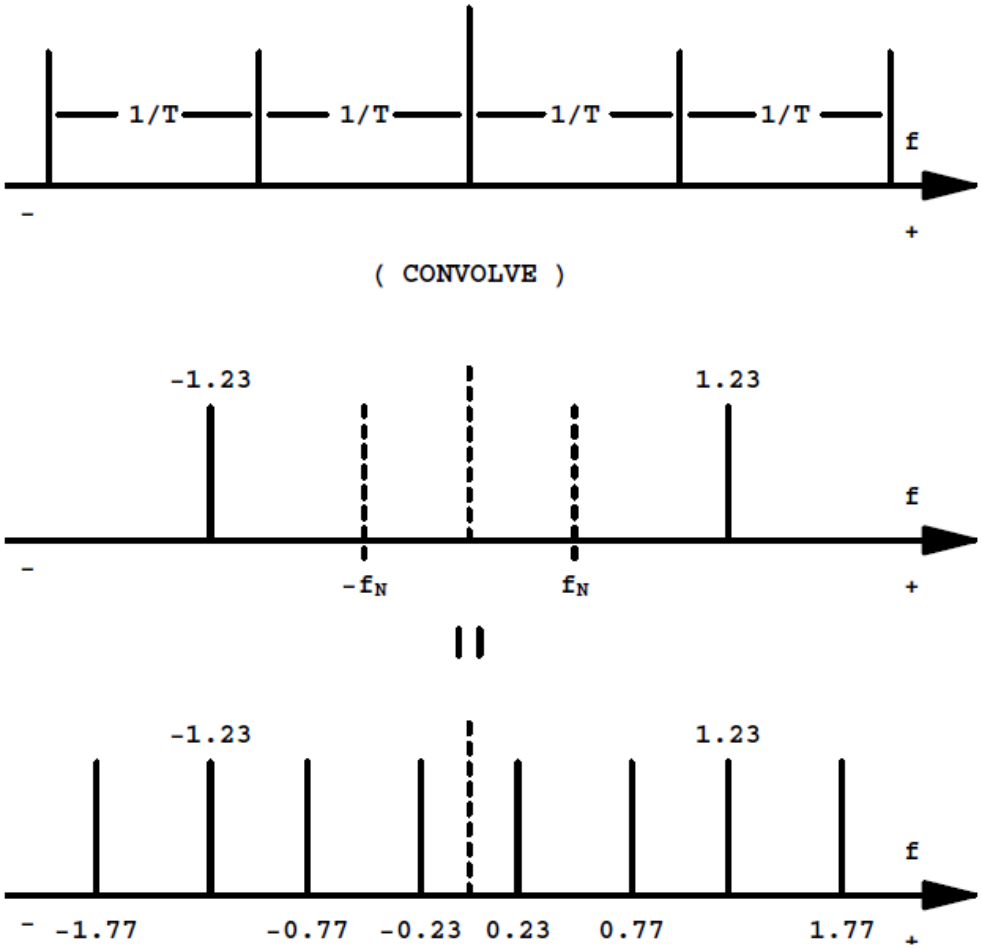
a) Hill stability the ordering of the two planets in terms of distance from the central star is conserved.

b) Lagrange stability includes Hill stability plus the planets remain bound to the star and the semimajor axis and eccentricity remain bounded.

Result for HD 208487: both two period solutions are Lagrange stable.

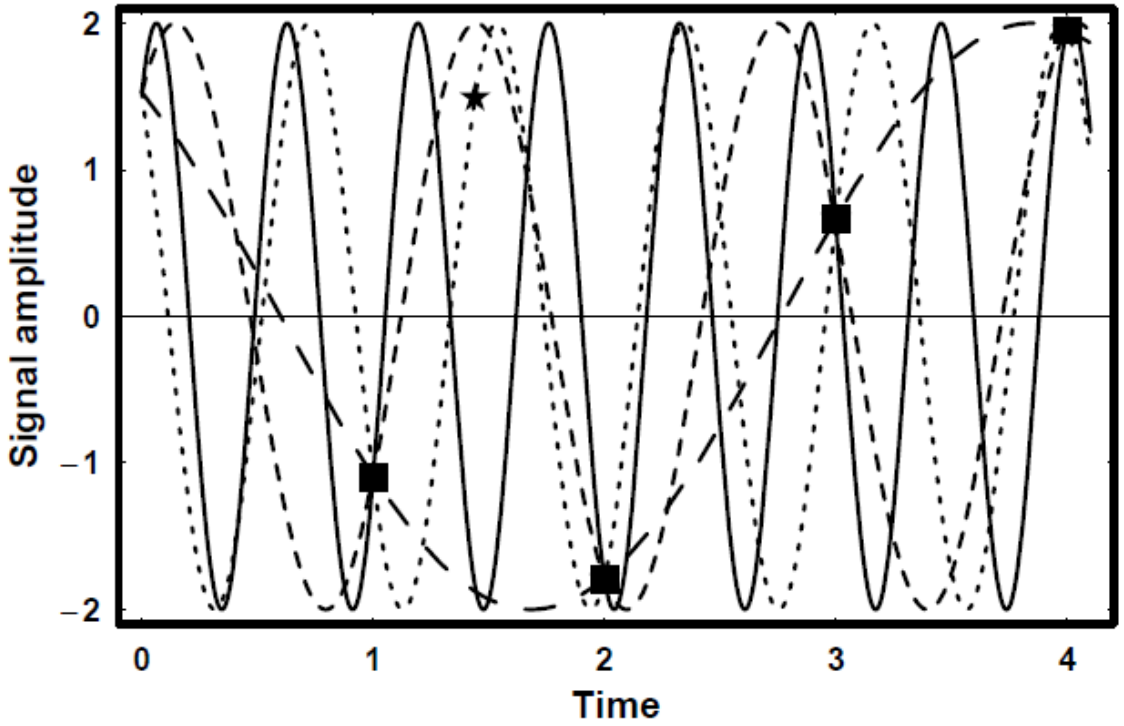
Aliases

Aliases



How aliasing arises. Uniform sampling at an interval T in the time domain corresponds to convolution in the frequency domain. The upper panel shows the Fourier Transform (FT) of the sampling. The middle panel shows the FT of the signal together with the Nyquist frequency. The lower panel shows the resulting convolution. There are 3 aliased signals at $f = 0.23$, $f = 0.77$, and $f = 1.77$, only one of which, at $f = 0.23$, is below the Nyquist frequency.

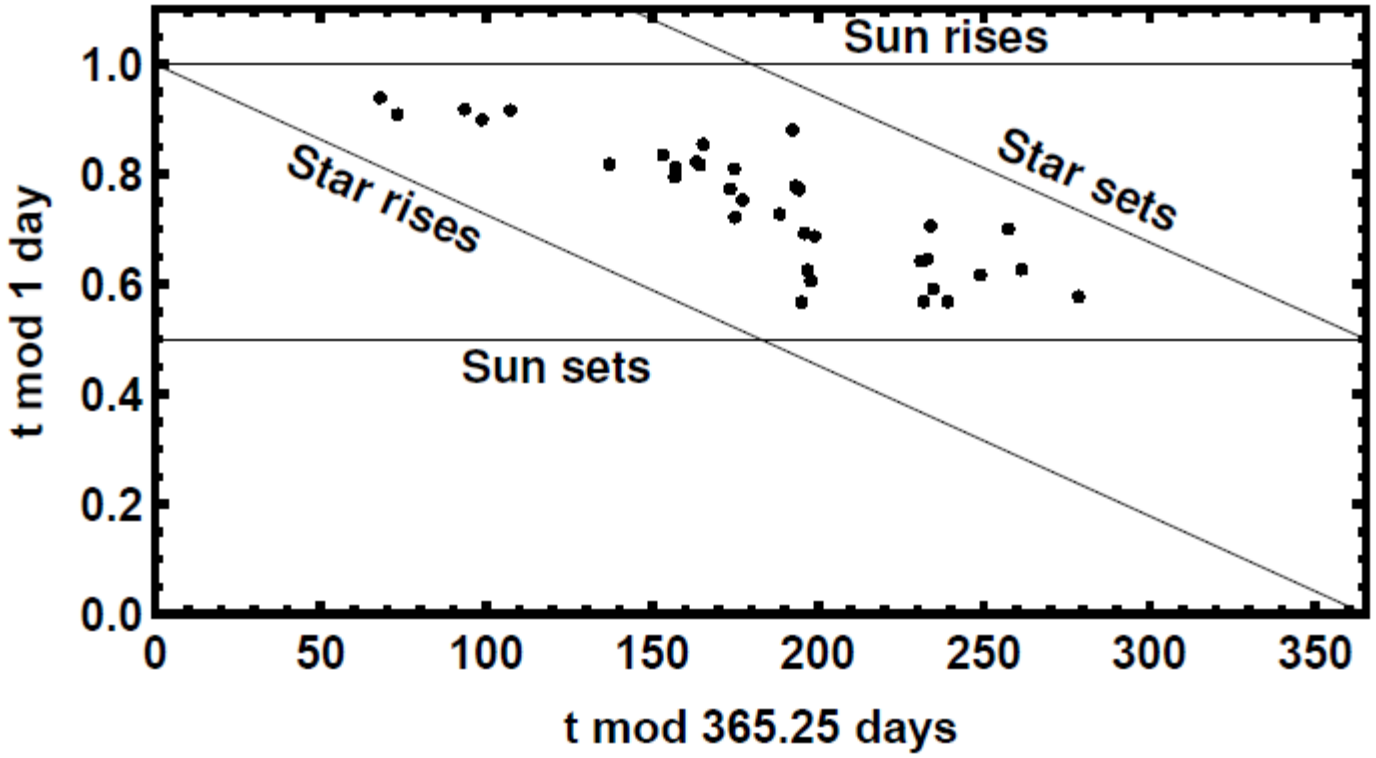
Aliases



An illustration of how four different frequencies can all pass through the same set of four uniformly sampled data points (boxes) but only one passes through all the points when one sample is replaced by a non-uniform sample (star).

There is effectively no Nyquist frequency for unevenly sampled data. Since RV sampling is very non-uniform, do we need to worry about Aliases?

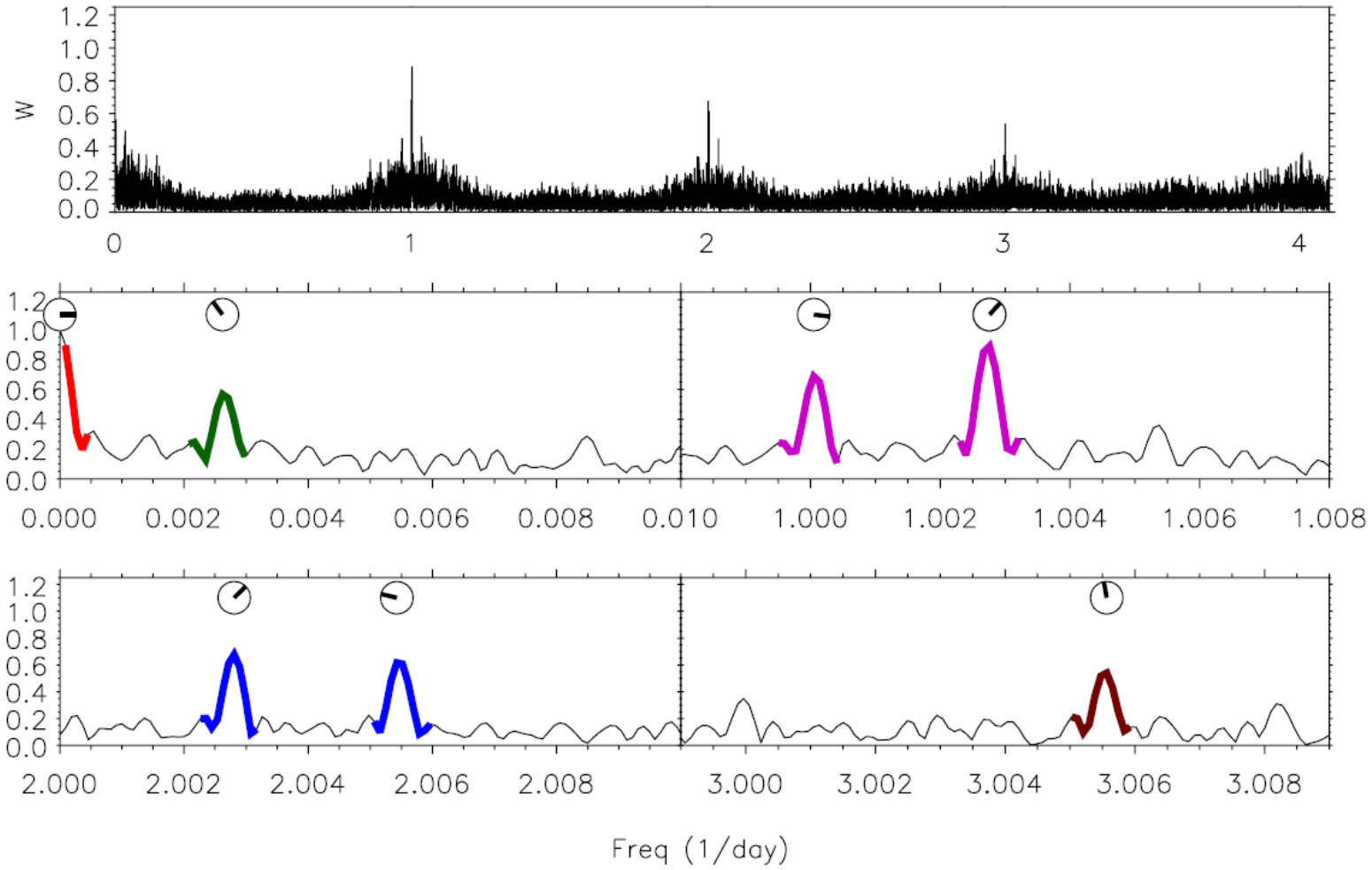
Aliases



Rebekah I. Dawson and Daniel C. Fabrycky (ApJ, 722:937, 2010) demonstrated how ground based astronomy has inherent periodicities in the sampling governed by when we can observe that give rise to significant aliases. Above is an example for the HD 208487 data.

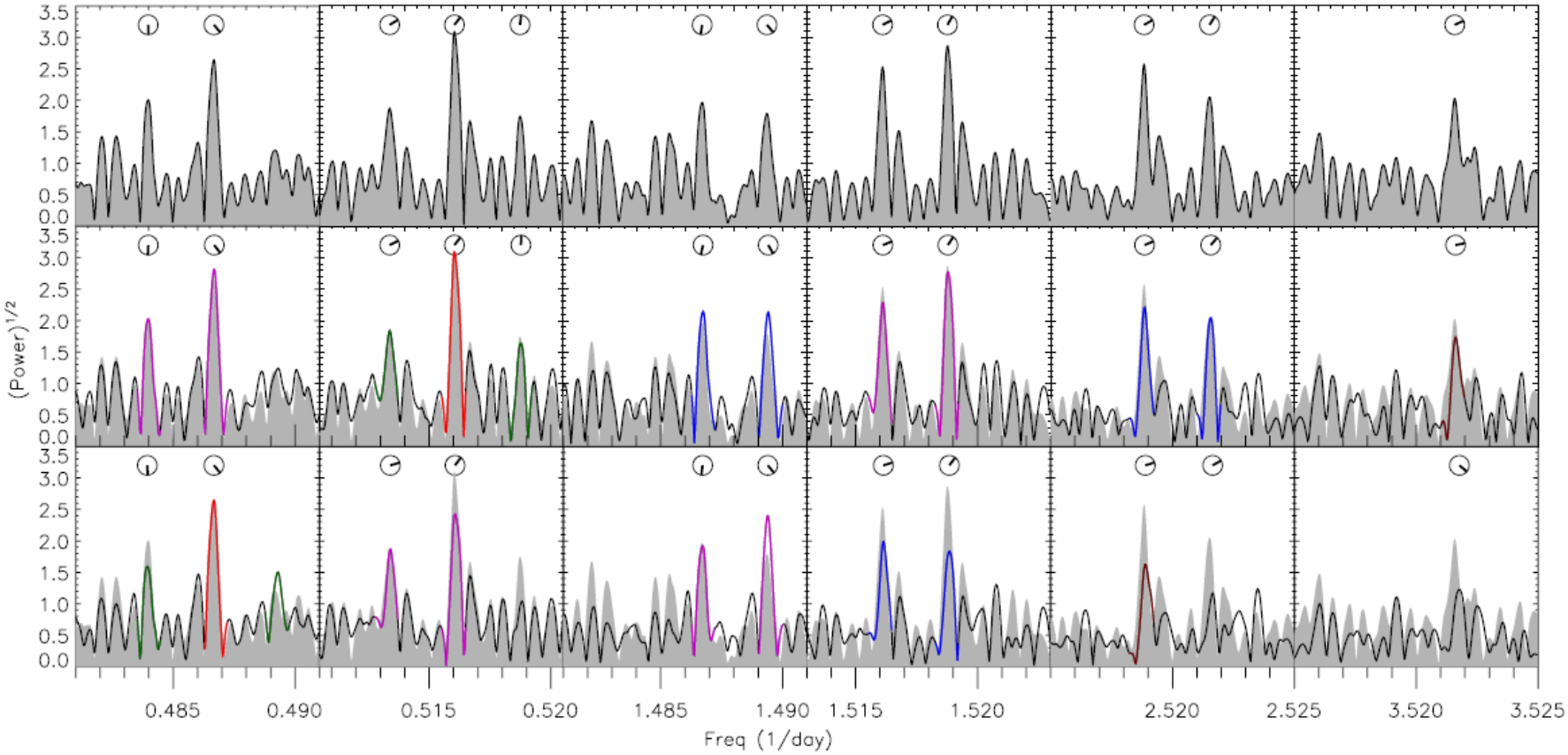
The aliases arise from the convolution of the true spectrum with the window function = Fourier transform of the sampling times.

Rebekah I. Dawson and Daniel C. Fabrycky ApJ, 722:937, 2010



Spectral window function of radial velocity measurements of GJ 876 (Rivera+ 2005). Major features of the spectral window function are colored: red (at 0 day^{-1}), green (yearly feature), fuschia (daily features) , blue (2 day^{-1}), and brown (3 day^{-1}) .

Rebekah I. Dawson and Daniel C. Fabrycky ApJ, 722:937, 2010



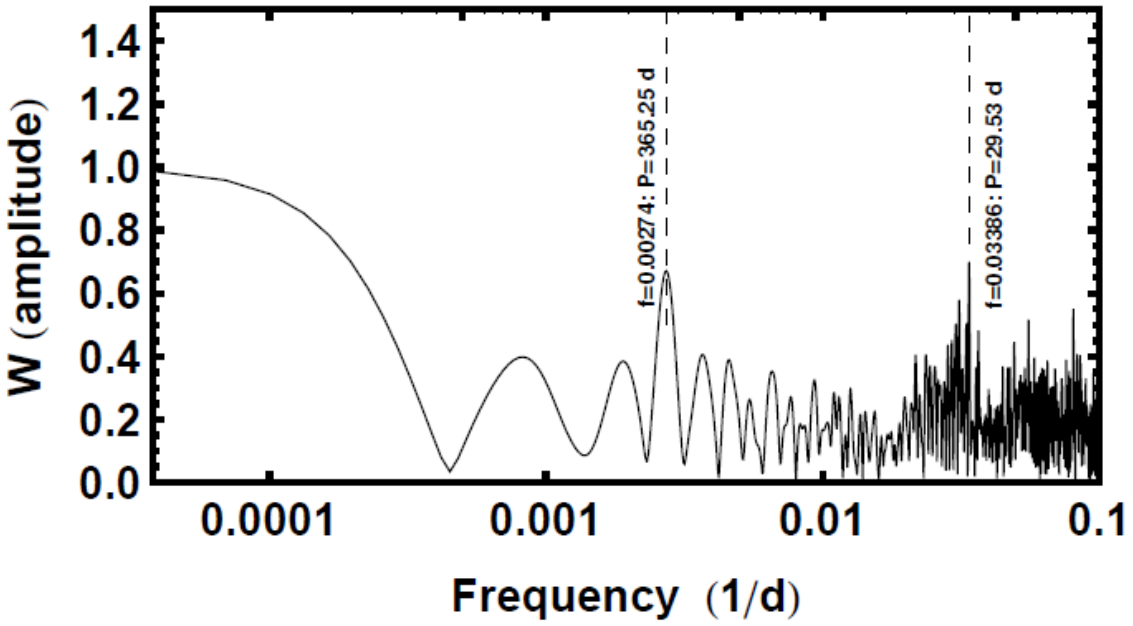
Top row = periodogram of the GJ 876 data. The second row = periodogram of sinusoid with P=1.94 d (frequency 0.516 day^{-1}) sampled at the times of the real data sets as solid lines plus a repeat of the periodogram of the data as a gray background, for comparison. Third row = same as second row but for P=2.05 d (frequency 0.487 day^{-1}). Dials above the peaks show the phase at each peak. Colors correspond to the feature in the window function that creates the particular alias, with red =candidate frequency, the green = yearly aliases, and the fuschia = daily, blue = 2 day^{-1} , and brown = 3 day^{-1}.

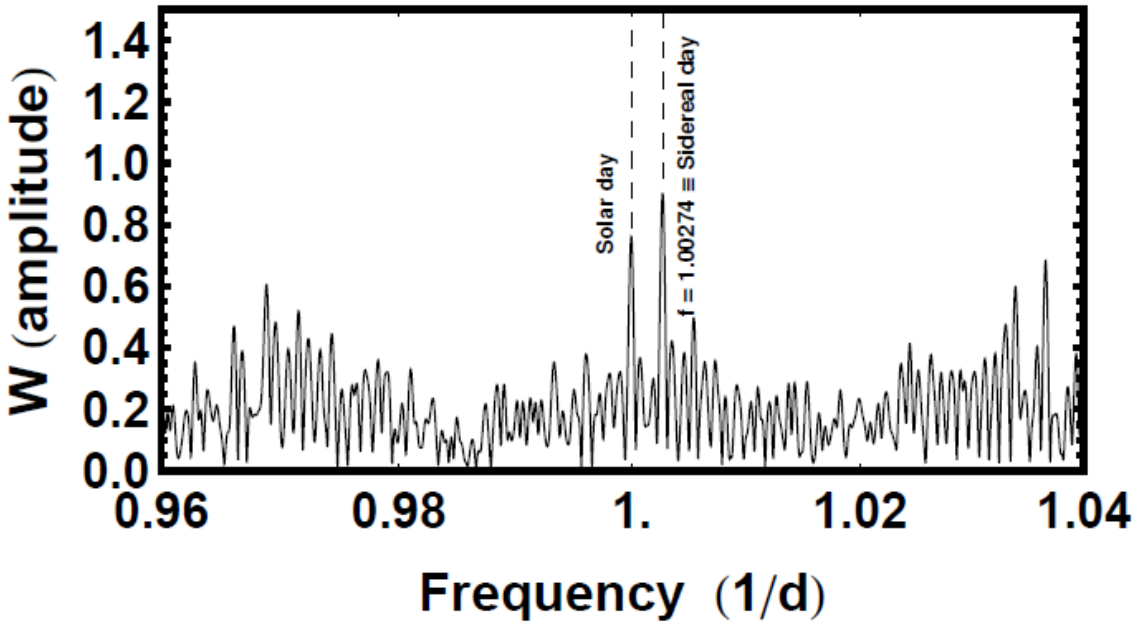
P=1.94 days matches the heights and phases of the peaks much better, both for the yearly aliases on either side of the main peak in Column 2 and the daily aliases in the other columns. Candidate frequencies have different types of aliases at different locations, allowing us to break the degeneracy.

Window function HD 208487

$$W(f) = N^{-1} \sum_{k=1}^{N} e^{i2\pi f t_k}$$

Synodic month = 29.53 d, measured from a lunar phase until the return of that same phase

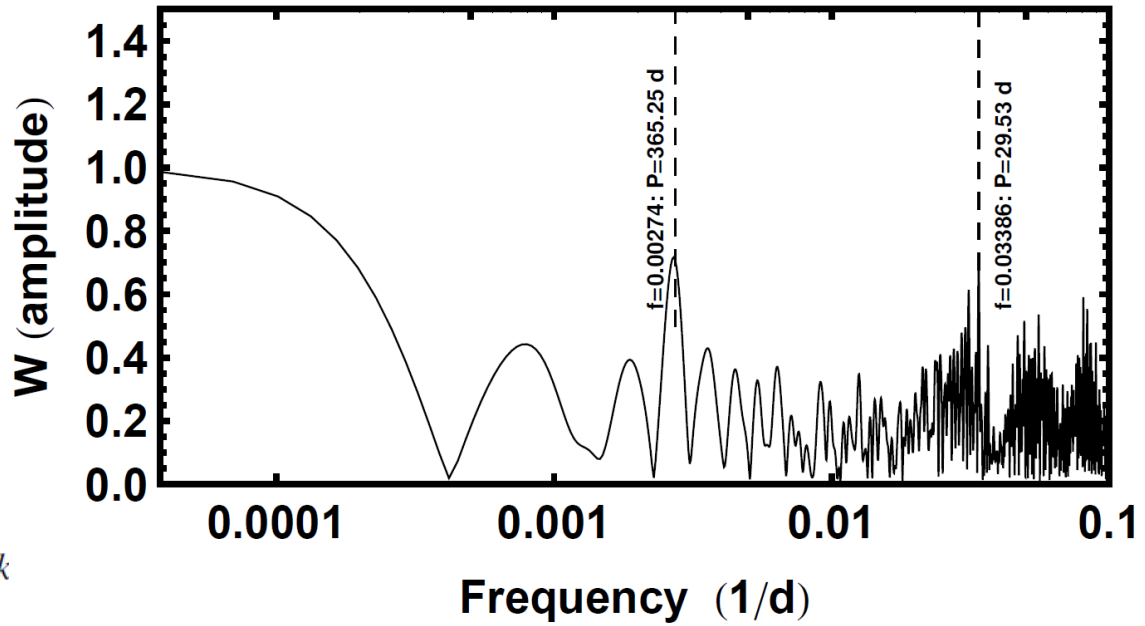


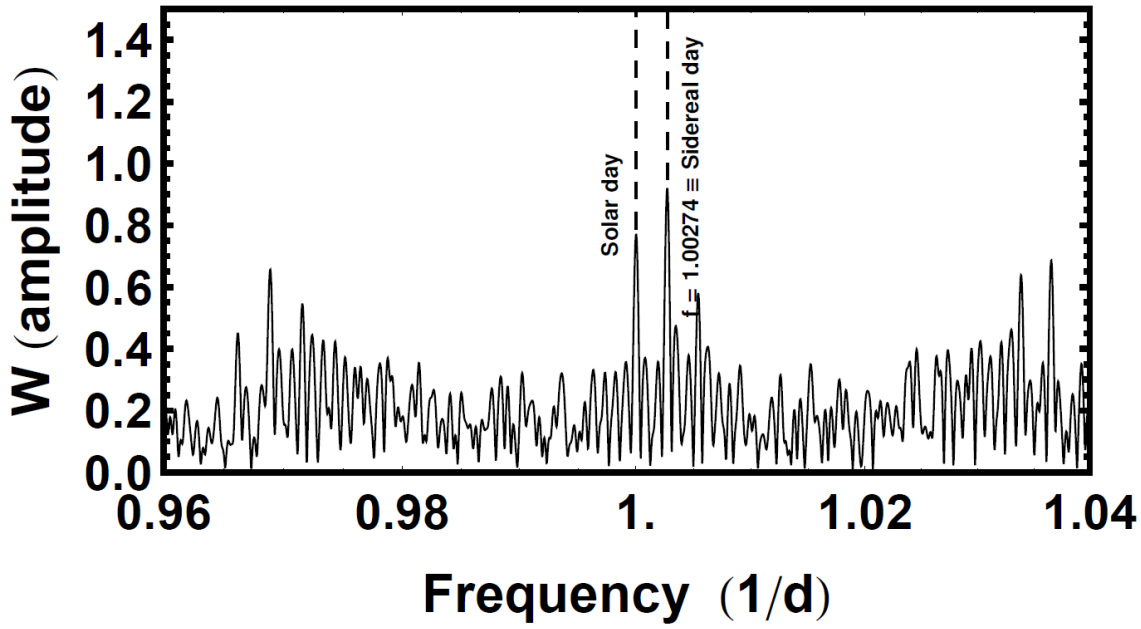


Weighted Window function HD 208487

$$W(f) = wt_{\text{sum}}^{-1} \sum_{k=1}^{N} wt_k \; e^{i2\pi f t_k}$$

Synodic month = 29.53 d, measured from a lunar phase until the return of that same phase)

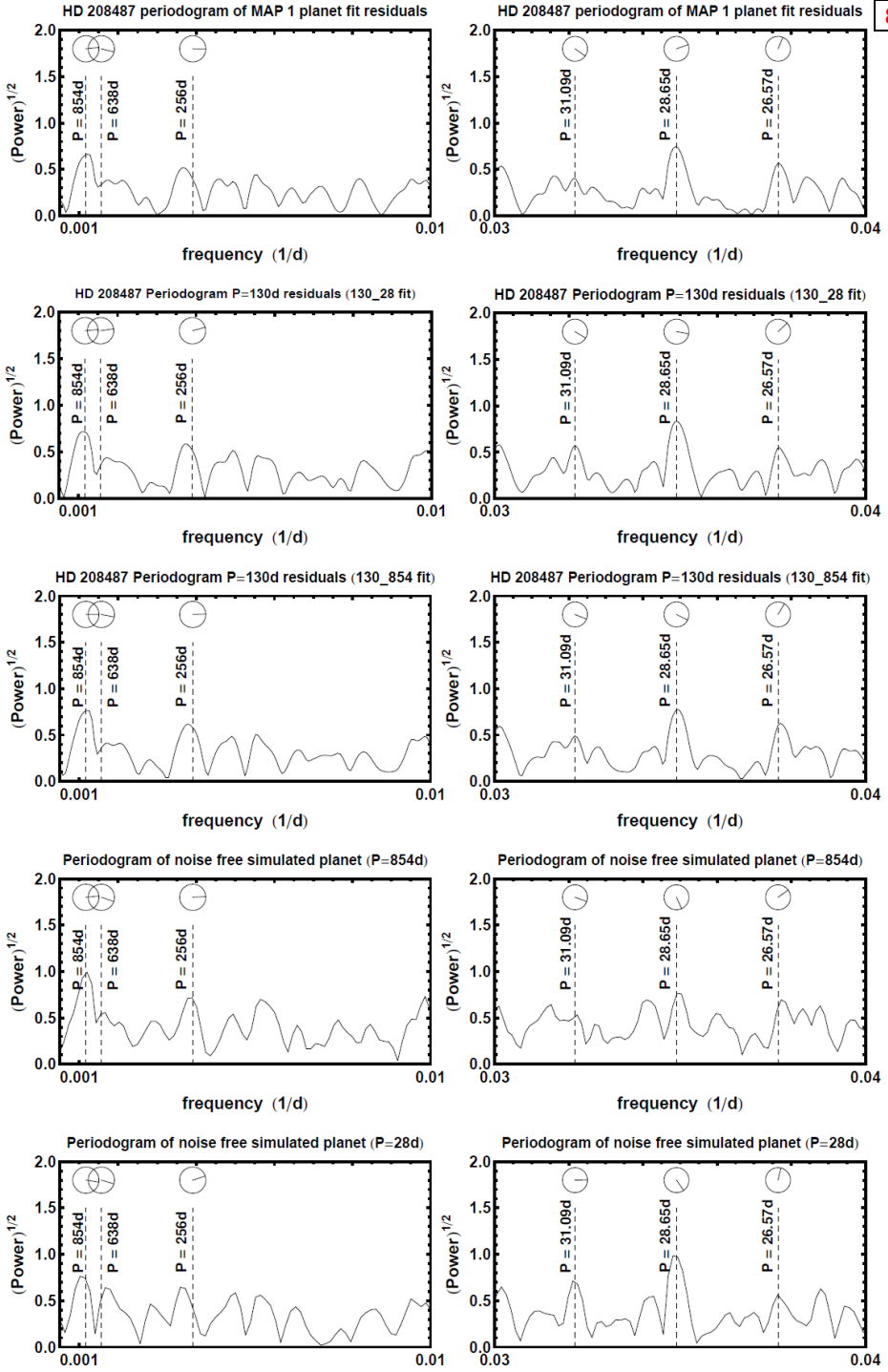




HD 208487

Alas, no clear cut conclusion from the periodogram and phase circle plots.

Model selection discussed shortly, favors the ~ 900 d signal over 28 d signal by a factor ~ 10.



FMCMC software demo

Gliese 581

Gliese 581 the star with two possible habitable zone planets

History

1) 2005 to 2009,
   Planet e is 1.9 Earth mass
   Planet b is 16 Earth mass
   Planet c is 5 Earth mass
   Planet d is 7 Earth mass (HZ)

M. Mayor et al., A&A, 507, p. 487, Nov. 2009

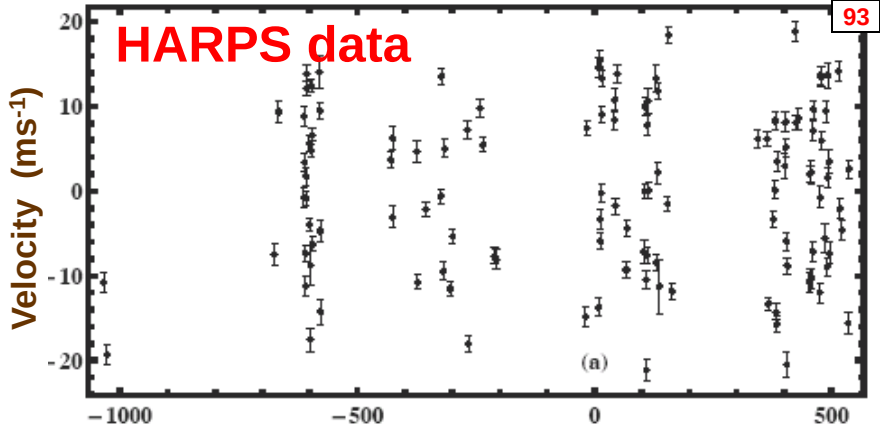
2) 2010,
   Planet f is 7 Earth mass
   Planet g is 3.1 Earth mass (HZ)

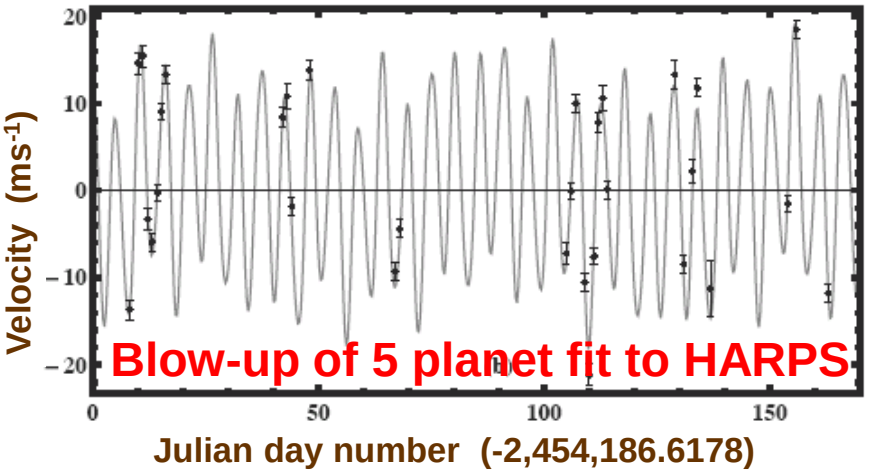
Steven Vogt et al., ApJ, 723, p. 954, 2011, their analysis assumes all circular orbits.

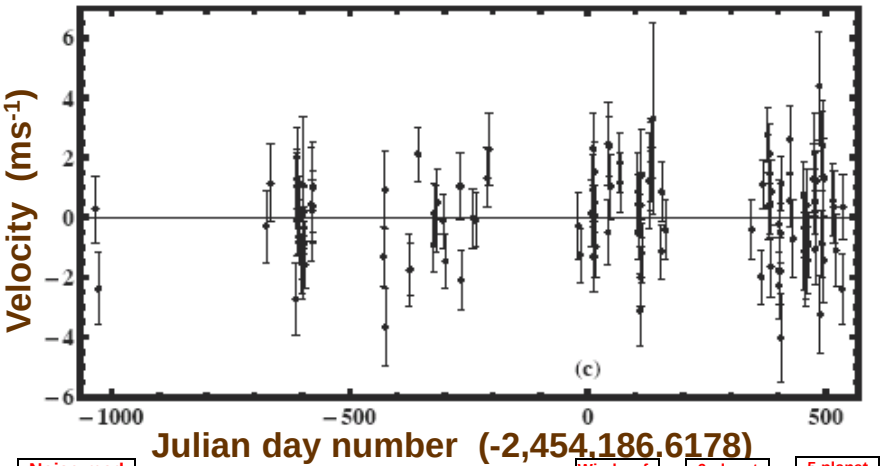
3) 2011, Gregory, P. C., MNRAS, 415, 2523

Borderline evidence for 5 planets.
Don't support claim for planet 581g.

Find evidence that Keck HIRES uncertainties are much larger than the quoted values by an extra 1.8 m s^{-1} added in quadrature.

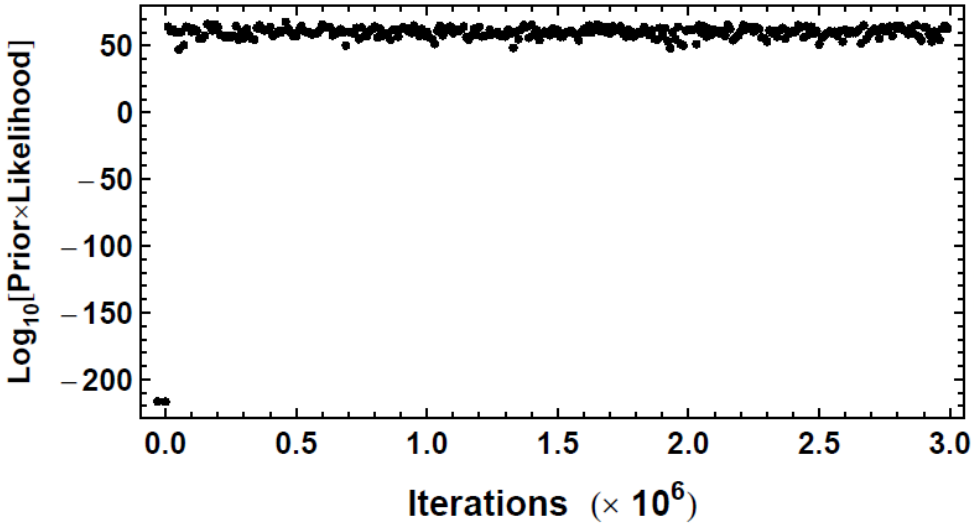


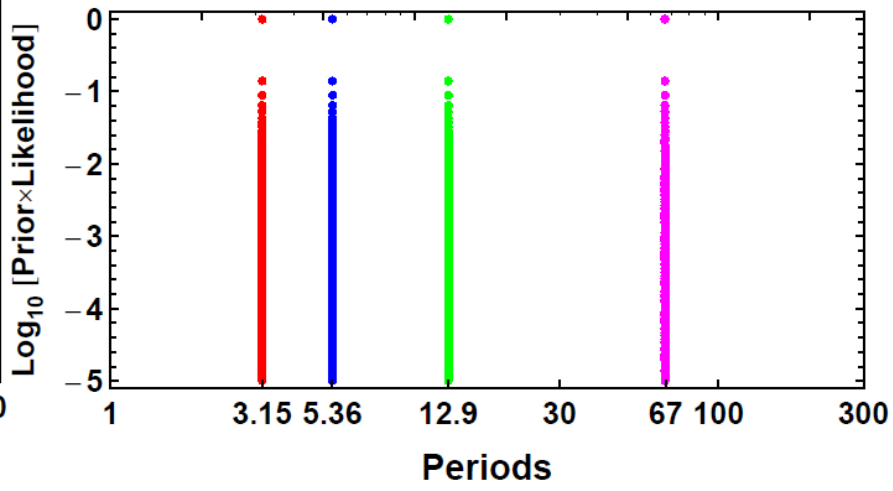


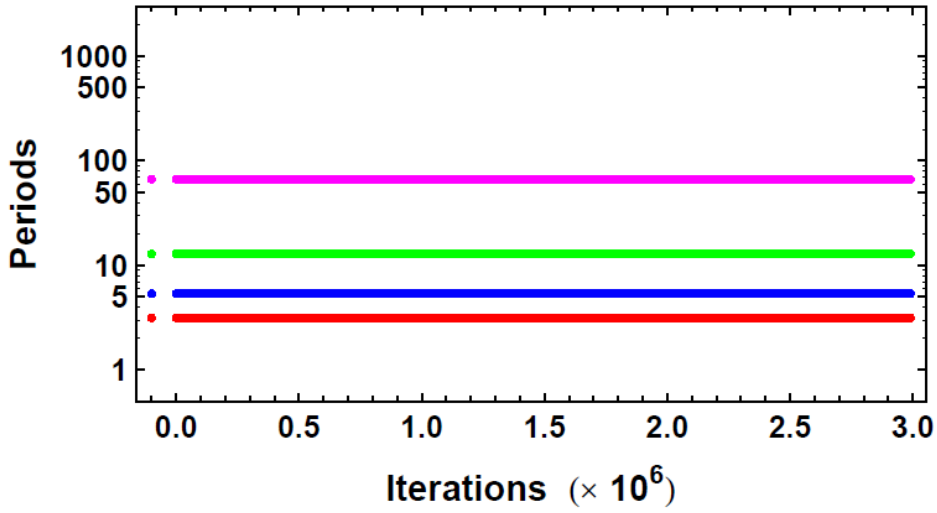


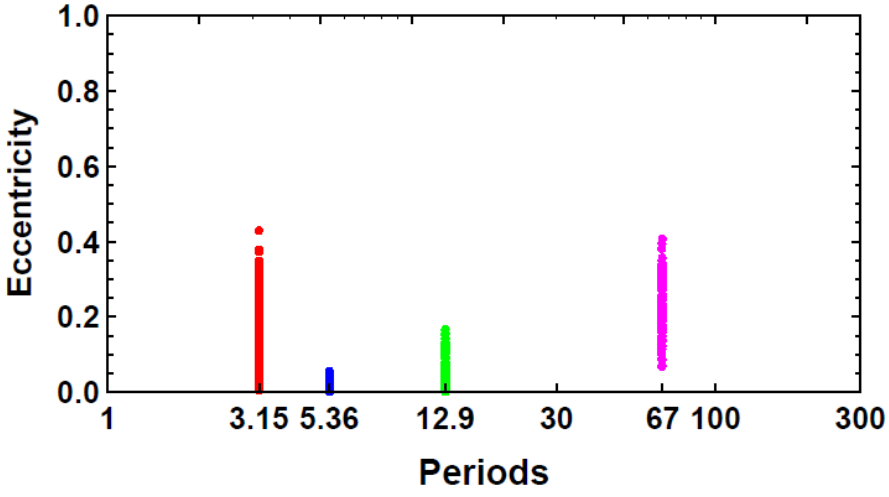
Gliese periodogram plots (4 planet model)

$p(f|M,I) \propto 1/f$



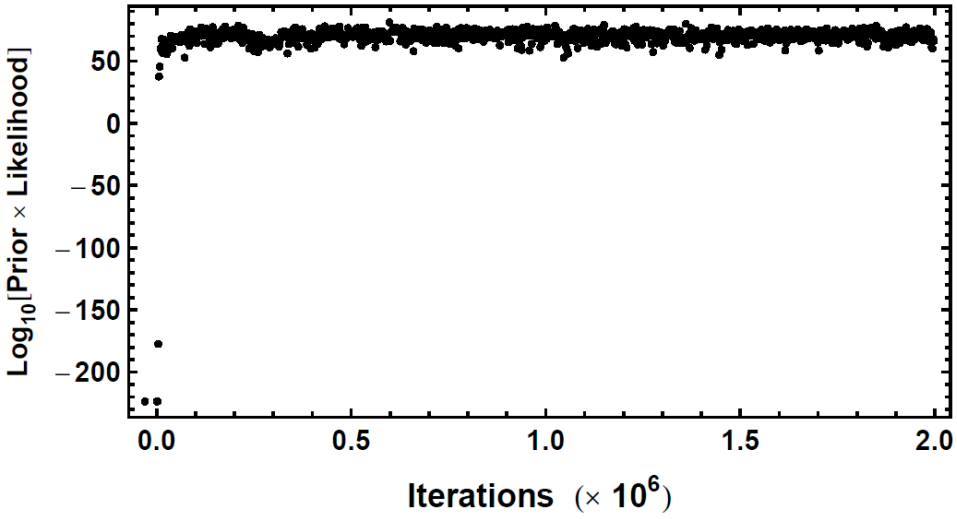


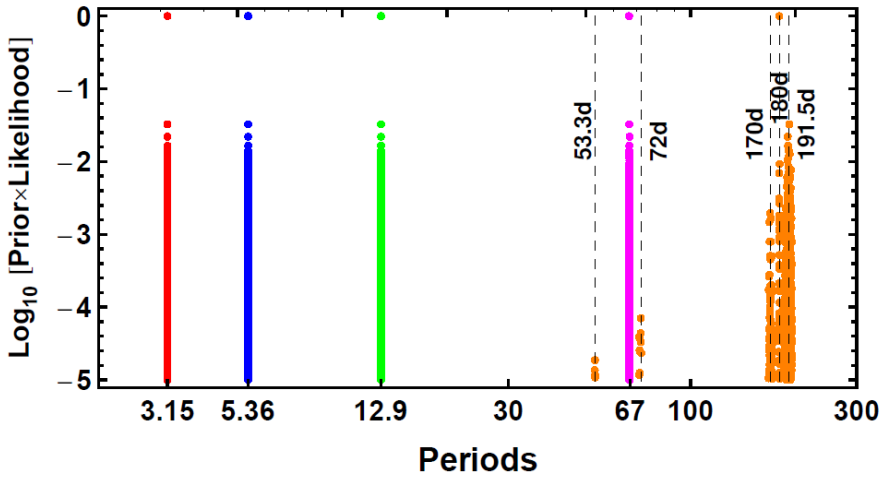


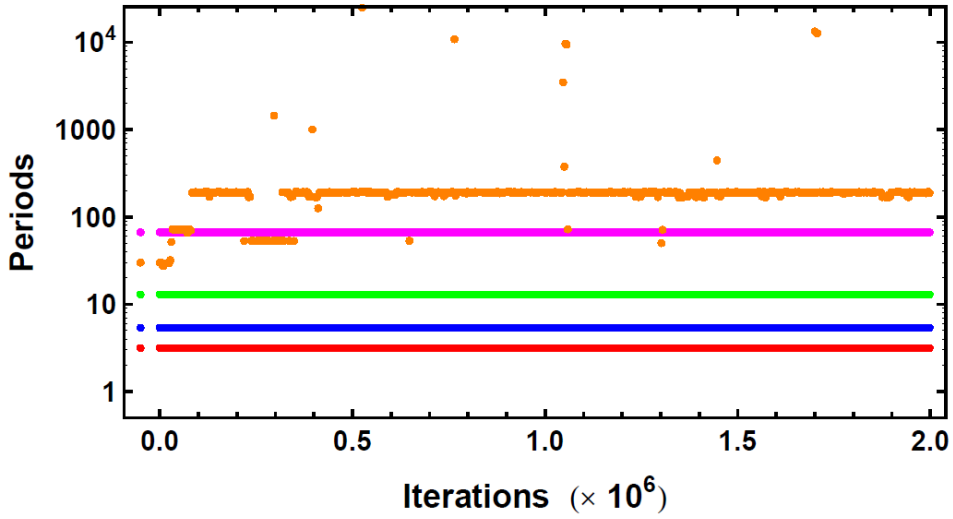


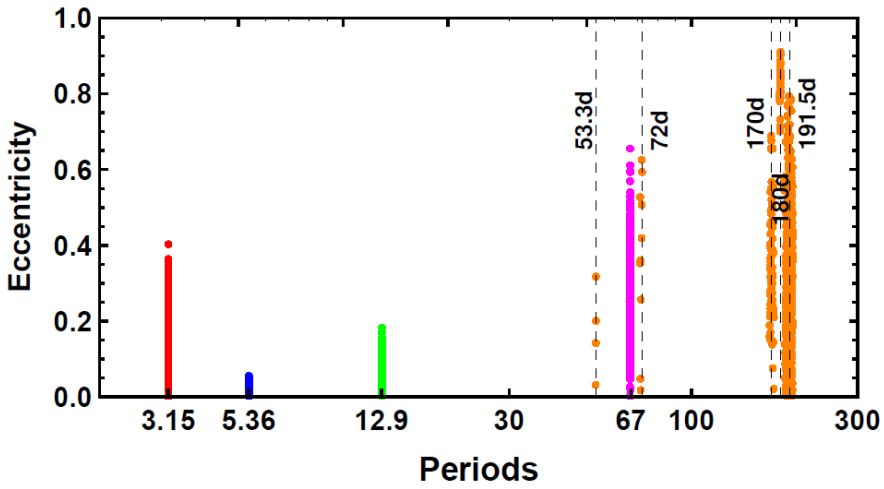
Gliese periodogram plots (5 planet model)

$p(f|M,I) \propto 1/f$



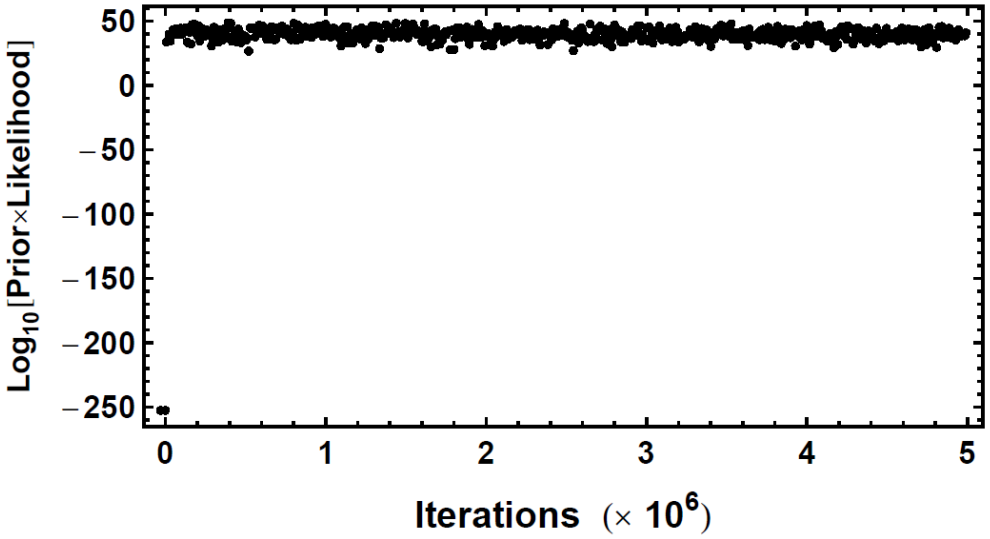


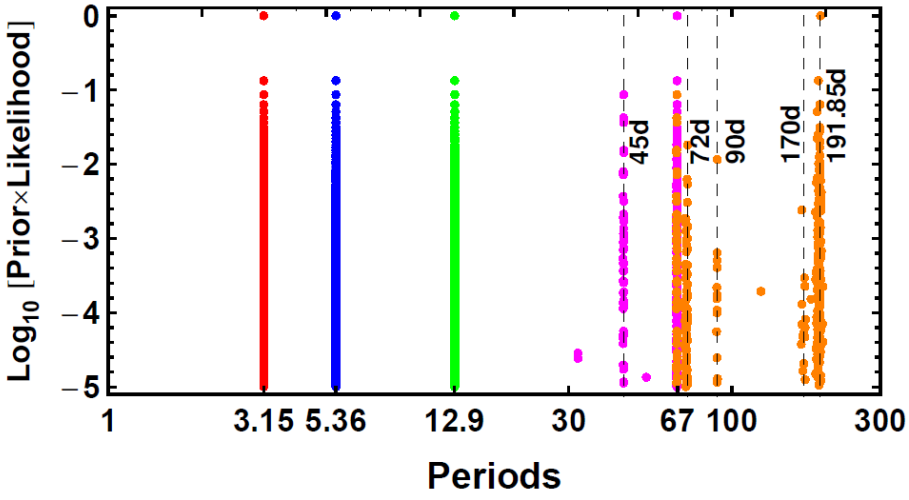


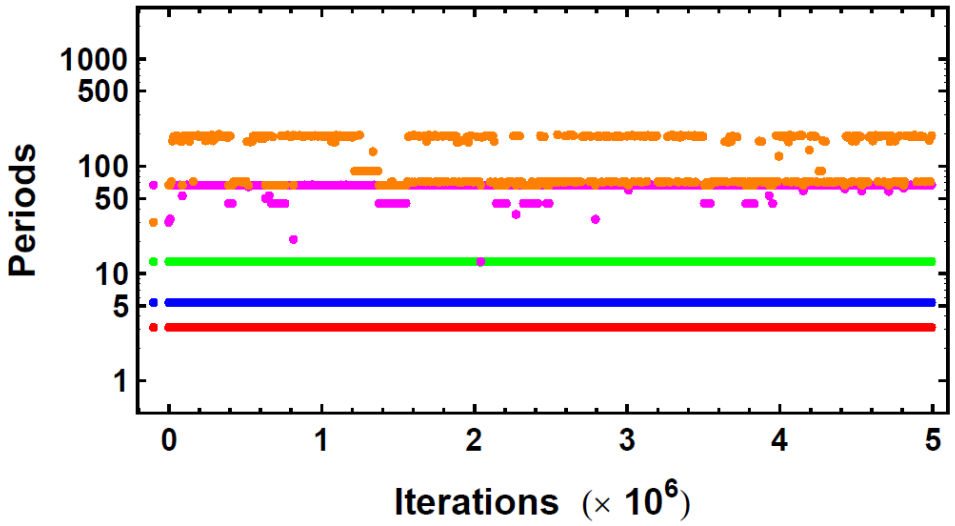


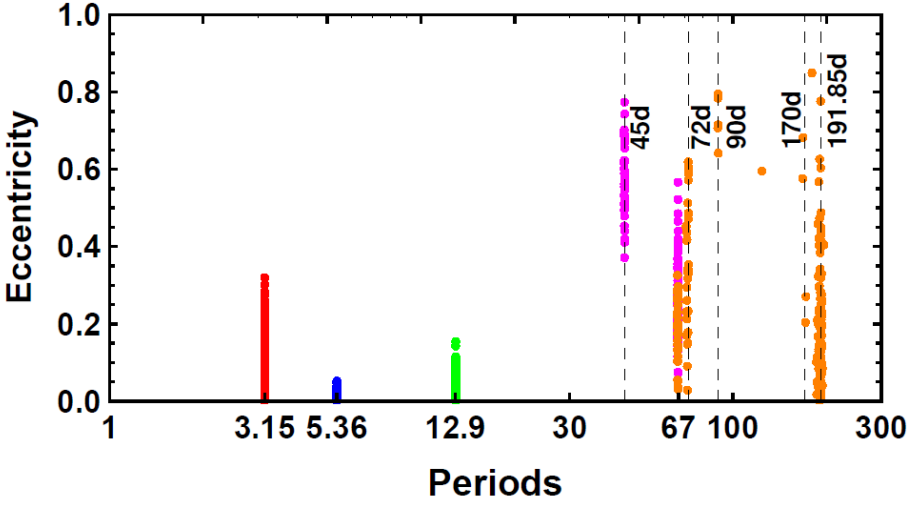
Gliese periodogram plots (5 planet model)

$p(f|M,I) \propto 1/\sqrt{f}$

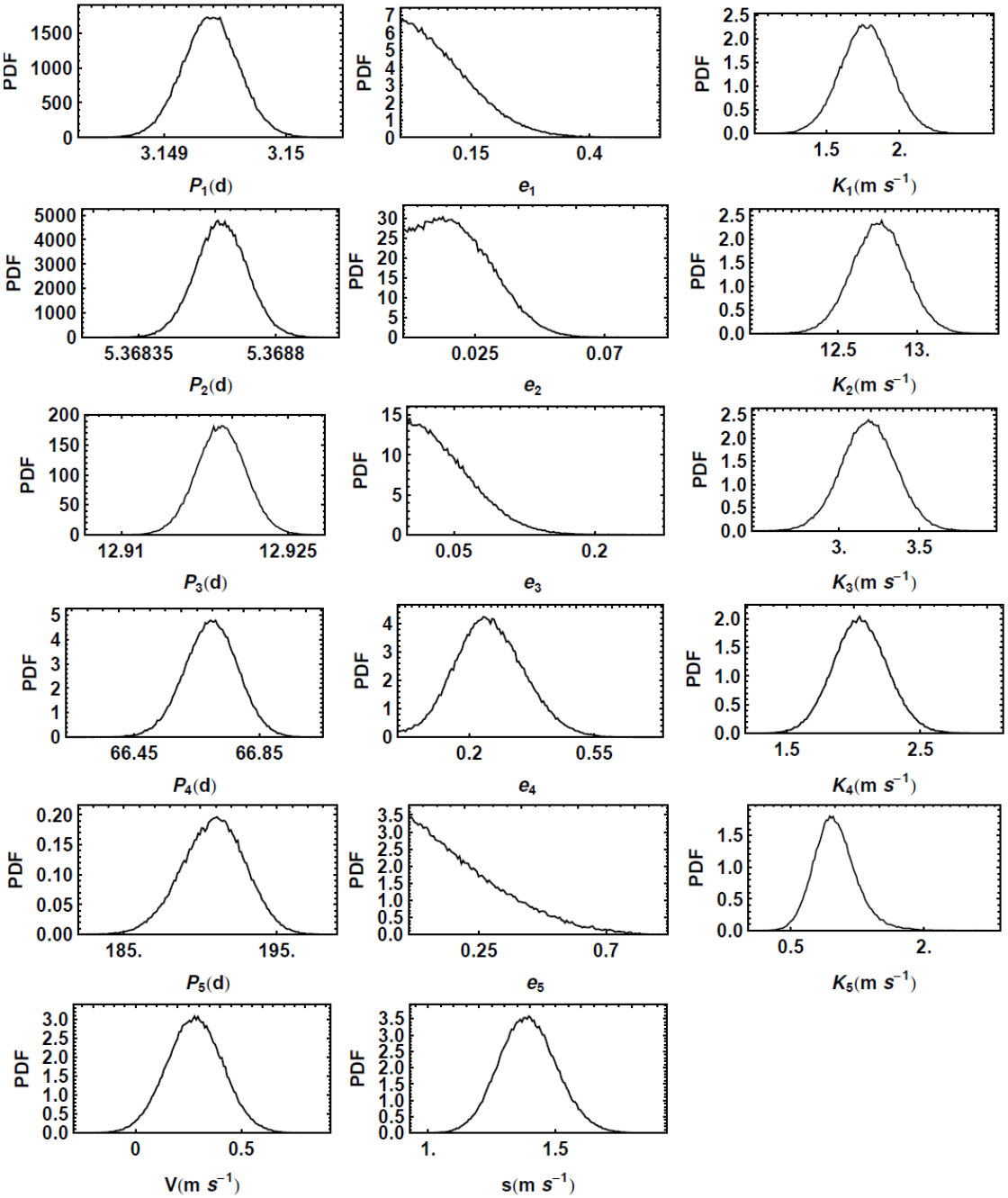








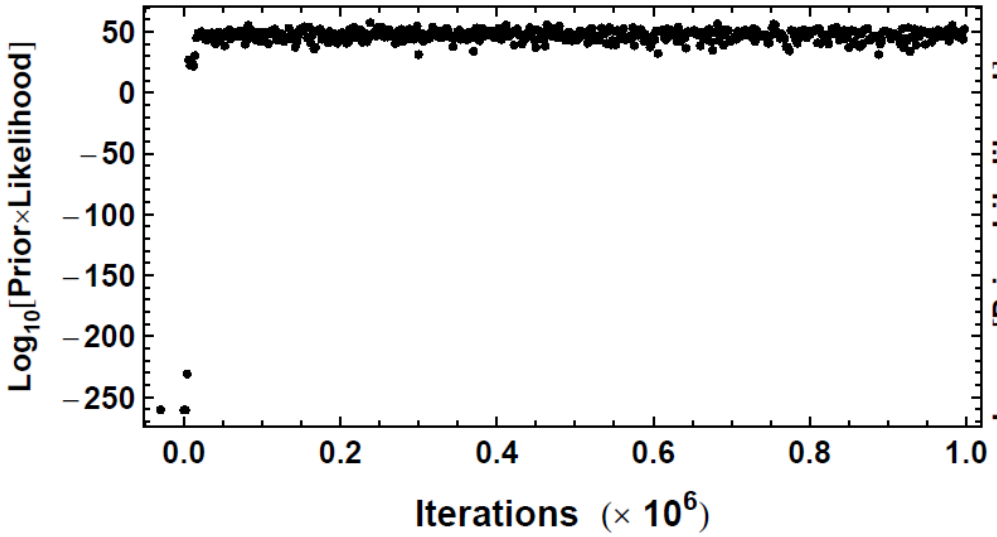
Gliese parameter marginals (5 planet model) $p(f|M,I) \propto 1/\sqrt{f}$

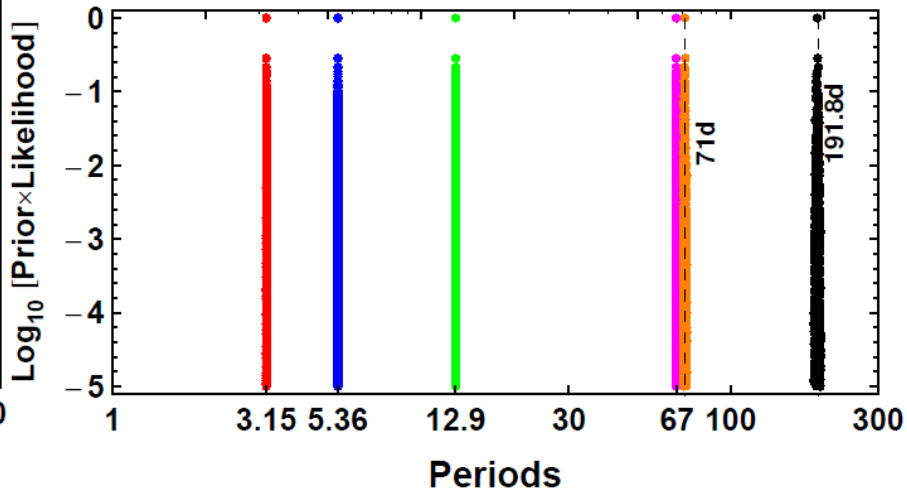


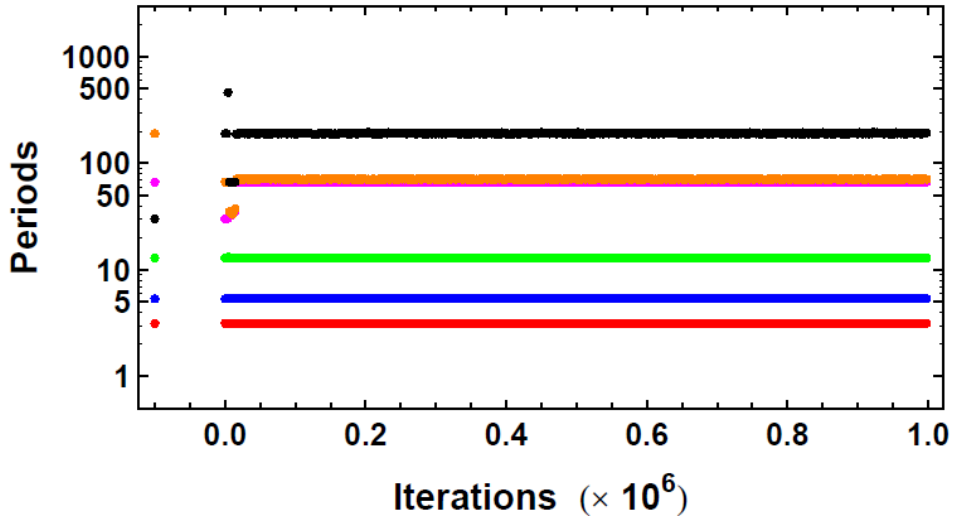
After filtering out the post burn-in FMCMC iterations that correspond to the 5 dominant period peaks at 3.15, 5.37, 12.9, 66.9, and 192d

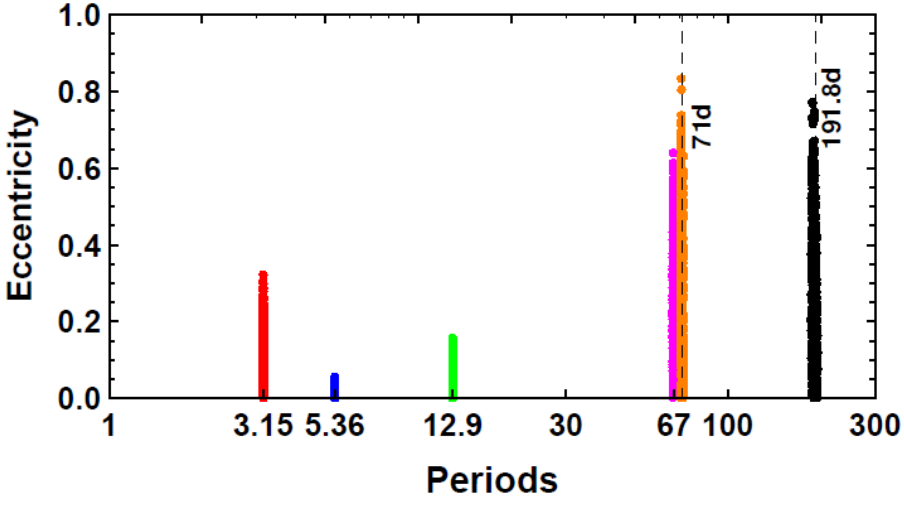
Gliese periodogram plots (6 planet model)

$p(f|M,I) \propto 1/\sqrt{f}$









Model Selection

Bayesian Model Selection

Compare two models by computing the ratio of their posterior probabilities (odds ratio) which automatically incorporates a quantitative Occam's razor.

Expand with Bayes' theorem

$$\text{Odd ratio} = \frac{p(M_1|D,I)}{p(M_0|D,I)} = \frac{\dfrac{p(M_1|I)\,p(D|M_1,I)}{p(D|I)}}{\dfrac{p(M_0|I)\,p(D|M_0,I)}{p(D|I)}} = \frac{p(M_1|I)}{p(M_0|I)} \frac{p(D|M_1,I)}{p(D|M_0,I)}$$

posterior probability ratio — **prior probability ratio** — **Bayes factor**

Bayes factor, $B_{10} = \dfrac{p(D|M_1,I)}{p(D|M_0,I)} = \text{marginal likelihood ratio}$

marginal likelihood for M_1

$$p(D|M_1,I) = \int d\theta\; p(\theta|M_1,I)\; p(D|\theta,M_1,I)$$

In words: the marginal likelihood for a model is the weighted average likelihood for its parameter(s). The weighting function is the prior for the parameter.

Occam return

Bayesian Model Selection and Extrasolar Planets

Eric B. Ford and Philip C. Gregory

in 'Statistical Challenges in Modern Astronomy IV', G. J. Babu and E. D. Feigelson,eds, Astron. Soc. of the Pacific Conference Series, 371, pp. 189-205 (2007)

We compared a wide variety of estimators of the marginal likelihood and feature those that display desirable convergence properties based on the analysis of a sample data set for HD 88133. This was the focus of a SAMSI workshop on exoplanets. Some of the estimators considered:

- Restricted Monte Carlo
- Partial linearization and Laplace Approximation
- Harmonic mean
- Weighted harmonic mean
- Basic importance sampling
- Importance sampling with a mixture of multivariate normals
- Ratio estimator
- Parallel tempering

Since then I have added one new method:

- Nested Restricted Monte Carlo

and compared to Parallel tempering & Ratio estimator methods

Bayesian model selection from parallel tempering MCMC

Markov chain Monte Carlo analysis produces samples in model parameter space in proportion to the posterior probability distribution. This is fine for parameter estimation.

For model selection we need to determine the proportionality constant to evaluate the marginal likelihood $p(D|M_i, I)$ for each model.

One solution is to use the MCMC results from all the parallel tempering chains spanning a wide range of β values.

$$\pi(X|D, M, \beta, I) = p(X|M, I)\ p(D|X, M, I)^{\beta} \quad (0 < \beta \leqslant 1)$$

Prior

Likelihood

Here X represents the model parameters and β = 1 corresponds to our desired target distribution. Others values of β correspond to progressively flatter probability distributions.

Model probabilities from parallel tempering chains commonly known as thermodynamic integration

$$\ln[p(D|M_i, I)] = \int d\beta \langle \ln[p(D|M_i, X, I)] \rangle_\beta$$

where $\langle \ln[p(D|M_i, X, I)] \rangle_\beta$

= expectation value of $\ln[p(D|M_i, X, I)]$ for a given β

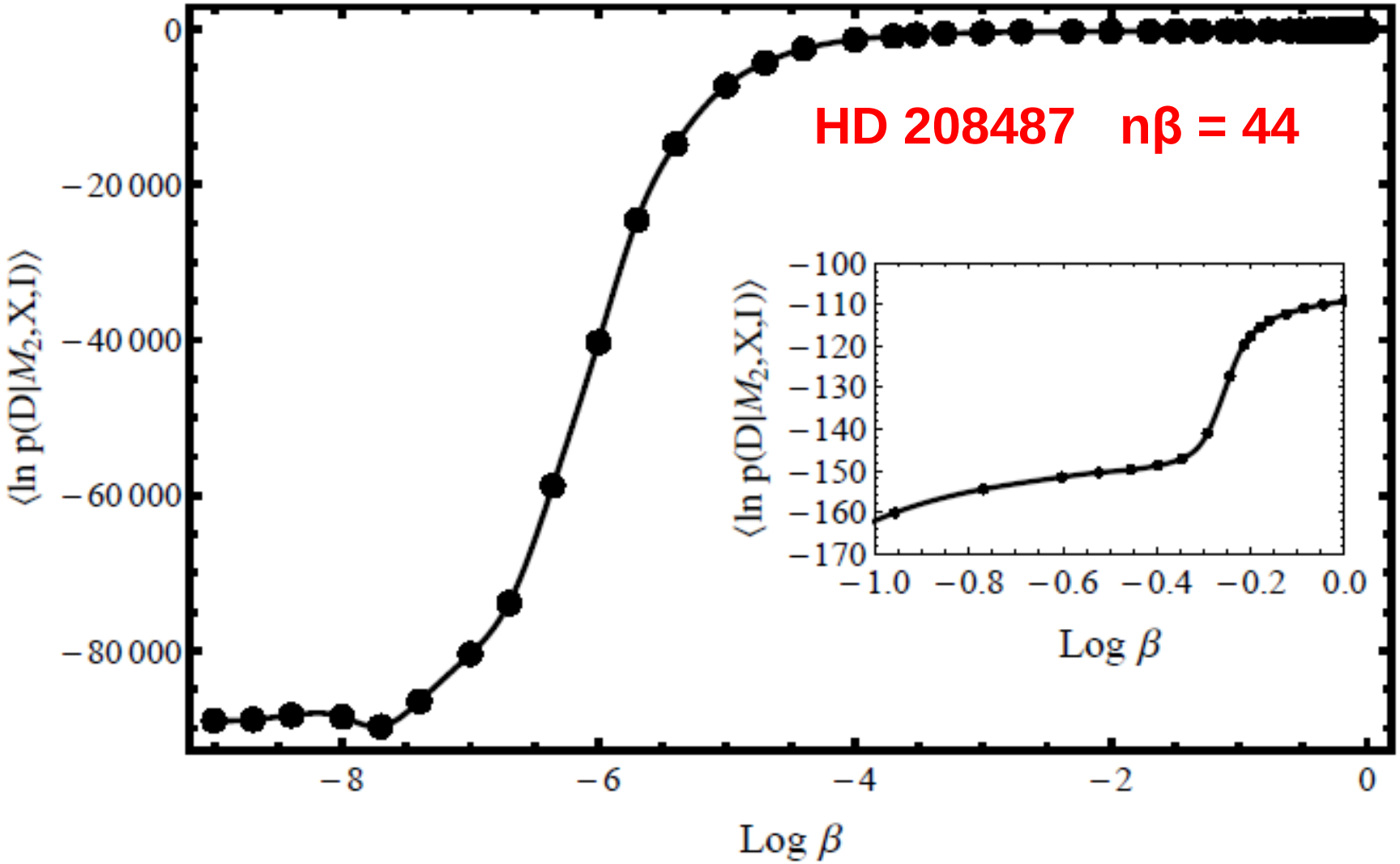
$$= \frac{1}{n} \sum_t \ln[p(D|M_i, X_{t,\beta}, I)]$$

n = number of MCMC iterations

and $p(D|M_i, X_{t,\beta}, I)$ is the likelihood of the set of parameter choices for iteration t and for the given value of β.

$$\ln[p(D \mid M_2, I)] = \int \langle \ln[p(D \mid M_2, X, I)] \rangle_\beta \, d\beta$$

used in the calculation of model probabilities



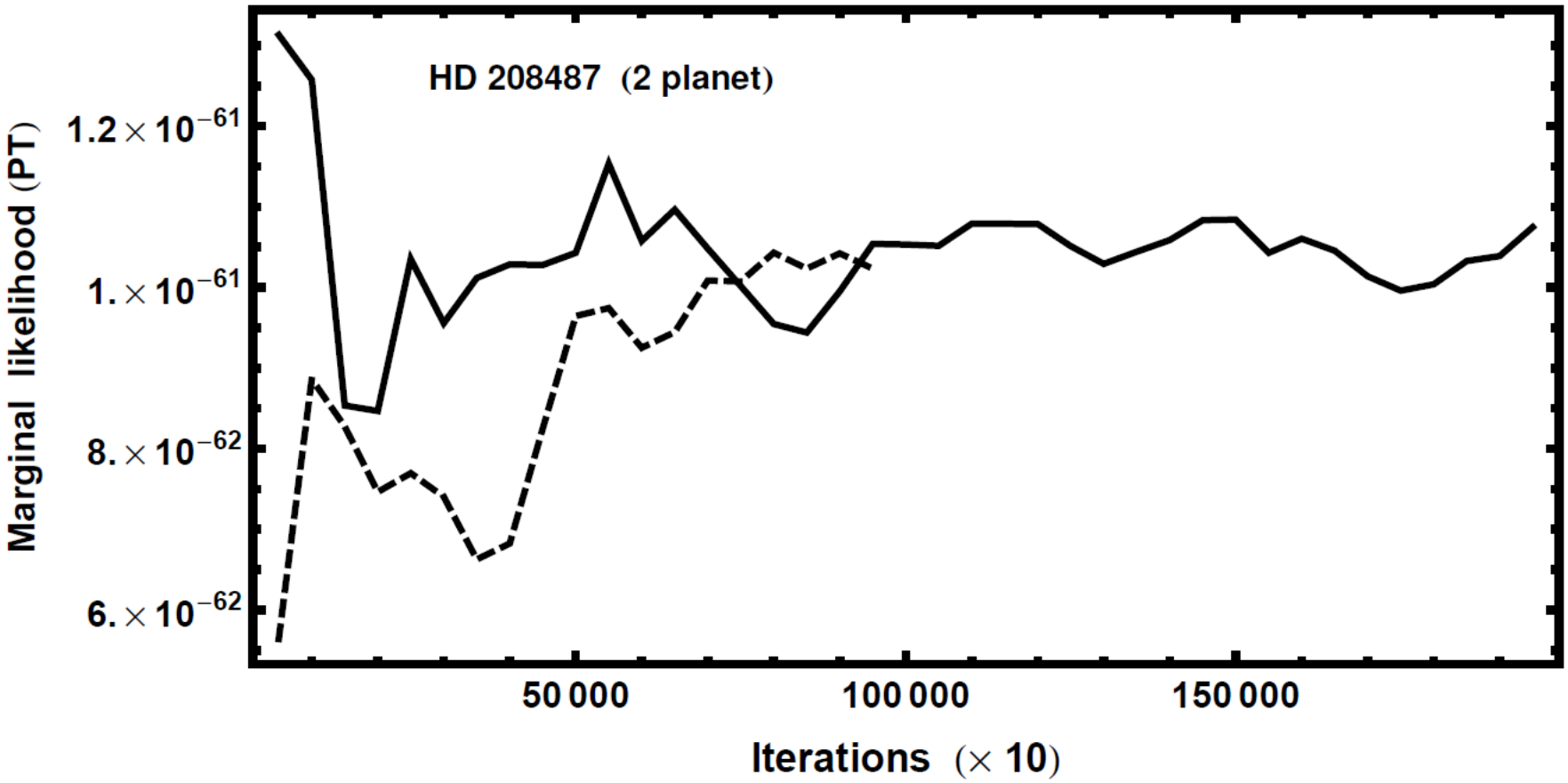
Fractional error in marginal likelihood versus parallel tempering β value

β range	$p(D\|M_2, I)_{PT}$	Fractional error
$10^{-1} - 1.0$	3.290×10^{-52}	3×10^{9}
$10^{-2} - 1.0$	4.779×10^{-60}	43
$10^{-3} - 1.0$	2.817×10^{-61}	2.6
$10^{-4} - 1.0$	1.635×10^{-61}	0.51
$10^{-5} - 1.0$	1.306×10^{-61}	0.21
$10^{-6} - 1.0$	1.148×10^{-61}	0.06
$10^{-7} - 1.0$	1.091×10^{-61}	0.008
$10^{-8} - 1.0$	1.083×10^{-61}	0.0008
$10^{-9} - 1.0$	1.082×10^{-61}	0.0

The third column gives the fractional error in the marginal likelihood that would result if this decade of β was not included, which indicates the sensitivity of the result to that decade.

return 208487

Marginal likelihood estimate versus MCMC iteration number



A plot of the marginal likelihood, $p(D|M_2, X, I)_{PT}$, versus FMCMC iteration for the two planet HD 208487 model results for two trials.

Nested Restricted Monte Carlo

One method I developed to estimate the marginal likelihoods is Nested Restricted Monte Carlo (NRMC) integration. For large parameter spaces, Monte Carlo (MC) integration is hopelessly inefficient in exploring the whole prior parameter range. The fraction of the prior volume containing significant probability rapidly declines as the number of dimensions increase.

In Restricted MC (RMC) this is less of a problem because the volume of parameter space sampled is greatly restricted to a region delineated by the outer borders of the marginal distributions of the parameters obtained from the MCMC run.

In Nested RMC (NRMC) integration, multiple boundaries are constructed based on credible regions ranging from 30% to > 99%, as needed. We are able to compute the contribution to the total integral from each nested interval and sum these contributions. For example, for the interval between the 30% and 60% credible regions, we generate random parameter samples within the 60% region and reject any sample that falls within the 30% region. Using the remaining samples we can compute the contribution to the NRMC integral from that interval.

Nested Restricted Monte Carlo (NRMC) Integration

Construction of hypercubes

In NRMC integration, multiple boundaries of a restricted hypercube in parameter space are constructed based on credible regions ranging from 30% to ≥ 99%, as needed.

To construct the *x*% hypercube we compute the *x*% credible region of the marginal distribution for each parameter of the particular model.

The *x*% hypercube is delineated by the *x*% credible range of the marginal for each parameter.

Note: the actual fraction of the total probability of the joint posterior distribution contained within a hypercube defined in this way will be greater than x%.

Nested Restricted Monte Carlo (NRMC) Integration

Construction of hypercubes

In NRMC integration, multiple boundaries of a restricted hypercube in parameter space are constructed based on credible regions ranging from 30% to ≥ 99%, as needed.

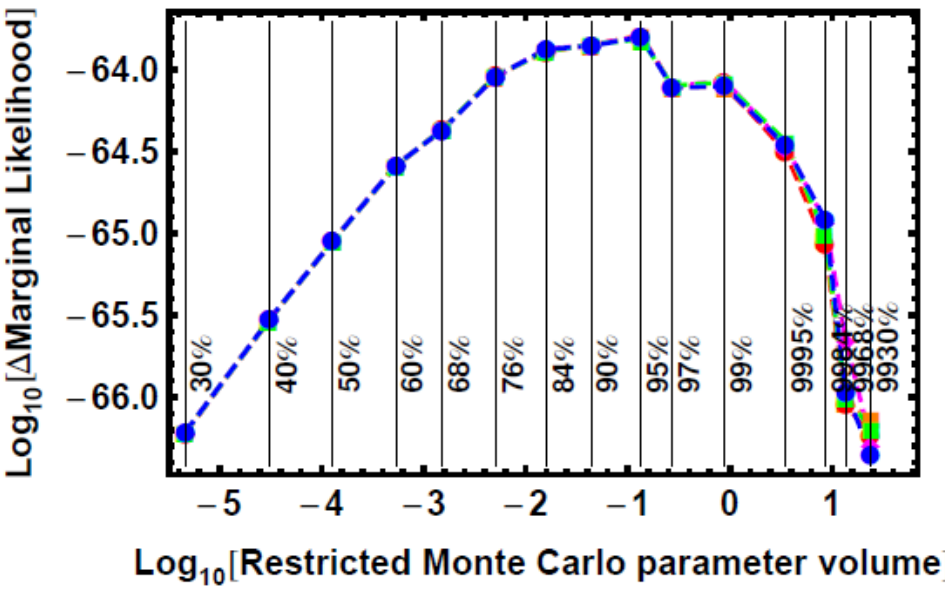
To construct the *x*% hypercube we compute the *x*% credible region of the marginal distribution for each parameter of the particular model.

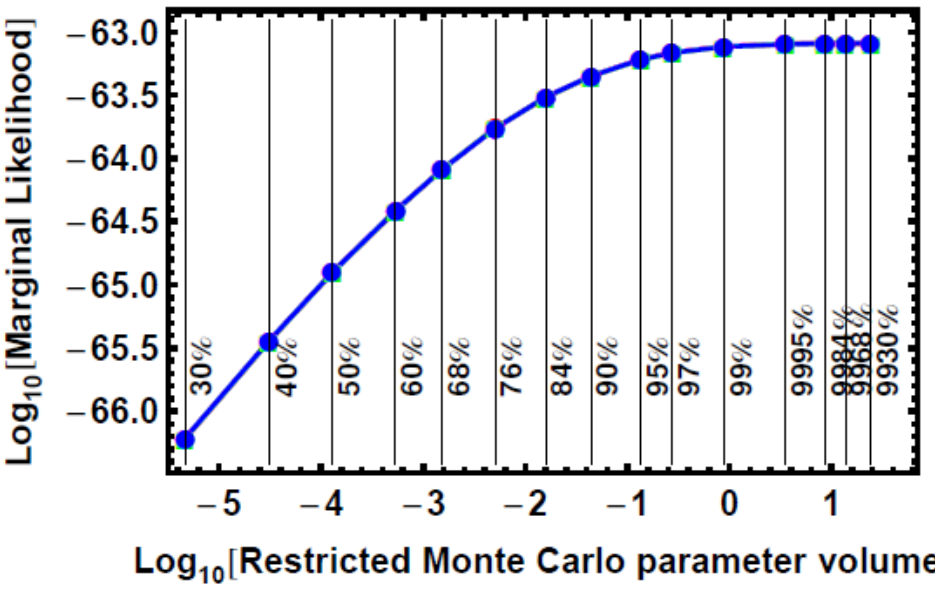
The *x*% hypercube is delineated by the *x*% credible range of the marginal for each parameter.

Contribution from each nested interval

For example, for the interval between the 30% and 60% hypercubes, generate random parameter samples within the 60% hypercube and reject any sample that falls within the 30% hypercube. Using the remaining samples we can compute the contribution to the NRMC integral from that interval.

NRMC Integration HD 208487 (1 planet)





The left panel shows the contributions from the individual intervals for 5 repeats of the NRMC evaluation for the 1 planet HD 208487 model. The right panel shows the summation of the individual contributions versus the volume of the credible region.

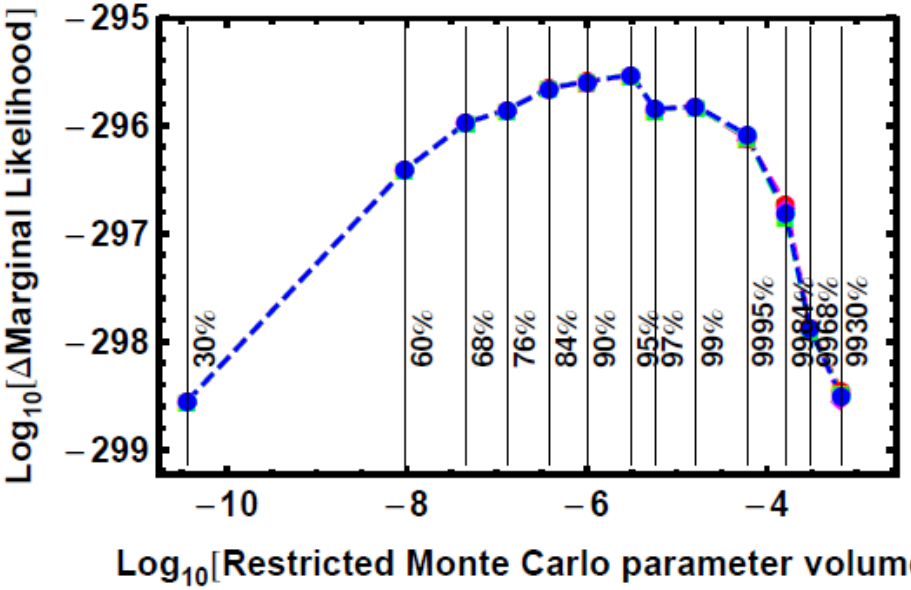
The 9995% boundary is defined as follows. Let XU99 and XL99 correspond to the upper and lower boundaries of the 99% credible region, for a particular parameter. Similarly, XU95 and XL95 are the upper and lower boundaries of the 95% credible region for the parameter.

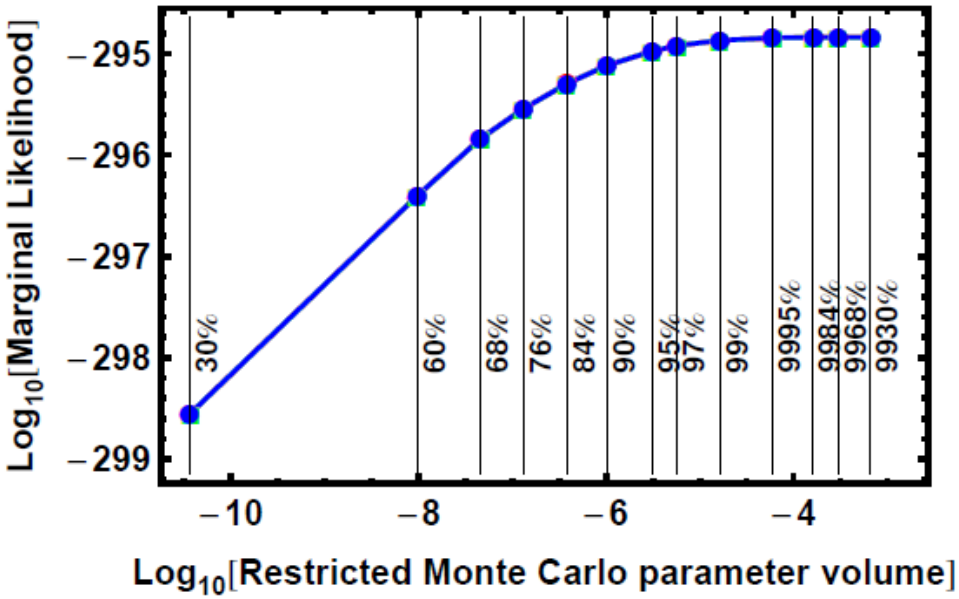
Then XU9995 = XU99 +(XU99−XU95),

XL9995 = XL99 +(XL99−XL95).

Similarly, XU9984 = XU99 + (XU99 −XU84).

NRMC Integration Gliese 581 (1 planet)

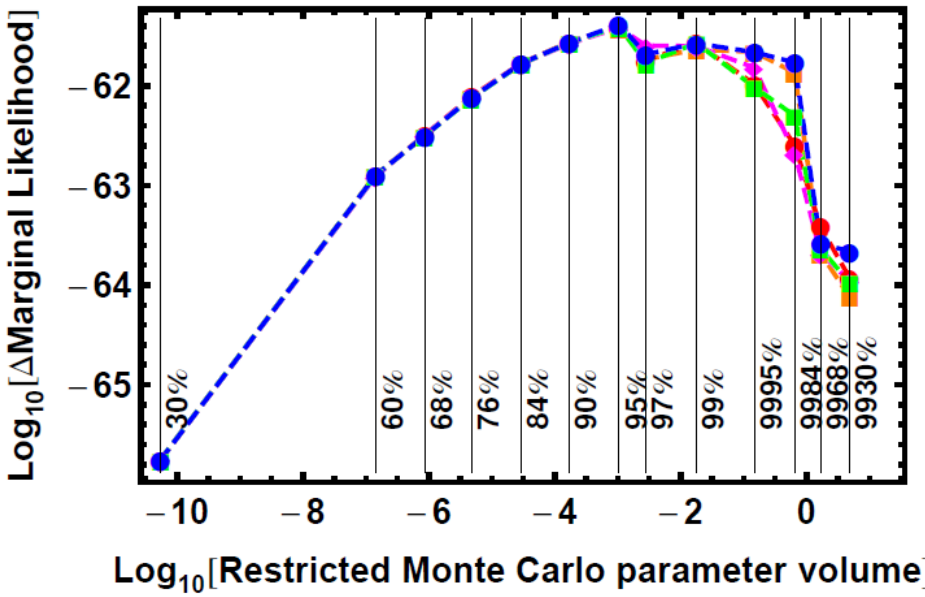


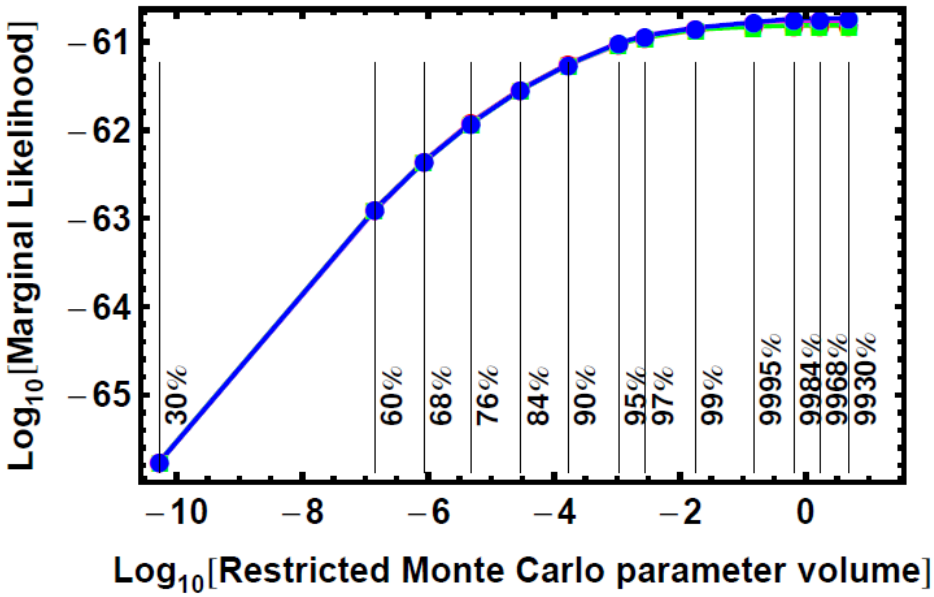


The left panel shows the contributions from the individual intervals for 5 repeats of the NRMC evaluation for the 1 planet Gliese 581 model. The right panel shows the summation of the individual contributions versus the volume of the credible region.

The mean value of the prior × likelihood within the 30% credible region is 2×10^5 larger than the mean in the shell between the 97 and 99% credible regions. However, the volume of the shell between 97 and 99% is 8×10^{11} larger so the contribution from the 30% credible region to the marginal likelihood is negligible.

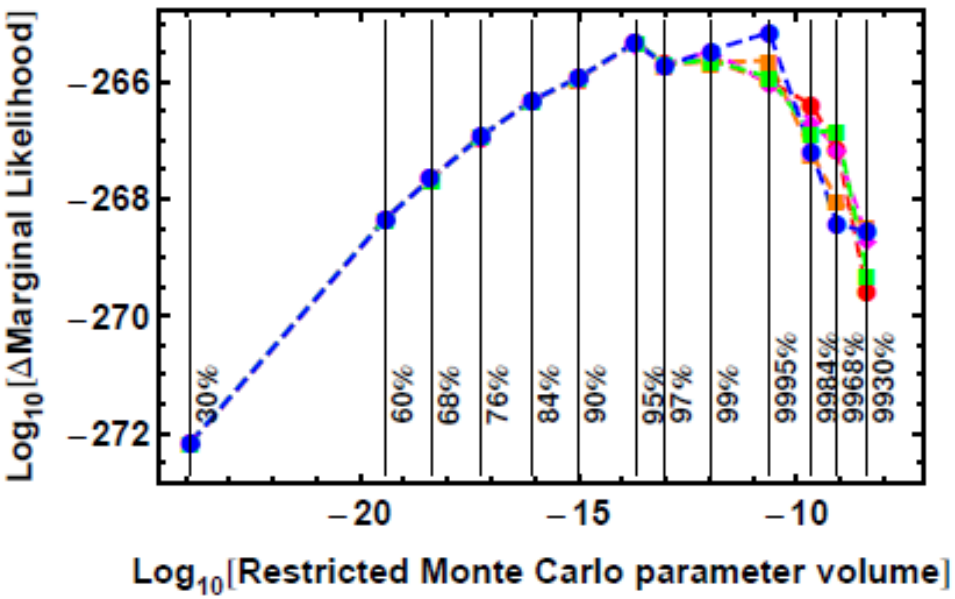
NRMC Integration HD 208487 (2 planet)

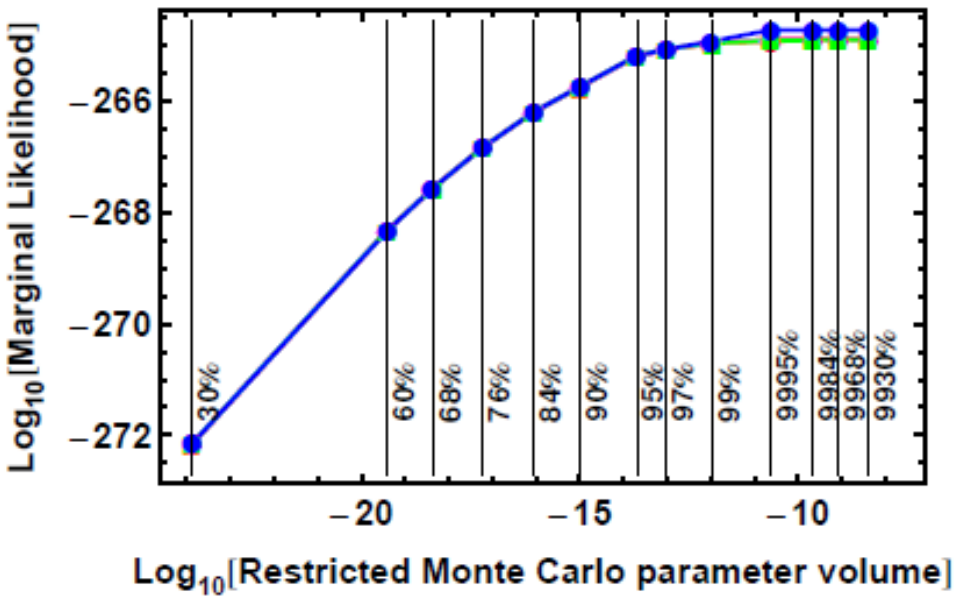




The left panel shows the contributions from the individual intervals for 5 repeats of the NRMC evaluation for the 2 planet Gliese 581 model. The right panel shows the summation of the individual contributions versus the volume of the credible region.

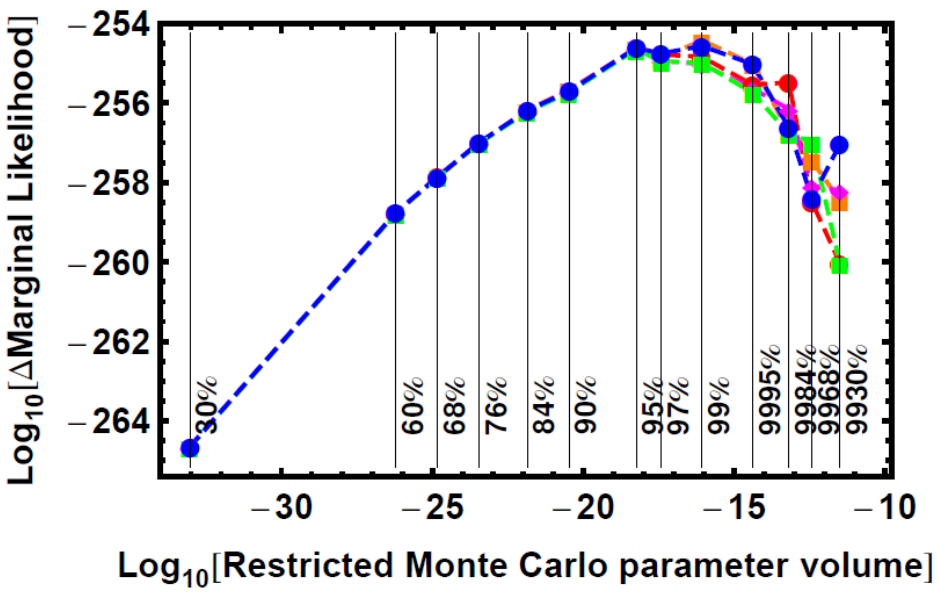
NRMC Integration Gliese 581 (3 planet)

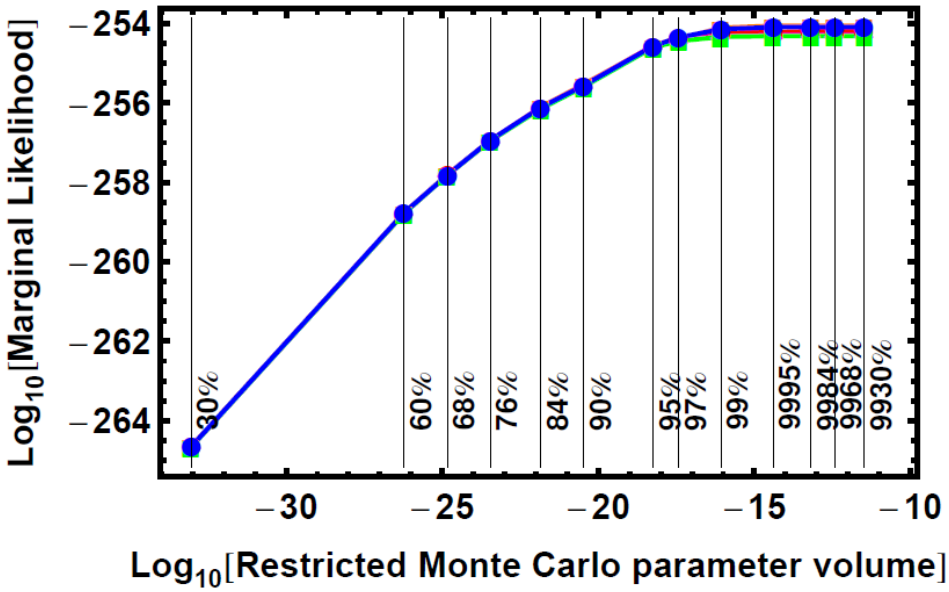




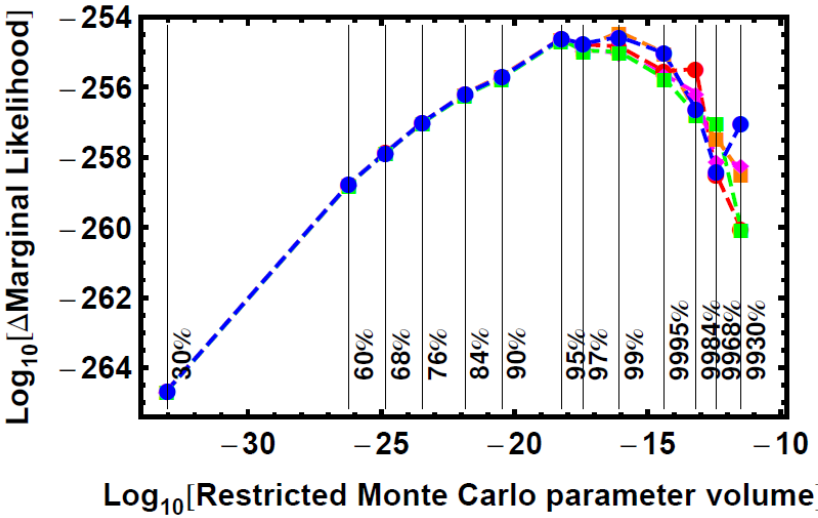
The left panel shows the contributions from the individual intervals for 5 repeats of the NRMC evaluation for the 3 planet Gliese 581 model. The right panel shows the summation of the individual contributions versus the volume of the credible region.

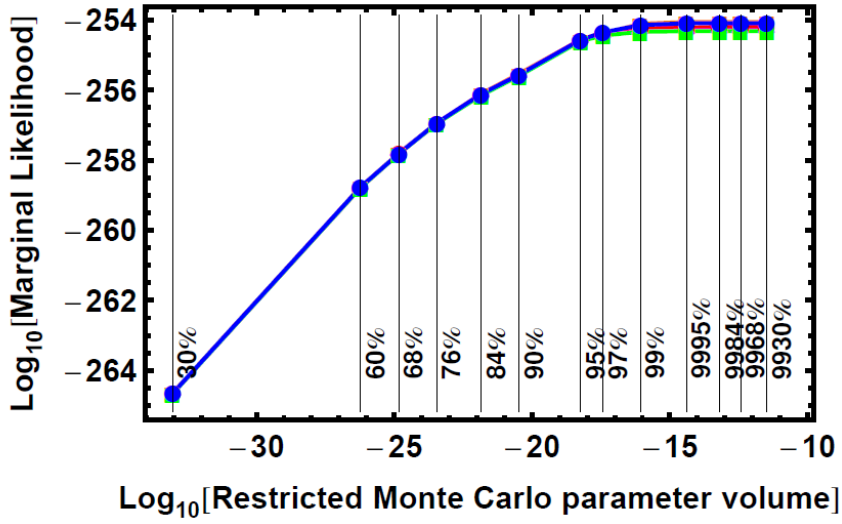
NRMC Integration Gliese 581 (4 planet)

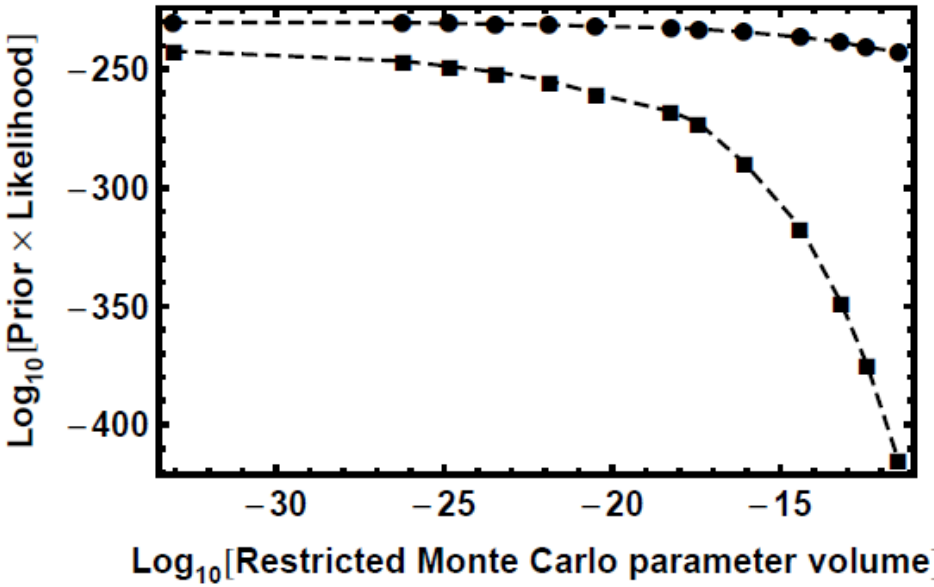


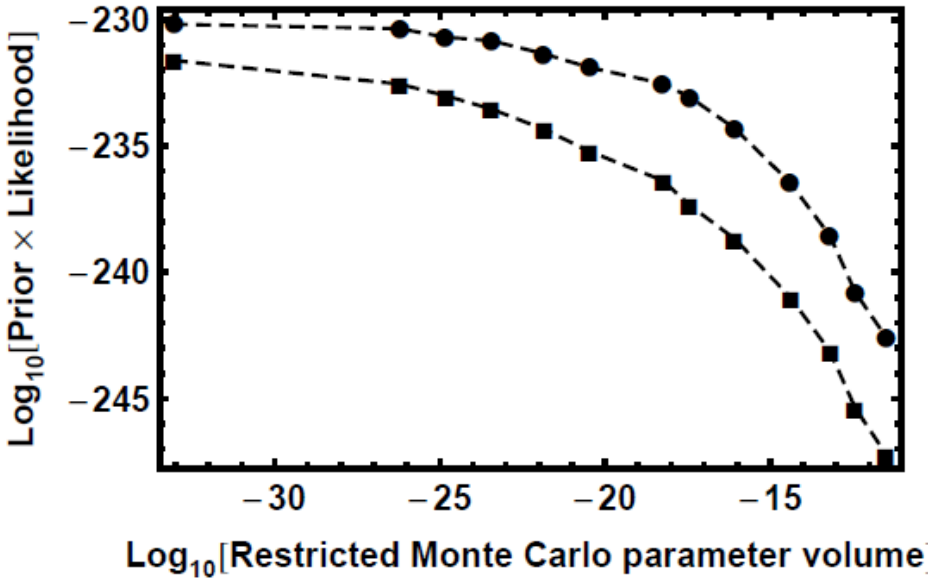


The left panel shows the contributions from the individual intervals for 5 repeats of the NRMC evaluation for the 4 planet Gliese 581 model. The right panel shows the summation of the individual contributions versus the volume of the credible region.



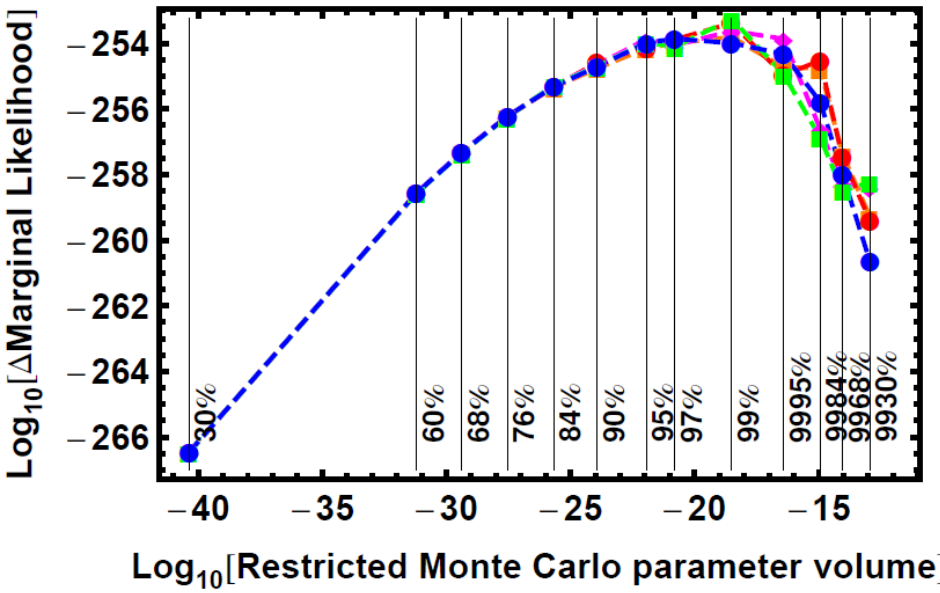


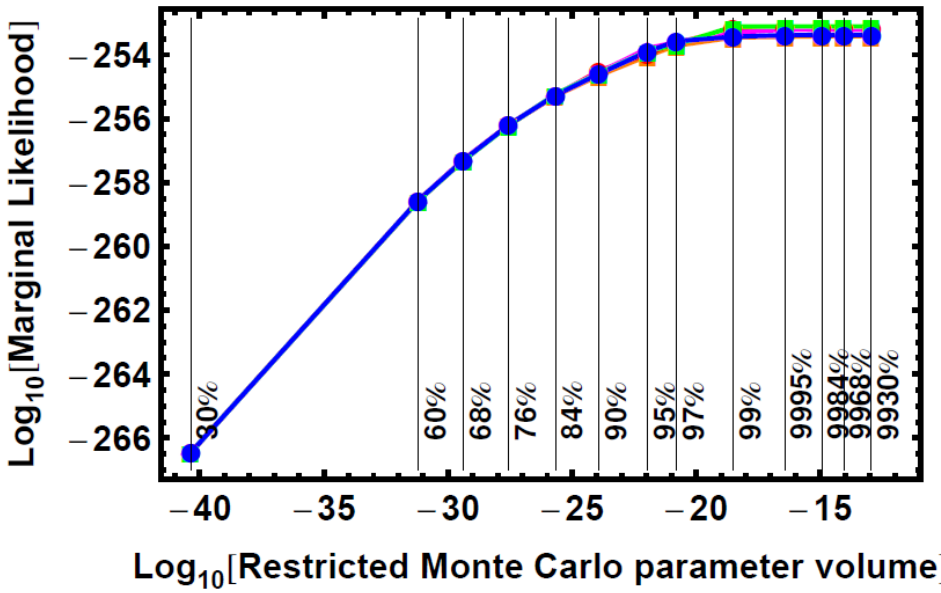




Left shows the maximum and min values of the Log10[prior × likelihood] for each interval of credible region versus parameter volume for the NRMC 4 planet fit samples. The right shows the maximum and mean values of the Log10[prior × likelihood] versus the parameter volume.

NRMC Integration Gliese 581 (5 planet)





Left panel: Contribution of the individual nested intervals to the NRMC marginal likelihood for 5 planet Gl 581 model for 5 repeats. The right panel: the integral of these contributions versus the parameter volume of the credible region.

return

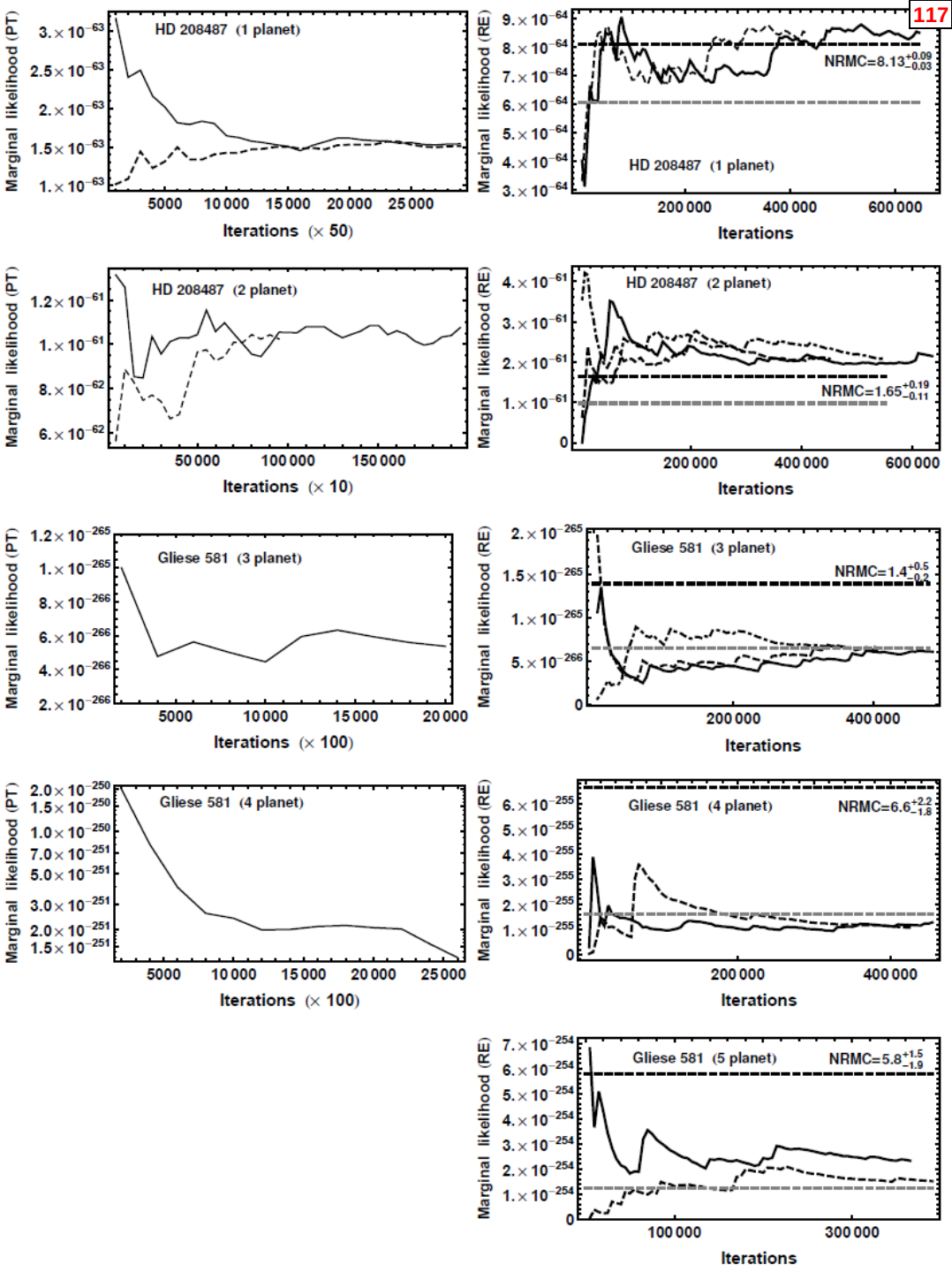
Comparisons of marginal likelihood estimators versus iteration for one to five planet model fits.

-The left hand column of plots show parallel tempering marginal likelihoods versus iteration number.

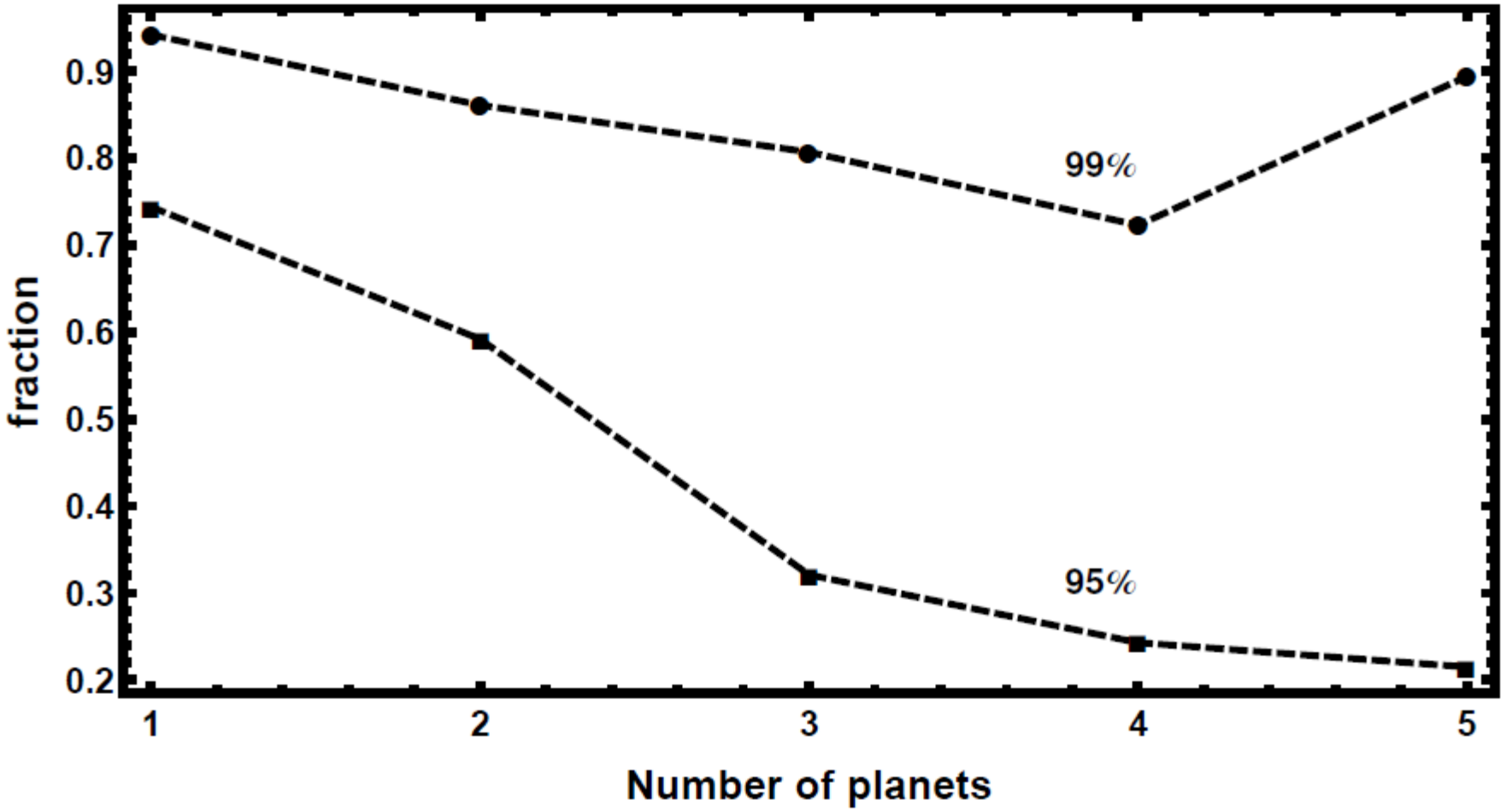
-The curves in the right hand column of panels show ratio estimator marginal likelihoods.

-The horizontal black dashed lines are the NRMC marginal likelihoods with the numerical value of the mean and range of 5 repeats.

-The horizontal gray dashed lines are the NRMC marginal likelihood value within the MCMC 95% credible region of the model parameters.



NRMC Integration Gliese 581



The fraction of the total NRMC marginal likelihood within the MCMC 95% and 99% credible regions versus the number of planets.

Comparison of the 3 marginal likelihood estimates

Model	RE	:	NRMC	:	PT
1 planet	1.0	:	0.96	:	1.82
2 planet	1.0	:	0.75	:	0.52
3 planet	1.0	:	2.22	:	0.94
4 planet	1.0	:	5.6		
5 planet	1.0	:	2.4		

Conclusions

1) For up to 3 planets (17 parameters) agreement within factor ~ 2.

2) The contribution to NRMC estimate from low probability density regions increase markedly with increasing model parameters.

3) There is an indication that the RE estimates and others directly derived from the MCMC posterior samples may be dynamic range limited by the number of iterations, leading to an under estimate in large parameter spaces.

4) NRMC method is conceptually simple and fast to compute.

Marginal Likelihood Ratio Estimator

Starting point is Bayes' theorem

$$p(\vec{X} \mid D, M_i, I) = \frac{p(\vec{X} \mid M_i, I)\, p(D \mid M_i, \vec{X}, I)}{p(D \mid M_i, I)}$$

Re-arrange and multiply through by additional sampling distribution $h(\vec{X})$

$$p(D \mid M_i, I)\, p(\vec{X} \mid D, M_i, I)\, h(\vec{X}) = p(\vec{X} \mid M_i, I)\, p(D \mid M_i, \vec{X}, I)\, h(\vec{X})$$

Integrate both sides over the prior range for X.

$$p(D \mid M_i, I)_{re} \int p(\vec{X} \mid D, M_i, I)\, h(\vec{X})\, d\vec{X} = \int p(\vec{X} \mid M_i, I)\, p(D \mid M_i, \vec{X}, I)\, h(\vec{X})\, d\vec{X}$$

The ratio estimator of the marginal likelihood, which we designate by $p(D|M_i,I)_{re}$, is given by

$$p(D \mid M_i, I)_{re} = \frac{\int p(\vec{X} \mid M_i, I)\, p(D \mid M_i, \vec{X}, I)\, h(\vec{X})\, d\vec{X}}{\int p(\vec{X} \mid D, M_i, I)\, h(\vec{X})\, d\vec{X}}$$

Marginal Likelihood Ratio Estimator

The ratio estimator of the marginal likelihood, which we designate by $p(D|M_i,I)_{re}$, is given by

$$p(D \mid M_i, I)_{re} = \frac{\int p(\vec{X} \mid M_i, I)\, p(D \mid M_i, \vec{X}, I)\, h(\vec{X})\, d\vec{X}}{\int p(\vec{X} \mid D, M_i, I)\, h(\vec{X})\, d\vec{X}}$$

The arbitrary function $h(\vec{X})$ was set equal to a multivariate normal with a covariance matrix equal to twice the covariance matrix computed from a sub-sample of the β = 1 MCMC draws.

Interpret the numerator as the weighted average of the prior x Likelihood, weighted by $h(\vec{X})$. Similarly, interpret the denominator as the weighted average of $h(\vec{X})$ weighted by the posterior.

To obtain the marginal likelihood ratio estimator, $p(D|M_i,I)_{re}$, we approximate the numerator by drawing samples $\tilde{\vec{X}}_1, \ldots, \tilde{\vec{X}}_{n_s'}$ from $h(\vec{X})$ and approximate the denominator by drawing samples $\vec{X}_1, \ldots, \vec{X}_{n_s}$ from the remainder of the MCMC draws.

$$p(D \mid M_i, I)_{re} \approx \frac{\frac{1}{n_s'} \sum_{j=1}^{n_s'} p(\tilde{X}_j \mid M_i, I)\, p(D \mid M_i, \tilde{X}_j, I)}{\frac{1}{n_s} \sum_{j=1}^{n_s} h(\vec{X}_j)}$$

Marginal Likelihood Ratio Estimator

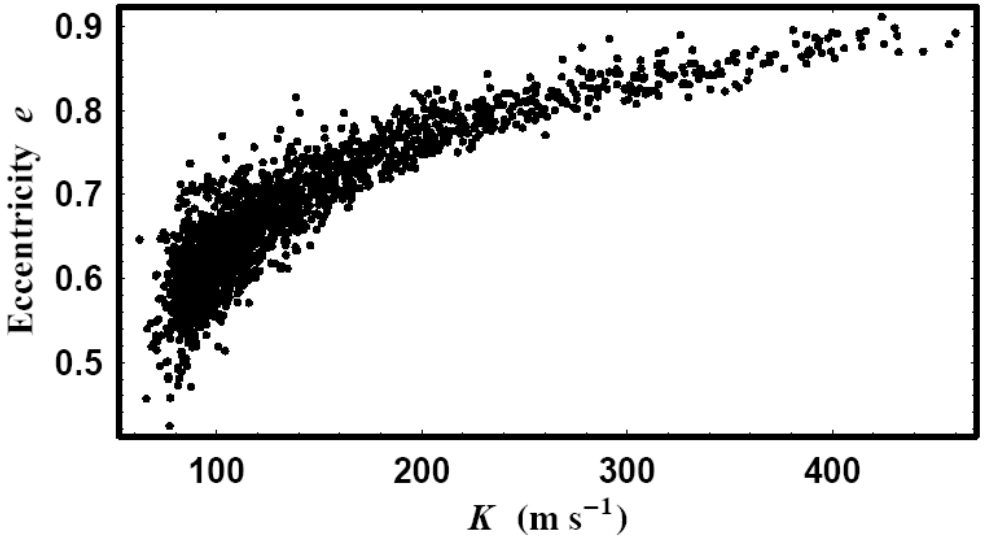
$$p(D \mid M_i, I)_{re} = \frac{\sum_{j=1}^{n_s'} p(\tilde{X}_j \mid M_i, I)\, p(D \mid M_i, \tilde{X}_j, I)}{\sum_{j=1}^{n_s} h(\vec{X}_j)}$$

This estimator is particularly useful, since there is no risk of a small denominator leading to a large variance, as in importance sampling.

Detail: Some of the samples from a multivariate normal $h(\vec{X})$ can have nonphysical parameter values (e.g. K < 0). Rejecting all nonphysical samples corresponds to sampling from a truncated multivariate normal. The factor required to normalize the truncated multivariate normal is just the ratio of the total number of samples from the full multivariate normal to the number of physical valid samples. Of course we need to use the same truncated multivariate normal in the denominator so the normalization factor cancels.

Improved Marginal Likelihood Ratio Estimator

For some data sets the posterior samples are not well modeled by a single multivariate normal.



Better to replace $h(\vec{X})$ **by** $g(\vec{X})$ **a mixture of multivariate normals, where**

$$g(\vec{X}) = \sum_{k=1}^{n_c} g_k(\vec{X})$$

We randomly choose n_c = 100-150 samples to be removed from the original MCMC posterior draws to be used as the locations for the mixture components. We also determine the covariance matrix for each mixture component using another sub-sample of the posterior draws.

$$p(D \mid M_i, I)_{re} = \frac{\sum_{j=1}^{n_s}{}' p(\tilde{X}_j \mid M_i, I)\, p(D \mid M_i, \tilde{X}_j, I)}{\sum_{j=1}^{n_s} g(\vec{X}_j)}$$

Bayesian false alarm probability

In the context of claiming the detection of *m* planets the false alarm probability (FAP_m) is the probability that there are actually fewer than *m* planets, i.e., *m* − 1 or less.

$$\mathrm{FAP}_m = \sum_{i=0}^{m-1} (\text{prob.of i planets})$$

If we assume *a priori* that all models under consideration are equally likely, then the probability of each model is related to the Bayes factors by

$$p(M_i \mid D, I) = \frac{B_{i1}}{\sum_{j=0}^{N} B_{j1}}$$

where *N* is the maximum number of planets in the hypothesis space under consideration, and of course *B*11 = 1. For the purpose of computing FAP_m we set *N* = *m*.

For *m* = 2

$$\mathrm{FAP}_2 = \frac{(B_{01} + B_{11})}{\sum_{j=0}^{2} B_{j1}}$$

Model Selection HD 208487

Model	Periods (d)	Marginal Likelihood	Bayes factor nominal	False Alarm Probability
M_0		1.44×10^{-68}	1.77×10^{-5}	
M_1	(130)	$(8.13^{+0.09}_{-0.03}) \times 10^{-64}$	1	1.4×10^{-4}
M_{2a}	(29, 130)	$(1.83^{+0.05}_{-0.03}) \times 10^{-62}$	22.5	0.90
M_{2b}	(130, 900)	$(1.65^{+0.19}_{-0.11}) \times 10^{-61}$	203	0.10
M_2	(29, 130) or (130, 900)	$(1.83^{+0.19}_{-0.11}) \times 10^{-61}$	225	4.4×10^{-3}

$$\mathrm{FAP}_2 = \frac{(B_{01} + B_{11})}{(B_{01} + B_{11} + B_{2a1} + B_{2b1})} = 4.4 \times 10^{-3}$$

Model Selection Gliese 581

Model	Periods (d)	Marginal Likelihood	Bayes factor nominal	False Alarm Probability
M_0		5.32×10^{-393}	7.9×10^{-139}	
M_1	(5.37)	$(1.45 \pm 0.004) \times 10^{-295}$	2.2×10^{-41}	3.7×10^{-98}
M_2	(5.37, 12.9)	$(5.55^{+0.26}_{-0.09}) \times 10^{-273}$	2.6×10^{-19}	2.6×10^{-23}
M_3	(5.37, 12.9, 66.9)	$(1.40^{+0.5}_{-0.15}) \times 10^{-265}$	2.1×10^{-11}	3.9×10^{-8}
M_4	(3.15, 5.37, 12.9, 66.9)	$(6.7^{+2.2}_{-1.8}) \times 10^{-255}$	1.0	2.1×10^{-11}
M_{5a}	(3.15, 5.37, 12.9, 66.9, 192)	$(5.8^{+1.5}_{-1.9}) \times 10^{-254}$	8.7	0.19
M_{5b}	(3.15, 5.37, 12.9, 66.9,not 192)	$(0.7^{+0.18}_{-0.22}) \times 10^{-254}$	1.0	0.90
M_5	(3.15, 5.37, 12.9, 66.9,all)	$(6.5^{+1.7}_{-2.1}) \times 10^{-254}$	9.7	0.093
M_6	(3.15, 5.37, 12.9, 66.9, 71, 190)	$(5.21^{+1.5}_{-1.7}) \times 10^{-252}$	778	0.014

Results for M5b are based on the ratio of post burn-in samples that were not in the 192d peak to the samples in the 192 d peak.

Detecting extrasolar planets from stellar radial velocities Using Bayesian evidence. (MNRAS, 415, 3462, 2011)

F. Feroz⋆, S. T. Balan and M. P. Hobson
Astrophysics Group, Cavendish Laboratory

They have analyzed RV data using a MULTINEST version of John Skilling's nested sampling algorithm.

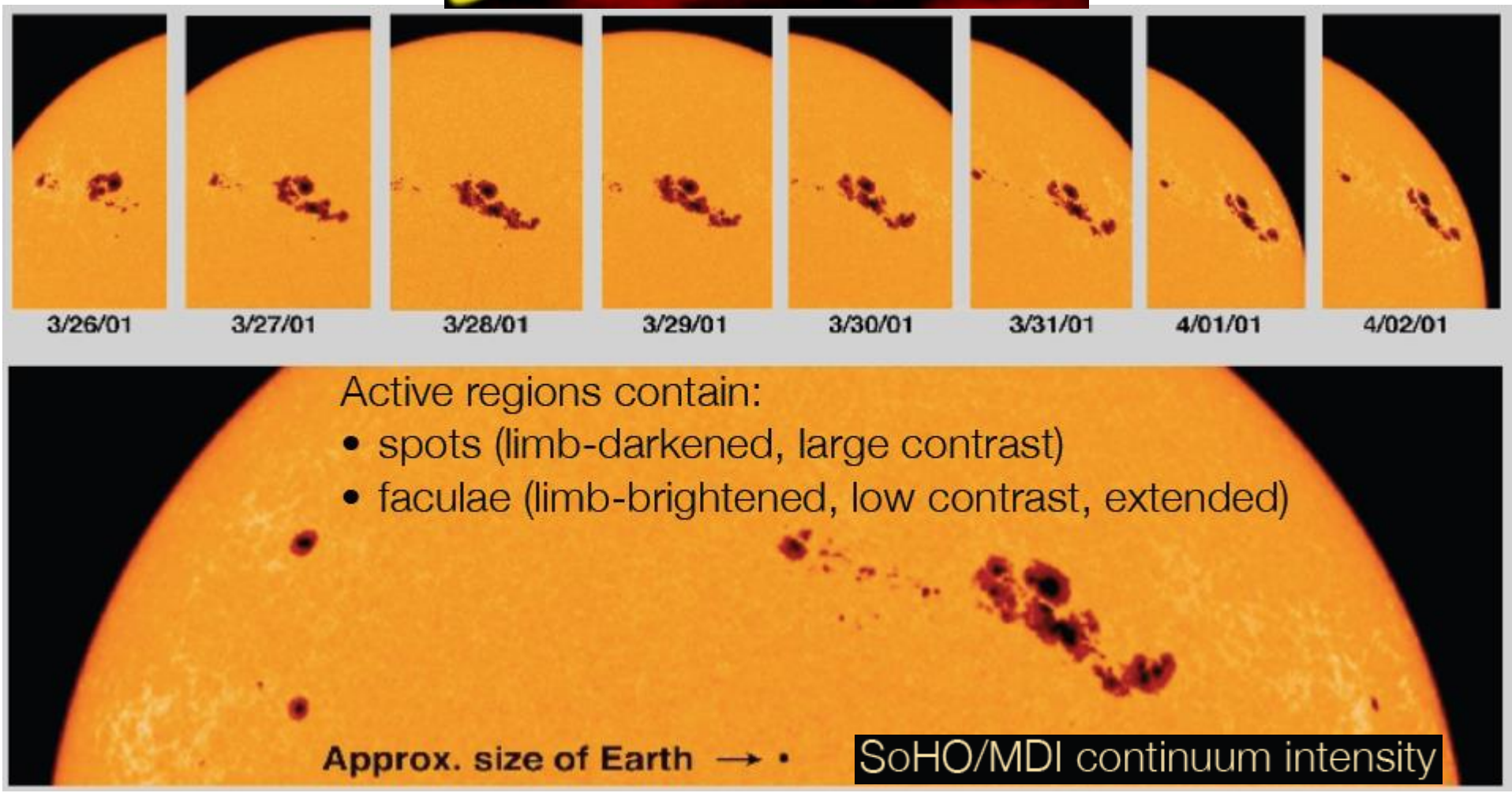
They found that MultiNest is not able to calculate the evidence values directly for systems with more than 3 planets.

Instead they base their model selection calculations for the n planet versus n+1 planet model on zero and one planet model fits to the n planet residuals. These results must necessarily be considered only approximate because in general inclusion of an additional planet (to n+1) will effect the best fit parameters for the n planet subset which will not be reflected in the n planet residuals.

Stay tuned – I hear new developments are underway

Stellar activity
Induced RV

In this section I present results of a preliminary investigation aimed at identifying and/or correcting for stellar activity induced RV signals



Sunspots rotate with the Sun (every 27 days).
They typical last from days to months.

SOHO is a mission of international cooperation between ESA and NASA

Stellar activity

Time Scale	Vel. noise	Type of activity	Partial solutions
~ 10 years	1 – 20 m/s	Magnetic cycle	correlation
10 – 50 d	few m/s	Active regions spots and plages	a) correlation b) FF' analysis + Gaussian process
15 min – 2 d	few m/s	Granulations	ave. 3x10 min/night reduce to ~ 0.5 m/s
~ 1 hr	< 1 m/s	Flares	
< 15 min	few m/s	Oscillations	ave. for 15 min reduce to ~ 0.2 m/s

Spectral line diagnostics for RV measurements (D. Queloz et al., 2009 A&A 506, 303)

Besides accurate RV measurements, HARPS provides additional information on the spectral line shapes that are extracted from the CCF, the average shape of all spectral lines of the star.

Any changes in the CCF that are correlated with the RV can be attributed to pulsation effects or stellar spots both of which affect the shapes and thus the centroids of the spectral line.

Two simple diagnostics used to look for changes in the shape of the CCF: (a) width (FWHM) and (b) its bisector span.

HARPS' wavelength coverage includes the Ca II H&K lines, so the activity S-index is computed as well. The spectroscopic S-index is sensitive to active regions on the stellar surface. The Ca II H& K flux that is converted to the Mount Wilson system according to Santos et al. (2000)and corrected for the photospheric flux is known as log(R(HK)) index.

In summary, HARPS provides three diagnostic measurements of stellar variability: (a) the bisector span of the CCF, (b) the FWHM of the CCF, and (c) the Ca II S-index.

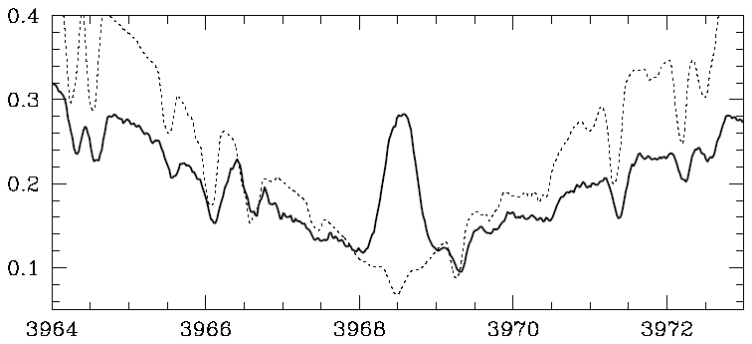
If star's velocity is due to a planetary companion, then the shape of the CCF should remain constant. On the other hand, RV variations caused by stellar variability should correlate with changes in the spectral line shapes and/or Ca II.

return

No planet for HD 166435

A&A 379, 279, 2001

D. Queloz, G. W. Henry, J. P. Sivan, S. L. Baliunas, J. L. Beuzit, R. A. Donahue, M. Mayor, D. Naef, C. Perrier, and S. Udry



An emission feature is clearly visible in the core of the Ca II H line. For comparison, a solar integrated spectrum with the same resolution has been superimposed (hatched line).
An emission feature is clearly visible in the core of the line.
Note: larger line broadening of HD 166435 compared to the Sun.

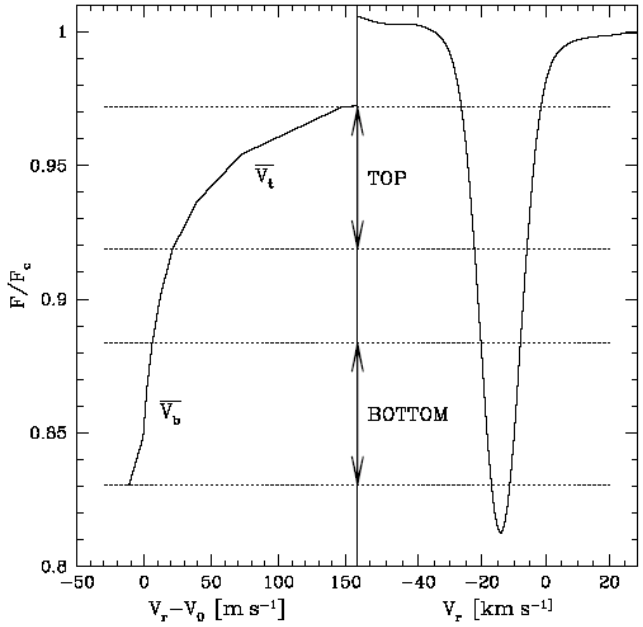


Fig. 5. Right: the mean CCF function of HD 166435's spectra constructed with a template selecting only the weak and non-saturated lines. This profile represents the mean spectral-line profile of the lines selected by the template. Left: the bisector of the CCF. V_0 is an arbitrary offset. Note the definition of the boundaries for the computation of ($\overline{V_t}$ and $\overline{V_b}$).

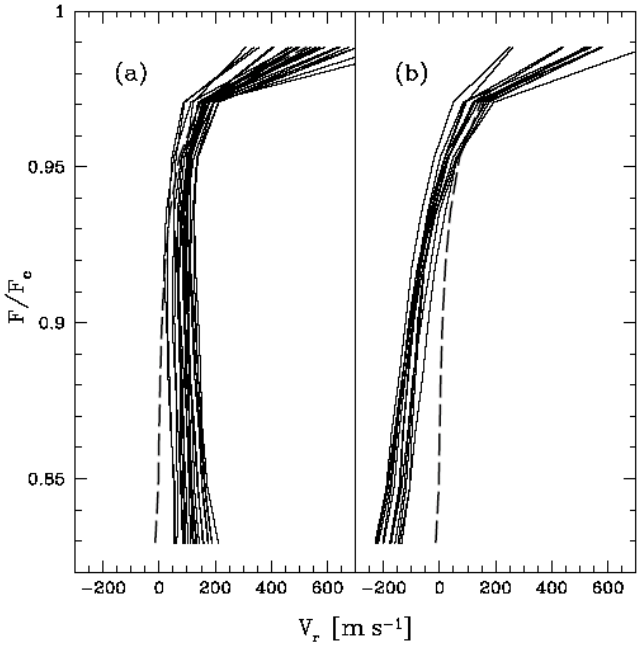


Fig. 6. Individual bisectors for two sets of spectra selected at opposite phases of the radial-velocity cycle. **a)** Spectra measured at $\phi = 0.0 \pm 0.1$. **b)** Spectra measured at $\phi = 0.5 \pm 0.1$. The hatched line illustrates the mean bisector computed by averaging all spectra.

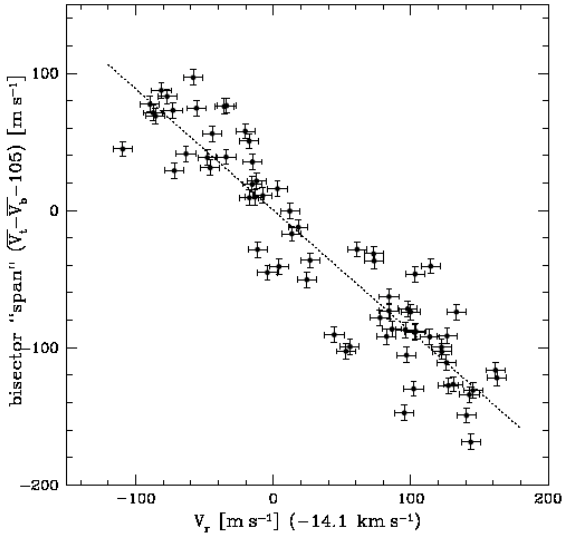


Fig. 7. Radial velocity of each CCF versus the bisector span ($\overline{V_t} - \overline{V_b}$) of the CCF profile. The dotted line is the best linear fit to the data.

No planet for HD 166435

D. Queloz, G. W. Henry, J. P. Sivan, S. L. Baliunas, J. L. Beuzit, R. A. Donahue, M. Mayor, D. Naef, C. Perrier, and S. Udry

A&A 379, 279, 2001

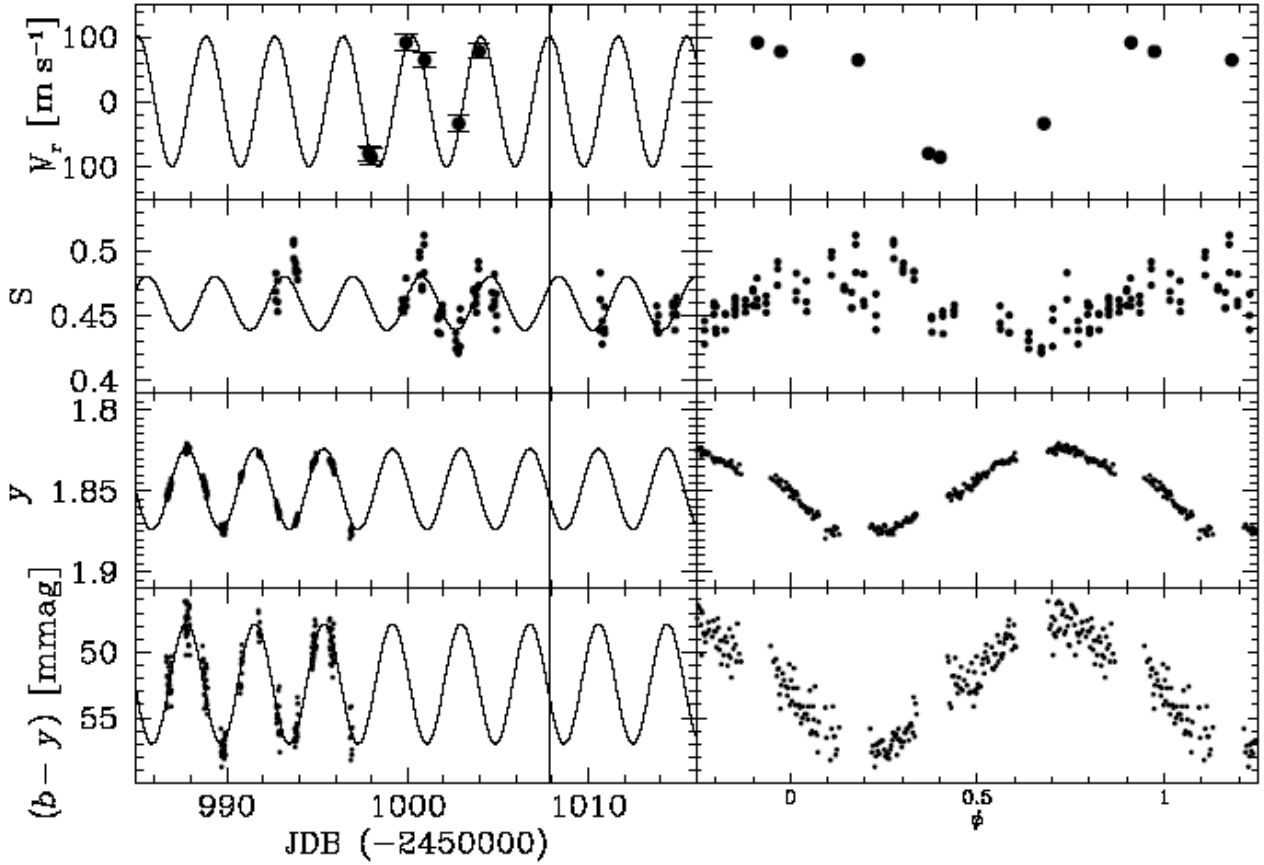


Fig. 8. Left: simultaneous observations of (from the top) radial velocity, S index, delta y magnitude, and delta $(b-y)$ color, of HD 166435 over a time span of 30 days. A best-fit sine-curve with a period fixed at 3.798 days is shown. A vertical line is drawn at an epoch of maximum radial velocity to help visualize the phase offsets between data sets. Right: same data but phase folded with $P = 3.798\,\mathrm{d}$ and $T_0 = 2\,450\,996.5$.

Gliese 667C isolated M dwarf ($M = 0.31\ M_S$)

component of triple star system D = 22.1 ly

History

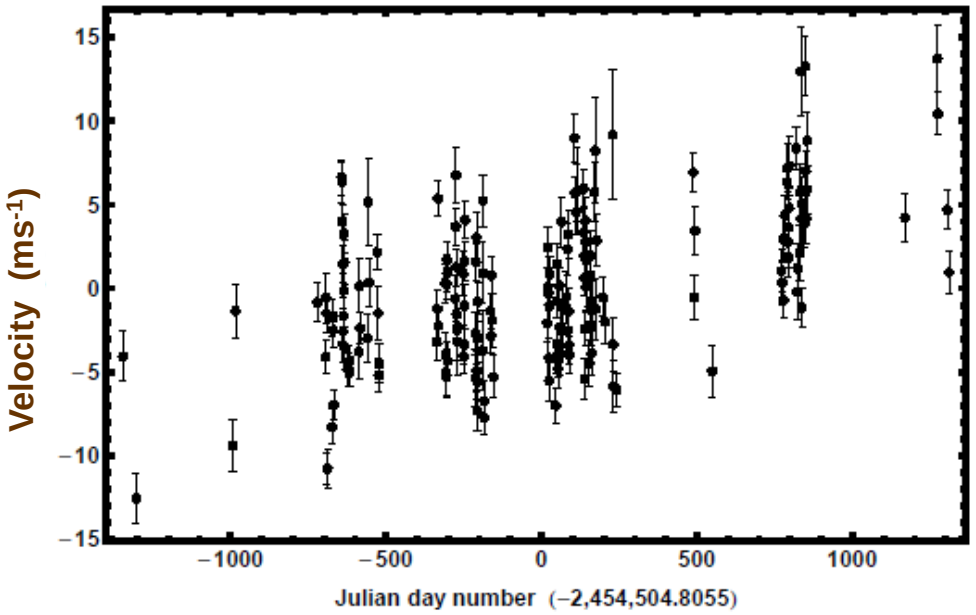
1) 2011, Planet b Per = 7.2 d
M sin i = 5.9 M_{Earth}
+ two other interesting periods at 90 & 28d (habitable zone orbit)
Bonfils et al., arXiv:1111.5019v2

2) 2012, Anglada et al. & Delfosse et al.,
Confirm planet b & report
planet c 28d, 4.3 M_{Earth} (HZ)

3) 2012, Gregory, P. C. arXiv: 1212.4058V2

"Additional Keplerian signals - - -"
Periods = 7.2, 28, 31, 39, 53, 91.5 d

HARPS data



Likely explanation of the slope is the orbital motion (P ~ 3900 yr) of star about the CM of the Gl 667ABC triple system.

4) 2013, Anglada-Escuda et al. Ap.J. 751, L16 (HARPS, PFS, HIRES data)

"A dynamically-packed planetary system around GJ 667C with 3 super-earths in its habitable zone,"
Additional P = 39, 62, 92 , 260 d + tantalizing evidence for a ~ $1M_E$ with P = 17 d

Exoplanet RV analysis

RV data

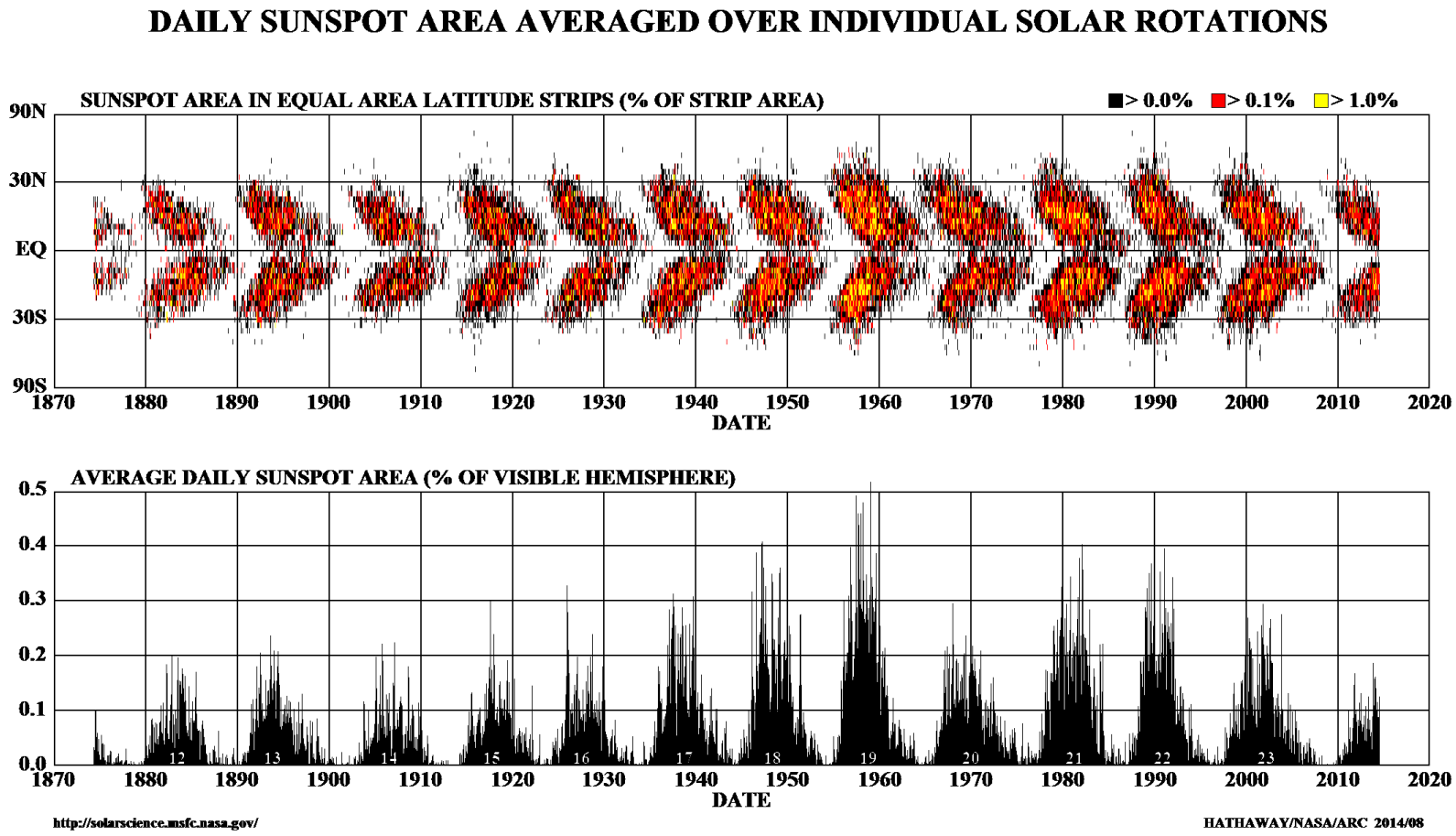
Multi-planet
Kepler periodograms

Diagnostic data on stellar activity

FWHM of CCF S index Bisector span Hα

-Lomb-Scargle periodogram
-Generalized Lomb-Scargle
(floating offset and weights)

} single sine wave model



Periodogram based on an apodized sine model

$$y = A \exp\left[-\frac{(t - t_1)^2}{2\,{\tau_1}^2}\right] \cos[2\pi f t + \theta]$$

Amplitude drops to $0.606\,A \;\; at \;\; (t_1 - \tau_1) \;\; \& \;\; (t_1 + \tau_1)$

Exoplanet diagnostic analysis

RV data

Multi-planet Kepler periodograms

-Lomb-Scargle periodogram
-Generalized Lomb-Scargle (floating offset and weights)

} single sine wave model

Explore other types of periodograms ⇒

-Multiple sine wave models
-Apodized sine wave models
-Apodized chirp models
-Combination of chirp + sine models

Diagnostic data on stellar activity

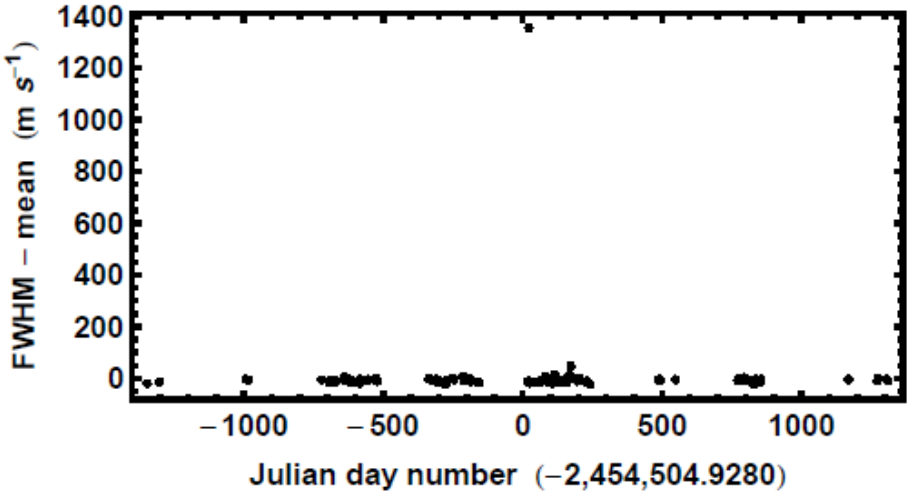
FWHM of CCF | S index | Bisector span | Hα

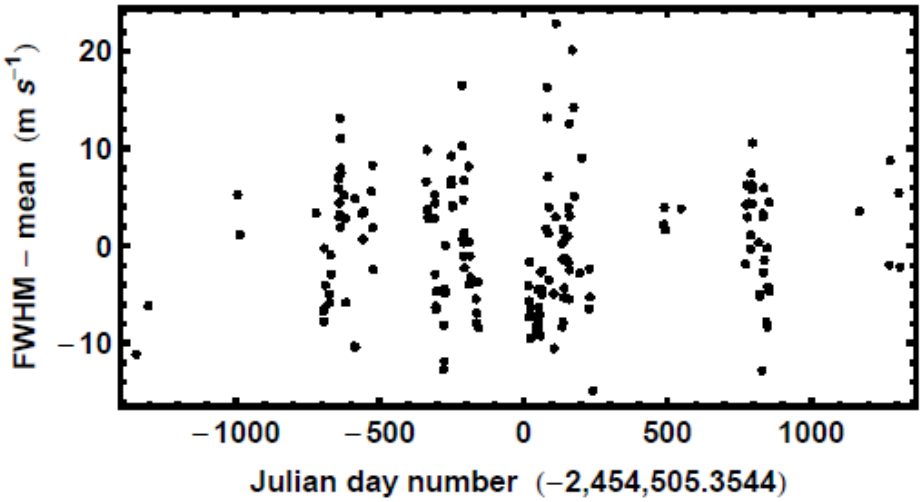
Correlation analysis between RV & diagnostics

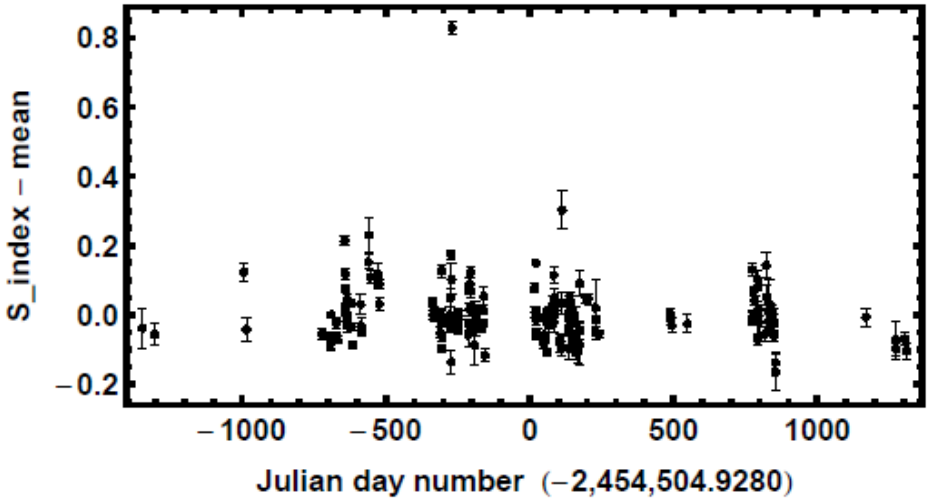
-Recent work by Paul Robertson et al. on Gliese 581 show seasonal dependent correlations between RV and Hα stellar activity. The artifact RV component removed by linear regression eliminated planet d. (Science 2014, arXiv 1407.1049R)

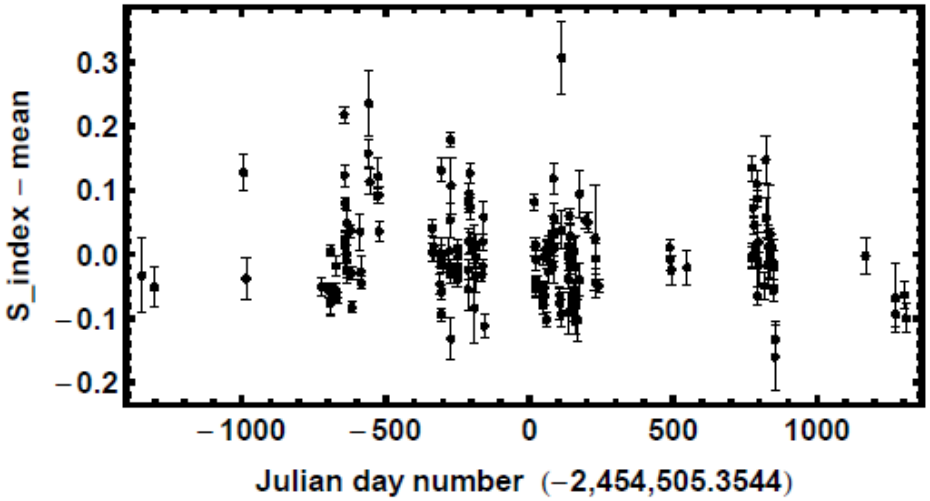
-I will show an example of this in Gliese 667C and demonstrate how a hierarchical (multi-level) Bayesian analysis can effectively deconvolve the effects of the measurement errors to better estimate the underlying regression relation.

FWHM and S-index diagnostics before and after removal of outliers









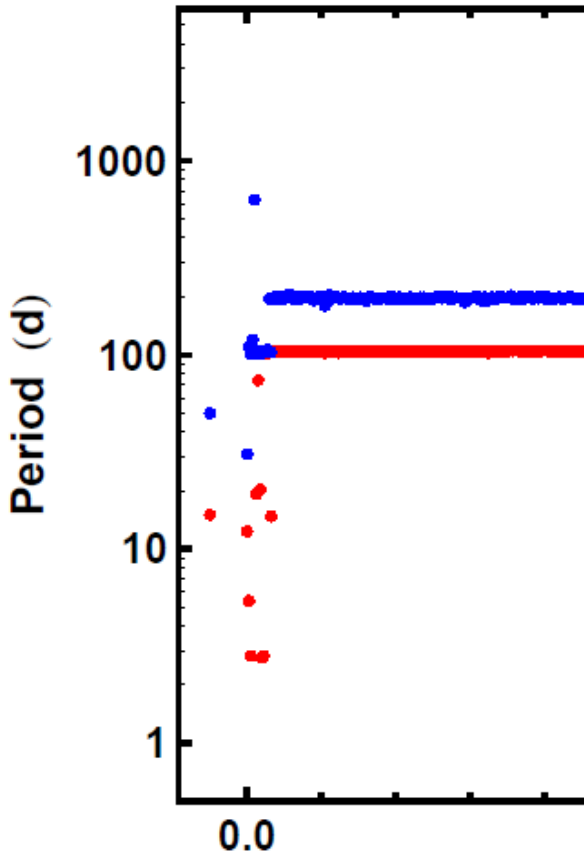
Relative probabilities via Bayes factors of different periodogram models fit to the FWHM of CCF diagnostic

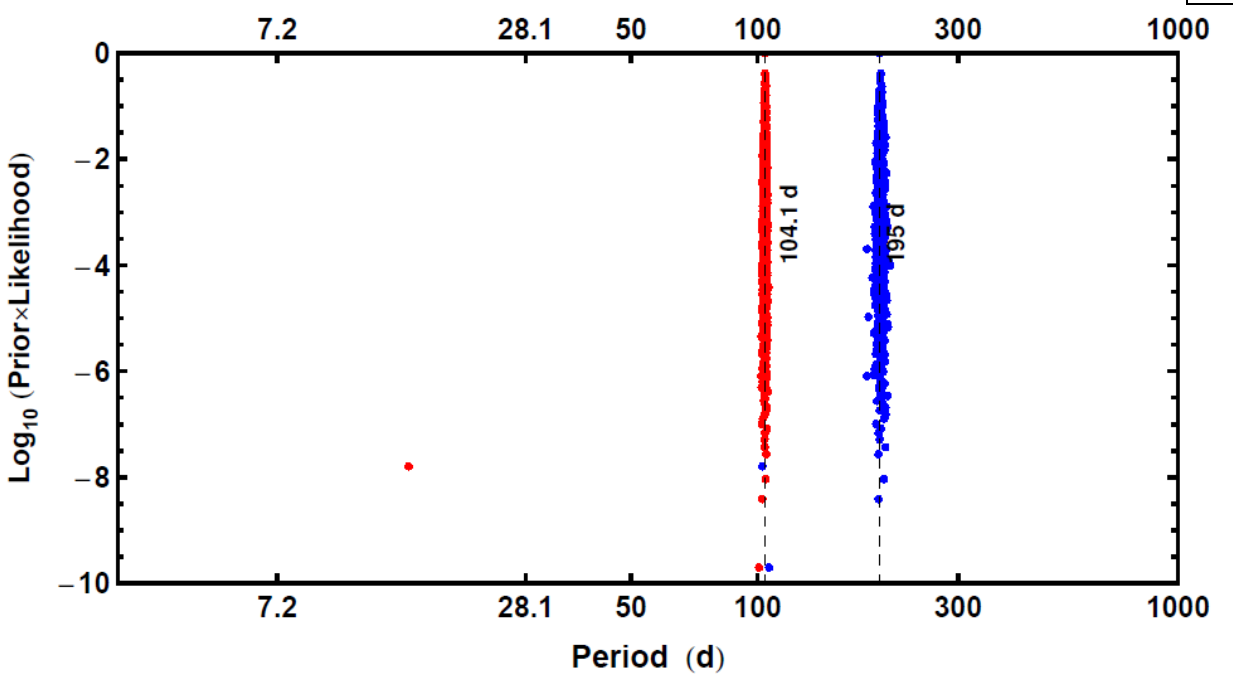
k – number of signals	sine waves	apodized sine waves	apodized chirp signals	1 chirped +1 sine (both apodized)
0	3.8×10^{-14}	3.8×10^{-14}	3.8×10^{-14}	3.8×10^{-14}
1	5.9×10^{-5}	3.7×10^{-4}	4.1×10^{-3}	
2	5.2×10^{-3}	1.0 [1]	0.15	0.22 [2]
3	1.5×10^{-2}	0.1 [4]	0.05 [3]	

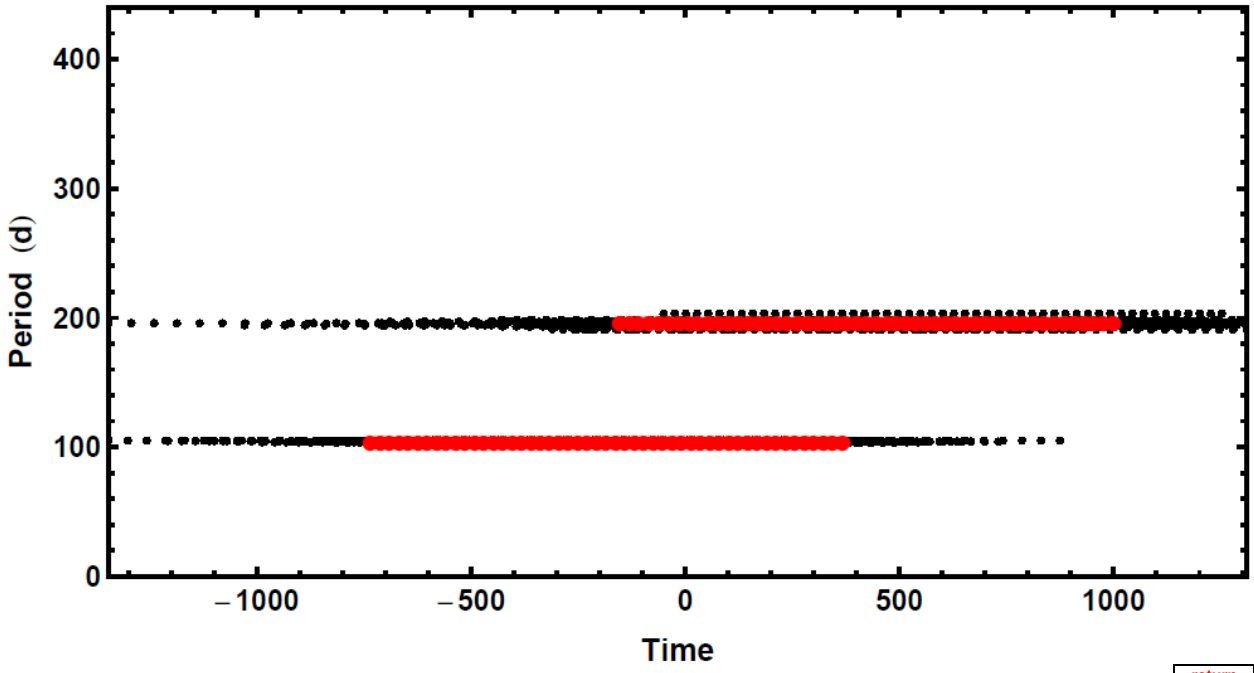
Relative probabilities via Bayes factors of different periodogram models fit to the S-index diagnostic

k – number of signals	sine waves	apodized sine waves	apodized chirp signals	
0	0.99	0.99		
1	1.0 [6]	2.0 [5]		

Two apodized sine periodogram of FWHM diagnostic

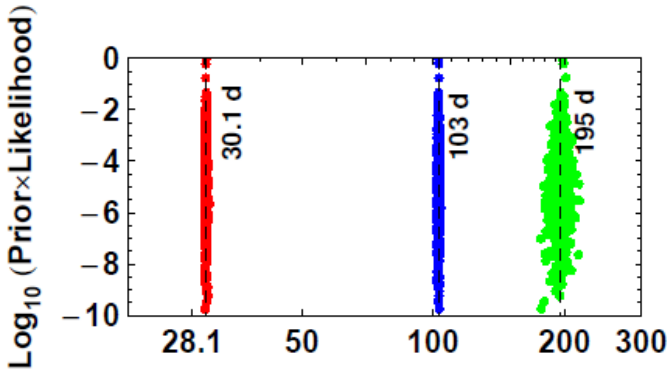


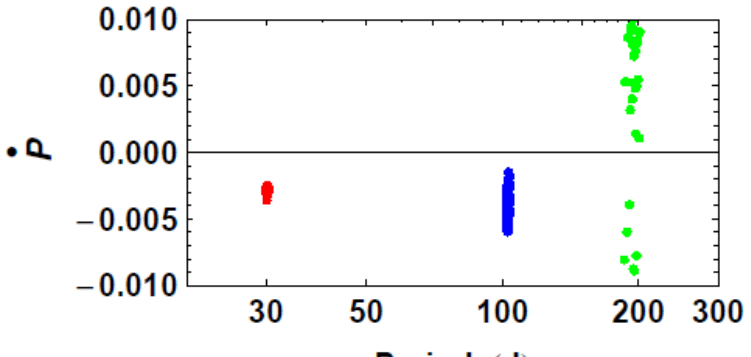




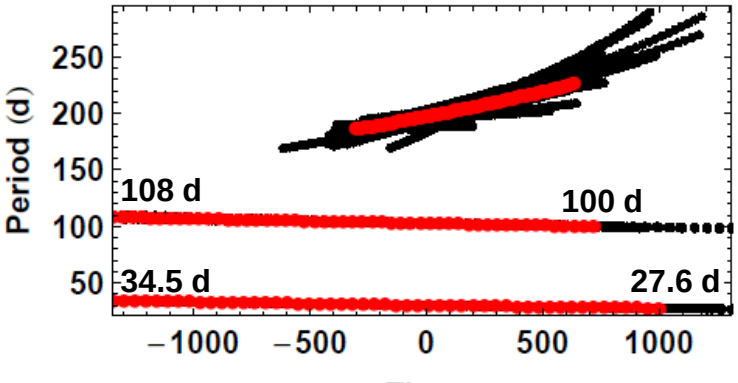
return

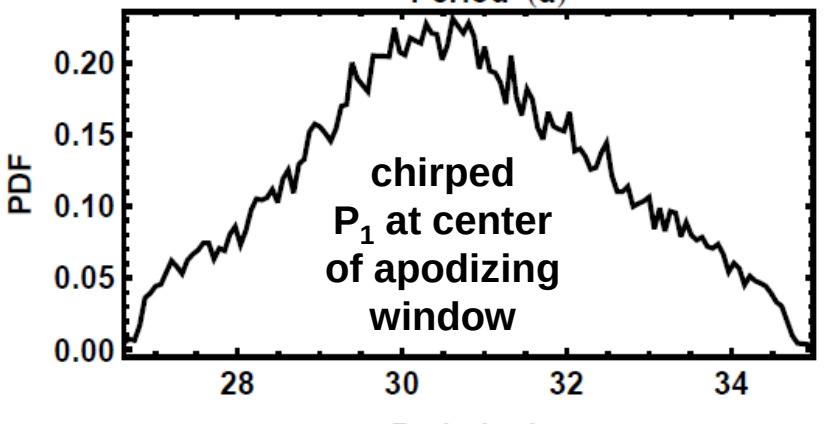
3 apodized chirp periodogram of FWHM diagnostic

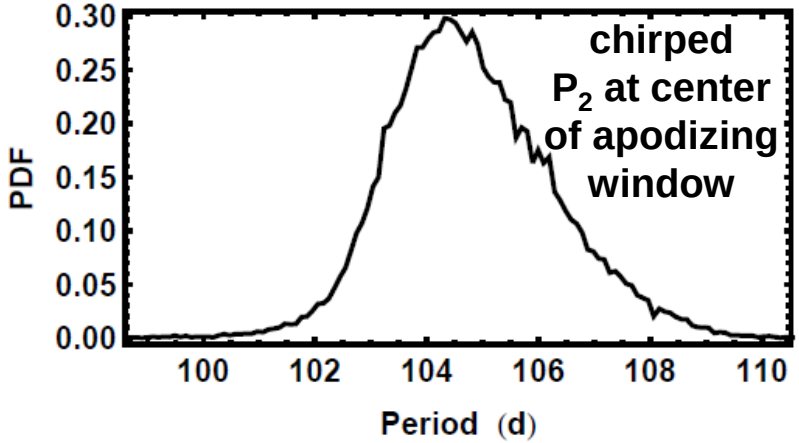


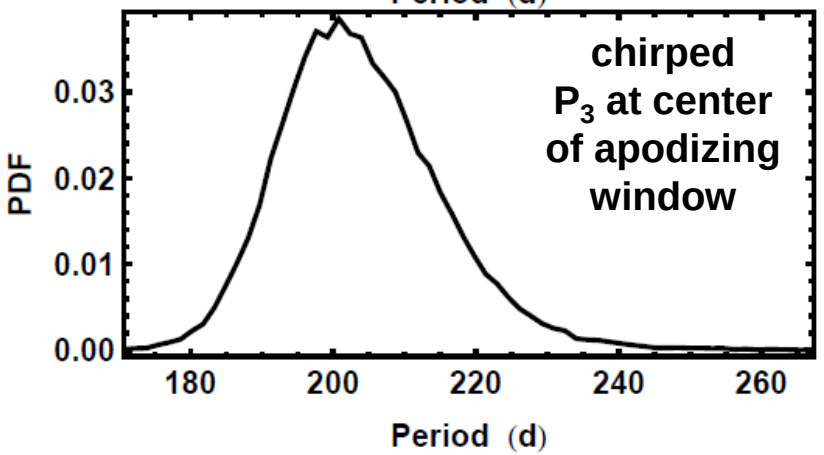


Chirped P at t = 0



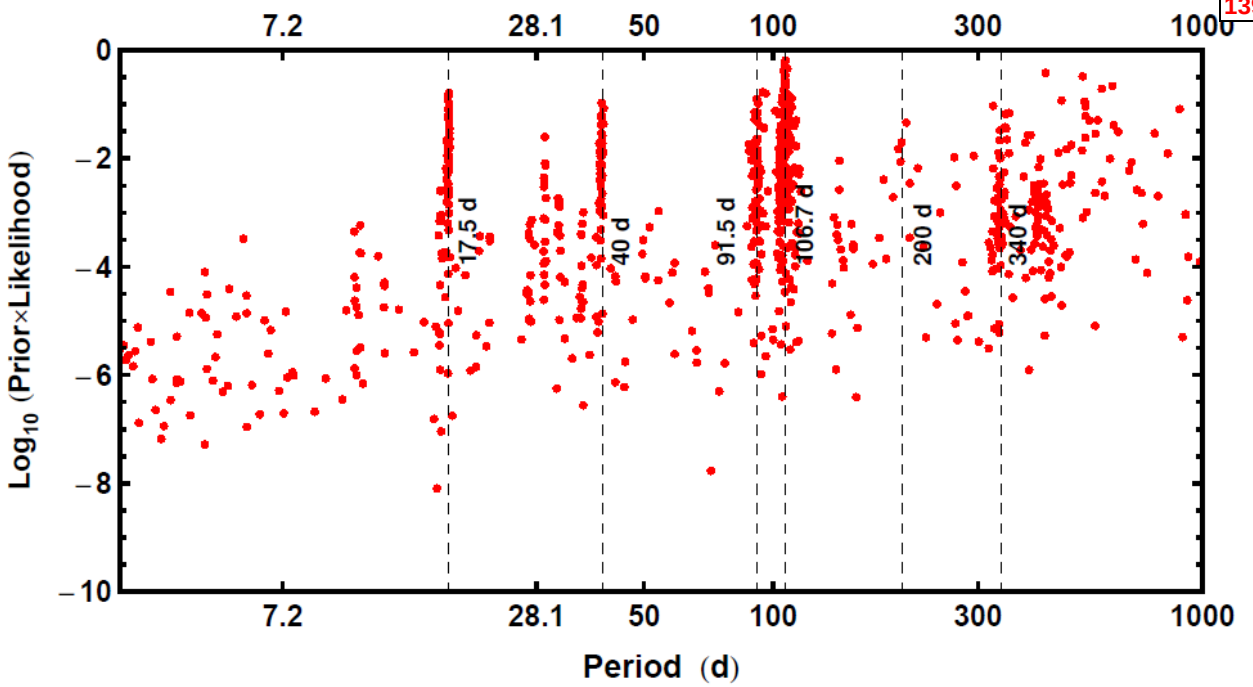


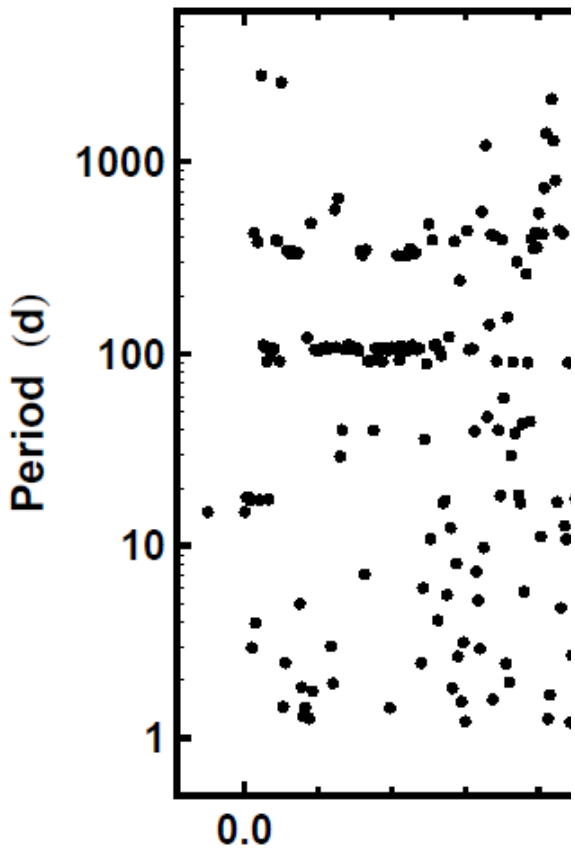


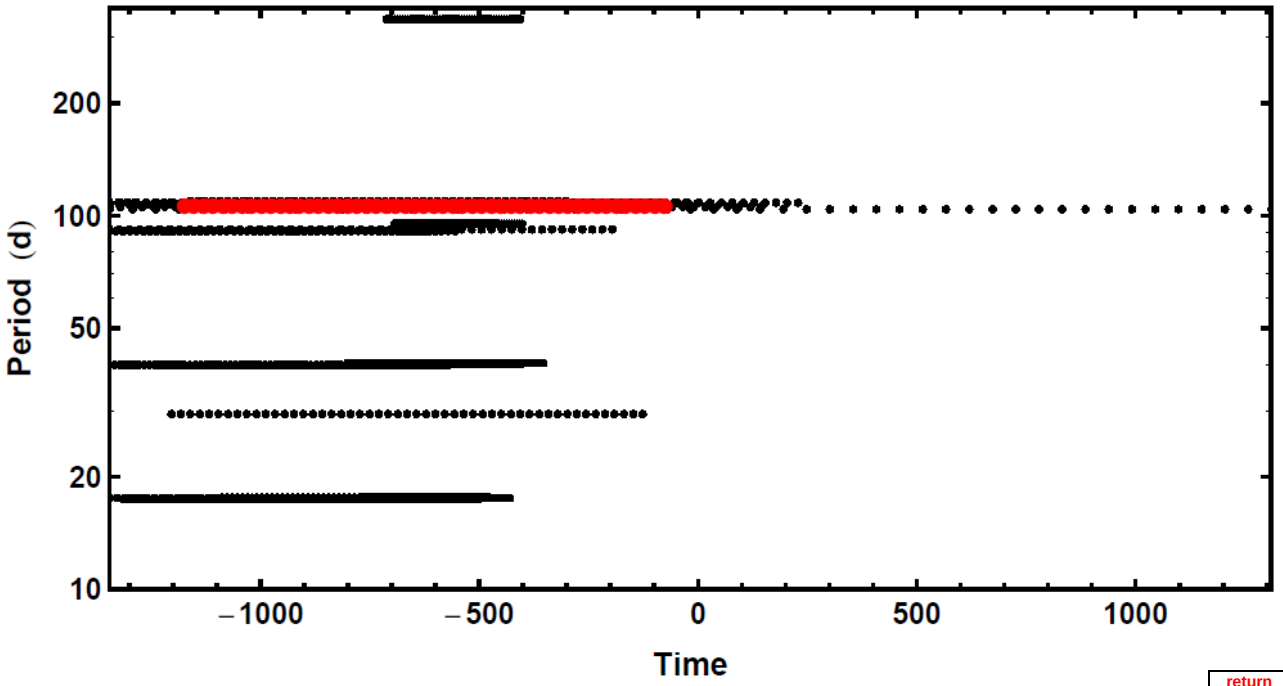


return

One apodized sine periodogram of S-index diagnostic







return

Part 1 Conclusions

For FWHM diagnostic

1) Apodized models are favored.

2) Strongest period around 105 d.
Thought to be the star's rotation period.

3) Chirped models yield period derivatives of the same sign and with a magnitude ≈ the maximum solar value for P = 34.5 → 27.6 d and 108 → 100 d.

For S-index diagnostic

1) Apodized model very slightly favored.

2) Strongest period around 106 d.

3) Other suspicious artifacts around 17.5, 40 & 91.5 d.

Exoplanet diagnostic analysis

RV data

Multi-planet Kepler periodograms

Explore other types of periodograms

Diagnostic data on stellar activity

FWHM of CCF S index Bisector span Hα

-Lomb-Scargle periodogram
-Generalized Lomb-Scargle (floating offset and weights)
} single sine wave model

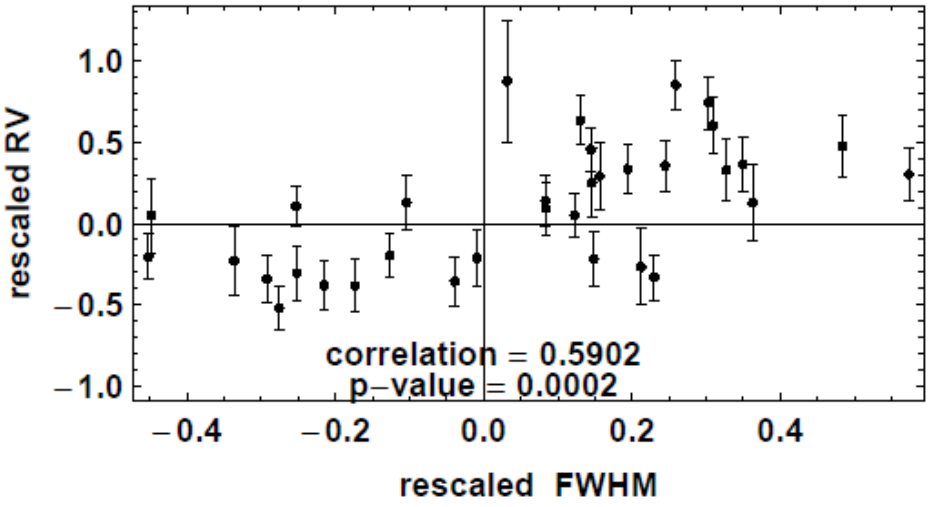
-Multiple sine wave models
-Apodized sine wave models
-Apodized chirp models
-Combination of chirp + sine models

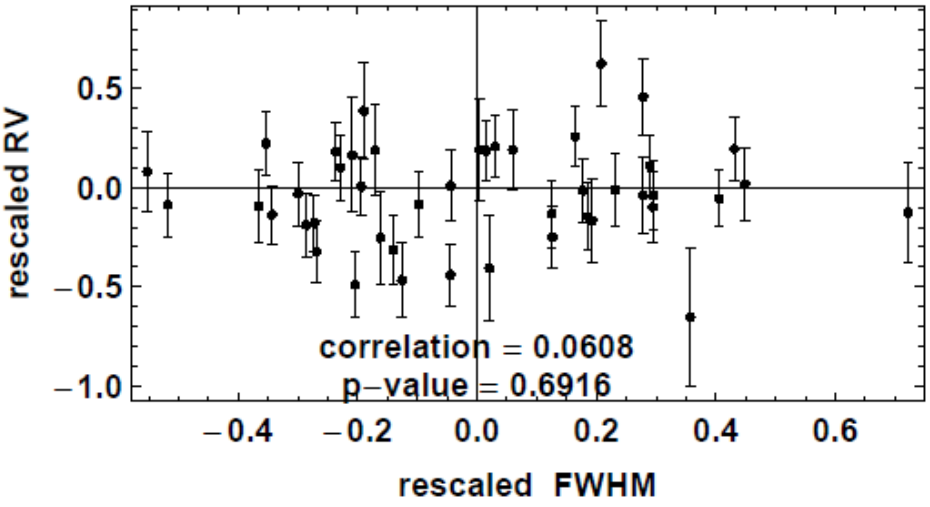
Correlation analysis between RV & diagnostics

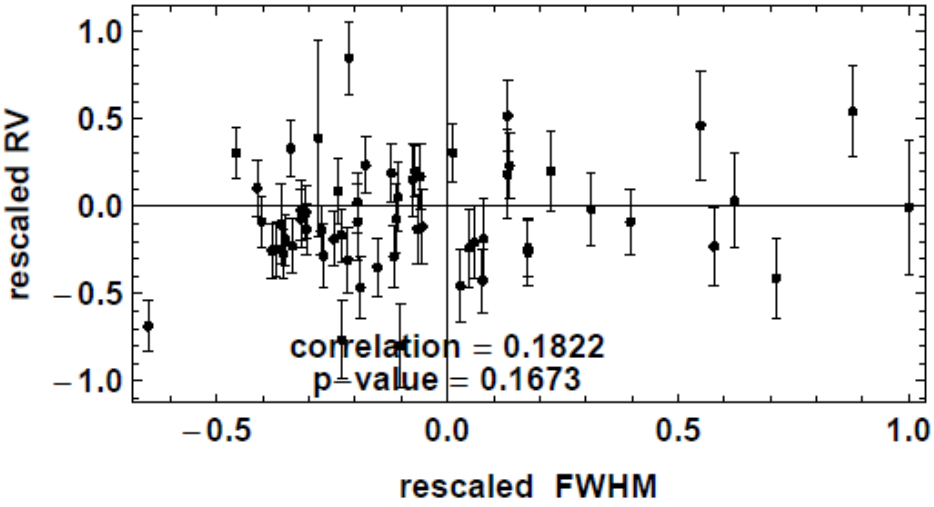
-Recent work by Paul Robertson et al. on Gliese 581 show seasonal dependent correlations between RV and Hα stellar activity. The artifact RV component removed by linear regression yields one less planet. (Science 2014, arXiv 1407.1049R)

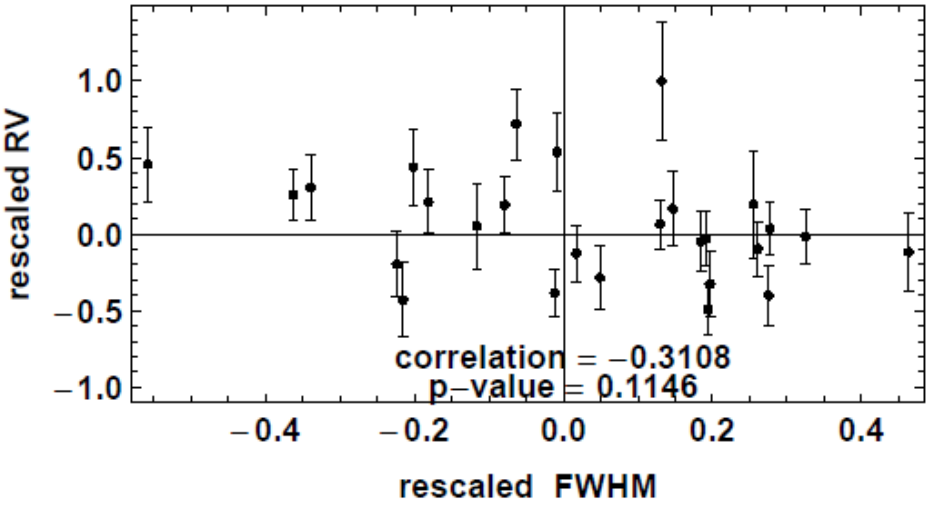
-I will show an example of this in Gliese 667C and demonstrate how a hierarchical (multi-level) Bayesian analysis can effectively deconvolve the effects of the measurement errors to better estimate the underlying regression relation.

Pearson correlation coefficient between 2 planet RV residuals and FWHM diagnostic for 4 different observing seasons.

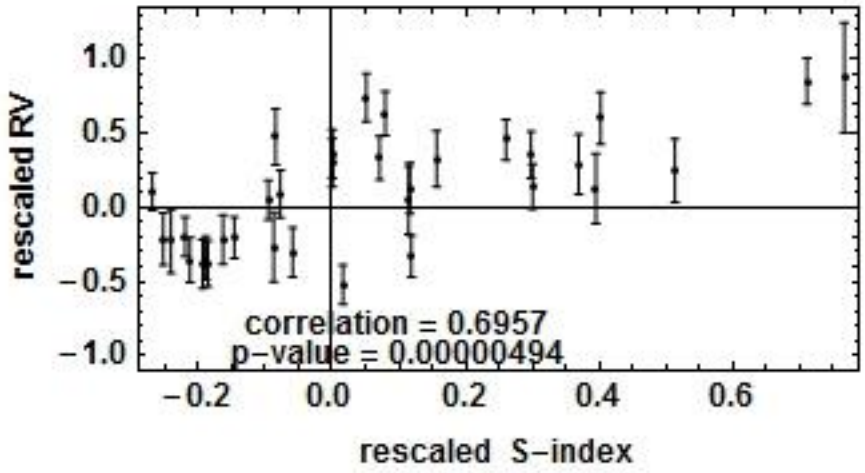


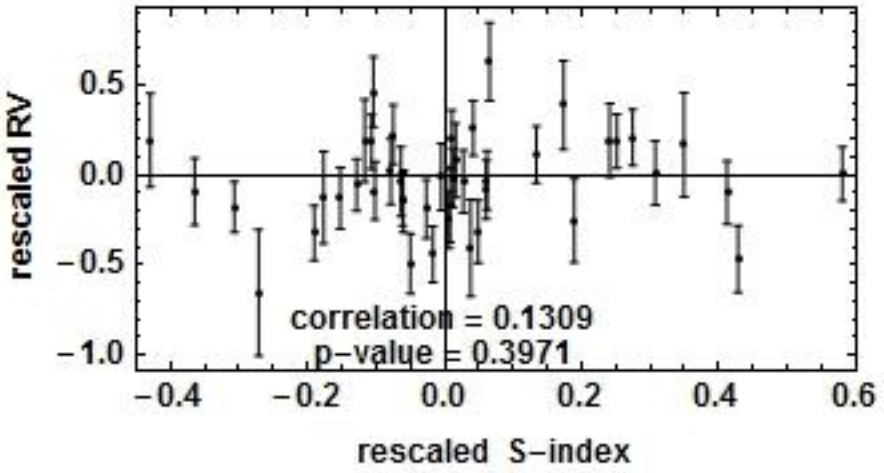


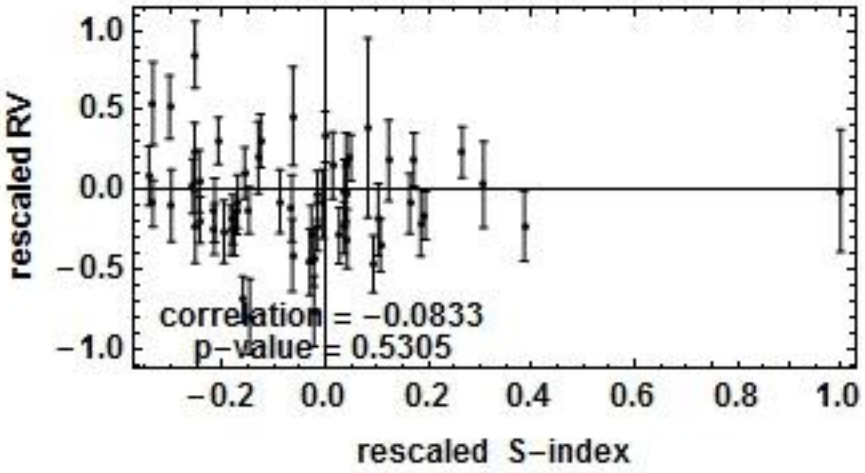


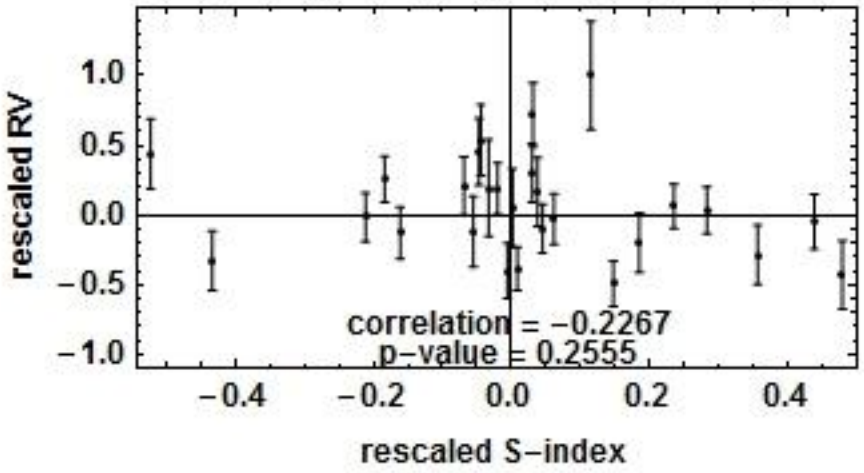


Pearson correlation coefficient between 2 planet RV residuals and S index for 4 different observing seasons.



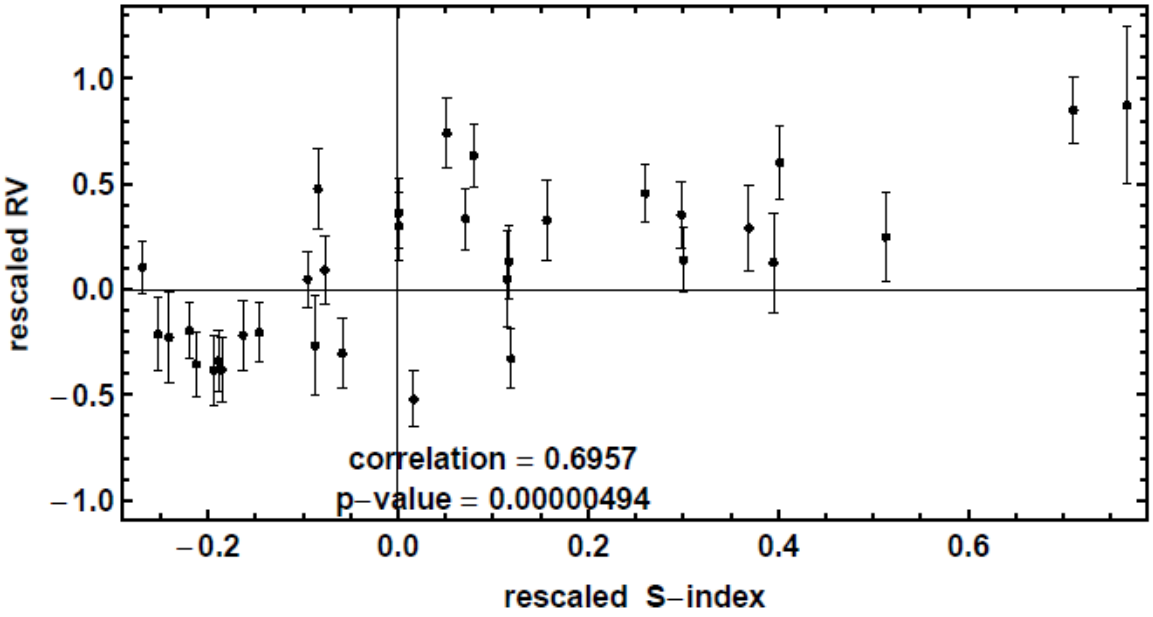


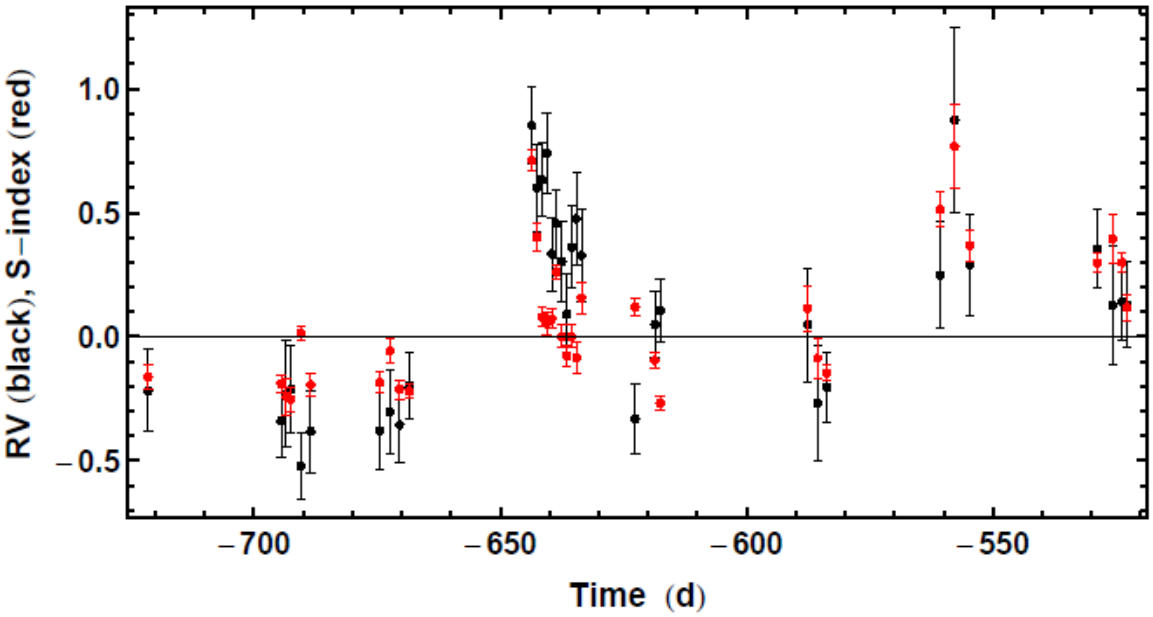




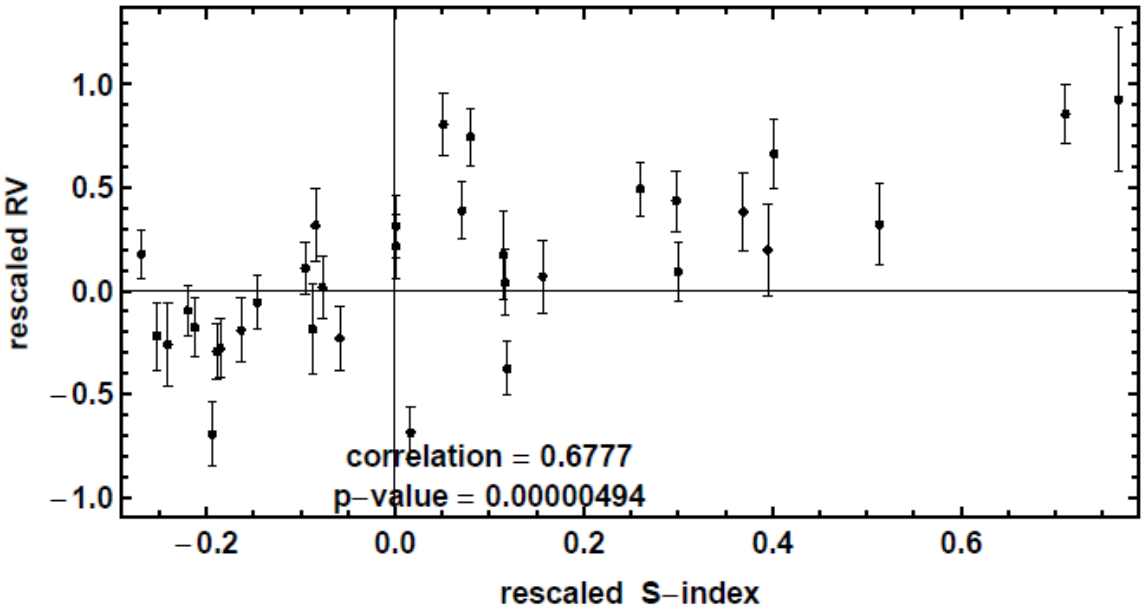
Only one season exhibits a significant correlation.
Correlation strongest with S-index

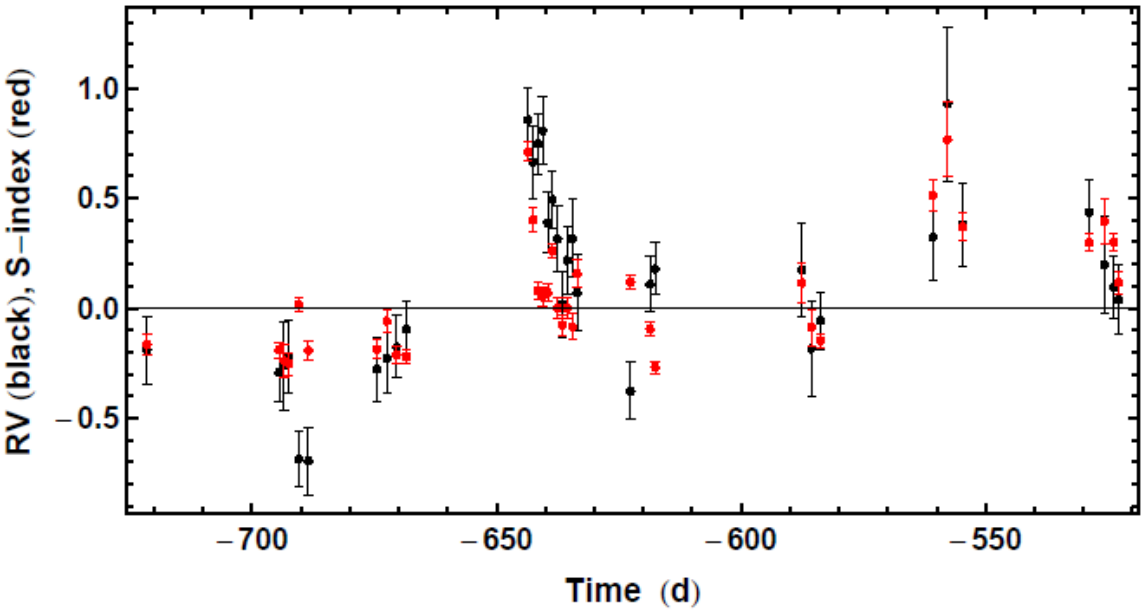
Pearson correlation coefficient between 2 planet RV residuals and S index for most correlated season.





Pearson correlation coefficient between 1 planet RV residuals and S index for most correlated season.





return

Hierarchical/multilevel Bayes regression analysis *

A common problem in astronomy is to explore whether there is a correlation (i.e., a straight line relationship, commonly referred to as the regression line), between the dependent and independent variables.

There is often some intrinsic scatter about the regression line. The intrinsic scatter arises from variations in the physical properties that are not completely captured by the independent variables included in the regression, in this case the single independent variable, the S-index.

Because of measurement uncertainties we don't know the true values of RV and S only their measured values. We represent their true values by y_t and x_t , *frequently referred to as hidden or latent variables in hierarchical Bayes (also known as multilevel modeling)*.

The effect of measurement error in the independent variable, x, is to bias the slope towards zero and reduce the magnitude of the observed correlation. Measurement error in the response, y, also reduces the magnitude of the correlation.

*** Ref: Brandon Kelly, Ap.J., 665, 875, 2007**

Hierarchical Bayes regression analysis

Can allow for an intrinsic scatter in the regression relationship between the hidden true values of x_t and y_t according to the additive noise model

$$y_{ti} = \alpha + \beta\, x_{ti} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and σ is unknown and treated as a model parameter.

We now allow for independent measurement errors in both the dependent and independent variables according to

$$x_i = x_{ti} + e_{x,i},$$

where $e_{xi} \sim N(0, \sigma^2_{x,i})$ and $\sigma_{x,i}$ is assumed known. Also

$$y_i = y_{ti} + e_{y,i},$$

where $e_{yi} \sim N(0, \sigma^2_{y,i})$ and $\sigma_{y,i}$ is assumed known.

In a Bayesian analysis we can eliminate the hidden values by a process called marginalization ⇨ need to specify a prior for each and integrate the joint posterior probability distribution over their possible values.

How to handle hidden variables in a Bayesian analysis

To handle hidden true values, x_t , we must first specify our prior information. As a first guess we might assume an independent uniform prior and specify some large prior range between -C and C that we are confident the x_t value falls within. Suppose that the first n samples of the observed values, x_i, fall within the much smaller range 0.01 C and 0.1 C. Do we really believe the next x_{ti} is likely to be anywhere in the range -C and C?

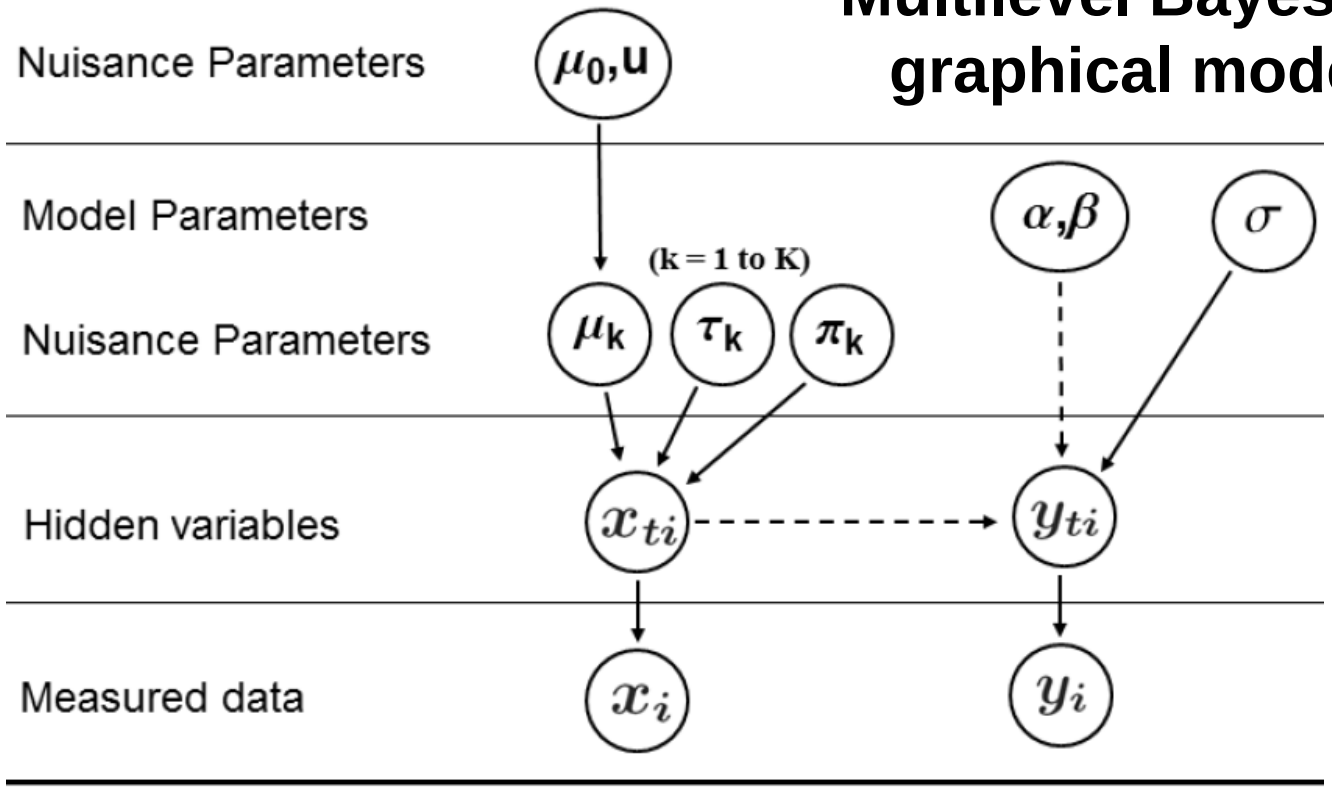
Another choice for the prior is to choose what we will refer to as an informative prior like a Gaussian (or a mixture of Gaussians) and learn about the mean and variance of the Gaussian(s) from the measured sample.

This leads to a probabilistic dependence among the x_{ti} values that implements a pooling of information that can improve the accuracy of inference. Each x_i measurement bears on the estimation of the unknown mean and variance of the population of x_t values, and thus indirectly, each measured x_i bears on the estimation of every other x_{ti}, via a kind of adaptive correction.

This is referred to as borrowing strength from each other and avoids the biased estimates of the intercept and slope common to ordinary least-squares analysis of this situation.

Multilevel models can have enough parameters to fit the data well, while the choice of an informative prior for the hidden parameters structures some dependence into these parameters thereby avoiding problems of over fitting.

Multilevel Bayesian graphical model



Solid arrows denote conditional dependencies.

Dashed arrows represent deterministic conditionals.

Absence of a connection denotes conditional independence.

$$p(\mu_k|\mu_0,u,I) \sim N(\mu_0,u^2); \quad p(\tau_k|I) = \frac{(\tau_k+\langle\sigma_x\rangle)^{-1}}{\ln(1+\frac{\tau_{max}}{\langle\sigma_x\rangle})}$$

$$p(\pi_k|I) \sim \text{Dirichlet}(1,1,\cdots,1); \quad p(u|I) = \frac{(u+\langle\sigma_x\rangle)^{-1}}{\ln(1+\frac{u_{max}}{\langle\sigma_x\rangle})}$$

$$p(x_{ti}|\mu_1,\tau_1,\cdots,\mu_K,\tau_K,I) = \sum_{i=1}^{K} \frac{\pi_k}{\sqrt{2\pi\tau_k^2}} \exp[-\frac{(x_{ti}-\mu_k)^2}{2\tau_k^2}]$$

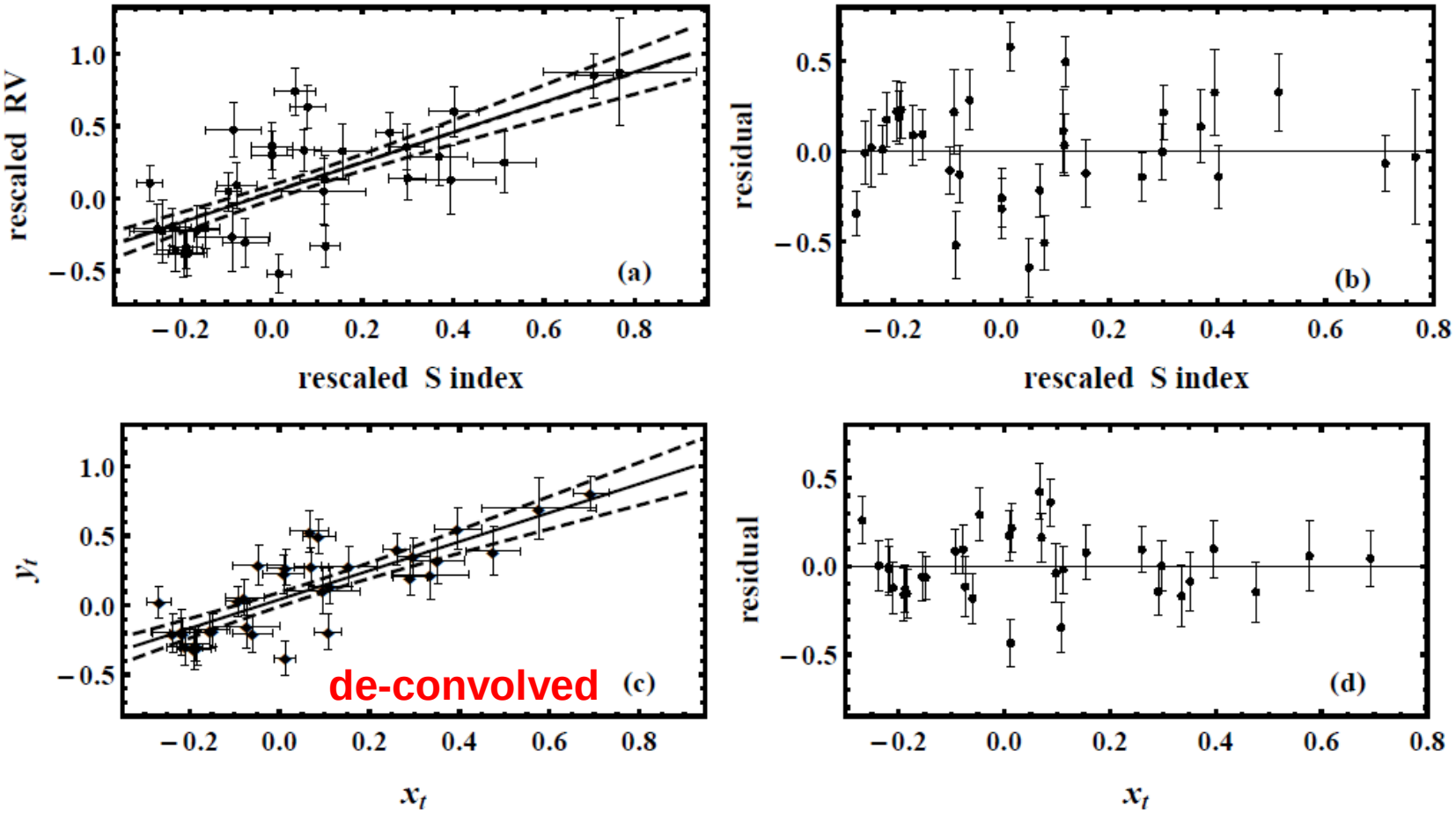
$$p(y_{ti}|x_{ti},\alpha,\beta,\sigma,I) \sim N(\alpha+\beta\, x_{ti},\sigma^2)$$

$$p(x_i|x_{ti},I) \sim N(x_{ti},\sigma_{x,i}^2); \quad p(y_i|y_{ti},I) \sim N(y_{ti},\sigma_{y,i}^2)$$

Conditional probabilities

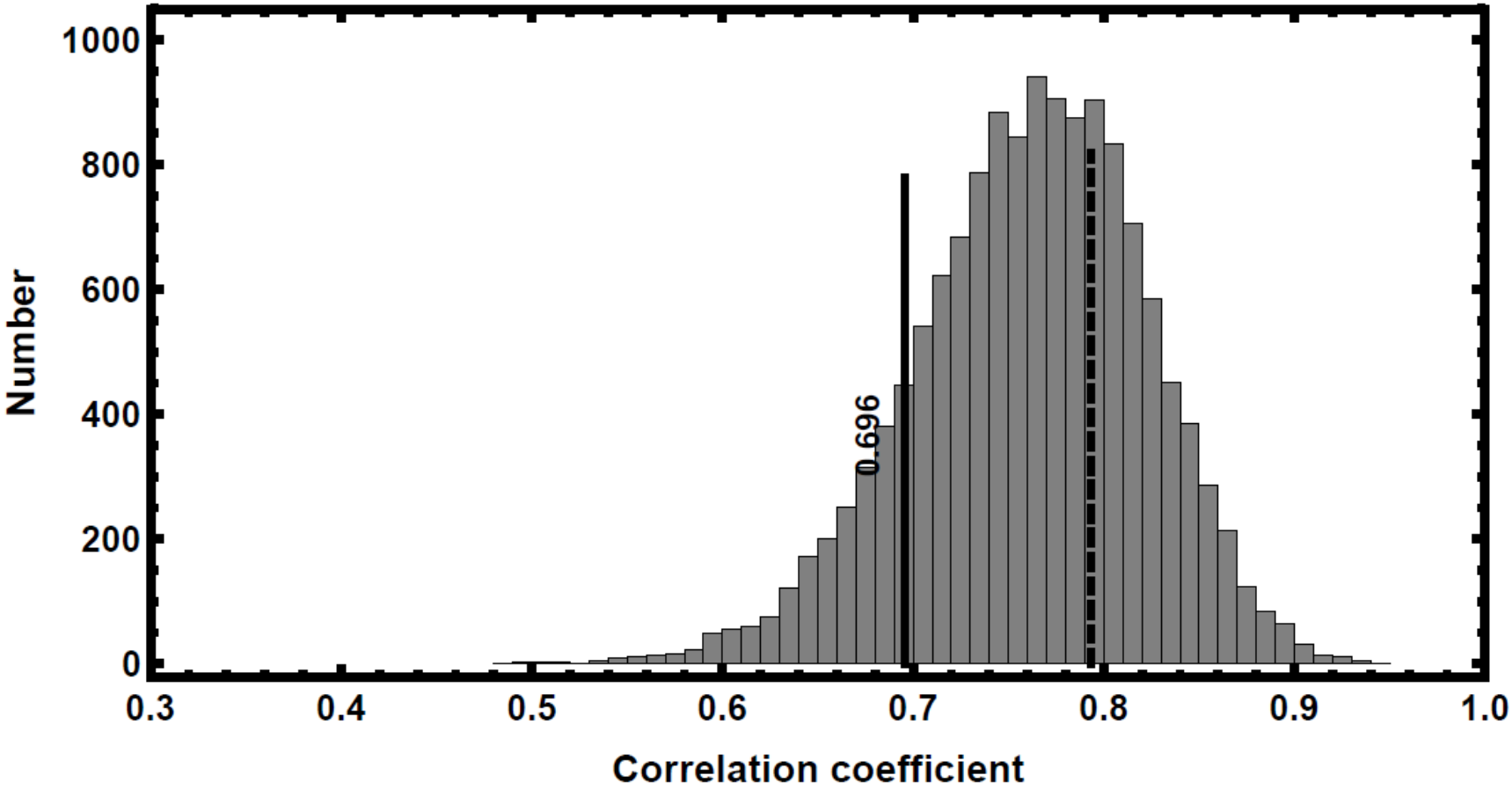
Hierarchical Bayes Regression Line

Yields representative samples of the underlying regression, effectively deconvolves blurring effects of measurement errors.



Panel (a) shows the raw data. In panel (c), the points and error bars are the mean and standard deviation of the MCMC estimates of the true coordinates. In both the solid line is the same mean regression line and the dashed lines show ±1 std. dev. fit uncertainty. The dot-dashed line is the MAP fit. Panels (b) & (d) show residuals.

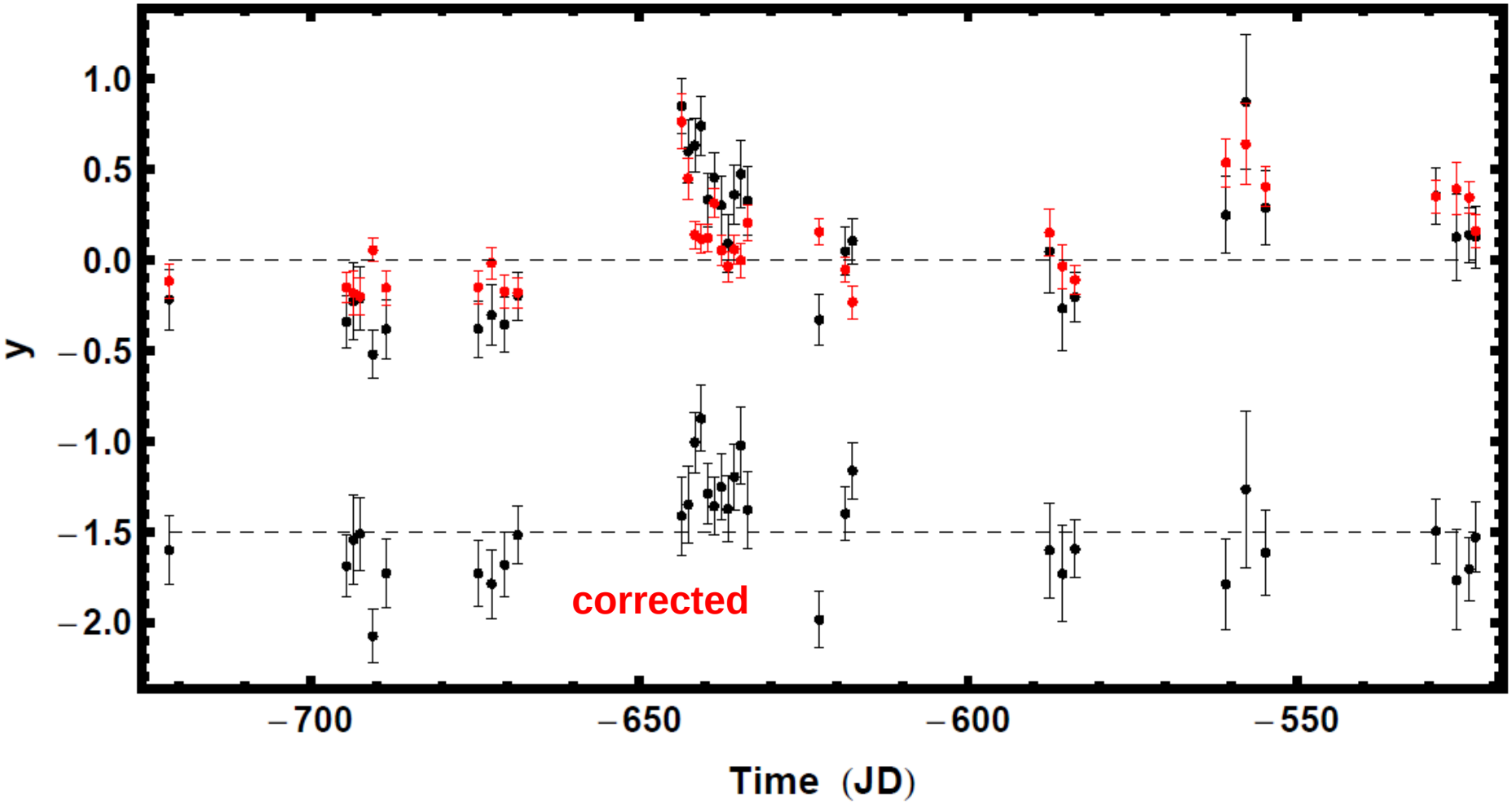
Correlation coefficient between 2 planet RV residuals and S-index



The solid vertical line shows the correlation coefficient derived from the raw RV residuals and S-index values. The histogram is the marginal distribution of the correlation coefficient for the estimated true values of RV residuals and S-index from our hierarchical Bayesian analysis. The dashed line is predicted correlation from raw data and errors.

return

Comparison of measured (black) and predicted RV from regression line (red). Residual in black below.



The standard deviation of the corrected RV is reduced by a factor of 1.42.

return

Before correction of radial velocity data.

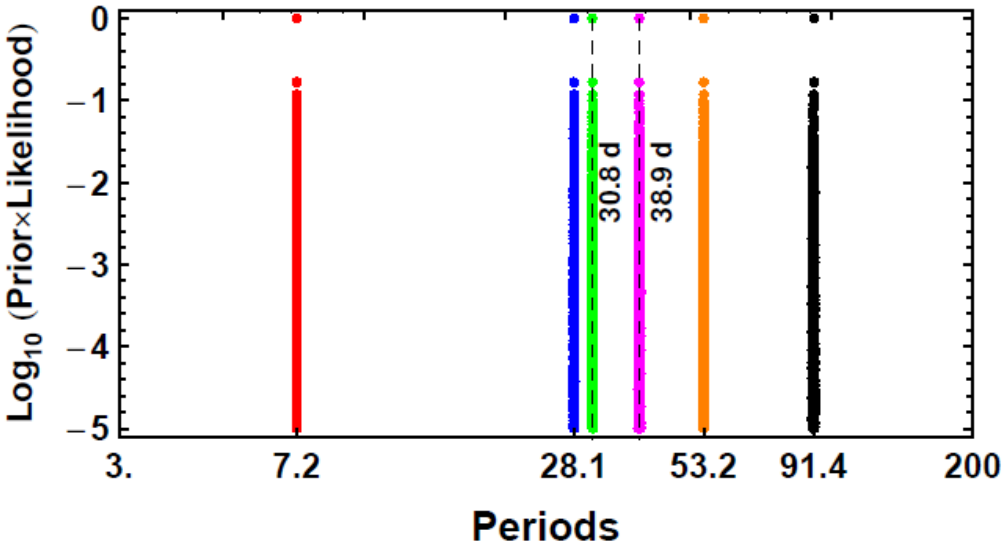
Relative probabilities via Bayes factors of different models analysis of Gliese 667C HARPS radial velocity data.

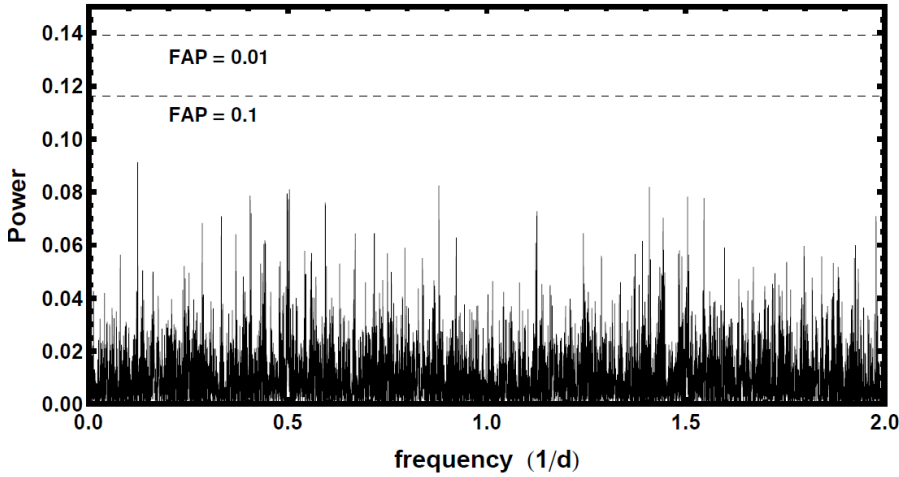
Number of signals	Kepler model	
0	3.6×10^{-38}	
1	3.1×10^{-16}	
2	7.7×10^{-12}	
3	2.7×10^{-6}	
4	4.4×10^{-5}	
5	0.003	
6	1.0	

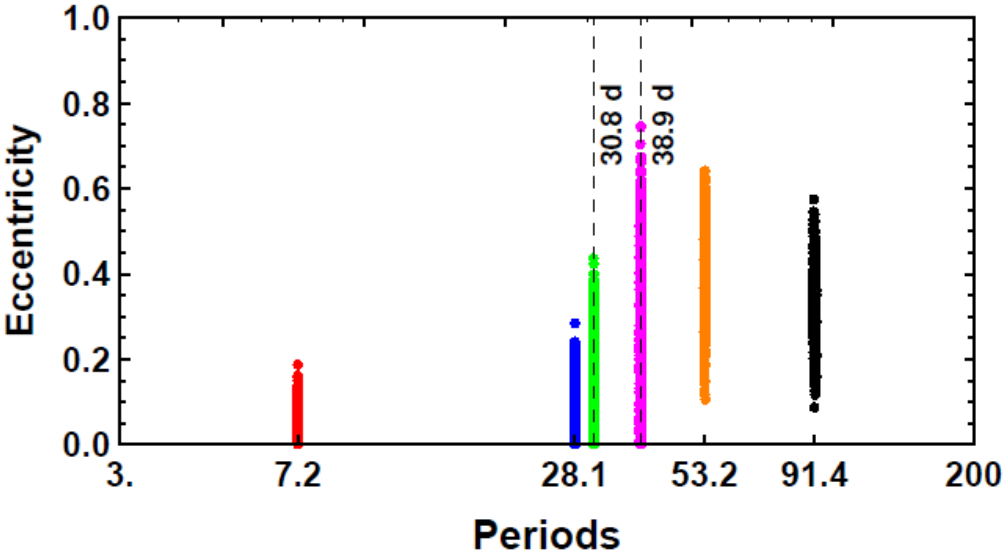
1

Return

6 planet Kepler periodogram of raw RV data (Gregory 2012)







Generalized Lomb-Scargle periodogram for the 6 planet fit residuals. Spectrum consistent with white noise.

return

After correction of radial velocity data.

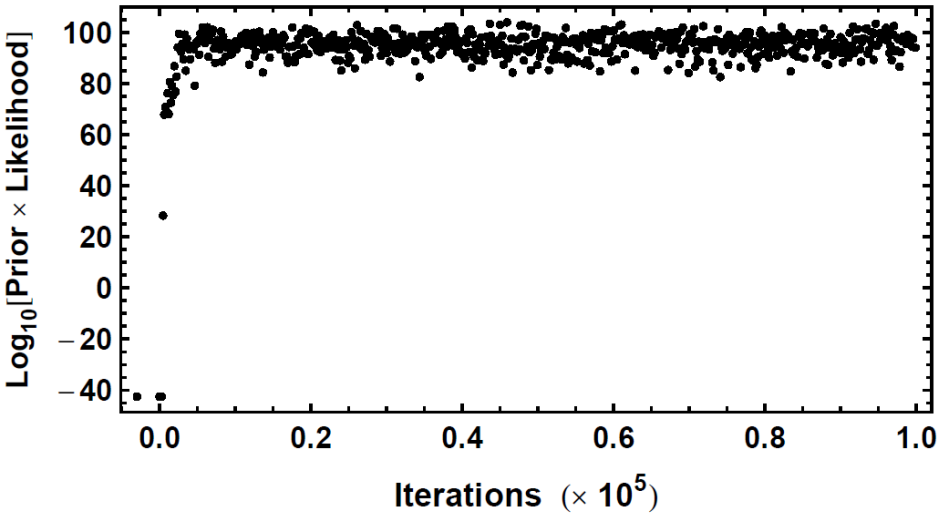
Relative probabilities via Bayes factors of different models for Gliese 667C corrected radial velocities.

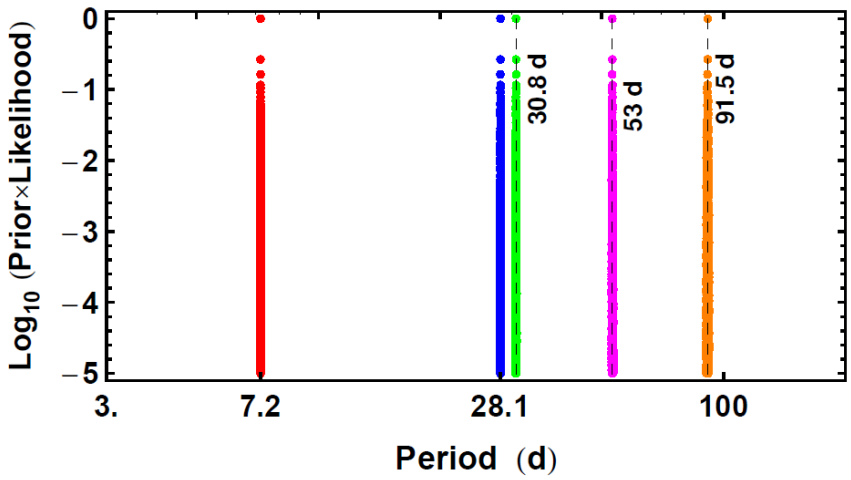
Number of signals	Kepler model	Apodized Kepler model
0	4.2×10^{-37}	4.2×10^{-37}
1	2.5×10^{-12}	1.7×10^{-12}
2	6.4×10^{-8}	3.8×10^{-8}
3	1.8×10^{-5} 3	2.3×10^{-4} 4
4	2×10^{-4}	6.3×10^{-5} 6
5	1.0 1	0.24 2 sim
6	0.44 5	8.4×10^{-4} 7

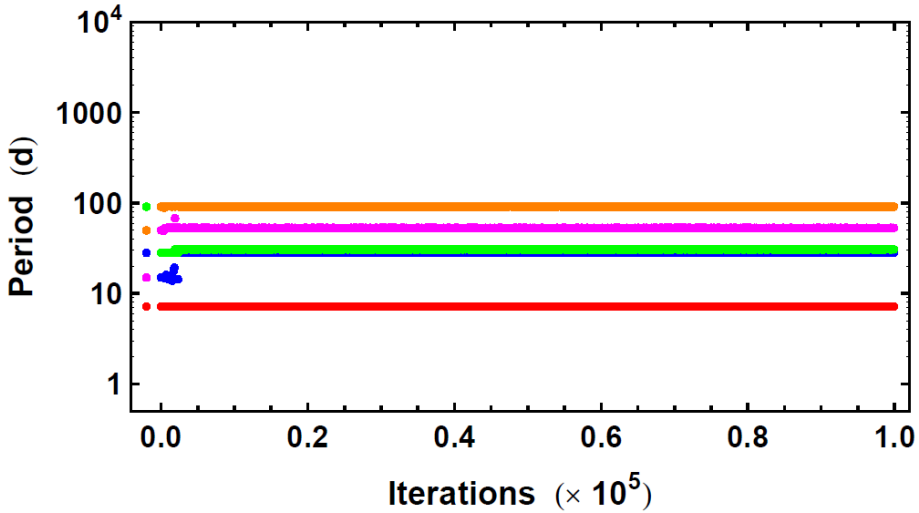
39 d signal no longer significant

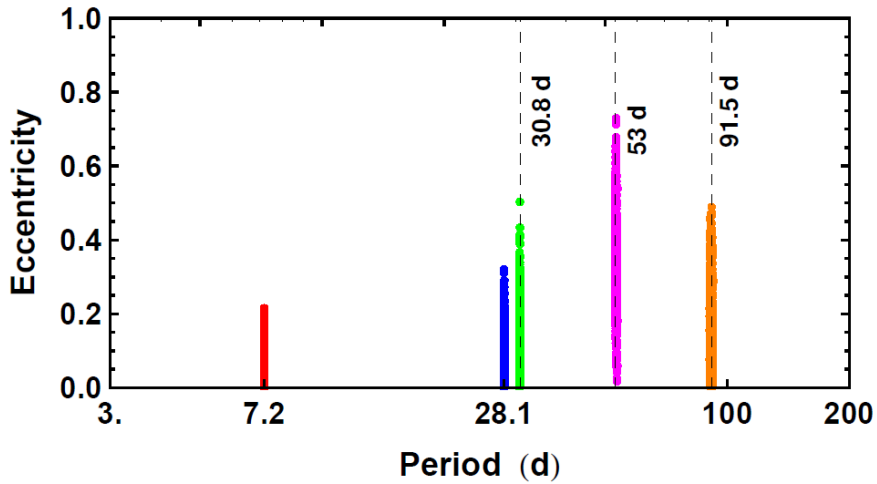
Of those remaining, the 53 d signal is 2nd harmonic of rotation period + periods of 28, 31 & 91.5 d are suspect from their appearance in the apodized periodogram results of diagnostics shown above.

5 signal Kepler periodogram of corrected RV data









Of the remaining signals , the 53 d signal is 2nd harmonic of rotation period. Periods of 28, 31 & 91.5 d are suspect from their appearance in the apodized periodogram results of diagnostics shown above.

return

Preliminary conclusions based on HARPS data

1) Alternative periodogram models provide some insights to distinguishing signals from stellar artifacts.

2) A hierarchical (multilevel) Bayes regression analysis can effectively de-convolve the blurring effect of measurement errors.

3) Based on these preliminary findings, the only signal that I would wager a bet on to be planetary in origin is the 7.2 d period. The close proximity of the 28 and 31 d signals together with the period derivative hinted by the chirp analysis calls both into question.