TIME SERIES CENTER

Harvard University Initiative in Innovative Computing

I. Who

II. What

III. Plan

IV. Projects-Challenges

Pavlos Protopapas -CfA-IIC

Basic IDEA

Big idea, vague idea, promises?

Recipe

- Get data
- Get people that are interested in the science
- Get people with skills
- Get hardware

DATA-PROJECTS

Right now we have only astronomical data.

- **MACHO** - 66 million objects. 1000 flux observations per object in 2 bands (wavelengths)
- **SuperMACHO** - Close to a million objects. 100 flux observations per objects.
- **TAOS** - 100000 objects. 100K flux observations per object. 4 telescopes.
- **ESSENCE** - Thousands obejcts, hundred observations.
- **MPC** - Few hundred objects. Few hundred observations
- **Pan-STARRS**. Billions of objects. Hundred observations per object.

ASTRONOMY

Extra-solar planets. Either discovery of extra solar planet or statistical estimates of the abundance of planetary systems

Dark matter (Baryonic). Pan-STARRS will discover more lensing events in a single year than the combination of all monitoring programs which have been active to date. This is because it covers a larger area of the sky and goes deeper. Pan-STARRS data taken over an interval of several years can therefore provide the opportunity to derive reliable limits on Galactic dark matter.

Cosmology. SN from PanStarrs will help determine cosmological constants.

New class of variable star. Finding a new class or subclass of variable stars will be of tremendous value to astronomy.

Asteroids, KBO etc. Light curves can tell us about orbits, mass.

Understanding of the solar system. Killer asteroids.

COMPUTER SCIENSE-STATISTICS

- **Outlier/anomaly detection**
- **Clustering**
- **Identification of time series types**
- **Predicting properties of series**

In either case, analyzing a large data set requires efficient algorithms that scale linearly in the number of time series because even quadratic scaling incurs unrealistic run times.

– The feature space in which to represent the time series (Discrete Fourier Transform, Wavelets, Piecewise Linear, and symbolic methods)

– A distance metric for determining similarities in time series

COMPUTATIONAL QUESTIONS

The sizes of data sets in astronomy, medicine and other fields are presently exploding. The light curve center needs to be prepared for data rates starting in the 10's of gigabytes per night, scaling

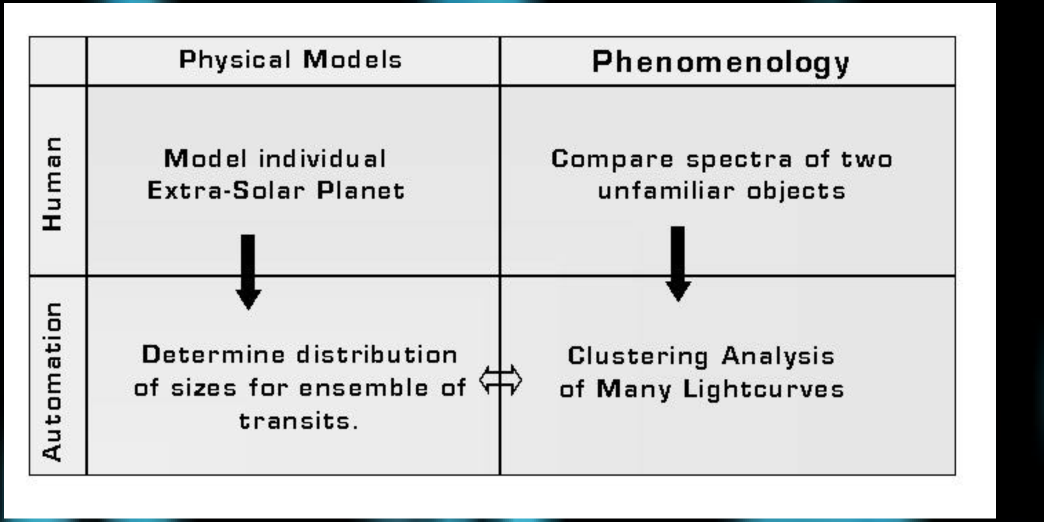
up to terabytes per night by the end of the decade.

Interplay between the algorithms used to study the time series, and the appropriate database indexing of the time series itself.

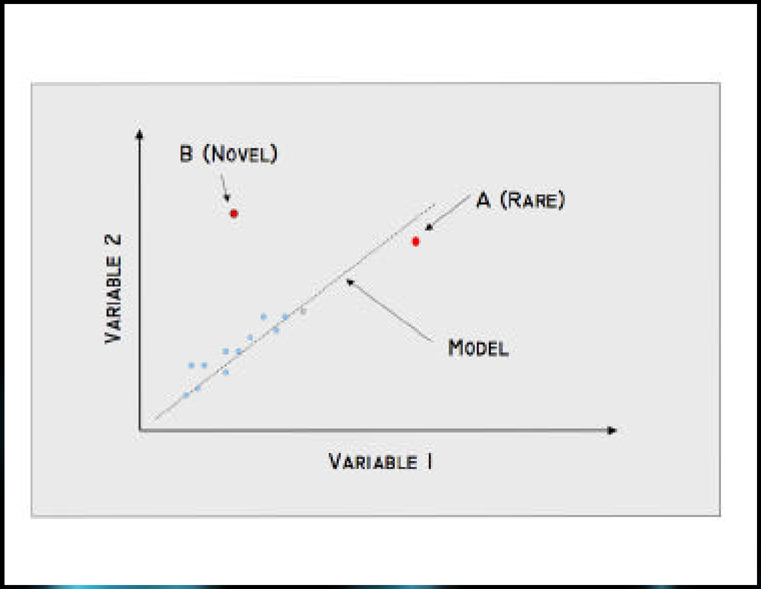
Real-time access

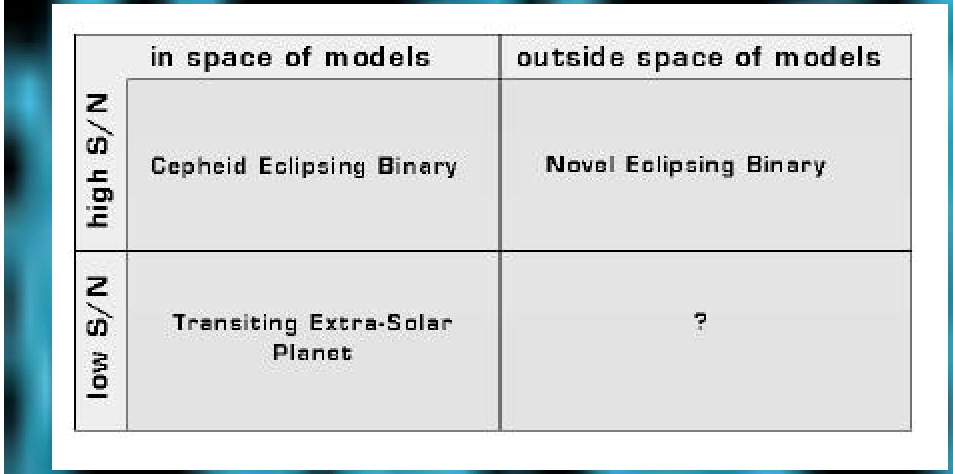
Distributed Computing

- VO standard.
- active query
- subscription

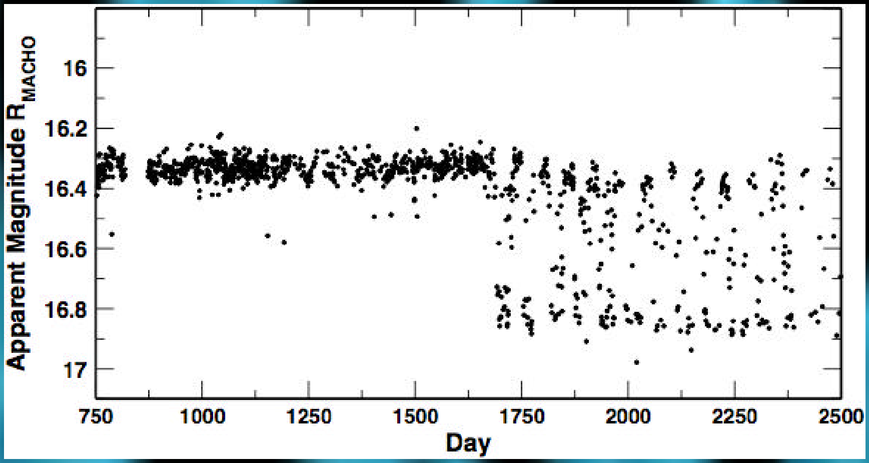


	Physical Models	Phenomenology
Human	Model individual Extra-Solar Planet	Compare spectra of two unfamiliar objects
Automation	Determine distribution of sizes for ensemble of transits. ⇔	Clustering Analysis of Many Lightcurves





	in space of models	outside space of models
high S/N	Cepheid Eclipsing Binary	Novel Eclipsing Binary
low S/N	Transiting Extra-Solar Planet	?



WHO

- Astronomers: C. Alcock, R. DiStefano, C. Stubbs, P. Protopapas
- CS: C. Brodley, R. Khardon, U. Rebbapragada
- Computational: R. Dave
- Statisticians: J. Rice

PLAN - KEY TO SUCCESS

DATA DATA DATA DATA.

Key to success is to get data that discoveries can be made.

All the kings algorithms and all the kings hardware can not put discoveries together.

PanStarrs is a key dataset.

Plan: 3 way

1. Get the data and parse them and made them available to people.
2. Prepare algorithms by CS
3. Prepare the questions by astronomers

DREAM

How about if the first earth like planet outside the solar system were discovered at IIC ?

How about if the first extra terrestrial life was detected from work at IIC ?

Dreaming ? There is as good chance to be part of this as anybody else.

Discoveries is the KEY

Projects underway

- Anomaly detection.
  1. Few outliers
  2. Class of outliers
- Extra Solar planets
- Temporal symmetries/asymmetries
- Binary Asteroids
- Microlensing searches
- Moving objects

Anomaly detection

- Only periodic light curves for now.

$$t \,?\, ?\frac{t \,?\, t_0}{T}?$$

- Need to worry about phase
- Define similarity. Pair wise correlation. Adjust for observational error

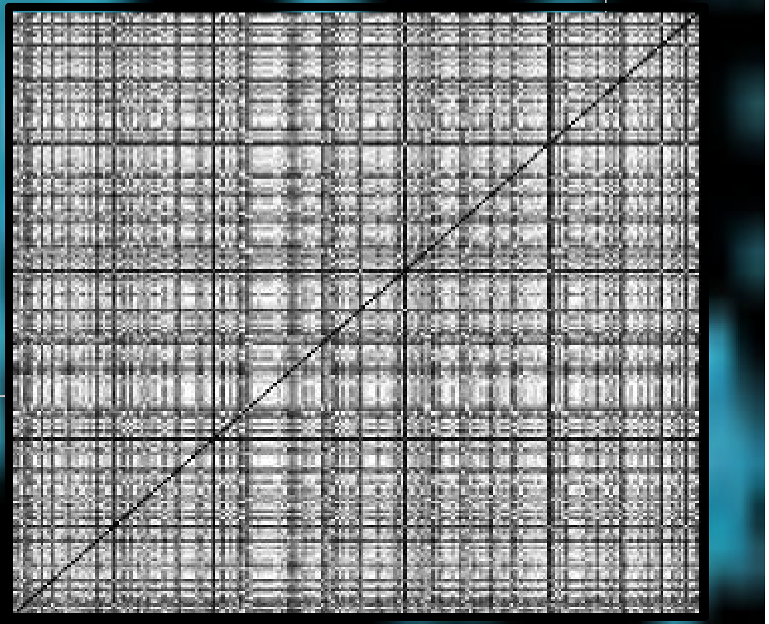
$$r_{ab} \,?\, \frac{\underset{n}{?} \,?\, f_n^A \,?\, \overline{f^A} \,?\, ?\, f_n^B \,?\, \overline{f^B} \,?}{\mathrm{var}(f^A)\mathrm{var}(f^B)}$$

Time warping method.

- Construct similarity matrix

- Construct similarity matrix
- Find outliers [weighted] averaging

  Question: How many and where to stop ?
- Extension 1: Compare to a centroid. Scales nicely but does not work well with not well define phase.
- Extension 2: Compare to multiple centroids. Redefine K-MEANS



difficulty: each pair has a an optimal relative phase.

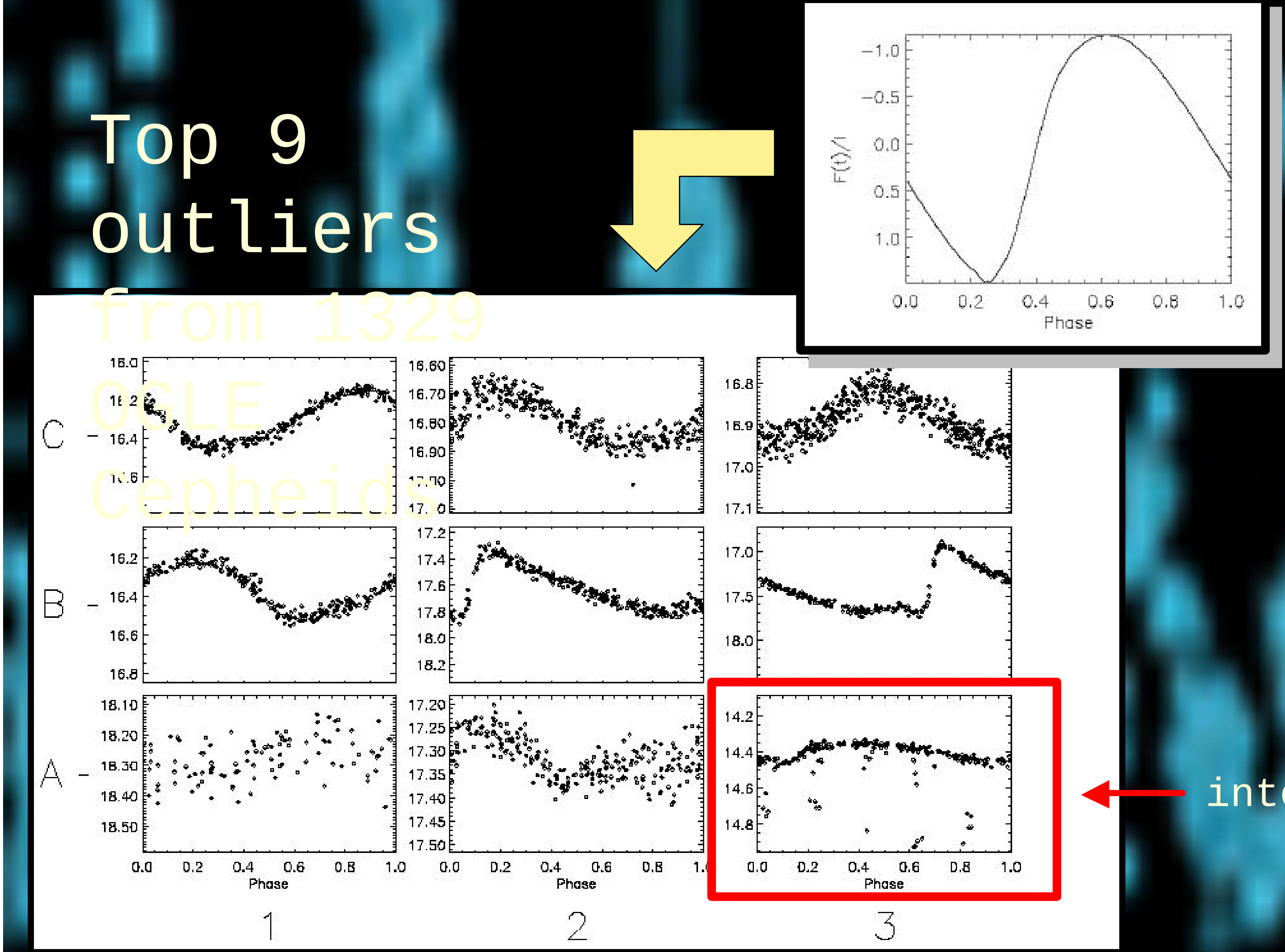
solution: Pk-means, which stands for Phased K-means, is a modification of the k-means clustering algorithm which takes into consideration the phasing of the time-series. Scales as O(N)

```
Algorithm 1 Pk-means(Lightcurves lc, Number of centroids)
Initialize centroids cen
while not Convergence do
        (closest_centroids, rephrased_lightcurves)?  CalcDistance(lc, cen)
        clusters ?  AssembleClusters(rephased_lightcurves,closest_centroids)
        centroids ?  RecalcCentroids(clusters)
end while
return centroids
```

```
Algorithm 2 CalcDistance(Lightcurves lc, Centroids cen)
for each lightcurve lc do
        for each centroid cen do
                (corr,phase) ?  CalcCorrelationUsingFFT(lc,cen)
                find max correlation ?  best phase, closest_centroid
        end for
        lc_phased ?  UpdatePhase(lc, best_phase)
end for
return closest_centroids, lc_phased
```

Cepheid centroid

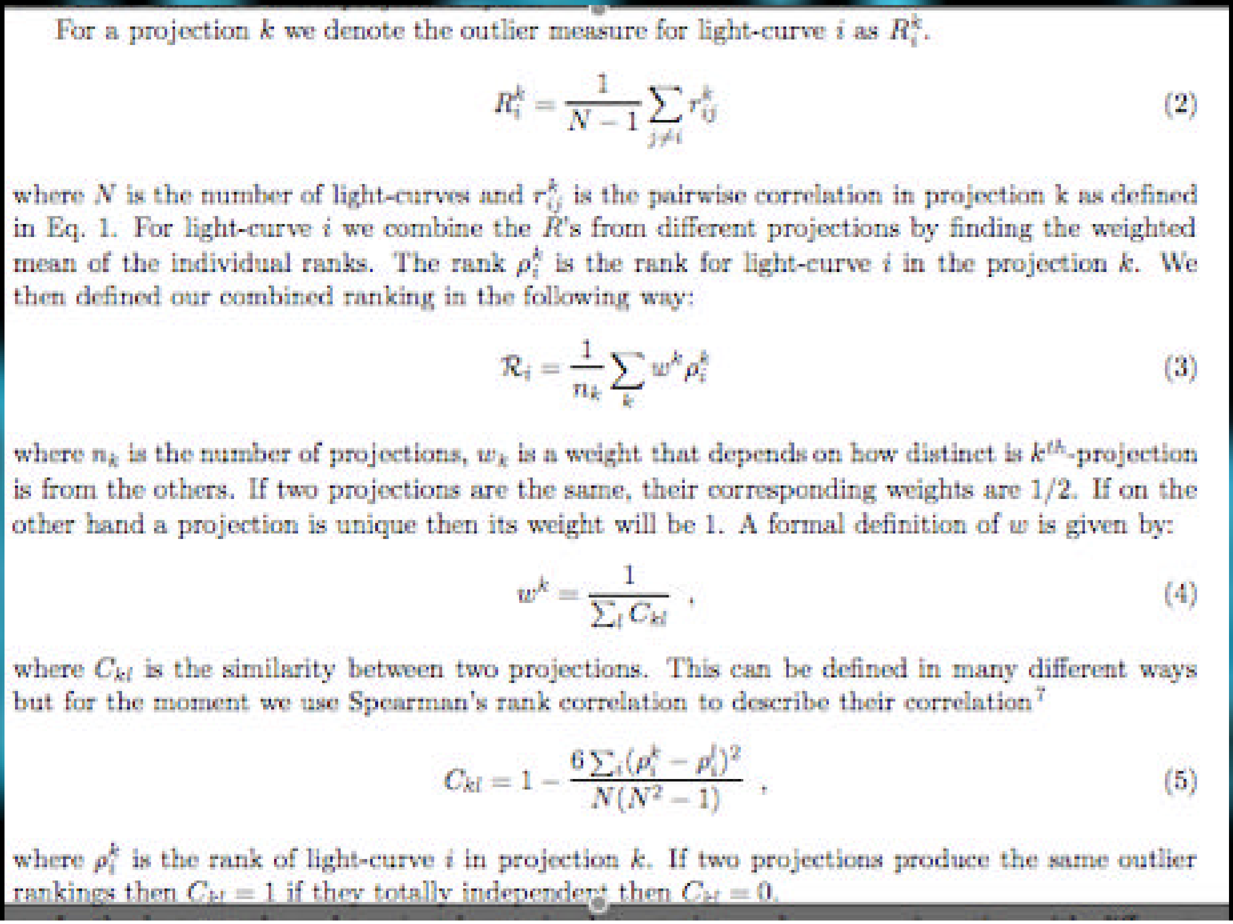
Top 9 outliers from 1329 OGLE Cepheids



interesting

Anomaly detection-EXTENSIONS

- Do the same not just with periodic light curves
- Different projections. Combine projections



For a projection k we denote the outlier measure for light-curve i as R_i^k.

$$R_i^k = \frac{1}{N-1} \sum_{j \neq i} r_{ij}^k \tag{2}$$

where N is the number of light-curves and r_{ij}^k is the pairwise correlation in projection k as defined in Eq. 1. For light-curve i we combine the R's from different projections by finding the weighted mean of the individual ranks. The rank ρ_i^k is the rank for light-curve i in the projection k. We then defined our combined ranking in the following way:

$$\mathcal{R}_i = \frac{1}{n_k} \sum_k w^k \rho_i^k \tag{3}$$

where n_k is the number of projections, w_k is a weight that depends on how distinct is k^{th}-projection is from the others. If two projections are the same, their corresponding weights are 1/2. If on the other hand a projection is unique then its weight will be 1. A formal definition of w is given by:

$$w^k = \frac{1}{\sum_l C_{kl}} \quad , \tag{4}$$

where C_{kl} is the similarity between two projections. This can be defined in many different ways but for the moment we use Spearman's rank correlation to describe their correlation[7]

$$C_{kl} = 1 - \frac{6 \sum_i (\rho_i^k - \rho_i^l)^2}{N(N^2-1)} \quad , \tag{5}$$

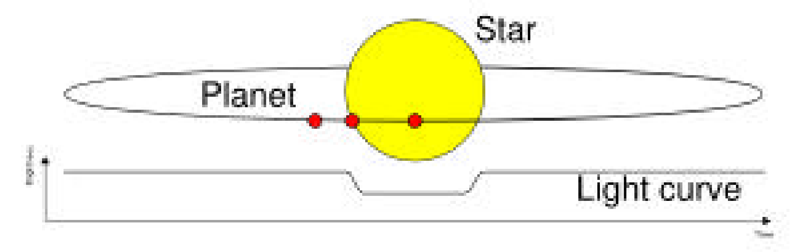
where ρ_i^k is the rank of light-curve i in projection k. If two projections produce the same outlier rankings then $C_{kl} = 1$ if they totally independent then $C_{kl} = 0$.

- Do the same not just with periodic light curves
- Different projections. Combine projections

- Find outlier clusters. Redefine “outliers”.
- Clustering methods.
- Define variability. Need a statistical test of variability. I am using wavelet decomposition. All coefficients must be zero.

Transit method-Extra solar planets searches

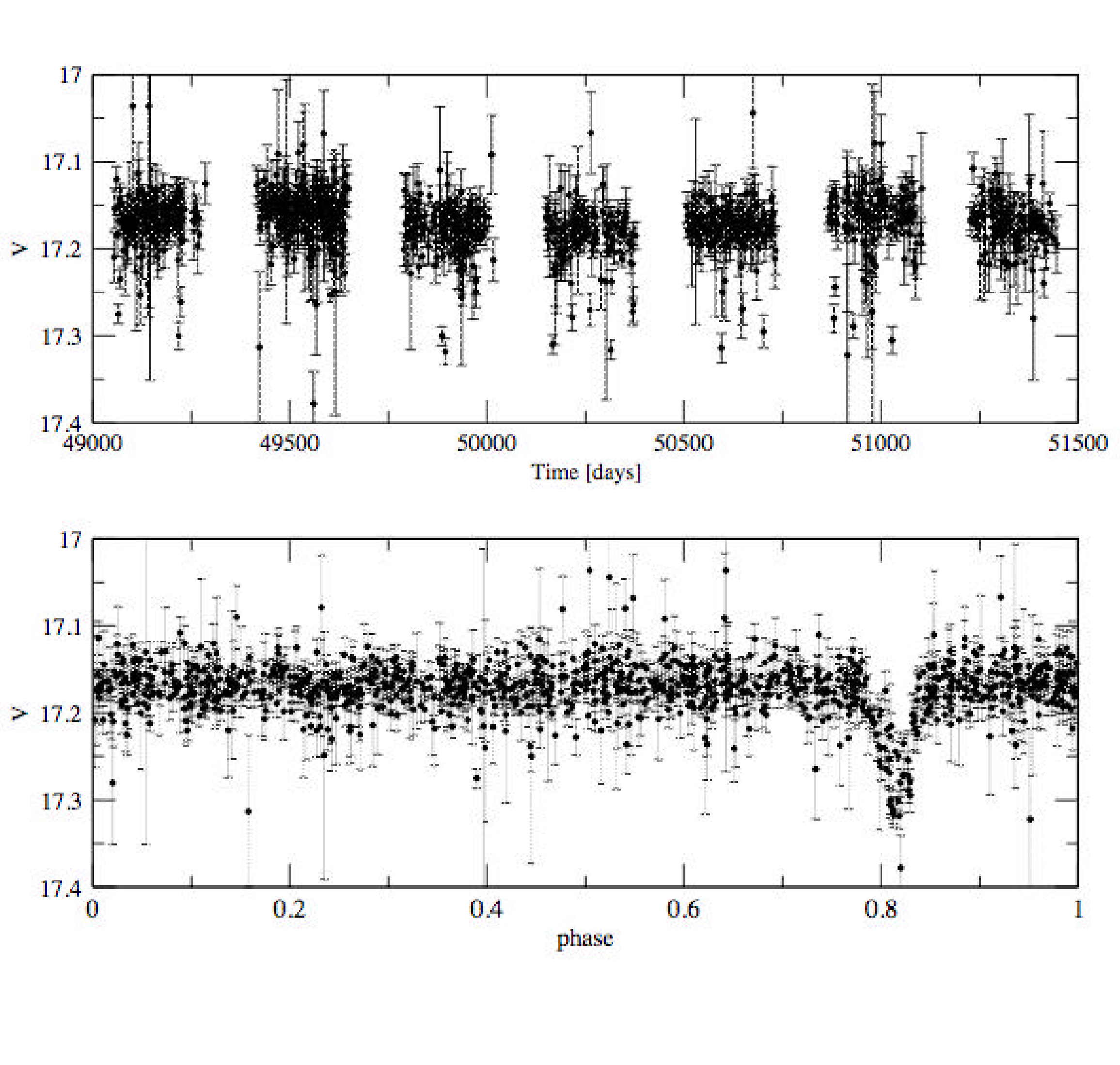
Looking for planets at other solar systems. Transit method when a planet goes in front of the star the light from the star is blocked.



Our job is to confirm that.

If the survey is designed for transit searches then the problem is simple. If not then the likelihood surface is erratic.

? ?typical light curve with non optimal sampling may look like anything



Multiple Optimized Parameter Estimation and Data Compression MOPED

Method to compress data by Heavens et al. (2000)

Given data x (our case a light-curve) which includes a signal part μ and a noise n

$$x ? ? ? n$$

The idea is to find weighting vector b_m (m runs from 1 to number of parameters)

$$y_m ? b_m x$$

that contains as much information as possible about the parameters (period, duration of the transit etc.).

These numbers y_m are then used as the data set in a likelihood analysis with the consequent increase in speed at finding the best solution. In MOPED, there is one vector associated with each parameter.

MOPED

Find the proper weights such as the transformation is lossless.

Lossless is defined as the Fisher matrix remains unchanged at the maximum likelihood.

The Fisher matrix is defined by:

$$\mathrm{F}_{ab} \; ? \; -\left\langle \frac{?^{2} \ln L}{??_{a} ??_{b}} \right\rangle$$

The posterior probability for the parameters is the likelihood, which for Gaussian noise is (alas needs to be Gaussian)

$$L(?_{a}) \; ? \; \exp \begin{matrix}?\\?\\?\end{matrix}? \frac{1}{2} \underset{i,j}{?} (x_{i} \; ? \; ?_{i}) C_{ij}^{?1} (x_{j} \; ? \; ?_{j}) \begin{matrix}?\\?\\?\end{matrix}$$

If we had the correct parameters then this can be shown to be perfectly lossless. Of course we can not know the answer a priory. Nevertheless Heavens et al (2000) show that when the weights are appropriate chosen the solution is still accurate.

MOPED

The weights are (complicated as it is)

$$b_1 \ ? \ \frac{C^{?1}?_{,1}}{\sqrt{?^t_{,1}C^{?1}?_{,1}}}$$

and

$$\mathrm{b_m} \ ? \ \frac{C^{?1}?_{,m} \ ? \ \mathop{?}\limits_{q?1}^{m?1}(?^t_{,m}b_q)b_q}{\sqrt{?^t_{,m}C^{?1}?_{,m} \ ? \ \mathop{?}\limits_{q?1}^{m?1}(?^t_{,m}b_q)^2}}$$

Where comma denotes derivatives.

Note:

C is the covariance matrix and depends on the data

? is the model and it depends on the parameters.

Need to choose a fiducial model for that

MOPED

Now what we do with that? Write the new likelihood

$$\ln L(?_{a}) ? \; ? \underset{i,j}{?} ?b_{m}(q_{f}) ?x ? b_{m}(q_{f}) ?? (q) ?$$

Where q_f is the fiducial model and q is the model we are trying out.

We choose q and calculate the log likelihood in this new space.

WHY ?

If the covariant matrix is known (or stays significantly same) then the second term needs to be computed only once for the whole dataset (because it depends on fiducial model and trial models)

So for each light-curve I compute the dot-product and subtract.

But there is more (do not run away)

Transit models

We need to choose a model for our transits

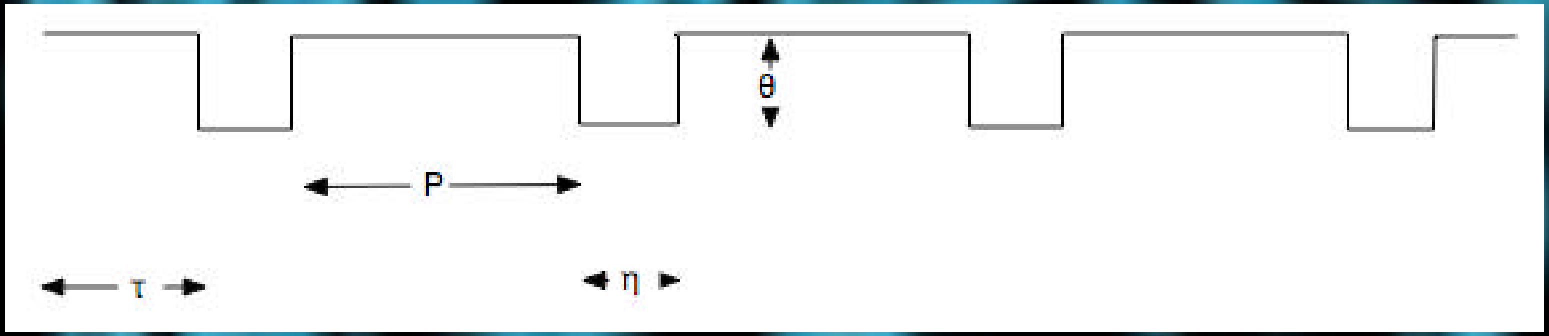
Four free parameters:

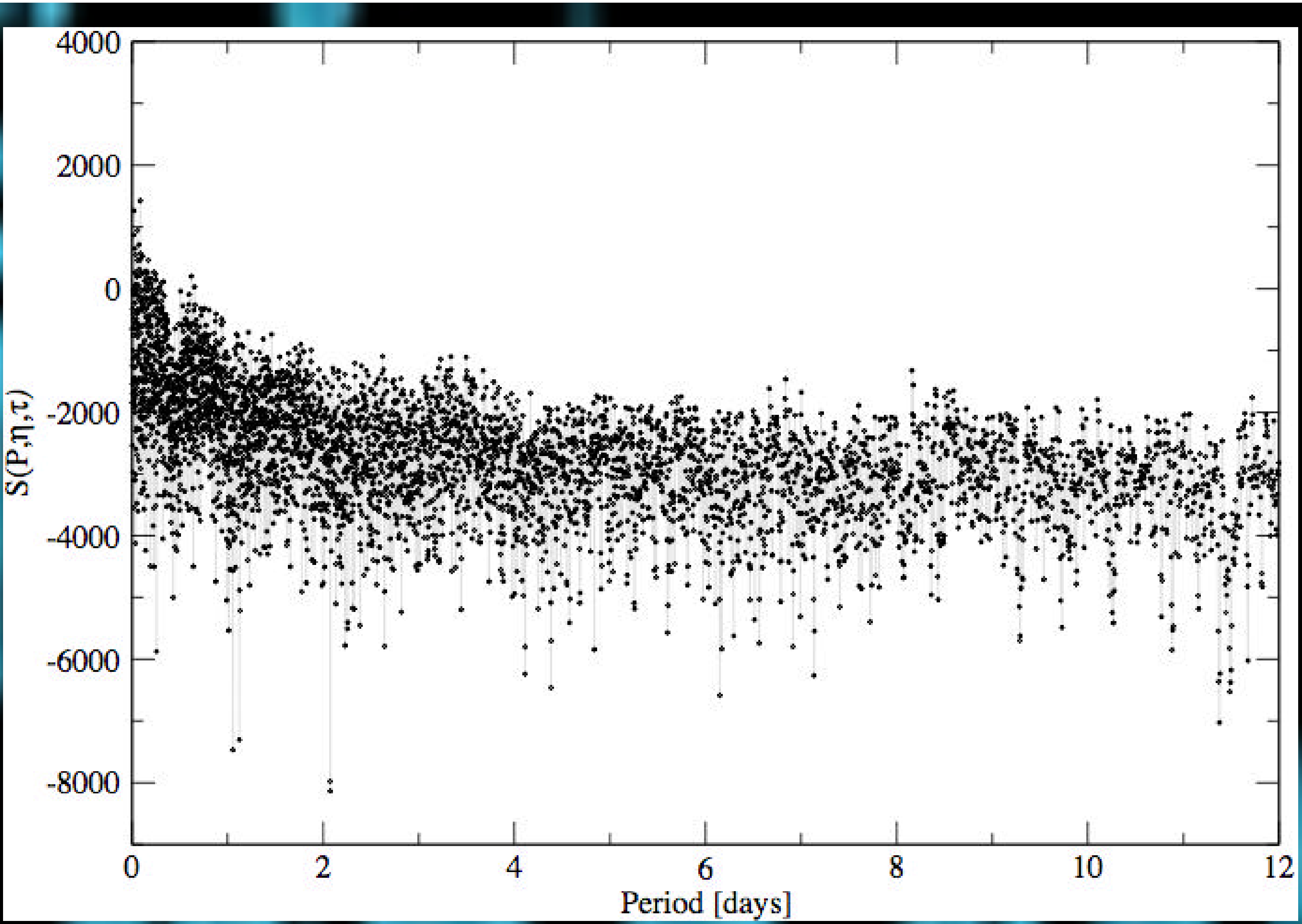
1. Period, P
2. Depth, ?

?? Duration, ?

?? Epoch, ?

Note: A more realistic model can easily be made using tanh




4000
2000
0
-2000
-4000
-6000
-8000
S(P,η,τ)
0
2
4
6
8
10
12
Period [days]

Multiple Fiducial models

For an arbitrary fiducial model the likelihood function will have several maxima/minima.

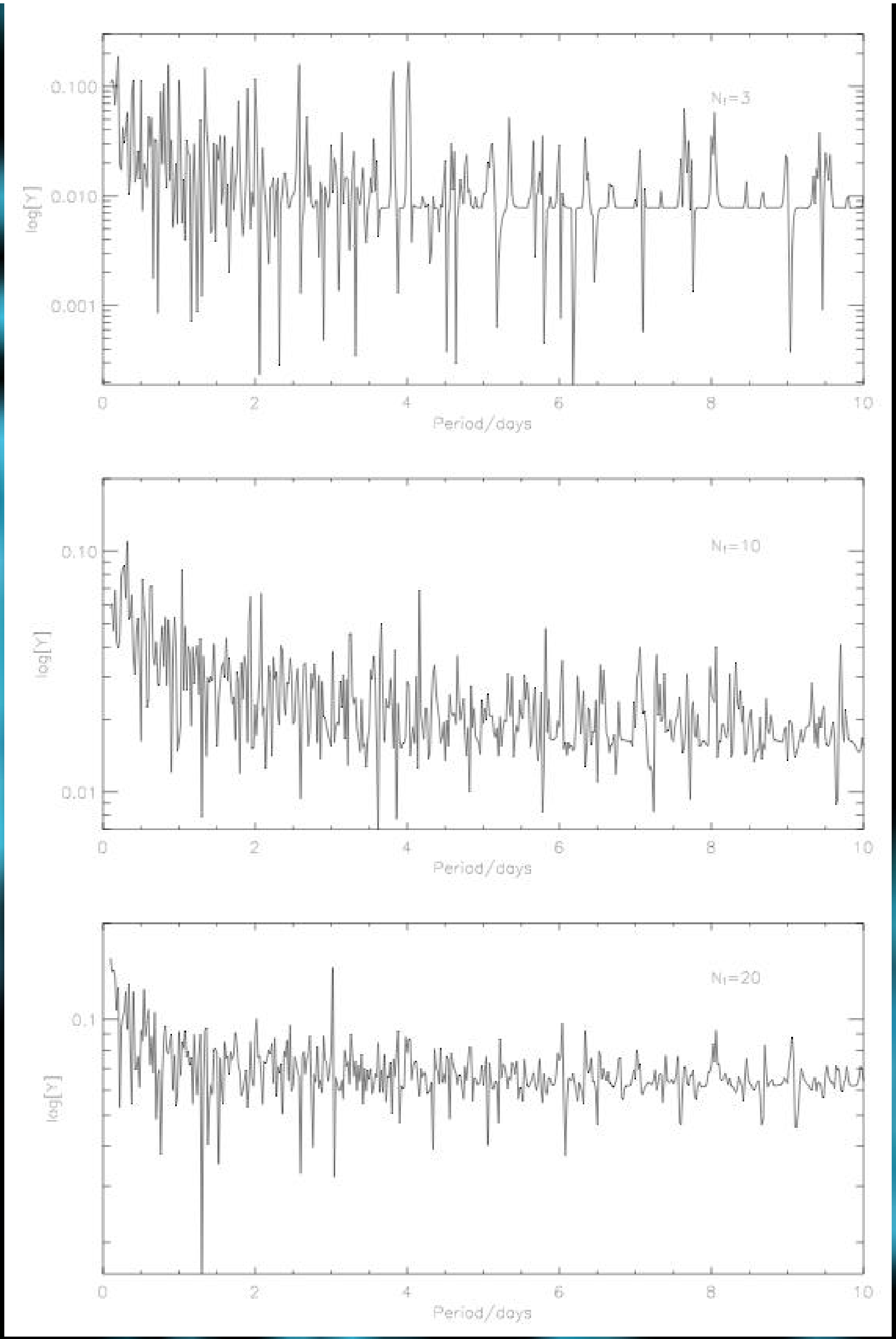
One of those maxima is guaranteed to be the true one. If there was no noise this would have been exact.

For an another fiducial model there again several maxima/minima. One of those maxima is guaranteed to be the true one

Combine several fiducial models and eliminate all but the true solutions.

We define a new measure

$$Y ? \frac{1}{N_f} \underset{\{q_f\}}{?} L$$



Y as a function of period.

First panel is after 3 fiducial models

Second panel after 10 fiducial models

Third panel after 20 fiducial models.

Synthetic light curves:

One 5 measurements/hour, total of 4000 measurements.

S/N=5

Confidence levels

Assume Gaussian error (can be done with Poisson)

No transit signal

Y follows a non-central $?^2$ distribution mean and variance

$$? \; ? \; r \; ? \; ?$$

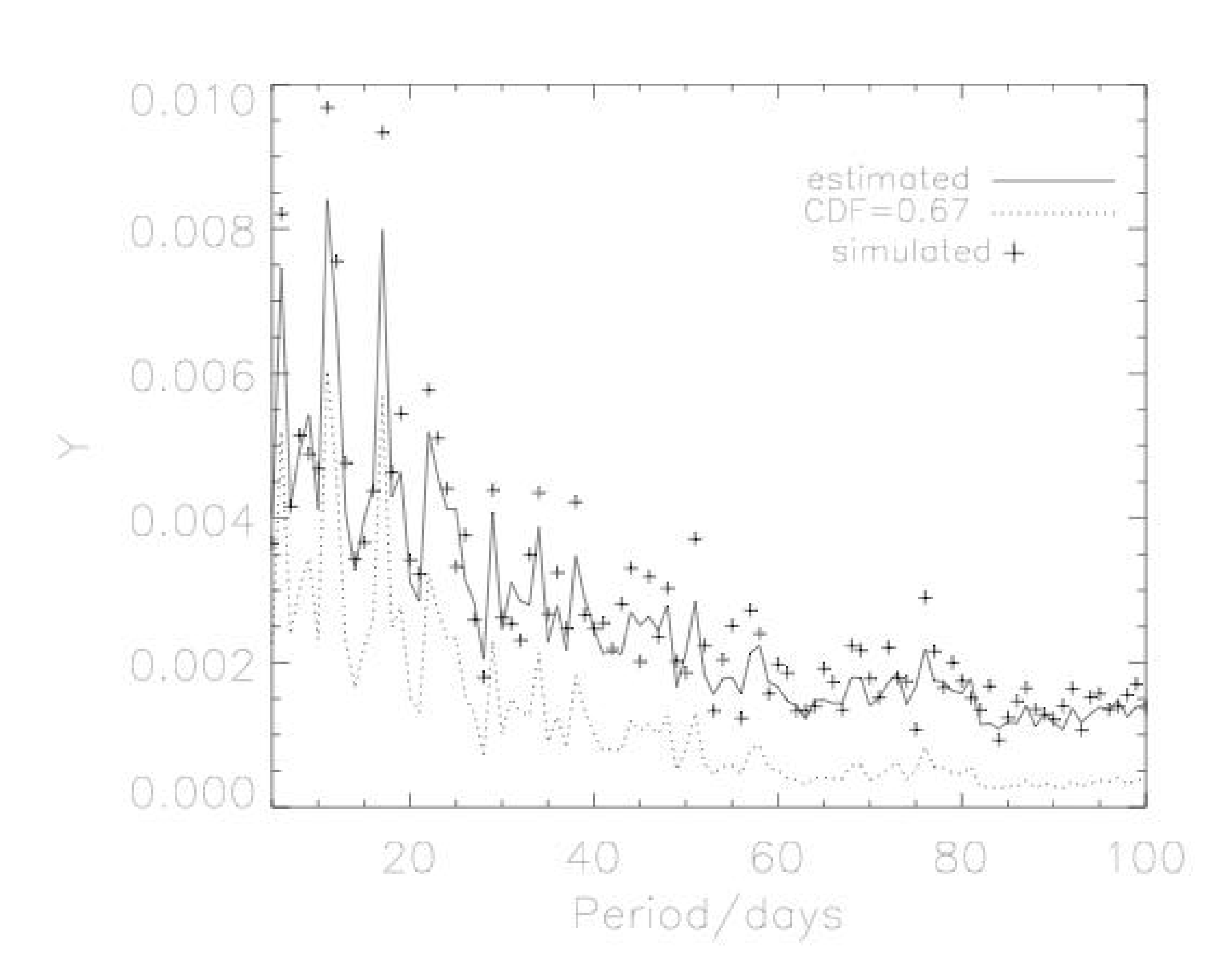
$$?^2 \; ? \; 2(r \; ? \; 2?)$$

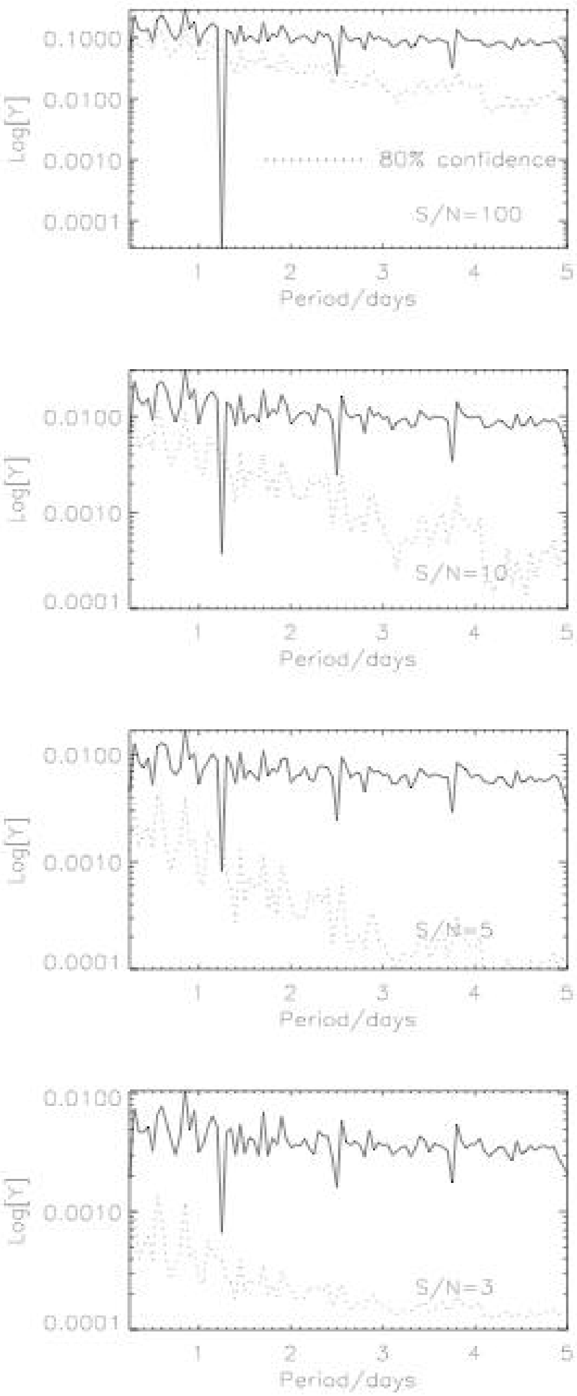
$$r \; ? \; 4$$

$$? \; ? \; \frac{E^2 ?? ?}{\mathrm{var} ?? ?}$$

We can estimate the error and thus the confidence of our results. But before lets make sure that I did this right.

Estimated vs. Real (simulated) Y for the null case





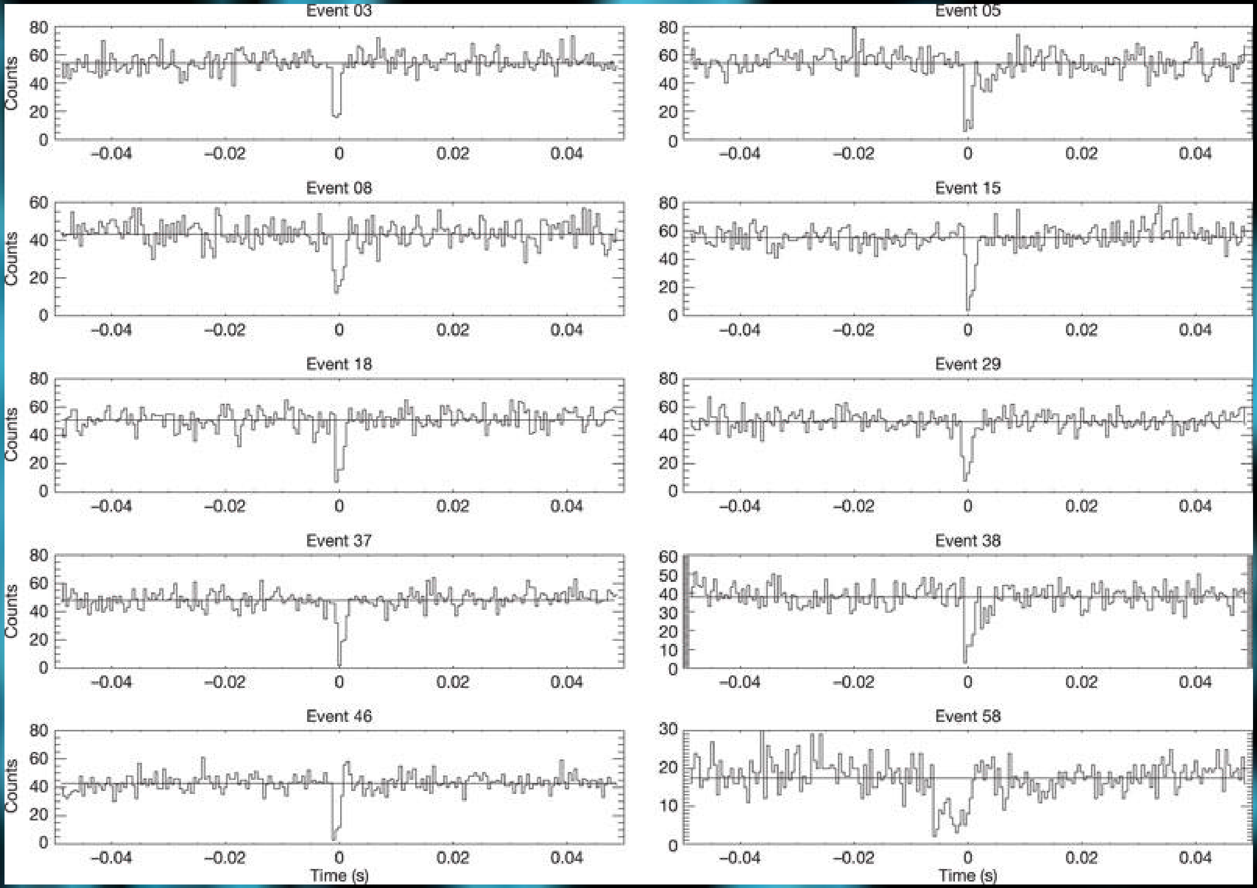
Y as a function of period for synthetic light curve.

Each panel shows different S/N

Dotted line shows 80% confidence level.

KBO-Temporal Symmetry/Assymetry

Hsiang-Kuang Chang, Sun-Kun King, Jau-Shian Liang, Ping-Shien Wu, Lupin Chun-Che Lin and Jeng-Lun Chiu, Nature 442, 660-663(10 August 2006)
X RAY data from RXTE (high time resolution data) from SCO-X1 (the second brighter x-rays source)
A trans-Neptunian object passes in front of a star, thus occulting the light.



Looking for a statistical test for temporal asymmetry

My method (under development). Assume time symmetry at ?

$$f^{A} ? s^{A} ? n$$

$$f^{B} ? s^{B} ? n$$

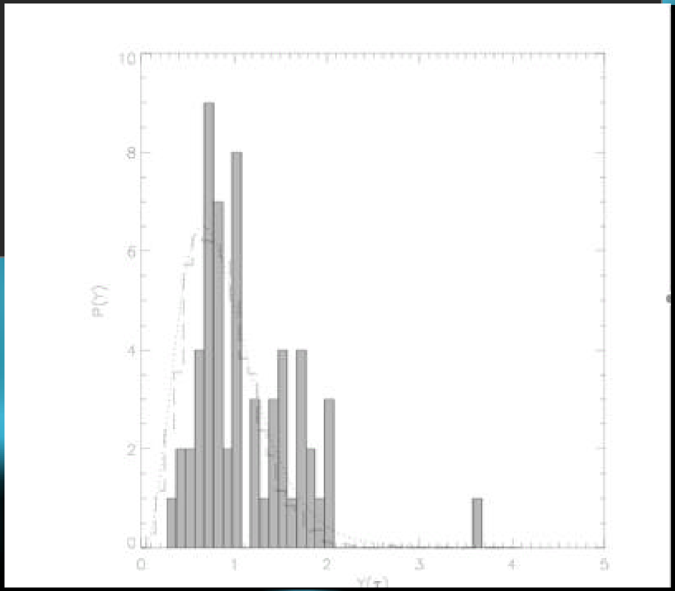
$$Q(?) ? \frac{1}{n ? 1} \sum_{i ? 1}^{n} \frac{? f_i^{A}(?) ? f_i^{2}(?) ?}{?^{A^2} ? ?^{B^2}}$$

If symmetric then Q follows a chi-square distribution.

Assume errors are Gaussian

They are definitely not !

QUESTIONS: What do I do ?



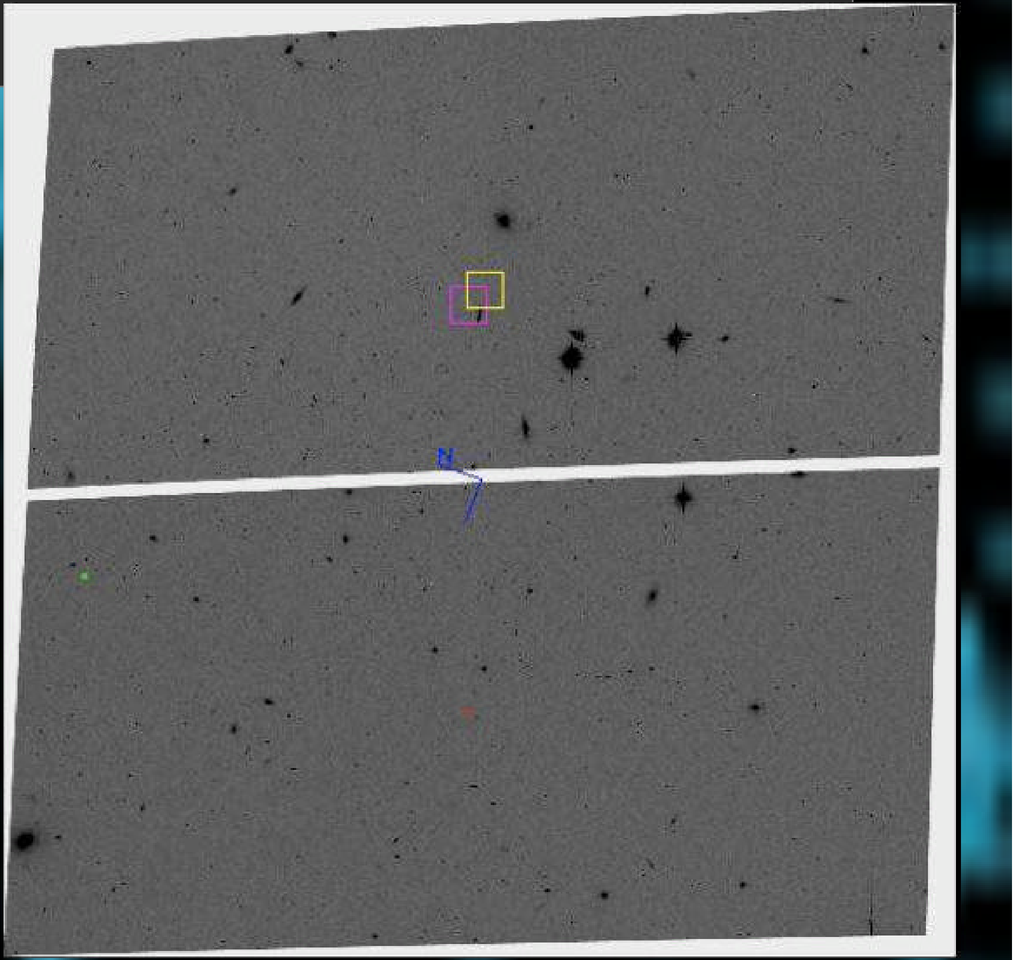
Binary Asteroids.

Looking for binary asteroids.

Look for tracks in the HST archive

Bayesian approach !

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.



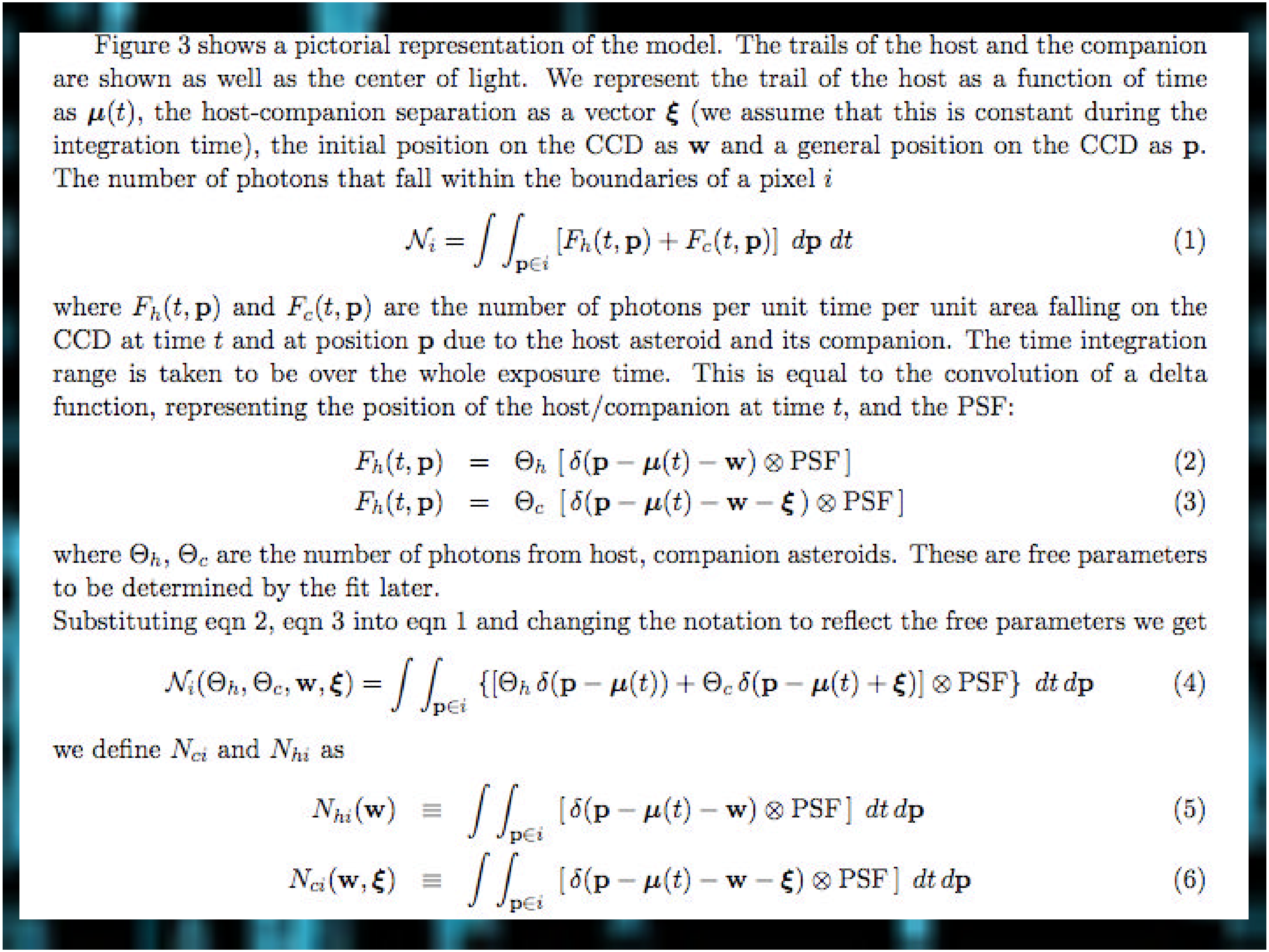


Figure 3 shows a pictorial representation of the model. The trails of the host and the companion are shown as well as the center of light. We represent the trail of the host as a function of time as $\boldsymbol{\mu}(t)$, the host-companion separation as a vector $\boldsymbol{\xi}$ (we assume that this is constant during the integration time), the initial position on the CCD as $\mathbf{w}$ and a general position on the CCD as $\mathbf{p}$. The number of photons that fall within the boundaries of a pixel i

$$\mathcal{N}_i = \int \int_{\mathbf{p}\in i} \left[F_h(t,\mathbf{p}) + F_c(t,\mathbf{p})\right] \, d\mathbf{p} \, dt \tag{1}$$

where $F_h(t,\mathbf{p})$ and $F_c(t,\mathbf{p})$ are the number of photons per unit time per unit area falling on the CCD at time t and at position $\mathbf{p}$ due to the host asteroid and its companion. The time integration range is taken to be over the whole exposure time. This is equal to the convolution of a delta function, representing the position of the host/companion at time t, and the PSF:

$$\begin{aligned} F_h(t,\mathbf{p}) &= \Theta_h \left[\delta(\mathbf{p} - \boldsymbol{\mu}(t) - \mathbf{w}) \otimes \mathrm{PSF}\right] & (2) \\ F_h(t,\mathbf{p}) &= \Theta_c \left[\delta(\mathbf{p} - \boldsymbol{\mu}(t) - \mathbf{w} - \boldsymbol{\xi}\,) \otimes \mathrm{PSF}\right] & (3) \end{aligned}$$

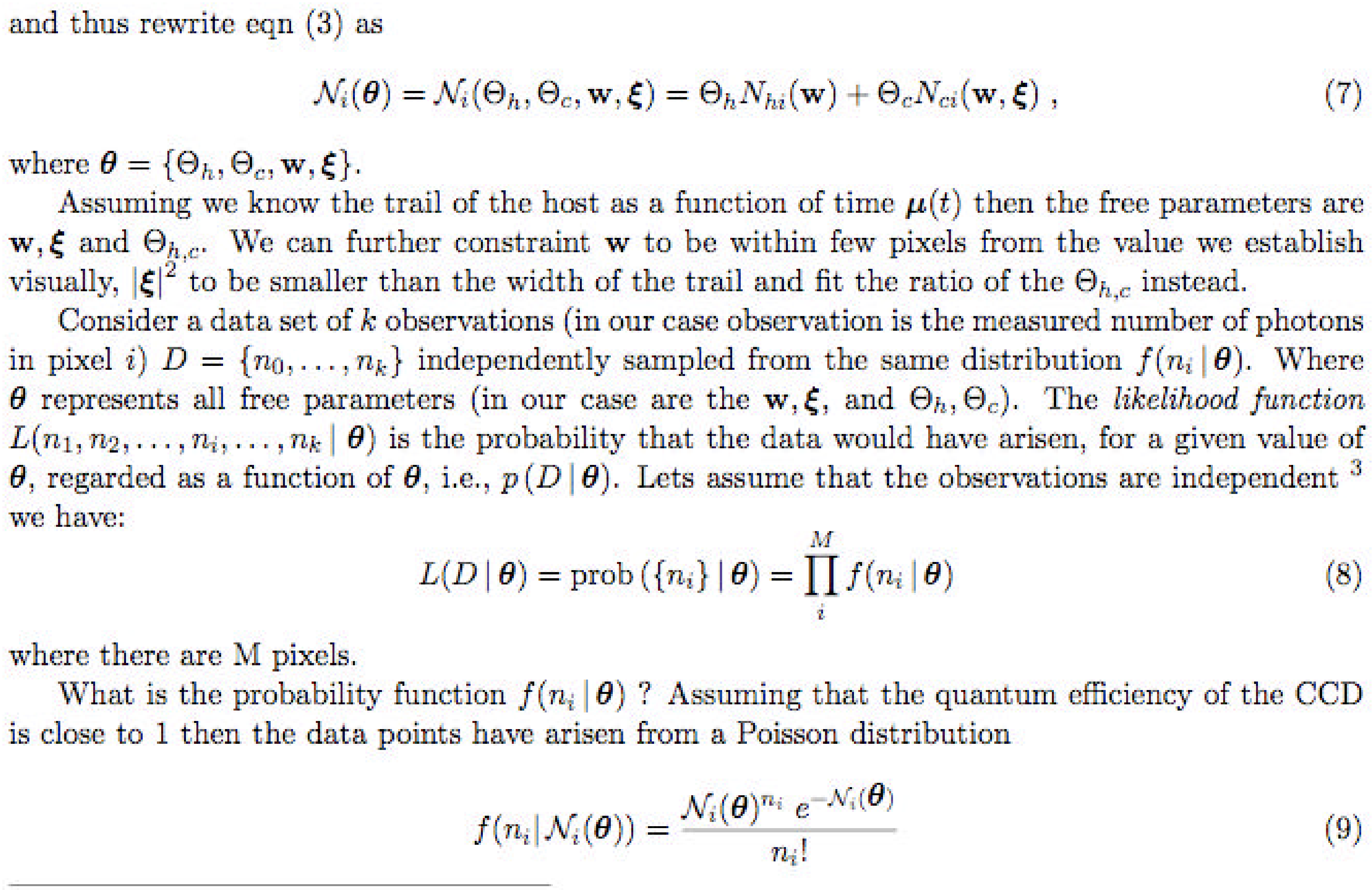
where Θ_h, Θ_c are the number of photons from host, companion asteroids. These are free parameters to be determined by the fit later.

Substituting eqn 2, eqn 3 into eqn 1 and changing the notation to reflect the free parameters we get

$$\mathcal{N}_i(\Theta_h, \Theta_c, \mathbf{w}, \boldsymbol{\xi}) = \int \int_{\mathbf{p}\in i} \left\{\left[\Theta_h \, \delta(\mathbf{p} - \boldsymbol{\mu}(t)) + \Theta_c \, \delta(\mathbf{p} - \boldsymbol{\mu}(t) + \boldsymbol{\xi})\right] \otimes \mathrm{PSF}\right\} \, dt \, d\mathbf{p} \tag{4}$$

we define N_{ci} and N_{hi} as

$$\begin{aligned} N_{hi}(\mathbf{w}) &\equiv \int \int_{\mathbf{p}\in i} \left[\delta(\mathbf{p} - \boldsymbol{\mu}(t) - \mathbf{w}) \otimes \mathrm{PSF}\right] \, dt \, d\mathbf{p} & (5) \\ N_{ci}(\mathbf{w}, \boldsymbol{\xi}) &\equiv \int \int_{\mathbf{p}\in i} \left[\delta(\mathbf{p} - \boldsymbol{\mu}(t) - \mathbf{w} - \boldsymbol{\xi}) \otimes \mathrm{PSF}\right] \, dt \, d\mathbf{p} & (6) \end{aligned}$$



and thus rewrite eqn (3) as

$$\mathcal{N}_i(\boldsymbol{\theta}) = \mathcal{N}_i(\Theta_h, \Theta_c, \mathbf{w}, \boldsymbol{\xi}) = \Theta_h N_{hi}(\mathbf{w}) + \Theta_c N_{ci}(\mathbf{w}, \boldsymbol{\xi}) \ , \tag{7}$$

where $\boldsymbol{\theta} = \{\Theta_h, \Theta_c, \mathbf{w}, \boldsymbol{\xi}\}$.

Assuming we know the trail of the host as a function of time $\boldsymbol{\mu}(t)$ then the free parameters are $\mathbf{w}, \boldsymbol{\xi}$ and $\Theta_{h,c}$. We can further constraint $\mathbf{w}$ to be within few pixels from the value we establish visually, $|\boldsymbol{\xi}|^2$ to be smaller than the width of the trail and fit the ratio of the $\Theta_{h,c}$ instead.

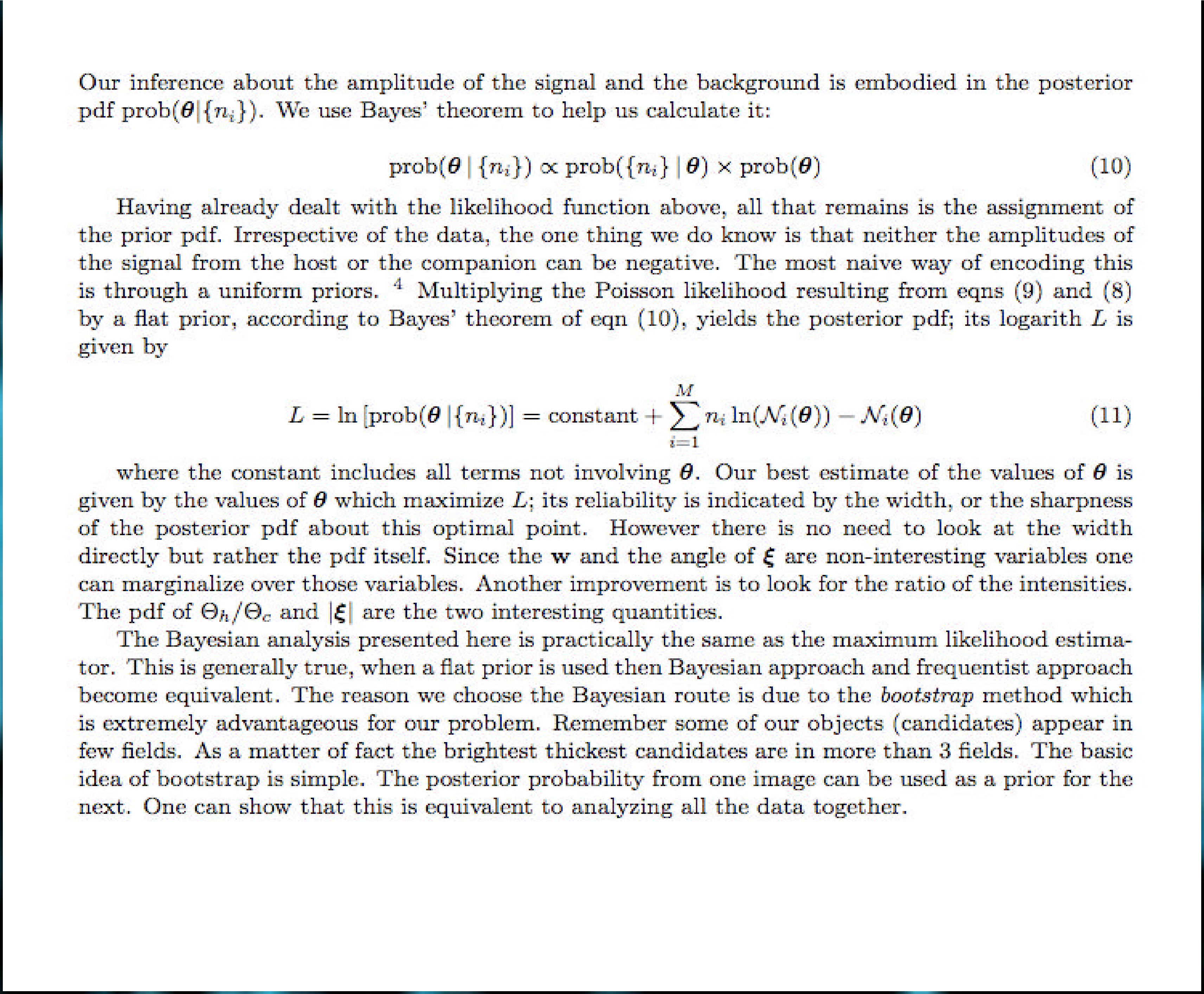
Consider a data set of k observations (in our case observation is the measured number of photons in pixel i) $D = \{n_0, \ldots, n_k\}$ independently sampled from the same distribution $f(n_i \,|\, \boldsymbol{\theta})$. Where $\boldsymbol{\theta}$ represents all free parameters (in our case are the $\mathbf{w}, \boldsymbol{\xi}$, and Θ_h, Θ_c). The *likelihood function* $L(n_1, n_2, \ldots, n_i, \ldots, n_k \,|\, \boldsymbol{\theta})$ is the probability that the data would have arisen, for a given value of $\boldsymbol{\theta}$, regarded as a function of $\boldsymbol{\theta}$, i.e., $p\,(D \,|\, \boldsymbol{\theta})$. Lets assume that the observations are independent [3] we have:

$$L(D \,|\, \boldsymbol{\theta}) = \mathrm{prob}\,(\{n_i\} \,|\, \boldsymbol{\theta}) = \prod_i^M f(n_i \,|\, \boldsymbol{\theta}) \tag{8}$$

where there are M pixels.

What is the probability function $f(n_i \,|\, \boldsymbol{\theta})$? Assuming that the quantum efficiency of the CCD is close to 1 then the data points have arisen from a Poisson distribution

$$f(n_i | \mathcal{N}_i(\boldsymbol{\theta})) = \frac{\mathcal{N}_i(\boldsymbol{\theta})^{n_i} \; e^{-\mathcal{N}_i(\boldsymbol{\theta})}}{n_i!} \tag{9}$$



Our inference about the amplitude of the signal and the background is embodied in the posterior pdf prob($\boldsymbol{\theta}|\{n_i\}$). We use Bayes' theorem to help us calculate it:

$$\text{prob}(\boldsymbol{\theta} \mid \{n_i\}) \propto \text{prob}(\{n_i\} \mid \boldsymbol{\theta}) \times \text{prob}(\boldsymbol{\theta}) \tag{10}$$

Having already dealt with the likelihood function above, all that remains is the assignment of the prior pdf. Irrespective of the data, the one thing we do know is that neither the amplitudes of the signal from the host or the companion can be negative. The most naive way of encoding this is through a uniform priors. [4] Multiplying the Poisson likelihood resulting from eqns (9) and (8) by a flat prior, according to Bayes' theorem of eqn (10), yields the posterior pdf; its logarith L is given by

$$L = \ln\left[\text{prob}(\boldsymbol{\theta} \,|\{n_i\})\right] = \text{constant} + \sum_{i=1}^{M} n_i \ln(\mathcal{N}_i(\boldsymbol{\theta})) - \mathcal{N}_i(\boldsymbol{\theta}) \tag{11}$$

where the constant includes all terms not involving $\boldsymbol{\theta}$. Our best estimate of the values of $\boldsymbol{\theta}$ is given by the values of $\boldsymbol{\theta}$ which maximize L; its reliability is indicated by the width, or the sharpness of the posterior pdf about this optimal point. However there is no need to look at the width directly but rather the pdf itself. Since the $\mathbf{w}$ and the angle of $\boldsymbol{\xi}$ are non-interesting variables one can marginalize over those variables. Another improvement is to look for the ratio of the intensities. The pdf of Θ_h/Θ_c and $|\boldsymbol{\xi}|$ are the two interesting quantities.

The Bayesian analysis presented here is practically the same as the maximum likelihood estimator. This is generally true, when a flat prior is used then Bayesian approach and frequentist approach become equivalent. The reason we choose the Bayesian route is due to the *bootstrap* method which is extremely advantageous for our problem. Remember some of our objects (candidates) appear in few fields. As a matter of fact the brightest thickest candidates are in more than 3 fields. The basic idea of bootstrap is simple. The posterior probability from one image can be used as a prior for the next. One can show that this is equivalent to analyzing all the data together.