PRINCETON UNIVERSITY

Summer Student 2016 - University of Pisa

Search for candidates of ring-like extrasolar systems

Statistical methods and data analysis of transit events

Marco Cilibrasi

Supervisors
Dr. Joel Hartman
Prof. Gaspar Bakos
Department of Astrophysical Sciences, Princeton University - Princeton, New Jersey (USA)

October 8, 2016

Abstract

This report is the results of my work during my visit at the Department of Astrophysical Sciences at Princeton University with Dr. Joel Hartman and Prof. Gaspar Bakos. The goal of the project was to find some interesting candidates of ring-like system from the HATNet data, using some algorithms of common use and some of them developed by the HATNet and HATSouth group. In the first part of this report I introduce the exoplanets transits method, its potential and its limits and all the features of the instruments of the group, HATNet and HATSouth telescopes. In the second part I explain how we filter the data coming from these telescopes. In particular I use three different processes called *FIT*, *EPD* and and *TFA*. I will spend some time on this last method because it is the most interesting and above all it is developed by the Princeton group. In the third part I show how to search for transits. In particular I focus on the *BLS* method, that is a very powerful method developed by the group. I will also show how to detect and delete false transits (due to instrumental errors) via two procedures: a so called *histogram procedure* and the *PCA* procedure. In the last part I explain how to go on once a transit candidate is detected and I also show a model developed by Kenworthy & Mamajek to fit a ring-like system transit.

Contents

Chapter 1

Introduction

1.1 Exoplanet transits

Our knowledge about extrasoalr planets is mainly due to two important detection methods, that are in some sense complementary. In fact while the so called radial velocity method (we won't discuss about this) provides information about a planet's mass, the photometric (or transit) method can determine the planet's radius. If a planet crosses in front of its parent star's disk, then the observed visual brightness of the star drops by a small amount; depending on the relative sizes of the star and the planet.

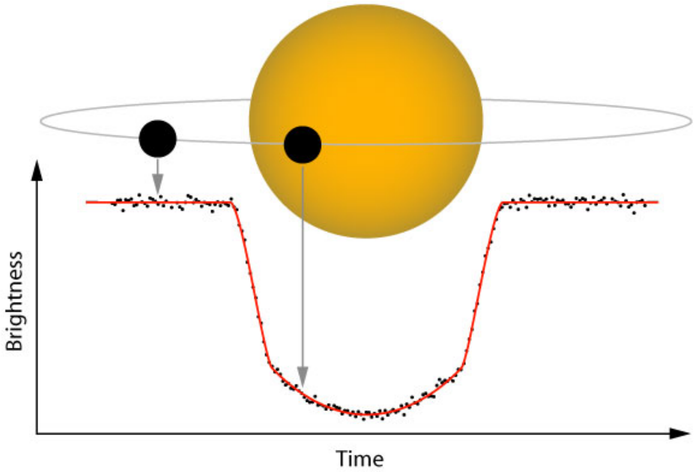


Figure 1.1: Scheme of an exoplanet transit.

This method has two major disadvantages. First, planetary transits are only observable when the planet's orbit happens to be perfectly aligned from the astronomers' vantage point. The probability of a planetary orbital plane being directly on the line-of-sight to a star is the ratio of the diameter of the star to the diameter of the orbit (in small stars, the radius of the planet is also an important factor). About 10% of planets with small orbits have such an alignment, and the fraction decreases for planets with larger orbits. For a planet orbiting a Sun-sized star at 1 AU, the probability of a random alignment producing a transit is 0.47%. Therefore, the method cannot guarantee that any particular star is not a host to planets. However, by scanning large areas of the sky containing thousands or even hundreds of thousands of stars at once, transit surveys can find more extrasolar planets than the radial-velocity method. Several surveys have taken that approach, such as the ground-based MEarth Project, SuperWASP and HATNet and the space-based COROT and Kepler missions.

The second disadvantage of this method is a high rate of false detections. A 2012 study found that the rate of false positives for transits observed by the Kepler mission could be as high as 40% in single-planet systems. For this reason, a star with a single transit detection requires additional confirmation, typically from the radial-velocity method. Radial velocity method is especially necessary for Jupiter-sized or larger planets as objects of that size encompass not only planets, but also brown dwarfs and even small stars. As false positive rate is very low in stars with two or more planet candidates, some of them can also be confirmed through the transit timing variation method [1].

The main advantage of the transit method is that the size of the planet can be determined from the lightcurve. When combined with the radial-velocity method (which determines the planet's mass) one can determine the density of the planet, and hence learn something about the planet's physical structure. The planets that have been studied by both methods are by far the best-characterized of all known exoplanets.

The transit method also makes it possible to study the atmosphere of the transiting planet. When the planet transits the star, light from the star passes through the upper atmosphere of the planet. By studying the high-resolution stellar spectrum carefully, one can detect elements present in the planet's atmosphere. Additionally, the secondary eclipse (when the planet is blocked by its star) allows direct measurement of the planet's radiation and helps to constrain the planet's eccentricity without the presence of other planets. It is then possible to measure the planet's temperature and even to detect possible signs of cloud formations on it.

1.1.1 Ring-like systems transits

In our case we are interested in detecting a special kind of transits that is a transit of a planet with a big rings system. This system must be big because we expect that only large rings could effect the light curve shape in order to be seen by telescopes. In particular then we expect to see very long events (even several days of transit) in our lightcurve, with a peculiar shape. In fact we expect to see a jagged shape of the transit due to the change in the composition (then in the optical depth) of the rings. We also expect to see some symmetry in the pattern (even if the possible inclination of the system would produce some sort of asymmetry), as we will see in the

[1] Duration variation refers to changes in how long the transit takes. Duration variations may be caused by an exomoon, apsidal precession for eccentric planets due to another planet in the same system, or general relativity.

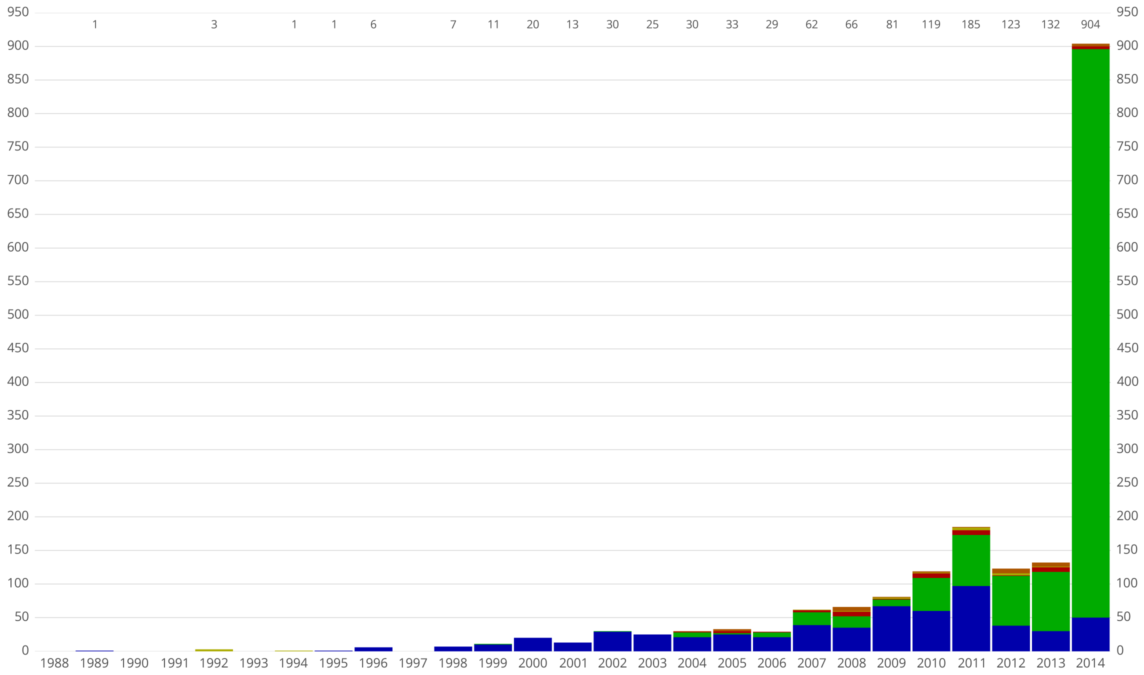


Figure 1.2: In this figure we show the number of exoplanets discovered by different methods: transits, radial velocity, direct detection, gravitational lensing and timing variation. Note how important is the contribute of transit method, especially in the last five years.

section 4.2.

The nature of our target implies a very important and tricky problem. Since we are looking for big ring system we have to estimate the conditions that allow such a big system to be dynamically stable. At the first order we can assume that a rings system is stable if it is entirely inside the so called *Hill sphere*. This is the region in which the planet dominates the attraction of satellites and rings. To be retained by a planet, a body must have an orbit that lies within the planet's Hill sphere. In more precise terms, the Hill sphere approximates the gravitational sphere of influence of a smaller body (the planet) in the face of perturbations from a more massive body (the star).

If the mass of the smaller body is m, and it orbits a heavier body of mass M with a semi-major axis a and an eccentricity of e, then the radius R_H of the Hill sphere for the smaller body is, approximately

$$R_H = a(1-e)\sqrt[3]{\frac{m}{3M}} \tag{1.1}$$

Then if we want to have stable large rings we must have a very big R_H that implies, once the masses are fixed, that a should be very big. As a consequence the period P of the planet would be very long[2], a single orbit could take dozens of years to be completed. This is an important fact because it means that in general we are not able to see the transit event more than once (we informally call those *non periodic* events). As we will see later this brings a lot of complications to

[2] Third Kepler law: $P^2 \propto a^3$

our problems because we don't have all the statistic that is used in the case of "normal" transits search and above all we are not able to predict when another event will occur, then generally we can't send proposal to bigger instruments in order to observe the event in a finer way and get more precise results.

In fact, as we will see in the next section, the group I worked with owns some small semi-professional telescopes and usually works with those. The convenience is that the group can work with always available instruments (they don't need to send proposals to other facilities) paying a not to high amount of money. The handicap is that the data are not that good so a lot of analysis work is required and often it is possible only to detect some *candidates* of transiting planets, if the data are not good enough (for example they have a low signal-to-noise ratio). Then it is fundamental to be able to predict future events in order to get observational time once some interesting candidates are detected. As we said this is not possible for ring-like systems.

1.2 HATNet Exoplanet Survey

The Hungarian-made Automated Telescope Network (HATNet) Exoplanet Survey [3] is a geographically distributed network of 7 small telescopes optimized for detecting transiting exoplanets. Since first light in 2003, they have discovered 60 exoplanets to date.

They detect exoplanets by monitoring large parts of the sky and attempting to determine if the light from a star periodically dims for a short time. This short "dip" in the light received may be caused by a small planetary-sized object moving in its orbit in front of the star as seen from our vantage point on Earth. This is a so-called exoplanet transit signal. The length and shape of this transit signal can be related to the radius and orbital period of the object passing in front of the star. Extensive follow-up observations are usually carried out to determine if the object detected in this way has a radius and mass consistent with that of a planet. Exoplanet detection by this method has proven to be enormously successful, and is responsible for a large fraction of all discovered exoplanets over the past twenty years.

HATNet telescopes are located at the Fred Lawrence Whipple Observatory (FLWO) at Mount Hopkins in Arizona, USA (5 telescopes), and at the Mauna Kea Observatory in Hawaii, USA (2 telescopes). The large separation in longitude allows them to seamlessly monitor the sky over the better part of 24 hours, reducing their susceptibility to false-positive signals caused by interruptions in observing. They use off-the-shelf 200-mm f/1.8 lenses attached to large format CCD cameras as the basis of our instruments, coupled with purpose-built robust telescope mounts and dome enclosures. All operations are automated; their telescopes are fully robotic and make decisions about when and what to observe based on weather conditions and pre-programmed observing priorities.

They and their collaborators have developed sophisticated algorithms to tease out such planetary transit signals from large time-series datasets, including the Trend Filtering Algorithm (TFA), and the Box Least-Squares (BLS) algorithm to find periodic signals in noisy data, that we are going to deal with later in this report.

1.3 HATSouth Exoplanet Survey

The Hungarian-made Automated Telescope Network-South (HATSouth) Exoplanet Survey [4] is a network of 6 astrograph telescope systems designed to detect transiting exoplanets in orbit around relatively bright stars visible from the Southern hemisphere. The telescopes are distributed over three continents; South America, Africa, and Australia. Since first light in 2009, they have discovered 35 exoplanets to date.

HATSouth telescope systems are located at the Las Campanas Observatory (LCO) in Chile, at the High Energy Stereoscopic System (HESS) site in Namibia, and at the Siding Spring Observatory (SSO) in Australia. Each site has two HATSouth stations; each station consists of four 180-mm f/2.8 Takahashi astrographs attached to large-format CCD cameras, accompanied by purpose-built heavy-duty mounts and dome enclosures. All operations are automated, similarly to HATNet.

Again, the large separation in longitude of the HATSouth sites allows them to seamlessly monitor the sky over the better part of 24 hours. This is especially advantageous during the Southern summer months, when a target star field can be observed in a continuous relay by telescopes at all three sites. During the Southern winter, a target star field will still be observed by telescopes at two sites near-continuously, thus ensuring dense coverage year-round.

Chapter 2

First lightcurve analisys

2.1 Data collection

First of all the data I worked on are collected looking at the sky with the telescopes of HATNet. They divide the sky into about 200 fields and they look at one by one and, as we sai before, these telescopes are fully robotic and make decisions about when and which field to observe based on weather conditions and pre-programmed observing priorities. The data we receive from the observatories are collected in some *.txt* files, one for each star[1], in which we can see 20 different columns, each of them with as many lines as the number of points taken by the telescopes in that field. In table 2.1 we can see what is reported in that colomns.

As we can see the data are taken using three different aperture of the objective, in order to have good data for all the range of brightness of the stars in the field. Obviously we would like to take measurements of the luminosity of bright stars with smaller apertures in order to avoid saturation and to take measurements of faint stars with bigger apertures in order to receive more photons. There is an automatic procedures that calculate the proper aperture for each star and produces a text file in which the user can check what this apertures are so that he can consider only the best data.

Another thing that we notice is the presence of a quality flag. In fact if something goes wrong with the measurements (for example we detect a saturation in the image of the star) the line si marked with a X flag, while if something works well the line is marked with a G flag. In our analysis we use only the proper aperture data, indicated by the automatic procedure described before, and we deleted from the curves all the data flagged with an X.

[1] For example I worked on field 144 that has about 18 thousands of stars.

Colomn	Content
1	Number of the instrument that has taken the measurement (from 1 to 12) and serial number of the specific image (for example '$10-82919$')
2	Time at which the image was taken in HJD
3	Raw magnitude with aperture 1
4	Error on the raw magnitude with aperture 1
5	Quality flag for aperture 1
6	Raw magnitude with aperture 2
7	Error on the raw magnitude with aperture 2
8	Quality flag for aperture 2
9	Raw magnitude with aperture 3
10	Error on the raw magnitude with aperture 3
11	Quality flag for aperture 3
12	FITMag magnitude with aperture 1
13	FITMag magnitude with aperture 2
14	FITMag magnitude with aperture 3
15	EPDMag magnitude with aperture 1
16	EPDMag magnitude with aperture 2
17	EPDMag magnitude with aperture 3
18	TFAMag magnitude with aperture 1
19	TFAMag magnitude with aperture 2
20	TFAMag magnitude with aperture 3

Table 2.1: In this table we can see what is the content of all the coloms of the data files. Note that the time in HJD (Heliocentric Julian Date) is dependent on the position of the object in the sky, this will affect our procedures because we will have the use the image number as a reference and not the time value. This may cause some counterintuitive results.

2.2 FITMag and EPDMag

Since we riceive the data from the telescopes and we get the text files we have to apply some algorithm and methods to filter the data in order to avoid instrumental and other systematic errors. Starting from the raw data the first thing we want to obtain is what we call *FITMag* data. This FITMag data are the magnitudes after applying an ensemble zero-point correction based on all of the stars in an image. This corrects for large systematic variations in the raw brightness. These are our preliminary light curves, they still may have large systematic errors, but for variable stars with high amplitude and long time duration variations, these are probably the best values to work with because, as we will see, the next steps in the data 'filtering' could remove the dips we are looking for.

We usually apply another filtering procedure after that in order to avoid any kind of error, we want to remove some outliers from the curves. The procedure is about deleting the points that are less or greater than the mean of the other closest points more than an arbitrary magnitude (in our case we use a threshold of $1\,mag$). This should give us better light curves and it should not delete interesting points because we expect that the transit duration would be longer than the 4 or 5 points we use to compute the mean (the points are taken every 5-6 minutes).

The second step is the computation of the so called *EPDMag* light curves, as we can see in table 2.1. These are the FitMag light curves after applying an additional filtering to remove variations that are correlated with changes in instrumental parameters, such as the x and y position of the star on each image, the sub-pixel position of the star on each image, the background, the scatter in the background, the shape of the point spread function, the hour angle of the observations, and the airmass of the observations. The decorrelation is done on a star-by-star basis. These light curves have fewer systematic errors than the FitMag light curves, but high amplitude or long-duration astrophysical variations may be distorted or removed. For searching for short (for example less than one day) and/or low amplitude (less than a few %) variations, it is usually better to use than the FitMag light curves, in our case we consider also the EPDMag lightcurves but we expect to find our ring-like system transits in the FITMag lightcurves.

2.3 TFAMag

In the end there is another algorithm developed by the HATNet group that can be used to reduce photometric data. They call this algorithm *trend filtering algorithm* (TFA). In the original paper [5] they show that various systematics related to certain instrumental effects and data reduction anomalies in wide-field variability surveys can be efficiently corrected by the TFA applied to the photometric time-series produced by standard data pipelines.

The idea is that many of the systematic variations in a given light curve are shared by light curves of other stars in the same data set, due to common effects such as colour-dependent extinction, or blending of two or more star images (which could produce similar light-curve variations in all of them), etc. A solution to remove or reduce these common systematic variations can therefore be devised using the already reduced data. For each target star, one must identify objects in the field that suffer from the same type of systematics as the target, and apply some kind of optimum filtering of the target light curve based on the light curves of a set of template stars as follows.

Let me assume that all time-series are sampled in the same moments and contain the same number of data points N. Let our filter be assembled from a subset of M time-series, distributed nearly uniformly in the field. Because we have no a priori knowledge on which stars are authentic variables, the above selection is almost random, except that stars with a low number of data points, low brightness and high standard deviation are not selected. Once the template set is selected, it is fixed throughout the analysis.

The filter $\{F(i); i = 1, 2, ..., N\}$ is built up from the following linear combination of the template time-series $\{X_j(i); i = 1, 2, ..., N; j = 1, 2, ..., M\}$

$$F(i) = \sum_{j=1}^{M} c_j X_j(i) \tag{2.1}$$

It is assumed that the template time-series are zero-averaged (i.e. $\sum_{i=1}^{N} X_j(i) = 0$ for all j), therefore we usually subtract the mean of the time series in order to add it again in the end of the process. The coefficients $\{c_j\}$ are determined through the minimization of the following expression:

$$D = \sum_{i=1}^{N} [Y(i) - A(i) - F(i)]^2 \tag{2.2}$$

where $\{Y(i); i = 1, 2, ..., N\}$ denotes the target time-series being filtered, and $\{A(i); i = 1, 2, ..., N\}$ denotes the current best estimate of the detrended light curve. The goal of this procedure is then to find, thanks to a least square procedure, the linear combination of time-series that fit best the target time-series.

We start from the assumption that the time-series is dominated by systematics and noise, and we have no a priori knowledge of any real (periodic or aperiodic) signal in the light curve. Consequently, initially we set $\{A(i)\}$ equal to the average of the target time-series, i.e. $A(i) =< Y >= \frac{1}{N} \sum_{k=1}^{N} Y(k)$.

The filtered time-series $\{\hat{Y}(i) \equiv Y(i) - F(i)\}$ is then phase-folded and binned, then remapped to the original time-base to give a new estimate of $\{A(i)\}$, which is in turn fed into equation 2.2 to compute a new set of filter coefficients $\{c_j\}$. The new filter leads to a better determination of $\{A(i)\}$, and the iteration continues until some convergence criterion is satisfied, for example $\sigma < 10^{-2}$ where

$$\sigma^2 = \frac{D}{N - M} \tag{2.3}$$

At a more technical level, the main steps of the TFA are the following.
Compute the normal matrix from the above template time-series

$$g_{j,k} = \sum_{i=1}^{N} X_j(i) X_k(i) \ \ ; \ \ j, k = 1, 2, ..., M \tag{2.4}$$

and compute the inverse of it: $\{G_{j,k}\}$. For each lightcurve compute scalar products of the target and template time-series:

$$h_j = \sum_{i=1}^{N} \bar{Y}(i) X_j(i) \tag{2.5}$$

where $\{\bar{Y}(i) \equiv Y(i) - A(i)\}$. Solving the equation $\frac{\partial D}{\partial c_j} \equiv 0$ for all j it comes out that the solution is:

$$c_j = \sum_{k=1}^{M} G_{j,k} h_k \tag{2.6}$$

The corrected time-series, that becomes our new estimate for $\{A\}$ in the iteration is computed from

$$\hat{Y}(i) = Y(i) - \sum_{k=1}^{M} c_k X_k(i) \tag{2.7}$$

Statistical tests, performed on the data base of the HATNet project, show that by the application of this filtering method the cumulative detection probability of periodic transits increases by up to 0.4 for variables brighter than 11 *mag*, with a trend of increasing efficiency toward brighter magnitudes. The improvement in the S/N ratio is often spectacular, leading to secure discoveries from signals originally completely hidden in the systematics. Once the period is found, the signal can be reconstructed by applying the TFA iteratively, as mentioned above.

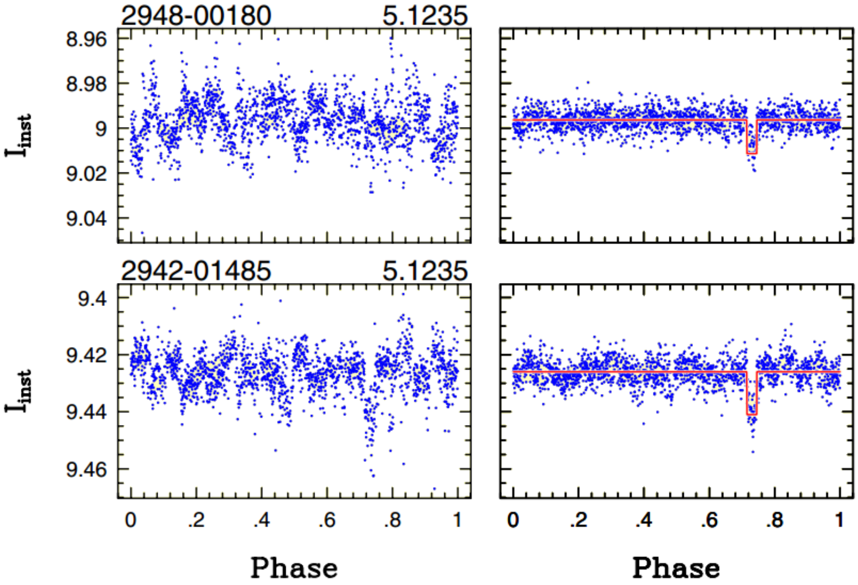


Figure 2.1: Reconstruction of their standard box-shaped test signal 1 with the aid of iterative TFA filtering. Left column: original folded signal. Right column: TFA-filtered signal with the synthetic signal shown by a solid line. Identifiers (GSC numbers) of the stars employed in the generation of the test signals and the corresponding periods are given in the headers of the panels on the left. They used 595 template light curves and bin averaging in the TFA filtering.
A trend filtering algorithm for wide-field variability surveys [5]

In out project we use all the kind of lightcurves (FIT, EPD and TFA) because, as said before for EPD curves, we noticed that both EPD and TFA alogithms tend to delete or distort high amplitude or long-duration astrophysical variations if not periodic and, unfortunately, this is exactly the kind of event we are looking for, although EPD and TFA curves have certainly fewer systematic errors.

Chapter 3

Searching algorithms

Once we have filter the lightcurves with the statistical methods above we have to use some sort of procedure to find the transit events. The easiest thing to do is to study the box fitting algorithm (BLS [6]) that the HATNet group has developed for periodic transit search and modify it in order to look for non-periodic and long duration transits.

3.1 BLS algorithm

The BLS (Box Least Square) algorithm searches for signals characterized by a periodic alternation between two discrete levels, with much less time spent at the lower level. In the original paper ([6]) it is shown that the crucial parameter is the effective signal-to-noise ratio – the expected depth of the transit divided by the standard deviation of the measured photometric average within the transit, when this parameter exceeds the value of 6 we can expect a significant detection of the transit. It is also shown that the box-fitting algorithm performs better than other methods available in the astronomical literature, especially for low signal-to-noise ratios as the Discrete Fourier Transformation (DFT), multyfrequency DFT or the Phase Dispersion Minimization (PDM). Here I show the technical details of the algorithm.

We assume a strictly periodic signal, with a period P_0, that takes on only two discrete values, H and L. The time spent in the transit phase L is qP_0, where the fractional transit length q is assumed to be a small number ($\sim 0.01-0.05$). For any given set of data points, the algorithm aims to find the best model with estimators of five parameters: P_0, q, L, H and t_0, the epoch of the transit. Actually, if we assume the average of the signal is zero (then again we subtract the mean of the signal and we readd it at the end), we have $H = -L\frac{q}{1-q}$, and the number of parameters of the model is reduced to four. We denote the data set by $\{x_i, i = 1, 2, ..., n\}$, we assume that the standard error σ_i is constant so we use a least square procedure instead of a least χ^2 procedure.

For a given trial period we consider a folded time series, which is a permutation of the original time series. This series is denoted by $\bar{x}_i$. We fit a step function to the folded time series with the following parameters:

- $\hat{L}$ - the level in $[i_1, i_2]$
- $\hat{H}$ - the level in $[1, i_1)$ and $(i_2, n]$

and the relative time spent at level $\hat{L}$ is characterized by $r = \frac{i_2 - i_1}{N}$. For any given (i_1, i_2) we minimize the expression of the reduced least square

$$D = \frac{1}{N}\left[\sum_{i=1}^{i_1}(\bar{x}_i - \hat{H})^2 + \sum_{i=i_1}^{i_2}(\bar{x}_i - \hat{L})^2 + \sum_{i=i_2}^{n}(\bar{x}_i - \hat{H})^2\right] \tag{3.1}$$

Minimization of D yields simple weighted arithmetic averages over the proper index regimes

$$\hat{L} = \frac{s}{r} \quad ; \quad \hat{H} = -\frac{s}{1-r} \tag{3.2}$$

where

$$s = \frac{1}{N}\sum_{i=i_1}^{i_2}\bar{x}_i \tag{3.3}$$

With these formulae, the average squared deviation of the fit becomes

$$D = \frac{1}{N}\sum_{i=1}^{n}\bar{x}_i^2 - \frac{s^2}{r(1-r)} \tag{3.4}$$

Once this expression is evaluated, one has to repeat the computation with other (i_1, i_2) values and find the absolute minimum of D for any given period. The first term on the right hand side of equation 3.4 does not depend on the trial period, and therefore one can use the second term alone to characterize the quality of the fit. We define the Box-fitting Least Squares (BLS) frequency spectrum by the amount of Signal Residue of the time series at any given trial period:

$$SR = MAX\left\{\sqrt{\frac{s^2(i_1, i_2)}{r(i_1, i_2)[1 - r(i_1, i_2)]}}\right\} \tag{3.5}$$

Here, the maximization goes over the values $i_1 = 1, 2, ..., n^\star$, while the i_2 values satisfy the inequality $\Delta i_{min} < i_2 - i_1 < \Delta i_{max}$, where $\Delta i_{min/max}$ are determined by the minimum/maximum fractional transit length suspected to be present in the signal. The maximum lower index $n^\star$ depends on $i_2 - i_1$ and covers the range of $[n - \Delta i_{max}, n - \Delta i_{min}]$. We can also define $\delta \equiv H - L$ for the transit depth and we find that SR at the correct test period yields also an estimate of δ, i.e., $SR = \hat{\delta}\sqrt{r(1-r)}$.

To characterize the signal we can introduce two differente quantities, the signal to noise ratio $SNR = \delta/\sigma$ and, in order to characterize the peaks of the BLS spectrum, the Signal Detection Efficiency:

$$SDE = \frac{SR_{peak} - < SR >}{\sigma(SR)} \tag{3.6}$$

where SR_{peak} is the SR at the highest peak, $< SR >$ is the average, and $\sigma(SR)$ is the standard deviation of SR over th frequency band tested.

In a practical implementation of the above procedure, it is suggested to divide the folded time series into m bins and evaluate SR by using these binned values. This approach is very efficient computationally, and yields an exact LS solution with a time resolution defined by the number of bins. Although the lower resolution affects the efficiency of the signal detection, in all interesting cases a good compromise can be made between computational constraints and the effectiveness of signal detection.

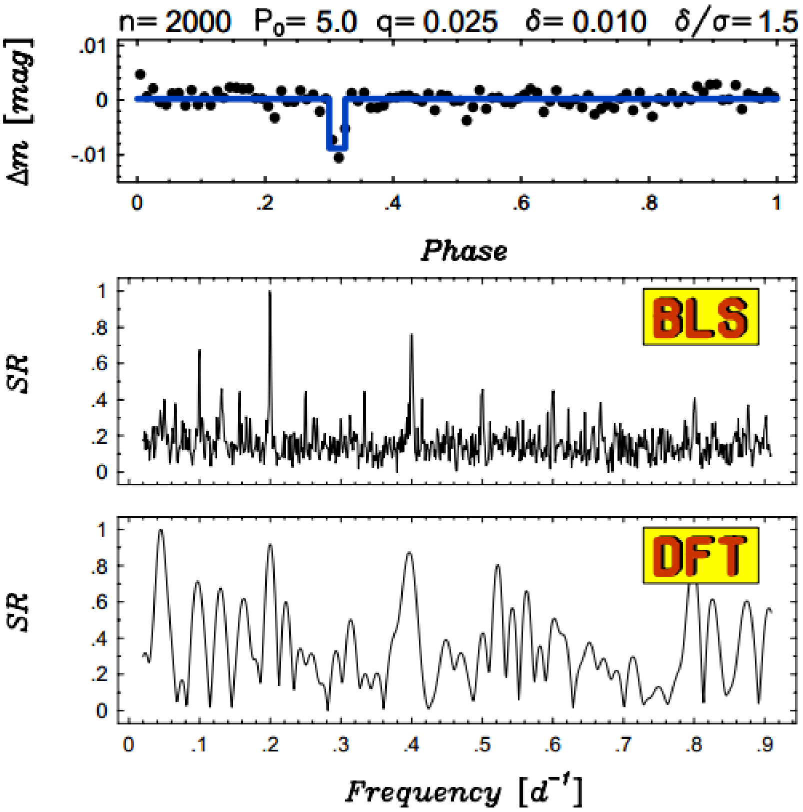


Figure 3.1: Comparison of the BLS and DFT methods for a realization of a signal with parameters shown in the header (other parameters are standard). The uppermost panel shows the folded/binned time series (dots) with the period and the fit (continuous line) obtained by the BLS method.
A box-fitting algorithm in the search for periodic transits [6]

In our case we don't expect to find periodic events, but we look for non-periodic events with probably a long duration and a large depth. In my work I modify the BLS algorithm in order to accomplish our constraint. First of all we don't want the algorithm to go through a lot of frequencies but we only want to compute SR for one frequency that is simply $\frac{1}{T}$ where T is the time duration of the processed series, i.e. $T = t(n) - t(1)$. This obvoiusly makes the process much faster, although we usually set a quite high value for $\Delta imax$ (for example 0.5) in order to look for long transit events, that makes the process a little bit slower.

The real change is that we don't have anymore the SDE quantity to characterize the detected event and to determine "how transit-like" a light curve is. We have then to define a new quantity that allow us to sort the light curves from the best transit-like to the worst and even to set a threshold above that we can say to have detected a transit (and as a consequence a candidate for a ring-like system). The quantity chosen is the so called χ^2_{ratio}, that is defined as $\chi^2_{ratio} = \frac{\chi^2_{lin}}{D_{min}}$, where χ^2_{lin} is simply the reduced least square[1] of a linear fit of the data series

$$\chi^2_{lin} = \frac{1}{N}\sum_{i=1}^{N}[x_i - mean(x)]^2 \tag{3.7}$$

while

$$D_{min} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i^2 - SR^2_{MAX} \tag{3.8}$$

that in fact is the least squares sum of the best-fitting box function. In analogy we can also use the quantity

$$\frac{\chi^2_{lin} - D_{min}}{\chi^2_{lin}} \equiv 1 - \frac{1}{\chi^2_{ratio}} \tag{3.9}$$

It is clear than higher is the χ^2_{ratio} for a given series and higher is the probability that that series present a non-periodic transit with tha caracteristic we are looking for. After having run the BLS algorithm on all the lightcurves we sort them by the χ^2_{ratio} and we obtain some interesting results with the TFA curves (see figure 3.2), while for EPD and FIT curves with have a lot of good transits detected, but it seems that most of the best 300 curves have the same transit patterns, obviously because we have not processed the curves with the TFA algorithm, and that in all kinds of curves we detect some variable stars that make the BLS alogirthm produce some ambiguos results (the BLS works well when the lightcurve looks actually like a boc, as described in the previous paragraphs).

What we want to do is then to use some sort of procedures to detect and delete all these common patterns or these sistematic errors in our data in the FIT and EPD curves because we think that the TFA algorithm tend to delete the majority of long and deep non periodic transits.

3.2 Histogram algorithm

The simplest way to do that is what we call *histogram algorithm.* This algorithm is based on the idea that the distribution of the number of transits is poissonian and now I explain what this sentence means exactly.

With the BLS algorithm we have recognize a transit for each lightcurve (some of the have a high value for $\chi ratio$, other have a low number). We can flag all the points of a lightcurve with a boolean value, 1 if that point is in the transit interval, 0 if it is not. This way we can count

[1] Actually the name χ^2 is improper because it is in fact a least squares sum, but χ is much easier to write.

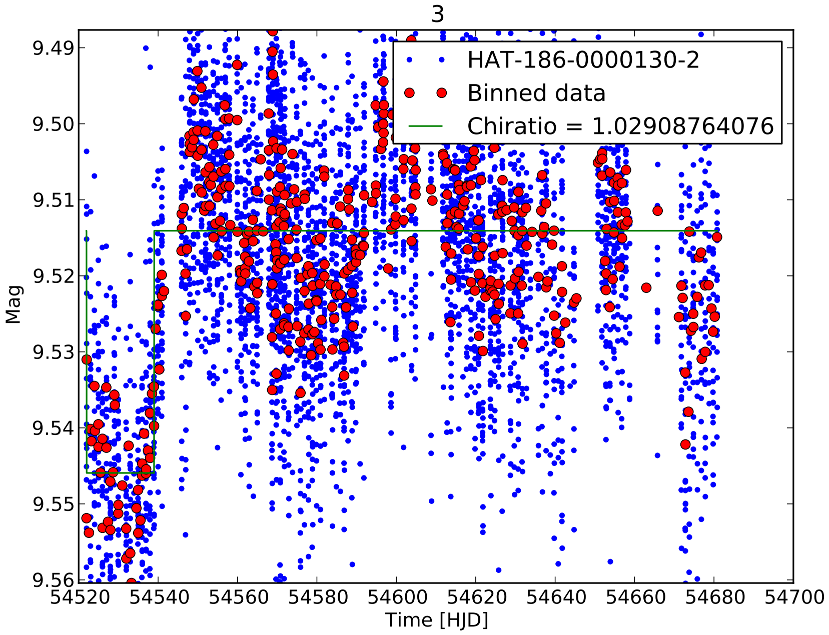


Figure 3.2: In figure there's an example of a TFA lightcurve with all the parameters. The 3 in the title means that this curves is the third best chiratio lightcurves in the TFAs, $HAT - 186 - 0000130 - 2$ means that the star is the number 130 in the $186th$ field (I work on $144th$ field, but there are also some neighbour stars) and the points are taken with the aperture 2. The blue dots are the observational data, the red ones are the chosen binning for BLS, the green line is the box-fit function and in the legend is reported also the $\chi ratio$ of the fit.
HATNet data [1]

how many stars have a transit detected with BLS in a given image, it is easy to understand if we imagine that our data are organized in a matrix: the lines are the different stars of the field and the columns are the different images taken by telescopes and in the matrix we can have the magnitude measured or, in this case, the boolean value for the transit.

We can associate to each image a transit rate (from 0 to 1) computing the ratio beetwen the number of transits detected in that image and the total number of stars that have a good measurement in that image (not all the stars have points in all the image because the filtering procedures described before). If time and number of the image were in a 1 to 1 relation we could plot this ratio *vs* time and observe if there is some time interval with an overabundance or a poorness of transits in order to recognize some sistematic errors and correct them. That's not the case, then we can go further in our analysis and think about another way to solve this problem.

What we do is to plot histograms with the binned transits rates (usually we take the square root of the number of images for the bins number[2]) on the x axis, while on the y axis we plot the number of images with a rate within a certain bin. We state that the distribution that we observe should be a Poisson distribution with a certain μ that we don't estimate, but we know that exists. For example the histogram that we obtain with TFA lightcurves is shown in figure 3.3 and it is

[2] In *Python* for example it will be $int(sqrt(N))$.

actually very similar to a Poisson distribution.

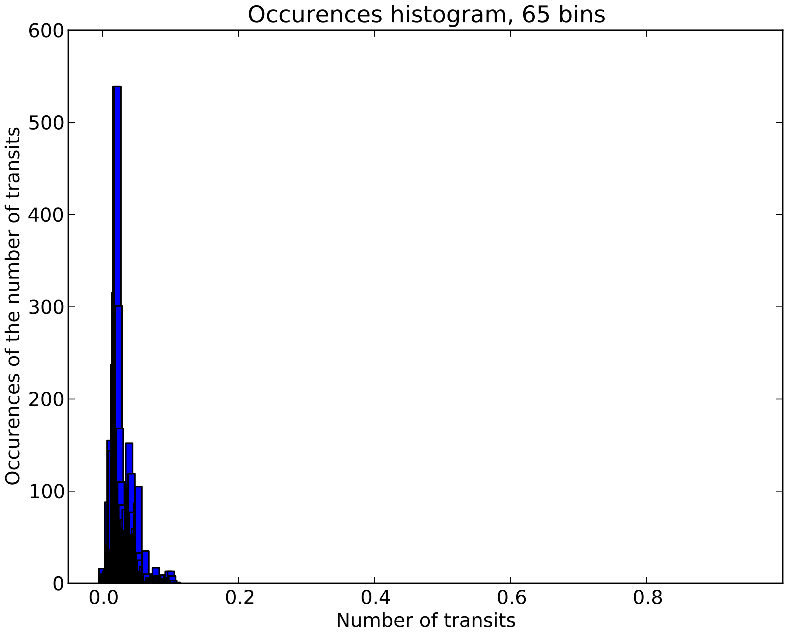


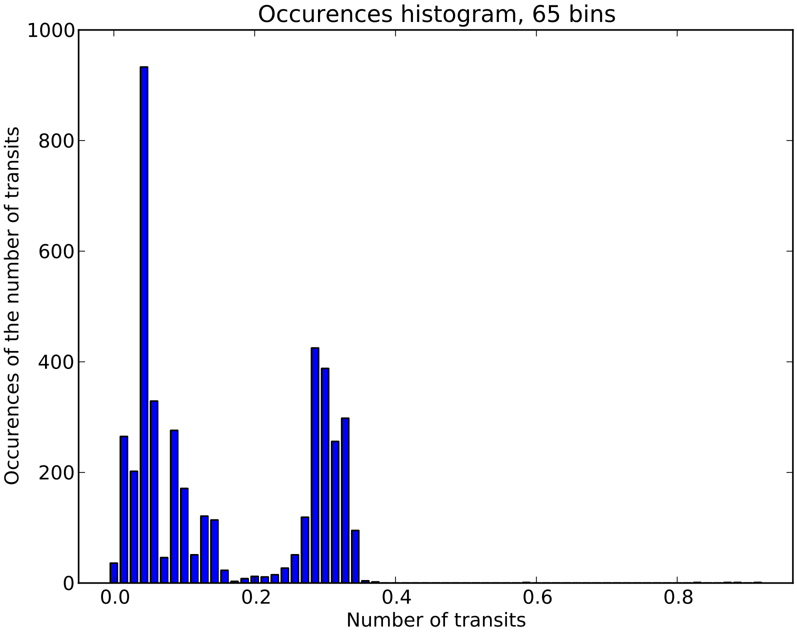
Figure 3.3: Result of the histogram algorithm for TFA lightcurves. We notice that the distribution shape is very similar to a Poisson shape with a μ close to 0.

On the contrary for FIT and EPD we obtain a very differente result, shown in figure 3.4. For FIT light curves we notice that the distribution is similar to a Poisson distribution with μ greater than μ_{TFA}, with another peak close to 0.3. For EPD lightcurves we have a strange shape that reminds of a Poisson distribution with other peaks.

The idea now is to take the images with a transits rate greater than a chosen threshold (could be ~ 0.2 for FIT and ~ 0.1 for EPD) and delete them from our data. We think that what we delete are the points with sistematic errors and this could be give us better data to analize, even if this procedure is highly depending on the chosen threshold (and we can't apply this procedure automatically to all the other fields because we will have tho choose a threshoold for each field) and even if deleting all the data of a certain image in all the lightcurve we are going to delete also some "good" data only because they occur in the same image of some "bad" data. In fact we think that these bad data are due to some focusing errors or some mistakes in the PSF[3] fitting with star at the edge of the field, while measurements of the magnitude of stars in the center of the fields should be good.

Another problem with this procedure is that we don't distinguish yet between good transits (high $\chi ratio$) and bad transits (low $\chi ratio$) and this could falsify our results. Another thing we should be aware of is the fact that the μ's of the Poisson ditributions for different kinf of lightcurves (FIT, EPD, TFA) are different, so this should be taken into account somehow. We could also find

[3] Point Spread Function



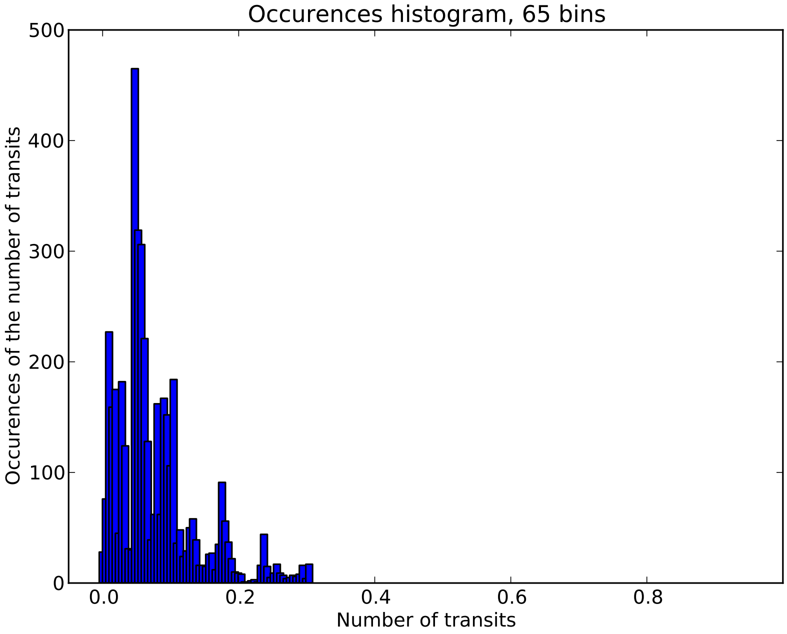


Figure 3.4: Result of the histogram algorithm for FIT and EPD lightcurves. For FIT we notice that the distribution is similar to a Poisson distribution with μ greater than μ_{TFA}, with another peak close to 0.3. For EPD we have a strange shape that reminds of a Poisson distribution with other peaks.

a way to estimate the value of μ before running the histogram procedures, or we can find a precise μ for the distributions above with some sort of best fit procedure in order to understand better what is happening. We are still working on that.

3.3 PCA algorithm

Another procedure that we can follow to detect and delete the common patterns in the light curves is the so callaed *Principal Component Analysis (PCA) algorithm.* Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

The primary goal of this techique is the reduction of an high number of variables to some less latent variables (feature reduction). We use this algorithm in a different way. In fact the first principal component simply represents the most common trend in the data. What we do is to compute a *first principal component correlation value* for all the best[4] lightcurves and to delete from our data those that has a correlation value higher than a certain threshold. Now we show the procedure we follow in our work to find the principal components defined before, withouth giving a mathematical proof of it.

Let X_{ij} be the j^{th} observed point of the light curve of the i^{th} star, $1 \leq j \leq P$, $1 \leq i \leq N$. In our procedure we usually subract the mean magnitude from the light curve and we readd it at the end of the process, otherwise our principal component will mainly represent the continuos part (offset) of the curves. We can compute the *correlation matrix*

$$S_{jk} = \frac{1}{N} \sum_{i=1}^{N} X_{ji} X_{ki} \tag{3.10}$$

This measures the correlation between the j^{th} and the k^{th} point on the light curve, averaged over all N stars in the sample. It can be shown[5] that in order to maximise the variance between a set of new orthogonal axes $\mathbf{u}^t$ it is necessary to solve the eigenvalue equation

$$S\,\mathbf{u} = \lambda\,\mathbf{u} \tag{3.11}$$

This equation results in a set of vectors $\mathbf{u}^t$ ($1 \leq t \leq P$), the eigenvectors, which are the new axes we seek, and their corresponding eigenvalues, λ^t. We can project each observed light curve onto a given axis by forming the dot product,

$$PC_t(i) = \sum_{j=1}^{P} X_{ij} u_j^t \tag{3.12}$$

[4] Best in the sense of sorted by $\chi ratio$.

[5] As said before we won't give a mathematical proof of that in this report.

and obviously the observed light curves can be reproduced by the formula

$$\sum_{t=1}^{P} PC_t(i)\mathbf{u}^t \tag{3.13}$$

The last equation shows that an observed light curve is expressed as a linear combination of "basis" light curves, which are the igenvectors $\mathbf{u}^t$ of the correlation matrix S. The quantity λ_t^2, after suitable normalisation, shows the percentage variance in the sample explained by th t^{th} principal component. A plot of $PC_t(i)$ against period would indicate how that particular coefficient changes with time. In figure 3.5 is shown an example of how PCA works better than Fourier Analysis in reconstruct light curves, in this case of a variable star.

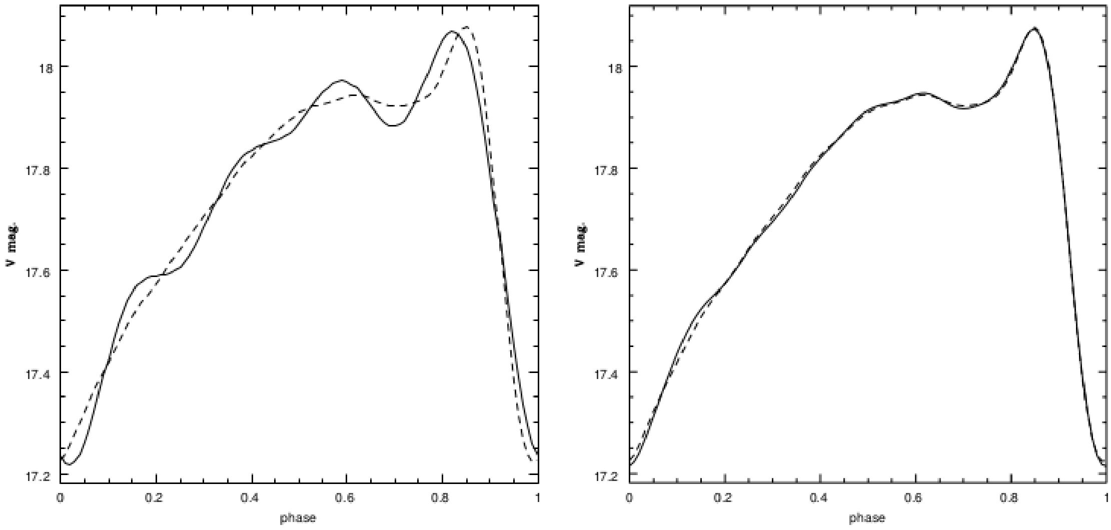


Figure 3.5: RRab light curve reproduction using Fourier (solid lines) and PCA (dashed lines) methods. The left panel is a fourth order (9 parameters) Fourier fit (solid) and an eight order PCA (9 parameters) fit (dashed). The right panel is an eight order (17 parameters) Fourier fit (solid) and an eight order PCA (9 parameters) fit (dashed).
Principal Component Analysis of RR Lyrae light curves [8]

In our analysis we assume that the first principal component is also the most common trend of the data. We can then go on with two different procedures. One thing we can do is to subtract the projection of a lightcurve on the first principal component (or on the first two, or three, it depends on us and on the particular shape of data) from the lightcurve it self, in other words

$$X_{ij}^{fin} = X_{ij} - \sum_{k=1}^{n} < \mathbf{X}_i, \mathbf{u}^k > u_j^k \tag{3.14}$$

where n can be $1, 2, ...$ depending on us. This procedure is in fact very similar to the TFA algorithm and we are afraid that this subtraction could distort our data, create transits where they are not and delete some real transits. Then we use a procedure that may cause a loss of good data, but certainly it doesn't distorte them.

What we do is to compute a correlation value for each lightcurve with the first n eigenvectors. For example we can compute

$$C_i^t =< \mathbf{X}_i, \mathbf{u}^t >= \sum_{j=1}^{P} X_{ij} u_j^t \quad ; \quad t = 1, 2, ..., n \tag{3.15}$$

Then we can simply decide a value for n and delete all the lightcurves that have a value of C greater than a certain threshold, always arbitrary and depending on our choice and on the particular shape of the data. This way we unfortunately delete entire lightcurves and then also possible good data from those.

This is a very tricky procedure because we have to define a lot of arbitrary quantities (n, the thresholds for C and so on), then we also lose a lot of possible good data and we observe that it is not so efficient in recognizing common pattern in our data, maybe because the data are not uniformly distributed in time. We will try in future to find a correct set of values for n and the thresholds and above all we have to find an automatic way to set these values for all the other fields. We could also use some sort of modified TFA procedures for this point, since they are quite similar. We are still working on that.

Chapter 4

Future developments

4.1 Candidates

Our goal with all these procedures is to find some candidates of transiting ring-like systems. Usually, for normal transit periodic events, the best thing to do, once some candidates are recognized, is to write a paper presenting the results and to prepare a proposal for some better observations, for example with bigger telescopes or even with space telescopes. We can get the period of our events from our analysis so it is easy to foresee when we can observe a transit event (or some transit events) again.

In our case things are more complicated. First of all candidates detection is harder because of the lack of periodicity. This obviously make impossibile to foresede if and when it will be possible to observe another transit event. It is easy to understand that it is impossible then to send a proposal to another telescope if we are not able to say when we will need the observation time.

At this point the best thing to do is to check in the literature if someone with a better instrument has looked at the same star in the exact moment of the transit we detect, or also in the past (if we are lucky the period of the ring-like system we are looking for is not too long). If data are available we can make a comparison and maybe do a more precise analysis of that event. In literature we can also find some examples of ring transit modelling, and our ideal goal is, if we are lucky with available data in literature, to get some physical parameters of the system from the transit lightcurves.

4.2 Ring transit modelling

To conclude we can present an example of a ring transit model found in literature ([9]). In the paper they analyze the light curve of $1SWASP\ J140747.93-394542.6$ (or $J1407$), a ~ 16 Myr old star, that underwent a complex series of deep eclipses that lasted 56 days, centered on 2007 April. This light curve is interpreted as the transit of a giant ring system that is filling up a fraction of the Hill sphere of an unseen secondary companion, $J1407b$. They fit the light curve with a model of an azimuthally symmetric ring system, including spatial scales down to the temporal limit set by the star's diameter and relative velocity.

Their ring model is composed of two parts, which they solve sequentially for an observed transiting ring system. They assume that the primary star and ring system are at a similar distance from the Earth, and that the ring system is at least several times larger than the angular size of the primary star. They first solve for the orientation of the plane of the ring system relative to our line of sight, and they then solve for the transmission of the rings as a function of radius from the secondary companion, given the geometry of the ring system derived in the previous step.

The primary is a star at a distance of d parsecs, with radius R . They approximate the orbit of the secondary for the duration of the eclipse as being a straight line, with constant relative velocity of v. Surrounding the secondary companion is a ring system in a plane that contains both the rings and the equatorial plane of the secondary companion, which they refer to as the ring plane. The rings are composed of individual particles that orbit the secondary companion in Keplerian orbits, assumed circular. These particles scatter light out of any incident beam and in aggregate are approximated by a smooth screen with an optical transmission of $\tau(r)$ that varies as a function of radial distance r from the center of the secondary companion. The rings are assumed to be azimuthally symmetric; the inclination of the ring plane as seen from the Earth is i_{disk} (with $0°$ being face-on), and the projected angle between the normal of the secondary companion's orbit and the normal of the ring plane is ϕ_{disk} .

The parametric equation for a projected ring is then

$$\begin{cases} x(p) = r(cos\, p\; cos\, \phi_{disk} - sin\, p\; cos\, i_{disk}\; sin\, \phi_{disk}) \\ y(p) = r(cos\, p\; sin\, \phi_{disk} - sin\, p\; cos\, i_{disk}\; cos\, \phi_{disk}) \end{cases} \tag{4.1}$$

where x and y are coordinates on the ring at radius r at a given value of the parametric variable p, which has a value from 0 to 2π radians. Considering that the star moves along a line parallel to the x-axis with a speed v ($x = v(t - t_b)$ and $y = b$) we can obtain a relation for $r(t)$ and therefore for $G(t) = \frac{dr}{dt}$ that we show in figure 4.1. Knowing $G(t)$ we can also in principle know $g(t) = \frac{dI(t)}{dt}$, that is a quantity we can measure with photometry methods, starting from a ring opacity model $I(r) = I_0 e^{-\tau(r)}$, where $\tau(r)$ is the optical depth, and getting $g(t) = \frac{dI(r)}{dr}\frac{dr}{dt} = -\tau(r(t))I(r(t))G(t)$.

The function $G(t)$ represents also the maximum flux change possible between a fully transparent and fully opaque ring. For rings that have intermediate values of transmission, the resultant light gradients will lie underneath the curve in figure 4.1, and so the curve represents an upper bound on the light gradient for a given ring orientation. Since we do not know $\tau(r)$, we can use $G(t)$ as an upper limit, and we search for ring orientations that have all measured gradients lying underneath this curve. For example we have n measurements of the light-curve gradient $g(t)$ at time t. The model gradient $G(t)$ is calculated as above. In the paper they introduce a cost function that

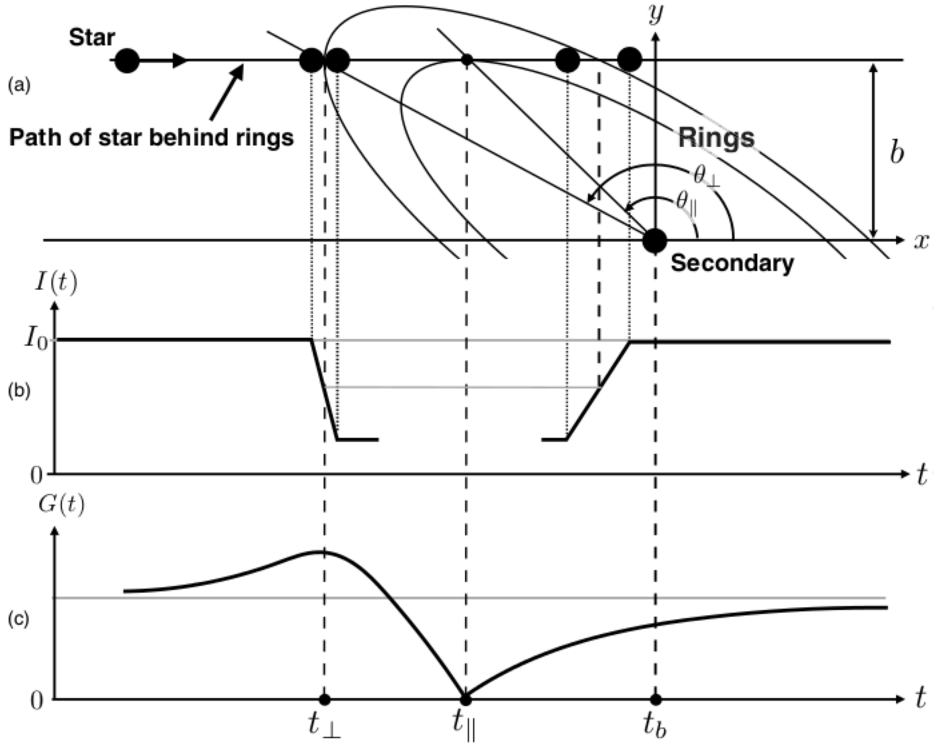


Figure 4.1: Geometry of the ring model. Panel (a) shows a ring system inclined at an angle of idisk and rotated from the line of relative velocity by ϕ_{disk} . The star passes behind the ring system with impact parameter b at time t_b . Panel (b) shows the resultant light curve $I(t)$ of the star as a function of time, demonstrating how the local ring tangent convolved with the finite-sized disk of the star produces light curves with different local slopes. Panel (c) highlights the three significant epochs in the rate of change of ring radius $G(t) = \frac{dr}{dt}$: t_b , at closest projected separation of the star and the secondary; $t_\perp$, where the ring tangent is perpendicular to the direction of stellar motion; and $t_\parallel$, where stellar motion is tangent to the ring. $t_\parallel$ also marks where the stellar path touches the smallest ring radius.
Modelling giant extrasolar ring systems in eclipse and the case of J1407B: sculpting by exomoons? [9]

minimizes the difference between the model and measured gradients and penalizes heavily if the measured point goes above the model point. If we define $\delta_t = G(t) - g(t)$, then the cost function Δ is

$$\Delta = \sum_{t=1}^{n} \begin{cases} \delta_t & if \;\; \delta_t > 0 \\ -50 \cdot \delta_t & otherwise \end{cases} \tag{4.2}$$

In figure 4.2 is shown the results for $1SWASP\ J140747.93 - 394542.6$, from those we can obtain all the parameters of the ring system minimizing the cost function, reported in table 4.1.

b	t_b	i_{disk}	ϕ_{disk}	$t_\parallel$	v
days	days	deg	deg	days	km s^{-1}
3.92	54225.46	70.0	166.1	54220.65	33.0

Table 4.1: Rings Model Parameters

Once we have all the mechanicals parameters of the system we can fit the ring structure directly

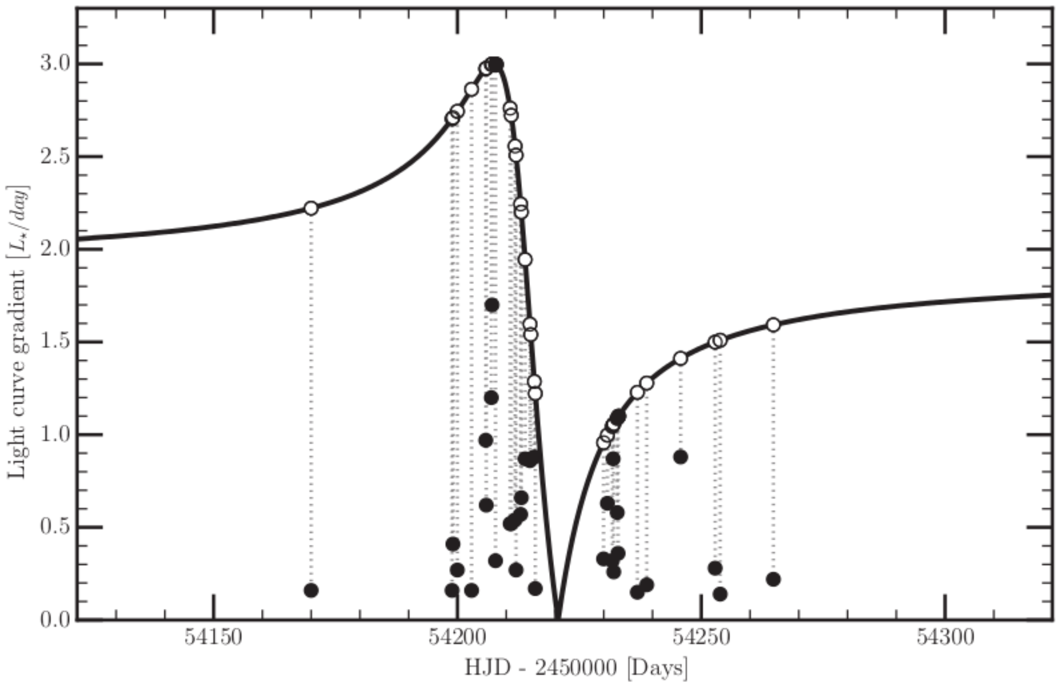


Figure 4.2: Measured gradients in the light curve of $J1407$ plotted as a function of MJD of observation. The line shows the maximum allowed gradient of the light curve $G(t)$ for a given set of disk parameters $t_{\|}$, $t_{\perp}$, i_{disk}, and ϕ_{disk}. Dotted lines connect the black circle measured values to the maximum allowed open circle theoretical maximum values on the black line.
Modelling giant extrasolar ring systems in eclipse and the case of J1407B: sculpting by exomoons? [9]

on the $I(t)$ plot obtaining finally the optical depth of the rings (and as a consequence we can make some statement about density, opacity and so on). The number of ring edges in the light curve is estimated by counting the number of slope changes identified in the light curve and indirectly implied by the change of the light curve during daylight hours. The final fit results are shown in fig 4.3.

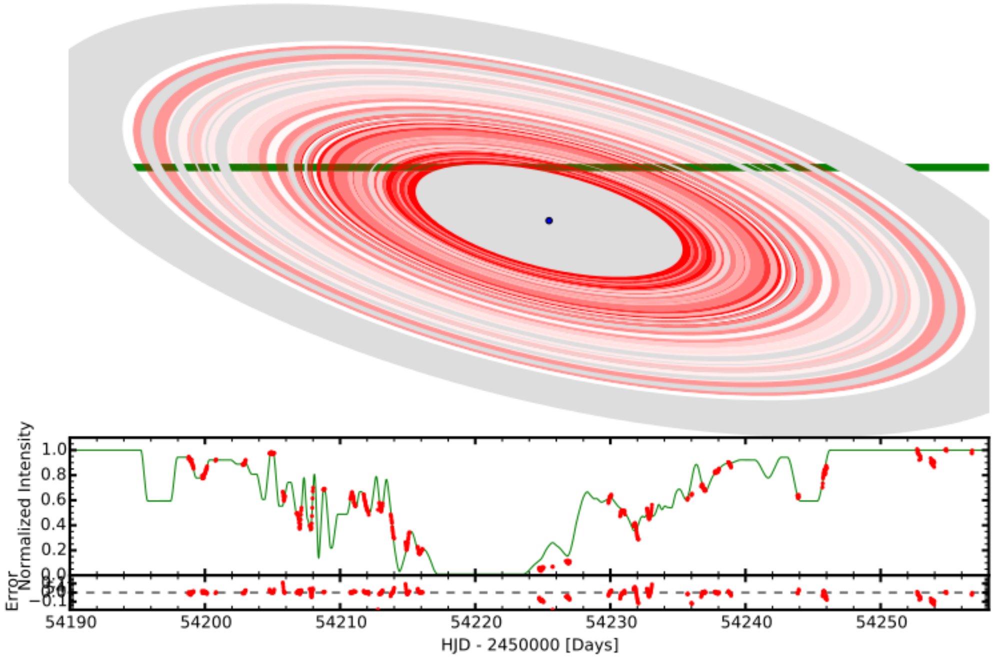


Figure 4.3: Model ring fit to $J1407$ data. The image of the ring system around $J1407b$ is shown as a series of nested red rings. The intensity of the color corresponds to the transmission of the ring. The green line shows the path and diameter of the star $J1407$ behind the ring system. The gray rings denote where no photometric data constrain the model fit. The lower graph shows the model transmitted intensity $I(t)$ as a function of HJD. The red points are the binned measured flux from $J1407$ normalized to unity outside the eclipse. Error bars in the photometry are shown as vertical red bars.
Modelling giant extrasolar ring systems in eclipse and the case of J1407B: sculpting by exomoons? [9]

Bibliography

[1] *HATNet website*, http://hatnet.org/

[2] *HATSouth website*, http://hatsouth.org/

[3] *Wide-Field Millimagnitude Photometry with the HAT: A Tool for Extrasolar Planet Detection*, G. Bakos et al., the Astronomical Society of the Pacific, **116**, 266–277, (2004)

[4] *HATSouth: A Global Network of Fully Automated Identical Wide-Field Telescopes*, G. Bakos et al., the Astronomical Society of the Pacific, **125**, 154–182, (2013)

[5] *A trend filtering algorithm for wide-field variability surveys*, G. Kovacs et al., Mon. Not. R. Astron. Soc., **356**, 557–567 (2005)

[6] *A box-fitting algorithm in the search for periodic transits*, G. Kovacs et al., A&A, **391**, 369–377 (2002)

[7] *The Use of Principal Components Analysis in Analysing Variable Star Data*, S. M. Kanbur et al., The Impact of Large-Scale Surveys on Pulsating Star Research, ASP Conference Series, **203**, 56-59 (2000)

[8] *Principal Component Analysis of RR Lyrae light curves*, S. M. Kanbur & H. Mariani, Mon. Not. R. Astron. Soc. (2004)

[9] *Modelling giant extrasolar ring systems in eclipse and the case of J1407B: sculpting by exomoons?*, M. A. Kenworthy & E. E. Mamajek, The Astrophysical Journal, **800**, 126-135 (2015)