SAMSI Astrostatistics Tutorial

Models with Gaussian Uncertainties (lecture 2)

Phil Gregory

University of British Columbia

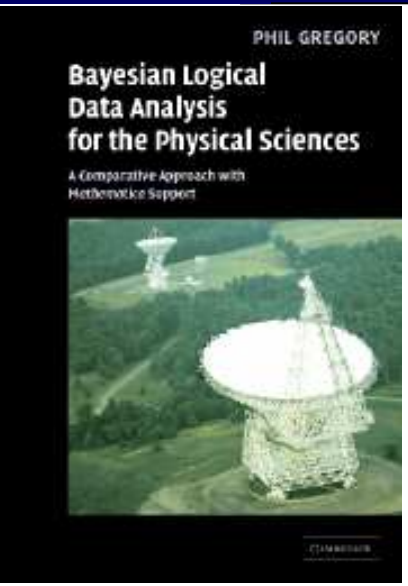
2006

The rewards of data analysis:

'The universe is full of magical things, patiently waiting for our wits to grow sharper.‘

Eden Philpotts (1862-1960)

Author and playwright



Resources and solutions

Book Preface (11 Kb)

Mathematica Tutorial (1415 Kb)

Errata (206 Kb)

Revisions (268 Kb)

Book website: www.cambridge.org/052184150X

Bayesian Logical Data Analysis for the Physical Sciences

outline

Contents:

Outline

Gaussian uncertainties

According to the Central Limit Theorem the probability distribution of the average of m independent measurements of some quantity tends to a Gaussian as m increases regardless of the probability distribution of individual measurements of the quantity, provided it has a finite variance.

Generalization of the Central Limit Theorem: "Any quantity that stems from a large number of sub-processes is expected to have a Gaussian (normal) distribution."

The reason that the normal (Gaussian) distribution occurs so often is because measured quantities are frequently the result of a large number of effects, i.e., is some kind of averaged resultant of these effects.

Since the distribution of the average of a collection of random variables tends to a Gaussian, this is often what we observe.

Of course, by working with average values we can ensure the Likelihood function is well described by a Gaussian.

Gaussian uncertainties, alternate justification

Chapter 8

outline

In Bayesian inference, probabilities are a measure of our state of knowledge about nature, not a measure of nature itself.

equivalently

In Bayesian inference, probabilities are a measure of our state of ignorance about nature, not a measure of nature itself.

What if we have very limited knowledge about the choice of sampling distribution to be used in calculating the likelihood and not enough measurements (<5) to ensure a Gaussian distribution by averaging?

We appeal to the Maximum Entropy Principle. It says that unless we have some additional prior information which justifies the use of some other sampling distribution, then use a Gaussian sampling distribution. It makes the fewest assumptions about the information you don't have and will lead to the most conservative estimates (i.e., greater uncertainty than you would get from choosing a more appropriate distribution based on more information).

The Maximum Entropy Principle, developed by E. T. Jaynes, is based on Claude Shannon's landmark 1948 paper on the use of entropy to quantify the uncertainty of a discrete probability distribution.

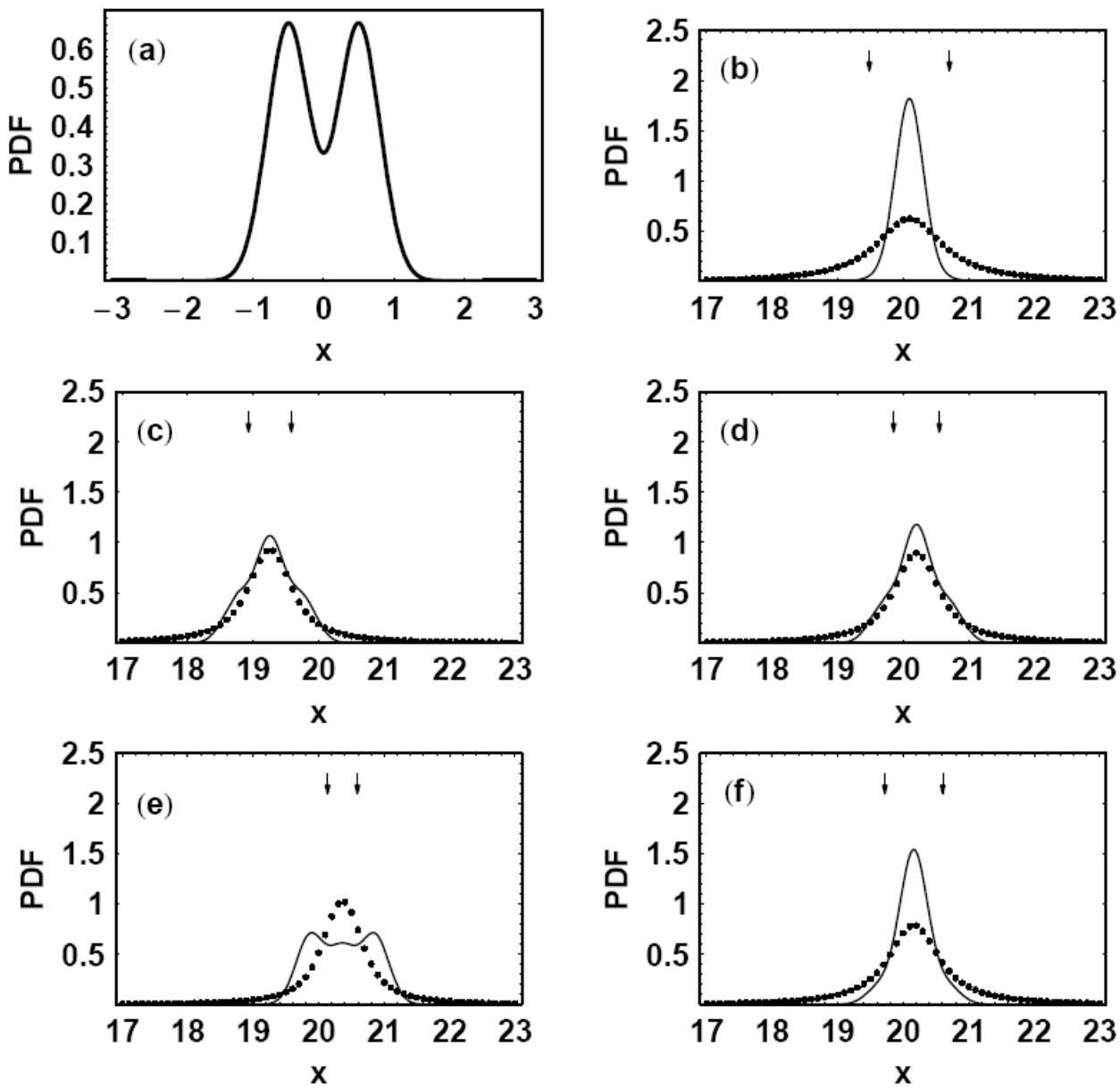
Panel (a) shows the true bimodal distribution of measurement errors for some instrument.

Panels (b), (c), (d), (e) & (f) show a comparison of the posterior PDF's derived from five simulated data derived from (a).

Each sample consists of the two data points indicated by the two arrows at the top of each panel.

The solid curve shows the result obtained using the true sampling distribution.

The dotted curve shows the result using a Gaussian with unknown σ and marginalizing σ.



A Bayesian Revolution in Spectral Analysis

outline

1) Fourier Power Spectrum (periodogram)

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.

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

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

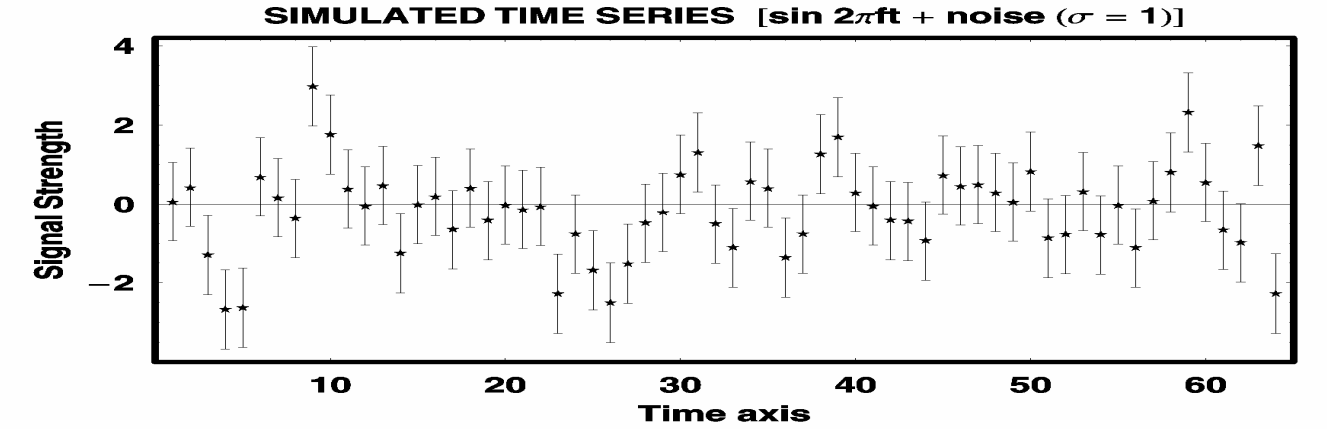
$$p(f_n \mid D, I) \propto \exp\left\{\frac{C(f_n)}{\sigma^2}\right\}$$

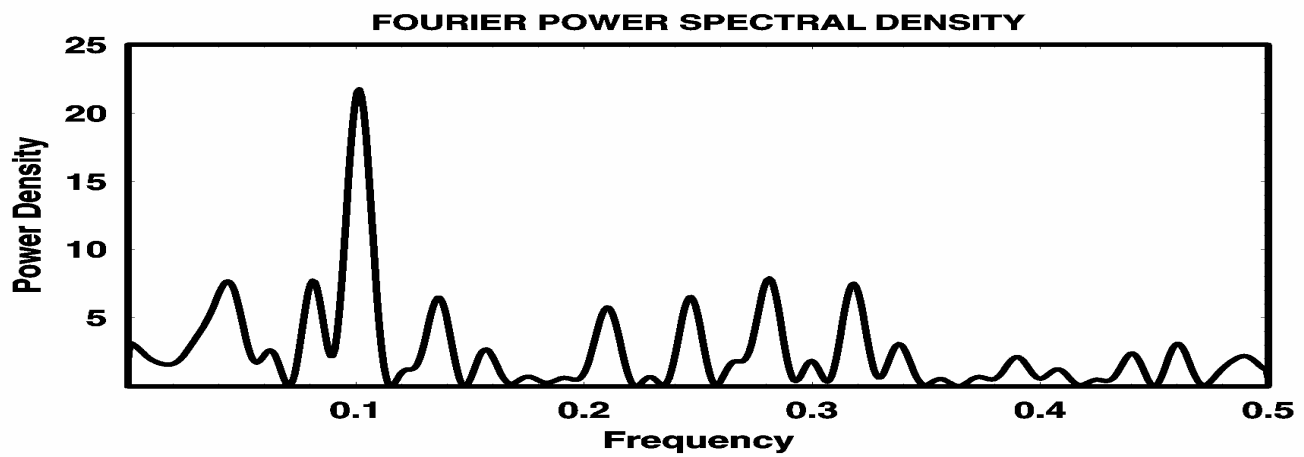
Thus $C(f_n)$ is indeed fundamental to spectral analysis but not because it is itself a satisfactory spectrum estimator.

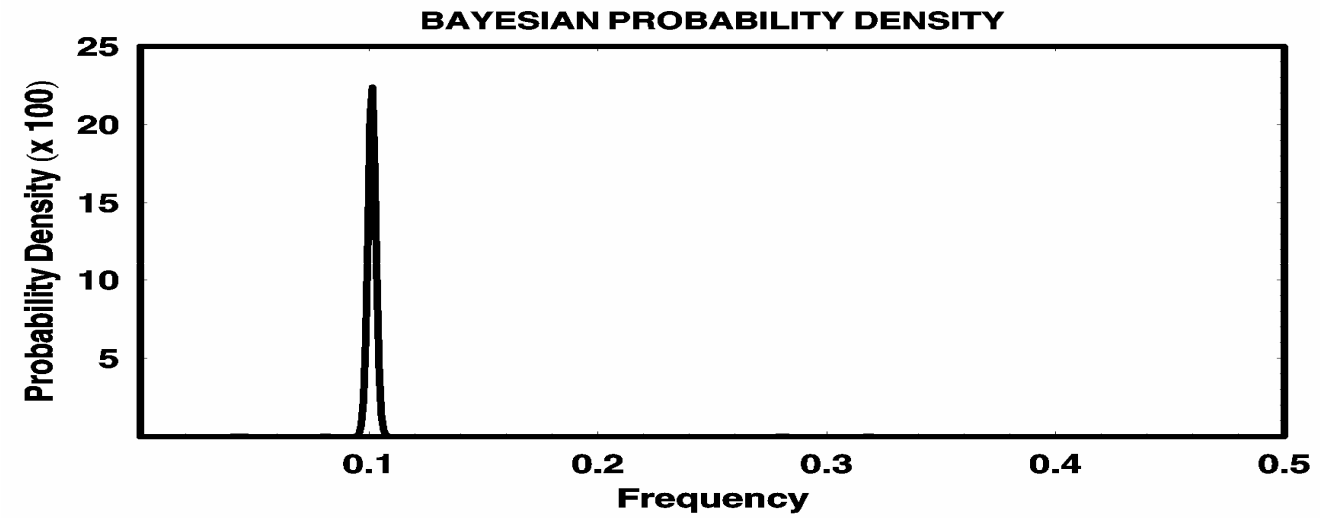
The proper algorithm to convert $C(f_n)$ to $p(f_n|D,I)$ involves first dividing $C(f_n)$ by the noise variance and then exponentiation.

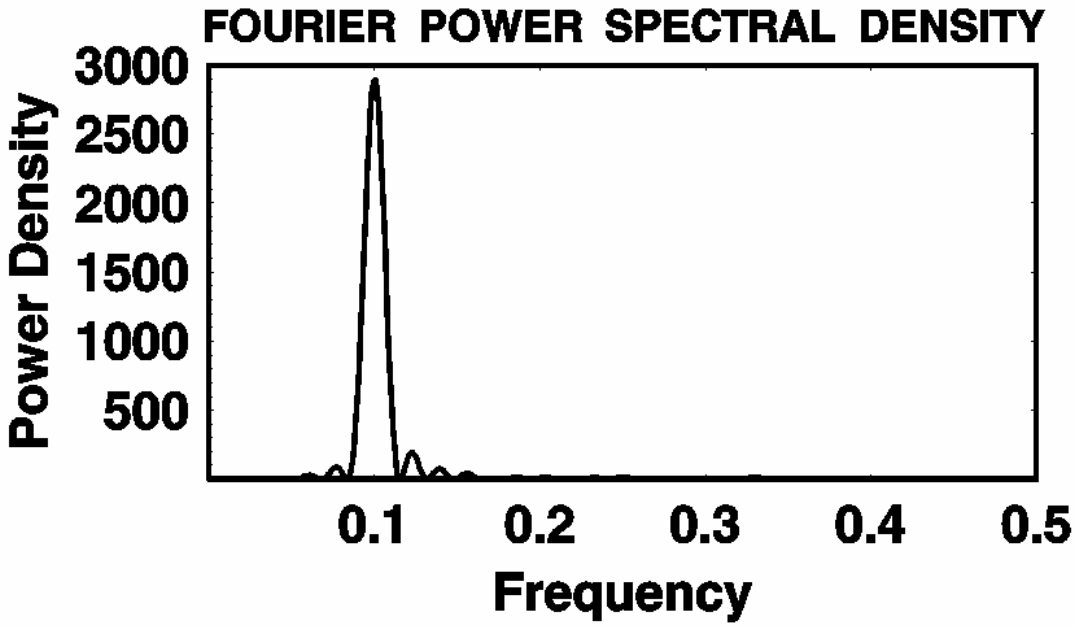
This naturally suppresses spurious ripples at the base of the periodogram as well as linear smoothing; but does it by attenuation rather than smearing, and therefore does not lose any resolution.

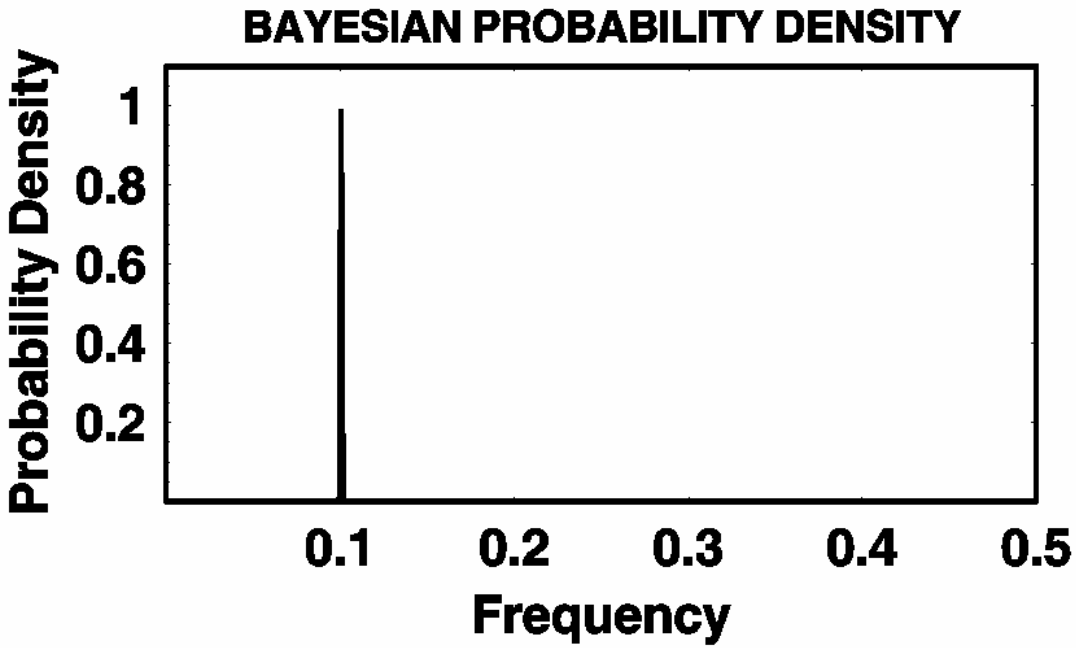
The Bayesian nonlinear processing of $C(f_n)$ also yields, when the data give evidence for them, arbitrarily sharp spectral peaks.











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 saw how to do that when estimating a mean in the previous lecture.

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

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

The Bretthorst periodogram:
A Bayesian generalization of the Lomb-Scargle periodogram

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

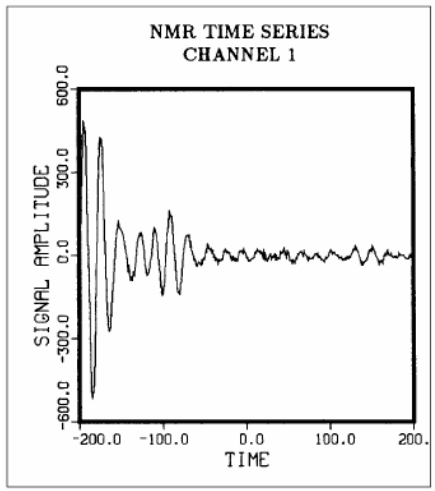
Bretthorst (2001) generalized Jaynes' insights to a broader range of single-frequency estimation problems and sampling conditions.

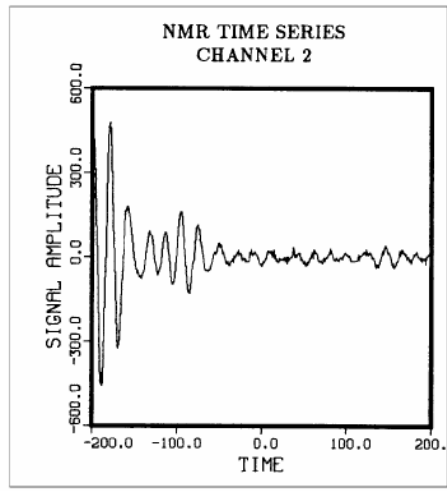
In the course of this development, Bretthorst established a connection between the Bayesian results and an existing frequentist statistic known as the Lomb-Scargle periodigram, which is a widely used replacement for the Schuster periodogram in the case of non-uniform sampling.

Bretthorst'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

Let $d_R(t_i)$ denote the real data at time t_i and $d_I(t'_j)$ denote the imaginary data at time . There are N_R real samples and N_I imaginary samples for a total of $N = N_R + N_I$ samples.

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

The analysis does not require the t_i and t'_j to be simultaneous and successive samples can be unequally spaced in time.

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

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

$$d_I(t'_j) = A\cos(2\pi f\, t'_j - \theta)\, Z(t'_j) + B\sin(2\pi f\, t'_j - \theta)\, Z(t'_j) + e_I(t'_j),$$

The function $Z(t_i)$ describes an arbitrary modulation of the amplitude, e.g., exponential decay as exhibited in NMR signals. $Z(t_i)$ is sometimes called a *weighting function* or *apodizing* function.

In this analysis $Z(t_i)$ is assumed to be known, but in principle it might 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 parameter of interest is the frequency f. The other nuisance parameters are the amplitudes A, B as well as σ, the standard deviation of the Gaussian noise probabilities used to assign the likelihoods.

To compute p (f |D,I) we need to marginalize over the three nuisance parameters A, B, σ.

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

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

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

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* (assuming independent uniform priors), 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

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

The Bretthorst periodogram

outline

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

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).
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.

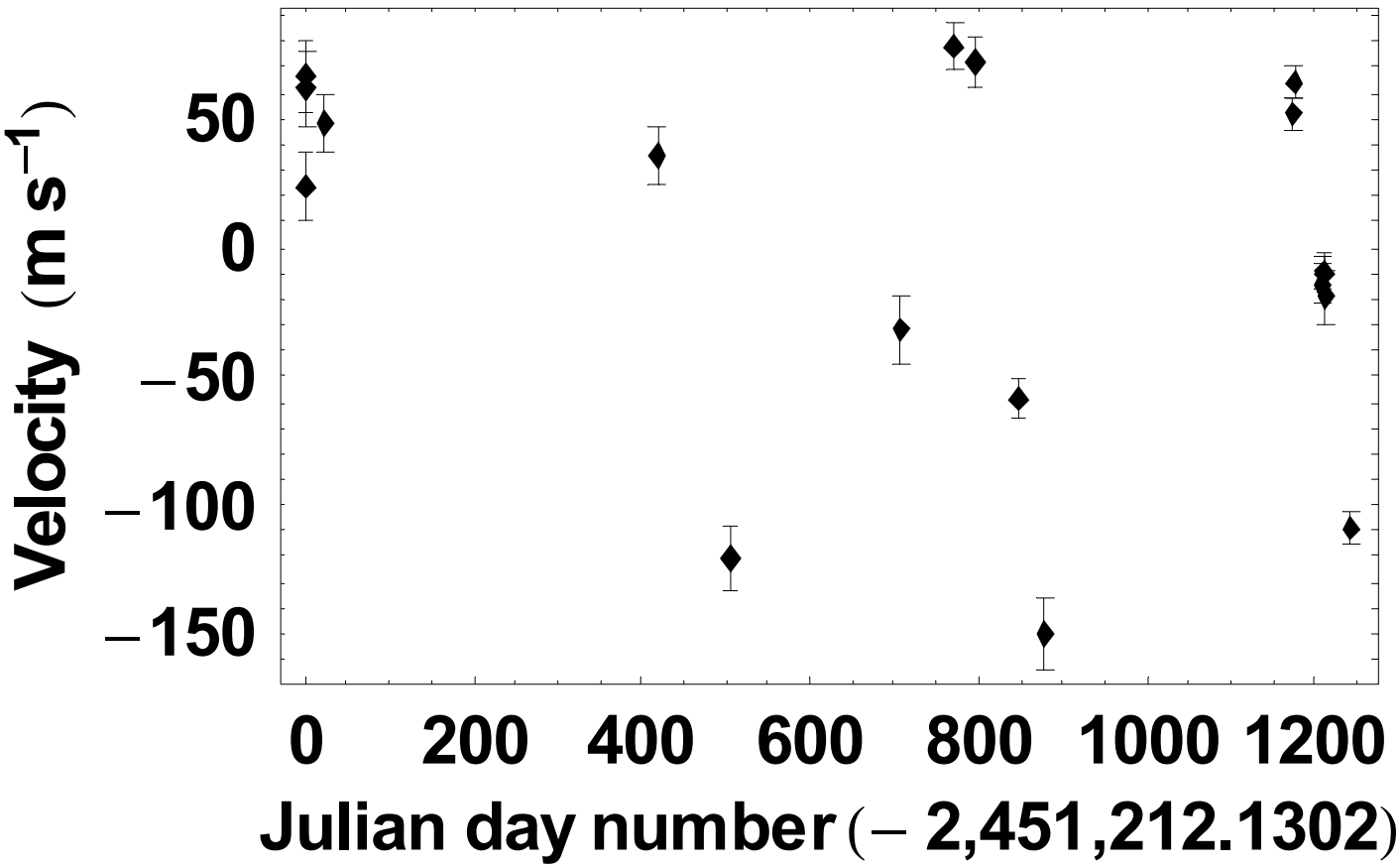
$$p(f_n \mid D, I) \propto \left[1 - \frac{2C(f_n)}{N\overline{d^2}}\right]^{\frac{2-N}{2}}$$

The Schuster periodogram is not a sufficient statistic for frequency estimation in real, i.e., nonquadrature, data. However, the Schuster periodogram is often an excellent approximation and is much faster to compute.

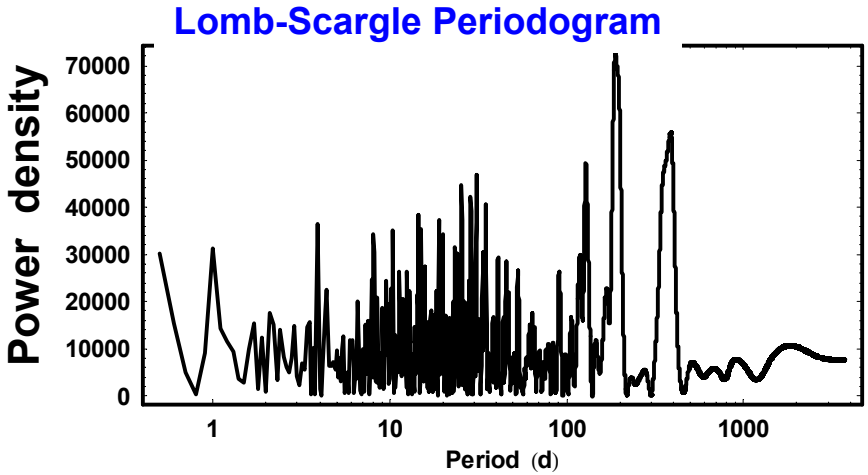
Application to extra-solar planet data

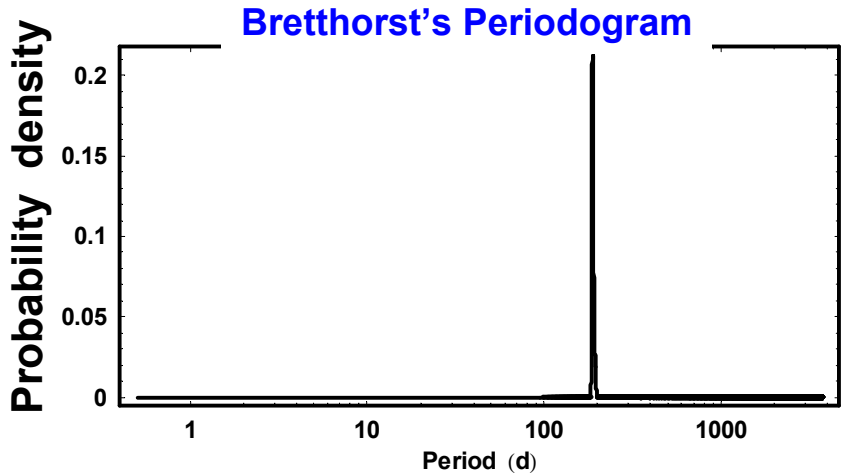
Precision radial velocity measurements for HD 73526

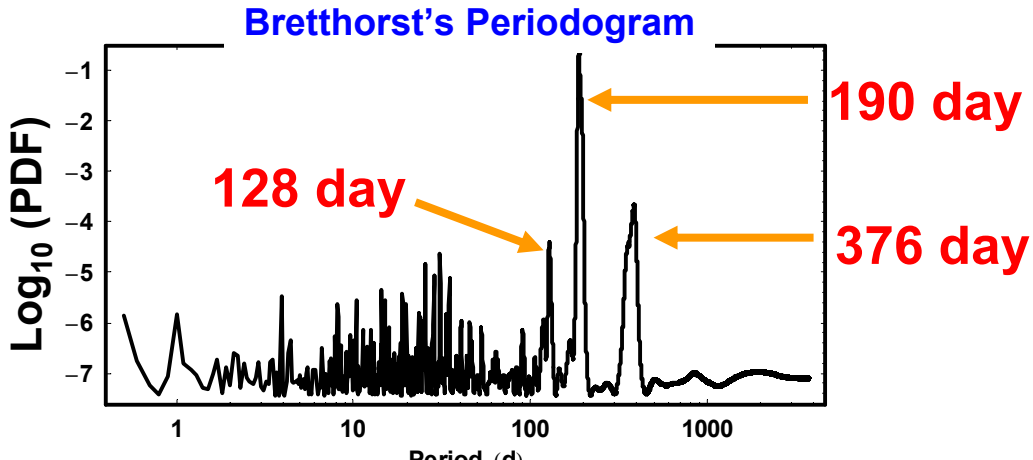
(ref. Tinney, G. C. 2003, Astrophsical Journal, 587, p. 423)



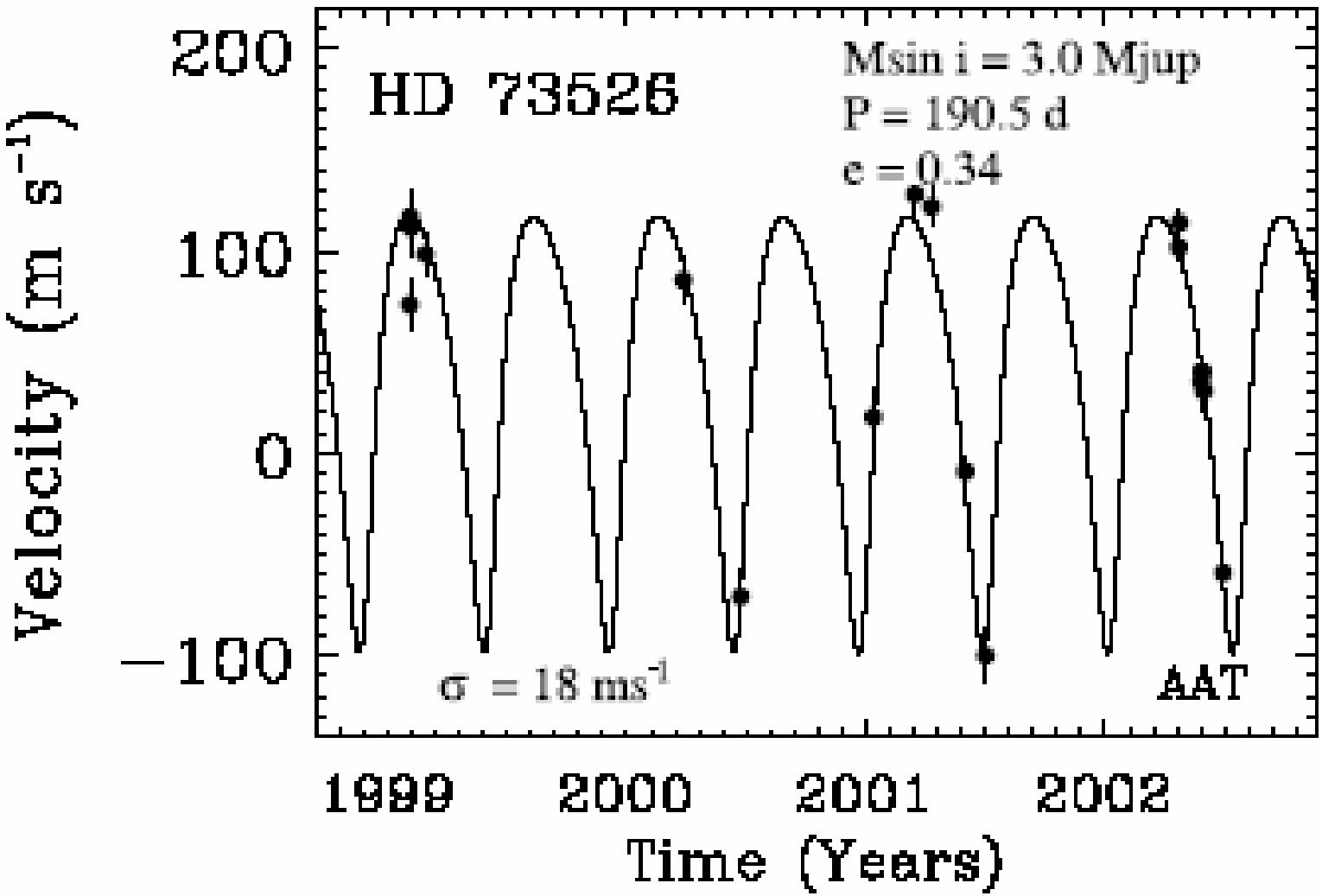
Comparison of Lomb-Scargle to Bretthorst's Bayesian generalization of Lomb-Scarge.







HD 73526 orbital solution of Tinney et al. (2003)
Obtained from a nonlinear least squares fit procedure.



THE ASTROPHYSICAL JOURNAL, 587:423–428, 2003 April 10

Problem

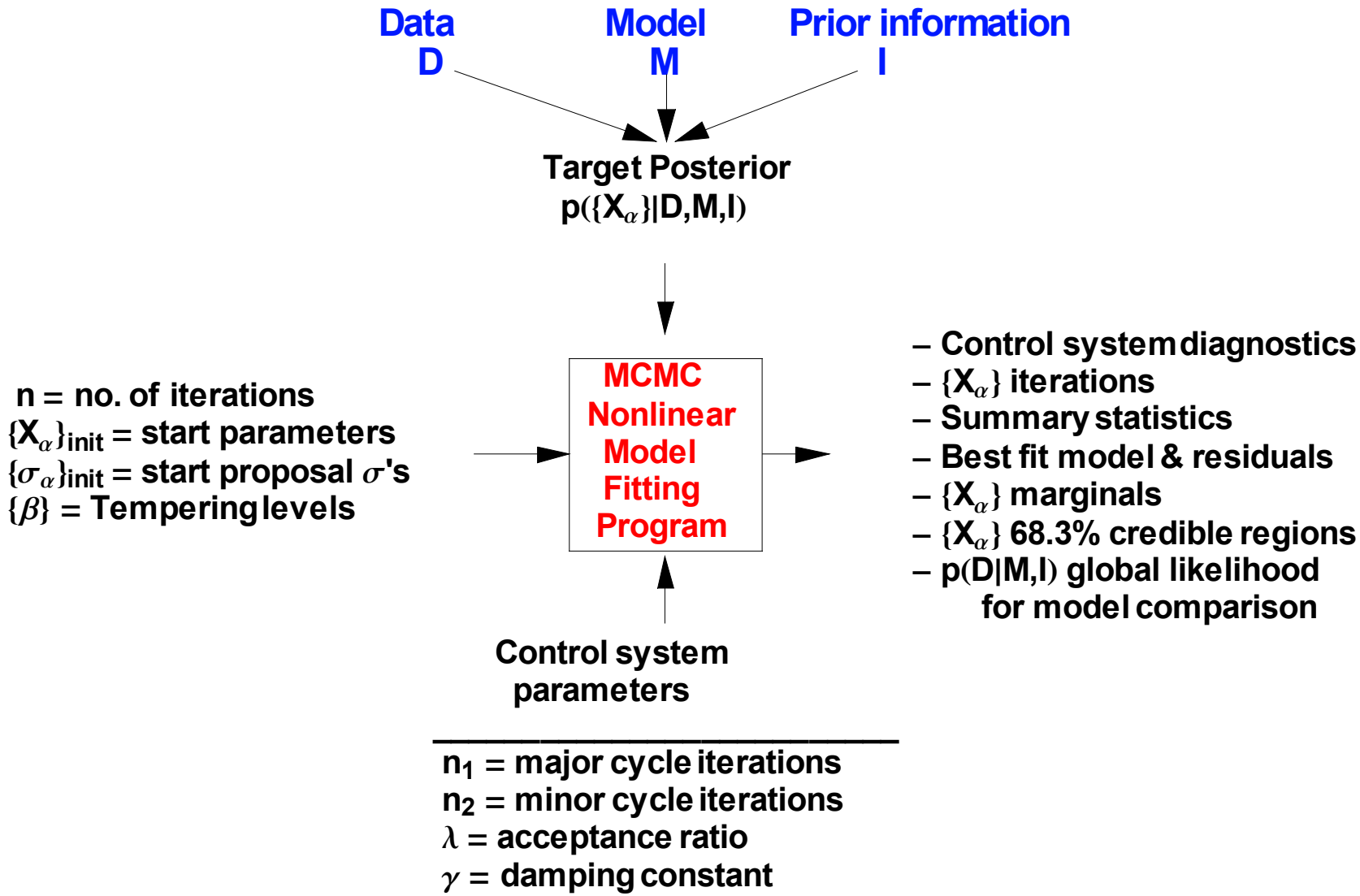
The Lomb-Scargle periodogram, and Bretthorst's Bayesian generalization, assume a sinusoidal signal which is only optimum for circular orbits.

Why not develop a Bayesian Kepler periodogram designed for all Kepler orbits. This has recently been accomplished using Markov chain Monte Carlo (MCMC) algorithms by two different individuals.

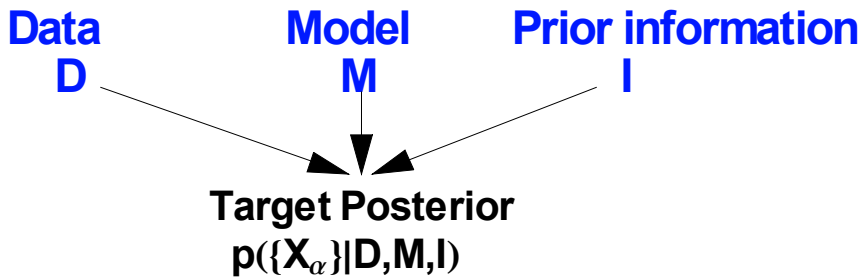
Ford, E. B. (2005), AJ, 129, 1706

Gregory, P. C. (2005), Ap. J., 631, 1198, 2005

Gregory, P. C. (2005), AIP Conference Proceeding 803, p. 139



Schematic of a Bayesian Markov chain Monte Carlo program for nonlinear model fitting. The program incorporates a control system that automates the selection of Gaussian proposal distribution σ's.



If you input a Kepler model the algorithm becomes

A Kepler periodogram

Optimum for finding all Kepler orbits and evaluating their probabilities.

Capable of simultaneously fitting multiple planet models.

A multi-planet Kepler periodogram

Advantages of the Bayesian Multi-planet Kepler Periodogram

1) Good initial guesses of the parameter values are not required. The algorithm is capable of efficiently exploring all regions of joint prior parameter space having significant probability.

2) The mathematical form of a planetary signal embedded in the host star's reflex motion is well known and is built into the Bayesian analysis as part of the prior information. There is thus no need to carry out a separate search for possible orbital period(s).

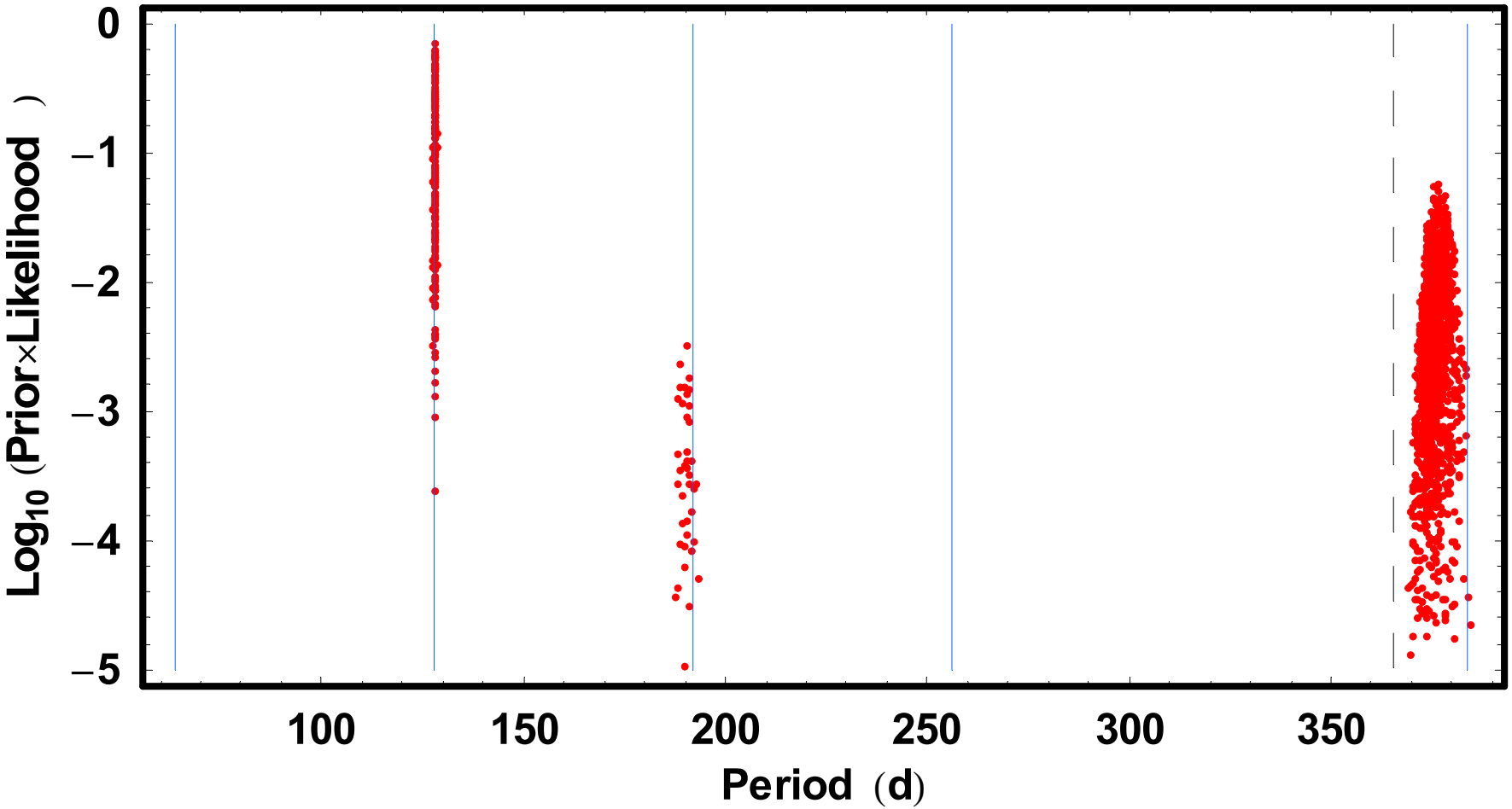
3) The method is capable of fitting a portion of an orbital period, so search periods longer than the duration of the data are possible.

4) More complicated (two or more planet) models contain larger numbers of parameters and thus incur a larger Occam penalty. A quantitative Occam's razor is automatically incorporated in a Bayesian model selection analysis. The analysis yields the relative probability of each of the models explored.

5) The analysis yields the full marginal posterior probability density function (PDF) for each model parameter, not just the maximum *a posterior* (MAP) probability values.

HD 73526 results for a one planet model

outline



A plot of Log_{10}(prior x likelihood) versus orbital period P for the β = 1 simulation. The upper envelope of the points is the projection of the joint posterior for the parameters onto the period-Log_{10}(prior x likelihood) plane. The solid vertical lines are located at 1/2, 1, 3/2, 2, 3 times the 128 day period of the highest peak. The dashed vertical line corresponds to the earth's orbital period.

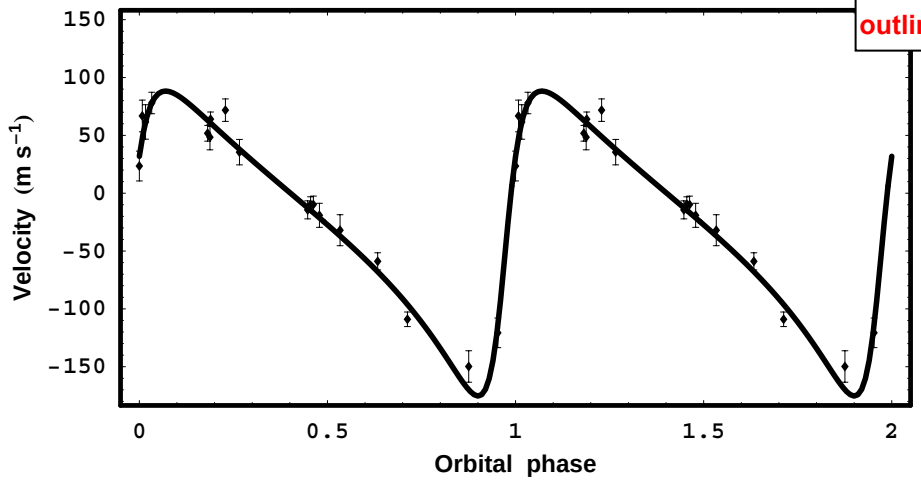
Best fit orbits versus phase

outline

P = 128 days

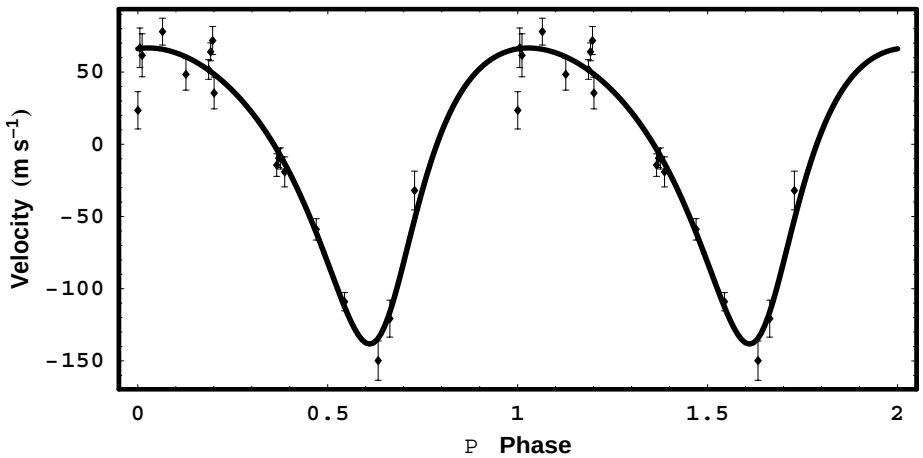
RMS residual = 11 m s^{-1}

Mean data error = 10 m s^{-1}



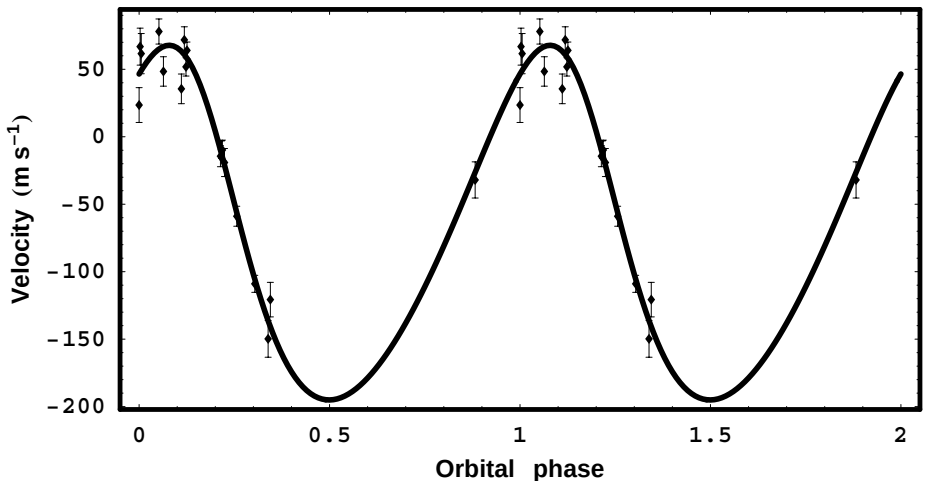
P = 190 days

RMS residual = 15 m s^{-1}



P = 376 days

RMS residual = 14 m s^{-1}



Bayesian Spectrum Analysis with Weak Prior Information of the Signal Model

In the early 1990's, Tom Loredo and I attached the question of

"How to detect a signal of unknown shape in a time series"

We developed a useful Bayesian solution to this problem.

We were initially motivated by the problem of detecting periodic signals in X-ray astronomy data. In this case the time series consisted of individual photon arrival times where the appropriate sampling distribution is the Poisson distribution.

X-ray pulsars exhibit a wide range of pulse shapes.

Gregory-Loredo (GL) Method (Poisson case Ap.J. 398,1992)

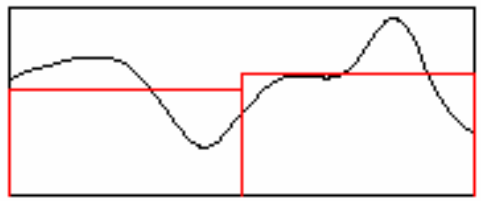
outline

To address the detection problem we compute the ratio of the probabilities (odds) of two models M_{Per} and M_1 .

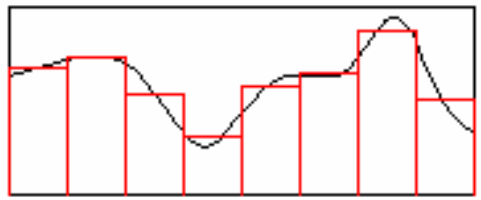
Model M_{per} is a family of periodic models capable of describing a background + periodic signal of arbitrary shape.

Each member of the family is a histogram with m bins where m > or = 2. Three examples are shown .

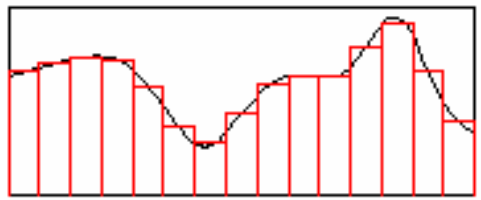
M_m represents the periodic model with m bins.



m = 2



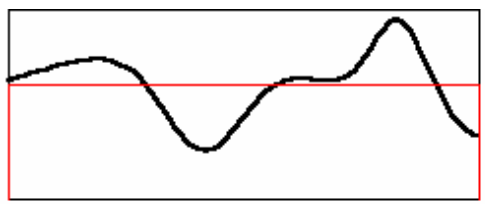
m = 8



m = 15

Model M_1 assumes the data is consistent with a constant event rate A.

M_1 is a special case of M_m with m = 1 bin



m = 1

GL Method continued

Search parameters for one of the m bin periodic models

unknown period, phase + m shape parameters

Special feature of the histogram models is that the search in the m shape parameters can be carried out analytically, permitting the method to be computationally tractable.

The Bayesian posterior probability for a particular periodic model, M_m, contains a term* which quantifies Occam's razor, penalizing successively more complicated periodic models (increasing m) for their greater complexity.

*** This term arises automatically in any Bayesian model comparison. It is not added in some a hoc fashion.**

The calculation thus balances model simplicity with goodness of fit, allowing us to determine both:

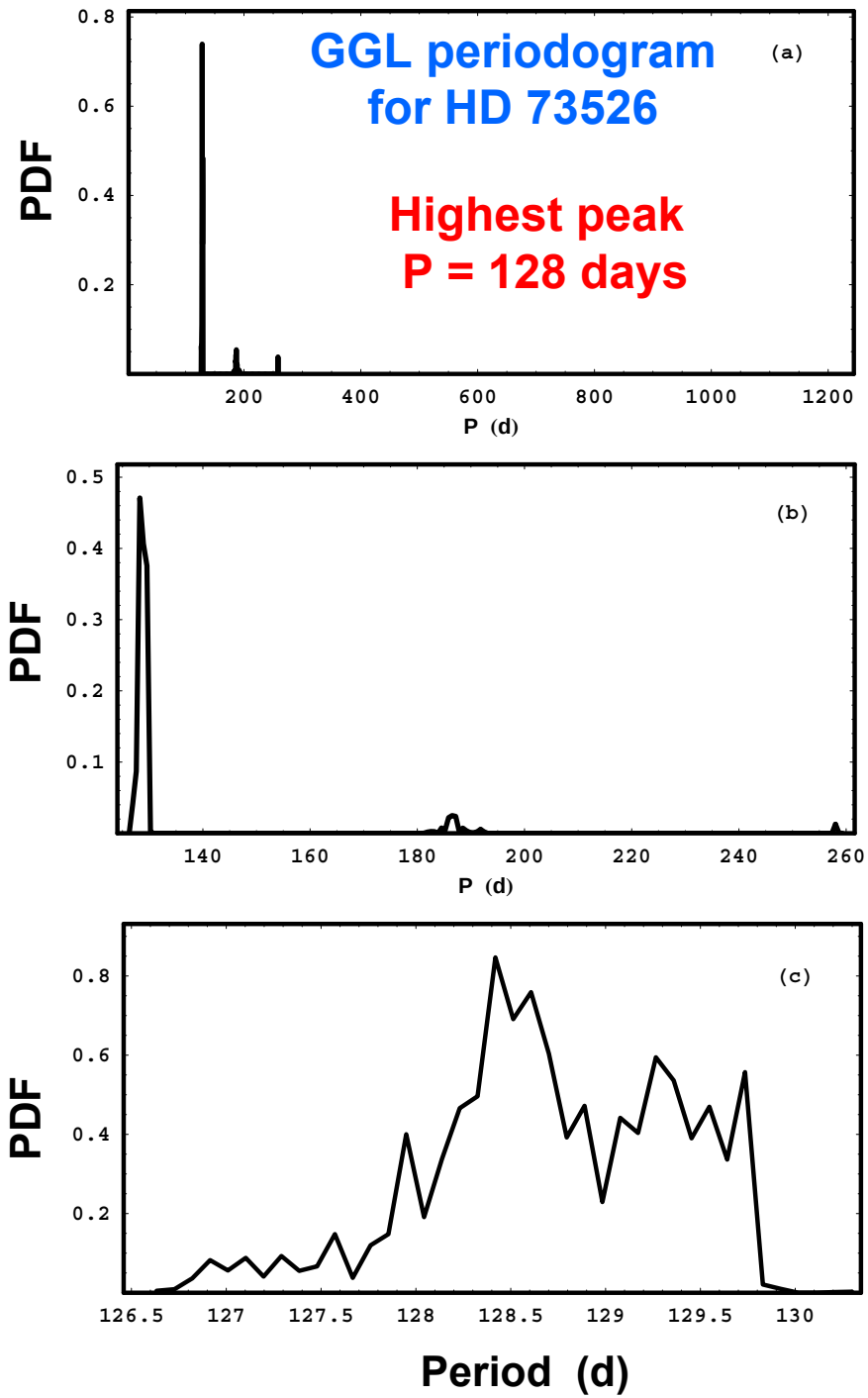
- **whether there is evidence for a periodic signal, and**
- **the optimum number of bins for describing the structure in the data.**

The Bayesian solution in the Poisson case is very satisfying:
In the absence of knowledge about the shape of the signal the method identifies the most organized (minimum entropy) significant periodic structure in the model parameter space. What structure is significant is determined through built in quantified Occam's penalties in the calculation.

The GGL periodogram[#] is a version of the original Gregory-Loredo algorithm modified for Gaussian uncertainties.

Both assume the shape of a periodic signal is unknown but can be modelled with a family of histograms with differing In the number of bins.

Allows for different Gaussian noise σ for each point.

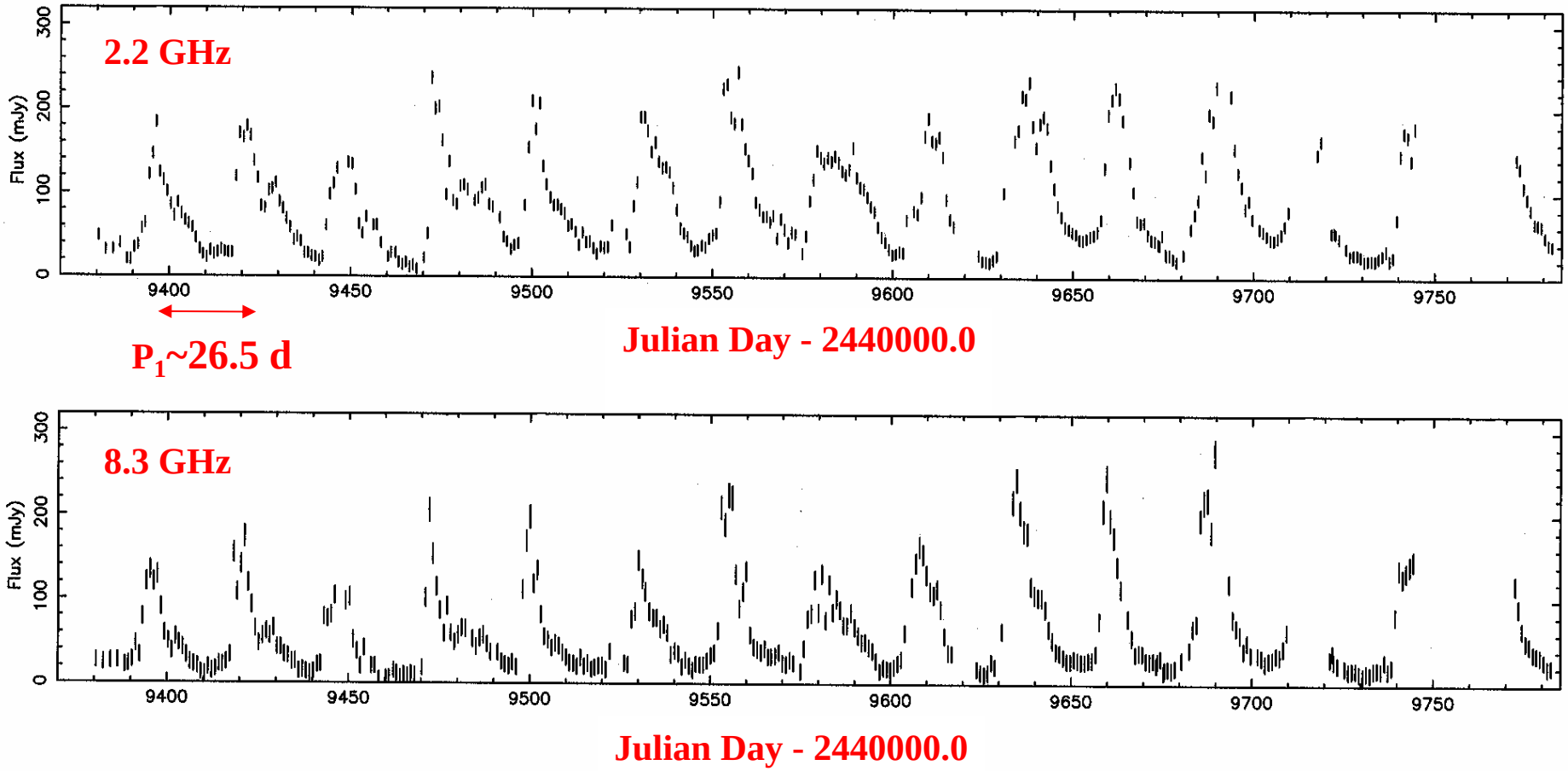


[#] Gregory, Ap. J. 520, pp. 361-375, 1999

Periodic Radio Outbursts from LSI+61^0 303

outline

Green Bank Interferometer-NASA-Naval Observatory Radio Monitoring



P_1~26.5 d

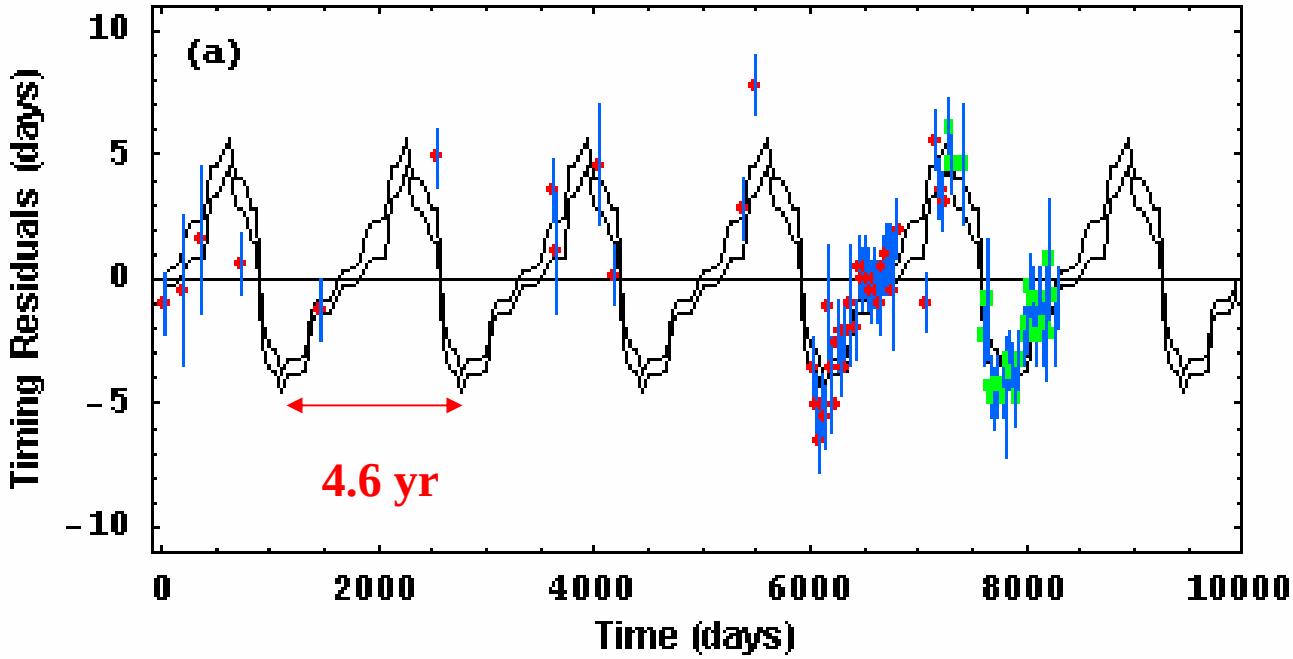
1 year

Return

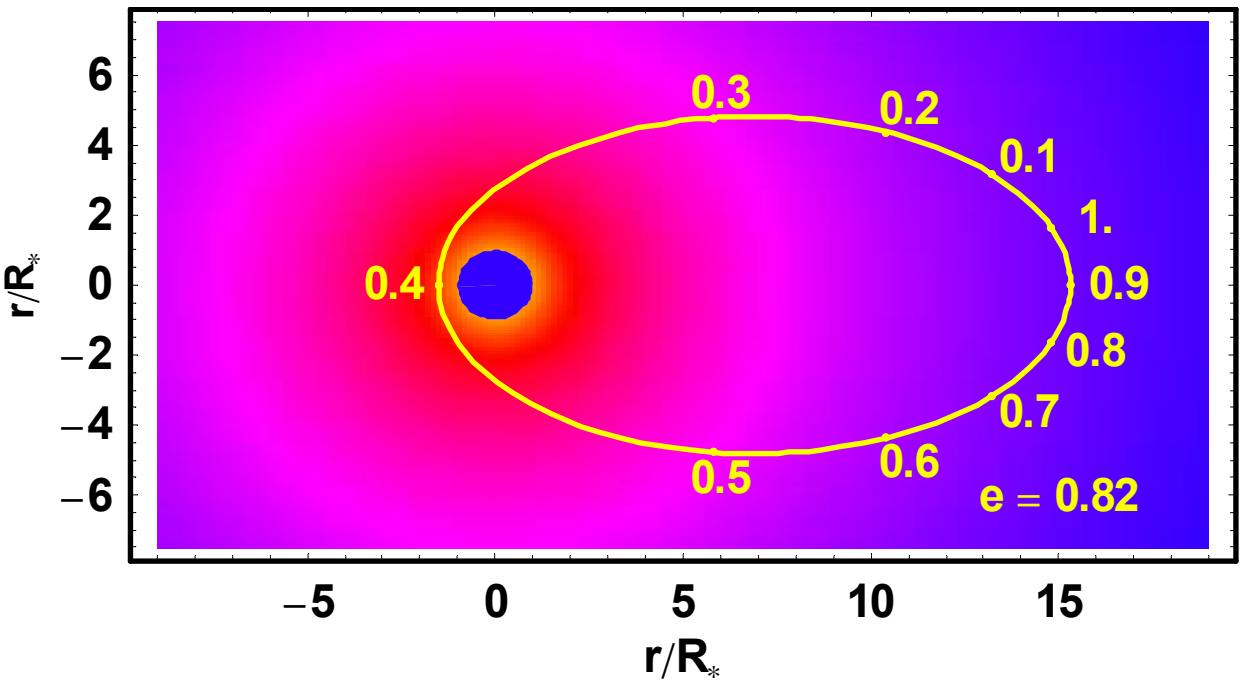
23 years of data since discovery in 1977

The last 6 years are from the GBI radio monitoring program

Discovery of a 4.6 year modulation of the phase of the radio outbursts



Outburst timing residuals span ~ 0.5 of an orbit



Conclusions

1. In these examples I have tried to demonstrate the power of the Bayesian approach in the arena of spectral analysis.
2. A Bayesian analysis can sometimes yield orders of magnitude improvement in model parameter estimation, by the inclusion of valuable prior information such as our knowledge of the signal model.
3. For some problems a Bayesian analysis may lead to a familiar statistic. Even in this situation (as we saw with the periodogram) it often leads to new insights concerning the interpretation and generalization of the statistic.
4. As a theory of extended logic it can be used to address the question of what is the optimal answer to a particular scientific question for a given state of knowledge, in contrast to a numerical recipe approach.
5. With all useful theoretical advances we expect to gain powerful new insights. This has already been amply demonstrated in the case of BPT and it is still early days.

One of these important new insights :

BPT provides a means of assessing competing theories at the forefront of science by quantifying Occam's razor,