Estimation of periodicity in non-uniformly sampled astronomical data using a 2D kernel in correntropy

Conference Paper · September 2011

DOI: 10.1109/MLSP.2011.6064635

CITATIONS
0

READS
41

4 authors, including:

Jose C Principe
University of Florida
1,182 PUBLICATIONS **23,475** CITATIONS
SEE PROFILE

Pablo Estevez
University of Chile
156 PUBLICATIONS **2,734** CITATIONS
SEE PROFILE

Pavlos Protopapas
Harvard School of Engineering and Applied Sciences
149 PUBLICATIONS **1,964** CITATIONS
SEE PROFILE

2011 IEEE International Workshop on Machine Learning for Signal Processing
September 18-21, 2011, Beijing, China

ESTIMATION OF PERIODICITY IN NON-UNIFORMLY SAMPLED ASTRONOMICAL DATA USING A 2D KERNEL IN CORRENTROPY

Bibhu Prasad Mishra[1], *Jose C. Principe*[1], *Pablo A. Estévez*[2] and *Pavlos Protopapas*[3]

[1] University of Florida, Computational NeuroEngineering Laboratory
[2] Universidad de Chile, Electrical Engineering Department
[3] Harvard University, IIC, School of Engineering and Applied Sciences

ABSTRACT

Estimation of periodicity in non-uniformly sampled time series data is frequently a goal in astronomical data analysis. There are various problems faced: Firstly, data is sampled non-uniformly which makes it difficult to use simple Fourier transform for performing spectral analysis. Secondly, there are large gaps in data which makes it difficult to interpolate the signal for re-sampling. Thirdly, in data sets with smaller time periods the non-uniformity in sampling and noise in data pose even greater problems because of the lesser number of samples per period. Finally, recent use of CCD technology has enabled collection of vast amounts of data from various sources. In order to process this huge amount of data we also need to remove human intervention from the process of periodicity estimation to make the algorithm more efficient. In the present work we focus on correntropy and design a new spatio-temporal kernel to accurately estimate the time period of the data without any human intervention.

I. INTRODUCTION

Astronomical observations using visual wavelengths are called light curves (i.e. brightness magnitude over time) and are used to quantify either the intrinsic variation of stars such as pulsations and eruptions, or extrinsic variations such as binary stars and planetary transits. Of particular interest is the quantification of the period of light curves, a telltale of objects such as eclipsing binaries, RRLs (pulsating variable stars named after RR Lyrae), cepheids (intrinsically variable stars with exceptionally regular periods of light pulsation), etc [1]. There are several difficulties that need to be addressed in this area. Firstly, the data set normally consist of samples which have been taken at non-uniformly spaced time instants. This prevents the direct usage of Fourier transformation or correlation to study the spectral composition or the periodicity of the signal. One possible alternative is to interpolate the data and re-sample it periodically before applying the method of choice. The presence of gaps in the time series creates further problem as even use of interpolation does not give results in an acceptable range. This problem is avoided by framing the time series data and using the frames which don't have gaps or have very few missing consecutive points. Generally time series data with larger time periods allow more missing points in a frame and also the frame length is larger, whereas for data sets with smaller time periods smaller frame length is used and fewer number of missing consecutive samples are allowed. There is also the problem of noise as each sample point at each time instant has an associated error variance. Finally, although framing and interpolating enable usage of simple standard techniques such as correlation or Fourier Transform this method no longer uses the originally sampled data points directly. This further introduces interpolation noise which can further compromise the precision of period determination.

The time series data analyzed in this work comes from photometric astronomical surveys. These are basically time series of intensity of light collected from various channels like telescopes, different spectral bands or various instruments. The MACHO (Massive Compact Halo Object) survey [2] is operated with the purpose of searching for the missing dark matter in the galactic halo, like brown dwarfs or planets. In MACHO the light amplification is caused by bending of space around a heavy object due to the phenomenon known as microlensing. Due to variations in atmosphere and the sky conditions the data collected is non-uniform in nature and is noisy. Existing techniques mostly use Lomb-Scargle (LS) periodogram [3], [4] which is an extension of classical periodogram techniques but it works with non-uniformly sampled data. The estimated period given by the LS periodogram is used to fold the time series modulo the estimated value for the period so that the periodic nature of data is clearly seen. Then, the estimated period is trimmed such that the scatter of the folded plot is reduced. Once this is achieved it is possible to perform calculations to

THIS RESEARCH WAS SUPPORTED BY NSF UNDER GRANT ECCS 0856441.

978-1-4577-1623-2/11/$26.00

obtain a more precise estimate of the period. This final step known as analysis of variance (AoV) in astronomy is due to [5]. This process is computationally intensive and with data being collected from billions of astronomical objects we need a technique which is more efficient and accurate at the same time. Recently a method proposed in [6] uses slotted correntropy for estimation of periodicity of the light curves. First of all this technique selects a frame of sample points spanning over a period equal to half of the total duration of the light curve and removes samples which have an associated error value greater than a certain threshold. Then slotted correntropy is applied on the samples as defined in [6] and Fourier analysis is further performed on the slotted correntropy result. A fixed number of highest peaks are obtained from the spectral plot. For each of these peaks a range of trial periods values in an interval around the maxima are chosen. Using each of the trial period values folding of the light curve is done and then the folded light curve is divided into dynamically chosen bins and then Information Potential (IP) is calculated on each of these bins and average is taken. The final metric is used to choose the estimate of the period. In the step where slotted correntropy is performed 25 different kernel sizes are used and highest peaks of the spectral plot in each of these cases are used. This method has been found to give better accuracy than the existing techniques but the drawback is that it involves a lot of steps, does not make use of the temporal information while selecting trial periods and involves use of various kernel sizes. Hence in the current work we look into a method which makes use of the temporal information directly and fewer steps along with use of a single kernel size chosen according to the light curve data. This inherent difficulty of the problem requires computationally intelligent techniques [7], [8], [9], [10], [11] to solve the problem. The present work proposes an algorithm using Information Theoretic approach based on correntropy [12], [13], [14] with a new spatio-temporal kernel. We will be comparing the results of the current work with the algorithm proposed in [6].

The rest of the paper is organized as follows: Section II describes necessary theoretical background for this study. Section III illustrates the new proposed method in detail. Next, the section IV shows experimental results with discussion on the same. Finally, section V draws the principal conclusion.

II. THEORETICAL BACKGROUND

Correntropy uses a positive definite kernel to calculate a generalized correlation function. Correntropy is a function of two arguments similar to correlation but with the addition of higher order moments introduced by the kernel function. There are various types of kernel functions used like Gaussian, Spline or Sigmoid but in this particular case we have used the Gaussian kernel. Correntropy has been found to produce sharper and narrower peaks corresponding to similarity estimation as compared to correlation function. Given a random process $\{x_t : t \epsilon T\}$ where t denotes time and T the index set of interest, correntropy function is defined as;

$$V(t,s) = E[\kappa(x_t, x_s)] \tag{1}$$

$$\kappa(x_t, x_s) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{(x_t - x_s)^2}{2\sigma^2}\}} \tag{2}$$

Applying Taylor series expansion to the Gaussian kernel we can express the correntropy function as;

$$V(t,s) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2\sigma^2)^k k!} E[(x_t - x_s)^{2k}] \tag{3}$$

To obtain a univariate correntropy function we must require that the even moment terms are shift invariant which is a stronger condition than the wide sense stationary condition required by correlation function. Correntropy can be estimated directly from samples using the ergodicity assumption as;

$$V(\tau) = V(t+\tau, t) = \frac{1}{N} \sum_{n=0}^{N-1} \kappa(x_n, x_{n+\tau}) \tag{4}$$

and a strict stationarity on even moments is sufficient when Gaussian kernel is used. Another well defined metric which correntropy induces in the input space is correntropy induced metric (CIM) [15]. CIM is defined as;

$$CIM(X,Y) = (\kappa(0) - V(X,Y))^{1/2} \tag{5}$$

For Gaussian kernel it has been observed that CIM behaves as L2 norm when two vectors are close, as L1 norm outside L2 norm and as they go farther apart it becomes insensitive to distance i.e. behaves as L0 norm. The space over which it behaves as L2 norm or L0 norm directly depends on the kernel size σ. This unique property of CIM is very useful in rejecting outliers.

III. PROPOSED METHOD

This section defines a 2 dimensional kernel based correntropy and then uses it for quantification of the likely period. Before describing the steps first we discuss the idea behind this proposed technique.

A periodic signal repeats itself after a fixed interval of time. If we compare two samples which have been collected at intervals equal to a multiple of the period of the signal then it is expected that these values are equal in magnitude. In our case this happens rarely because, first of all, the signal is non-uniformly sampled with gaps and there are lots of noise and modulations. But still if we take two samples at an interval close to the multiple of the actual time period then the magnitude will be comparable too. This idea suggests that one should be folding the observations to the principal argument of the period. Thus if we know the period

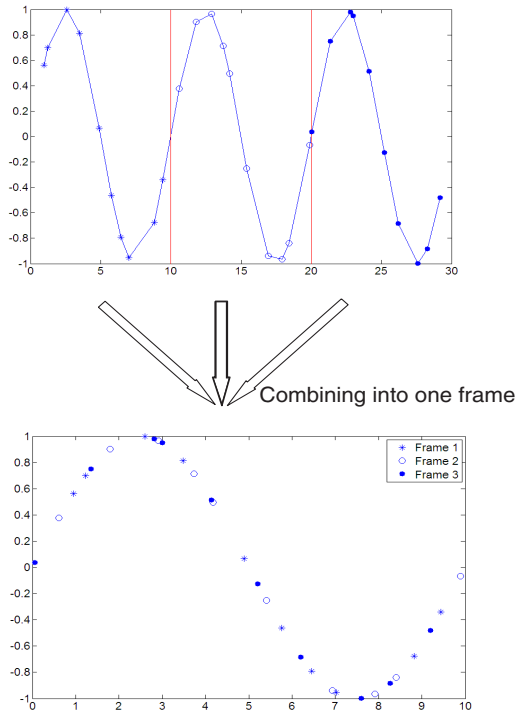


Fig. 1. Reconstruction of single period of the signal by breaking the original signal into frames of length equal to the actual Time period of the signal

we can reconstruct one period data as $x(t) = x(t + nT)$ where T is the period and n is an integer. This idea is illustrated in Figure 1 where the signal with a time period of 10 units and average sampling time of 1 unit is used to reconstruct a single period. If we fold the data using a value of T which is not a multiple of the actual period then the actual signal would not be obtained. It is easy to see that the period T will yield the smoothest representation in the principal argument domain. Therefore one needs to find a methodology to compare the similarity of the samples both in time and in amplitude, which will be implemented with a two dimensional kernel. We saw how we can create a single period of the signal by knowing the actual period. Unfortunately this method is greedy, and many possible values of trial period needs to be evaluated to obtain the period for which the similarity is the highest.

More concretely, we define a two dimension vector $\mathbf{h}$ which has time value in one dimension and magnitude value in the other. It is expressed as $\mathbf{h_a} = [t_a, x_a]^T$ and $\mathbf{h_b} = [t_b, x_b]^T$. The product kernel κ is defined as;

$$\kappa(\mathbf{h_a}, \mathbf{h_b}) = \kappa_1(t_a, t_b) \times \kappa_2(x_a, x_b) \tag{6}$$

where κ_1 and κ_2 are both Gaussian kernel as defined in equation 2 defined on time (t) and magnitude (x) component of the data set respectively. This kernel is still positive definite, being effectively a Gaussian kernel with diagonal covariance matrix with first diagonal component σ_1 dealing with time component t_k and second diagonal component σ_2 dealing with magnitude of data x_k at that time instant. Using the newly defined kernel the correntropy equation is defined as follows;

$$V_P = \frac{1}{N-1} \sum_{i=1}^{N-1} \kappa(\mathbf{h_i}, \mathbf{h_{i+1}}) \tag{7}$$

where $\mathbf{h_i}$ is a fixed sequence of vectors denoted by P.

The algorithm for the period T estimation is as follows:

1) Let $H = \{\mathbf{h_k} = [t_k, x_k]^T, 1 < k < N\}$ where N is the total number of data points obtained by selecting frames of the light curve.
2) For the trial period $T = p$, the transformation ϕ_p on H is such that $\phi_p(H) = Y$ where $Y = \{\mathbf{\Psi_k} = [\tau_k, x_k]^T, 1 < k < N\}$ s.t. $\tau_k = (t_k - \left\lfloor \frac{t_k}{p} \right\rfloor p)/p$ where $\lfloor \cdot \rfloor$ is floor function.
3) Then we order the transformed vectors such that $\mathbf{\Psi_{k_i}}$ precedes $\mathbf{\Psi_{k_{i+1}}}$ if $\tau_{k_i} <= \tau_{k_{i+1}}$. If $\tau_{k_i} = \tau_{k_{i+1}}$ we order the amplitudes s.t. $x_{k_i} <= x_{k_{i+1}}$
4) Calculate correntropy with the 2D kernel $V(p)$ using transformed vectors as shown in Equation 7
5) Calculate correntropy with the time kernel only, as a normalizing factor $U(p)$ where $U(p) = \frac{1}{N-1} \sum_{i=1}^{N-1} \kappa_1(\tau_{k_i}, \tau_{k_{i+1}})$
6) Vary the value of p over a range and repeat from step 2 to step 5.
7) The value of p which gives first significant peak in the plot of $V(p)/U(p)$ is the desired period.

In the above algorithm the range depends on some apriori knowledge of the periods of interest. The range of 0.25 to 200 is used and step size of $10^{(-4)}$ is used. For lower periods i.e. for values less than 2 days step size of $10^{(-5)}$ is used. The reason behind different step sizes is that for lower values of p a small deviation in the estimated period value can give noisy period reconstruction as the number of cycles is larger in the given data set. A significant peak is defined as a peak which exceeds 90% of the dynamic range of the plot. In the method proposed in the current work we see that folded light curve is used to compute the correntropy measure whereas in case of [6] folded light curve was used to obtain the IP based measure. In the present work the trial period values are directly used with a fixed kernel size for a particular light curve whereas in case of [6] the trial period values are selected from the Correntropy spectral density (CSD) plot for various kernel sizes. Also an interesting aspect of the proposed method is that it is able to use the non-uniform sampling to its advantage to detect periodicities corresponding to frequencies above the Nyquist rate. This is possible because of the folding of the light curves along with the usage of the

Table I. Comparison of proposed technique and method presented in [6] with results published by Time Series Center, Harvard University being used as the golden standard in a subset of 200 EB light curves from MACHO survey

Method	Hits[%]	Multiples[%]	Misses[%]
Proposed Technique	*69.0*	*31.0*	*0.0*
Slotted Correntropy	*74.0*	*25.5*	*0.5*

Table II. Comparison of proposed technique and method presented in [6] with results published by Time Series Center, Harvard University being used as the golden standard in a subset of 400 Cepheid and RRL light curves from MACHO survey

Method	Hits[%]	Multiples[%]	Misses[%]
Proposed Technique	*94.25*	*3.5*	*2.25*
Slotted Correntropy	*97.0*	*2.75*	*0.25*

temporal information of the folded light curve in the method.

IV. RESULTS AND DISCUSSION

The results are obtained by applying our proposed correntropy based technique on light curves from the MACHO survey. We use the results published by Time Series Center, Harvard University as a golden standard. These results have been obtained by using AoV and visually inspected by the Time Series Center team. In this section we first look into the selection of the kernel sizes. The value of σ_1 is considered w.r.t. average sampling period (determined by dividing the time interval over which all transformed vectors are spread by total number of vectors) for choosing an appropriate value of standard deviation for the time kernel. We can observe in Figure 2(a) that the peak becomes more prominent by increasing $\sigma_1 \times (Average\ Sampling\ Rate)$ and we also take into consideration the fact that consecutive vectors in time passed through the kernel should be given more importance as compared to vectors which are far apart from each other in time. Giving more weight to consecutive vectors is especially more significant as we are trying to measure similarity between the transformed 2D vectors which are consecutive in time, during the implementation of our proposed technique. To give more importance to vectors which are closer in time we need to reduce the kernel size. So a trade-off is considered between these two opposing factors and we have $\sigma_1 \times (Average\ Sampling\ Rate) = 1$. One thing to be noticed is that the average sampling rate is always fixed for all values of trial period while scanning over a range because in the proposed technique we scale all the 2D vectors in the time range $0-1$ after performing the modulo operation and the total number of vectors is fixed.

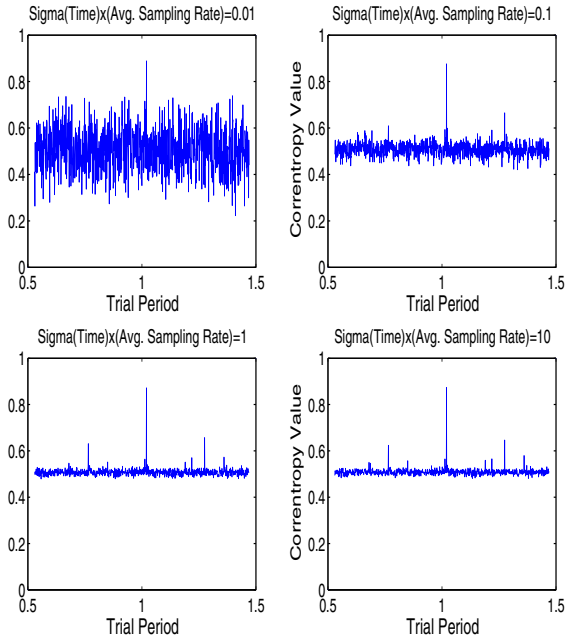
Similarly for magnitude the value of σ_2 is considered w.r.t. amplitude dynamic range. Choosing a very large kernel size means any two magnitude values from the corresponding vectors passed through the kernel will give similar output as the kernel tapers very slowly. Choosing a very small kernel size would give an output of 1 only when we have equal magnitude values and give output close to zero for any other pair of amplitude values. This is clearly reflected in Figure 2(b) where in the plot of $\sigma_2/(Magnitude\ Dynamic\ Range)$ vs $Trial\ Period$ we see a larger kernel size gives a flat plot having a value close to one irrespective of the trial period value and where a small kernel size gives a plot having value close to zero. Therefore to obtain a sharper peak at the true period we choose $\sigma_2/(Magnitude\ Dynamic\ Range) = 0.1$ as the optimum value.

For simplicity and to have the plot values restricted between 0 and 1 we drop the normalizing factor for unit integral in the Gaussian kernel.

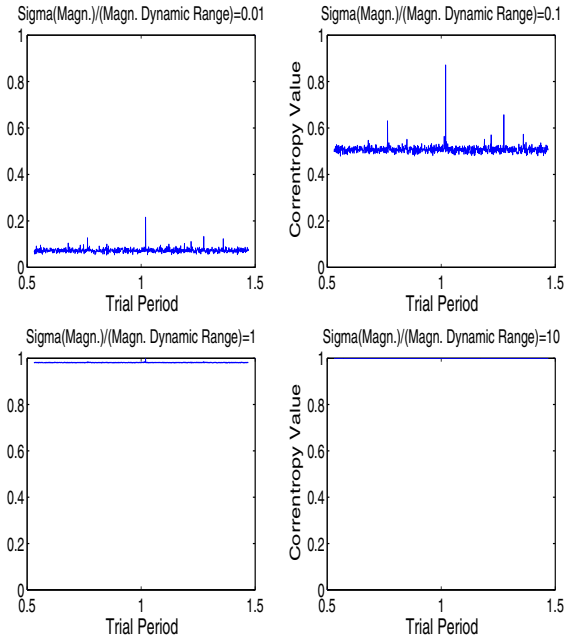
Now we present the results obtained by testing and comparing our correntropy based proposed algorithm to the algorithm presented in [6]. In Tables I and II the column indicating multiples means the estimated period in those percentage of cases were integral multiples or sub-multiple of the true period. In Table I we see that the accuracy obtained for the proposed method and also the slotted correntropy is less compared to accuracy obtained in Table II for Cepheids and RRL light curves. Large percentage of EB light curves give an estimated period value which is sub-multiple of the true period. This means peaks have been obtained at a sub-multiple of the actual time period and the difference between the maximum peak value and the peak value at sub-multiple is less than 10% of dynamic range of the plot obtained using the proposed technique. This is illustrated in Figure 3(a). We see a distinct but smaller peak at half of the true period. This is true for all the EB light curves which do not give the accurate estimate of the true period. The proposed algorithm in fact produces multiple peaks at integral multiple of the true period. The reason for getting a peak at sub-multiple of true period in case of EB light curves is due to the shape of the signal which can be seen in Figure 3(b). We see the modulation effect inside a period which is responsible for the peak at a value which is half of the true period. Therefore this will be very difficult to discern with the current correntropy technique and further processing will be necessary to cope with this phenomenon. Ideas from pitch detection in speech may be very appropriate, but will be object of further research.

V. CONCLUSION

In this work we have introduced a novel spatio-temporal kernel in correntropy which gives lower but comparable accuracy to the recently proposed slotted correntropy technique. The proposed method is able to overcome the difficulties faced due to non-uniform sampling, large gaps in



(a) Correntropy plot with varying standard deviation values for time kernel



(b) Correntropy plot with varying standard deviation values for magnitude kernel

Fig. 2. Determination of kernel size using light curve 1.3442.172 with true period of 1.02 days as an example

data and light curves with smaller time periods and along with it removes the need of manual verification of the data. The most interesting aspect of the proposed method is that it uses the temporal information of the data in the definition of the kernel function for correntropy, whereas most existing methods tend to avoid the use of temporal information. Popular techniques such as Lomb Periodogram which uses the temporal information tend to give very low accuracy (close to 10%) in case of EB light curve data due to their inherent modulation. Another advantage is that it provides better accuracy to light curves with periodicities that are close to the sampling frequency because of the intrinsic folding that is created by the kernel. Essentially we can estimate frequencies above the Nyquist rate. Finally, when compared with the slotted correntropy, the proposed method is much more straight forward to apply because the weighting is again included in the kernel definition. This is also due to the fact that in [6] the slotted correntropy method approximates the time values according to time slots. Hence it needs multiple peaks and various kernel sizes to help capture a value close to the true period for folding operation prior to using information potential. However one drawback of the proposed method is that it is greedy as it scans through the value of trial periods to give a similarity measure over a certain range which is used to identify the true period of the light curve. Future work will deal with the issue of kernel design which we believe is the source of the lesser performance w.r.t. slotted correntropy. In fact, the Gaussian in time may be too broad to provide sufficient accuracy whereas a Laplacian kernel may be more appropriate. We also need to reduce the search for the trial period and also develop methods to identify the peaks associated with the true period more accurately.

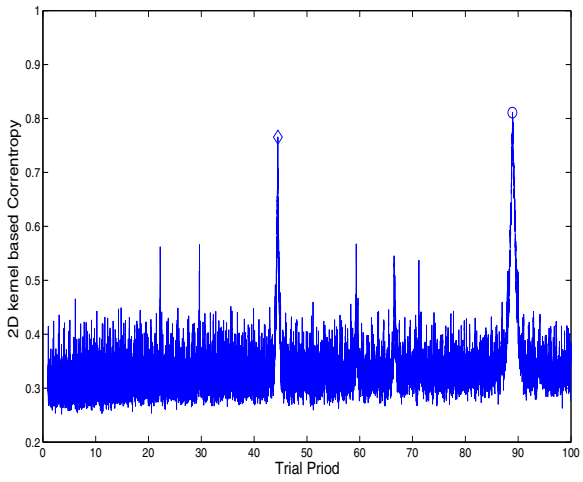
VI. REFERENCES

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

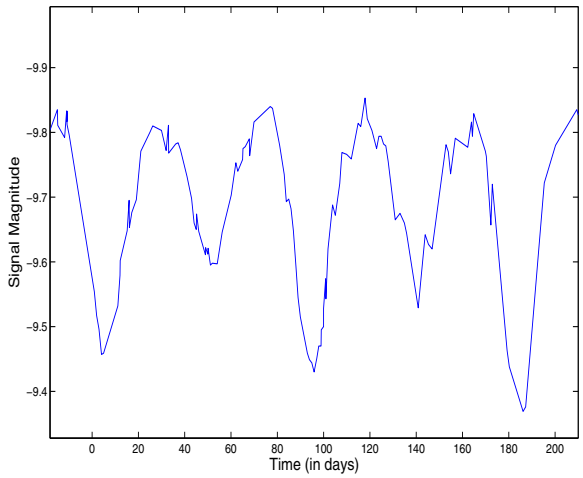
[2] C. Alcock, R.A. Allsman, D.R. Alves, T.S. Axelrod, A.C. Becker, D.P. Bennett, K.H. Cook, N. Dalal, A.J. Drake, K.C. Freeman, M. Geha, K. Griest, M.J. Lehner, S.L. Marshall, D. Minniti, C.A. Nelson, B.A. Peterson, P. Popowski, M.R. Pratt, P.J. Quinn, C.W. Stubbs, W. Sutherland, A.B. Tomaney, T. Vandehei and D. Welch, "The MACHO Project: Microlensing Results from 5.7 Years of LMC Observations," Astrophysical Journal, vol. 542, pp. 281-307, 2000.

[3] N.R. Lomb, "Least Square Frequency Analysis of Unequally Spaced Data", Astrophysics and Space Science, vol. 39, Feb. 1976, p. 447-462.

[4] J. D. Scargle, "Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data," The Astrophysical Journals,



(a) Correntropy plot for the light curve 1.3810.19 obtained by our proposed technique



(b) Plot of signal from light curve 1.3810.19

Fig. 3. Plot of light curve 1.3810.19 with true period at 88.94 days and its correntropy plot using the proposed algorithm

vol. 263, pp. 835-853, December, 1982.

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

[6] Pablo Huijse, Pablo A. Estévez, Pablo Zegers, Jose C. Principe, and Pavlos Protopapas, "Period Estimation in Astronomical Time Series Using Slotted Correntropy," IEEE Signal Processing Letters, vol. 18, no. 6, June 2011

[7] B. E. Boser, 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.

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

[9] P. Protopapas, J. M. Giammarco, L. Faccioli, M. F. Struble, 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.

[10] G. Wachman, R. Khardon, P. Protopapas and C. Alcock, "Kernels for Periodic Time Series Arising in Astronomy," Proceedings of the European Conference on Machine Learning, Lecture Notes in Computer Science, Vol. 5782, pp. 489-505, 2009.

[11] T.-F. Wu, 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.

[12] Jian-Wu Xu, Puskal P. Pokharel, Antonio R.C.Paiva and Jose C. Principe, "Non-Linear Component Analysis based on Correntropy",IJCNN, July 16-21,2006

[13] A. Gunduz and J. C. Principe, "Correntropy as a Novel Measure for Nonlinearity Tests," Signal Processing, vol. 89, pp. 147-23, 2009.

[14] I. Santamaŕia, P. P. Pokharel and J. C. Principe, "Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization," IEEE Transactions on Signal Processing, vol. 54, no. 6, pp. 2187-2197, June, 2006.

[15] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," IEEE Transactions on Signal Processing, vol. 55, no. 11, pp. 52865298, 2007.