



UNIVERSIDAD DE CHILE  
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS  
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

FINDING PERIODICITIES IN ASTRONOMICAL LIGHT CURVES USING  
INFORMATION THEORETIC LEARNING

TESIS PARA OPTAR AL GRADO DE DOCTOR EN  
INGENIERÍA ELÉCTRICA

PABLO ANDRÉS HUIJSE HEISE

PROFESOR GUÍA:  
PABLO A. ESTÉVEZ VALENCIA

MIEMBROS DE LA COMISIÓN:  
JOSÉ C. PRÍNCIPE  
SUSANA EYHERAMENDY DUERR  
JORGE F. SILVA SÁNCHEZ

Este trabajo ha sido parcialmente financiado por CONICYT-CHILE bajo los proyectos FONDECYT 1110701 y 1140816, y su programa de becas de doctorado.

SANTIAGO DE CHILE  
2014



RESUMEN DE LA TESIS PARA OPTAR AL  
GRADO DE: Doctor en Ingeniería Eléctrica  
POR: Pablo Andrés Huijse Heise  
FECHA: 31/05/2041  
PROFESOR GUÍA: Pablo Antonio Estévez Valencia.

## FINDING PERIODICITIES IN ASTRONOMICAL LIGHT CURVES USING INFORMATION THEORETIC LEARNING

The analysis of time-variable astronomical phenomena is of great interest as it helps to improve our understanding of the structure and topology of our Universe, the mechanisms of galaxy and stellar evolution, etc. The basic tool to study variability in the sky is the light curve. Light curves are time series of stellar brightness and their analysis reveals key information about the physics behind the variable phenomena. Periodic variable stars are particularly interesting. Periodic variable stars are used to estimate the size and distance-scales of our Universe, and the period is a key parameter for stellar parameter estimation, stellar classification and exoplanet detection. The precise estimation of the period is critical in order to accomplish these scientific tasks. Astronomy is experiencing a paradigm change due to the extent volumes of data generated by current astronomical surveys. In less than 10 years, hundreds of Petabytes of astronomical images and time series catalogs will be produced. Conventional astronomy does not possess the tools required for this massive data mining operation. Nowadays there is a growing need for methods with solid statistical background to do automatic astronomical time series analysis. These methods need to be robust, fully-automated and computationally efficient.

In this doctoral research I developed methods for periodicity detection and period estimation in light curves that are based on information theoretic concept of correntropy and advanced signal processing techniques. These methods are intended for automatic and efficient periodic light curve discrimination in large astronomical databases. Correntropy is a generalization of the conventional correlation to higher order statistics. In this thesis I propose the slotted correntropy estimator, the correntropy kernelized periodogram (CKP) and the correntropy non-negative matrix factorization spectrum (CNMFS). The slotted correntropy extends correntropy to unevenly sampled time series such as light curves. The CKP is a generalized periodogram that can be computed directly from the samples without regards on their sampling. The CNMFS is a high resolution spectrum that is localized on the fundamental frequency of the process.

The results presented in this thesis show that information theoretic based criteria perform better than conventional methods used in astronomy such as the LS periodogram, analysis of variance, string length and the slotted autocorrelation function (second-order methods). Including the higher-order moments of the time series into the estimation makes the proposed information-theoretic methods more robust against noise and outliers, giving them the upper hand in term of the precision of the detected periods. The proposed methods are also general as they do not pose any assumption on the underlying periodic signal (e.g. sum of sine-waves), and can be adapted heuristically (CKP) or automatically (CNMFS) to different periodic light curve shapes. The proposed methods are less prone to return a harmonic, sub-harmonic or an alias of the underlying period, a usual problem with conventional methods. The results also show that the proposed methods are more robust and less dependant on the number of samples and the time span of the light curve, i.e. the period can be recovered even if few samples or only a short piece of the light curve is available. This suggests that these methods may outperform conventional methods for early or online periodicity discrimination on surveys that are currently operating (VVV, DECam).



# Resumen

El análisis de fenómenos astronómicos variables en el tiempo es de gran interés científico pues ayuda a mejorar nuestro entendimiento de la estructura y topología de nuestro Universo, los mecanismos de evolución estelar, etc. La herramienta básica para estudiar variabilidad celeste es la curva de luz. Las curvas de luz son series de tiempo de brillo estelar y su análisis revela información clave sobre los procesos físicos tras el fenómeno variable. Las estrellas variables periódicas son particularmente interesantes, pues se usan para estimar el tamaño y las escalas de distancia en nuestro Universo, y su período es un parámetro clave para la estimación de otros parámetros estelares como la masa y el radio, para la clasificación estelar y la detección de exoplanetas. Una estimación precisa del período es crítica para el cumplimiento de estas tareas científicas. La astronomía está experimentando un cambio de paradigma debido a los extensos volúmenes de datos generados por los sondeos astronómicos actuales. En menos de 10 años, se producirán cientos de Petabytes de imágenes astronómicas y catálogos de series tiempo. La astronomía convencional no posee las herramientas que se requieren para esta operación masiva de minería de datos. Hoy en día existe una creciente necesidad por métodos con sólidas bases estadísticas para realizar análisis automático de series de tiempo astronómicas. Los métodos han de ser robustos, completamente automáticos y computacionalmente eficientes.

En esta investigación doctoral he desarrollado métodos para detección de periodicidad y estimación de período en curvas de luz que están basados en conceptos de teoría de la información de correntropía y técnicas avanzadas de procesamiento de señales. Estos métodos fueron desarrollados teniendo en mente el procesamiento eficiente de grandes bases de datos de curvas de luz. La correntropía es una generalización de la correlación convencional a estadísticos de alto orden. En esta tesis propongo la correntropía ranurada, el periodograma kernelizado de correntropía (CKP) y el espectro de correntropía mediante factorización no-negativa de matrices (CNMFS). La correntropía ranurada extiende la correntropía a series de tiempo con muestreo irregular tales como las curvas de luz. El CKP es un periodograma generalizado que puede computarse directamente de las muestras sin importar su muestreo. El CNMFS es un espectro de alta resolución que está localizado en la frecuencia fundamental del proceso.

Los resultados presentados en esta tesis muestran que los criterios basados en teoría de la información tienen un desempeño superior a los métodos convencionales usados en astronomía tales como el periodograma LS, análisis de varianza, string length y la función de correlación ranurada (métodos de segundo orden). Incluir los momentos de alto orden de la serie de tiempo hace que los métodos propuestos sean más robustos al ruido y a los outliers,

lo cual a su vez se traduce en una mayor precisión en la detección de período. Los métodos propuestos son generales, en el sentido de que no hacen supuestos sobre la señal periódica subyacente (e.g. suma de señales sinusoidales), y pueden ser adaptados heurísticamente (CKP) o automáticamente (CNMFS) a diferentes tipos de periodicidad. Los métodos propuestos son menos propensos a entregar un armónico, sub-armónico o alias del período subyacente, un problema usual de los métodos convencionales. Los resultados muestran que los métodos propuestos son más robustos y menos dependientes del número de muestras y del tiempo total de la curva de luz, es decir, el período puede ser recuperado incluso si pocas muestras o un segmento corto de la curva de luz está disponible. Esto sugiere que los métodos propuestos pueden funcionar mejor que los métodos convencionales para discriminación temprana u online de periodicidad en sondeos que están operando actualmente (VVV, DECam).

# Contents

<b>1. Introduction</b>	<b>1</b>
1.1. Astronomical Background	2
1.1.1. Variable Stars and other time-variable phenomena	4
1.2. Brief description of the scientific problem	6
1.3. Purpose of the study	7
1.3.1. Hypothesis	8
1.3.2. General objective	8
1.3.3. Specific objectives	8
<b>2. Literature review</b>	<b>12</b>
2.1. Time series analysis in the frequency domain	12
2.1.1. Lomb-Scargle periodogram	14
2.1.2. Generalized Lomb-Scargle periodogram	15
2.1.3. Direct Quadratic Spectrum Estimator	16
2.1.4. Nyquist frequency for unevenly sampled time series	16
2.2. Time series analysis in the time domain	17
2.2.1. Slotted autocorrelation	18
2.2.2. Kernel windows for the slotting technique	20
2.3. Statistical methods for unevenly sampled time series analysis based on epoch folding	21
2.3.1. Epoch Folding	21
2.3.2. String Length	23
2.3.3. Analysis of variance	24
2.3.4. Phase dispersion minimization	25
2.3.5. Shannon's entropy minimization	26
2.3.6. Conditional Entropy Minimization	27
2.4. Statistical significance analysis for finding periodicities	27
2.4.1. Periodicity test using the maximum periodogram ordinate	28
2.5. Surrogates methods	29
2.6. Kernel methods	30
2.6.1. Translation-invariant kernel functions	31
2.6.2. Kernel construction	32
2.6.3. Periodic kernel functions	32
2.7. Spectral estimation using Basis Pursuit	33
2.7.1. Overcomplete dictionaries	34
2.7.2. L1 norm minimization for BP	35

2.7.3.	Basis pursuit de-noising . . . . .	36
2.8.	Non-negative matrix factorization . . . . .	36
2.9.	Higher order moments . . . . .	38
2.10.	Information Theoretic Learning . . . . .	39
2.10.1.	An entropy measure directly estimated from data . . . . .	40
2.10.2.	Generalized correlation function: Correntropy . . . . .	42
<b>3.</b>	<b>Methods</b>	<b>46</b>
3.1.	Slotted correntropy . . . . .	47
3.1.1.	IP metric for CSD peak discrimination . . . . .	49
3.1.2.	A pipeline for period estimation using the Slotted Correntropy and the IP metric . . . . .	51
3.2.	Correntropy Kernelized Periodogram . . . . .	52
3.2.1.	A statistical test based on the CKP for periodicity discrimination . . . . .	55
3.2.2.	Periodicity detection using the CKP as the test statistic . . . . .	56
3.3.	An efficient pipeline for periodic light curve discrimination on large astronomical databases using the CKP . . . . .	57
3.3.1.	An heuristic rule to select the kernel size of the periodic kernel function . . . . .	58
3.3.2.	Normalizing the CKP against sample size and kernel sizes . . . . .	58
3.3.3.	Trial period extraction, the bands method . . . . .	59
3.3.4.	A procedure to filter spurious periods present in the EROS-2 light curves . . . . .	61
3.3.5.	Obtaining the periodicity discrimination thresholds . . . . .	62
3.3.6.	Detecting additional periodic components . . . . .	64
3.3.7.	Pipeline for periodic light curve discrimination in large astronomical databases . . . . .	65
3.3.8.	About GPGPU implementations and GPU cluster environments . . . . .	65
3.4.	Correntropy NMF Spectrum . . . . .	69
3.4.1.	Procedure for period estimation using the CNMFS . . . . .	73
3.5.	Databases . . . . .	74
3.6.	Performance criteria . . . . .	75
<b>4.</b>	<b>Results</b>	<b>79</b>
4.1.	Slotted autocorrentropy for period estimation on MACHO light curves . . . . .	79
4.2.	CKP for periodic light curve discrimination on the MACHO database . . . . .	80
4.3.	CKP for periodic light curve discrimination on the EROS-2 database . . . . .	84
4.3.1.	Efficiency in the synthetic light curve set . . . . .	85
4.3.2.	Results for selected EROS-2 fields . . . . .	86
4.3.3.	Results on EROS-2 LMC and SMC fields . . . . .	87
4.3.4.	Multimodes in the EROS-2 survey . . . . .	96
4.3.5.	Computational efficiency and processing times . . . . .	99
4.4.	Period estimation using the CNMFS . . . . .	102
4.4.1.	Setup and performance criteria . . . . .	102
4.4.2.	Period estimation in synthetic time series . . . . .	104
4.4.3.	Period estimation in light curves . . . . .	106
<b>5.</b>	<b>Conclusions</b>	<b>111</b>
5.1.	Future Work . . . . .	114

<b>Bibliography</b>	<b>116</b>
<b>Appendices</b>	
<b>Appendix A. Generation of a synthetic light curve database</b>	<b>126</b>
<b>Appendix B. Periodic light curves found in the EROS-2 survey data from the Large and Small Magellanic clouds</b>	<b>131</b>



# Chapter 1

## Introduction

Recent advances in observing, storage, and processing technologies have facilitated the evolution of astronomical surveys from observations of small and focused areas of the sky (MACHO [1], EROS [2], OGLE [3]) to deep and extended panoramic sky surveys (SDSS [4], Pan-STARRS [5], CRTS [6]), DECam [7], VVV [8]. Several new grand telescopes are planned for the next decade [9], among which is the Large Synoptic Survey Telescope (LSST; [10]) under construction in northern Chile and expected to begin operation by 2022. The LSST will generate a 150 Petabytes imaging database, and a 40 PetaBytes worth catalog associated with 50 billion astronomical objects during 10 years [11]. The LSST will stream data at rates of 2 TeraByte per hour, effectively capturing an unprecedented movie of the sky. Time-domain astronomy, *i.e.* the study of everything that varies in the sky in position or time, will experience a revolution in terms of scientific opportunities. Analyzing the LSST data will greatly improve our understanding of the known time-variable astrophysical phenomena. The LSST is pushing the limits on resolution, scale and cadence, hence it is also expected that a plethora of yet unseen phenomena will be discovered.

Conventional astronomy does not have the tools required to efficiently mine this deluge of data. In this data-intensive era of astronomy a multi-disciplinary approach that combines astronomy, statistics, informatics, data mining, machine learning and engineering is needed to cope with the challenges imposed by projects such as the LSST. In this approach, astronomy and astrophysics drive the scientific goals; statistics, signal processing, data-mining and machine learning provide the tools to analyze and interpret vast amounts of data; applied computer science and high-performance computing provide the techniques to implement efficient methods and the pipelines needed to process astronomical data in a feasible time. Nowadays, there is a growing need for fully-automated methods with solid statistical foundations for time series analysis in Peta-scale astronomical databases.

The study of variable stars is of great importance in time-domain astronomy. The analysis of variable astronomical objects paves the way towards the understanding of astrophysical phenomena, and provides valuable insights in topics such as galaxy and stellar evolution, universe topology, and others. There are certain types of variable stars [12, 13, 14] whose brightness varies following regular cycles. Examples of this kind of stars are the pulsating variables and eclipsing binary stars. Pulsating stars, such as Cepheids and RR Lyrae, ex-

pand and contract periodically effectively changing their size, temperature and brightness. Eclipsing binaries are systems of two stars with a common center of mass whose orbital plane is aligned to Earth. Periodic drops in brightness are observed due to the mutual eclipses between the components of the system. Although most stars have at least some variation in luminosity, current ground based survey estimations indicate that 3% of the stars are varying more than the sensitivity of the instruments and  $\sim 1\%$  are periodic [15].

Detecting periodicity and estimating the period of stars is of high importance in astronomy. The period is a key feature for classifying variable stars [16, 17]; and estimating other parameters such as mass and distance to Earth [18]. Light curve analysis is a particularly challenging task. Astronomical time series are unevenly sampled due to constraints in the observation schedules, the day-night cycle, weather conditions, equipment positioning, calibration and maintenance. Light curves are also affected by several noise sources such as light contamination from other astronomical sources near the line of sight (sky background), the atmosphere, the instruments and particularly the CCD detectors, among others.

Currently, most periodicity finding schemes used in astronomy are based on grid searches that take considerable computational time, make strong assumptions on the underlying signal and/or rely somehow on human visual inspection. In this thesis, automated and efficient computational methods, based on information theoretic concepts, and capable of performing robust periodicity analysis for large light curve databases are proposed. In the following sections a brief astronomical background on variable stars and a description of the scientific problem are given. In Chapter 2 the literature on both time domain and frequency domain light curve analysis is reviewed. The methods proposed in this thesis are presented in Chapter 3. The results are reported in Chapter 4. Finally, in Chapter 5, the conclusions are drawn.

## 1.1. Astronomical Background

Photometry is the branch of astronomy dedicated to the precise measurement of visible electromagnetic radiation from astronomical objects. To achieve this, several techniques and methods are applied to transform the raw data from the astronomical instruments into standard units of flux or intensity [19]. Recently, photometry have had a revolution due to the increasing availability and refinement of CCD (charged-coupled devices) technology. In comparison with conventional photoelectric photometers, CCD sensors offers a better quantum efficiency<sup>1</sup>; a wider spectral response; robustness to noise; and digitized output. Using CCD sensors it is possible to capture astronomical images of greater quality with shorter exposition times. This has entailed an increase in both the number and extension of astronomical surveys such as MACHO [1], EROS [2], OGLE [3], SDSS [4], Pan-STARRS [5], CRTS [6] and more recently the LSST [10]. The objectives of these astronomical surveys are among others to map regions of the sky, to classify celestial objects and to search for astronomical transient events.

Using the data obtained from CCD sensors and the photometry techniques, the apparent brightness of stellar objects can be obtained. By comparing this apparent brightness with

---

<sup>1</sup>Percentage of detected photons.

those of other “control stars” in the same digital image an absolute brightness value is obtained (differential photometry). Investigations of the brightness of distant stellar objects are necessary for understanding their structure and behavior.

The basic tool in the analysis of astronomical brightness variations is the **light curve**. A light curve is a plot of the magnitude of an object’s electromagnetic radiation (in the visible spectrum) as a function of time. The assignment of magnitudes comes from a logarithm system of ranking brightness, originally developed by Hipparchos (140 AD). The magnitude is inversely proportional to the brightness, so stars with higher magnitudes are fainter than those with smaller magnitude values. The time in light curves is commonly expressed in Julian dates. By analyzing the light curves, astronomers can gain a better understanding of stellar objects and perform tasks such as transient event detection, parameter estimation, and variable star detection and classification.

The challenges of astronomical time series analysis are not related exclusively to the sheer size of the databases but also to the characteristics of the data itself. Astronomical time series are unevenly sampled due to constraints in the observation schedules, telescope allocations and other limitations. When observations are taken from Earth the resulting light curves will have periodic one-day gaps. The sampling is randomized because observations for each object happen at different times every night. The cycles of the moon, bad weather conditions and sky visibility impose additional constraints which translate into data gaps of different lengths. Space observations are also restricted as they are regulated by the satellite orbits, also additional processing is required in order to compensate for the drift of the telescope. Discontinuities in light curves can also be caused by technical factors: repositioning of the telescopes, calibration of equipment, electrical, and mechanical failures, etc.

Astronomical time series are also affected by several noise sources. These noise sources can be broadly categorized into two classes. The first class is related to observations, such as the brightness of closer astronomical objects, and atmospheric noise due to refraction and extinction phenomena (scattering of light due to atmospheric dust). On the other hand, there are noise sources related to the instrumentation, in particular to the CCD cameras, such as sensitivity variations of the detector, and thermal noise. In general, errors in astronomical time series are non-Gaussian and heteroscedastic, *i.e.* the variance of the error is not constant, and changes along the magnitude axis [20].

Other common problematic situations arising in time-domain astronomy are the sample-selection bias and the lack of balance between classes. Generally the astrophysical phenomena of interest represents a small fraction of the observable sky, hence the vast majority of the data belongs to the “background class”. This is especially noticeable when the objective is to find unknown phenomena, a task known as novelty detection. Sufficient coverage and exhaustive labeling are required in order to have a good representation of the sample, and to assure capturing the rare objects of interests.

In time series analysis, mathematical and computational techniques are applied in order to extract the most information from the observed system. This information is used to: explain the variability of the system, find similarities with other known systems, find the limits of the system, predict its behavior, etc. If a time series repeats itself in time, then the most condensed information that could model the system is the oscillation period. Detecting

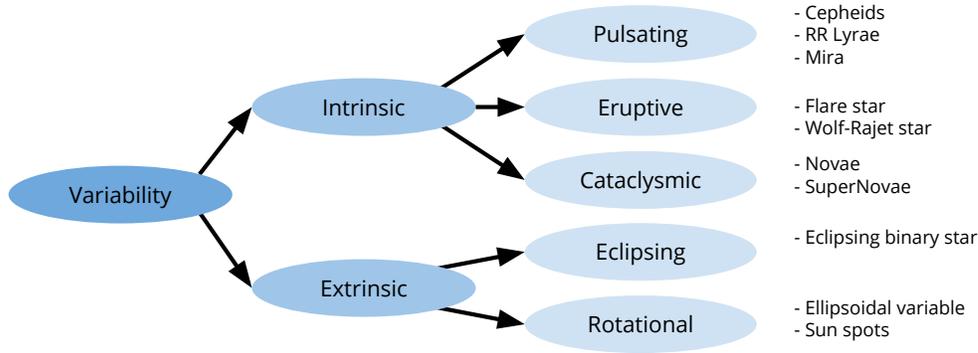


Figure 1.1: Variable star topological classification.

periodicity and estimating the oscillation period are fundamental tasks in astronomical time series analysis.

In the following section several time-domain astronomical phenomena are presented, with emphasis on periodic phenomena that varies in the optical spectrum. The scientific interest behind these astronomical objects and events is also explained.

### 1.1.1. Variable Stars and other time-variable phenomena

Among the “observable stars” there is a particular group called the **variable stars** [12, 13, 14]. Variable stars correspond to stellar objects whose brightness, as observed from Earth, fluctuates in time above a certain variability threshold defined by the sensitivity of the instruments. Although most stars have at least some variation in luminosity, current estimations indicate that 3% of the stars are varying more than the sensitivity of the instruments and  $\sim 1\%$  are periodic [15].

Variable star analysis is a fundamental pivot in the study of stellar structure and properties, stellar evolution and the distribution and size of our Universe. In what follows the major categories of variable stars are briefly described while trying to emphasize the scientific interest behind each of them. For a more in-depth definition of the objects and their mechanisms of variability, the reader can refer to [13]. The relation between different classes of variable stars, according to [14], is summarized by the tree diagram shown in Figure 1.1. The class definitions presented here follow the one given by the General Catalog of Variable Stars (GCVS) [21].

Variable stars can be broadly classified as either intrinsic or extrinsic variables depending on the phenomena that explain their fluctuations in brightness. In the intrinsic case, internal physical processes occurring within the star are the reasons behind their variability. On the other hand, the change in brightness of the extrinsic variables is explained by the orbital and/or rotational configuration of the star, the existence of companion stars or broadly speaking the relation between neighboring stars. Notable examples of variable stars belonging to these categories are described in Table 1.1.

The analysis of intrinsic variable stars is of great importance for the study of stellar nuclei and evolution. Some classes of intrinsic variable stars can be used as distance markers to study the distribution and topology of the Universe. Cepheid and RR Lyrae stars are considered standard candles because of the relation between their pulsation period and their absolute brightness (period-luminosity relationship). With the period and the apparent brightness measured from the telescope it is possible to estimate the distance from these stars to Earth [22]. Type 1A Supernovae are also standard candles, although they can be used to trace much longer distances than Cepheids and RR Lyrae [23]. In fact the 2011 Nobel prize in Physics was bestowed to three scientists for the discovery of the accelerating expansion of the Universe through observations of distant supernovae<sup>2</sup>. The period of the eclipsing binary (EB) is a key parameter in astrophysics studies as it can be used to calculate the radii and masses of the components [18] and the distance to Earth of the binary system with great accuracy [24]. Rotational extrinsic variables change their brightness due to inhomogeneities in their surface and their study is of great importance in the field of Helioseismology.

The period of pulsating variables (Fig. 1.2a), eclipsing binaries (Fig. 1.2b), and other periodic stars is also critical to characterize these celestial objects and obtain valuable insight regarding their internal structure and evolution. Using the period it is possible to classify a variable star into one of the known classes with great accuracy [16, 17, 25], or to discriminate new classes of periodic variable stars [26, 27]. The shape of a periodic variable star can be studied in great detail using a technique called epoch folding. In this technique one uses a candidate period to transform the time axis of a light curve so that the behavior occurring during one phase is exposed. If the candidate period is correct then a clear shape will appear in the phase diagram. The epoch folding technique is described in detail in Chapter 2. Examples of light curves and phase diagrams of periodic variable stars are shown in Fig. 1.3.

Table 1.2 describes other variable phenomena which are not related to variable stars, such as the gravitational lensing, Active Galactic Nuclei (AGN) and transiting extrasolar planets. Gravitational lensing events in light curves may reveal the presence of dark massive planets. The detection of massive astrophysical compact halo objects (MACHOs) through microlensing was the objective of the MACHO [1], EROS [2] and OGLE [3] surveys. The scientific objective of this search was to study if the MACHOs represent a relevant fraction of the missing “dark matter” of the Universe.

Extrasolar planets [28] or exoplanets are planets outside our solar system. The detection and characterization of exoplanets is a topic of great scientific interest. Exoplanet research is broad and involves astrophysics and astronomy but also new sciences such as astrochemistry (exoplanet atmospheric composition), astrobiology (detection of biological signatures), and astrogeology among others. Current methods for exoplanet detection include the transit method, gravitational microlensing, radial velocity and astrometry [29].

If the orbital plane of an extrasolar planet is aligned with the Earth, then periodic drops in the brightness of its host star will appear on its light curve. This indirect way of finding extrasolar planets from the light of its host star is called the transit method. The periodic signature of the transiting exoplanet can be very difficult to find, specially when the planet is very small with respect to its host star. Robust methods for period estimation and periodicity

---

<sup>2</sup>[http://www.nobelprize.org/nobel\\_prizes/physics/laureates/2011/](http://www.nobelprize.org/nobel_prizes/physics/laureates/2011/)

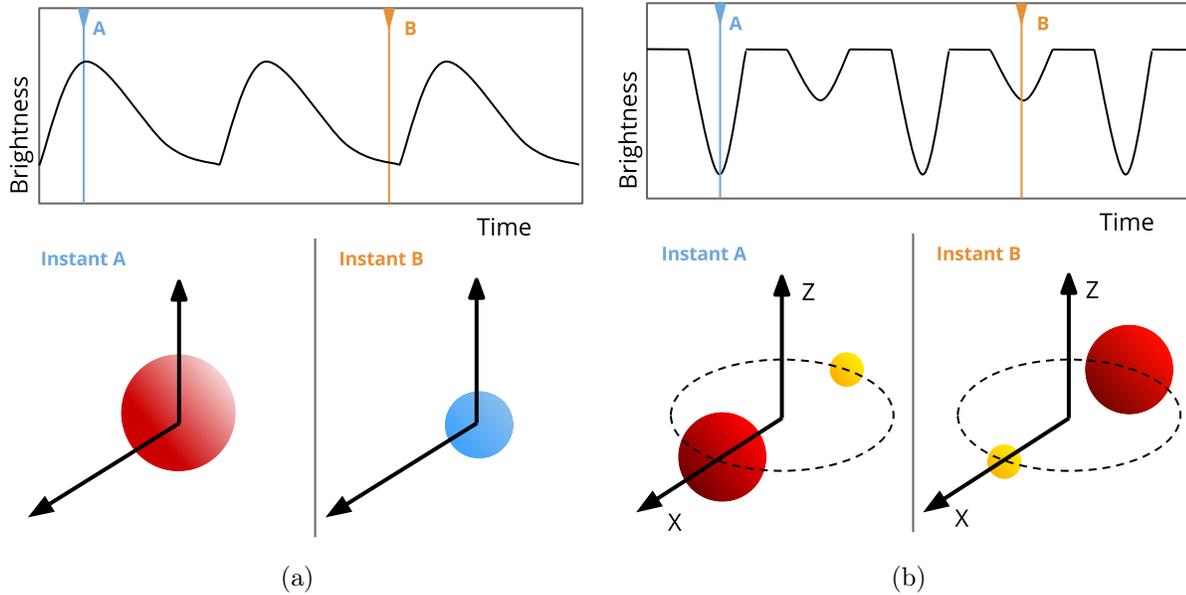


Figure 1.2: (a) Light curve of a pulsating variable star (upper left panel), such as a Cepheid or RR Lyrae. The star pulsates periodically changing in size, temperature and brightness which is reflected on its light curve. (b) Light curve of eclipsing binary star (upper right panel). The lower panels show the geometry of the binary system at the instants were the eclipses occur. The periodic pattern in the light curve is observed because the Earth (X axis) is aligned with the orbital plane of the system (Z axis).

detection are critical for exoplanet detection, particularly in the case of the transit and radial velocity methods.

## 1.2. Brief description of the scientific problem

Light curve analysis is invaluable in the understanding of the nature of variable stars. By analyzing the brightness fluctuations and the morphology of light curves it is possible to extract relevant parameters of the studied stellar objects. These parameters are then used to classify a stellar object, thus helping in the development of a cosmic census. Certain astronomical events detected from the light curves could give us a better understanding of the structure of the Universe. Extrasolar planets leave a signature in the light curves of their corresponding stars which can be detected with period search techniques.

Computational intelligence methodologies can be applied to astronomical time series databases in order to:

- a) Discriminate between periodic, quasi-periodic and non-periodic light curves

---

<sup>3</sup>Such as the semi-regular (characterized by their quasi-periodic variability) and the slowly irregular variables.

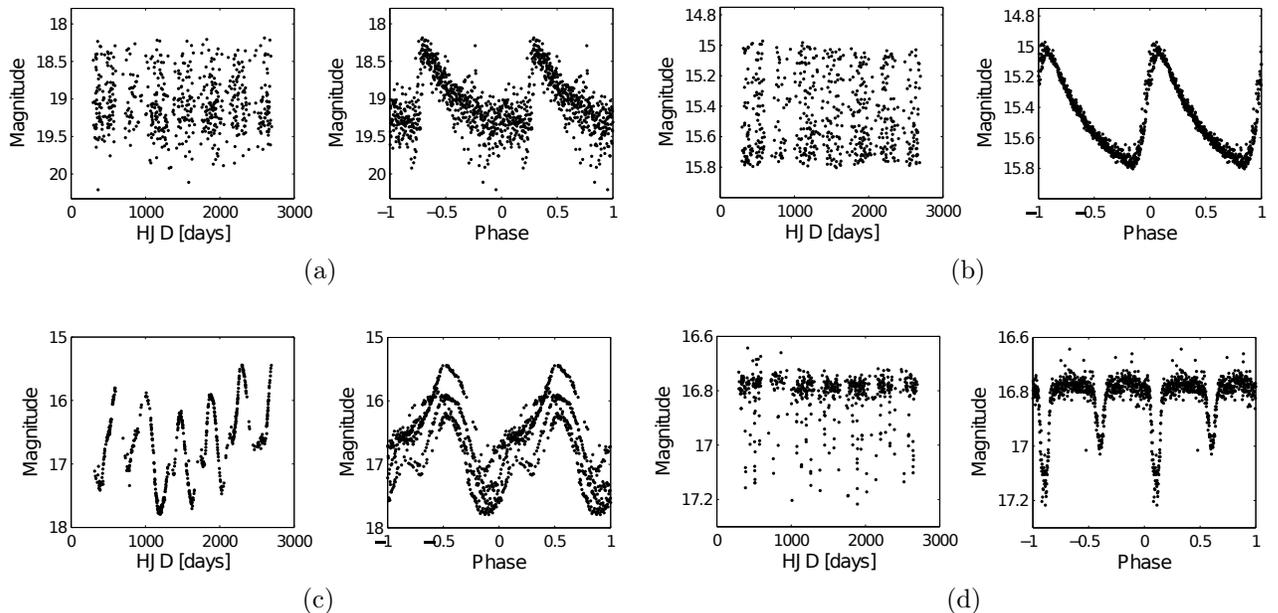


Figure 1.3: Light curve and phase diagram of an RR Lyrae (a), Cepheid (b), Mira (c) and eclipsing binary star (d), respectively. The phase diagram is obtained using the underlying period of the light curves and the epoch folding transformation [13]. If the folding period is correct a clear profile of the periodicity will appear in the phase diagram.

b) Estimate the fundamental period of periodic variable stars

But there are two main difficulties, related to

1. **Astronomical time series:** Light curves are noisy, unevenly sampled and have data gaps of different sizes.
2. **Big data analysis:** Light curves databases are growing in number and size. We are approaching the Petabyte era of astronomical surveys [10], in which billions of objects will be followed every night producing an unprecedented number of light curves. Light curve generation rates will keep increasing in the near future.

There is a growing need for fully-automated and computationally-efficient methods capable of performing robust light curve analysis techniques for very large astronomical databases.

### 1.3. Purpose of the study

In this thesis, a framework for astronomical time series analysis based on information theoretic learning (ITL) [31] is developed. This framework is focused on the tasks of periodicity detection and period estimation in light curves. The proposed framework intends to provide a systematic and efficient way to detect periodic variable stars and estimate their fundamental period in large astronomical time series databases.

The main tools proposed in this thesis are the slotted autocorrentropy estimator, the Correntropy kernelized periodogram (CKP) and the Correntropy Nonnegative Matrix Factorization Spectrum (CNMFS). These methods are based on the Correntropy function [31], a generalized correlation function that measures similarities in random processes. Contrary to second-order statistics such as the conventional correlation, correntropy extracts information directly from the probability density function (PDF) of the random process, which gives correntropy a higher discriminate power and robustness to noise.

- The slotted autocorrentropy estimator is a time-domain metric to estimate the period of a light curve. Contrary to the conventional autocorrentropy estimator the slotted correntropy can be applied on unevenly sampled random processes.
- The CKP is a frequency-domain metric that combines the correntropy function with a periodic kernel in order to assess periodicity directly from the samples of a light curves.
- The CNMFS is a spectrum estimator for the slotted autocorrentropy with advantages over classical Fourier-based estimators: it has a higher frequency resolution, it is well localized in the fundamental frequencies (less harmonic content), and is less affected by aliasing and spurious periodicities due to noise.

Established methods for light curve analysis are reviewed in Chapter 2. Current methods do not perform well for all kinds of variable stars, and require the supervision of astronomers to discriminate between the obtained results (by visual inspection). A method that requires human intervention is not appropriate to analyze large astronomical databases. The proposed methods are presented in Chapter 3.

### 1.3.1. Hypothesis

Metrics based on ITL, and particularly the slotted autocorrentropy, CKP and CNMFS, will help to develop pipelines for light curve analysis that are more precise, more efficient and more robust than conventional and established methods used in astronomy. This is because ITL metrics use the whole information contained in the data samples, while most current methods are based on second-order statistics.

### 1.3.2. General objective

To develop an fully-automated, computationally-efficient and robust framework for finding periodicities and estimating periods in astronomical time series based on information theoretic learning, kernel methods, signal processing tools and statistical criteria.

### 1.3.3. Specific objectives

1. To design a method to detect periodicities in light curves and estimate the underlying period in periodic light curves.

2. To compare the proposed methods with established methods in an astronomical time series dataset with available reference periods. In periodicity detection tasks, performance will be measured in terms of the accuracy (correct classification rate), where the classes are periodic and non-periodic stars. In the period estimation task the performance will be measured in terms of the estimated period precision by comparing it with the reference periods.
3. To apply the proposed astronomical light curve analysis framework in a large untested astronomical time series database. Performance will be measured in terms of the number of true periodic variable stars detected or correctly estimated and the computational time per light curve.

Table 1.1: Examples of variable stars

Object name	Variability	Description
Cepheid	Intrinsic	Radially pulsating supergiant star. It expands and contract periodically changing its size, temperature and brightness. The period ranges between 1 and 100 days. Fig. 1.2a shows a diagram of a pulsating variable.
RR Lyrae	Intrinsic	Radially pulsating star. Older, with a lower mass, and less luminous than Cepheids. The period ranges from 0.3 to 1.2 days.
Mira	Intrinsic	Pulsating red giant star. These variables are a thousand times brighter than our sun and have periods longer than 80-days. The Miras and other giant stars <sup>3</sup> belong to the Long Period Variable (LPV) class.
Flare star	Intrinsic	Eruptive variable star. Material is violently ejected from their corona (the most outer region of a star) in a non-periodic basis. The brightness increases during this event.
Nova	Intrinsic	Cataclysmic binary star system with very close components. The hotter component “steals” material from the cooler component which then becomes the catalyst of the explosive reaction. The brightness of the Nova increases in $\sim 10$ orders of magnitude during the course of hundreds of years and then gradually returns to its former brightness.
SuperNova	Intrinsic	The final explosion of a massive star. During this transient event, the brightness of the system increases by $\sim 20$ orders of magnitude. The material ejected from the explosion forms a gas cloud (the basis from which new stars will be formed) and the remnant of the star will evolve into either a neutron star or a black hole depending on the its original mass.
Eclipsing Binary	Extrinsic	Binary star system with orbital plane aligned with Earth. The mutual eclipses appear as periodic brightness drops in the light curve. A diagram showing the geometrical configuration of the system is shown in Fig. 1.2b.
Pulsar	Extrinsic	Highly magnetized and dense neutron stars with fast rotational speeds. Pulsars emit electromagnetic radiation with periods that range between 1ms and 10s. These emissions can be detected if the emission axis of the pulsar is aligned with Earth.

Table 1.2: Examples of non-stellar variable sources

Name	Description
Gravitational lensing	An increase of several orders of magnitude in the observed brightness of a star due to a massive dark object passing in front of it and acting as a lens. The dark object bends the light of the source. If the dark object is of planetary size the effect is called microlensing [30].
Active Galactic Nuclei	A compact region in the center of a galaxy characterized by their variable, strong and broad electromagnetic emissions. The most studied AGN is the quasar (quasi stellar radio sources). Quasars are one of the most luminous objects in the sky and are characterized by their strong stochastic variability across wavelengths and on timescales.
Transiting extrasolar planet	Planet outside our solar system, orbiting a star different than our sun. If the orbital plane of the exoplanet is aligned with the Earth, periodic dips will appear in the light curve of its main star. Small planets may induce a shallow minimum that needs to be discriminated against the noise of the light curve.

# Chapter 2

## Literature review

In this section the theoretical background for the methods proposed in this thesis is established. First, the literature for astronomical light curve periodicity detection is reviewed. The methods found in the literature are categorized depending on the domain in which they operate. Frequency-domain methods are based in the discrete Fourier transform. These methods analyze the response of the process to given set of frequencies. Time-domain methods are based on the correlation function, *i.e.* they try to find second-order similarities within the time series. Statistical methods based on the epoch folding transformation are also reviewed. The epoch folding technique produces a phase diagram of the light-curve using a trial period. Different statistical methods can be applied on the phase diagram in order to test the validity of the trial period. Secondly, the literature for statistical tests on spectral ordinates and surrogate tests for unevenly sampled time series is reviewed. Thirdly, a brief introduction to kernel methods and kernel construction is given. After that, methods for sparse decomposition of time series on overcomplete dictionaries are reviewed. Finally, the framework of Information Theoretic Learning (ITL) is introduced in order to present the correntropy function, a generalized correlation function for random processes.

### 2.1. Time series analysis in the frequency domain

Fourier analysis can be described as a technique in which the time series is represented by an infinite number of sine and cosine functions using different amplitudes, frequencies and phases. For a time series, the magnitudes of these components are calculated using the Fourier transform. The time series is then mapped to frequency space and the result is called the Fourier spectrum of the time series. Usually the mapping is done using the discrete Fourier transform (DFT). The DFT of a discrete signal  $\{x_n\}$  with  $n = 0, \dots, N - 1$  is defined as

$$X_k = \mathcal{F}[[x_n]] = \sum_{n=0}^{N-1} x_n \cdot \exp\left(-j2\pi n \frac{k}{N}\right) \quad k = 0, 1, \dots, N - 1 \quad (2.1)$$

where the frequencies of interest are  $f = \frac{k}{N}F_s$ , and  $F_s$  is the sampling frequency of the time series. To represent a function accurately, the original function must be sampled at a sufficiently high rate. The appropriate rate for a uniformly sampled time series is determined by the Sampling Theorem [32]. This theorem provides a cut-off frequency called the Nyquist frequency  $f_N$  which is defined as

$$f_N = \frac{F_s}{2}, \quad (2.2)$$

where  $F_s$  is the sampling frequency. Any frequencies greater than  $f_N$  present in the signal will be aliased<sup>1</sup> at lower frequencies in the Fourier spectrum. If the signal is filtered in  $[-f_N, f_N]$  and regularly sampled (constant  $F_s$ ) then, in accordance to the Sampling Theorem, no aliasing will occur [32].

For a discrete time but continuous frequency representation the discrete time Fourier transform (DTFT) is used. The DTFT of a discrete signal  $\{x_n\}$  with  $n \in \mathbb{Z}$  and  $|n| \leq M$  is

$$X_M(f) = \mathcal{F}[\{x_n\}] = \sum_{n=-M}^M x_n \cdot \exp\left(-j2\pi f \frac{n}{F_s}\right), \quad (2.3)$$

where the range of frequencies  $f \in [-f_N, f_N]$ .

In time series analysis it is often more convenient to study the power spectral density function rather than the Fourier transform of the signal. The power spectrum represents the distribution of the signal power over the frequency interval. For a discrete signal  $\{x_n\}$  its power spectral density (PSD) is defined as

$$\begin{aligned} P_{XX}(f) &= \lim_{M \rightarrow \infty} \frac{1}{2M+1} |X_M(f)|^2 \\ &= \lim_{M \rightarrow \infty} \frac{1}{2M+1} \left| \sum_{n=-M}^M x_n \cdot \exp\left(-j2\pi f \frac{n}{F_s}\right) \right|^2 \end{aligned} \quad (2.4)$$

where  $f \in [-f_N, f_N]$  and  $X_M(f)$  is the DTFT of  $\{x_n\}$ .

Power spectral density estimation is a field on its own. Methods for power spectral density can be classified as nonparametric or parametric. Parametric estimation methods rely on the development of models for signal generation. Nonparametric methods rely on the direct use of the available data. In what follows the sample power spectrum, also known as the periodogram, is reviewed. The periodogram is a nonparametric estimator of the PSD. The periodogram of a discrete signal  $\{x_n\}$  is defined as

$$\hat{P}_{XX}(f) = \frac{1}{M} \left| \sum_{n=0}^{M-1} x_n \cdot \exp\left(-j2\pi f \frac{n}{F_s}\right) \right|^2 \quad -f_N < f < f_N, \quad (2.5)$$

---

<sup>1</sup>The aliasing phenomenon can be described as a leakage of power from high frequencies to much lower frequencies.

which corresponds to the DTFT of the signal  $\{x_n\}$  multiplied by a rectangular window in time  $w_R(n)$  of length  $M$ . This translates in frequency domain to

$$\widehat{P}_{XX}(f) = \mathcal{F}[[w_R(n)]] * P_{XX}(f) = \frac{\sin(M\pi f)}{\sin(\pi f)} * P_{XX}(f), \quad (2.6)$$

where  $*$  is the convolution operator and  $P_{XX}(f)$  is the PSD. Eq (2.6) shows that the periodogram is a smoothed sampled PSD. If  $M$  tends to infinity the expected value of the periodogram equals that of the PSD (it is an unbiased estimator), but the variance of the periodogram does not approach zero[32].

The quality of an estimator is measured in terms of its bias<sup>2</sup> and variance, but usually there is a tradeoff between them. Ideally, a good estimator should have a small bias and a small variance. The bias and variance of the periodogram can be controlled by selecting an appropriate window function and by periodogram smoothing. Different spectral window functions<sup>3</sup> have different associated bias and variance characteristics. Periodogram smoothing consist of averaging several periodograms computed from disjoint segments of the time series. Reducing the length of the segments will reduce variance and increase bias of the estimator [32].

The fundamental period of a periodic time series can be extracted from its periodogram. In ideal conditions, the fundamental period is the inverse of the frequency associated with the global maximum (highest peak) of the periodogram.

### 2.1.1. Lomb-Scargle periodogram

The Lomb-Scargle (LS) periodogram is an extension of the conventional periodogram (Eq. (2.5)) for unevenly sampled time series. It was proposed independently in [33] and [34]. This technique evaluates the sample power spectrum only in the available random sampled points of the time series. Lomb proposed to fit the time series with a trigonometric model in a least squares sense, which is done minimizing

$$E(\omega) = \sum_{i=0}^N (x_i - A \cos(\omega t_i - \phi) - B \sin(\omega t_i - \phi))^2, \quad (2.7)$$

where  $\{t_i, x_i\}$  with  $i = 0, \dots, N$  corresponds to the irregularly sampled data points of the time series,  $\omega$  is the angular frequency and  $\phi$  is the phase. The minimization is done with respect to constants  $A$  and  $B$ . Lomb imposed an additional restriction

$$\tan(2\phi) = \frac{\sum_{i=0}^N \sin(2\omega t_i)}{\sum_{i=0}^N \cos(2\omega t_i)}, \quad (2.8)$$

---

<sup>2</sup>The bias is the difference between the expected value of the estimator and the true value (PSD).

<sup>3</sup>*e.g.* rectangular, Tukey, Hamming, Parzen, Blackman, Kaiser, etc.

which defines the phase parameter and guarantee the orthogonality of the trigonometric model for the sampling times  $t_i$ . Finally, the LS periodogram as a function of the angular frequency  $\omega$ , is computed as

$$P_{LS}(\omega) = \frac{1}{2\sigma^2} \left( \frac{\left[ \sum_{i=0}^N x_i \cos(\omega t_i - \tau) \right]^2}{\sum_{i=0}^N \cos^2(\omega t_i - \tau)} + \frac{\left[ \sum_{i=0}^N x_i \sin(\omega t_i - \tau) \right]^2}{\sum_{i=0}^N \sin^2(\omega t_i - \tau)} \right) \quad (2.9)$$

where  $\sigma$  corresponds to the standard deviation of the time series and  $\phi$  is calculated using Eq. (2.8). The maximum of the LS power spectrum corresponds to the angular frequency whose model best fitted the time series in a least squares sense. The original LS periodogram requires  $(10N)^2$  operations to analyze  $N$  points. In [35] an alternative implementation based on the fast Fourier transform was proposed. This new implementation reduces the number of operations to  $10^2 N \log(N)$ .

Although the LS periodogram has a strong theoretic background, it presents the following drawbacks:

- The LS periodogram is optimized to identify sinusoidal periodic fluctuations in time series. The LS periodogram is not optimal when the time series contains non-sinusoidal periodic fluctuations, such as eclipsing binary stars light curves.
- Discriminating the underlying period from the peaks of the LS periodogram is not easy. The periodogram is contaminated with spectral peaks associated to harmonics and aliases of the underlying period, and spurious periodicities due to the sampling and the noise.

### 2.1.2. Generalized Lomb-Scargle periodogram

In [36] a generalization of the Lomb-Scargle periodogram is presented. The generalized Lomb-Scargle (GLS) periodogram solves two shortcomings of the original LS periodogram. Firstly, the GLS introduces an offset constant which lifts the assumption that the mean of the data is equivalent to the mean of the fitted sine functions. Hence the GLS performs a full sine-wave fitting of the form  $y(t) = A \cos(\omega t) + B \sin(\omega t) + C$ , where  $C$  is the newly introduced offset. Secondly, the GLS takes into account the measurement errors by introducing a weighted sum. The GLS achieves to minimize the square difference between the data and the model

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - y(t_i))^2}{\sigma_i^2},$$

where  $\sigma_i$  are the measurement errors ( $\chi^2$  fit). In comparison with the original LS periodogram, the GLS periodogram is less susceptible to aliasing and gives a better estimation of the spectral ordinates.

### 2.1.3. Direct Quadratic Spectrum Estimator

The Direct Quadratic Spectrum Estimator (DQSE), defined in [37], was developed for time series with irregular sampling or missing data. For an irregularly sampled process  $\{t_i, x_i\}$  with  $i = 1, \dots, N$ , the DQSE is defined as

$$Z(f) = \frac{1}{T} \sum_{i=1}^N \sum_{j>i}^N x_i x_j \cos(2\pi f(t_j - t_i)) D(t_j - t_i) F(t_i, t_j), \quad (2.10)$$

where  $T$  is the effective record length,  $D(\cdot)$  is a lag window, and  $F(\cdot, \cdot)$  is the data spacing factor function. In the original definition a Hanning window is used as the lag window. The data spacing factor corresponds to the area that the product  $x_i x_j$  (covariance) represents in the double integration. For regularly sampled processes  $F(t_i, t_j) = (\Delta t)^2$ , where  $\Delta t$  is the sampling rate of the time series. In the case of arbitrary irregular spacing, *i.e.* not following any particular distribution, a simple trapezoidal factor can be used,

$$F(t_i, t_j) = \left( \frac{t_{i+1} - t_i}{2} - \frac{t_i - t_{i-1}}{2} \right) \left( \frac{t_{j+1} - t_j}{2} - \frac{t_j - t_{j-1}}{2} \right) = \delta_i \delta_j, \quad (2.11)$$

which corresponds to a local definition of the sampling rate.

The DQSE provides a way to estimate the spectrum using all the available samples of the irregularly sampled time series. Contrary to the LS periodogram it does not require fitting a model to the data, hence it is more general as it makes no assumption about the process. One disadvantage of the DQSE, with respect to the LS periodogram, is its  $O(N(N-1)/2)$  computational complexity.

### 2.1.4. Nyquist frequency for unevenly sampled time series

The range of frequencies in which the spectral density is estimated has to be carefully selected. If not, important information may appear as aliases in the spectrum. When dealing with unevenly sampled time series the original definition of the Nyquist frequency (Eq. 2.2) does not apply.

One option is to search for frequencies up to the inverse of half of the average sampling frequency of the time series [34]. The maximum frequency, for an unevenly sampled time series  $\{t_i, x_i\}$  with  $i = 1, \dots, N$  was defined in [34] as

$$f_{max} = \frac{1}{\frac{2}{N} \sum_{i=1}^{N-1} (t_{i+1} - t_i)} = \frac{N}{2(t_N - t_1)}.$$

This option has been widely used although it is only intuitively correct.

In [38] it has been proved that the Nyquist frequency for unevenly sampled time series is best estimated as:

$$f_{max} = \frac{1}{2p} \geq \frac{N}{2(t_N - t_1)}$$

where  $p$  is the greatest common divisor (gcd) for all  $(t_i - t_1)$ .

## 2.2. Time series analysis in the time domain

A set of independent samples  $\{x_i\}$  of a random variable  $X$  with Gaussian (normal) distribution can be fully characterized by its mean and variance. If this set of samples form part of a time series then it is also necessary to measure its time structure in order to fully characterize it. To do this the concept of neighborhood is added, where neighboring values of the time series are expected to be correlated.

The correlation of a signal with itself is called the time autocorrelation function. The true autocorrelation of a random process  $X(t)$  with  $t \in T$  and  $T$  being an index set, is defined as

$$r_X(\tau) = \mathbb{E}[X(t)X(t + \tau)], \quad (2.12)$$

where  $\mathbb{E}$  is the mathematical expectation operator. The autocorrelation is a function of the time lag  $\tau$ .

For a discrete time series  $\{x_n\}$  with  $n = 0, \dots, N - 1$ , the autocorrelation function (2.12) can be estimated through the sample mean, obtaining the following expression

$$\hat{r}_X(m) = \frac{1}{N - m + 1} \sum_{n=m}^{N-1} x_n \cdot x_{n-m}, \quad (2.13)$$

where  $m$  is the discrete time lag. Eq. (2.13) is called the sample autocorrelation and it is an unbiased estimator of the autocorrelation function.

The sample autocorrelation can be used as a similarity metric between different segments of the time series. This means that the autocorrelation is appropriate for the detection of repetitive patterns or periodicities in the time series. For example, if the time series' values separated by time lag  $\tau$  are similar, then the sample autocorrelation at lag  $\tau$  will reach a local maximum.

The sample autocorrelation provides a way to do time series analysis without leaving the time domain. Time domain analysis methods often have an easier interpretability and less computational costs than frequency domain analysis methods. On the other hand, time series presenting a multiperiodic behavior are more difficult to evaluate using time domain methods.

The Wiener–Khinchin theorem [32] states that the power spectral density (PSD) of a wide-sense stationary random process is the Fourier transform of the autocorrelation function (2.12), *i.e.*

$$P_{XX}(f) = \sum_{n=-\infty}^{\infty} r_X(n) \cdot \exp\left(-j2\pi f \frac{n}{F_s}\right) = \mathcal{F}\llbracket r_X(n) \rrbracket.$$

In the same way, the periodogram estimator of the PSD is the Fourier transform of the sample autocorrelation (2.13). The Wiener–Khinchin theorem provides an alternative for PSD estimation in the time domain.

However the sample autocorrelation (Eq. 2.13) cannot be computed if the time series is unevenly sampled. The following options have been used for dealing with unevenly sampled time series:

- Resample the unevenly sampled time series to a regularly sampled time grid. Then, compute the sample autocorrelation of the interpolated time series. Methods such as linear interpolation, polynomial interpolation or splines [39] can be used to resample the time series.
- Select a slot size to compute the slotted autocorrelation directly from the unevenly sampled data. The slotting technique [40] is presented in the following section.

### 2.2.1. Slotted autocorrelation

In [40] a method to estimate the autocorrelation function of an unevenly sampled time series was proposed. This method was called the slotting technique and was applied to estimate the autocorrelation function and the power spectral density from time series of Laser Doppler Anemometry (LDA) data. LDA is a common tool in fluid dynamics research and LDA data is characterized as being unevenly sampled. The slotting technique for discrete correlation estimation has been extended considerably in the LDA field [41, 42, 43]. The classical slotting technique and some of its extensions are reviewed in what follows.

For a random process  $\{X_i\}$  with  $i = 1, \dots, N$ , the applications of the slotting technique can be summarized as

1. Select the slot size parameter  $\Delta\tau$ .
2. Generate a set of discrete lags as:  $k \cdot \Delta\tau$ , where  $k = 0, 1, \dots, [\frac{\tau_{max}}{\Delta\tau}]$ ,  $\tau_{max}$  is the maximum lag and  $[\cdot]$  is the nearest integer function.
3. For each lag  $k\Delta\tau$  and each pair of indexes  $i$  and  $j$  associated to samples  $(t_i, x_i)$  and  $(t_j, x_j)$  respectively, compute the slotting condition  $B_{k\Delta\tau}$  as

$$B_{k\Delta\tau}(t_i, t_j) = \begin{cases} 1 & \text{if } |(t_i - t_j) - k\Delta\tau| < \Delta\tau/2 \\ 0 & \text{otherwise} \end{cases} \quad (2.14)$$

and calculate the slotted autocorrelation estimator as

$$\hat{r}_X[k\Delta\tau] = \frac{\sum_{i=1}^N \sum_{j=i+1}^N x_i \cdot x_j \cdot B_{k\Delta\tau}(t_i, t_j)}{\sum_{i=1}^N \sum_{j=i+1}^N B_{k\Delta\tau}(t_i, t_j)}. \quad (2.15)$$

Fig. 2.1 shows a diagram of the slotting technique to estimate the autocorrelation of an unevenly sampled time series. The slot size  $\Delta\tau$  corresponds to the sampling step of

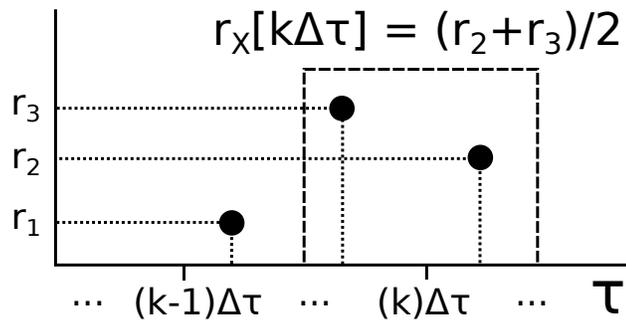


Figure 2.1: The slotted correlation is a correlation estimator for unevenly sampled time series. In this example there are two pairs of samples, with crossproducts  $r_2$  and  $r_3$ , whose time difference fall within slot  $k\Delta\tau$ . A rectangular slotting function is considered (Eq. 2.14). The slotted correlation at lag  $k\Delta\tau$  is then estimated as the average between the crossproducts falling in the slot. The crossproduct  $r_1$  contributes for the average of slot  $(k-1)\Delta\tau$ .

the slotted autocorrelation function. The slot size has to be set carefully. If the slot size is overestimated, important structures in the time series may be overlooked by the slotted autocorrelation function. If the slot size is underestimated empty slots may appear. If there are no samples that comply the slotting condition (Eq. 2.14) for a certain slot, the slotted autocorrelation function will be undefined at that slot.

The slotted autocorrelation function was independently proposed in [44] with the name of discrete correlation function. In [44] the slotting technique was used to estimate the discrete correlation function of light curves from periodic variable stars (Cepheids and RR Lyrae), where it proved to be a valuable alternative to the LS-periodogram.

The slotting technique determines an equidistant autocorrelation estimate of the data. A spectral density estimator (slotted periodogram) can be obtained by taking the Fourier transform of the windowed slotted correlation estimator. But, the slotted autocorrelation estimator has a major drawback: it does not comply the semi-positive definite property. This means that the slotted periodogram is not guaranteed to be positive at all frequencies, hence it cannot be interpreted as the distribution of power over frequencies.

In [41] two modifications to the original slotting technique were proposed: a) variable spectral window and b) local normalization scheme. These modifications could help in obtaining a slotted periodogram that is positive at all frequencies. The variable window technique consists of taking a wide window at low frequencies and a small one at high frequencies. The local normalization scheme takes into account the variance of the observations that contribute to the correlation at a particular slot. Only by using both modifications and selecting an appropriate window size a positive periodogram can be obtained. But, choosing an appropriate variable window requires *a priori* knowledge of the spectrum. So far, there is no slotted autocorrelation estimator yet that can guarantee to be semi-positive definite.

In what follows two more modifications to the original slotting technique are reviewed. These modifications do not tackle the semi-positive definite issue but they can help to improve the performance of the slotted correlation estimator.

The first modification is called fuzzy slotting. The fuzzy slot condition was proposed in [42] and is defined as

$$\hat{B}_{k\Delta\tau}(t_i, t_j) = \begin{cases} 1 - |(t_i - t_j) - k\Delta\tau|/\Delta\tau & \text{if } |(t_i - t_j) - k\Delta\tau| < \Delta\tau \\ 0 & \text{otherwise} \end{cases} \quad (2.16)$$

which corresponds to a triangular slot with a greater acceptance area than the original slotting condition (Eq. 2.14). A pair of samples might partially contribute to the slotted correlation at neighboring lags, as in fuzzy logic schemes.

The second modification was proposed in [43] and is called local time estimation. The time lags generated to compute the slotted autocorrelation estimator ( $k\Delta\tau$ ) are defined as the mid points of their corresponding slots. This translates to a higher weighting of the cross-products with larger time lags within a specific bin [43], hence increasing the bias of the slotted correlation estimator. To decrease the bias the author proposed to compute the time lags of the slotted correlation estimator as follows:

$$\hat{\tau}_k = \frac{\sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} (t_j - t_i) \cdot B_{k\Delta\tau}(t_i, t_j)}{\sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} B_{k\Delta\tau}(t_i, t_j)}. \quad (2.17)$$

The slotted correlation estimator is computed using the original algorithm. The difference is that the slotted correlation using local time estimation is a function of  $\hat{\tau}_k$  instead of  $k\Delta\tau$  (fixed original lags). Note that the slotted correlation estimator with local time estimation is unevenly sampled because the lags  $\hat{\tau}_k$  are not equally spaced.

### 2.2.2. Kernel windows for the slotting technique

In the original slotting technique [40, 44] a rectangular window is used to obtain the mean value of the sample pairs in time (Eq. 2.14). The slotted autocorrelation with rectangular window is known for its high variance, specially at high frequencies [45]. By replacing the rectangular window for a weighted average the undesirable cut-off effect of the slots is prevented, reducing the variance of the estimator. The weighted average of the observations can be performed using a symmetric kernel function that tends to zero for time differences that are larger than a certain threshold (slot size). The fuzzy triangular window [42], reviewed in the previous section, is an example of this. Other kernel functions used to estimate the slotted autocorrelation are the sinc kernel [46]

$$\hat{B}_{k\Delta\tau}(t_i, t_j) = \frac{1}{N} \frac{\sin(\pi\Delta\tau(k - |t_i - t_j|))}{\pi\Delta\tau(k - |t_i - t_j|)}, \quad (2.18)$$

and the Gaussian kernel [47]

$$\hat{B}_{k\Delta\tau}(t_i, t_j) = \frac{1}{2\pi\Delta\tau} \exp\left(-\frac{(k - |t_i - t_j|)^2}{2\Delta\tau^2}\right), \quad (2.19)$$

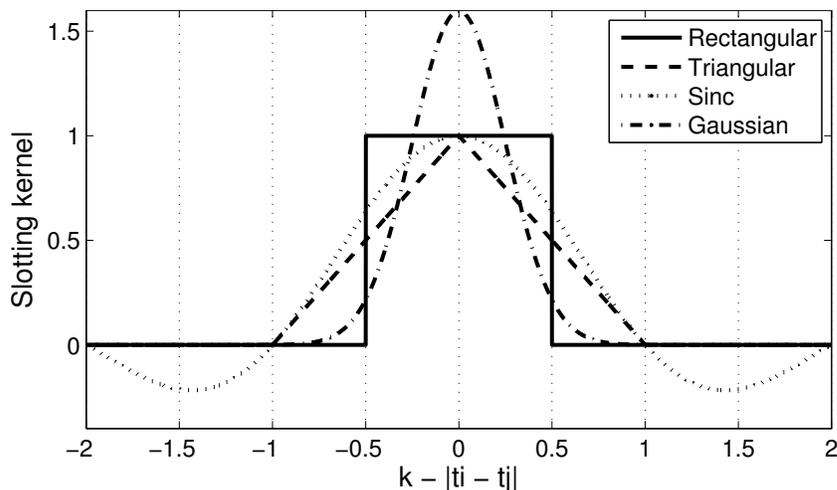


Figure 2.2: Slotting functions calculated using the kernel sizes in Table 2.1. The rectangular window gives the same weight to all the samples that fall in a given slot, which might introduce artifacts to the correlation estimator. The triangular, sinc and Gaussian window functions weight the samples according to how close they are to the corresponding lag.

Table 2.1: Standard choice of kernel bandwidth ( $\Delta\tau$ ) for the rectangular, sinc and Gaussian kernel functions.  $\Delta t$  corresponds to the mean sample rate of the time series.

Kernel	Definition	Standard choice for $\Delta\tau$
Rectangular	Eq. (2.14)	$\Delta t/2$
Sinc	Eq. (2.18)	$\Delta t$
Gaussian	Eq. (2.19)	$\Delta t/4$

where  $\Delta\tau$  is the kernel size. Table 2.1 shows the standard choice of kernel bandwidth for these kernels [48]. The slotting kernel functions are shown in Fig. 2.2. Smooth kernel functions effectively “use” observations whose inter-sampling time difference is close to the lag for which correlation is estimated. The rectangular kernel gives the same weight to all the observations falling on a given interval, instead the Gaussian and sinc kernel weight the products smoothly according to the difference between observation interval and desired lag. The results presented in [48] show that the Gaussian kernel performs consistently better than linear interpolation, slotting with rectangular window, slotting with sinc window and the Lomb-Scargle periodogram for autocorrelation estimation of unevenly sampled time series.

## 2.3. Statistical methods for unevenly sampled time series analysis based on epoch folding

### 2.3.1. Epoch Folding

Epoch folding is a widely used technique in astronomical time series analysis. The application of this technique can be summarized in three fundamental steps, for a time series

$\{t_i, x_i\}$  with  $i \in 1, \dots, N$

1. Propose a trial period  $P_t$ .
2. Transform the time instants  $t_i$  to phases  $\phi_i(P_t)$  using the folding transformation

$$\phi_i(P_t) = \frac{t_i \bmod P_t}{P_t}, \quad (2.20)$$

where mod is the modulo operation.

3. Re-order the magnitudes following the ascending order of phases  $\phi_i(P_t)$ .

The time space is mapped to a phase space defined by the trial period  $P_t$ . The pairs  $\{\phi_i(P_t), x_i\}$  are then sorted in ascending order according to their phases. The transformed and sorted time series  $\{\phi_k(P_t), x_k\}$  with  $k = 1, \dots, N$  is called folded time series. The epoch folding transformation is equivalent to partitioning the time series into consecutive segments of length  $P_t$  and then putting the samples of each of the segments on top of each other effectively folding the time series.

In practice, when the trial period is close to the underlying period (or an integer multiple of it) of the time series, the resulting folded light curve will look more ordered. In the other hand, if the trial period is different from the underlying period, the resulting folded light curve won't show a clear structure and will look like noise. Figure 2.3b shows a light curve folded using its underlying period and Fig. 2.3c illustrates the case of using an incorrect period for folding.

The ratio between the time length of the time series and the trial period gives the number of segments to be overlapped. Small errors on the estimation of the underlying period are propagated through when the epoch folding technique is applied. In general, the smaller the period the most precise estimation is needed in order to obtain an ordered folded curve.

A set of trial periods can be discriminated by analyzing the corresponding folded time series. The “quality” of the folded curve can be assessed visually or by using discrimination metrics. In this chapter several methods that use the epoch folding technique are presented.

Some comments about the epoch folding technique for unevenly sampled time series:

- It needs a trial period. The trial period can be obtained from another method, such as the LS periodogram.
- Contrary to Fourier based techniques, it does not pose assumptions on the light curve's shape.
- Usually the quality of the trial period is visually assessed. It requires an additional metric to be used in an automated method of period estimation.

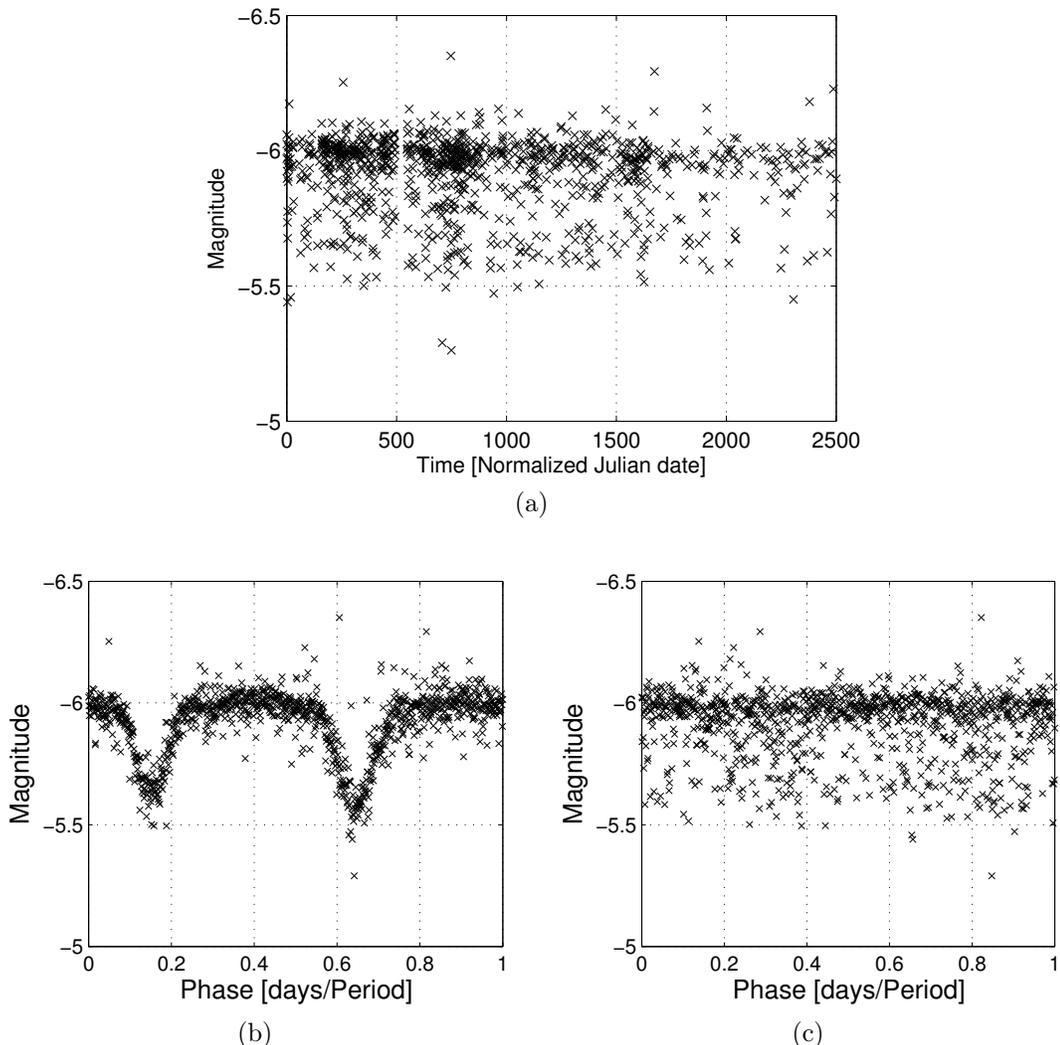


Figure 2.3: Light curve of an eclipsing binary star (a), folded with its underlying period (b), and folded with an incorrect period (c).

### 2.3.2. String Length

The epoch folding technique is quite useful for irregularly sampled time series that have non-sinusoidal variations. But, it lacks of a quantitative way to measure the quality of the folding period (it relies on visual inspection).

String length is a family of methods applying distance metrics to the folded time series. In what follows, the original string length implementation proposed in [49], is described. For a trial period  $P_t$  and a folded time series  $\{\phi_k(P_t), x_k\}$  with  $k = 1, \dots, N$  the total distance between adjacent samples is computed as

$$d_{Total}(P_c) = \sqrt{(\phi_1 - \phi_N)^2 + (x_1 - x_N)^2} + \sum_{k=2}^N \sqrt{(\phi_k - \phi_{k-1})^2 + (x_k - x_{k-1})^2}.$$

The string length criterion corresponds to the minimization of the total distance as a function

of the trial period  $P_c$ . Thus, an array of trial periods has to be tested when using the string length metric. Intuitively, the most ordered folded curve will achieve the minimum total distance, this should correspond to the one folded with the underlying period. Although this criterion has a clear and intuitive explanation it lacks of statistical sense.

An improved string length criterion was presented by [50]. This criterion is based in the minimization of the regularized string length Lafler-Kinman statistic (SLLK). For a trial period  $P_t$  and a folded time series  $\{\phi_k(P_t), x_k\}$  with  $k = 1, \dots, N$  the SLLK statistic is computed as

$$\Theta_{SLLK}(P_t) = \frac{N-1}{2N} \frac{\sum_{k=1}^N (x_{k+1} - x_k)^2}{\sum_{k=1}^N (x_k - \bar{x})^2}, \quad (2.21)$$

where  $\bar{x}$  is the mean value of the magnitudes of the time series and  $x_{N+1} = x_1$ . The SLLK statistic is a ratio between the squared differences of adjacent magnitudes values and the sample variance of the time series. In [50] it is mentioned that scaling by  $\frac{N-1}{N}$  removes the sample-size bias. Scaling by  $1/2$  normalizes the metric so that  $\Theta_{SLLK}(P_t) \approx 1$  when incorrect periods are used to fold the time series. The SLLK statistics depends only on the trial period and the amplitude-to-noise ratio of the measurements. In [50] it was stated that the development of a statistical test for period detection using the Eq. (2.21) as test statistic should be straight forward.

### 2.3.3. Analysis of variance

In [51] an statistical test based on analysis of variance (AoV) was formulated for discriminating among trial periods used to fold a light curve. In what follows this method is explained and commented.

First, a trial period is proposed and the light curve is folded using Eq. (2.20). Then, the phase space is partitioned into  $r$  equally spaced and disjoint segments. The objective of the AoV test is to prove the validity of the null hypothesis where all the mean values of the bins are statistically equals. To do this the following statistic is computed:

$$\Theta_{AoV} = \frac{s_1^2}{s_2^2}, \quad (2.22)$$

$$s_1^2 = \frac{1}{r-1} \sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2,$$

$$s_2^2 = \frac{1}{N-r} \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

where  $\bar{x}$  is the global mean of the time series,  $\bar{x}_i$  is the mean of segment  $i$ ,  $n_i$  is the number of elements that belong to segment  $i$  and  $N$  is the number of elements of the time series.

$s_1^2$  measures the variance between the bins while  $s_2^2$  measures the variance within bins. The statistics  $s_1^2$  and  $s_2^2$  follows a  $\chi_2$  distribution and are independent. This implies that  $\Theta_{AoV}$  is a Fisher-Snedecor statistic.

The statistical test consist of analyzing if  $\Theta_{AoV}$  is greater that the critical value for the  $F$  statistic with parameters  $N$  and  $r$ . If this is true, then the null hypothesis is rejected, *i.e.* at least one of the local means is statistically different. If all the local means are equals then the light curve was folded using an incorrect period. Not-satisfying the null hypothesis means that the trial period is a good guess for the underlying period.

In order to find the underlying period using AoV an array of trial periods has to be generated in advance. The  $\Theta_{AoV}$  statistic is computed for each trial period producing a periodogram. If the maximum value of the periodogram is greater than the critical value, then the null hypothesis is rejected. The period associated to the maximum  $\Theta_{AoV}$  is taken as the best guess for the underlying period.

Some comments about the AoV periodogram for unevenly sampled time series:

- It has a clear statistical meaning.
- Usually it outperforms other methods, such as the LS Periodogram, in detecting the underlying period of light curves with non-sinusoidal fluctuations [51].
- It requires sweeping an extent array of trial periods.
- It depends on the position and size of the bins.

### 2.3.4. Phase dispersion minimization

Phase dispersion minimization (PDM) [52] is a method that is similar to AoV. PDM has been used in the estimation of period from astronomical time series for eclipsing binary stars [53]. PDM requires a trial period to fold the light curve using Eq. (2.20). Then, the phase space is partitioned in  $m$  equally spaced and disjoint segments. The following statistics is computed

$$s_{PDM}^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{n_i - 1} \quad (2.23)$$

where  $\bar{x}_i$  is the mean value of segment  $i$  and  $n_i$  is the number of samples that belong to the segment  $i$ . The PDM statistic  $s_{PDM}^2$  is minimized in order to search for the partitioned folded light curve with minimum variance within bins, this should be the one folded with the underlying period.

The PDM statistic is quite similar to the variance within bins defined in AoV ( $s_2^2$ ). But, its distribution for the null hypothesis is not  $F$  nor  $\chi_2$ . In [51] it is argued that PDM should not do any better or have less computational cost than AoV.

### 2.3.5. Shannon's entropy minimization

Shannon's entropy of a random variable  $X$  is defined as

$$H_S = - \sum_{i=1}^N p(x_i) \log(p(x_i)), \quad (2.24)$$

where  $x_i$  are instances of  $X$  and  $p(x_i)$  is the probability mass function (PMF) of  $X$ . The concept of entropy is further discussed in Section 2.10.1.

In [54] a period discrimination method for unevenly sampled astronomical time series based on the minimization of Shannon's entropy was proposed. The Shannon's entropy was used to quantify the smoothness of the folded light curve. As explained by the authors, it is assumed that the smoothest folded light curve is obtained when the underlying period is used. For a time series  $\{t_i, x_i\}$  with  $i = 1, \dots, N$  and a trial period  $P_t$ :

1. Fold the light curve using Eq. (2.20) with trial period  $P_t$ .
2. Normalize the folded light curve to the unit square ( $\phi_i \in [0, 1]$  and  $x_i \in [0, 1] \forall i$ ).
3. Split the normalized folded light curve using a partition scheme  $\alpha = \{a_j, j = 1 \dots q\}$  where  $a_j$  are two-dimensional, disjoint and equally spaced bins and  $q$  is the number of bins.
4. Estimate the probability of occupation of bin  $a_j$  as

$$\mu_p(a_j) = \frac{1}{N} \int_{a_j} \sum_{i=1}^N \delta[(\phi, x) - (\phi_i, x_i)] d\phi dx,$$

where  $(\phi_i, x_i)$  are the samples of the normalized folded light curve and  $\delta[\cdot]$  is the Dirac delta function.

5. Compute Shannon's entropy as

$$H_P(\alpha) = - \sum_{j=1}^q \mu_p(a_j) \log(\mu_p(a_j))$$

which depends on the partition scheme  $\alpha$ .

The authors identified three possible outputs for the method. When the light curve is folded using an incorrect period, samples will be distributed randomly and uniformly in the unit square, and Shannon's entropy will be maximum. When the light curve is folded using its underlying period, Shannon's entropy reaches its global minimum, *i.e.* the folded light curve is at its most ordered state [54]. If a multiple of the underlying period is used to fold the light curve, the Shannon's entropy will reach a local minima.

Some comments about the Shannon's entropy minimization method for unevenly sampled time series:

- It requires a trial period.
- It is sensitive to the partition scheme applied on the folded light curve.
- It has problems discriminating spurious periods produced by the aliasing phenomenon [54].
- For some light curves, such as eclipsing binaries, an incorrectly folded light curve won't fill the unit square uniformly.

### 2.3.6. Conditional Entropy Minimization

Another example of information theoretic concepts used for periodicity detection in light curves can be found in [55]. In this work the Shannon's conditional entropy of a light curve is computed from a binned phase diagram obtained for a given period candidate. A partition of the phase diagram is obtained as described in the previous section. The Shannon's conditional entropy is defined as

$$H_c(x|\phi) = \sum_{i,j} p(x_i, \phi_j) \ln \left( \frac{p(\phi_j)}{p(x_i, \phi_j)} \right)$$

where  $p(x_i, \phi_j)$  is the occupation probability of the  $(i, j)$  bin and  $p(\phi_j) = \sum_i p(x_i, \phi_j)$  is the occupation probability of the  $j$  column in the magnitude-phase grid, respectively.

The conditional entropy is computed over a grid of trial periods, each one yielding a different phase diagram, obtaining a conditional entropy periodogram. The minimum value of the periodogram is selected as the underlying period of the light curve.

One of the weaknesses of the Shannon's entropy minimization method [54] is that it is sensitive to bad partitions caused by sampling artifacts, data gaps, inhomogeneous sample concentration and aliasing. According to [55], using the conditional entropy instead provides robustness against these cases.

The problem of period harmonics is also treated in [55]. In the case of eclipsing binary light curves, due to their particular shape, it is not uncommon for conventional methods to find subharmonics of the true period. In [55] smoothing splines are fitted to the light curve folded with the trial period and its harmonics in order to reveal the presence of the two dips caused by the eclipsing components. Once the eclipsing binary has been pre-classified with this procedure, the period is corrected using simple rules.

## 2.4. Statistical significance analysis for finding periodicities

If a time series is periodic with an oscillation frequency  $f$  then its periodogram  $P_{Per}(f)$  will exhibit a peak at that frequency with high probability. But the inverse is not necessarily true, a peak in the periodogram does not imply that the time series is periodic. Spectral peaks may be produced by measurement errors, random fluctuations, aliasing or noise.

Statistical tests have been used to prove if a peak in the power spectrum corresponds to a true period or if the time series is periodic at all [56, 57, 58]. A test of significance is a statistical test closely related to hypothesis testing. The general objective of the significance test is to prove if a realization of a certain experiment is statistically significant. The realization is said to be statistically significant depending on how inconsistent it is with respect to the null hypothesis. The null hypothesis is associated with a contradiction of what is to be proved.

An statistical significance test for spectral peaks may be used for

- **Spectral peak validation:** Spurious spectral components due to sampling or noise may be discarded by using an statistical criteria. An appropriate null hypothesis would be that the spectral peak is due to random noise.
- **Periodicity detection:** The highest peak in the spectrum is tested to assess if the time series is indeed periodic. The periodicity is confirmed with a certain confidence degree. An appropriate null hypothesis is to state that there are no periodic components in the time series.

### 2.4.1. Periodicity test using the maximum periodogram ordinate

In [56] Ronald Fisher presented an statistical significance test to assess periodicity in time series. The Fisher periodicity test is based on the maximum periodogram ordinate. The null hypothesis of the periodicity test is that the time series has no periodic components. The alternative hypothesis is that the time series contains at least one significant periodic component. The statistic of the periodicity test is the Fisher g-statistic. For a time series, its periodogram is computed using Eq. (2.5) over a range of discrete frequencies. The g-statistic is then computed as

$$g = \frac{\max_k \hat{P}_{XX}(f_k)}{\sum_k \hat{P}_{XX}(f_k)}, \quad (2.25)$$

where  $\hat{P}_{XX}(f_k)$  is the spectral value of the periodogram at discrete frequency  $f_k$ . Large values of  $g$  indicates a nonrandom periodicity. The p-values  $P(g \geq g^*)$  of a realization  $g^*$  of  $g$ , were calculated in [56] based on the distribution of the g-statistic under the null hypothesis and the Gaussian noise assumption. The null hypothesis is then rejected for a certain confidence level  $\alpha$  if  $P(g \geq g^*) < \alpha$ .

The Fisher test for periodicity detection have had several extensions [59]. Of particular interest is the extension of multi-periodic component assessment. In this test the null hypothesis of no periodic components is compared to the alternative hypothesis that there are  $r$  harmonic components in the time series. A set of statistics [59] are used to test the  $r$  highest spectral peaks of the periodogram.

## 2.5. Surrogates methods

Surrogate methods provide a way to estimate the unknown distribution of an statistic in hypothesis testing. Surrogate methods have been successfully used in the development of hypothesis tests for time series [60, 61]. The steps of a conventional hypothesis test using surrogates [61] are:

1. Formulate an appropriate null and alternative hypotheses.
2. Propose a test statistic in accordance to the null hypothesis, *i.e.* it has to react to the null hypothesis. It may reach a maximum or a minimum when the hypothesized process is strong.
3. Generate an appropriate Monte-Carlo sample for the null hypothesis. This step is known as surrogate generation. Surrogate algorithms generate an ensemble of data sets that are like the original data but complying with the null hypothesis. In general terms, a surrogate is what the original data is expected to be if it is consistent with the null hypothesis.
4. Estimate the distribution of the test statistic values in the surrogates and compare it with the value of the test statistic in the original data. The null hypothesis may be rejected, with a certain confidence degree  $(1 - \alpha) \cdot 100\%$ , if the original test statistic value is greater than (or lesser than) the value of the test statistic in  $M$  of the generated surrogates.

The amount of surrogates  $M$  that needs to be generated depends on the confidence degree. For a one-sided test at least  $M = \frac{1}{\alpha} - 1$  surrogates are required [61]. In a one-sided test the values that can reject the null hypothesis are located entirely in one tail of the probability distribution. For a two-sided test at least  $M = \frac{2}{\alpha} - 1$  surrogates are required [61]. In a two-sided test the values that can reject the null hypothesis are located in both tails of the probability distribution. Using a number of surrogates greater than  $M$  can increase the discrimination power of the test.

In what follows several surrogates generation algorithms and their application to unevenly sampled time series are reviewed. Refer to [61, 62, 63, 64, 65] for a more detailed description of the following methods.

- **Noise generator:** The surrogates are sampled at the same instants of the original time series. The magnitude values of the surrogates are produced using a white noise (or Gaussian noise) generator with the same mean and variance than the original time series. The time correlations and the probability density function (PDF) of the original time series are not preserved by the surrogates.
- **Shuffling:** The surrogates are generated by random shuffling of the magnitude values of the original time series. The time instants of the original time series are used. The length and the PDF of the time series are preserved by the surrogates. Shuffling assumes independent and identically (i.i.d) distributed random variables. The surrogates generated with shuffling lose the time correlations of the original time series.

- **Fourier based surrogates:** Basically, Fourier based surrogates are obtained by taking the Fourier transform (FT) of the original time series, randomizing the Fourier phases and applying the inverse FT. The Fourier magnitudes of the original time series are preserved by the surrogates, *i.e.* the statistical moments up to the second and the time correlations are preserved. This method cannot be applied directly to unevenly sampled time series because it utilizes the FT and IFT. In [61] alternatives for the unevenly sampled case are given, such as using the LS periodogram or the slotted correlation to generate constraints for the surrogates.
- **Block Bootstrap:** The bootstrap method for surrogate generation [64] preserves the time structure of the time series. Set a block length  $L$  and divide the original time series in  $M = \lfloor N/L \rfloor$  disjoint segments, where  $N$  is the length of the time series. The surrogates are generated by resampling  $M$  blocks with replacement, *i.e.* a particular block may be selected more than once. The time structure of the original time series is preserved within the blocks. The block length has to be carefully selected. In unevenly sampled time series selecting a fixed sample length will cause blocks to have different time length and vice versa.
- **Stationary Bootstrap:** The stationary bootstrap [65] is a modification of the original block bootstrap algorithm. In the stationary bootstrap the block length  $L$  is a random variable with a geometric distribution.

## 2.6. Kernel methods

In kernel methods [66] the input data, usually belonging to a low dimensional space  $\mathcal{X} \subseteq \mathbb{R}^N$ , is mapped to a Hilbert space<sup>4</sup>  $\mathcal{H}$  of greater dimensionality using a nonlinear transformation. Kernel methods solve nonlinear problems using linear algorithms and inner product evaluations in the feature space. Support Vector Machines (SVM) [67], Kernel Principal Component Analysis (KPCA) and Kernel Discriminant Analysis (KLD) are examples of kernel methods.

Kernel methods use a nonlinear transform  $\phi(\cdot) : \mathcal{X} \rightarrow \mathcal{H}$  to map the input data to feature space. However, the operations are usually directly done in input space using the “kernel trick” [67]

$$\kappa(x, z) = \langle \phi(x), \phi(z) \rangle_{\mathcal{H}},$$

where  $x \in \mathcal{X}$ ,  $z \in \mathcal{X}$ , and  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  is the inner product in feature space. Kernel is the general name of a continuous, bivariate function in  $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . A kernel function is a valid inner product in  $\mathcal{H}$  if it is symmetric and also positive definite. The nonlinear transform  $\phi(\cdot)$  is guaranteed to exist if the kernel function  $\kappa(\cdot, \cdot)$  complies the Mercer condition [68].

**Mercer condition:** There exists a nonlinear transform  $\phi(\cdot)$  in  $\mathcal{H}$  if and only if, for all square-integrable function  $g(x)$ ,

$$\int \int \kappa(x, y) g(x) g(y) dx dy \geq 0.$$

---

<sup>4</sup>A Hilbert is a generalized Euclidian space. A Hilbert space is any vector space, provided with an inner product, that is complete wrt the norm defined by that inner product.

which is true for any continuous, symmetric and positive-definite kernel function.

### 2.6.1. Translation-invariant kernel functions

Translation-invariant kernel functions are the most widely used class of reproducing kernels. A kernel  $\kappa(\cdot, \cdot)$  is said to be translation-invariant if

$$\kappa(x - a, z - a) = \kappa(x, z) \quad \forall x, z, a \in \mathcal{X}.$$

The radial basis functions (RBF) are an important class of translation-invariant kernels. The RBF kernel family is known for its good generalization and smooth interpolation capabilities. An RBF kernel is defined by a function  $\varphi : [0, \infty) \rightarrow \mathbb{R}$  such that

$$\kappa(x, z) = \varphi(\|x - z\|),$$

where  $x, z \in \mathcal{X}$  and  $\|\cdot\|$  is the Euclidian norm in  $\mathcal{X}$ . Examples of RBF are the Gaussian kernel, Laplacian kernel and the inverse multiquadratic kernel [69].

The Gaussian kernel function is defined as

$$G_\sigma(x - z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \quad (2.26)$$

where  $\sigma$  is the kernel bandwidth or kernel size. The kernel size has to be carefully estimated. If overestimated, the exponential function in the Gaussian kernel will behave almost linearly. In the other hand, if underestimated, the kernel function will be highly sensitive to noise in data. The selection of the kernel size is problem-driven, although the Silverman's rule of thumb [70] provides a good first estimate. For a random process  $X_n$  with  $n = 1, \dots, N$  the Silverman's kernel size is defined as [70]

$$\sigma_s = 1.06 \cdot \min\left[\hat{\sigma}, \frac{R}{1.34}\right] \cdot N^{-1/5} \quad (2.27)$$

where  $\hat{\sigma}$  is the standard deviation,  $R$  is the interquartile range<sup>5</sup> and  $N$  is the number of samples.

The Laplacian kernel is defined as

$$L_\sigma(x - z) = \exp\left(-\frac{\|x - z\|}{\sigma}\right) \quad (2.28)$$

where  $\sigma$  has the same interpretation than in the Gaussian kernel. Fig. 2.4 shows the Gaussian kernel and the Laplacian kernel as a function of  $\|x - z\|$ . The Laplacian kernel is peakier at  $x \approx z$  and has wider tails than the Gaussian kernel. The Laplacian kernel is less sensitive to changes in  $\sigma$ .

---

<sup>5</sup>The interquartile range is a measure of statistical dispersion that is defined as the difference between the third and first quartiles. If the data is arranged in ascending order, the quartiles are three points of the data set that divide the data in four groups, each one representing a 25% of the sampled population. The second quartile is the median.

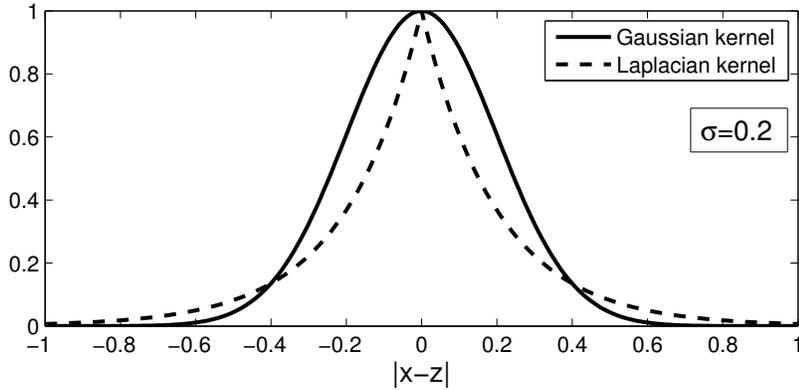


Figure 2.4: Gaussian kernel (solid line) and Laplacian kernel (dotted line) as a function of  $\|x - z\|$ . A kernel size  $\sigma = 0.2$  is used for both kernels.

### 2.6.2. Kernel construction

The selection of an appropriate kernel function is often a problem-driven issue. To propose a bivariate function as a valid kernel the properties mentioned in the previous section have to be fulfilled. The positive-definite requirement of a valid kernel function is often difficult to verify. The following properties [66, 71] can be used to device valid Mercer kernels. For two valid kernels  $\kappa_1 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\mathbb{N}}$  and  $\kappa_2 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^{\mathbb{N}}$ ,  $c \in \mathbb{R}$ ,  $f(\cdot)$  a real valued function on  $\mathcal{X}$ , and a transformation  $\phi : \mathcal{X} \rightarrow \mathbb{R}^{\mathbb{N}}$  with a valid kernel  $\kappa_3$  in  $\mathbb{R}^{\mathbb{N}} \times \mathbb{R}^{\mathbb{N}}$ , then the following are also valid kernel functions:

1.  $\kappa(x, y) = \kappa_1(x, y) + \kappa_2(x, y)$
2.  $\kappa(x, y) = \kappa_1(x, y) + c$
3.  $\kappa(x, y) = c \cdot \kappa_1(x, y) + \kappa_2(x, y)$
4.  $\kappa(x, y) = \kappa_1(x, y) \cdot \kappa_2(x, y)$
5.  $\kappa(x, y) = f(x)f(y)$
6.  $\kappa(x, y) = \kappa_3(\phi(x), \phi(y))$
7.  $\kappa(x, y) = \frac{\kappa_1(x, y)}{\sqrt{\kappa_1(x, x) \cdot \kappa_1(y, y)}}$

### 2.6.3. Periodic kernel functions

A kernel function is periodic with period  $P$  if it repeats itself for input vectors separated by  $P$ . Periodic kernel functions are appropriate for nonparametric estimation, modelling and regression of periodic time series [72]. Periodic kernel functions has also been proposed in the Gaussian processes literature [71, 73].

It is possible to build periodic kernel functions using the following property of translation-invariant kernels [66]. Given a translation-invariant kernel  $\kappa(x)$ , the kernel function

$$\kappa_T(x) = \sum_{n \in \mathbb{Z}} \kappa(x + nT), \quad (2.29)$$

is a periodic kernel function with period  $T$ .

A periodic kernel function can also be obtained by applying a nonlinear mapping (or warping)  $u(x)$  to the input vector  $x$ . In [73] a periodic kernel function was constructed by mapping a unidimensional input variable  $x$  using a periodic-twodimensional warping function defined as

$$u_P(x) = \left( \cos\left(\frac{2\pi}{P}x\right), \sin\left(\frac{2\pi}{P}x\right) \right). \quad (2.30)$$

The periodic kernel function  $G_{\sigma;P}(x_i - x_j)$  with period  $P$ , is obtained by applying the warping function (2.30) to the inputs of the Gaussian kernel function (2.26). The periodic kernel function is defined as

$$G_{\sigma;P}(x_i - x_j) = G_{\sigma}(u_P(x_i) - u_P(x_j)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{2 \sin^2\left(\frac{\pi}{P}(x_i - x_j)\right)}{\sigma}\right), \quad (2.31)$$

where the following was used

$$\|u_P(x_i) - u_P(x_j)\|^2 = 4 \sin^2\left(\frac{\pi}{P}(x_i - x_j)\right).$$

The periodic kernel is a valid Mercer kernel function (property 6 of Mercer kernel construction). Note that the periodic kernel is a function of time  $\delta t = (t_z - t_y)$  and frequency (the inverse of the period). The Taylor series expansion at  $\delta t = 0$  of Eq. (2.31) is defined as

$$G_{\sigma}^P(f, \delta t) = \lim_{N \rightarrow \infty} \sum_{k=0}^N \frac{(-1)^k}{k! \sigma_t^{2k} 2^{k-1}} \left[ \sum_{m=0}^k \binom{2k}{k-m} (-1)^m g_m \cos(2\pi m f \delta t) \right], \quad (2.32)$$

where

$$g_m = \begin{cases} 1/2, & \text{if } m = 0. \\ 1, & \text{otherwise.} \end{cases} \quad (2.33)$$

Note that for large values of  $\sigma$ , only the first terms contribute to the sum and thus the periodic kernel tends to a constant plus  $\cos(2\pi f \delta t)$ , which corresponds to the real part of the Fourier basis.

## 2.7. Spectral estimation using Basis Pursuit

Basis pursuit (BP) [74, 75] is a technique to decompose a signal or time series  $\{t_i, x_i\}_{i=1, \dots, N}$  into a set of elementary functions  $\phi_k(t_i)$  where  $i \in [1, N]$ . The  $k$ -indexed functions are called atoms and the set of atomic functions is called Dictionary. The signal is decomposed as

$$x_i = \sum_{k \in \mathcal{D}} \alpha_k \phi_k(t_i), \quad (2.34)$$

where  $\{\alpha_k\}_{k \in \mathcal{D}}$  is a set of coefficients. The goal of the BP algorithm is to find a set of coefficients such that

$$\min \sum_{k \in \mathcal{D}} |\alpha_k| \quad \text{subject to } x_i = \sum_{k \in \mathcal{D}} \phi_k(t_i) \alpha_k \quad \forall i \in [1, N]. \quad (2.35)$$

If we define  $\alpha$  the coefficient vector,  $t$  the time vector,  $x$  the signal vector, and  $\Phi$  the Dictionary matrix we can rewrite Eq. (2.35) in matricial form as follows

$$\min \|\alpha\|_1 \quad \text{subject to } x(t) = \Phi(t)\alpha. \quad (2.36)$$

The basis pursuit representation is optimal in the sense of having the smallest  $l_1$  norm of the coefficient vector. This forces the representation to be sparse, *i.e.* the signal will be decomposed using as few atoms as possible, enhancing the interpretability of the representation.

The characteristics of the dictionary define the representation. In [76] a pure-frequency Fourier dictionary was used to perform spectral estimation. The even and odd Fourier dictionaries are defined as

$$\Phi_E(f_k) = \cos(2\pi t f_k) \quad \text{and} \quad \Phi_O(f_k) = \sin(2\pi t f_k), \quad (2.37)$$

respectively. The frequency grid dictates the number of atoms in the dictionaries and is defined as

$$f_k = \frac{k}{T}, \quad k = 0, \dots, \frac{N}{2}. \quad (2.38)$$

The frequency dictionary can be used to define the following BP frequency decomposition

$$S(f_k) = |\alpha_k^E| + |\alpha_k^O|, \quad (2.39)$$

where  $\alpha_k^E$  and  $\alpha_k^O$  are the  $k$ -th coefficients associated to the cosine (even) and sine (odd) dictionaries, respectively. The Fourier dictionary size is  $N$ . The atoms of the basis are all mutually orthogonal.

The dictionary can be extended to the time-frequency domain using Wavelets, Gaborlets, cosine packets, and/or Chirplets atomic functions [76]. The user has the flexibility of choosing a given dictionary or merging different dictionaries depending on the phenomena that needs to be characterized.

### 2.7.1. Overcomplete dictionaries

Traditional methods of analysis and reconstruction, such as the Fourier transform, decompose a signal of length  $N$  into a basis composed of  $N$  atoms. If the dictionary forms a basis of linearly independent atoms then the representation will be unique. In the basis pursuit methodology one may use overcomplete dictionaries that do not necessarily form a basis [75, 76]. Overcomplete dictionaries can be constructed by increasing the number of atoms of a basis. In the case of the Fourier dictionary (or other frequency-indexed dictionaries), this can be achieved by decreasing the frequency spacing between atoms. To do this the frequency parameter of the dictionary (Eq. 2.40) is redefined as

$$f_k = \frac{k}{TL}, \quad k = 0, \dots, \frac{LN}{2}, \quad (2.40)$$

where  $N$  and  $T$  are the number of samples and time span of the input signal, and  $L$  is the overcompleteness parameter. The parameter  $L$  controls the number of atoms in the pure-frequency dictionary. For  $L = 1$  the frequency dictionary is complete. By increasing  $L$ , *i.e.* sampling the frequencies more finely, an overcomplete dictionary is obtained. By using overcomplete dictionaries, sparse and high-resolution spectral representations can be obtained. It is important to distinguish between frequency spacing and frequency resolution. The former refers to the separation between frequency components in the grid defined by Eq. (2.40). The later refers to the width of the peaks appearing in the spectrum. For example, when the discrete Fourier transform is used, even if the frequencies are finely spaced, the width of the peaks will be proportional to  $1/T$ , where  $T$  is the length of the time series.

Overcomplete dictionaries can also be obtained by merging complete dictionaries [76]. Note that when overcomplete dictionaries are used, an underdetermined system of equations will be obtained and the solution will be non-unique. BP finds the optimal solution in the sense of the  $l_1$  of the coefficients. The advantages of BP using overcomplete dictionaries are better sparsity and super-resolution [76].

## 2.7.2. L1 norm minimization for BP

In this subsection guidelines to solve the the basis pursuit optimization problem are presented. Eq. (2.36) can be solved with an interior-point linear programming algorithm by adding auxiliary variables [75] leading to the following formulation

$$\min \mathbb{1}^T \begin{pmatrix} u \\ v \end{pmatrix} \text{ subject to } s(t) = (\Phi(t), -\Phi(t)) \begin{pmatrix} u \\ v \end{pmatrix}, \quad (2.41)$$

where the coefficients of the original formulation can be obtained as  $\alpha = u - v$ . Another solution is to make a change of variables  $z_k = |\alpha_k|$  in order to transfer the non-linearities of the  $l_1$  norm to a set of additional constraints. The new problem is stated as follows

$$\min \mathbb{1}^T z \text{ subject to } s(t) = \Phi(t)\alpha \text{ and } \alpha_k \geq -z_k \forall k \text{ and } \alpha_k \leq z_k \forall k, \quad (2.42)$$

the new constraints can be written in matricial form as

$$\min \mathbb{1}^T z \text{ subject to } s(t) = \Phi(t)\alpha \text{ and } I\alpha + Iz \geq 0 \text{ and } I\alpha - Iz \leq 0, \quad (2.43)$$

where  $I$  is the identity matrix. Note that this introduces a large set of constraints. The three macro-constraints can be transformed into three variables by using the dual representation [77] of the optimization problem,

$$\max s^T a \text{ subject to } \Phi(t)^T a - 2b = -\mathbb{1} \text{ and } \mathbb{1} \geq b \geq 0, \quad (2.44)$$

where the new variables are  $a$ ,  $b$  and  $c$ , one for each of the constraints in Eq (2.43). While solving the dual one gets that  $c$  can be expressed in terms of  $a$  and  $b$ . Because of this,  $c$  is omitted in the final expression given by Eq. (2.44).

MATLAB “linprog” function, which is based on LIPSOL [78], a primal-dual interior-point method [79], can be used to solve the optimization problem. Interior-point methods search

for the solution inside the feasible region rather than at the edge of the region (simplex). Primal-dual methods take steps in both the primal and the dual at the same time using a variant of Newton’s method for solving nonlinear equations (equality constraints) and barrier methods (inequality constraints).

### 2.7.3. Basis pursuit de-noising

In [75] BP was extended to the case of noisy signals of the form

$$x = y + \sigma z,$$

where  $z$  is white Gaussian noise with amplitude  $\sigma > 0$ , and  $y$  is the clean signal. In this setting,  $x$  is known and  $y$  is unknown. To solve this problem the authors propose to change the objective function given in Eq. (2.36) as follows

$$\min_{\alpha} \lambda \|\alpha\|_1 + \frac{1}{2} \|x(t) - \Phi(t)\alpha\|_2^2, \quad (2.45)$$

which has no constraints. In this formulation the  $l_2$  difference between the signal and its representation is minimized where in the original formulation the terms are forced to be equal through the constraint. The optimal solution  $\alpha^\lambda$  depends on  $\lambda$  which is the regularization parameter. Assuming that the dictionary is normalized ( $\|\phi_k(\cdot)\|_2 = 1 \forall k$ ), the following rule of thumb [75] may be used to set the regularization parameter

$$\lambda = \sigma \sqrt{2 \log(p)},$$

where  $p$  is the cardinality of the dictionary (number of atoms). Note that estimating the optimal parameters of the decomposition in a least square sense subject to a  $l_1$  penalty was presented independently under the name Least Absolute Shrinkage and Selection Operator (LASSO) [80] in the statistics community.

## 2.8. Non-negative matrix factorization

Non-negative matrix factorization (NMF) [81, 82, 83] is a technique to learn localized representations for data in an unsupervised way. In NMF, a data matrix  $V \in \mathbb{R}^{M \times P}$  is decomposed as follows

$$V \approx \widehat{V} = WH, \quad (2.46)$$

where  $W \in \mathbb{R}^{M \times K}$  is the dictionary matrix,  $H \in \mathbb{R}^{K \times P}$  is the coefficient matrix and  $\widehat{V} \in \mathbb{R}^{M \times P}$  is the NMF reconstruction of the data matrix. For dimensionality reduction purposes, the number of columns of the dictionary  $K$ , is usually set smaller than both  $M$  and  $P$ . The decomposition shown in Eq. (2.46) is obtained by solving

$$\min_{W, H} L(V || WH) \text{ s.t. } W \geq 0, H \geq 0, \quad (2.47)$$

where  $L(\cdot||\cdot)$  is a cost function or distance metric for matrices and the inequality constraints are component-wise. The Frobenius norm, Information divergence and Kullback-Leibler divergence [82, 84, 85] are among the most commonly used cost functions in the NMF literature. In what follows we focus on the Frobenius norm because it is the simplest and the most inexpensive to compute

$$\begin{aligned} D_F(V||WH) &= \frac{1}{2} \|V - WH\|_F^2 \\ &= \frac{1}{2} \sum_{ij} (V_{ij} - [WH]_{ij})^2. \end{aligned} \quad (2.48)$$

NMF can be compared with other decomposition techniques such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Vector Quantization (VQ). The key difference of NMF is the imposed non-negativity constraint, which forces the components to be positive but most importantly it forces the combinations between dictionary elements to be purely additive. This enhances the physical interpretability of the NMF decomposition with respect to other decomposition techniques. The non-negativity constraint is also responsible for the inherent sparseness that is usually found in the NMF representations. Note that the sparseness is not actively enforced in Eq. (2.47) and thus cannot be guaranteed or controlled. The NMF with sparseness constraint [86] was developed with such constraint in mind. The non-negativity makes the NMF particularly appealing and highly compatible in spectral analysis because of the positive nature of spectral decompositions for real-time time series.

The problem stated in Eq. (2.47) is non-convex with respect to  $W$  and  $H$ , may have several local minima, and it has been proven to be NP-hard [87]. A popular approach to solve the NMF is the alternating non-negative least squares (ANLS) [88, 89] in which Eq. (2.47) is divided into two NLS sub-problems which are easier to handle

### I Dictionary learning:

$$\min_{W \geq 0} \frac{1}{2} \|H^T W^T - V^T\|_F^2 \quad \text{with fixed } H. \quad (2.49)$$

### II Sparse coding:

$$\min_{H \geq 0} \frac{1}{2} \|V - WH\|_F^2 \quad \text{with fixed } W. \quad (2.50)$$

Note that in step I the coefficients are fixed, while in step II the dictionary is fixed. The sub-problems are convex, thus optimal solutions can be found for each of them. In most methods steps I and II are solved iteratively until a stopping criterion is met. It is important to note that there is no guarantee that a global minimum of Eq. (2.47) can be reached by ANLS.

Commonly, the NLS sub-problems have been solved using either gradient descent or multiplicative based approaches. The multiplicative rules [82, 84, 85, 90] were proposed as an alternative to gradient descent methods (which are in contrast additive rules). The multiplicative rules for the Frobenius norm cost function are defined as

$$H_{ij} \leftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T W H)_{ij}}, \quad (2.51)$$

for the coefficient matrix and

$$W_{ij} \leftarrow W_{ij} \frac{(V H^T)_{ij}}{(W H^T H)_{ij}}, \quad (2.52)$$

for the dictionary matrix. Note that Eq. (2.51) is equivalent to

$$H_{ij}^{t+1} = H_{ij}^t - \eta_{ij} \nabla_H D_F(V || W H), \quad (2.53)$$

where the step size correspond to

$$\eta_{ij} = \frac{H_{ij}^t}{(W^T W H)_{ij}^t}, \quad (2.54)$$

and the same can be shown for Eq. (2.52), hence the multiplicative rules are a special case of gradient descent with an adaptive step size. In [90] it was shown that with minor corrections any limit point given by the multiplicative rules is a stationary point of Eq. (2.47).

Other successful methods to solve the NLS sub-problems are the interior-point gradient [91], projected gradient [88] and active-set like method [89]. These methods are more efficient and converge faster than the multiplicative rules under the assumption that  $K < \min(M, P)$ .

## 2.9. Higher order moments

Random variables can only be completely described by their probability density functions (PDFs). Statistical moments are descriptors of the shape of the PDF of the random variable. The k-th central moment of a real-valued random variable  $X$  is defined as

$$\mu_k = \mathbb{E} [(X - \mu)^k],$$

where  $\mathbb{E}[\cdot]$  is the mathematical expectation operator and  $\mu = \mathbb{E}[X]$  is the first order moment of  $X$ . The first order moment is called the mean and is a measure of the central tendency of the random variable. The mean describes the geometric center of the PDF. The second central moment is defined as  $\sigma^2 = \mathbb{E} [(X - \mu)^2]$ , and is called the variance. The variance is a measure of the dispersion of the random variable. The autocorrelation and the power spectrum are second-order statistics of random processes.

If a random variable is Gaussian (normal) distributed then it can be completely described through its first and second order moments. But, real-world data is often not Gaussian-distributed, hence higher-order moments are needed to fully characterize the statistical distribution.

Examples of higher-order statistics are the Skewness, Kurtosis, cumulants and the polyspectra [92, 93]. The skewness is the third central moment and is a measure of the symmetry of the PDF of the random variable. The Kurtosis is the fourth central moment and describes

flatness of the PDF. The polyspectra are higher order extensions of the power spectrum and provide supplementary information to it. The bispectrum is the third-order polyspectrum and is related to the process's skewness. The trispectrum is the fourth-order polyspectrum and is related to the process' Kurtosis. The importance of the higher order moments for the analysis of non-Gaussian random processes was discussed as an introductory concept for the following section.

## 2.10. Information Theoretic Learning

Information Theory combines the electrical engineering and applied mathematics disciplines in order to develop expressions capable of measuring the information contained in random variables and processes. Some of these measurements are the entropy, which measures the uncertainty of a random variable; and the mutual information, which measures the amount of information that one random variable has of another random variable. For a certain message (random variable) whose content is completely known *a priori* (zero uncertainty), no interesting information would be found, and hence its entropy would be minimum. Information Theory has impacted successfully in the areas of statistics and telecommunications. Particularly in telecommunications, IT metrics has been used to develop optimum codification schemes and optimize the amount of information transmitted in a limited channel (bandwidth estimation).

In this thesis, methods for astronomical time series analysis based on information theoretic criteria are proposed. The Information Theoretic Learning (ITL) [31] framework was developed by investigators of the computational neuro-engineering laboratory of the University of Florida. The major objective of the ITL research is the development of higher-order statistical descriptors to substitute conventional second-order statistical descriptors such as variance and correlation. In what follows the main concepts and some applications of the ITL research are briefly described.

Training learning machines, such as neural networks, is a process in which the machine integrates in its parameters all the information that is extracted from the data that represents the problem to be solved. Second-order statistics, such as the mean square error (MSE) and correlation, are typically used to extract information from the input data. But, second-order functionals are limited, because they assume that the data complies with the properties of linearity and gaussianity, which is not true in the majority of practical cases. The premise of ITL is that there exists information in the data that cannot be captured by second-order statistics. To obtain this information, the ITL framework uses information theoretic criteria that convey information from the whole probability density function (PDF) of the data, *i.e.* all the higher-order moments are included. Usually using these higher-order criteria improve the performance of the trained learning machines.

In the ITL framework, several criteria to train learning machines, such as MaxEnt and MinXEnt, have been proposed. These criteria adjust the parameters of the learning machine by means of the optimization of a certain information theoretic functional. For example, in MaxEnt, the entropy of the output of the learning machine is maximized. This is equivalent

to search for the output PDF that conveys the maximum amount of information possible. In MinXEnt the relative entropy (divergence) between different outputs (or between the output and other signals) of the learning machines is minimized.

These criteria allow us to obtain much more information than those based on second order statistics. The reason behind this is that they operate using the whole PDF of data. But estimating the PDF is not trivial. This is often solved parametrically by assuming a certain statistical distribution and then calculating its corresponding parameters. Yet, assuming an *a priori* PDF for the data is a restriction that could gravely affect the quality of the results obtained. The non-parametric estimation of the PDF directly from data is one of the advantages of the ITL framework.

### 2.10.1. An entropy measure directly estimated from data

In the information theory framework, entropy is the name given to a family of functionals capable of quantifying the uncertainty of a system. Uncertainty is a concept that is closely related to information. Information have been viewed as the capability of reducing uncertainty and hence these concepts are inversely proportional.

The most well known information-entropy measure is the Shannon's entropy which is defined as

$$H(X) = - \sum_k p(x_k) \log(p(x_k)) = -E[\log_2 p(X)], \quad (2.55)$$

where  $X$  is a discrete random variable and  $p(x_k)$  is its probability mass function (PMF). Shannon's entropy is a function of the PMF of  $X$ . Therefore in order to compute entropy the PMF has to be estimated.

In the ITL framework [31], the estimation of entropy directly from data has been done using the Parzen window estimator [94]. The Parzen window provides a way to do a non-parametric estimation of the PDF (or PMF) of a random variable. For a random variable  $X$ , its estimated PDF, using the Parzen window method, is

$$\hat{p}_X(x) = \frac{1}{N \cdot h} \sum_{i=1}^N \kappa \left( \frac{x - x_i}{h} \right), \quad (2.56)$$

where  $\kappa(\cdot)$  is a normalized, symmetric, positive definite kernel function and  $h$  is the smoothing parameter or kernel bandwidth. If the Gaussian kernel function (2.26) is used in (2.56) the following expression is obtained

$$\hat{p}_X(x) = \frac{1}{N} \sum_{i=1}^N G_\sigma(x - x_i). \quad (2.57)$$

As stated in [31], Shannon's entropy is not appropriate for applying the Parzen window method. This is because the Shannon's entropy is a sum of pondered logarithm functions. In order to obtain an entropy expression that can be directly estimated from data, the

generalized Renyi's entropy is used [31]. The family of generalized Renyi's entropies is defined as

$$H_{R\alpha}(X) = \frac{1}{1-\alpha} \log \left( \int p^\alpha(x) dx \right), \quad (2.58)$$

where  $X$  is a continuous random variable,  $p(x)$  is its PDF and  $\alpha$  is the order of Renyi's entropy. It can be proved, using the L'Hôpital limit theorem, that Renyi's entropy tends to Shannon's entropy for  $\alpha = 1$ . Renyi's quadratic entropy (RQE) corresponds to a particular case of (2.58) when  $\alpha = 2$ . If Eq. (2.57) is replaced as the PDF in Eq. (2.58) the following expression is obtained [31]

$$\begin{aligned} H_{R2}(X) &= -\log \left( \int_{-\infty}^{+\infty} p^2(x) dx \right) \\ &= -\log \left( \frac{1}{N^2} \int_{-\infty}^{+\infty} \sum_{i=1}^N \sum_{j=1}^N G_\sigma(x-x_i) \cdot G_\sigma(x-x_j) dx \right). \end{aligned} \quad (2.59)$$

Because the term inside the logarithm is squared, it is possible to use the convolution property of Gaussian functions. The convolution of two Gaussian functions gives result to another Gaussian function centered in the mid point between the centers of the original Gaussian functions, *i.e.*

$$\int_{-\infty}^{+\infty} G_{\sigma_1}(z-x_i) \cdot G_{\sigma_2}(z-x_j) dz = G_{\sqrt{\sigma_1^2+\sigma_2^2}}(x_i-x_j). \quad (2.60)$$

Then, by applying this property, Expression (2.59) can be further developed as follows:

$$\begin{aligned} H_{R2}(X) &= -\log \left( \frac{1}{N^2} \int_{-\infty}^{+\infty} \sum_{i=1}^N \sum_{j=1}^N G_\sigma(x-x_i) \cdot G_\sigma(x-x_j) dx \right) \\ &= -\log \left( \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \int_{-\infty}^{+\infty} G_\sigma(x-x_i) \cdot G_\sigma(x-x_j) dx \right) \\ &= -\log \left( \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G_{\sqrt{2}\sigma}(x_i-x_j) \right). \end{aligned} \quad (2.61)$$

The resulting expression of the Renyi's quadratic entropy estimator [31] for a random variable  $X$  is

$$\hat{H}_{R2}(X) = -\log \left( \hat{V}(X) \right), \quad (2.62)$$

where

$$\hat{V}(X) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|x_i-x_j\|^2}{2\sigma^2} \right), \quad (2.63)$$

and  $N$  is the number of samples of  $X$  and  $\sigma$  is the kernel size of the Gaussian kernel function. Expression (2.62) is an entropy estimator which can be computed directly from the samples. Equation (2.63) is the argument of the logarithm in Eq. (2.62) and is called the quadratic Information Potential (IP). The quadratic IP estimator is a function of the kernel size  $\sigma$  (Gaussian kernel).

## 2.10.2. Generalized correlation function: Correntropy

In [31, 95] an information theoretic functional capable of measuring the statistical magnitude distribution and the time structure of random processes was presented. This generalized correlation function (GCF) measures similarities between feature vectors separated by a certain time delay in input space. The similarities are measured in terms of inner products in a high-dimensional kernel space. The GCF conveys information of the PDF of the random process via the RQE. Because of this the GCF was given the name of **correntropy** and **autocorrentropy** for its univariate version.

For a random process  $\{X_t, t \in T\}$  with  $T$  being an index set, the correntropy function is defined as

$$V(t_1, t_2) = \mathbb{E}[\kappa(x_{t_1}, x_{t_2})], \quad (2.64)$$

and the centered correntropy is defined as

$$U(t_1, t_2) = \mathbb{E}_{x_{t_1} x_{t_2}}[\kappa(x_{t_1}, x_{t_2})] - \mathbb{E}_{x_{t_1}} \mathbb{E}_{x_{t_2}}[\kappa(x_{t_1}, x_{t_2})], \quad (2.65)$$

where  $\mathbb{E}[\cdot]$  denotes the expected value and  $\kappa(\cdot, \cdot)$  is any positive definite kernel. In [31, 95], the properties of the correntropy function are extensively explored for the case of translation invariant kernel such as the Gaussian kernel function. Using the Taylor series of the Gaussian kernel, expression (2.64) can be expanded as

$$V(t_1, t_2) = \mathbb{E}[G_\sigma(x_{t_1} - x_{t_2})] = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n \sigma^{2n} n!} \mathbb{E}[\|x_{t_1} - x_{t_2}\|^{2n}], \quad (2.66)$$

where  $\sigma$  is the Gaussian kernel size. By analyzing Eq. (2.66) it is possible to note that the correntropy function

- includes all the even-order statistics of the random variable  $(x_{t_1} - x_{t_2})$
- includes the information of conventional correlation (for  $n = 1$ )

When the Gaussian kernel is used, the kernel size  $\sigma$  controls the emphasis given to higher-order moments with respect to second-order moments. In [31, 95] the kernel size was interpreted as the resolution in which the correntropy function search for similarities in the high-dimensional kernel feature space. For large values of the kernel size, the second-order moments have more relevance and the correntropy function approximates the conventional correlation. On the other hand if the kernel size is set too small, the correntropy function will not be able to discriminate between signal and noise.

Correntropy quantifies similarity using the correntropy induced metric (CIM) defined as

$$CIM(x, y) = (\kappa(0, 0) - \mathbb{E}[\kappa(x, y)])^{1/2}. \quad (2.67)$$

The CIM is a metric very different from the  $L_p$  norms that define the Minkowski spaces where the distances are always weighted the same (Fig 2.5)<sup>6</sup>. This means that distances

---

<sup>6</sup>For  $\vec{x} \in \mathbb{R}^n$ , the  $L_p$  norms are defined as  $L_p = \|\vec{x}\|_p = \left(\sum_{i=1}^N x_i^p\right)^{\frac{1}{p}}$ ,  $p \in (0, \infty)$ . In the limit  $p \rightarrow 0$ , the  $L_0$  norm is defined as the number of non-zero components in the vector (counting norm).

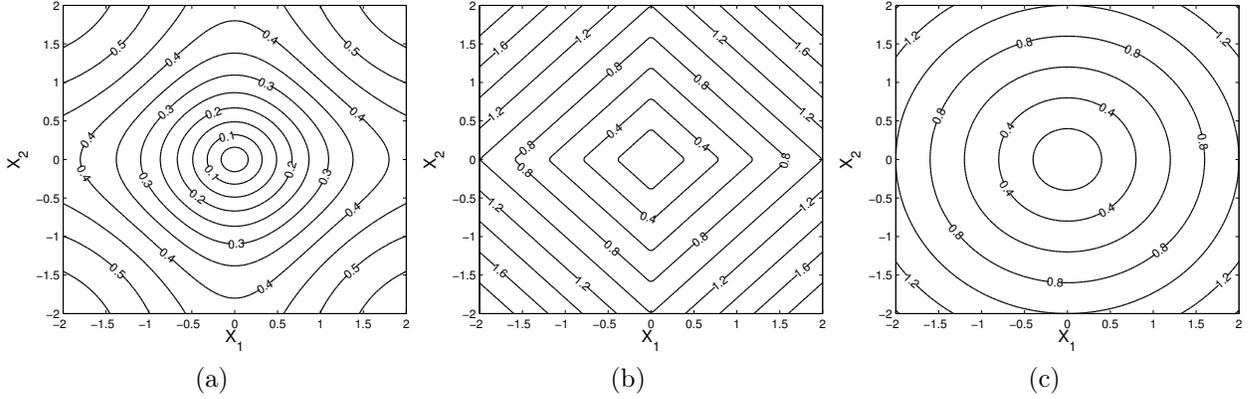


Figure 2.5: Distances to the origin (contours) in a bidimensional sample space using the CIM( $X,0$ ) (a),  $L_1$  norm (b) and  $L_2$  norm (c). For the CIM (Eq. 2.67) a Gaussian kernel function with  $\sigma = 1$  is considered. Note how the CIM incorporates the  $L_1$ ,  $L_2$  and  $L_0$  norms at different scales.

between the arguments of the CIM are weighted nonuniformly, i.e. if the distance between the arguments is small then the CIM approximates the  $L_2$  norm, but if the difference is larger then it will approximate the  $L_1$  norm, and for very large difference between the arguments, the CIM tends to the  $L_0$  norm. The transitions between the norms are smooth, and the assessment of ‘small’ and ‘large’, the scale in this space is controlled by the kernel size, which impacts drastically on the assessment of similarity.

For a discrete strictly stationary random process  $\{X_n\}$ , the univariate correntropy function or autocorrentropy can be defined as

$$V[m] = \mathbb{E}[\kappa(x_n, x_{n-m})],$$

which can be estimated through the sample mean

$$\begin{aligned} \hat{V}[m] &= \frac{1}{N-m+1} \sum_{n=m}^N G_\sigma(x_n - x_{n-m}) \\ &= \frac{1}{N-m+1} \frac{1}{\sqrt{2\pi}\sigma} \sum_{n=m}^N \exp\left(-\frac{\|x_n - x_{n-m}\|^2}{2\sigma^2}\right), \end{aligned} \quad (2.68)$$

likewise, the estimator of the univariate centered correntropy function (Eq. 2.65) is

$$\hat{U}[m] = \frac{1}{N-m+1} \sum_{n=m}^N G_\sigma(x_n - x_{n-m}) - \frac{1}{N^2} \sum_{n=1}^N \sum_{m=1}^N G_\sigma(x_n - x_m), \quad (2.69)$$

where the translation-invariant Gaussian kernel with kernel size  $\sigma$  is used,  $N$  is the number of samples of  $\{X_n\}$  and the discrete lag  $m \in [1, N]$ . In practice, the maximum lag should be chosen so that there are enough samples to estimate correntropy at each lag. The second term in Eq. (2.69) corresponds to the estimator of the quadratic Information potential (2.63).

In [31, 95] the correntropy function was proved to be symmetric, positive definite and always maximum at the origin (zero lag). The close relationship of the correntropy function

and the RQE (2.62) was also proven. The mean value of the correntropy function over the lags is a biased estimator of the Information Potential (2.63) (the argument of the logarithm in the RQE).

As shown in [31, 95, 96], the Fourier transform of the centered autocorrentropy function is called correntropy spectral density (CSD) and is defined as

$$P_\sigma[f] = \sum_{m=-\infty}^{\infty} U[m] \cdot \exp\left(-j2\pi f \frac{m}{F_s}\right) \quad (2.70)$$

where  $U[m]$  is the univariate centered correntropy function. The CSD can be considered as a generalized PSD function, although it is a function of the kernel size and it does not measure power. As with correntropy, the kernel size controls the influence of the higher-order moments versus the second-order statistical descriptors. Particularly, for large values of the kernel size, the CSD approximates the conventional PSD.

It is appropriate to present a synthetic example to illustrate the difference between autocorrelation and autocorrentropy in assessing similarity over time, and also to elucidate the role of the kernel size. Let us take the case of the stochastic process with uniform random amplitude in  $[-A, A]$  and a random phase in  $[-\pi, \pi]$  defined as  $x_n = A \sin(w_0 n + \varphi)$ . As it is well known, the autocorrelation function of a sinewave is a sinewave of the same period. But should it be a sinewave if we are interested in assessing the degree of similarity of the signal time structure? Since the sinewave is periodic, the similarity is maximum when the delay is exactly one period, but for intermediate shifts, the two functions are very dissimilar, and autocorrelation does not show this very clearly (and the similarity is not normalized nor always positive, hence the use of the correlation coefficient). Therefore, if we are seeking a discriminative measure of similarity, the autocorrelation function is not exploiting optimally the information available in the statistics of the data. It turns out that correntropy is more discriminative, as shown in Fig 2.6, i.e. the autocorrentropy of a sinewave (or any other periodic function) is a periodic pulse train with period given by the data, where the pulses can be made arbitrarily sharp by decreasing the kernel size to zero. This can be easily explained by observing Eq. (2.68). When  $x_n$  and  $x_{n-m}$  are similar the argument is close to zero and the Gaussian yields a value close to the argument square; when the difference increases, the Gaussian function produces smaller results proportional to the difference in arguments around the value  $(x_n - x_{n-m})^2/\sigma^2 = 1$ ; and for larger differences, the Gaussian gives back very small values close to zero effectively exponentially decreasing the differences (see Fig 2.5a). Of course if white noise is added to the sinewave, one immediately sees that the kernel size can not be made arbitrarily small, otherwise the correntropy becomes always very small, not capturing the periodic nature of the noisy signal. But if the kernel size needs to be made very large to accommodate large noises, then the autocorrentropy approaches the autocorrelation function.

In [96] the correntropy function and the CSD were used to solve the problem of detecting the fundamental frequency in speech signals. Correntropy outperformed conventional correlation, showing greater discriminatory and robustness to noise. In [97] correntropy was used to solve the blind source separation (BSS) problem, successfully separating signals coming from independent and identically distributed sources and also Gaussian sources. Correntropy outperformed methods that also make use of higher-order statistics such as Indepen-

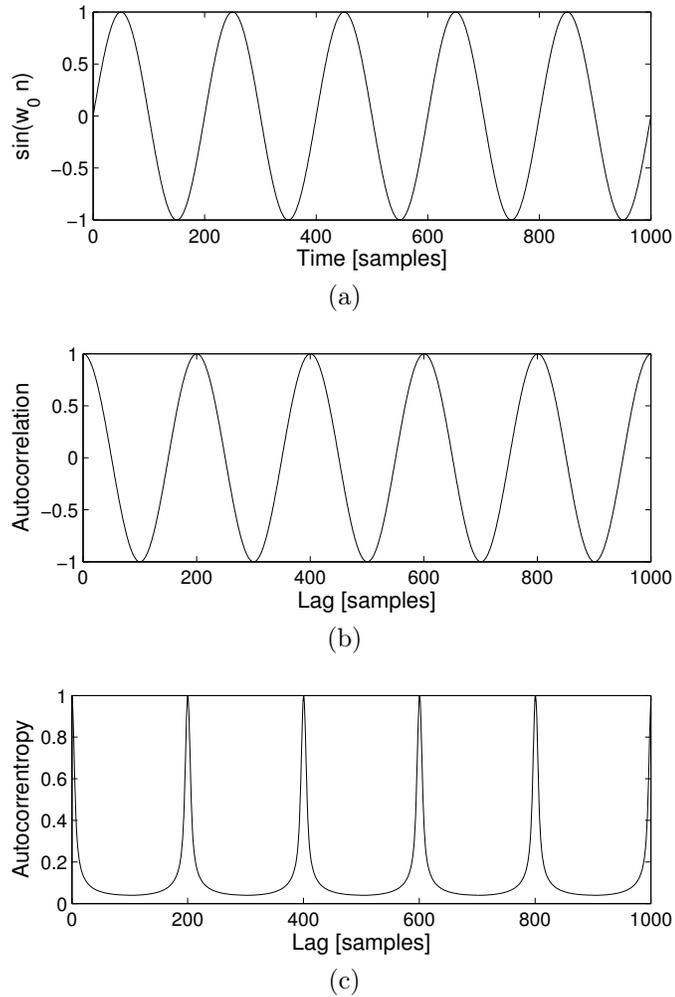


Figure 2.6: (a) Plot of  $x_n = A \sin(w_0 n + \varphi)$  with unit amplitude,  $w_0 = 2\pi/200$  and where  $\varphi$  is a random variable uniformly distributed in  $[-\pi, \pi]$ . (b) Autocorrelation of  $x_n$ , note that the autocorrelation function of a sinewave is a sinewave. (c) Autocorrentropy of  $x_n$ , note that the autocorrentropy of a sinewave is a train pulse in which the periodicity is represented by the peaks. The sharpness of the peaks can be controlled by changing the kernel size  $\sigma$ .

dent Component Analysis (ICA). In [98], correntropy was used as a discriminatory metric for the detection of nonlinearities in time series, outperforming traditional methods such as the Lyapunov exponents.

# Chapter 3

## Methods

In Section 2.10.2 the correntropy function for random processes was introduced. The autocorrentropy function measures similarities between segments of a time series and it is appropriate for recurrence detection and period estimation tasks. The original definition of the autocorrentropy function (Eq. 2.68) is only applicable for time series which are regularly sampled in time, which is not the case for astronomical light curves. In order to use Eq. (2.68) one would need to interpolate the unevenly sampled time series into a regular time grid. This creates a new time series which is regular in time. The main disadvantage of interpolation schemes is that the original data of the process is not used by the correntropy estimator. In the particular case of light curves, simple interpolation schemes might introduce spurious content due to the data gaps and the changes in sample density along the time axis. Third order splines [39] were used to interpolate light curves for a period estimation pipeline that we published in [99]. In [99] we identified the aforementioned pitfalls of interpolation in the case of light curves. Instead of exploring or developing more advanced interpolation schemes, an approach that uses the actual light curve data was preferred.

The slotting technique, reviewed in Section 2.2.1, provides an alternative to develop autocorrelation estimators for unevenly sampled processes. In this thesis an autocorrentropy estimator for unevenly sampled time series, that is based on the slotting technique, is proposed [100]. A Gaussian kernel is considered for the slotting function (Sec. 2.2.2). The details of this method and a period estimation pipeline for light curves based on the slotted correntropy, are presented in Section 3.1. In [100] we showed that the slotted correntropy performed better than the slotted correlation and conventional methods used in astronomy for period estimation. One disadvantage of the slotted correntropy is its dependence on the slot size. The slot size defines the resolution of the time lag axis and it represents a trade-off between the variance of the estimator and the ability to capture short periods (fast frequencies).

Motivated by this we proposed the correntropy kernelized periodogram (CKP) [101], a generalized periodogram developed for periodic light curve discrimination. The CKP does not require resampling, slotting or folding schemes, as it is estimated directly from the available samples of the light curve following the approach of the direct quadratic spectrum estimator (Sec. 2.1.3). The CKP combines the correntropy function and the periodic kernel function. Correntropy assess similarities between magnitudes and the periodic kernel assess similarities

between time instants for a given period. The details on the CKP metric and an statistical test for light curve periodicity based on the CKP are presented in Section 3.10. In [101] we showed that the CKP outperforms the slotted correntropy and conventional methods for period estimation and periodicity discrimination in light curves. One disadvantage of the CKP is the high presence of harmonics, sub-harmonics and aliases of the underlying period in the periodogram. This difficults the task of discriminating two or more periodic signals embedded in the same light curve.

With this in mind, and with the idea of increasing the frequency resolution of the periodogram the Correntropy Non-negative Matrix Factorization Spectrum (CNMFS) is proposed. The CNMFS is a decomposition of the correntropy function into an adaptive dictionary of periodic functions. The decomposition is obtained with a modified implementation of the NMF algorithm. The details on the CNMFS, and its application for the case of astronomical light curves, are presented in section 3.4. Using the CNMFS sparse periodograms with high frequency resolution and high peak resolvability can be obtained.

### 3.1. Slotted correntropy

The slotted autocorrentropy function, for a discrete random process  $\{X_n\}$  with  $n = 1, \dots, N$ , is computed as follows:

1. Select an appropriate value for the slot size parameter  $\Delta\tau$ . The standard choice for the Gaussian slotting function is  $\Delta\tau = \Delta t/4$ , where  $\Delta t$  is the average sample rate of the time series.
2. Generate a set of discrete lags as:  $k \cdot \Delta\tau$ , where  $k = 0, 1, \dots, N_S$ ,  $N_S = \lceil \frac{\tau_{max}}{\Delta\tau} \rceil$ ,  $\tau_{max}$  is the maximum lag and  $\lceil \cdot \rceil$  is the nearest integer function. The maximum lag has to be set in order to ensure that the autocorrentropy at higher lags is not estimated with too few samples. Usually the maximum lag is set to  $\tau_{max} = 0.1N$ , *i.e.* ten percent of the total length of the time series, which is a rule of thumb in autocorrelation estimation.
3. For each lag  $k\Delta\tau$  and each pair of indexes  $i$  and  $j$  associated to samples  $(t_i, x_i)$  and  $(t_j, x_j)$  respectively, compute the Gaussian slotting function  $B_{k\Delta\tau}$  as

$$\hat{B}_{k\Delta\tau}(t_i, t_j) = G_{\Delta\tau}(k - |t_i - t_j|), \quad (3.1)$$

and calculate the slotted autocorrentropy estimator at lag  $k$  as

$$\hat{V}_S[k\Delta\tau] = \frac{\sum_{i=1}^N \sum_{j=i+1}^N G_{\sigma}(x_i - x_j) \cdot B_{k\Delta\tau}(t_i, t_j)}{\sum_{i=1}^N \sum_{j=i+1}^N B_{k\Delta\tau}(t_i, t_j)} \quad \text{with } k = 0, 1, \dots, N_S. \quad (3.2)$$

where  $G_{\sigma}(\cdot)$  is the Gaussian kernel function with kernel size  $\sigma$ . The slot size  $\Delta\tau$  and the kernel size  $\sigma$  are used-defined parameters that need to be set accordingly to the data. The slot size  $\Delta\tau$  defines the time lag resolution of the correntropy function. If it is set too small, only a few samples will satisfy the slotting condition and the variance of the estimator

will increase. If it is set too large, interesting high-frequency behavior will be lost from the correlogram function, *i.e.* short periods may not be found. The average sampling rate, the sample density along the time axis and the histogram of data gaps provide valuable insight to select this parameter.

The kernel size of the correlogram function controls the influence of the higher order moments, if it is set too high then correlogram approximates the conventional correlation (second order moment dominates). If it is set too small, then it might not be possible to discriminate between the signal of interest and the noise present in the light curve. The Silverman's rule of thumb (Eq. 2.27) provides an initial guess for this parameter. In the case of light curves, it is common that an estimation of the photometric error per sample is available. Statistics of the photometric error can be used to define a lower bound for the kernel size.

Listing 3.1 shows an efficient Matlab implementation of the slotted correlogram using a Gaussian slotting window. This function receives two  $1 \times N$  vectors, the time instants and the magnitudes of the light curve. The kernel size is estimated using the Silverman's rule (line 3). The maximum lag is set to 10% of the total time span (line 6). All the differences between time instants and magnitudes are computed (lines 7 to 12). After that, the pairs are sorted in ascending order according to their time differences (this speeds up the search for pairs in the next step). For each time lag, the time differences in a neighbourhood of three kernel sizes ( $3\sigma$ ) are selected<sup>1</sup> (line 19). The relative error of this approximation with respect to using all the pairs is  $\sim 10^{-5}$ , and the speed-up is  $\sim 10$ . The Gaussian slotting window and the correlogram value for that lag are computed using only the selected pairs (lines 20-22).

Listing 3.1: Matlab implementation of the slotted correlogram

```

1  function [lag,v] = slottedCorrelogram(t,x)
2  N=length(t);
3  sigma = 1.06*std(x)*power(N,-0.2);
4  msr = (t(end) - t(1))/N;
5  ss = msr*0.25;
6  max_lag = 0.1*(t(end)-t(1))/ss;
7  dt = repmat(t,N,1);
8  dt = (dt - dt')';
9  dt = dt(tril(true(size(dt))));
10 dx = repmat(x,N,1);
11 dx = (dx - dx')';
12 dx = dx(tril(true(size(dx))));
13 [dt,I] = sort(dt,);
14 dx = dx(I);
15 Gx = exp(-0.5*dx.^2/sigma^2);
16 v = zeros(size(0:max_lag));
17 lag = zeros(size(0:max_lag));
18 for k = 0:max_lag
19     index = find(abs(dt - k*ss) < ss*3);
20     Gw = 1/(2*pi*ss)*exp(-0.5*(dt(index) - k*ss).^2/ss^2);
21     lag(k+1) = sum(Gw.*dt(index))/sum(Gw);
22     v(k+1) = sum(Gw.*Gx(index))/sum(Gw);
23 end
24 end

```

<sup>1</sup>Samples outside this neighbourhood contribute less than 0.3% to the Gaussian weighted average.

---

Using the slotted correntropy function it is possible to compute the correntropy spectral density (CSD) of the light curve. The periodicities detected with the correntropy function appear as peaks in the CSD. A strong peak at a given frequency  $f$  suggest that a periodicity with a period  $1/f$  exists within the light curve. But a peak might also be related to harmonics of the fundamental frequency, aliasing, or noise. In the next subsection an information theoretic metric to discriminate the CSD peak associated to the fundamental frequency of the light curve is presented.

### 3.1.1. IP metric for CSD peak discrimination

Spurious peaks associated to noise, aliasing or the sampling may appear in the CSD of the slotted autocorrentropy. In order to discriminate peaks that are associated to real periodic behavior a metric that uses the Information Potential (Eq. 2.63) of the folded light curve is proposed. The trial periods used to fold the light curve are obtained as the inverse of the frequencies associated to the 10 highest peaks of the CSD. The procedure to compute the IP-based discrimination metric for a certain trial period  $P_c$  is as follows:

1. Transform the time axis of the light-curve using the epoch folding transformation as follows

$$\phi_n(P_c) = \frac{(t_n \bmod P_c)}{P_c}, \quad n = 1, \dots, N$$

where *mod* stands for the modulo operation. With this procedure a folded light curve  $\{\phi_n(P_c), x_n\}$  is obtained.

2. Smooth the folded light curve by taking a moving average of 20 samples. Search for local maxima and minima using samples in windows of size  $M = N/10$  and an overlap of half the window size. The value of M is set as a compromise between detecting spurious peaks as local maxima and missing true peaks.
3. Segment the folded light curve (non-smoothed) into bins so that each local optima corresponds to the center of a different bin. The boundaries of the bins are chosen as the mid-points between adjacent local optima. This procedure results in an adaptive number of bins  $H$ .
4. Compute the discrimination metric as the average of the squared differences between the IP of each individual bin,  $h$ , and the global IP,

$$Q(P_c) = \frac{1}{H} \sum_{h=1}^H [IP(\{x_n\}_{n \in h}) - IP(\{x_n\})]^2. \quad (3.3)$$

If the light curve is folded with a period that is unrelated to the underlying period of the star, a noisy pattern will appear in the phase diagram and the data will appear uniformly distributed across the bins. In this case, the distribution of the binned data will be very

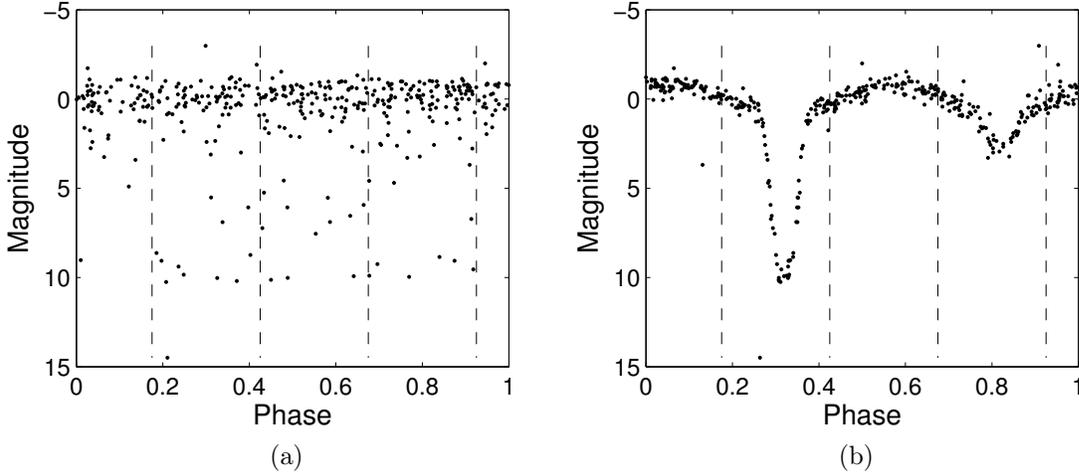


Figure 3.1: (a) Light curve folded with a period that is unrelated to the underlying period of the star. The distribution across the bins is uniform and almost equal to the distribution of the complete light curve, which translates into a low IP metric value. (b) In this case the same light curve is folded with its true period, revealing a phase diagram of an eclipsing binary. In this case the IP metric yields a high value, because the distribution of the bins differs from the distribution of the complete light curve.

similar to the distribution of the complete light curve which translates into a low IP metric value. On the other hand, if the underlying period or one of its multiples is used to fold the light curve, a clear pattern will be revealed in the phase diagram. In this case the data bins will have dissimilar distributions and a high IP metric value will be obtained.

In this sense, the IP metric can be considered as a generalization of the AoV statistic for light curves (Section 2.3.3), where the variance (second order statistic) has been replaced by the IP, a descriptor that incorporates higher order statistics of the distribution of the binned data. The adaptive bin selection criterion was introduced for the case of eclipsing binary stars in order to decrease the chance of getting harmonics of the true period. Maximizing Eq. (3.3) is equivalent to searching for the largest difference between the information content in the dynamically chosen bins and the complete light curve. The maximum occurs when the light curve is folded using its underlying period. Fig. 3.1 shows the phase diagrams and the bin edges from which the IP metric is computed for the case of a wrong period (Fig. 3.1a) and the underlying period of the star (Fig. 3.1b).

Maximizing the IP metric yields the periodicity that produces the most ordered folded light curve from the peaks of the CSD, but it does not tell if a light curve is actually periodic. For example, if a light curve has no significant periodicity (non-spurious) the proposed method would still return a candidate period. An statistical test using the IP metric is needed to prove that the final candidate period corresponds to a real periodicity. In the following sections an statistical periodicity test based on the correntropy generalized periodogram is presented.

### 3.1.2. A pipeline for period estimation using the Slotted Correntropy and the IP metric

In what follows the pipeline for light curve period estimation that we published in [100] is described. This pipeline was tested on a subset of 600 light-curves drawn from the MACHO survey [1]. The subset contains 200 light-curves from each of the following three types of variable stars: EBs, Cepheids and RR Lyrae, whose periods range from 0.2 days to 200 days. Each light-curve has approximately 1000 samples and contains 3 data columns: time, magnitude and an error estimation for the magnitude. The periods of these light-curves were estimated by expert astronomers from the Harvard Time Series Center (TSC) using epoch folding, AoV, and visual inspection. In our work the TSC periods are considered to be the gold standard. The procedure to find the period of a given light curve is as follows:

1. Normalize the light curve’s magnitude to have zero mean and unit standard deviation. Discard samples having an error estimation greater than the mean error plus two times its standard deviation (usually less than 1% of the samples of a light curve are discarded).
2. Select a window of half the length of the light curve containing the maximum number of samples per day. The selected window and the whole light curve are independently analyzed in the next stages in order to generate the pool of trial periods.
3. Compute the slotted correntropy using Eq. (3.2). The maximum lag is set to  $0.1N$ , where  $N$  is the light curve length. This value is chosen as a trade-off between having enough samples to estimate correntropy with longer lags and bounding the longer period that could be detected.
4. Compute centered slotted correntropy by removing its mean value in the interval  $[-0.1N, 0.1N]$ . Multiply this result by a Hamming window in the same interval, and compute the CSD.
5. Store the periods associated with the  $M$  highest peaks of the CSD. For each trial period  $P_c$  compute the IP metric in the interval  $[P_c - 0.5, P_c + 0.5]$  with step size  $10^{-3}$ . Save the fine-tuned period that maximizes Eq. (3.3).
6. Save the fine-tuned period with highest  $Q$  among all the periods obtained from both windows selected in Step 2.

The pipeline has three user defined parameters, the kernel size, the slot size and the number of trial periods extracted from the CSD. In what follows the effects that these parameters have on the analysis and the heuristic rules used to set them are described.

- Gaussian kernel size ( $\sigma$ ): If is set too large the higher-order moments in correntropy will be ignored (correntropy approximates correlation). If it is set too small then correntropy will not be able to discriminate between signal and noise. Steps 1-6 of our method are repeated for 25 values of the kernel size in the interval  $[0.01, 5]$  (logarithmically spaced). Each kernel size provides a set of trial periods.

- Correntropy slot size ( $\Delta\tau$ ): The slot size defines the time lag resolution of correntropy. If it is set too small, it will be harder to satisfy condition (3.1). If it is set too large short periods may not be found. We set  $\Delta\tau = 0.25$  days in order to capture the shorter periods present in the data set.
- Number of peaks analyzed from the CSD ( $M$ ): In our experiments we found that setting this value to  $M = 10$ , is a good trade-off between obtaining the highest hit rate and having less computational load.

The results obtained with this pipeline are presented in Section 4.1.

## 3.2. Correntropy Kernelized Periodogram

For a discrete unidimensional random process  $\{x_n, n = 1, \dots, N\}$  with kernel sizes  $\sigma_t$  and  $\sigma_y$ , and a period  $1/f$ , the CKP is computed as:

$$\text{CKP}_{\{\sigma_t, \sigma_y\}}(f) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (G_{\sigma_y}(y_i - y_j) - IP_{\sigma_y}) G_{\sigma_t}^P(f, t_i - t_j), \quad (3.4)$$

where  $G_{\sigma_y}(\cdot)$  is the Gaussian kernel function (Eq. 2.26),  $G_{\sigma_t}^P(\cdot, \cdot)$  is the periodic kernel function (Eq. 2.31), and  $IP_{\sigma_y}$  is the information potential

$$IP_{\sigma_y} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N G_{\sigma_y}(y_i - y_j). \quad (3.5)$$

Note that Eq. (3.4) is similar to the CSD (Eq. 2.70), but with two main differences: a) the CKP is estimated in a direct quadratic approach and b) the basis functions,  $\exp(-i 2\pi fm/F_s)$  have been replaced by the periodic kernel. In this sense the CKP can be interpreted as the result of transforming the autocorrentropy function through a basis defined by the periodic kernel<sup>2</sup>.

By comparing magnitude values through the autocorrentropy function the CKP is effectively using a CIM (Eq. 2.67) like metric to measure magnitude distances. The kernel size  $\sigma_y$  has influence in the assessment of magnitude similarities as explained in the previous section. The CKP compares time differences with the trial period through the periodic kernel. The periodic kernel size  $\sigma_t$  allows the user to choose how this comparison is made.

By summing both in the time and magnitude indexes, a function of the trial period is obtained, thus the CKP is considered a generalized periodogram. If all the magnitude values separated by a given trial period  $1/f$  are similar, it will be reflected as a peak in the CKP located at  $f$ . Consequently, in order to detect periods in lightcurves the CKP is maximized over the frequency (inverse of the period) for a given combination of parameters, namely the two kernel bandwidths  $(\sigma_y, \sigma_t)$ .

---

<sup>2</sup>Although, the periodic kernel does not form a basis as it does not comply the orthogonality condition.

One of the major advantages of the CKP over conventional methods is its adaptability given by the kernel parameters. In what follows, we describe heuristic approaches that use the available information on the lightcurve to set the kernel sizes. These approximations are necessary when dealing with large datasets. Without them the maximization of the CKP would have been a very expensive procedure.

- The kernel bandwidth,  $\sigma_y$ , controls the observation window that is used to compare the magnitude values of the lightcurve. This parameter needs to be set small enough so that outliers are filtered, but large enough to compensate for the observational and other measurements errors. Conveniently those errors are usually available for most measurements in light curves (these are the magnitude errors). For a given lightcurve the Gaussian kernel bandwidth can be selected as

$$\sigma_y = \text{med}(\{e\}), \quad (3.6)$$

where  $\text{med}$  is the median, and  $\{e\}$  are the error bars of the measurements in light curve. In reality the observational errors are not constant and therefore Eq. 3.6 should not be the same for all pairs and should be a combination of the two observational errors added in quadrature. Practically, though the variations of the error bars in real dataset are not that significant and therefore the difference of this approximation and the correct approach is insignificant.

- The kernel bandwidth,  $\sigma_t$ , controls the observation window that is used to compare the time differences of the light curve with the trial period. When  $\sigma_t \rightarrow 0$  only the samples whose time differences are equal to the trial period will be picked by the periodic kernel. The smaller the  $\sigma_t$  is, the more precise the estimation will be, although in practice fewer samples will be available. When  $\sigma_t$  grows large, the exponential in Eq. (2.31) takes less relevance and the periodic kernel tends to a sinusoidal function<sup>3</sup>. Intuitively, this parameter has influence on the periodicity's shape. A smaller  $\sigma_t$  is beneficial to pick up shapes that have many features or abrupt changes, such as the narrow eclipses of an Algol-type eclipsing binary. On the contrary a large  $\sigma_t$  is used for smoother shapes, *i.e.* wiggles and high derivatives are ignored. In summary the  $\sigma_t$  needs to be set small enough so that the features of the periodicity will not be missed, but large enough so that there will be enough samples representing the period and to avoid picking up structures due to the noise.

Fig. 3.2 shows an example using a synthetic time series to illustrate the effect of the proposed metric. Fig. 3.2a shows a synthetic time series  $y_i = \sin(2\pi t_i/P) + 0.8 \cdot \varepsilon_i$ , with  $t_i = \frac{T_{max}}{N}(i + 0.5 \cdot \varepsilon_i)$ , where  $\varepsilon_i$  and  $\varepsilon_i$  are normally distributed random variables with zero mean and unit standard deviation. The noise in time simulates uneven sampling. In this example  $N = 400$ ,  $T_{max} = 25$ , and the underlying period is  $P = 2.456$  seconds. Fig. 3.2b shows the kernel coefficients,  $G_{\sigma_y}(y_i(t_i) - y_j(t_j))$  and  $G_{\sigma_t}^P(f, t_i - t_j)$ , as a function of the time differences collected from the time series. Fig. 3.2c shows the CKP for a range of periods, the location of the underlying period in the periodogram is marked with a dotted line. The CKP reaches a global maximum at the corresponding underlying period  $P = 2.456$ .

Listing 3.2 shows a simple and efficient Matlab implementation of the CKP. This function receives two  $N \times 1$  vectors, the time instants and the magnitudes of the light curve and two

---

<sup>3</sup>As shown in Section 2.6.3 through the Taylor expansion of the periodic kernel (Eq. 2.31).

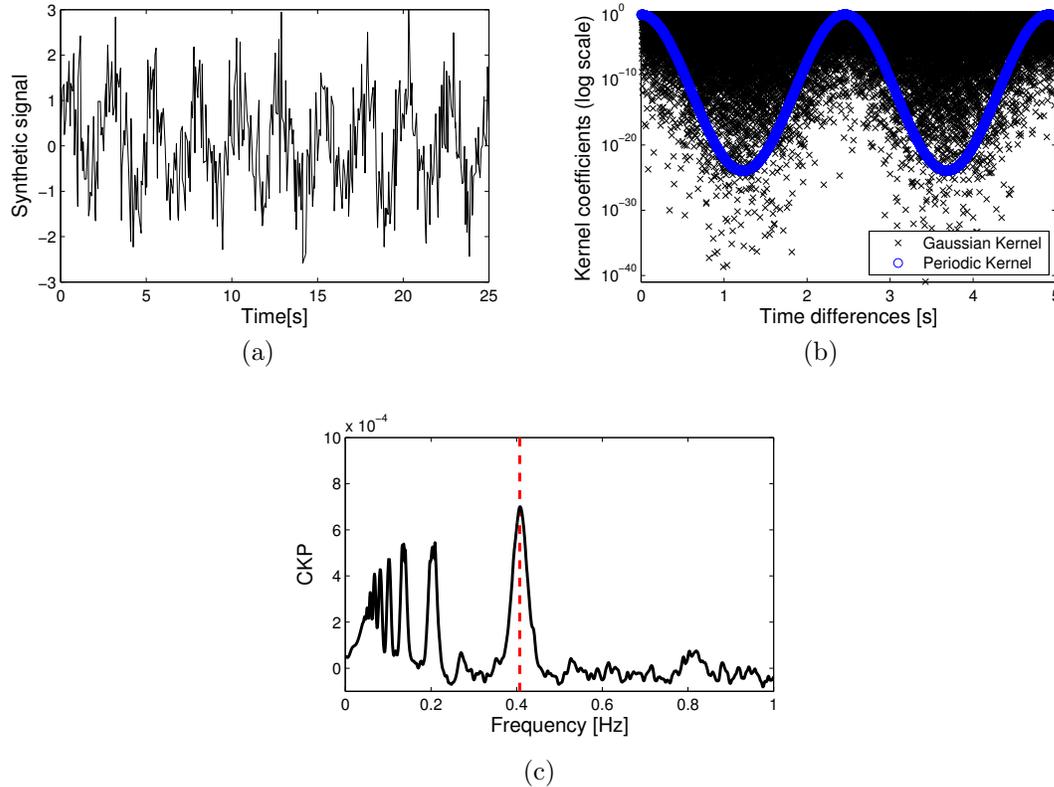


Figure 3.2: (a) Synthetic time series  $\sin(2\pi t/P)$  with period  $P = 2.456$  seconds plus Gaussian noise. The time instants have been randomly perturbed to simulate uneven sampling. (b) Kernel coefficients,  $G_{\sigma_y}(y_i(t_i) - y_j(t_j))$  and  $G_{\sigma_t}^P(f, t_i - t_j)$ , as a function of the time differences. The CKP is the pointwise product between the centered Gaussian kernel coefficients and the periodic kernel coefficients. (c) CKP as a function of the trial frequency, the dotted line marks the location of the true period. The global maximum of the CKP corresponds to the underlying period.

scalar values (kernel sizes). All the differences between time instants and magnitudes are computed. After that, for each frequency value the periodic kernel is computed and the periodogram value is obtained by sum-reducing the multiplied kernel matrices. The limits and resolution of the frequency grid have to be chosen by the user (application dependant). The minimum frequency can be set as the inverse of half the total time span (so that the period appears at least two times in the observable data). The maximum frequency is given by the Nyquist Frequency (see Section 2.1.4 for details on how to select this limit for unevenly sampled light curves).

#### Listing 3.2: Matlab implementation of the CKP

```

1 function [f,periodogram] = CKP(t,x,sx,st)
2 %sx is the magnitude's kernel size, and st is the time's kernel size
3 N = length(t);
4 dt = repmat(t,1,N);
5 dx = repmat(x,1,N);
6 dt = dt - dt';
7 dx = dx - dx';
8 Gx = exp(-0.5*dx.^2/sx^2);

```

```

9     IP = mean(mean(Gx));
10    f = linspace(0,2,1000);
11    periodogram = zeros(length(f),1);
12    for i=1:length(f);
13        Gt = exp(-2*sin(pi*abs(dt)*f(i)).^2/st^2);
14        periodogram(i) = 1/N^2*sum(sum((Gx -IP).*Gt));
15    end

```

---

### 3.2.1. A statistical test based on the CKP for periodicity discrimination

For a periodic time series with an oscillation frequency  $f$ , its periodogram will exhibit a peak at that frequency with high probability. But the inverse is not necessarily true, a peak in the periodogram does not imply that the time series is periodic. Spurious peaks may be produced by measurement errors, random fluctuations, aliasing or noise.

In this section, a statistical test for periodicity is introduced, using the global maximum of the CKP as test statistic. The null hypothesis is that there are no significant periodic components in the time series. The alternative hypothesis is that the CKP maximum corresponds to a true periodicity. To obtain the p-values, the distribution of the test statistic is required. When the distribution of test statistic is unknown Monte-Carlo sampling and surrogate time series can be used to obtain it. The distribution of the maximum value of the CKP is obtained through Monte-Carlo simulations. Surrogate time series [60, 61] are used to test the null hypothesis. The surrogate generation algorithm has to be consistent with the null hypothesis. To achieve this, the block bootstrap method [64], which breaks periodicities preserving the noise characteristic and time correlations of the light curve, is used. The procedure used to construct an unevenly sampled surrogate using the block bootstrap method is as follows

1. Obtain a data block from the light curve by randomly selecting a block of length  $L$  and a starting point  $j \in [1, N - L]$ .
2. Subtract the first time instant of the block, so that it starts at 0 days.
3. Add the value of the last time instant of the previous block to the time instants of the current block.
4. Parse the current block to the surrogate time series.
5. Repeat steps 1-4 until the surrogate time series have the same amount of samples of the original light curve.

For a given significance level  $\alpha$  and kernel sizes  $\sigma_t$  and  $\sigma_m$ , the null hypothesis is rejected if

$$\max_f \text{CKP}_{\{\sigma_t, \sigma_y\}}(f) > \text{CKP}_{\{\sigma_t, \sigma_y\}}^\alpha(f),$$

where for  $N$  light curves  $\text{CKP}_{\{\sigma_t, \sigma_y\}}^\alpha(f)$  is pre-computed as follows:

1. Generate  $M$  surrogates from each light curve using block bootstrap.
2. Save the maximum CKP ordinate value of each surrogate.
3. Find  $P_\alpha$  such that a  $(1 - \alpha)\%$  of the ordinate values saved from the surrogates are below this threshold (one-tailed distribution).
4. Compute  $\text{CKP}_{\{\sigma_t, \sigma_y\}}^\alpha(f)$  as  $\widehat{P}_\alpha$ , the average  $P_\alpha$  among the surrogates, and its corresponding error bars as the standard deviation of  $P_\alpha$  for the  $N$  light curves ( $N \cdot M$  surrogates).

### 3.2.2. Periodicity detection using the CKP as the test statistic

In this section the pipeline for periodicity detection that we published in [101] is described. This pipeline uses the CKP and the statistical test described in Section 3.2.1 in order to discriminate if a light curve is periodic or not. This pipeline was tested on 5,000 light curves from the MACHO project. The subset contains 1,500 periodic light curve and 3,500 light curves from variable stars that are not periodic. The underlying periods of the periodic variable stars were estimated by investigators from the Harvard Time Series Center (TSC) using epoch folding, AoV, and visual inspection. The TSC periods are considered to be the gold standard. There are two light curves per stellar object: channels blue and red. Only the blue channel light curves were used. Each light-curve has approximately 1000 samples and contains 3 data columns: time, magnitude and an error estimation for the magnitude.

For a given light curve and parameters  $\sigma_m$  and  $\sigma_t$

1. **Cleaning:** The light curve’s blue channel is imported. The mean  $\bar{e}$  and the standard deviation  $\sigma_e$  of the photometric error are computed. Samples that do not comply with  $e_i < \bar{e} + 3 \cdot \sigma_e$ , where  $e_i$  is the photometric error of sample  $i$ , are discarded.
2. **Computing the periodogram** The CKP (Eq. 3.4) is computed on 20,000 logarithmically spaced periods between 0.4 days and 300 days. The periods associated to the ten highest local maxima at the periodogram are saved as trial periods for the next step.
3. **Fine-tuning of trial periods:** The CKP is used to fine tune the ten trial periods. Each trial period is fine-tuned around a 0.5% of its value ( $[1.0025 \cdot f_{trial}, 0.9975 \cdot f_{trial}]$ ), using a step size of  $df = \frac{0.01}{T}$  in frequency, where  $T$  is the total time span of the light curve.
4. **Selection of the best trial period:** The trial periods are sorted in descending order following its CKP ordinate value. The best trial period  $P_{best}$  is selected as the one with the highest value of  $\text{CKP}_{\{\sigma_t, \sigma_y\}}(\cdot)$ , that is not a multiple of a spurious period. The spurious periods are detected using the spectral window as described below.
5. Finally, if the best period comply with  $\text{CKP}_{\{\sigma_t, \sigma_y\}}(1/P_{best}) > \theta$  then the light curve is labeled as periodic, where  $\theta$  is the periodogram threshold for periodicity. The confidence

associated to  $\theta$  is obtained using the procedure described in Section 3.2.1.

To obtain the spurious periods the following spectral window function is used

$$W(f) = \frac{1}{N} \left| \sum_{i=1}^N \exp(j2\pi f t_i) \right|^2, \quad (3.7)$$

where  $t_i$  with  $i = 1, \dots, N$  are the time instants of the light curve. Eq. (3.7) is the periodogram of the sampling pattern of the light curve. The frequencies associated to the peaks of the spectral window are related to spurious periodicities caused by the sampling. In most cases, the spurious periods obtained from the spectral window are multiples of the sidereal day (0.99727 days) and yearly periodicities (365.25 days). The moon phase period (29.53 days) is also added to the list of spurious periods. The moon phase has no relation to the sampling but it is intrinsic to the data. The daily sampling produces alias peaks of the true periods ( $P$ ) in the CKP at  $(1/P + k) 1/\text{days}$ , where  $k \in \mathbb{Z}$ . The CKP values of the aliases are very low and therefore they do not need to be filtered out.

### 3.3. An efficient pipeline for periodic light curve discrimination on large astronomical databases using the CKP

In this Section the pipeline for period light curve discrimination that we published in [102] is described. This pipeline is based on the one presented in Section 3.2.2, with the following major changes:

1. An algorithm to select the kernel size of the periodic kernel function that is based on the skewness of the light curve is proposed.
2. A fast trial period search algorithm is introduced. This reduces the amount of periods that have to be tested with the CKP, alleviating the computational load of the pipeline.
3. A normalization term for the CKP is added. A normalization term is needed in order to make ensemble comparisons between light curves with different number of samples and different kernel sizes.
4. A new way of computing the periodicity discrimination thresholds is proposed. With this new approach, the thresholds are computed for the whole survey as a function of the signal-to-noise ratio, instead of the per-light curve thresholds that were previously used (Section 3.2.1).
5. The CKP and fast trial period extraction method are implemented in a cluster GPGPU environment taking into consideration the latest optimizations pointed by the GPU manufacturer.

Most of these modifications are focused on decreasing the computational time of the pipeline, in order to process large astronomical databases efficiently. The new steps and

methods of the pipeline are described in the following subsections. The complete pipeline is presented in Section 3.3.7.

### 3.3.1. An heuristic rule to select the kernel size of the periodic kernel function

Since  $\sigma_t$  describes the smoothness of the shape of the light curve, a simple way to estimate  $\sigma_t$  is to find the variation of  $\delta t$ 's in a given y-band. A method to do this was examined, but it was not robust against low sampled light curves.

Empirically, we observed that for almost all periodic light curves, the CKP is maximized for  $\sigma_t \sim 0.1 - 0.6$  and that the value of  $\sigma_t$  is strongly correlated with the third moment or the skewness of the distribution of the magnitudes of the light curves. For a given lightcurve the periodic kernel bandwidth is selected as

$$\sigma_t = \lambda + 0.5 \exp(-\mu S(\{x\})^2), \quad (3.8)$$

where  $S(\{x\})$  is the quartile estimator of the skewness [103]

$$S(\{x\}) = \frac{Q_3(\{x\}) + Q_1(\{x\}) - 2Q_2(\{x\})}{Q_3(\{x\}) - Q_1(\{x\})}, \quad (3.9)$$

and  $\lambda = 0.1, \mu = 12$  were estimated empirically for the EROS-2 database using synthetic light curves. Light curves with skewed distributions, such as those corresponding to eclipsing binaries (Fig. 3.3a), get a small  $\sigma_t$  value. On the other hand, light curves with very symmetric distributions (Fig. 3.3b) will get a larger  $\sigma_t$ .

### 3.3.2. Normalizing the CKP against sample size and kernel sizes

In what follows we use the CKP for period estimation and periodicity discrimination for light curves for the EROS-2 dataset. To accomplish the task for periodicity discrimination, light curves are compared through their CKP values. The kernel sizes for the Gaussian kernel and periodic kernel are selected for each light curve differently as described above using Eq. (3.6) and Eq. (3.8), respectively. In order to compare different light curves, the CKP is required to be invariant under  $\sigma_y, \sigma_t$  and the sample size  $N$ .

For that we propose a properly normalized CKP metric as:

$$\text{nCKP}_{\{\sigma_t, \sigma_y\}}(f) = \frac{\sqrt{N}\sigma_t}{IP_{\sigma_y}} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (G_{\sigma_y}(\Delta y_{ij}) - IP_{\sigma_y}) G_{\sigma_t}^P(f, \Delta t_{ij}), \quad (3.10)$$

where  $1/IP_{\sigma_y}$  normalizes against  $\sigma_y$ ,  $\sqrt{\sigma_t}$  normalizes against  $\sigma_t$  and  $\sqrt{N}$  normalizes against the number of samples. The normalization factors were confirmed empirically by comparing the distribution of the CKP across different sets of surrogate light curves. Fig. 3.4a shows a histogram of  $\max \text{CKP}_{\{\sigma_t, \sigma_y\}}(f)$  for three sets of surrogates generated with different  $N$  values.

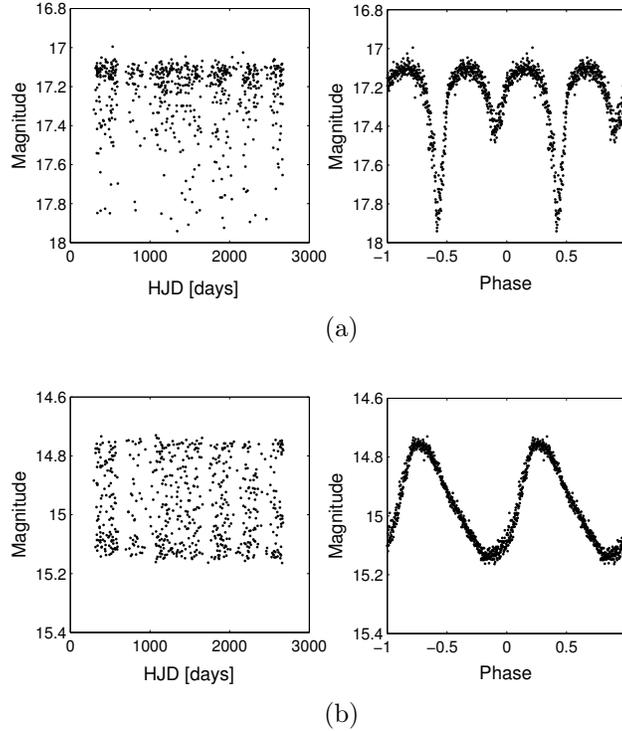


Figure 3.3: (a) Light curve lm0090l7821 folded with a period of 1.4255 days. This light curve has a highly positive skewed distribution. A time kernel bandwidth of  $\sigma_t = 0.115$  is selected for this lightcurve. (b) light curve lm0090n9337 folded with a period of 4.3949 days. This light curve has a symmetric distribution. In this case a time kernel bandwidth of 0.475 is selected.

In this figure the unnormalized CKP is used (Eq. 3.4). For the histogram shown in Fig. 3.4b the normalized CKP (Eq. 3.10) is illustrated, in this case the distribution of the CKP is equivalent, thus it is invariant to the different  $N$  of the surrogates.

### 3.3.3. Trial period extraction, the bands method

The final parameter to be estimated by maximizing the CKP is the period. Unfortunately the dependence on the CKP to period is not uniform and difficult to model [101] and therefore even clever optimization techniques fail to converge faster than the brute force approach.

To alleviate this problem, a fast search algorithm is adopted. The basic idea is that two points in an ideal lightcurve having the same magnitude, have to be apart in time by an integer multiple of the period. For the ideal lightcurve case, finding the period is as simple as finding the greatest common divisor of the times of two points with the same magnitude<sup>4</sup>. However, the ideal case is not applicable to astronomical data because: a) the observations are not performed continuously and b) measurements are not perfect but suffer from observational errors.

<sup>4</sup>This is the oldest known algorithm, namely the Euclid algorithm.

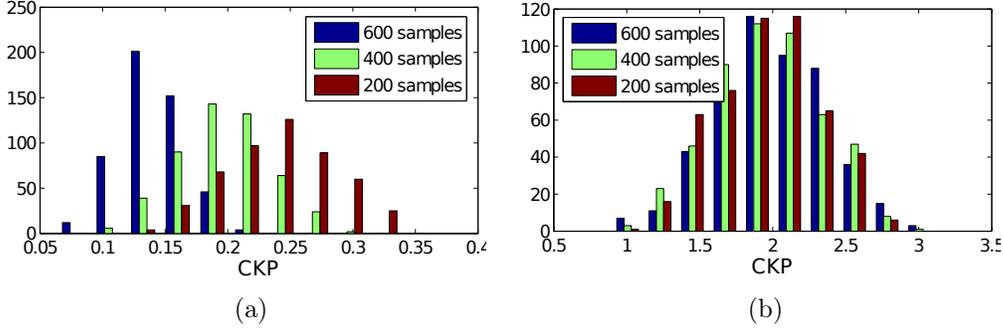


Figure 3.4: Distribution of the maximum CKP values on a set of 1500 synthetic light curves. The light curves are generated with the same period and SNR but using different number of samples ( $N$ ). Three sets of 500 light curves are generated using 200, 400 and 600 number of samples, respectively. Fig (a) shows the distribution of the unnormalized CKP. It is clear the CKP is not invariant to  $N$ . Light curves with higher  $N$  have higher CKP values. Fig (b) shows the distribution of the normalized CKP.

What follows, is an approximation to the ideal case tailored for real light curves. The idea is that instead of looking at pairs of points with the same magnitude, we select subsets of points with similar magnitudes. These subsets, called bands, should contain points that have time differences that are multiples of the period, and therefore, in Fourier space these periods are enhanced. To avoid bands where the light curve is in its nominal state we select bands where the derivatives are higher.

The trial period extraction method, referred as the bands method from here on, is described in what follows. For an unidimensional time series  $\{t_i, x_i\}$  with  $i = 1, \dots, N$

- Compute the first derivatives  $d_i = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$ .
- Divide the ordinate axis in  $m$  uneven-width bands, such that each band has  $N/m$  samples.
- Compute the sum of the first derivatives that belong to band- $j$  ( $B_j$ ),  $D_j = \sum_{i \in B_j} |d_i|$ , with  $j = 1, \dots, m$ .
- Sort the bands in descending order of  $D_j$  and keep the first  $N_b$  bands.
- For each band compute the spectral window function [32] on a linearly spaced frequency grid from 0.00125 1/days to 3 1/days (periods between 0.3 days and 800 days),

$$S_j(f) = \left| \sum_{i \in B_j} \exp(j2\pi f t_i) \right|^2 \quad (3.11)$$

- Save the frequencies associated with the  $N_t$  highest local maxima of  $S_j(f)$ . Periods that comply with  $\|P - 1\| < 1e - 4$  are omitted<sup>5</sup>. This gives a total of  $N_b N_t$  trials

The number of bands,  $N_b$ , and the amount of trial periods extracted per band,  $N_t$ , are user-defined parameters. These parameters represent a trade-off between hit rate<sup>6</sup> and com-

<sup>5</sup>The one day pseudo sampling period is strongly represented in all the bands.

<sup>6</sup>Finding the underlying period of the light curve.

putational time. The correct period is expected to appear in the first sorted bands, however the true period may be captured by different bands although with different amplitudes, *i.e.* the rank of true period may vary across bands. For example the true period may be ranked 100<sup>th</sup> in the first band and 10<sup>th</sup> in the third band.

The values of these parameters were selected through extensive testing using synthetic light curves. For the EROS-2 database the best operation point was found to be  $N_b = 3$  and  $N_t = 150$ . For more details please refer to [102].

Fig. 3.5a shows a plot of an EROS-2 lightcurve, lm0090m4818. Fig. 3.5b shows the same lightcurve folded with a period of 1.54192 days. The black dotted lines mark the band divisions on the magnitude axis. The shaded region shows the best band in terms of the first derivatives criterion. Fig. 3.5c shows a plot of the spectral window function of the time instants extracted from the best band of lm0090m4818. The true period of the lightcurve is associated with the eighth highest local maximum of the spectral window. In this case, if  $N_t > 8$  then the underlying period will be within the trial period set that is to be evaluated by the CKP in the next step of the pipeline.

### 3.3.4. A procedure to filter spurious periods present in the EROS-2 light curves

The list of trial periods extracted with the bands method and analyzed with the CKP contain spurious periods related to the solar day, sidereal day, the moon phase, the year, their multiples (harmonics and subharmonics) and aliases between them. These periodicities are embedded in the EROS-2 light curves and are common in ground-based astronomical observations. Even light curves from non-variable stars exhibit these periodicities. In order to discriminate periodic variability due to stellar phenomena these spurious periods must be filtered.

A complete list of spurious periodicities found in the EROS-2 survey light curves is given in Table 3.1. This list of spurious periods was found by analyzing the histogram of the periodic light curves detected with the pipeline (Fig. 3.6a). For each of these spurious periods a filter, shaped as a Gaussian kernel in the period vs CKP space, is created. The standard deviation and the amplitude of the masks are set so that the associated spurious peak in the period histogram is flattened (Fig. 3.6b). The trial periods and their associated CKP values are tested to check if they fall into any of the spurious periods masks. Fig 3.7 shows a Gaussian mask used to filter periods associated to the sidereal day. The circles corresponds to trial periods that fall into the mask, *i.e.* they are filtered as spurious. If all the trial periods for a given light curve are filtered by the masks, the light curve is immediately labeled as non-periodic. The crosses correspond to trial periods that are not related to the sidereal day, according to the filter. The trial period that maximizes the CKP and passes all the tests is selected as the best trial period for the light curve, and continues to the following stages of the pipeline.

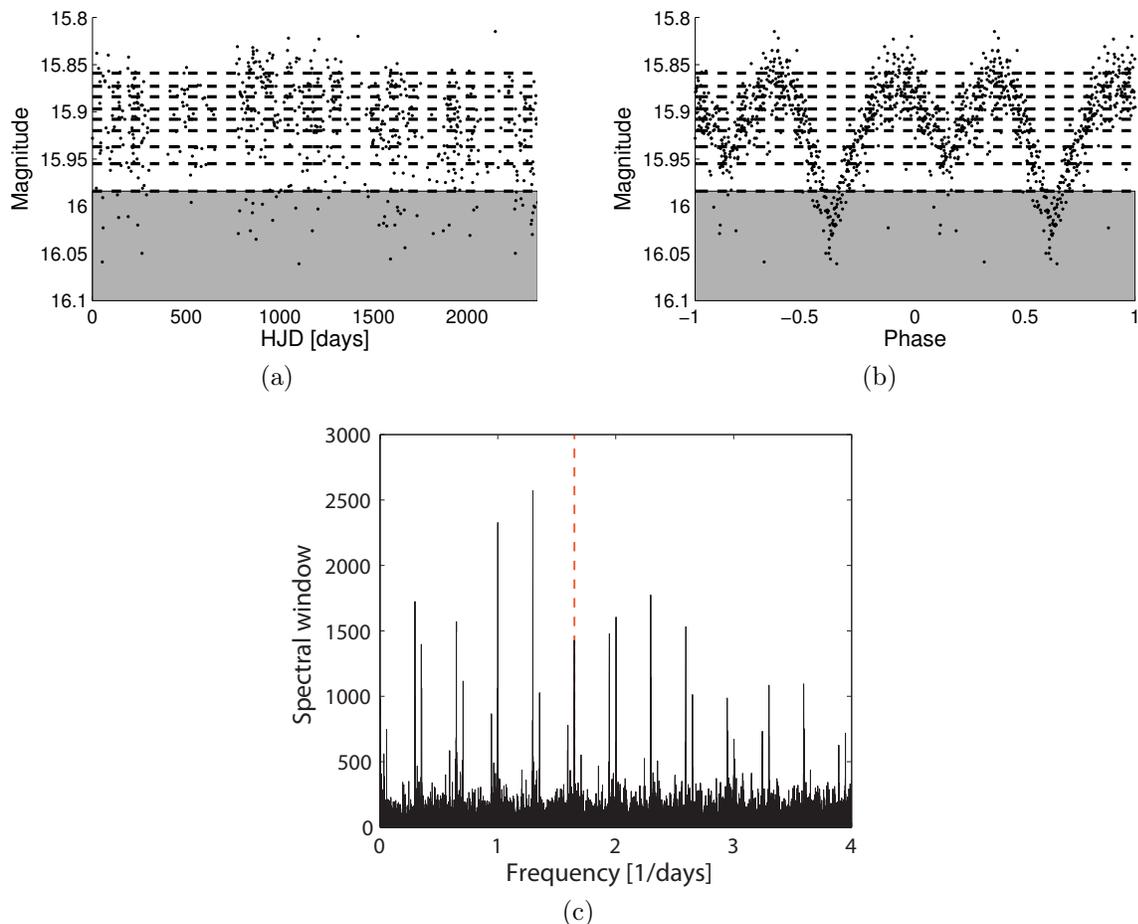


Figure 3.5: (a) EROS-2 lightcurve  $lm0090m4818$ . The dotted lines show the band divisions. The shaded region shows the best band in terms of the first derivatives criterion. (b) Same lightcurve folded with a period of 1.54192 days. (c) Spectral window of the tenth band from lightcurve  $lm0090m4818$ . The red dotted line shows the location of the underlying period ( $1/P = 0.6485$ ). The underlying period is associated to the eighth highest local maximum of the spectrum.

### 3.3.5. Obtaining the periodicity discrimination thresholds

A light curve is labelled as periodic if the CKP value associated to its best trial period is above a given periodicity discrimination threshold. The threshold is determined by optimizing the  $F_1$  score (Eq. 3.24) with a training set created as described in Appendix A. The periodicity threshold is obtained as a function of the Signal-to-Noise ratio, which in this case is defined as

$$SNR = \frac{0.7413 \text{ iqr}(\{y\})}{\text{med}(\{e\})}, \quad (3.12)$$

where  $\text{iqr}(\cdot)$  is the interquartile range,  $\text{med}(\cdot)$  is the median and  $y$  and  $e$  are the magnitudes and error bars of the light curve, respectively.

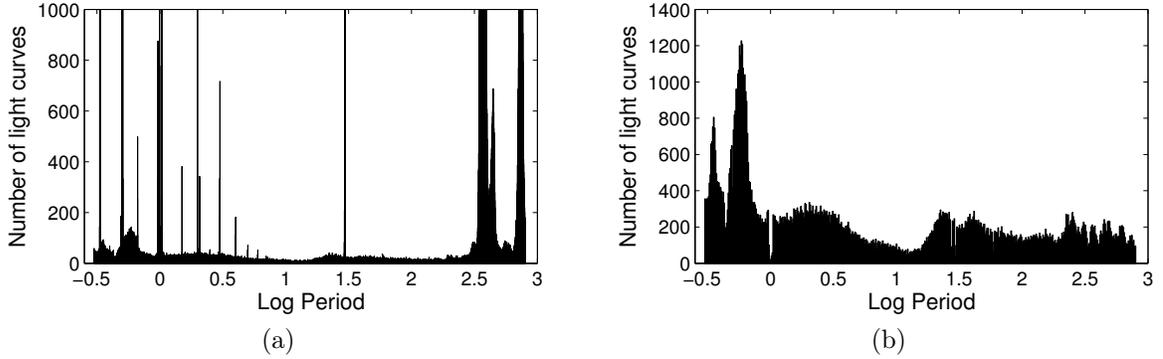


Figure 3.6: (a) Histogram of the periodic light curves detected with the proposed method on the EROS-2 light curves from the LMC. The spurious periods have not been filtered in these results. The vertical lines corresponds to the spurious periods, their multiples and aliases. (b) Histogram of the periodic light curves detected in the LMC after carrying out the spurious period removal scheme.

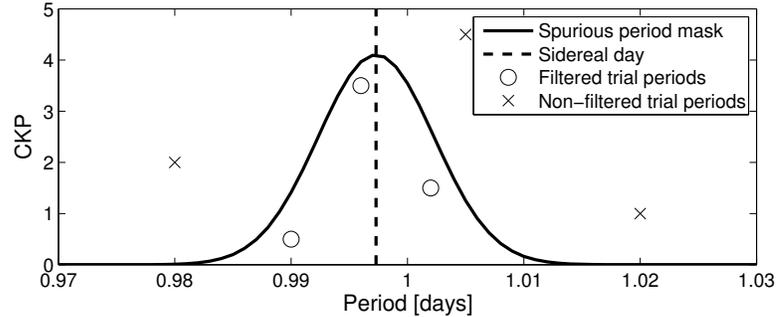


Figure 3.7: Gaussian mask (solid line) used to filter periodicities associated to the sidereal day (dotted line). Trial periods that fall into the mask are marked with circles. Trial periods that are not related to the sidereal day are marked with crosses, these periods continue to the following stages of the pipeline.

The SNR values are discretized in eight bins:  $S = \{[0, 1.5], [1.5, 2], [2, 2.5], [2.5, 3.5], [3.5, 5], [5, 10], [10, 20], [20, \infty]\}$ . The periodicity threshold is computed according to the following procedure:

- Evaluate the CKP values for each light curve in the training set whose SNR fall in bin  $S$ .
- Construct a threshold array of 5000 points in  $[\min(\text{CKP}), \max(\text{CKP})]$ .
- Compute the  $F_1$  score (Eq. 3.24) at each threshold value.
- Select the threshold  $th(S)$  as the CKP value that maximizes the  $F_1$  score.

Once the thresholds have been computed, a light curve whose SNR falls in bin  $S$  is labelled as periodic if:

$$CKP(P_{best}) > th(S),$$

where  $P_{best}$  is the detected period that maximizes the CKP for the given light curve.

Table 3.1: Description of the spurious periods

Period [days]	Description
1	Solar day ( $P_d$ )
29.5305	Moon phase or Synodic month ( $P_m$ )
365.24	Tropical year ( $P_y$ )
2,335	Average time span of EROS-2 lightcurves ( $T$ )
0.4917	$((P_d/2)^{-1} + P_m^{-1})^{-1}$
0.5086	$((P_d/2)^{-1} - P_m^{-1})^{-1}$
0.9672	$(P_d^{-1} + P_m^{-1})^{-1}$
1.0351	Lunar day, $(P_d^{-1} - P_m^{-1})^{-1}$
0.9973	Sidereal day, $(P_d^{-1} + P_y^{-1})^{-1}$
1.0027	$(P_d^{-1} - P_y^{-1})^{-1}$
27.31	Sidereal month, $(P_m^{-1} + P_y^{-1})^{-1}$
32.13	$(P_m^{-1} - P_y^{-1})^{-1}$
315.65	$(P_y^{-1} + T^{-1})^{-1}$
432.63	$(P_y^{-1} - T^{-1})^{-1}$

### 3.3.6. Detecting additional periodic components

It is known that periodic stars might exhibit multimode oscillations which is manifested in the morphology of the light curves [104]. For example, double mode Cepheids and RRd Lyrae stars pulsate regularly at their fundamental frequency (period  $P_0$ ) and first overtone (period  $P_1$ ). The ratio  $P_1/P_0$  can be used to discriminate the class of a double mode star [104]. Periods associated to the second and subsequent oscillation modes are expected to appear in the periodogram of the light curve. In this section a procedure to find additional periodic components on a periodic light curve is presented. This procedure is applied as a post-processing stage in the pipeline.

For each periodic light curve, the strongest period<sup>7</sup> found,  $P_0$ , is used to ‘remove’ the periodic signal. By removing this component from the time series, the peaks associated to  $1/P_0$  and all its harmonics and sub-harmonics will be eliminated from the CKP. This facilitates the process of finding additional periodicities, specially in the case of very strong fundamental frequencies. This procedure is performed as follows:

1. Fold the light curve with  $P_0$ .
2. Obtain a template of the periodicity by smoothing the folded light curve using a moving average of 30 samples.
3. Subtract the template from the folded light curve.
4. Rearrange the light curve samples to their original time order.

If the whitened light curve is found to be periodic with period  $P_1$ , that is not multiple/sub-multiple or alias of  $P_0$ , then the light curve is selected as a dual mode candidate. Subsequent

<sup>7</sup>the one with the highest associated CKP value.

oscillation modes can be found by repeating the procedure above.

### 3.3.7. Pipeline for periodic light curve discrimination in large astronomical databases

In this section the steps of the pipeline published by us in [102] are enumerated. The major steps are described in previous sections. For a given EROS-2 light curve:

1. Outlier samples having a photometric error larger than the mean plus three times the standard deviation are removed.
2. A least square linear fit is applied to the light curve. If the correlation coefficient between the linear fit and the light curve is above 0.5, the linear fit is subtracted from the light curve. This is equivalent to removing a linear trend from the light curve.
3. The band method is used to obtain 450 trial periods (Three bands and 150 trial periods per band).
4. The kernel sizes for the light curve are estimated using Eq. (3.6) and Eq. (3.8).
5. The CKP is used to evaluate the trial periods. The trial periods are sorted in decreased order following their CKP value. Spurious periods related to the sidereal day, Synodic month, tropical year, total time span and their aliases are removed from the list, as described in Section 3.3.4.
6. The period with the highest CKP that is not spurious is tested against the periodicity threshold. If its CKP is higher than the threshold the light curve is labeled as periodic.
7. If the light curve is labeled as periodic, the period is removed from the time series and pipeline is run again in order to find additional oscillation modes.

### 3.3.8. About GPGPU implementations and GPU cluster environments

In order to achieve competitive computational times the following three strategies have been used.

- Programming based on Graphical Processor Units (GPUs). GPU programming also called GPGPU (general purpose computing in GPUs) is a new paradigm for highly parallel applications where high-complexity calculations are offloaded to the GPU (co-processor). GPUs are inherently parallel harnessing up to 512 processing cores. The GPU manufacturer, NVIDIA, provides a toolkit and compiler known as CUDA [105] (Compute Unified Device Architecture) for parallel programming on GPUs. The CUDA programming language is a variation of C. An open source alternative to CUDA is OpenCL which is supported by most GPU manufacturers.

- A limited search for period candidates is done. A sweep over the whole range of periods usually takes 20,000 trial periods, and therefore 20,000 CKP evaluations per light curve. We have reduced the search to 600 trial periods per light curve by using the bands method. The description and evaluation of the trial period search algorithm can be found in Section 3.3.3 and in greater detail in [102].
- Running the program in a computer cluster. In order to speed up the computation of tens of millions of light curves, we have used the Dell/NVIDIA Cluster called FORGE of the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign. We used a queue with 12 machines (nodes) each having 16 CPU cores and 8 Tesla C2070 GPUs, having in total 96 GPUs. For a given node we launch 8 OpenMP threads each one controlling one GPU. We compute one observational field of  $\sim 17,000$  light curves per GPU, i.e. we process simultaneously 96 independent fields. Fig. 3.8 shows a flow diagram of this parallel computation operation.

Today, modern CPUs offer up to 240 GFlops (latest Intel Xeon E5 2690 released in march 2012) of numerical performance while modern GPUs offer approximately 4 TFlops (Nvidia GTX 680), equivalent to a super-computer from 10 years ago. Due to this big difference of orders of magnitude, people from the HPC community are moving towards GPU computing [106]. But, not all applications benefit from GPGPU. Explicit thread and data parallelism must be exploited in order to get the theoretical speed-ups of GPUs over CPUs. Problems that require looping or conditional branching are not well suited for GPGPU.

To evaluate the CKP metric (Eq. 3.4), one requires the  $N(N - 1)/2$  interactions between the  $N$  samples of the time series<sup>8</sup>. The CKP can be computed efficiently by mapping each of these interactions to a single GPU thread. The final value of the CKP is obtained through a  $\log(N)$ -step sum reduction performed on the GPU. The pseudocode to compute the CKP using GPU is given in Section 3.3.8.

## Brief review of the NVIDIA GPU architecture

CUDA uses a massive parallelism programming model, where each of the SIMD (single instruction, multiple data) processors on the GPU executes the same instruction over different data elements in parallel. The GPU can be viewed as a co-processor for the CPU. A function that is executed by the GPU threads is called a kernel. The CUDA programming model [105] is shown in Fig. 3.9a. The smallest unit of the programming model is the thread. Threads are packed into blocks, and blocks are packed into a grid. Before executing a CUDA kernel one has to define the size of the grid and blocks. The number of threads can surpass the number of actual physical execution units in the GPU. Synchronized stops can be used to share data and results between threads.

Each thread has a unique index which can be used to access a particular memory bin. The memory model of the GPU [105] is shown in Fig. 3.9b. The global memory is a high latency memory that can be accessed by any thread at any moment and is also the only memory which accept transfers from the host memory. The shared memory is a low latency

---

<sup>8</sup>The kernel matrices given by Eq. (2.26) and Eq. (2.31) are symmetric, thus only the upper triangular part needs to be computed. The diagonal of the kernel matrices is constant and is omitted from the computations.

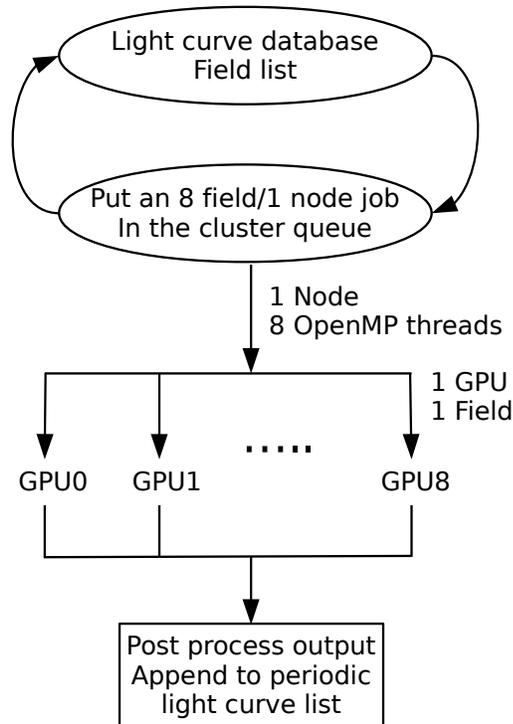


Figure 3.8: Flow diagram of parallel computation of independent fields, using a GPU cluster.

memory that can be accessed by threads of the same block. Register memory is exclusive for each thread. Constant memory is designed for broadcasting the same data to all the threads.

### Pseudocode to compute CKP in GPU and GPU optimization highlights

In this Section, a descriptive pseudocode of the GPGPU implementation of the CKP is given. GPGPU optimizations targeting the NVIDIA Fermi GPU architecture are highlighted.

- 1: **for** each EROS-2 light curve: **do**
- 2:   Transfer magnitudes and time instants of samples to the GPU global memory
- 3:   Transfer the kernel sizes for magnitudes and time to the GPU constant memory
- 4:   Generate two arrays containing the indexes of pairs of samples used to compute the Gaussian and periodic kernels, and transfer them to global memory. Sort the indexes taking into account that the kernel matrices are symmetric and their diagonals are constant.
- 5:   Compute the Gaussian kernel in the GPU, and save an array of size  $N \cdot (N - 1)/2$
- 6:   Reduce the Gaussian kernel in the GPU. Each block of 512 threads reduces a matrix section. The reduction is made on powers of two, i.e. from 512 to 256, 256 to 128, etc. The reduced values are saved in the GPU shared memory. At the end of this computation, the first block element contains the overall sum of the block. Accumulate the reductions of all blocks. The final reduction ( $R$ ) is used to compute the information potential:  $IP = (N + 2 \cdot R)/N^2$
- 7:   Subtract IP from the Gaussian kernel matrix to normalize it

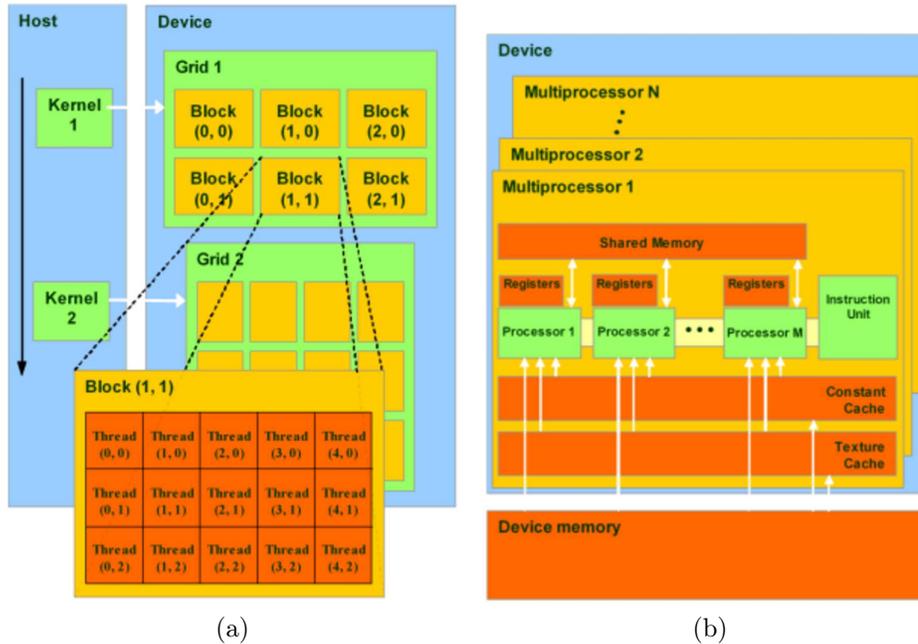


Figure 3.9: The CUDA programming model (a), and the GPU memory model (b). Taken from [105].

```

8:   for every period trial: do
9:     Compute the periodic kernel matrix in the GPU and save it in an array of size
       $N \cdot (N - 1)/2$ 
10:    Carry out a point-wise multiplication with the Gaussian kernel array
11:    Reduce the obtained array in the GPU using the same algorithm utilized for reducing
      the information potential
12:    Transfer back the reduction result to CPU memory. A CKP value is obtained for
      each trial period
13:  end for
14:  Output the periods associated with the 10 highest CKP values
15: end for

```

The following GPU optimizations are considered in this implementation:

1. Efficient access to global memory. The Fermi NVIDIA GPU architecture [107] introduced a cache line for the global memory. Readings to global memory can be optimized by choosing an access pattern that makes use of the cache. For example, within a thread block it is better to calculate  $G_\sigma(x_1 - x_2)$  and  $G_\sigma(x_1 - x_3)$ , because in this way, the first element  $x_1$  stays in the cache, and it is requested only once.
2. Use of lower-latency memory. Accesses to the GPU shared memory (on-chip) are a hundred times faster than accesses to global memory (off-chip). But the shared memory has many restrictions. The shared memory scope is the thread block, meaning that it can not be used to share data between blocks. If two threads read the same shared memory bin, the accesses are serialized. Shared memory is used to save the temporal sums when reducing the kernel matrices. We have followed the NVIDIA guidelines for

efficient reduction using shared memory.

3. Use of low latency GPU constant memory for parameters that are broadcasted to all the threads.
4. Use of GPU streams to overlap transfers to GPU global memory with GPU computation.

### 3.4. Correntropy NMF Spectrum

The CKP has better periodicity discrimination capabilities than conventional methods [101], but is in general noisy, presenting a high presence of harmonics, sub-harmonics and aliases of the underlying period in the periodogram. In this section we propose a method to obtain high-resolution frequency decomposition of the correntropy function based on Non-negative Matrix Factorization (NMF; Section 2.8). These frequency decompositions are sparse, localized on the fundamental frequency of the signal and are also less affected by noise and aliasing.

The correntropy function is decomposed into a dictionary of adaptive frequency atoms built using the periodic kernel function (Eq. 2.31). NMF is used to obtain a part-based and sparse decomposition of the correntropy function into the basis functions. The main difference with the standard NMF procedure is that the dictionary matrix is constrained to maintain a frequency-indexed structure, *i.e.* the atoms are always periodic functions. Then, the NMF coefficients can be treated as a generalized periodogram of the correntropy function. With respect to conventional Fourier representations, this new spectral decomposition is sparser, more localized (less harmonic content), less affected by noise, and able to reach a higher resolution in frequency space (superresolution). The conventional NMF algorithms were presented in Section 2.8, in what follows the modifications required to obtain sparse spectral decompositions for the correntropy function using NMF are presented.

In this case the data matrix corresponds to the correntropy function, calculated from Eq. (2.68) (or Eq. 3.2 in the case of unevenly sampled processes),  $V \in \mathbb{R}^{M \times 1}$  for all  $M$  lags. Two dictionaries are used. The first one,  $W_P \in \mathbb{R}^{M \times K}$ , is a pure-frequency dictionary with each element being defined as

$$w_P(n, k) = \exp\left(-\frac{2 \sin^2(\pi f_k \tau_n)}{\sigma_k^2}\right), \quad (3.13)$$

where  $\sigma_k$  and  $f_k$  are the kernel size and frequency of the  $k$ -th atom, respectively and  $\tau_n$  is the  $n$ -th lag of the correntropy function. An overcomplete frequency dictionary is considered (Sec. 2.7.1). The frequency grid for the dictionary is defined as in Eq. (2.40)<sup>9</sup>, with an overcompleteness parameter  $L > 1$ . The frequency dictionary atoms (Eq. 3.13), which are based on the periodic kernel function (Sec. 2.6.3), are non-negative, even and have a free parameter  $\sigma_k$ . An odd frequency dictionary is not required as correntropy is a symmetric

---

<sup>9</sup>The number of samples and time span still refer to the input signal, *i.e.* the signal from which correntropy is calculated.

function of the lag. The kernel size of the atoms plays a crucial role in the adaptation of the dictionary, as shown in the following section.

The second dictionary,  $W_K \in \mathbb{R}^{M \times M}$  is a Kronecker delta dictionary where each element is defined as

$$w_K(i, j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j. \end{cases} \quad (3.14)$$

The Kronecker delta dictionary is included to compensate for impulsive noise [76] that may be present in the correntropy function. Fitting the noise using the pure-frequency dictionary would reduce the sparseness of the representations as several periodic components would be required to model it.

Eq. (2.46) can be restated for our particular case as follows

$$V \approx \hat{V} = [W_P, W_K][H_P^T, H_K^T]^T, \quad (3.15)$$

where  $H_P$  and  $H_K$ , are the coefficients associated to the pure-frequency and Kronecker delta dictionaries, respectively, and  $[\cdot, \cdot]$  is the horizontal concatenation operator. The NMF frequency domain representation of correntropy corresponds to

$$S(f_k) = H_P(k), \quad (3.16)$$

and it will be referred to as the correntropy non-negative matrix factorization spectrum (CNMFS). The  $H_K$  coefficients are discarded because they are associated to the impulsive noise.

To find the CNMFS an ANLS scheme as described in Section 2.8 is used. In the **sparse coding** step, the coefficients of the representation are updated using the multiplicative rule for the Frobenius norm (Eq. 2.48). Compared with the conventional NMF literature, no major changes are required for the sparse coding step.

In the **dictionary learning** step, the dictionary is updated without breaking the frequency structure, which is a major departure from the NMF literature. In the original NMF framework, the dictionary learning step imposes no constraint to the dictionary matrix, except for the non-negativity. This means that performing updates to the dictionary using either the multiplicative or gradient descent based rules will not preserve the frequency structure of the dictionary. If the goal is to build frequency representations, it is necessary to preserve the frequency indexing of the atoms.

A frequency-indexed dictionary based on the periodic kernel function (Eq. 3.13) is used. The atoms in this dictionary are characterized by their frequency and kernel size. The kernel size dictates the shape of the atom, all belonging to the Gaussian family ranging from a pulse train to a smooth sinusoidal like shape, as shown in Figure 3.10. The kernel sizes can be viewed as knobs that may be used to adjust the shape of the atoms. This adjustment does not modify the frequency indexing of the atoms.

We propose using a dictionary learning step where adaptation is performed on the kernel sizes of the atoms rather than on the atoms themselves. By doing so, the frequency meaning

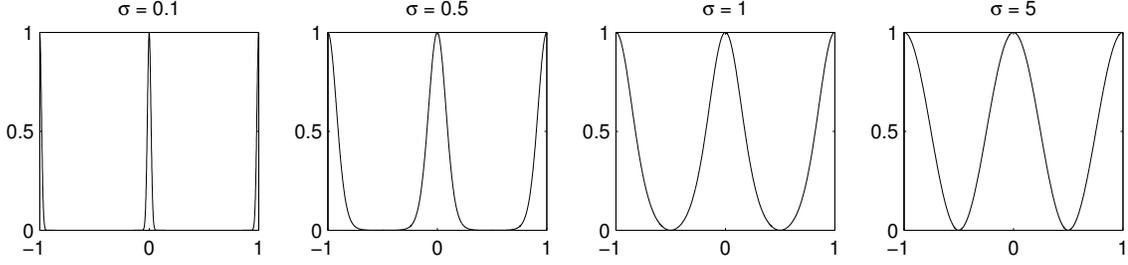


Figure 3.10: Example of a frequency-indexed atom built using Eq. (??). The kernel size dictates the shape of the atom, which can be varied from a pulse train to a sine wave.

of the atoms is preserved and at the same time the fitting that is being performed to the correntropy function is improved. As a consequence, sparser and more localized representations are obtained.

To adapt the kernel sizes of the pure-frequency dictionary a simple additive gradient-descent scheme is used. The gradient of the kernel size of the  $k$ -th atom of  $W_1$  is given by

$$\nabla_k = \frac{4h_1(k)}{\sigma_k^3} \sum_{m=1}^M \varepsilon_m w_1(m, k) \sin^2(\pi \tau_m f_k), \quad (3.17)$$

with

$$\varepsilon_m = \sum_{i=1}^K h_1(i) w_1(m, i) + \sum_{j=1}^M h_2(j) w_2(m, j) - v_m, \quad (3.18)$$

where  $v_m$  is the correntropy value at lag  $\tau_m$ . The kernel size is then updated as

$$\sigma_k^{t+1} - \sigma_k^t = \Delta \sigma_k^{t+1} = \alpha \Delta \sigma_k^t - \mu \nabla_k, \quad (3.19)$$

where  $\alpha$  is the momentum rate and  $\mu$  is the learning rate. After a kernel size update, the atoms are normalized to the range  $[0, 1]$ .

It is necessary to define a starting point for the dictionary and coefficient matrices. Usually they are filled randomly, or by feeding a solution from another method. One of the weaknesses of the multiplicative method for NMF is that the local minimum obtained has a strong dependence on the starting point. A bad starting point can yield a very poor local minimum.

Here we propose to define the starting point of the dictionary by the kernel size used to estimate the correntropy function for all the atoms in the pure-frequency dictionary. The Kronecker delta dictionary, being parameter-free, does not require initialization. The starting point for the frequency coefficients is given by

$$H^0 = [W_1^0, W_2^0]^T V. \quad (3.20)$$

The elements of the dictionary matrix are defined in  $[0, 1]$ . In order to facilitate the fitting, the autocorrentropy function is normalized by subtracting its minimum value, so that it also starts at zero.

Now that the method has been outlined let's consider a simple example using a 4 Hz sine wave with a duration of 10 seconds. Figures 3.11a and 3.11c show the autocorrelation and autocorrentropy function of this signal, respectively. A maximum lag of 1 second is used. Eq. (2.27) is used to estimate the kernel size of correntropy. The correlation function of a sine-wave is a sine-wave having the period of the signal (0.25 seconds). On the other hand the autocorrentropy function resembles a pulse train being much peakier at the corresponding period (Fig. 3.11c). It is clear that in the time domain the autocorrentropy function emphasizes much better the periodicity of the signal.

Now let's consider the frequency domain representations of the correlation and correntropy functions. Fig. 3.11b shows the PSD obtained by taking the discrete Fourier transform (DFT) of the correlation function. As expected, the PSD exhibits a single peak located at 4Hz. The mid-amplitude width of the peak is 1Hz, which is equivalent to one over the maximum lag. Fig. 3.11d shows the CSD computed using Eq. (2.70). In this case a large amount of frequency harmonics arise in the spectrum. There is no difference in terms of frequency resolution with respect to the PSD. Computing the CSD is equivalent to decomposing the correntropy function in a sine-wave basis. A pulse-train like signal will need a large amount of harmonics to be represented in this basis.

In this simple noiseless example, there is no advantage in using correntropy over correlation for frequency domain analysis (Fourier). In fact the large amount of harmonics encountered may hinder periodicity estimation applications. This situation was already mentioned in [96], where the correntropy function was used for pitch tracking on speech signals. In this application a time-domain method was preferred over a frequency-domain method due to the excessive harmonic content of the CSD [96].

Now let's study the CNMFS for the same synthetic time series. Figures 3.11e and 3.11f show the reconstruction of correntropy and the CNMFS for the 4 Hz sinewave, respectively. These results are obtained after 50 iterations, using  $\mu = 0.02$ ,  $\alpha = 0.5$  and  $L = 10$ . In comparison to the CSD (Fig. 3.11d), the CNMFS is highly concentrated on the true periodicity of the signal.

By increasing the dimensionality of the frequency dictionary, beyond the number of samples of the signal, overcomplete representations are obtained in an approach similar to basis pursuit. When using overcomplete dictionaries, the solution to the decomposition may not be unique anymore, which is a shortcoming for spectral estimation. In contrast, what one gains is the ability to super-resolve, *i.e.* the frequency resolution can be increased with respect to methods based on complete dictionaries. By adapting the frequency atoms the harmonic content in the representation is reduced. But the adaptation of the kernel size has also an effect on the frequency resolution of the spectral peaks, as illustrated in the following example.

Fig. 3.12 shows the width of the peak located at the fundamental frequency after 50 iterations, as a function of the kernel size normalized by its optimal value. The width of the peak is measured at mid-amplitude. The optimal kernel size value is found using the dictionary adaptation procedure. The curves correspond to dictionaries with different  $L$  values. The time series of the previous example is used and here the kernel size of the dictionary remains fixed, since it will be scanned in the plots. The Fourier's frequency

resolution corresponds to 0.5 Hz, the inverse of the record length (2s). For the CNMFS, increasing the overcompleteness decreases the width of the spectral peak as expected. When  $L = 1$  the resolution is equivalent to Fourier methods. For  $L < 5$  the frequency resolution appears to be constant. However, this behavior changes when the frequency spacing is increased further revealing a clear dependency between the kernel size and the frequency resolution. The width of the peak is at its minimum when the kernel size approximates its optimal value. In practice, the method finds this optimal value in the dictionary learning step. Intuitively, when the atoms are adapted to the correntropy function less neighboring atoms will be needed to characterize the decomposition. The frequency resolution reaches a limit value for  $L > 10$ . This contributes to the enhanced spectral localization of the CNMFS as seen in the experimental section.

### 3.4.1. Procedure for period estimation using the CNMFS

In this section a procedure for period estimation based on the CNMFS is presented. The fundamental period of a stationary time series is estimated by applying the following steps:

1. **Pre-processing:** The time series is normalized by subtracting its mean and then dividing it by its standard deviation.
2. **Correntropy:** The correntropy function of the time series is estimated using Eq. (2.68). If the time series is unevenly sampled, the slotted correntropy (Eq. 3.2) is used instead. The maximum lag is set to a 10% of the total time span of the time series. The kernel size is estimated using Eq. (2.27).
3. **Frequency domain representation:** The CNMFS is obtained as follows
  - I Initialize dictionary and coefficients as described in Section 3.4.
  - II Verify stopping criteria
  - III Update the coefficients vector using Eq. (2.51).
  - IV Update the dictionary matrix using Eq. (3.17) and Eq. (3.19).
  - V Go back to step ii.
4. **Fundamental frequency estimation:** Find the maximum of  $H_1$ , the inverse of the associated frequency is saved as the fundamental period.

The NMF routine stops if the objective function at instant  $t$  is below a given threshold

$$D_F(V|| (WH)^t) = D_F^t < \varepsilon_1, \quad (3.21)$$

and the improvement between iterations is negligible

$$\frac{|D_F^t - D_F^{t-1}|}{D_F^t} < \varepsilon_2. \quad (3.22)$$

The thresholds  $\varepsilon_1$  and  $\varepsilon_2$  should be set accordingly to the data. In our case these values were set to  $\varepsilon_1 = 0.01$  and  $\varepsilon_2 = 0.001$ . These values were obtained after extensive testing on synthetic data. This simple stopping condition reflects the symptoms of reaching a stationary point. This condition is referred to as the naive stationarity condition.

Additionally, the routine is also set to stop after a predefined number of iterations have been reached. This condition is added as a safeguard for cases where the time series could not be modeled by the dictionary. In our experiments the maximum number of iterations is set to 200, high enough so that in practice the routine only stops due to the first condition.

Other parameters to be considered are  $L$ ,  $\mu$  and  $\alpha$ . The overcompleteness of the dictionary  $L$  should be set accordingly to the desired frequency resolution. Bear in mind that  $L$  controls the dimensionality of the dictionary which greatly influences the computational complexity of the pipeline. The learning rate  $\mu$  and the momentum rate  $\alpha$  are set accordingly to the data in order to avoid instabilities during the learning process without sacrificing convergence speed.

### 3.5. Databases

The methods described in this thesis were tested on the MACHO [1] and EROS-2 [108, 2] astronomical survey databases. Both projects surveyed the Magellanic clouds. The Large and Small Magellanic clouds are irregular dwarf galaxies that belong to the local group along with the Milky Way and the Andromeda Galaxy. In this section these surveys are described.

The MACHO project [1] was designed to search for gravitational microlensing events in the Magellanic Clouds and the galactic bulge. The project started in 1992 and concluded in 1999. More than 20 million stars were surveyed. The MACHO project has been an important source for finding variable stars. The complete light curve database is available through the MACHO project's website<sup>10</sup>. There are two light curves per stellar object: channels blue and red. Each light-curve has approximately 1000 samples and contains 3 data columns: time, magnitude and an error estimation for the magnitude. Astronomers from the Harvard Time Series Center (TSC) build a catalog of periodic variable stars from the MACHO survey. The underlying periods of the periodic variable stars were estimated using epoch folding, AoV, and visual inspection.

The EROS-2 project [108, 2]<sup>11</sup> was designed to search for gravitational microlensing events caused by massive compact halo objects (MACHOs) in the halo of the Milky Way. To do this, 32.8 million stars in the Magellanic clouds were surveyed over 6.7 years. The objective of the EROS-2 survey was to test the hypothesis that MACHOs were a major component of the dark matter present in the Halo of our galaxy. The EROS-2 project surveyed 28.8 million stars in the Large Magellanic Cloud (LMC) and 4 million stars in the Small Magellanic Cloud (SMC), distributed in 88 and 10 observational fields, respectively. Each field is divided in 32 chips (8 CCDs and 4 quadrants per CCD). Each light curve file has 5 columns: time instant,

---

<sup>10</sup><http://wwwmacho.anu.edu.au/>

<sup>11</sup><http://eros.in2p3.fr/>

red channel magnitude, red channel error bars, blue channel magnitude and blue channel error bars. The average number of samples per lightcurve is 430 and 780 in the LMC and SMC, respectively.

The OGLE-3 project [109] is also a microlensing survey aimed at the Magellanic clouds. A catalog of variable stars found by the OGLE collaboration is available to the public<sup>12</sup>. The catalogs of periodic variable stars are used to compare results and perform crossmatching analysis.

### 3.6. Performance criteria

In what follows we define performance measures for the problems of periodic light curve discrimination and period estimation. The former consists of discriminating between periodic and non periodic light curves. The latter consists in estimating the true periods of periodic light curves. For the period estimation problem the classification is done by using the TSC periods as golden standard. An estimated period  $P_{est}$  is classified as either a Hit, a Multiple or a Miss with respect to the reference period  $P_{ref}$  according to the following criteria:

- Hit if  $|P_{ref} - P_{est}| < \varepsilon \cdot P_{ref}$
- Multiple if  $P_{est} > P_{ref}$  and

$$\left| \frac{P_{est}}{P_{ref}} - \left[ \frac{P_{est}}{P_{ref}} \right] \right| < \varepsilon,$$

or if  $P_{est} < P_{ref}$  and

$$\left| \frac{P_{ref}}{P_{est}} - \left[ \frac{P_{ref}}{P_{est}} \right] \right| < \varepsilon,$$

where  $[x]$  is the largest integer less than or equal  $x$ .

- Miss if it does not belong to any of the other categories.

The tolerance value  $\varepsilon$  controls the accepted relative error between the estimated period and the reference period. A value of  $\varepsilon = 0.005$ , *i.e.* a relative error of 0.5% will be considered, small enough to obtain a clean folded curve from the estimated period.

The task of discriminating periodic light curves can be viewed as a binary classification problem where the classes are periodic (true) and non-periodic (false) light curves. In this case: true positives (TP) are the periodic light curves classified as periodic, false positive (FP) are the non-periodic lightcurves classified as periodic, true negative (TN) are the non-periodic light curves classified as non-periodic and false negative (FN) are the periodic light curves classified as non-periodic. Confusion matrix and Receiver Operating Characteristic (ROC) curves are used to evaluate the periodicity discrimination method. An ROC curve is a plot of the true positive rate (TPR) as a function of the false positive rate (FPR). Different points in the ROC curve are obtained by changing the threshold value at the output of the classifier. The TPR represents the proportion of periodic light curves that are correctly identified as such. The FPR represents the proportion of non-periodic light curves that are

<sup>12</sup><http://ogledb.astrouw.edu.pl/~ogle/CVS/>

incorrectly classified as periodic. To evaluate the performance of our method we also use the definitions of recall,  $r$ , precision  $p$

$$r = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad p = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (3.23)$$

and F-score

$$F_\beta = \frac{(1 + \beta^2)pr}{\beta p + r}. \quad (3.24)$$

The denominator of  $r$  in Eq. (3.23) corresponds to the number of periodic lightcurves in the dataset. Recall, or completeness, is the ratio of recovered periodic lightcurves over the total number of periodic lightcurves in the dataset. The denominator of  $p$  in Eq. (3.23) corresponds to the number of lightcurves that are classified as periodic. Precision or completeness, is the ratio of recovered periodic lightcurves over the total amount of lightcurves that are classified as periodic. Note that recall is equivalent to the TPR. The F-score (Eq. 3.24) is a weighted average of recall and precision. The parameter  $\beta$  controls the importance of recall over precision on the weighted average. In what follows we use the  $F_1$  score ( $\beta = 1$ ).

We also define hit rate as:

$$HR = \frac{\text{TP}^*}{\text{TP}^* + \text{FN}}, \quad (3.25)$$

where  $\text{TP}^*$  are the periodic lightcurves classified as periodic and at the same time the true period is recovered<sup>13</sup>.

---

<sup>13</sup>Note that a light curve can be classified as periodic even if the true period is not recovered, such as when a multiple of the true period is found.

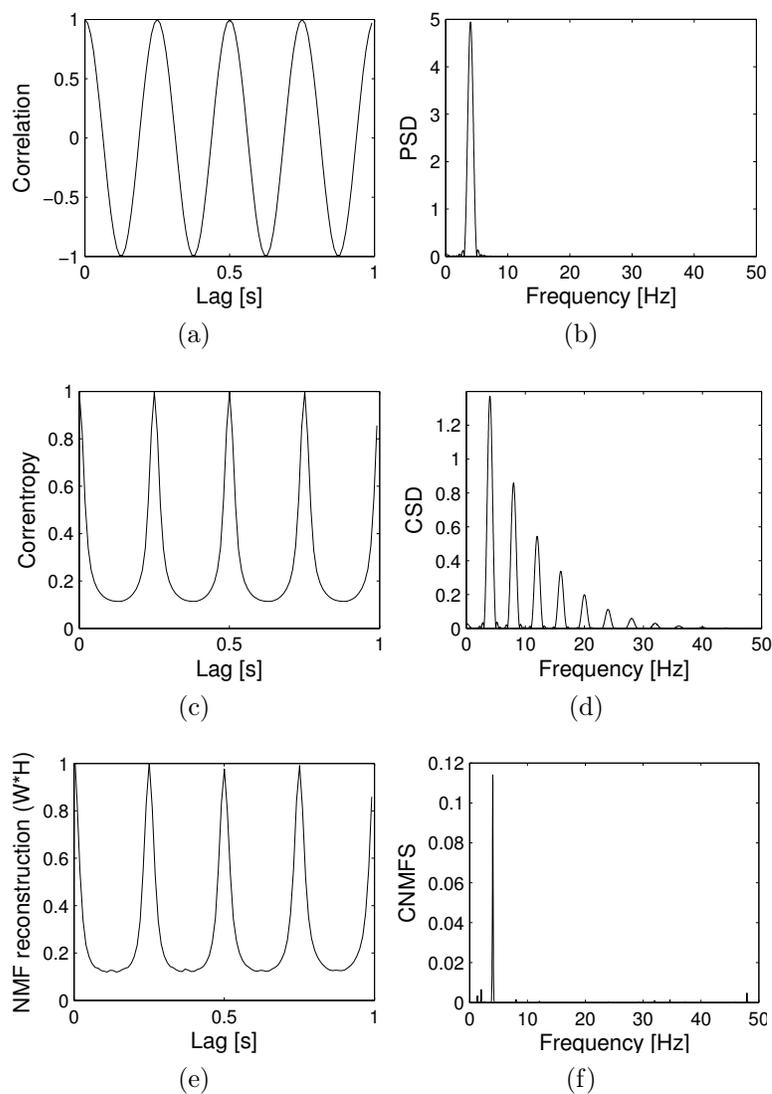


Figure 3.11: The autocorrelation (a) and autocorrentropy (c) functions of a 4 Hz sine wave. The correntropy emphasizes better the 0.25 s period of the signal. The PSD (b) and CSD (d) translate the autocorrelation and autocorrentropy to the frequency domain, respectively. Due to the pulse-train like waveform of correntropy a large amount of harmonics appear in the CSD. In this example, there is no clear advantage of correntropy over correlation when doing frequency domain analysis. (e) The reconstructed correntropy function and (f) the CNMFS for the 4 Hz sine wave used in the previous example. This result was obtained after 10 iterations, using  $\mu = 0.02$ ,  $\alpha = 0.5$  and  $L = 10$ . In comparison to the CSD, the CNMFS is highly concentrated on the true periodicity of the signal.

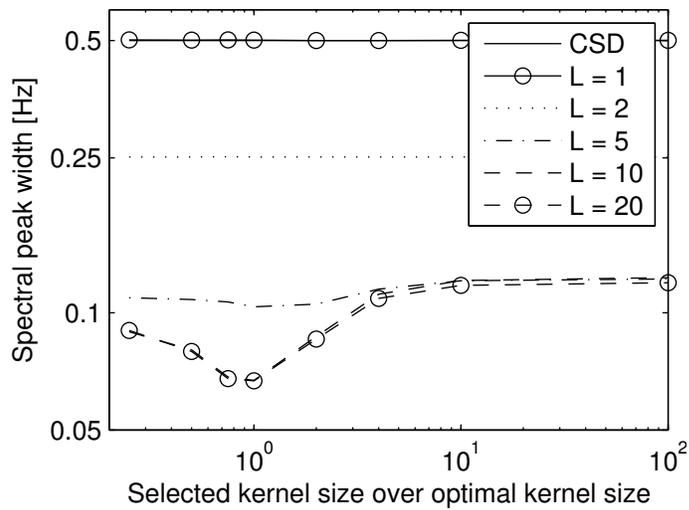


Figure 3.12: Frequency resolution as a function of the ratio between the kernel and its optimal value. For  $L = 1$  the resolution is equivalent to Fourier based methods. For  $L < 5$  resolution appears to be constant. When the spacing increases ( $L \geq 5$ ) a clear dependency between the kernel size of the dictionary and the frequency resolution is revealed. The best resolution is achieved when the kernel size is close to its optimal value. The proposed dictionary adaptation method reaches this value automatically.

# Chapter 4

## Results

### 4.1. Slotted autocorrentropy for period estimation on MACHO light curves

The slotted autocorrentropy is tested on a subset of 600 light curves from the MACHO survey. This subset contains 200 light curves from each of the following three types of variable stars: eclipsing binaries (EB), Cepheids and RR Lyrae, whose periods range from 0.2 days to 200 days. The pipeline described in Section 3.1.2 is applied on this dataset. The performance of the period estimation pipeline was compared with widely used software solutions for astronomical time series analysis: VarTools [110] and SigSpec [111]. VarTools includes LS and AoV light curve analyses. SigSpec combines Fourier based methods with statistical metrics of spectral significance. The SLLK string length method [50] is also considered. For Vartools LS, the highest peak of the periodogram gives the estimated period. In the variant Vartools LS+IP, the IP metric is used to discriminate among the top 10 peaks obtained with the LS periodogram. Likewise, in SLLK+IP, the IP metric is used to discriminate among the top 10 shortest strings found by SLLK. For SLLK and Vartools AoV, the corresponding statistics are minimized in an array of periods ranging from 0.2 to 200 days with a step size of  $1e-4$ . For the slotted correlation+IP we compute the top 10 peaks of the PSD, and then apply the IP metric to discriminate among them.

Table 4.1 shows the results obtained with each of the described methods in a subset of 200 EB light curves. The slotted correntropy+IP method obtained the highest hit rate (74%), followed by SLLK+IP (65%), and slotted correlation+IP (50%). Table 4.2 shows the results obtained in a subset of 400 cepheids and RR Lyrae light curves. The slotted correntropy+IP method, Vartools LS and Vartools AoV obtained the highest hit rate (97%), followed by SigSpec (95.5%). SLLK enhanced by the IP metric achieved 90.25%. The proposed method outperforms its competitors on EB period estimation and obtains the same performance in the case of the pulsating variables.

In the EB case, conventional methods based on Fourier analysis tend to find a harmonic

Table 4.1: Performance of the proposed period estimator (slotted correntropy+IP) versus conventional techniques in a subset of 200 EB light curves from the MACHO survey

Period estimation methods	Hits[%]	Multiples[%]	Misses[%]
<b>Slotted correntropy + IP</b>	74.0	25.5	0.5
Slotted correlation + IP	50.0	48.5	1.5
VarTools LS	11.0	89.0	0.0
VarTools LS + IP	18.0	82.0	0.0
VarTools AoV	39.5	60.5	0.0
SigSpec	11.0	88.5	0.5
SLLK	42.5	54.5	3.0
SLLK +IP	65.0	34.5	0.5

Table 4.2: Performance of the proposed period estimator (slotted correntropy+IP) versus conventional techniques in a subset of 400 cepheids and RRL light curves from the MACHO survey

Period estimation methods	Hits[%]	Multiples[%]	Misses[%]
<b>Slotted correntropy + IP</b>	97.00	2.75	0.25
Slotted correlation + IP	93.00	5.75	1.25
VarTools LS	97.00	2.75	0.25
VarTools LS + IP	97.00	2.75	0.25
VarTools AoV	97.00	2.75	0.25
SigSpec	95.50	4.25	0.25
SLLK	68.50	28.00	3.50
SLLK +IP	90.25	4.25	0.25

of the underlying period<sup>1</sup>. EB light curves exhibit a double-bump shape that is difficult to model using sine-waves, which explains why Fourier methods obtain integer submultiples of the true period. Additionally, if the components of the binary system are similar, the eclipses might not be clearly distinguishable, even for the expert astronomer. Correntropy and the IP metric are based on higher order statistics and make no assumption on the underlying structure of the data, which is reflected on the obtained results.

## 4.2. CKP for periodic light curve discrimination on the MACHO database

The CKP is tested on a subset of 1500 periodic light curves (500 Cepheids, 500 RR Lyrae and 500 eclipsing binaries) and 3500 non-periodic light curves drawn from the MACHO survey. The subset was divided into a training set for parameter adjustment and a testing set for comparison with other methods. The training set consisted of 2500 light curves (750 periodic and 1750 non periodic) randomly selected from the available classes. The remaining 2500 light curves were used for testing purposes. Note that there is a natural imbalance

<sup>1</sup>In most cases the closest integer submultiple, *i.e.* half of the true period

Table 4.3: Statistical significance thresholds of the CKP.

$1 - \alpha$	$\widehat{P}_\alpha$	Error bar
0.99	3.59e - 4	6.09e - 6
0.95	3.12e - 4	5.66e - 6
0.90	2.80e - 4	5.58e - 6

between periodic and aperiodic classes of stars. Only 3% of the surveyed stars are expected to be variables and 1% to be periodic. Due to this, when detecting periodic behaviour, we have to achieve a false positive rate less than 0.1%.

**Statistical significance test:** Using the procedure described in Section 3.2.1 statistical significance thresholds for the CKP were computed. Table 4.3 shows the significance thresholds and their corresponding CKP ordinate values for the best combination of kernel sizes (more details on [101]). The thresholds were computed using the light curve training set ( $N = 2500$ ) and five hundred surrogates per light curve ( $M = 500$ ). Fig. 4.1 shows the location of these thresholds in the ROC curve of the testing dataset. FPR rates below 1% are associated with confidence levels between 95% and 99%.

Fig. 4.2 shows two light curves in which the CKP ordinate value associated with the fundamental period has a confidence level higher than 99%. In the folded light curves the periodic nature of the light curve can be clearly observed. In period detection schemes based on visual inspection these light curves would be undoubtedly labeled as periodic. Fig. 4.3 shows two light curves in which the CKP ordinate values associated with the fundamental period have a statistical confidence between 90% and 95%. These light curves are indeed periodic although compared to the previous two (Fig. 4.2), their periodicity is less clear as their signal to noise ratio is smaller.

By associating a statistical level of confidence to the detected periods, we have obtained a way to set a period detection threshold and also a better interpretation of period quality. The level of confidence on the detected period could be used in later (post-processing) stages of the period detection pipeline. For example, periods with lower confidence levels may be selected for additional analysis stages in which finer resolution or different parameter combinations may be used.

**Comparison with other methods:** The performance of the CKP method was compared with the slotted correntropy and other widely used techniques in astronomy. The software VarTools [110, 112] was used to perform a Lomb-Scargle periodogram and Analysis of Variance (AoV) analysis. The regularized Lafler-Kinman string length (SLLK) statistic and the slotted autocorrelation were also considered. For Vartools LS, the period associated with the highest peak of the LS periodogram, that is not a multiple of the known spurious periods (sidereal day, moon phase, etc.), gives the estimated period. A periodogram resolution of  $0.1/T$  and a fine tune resolution of  $0.01/T$ , where  $T$  is the total time span of the light curve, were used. For Vartools AoV and SLLK, the corresponding statistics are minimized in an array of periods ranging from 0.4 to 300 days with a step size of  $1e-4$ . For AoV the default value of 8 bins is used. For AoV and SLLK, the period that minimizes the corresponding metrics, that is not a multiple of the known spurious periods, is selected

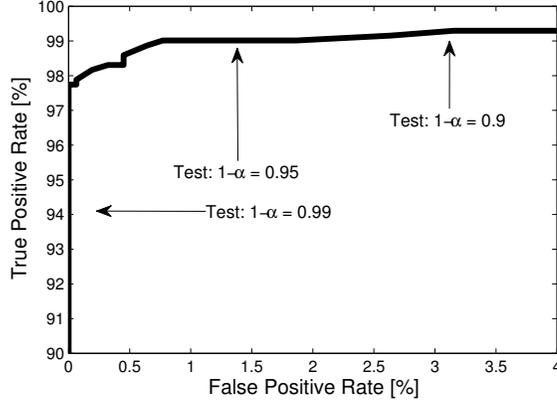


Figure 4.1: ROC curve of the CKP in the training and test subsets. The significance thresholds of the CKP are shown in the ROC curves.

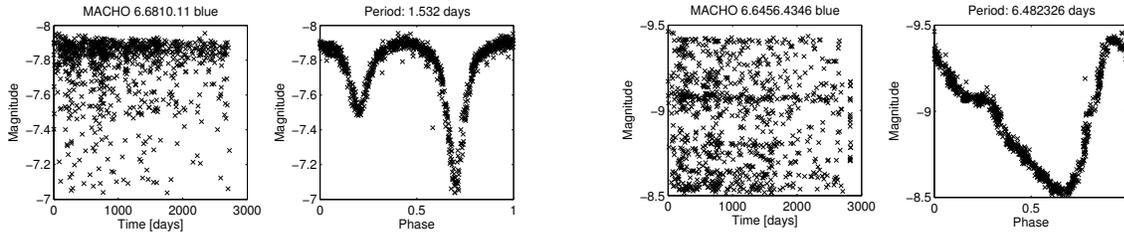


Figure 4.2: Examples of periodic light curves detected by the CKP method with a level of confidence greater than 99%. The original light curves are shown on the left column. The right column shows the same light curves folded with the estimated period.

as the estimated period. For the slotted autocorrelation/autocorrentropy the highest peak of the PSD/CSD estimator function, that is not associated to the known spurious periods, delivers the estimated period. For the slotted autocorrentropy/autocorrelation a slot size  $\Delta\tau = 0.25$  was considered [100]. For the CKP method the best combination of kernel sizes ( $\sigma_t = 0.1$  and  $k = 1$ ) obtained with the training dataset is used [101]. The influence of the higher-order statistical moments is assessed by comparing the CKP with a linear version of the proposed metric. In this linear version, the Gaussian kernel used to compare magnitude values is replaced by a linear kernel. The periodic kernel remains unchanged. The procedure to detect a period is the same as explained in Section 3.2.2, with an additional pre-processing step where the data vector is zero-mean normalized.

The results for period estimation on the testing subset are shown in Table 4.4. The CKP method obtained the highest hit rate (88%), followed by the linear version of the CKP (80%), the slotted correntropy (78%) and the AoV periodogram (75%). The CKP obtained 7.6% more hits and 57% less misses than its linear version. This is because the Gaussian kernel incorporates all the even-order moments of the process and gives robustness to outlier data samples which are common in astronomical time series. The CKP obtained 9.2% more hits and 72.7% less misses than the slotted correntropy. This is because in the slotted correntropy, kernel coefficients are averaged on time slots, therefore the actual time differences between samples are not used.

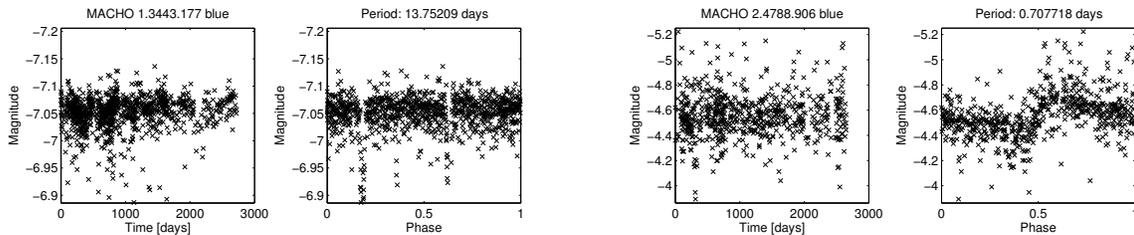


Figure 4.3: Examples of periodic light curves detected by the CKP method with a level of confidence between 90% and 95%. The original light curves are shown on the left column. The right column shows the same light curves folded with the estimated period.

The proposed method obtained the lowest miss rate (0.4%). In all these missed cases<sup>2</sup>, the true period was correctly found by the proposed metric but they were filtered out for being too near to the sidereal day. More accurate ways of filtering out spurious periods, using the data samples instead of a straightforward comparison of the detected periods, could be implemented to recover such missed cases.

Out of the cases where the correct period is found by the CKP but not by the AoV periodogram, an 84% corresponds to eclipsing binaries. This is expected as conventional methods perform well on pulsating variables [100]. Eclipsing binaries light curves are typically more difficult to analyze as their variations are non-sinusoidal and due to their morphology/shape characteristics most methods tend to return harmonics or sub-harmonics rather than the true period. The remaining 16% corresponds to light curves with low signal-to-noise ratio (SNR). It is important to consider that the MACHO subset light curves are well sampled, they represent only three classes of periodic variables and their periods were found using visual inspection. There are not many examples of periodicities found in low SNR regimes, which is clearly a bias caused by the human expert. In a real world scenario most light curves will have low SNR. For a correct characterization of the CKP, a larger and more heterogeneous set of light curves is required. This is addressed in the next section through a test using synthetic light curves and 32 million light curves from the EROS-2 survey.

Fig. 4.4 shows ROC curves for the task of periodic versus non-periodic discrimination in the 2500 light curve testing subset. The CKP is compared with the LS and AoV periodograms. The proposed method clearly outperforms its competitors in the FPR range of interest (below 1%). It is worth noting that even if a harmonic of the true period is found, periodicity can still be detected. This is true for all the methods as periodicity detection rates are comparatively better than Hit rates obtained for period estimation.

<sup>2</sup>Light curves 1.4539.37, 3.6605.124 and 6.5726.1276, with periods 2.9955 (three times sidereal day), 3.9813 (four times the sidereal day) and 0.99676, respectively.

Table 4.4: Period estimation performance of the CKP method versus other techniques for the testing database (2,500 light curves)

Method	Hits[%]	Multiples[%]	Misses[%]
<b>CKP</b>	<b>88.00</b>	<b>11.60</b>	<b>0.4</b>
CKP (linear kernel)	80.40	18.67	0.93
Slotted correntropy	78.80	19.73	1.47
Slotted correlation	70.00	28.80	1.20
VarTools LS	61.73	36.00	2.27
VarTools AoV	75.33	23.60	1.07
SLLK	71.47	26.27	2.27

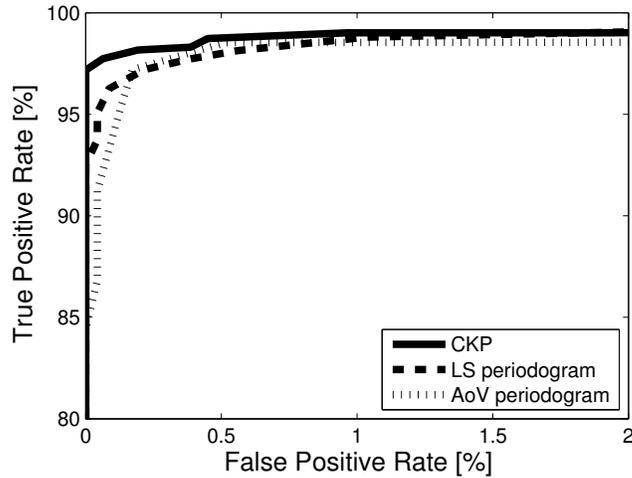


Figure 4.4: Receiver operating characteristic curves of the CKP, the LS periodogram and the AoV periodogram.

### 4.3. CKP for periodic light curve discrimination on the EROS-2 database

The pipeline presented in Section 3.3.7 is used to process 32.8 million light curves drawn from the EROS-2 survey. In this section results obtained on synthetic and real EROS-2 light curves are presented. The synthetic light curves are generated from real EROS-2 light curves using the procedure described in Appendix A. Using the synthetic set, where the parameters are known, the performance of the pipeline can be assessed. The synthetic set is also used as a training set to calibrate the parameters of the method. The pipeline is then used to process the EROS-2 light curves. This generates a list of periodic stars from the surveyed fields. The EROS-2 light curves have not been analyzed for periodicity detection before, hence there are no previous results to compare. Catalogues of periodic variable stars from the OGLE-3 survey, a project that surveyed the same area of the sky as EROS-2, are used to crossmatch and compare.

### 4.3.1. Efficiency in the synthetic light curve set

In the following tests we assess the efficiency of the periodicity detection pipeline as a function of the parameters of the synthetic lightcurves. The Hit rate (Eq. 3.25) is measured as a function of the total time span divided by the period ( $T/P$ ), number of samples ( $N$ ), smoothness ( $\sigma$ ), and SNR in a set of the 10,000 synthetic periodic light curves. Hit rates are computed as a function of one of the parameters at a time (summing for the other three). The CKP is compared with the LS periodogram on each test.

Fig. 4.5a shows a plot of the Hit Rate (HR) as a function of the ratio between the total time span of the light curve and its period ( $T/P$ ). The total time span of the light curves in EROS-2 survey is approximately 2500 days, and the sampling rate is approximately 1.2 samples per day. The ratio  $T/P$  can be viewed as the number of times that the underlying period repeats itself within the time series. The period range in the training set goes from 0.4 days to 1000 days. The HR is stable across the given range except for  $T/P$  below 10 and above 2300. Intuitively, the fewer times a signal is repeated across  $T$  the more difficult is to assess its periodicity. This can be seen in the plot for periods above 280 days. There is also a limit in the resolution associated to the sampling rate, which affects the detection of very short periods (fast frequencies). This is reflected as a HR drop for periods below 0.5 days. The same HR drops can be observed for the LS periodogram. The CKP achieves a higher HR than the LS periodogram for all the  $T/P$  values considered.

Fig. 4.5b shows a plot of HR as a function of the number of samples ( $N$ ) of the synthetic light curves. The HR increases with the  $N$ . The hit rate rises by 5% when the number of samples increases from 300 to 600. In comparison with the LS periodogram, the CKP is less affected by  $N$ . Intuitively, the less information available on the process the harder is to assess its periodicity.

Fig. 4.5c shows a plot of the HR as a function of the smoothness ( $\sigma$ ) of the synthetic light curves. The HR is stable across the given range, decreasing slowly for the very large and very small values of  $\sigma$ . Overall, the smoothness does not have great influence on the HR obtained by the CKP. The HR obtained with the LS periodogram increases with  $\sigma$ . This is expected, as smaller values of  $\sigma$  produce light curves with highly non-sinusoidal shapes (see Appendix A).

Finally, Fig. 4.5d shows a plot of the HR as a function of the SNR of the synthetic light curves. The HR is stable for the given SNR range, dropping abruptly for SNR below 1.8. For SNR of 1.2 hit rate has decreased by a almost 25%. A similar behaviour can be seen for the LS periodogram, although the HR obtained with this method is always lower than the one obtained with the CKP.

From these results, it is clear the CKP is more robust and less dependent on the number of samples, the shape of the periodicity (smoothness), the noise in the light curve and the value of the period. These advantages are explained by the higher-order moments acting in the CKP and the adaptiveness of the metric to different noise-regimes and light curve shapes.

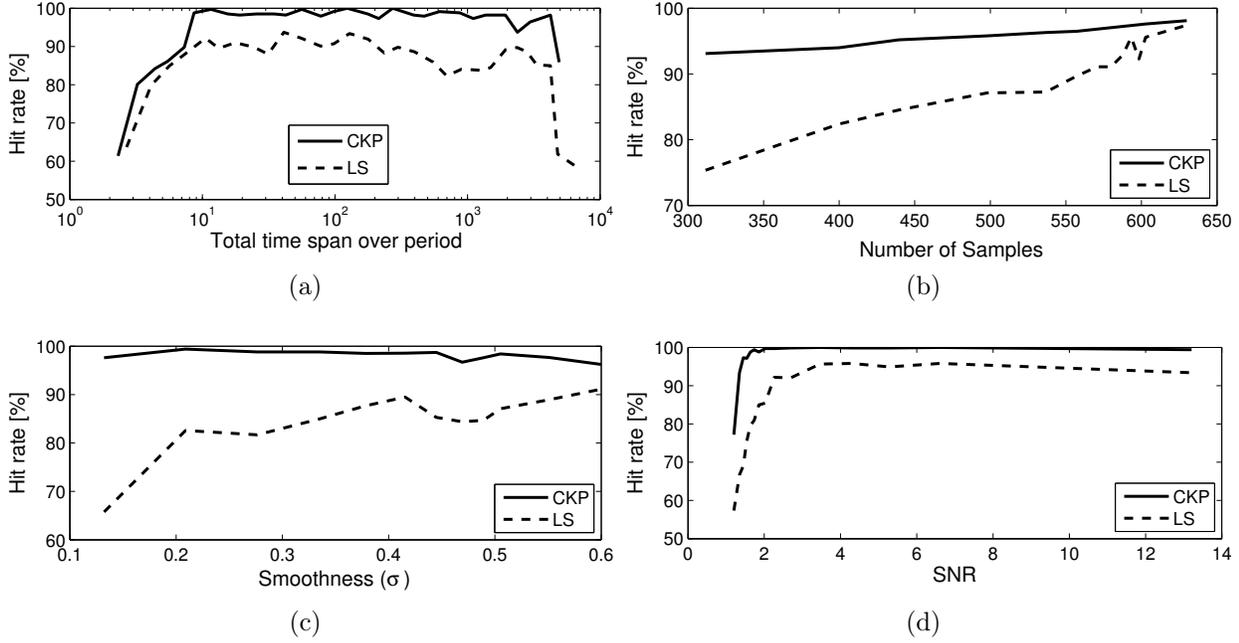


Figure 4.5: Hit rate in the synthetic periodic light curves as a function of the value of the parameters used to generate the dataset. The parameters are the period over total time span (a), number of samples (b), the smoothness (c) and SNR (d). The proposed method (solid line) is compared with the LS periodogram (dotted line). The CKP performs better than the LS periodogram in the four tests.

### 4.3.2. Results for selected EROS-2 fields

In this experiment the periodicity discrimination pipeline (Section 3.3.7) is evaluated on three fields from the EROS-2 survey. The objectives are to measure the accuracy of the method and to compare the number of periodic light curves in the fields with the expected number of periodic light curves computed from the synthetic results by performing visual inspection to a large but manageable number of light curves. The first six chips from fields lm009, lm012 and sm001 are used in this experiment. Table 4.5 shows the number of light curves, the average number of samples and the average SNR from the selected fields.

Table 4.6 shows the results obtained for the selected fields. Column two ( $\tilde{N}_p$ ) corresponds to the number of light curves labelled as periodic by our method. These light curves are folded with the detected period and visually checked in order to find the number of false positives (column three). Column four is the precision in the detected periodic light curves set. Column five gives an estimate of the false negatives (FN) in the field. The FNs are estimated by visually inspecting the folded light curves of the objects that are below the periodicity thresholds. Because it is impracticable to check all the non-periodic objects, the search for FNs is stopped if 50 consecutive non-periodic light curves are found for each SNR bin. Column six is the recall calculated using the observed number of true positives ( $\tilde{N}_p$  - FP) and the FN. Column seven corresponds to the observed number of periodic light curves ( $\tilde{N}_p$  - FP + FN). Column eight shows an estimation of the true number of periodic variables ( $N_p$ ) using the synthetic precision and recall values [102]. Column seven is also an estimation of

Table 4.5: Characteristics of selected fields.

Field	Number of lightcurves	Average N	Average SNR
lm009	109,802	548	1.628
lm012	95,010	447	0.959
sm001	92,666	830	1.505

Table 4.6: Results in the selected EROS-2 survey fields.

Field	$\tilde{N}_p$	FP	Prec. [%]	FN	Recall [%]	Observed $N_p$	Estimated $N_p$
lm009	1160	41	96.47	66	94.43	1185	1189
lm012	718	30	95.82	51	93.10	739	743
sm001	1564	69	95.59	99	93.79	1594	1637

$N_p$  because the true amount of FNs is not known.

A grand total of 1160 periodic lightcurves is recovered from field lm009, which corresponds to a 1.06% of the field. The percentage of periodics lightcurves in lm012 and sm001 is 0.75% and 1.69%, respectively<sup>3</sup>. The overall precision and recall in all the fields is within 2% of the overall precision and recall found in the synthetic dataset. For comparison we ran the Lomb-Scargle periodogram<sup>4</sup> on the lm009 field. The spurious periods are filtered as described in previous Sections. The filtered periods found with the LS periodogram are sorted according to their normalized LS statistic. By imposing a threshold on this statistic the same periodic light curves obtained by the CKP plus 298 false positives and 14 additional true positives are obtained. This corresponds to a drop of 16.5% in precision and a negligible increase in recall (1%) with respect to the CKP.

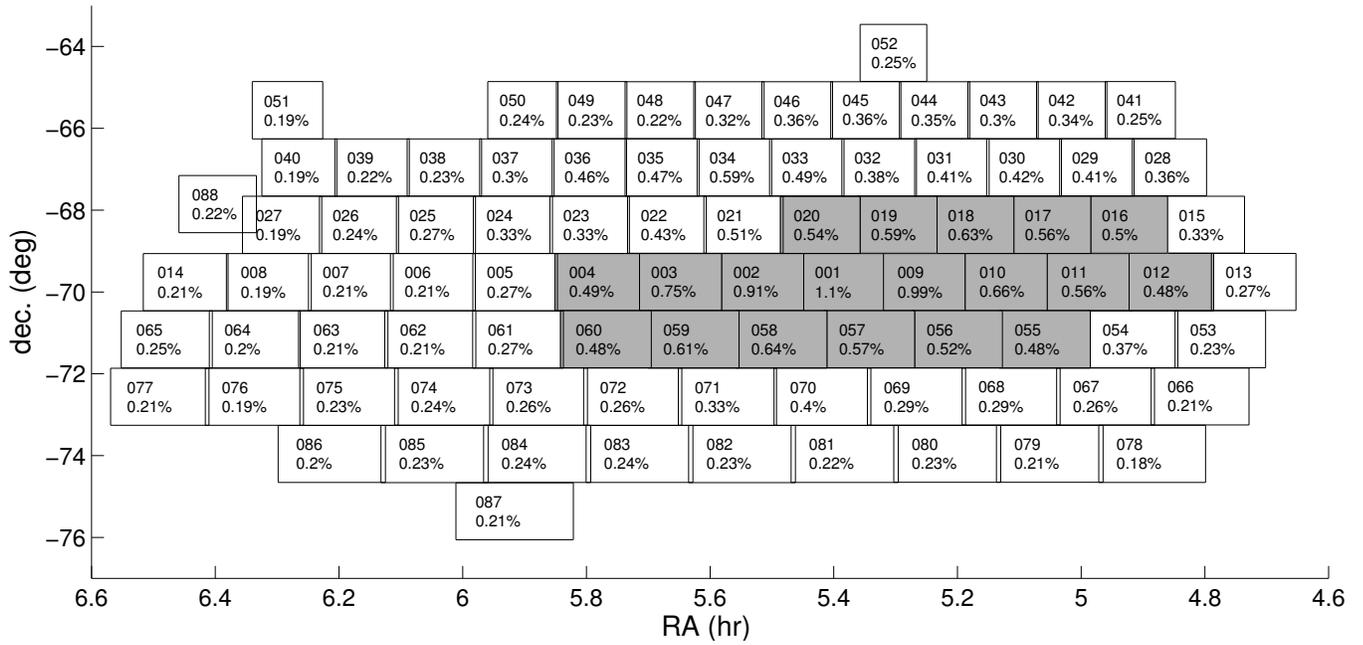
It is important to note that there are periodic behaviors that are not captured in the proposed synthetic light curve set. Examples of these are periodicities mounted on polynomial trends, objects with more than one oscillation period, objects that are not periodic in the whole time span and objects whose oscillations amplitude change irregularly or follow a modulation pattern, such as semi-regular and irregular LPVs. These cases are considered as non-periodic during the inspection. Although, the periodicity discrimination pipeline was not designed to discriminate quasi-periodicities and other irregular variable phenomena, some of them are detected.

### 4.3.3. Results on EROS-2 LMC and SMC fields

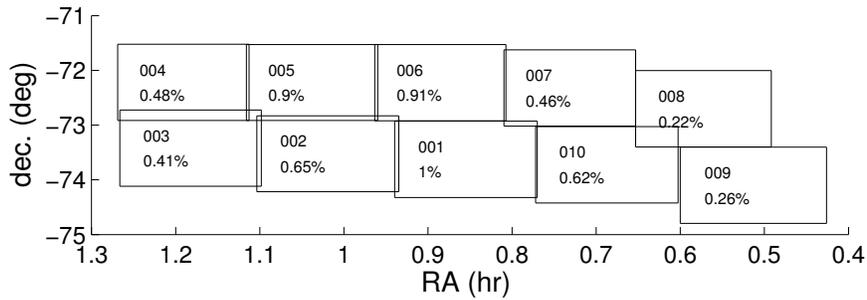
A total of 32.8 million light curves from the EROS-2 survey were processed with the CKP based periodicity discrimination pipeline, 28.8 million from the LMC and 4 million from the SMC. Table 4.7 shows the summary of the results for the LMC and SMC.  $\tilde{N}_p$  corresponds to the number of light curves labeled as periodic by the method. The *Discarded* column corresponds to the number of periodic light curves that appear twice in the list, due to field

<sup>3</sup>These chips have a higher number of periodics than the average found in the LMC and SMC as it can be seen in Fig. 4.6a. This issue is discussed in the next section.

<sup>4</sup>The vartools software with the -LS option is used.



(a)



(b)

Figure 4.6: Maps of the EROS-2 LMC (a) and SMC (b) fields, respectively. The greyed zone in the LMC corresponds to the fields associated with the LMC bar. The percentage of periodic light curves found by the CKP is shown below the number of each field.

overlapping and blending. Column  $N_p$  corresponds to an estimation of the true number of periodic variables using the synthetic precision and recall [102].

To select the “duplicate” light curves, the nearest neighbor for each object in terms of angular distances is firstly identified. If the distance to the nearest neighbor is less than 10 arcsec and both objects have the same period, then the light curve with the lowest magnitude is added to the discarded set. Using this criterion 2663 pairs of light curves are selected from the LMC. From this set 336 correspond to light curves that reside in different chips. The average delta magnitude in this set is 0.281 and the average delta CKP is 0.744. Each pair of light curves correspond to the same star which appears twice in the survey due to the overlapping in the observational fields. The other 2327 cases correspond to light curves that are neighbours in the same chip. The average delta magnitude in this set is 2.15 and the average delta CKP is 3.02, much higher than the previous set. In this set the more luminous star of the pair injects its periodicity in the light curve of the less luminous star (blending). Fig. 4.7 shows an example of an overlapped pair and a blended pair. It is interesting to note that 72% of the blended light curves are found in the fields within the LMC bar where the star density is the highest, while the overlapped light curves are equally distributed between bar and non-bar fields.

In the SMC 1817 pairs of light curves are selected to be discarded. In this case 386 are due to field overlapping and 1431 are due to blending. The average delta magnitude in the overlapped light curves is 0.21 and the average delta CKP is 0.78. The average delta magnitude in the blended light curves is 2.34 and the average delta CKP is 4.86. The percentage of discarded light curves in the SMC is 7.2% which is higher than the 2.3% found in the LMC.

Fig 4.6a shows a map of the 88 fields of the LMC. The shaded fields correspond to the LMC bar. The percentage of periodic light curves is shown for each field below its name. The fields corresponding to the LMC bar have a higher percentage of periodics. The percentage of periodics tends to drop the further the field is from the LMC bar. Fig 4.6b shows a map of the 10 fields of the SMC where the same pattern is apparent. Because the cores of the LMCs have older population of stars it is known that one would expect more periodic stars in those regions.

A grand total of 118,320 and 23,103 periodic light curves were found from the LMC and SMC blue channel data, respectively. Using the recall and precision from the training dataset the true number of periodic light curves is estimated to be 121,147 for LMC and 24,855 for the SMC. A 0.42% of the light curves of the LMC is periodic and a 0.61% of the light curves in the SMC is periodic.

Figures 4.8a shows the histogram of the periods found in the LMC blue channel data. Some of the known populations of periodic variables are identified in the histogram. The most notable populations correspond to c-type RR Lyrae (period centered in 0.3 days) and ab-type RR Lyrae (period centered in 0.6 days). These results are consistent with the RR Lyrae period histogram from the MACHO survey results on the LMC [113].

Fig. 4.9a shows a color magnitude diagram of the periodic light curves found in the LMC blue channel. The third axis corresponds to the detected period. The regions of interest are

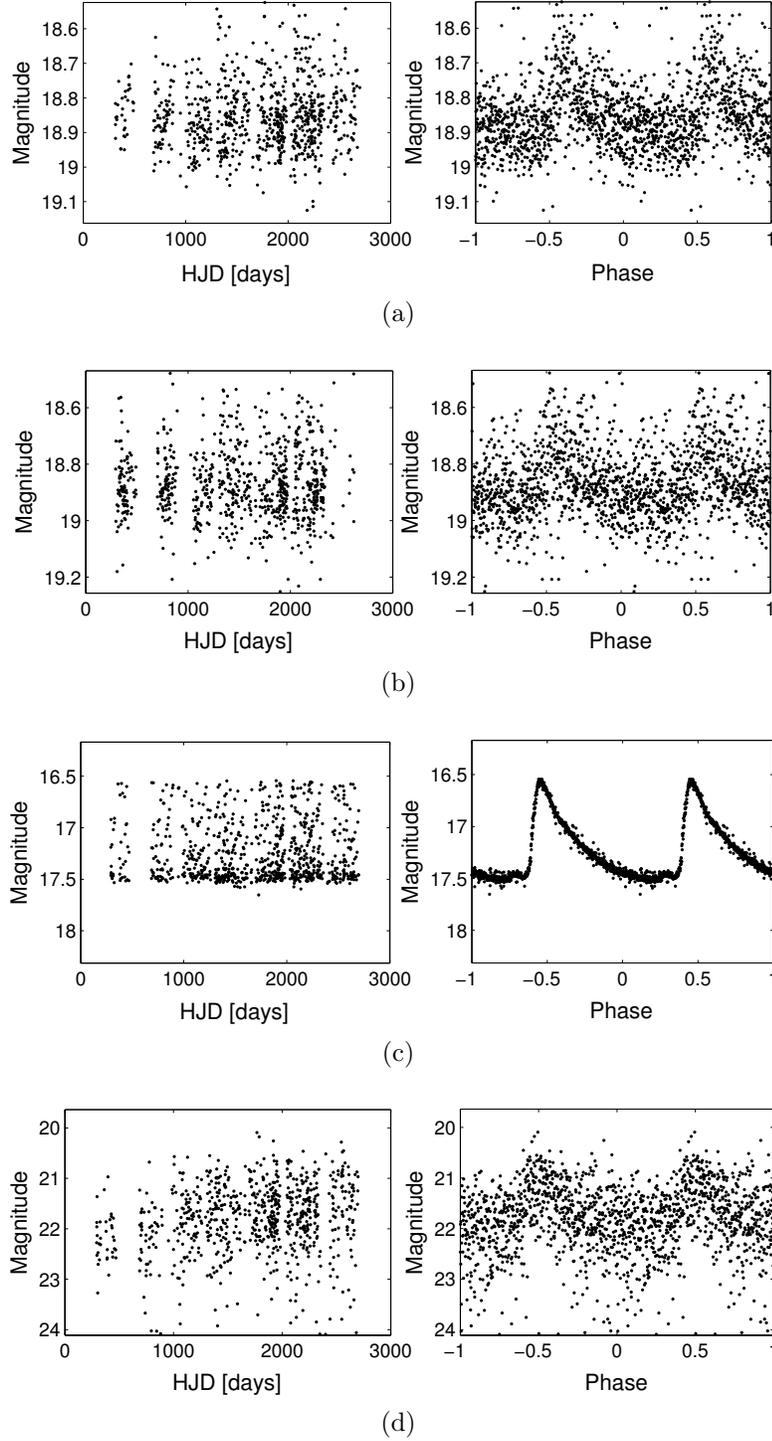
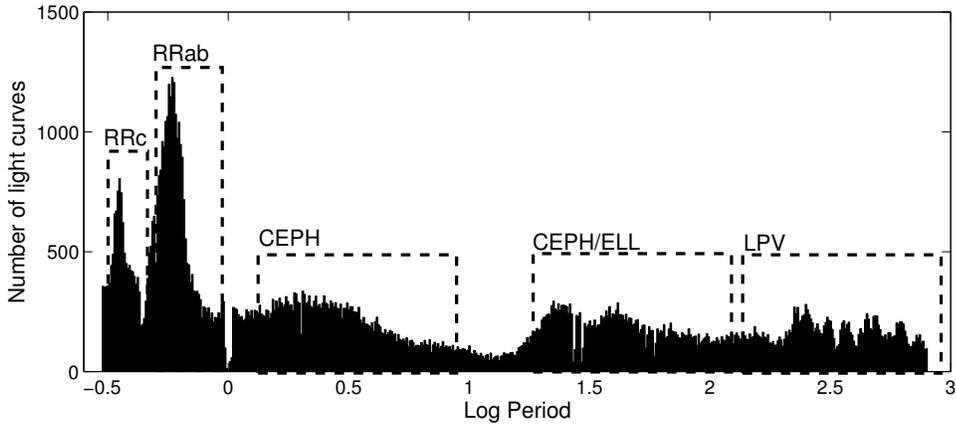
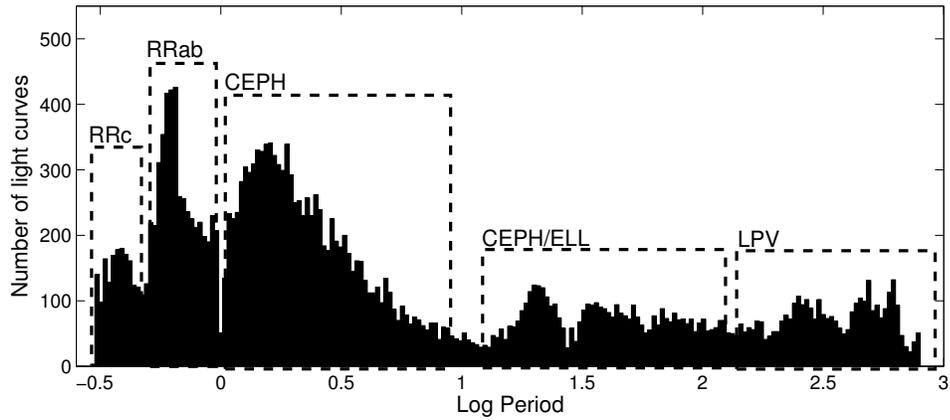


Figure 4.7: Examples of overlapping and blending. A period of 2.4796 days is detected for lightcurves sm0077n17908 (a) and sm0010k3199 (b). The angular distance between these lightcurves is 0.5 arcsec. Their difference in magnitude and CKP value is 0.03 and 0.23, respectively. These lightcurves are associated to a star that is in an overlapped region between fields sm001 and sm007. lightcurves sm0023n10183 (c) and sm0023n10325 (d) are also found to have the same period (1.2535 days), but they reside in the same field. Their angular distance,  $\delta$ -magnitude and  $\delta$ -CKP is  $4.9''$ , 4.5 and 4.1, respectively. In this case the light from sm0023n10183 (c) introduces a periodicity in its neighbour (d).



(a)



(b)

Figure 4.8: Histogram of the periods found in the LMC (a) and SMC (b) blue channel data. The regions marked with dotted boxes are associated to clusters of a given type periodic variable star.

marked with black squares. Examples of the periodic variable stars found in these regions are presented in Appendix B. These results are consistent with the color magnitude diagram of the LMC periodic variables from the OGLE survey [114].

Fig. 4.8b and 4.9b show the histogram of periods and the color magnitude diagram of the periodic lightcurves found in the SMC blue channel, respectively. By comparing the histogram and color magnitude diagram with those of the LMC, the following differences arise: the relative size of the Cepheid population is larger in the SMC, the relative size of the c-type RR Lyrae population is larger in the LMC.

The red channel light curves were also analyzed for comparison purposes. A grand total of 87,025 and 14,501 periodic light curves is collected from the LMC and SMC red channel data, respectively. This represents a decrease of 30% with respect to the amount of periodics collected from the blue channel. By cross-matching the lists obtained from the blue and red channels in the LMC we found that 68,179 objects appear in both lists, 50,141 objects are found only in the blue channel, and 18,846 objects are found only in the red channel. For the SMC, 12,536 objects appear in both lists, 1,965 appear exclusively in the red and 10,567

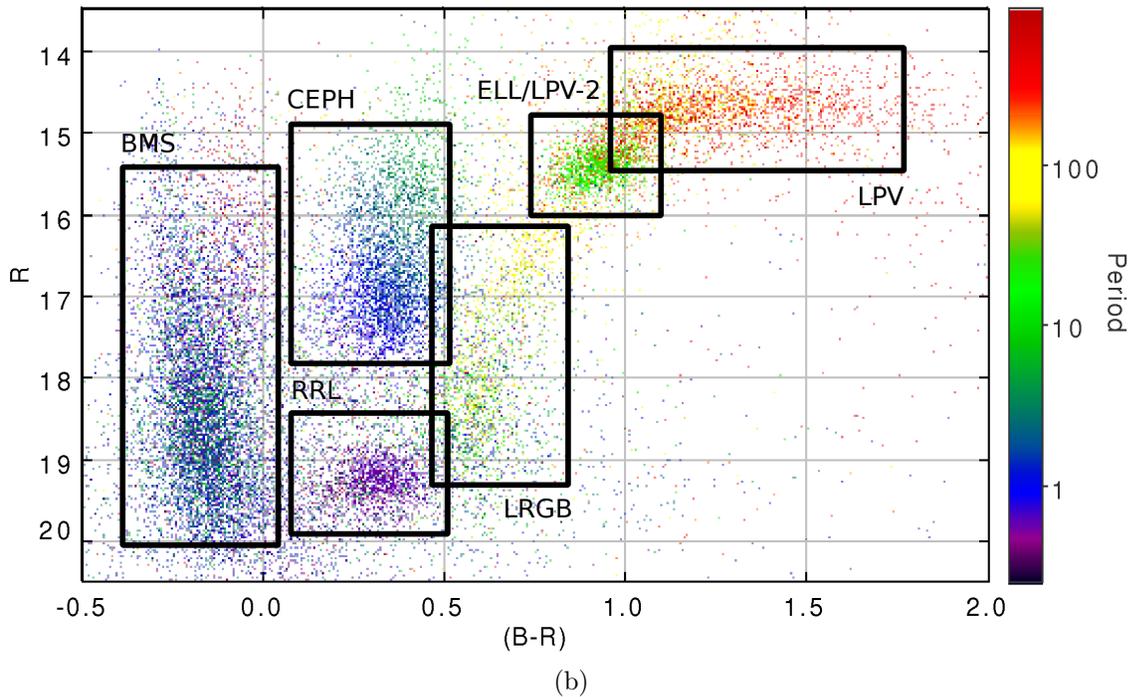
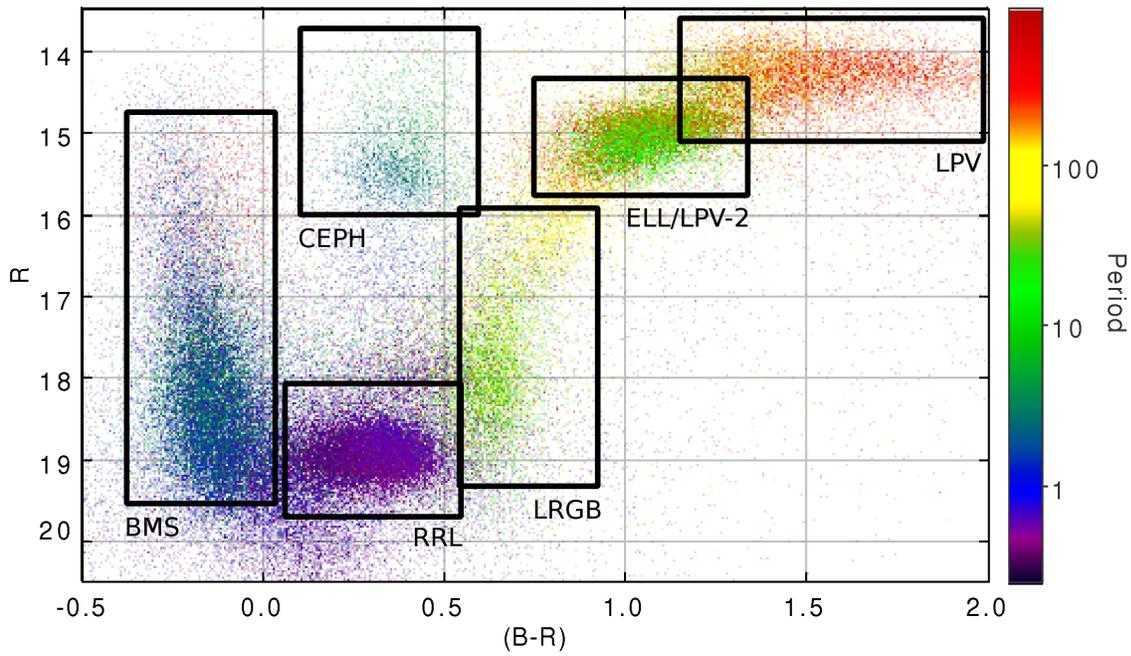


Figure 4.9: Color magnitude diagram showing the periodic lightcurves found in the LMC (a) and SMC (b). BMS corresponds to blue main sequence. LRGB corresponds to the lower red giant branch. Black boxes mark the location of Cepheid, RR Lyrae, LPV and ellipsoidal variable populations.

Table 4.7: Periodic lightcurve discrimination results summary on the EROS-2 survey.

	$N_{LC}$	$\tilde{N}_p$	Discarded	$N_p$	Periodics [%]
LMC	28,797,305	120,983	2,663	121,147	0.42
SMC	4,064,179	24,920	1,817	24,855	0.61

appear exclusively in the blue. For a given object the SNR may change between channels as shown in the examples of Fig. 4.10. By inspecting the histogram of the color  $(B - R)_{eros}$  of the EROS-2 light curves, it is clear that it is skewed to the blue side. The average color value in the LMC and SMC is 0.46 and 0.31, respectively and therefore the SNR is higher in the blue channel which explains why more periodics are found in the blue channel data<sup>5</sup>.

The catalogs are compared with existing periodic variable star catalogs for the LMC and SMC. We first tested against the published OGLE catalogs for Cepheids [109, 115], type II Cepheids [116, 117], RR Lyrae [118, 119] and LPV [120, 121] in the LMC and SMC. The OGLE team performed an extensive period search using Fourier based methods, analysis of variance and visual inspection. In our test the objective is to reveal how many of the periodic variables reported by the OGLE team can be found in our catalogs and to analyze the discrepancies between the detected periods. Table 4.8 summarizes the results of the crossmatching. First, for each OGLE object, a nearest neighbor in the EROS catalog is found. Neighbors with a separation larger than 1.5 arcsec are not considered. Column  $N_{inEROS}$  corresponds to the number of OGLE objects that were found in the EROS set within the search distance. The OGLE objects that did not have an EROS neighbor were either out of EROS bounds, located on inter-chip EROS zones or located on corrupted EROS chips. Column  $N_{match}$  corresponds to the number of crossmatched OGLE-EROS objects that appear in our periodic variable catalog. The differences between  $N_{inEROS}$  and  $N_{match}$  are due to OGLE objects whose CKP is below the periodicity threshold (low SNR light curves). There are cases in which the true period is within the spurious filters areas and was missed in our search. Finally the periods reported by OGLE are compared to the periods found with the CKP based periodicity discriminator. The agreement column corresponds to the percentage of light curves in which the OGLE period is equal to the period found in EROS (a 1% relative error is considered). The multiple column corresponds to the cases in which the reported period is either a multiple, sub-multiple or alias of the OGLE period. The disagreement column corresponds to the cases in which the reported period is not related to the OGLE period.

There is a high level of agreement between the reported and OGLE periods for Cepheids, type II Cepheids and RR Lyrae classes, in both the LMC and SMC. The periods labeled as multiples were visually inspected. In these cases the OGLE period is the correct period, but it was not found by the proposed method because it was either below 0.3 days or filtered in the spurious period rejection stage. Examples of the light curves in which the reported period is in disagreement with the OGLE period are shown in Fig. 4.11.

For the LPV class the difference between  $N_{inEROS}$  and  $N_{match}$  is larger than in other classes (i.e. more objects with CKP below periodicity threshold). This is expected as the

<sup>5</sup>Another reason could be related to the training scheme, in which only blue channel light curves were used to create the synthetic database.

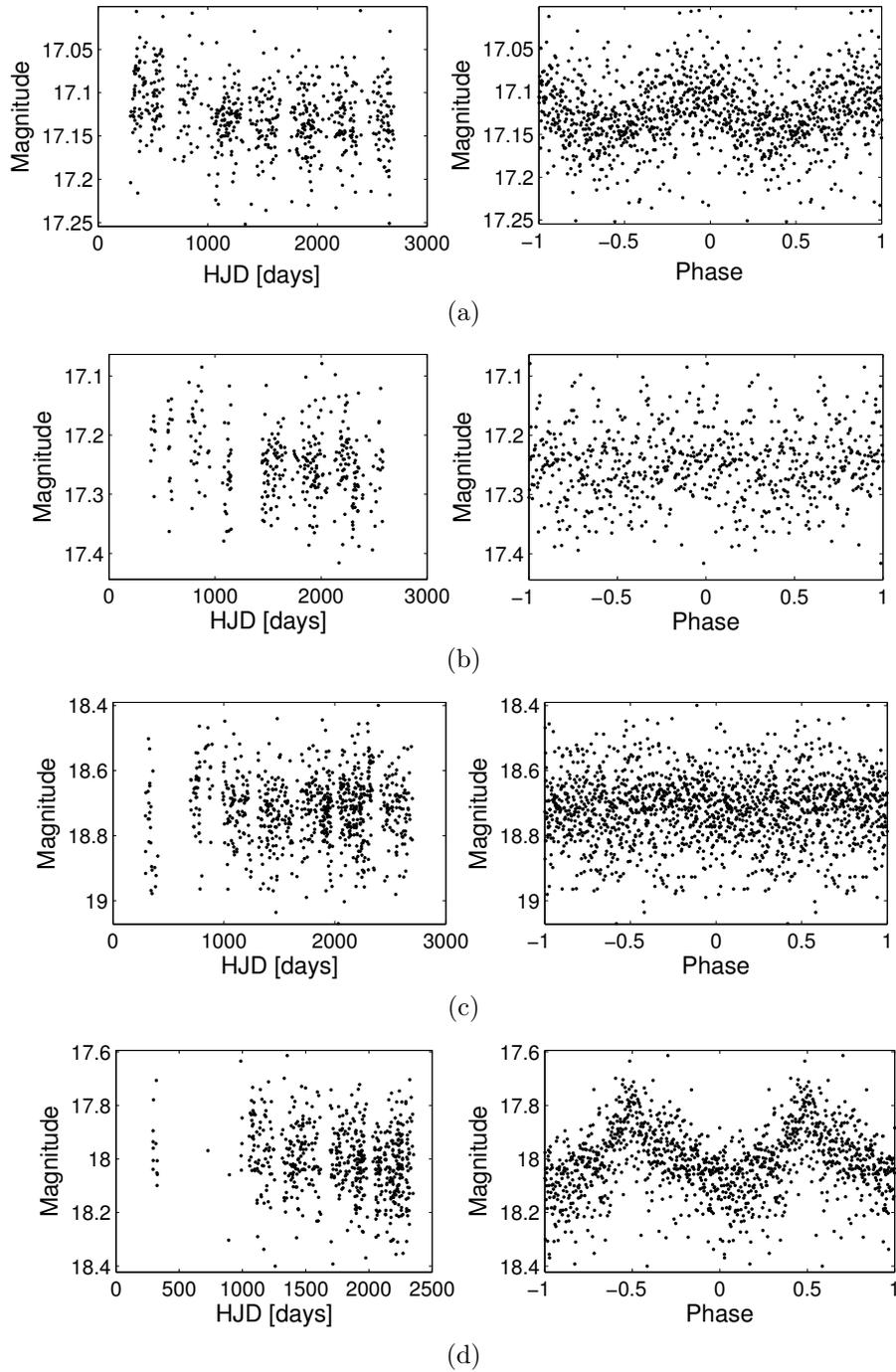


Figure 4.10: Examples of periodic light curves detected only in one of the EROS-2 channels. Fig (a) and (b) correspond to light curve lm0012k17912. Figure (a) shows the blue channel light curve folded with the detected period of 0.48004 days. On the red channel data no strong periodicity was found. Fig (b) shows red channel light curve folded with the 0.48004 days periods. Figure (c) and (d) correspond to light curve sm0010l10270. Fig. (d) shows the red channel data folded with the detected period of 10.4453 days. Using the blue channel data no strong periodicity was found. Figure (c) shows the blue channel data folded with the period detected in the red channel. In both cases the periodicity is only assessed in one of the channels.

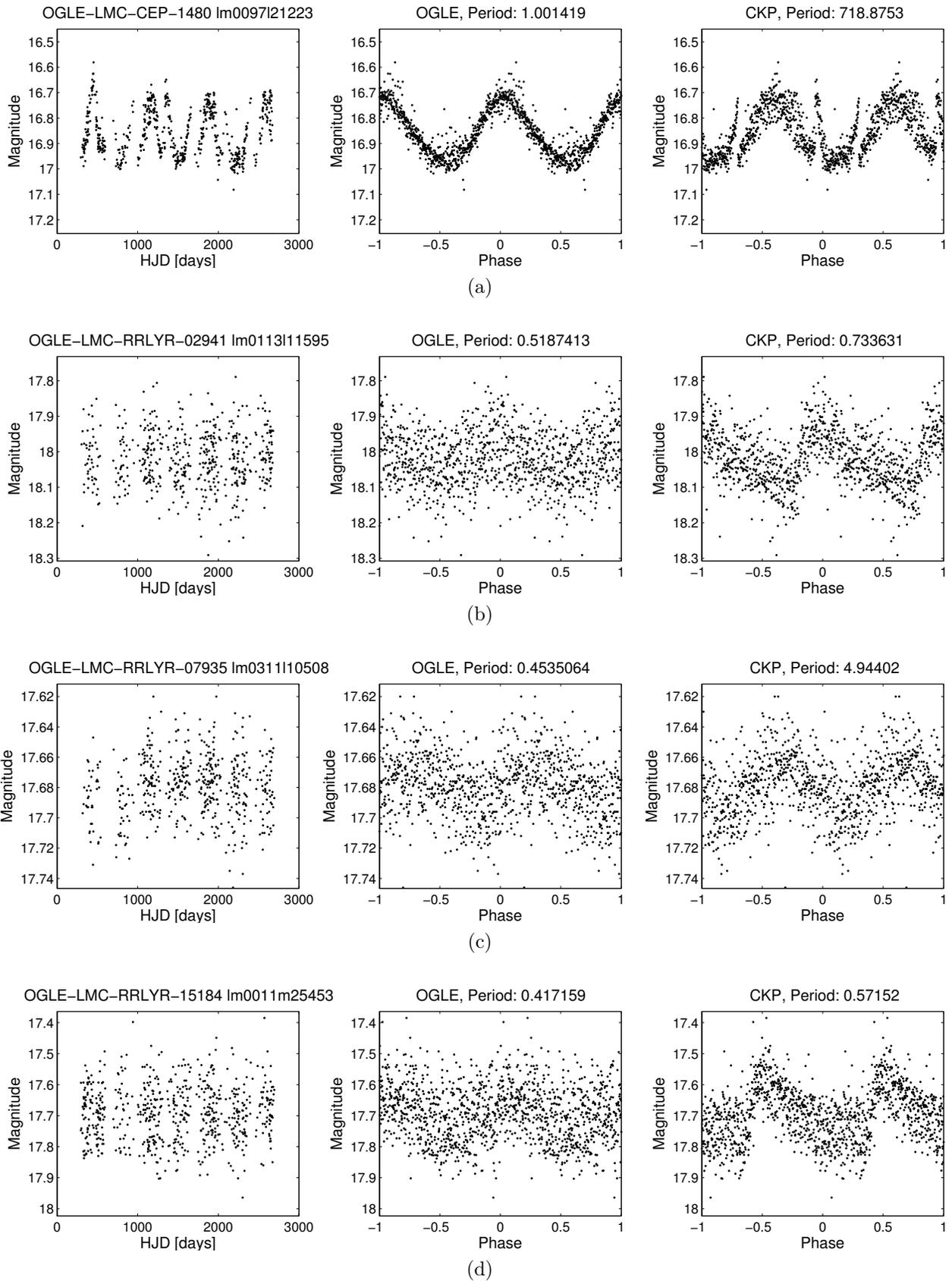


Figure 4.11: Light curves in which the reported period is in disagreement with the OGLE period. The EROS and OGLE labels, along the periods are shown in the title of each light curve.

Table 4.8: Crossmatching with OGLE periodic variable catalogs in the LMC and SMC.

OGLE catalog	$N_{catalog}$	$N_{inEROS}$	$N_{match}$	Agree [%]	Multiple [%]	Disagree [%]
OGLE-LMC-CEPH	3,375	2,727	2,711	98.8	1.0	0.2
OGLE-LMC-t2CEPH	203	161	148	94.6	4.1	1.3
OGLE-LMC-RRLyrr	24,906	18,092	17,272	92.0	6.8	1.2
OGLE-LMC-LPV	91,995	74,960	20,430	77.2	2.0	20.8
OGLE-SMC-CEPH	4,630	3,413	3,395	99.3	0.6	0.1
OGLE-SMC-t2CEPH	43	30	30	93.4	3.3	3.3
OGLE-SMC-RRLyrr	2,475	1,392	1,360	97.7	1.7	0.6
OGLE-SMC-LPV	19,384	14,103	4,413	70.3	2.6	27.1

Table 4.9: Crossmatching with EROS-2 *beat* Cepheid catalogs for the LMC and SMC.

Beat Cepheids catalog	$N_{catalog}$	$N_{match}$	Agree [%]	Multiple [%]	Disagree [%]
F/FO pulsation	115	109	100.0	0.0	0.0
FO/SO pulsation	302	300	99.0	0.66	0.33

LPVs are known to suffer from irregularities that affect their period. Additionally, the level of agreement between periods is lower than the other classes. Fig. 4.12 shows examples of disagreeing periods in the LPV class.

There are 80,304 objects in our periodic catalog that do not have a neighbor from the OGLE periodic variable catalogs (within 2.5 arcsec). Some of these objects may have not been surveyed by the OGLE project, or they could belong to classes with currently not available catalogs such as eclipsing binaries. A 60% of these light curves have a low CKP value which translates roughly to low SNR. This could indicate that the proposed method is more sensitive than the method used by the OGLE team. Fig. 4.13 shows examples of periodic light curves found in the EROS catalog that do not appear in the OGLE catalogs.

The periodic variable catalogs are also compared to the lists of beat Cepheids found in the EROS-2 data by [122]. The catalog contains Cepheids pulsating on their fundamental and first overtone (F/FO) and first and second overtone (FO/SO), respectively. The periods were obtained using a combination of Fourier decomposition, Analysis of Variance and visual inspection. The results are summarized in Table 4.9. There are eight cases that do not appear in our catalog due to their CKP value being below the threshold. In the remaining 409 cases, only three cases show disagreement with the reported period. The one case in which the period is not a multiple of the EROS-2 period was shown in Fig. 4.11a.

#### 4.3.4. Multimodes in the EROS-2 survey

The post-processing procedure presented in Section 3.3.6 is applied on 34,000 periodic light curves from the LMC with CKP values above 2.0<sup>6</sup>. From this set 1165 light curves are selected as dual mode candidates. After evaluating the double mode candidates, 116 are

<sup>6</sup>Only the most prominent periodic light curves are selected.

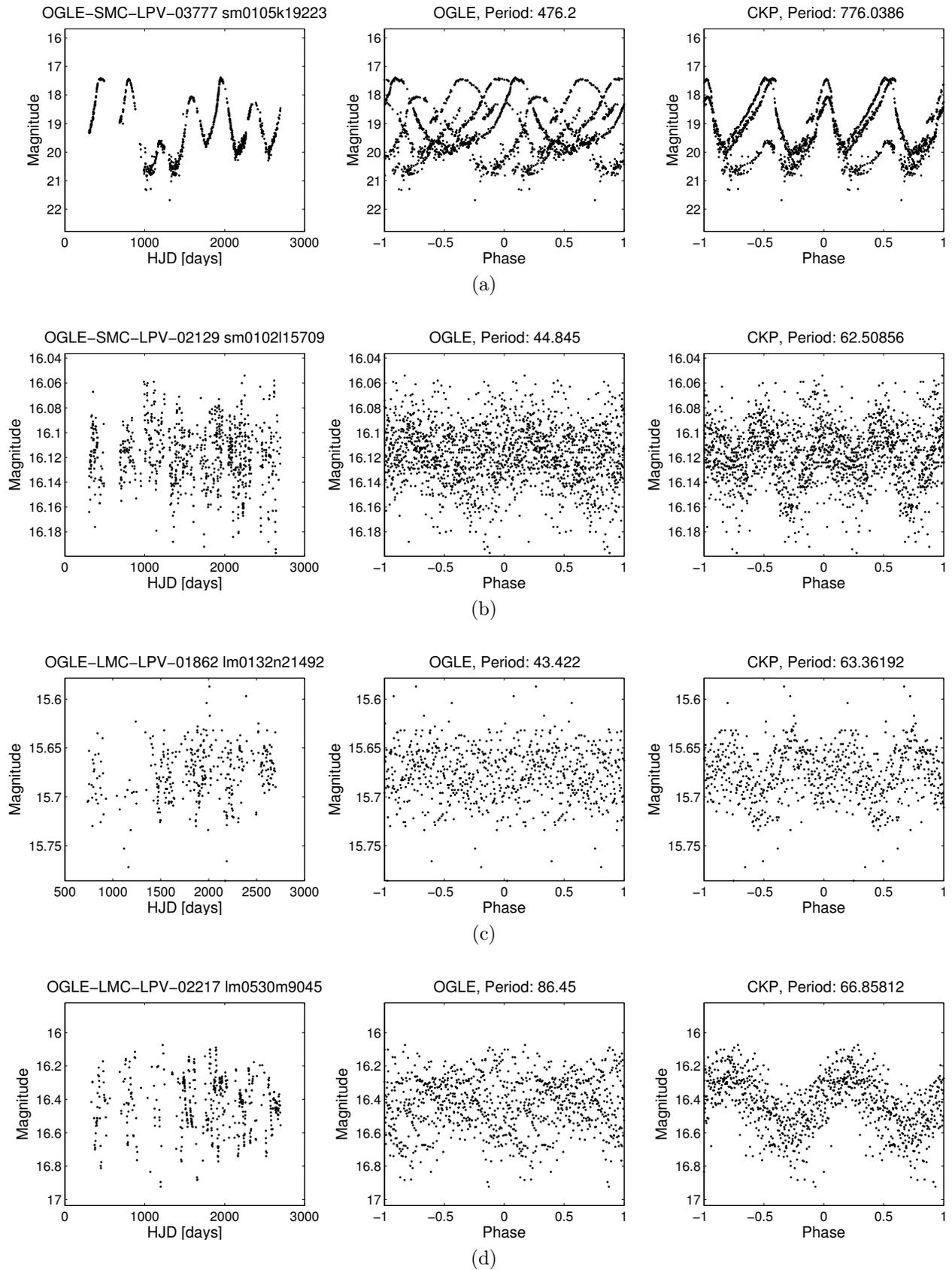
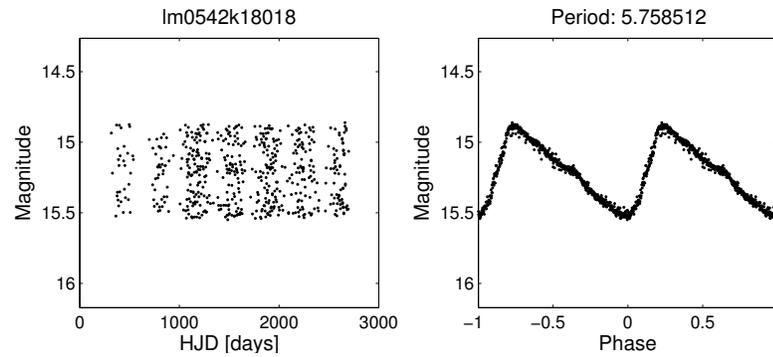
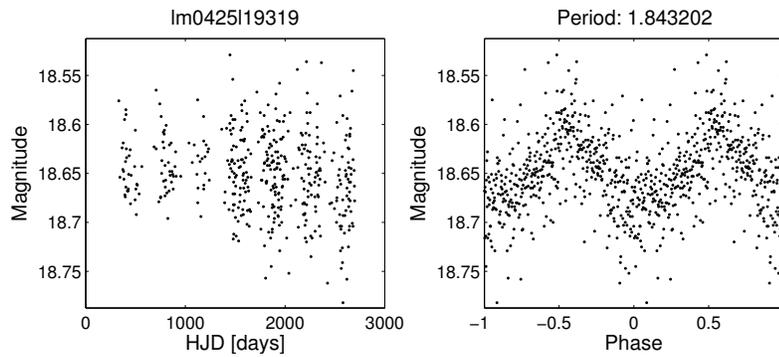


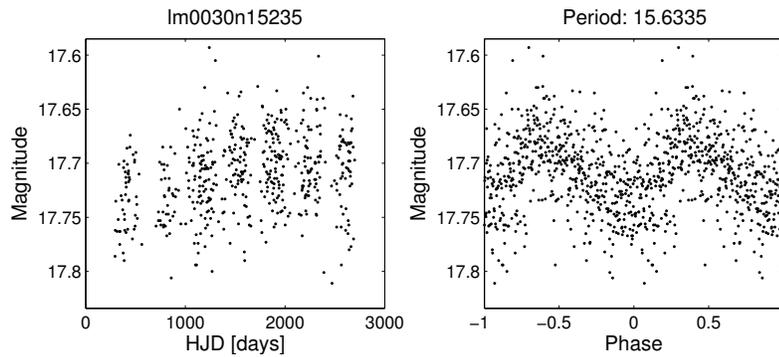
Figure 4.12: Examples of LPVs in which the reported period is in disagreement with the OGLE period. The EROS and OGLE labels, along the periods are shown in the title of each light curve.



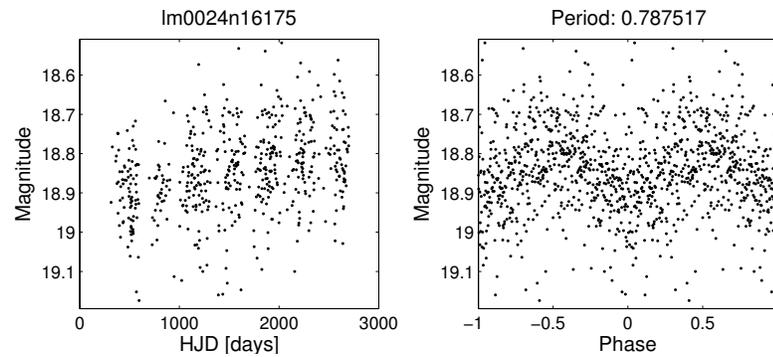
(a)



(b)



(c)



(d)

Figure 4.13: Examples of periodic light curves not found by OGLE. Fig (a) corresponds to a Cepheid variable with high SNR not found by OGLE. The majority of these light curves have a low CKP value which translates roughly to low SNR. Figures (b), (c) and (d) are low SNR examples.

found to have a third oscillation mode. Examples of dual mode and triple mode candidates are shown in Figures 4.14 and 4.15, respectively.

Fig. 4.16 shows a Petersen diagram of the 1165 light curves selected as dual modes candidates. The triangles in the plot mark the 116 light curves in which a third mode was found. The periods are sorted so that  $P_0 > P_1$  in all cases. The triple mode candidates occupy two horizontal lines at period ratios of 0.72 and 0.8. These values are close to the known ratios associated to the first and second overtones [104]. A prominent horizontal line appears at  $P_1/P_0 \sim 2/3$  for fundamental periods above 10 days. According to [123] this ratio is associated to the period doubling phenomenon.

### 4.3.5. Computational efficiency and processing times

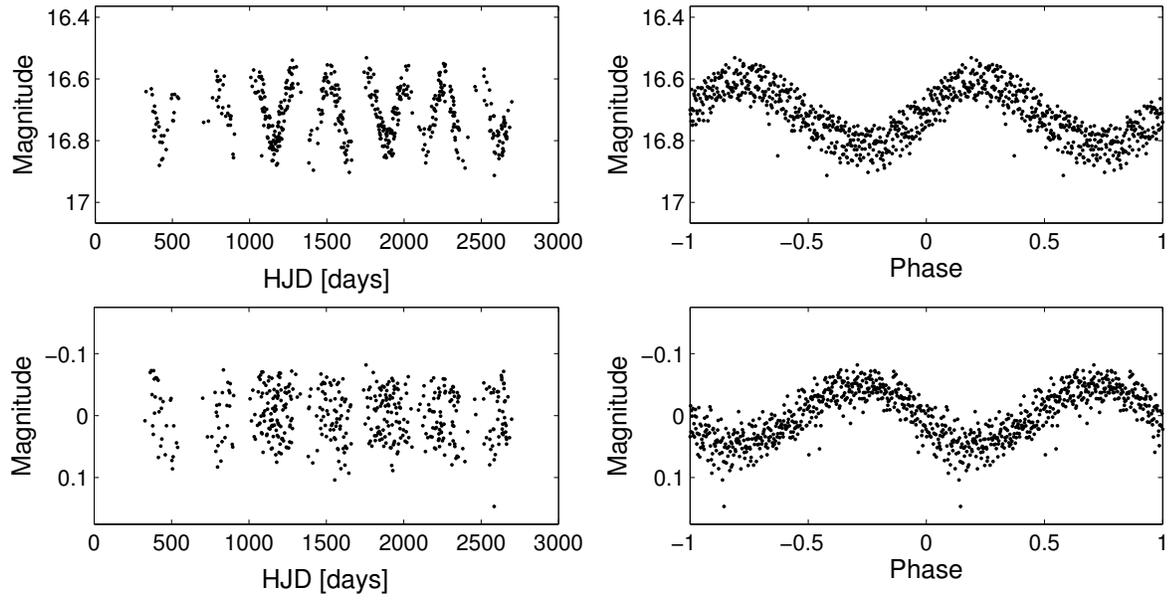
The EROS-2 periodicity discrimination pipeline is programmed for computational architectures based on graphical processing units (GPUs). Topics related to General Purpose Computing on GPU (GPGPU) and the pseudocode for the GPGPU CKP implementation can be found in Section 3.3.8. The computational time required to analyze one light curve using the GPU implementation of the pipeline is shown in Fig. 4.17a. These times include the importation and transferring of the light curves to the GPU device. Times were measured on a NVIDIA Tesla C2070 GPU. Fig. 4.17b shows the speedup as a function of number of samples. Speedup is defined as the CPU time over the GPU time. For the task of CKP computation one GPU is roughly equivalent to a 38/48 cores CPU cluster. The speedup increases with the number of samples, *i.e.* the GPU implementation scales better with this parameter.

The 32.8 million light curves from the EROS-2 survey are processed on the NSCA Dell/NVIDIA cluster Forge. Forge is part of the Extreme Science and Engineering Discovery Environment (XSEDE). Forge has a total of 288 NVIDIA Tesla C2070 accelerators distributed on 44 nodes, however the maximum number of nodes that can be used at a time is 26. The cluster has two queues. The first queue has 18 nodes available with 6 GPUs per node. The second queue has 8 nodes available with 8 GPUs per node<sup>7</sup>. For a given node,  $M$  OpenMP<sup>8</sup> threads are launched, where  $M$  is the number of available GPUs. Each thread controls one GPU. Each GPU processes one chip from EROS-2. Table 4.10 shows the total computational time required to process the 32.8 million light curves from the LMC and SMC. These results do not include the time required to transfer the entire dataset to the cluster nor the time a job is waiting on the queue. Due to sharing of the cluster resources an average of 12 nodes at a time could be used. Using 12 nodes the EROS-2 dataset is processed in approximately 18 hours.

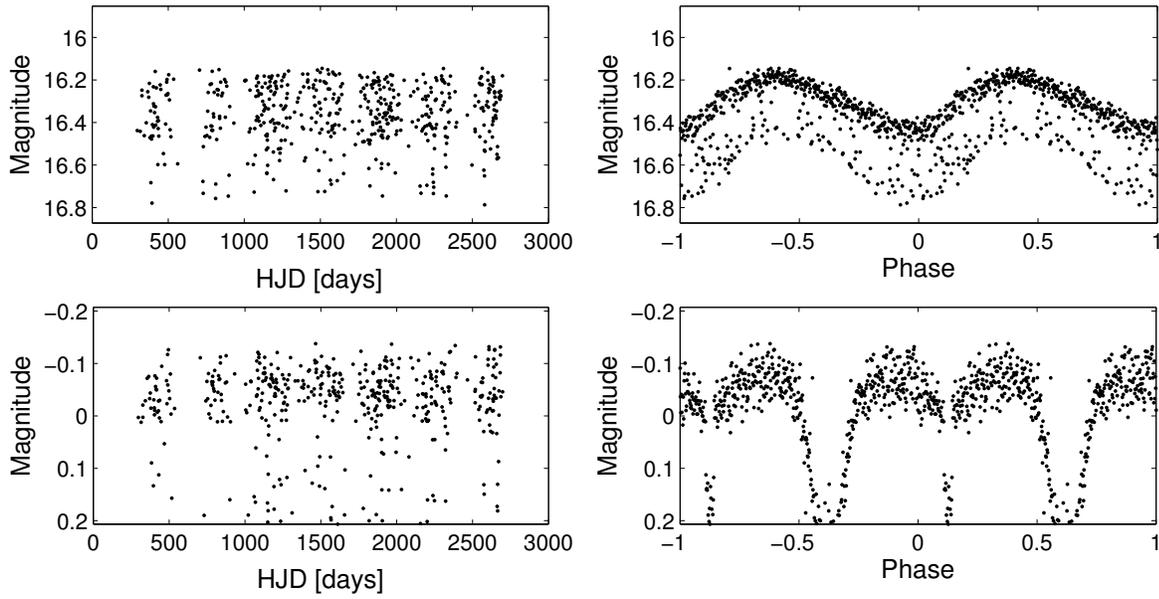
---

<sup>7</sup>More detailed information can be found at <https://www.xsede.org/web/guest/nscsa-forge>.

<sup>8</sup>API for multiprocessing programming in C++, more information at <http://openmp.org/wp>.



(a)



(b)

Figure 4.14: Light curves lm0356k24082 (a) and lm0100m7313 (b) are selected as dual mode candidates. On each plot, the first and second rows correspond to the original and whitened light curve, respectively. In (a) the original light curve is folded with  $P_0 = 244.06$  days. The whitened light curve is folded with  $P_1 = 3.6399$  days. In (b) the original light curve is folded with  $P_0 = 6.3419$  days. The whitened light curve is folded with  $P_1 = 84.19$  days.

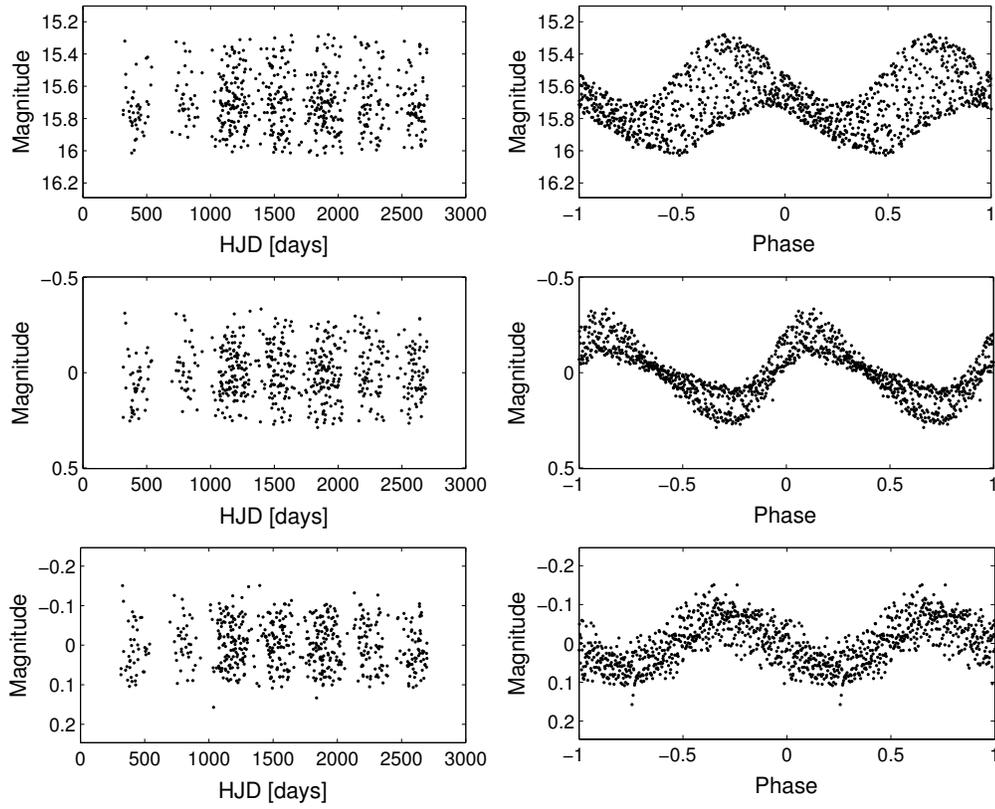


Figure 4.15: Light curve lm056518888 is selected as a triple mode candidate. In the plot the first, second and third rows correspond to the original, first whitened and second whitened light curves, respectively. The original light curve is folded with the detected period  $P_0 = 2.4725$  days. The first whitened light curve is folded with  $P_1 = 3.4455$  days. The second whitened light curve is folded with  $P_2 = 1.4395$  days.

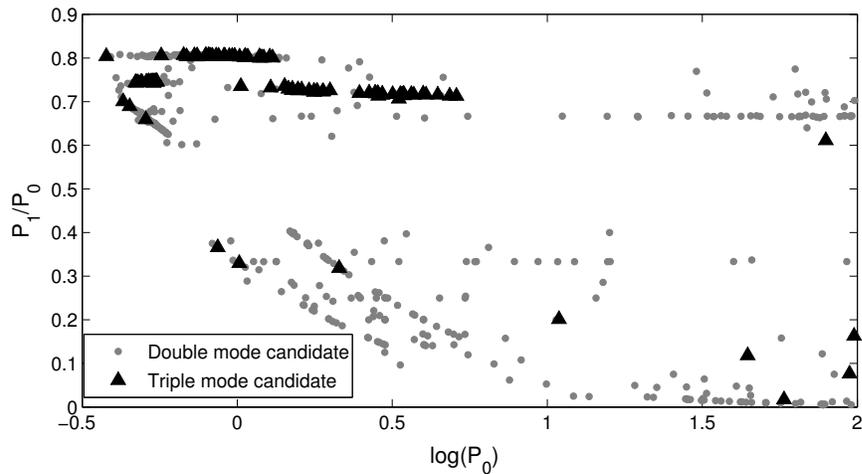


Figure 4.16: Petersen diagram of the 1165 dual mode candidates found in the LMC. The triangles mark the location of 116 triple mode candidates. Clear structures arise in the diagram.

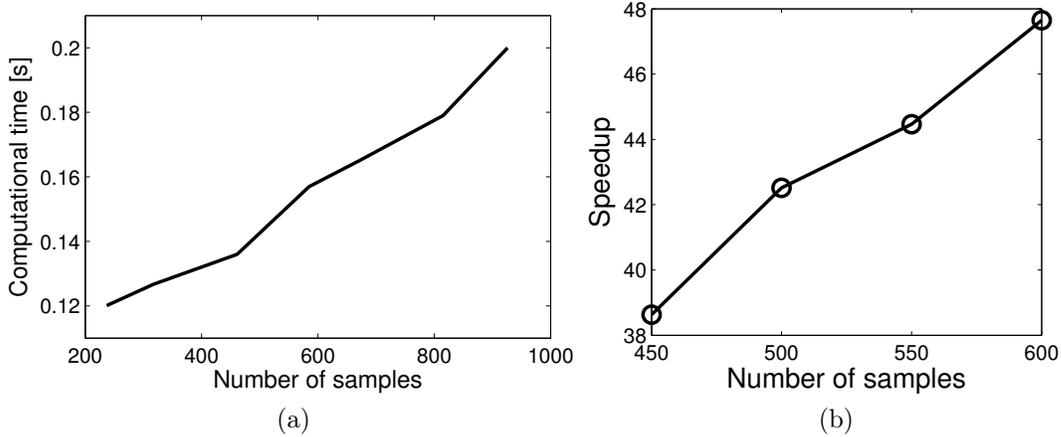


Figure 4.17: (a) Computational time required to process one light curve using CKP implementation in GPGPU, as a function of the number of samples. (b) Speedup of the GPU versus the CPU implementation of the pipeline.

Table 4.10: Total computational time required to process the 32.8 million EROS-2 light curves (LMC and SMC) on XSEDE Forge cluster. GPUs in all nodes are NVIDIA Tesla C2070.

Hardware	Computational time
Using 1 GPU	52.2 days
Using 6 GPUs (1 node)	8.71 days
Using 12 nodes (6 GPUs/node)	17.41 hours
Using all available nodes	7.28 hours

## 4.4. Period estimation using the CNMFS

### 4.4.1. Setup and performance criteria

In this section the experiments used to characterize and evaluate the period estimation performance of the CNMFS are described. First, the CNMFS is tested on a set of synthetic periodic time series created with the following procedure:

1. Define a time span  $T$ , sampling frequency  $F_s$ , fundamental frequency and signal-to-noise ratio (SNR).
2. Generate a regularly sampled time vector using  $T$  and  $F_s$ , the number of points of the signal is defined as  $N = TF_s$ .
3. Generate a N-length random periodic signal using the periodic kernel function (Eq. ??) as a covariance matrix for a multivariate normal random number generator. Normalize the signal by subtracting its mean and dividing by its standard deviation.
4. Generate a N-length noise vector using a random number generator. A heavy tailed

t-student random number generator is used.

5. Scale the amplitude of the signal so that it matches the square root of the desired SNR times the standard deviation of the noise.
6. Add the noise vector to the signal vector.

The following experiments are considered

- **E-1:** Synthetic time series with varying SNR and frequency.
- **E-2:** Synthetic time series with varying time span and frequency.
- **E-3:** Astronomical light curves.

In the first experiment **E-1** a set of 400 synthetic periodic time series with different frequencies and signal-to-noise ratios (SNRs) is used. The time series have a fixed time span of  $T = 2$  seconds and are sampled at  $F_s = 500$  Hz<sup>9</sup>. The synthetic set is divided in four groups of 100 time series each. The fundamental frequency is constant along each group, and the following values are used: 9 Hz, 15 Hz, 55 Hz and 121 Hz. The groups are separated in 10 sub-groups each having a constant SNR. The values for the SNR are 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4 and 5.

In the second experiment **E-2** a set of 400 synthetic periodic time series with different frequencies and time spans is used. These time series have a fixed SNR = 5 and are sampled at  $F_s = 500$  Hz. Again, the set is divided into 4 groups, each having 100 time series and a particular fundamental frequency. The groups are then separated into 5 sub-groups each having a particular time span  $T$ . The time spans considered are 0.1, 0.2, 0.5, 1 and 2 seconds.

The objective of **E-1** and **E-2** is to assess the influence of the adaptive frequency and the Kronecker delta dictionaries, as well as to compare the performance of the CNMFS with other frequency domain representations for correntropy, such as the CSD and the CKP. For reference, the correlation NMF spectrum and the PSD are also included in the comparison. The overcompleteness of the dictionary is set to  $L = 10$ . After a preliminary evaluation, the values for the learning rate and momentum rate are set to  $\mu = 0.02$  and  $\alpha = 0.5$ , respectively.

In the third experiment **E-3** a set of 1500 periodic light curves drawn from the MACHO project [1] is used. The set corresponds to 500 Cepheids, 500 RR Lyrae and 500 eclipsing binary stars. The objective of this experiment is to compare the period estimation performance of the CNMFS with conventional methods used in astronomy. The Lomb-Scargle (LS) [34] and Analysis of Variance (AoV) [51] periodograms are considered. The implementation of these methods is provided by the astronomical software package Vartools [110]. For reference the CSD and CKP are also included in the comparison. For all the methods, the frequency axis spans from 0 [1/days] to 2 [1/days] with a resolution of  $0.1/T$ . Eight bins are used for the AoV periodogram. The periods of these light curves were estimated by astronomers of the Harvard Time Series Center (TSC). These target periods are considered as the gold standard. The light curves are irregularly sampled, having 1000 points in average.

The methods are compared in terms of

---

<sup>9</sup>The Nyquist frequency is 250 Hz and the number of samples is 1000.

- **Hit rate:** Percentage of cases in which the frequency associated to the global maximum of the spectrum corresponds to the fundamental frequency of the time series. A tolerance of 2% is considered.
- **Spectral localization:** Percentage of “energy” allocated at the fundamental frequency. For a given spectrum this metric is calculated as

$$q_1 = \frac{S(f^*)}{\sum_{i=1}^{N_f} S(f_i)}, \quad (4.1)$$

where  $N_f$  is the length of the spectrum and  $S(f^*)$  is the spectral ordinate at the fundamental frequency. The average of  $q_1$  along the dataset is considered.

- **Spectral sparseness:** Percentage of spectral ordinates with a value lower than 10% of the global maximum of the spectrum. For a given spectrum this metric is calculated as

$$q_2 = \frac{1}{N_f} \sum_{i=1}^{N_f} I_i, \quad (4.2)$$

where

$$I_i = \begin{cases} 1 & \text{if } S(f_i) \leq 0.1 \max(S(f)) \\ 0 & \text{otherwise.} \end{cases} \quad (4.3)$$

The average of  $q_2$  along the dataset is considered.

- **Fitting error:** This corresponds to the value of the objective function (Eq. 2.48) at the end of the NMF routine. This metric is used to compare different NMF implementations.

#### 4.4.2. Period estimation in synthetic time series

In this section the results obtained on the synthetic data set are presented and analyzed. First, the influence of the adaptive frequency dictionary and the Kronecker delta dictionary on the **E-1** set is evaluated.

Fig. 4.18a shows the hit rate as a function of the number of iterations. In general, the hit rate metric is not too dependent on the characteristics of the tested dictionaries. All the CNMFS routines, having different dictionaries, achieve a hit rate of 95%, although not at the same time. When the adaptive frequency and Kronecker delta dictionaries are used the maximum hit rate is obtained in less iterations with respect to the other combinations. Using the adaptive dictionary, the correntropy can be fitted using fewer atoms. Fig. 4.18b shows the spectral localization metric as a function of the number of iterations. The best results are obtained when the adaptive frequency and Kronecker dictionaries are used. For this metric, the adaptiveness of the frequency dictionary carries more weight than the inclusion of the Kronecker delta dictionary. An adaptive frequency dictionary helps reducing the number of harmonics required to fit the periodic components in the correntropy function, hence increasing the spectral localization. When the Kronecker delta dictionary is not used the non-periodic components of the correntropy are fitted by a large combination of high-frequency

atoms, hence spectral localization is decreased. The Kronecker delta fits the non-periodic behavior filtering it from the frequency representation.

The spectral localization metric (Eq. 4.1) measures the importance of the atom associated to the fundamental frequency with respect to the total “spectral energy”. In general, the spectral baseline is what dominates the total spectral energy, hence this metric measures quality in a global sense. The sparseness metric (Eq. 4.2) is proposed as a complement for the localization metric. The sparseness is designed to measure the contribution from those spectral ordinates that rise above the baseline. The ideal frequency representation is one with high localization and sparseness, *i.e.* one where the only relevant peaks are those strictly associated to fundamental frequencies of the signal. Fig. 4.18c shows the sparseness of the NMFS as a function of the number of iterations. The sparseness metric benefits greatly from the adaptiveness of the frequency dictionary. Harmonics have large spectral ordinates, hence removing them improves sparseness. On the other hand the inclusion of the Kronecker delta dictionary is irrelevant in terms of this metric. As mentioned before the inclusion of the Kronecker delta dictionary decreases the number of high frequency atoms added by the routine in order to fit non-periodic behavior. The spectral ordinates of these atoms are usually low, hence the Kronecker dictionary has very low influence in terms of sparseness.

Fig. 4.18d shows the fitting error as a function of the number of iterations. From this figure it is clear that the Kronecker delta dictionary has great influence over the fitting error. This is expected as the non-periodic behavior is rather difficult if not impossible to fit completely solely with periodic atoms. On the other hand the choice regarding the frequency dictionary seems to be irrelevant in terms of this metric. This is also expected, as periodic behavior can be fitted equally well using either one correctly adapted atom or many static atoms.

Table 4.11 summarizes the results of experiment **E-1** for the CNMFS, correlation NMFS, PSD, CSD and CKP. In this experiment the period is always well-sampled with respect to the total time span, so the difficulty on detecting the period comes from the noise regime in the time series. The maximum hit rate is obtained by the CNMFS and CKP, followed by the correlation NMFS, CSD and PSD. The CNMFS, CSD and CKP are based on the correntropy function, hence have access to information on the second and the higher order moments of the input signal. On the other hand the correlation NMFS and the PSD, computed from the correlation function, are based on second-order statistics only. The PSD obtains the lowest hit rate which is explained by its inferior discriminatory capabilities with respect to the correntropy based estimators. There is a 20% difference in hit rate between the CSD and the CNMFS. In most cases, this difference is due to the CSD getting a multiple of the fundamental frequency as the global maximum of the spectrum. The CNMFS and CKP, which are based on the periodic kernel function, are less prone to this kind of error. In this experiment, the CNMFS and CKP obtained the same hit rate. Both methods failed exactly on the same cases, which correspond to very low SNR time series. This behavior is expected as the experiment was designed to compare the robustness against noise, exclusively. In this setup the periods are well-sampled, the synthetic light curves have only one periodic component and there are no embedded spurious periodicities, hence an increased frequency resolution (CNMFS) will not yield much increase in hit rate.

Table 4.12 summarizes the results of experiment **E-2** for the CNMFS, correlation NMFS,

PSD, CSD and CKP. In this experiment the time series have a high SNR, but decreasing time spans. The difficulty in detecting the period increases when the time span approximates the period, either because the period is large, or the time span is short. The highest hit rate is obtained by the CNMFS followed by the correlation NMFS, CSD, CKP and PSD. The relative hit rate difference between correntropy and correlation based methods is smaller than in the previous experiment, which is due to the high SNR of the time series (less influence of the noise). The increased frequency resolution of the NMF spectra helps them to achieve a more precise estimation than Fourier methods even when the period is very close to the time span.

In all the experiments, the highest performance in terms of localization and sparseness is obtained by the CNMF spectra. With the CNMFS the user is able to select a frequency spacing according to a desired frequency resolution. On the other hand, the frequency resolution of the Fourier transform is limited by the total time span (number of samples) of the signal. The PSD performs reasonably well in terms of sparseness, because a few “sine-wave atoms” are needed to reproduce the autocorrelation function. But, due to the limits in the frequency resolution, the peaks in the PSD are much broader than the NMFS peaks. The baseline of the PSD are proportionally much higher than the NMFS baseline. These reasons explain the difference in terms of spectral localization. The CSD performs worse than the PSD in terms of localization and sparseness. As explained in previous sections the sine-wave basis is a poor choice for the correntropy function, as several sine-wave atoms are needed to represent it, hence the large difference in sparseness with respect to the PSD. The CKP has the worst performance in terms of localization and sparseness.

The CKP estimator presents a high variance, and its spectral baseline is proportionally higher with respect to its competitors. The periodic kernel function does not yield an orthogonal basis. In practice, this translates as high contributions due to harmonics and sub-harmonics of the fundamental frequency plus aliasing with respect to sub-multiples of the Nyquist frequency<sup>10</sup>. A considerable amount of components with large spectral ordinate values arise in the CKP which explain its poor sparseness and localization. On other note, the bandwidth of the periodic kernel is rather difficult to calibrate. Note that an incorrect kernel bandwidth may enhance the unwanted spectral components.

Having a sparse, adaptive, high-resolution and localized spectrum estimator is desirable in a real-world scenario, in which light curves may have more than one underlying periodic signal, embedded spurious periodicities, high SNR and/or few available samples.

### 4.4.3. Period estimation in light curves

As an example let’s consider light curve MACHO 1.3449.948, which corresponds to an eclipsing binary star with a period of 14.0044 days. A plot of this light curve is shown in Fig. 4.19a. For these stellar objects conventional methods have difficulties discriminating the correct orbital period from their harmonics. Figures 4.19b and 4.19c show the slotted

---

<sup>10</sup>The aliasing in the CKP is greatly reduced when the sampling is irregular, which corresponds to the case for which the CKP was designed for.

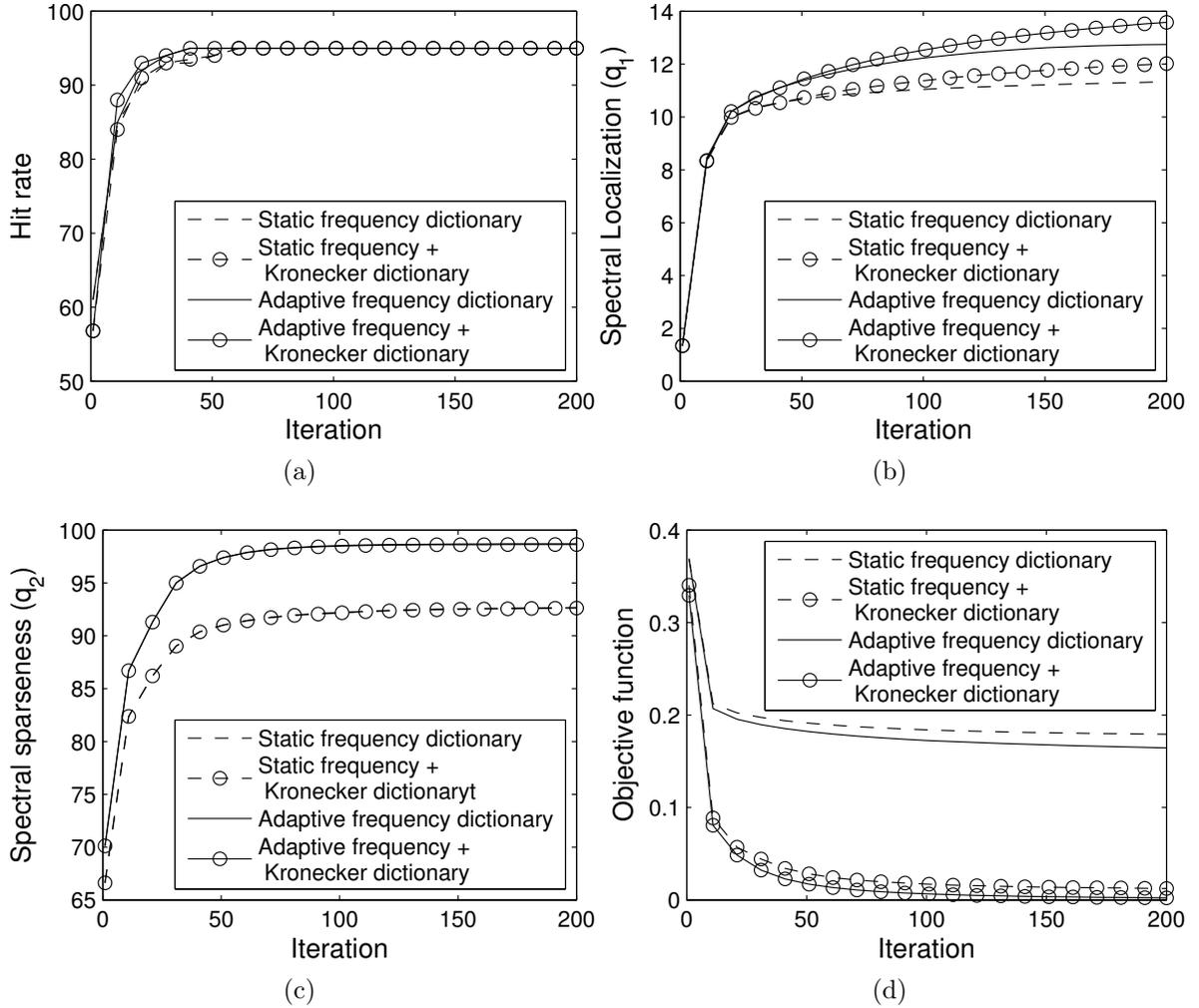


Figure 4.18: (a) Hit rate, (b) spectral localization, (c) spectral sparseness and (d) fitting error as a function of the number of iterations of the NMF routine. The best performance in terms of the four metrics is obtained when the adaptive frequency and Kronecker dictionaries are used.

estimators of the autocorrelation and autocorrentropy of the signal, respectively. The correntropy function is estimated using the kernel size given by the Silverman rule (Eq. 2.27). The slot size of the slotted estimators is set to 0.25 [days] following [100]. The kernel size of the CKP is set following the guidelines given in [101]. Fig. 4.19d shows the PSD, obtained by computing the DFT of the slotted autocorrelation. The true period is clearly missing from the spectrum, and the maximum corresponds to the closest multiple in frequency (7 days). Other peaks associated to multiples arise in the spectrum. Fig. 4.19e shows the CSD, computed from the DFT of the slotted autocorrelation. In this case the maximum of the spectrum is also the closest frequency multiple, but a small contribution at the true period can be seen. At first glance, the Fourier basis appears to be unable to highlight to true periodicity in the frequency domain. The CKP for this signal is shown in Fig. 4.19f. Unlike the PSD and CSD, the CKP is able to pick the true frequency of the signal. The periodic kernel gives the CKP the upper hand in terms of discriminability. But the frequency representation is far from ideal as it has abundant contribution at harmonics, sub-harmonics and aliases of

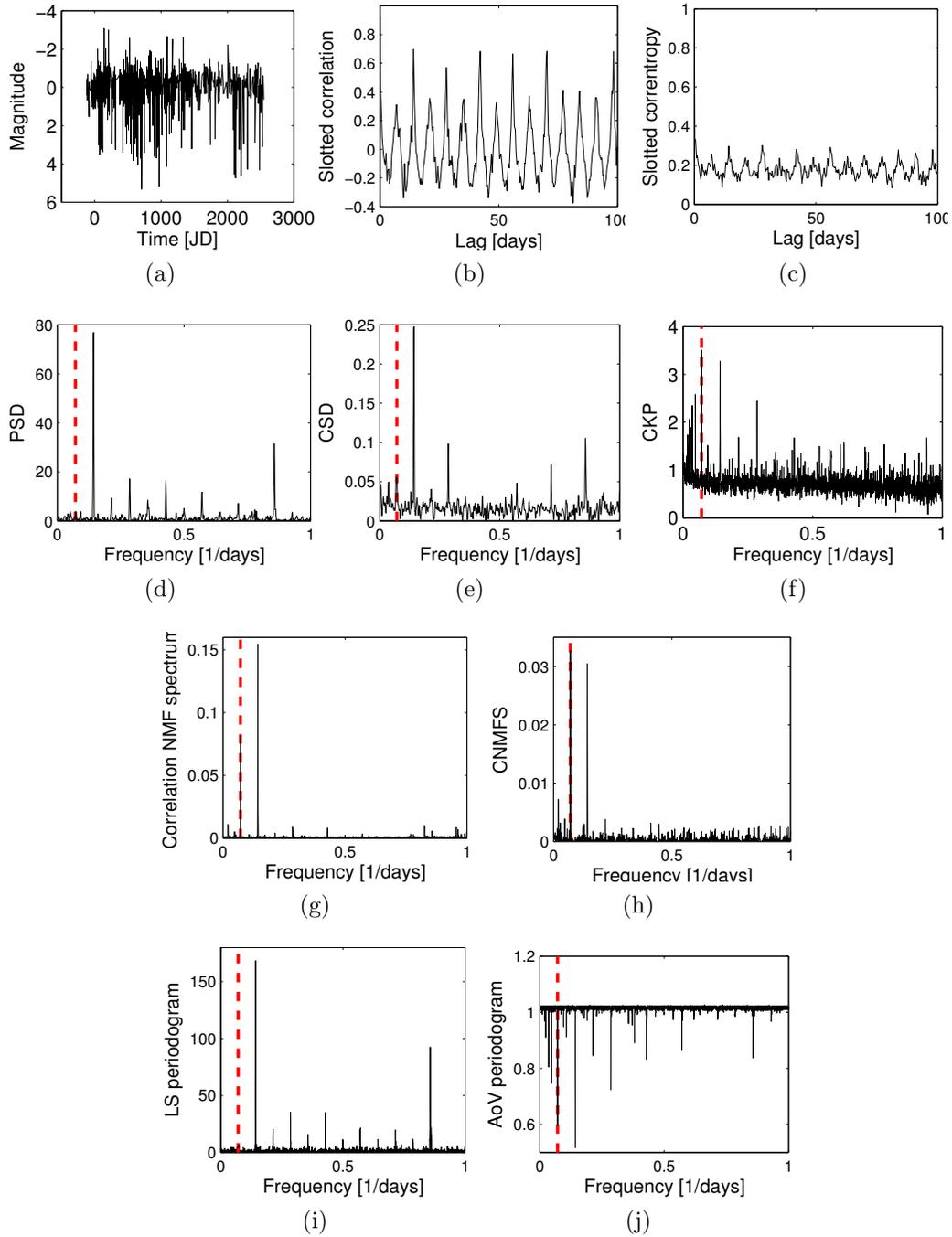


Figure 4.19: (a) Light curve 1.3449.948 from the MACHO project, it corresponds to an eclipsing binary star with a 14 days period. The red dotted line marks the location of the underlying period. The slotted correlation (b) and slotted correntropy (c) of this time series, respectively. The true period does not appear on the PSD (d) and it appears weakly on the CSD (e). The CKP (f), gets the correct period but spectral localization is very poor due to the noisy baseline and high harmonic content. The NMF spectra of the (g) correlation and (h) correntropy show a large contribution at the true period. In the CNMFS the true period is in fact the maximum of the spectrum. The LS periodogram and (i) AoV periodogram (j) are shown for reference, the underlying period is not the global maximum/minimum in neither of them.

Table 4.11: Period estimation performance of the CNMFS and related methods in the synthetic data set **E-1** with different SNRs

Method	Hit rate	Localization	Sparseness
CNMFS	95.0	21.7	97.5
Correlation NMFS	82.0	23.2	98.2
CKP	95.0	0.2	0.5
CSD	74.5	0.3	26.6
PSD	68.2	1.5	67.6

Table 4.12: Period estimation performance of the CNMFS and related methods in the synthetic data set **E-2** with reduced time spans

Method	Hit rate	Localization	Sparseness
CNMFS	88.0	19.6	94.1
Correlation NMFS	85.4	19.1	92.6
CKP	75.8	0.3	0.4
CSD	76.4	0.5	21.3
PSD	73.2	2.2	52.5

the fundamental frequency. In addition, its baseline exhibits higher variance with respect to the PSD and CSD.

NMF can be used to obtain frequency representations for the slotted autocorrelation and slotted autocorrentropy. Fig. 4.19g shows the correlation NMFS, in which a strong component associated to the true period arises. The maximum of the spectrum corresponds to the closest frequency multiple. A large kernel size is used as a starting point of the dictionary, which yields a Fourier like dictionary. The updates on the kernel sizes along the iterations are negligible. Fig. 4.19h shows the proposed CNMFS. In this case the maximum of the spectrum corresponds to the true period. There is also a high contribution of the closest submultiple. By using the correntropy function and the adaptive dictionary a frequency representation that is more faithful to the characteristics of the input signal is obtained.

The Lomb-Scargle and AoV periodograms are shown in Figures 4.19i and 4.19j, respectively. The LS periodogram closely resembles the PSD of the slotted autocorrelation. In the AoV periodogram, the global minimum corresponds to the closest sub-multiple. A peak corresponding to the true period appears as the second largest of the periodogram. In contrast to the CNMFS, the AoV periodogram shows peaks at several frequency multiples and sub-multiples.

Table 4.13 shows the results obtained on the subset of 1,500 periodic MACHO lightcurves. The CNMFS and CKP obtain the highest hit rate, followed by the AoV periodogram and the LS periodogram. In most of the cases where the methods fail to get the true period a multiple is obtained instead. Both, the CNMFS and CKP obtain the same hit rate, failing in the same light curves. This agreement in hit rate is equivalent to what was seen in the results of experiment **E-1**. The periodic light curves in this set represent only three classes of variable stars, have only one oscillation mode, their periods are well-sampled with respect

Table 4.13: Period estimation performance of the CNMFS and related methods in the MA-CHO light curve data set

<b>Method</b>	Hit rate	Localization	Sparseness
CNMFS	91.0	13.4	89.1
LS periodogram	63.4	4.5	71.0
AoV periodogram	78.6	2.2	56.2
CKP	91.0	0.1	0.2

to the time span, and they were estimated through visual inspection (human expert bias). A test for the CNMFS on a larger and more heterogeneous set is currently in preparation. The CNMFS is expected to attain higher hit rates due its increased frequency resolution and sparseness.

The LS periodogram uses the Fourier basis and is the most prone to picking multiples specially in the subset of eclipsing binary light curves. In this sense the AoV periodogram is better suited for more general and non-sinusoidal periodicities. The best results in terms of spectral localization and sparseness are obtained by the CNMFS. The second best is the LS periodogram. The CKP performs poorly in terms of these metrics for the same reasons given in previous sections. The AoV periodogram has important contribution at harmonics and subharmonics of the fundamental frequency which decreases the sparseness of the representation.

The average computational time required to process one light curve using the CNMFS is  $\sim 9.4s$ <sup>11</sup>. In comparison, the LS and AoV periodogram take a tenth of a second to process one light curve (VarTools, C implementation). The gradient-descent performed to adapt the dictionary consumes 90% of the computational time. Improving the efficiency and/or reducing the computational complexity of this step is an issue that will be addressed in the future. The following strategies are considered to decrease the computational time

- To implement a more focused adaption scheme, for example to adapt only the atoms with high coefficients. This criterion should not be used in the initial iterations to avoid falling into local minima.
- To try more sophisticated optimization methods for the dictionary, such as gradient descent with line-search, conjugate gradient or second-order gradient descent.
- To port the code to highly parallel hardware and/or use more efficient software libraries.

It is expected that with these modifications, competitive computational times will be achieved.

---

<sup>11</sup>Using Matlab R2011, measured on an Intel i7-2670QM @3.10Ghz

# Chapter 5

## Conclusions

In this thesis, information theoretic criteria for light curve analysis have been presented and tested. These methods are based on the information theoretic concept of correntropy, a generalized correlation function that quantifies similarities in random processes. Correntropy supersedes the conventional autocorrelation function as it is not limited to second order moments (variance and power). Correntropy measures similarities through a kernel function, for example when the Gaussian kernel is used, correntropy incorporates information of all the even-order moments of the process. This gives correntropy a higher discriminative power and robustness to noise and outliers. The autocorrentropy function can be used to find repeating patterns within a time series, hence it is appropriate for the tasks of period or fundamental frequency estimation.

The methods proposed in this thesis focus on the tasks of periodic light curve detection, *i.e.* discriminating if a light curve is periodic or not, and light curve period estimation, *i.e.* finding the underlying period of a light curve with a high precision. Using these methods, efficient periodicity detection pipelines with solid information theoretic background have been developed. The core of these pipelines and the main contributions of this research are the slotted autocorrentropy, the correntropy kernelized periodogram (CKP) and the correntropy non-negative matrix factorization spectrum (CNMFS).

The slotted correntropy estimator is an extension of the correntropy function for unevenly sampled time series, and it is designed to estimate the period of astronomical light curves. The slotted correntropy outperformed the slotted correlation and conventional methods used in astronomy in a period estimation task on a subset of periodic variable stars from the MACHO survey. This is specially noticeable in the case of non-sinusoidal light curves (eclipsing binary stars), because the slotted correntropy does not pose any assumptions on the data and uses more information from the process (higher-order moments). One disadvantage of the slotted correntropy is that it is highly-dependent on the slot size parameter.

To overcome this dependency the CKP was proposed. The CKP is a frequency-domain metric that combines the correntropy function with a periodic kernel. The CKP is able to assess periodicity directly from the available samples of the light curve as it does not require interpolation, folding or slotting schemes. The CKP has two parameters that can be used to

adapt the metric to different noise regimes and outlier rejection conditions, and it can also be tuned to be more discriminative towards particular periodicities. An statistical test for periodicity detection using the CKP as test-statistic and surrogate light curves was proposed. Using this test the confidence of a particular period found in the periodogram can be assessed. The results show that the statistical test based on the CKP is able to discriminate periodic light curves at lower false positive rates (97% true positive rate at 0.1% false positive rate) than its competitors, including the slotted correntropy. The problem of periodic light curve detection is unbalanced in the sense that non-periodic and non-variable stars are much more abundant than the periodic ones. Fully-automated methods for periodic light curve detection must achieve for minimal false positive rates in order for their results to be useful.

The CKP was used as the core of a periodic light curve discrimination pipeline that was applied to the EROS-2 survey catalog, a database that has not been analyzed before for periodic variables. In this work 32.8 million light curves from the Large and Small Magellanic clouds were analyzed. The main result is a set of catalogs of periodic variable stars found in the Magellanic clouds, in total 121,147 and 24,855 periodic stars were found in the LMC and SMC, respectively. Light curves having multiple valid periodicities were also found and catalogued. Approximately a 0.5% of the stars in the EROS-2 survey are found to be periodic. The efficiency of the pipeline was characterized using synthetic time series modeled after the EROS-2 light curves. Through this test we estimate that the pipeline achieves a recovery rate (recall) of 93% at a precision (purity) of 95% on the real database. The pipeline was also compared to an equivalent implementation that uses the LS periodogram instead of the CKP. The results show that the CKP pipeline is more robust against the number of samples, the value of the period, the shape of the periodicity and the SNR of the light curves. This is explained by the higher-order moments included in the CKP and its adaptiveness via the kernel parameters. The periodic stars recovered by the CKP pipeline were crossmatched with the OGLE-3 catalogs, obtaining a high level of agreement in the periods and also detecting new periodic light curves not found by the OGLE team. The histogram and the distribution in the color magnitude diagram of the periodic light curves show that the results obtained are consistent with the literature. The Petersen diagram for multimodal light curves also shows consistent results. Different populations of periodic variable stars were found by inspecting the color magnitude diagram and the period detected by the pipeline. Several periodic light curves that do not fall into any known category were also found while visually inspecting the light curves. The results obtained by the pipeline provide fertile ground for stellar classification, novelty detection schemes and other higher order analyses that require the period as input. The pipeline built is fully automated and the parameters of the CKP are selected using heuristic rules in order to adapt the metrics to the data. The pipeline is computationally efficient taking 0.16 seconds to discriminate if a light curve is periodic or not, and processing the whole 32.8 million EROS-2 light curves in less than 24 hours using GPGPU cluster resources. This suggests that using newer hardware and with a few additional optimizations the pipeline may scale well for more modern and larger databases (VVV, LSST).

One disadvantage of the CKP is that it is dense, in the sense that it presents high spectral activity associated to harmonics, sub-harmonics and aliases of the underlying period. To compensate this, a spectral cleaning stage had to be added to the pipeline. This issue also difficults the tasks of discriminating light curves with more than one underlying periodic sig-

nal embedded. To overcome this and to increase the frequency resolution of the periodogram, the CNMFS was proposed. The CNMFS is obtained by decomposing the autocorrentropy estimator into a dictionary of frequency-indexed atoms using a modification of the NMF algorithm. The CNMFS takes the information of the autocorrentropy function and represents it in a sparse, localized and high-resolution frequency spectrum. The CNMFS is obtained using a simple gradient-descent based approach. An adaptation scheme for the dictionary that preserves its frequency structure was also proposed. By using an overcomplete set of periodic functions it is possible to super-resolve the spectrum, *i.e.* to push the limits on frequency resolution. The shape of the atoms is modified during the adaptation in order to improve the fitting of the correntropy function and increase the frequency resolution even further. The dictionary learning scheme selects the kernel size of the periodic kernel automatically, *i.e.* the user does not need to set this parameter anymore. Also, the selection of the kernel size is performed individually for each atom. This is more general, as different periodicities found in the correntropy function may not necessarily have the same shape. Using individual kernel sizes optimizes the fit for each component. Different dictionaries of functions can be joined in order to model different behavior present in the correntropy function. For example, by adding a Kronecker delta dictionary, impulsive noise can be easily removed from the spectrum. This shows the flexibility of the dictionary of functions approach, although one should add more atoms carefully as it increases the computational complexity of the algorithm.

The result of this iterative optimization process is a high frequency resolution spectral representation that is less affected by noise and aliasing. The CNMFS preserves the higher discrimination power of the correntropy function and translates it faithfully to the frequency domain. Contrary to the CKP and conventional spectrum estimators, the CNMFS is well localized in the fundamental frequency of the time series having less contribution from harmonics and sub-harmonics, which facilitates the task of discriminating true periodicities and detecting multiple valid periodicities. The overcompleteness of the dictionary and the adaptiveness of the atoms give the CNMFS enhanced spectral resolvability and increased robustness in low sample size and short time span scenarios. Another advantage of the CNMFS is that its parameters are adapted automatically without input from the user. One disadvantage of the CNMFS with respect to the CKP is its higher computational complexity. A more efficient computational implementation for the CNMFS is currently being developed. This new implementation considers replacing the simple gradient-descent rule for the dictionary with more sophisticated optimization methods, using fast linear algebra software libraries and porting the code to highly parallel hardware architectures (*e.g.* GPUs). A modification on the adaptation strategy that involves selective atom adaption is also under study. It is expected that with these modifications, competitive computational times for the CNMFS will be achieved.

The results presented in this thesis show that information theoretic based criteria perform better than conventional methods used in astronomy such as the LS periodogram, analysis of variance, string length and the slotted autocorrelation function (second-order methods). Including the higher-order moments of the time series into the estimation makes the proposed information-theoretic methods more robust against noise and outliers, giving them the upper hand in term of the precision of the detected periods. The proposed methods are also general as they do not pose any assumption on the underlying periodic signal (*e.g.* sum of sine-waves), and can be adapted heuristically (CKP) or automatically (CNMFS) to different periodic light

curve shapes. The proposed methods are less prone to return a harmonic, sub-harmonic or an alias of the underlying period, a usual problem with conventional methods. The results also show that the proposed methods are more robust and less dependant on the number of samples and the time span of the light curve, *i.e.* the period can be recovered even if few samples or only a short piece of the light curve is available. This suggests that these methods may outperform conventional methods for early or online periodicity discrimination on surveys that are currently operating (VVV, DECam).

Chile is a country favored by their northern skies, the clearest in the planet. Because of this, Chile is becoming an astronomical pole, attracting the most ambitious and potent astronomical observatories and facilities up to date, and also the most renowned astronomers, researchers, and institutions all over the world. A ten percent of the observation time of all these facilities is reserved for meritorious chilean researchers and institutions. The methods presented in this thesis are intended for astronomers, and particularly for researchers residing in the country, to do their science more efficiently and to make the most out of the large amounts of data to be collected in the near future.

## 5.1. Future Work

A possible extension of the methods presented in this thesis involves assessing and discriminating non-stationary behavior and quasi-periodic variability. The discrimination of periodic from non-periodic variable stars is a simplification of the real problem, as there are also stars, such as the Long Periodic Variables, that present quasi-periodic variability, *i.e.* irregular periodicities where the patterns do not repeat exactly, and also transient periodic behavior, *i.e.* periodicities that last for a given time frame and periods that change through time. In order to capture these behaviors the methods should be extended to the time-frequency domain. One alternative is to follow an spectrogram-like approach and evaluate the metrics on segments of the time series in order to reveal local periodicities in the windows. There is a high synergy between the spectrogram approach and the CNMFS, because very short time windows can be used (high time resolution) without losing much frequency resolution in comparison with Fourier-based methods. Another approach involves developing kernels that model quasi-periodic phenomena and plug them into the NMF dictionary of functions or into the kernelized periodogram. A thoughtful study on the different quasi-periodic signatures found on LPVs is required to develop these kernel functions. Another extension for the methods involves the detection of trends and tendencies that break the stationary assumption. Methods to effectively remove trends of different natures or kernels to fit the trends embedded in the light curves are also required if one wants the pipeline to work under any condition. The median of the photometric error in the light curves was used to select the kernel size of the Gaussian kernel, which has control over the observation window and the outlier rejection criterion. Making better use of individual photometric error values is also considered as an extension for the proposed methods. An alternative to be tested could be a local kernel size selection scheme, where sample pairs contribute to the estimation inversely proportional to their joint error.

Testing the proposed pipelines on larger light curve databases from modern surveys is also

considered for future work. Astronomical surveys such as the VVV, the CRTS and PAN-STARRS surpass the amount of light curves EROS-2 had in at least one order of magnitude. Improving computational efficiency and making use of more modern HPC resources is required in order to process these surveys in feasible computational times. In the near future, the LSST will push the constraint on computational time even further. These surveys are also more heterogeneous, having several frequency bands per light curve. Processing these bands in parallel and joining the results obtained from each of them is also an open challenge that should be addressed.

# Bibliography

- [1] C. Alcock *et al.*, “The MACHO project: Microlensing results from 5.7 years of lmc observations,” *The Astrophysical Journal*, vol. 542, pp. 281–307, 2000.
- [2] Y. R. Rahal *et al.*, “The EROS2 search for microlensing events towards the spiral arms: the complete seven season results,” *Astronomy & Astrophysics*, vol. 500, pp. 1027–1044, 2009.
- [3] A. Udalski, M. Kubiak, and M. Szymanski, “Optical gravitational lensing experiment. OGLE-II – the second phase of the OGLE project,” *Acta Astronomica*, vol. 47, pp. 319–344, 1997.
- [4] D. G. York *et al.*, “The sloan digital sky survey: Technical summary,” *The Astronomical Journal*, vol. 120, no. 3, p. 1579, 2000.
- [5] N. Kaiser *et al.*, “Pan-STARRS: A large synoptic survey telescope array,” *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 4836, pp. 154–164, 2002.
- [6] S. Larson *et al.*, “The CSS and SSS NEO surveys,” in *AAS/Division for Planetary Sciences Meeting Abstracts #35*, vol. 35 of *Bulletin of the American Astronomical Society*, p. 982, May 2003.
- [7] T. Diehla, “The dark energy survey camera (DECam),” *Physics Procedia*, vol. 37, no. 0, pp. 1332 – 1340, 2012. Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011).
- [8] D. Minniti *et al.*, “VISTA Variables in the Via Lactea (VVV): The public ESO near-IR variability survey of the Milky Way,” *New Astronomy*, vol. 15, pp. 433–443, July 2010.
- [9] J. Tyson and K. Borne, “Future sky surveys, new discovery frontiers,” in *Advances in Machine Learning and Data Mining for Astronomy* (M. Way, J. D. Scargle, K. Ali, and A. Srivastava, eds.), ch. 9, pp. 161–181, CRC Press, 2012.
- [10] Z. Ivezić *et al.*, “LSST: from science drivers to reference design and anticipated data products,” *ArXiv e-prints*, June 2011. Living document found at: <http://www.lsst.org/lsst/overview/>.
- [11] K. Borne, “Virtual observatories, data mining and astroinformatics,” in *Planets, Stars*

*and Stellar Systems. Astronomical Techniques, Software, and Data* (T. Oswalt and H. Bond, eds.), vol. 2, pp. 404–443, Wiley, 2013.

- [12] M. Petit, *Variable Stars*. Reading, MA: New York: Wiley, 1987.
- [13] J. Percy, *Understanding Variable Stars*. Cambridge University Press, 2007.
- [14] L. Eyer and N. Mowlavi, “Variable stars across the observational hr diagram,” *Journal of Physics: Conference Series*, vol. 118, no. 1, p. 012010, 2008.
- [15] L. Eyer, “First Thoughts about Variable Star Analysis,” *Baltic Astronomy*, vol. 8, pp. 321–324, 1999.
- [16] J. Debosscher *et al.*, “Automated supervised classification of variable stars. I methodology,” *Astronomy & Astrophysics*, vol. 475, pp. 1159–1183, 2007.
- [17] G. Wachman, R. Khardon, P. Protopapas, and C. Alcock, “Kernels for periodic time series arising in astronomy,” in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, (Bled, Slovenia), pp. 489–505, 2009.
- [18] D. Popper, “Stellar masses,” *Annual review of astronomy and astrophysics*, vol. 18, pp. 115–164, 1980.
- [19] B. D. Warner, *A Practical Guide to Lightcurve Photometry and Analysis*. New York, NY: Springer, 2006.
- [20] N. Ball, “Techniques for massive-data machine learning in astronomy,” in *Statistical Challenges in Modern Astronomy V* (E. Feigelson and G. Babu, eds.), vol. 209 of *Lecture Notes in Statistics*, ch. 44, pp. 473–478, Wiley, 2013.
- [21] N. Samus *et al.*, “General catalogue of variable stars,” 2012. (Samus+ 2007-2012), VizieR On-line Data Catalog: B/gcvs.
- [22] A. Walker, “Distances to local group galaxies,” in *Stellar Candles for the Extragalactic Distance Scale* (D. Alloin and W. Gieren, eds.), vol. 635 of *Lecture Notes in Physics*, pp. 265–279, Springer Berlin Heidelberg, 2003.
- [23] F. E. Olivares and M. Hamuy, *Core-Collapse Supernovae As Standard Candles*. Lambert Academic Publishing, 2011.
- [24] D. Graczyk, G. Pietrzyński, B. Pilecki, I. B. Thompson, W. Gieren, P. Konorski, A. Udalski, and I. Soszyński, “The distance to the Small Magellanic Cloud from eclipsing binaries,” in *IAU Symposium* (R. de Grijs, ed.), vol. 289 of *IAU Symposium*, pp. 222–225, Feb. 2013.
- [25] J. W. Richards *et al.*, “On machine-learned classification of variable stars with sparse and noisy time-series data,” *The Astrophysical Journal*, vol. 733, no. 1, p. 10, 2011.

- [26] P. Protopapas *et al.*, “Finding outlier light curves in catalogues of periodic variable stars,” *The Royal Astronomical Society, Monthly Notices*, vol. 369, pp. 677–696, 2006.
- [27] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, “Finding Anomalous Periodic Time Series: An Application to Catalogs of Periodic Variable Stars,” *Machine Learning*, vol. 74, pp. 281–313, May 2009.
- [28] M. Mayor and D. Queloz, “A Jupiter-mass companion to a solar-type star,” *Nature*, vol. 378, pp. 355–359, 1995.
- [29] J. T. Wright and B. S. Gaudi, “Exoplanet Detection Methods,” in *Planets, Stars and Stellar Systems. Volume 3: Solar and Stellar Planetary Systems* (T. D. Oswalt, L. M. French, and P. Kalas, eds.), p. 489, Springer, 2013.
- [30] C. Alcock, C. W. Akerlof, R. A. Allsman, T. S. Axelrod, D. P. Bennett, S. Chan, K. H. Cook, K. C. Freeman, K. Griest, S. L. Marshall, H.-S. Park, S. Perlmutter, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, and W. Sutherland, “Possible gravitational microlensing of a star in the Large Magellanic Cloud,” *Nature*, vol. 365, pp. 621–623, Oct. 1993.
- [31] J. Principe, *Information Theoretic Learning: Renyi’s Entropy and Kernel Perspectives*. New York: Springer Verlag, 2010.
- [32] G. M. Jenkins and D. G. Watts, *Spectral analysis and its applications*. Holden-day, 1968.
- [33] N. Lomb, “Least-squares frequency analysis of unequally spaced data,” *Astrophysics and Space Science*, vol. 39, pp. 447–462, 1976.
- [34] J. Scargle, “Studies in astronomical time series analysis. ii. statistical aspects of spectral analysis of unevenly spaced data,” *The Astrophysical Journal*, vol. 263, pp. 835–853, 1982.
- [35] W. Press and G. Rybicki, “Fast algorithm for spectral analysis of unevenly sampled data,” *The Astrophysical Journal*, vol. 338, pp. 277–280, 1989.
- [36] Zechmeister, M. and Kürster, M., “The generalised lomb-scargle periodogram,” *Astronomy & Astrophysics*, vol. 496, no. 2, pp. 577–584, 2009.
- [37] D. Marquardt and S. Acuff, *Direct Quadratic Spectrum Estimation with Irregularly Spaced Data*, pp. 211–223. Springer-Verlag, 1984.
- [38] L. Eyer and P. Bartholdi, “Variable stars: which nyquist frequency?,” *Astronomy and Astrophysics Supplement Series*, vol. 135, pp. 1–3, 1998.
- [39] C. de Boor, *A practical guide to splines*. New York: Springer Verlag, 1978.
- [40] W. Mayo, “Spectrum measurements with laser velocimeters,” in *Proceedings of dynamic flow conference, DISA Elektronik A/S DK-2740*, (Skoolunder, Denmark), pp. 851–868,

1978.

- [41] M. Tummers and D. Passchier, “Spectral estimation using a variable window and the slotting technique with local normalization,” *Measurement Science and Technology*, vol. 7, pp. 1541–1546, 1996.
- [42] H. Nobach, E. Müller, and C. Tropea, “Efficient estimation of power spectral density from laser doppler anemometer data,” *Experiments in Fluids*, vol. 24, no. 5, pp. 499–509, 1998.
- [43] H. Nobach, “Local time estimation for the slotted correlation function of randomly sampled lda data,” *Experiments in Fluids*, vol. 32, no. 3, pp. 337–345, 2002.
- [44] R. A. Edelson and J. Krolik, “The discrete correlation function: A new method for analyzing unevenly sampled variability data,” *The Astrophysical Journal*, vol. 333, pp. 646–659, 1988.
- [45] L. Benedict, H. Nobach, and C. Tropea, “Benchmark tests for the estimation of power spectra from lda signals,” in *Proc. 9th Int. Symp. on Applications of Laser Technology to Fluid Mechanics (Lisbon)*, 1998.
- [46] P. Stoica and N. Sandgren, “Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approaches,” *Digital Signal Processing*, vol. 16, no. 6, pp. 712 – 734, 2006.
- [47] O. Bjørnstad and W. Falck, “Nonparametric spatial covariance functions: Estimation and testing,” *Environmental and Ecological Statistics*, vol. 8, no. 1, pp. 53–70, 2001.
- [48] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, “Comparison of correlation analysis techniques for irregularly sampled time series.,” *Nonlinear Processes in Geophysics*, vol. 18, no. 3, 2011.
- [49] M. Dworetzky, “A period finding method for sparse randomly spaced observations,” *Monthly Notices of the Royal Astronomical Society*, vol. 203, pp. 917–923, 1983.
- [50] D. Clarke, “String/rope length methods using the lafler-kinman statistic,” *Astronomy & Astrophysics*, vol. 386, pp. 763–774, 2002.
- [51] A. Schwarzenberg-Czerny, “On the advantage of using analysis of variance for period search,” *Monthly Notices of the Royal Astronomical Society*, vol. 241, pp. 153–165, 1989.
- [52] R. Stellingwerf, “Period determination using phase dispersion minimization,” *The Astrophysical Journal*, vol. 224, pp. 953–960, 1978.
- [53] A. Derekas, L.L.Kiss, and T.R.Bedding, “Eclipsing binaries in MACHO database: new periods and classifications for 3031 systems in the large magellanic cloud,” *The Astrophysical Journal*, vol. 663, pp. 249–257, 2007.

- [54] P. M. Cincotta, A. Helmi, M. Mendez, J. A. Nunez, and H. Vucetich, “A search for periodicity using the shannon entropy,” *Royal Astronomical Society, Monthly Notices*, vol. 302, pp. 582–586, 1999.
- [55] M. J. Graham, A. J. Drake, S. G. Djorgovski, A. A. Mahabal, and C. Donalek, “Using conditional entropy to identify periodicity,” *Monthly Notices of the Royal Astronomical Society*, vol. 434, pp. 2629–2635, Sept. 2013.
- [56] R. Fisher, “Tests of significance in harmonic analysis.,” *Proceedings of the Royal Society*, vol. 125, pp. 54–59, 1929.
- [57] C. Koen, “Significance testing of periodogram ordinates,” *The Astrophysical journal*, vol. 348, pp. 700–702, 1990.
- [58] M. Ahdesmaki, H. Lahdesmaki, and O. Yli-Harja, “Robust fisher’s test for periodicity detection in noisy biological time series,” in *Genomic Signal Processing and Statistics, 2007. GENSIPS 2007. IEEE International Workshop on*, pp. 1–4, 2007.
- [59] M. Artis, M. Hoffmann, D. Nachane, and J. Toro, “The detection of hidden periodicities: A comparison of alternative methods,” Tech. Rep. ECO2004/10, European University Institute, 2004.
- [60] A. Schmitz and T. Schreiber, “Testing for nonlinearity in unevenly sampled time series,” *Phys. Rev. E*, vol. 59, no. 4, pp. 4044–4047, 1999.
- [61] T. Schreiber and A. Schmitz, “Surrogate time series,” *Physica D: Nonlinear Phenomena*, vol. 142, pp. 346–382, 1999.
- [62] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, “Testing for nonlinearity in time series: the method of surrogate data,” *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 77 – 94, 1992.
- [63] J. Theiler and D. Prichard, “Constrained-realization monte-carlo method for hypothesis testing,” *Physica D*, vol. 94, pp. 221–235, 1995.
- [64] P. Buhlmann, “Bootstraps for time series,” *Statistical Science*, vol. 17, no. 1, pp. 52–72, 1999.
- [65] D. N. Politis and J. P. Romano, “The stationary bootstrap,” *Journal of the American Statistical Association*, vol. 89, no. 498, pp. 1303–1313, 1994.
- [66] J. S. Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [67] B. Boser, I. Guyon, and V. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the 5th Annual ACM Workshop on COLT*, (Pittsburgh, USA), pp. 144–152, 1992.
- [68] J. Mercer, “Functions of positive and negative type, and their connection with the the-

- ory of integral equations,” *Philosophical Transactions of the Royal Society of London*, vol. 209, pp. 415–446, 1909.
- [69] M. Buhmann, *Radial Basis Functions: Theory and Implementation*. Cambridge University Press, 2003.
- [70] B. W. Silverman, *Density estimation for statistics and data analysis*. Chapman & Hall, 1 ed., 1986.
- [71] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.
- [72] M. Michalak, “Time series prediction using periodic kernels,” in *Computer Recognition Systems 4*, pp. 136–146, Berlin: Springer Verlag, 2010.
- [73] D. Mckay, *Introduction to Gaussian Processes*, vol. 168, pp. 133–165. Springer, Berlin, 1998.
- [74] S. Chen and D. Donoho, “Basis pursuit,” in *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, vol. 1, pp. 41–44 vol.1, Oct 1994.
- [75] S. S. Chen, D. L. Donoho, Michael, and A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.
- [76] S. S. Chen and D. Donoho, “Application of basis pursuit in spectrum estimation,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 3, pp. 1865–1868 vol.3, 1998.
- [77] H. Hindi, “A tutorial on convex optimization ii: duality and interior point methods,” in *American Control Conference, 2006*, pp. 11 pp.–, June 2006.
- [78] Y. Zhang, “Solving large-scale linear programs by interior-point methods under the matlab environment†,” *Optimization Methods and Software*, vol. 10, no. 1, pp. 1–31, 1998.
- [79] P. Carbonetto, “Intuition behind primal-dual interior-point methods for linear and quadratic programming.” Available online at: [www.cs.ubc.ca/~pcarbo/lp.pdf](http://www.cs.ubc.ca/~pcarbo/lp.pdf).
- [80] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [81] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.
- [82] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

- [83] Y.-X. Wang and Y.-J. Zhang, “Nonnegative matrix factorization: A comprehensive review,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 6, pp. 1336–1353, 2013.
- [84] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *In NIPS*, pp. 556–562, MIT Press, 2001.
- [85] Z. Yang and E. Oja, “Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization,” *Trans. Neur. Netw.*, vol. 22, pp. 1878–1891, Dec. 2011.
- [86] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.
- [87] S. A. Vavasis, “On the complexity of nonnegative matrix factorization,” *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.
- [88] C. Lin, “Projected Gradient Methods for Nonnegative Matrix Factorization,” *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.
- [89] J. Kim and H. Park, “Fast nonnegative matrix factorization: An active-set-like method and comparisons,” *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2008.
- [90] C.-J. Lin, “On the convergence of multiplicative update algorithms for nonnegative matrix factorization,” *Neural Networks, IEEE Transactions on*, vol. 18, no. 6, pp. 1589–1596, 2007.
- [91] M. Merritt and Y. Zhang, “Interior-Point Gradient Method for Large-Scale Totally Nonnegative Least Squares Problems,” *Journal of Optimization Theory and Applications*, vol. 126, pp. 191–202, 2005.
- [92] C. Nikias and J. Mendel, “Signal processing with higher-order spectra,” *Signal Processing Magazine, IEEE*, vol. 10, no. 3, pp. 10–37, 1993.
- [93] G. Scarano, R. Cusani, and A. Laurenti, “Spectral analysis by mixtures of higher order moments,” in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, vol. 5, pp. 2395–2398, 1990.
- [94] E. Parzen, “On the estimation of a probability density function and the mode,” *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.
- [95] I. Santamaria, P. Pokharel, and J. Principe, “Generalized correlation function: Definition, properties, and application to blind equalization,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006.
- [96] J. Xu and J. Principe, “A pitch detector based on a generalized correlation function,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1420–1432, 2008.

- [97] R. Li, W. Liu, and J. C. Principe, “A unifying criterion for instantaneous blind source separation based on correntropy,” *IEEE Transactions on Signal Processing*, vol. 87, no. 8, pp. 1872–1881, 2007.
- [98] A. Gunduz and J. C. Principe, “Correntropy as a novel measure for nonlinearity tests,” *IEEE Transactions on Signal Processing*, vol. 89, no. 1, pp. 14–23, 2009.
- [99] P. Estévez, P. Huijse, P. Zegers, J. C. Principe, and P. Protopapas, “Period detection in light curves from astronomical objects using correntropy,” in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, (Barcelona, Spain), pp. 1–7, 2010.
- [100] P. Huijse, P. A. Estevez, P. Zegers, J. C. Principe, and P. Protopapas, “Period estimation in astronomical time series using slotted correntropy,” *IEEE Signal Processing Letters*, vol. 18, no. 6, pp. 371–374, 2011.
- [101] P. Huijse, P. A. Estévez, P. Protopapas, P. Zegers, and J. C. Príncipe, “An information theoretic algorithm for finding periodicities in stellar light curves,” *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5135–5145, 2012.
- [102] P. Protopapas, P. Huijse, P. A. Estévez, J. C. Príncipe, P. Zegers, and J. B. Marquette, “A novel, fully automated pipeline for period estimation in the EROS-2 data set.” Manuscript submitted to A&J, 2014.
- [103] A. Bowley, *Elements of Statistics*. New York: Scribner, 1920.
- [104] P. Moskalik, “Multi-periodic Oscillations in Cepheids and RR Lyrae-Type Stars,” in *Advances in Solid State Physics* (J. C. Suárez, R. Garrido, L. A. Balona, and J. Christensen-Dalsgaard, eds.), vol. 31 of *Advances in Solid State Physics*, p. 103, 2013.
- [105] NVIDIA, *CUDA C Programming Guide version 4.2*. NVIDIA, 2012.
- [106] V. Kindratenko and P. Trancoso, “Trends in high-performance computing,” *Computing in Science and Engineering*, vol. 13, no. 3, pp. 92–95, 2011.
- [107] NVIDIA, *Tuning CUDA applications for Fermi*. 2012.
- [108] P. Tisserand *et al.*, “Limits on the macho content of the galactic halo from the EROS-2 survey of the magellanic clouds,” *Astronomy & Astrophysics*, vol. 496, pp. 387–404, 2007.
- [109] I. Soszynski, R. Poleski, A. Udalski, M. K. Szymanski, M. Kubiak, G. Pietrzynski, L. Wyrzykowski, O. Szewczyk, and K. Ulaczyk, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. I. Classical Cepheids in the Large Magellanic Cloud,” *Acta Astronomica*, vol. 58, pp. 163–185, Sept. 2008.
- [110] J. D. Hartman, B. S. Gaudi, M. J. Holman, B. A. McLeod, K. Z. Stanek, J. A. Barranco, M. H. Pinsonneault, and J. S. Kalirai, “Deep MMT transit survey of the open cluster M37. II variable stars,” *The Astrophysical Journal*, vol. 675, pp. 1254–1277, 2008.

Software available online at: <http://www.cfa.harvard.edu/~jhartman/vartools/>.

- [111] P. Reegen, “Sigspec i. frequency- and phase-resolved significance in fourier space,” *Astronomy & Astrophysics*, vol. 467, pp. 1353–1371, 2007.
- [112] J. Devor, “Solutions for 10,000 eclipsing binaries in the bulge fields of OGLE II using DEBiL,” *The Astrophysical Journal*, vol. 628, pp. 411–425, 2005.
- [113] K. H. Cook, C. Alcock, H. A. Allsman, T. S. Axelrod, K. C. Freeman, B. A. Peterson, P. J. Quinn, A. W. Rodgers, D. P. Bennett, J. Reimann, K. Griest, S. L. Marshall, M. R. Pratt, C. W. Stubbs, W. Sutherland, and D. L. Welch, “Variable Stars in the MACHO Collaboration Database,” in *IAU Colloq. 155: Astrophysical Applications of Stellar Pulsation* (R. S. Stobie and P. A. Whitelock, eds.), vol. 83 of *Astronomical Society of the Pacific Conference Series*, p. 221, 1995.
- [114] M. Spano, N. Mowlavi, L. Eyer, and G. Burki, “Variability Morphologies in the Color-Magnitude Diagram Searching for Secular Variability,” in *American Institute of Physics Conference Series* (J. A. Guzik and P. A. Bradley, eds.), vol. 1170 of *American Institute of Physics Conference Series*, pp. 324–326, Sept. 2009.
- [115] I. Soszyński, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, and K. Ulaczyk, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VII. Classical Cepheids in the Small Magellanic Cloud,” *Acta Astronomica*, vol. 60, pp. 17–39, Mar. 2010.
- [116] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. II. Type II Cepheids and Anomalous Cepheids in the Large Magellanic Cloud,” *Acta Astronomica*, vol. 58, p. 293, Dec. 2008.
- [117] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, and R. Poleski, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VIII. Type II Cepheids in the Small Magellanic Cloud,” *Acta Astronomica*, vol. 60, pp. 91–107, June 2010.
- [118] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. III. RR Lyrae Stars in the Large Magellanic Cloud,” *Acta Astronomica*, vol. 59, pp. 1–18, Mar. 2009.
- [119] I. Soszyński, A. Udalski, M. K. Szymański, J. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, and R. Poleski, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IX. RR Lyr Stars in the Small Magellanic Cloud,” *Acta Astronomica*, vol. 60, pp. 165–178, Sept. 2010.
- [120] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IV. Long-Period Variables in the Large Magellanic Cloud,” *Acta Astronomica*, vol. 59, pp. 239–253, Sept. 2009.

- [121] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, S. Kozłowski, and P. Pietrukowicz, “The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XIII. Long-Period Variables in the Small Magellanic Cloud,” *Acta Astronomica*, vol. 61, pp. 217–230, Sept. 2011.
- [122] J. B. Marquette, J. P. Beaulieu, J. R. Buchler, R. Szabó, P. Tisserand, S. Belghith, P. Fouqué, É. Lesquoy, A. Milsztajn, A. Schwarzenberg-Czerny, C. Afonso, J. N. Albert, J. Andersen, R. Ansari, É. Aubourg, P. Bareyre, X. Charlot, C. Coutures, R. Ferlet, J. F. Glicenstein, B. Goldman, A. Gould, D. Graff, M. Gros, J. Haïssinski, C. Hamadache, J. de Kat, L. Le Guillou, C. Loup, C. Magneville, É. Maurice, A. Maury, M. Moniez, N. Palanque-Delabrouille, O. Perdereau, Y. R. Rahal, J. Rich, M. Spiro, and A. Vidal-Madjar, “The beat Cepheids in the Magellanic Clouds: an analysis from the EROS-2 database,” *Astronomy and Astrophysics*, vol. 495, pp. 249–256, Feb. 2009.
- [123] R. Smolec, I. Soszyński, P. Moskalik, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, S. Kozłowski, and P. Pietrukowicz, “Discovery of period doubling in BL Herculis stars of the OGLE survey. Observations and theoretical models,” *Monthly Notices of the Royal Astronomical Society*, vol. 419, pp. 2407–2423, Jan. 2012.
- [124] A. C. Fabian, “Serendipity in Astronomy,” *ArXiv e-prints*, Aug. 2009.

# Appendix A

## Generation of a synthetic light curve database

Assessing the performance of a periodicity discrimination method when only unlabeled or untested data is available is not an easy task. Perhaps the most straightforward approach in this cases is to grab a small fraction of the database and use the skills of a human expert to manually inspect the light curves. This approach is time-consuming, expensive and also prone to errors. Additionally, if the selected fraction is not large enough then the manually inspected dataset might not be representative of the whole.

In this Appendix, guidelines to construct a synthetic light curve dataset that captures the characteristic of a given database are presented. The synthetic light curve dataset can then be used to calibrate parameters of the method, assess the completeness and efficiency of the pipeline, and obtaining estimates of the output of the method in the real database. This method was used to calibrate a pipeline for periodicity discrimination on the EROS-2 database [102]. The synthetic light curves are generated having different periods, number of samples, kernel sizes and Signal-to-Noise ratio (SNR). Error bars and time instants from EROS-2 light curves are used to generate the synthetic light curves, hence preserving the error and sampling distribution.

### A procedure to generate periodic synthetic light curves

The periodic synthetic light curves are generated using a multivariate Gaussian generative model with a covariance matrix similar to the periodic kernel in Eq. 2.31. To generate a periodic synthetic light curve, with period  $P$ , signal-to-noise ratio  $S$ , and smoothness  $\sigma$  we follow the procedure below.

1. Randomly select a light curve from the database and extract its time instants  $\{t_i\}$  and error bars  $\{e_i\}$ . This defines the number of samples,  $N$ , of the generated lightcurve.
2. Use the time instants  $\{t_i\}$ , period  $P$ , smoothness  $\sigma$  and generate an  $N \times N$  covariance matrix as,

$$\Sigma_1(i, j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{2 \sin^2(\pi(t_i - t_j)/P)}{\sigma^2}\right).$$

3. Generate a random periodic vector,  $Y_s$ , of length  $N$  using a multivariate normal random generator with  $N \times 1$  zero mean vector and  $\Sigma_1$  covariance matrix.
4. Use the error bars to generate a  $N \times N$  diagonal covariance matrix with diagonal elements,

$$\Sigma_2(i, i) = e_i^2$$

5. Generate a random noise vector  $Y_n$  of length  $N$  using a multivariate normal random generator with a  $N \times 1$  zero mean vector and  $\Sigma_2$  covariance matrix.
6. The synthetic light curve  $Y$  is obtained by summing the noise vector and the signal vector as follows

$$Y = S \frac{\text{med}(e_i)}{0.7413 \text{ iqr}(Y_s)} Y_s + Y_n, \quad (\text{A.1})$$

where  $S$  is the desired signal-to-noise ratio,  $\text{med}$  is the median function and  $\text{iqr}$  is the interquartile range. Note that the resulting light curve has signal-to-noise ratio  $S$  by construction.

For our purpose we generated a set of 10,000 synthetic periodic lightcurves, generated using the following parameter ranges,

- Ten linearly spaced values for  $\sigma$  in the range  $[0.1, 0.6]$ .
- Twenty logarithmically spaced values for  $P$  in the range  $[0.4, 1000]$  days.
- Ten values for  $S$  extracted from the distribution of the signal-to-noise ratio of EROS-2 lightcurves.

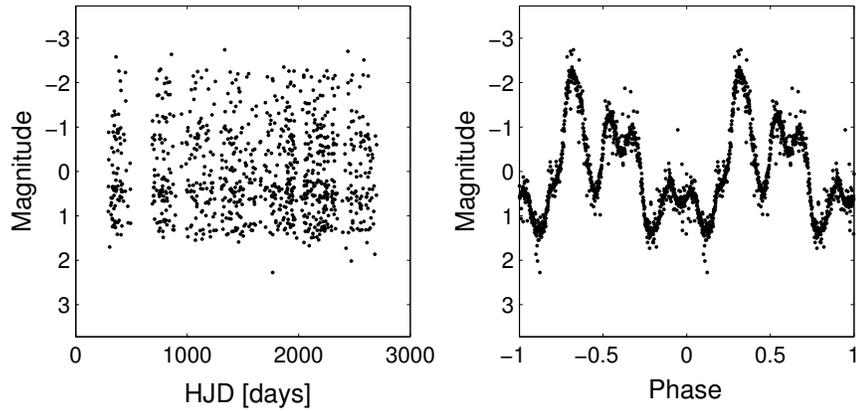
Five synthetic lightcurves are generated for each combination of  $S$ ,  $P$  and  $\sigma$ .

We present examples of the synthetic light curves generated using this procedure in Fig A.1. Fig A.1a shows a synthetic light curve with a period of 2.432 days, a smoothness value of 0.2 and a SNR of 10. Using a low smoothness value yields a shape with many features. Due to the high SNR the periodicity is very clear. Fig A.1b shows a synthetic light curve with a period of 10.42 days, smoothness of 0.5 and SNR of 4. In this case, a higher  $\sigma$  value yields a smoother shape as seen in the folded light curve. Fig A.1c shows a synthetic light curve with a period of 154 days, smoothness of 0.4 and SNR of 2.

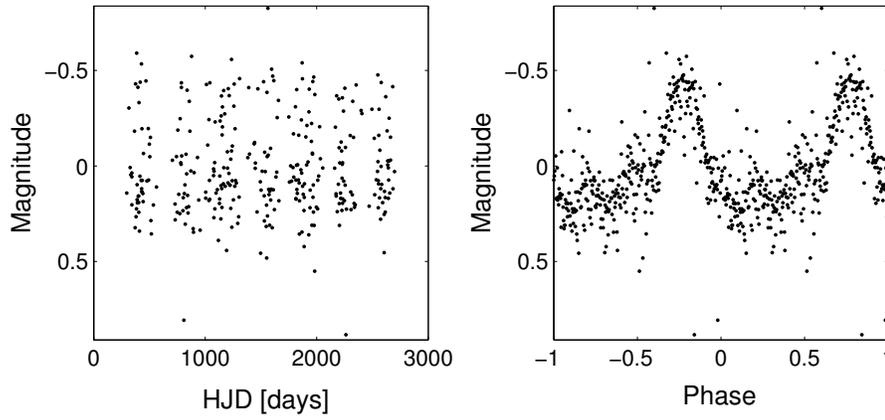
### **A procedure to generate non-periodic light curves**

The non-periodic synthetic light curves are generated using block-bootstrap surrogates [60, 61, 64]. The procedure to generate a non-periodic synthetic light curve is as follows

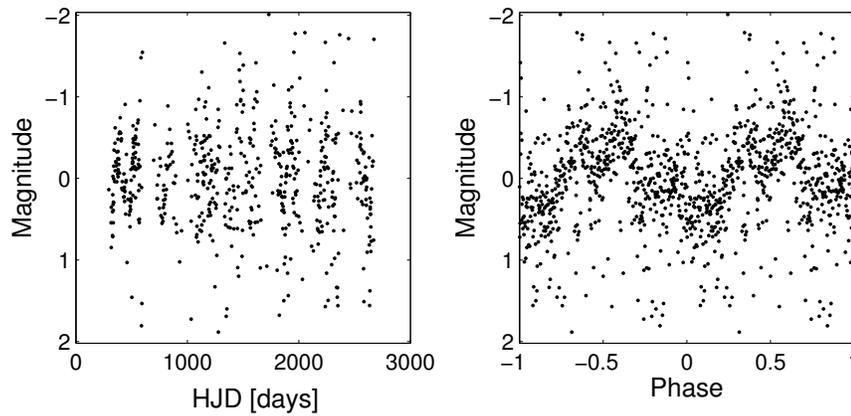
1. Randomly select a light curve and extract its time instants  $\{t_i\}$  and error bars  $\{e_i\}$ . This defines the number of samples  $N$  of the generated lightcurve.
2. Compute slotted autocorrelation function (ACF) [44] of the lightcurve.
3. Find the time lag associated to the ACF value of  $\exp(-1)$ , this time lag is used as the block length (BL) for the block bootstrap method below.



(a)



(b)



(c)

Figure A.1: Examples of synthetic periodic light curves. (a) light curve created using  $P=2.432\text{d}$ ,  $\sigma_t = 0.2$ ,  $\text{SNR} = 10$  and  $N= 642$ . (b) light curve created using  $P= 10.24\text{d}$ ,  $\sigma_t = 0.5$ ,  $\text{SNR} = 4$  and  $N= 342$ . (c) light curve created using  $P= 154\text{d}$ ,  $\sigma_t = 0.4$ ,  $\text{SNR} = 2$  and  $N= 932$ .

4. Until at least  $N$  magnitude values have been created, do

- a) Randomly select the block starting point  $i_s$ , such that  $i_s \in [1, N - N']$ . Find  $N'$  as the last light curve sample that complies with

$$t(N) - t(N') > BL$$

- b) Find the end point of the block  $i_e$  as the first time instant that complies with

$$t(i_e + 1) - t(i_s) > BL$$

- c) Grab the time instants, magnitudes, and error bars of the original light curve segment in  $[i_s, i_e + 1]$ .
- d) Subtract the initial time  $t_{i_s}$  to the selected time instants. After this the block starts at zero days.
- e) Add the time from the previous block  $t_{PB}$  to the selected time instants ( $t_{PB} = 0$  for the first block). After this the block starts where the last block ended.
- f) Update  $t_{PB} = t(i_e + 1)$ . Delete the time instant, magnitude and error bar of sample  $i_e + 1$  from the block.
- g) Add the newly constructed block to the surrogate.

For each EROS-2 light curve selected, ten surrogates were created. Ten thousand EROS-2 lightcurves were used to create a training set of 100,000 non-periodic synthetic light curves. To demonstrate that the resulting surrogates are not periodic and retain the same spectra characteristics as the originals light curves, we perform the procedure described above with a light curve of a periodic star. Fig A.2a shows EROS-2 lightcurve lm0090127524 folded with a period of 0.337443 days. The associated CKP value is 2.7424. The block bootstrap method was used to create a non-periodic synthetic light curve. Fig. A.2b shows the slotted ACF and the block length selected for this light curve is 3.67 days. Ten surrogates are generated using the procedure described above. Fig. A.2c shows one of the surrogates. The surrogate is folded with its best period and clearly the periodicity of the original light curve is not retained by the surrogate.

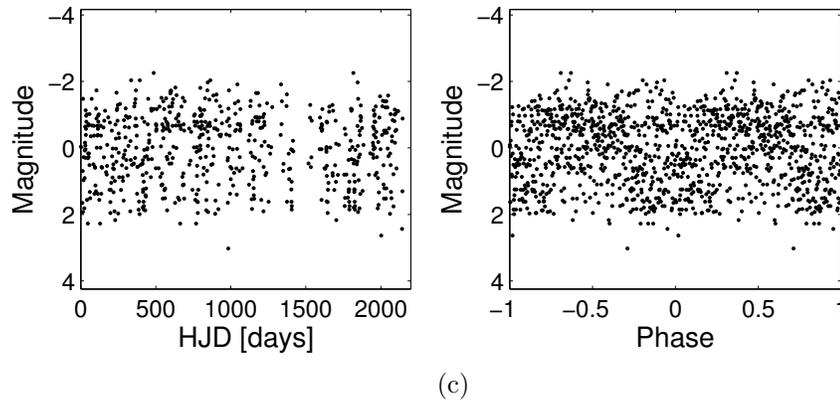
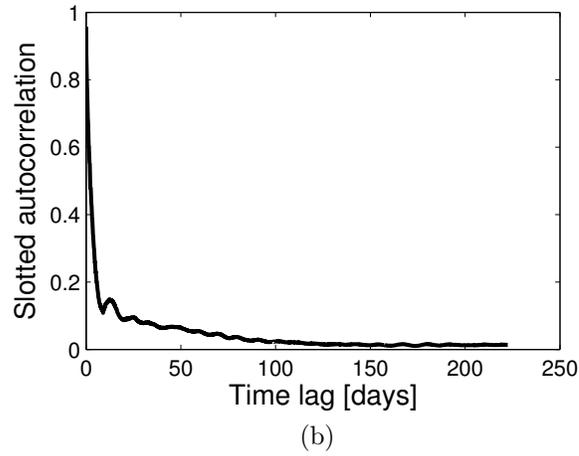
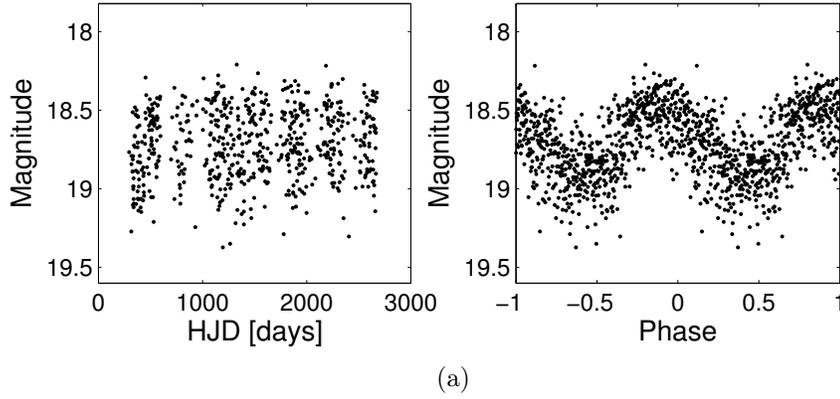


Figure A.2: (a) Periodic light curve EROS-2 Im0090127524 folded with the period of 0.337443 days, this period has a CKP value of 2.7424. (b) Slotted autocorrelation function of light curve Im0090127524. Using the slotted ACF, a window length of 3.67 days is selected to create the surrogates. (c) A surrogate created from Im0090127524. The CKP value of the surrogate is 0.4532, which is below the corresponding periodicity threshold.

# Appendix B

## Periodic light curves found in the EROS-2 survey data from the Large and Small Magellanic clouds

Using the underlying period and the epoch folding technique the shape of the periodicity hidden in the light curve is revealed. Analyzing the morphology of periodic light curves, astronomers gain insight on the phenomena behind their variability. The shape of the light curve is also a distinctive and strong feature to discriminate between different classes of periodic variable stars [16, 17]. Precise estimations of the periods are required in order to accurately classify periodic variable stars.

In this Section, folded light curves of periodic variable stars found in the EROS-2 survey data using the pipeline described in Section 3.3.7 are shown. Figure B.1 show examples of Cepheid variables, bright radial pulsators with periods ranging from 1 to 50 days. Figure B.2 contain examples of RR Lyrae variable, pulsating variable that are older and dimmer than Cepheids, with periods shorter than a day. Figure B.3 show examples of Long Period Variables, pulsating giant stars characterized by their semi-regular variability and long periods. The light curves in Figure B.4 correspond to eclipsing binary stars, binary systems whose orbital plane is aligned with the Earth. The alternating drops in brightness correspond to the mutual eclipses performed by the system as observed from Earth. These examples were classified by visually inspecting the folded light curves using the periods obtained by the proposed methods. The periods and the distinctive shapes revealed in the phase diagrams can be used to train machine learning algorithms in order to classify periodic variable stars autonomously [16, 17]. Stars whose periodic behavior do not fall in one of the known variable star categories were also found and are presented in Figure B.5. Further astronomical analysis is required in order to understand the mechanism behind these light curves. Unsupervised methods for novelty detection could be used to discriminate novel and rare periodic phenomena automatically [124, 26, 27]. The results obtained by the methods for periodicity detection proposed in this thesis open the door for large-scale variable star analyses such as stellar classification and the study of rare stellar phenomena.

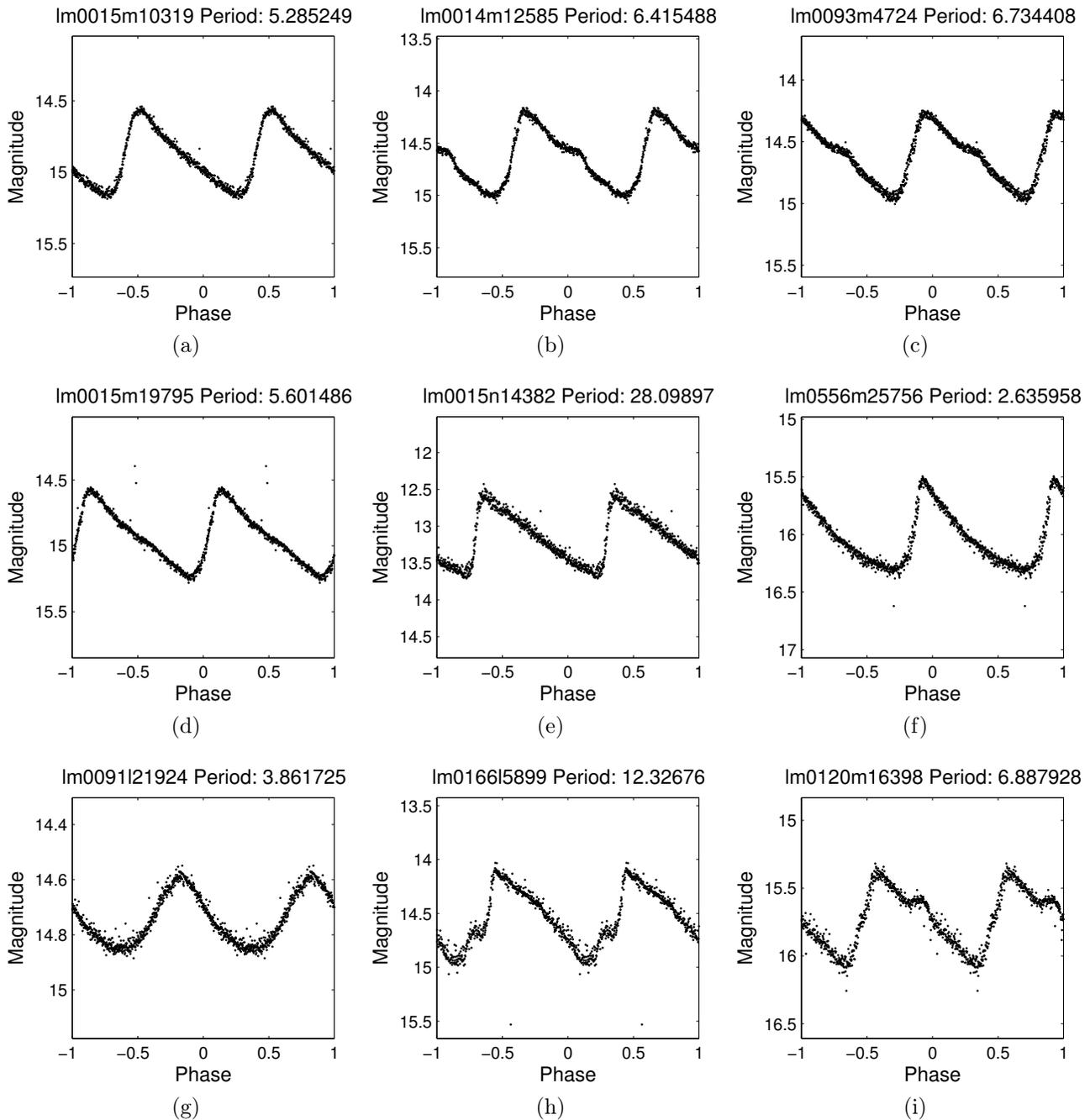


Figure B.1: Examples of EROS-2 periodic lightcurves folded with their estimated period. These light curves correspond to Cepheid variables, very bright pulsating stars with periods between 1 and 50 days.

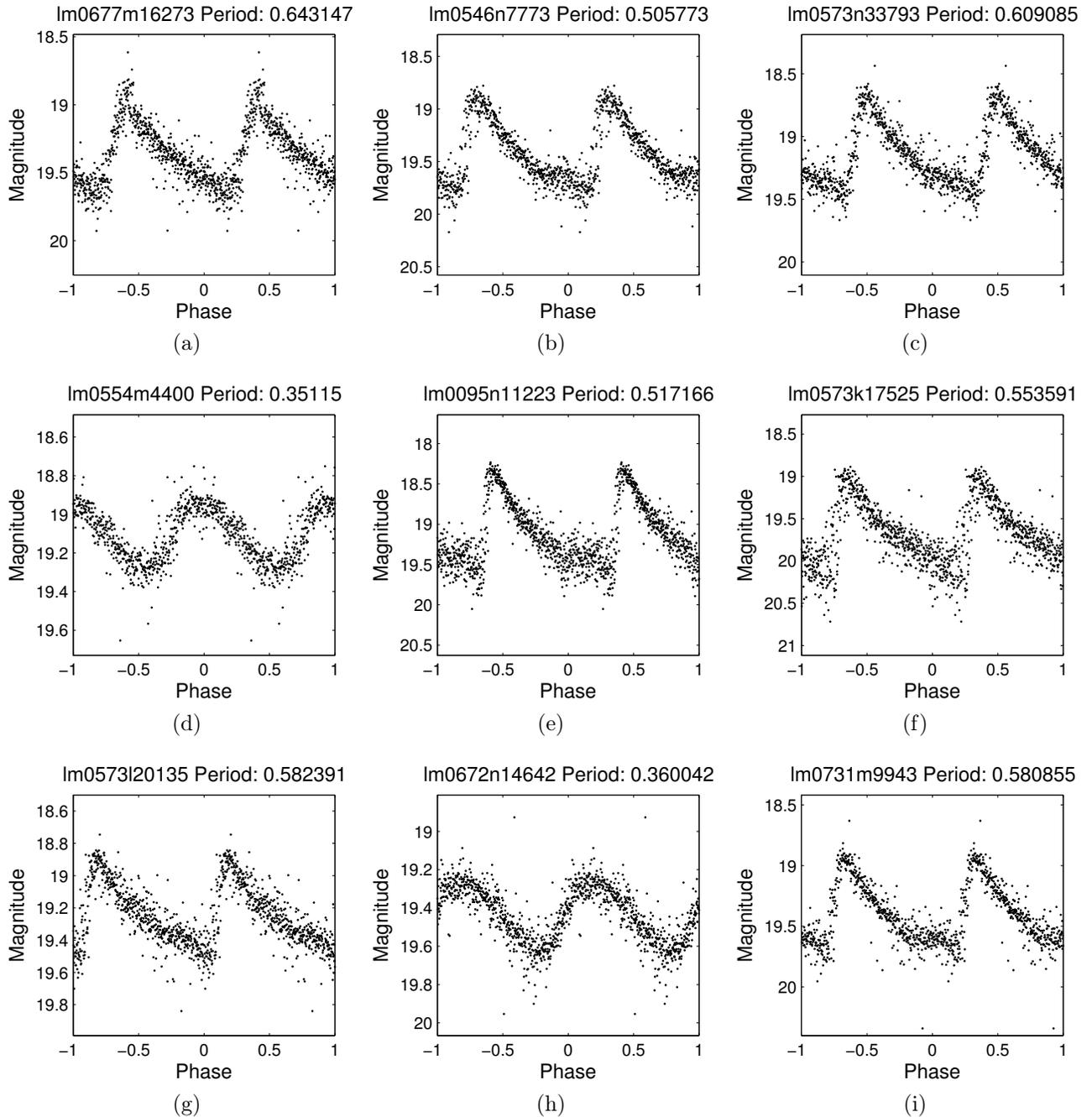


Figure B.2: Examples of EROS-2 periodic lightcurves folded with their estimated period. These light curves correspond to RR Lyrae variables, radial pulsators that are older and dimmer than Cepheids with periods shorter than 1 day.

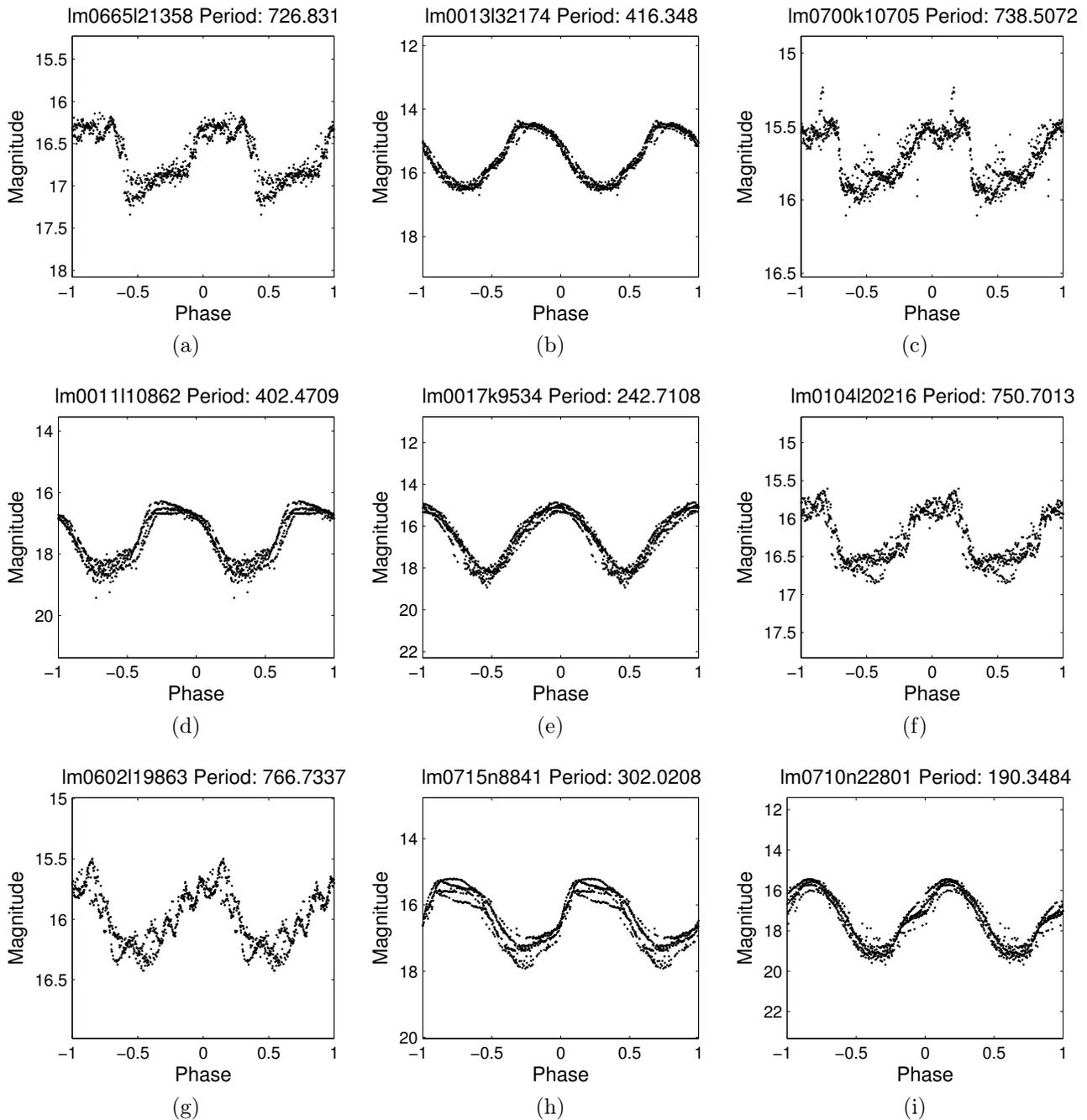


Figure B.3: Examples of EROS-2 periodic lightcurves folded with their estimated period. These light curves correspond to long period variables, pulsating giant variable stars characterized by their long periods and their semi-regular periodicities.

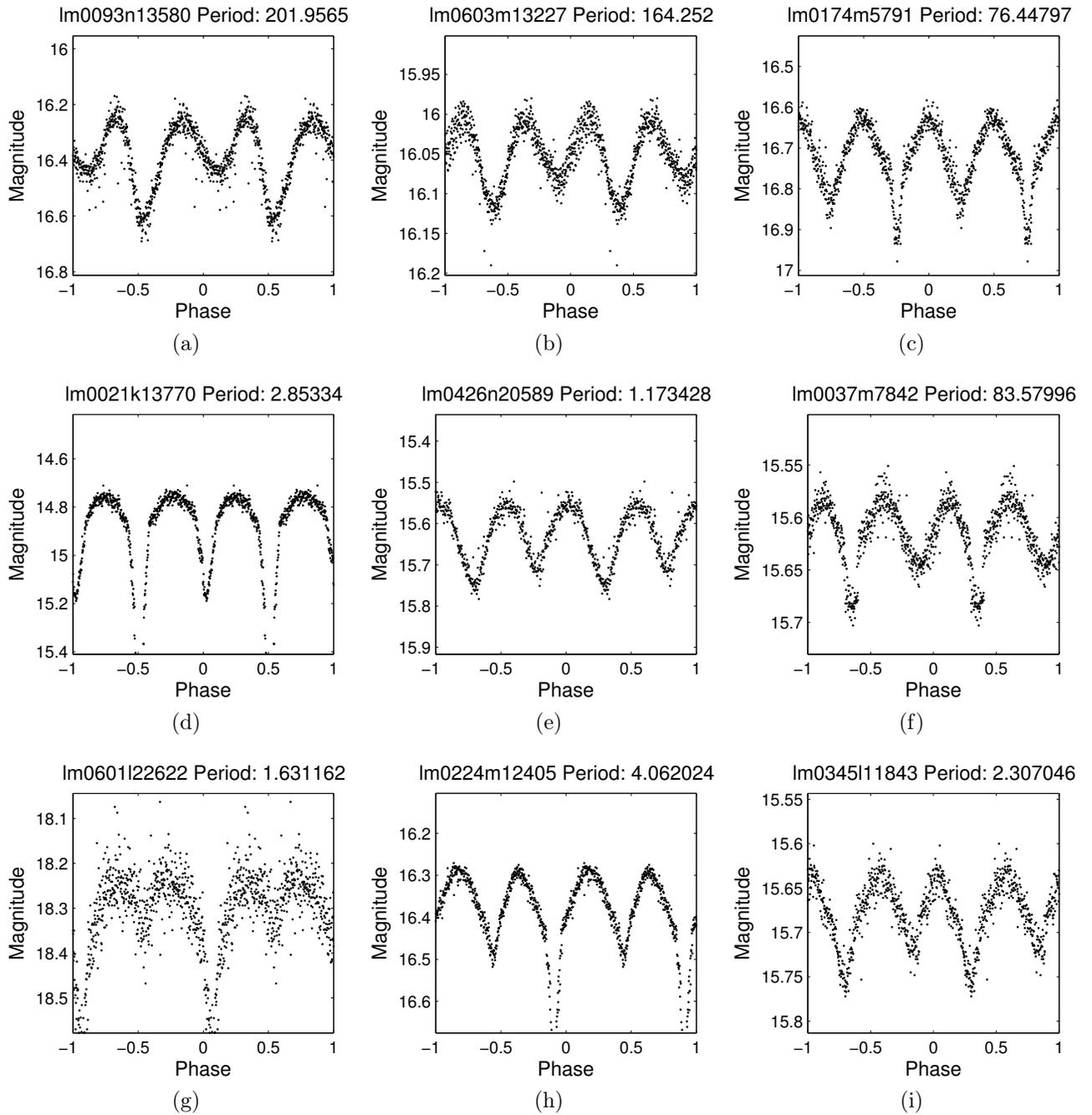


Figure B.4: Examples of EROS-2 periodic lightcurves folded with their estimated period. These light curves correspond to eclipsing binary stars, binary systems whose orbital plane is aligned with the Earth. The alternating drops in brightness correspond to the mutual eclipses performed by the system as observed from Earth.

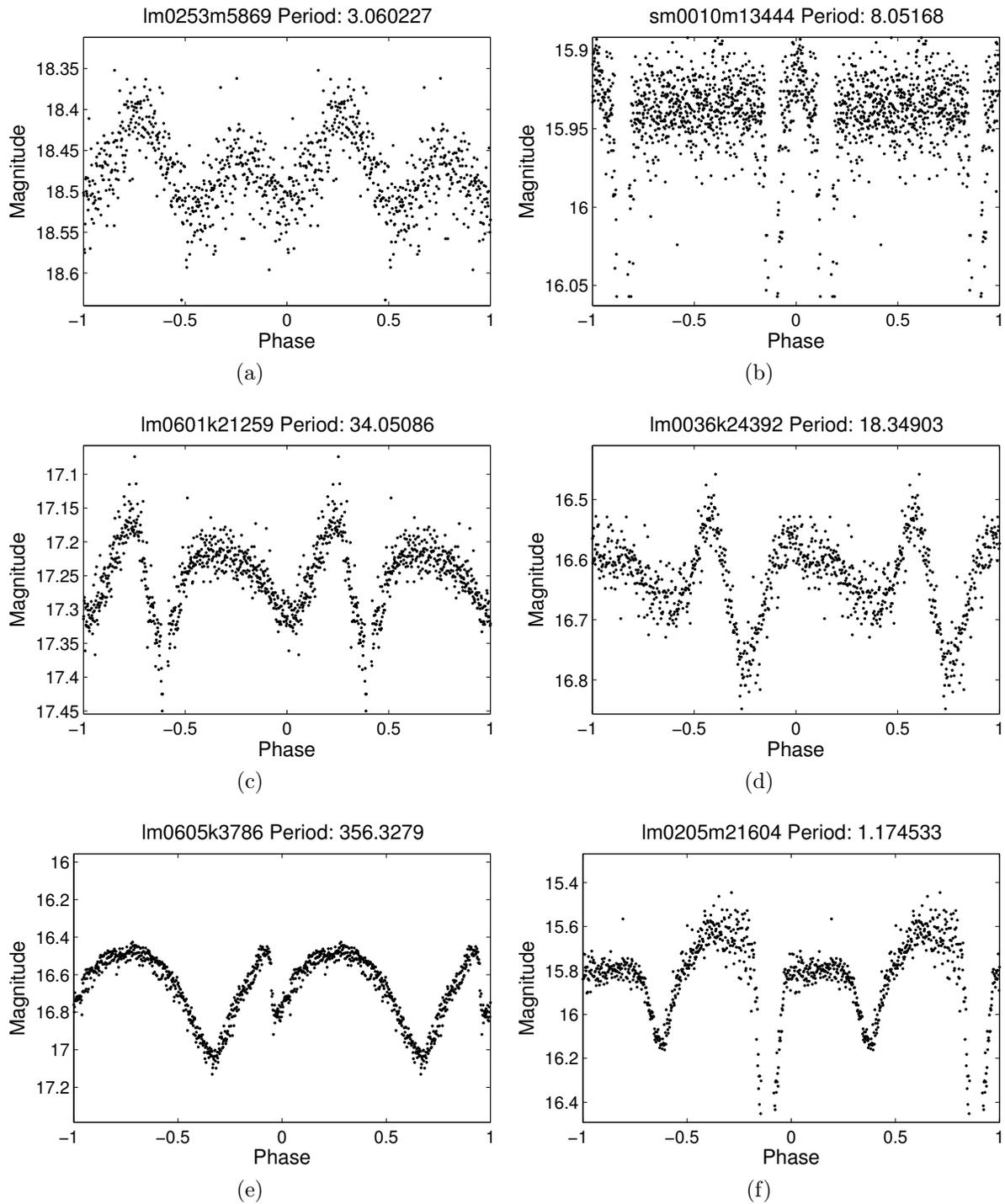


Figure B.5: Examples of EROS-2 periodic lightcurves folded with their estimated period. A priori these objects cannot be attributed to any known class.