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

DETECCIÓN DE PERÍODO EN SERIES DE TIEMPO ASTRONÓMICAS USANDO CORRENTROPÍA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA

PABLO ANDRÉS HUIJSE HEISE

PROFESOR GUÍA:
PABLO ESTÉVEZ VALENCIA

MIEMBROS DE LA COMISIÓN:
HECTOR AGUSTO ALEGRÍA
PABLO ZEGERS FERNÁNDEZ

SANTIAGO DE CHILE
OCTUBRE 2010

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELECTRICISTA
POR: PABLO ANDRÉS HUIJSE HEISE
FECHA: 26 DE AGOSTO DEL 2010
PROF. GUÍA: PABLO ESTÉVEZ VALENCIA

DETECCIÓN DE PERÍODO EN SERIES DE TIEMPO ASTRONÓMICAS USANDO CORRENTROPÍA

Las series de tiempo astronómicas estudiadas, también conocidas como "Curvas de Luz", contienen información de brillo estelar versus tiempo y se caracterizan por estar irregularmente muestreadas, poseer múltiples espacios sin datos (*gaps*) y ser ruidosas. El análisis de curvas de luz es importante pues permite estudiar el comportamiento del objeto estelar en cuestión, detectar eventos de interés, despejar otros parámetros del objeto y realizar tareas de clasificación estelar. En la actualidad son múltiples los catálogos de curvas de luz producidos por sondeos astronómicos alrededor del mundo. La extensión de estos catálogos sumado a las características ya mencionadas de las curvas de luz, hacen que el análisis de series de tiempo astronómicas sea una tarea difícil y costosa. Es por esto que existe un gran interés por el desarrollo de métodos inteligentes que sean capaces de analizar automáticamente los extensos catálogos de curvas de luz.

En el presente trabajo de memoria se propone y prueba una metodología para la detección automática de período en series de tiempo astronómicas basada en la correntropía y su espectro de potencia. La correntropía es una medida de similitud en un espacio de alta dimensionalidad, y puede describirse como la generalización de la función de autocorrelación pues incluye los estadísticos de alto orden, es capaz de detectar no linealidades y no está restringida a procesos Gaussianos. El método propuesto combina la densidad espectral de potencia de la correntropía con la técnica de *folding* (usada en astronomía para analizar curvas de luz), la técnica del ranurado (para el computo de la correntropía a partir de datos muestreados irregularmente), y otras técnicas convencionales de análisis de señales.

La aplicación propuesta en esta memoria de título está enfocada a un aspecto del análisis de curvas de luz, el cual corresponde a la detección de período de estrellas variables periódicas. El método desarrollado se prueba sobre una base de datos de 193 curvas de luz de estrellas binarias eclipsantes obtenidas del proyecto MACHO y cuyo período es conocido. La mejor versión de la implementación desarrollada obtiene un 58 % de aciertos (91 % si se consideran los múltiplos), superando a la autocorrelación convencional (20 % de aciertos) y a una batería de aplicaciones de detección de período para curvas de luz compuesta por Period04, SigSpec y VarTools (todas disponibles en la literatura). Sin embargo, el desempeño del método debe ser mejorado si se desea aplicar en bases de datos de tamaño real, por lo que se propone una serie de modificaciones, entre las que se incluye: Incorporar una etapa de ajuste fino de períodos candidatos; diseñar una métrica de calidad basada en la entropía cuadrática de Renyi; y diseñar una estrategia para escoger el parámetro de kernel en base a su distribución en los casos correctos.

Índice general

Capítulo 1

Introducción

El proyecto MACHO[2], realizado entre los años 1996 y 2002, fue un sondeo astronómico dedicado a la búsqueda de eventos de microlente gravitacional (aumento repentino del brillo de una estrella). La detección de dicho evento permitiría probar teorías con respecto al porcentaje de materia oscura que existe en el Universo. Para lograr este objetivo, se capturó información fotométrica de millones de estrellas durante los 8 años del proyecto, formando una extensa base de datos de series de tiempo o "curvas de luz" (brillo estelar en función del tiempo) como son llamadas en astronomía.

Las curvas de luz son la herramienta básica en el estudio de las Estrellas Variables[1], y su análisis proporciona información valiosa de las características y procesos físicos que realizan las estrellas. Dentro del conjunto de estrellas variables existen algunas como las variables Cefeidas, las RR Lyrae y las binarias eclipsantes que se caracterizan por realizar procesos periódicos los cuales se ven reflejados en sus curvas de luz. La detección del período de estas estrellas es importante, ya que puede usarse para obtener otros parámetros de la misma y/o realizar tareas de clasificación.

Sin embargo, el análisis de las curvas de luz no es una tarea sencilla pues estas series de tiempo se caracterizan por tener un muestreo irregular y un alto contenido de ruido, además se debe considerar la extensión de los catálogos de curvas de luz producidos por los diversos sondeos astronómicos. Estas son las razones por las que el análisis en curvas de luz es una tarea que requiere una gran cantidad de tiempo y esfuerzo por parte de los astrónomos

expertos.

En esta memoria de título se propone un método automático para la detección de período en series de tiempo astronómicas, el cual está basado en la correntropía[21, 22, 23, 24, 25] o correlación generalizada. La correntropía es un funcional desarrollado en el marco de la Teoría de la Información para el Aprendizaje de Máquinas (*Information Theoretic Learning*) y puede definirse como una medida de similitud en un espacio de alta dimensionalidad, para muestras que están separadas temporalmente en su espacio original, por lo que es apropiada para resolver el problema de detección de período.

Para probar el método diseñado, se realizan pruebas sobre 193 curvas de luz de estrellas binarias eclipsantes del proyecto MACHO. Los resultados obtenidos se comparan con los entregados por la correlación convencional y por aplicaciones normalmente usadas en astronomía para la detección de período en curvas de luz.

1.1. Objetivo General

Diseñar un método inteligente y automático para detectar período en series de tiempo astronómicas usando la correntropia, funcional definido en el marco ITL. El método ha de tener en consideración las dificultades inherentes del trabajo con series de tiempo astronómicas, es decir el muestreo irregular, *gaps*, *outliers* y un alto contenido de ruido.

1.2. Objetivos Específicos

1. Comparar los resultados obtenidos por la implementación desarrollada y una implementación paralela en que se reemplace la correntropía por la autocorrelación convencional.

2. Comparar los resultados obtenidos por la implementación desarrollada y programas que ejecuten las técnicas actuales usadas para resolver el problema de detección de período en series de tiempo astronómicas. Los programas que se usarán son VarTools[11], Period04[12] y SigSpec[13].

Capítulo 2

Antecedentes

2.1. Fotometría, Curvas de Luz y Estrellas Variables

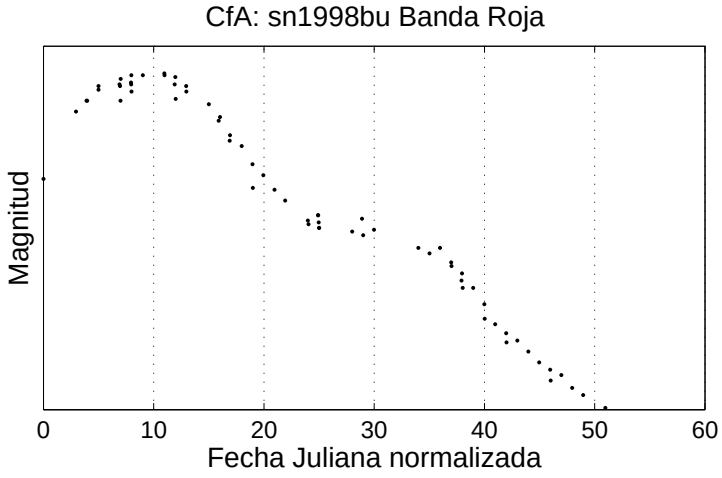
Dado que todos los objetos astronómicos se encuentran demasiado lejos de nosotros, una de las pocas maneras que tenemos para estudiarlos es mediante la observación de la luz que emiten, es en este campo en que se desempeña la fotometría. La fotometría, una de las ramas fundamentales de la Astronomía, está dedicada a la medición precisa de la radiación electromagnética en el espectro visible (magnitud y flujo) de los objetos estelares. Para lograr este objetivo, se aplican una serie de técnicas para transformar los datos obtenidos por los instrumentos astronómicos a unidades estándar de flujo o intensidad. Recientemente, la fotometría ha atravesado una revolución gracias a la masificación y refinamiento de los sensores basados en la tecnología CCD (*charged-coupled devices*). En comparación a los fotómetros fotoeléctricos convencionales, los sensores CCD ofrecen una mayor eficiencia cuántica[1], una respuesta espectral más ancha, mayor robustez al ruido y salida directamente digital. Usando sensores CCD es posible capturar imágenes astronómicas de mayor calidad con menores tiempos de exposición, lo cual se ha traducido en un aumento, tanto en número como en extensión, de los sondeos astronómicos realizados alrededor del mundo. Usando los datos obtenidos con los sensores CCD y las técnicas propias de la fotometría es posible obtener mediciones del brillo aparente de los astros. Estas mediciones son importantes pues son usadas por los astrónomos para despejar información de los objetos estudiados. Por ejemplo, se puede obtener la

[1]Porcentaje de fotones que caen sobre el CCD y que finalmente son detectados.

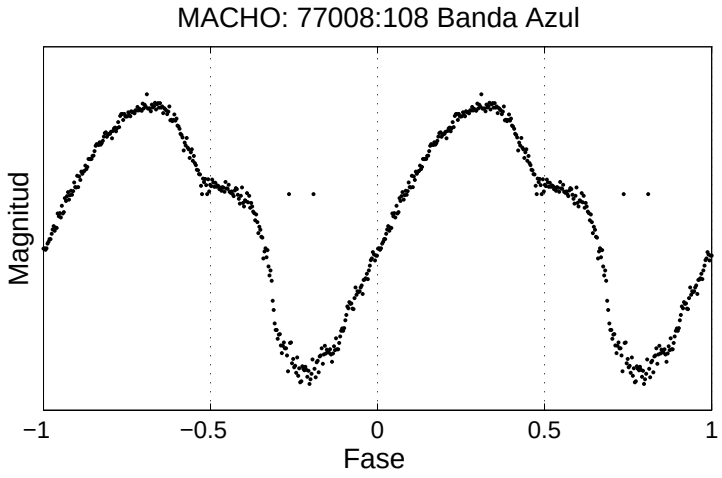
intensidad lumínica de un astro si se conoce su brillo aparente y su distancia a la Tierra (la distancia se puede calcular mediante la técnica de paralajes trigonométricos). Luego, la intensidad lumínica y la temperatura del astro (Ley de Stefann-Boltzmann) pueden usarse para estimar su área (o diámetro si se asume geometría esférica). La temperatura y la composición química de un astro pueden determinarse usando técnicas enmarcadas en el campo de la espectro-fotometría.

La fotometría también es usada para medir las variaciones lumínicas de ciertos objetos astronómicos. En estos casos, una herramienta sencilla pero de gran importancia es la denominada "Curva de Luz". La curva de luz es un gráfico de la magnitud (o flujo) de la radiación electromagnética (en el espectro visible), emitida por un astro (eje de las ordenadas) en función del tiempo (eje de las abscisas). El tiempo en las curvas de luz es expresando comúnmente en días Julianos. Estudiar el comportamiento de los fenómenos radiativos de un astro a través del tiempo, le permite a los astrónomos entender más sobre los procesos físicos internos que ocurren dentro de la estrella (o sistema estelar) en cuestión. La forma de la curva de luz, y en particular los parámetros que se pueden extraer de ella, pueden usarse para tareas importantes como clasificación estelar o detección de eventos.

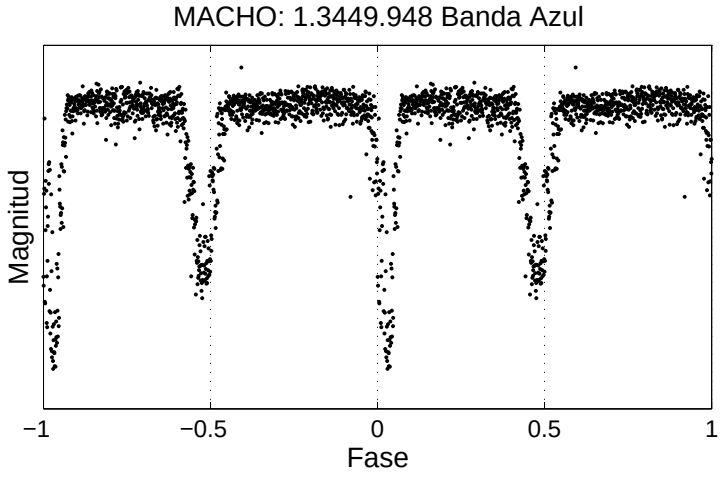
Tal como se mencionó, las curvas de luz son una herramienta de gran utilidad en el estudio de un tipo particular de objeto estelar denominado Estrella Variable[1]. La característica principal de este tipo de objeto astronómico es que su brillo aparente, detectado desde la Tierra, cambia a través del tiempo. Estas variaciones de luminosidad pueden deberse a cambios intrínsecos o extrínsecos, lo cual define dos categorías de estrellas variables. Las estrellas variables intrínsecas varían su luminosidad debido a cambios físicos internos propios de la etapa evolutiva en que se encuentra la estrella, por ejemplo las estrellas variables pulsantes, como las Cefeidas(Figura 2.1.b), se expanden y comprimen en ciclos regulares alterando consecuentemente la magnitud de sus emisiones electromagnéticas. Otro tipo de estrella variable intrínseca son las variables cataclísmicas, como las novas y supernovas (Figura 2.1.a), cuyo brillo puede aumentar hasta en 20 órdenes de magnitud durante el evento explosivo que las caracteriza. La segunda categoría mayor de estrellas variables la componen las de tipo extrínseco. Las variaciones que percibimos en la luminosidad de las estrellas variables extrínsecas son debidas a fenómenos relacionados a la rotación o a la influencia de otros objetos astronómicos (eclipses), siendo un ejemplo de este último caso las estrellas binarias



(a)



(b)



(c)

Figura 2.1: Estrellas Variables: Curva de Luz de una SuperNova (a) y Diagramas de Fase de una variable Cefeida (b) y una binaria eclipsante (c).

eclipsantes (Figura 2.1.c). Las estrellas binarias aparecen ante el observador como un punto único de luz, sin embargo corresponden a un sistema compuesto de dos estrellas, de características no necesariamente similares, que giran alrededor de un centro de gravedad común. La particularidad de las estrellas binarias eclipsantes radica en que su plano orbital se ubica perpendicular al plano del cielo, i.e. la Tierra se encuentra en su plano orbital. Debido a esto, se pueden observar variaciones periódicas en la radiación medida causadas por los eclipses estelares mutuos entre las estrellas del sistema.

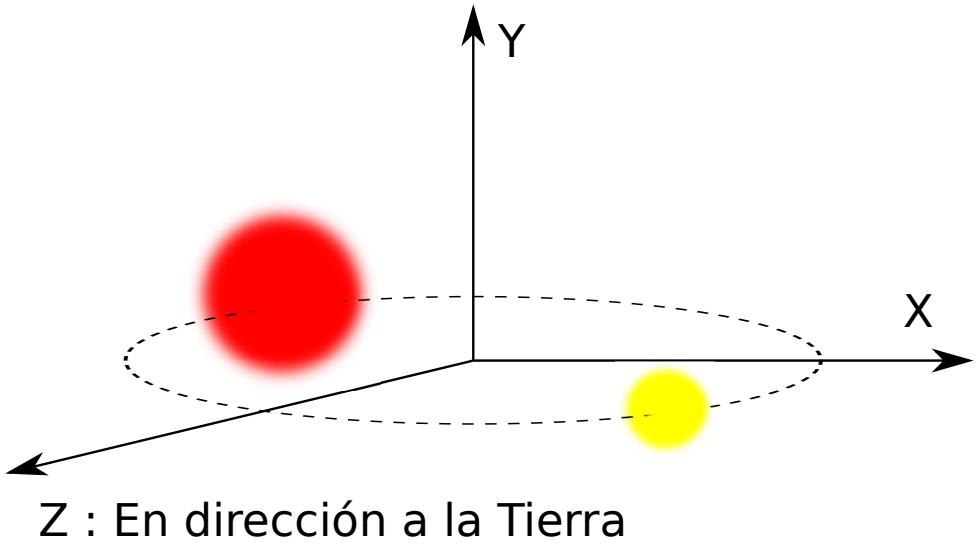


Figura 2.2: Esquema de una Estrella Binaria Eclipsante

Existe un subconjunto de estrellas variables cuyas emisiones electromagnéticas varían siguiendo patrones regulares, como por ejemplo las variables Cefeidas o las estrellas binarias eclipsantes. Las curvas de luz de estos objetos astronómicos tienen una alta presencia de componentes periódicos. El período de dichas curvas es un parámetro de sumo interés para los astrónomos, pues sirve para clasificar objetos estelares entre las categorías mencionadas. El período también puede servir para despejar otras variables de interés de la estrella o sistema, por ejemplo para un sistema binario, el período en conjunto con otros datos como la distancia a la Tierra y la magnitud de su brillo, permiten calcular información respecto de la masa y el radio de los objetos que componen el sistema(Leyes de Kepler).

2.2. Materia Oscura, Proyecto MACHO

Anteriormente se mencionó que el estudio de las curvas de luz puede llevar al descubrimiento de eventos estelares. Un ejemplo de evento estelar, que actualmente genera mucho interés, es el microlente gravitacional[2] (gravitational microlensing). Este fenómeno se da cuando un objeto compacto y oscuro se ubica entre el observador y una fuente luminosa (estrella) de origen extra-galáctico. Al ocurrir esto, el campo gravitacional del objeto oscuro actúa de forma similar a un lente, desviando y amplificando la radiación de la estrella. El evento es detectado por el observador como una notable amplificación en la magnitud aparente de la luminosidad de la fuente y por ende puede ser estudiado fotométricamente. Un punto muy importante a notar es que a través del microlente gravitacional, los astrónomos han sido capaces de estudiar los objetos que actúan como lente y que poseen un brillo demasiado tenue como para ser detectados con los métodos convencionales. Estos objetos son conocidos de forma genérica como "Objeto Astrofísico Compacto y Masivo del Halo de la Galaxia" (Massive Compact Halo Object abreviado MACHO). Hoy en día existe un enorme interés, de parte de las comunidades astronómicas, respecto a la detección de MACHOs y la razón de esto es que dichos objetos han sido postulados como probables candidatos de materia oscura. La materia oscura es una forma de materia que no puede ser detectada por la radiación electromagnética que emite. Las teorías astronómicas actuales indican que un 90 % de la masa total del universo conocido es oscura (no podemos percibir su luz). La masa total del Universo puede despejarse aplicando el Teorema del Virial[3], el dato de velocidad estimada de las galaxias y estrellas conocidas se obtiene a partir de las mediciones de Corrimiento Doppler cosmológico. Si se compara el valor obtenido al aplicar el teorema del Virial con la estimación de la suma de la masa de todas las estrellas conocidas, resulta que este último es muchísimo menor.

La mayor dificultad en la detección y observación de los MACHOs es la baja tasa de ocurrencia del fenómeno de microlente gravitacional. Sin embargo no han sido pocos los proyectos iniciados exclusivamente dedicados a detectar dichos eventos a través de extensos sondeos astronómicos. Uno de estos proyectos es MACHO[2], el cual surge como una colaboración entre científicos estadounidenses y australianos de los observatorios de Mt. Stromlo y Siding Spring, Centro de Astrofísica de Partículas de Sta. Barbara, Universidad de California y el

Laboratorio Nacional Lawrence Livermore. El principal objetivo del proyecto es probar la hipótesis de que una fracción significativa de la materia oscura ubicada en el halo de la vía láctea está conformada principalmente por MACHOs. Para esto, el proyecto cuenta con un sistema bi-canal de 8 cámaras CCD de 2048x2048 pixeles montadas en un telescopio de 50 pulgadas en Mt Stromlo, Australia, el cual captura varios Gbytes en imágenes de la Gran Nube de Magallanes y el Bulbo galáctico[2] cada noche. A través de un exhaustivo análisis fotométrico, el proyecto MACHO ha formado una base de datos de curvas de luz de aproximadamente 60 millones de estrellas. En este sentido, el proyecto MACHO se ha convertido en una importante fuente de curvas de luz astronómicas y en particular de estrellas variables con curvas de luz periódicas como las estrellas binarias eclipsantes.

2.3. Análisis de Curvas de Luz

El análisis de series de tiempo puede definirse como una disciplina en la que se aplican técnicas matemáticas y computacionales con el fin de comprender los procesos que ocurren en el sistema que se está observando. De forma resumida, lo que se desea es lograr extraer la mayor cantidad posible de información a partir de la serie de tiempo, con el fin de: explicar la variabilidad del sistema, encontrar semejanzas o diferencias con otros sistemas conocidos, encontrar los límites del sistema, predecir su comportamiento, etc. Si una serie de tiempo representa un fenómeno que se repite regularmente en el tiempo, la información más condensada que se puede obtener de dicha serie es su período de oscilación. Es por esto que la detección de período es una parte fundamental del análisis de series de tiempo.

El análisis de curvas de luz es una tarea compleja, pues estas se caracterizan por estar muestreadas irregularmente y por ser altamente ruidosas. Existen razones ineludibles para que la adquisición de los datos astronómicos sea irregular: el ciclo día-noche, la luz de la luna, el mal clima (nubosidad) y la ocultación de los objetos en estudio. Otras razones son de origen técnico: la necesidad de reenfocar y reposicionar el telescopio (tras el desplazamiento del objeto en estudio), la recalibración de los equipos o la ocurrencia de fallas eléctricas y mecánicas. Todas estas razones producen discontinuidades (*gaps*) de distinta duración en las

[2]Grupo de estrellas en el centro de nuestra galaxia.

curvas de luz.

El ruido que afecta a las curvas de luz puede ser dividido en dos clases. En primer lugar está el ruido propio de las observaciones, como por ejemplo: el brillo de objetos cercanos, el ruido de fondo debido al brillo del cielo nocturno y el ruido atmosférico causado por los fenómenos de refracción y extinción. En segundo lugar está el ruido propio de los instrumentos utilizados y en particular de las cámaras CCD como: las variaciones de sensibilidad en el detector, el ruido de corriente oscura asociado a la temperatura y el ruido de lectura asociado al diseño y a la calidad de los circuitos.

Son varios los métodos que han sido usados con el objetivo de detectar período en series de tiempo, incluso algunos han sido desarrollados en el ámbito astronómico y por ende tienen en cuenta las dificultades propias de las curvas de luz. A continuación se presentan algunos de estos métodos, poniendo énfasis en su aplicación al caso astronómico y en las ventajas y desventajas de cada uno.

2.3.1. Análisis de Fourier

El análisis de Fourier puede describirse como una técnica en la que una serie de tiempo es representada por infinitas funciones trigonométricas seno y coseno de distintas amplitudes, frecuencias y fases. Para una serie de tiempo, la amplitud de sus componentes frecuenciales se calculan mediante la aplicación de la transformada de Fourier. De esta forma la serie de tiempo es mapeada desde el espacio temporal al espacio frecuencial y el resultado obtenido es llamado espectro de frecuencia de la serie de tiempo. Dado que las series de tiempo son discretas el mapeo al espacio de frecuencias se ha de realizar usando la transformada de Fourier discreta (discrete Fourier transform abreviado como DFT o FFT para referirse a su implementación rápida),

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-jn2\pi\frac{k}{N}} \tag{2.1}$$

donde las frecuencias de interés se localizan en $f = \frac{k}{N} F_S$, con F_S la frecuencia de muestreo de la serie de tiempo. De esta forma, encontrar la frecuencia fundamental de oscilación

usando la DFT se reduce a definir un rango de frecuencias factibles y en seguida encontrar la componente frecuencial estadísticamente más significativa (dentro de ese rango). Si lo que se desea es encontrar el período fundamental de oscilación, entonces basta invertir la frecuencia fundamental obtenida. El rango de frecuencias factibles se define a través de sus cotas, la cota inferior de frecuencia viene dada por el máximo período que se ha de buscar, el cual corresponde a la duración total de la serie de tiempo[3]. La cota superior de frecuencia no está sujeta a interpretación alguna y viene dada por la frecuencia de Nyquist. La frecuencia de Nyquist equivale a la mitad de la frecuencia de muestreo de la serie de tiempo. Si la serie de tiempo está muestreada regularmente (frecuencia de muestreo constante) y cumple el Teorema del muestreo, entonces es posible asegurar que no se encontrarán frecuencias mayores que la dada por la frecuencia de Nyquist. Sin embargo, si la serie de tiempo no está regularmente muestreada, entonces no se cumple lo anterior y aparecerán componentes frecuenciales mayores a la frecuencia de Nyquist. Además, debido al fenómeno de *Aliasing*, aparecerán picos significativos además de los verdaderos componentes frecuenciales, cuya posición dependerá del muestreo de los datos.

Periodograma LS

Dado que la irregularidad en el muestreo es una situación común en las curvas de luz, no corresponde realizar detección de período mediante la aplicación de la DFT. Sin embargo, existe una modificación del periodograma convencional el cual es adecuado para este caso: el Periodograma LS, el cual está basado en las contribuciones independientes de N.R. Lomb[4] y J.D. Scargle[5]. Esta técnica ignora las discontinuidades en los datos y calcula directamente el espectro de potencia, con la salvedad de que la evaluación se hace sólo en los puntos conocidos. De forma resumida, la principal diferencia con el espectro de potencia convencional es que el periodograma LS se fundamenta en un cálculo "por puntos" en vez de "por intervalo de tiempo" lo que elimina los problemas de la transformada de Fourier ante datos muestreados irregularmente. Para definir esta generalización del periodograma convencional, Lomb propuso ajustar la serie de tiempo por un modelo de senos y cosenos en el sentido de mínimos cuadrados, es decir minimizó la expresión

[3]También se puede ser más riguroso y exigir que la cota equivalga al inverso del doble de la duración total de la serie, para asegurar que por lo menos el patrón "periódico" se repita una vez completo.

$$E(\omega) = \sum_{i=0}^{N} (x_i - A\cos(t_i - \phi) - B\sin(t_i - \phi))^2 \tag{2.2}$$

donde (ti, xi) con $i = 0...N$ corresponden a los puntos irregularmente muestreados de la serie de tiempo, ω es la frecuencia angular y ϕ es la fase. La minimización se realiza respecto de las constantes A y B. Sobre el resultado obtenido al minimizar (2.2), Lomb impuso la siguiente restricción

$$\tan(2\omega\phi) = \frac{\sum_{i=0}^{N} \sin(2\omega t_i)}{\sum_{i=0}^{N} \cos(2\omega t_i)} \tag{2.3}$$

la cual define el parámetro de fase y asegura la ortogonalidad de las funciones seno y coseno, del modelo trigonométrico usado, para los tiempos de muestreo t_i. Finalmente la expresión obtenida para el periodograma generalizado en función de la frecuencia angular resulta ser

$$P(\omega) = \frac{1}{2\sigma^2} \left(\frac{\left[\sum_{i=0}^{N} x_i \cos\omega(t_i - \phi)\right]^2}{\sum_{i=0}^{N} \cos^2\omega(t_i - \phi)} + \frac{\left[\sum_{i=0}^{N} x_i \sin\omega(t_i - \phi)\right]^2}{\sum_{i=0}^{N} \sin^2\omega(t_i - \phi)} \right) \tag{2.4}$$

donde σ corresponde a la desviación estándar de la serie de tiempo y ϕ se calcula usando (2.3). El máximo en el espectro de potencia corresponderá a la frecuencia angular cuyo modelo trigonométrico represente con mayor fidelidad a la serie de tiempo en el sentido de mínimos cuadrados. El método del periodograma LS podría clasificarse como lento pues requiere $(10N)^2$ operaciones para analizar N puntos, sin embargo en [6] se propuso una implementación alternativa, basada en la transformada rápida de Fourier (FFT) que reduce las operaciones a $10^2 N\log(N)$.

Finalmente, se identifican los puntos que caracterizan al método del periodograma LS:

- Está respaldado por teorías consolidadas y ampliamente estudiadas.
- Al igual que otros métodos basados en la transformada de Fourier, las series de tiempo se ajustan por senos y cosenos. Luego las variaciones no sinusoidales (como el caso de las curvas de luz de estrellas binarias eclipsantes) podrían no estar bien representadas.

- Proporciona múltiples períodos candidatos y no es trivial discriminar el período correcto.

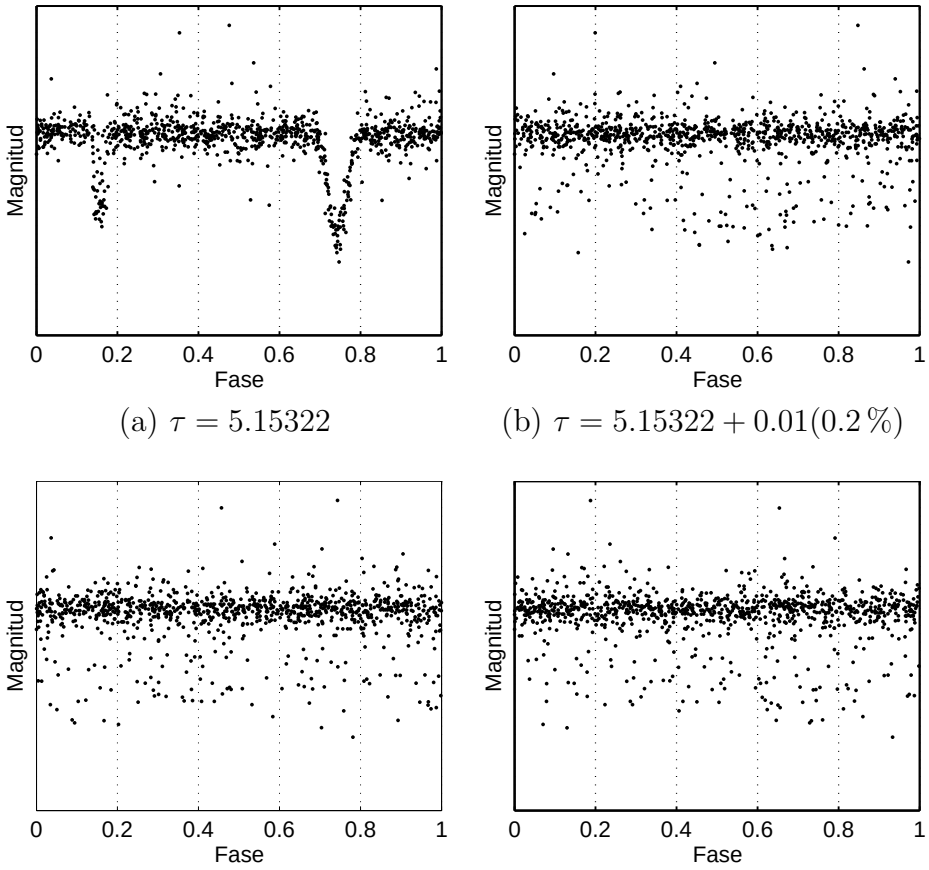
2.3.2. Epoch Folding

Otra técnica, ampliamente usada con series de tiempo astronómicas, es *epoch folding*. La aplicación de esta técnica se puede resumir en dos pasos fundamentales. El primer paso consiste en proponer un período candidato τ. Más adelante se comentará cómo se puede escoger u obtener dicho período. El segundo paso es modificar la componente temporal de la serie de tiempo según una transformación que es función del período candidato

$$t_{folded} = \frac{t \bmod \tau}{\tau} \tag{2.5}$$

en donde mod corresponde a la operación módulo o resto de la división. Esta transformación mapea la serie de tiempo a un espacio de fase definido por el período candidato τ, y es equivalente a particionar la serie de tiempo en intervalos adyacentes y disjuntos de largo τ que luego son superpuestos, "re-ordenando" los puntos de la serie. La superposición de los intervalos es lo que le da el nombre a la técnica pues es similar a doblar o plegar (fold) la serie de tiempo. En la práctica, si el período candidato usado para transformar la serie de tiempo es "cercano" al período real (o a un múltiplo o sub-múltiplo de este), el resultado en el espacio de fase será que los puntos de la serie tenderán a calzar, formándose una curva ordenada. En cambio, si el período candidato es "distinto" del período real, los puntos de la serie llenarán el espacio de fase, dando la impresión de que el resultado no es más que ruido. En general el período de una curva de luz es mucho menor que su largo, lo cual se traduce en que el número de intervalos disjuntos que se van a superponer usando (2.5) será grande. Esto implica que el resultado de la curva de *folding* no será bueno a menos de que la diferencia entre el período propuesto y el período real sea muy pequeña, pues los errores se propagan al doblar la serie. Un ejemplo de esta situación es presentado en la Fig. 2.3, donde la curva de luz de una estrella binaria eclipsante fue doblada con distintos períodos. Nótese que basta un error relativo de 0.2 % para que el resultado no sea el deseado.

Luego de lo explicado, podemos resaltar las fortalezas y debilidades de esta técnica al



(a) $\tau = 5.15322$ (b) $\tau = 5.15322 + 0.01(0.2\,\%)$

(c) $\tau = 5.15322 + 1$ (d) $\tau = 5.15322 + 10$

Figura 2.3: Curva de Luz de MACHO 1.3564.163 Azul (período real 5.15322 [días]) doblada usando distintos períodos candidatos.

momento de usarla en detección de período de curvas de luz:

- Es algorítmicamente sencilla.
- Es apropiada para curvas de luz cuyas variaciones no sean necesariamente sinusoidales.
- Necesita de un período candidato. La opción más sencilla y más costosa es realizar un barrido en período hasta encontrar la forma deseada en la curva doblada. Una opción más inteligente es aplicar otros métodos (por ejemplo Periodograma LS) que proporcionen "períodos de prueba" y usar *epoch folding* en torno a esos períodos.
- La calidad del resultado se mide visualmente. Por si sola, esta técnica no es suficiente para un sistema automático de detección de período.

String Length

Existen muchos métodos que usan la técnica de *epoch folding* como base, combinándola con métricas capaces de cuantificar la calidad de la curva de *folding* obtenida. Por ejemplo, en [8] se discutieron las propiedades de un método conocido como *String Length* (Largo de la Tira), para el caso específico de curvas de luz de estrellas binarias espectroscópicas[4]

A continuación se presenta el método de *String Length*. Para una serie de tiempo doblada $(\phi_i, x_i)_{i=1\ldots N}$ (la transformación de *folding* (2.5) reemplaza el tiempo t_i por la fase ϕ_i), se calcula la distancia total entre puntos adyacentes en el diagrama de fase, es decir la siguiente expresión

$$d_{TOT} = \sum_{i=2}^{N} \sqrt{(\phi_i - \phi_{i-1})^2 + (x_i - x_{i-1})^2}$$

El criterio de *String Length* corresponde a minimizar esta distancia total. Intuitivamente, la distancia total mínima, se obtendrá a partir de la curva de luz doblada con el período real, pues dicha curva es la más ordenada y por lo tanto sus puntos adyacentes serán siempre similares en magnitud.

[4]Sistema binario cuya relación se mide en base a su corrimiento al rojo (Efecto Doppler cosmológico) y al azul de su luz emitida

String Length provee una métrica para usar *epoch folding* sin la necesidad de analizar visualmente las curvas de luz dobladas. Sin embargo, se debe considerar que esta técnica se ve muy afectada por datos perdidos y outliers pues la métrica no es robusta a ellos.

2.3.3. Métodos estadísticos

En esta sección se presentará una rama de métodos que se basan en la aplicación de funcionales y criterios estadísticos sobre la curva de luz o sobre su transformación en fase según *Epoch Folding*. En primer lugar se hablará de la auto-correlación la cual es una alternativa clásica en el tema de detección de período en general. La función de auto-correlación se define como la correlación entre observaciones de una serie de tiempo como función del retardo temporal que las separa. La auto-correlación como función del retardo

$$C(\tau) = \frac{1}{N-\tau+1} \sum_{n=\tau}^{N} x_n x_{n-\tau} \tag{2.6}$$

para una serie x_i con $i = 0...N-1$ muestreada uniformemente en el tiempo. La auto-correlación puede usarse como una medida de similitud entre distintas partes de la serie de tiempo por lo cual es apropiada para detectar patrones repetitivos. Por ejemplo, si para cierto retardo τ los valores de la serie de tiempo tienden a calzar, la auto-correlación de la serie presentará un máximo notable en τ. Es decir, existen componentes periódicos en la serie y que su período de oscilación es τ. Dado que la auto-correlación actúa directamente en el espacio en que se encuentran los datos (espacio temporal), computarla tiene menores costos en tiempo de ejecución (y recursos computacionales) que los métodos basados en la transformada de Fourier. Sin embargo, la auto-correlación tiene desventajas importantes para el caso de las series de tiempo astronómicas:

- Si la serie de tiempo está muestreada irregularmente es necesario interpolar (con tal de re-muestrear los datos). Se verá en secciones posteriores que existe una versión de la autocorrelación para series muestreadas irregularmente, denominada autocorrelación ranurada [29]
- No es apropiada para series de tiempo con múltiples períodos.

Análisis de Varianza

Ciertos métodos evitan la necesidad de interpolar la serie de tiempo trabajando con la curva de luz doblada obtenida luego de aplicar *Epoch Folding*. Como se vio *epoch folding* es apropiado para estudiar series de tiempo irregulares, sin embargo carece de una medida cuantitativa para la calidad del resultado obtenido (es decir depende de inspección visual). En [7] se propuso usar "Análisis de Varianza de un solo factor" (*Analysis of Variance* abreviado como ANOVA) para formular una prueba estadística que permitiera discriminar entre períodos candidatos de series de tiempo astronómicas. El método usado será explicado y comentado a continuación.

En primer lugar se propone un período candidato con el cual se dobla la serie de tiempo usando (2.5). Una vez transformada la señal, el eje de fase es particionado en r cajones disjuntos y de igual largo que cubren todo el espacio. El objetivo de ANOVA es probar la validez de la hipótesis nula que afirma que todas las medias poblaciones (para este caso las medias de cada cajón) son estadísticamente idénticas. Para esto se computan los siguientes estadísticos:

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

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

$$\Theta_{AoV} = \frac{s_1^2}{s_2^2}$$

donde $\bar{x}$ es la media global de la serie de tiempo, $\bar{x}_i$ es la media del cajón i, n_i es el número de elementos que pertenecen al cajón i y N es el número de elementos de la serie de tiempo. s_1^2 mide las variaciones entre grupos (cajones) mientras que s_2^2 mide las variaciones intra-grupo (dentro de cada cajón). Dado que s_1^1 y s_2^2 tienen distribución χ_2 y son independientes se tiene que el estadístico Θ_{AoV} sigue una distribución de Fisher-Snedecor (es un estadístico F).

La prueba estadística consiste en analizar si Θ_{AoV} es mayor que el valor crítico para el estadístico F con parámetros N y r. De ser así, se prueba que no se cumple la hipótesis nula, es decir que al menos una de las medias es diferente. Que las medias sean iguales es indicio de

que la curva de luz fue doblada con un período incorrecto, por ende no satisfacer la hipótesis nula indica que el candidato puede ser el período real o uno de sus múltiplos.

Para encontrar el período real usando ANOVA se construye un arreglo de períodos candidatos y se computa Θ_{AoV} para cada uno de ellos formando un periodograma. Luego, se busca el máximo Θ_{AoV} en el periodograma y se verifica que no satisfaga la hipótesis nula. Si esto es cierto, se busca el período asociado a dicho Θ_{AoV} con lo que se tiene el período de la curva de luz.

Finalmente se resaltan las características de ANOVA para la detección de período de curvas de luz:

- Es un método robusto y de amplio respaldo teórico.
- Como se mostró en [7], tiene un desempeño superior que otros métodos (como el periodograma LS) en detección de período para series con variaciones no sinusoidales (señal triangular, pulsos angostos)
- Al igual que con *Epoch Folding*, requiere de un barrido de períodos candidatos lo cual lo hace computacionalmente pesado y lento.

Minimización de dispersión de fase

Minimización de dispersión de fase (*Phase Dispersion Minimization* abreviado PDM) es un método que posee muchas similitudes con ANOVA y que ha sido ocupado para detectar período en series de tiempo de estrellas binarias eclipsantes[9]. Al igual que con ANOVA, en PDM se requiere un período candidato con el que se ha de doblar la curva de luz según (2.5). Luego, el espacio de fase es particionado en m segmentos idénticos y disjuntos. A continuación se calcula el siguiente estadístico

$$s^2_{PDM} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x_i})^2}{n_i - 1}$$

donde $\bar{x_i}$ es la media del segmento i, n_i es el número de elementos que pertenecen al cajón i. El estadístico s^2_{PDM} es minimizado en búsqueda del período real de la serie de tiempo.

Se puede notar que el estadístico usado en PDM es muy similar a la varianza intra-grupos definida en ANOVA (s_2^2). Sin embargo su distribución para la hipótesis nula no es F o χ_2. Es por esto que en [7] se argumentó que PDM no debería ser mejor o menos costoso que ANOVA y que el test con PDM debería ser menos sensible a señales pequeñas.

Minimización de la Entropía de Shannon

Existe un precedente donde se usaron funcionales de Teoría de la Información para buscar períodos en series de tiempo astronómicas. En [10] se propuso usar la Entropía de Shannon

$$H_S = -\sum_{i=1}^{N} P(x_i) log(P(x_i)) \tag{2.7}$$

para verificar cuan "ordenada" se encontraba una curva de luz doblada y así tener una métrica de cuan correcto es el período usado para doblar la serie. A continuación se explicará con más detalle el método propuesto y se comentarán ciertos aspectos discutidos por los autores.

Al igual que ANOVA, el método involucra proponer un período candidato para doblar la señal mediante la transformación (2.5). Luego, la curva de luz doblada es normalizada al cuadrado unitario, el cual es particionado en q cajones bidimensionales disjuntos. Para computar la entropía de Shannon es necesario conocer la función de distribución de probabilidad (fdp) de los datos, por esto los autores definieron una fdp basado en la función Delta de Dirac

$$\rho_P(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_{P_i})$$

donde $x_{P_i} = (\phi_i, \eta_{P_i})$ con $i = 1...N$ son los puntos que componen la curva doblada con el período P y normalizada al cuadrado unitario. Con la fdp, es posible calcular la probabilidad de ocupación de cada cajón a_i del esquema de partición α

$$\mu_P(a_i) = \int_{a_i} \rho_P(x) dx$$

Finalmente, usando las q probabilidades de ocupación calculadas se puede computar la en-

tropía de Shannon para el esquema de partición

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

La idea básica en que se sustenta este método, es que mediante la entropía de Shannon debería ser posible medir el desorden (o la falta de información) en la curva de luz doblada. Luego, los autores identificaron tres casos para los resultados obtenidos con el método. Cuando la curva de luz era doblada usando un período incorrecto, los puntos de la curva se distribuían uniformemente en el cuadrado unitario y la entropía de Shannon alcanzaba un valor máximo constante. En cambio cuando la curva se doblada usando su período real la entropía de Shannon alcanzaba su mínimo global, pues la curva estaba en su estado de mayor orden. El tercer caso, en que los períodos corresponden a múltiplos racionales del período, se traducía como mínimos locales en la entropía.

A diferencia de otros algoritmos que usan la técnica de *epoch folding*, el método presentado en [10] fue probado con el rigor analítico que respalda a la Teoría de la Información. Sin embargo, al igual que *String length*, ANOVA y PDM, depende de la elección de un período candidato, el cual a priori es buscado mediante "fuerza bruta". Los autores también se manifestaron respecto a algunas de las debilidades del método, como por ejemplo, su excesiva dependencia en el esquema de partición y su sensibilidad al *aliasing* producido por la frecuencia de muestreo. Finalmente se debe notar que las pruebas realizadas fueron hechas usando datos artificiales y no curvas de luz reales, en particular el hecho de cubrir uniformemente el cuadrado unitario al doblar la curva de luz con períodos erróneos no es del todo cierto para algunas curvas de luz, como las procedentes de estrellas binarias eclipsantes.

2.4. Information Theoretic Learning

Debido a la vasta cantidad de estrellas en estudio y las enormes tasas de adquisición de datos de los sondeos astronómicos, las bases de datos estelares de la actualidad han crecido en proporciones que hacen poco práctico su análisis por medios convencionales. Este hecho ha generado una imperante necesidad por métodos automáticos e inteligentes para el análisis

de series de tiempo en Astronomía. Se espera que este tipo de métodos sean capaces de proporcionar información más profunda y detallada de las series de tiempo estudiadas y que además liberen de tareas tediosas a los astrónomos expertos.

En el área de la computación inteligente es donde se cree que pueden surgir los métodos que cumplirán con los requisitos pedidos. En la literatura es posible encontrar algunos ejemplos, como en [16], donde los autores desarrollaron un clasificador basado en redes bayesianas para catalogar estrellas en cinco categorías. En [17] los autores presentaron un método automático de detección de *outliers* dentro de sendos catálogos de curvas de luz periódicas obtenidas del proyecto MACHO, para esto implementaron una medida de similitud basada en la correlación cruzada, obteniendo resultados prometedores. Finalmente en [18], los autores pretenden resolver el problema de clasificación estelar automática a partir de curvas de luz, usando algoritmos desarrollados en el campo del Aprendizaje de Máquinas (Machine Learning). El trabajo realizado se basa en el diseño de una medida de similitud basada en métodos de kernel, dicha medida demostró un desempeño superior en comparación a la medida basada en correlación usada en [17].

En esta memoria se aborda el problema de detección de período en curvas de luz usando herramientas desarrolladas en el marco de la Teoría de la Información para el Aprendizaje de Máquinas [20](*Information Theoretic Learning* o ITL), campo que ha sido renovado gracias a los esfuerzos de los investigadores del Laboratorio de Neuro-Ingeniería Computacional de la Universidad de Florida. El entrenamiento de sistemas inteligentes (como por ejemplo las redes neuronales) es un proceso en que la máquina integra en sus parámetros la información obtenida a partir del conjunto de datos que representan el problema que se desea resolver. La obtención de información a partir de datos se ha realizado típicamente utilizando funcionales estadísticos de segundo orden basados en la correlación, como por ejemplo el error medio cuadrático. Sin embargo, dichos funcionales son sumamente limitados ya que asumen que los datos cumplen las propiedades de linealidad y gaussianidad, lo cual no es cierto en la mayor parte de los casos prácticos. La premisa de ITL es que existe información en los datos que escapa a los estadísticos de segundo orden; para obtener dicha información se propone la utilización de funcionales definidos en Teoría de la Información con el fin de trabajar directamente sobre la función de densidad de probabilidad (fdp) de los datos y así mejorar el desempeño de las máquinas inteligentes a entrenar.

En Teoría de la Información se han combinado las disciplinas de la Ingeniería Eléctrica y las Matemáticas Aplicadas, con el fin de desarrollar expresiones que permitan cuantificar la información contenida en procesos aleatorios. Es así que se han definido funcionales como la Entropía y la Información Mutua, donde la primera mide la incertidumbre que se tiene respecto de una variable aleatoria, mientras que la segunda mide la cantidad de información que una variable aleatoria contiene sobre otra. Nótese que la cantidad de información que guarda una variable aleatoria es inversamente proporcional a la certeza que se tiene sobre la misma, es decir si un mensaje se conoce a priori no existirá información interesante en él y su entropía será mínima. La Teoría de la Información ha impactado con éxito en áreas como la estadística y las telecomunicaciones, particularmente en esta última, los funcionales mencionados se usan para desarrollar sistemas óptimos de codificación, hacer estimaciones de ancho de banda y optimizar la cantidad de información a trasmitir a través de un canal.

En ITL se han definido criterios para el entrenamiento de máquinas inteligentes tales como MaxEnt y MinXEnt. Estos criterios ajustan los parámetros de la máquina mediante la optimización según un funcional de Teoría de la Información. Por ejemplo, cuando se usa el criterio MaxEnt, se maximiza la entropía en la salida de la máquina, lo cual es equivalente a buscar una fdp para la salida que contenga la mayor cantidad de información posible. Por otro lado, cuando se usa el criterio MinXEnt, lo que se busca es la minimización de la entropía cruzada (divergencia) entre las salidas de la máquina (o entre la salida y otras señales), lo cual equivale a encontrar las fdp de salida con una distancia (en el espacio de probabilidades) mínima entre si.

Estos criterios logran obtener mayor información que los estadísticos de segundo orden, pues trabajan sobre la fdp de los datos. Sin embargo su uso tiene un problema práctico: estimar la fdp no es trivial. Esto ha sido resuelto típicamente asumiendo cierta distribución estadística y luego calculando los parámetros que la definen. Sin embargo, asumir a priori una forma para la fdp de los datos es una restricción que puede afectar gravemente la calidad de los resultados obtenidos.

Para resolver este problema, en ITL se ha propuesto el uso del estimador de "Ventana de Parzen"[19]. Usando el método de Ventana de Parzen es posible obtener una estimación de la fdp de los datos a partir de las mismas muestras. Sin embargo, para usar este método fue

necesario reemplazar la entropía de Shannon por una expresión más general de entropía. Esto se debe a que la forma de la entropía de Shannon(2.7) (sumatoria ponderada del logaritmo de las probabilidades) no es apropiada para utilizar algoritmos simples de estimación. La expresión de entropía generalizada usada fue la entropía de Renyi[5], la cual se define como

$$H_{R\alpha} = \frac{1}{1-\alpha} \log \left(\sum_{x \in \aleph} p^{\alpha}(x) \right)$$

donde X es una variable aleatoria discreta, y $p(x)$ es su función de masa de probabilidad (fdp discreta). La entropía cuadrática de Renyi, que corresponde al caso $\alpha = 2$ de la expresión anterior, es apropiada si se desea obtener una estimación sencilla a partir del método de ventana de Parzen

$$H_{R2}(X) = -\log \left(\int_{-\infty}^{+\infty} f_X(z)^2 dz \right) \tag{2.8}$$

donde $f_X(z)$ es la fdp continua de la variable aleatoria unidimensional X. La ventana de Parzen proporciona un medio para realizar una estimación no paramétrica de la fdp de una variable aleatoria, por medio de una función kernel que satisface las propiedades de la fdp. La fdp estimada con el método de ventana de Parzen que para el caso continuo es

$$f_X(z, \{x\}) = \frac{1}{N} \sum_{i=1}^{N} \kappa \left(\frac{z - x_i}{h} \right)$$

donde $\kappa(\cdot)$ es la función de kernel usada y h es el parámetro de suavizado o ancho de banda del kernel. Recordemos que una función de kernel es un producto escalar en un espacio de Hilbert[6] de alta dimensionalidad (posiblemente infinita). La idea básica tras los métodos de kernel, como por ejemplo las Máquinas de Soporte Vectorial (*Support Vector Machines*), el Análisis de Componentes Principales basado en Kernel (*Kernel Principal Component Analysis*) o el Análisis Discriminante Lineal de Fisher basada en Kernel (*Kernel Fisher Linear*

[5]Es posible demostrar, mediante el Teorema de L'Hôpital que la entropía de Renyi tiende a la entropía de Shannon para el caso $\alpha = 1$.

[6]Un espacio de Hilbert es una generalización de un espacio Euclidiano y corresponde a cualquier espacio lineal, provisto de un producto punto, que sea completo con respecto a la norma definida por dicho producto punto

Discriminant), es que los datos, pertenecientes a un espacio de entrada de baja dimensionalidad, pueden mapearse a este espacio de características de dimensionalidad mayor[15] mediante una transformación no lineal $\phi(\cdot)$. La ventaja de realizar este mapeo es que en este espacio de mayor dimensionalidad es más probable que los datos sean linealmente separables.

La existencia de la transformación no lineal $\phi(\cdot)$ está garantizada siempre y cuando el kernel asociado cumpla el Teorema de Mercer. Dicho teorema afirma que existe $\phi(\cdot)$ si y solo si para toda función cuadrado integrable $g(x)$ se cumple que

$$\int\int \kappa(x,y)g(x)g(y)dxdy > 0$$

Si la función de kernel cumple la condición de Mercer entonces se cumple también la siguiente relación

$$\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

la cual es comúnmente conocida como el "Truco del Kernel" [15]. Gracias al truco del kernel no es necesario conocer explícitamente el mapeo $\phi(\cdot)$ pues basta conocer el producto punto o kernel el cual se expresa como una función de los vectores en el espacio de entrada. Un kernel muy usado en el marco ITL y que cumple con la condición de Mercer es el kernel Gaussiano

$$G_\sigma(x_i - x_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{2.9}$$

donde el parámetro σ es la varianza, también llamado tamaño o ancho de banda del kernel. Otros kernel que cumplen la condición de Mercer son el kernel polinomial homogéneo, polinomial inhomogéneo, Laplaciano, Gaussiano y sigmoide. Ahora, si reemplazamos el kernel Gaussiano en la expresión de la ventana de Parzen obtenemos

$$f_X(z, \{x\}) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(z - x_i)$$

Luego, es posible reemplazar la fdp dada por el estimador de Parzen en la expresión de la entropía cuadrática de Renyi (2.8). Dado que el término dentro del logaritmo está elevado

al cuadrado es posible usar la propiedad de convolución de dos funciones Gaussianas que da como resultado otra función Gaussiana centrada en el punto que corresponde a la diferencia entre los centros de las Gaussianas originales, i.e.

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

Finalmente, la expresión resultante para la entropía cuadrática de Renyi

$$H_{R2}(X|x) = -log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)\right) \tag{2.10}$$

donde X es una variable aleatoria con distribución $f_X(z)$ y N es el número de muestras de X. Notemos que con (2.10) se ha obtenido una expresión de entropía que, al contrario de la convencional entropía de Shannon, es fácil de computar pues puede obtenerse directamente de los datos.

2.5. Correntropía

Los funcionales de ITL presentados hasta ahora, como la entropía cuadrática de Renyi, sólo toman en consideración la distribución estadística de amplitudes de los procesos estocásticos estudiados, sin embargo en la práctica, es muy común que estos procesos dependan del tiempo y por ende posean una estructura temporal. Es con esta motivación que en [22, 21] fue presentada la función de correlación generalizada o **correntropía**. Este funcional, al contrario de la autocorrelación convencional, preserva los estadísticos de orden mayor y por ende también preserva las características no lineales de la señal. La correntropía toma ventaja de los métodos de kernel y puede definirse como una medida de similitud en un espacio de alta dimensionalidad entre vectores separados temporalmente en el espacio de entrada. En esta sección se presentará detalladamente la teoría y algunas de las aplicaciones de la correntropía, poniendo especial énfasis en las características que nos serán de utilidad para resolver el problema de detección de período en series de tiempo astronómicas.

Para un proceso estocástico la función de correlación generalizada según [21] se define

como

$$V(s,t) = E(\kappa(x_s, x_t)) \tag{2.11}$$

donde $E[\cdot]$ denota la esperanza. Si se usa el kernel Gaussiano definido en (2.9) es posible expandir la expresión (2.11) a partir de su serie de Taylor, con lo que se obtiene

$$V(s,t) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n \sigma^{2n} n!} E(\|x_s - x_t\|^{2n}) \tag{2.12}$$

De (2.12) podemos notar que:

- La correntropía incluye todos los estadísticos de orden par para $\|x_s - x_t\|$
- La correntropía incluye la información de la correlación convencional en $n = 1$
- Cuando se usa el kernel Gaussiano, el parámetro σ controla el énfasis que se le da a los estadísticos pares de orden mayor, con respecto a los de segundo orden.

En [21] los autores probaron extensamente las propiedades matemáticas de la correntropía. En particular se probó que la correntropía es simétrica, semi-definida positiva y que siempre es máxima para el retardo 0. Además consideremos que la información entregada por la correlación está incluida en la correntropía, como fue mostrado a partir de (2.12). Todas están razones llevaron a los autores de [21] a presentar la correntropía como una generalización de la correlación, lo cual es apropiado pues la correntropía extiende el estadístico de segundo orden a espacios no lineales y a procesos no Gaussianos. Además, la demostración de los puntos anteriores mostró la estrecha relación que existe entre la correntropía y la entropía cuadrática de Renyi estimada mediante Ventana de Parzen.

Las propiedades fundamentales de la correntropía fueron exploradas en [21], a continuación se describirán algunas de las aplicaciones de la correntropía presentadas en otras publicaciones. En [23] la correntropía fue usada para resolver el problema de "Separación ciega de señales", logrando separar con éxito señales provenientes de fuentes independientes e idénticamente distribuidas y también de fuentes Gaussianas con distintos espectros correlacionadas temporalmente. La correntropía superó en rendimiento a algoritmos que también

hacen uso de los estadísticos de mayor orden como "Análisis de Componentes Indepedientes" (*Independent Component Analysis*). En [24] las propiedades de la correntropía fueron estudiadas en el trato de señales con distribuciones no Gaussianas y particularmente en ambientes contaminados con ruido impulsivo, quedando demostrado las ventajas de ésta en comparación a los estadísticos de segundo orden. En [22], la correntropía fue usada como una medida discriminatoria en la detección de no linealidades en series de tiempo, reemplazando con éxito a funcionales tradicionales más costosos como el exponente de Lyapunov.

En [21] se explicó que los estadísticos pares de orden mayor son invariantes a corrimientos temporales. Con esta condición como base se definió la correntropía univariable también llamada autocorrentropía. La autocorrentropía para un proceso estocástico discreto de una variable $\{x_n\}$ con $x_n \in \mathbb{N}$ y $n = 1..N$, en función del retardo m, puede estimarse usando la media y el kernel Gaussiano (2.9)

$$\widehat{V}(m) = \frac{1}{N-m+1}\frac{1}{\sqrt{2\pi}\sigma}\sum_{n=m}^{N-1}\exp\left(-\frac{\|x_n - x_{n-m}\|^2}{2\sigma^2}\right) \tag{2.13}$$

la expresión anterior es válida para $0 < m < N-1$. Notemos que si (2.13) es promediada en función del retardo m se obtiene el argumento dentro del logaritmo en la expresión final de la entropía cuadrática de Renyi (2.10). Es por esta razón que la autocorrelación generalizada definida en [21] fue bautizada como correntropía.

Es importante recalcar que al usar el kernel Gaussiano en la expresión de estimación de correntropía (2.13) se tiene un parámetro libre, el cual corresponde al ancho de banda del kernel Gaussiano σ. Como se ha expresando en [21, 25], el ancho de banda del kernel puede interpretarse en la práctica como la resolución con que la correntropía busca las similitudes en el espacio de kernel de alta dimensionalidad. Si el ancho de banda crece, los estadísticos de segundo orden superan en importancia a los de mayor orden y por ende la correntropía aproxima a la correlación, perdiendo la capacidad de detectar no linealidades; por otro lado, si el ancho de banda se hace demasiado pequeño, se perderá el poder discriminatorio de la correntropía y el resultado tampoco será satisfactorio.

Como se mostró en [21, 25], tal como la transformada de Fourier de la autocorrelación define a la densidad espectral de potencia (power spectral density abreviado como PSD),

es posible obtener la densidad espectral de la correntropía (correntropy spectral density abreviado como CSD) calculando su transformada de Fourier discreta. Luego, la CSD puede definirse como

$$P[f] = \sum_{m=\infty}^{\infty} (\widehat{V}(m) - \langle \widehat{V}(m) \rangle) e^{-j2\pi m \frac{f}{F_s}} \tag{2.14}$$

donde $\langle \widehat{V}(m) \rangle$ es la media de la estimación de autocorrentropía y F_s es la frecuencia de muestreo de la serie de tiempo. Se debe indicar además que dicho funcional retiene las propiedades de la densidad espectral de potencia convencional.

En [25] la auto-correntropía (2.13) y la CSD (2.14) fueron usadas para resolver el problema de detección de frecuencia fundamental en la entonación de vocales. La detección de frecuencia fundamental en la voz es un problema de sumo interés en el área de las telecomunicaciones y son varios los codificadores de voz que incluyen un bloque detector de tono, con el fin de despejar la frecuencia fundamental de la persona por medio de la auto-correlación y su espectro de potencia. Los experimentos realizados en [25] fueron detección de vocal simple, detección de vocal doble y segregación de vocales en ambientes contaminados con ruido blanco Gaussiano. Los resultados obtenidos indicaron que la correntropía es robusta al ruido y que posee una resolución mayor que la correlación convencional. Un tema importante tratado en el análisis fue la necesidad de continuar la investigación en pos de determinar un método automático para seleccionar el ancho de banda del kernel Gaussiano. Este último punto, más el buen desempeño de la correntropía en los casos con ruido atañen directamente al método propuesto en esta memoria.

Como se ha mostrado hasta ahora, la correntropía supera claramente a los estadísticos de segundo orden. Esto se debe a que la correntropía trabaja sobre la fdp de los datos y por ende es capaz de detectar no linealidades. Además, la correntropía no requiere asumir que los datos sean Gaussianos y ha demostrado funcionar bien con series de tiempo ruidosas. Es por estas razones que se cree que la correntropía será de gran utilidad para resolver el problema de detección de período en curvas de luz.

Capítulo 3

Metodología e Implementación

3.1. Base de datos y Características de Software

La base de datos utilizada en las pruebas de esta memoria de título esta conformada por 193 curvas de luz de estrellas binarias eclipsantes con período único. La información contenida en estas curvas de luz fue producida en los sondeos astronómicos realizados por el proyecto MACHO[2].

Cada curva de luz posee dos identificadores, el primero es un código que la relaciona con uno de los objetos astronómicos estudiados por el proyecto MACHO, mientras que el segundo indica en que banda del espectro se midió la luminosidad de dicho objeto. El proyecto MACHO trabajó en dos intervalos del espectro electromagnético: 450 a 590 nm (banda azul) y 590 a 780 nm (banda roja). Las 193 curvas de luz usadas en esta memoria pertenecen a la banda azul.

Cada curva de luz posee tres dimensiones que se describen a continuación:

1. **Tiempo:** Corresponde a los instantes en que fueron tomadas las mediciones, se encuentra en días Julianos.

2. **Magnitud:** Corresponde al logaritmo de la intensidad luminosa del objeto estudiado, medida en los instantes indicados por el vector tiempo

3. **Error:** Corresponde a una estimación del error fotométrico para las mediciones tomadas en los instantes indicados por el vector tiempo.

Los períodos de las 193 curvas de luz consideradas fueron calculados en el Centro de Series de Tiempo de la Universidad Harvard usando una combinación de Análisis de Varianza, *Epoch Folding* e inspección visual por astrónomos expertos. Estos períodos se utilizan como referencia para evaluar el desempeño de la implementación automática de detección de período desarrollada en esta memoria de título. A continuación, en la tabla 3.1, se presenta un resumen de las características de la base de datos usada.

Característica	Valor
Número de muestras promedio por curva de luz:	1007.1 muestras
Tiempo total promedio por curva de luz	2715.8 días
Densidad de muestras promedio por curva de luz	0.3708 muestras/día
Período promedio de las curvas de luz	7.5562 días
Período mínimo:	0.2775 días
Período máximo:	158.4853 días

Tabla 3.1: Resumen de características de interés de la base de datos usada

La implementación para la detección de período en curvas de luz, desarrollada en esta memoria de título, se realiza utilizando Matlab R2008b, sin embargo operaciones pesadas como el computo de la autocorrentropía, CSD y la transformación de *folding* se programan en ANSI-C. Los programas escritos en ANSI-C se compilan usando mex (incluido en Matlab R2008b) obteniéndose funciones compatibles con Matlab.

3.2. Diagrama de bloques de la implementación

En la figura 3.1 se muestra un diagrama de bloques de la implementación programada. En las sub-secciones siguientes se especificarán las tareas realizadas por cada bloque.

3.2.1. Preprocesamiento y Selección

Como se mencionó en secciones anteriores, cada curva de luz posee un vector con estimaciones del error fotométrico para cada instante de medición. Usando esta información

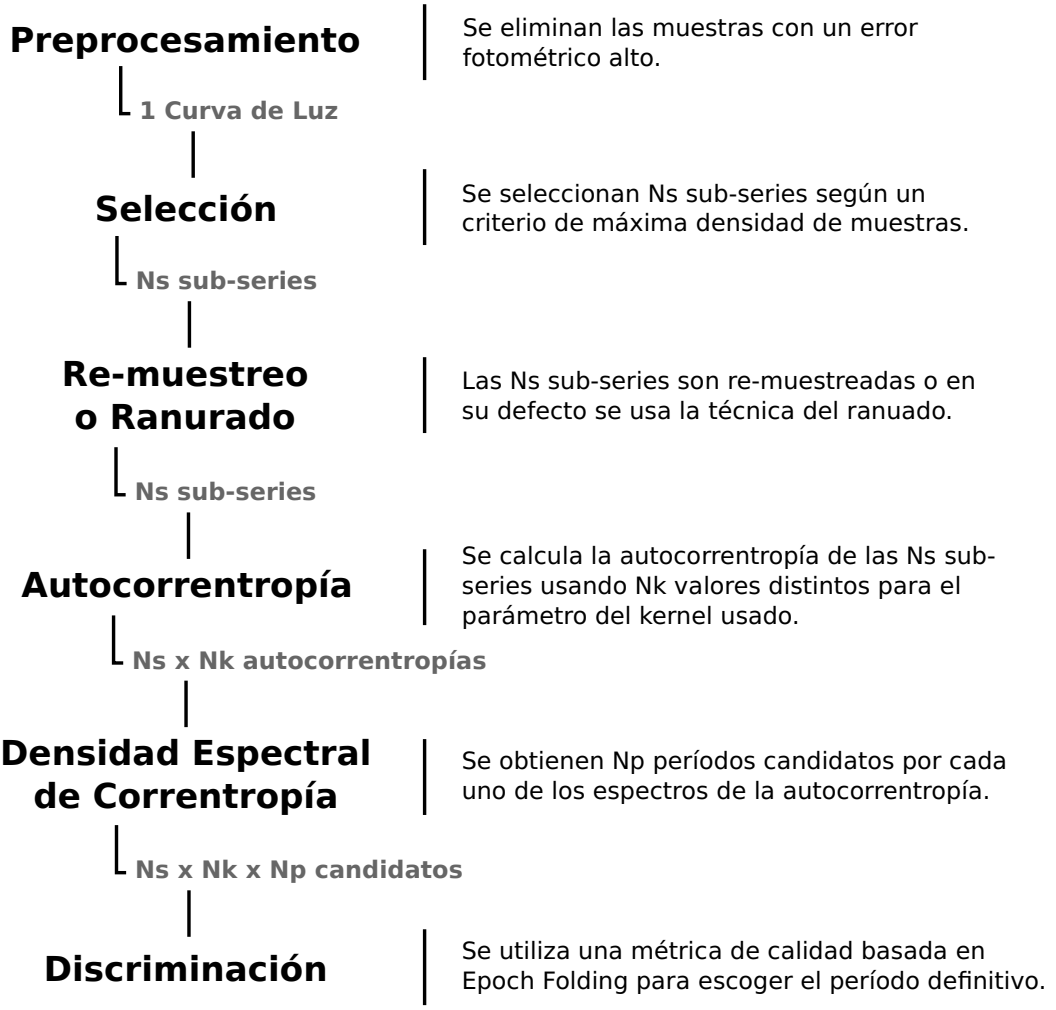


Figura 3.1: Diagrama de bloques de la implementación desarrollada.

se aplica un criterio estadístico sobre la curva de luz, borrando las muestras (es decir las tuplas tiempo;magnitud;error) cuyo error fotométrico asociado sea mayor que la media más dos desviaciones estándar, es decir que cumplan $e_i > \bar{e} + 2e_\sigma$ donde $\bar{e}$ es la media del error fotométrica y e_σ su desviación estándar.

Luego de aplicar el criterio anterior, se selecciona un grupo de sub-series de distinto largo a partir de la serie de tiempo original. El conjunto de sub-series posibles se define en función de su posición inicial dentro de la serie original y su largo (número de muestras). Para seleccionar las sub-series se definen cotas para su tamaño y luego se itera buscando la combinación de posición inicial y tamaño (permitido por las cotas) que maximice la densidad de muestras (número de muestras por unidad de tiempo dentro de la sub-serie). El algoritmo para calcular la densidad de muestras escrito en pseudo-código se muestra a continuación

```
Para NSS entre NMIN y NMAX
        Para I entre 1 y N-NSS
                Densidad = NSS/(t[I+NSS]-t[I])
        Fin
Fin
```

donde NSS representa los tamaños posibles e I representa las posiciones iniciales. Para cada curva de luz se estudian 5 sub-series más la serie de tiempo completa. Estas cinco sub-series se escogen siguiendo el algoritmo descrito y utilizando las siguientes cotas para su largo: $[200, 300]; [300, 400]; [400, 500]; [500, 600]; [600, 700]$, obteniéndose las ventanas más densas para cada uno de los intervalos usados.

Se espera que los períodos más pequeños (del orden de 1 día) se encuentren en sub-series más cortas pero más densas. Por el contrario, se espera que los períodos mayores (mayores que 50 días) se encuentren en las sub-series más largas o en su defecto al analizar la serie completa. Las etapas siguientes de la implementación se realizan en paralelo sobre cada una de las sub-series seleccionadas.

3.2.2. Métodos de Re-Muestreo y Técnica del Ranurado

Para computar la autocorrentropía usando (2.13) es necesario que los datos estén muestreados a una frecuencia constante. Como se ha explicado, este no es el caso de las series de tiempo astronómicas y por ende se usan dos alternativas basadas en técnicas convencionales de procesamiento de señales. La primera de estas alternativas consiste en usar métodos de re-muestreo para interpolar las curvas de luz a intervalos de tiempo regulares. La segunda alternativa se basa en el uso de la "Técnica del Ranurado" (Slotting Technique) [29], la cual ha sido utilizada para computar la autocorrelación en series de tiempo muestreadas irregularmente.

A priori no es claro cual de estas alternativas es más adecuada y por ende se realizará una prueba entre dichas familias de métodos. Manteniendo todos los parámetros de nuestro algoritmo constantes, se procederá a estudiar los resultados obtenidos usando distintos métodos

de re-muestreo y la técnica del ranurado. Para esto se utilizarán las 193 curvas de luz de estrellas eclipsantes binarias de la base de datos MACHO. En el caso de los métodos de interpolación, el procedimiento consiste en:

1. Seleccionar una frecuencia de muestreo F_s [1/días]
2. Generar un vector de tiempo, con nodos muestreados regularmente. La distancia temporal entre los nodos es $\frac{1}{F_s}$ días.
3. Seleccionar un método de re-muestreo e interpolar un valor de magnitud para la curva de luz en cada nodo.

Las técnicas de interpolación que forman parte de la batería de métodos de re-muestreo que se incorporarán en la comparación son:

- **Interpolación Lineal:** Se calculan rectas entre puntos sucesivos de la curva de luz.
- **Spline natural de orden 4, 5 y 6:** Obtener una spline corresponde a calcular polinomios por partes entre cada par de nodos de la curva de luz, imponiendo que los polinomios pasen por los datos originales que se usaron para interpolar. En particular la spline natural se construye tal que la segunda derivada en cada nodo resulte continua. La condición de borde de la spline natural es que la segunda derivada en los nodos inicial y final sea cero. El orden de la spline define el exponente de los polinomios que se utilizarán.
- **Interpolación mediante Spline cúbica con condiciones de borde de tipo "No-es-nodo":** La condición de borde conocida como "No-es-nodo" (not-a-knot) equivale a hacer cero la tercera derivada en el segundo y penúltimo nodo. Es sabido que esta condición impone menos restricciones (en comparación a la spline natural) sobre la forma de la serie de tiempo a interpolar[28].
- **Interpolación Polinomial de Hermite:** La curva de luz es interpolada usando splines cúbicas. La diferencia con la spline natural es que las pendientes en cada punto se escogen tal que se preserve la forma original de la serie de tiempo. Los resultados que se obtienen con splines cúbicas naturales son más suaves que los obtenidos con

interpolación polinomial de Hermite, sin embargo los resultados obtenidas con esta última poseen menos oscilaciones y sobresaltos [28].

- **Interpolación mediante Spline cúbica suavizada:** La curva de luz es aproximadamente ajustada usando splines cúbicas. La spline es construida según dos criterios opuestos: la spline debe ser cercana a los datos y la curvatura de la spline debe ser baja. Los resultados obtenidos pueden variar desde una recta ajustada por mínimos cuadrados (mínima curvatura) hasta la spline natural (mayor cercanía a los datos), en función de un parámetro de suavizado. Este parámetro fue escogido siguiendo los consejos dados en [27, 28] (el parámetro se computa como una función de la diferencia promedio entre nodos).

La frecuencia de muestreo que se usará es de 2 1/días (frecuencia de Nyquist resultante: 1 1/días). Para una descripción más detallada de los métodos presentados consultar [27], todos estos métodos forman parte del Toolbox de splines de Matlab. A continuación se presentará la técnica del ranurado dando énfasis a los cambios que se realizarán para usarla con la autocorrentropía.

Como se mencionó anteriormente, en [30] la técnica del ranurado fue usada para calcular la función de autocorrelación de una serie de tiempo irregularmente muestreada. Recordemos que para computar la autocorrelación para cierto retardo τ es necesario sumar todos los productos cruzados $x_i \cdot x_j$ que estén separados por un retardo temporal regular de largo τ. Si los datos están muestreados de forma irregular no es posible encontrar muestras que se encuentren separadas exactamente en τ unidades de tiempo. Como alternativa a esto, con la técnica del ranurado se utilizarán intervalos para cada uno de los retardos $\tau = k\Delta$ los cuales se definen como $[(k-0.5)\Delta, (k+0.5)\Delta]$, donde Δ es llamado "tamaño de ranura" y $k = 0...M-1$. El algoritmo para computar la autocorrelación ranurada (slotted autocorrelation) se muestra a continuación:

1. Seleccionar un tamaño de ranura Δ.
2. Para cada retardo $\tau = k\Delta$ con $k = 0, ..., \lfloor \tau_{max}/\Delta \rfloor$, y para cada par de índices i y j calcular la diferencia $(t_i - t_j)$.

3. Si se cumple que $[(k-0.5)\Delta < (t_i - t_j) < (k+0.5)\Delta]$ entonces el producto cruzado $x_i \cdot x_j$ contribuye a la suma de valores para el retardo $k\Delta$.

4. Normalizar el valor de la correlación en el retardo $k\Delta$ por el número de pares que cumple la condición $[(k-0.5)\Delta < (t_i - t_j) < (k+0.5)\Delta]$.

Notemos que una versión "ranurada" de la autocorrentropía podría computarse fácilmente, pues bastaría reemplazar el producto cruzado $x_i \cdot x_j$ por la función de kernel correspondiente $\kappa(x_i, x_j)$ en el paso 3 del algoritmo presentado.

Es importante mencionar que la técnica del ranurado introduce un nuevo parámetro a método: el tamaño de la ranura Δ. No existe una regla clara sobre la elección de dicho parámetro, sin embargo si se escoge un valor demasiado pequeño, puede darse que ciertos retardos no contengan muestras y por ende se indefinan. Por el contrario, si el valor del tamaño de la ranura se escoge demasiado grande, entonces no habrá suficiente resolución temporal entre los retardos y los períodos correctos podrían no detectarse. Finalmente, cabe destacar que como el tamaño de la ranura corresponde a la resolución temporal (inverso de la frecuencia de muestreo) entonces también define la frequencia de Nyquist ($1/2\Delta$).

3.2.3. Autocorrentropía y Densidad Espectral de Correntropía

Como se vio anteriormente, la autocorrentropía (2.13) es una función del retardo temporal τ, por lo que es necesario definir el parámetro $\tau_{\text{máx}}$ el cual corresponde al retardo máximo con que se computará la correntropía. El retardo máximo debe escogerse de manera tal que la cantidad de pares de muestras sea suficiente para cada retardo. Por ejemplo, si se escoge $\tau_{max} = N$ con N el número de muestras de la serie de tiempo, la estimación para el retardo máximo se basará sólo en un par de muestras (la primera y la última de la serie, pues son las únicas separadas en N). Para evitar producir estimaciones erróneas se usa un valor de $\tau_{max} = N/10$, es decir el retardo máximo equivale a un 10 % de las muestras de la sub-serie o serie de tiempo en estudio. Una alternativa a limitar el valor del máximo retardo sería utilizar los datos en forma cíclica, sin embargo esta opción no fue explorada. El cómputo de la autocorrentropía, para series de tiempo muestreadas regularmente, expresado como pseudo-código es como sigue:

```
Para i de 0 a Ntau
        Para j de i a N
                V[i] = V[i] + kern(x[j],x[j-i],params)
        Fin
        V[i]=V[i]/(N-i+1)
Fin
```

Así mismo, el computo de la autocorrentropía ranurada expresado como pseudo-código

```
Para i de 0 a N-1
        Para j de 0 a N-1
                Para k de 0 a Ntau
                        Si (k-0.5)*Delta < t[i] - t[j] < (k+0.5)*Delta
                                V[k] = V[k] + kern(x[i],x[j],params)
                                B[k] = B[k] + 1
                        Fin
                Fin
        Fin
Fin
V=V/B
```

No existe información a priori que nos indique cual función de kernel es más apropiada para el problema que se desea resolver. Es por esto que en esta investigación se prueban dos kernel de forma independiente: El kernel Gaussiano (2.9) y el kernel polinomial inhomogéneo, donde este último se define como:

$$\kappa(x_i, x_j) = (1 + x_i \cdot x_j)^p \tag{3.1}$$

donde p corresponde al orden del kernel polinomial. Notemos que ambas funciones de kernel poseen un parámetro: el kernel Gaussiano es función del ancho de banda o tamaño del

kernel σ; mientras que el kernel polinomial inhomogéneo es función del orden del kernel p. Estos parámetros tienen gran influencia en los resultados obtenidos y lamentablemente escoger un valor óptimo para ellos no es una tarea trivial. Desarrollar un método que seleccione automáticamente el valor de σ o p escapa a los alcances de esta memoria de título. Por lo tanto, en lugar de escoger un valor para el parámetro de la función kernel mediante reglas o heurísticas, se usará un arreglo predefinido de valores posibles. Esto se traduce en que para cada curva de luz estudiada se calcularán tantas autocorrentropías y espectros de potencia como valores se hayan incluido en el arreglo predefinido.

Para el kernel Gaussiano se utilizará un arreglo de 200 elementos tal que $\sigma \in [0.001, 10]$, donde los elementos del arreglo son equiespaceados dentro de cuatro intervalos de 50 muestras cada uno: $[0.001, 0.01] \wedge [0.01, 0.1] \wedge [0.1, 1] \wedge [1, 10]$. En tanto, para el kernel polinomial inhomogéneo se utilizará un arreglo de 30 valores enteros tal que $p \in [1, 30]$.

Antes de calcular la densidad espectral de la correntropía (CSD), la autocorrentropía es despojada de su media y multiplicada punto a punto por una ventana de Hamming[1]. Usando la ventana de Hamming se evita la aparición de ruido en el espectro causado por los efectos de borde. La densidad espectral de la autocorrentropía es calculada usando (2.14) sobre un arreglo de frecuencias en el intervalo $[0, F_S/2]$ (donde F_S es la frecuencia de muestreo) con una resolución espectral de 0.0001 1/días. A continuación se obtiene la densidad espectral de potencia de la correntropía (*Correntropy Power Spectral Density*) como el valor absoluto al cuadrado de la CSD.

A continuación se realiza una búsqueda de máximos locales en el espectro de potencia. Los máximos obtenidos son ordenados de mayor a menor según su magnitud y los primeros N_p son seleccionados. Finalmente, se guarda el inverso de las frecuencias asociadas a los máximos locales escogidos, formando un grupo de N_p períodos candidatos. En los experimentos que se realizan en esta memoria de título se considera un valor de $N_p = 10$. Se espera que el componente frecuencial asociado al período real de la curva de luz se manifieste al menos como un máximo local en el espectro de potencia.

En este punto se tienen $N_p = 10$ períodos candidatos por espectro, y un número de espectros que depende del kernel usado (30 para el kernel polinomial inhomogéneo y 200

[1]La ventana de Hamming se define como $w(n) = 0.54 - 0.46 \cdot \cos(\frac{2\pi n}{N-1})$ con $n = 0...N-1$

para el kernel Gaussiano). Estos períodos candidatos son evaluados usando una métrica de calidad, la cual se definirá en la sección siguiente. El mejor de los períodos candidatos se convierte en el período detectado para la sub-serie o serie de tiempo.

3.2.4. Métrica para evaluar la calidad de los períodos candidatos

Como se menciona en el apartado anterior, el período detectado de la serie de tiempo es seleccionado utilizando una métrica de calidad. En esta sección dicha métrica se presenta y describe. La métrica en cuestión se denomina "Cociente de Varianzas" y para calcularla se hace uso de la técnica de *Epoch Folding* (2.5). El algoritmo para calcular el cociente de varianzas se muestra a continuación

1. Doblar la curva de luz según (2.5) usando un período candidato $P_{candidato}$ obtenido del espectro de la autocorrentropía.
2. Calcular la varianza de la curva de luz doblada VAR_{Global}
3. Calcular la varianza promedio de la curva de luz doblada en torno a una media móvil VAR_{Local}. Para esto se crea una ventana de ancho 10 muestras y se estima la varianza dentro de cada ventana.
4. Obtener el "Cociente de Varianzas" como $\frac{VAR_{Local}}{VAR_{Global}}$

Cuando una curva de luz se dobla usando su período real, o un valor cercano a dicho período, el resultado producido es una señal ordenada y bien definida, lo que se traduce en una variabilidad local baja pues puntos adyacentes poseen valores similares. En cambio, si una curva de luz se dobla usando un período erróneo, el resultado obtenido es similar a ruido y la variabilidad local será similar a la variabilidad global. Luego, la regla de discriminación entre períodos candidatos se reduce a buscar aquel período que minimice el cociente de varianzas.

3.2.5. Resumen de parámetros

En la tabla 3.2 se presenta un resumen de los parámetros de la implementación desarrollada.

Parámetro	Valor
Cotas para el tamaño de las sub-series 1	$[200, 300]$
Cotas para el tamaño de las sub-series 2	$[300, 400]$
Cotas para el tamaño de las sub-series 3	$[400, 500]$
Cotas para el tamaño de las sub-series 4	$[500, 600]$
Cotas para el tamaño de las sub-series 5	$[600, 700]$
Frecuencia de Muestreo para interpolación	2 1/días
Tamaño de ranura (autocorrentropía ranurada)	0.5 días
Número de máximo de retardos (autocorrentropía)	10 % de la serie de tiempo
Ancho de banda o Tamaño del kernel Gaussiano	200 valores entre $[0.001, 10]$
Orden del kernel polinomial inhomogéneo	30 valores entre $[1, 30]$
Número de máximos locales extraídos de la CPSD	10
Resolución frecuencial para el calculo de la CPSD	0.0001 1/días

Tabla 3.2: Resumen de parámetros del método desarrollado.

3.3. Metodología de Pruebas

Para poder realizar las comparaciones propuestas en los objetivos de esta memoria es necesario desarrollar una metodología de pruebas. En esta sección se abarcarán dos puntos importantes. El primero corresponde a la forma en que se analizarán los resultados obtenidos, mientras que el segundo trata sobre las aplicaciones y métodos con los que se comparará nuestra implementación.

Para estudiar los resultados obtenidos se realizarán tablas de contingencia donde los períodos obtenidos serán clasificados en tres categorías: Acierto, Múltiplo, Fallo. La categoría de acierto corresponde a una detección que se aproxima al valor real reportado. La categoría de múltiplo corresponde a una detección que se aproxima a uno de los múltiplos (o sub-múltiplos) enteros del valor real reportado. Para verificar la pertenencia a estos categorías se fijará un diferencia umbral. Cuando un período detectado no supera esta diferencia umbral será clasificado como fallo. Esto puede resumirse como:

- Un período detectado es considerado como un acierto si: $|P_{real} - P_{detectado}| < \epsilon$
- Un período detectado es considerado como un múltiplo o sub-múltiplo del período real si:

$$\left| \frac{P_{real}}{P_{detectado}} - \left\lfloor \frac{P_{real}}{P_{detectado}} \right\rfloor \right| < \epsilon \quad \vee \quad \left| \frac{P_{detectado}}{P_{real}} - \left\lfloor \frac{P_{detectado}}{P_{real}} \right\rfloor \right| < \epsilon$$

- Si el período detectado no entra en las categorías anteriores es considerado como un fallo.

donde $\lfloor \cdot \rfloor$ es la operación entero más próximo. Como se menciona en secciones anteriores, los períodos considerados como reales son los obtenidos por los astrónomos del Centro de Series de Tiempo de la Universidad de Harvard. Para comparar los períodos obtenidos con los períodos reales se considera una tolerancia $\epsilon = 0.05$ [días].

Las tablas de contingencia otorgan una visión general de los resultados obtenidos por cada método. Sin embargo, casos particulares interesantes serán estudiados observando los espectros obtenidos.

3.3.1. Implementación paralela usando Autocorrelación

Las ventajas que posee la correntropía, respecto a los funcionales estadísticos de segundo orden, son expuestas detalladamente en el capítulo de contextualización de esta memoria de título. Es por esto que uno de los principales objetivos propuestos en esta memoria, es evidenciar dichas ventajas en función de los resultados obtenidos usando tanto la correntropía como la correlación convencional.

Para esto se programará una implementación alternativa en que se cambiará la correntropía por la autocorrelación (2.6). En reemplazo de la CSD se computará la densidad espectral de potencia (PSD), la cual corresponde a la transformada de Fourier de la autocorrelación. Dado que la autocorrelación no tiene parámetros, se calculará sólo un espectro por cada curva de luz. Las etapas de enventanado, re-muestro/ranurado y discriminación se mantendrán constantes.

3.3.2. Aplicaciones actuales para la detección de período en curvas de luz

Los resultados obtenidos con nuestra implementación serán comparados con los calculados por aplicaciones que ejecuten las algoritmos y técnicas actuales mencionadas en la contex-

tualización de esta memoria de título. Las aplicaciones escogidas para esta comparación son VarTools, Period04 y SigSpec, las cuales son brevemente descritas a continuación.

- **VarTools:** Este aplicación[11] fue creada por Joel D. Hartman del Centro de Astrofísica (CfA) de la Universidad de Harvard y del instituto Smithsonian. VarTools incluye un conjunto diverso de herramientas para el tratamiento y análisis de curvas de luz en las que se incluyen el Periodograma LS; Análisis de Varianza; Auto-correlación; *Epoch Folding*; entre otras. VarTools fue programado en ANSI-C y es compatible con cualquier sistema operativo basado en UNIX. Además cabe notar que la interfaz con el usuario es exclusivamente a través de la línea de comando de UNIX, por lo que es sencillo obtener resultados para una base de datos extensa mediante un archivo de procesamiento por lotes(*Shell Script*) apropiado. Para la comparación a realizar sólo se utilizará una de las herramientas de VarTools: El Periodograma LS (2.4). Esta herramienta es llamada mediante el siguiente comando:

  ```
  ./VarTools -LS minp maxp subsample Npeak LightCurve.dat
  ```

  Los parámetros *minp* y *maxp* especifican la cota inferior y superior con que se realiza la búsqueda de períodos. Se usarán 0.1 días y 200 días respectivamente. El parámetro *subsample* corresponde a la resolución frecuencial inicial, se usará un valor de 0.01 1/días. Finalmente el parámetro *Npeak* corresponde al número picos significativos recuperados del periodograma. Se usará un valor de Npeak=1 pues sólo se tomará en cuenta el pico más significativo.

- **Period04:** Esta aplicación[12] fue creada por Patrick Lenz integrante del Grupo de Teoría e Investigación de Estrellas Pulsantes (TOPS) del Instituto de Astronomía de la Universidad de Viena. Period04 posee herramientas basadas en la transformada de Fourier discreta para analizar el espectro de series de tiempo muestreadas de forma irregular. También ofrece la posibilidad de realizar ajustes multi-frecuenciales a las series de tiempo en el sentido de mínimos cuadrados. Period04 fue programado en JAVA y ha sido compilado en sistemas operativos Windows, UNIX y MAC OS. Un importante detalle de esta aplicación es que no puede considerarse como automática pues su orientación está hacia la interacción con el usuario (resultados gráficos). Por esta misma razón, no incluye opciones para analizar bases de datos completas.

- **SigSpec:** Esta aplicación[13] fue creada por Piet Reegen del Instituto de Astronomía de la Universidad de Viena. SigSpec obtiene la DFT de la serie de tiempo y luego computa una métrica original denominada como "Significancia" (*Significance*). La Significancia se define como el logaritmo de la probabilidad de que cierta amplitud A del espectro de la DFT exceda un límite dado, donde dicha probabilidad es calculada a partir de la fdp de las amplitudes de la DFT. El período de la serie de tiempo es obtenido buscando el valor máximo en la curva de Significancia (lo convencional sería buscar el valor máximo e la DFT). SigSpec fue escrito en ANSI-C por lo que es compatible con sistemas UNIX, sin embargo también se proporcionan binarios compatibles con sistemas operativos Windows. SigSpec cuenta con una interfaz de usuario a través de linea de comando e incluye opciones para tratar con bases de datos astronómicas completas de forma automática. La herramienta SigSpec se llamará usando el siguiente comando:

  ```
  ./SigSpec ufreq 5 LightCurve.dat
  ```

  donde se escoge un valor apropiado para el parámetro ufreq el cual corresponde al límite máximo de frecuencia (mínimo de período).

Capítulo 4

Resultados

En este capítulo se presentan y discuten los resultados de las pruebas realizadas a la implementación propuesta. En primer lugar se muestran los resultados que fueron usados para definir algunas de las etapas que componen el método. Posteriormente, se muestran las comparaciones realizadas entre los resultados obtenidos con la implementación desarrollada, la correlación convencional y aplicaciones actuales de detección de período en curvas de luz.

4.1. Resultados relacionados al desarrollo de la implementación propuesta

Comparación entre métodos de re-muestreo y la técnica del ranurado

En la sección 3.2.2 se muestran dos alternativas para calcular la correntropía a partir de una serie de tiempo irregularmente muestreada. La primera alternativa consiste en interpolar la curva de luz usando métodos de re-muestreo, con lo que se obtiene un nuevo vector de datos equiespaciados en el tiempo. Cuando la curva de luz es re-muestreada la correntropía puede ser calculada directamente a partir de los nuevos datos. La segunda alternativa es la técnica del ranurado (*slotting technique*), la cual permite trabajar sobre los datos originales de la curva de luz a cambio de definir intervalos para los retardos de la correntropía (normalmente

cada retardo es un valor único).

Para definir cual de estas alternativas será utilizada se realiza una prueba de detección de período sobre las 193 curvas de luz disponibles de estrellas eclipsantes binarias de MACHO. En la tabla 4.1 se presenta un resumen de los resultados obtenidos.

Método de re-muestreo	Aciertos [%]	Múltiplos [%]	Fallos [%]
Lineal	25.9067	50.2591	23.8342
Spline Cúbica con condición "no-es-nodo"	27.9793	49.2228	22.7979
Spline natural de orden 4	25.9170	51.8032	22.2798
Spline natural de orden 5	24.8704	52.8498	22.2798
Spline natural de orden 6	22.2797	55.4404	22.2798
Polinomial de Hermite	23.8342	50, 7772	25.3886
Spline suavizante.	26.4249	54.4041	19.1710
Sin re-muestreo	Aciertos [%]	Múltiplos [%]	Fallos [%]
Técnica del Ranurado $\Delta = 0.5$	49.2228	41.4508	9.3264

Tabla 4.1: Comparación entre métodos de re-muestreo y la técnica del ranurado.

Como se aprecia en la Tabla 4.1 la técnica del ranurado alcanza un porcentaje de aciertos mayor y un porcentaje de fallos menor que todos los métodos de interpolación. Una de las posibles razones por las que los métodos de re-muestreo son superados por la técnica del ranurado es que esta última permite calcular la correntropía usando los datos originales. Por otro lado, no olvidemos que las curvas de luz contienen una gran cantidad de espacios sin información (gaps), en dichos espacios los métodos de re-muestreo generan datos ruidosos, lo cual podría explicar su desempeño inferior.

El método de interpolación que obtuvo el mayor número de aciertos es la spline cúbica con condición de borde "no-es-nodo". En 3.2.2 se menciona que las condiciones de borde "no-es-nodo" permiten que la spline siga mejor la forma original de la curva de luz, lo cual podría explicar porque este método en particular supera a las splines de orden mayor (en particular la tasa de aciertos parece disminuir con el orden de la spline natural).

A continuación se presenta un análisis caso a caso en donde se compara la técnica del ranurado con la spline cúbica.

- La spline cúbica y la técnica del ranurado comparten un 20.7 % (40 curvas de luz) de los períodos marcados como Aciertos.

- La spline cúbica y la técnica del ranurado comparten un 22.3 % (43 curvas de luz) de los períodos marcados como Múltiplos.
- La spline cúbica y la técnica del ranurado comparten un 3.6 % (7 curvas de luz) de los períodos marcados como Fallos.
- 55 períodos (28.5 %) marcados como Múltiplos o Fallos usando la spline cúbica son marcados como aciertos al usar la técnica del ranurado.
- 11 períodos (5.7 %) marcados como Múltiplos al usar la spline cúbica son marcados como Fallos al usar la técnica del ranurado.
- 14 períodos (7.3 %) marcados como Aciertos al usar la spline cúbica son marcados como Múltiplos al usar la técnica del ranurado.
- Ningún período marcado como Acierto al usar la spline cúbica es marcado como Fallo al usar la técnica del ranurado.

En base a estos resultados se descarta el uso de métodos de re-muestreo en favor de la técnica del ranurado. Sin embargo, a pesar de la superioridad de la técnica del ranurado, existen 25 casos en que el resultado obtenido empeoró con respecto a la spline cúbica. Un análisis más acabado para estos casos muestra lo siguiente:

- Existen 14 casos de aciertos con la spline cúbica que resultaron en múltiplos con la correntropía ranurada. En estos casos se verifica que el período correcto está presente en el espectro de potencia de la correntropía ranurada. Sin embargo, la métrica de calidad descarta el período correcto en favor de múltiplos más precisos. La precisión de los múltiplos es menor en el caso de la spline cúbica por lo que la métrica no los prefiere sobre el período real. En secciones posteriores se discute respecto a como mejorar la métrica de calidad para evitar situaciones como esta.

- Existen 11 casos donde la spline cúbica detectó múltiplos mientras que la correntropía ranurada detectó fallos. Se verifica que los fallos detectados por la correntropía ranurada corresponden a múltiplos no enteros del período correcto (1.5 y 2.5 veces el período correcto).

Otro aspecto de interés es el origen de los períodos marcados como aciertos, es decir si el acierto es detectado al analizar la serie completa o una de las sub-series seleccionadas en el

bloque de enventanado. Para el caso de la correntropía estimada con la técnica del ranurado se tiene

- El 52.9412 % de los aciertos se detecta tanto en la serie completa como en las sub-series (período promedio detectado 6.6740 días).
- El 30.3922 % de los aciertos se detecta sólo en una de las sub-series analizadas (período promedio detectado 5.4966 días)
- El 16.6667 % de los aciertos se detecta sólo en la curva de luz completa (período promedio detectado 7.0038 días).

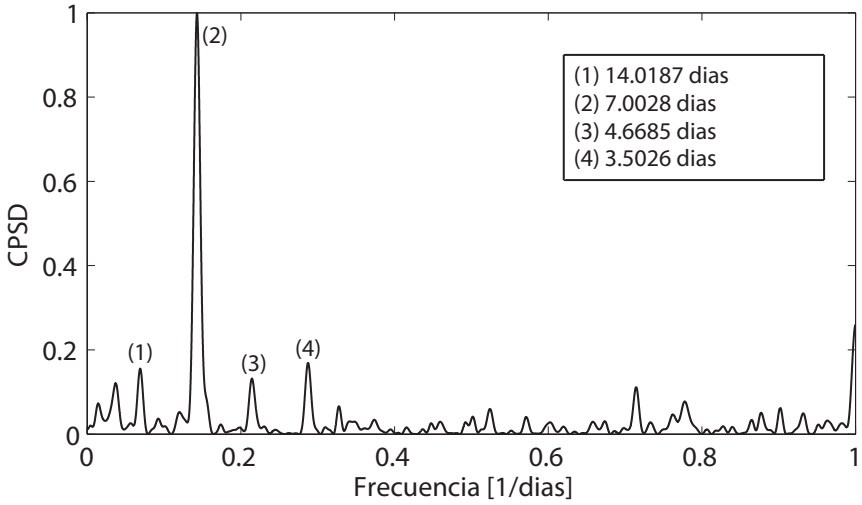
Los porcentajes mostrados están en función del total de aciertos y no del total de curvas de luz. De haber usado sólo la serie de tiempo completa el porcentaje de aciertos hubiera disminuido notablemente lo cual justifica la inclusión del bloque de enventanado.

Comparación entre el kernel Gaussiano y el kernel polinomial

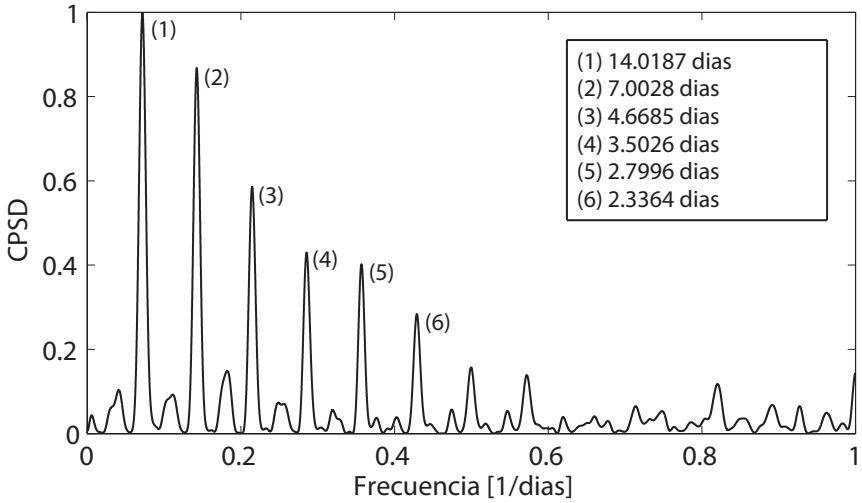
El kernel Gaussiano ha sido ampliamente usado en el marco ITL. Sin embargo, tal como se menciona en la sección 3.2.3, no existe prueba de que el kernel Gaussiano sea en realidad la función de kernel más apropiada para este problema. Para ilustrar las diferencias que pueden existir entre ambas estimaciones de correntropía se presenta el siguiente ejemplo.

Las Figuras 4.1(a) y 4.1(b) muestran los espectros de correntropía estimados usando el kernel Gaussiano y polinomial inhomogéneo respectivamente, para un caso particular. La curva de luz usada en este ejemplo corresponde al objeto 1.3449.948 del catálogo MACHO (período reportado 14,0063 días). Nótese como el período correcto (14 días) resalta frente a sus sub-múltiplos en el caso del kernel polinomial (Figura 4.1(b)). En el caso del kernel Gaussiano (Figura 4.1(a)), el componente del período correcto aparece menos intenso que el sub-múltiplo más próximo (7 días), además el espectro muestra una menor cantidad de componentes asociados a sub-múltiplos.

A continuación se presentan los resultados de una prueba sobre las 193 curvas de luz de estrellas binarias eclipsantes disponibles, en donde la estimación de la correntropía ranurada



(a) Kernel Gaussiano ($\sigma = 0.1$)



(b) Kernel polinomial inhomogéneo ($p = 6$)

Figura 4.1: Espectros de potencia para la curva de luz 1.3449.948 (período reportado 14,0063 días) usando distintas funciones de kernel

se realiza usando el kernel polinomial inhomogéneo (3.1). La Tabla 4.2 resume los resultados obtenidos y los contrasta con los del kernel Gaussiano.

Función de Kernel	Aciertos [%]	Múltiplos [%]	Fallos [%]
Gaussiano	49.2228	41.4508	9.3264
Polinomial	51.2953	32.6425	16.0622
Ambos kernels[1]	58.0311	32.1244	9.8446

Tabla 4.2: Resultados obtenidos con la correntropía ranurada usando el kernel Gaussiano, polinomial inhomogéneo y ambos combinados.

Se puede notar que el kernel polinomial inhomogéneo supera por un 2 % al kernel Gaussiano en la categoría de Aciertos, sin embargo este último supera al primero por un 7 % en la categoría de fallos. Si se analizan los resultados caso a caso se obtiene que

- Ambas funciones de kernel comparten un 36.2694 % de Aciertos.
- El kernel Gaussiano posee un 12.9534 % de aciertos que el kernel polinomial detecta como múltiplos o fallos.
- El kernel polinomial inhomogéneo posee un 15.0259 % de aciertos que el kernel Gaussiano detecta como múltiplos o fallos.
- Esto resulta en un total de 64.2487 % de aciertos, sin embargo la métrica de cociente de varianzas reemplaza un 6.2176 % de estos aciertos por múltiplos más precisos.

Los aciertos obtenidos con el kernel polinomial inhomogéneo no son exactamente los mismos que los obtenidos con el kernel Gaussiano, luego es posible combinar los resultados de ambos funcionales evaluándolos con la métrica del cociente de varianzas. Esto se muestra en la tabla 4.2 en la fila denominada "Ambos kernels". Mediante la combinación de resultados se obtiene un moderado aumento de 6.73 % en la tasa de aciertos. Es importante notar que la métrica usada evalúa mejor un múltiplo exacto que un valor medianamente cercano al período real. Para demostrar esta situación tómese como ejemplo la decisión de la métrica para la curva de luz 1.4047.536, cuyo período reportado es 2.05812 días. Para dicho caso el período que minimiza la métrica es de 1.02905 días (error absoluto de $2e-5$ días con la mitad del período reportado) el cual es detectado por el kernel polinomial. En tanto, el período desechado es de 2.05406 días (error absoluto de $4e-3$ días) el cual es detectado usando el

kernel Gaussiano. Esto indica que se debe dedicar esfuerzo al desarrollo de una métrica de calidad más inteligente si se desea mejorar el rendimiento de la implementación propuesta. Otra opción para mejorar el desempeño del método, que no involucra modificar la métrica, sería realizar una búsqueda fina en un intervalo pequeño para cada período candidato. Luego al usar la métrica se estarían comparando períodos con igual grado de precisión, evitando situaciones como las mostradas anteriormente.

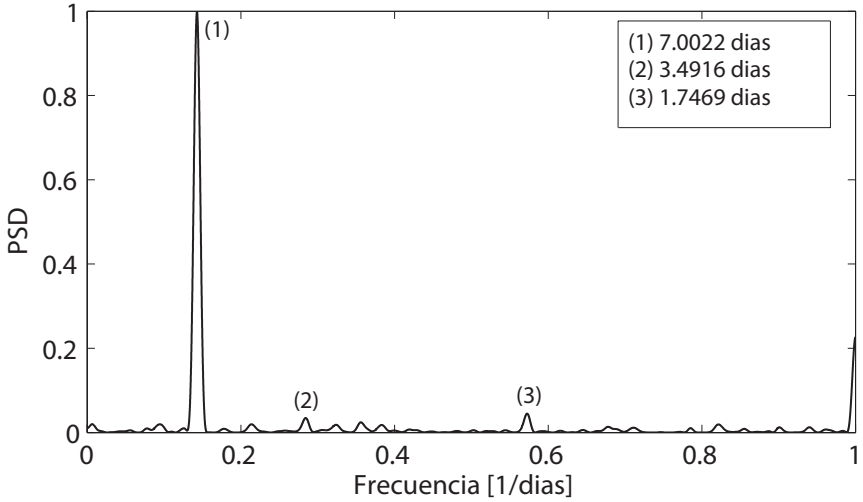
En las comparaciones que se presentan en las secciones siguientes se considerará la mejor versión de la implementación desarrollada la cual corresponde a la correntropía ranurada combinando los resultados obtenidos por el kernel polinomial inhomogéneo y kernel Gaussiano.

4.2. Comparación entre la Correntropía y la Autocorrelación ranurada

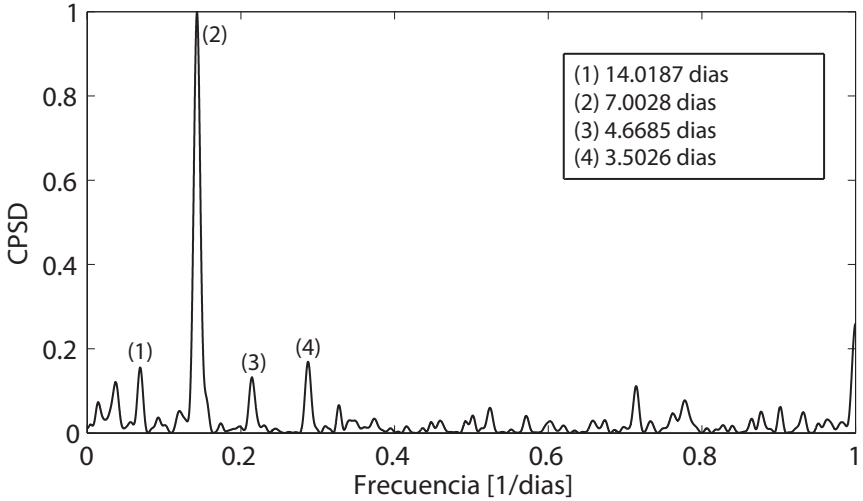
En esta sección se presentan los resultados de las prueba realizadas usando la autocorrelación convencional ranurada. Para esta prueba se usa una implementación idéntica a la propuesta en el capítulo de metodología, con la salvedad de que la correntropía y su densidad espectral (CSD) son reemplazadas por la autocorrelación y su espectro de potencia (PSD). El objetivo de esta prueba es poner en evidencia la ventaja que supone ocupar la correntropía en lugar de la autocorrelación para la resolver el problema de detección de período en curvas de luz.

En primer lugar se presentan ejemplos para dos casos particulares. En el primero de ellos se utiliza la curva de luz 1.3449.948 del catálogo MACHO cuyo período real es de 14.0062 días. En la Figura 4.2(a) y 4.2(b) se muestran las densidades espectrales de potencia calculadas a partir de la autocorrelación y la correntropía respectivamente. Nótese que el período real no aparece representado en la PSD, lo cual indica que la autocorrelación es incapaz de detectarlo. En cambio, la CSD si presenta un componente asociado al período correcto aunque no tan marcado como el sub-múltiplo más próximo (7 días). En general, se puede apreciar que la correntropía posee un espectro mucho más rico en múltiplos que el de la autocorrelación.

En el segundo ejemplo se utiliza la curva de luz 1.3566.63 del catálogo MACHO, cuyo período real es 1.60459 días. En la Figura 4.3(a) y 4.3(b) se muestran las densidades espectrales de potencia calculadas a partir de la autocorrelación y la correntropía respectivamente. Al igual que en el caso anterior, la autocorrelación es incapaz de encontrar el período correcto, lo cual queda demostrado al observar la PSD. Nótese que las períodos detectados en la PSD corresponden a los *alias* de 1 [día] de los múltiplos del período real. En cambio, el período real si aparece representado en la CSD aunque acompañado de los mismos componentes que aparecían en la PSD.

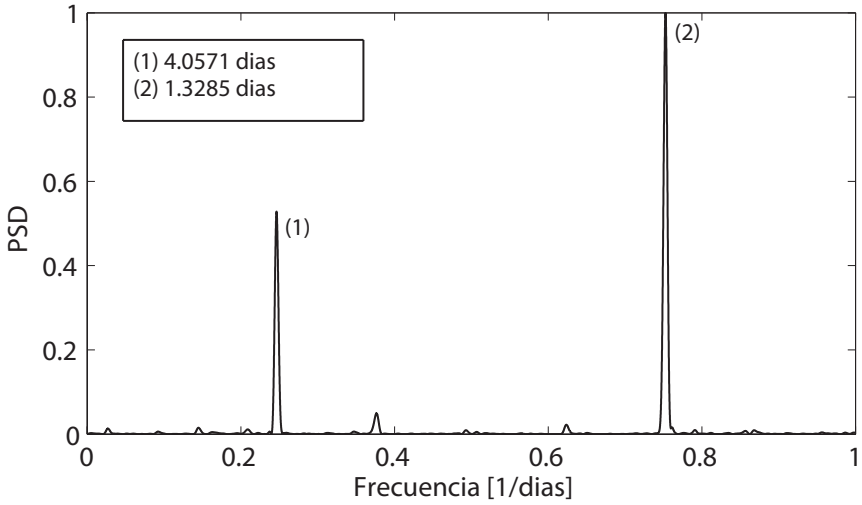


(a) Espectro de correlación

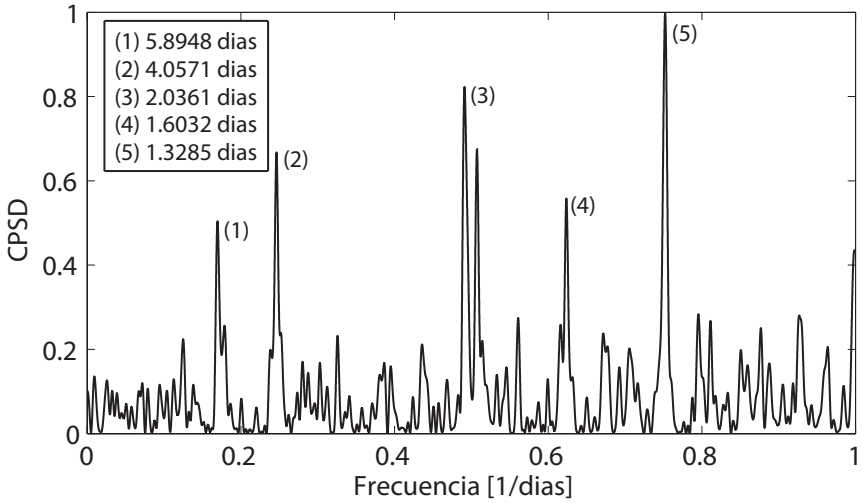


(b) Espectro de correntropía con kernel Gaussiano y $\sigma = 0.1$

Figura 4.2: Espectros de potencia para la curva de luz 1.3449.948 (período reportado 14,0063 días)



(a) Espectro de correlación



(b) Espectro de correntropía con kernel Gaussiano y $\sigma = 0.02$

Figura 4.3: Espectros de potencia para la curva de luz 1.3566.63 (período reportado 1,60459 días)

Se realiza una prueba usando las 193 curvas de luz de estrellas binarias eclipsantes disponibles. La Tabla 4.3 muestra un resumen de los resultados obtenidos por ambos funcionales.

Funcional	Aciertos [%]	Múltiplos [%]	Fallos [%]
Correlación	19.6891	54.4042	25.9067
Correntropía	58.0311	32.1244	9.8446

Tabla 4.3: Comparación entre las versiones ranuradas de la correlación convencional y la correntropía.

Si se analizan los resultados caso a caso se obtiene que

- Todos los aciertos de la correlación están incluidos en los aciertos de la correntropía.
- Entre los aciertos y múltiplos detectados por la correlación no se encuentra ningún fallo de la correntropia.
- Entre los aciertos de la correlación no se encuentra ningún múltiplo de la correntropía.

Esto nos indica que la correlación convencional no aporta información ajena a la entregada por la correntropía, lo cual era esperable pues como se indicó en la sección 2.5 el estadístico de segundo orden está incluido en la correntropía.

4.3. Comparación entre la implementación desarrollada y aplicaciones alternativas

En esta sección se muestra una comparación entre la implementación desarrollada y aplicaciones alternativas utilizadas actualmente para resolver el problema de detección de período en series de tiempo astronómicas. Las aplicaciones alternativas incorporadas en esta comparación son Vartools (periodograma LS), SigSpec y Period04 las cuales fueron descritas detalladamente en la sección 3.3.2. En la Tabla 4.4 se muestran los resultados obtenidos con las aplicaciones alternativas y la mejor versión de la implementación desarrollada (correntropía ranurada combinando los resultados obtenidos con el kernel Gaussiano y el kernel polinomial). En seguida, en la Tabla 4.5 se muestra un desglose de la categoría "Múltiplos" para cada uno de los métodos.

Aplicación	Aciertos [%]	Múltiplos [%]	Fallos [%]
VarTools	4.1451	95.8549	0
SigSpec	4.1451	95.8549	0
Period04	5.6995	94.3004	0
Correntropía	58.0311	32.1244	9.8446

Tabla 4.4: Comparación con otras aplicaciones de detección de período.

Método	Medio período [%]	Otro sub-múltiplo [%]	Otro múltiplo [%]
SigSpec	98.3784	1.6216	0
VarTools	98.3784	1.6216	0
Period04	96.7033	3.2967	0
Correntropía	65.7377	3.2787	30, 9836

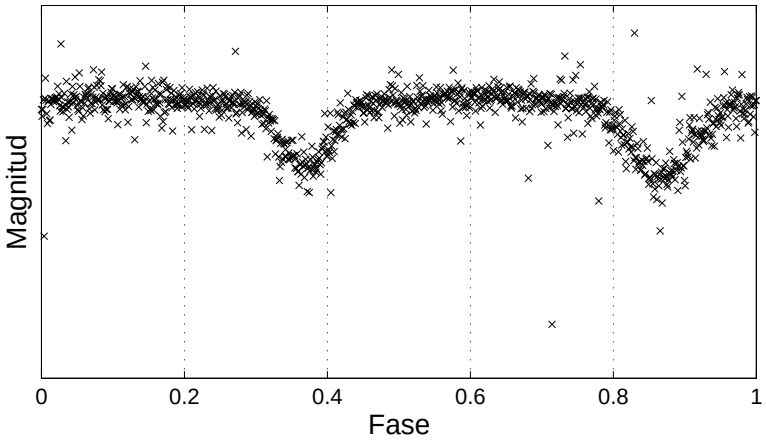
Tabla 4.5: Desglose de las detecciones marcadas como múltiplos

De las Tablas 4.4 y 4.5 se puede notar que

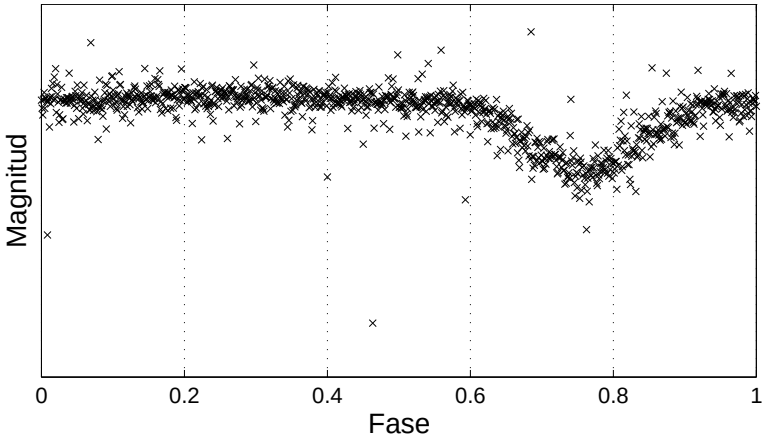
- Las aplicaciones alternativas obtienen resultados muy similares entre sí (en particular SigSpec obtiene los mismos resultados que el periodograma LS). Todas obtuvieron mayoritariamente múltiplos del período real y ninguna obtuvo fallos (la precisión de sus resultados es de 0.01 días, lo que es menor que la tolerancia usada)
- Ninguna de las aplicaciones alternativas fue capaz de alcanzar el porcentaje de aciertos obtenido por el método basado en la correntropía. Además, todos los aciertos obtenidos con VarTools, SigSpec y Period04 se encuentran entre los aciertos encontrados por la correntropía.
- Los métodos competidores sólo obtienen sub-múltiplos del período correcto. En cambio, el método propuesto detecta un porcentaje importante de múltiplos.
- El sub-múltiplo más recurrente para todos los métodos es el medio período.

Las aplicaciones alternativas utilizadas se basan en el ajuste de sinusoides en el sentido de mínimos cuadrados. Si bien esto funciona relativamente bien con otras curvas de luz, queda demostrado que no es apropiado para el caso de estrellas binarias eclipsantes (pues sus variaciones son claramente no sinusoidales). Para aumentar el porcentaje de aciertos de las aplicaciones competidores sería necesario incorporar criterios adicionales, como Análisis de Varianza o *String Length*, para discriminar los períodos presentes en los espectros obtenidos.

Es importante mencionar que detectar la mitad del período correcto no es una casualidad cuando se trabaja con curvas de luz de estrellas binarias eclipsantes. Recordemos que por cada ciclo que desarrolla una estrella de esta categoría ocurren dos eclipses (estrella primaria sobre secundaria y viceversa). Cuando la curva de luz de una estrella binaria eclipsante es doblada usando su período real es posible apreciar claramente ambos eclipses (Figura 2.1.c). Pero, cuando las características radiativas de las componentes del sistema binario son similares, diferenciar un eclipse de un ciclo puede aumentar de dificultad. Un ejemplo de esta situación puede apreciarse en la Figura 4.4, donde se muestran dos curvas de luz dobladas de la estrella binaria 1.3444.614 del catálogo MACHO (doblada con el período reportado (a) y con la mitad de dicho período (b)). Los eclipses mutuos son muy similares en magnitud y forma, por lo que no sería difícil (incluso para una persona) confundir un ciclo (dos eclipses) con un eclipse individual.



(a)



(b)

Figura 4.4: Diagramas de fase de la curva de luz 1.3444.614 del catálogo MACHO doblada con su período real (a) y con la mitad del período real (b).

Capítulo 5

Conclusiones

- Se desarrolla un método automático para la detección de período en series de tiempo astronómicas en base al funcional de correntropía, la técnica de *folding*, la técnica del ranurado y otros métodos de procesamiento de señales. Se realizan pruebas sobre una base de datos de 193 curvas de luz de estrellas binarias eclipsantes del proyecto MACHO cuyo período real es conocido. La mejor versión de la implementación desarrollada[1] logró detectar el período correcto en un 58 % de los casos. En el 42 % restante se obtuvo un múltiplo o sub-múltiplo del período correcto (32 %) o un valor sin relación clara con el período real (10 %).
- Se compara la correntropía con la correlación convencional (ambas en su versión ranurada). Las pruebas realizadas mostraron que la correlación obtiene un menor número de aciertos y un mayor número de fallos que la correntropía. Se verificó que los aciertos obtenidos usando la correlación están incluidos en los aciertos obtenidos con la correntropía, lo cual indica que para esta aplicación el estadístico de segundo orden no captura información que escape a la correntropía.
- Se compara la implementación desarrollada con aplicaciones usadas actualmente en astronomía para detección de período en curvas de luz. Las aplicaciones incluidas en la comparación son Period04, SigSpec y VarTools. En general, las aplicaciones competidoras son incapaces de detectar el período correcto retornando en su lugar el sub-múltiplo más cercano. Al contrario de la correntropía, las aplicaciones competidoras no obtienen

[1] Correntropía ranurada, combinando los resultados obtenidos con el kernel Gaussiano y polinomial inhomogéneo.

ningún resultado que caiga en la categoría de Fallo.

- Un 58 % de aciertos es un porcentaje bajo si se desea trabajar con bases de datos astronómicas de dimensiones reales (un millon curvas de luz o más). Para mejorar el rendimiento del método en el corto y mediano plazo, se propone lo siguiente:

    1. Analizar profundamente los casos en que no se detecta el período correcto. En los casos marcados como múltiplos (32 %) se debe verificar si el período correcto (o un valor cercano) está presente en los espectros analizados, si esto es cierto la métrica de calidad deberá ser revisada. También se deben estudiar las características que tienen en común las curvas de luz cuyos períodos detectados fueron marcados como fallos (10 %).

    2. Modificar la estrategia de enventanado. Hasta ahora se han analizado las sub-series obtenidas de las curvas de luz de forma independiente, escogiendo el mejor período candidato sin importar de cual sub-serie se obtuvo o en cuentas de ellas apareció. El período correcto debería aparecer en más de una de las sub-series, por lo que se propone añadir una verificación de consistencia en la etapa de discriminación de períodos candidatos. Esta idea podría extenderse si se usa una estrategia de ventana deslizante para seleccionar las sub-series reemplazando al criterio de máxima densidad (el cual sólo se preocupa de la cantidad y no de la calidad de las muestras).

    3. Seleccionar el parámetro de la función de kernel[2] de forma óptima, es un tema no tratado en esta memoria de título. Es por esto que se usaron valores predefinidos para dicho parámetro lo cual se tradujo en el computo de un extenso número de espectros para cada curva de luz. Esto es poco eficiente en términos de tiempo de computo lo cual es un punto crítico a considerar cuando se estudien bases de datos mayores. Estudiar como se distribuye el parámetro del kernel en los 193 casos analizados podría ayudar a encontrar patrones que permitan seleccionar el tamaño de kernel apropiadamente.

    4. Añadir una etapa previa a la discriminación con métrica de calidad, en que se realice una búsqueda fina en torno a los períodos candidatos con el fin de aumentar

[2]Orden p para el kernel polinomial y ancho de banda σ para el kernel Gaussiano

su precisión. De esta manera se evitarían los casos en que la métrica de calidad reemplaza un período correcto por un múltiplo más preciso.

5. El cociente de varianzas ha sido usado como métrica de calidad para distinguir el período correcto entre los candidatos obtenidos de los espectros de potencia de correntropía. Sin embargo, esto se contradice con lo propuesto en esta memoria pues el cociente de varianzas es un estadístico de segundo orden. Es posible que una métrica alternativa que considere los estadísticos de mayor orden proporcione mejores resultados que el cociente de varianzas.

6. La aplicación desarrollada trabaja bajo el supuesto de que la curva de luz estudiada corresponde a una estrella variable periódica. Sin embargo, la cantidad de estrellas variables en una base de datos real es baja en comparación al total. Por esta razón se debe considerar la inclusión de un paso previo que verifique variabilidad en la curva de luz, en particular si es periódica o no.

Bibliografía

[1] M. Petit, "Variable Stars", *New York: Wiley*, 1987.

[2] A.C.Becker (U.Washington), C.Alcock (LLNL), R.A.Allsman (ANUSF), D.Alves (STSCI), T.S. Axelrod (MSSSO), D.P. Bennett (Notre Dame), K.H. Cook (LLNL), A.J. Drake, K.C. Freeman (MSSSO), M. Geha (LLNL), K. Griest (UCSD), M.J. Lehner (Sheffield), S.L. Marshall (LLNL), D. Minniti (P. Universidad Catolica), C.A. Nelson (LLNL), B.A. Peterson (MSSSO), P. Popowski (LLNL), M.R. Pratt (Washington), P.J. Quinn (Notre Dame), A.W. Rodgers (), C.W. Stubbs (Washington), W. Sutherland (Oxford), A.B. Tomaney (Washington), T. Vandehei (UCSD), D.L. Welch (McMaster), "The MACHO Project: Microlensing Results from 5.7 Years of LMC Observations", *The Astrophysical Journal*, vol. 542, pp. 281-307, 2000.

[3] C. Miller, "Cosmic Hide and Seek: The Search for the Missing Mass", 1995, disponible en: http://www.eclipse.net/∼cmmiller/DM/

[4] N.R. Lomb, "Least-Squares Frequency Analysis of Unequally Spaced Data", *Astrophysics and Space Science*, vol. 39, pp. 447-462, February, 1976.

[5] J.D. 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, December, 1982.

[6] W.H. Press and G.B. Rybicki, "Fast Algorithm for Spectral Analysis of Unevenly Sampled Data", *The Astrophysical Journal*, vol. 338, pp. 277-280, 1989.

[7] A. Schwarzenberg-Czerny, "On the advantage of using analysis of variance for period search", *Royal Astronomical Society, Monthly Notices*, vol. 241, pp. 153-165, 1989.

[8] M.M.Dworetsky, "A period finding method for sparse randonmly spaced observations", *Royal Astronomy Soc*, vol. 203 pp. 917-924, 1983.

[9] A.Derekas, L.L.Kiss, 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.

[10] P. M. Cincotta, A. Helmi,M. Mendez, J. A. Nunez, H. Vucetich, "A search for periodicity using the Shannon entropy", *Royal Astronomy Soc*, vol. 302 pp. 582-586, 1999.

[11] 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 vailable online at: http://www.cfa.harvard.edu/ jhartman/vartools/.

[12] P. Lenz, M. Breger, "Period04 User Guide", *Communications in Asteroseismology*, vol. 146, pp. 53-136, 2005.

[13] P. Reegen, "SigSpec I. Frequency- and phase-resolved significance in Fourier space",*Astronomy & Astrophysics*, vol. 467, no. 3, pp. 1353-1371, 2007.

[14] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, "Numerical Recipes in C, 2nd ed.", *New York: Cambridge University Press*, 1992.

[15] B.E. Boser and I.M. Guyon and V.N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", *Proceedings of the 5th Annual ACM Workshop on COLT*, pp. 144-152, Pittsburgh, USA, 1992.

[16] J. Debosscher, L.M. Sarro, C. Aerts, J. Cuypers, B. Vandenbussche, R. Garrido, E. Solano, "Automated Supervised Classification of Variable Stars. I. Methodology", *Astronomy and Astrophysics*, vol. 475, pp. 1159-1183, Dec, 2007.

[17] P. Protopapas and J.M. Giammarco and L. Faccioli and M.F. Struble and R. Dave and C. Alcock, "Finding Outlier Light Curves in Catalogues of Periodic Variable Stars", *Monthly Notices of the Royal Astronomical Society*, vol. 369, pp. 677-696, June, 2006.

[18] G. Wachman and R. Khardon and P. Protopapas and C. Alcock, "Kernels for Periodic Time Series Arising in Astronomy", *Proceedings of the European Conference on Machine Learning*, 2009.

[19] E. Parzen, "On the estimation of a probability density function and the mode", *The Annals of Mathematical Statistics*, vol. 33, pp. 1065, 1962.

[20] J.C. Principe, D. Xu, J.W. Fisher III, "Information-Theoretic Learning, Unsupervised Adaptive Filtering", *S. Haykin, New York: Wiley*, 2000.

[21] I. Santamaría and P.P. Pokharel and J.C. Príncipe, "Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization", *IEEE Transactions on Signal Processing*, vol. 54, no. 6, June, 2006.

[22] A. Gunduz and A. Hegde and J.C. Príncipe, "Correntropy as a Novel Measure for Nonlinearity Tests", *Proceedings of the 2006 International Joint Conference on Neural Networks*, Vancouver, Canada, 2006.

[23] R. Li, W. Liu, J.C. Príncipe, "A unifying criterion for instantaneous blind source separation based on correntropy", *IEEE Transactions on on Signal Processing*, vol. 87, no. 8, pp. 1872-1881, 2007.

[24] W. Liu, P,P. Pokharel, J.C. Príncipe, "Correntropy: Properties and Applications in Non-Gaussian Signal Processing", *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286-5298, 2007.

[25] J.W. Xu and J.C. Príncipe, "A Pitch Detector Based on a Generalized Correlation Function", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no.8, 2008.

[26] T.-F. Wu and C.J. Lin and R.C. Weng, "Probability Estimates for Multi-Class Classification by Pairwise Coupling", *Journal of Machine Learning Research*, vol. 5, pp. 975-1005, 2004.

[27] Carl de Boor, "Spline toolbox (for use with matlab) user's guide"

[28] Carl de Boor, "A practical guide to splines", *Chapter XIV: Smoothing and least square approximations, Applied Mathematical Sciences, New York: Springer*, 1978.

[29] W.T. Mayo, "Spectrum measurements with laser velocimeters", *Proceedings of dynamic flow conference*, DISA Electronik A/S DK-2740, Skoolunder, Denmark, 1978.

[30] M.J. Tummers and D.M. Passchier, "Spectral estimation using a variable window and the slotting technique with local normalization", *Measurement Science and Technology* vol. 7, pp. 1541-1546, 1996.

[31] P.A. Estévez, P. Huijse, P. Zegers, J. C. Príncipe, P. Protopapas, "Period Detection in Light Curves from Astronomical Objects Using Correntropy", Congreso mundial de Inteligencia Computacional de la IEEE, Barcelona, España, Jul. 2010.