Classifying Periodic Astrophysical Phenomena from non-survey optimized variable-cadence observational data

Paul R. McWhirter[a,b], Abir Hussain[a], Dhiya Al-Jumeily[a], Iain A. Steele[b], Marley M. B. R. Vellasco[c]

[a] *Applied Computing Research Group, Department of Computer Science, LJMU, James Parsons Building, 3 Byrom Street, Liverpool, L3 3AF, UK*
[b] *Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool L3 5RF, UK*
[c] *Pontifícia Universidade Católica do Rio de Janeiro, R. Marquês de São Vicente, 225 - Gávea, Rio de Janeiro - RJ, 22451-900, Brazil*

Abstract

Modern time-domain astronomy is capable of collecting a staggeringly large amount of data on millions of objects in real time. Therefore, the production of methods and systems for the automated classification of time-domain astronomical objects is of great importance. The Liverpool Telescope has a number of wide-field image gathering instruments mounted upon its structure, the Small Telescopes Installed at the Liverpool Telescope. These instruments have been in operation since March 2009 gathering data of large areas of sky around the current field of view of the main telescope generating a large dataset containing millions of light sources. The instruments are inexpensive to run as they do not require a separate telescope to operate but this style of surveying the sky introduces structured artifacts into our data due to the variable cadence at which sky fields are resampled. These artifacts can make light sources appear variable and must be addressed in any processing method.

The data from large sky surveys can lead to the discovery of interesting new variable objects. Efficient software and analysis tools are required to rapidly de-

Email addresses: P.R.McWhirter@ljmu.ac.uk (Paul R. McWhirter), A.Hussain@ljmu.ac.uk (Abir Hussain), D.Aljumeily@ljmu.ac.uk (Dhiya Al-Jumeily), I.A.Steele@ljmu.ac.uk (Iain A. Steele), marley@ele.puc-rio.br (Marley M. B. R. Vellasco)

termine which potentially variable objects are worthy of further telescope time. Machine learning offers a solution to the quick detection of variability by characterising the detected signals relative to previously seen exemplars. In this paper, we introduce a processing system designed for use with the Liverpool Telescope identifying potentially interesting objects through the application of a novel representation learning approach to data collected automatically from the wide-field instruments. Our method automatically produces a set of classification features by applying Principal Component Analysis on set of variable light curves using a piecewise polynomial fitted via a genetic algorithm applied to the epoch-folded data. The epoch-folding requires the selection of a candidate period for variable light curves identified using a genetic algorithm period estimation method specifically developed for this dataset. A Random Forest classifier is then used to classify the learned features to determine if a light curve is generated by an object of interest. This system allows for the telescope to automatically identify new targets through passive observations which do not affect day-to-day operations as the unique artifacts resulting from such a survey method are incorporated into the methods.

We demonstrate the power of this feature extraction method compared to feature engineering performed by previous studies by training classification models on 859 light curves of 12 known variable star classes from our dataset. We show that our new features produce a model with a superior mean cross-validation F1 score of 0.4729 with a standard deviation of 0.0931 compared with the engineered features at 0.3902 with a standard deviation of 0.0619. We show that the features extracted from the representation learning are given relatively high importance in the final classification model. Additionally, we compare engineered features computed on the interpolated polynomial fits and show that they produce more reliable distributions than those fit to the raw light curve when the period estimation is correct.

Keywords: Astronomical Time-Series, Light Curve Analysis, Period Analysis, Variable Stars, Binary Stars, Random Forest Classification

1. Introduction

Time Domain Astronomy is a field of research addressing astronomical objects and phenomena responsible for the production of independent light sources that exhibit variation of timescales detectable by instrumentation. These objects can exhibit intrinsic variability due to changes in the structure of an object or extrinsic variability of separate structures. Analysis of these objects grants valuable information into physics and the wider universe. A subset of these variable objects that exhibit periodic variability that, if correctly identified, can be used to perform a variety of important tasks such as distance measurement through standard-candle methods. The ability to reliably observe these light sources is rapidly improving through the development of new technological solutions, both hardware and software based.

Advances in observational, storage and data processing technologies have allowed for extended sky surveys to be conducted and exploited. These surveys range from focused observations of specific regions of the sky such as the MACHO (Alcock et al., 2000), EROS (Rahal et al., 2009) and OGLE (Udalski et al., 1997) surveys to extended sky surveys probing large swathes of the night sky such as SDSS (York et al., 2000), Pan-STARRS (Kaiser, 2002) and CRTS (Larson, 2003). This progress continues to enhance observational capability with the construction of the Large Synoptic Survey Telescope (LSST) in northern Chile due to commence operations at the beginning of the next decade (Ivezic, 2014). With this constant improvement in capability, the fields of Astronomy, Computer Science, Computational Intelligence and Statistics are striving to develop efficient implementations of multiple algorithms that can describe the properties of observed light sources and correctly classify them.

Time domain astronomy is characterised by the large datasets generated by sky surveys containing time-series data (Vaughan, 2011). Time-series data contains information on the temporal component of measurements and the whole time-series contains multiple observations at differing times. Most data-gathering exercises outside of astronomy result in a large number of observations with consistent time intervals between individual observations. In Time-Domain Astronomy, it is common for these observations to have a significantly uneven distribution in time with inconsistent intervals between observations (Lomb, 1976, Scargle, 1982). Major causes of this include weather limitations that can prevent telescope operation for uncertain periods and limited access time to telescopes due to the volume of astronomers requiring observations. As a result, astronomy requires data processing capable of automated analysis of time-series data on individual objects that can contain a cluster of observations over the space of days followed by no additional observations for a period of months (Lomb, 1976, Scargle, 1982).

This paper will focus on our system for processing uneven-cadence time-series astronomical light curve data through representation learning of useful features using Principal Component Analysis (PCA) unsupervised learning on PolyFit interpolated phased light curves using a period determined by a genetic algorithm period estimation method and a Random Forest machine learning algorithm for classification. The data is in the form of a wide field object SQL database containing millions of stellar objects generated from observational images gathered by the Small Telescopes Installed at the Liverpool Telescope (STILT) instruments (Steele et al., 2004, Mawson et al., 2013). The database contains time-series data on the magnitude (i.e. brightness) of detected objects over a period from March 2009 to March 2012 (Mawson et al., 2013). The specific nature of the collection of this data using images captured by passive surveying, i.e. the instrument has no control over the telescope, is greatly advantageous as it does not take up telescope time. Unfortunately, it does introduce sampling based artifacts into the data that must be accurately determined by the processing pipeline. The successful development of a method to perform survey astronomy whilst compensating for sampling artifacts will allow a novel implementation of new wide field variability surveys at a relatively low cost to former variability surveys.

The remainder of this paper is organised as follows. §2 will discuss the gathering of data to produce light curves and properties of these for variable stars while §3 will discuss our methodology to extract descriptive light curve features independent of the sampling cadence. §4 demonstrates experimentally that the new features are atleast as useful as those from previous studies. Finally §5 discusses the possibilities for a variable star classification system based on the proposed method and concludes the paper with proposed further investigations.

2. Light Curves and Variable Stars

Photometry is a branch of instrumentation concerned with the precise measurement of the visible-wavelength electromagnetic radiation (light) captured by an appropriate instrument from a light source. Photometric data on a large number of objects can be generated through the production of wide-angle images of the sky. The intensity of the image pixels is determined by the activation of the Charge-Coupled Device (CCD) cameras pixels by incoming light from multiple astronomical objects with some background noise and detection bias from the camera (Mawson et al., 2013). As a result, each image contains important information about the brightness (magnitude) of the detected objects. By identifying objects in multiple images with different observation times, information on the change of the brightness of these objects can be determined. This task itself is non-trivial as the objects could be located in different regions of consecutive images due to the motion of the telescope. The resulting brightness-over-time data for each individual object is defined as the objects light curve (Lomb, 1976, Scargle, 1982, Huijse et al., 2012).

Light curves present a quantity of useful data on a light source in the form of a time-series. This time-series is univariate with magnitudes, magnitude error and the associated time instants of measurement. Magnitude is a logarithmic brightness scale used by astronomers as shown in Equation 1.

$$m - m_{\mathrm{ref}} = -2.5 \log_{10} \frac{F}{F_{\mathrm{ref}}} \tag{1}$$

Where m represents the apparent magnitude of a detected source (i.e. the magnitude of the object as it appears from Earth), m_{ref} represents the apparent magnitude of a suitably chosen reference source, F is the total flux of the detected source and F_{ref} is the total flux of the reference source. Flux is a measure of the quantity of light detected by the CCD instrument.

This data can be manually manipulated by experienced astronomers to reveal a wealth of properties associated with the light source object(s). However, the number of light curves being generated by successive extended sky surveys has already passed the point where it is unfeasible for these light curves to be manually analyzed. There are a number of problems associated with the extraction of useful information from light curve time-series in which computational intelligence algorithms are of extreme interest. These problems can be categorized as a parameter (feature) extraction process, an experience-based classification operation and an organizational method that attempts to identify structure across the large assortment of light curves (Richards et al., 2011b). These problems appear to be precisely positioned for exploitation by modern machine learning and computational intelligence methods.

The resultant databases from such extended sky surveys can be dauntingly large potentially containing the light curves of millions of individual light sources. Additionally, the data itself exhibits a number of characteristics that can prove greatly detrimental to the efficient and accurate analysis of the light curves. The dominant property of astronomical light curves is the sampling of these light curves. Whilst surveys will attempt to optimize for a specific sampling rate, a property named cadence, limitations in observational schedules and telescope limitations result in uneven sampling containing artifacts such as gaps in the dataset and non-integer deviations from the desired sampling rate. For example, Earth based observations have an unavoidable periodic one-day gap in observations due to the inability to observe during daytime hours. As well as sampling artifacts, there are also periodic light variations due to local cycles such as the orbit of the moon resulting in different phases, which periodically vary the background sky brightness through the monthly cycle. Additionally, a number of noise sources often affect astronomical data. The Earths atmosphere can result in noise in the coordinate positioning of light sources as well as refraction and extinction resulting in variations to the measured brightness. CCD cameras used to gather this data are subject to two major sources of noise. Each pixel on a CCD will have slight difference in light sensitivity called flat field error and there is a thermal noise caused by thermal photons produced by the instrument. These noise sources are usually limited through the production of flat field and dark frames as well as the cooling of the CCD cameras.

The analysis of variable astronomical objects is a major element in the understanding of stellar and galactic evolution as well as the topology of the universe (Richards et al., 2011b). Many astronomical objects exhibit brightness variability due to a large number of differing physical processes that uniquely influence an objects light curve. Therefore, the light curve can be used in the classification of variable objects based on the signature of these potentially periodic physical processes and the detection of unknown candidate objects or even unknown variability phenomena that might be due to previously unrecognized astrophysical processes (Protopapas et al., 2006). Figure 1 demonstrates the light curve of a pulsating variable star of classical Cepheid classification (Eyer and Mowlavi, 2008). Pulsating stars are unstable stars and undergo periods of pulsation where they grow and contract in size (Percy, 2008). These size oscillations produce changes to the stars temperature and brightness resulting in a measurable change upon the light curves (Lomb, 1976, Scargle, 1982, Huijse et al., 2012).

Another important type of variable object is the eclipsing binary (LaCourse et al., 2015). In these systems, two or more stars are in close proximity to each other and execute orbits around a common gravitational centre-point. The close proximity of the stars often means that they cannot be distinguished on an image and appear as a single source of light. Variations in these objects are caused by the plane of the orbit aligning with the view from Earth. As a result, one star periodically passes in front of another resulting in a change in either the brightness and/or the relative brightness in different colour filters of the source of light in the astronomical images. These different types of variable object must be catalogued from the initial surveys so that they can be used by

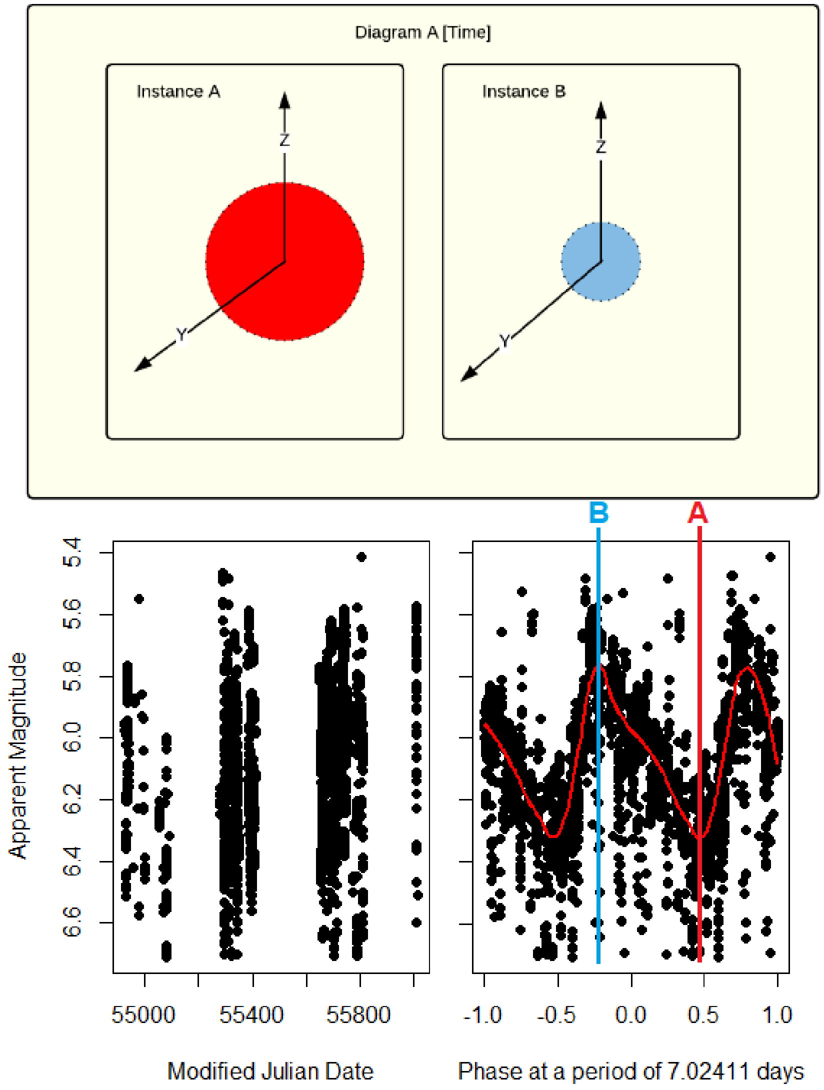


Figure 1: The light curve of the star U Aquilae, a pulsating classical Cepheid star. The star expands and contracts which produces a clear oscillation in the light curve when phased to the known period of 7.02 days.

astronomers for additional research.

2.1. Small Telescopes Installed at the Liverpool Telescope

The Liverpool Telescope is a fully robotic two-metre class telescope located at the Observatorio del Roque de los Muchachos on the island of La Palma, Canary Islands. It is administered by a collaboration between Liverpool John Moores University and the Instituto de Astrofisica de Canarias (Steele et al., 2004, Mawson et al., 2013). The Small Telescopes at the Liverpool Telescope (STILT) are a set of wide field imaging devices that complement the instrumentation available to the Liverpool Telescope (Mawson et al., 2013). The STILT instruments consist of three instruments with Andor Ikon-M DU934N-BV CCD cameras (Steele et al., 2004, Copperwheat et al., 2016) detecting unfiltered optical wavelength white light (all electromagnetic radiation across the visible spectrum). They are named SkycamA with a whole sky view, SkycamT with a $9^\circ \times 9^\circ$ (formerly $21^\circ \times 21^\circ$) field of view (FoV) and SkycamZ with a $1^\circ \times 1^\circ$ FoV. They have varying field of views and are mounted directly to the body of the main Liverpool Telescope aimed co-parallel with the main telescopes focus. These instruments have no control over the motion of the Liverpool Telescope and simply take exposures as directed by a small Asus eee pc-powered control unit (Mawson et al., 2013). The sky coverage and cadence of the Skycams is highly variable and is not optimised for any particular survey or science program (Mawson et al., 2013).

The STILT dataset is a Structured Query Language (SQL) database of multi-object photometry deployed on the MySQL platform. It contains 1.24 billion separate object observations of 27.74 million independent stellar objects. The database contains time-series data on the magnitude of detected objects over a period of time from March 2009 to March 2012 for SkycamT and July 2009 to March 2012 for SkycamZ (Mawson et al., 2013, McWhirter et al., 2016).

To create this MySQL database, the time-stamped observational images are processed by a data reduction pipeline (Mawson et al., 2013). Software using this methodology corrects the raw images using dark and flat frames.

The dark current and bias noise correction is accomplished by using a single reduction frame. This reduction frame consists of between 30 and 210 dark frames generated by obtaining exposures of the inside of the dome at midnight on nights where the weather prevents observing. These stacked frames must then be updated on a weekly basis. Upon the removal of these known sources of noise, the images are then fit to the World Coordinate System (WCS), a system that allows the location of the frame in the sky to be determined and recorded accurately. This is required as the Skycams are not linked to the Liverpool Telescopes computer systems and therefore they have no knowledge of the current coordinates of the telescopes primary field of view (Mawson et al., 2013). This necessitates the fitting of WCS information through the use of Blind Astrometric Calibration. This is accomplished through the use of two pieces of software, Source Extractor (SExtractor) and Astrometry.net. Source Extractor is capable of identifying the sources of light present in an image and outputting information about them such as their pixel coordinates, the Flux (intensity) of the sources, their ellipticity (how elliptical the light source is on the image) and properties of the size of the source such as the isophotal area (area of the same brightness) (Bertin and Arnouts, 1996). It is also capable of filtering out artifacts to maintain the purity of the sources identified. The second piece of software, Astrometry.net, uses the observing frame to determine its coordinates (Lang et al., 2010). This is accomplished by assigning each source extracted a unique hash key generated by a quadrilateral produced by the four nearby bright sources. This hash key is generated in a specific manner such that it is invariant to the images orientation and scale allowing it to function successfully for images with various fields of view. The authors of the software claim more than a 99% success rate for contemporary near-ultraviolet (near-UV) and optical survey data with zero false positives for fields with a well matched set of reference quadrilaterals (Lang et al., 2010).

All data from these images that pass all quality control checks is then stored in one of two MySQL databases, one for SkycamT images and one for SkycamZ images (Mawson et al., 2013, McWhirter et al., 2016). This data comprises

Source Extractor output as described previously, data from the FITS header (FITS is a common file type used for astronomy images) and catalogue information from the US Naval Observatory B catalogue based on coordinate matching the sources to known stars. At the end of each observing night, a check is performed to determine the quality of the observations recorded that night. The standard deviation of an objects magnitude values is then determined. A larger value of standard deviation indicates the data is of poor quality as these standard deviations are many times larger than those expected from even the most variable stars.

3. PolyFit Feature Representation

Previous research has resulted in a set of carefully engineered features for the description of light curves. The process of developing these features was powered by over a decade of work performed by experts in the analysis of light curves. The features they developed were designed for the current generation of survey data present at the time with many systems relying on OGLE (Udalski et al., 1997, Richards et al., 2011b), MACHO (Alcock et al., 2000, Kim et al., 2011), EROS2 (Rahal et al., 2009, Protopapas et al., 2015) and Kepler (LaCourse et al., 2015, Kugler et al., 2016, Matijevic, 2012, Parvizi et al., 2014, Neff et al., 2014) data. Some of the early classification features were based on a Fourier decomposition of the time-series data to generate a set of periodic features (Debosscher et al., 2007). These features were further extended using a set of non-periodic features previously identified as useful for variability detection. The full set of periodic and non-periodic features were then used to train a number of models using several different machine learning classifiers (Richards et al., 2011b). Additional non-linear features were introduced later and a subset were selected for variable star classification (Kim and Bailer-Jones, 2016). Unfortunately, despite the best of intentions, biases have been introduced into the classification process. This is due to the engineered features that are designed for and perform well at classification tasks for one survey does not guarantee good performance in other surveys with differing statistics (Benavente et al., 2017). The efficient production of a set of informative features is highly important (Huijse et al., 2014).

In regards to the classifiers, research has been conducted in mapping the models trained by one survey to the unclassified data of another such as considering different survey statistics to be a rotation in a higher dimensional feature space (Benavente et al., 2017). Other approaches aim to produce a set of highly capable classification models on a subset of object types with high performance and then combine them into a meta-classification model for improved multi-survey capability (Pichara et al., 2016). The features derived from the works of Richards et al. (Richards et al., 2011b) and Kim & Bailer Jones (Kim and Bailer-Jones, 2016) are useful in the classification of Skycam light curves however, many of the statistical features are unreliable. The physical reason for these features to be important for classification remains intact but the considerable noise in the Skycam data heavily poisons these features. As a result, research was conducted into the computation of a set of new features tuned for performance on the Skycam light curves. Representation Learning is a machine learning technique which extracts useful non-linear representation features of the raw data based on their performance at a given task, such as the classification of the variable star light curves (Bengio et al., 2013). Convolutional Neural Networks (CNNs) were used to attempt to extract features automatically from a two dimensional representation of the Skycam light curve data (McWhirter et al., 2017). Whilst the results of this method were shown to be inferior to the engineered features from previous studies, it did demonstrate that representation learning was possible on the Skycam data.

The goal of the representation learning is to produce features that model the shape of the folded light curve. Phase folding or epoch folding is a procedure that ‘folds’ multiple periodic waves together using equation 2.

$$\phi_i = \left| \frac{t_i - t_0}{P} \right|_1 \tag{2}$$

where ϕ_i is the phase of the i^{th} data point, t_i is the time instant (in time units) of the i^{th} data point, P is the period of the light curve in the same time units, t_0 is an arbitrary timestamp of phase 0 and $|x|_1$ is the modulus of x with 1, the decimal remainder of the function x.

This emphasises the shape of the variability over most sampling artifacts as long as the baseline (duration of time the source was observed) is much greater than the period of the variation and that the period is not close to a spurious sampling period. Equation 2 shows that the period is an important component of the resulting phased light curve. The period most be estimated from the data present in the light curve. We make use of GRAPE: Genetic Routine for Astronomcal Period Estimation designed for use on the Skycam light curves (McWhirter et al., 2018). This method uses a Bayesian Generalised Lomb Scargle (BGLS) periodogram (Mortier et al., 2015, Mortier and Cameron, 2017) optimised within a genetic algorithm to rapidly compute a candidate period and eliminate spurious results due to the unusual Skycam cadence. This method has the highest performance at correctly identifying candidate periods in the Skycam light curves (McWhirter et al., 2018).

A model can then be produced through the interpolation of a fitted model on the phase-folded data points. This model would remove much of the light curve noise and produce a fit which has the flexibility to correctly fit any possible phase-folded light curve shape whilst not overfitting on the noise. The chosen model for this interpolation on the Skycam light curves is the PolyFit model (Prsa et al., 2008). This model was developed for the fitting of eclipsing binary light curves and therefore is specifically designed to accurately reproduce the thin primary and secondary eclipses of detached binaries whilst still maintaining good performance on other light curve shapes such as pulsating variables (Prsa et al., 2008, Paegert et al., 2014, Parvizi et al., 2014).

The PolyFit algorithm is designed to outperform Fourier and Spline models when applied to any eclipsing binary light curve. The PolyFit algorithm is a method of fitting a polynomial chain $P(x)$ of smooth, piecewise n^{th} order polynomials which connect at a set of knots (Prsa et al., 2008). The algorithm has two main additions compared to normal piecewise polynomial methods to achieve the desired performance. First, unlike spline models, the polynomials are not required to be differentiable (although they remain continuous) at the knots allowing the modelling of sharp, narrow features such as eclipses. The second requirement is that the model cycles across the phase boundary between 0.5 and -0.5 when centred on zero. Our implementation of PolyFit utilises 4 knots with 4 2nd order piecewise polynomials fit using regularised polynomial regression through the implementation of the normal equation on the 4 subsets of data defined between each pair of consecutive knots. This produces a set of 16 parameters which fully describe the fitted PolyFit model, 4 knot locations with phases of $[-0.5, 0.5]$ and 12 polynomial parameters, a intercept, first order and second order for each of the four polynomials. This is substantially less free parameters than a Fourier model would require for similar eclipse fitting performance (Debosscher et al., 2007). In addition to these features, the PolyFit model is used to interpolate 99 magnitude values across the $[-0.5, 0.5]$ phase range for further analysis.

Figure 2 demonstrates the PolyFit algorithm applied to the Skycam light curve of the eclipsing binary RS Sagitarii. The black points indicate the light curve observations phased by a candidate period and phase binned into 100 bins, the red line indicates the fitted PolyFit model and the green crosses indicate the phase locations of the four knots. Figure 3 shows the capability of this method to accurately fit narrow eclipse features compared with spline and Fourier models. The top plot is the PolyFit model which fits the primary and secondary eclipses without substantially overfitting on the out-of-eclipse noise. The middle plot demonstrates a spline model with a *span* of 0.2 where the span defines the smoothness of the fitted spline polynomials. This shorter span results in a spline model which performs well on the deep eclipse but overfits the noise in the out-of-eclipse light curve. Increasing the span improves the out-of-eclipse performance at a cost of poorer eclipse modelling. Each eclipsing binary light curve will have an optimal value of span which compromises between eclipse and noise fitting yet it is unlikely that this optimal span value will perform as well as the PolyFit model. The bottom plot demonstrates a Fourier fit with eight harmonics and

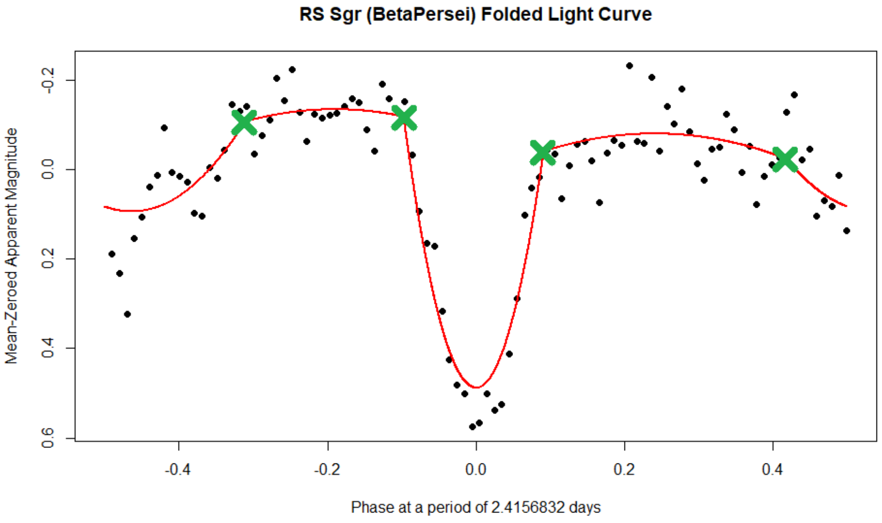


Figure 2: The PolyFit algorithm applied to the Skycam light curve of the Eclipsing Binary RS Sagitarii. The light curve has been phase-folded and phase binned into 100 bins. The red line indicates the fitted PolyFit model and the green crosses indicate the optimal knot points found by the optimisation algorithm. This method produces a superior fit to the narrow eclipse feature than the Fourier or spline models in figure 3.

an intercept with 17 parameters to the 16 PolyFit model parameters. As with the spline model with a low span argument, the eight harmonic Fourier model correctly models the deep primary eclipse but also overfits the out-of-eclipse noise. Reducing the number of harmonics in the Fourier model will result in a similar effect to increasing the span argument value. The PolyFit model is the only method of the three to fit the eclipses without overfitting on the noise.

The PolyFit algorithm is implemented by first selecting an initial state for the knots either by random or by a controlled method such as where the difference between the magnitudes of two data points crosses the mean magnitude. Using this initial set of knots x_k, $k = 1, \dots, 4$, the phase range of $[-0.5, 0.5]$ is partitioned into 4 intervals as shown in equation 3 (Prsa et al., 2008).

$$I_1 = [x_1, x_2), \quad I_2 = [x_2, x_3), \quad I_3 = [x_3, x_4), \quad I_4 = [x_4, x_1) \tag{3}$$

For the first phase interval I_1, use a regularised least-squares regression fit using the data points in this phase interval with 3 free parameters as shown in equation 4.

$$P_1(x) = a_0^{(1)} + a_1^{(1)}(x - x_1) + a_2^{(1)}(x - x_1)^2 \tag{4}$$

where $P_1(x)$ is the first polynomial as a function of phase x, x_1 is the phase of the first knot and $a_j^{(1)}$ are the fitted polynomial parameters where $j = 1, \dots, 3$. With the first three parameters computed, the next requirement is to compute $p_2(x)$ with respect to $p_1(x)$ and $p_3(x)$ with respect to $p_2(x)$ as shown in equation 5.

$$P_k(x) = a_0^{(k)} + a_1^{(k)}(x - x_k) + a_2^{(k)}(x - x_k)^2 \tag{5}$$

where $P_k(x)$ is the k^{th} polynomial of interval I_k. This must be computed whilst satisfying the constraint that the polynomial must connect with the previous polynomial at the knot x_k. This is shown in equation 6 and results in the computation of $p_2(x)$ and $p_3(x)$ being for 2 free parameters as the intercept $a_0^{(k)}$ where $k = [2, 3]$ has already been computed.

$$P_k(x_k) = p_{k-1}(x_k): \;\; a_0^{(k)} = a_0^{(k-1)} + a_1^{(k-1)}(x_k - x_{k-1}) + a_2^{(k-1)}(x_k - x_{k-1})^2 \tag{6}$$

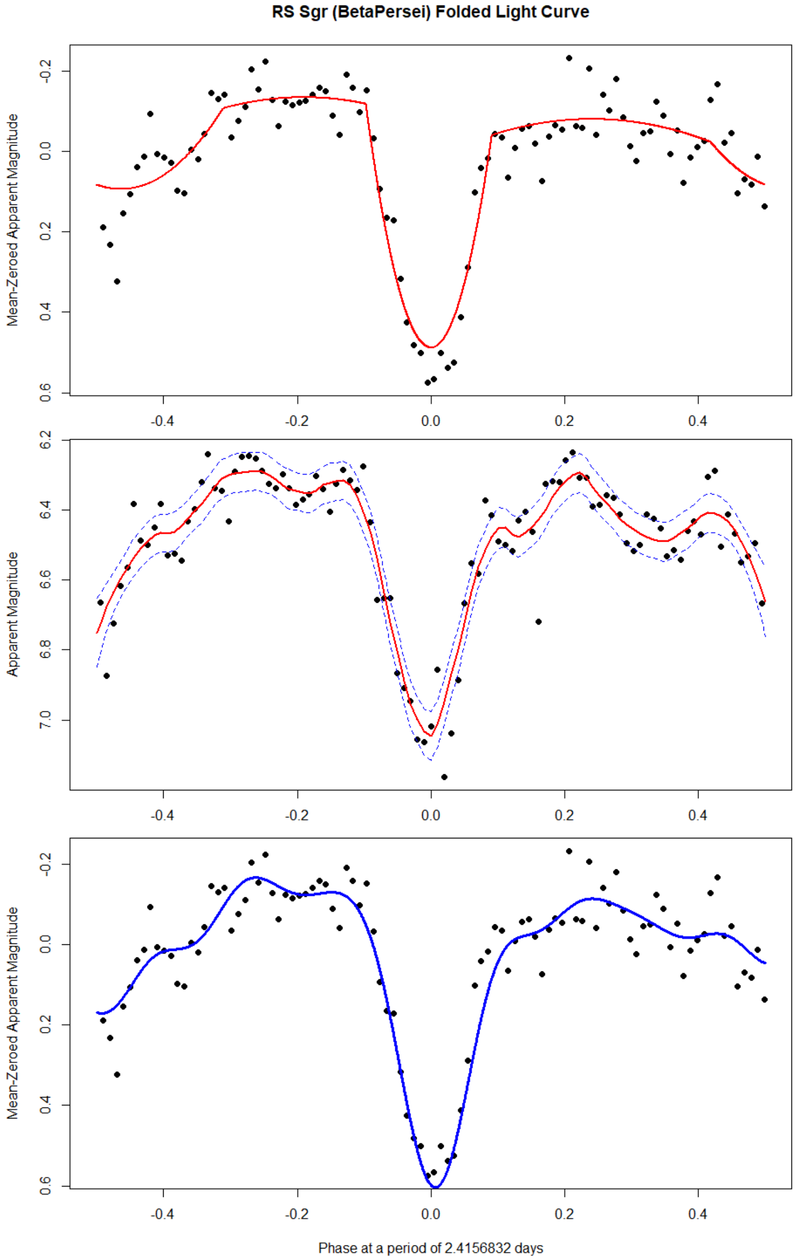


Figure 3: The PolyFit algorithm (top) fitted to a Skycam eclipsing binary light curve. The model provides a much more satisfactory fit than the spline model (middle) or the Fourier model (bottom) despite the Fourier model utilising more fitted parameters. The PolyFit model can accurately reproduce narrow eclipsing binary features whilst still providing good performance on pulsating and rotational light curves.

For the final phase interval I_4 there are two constraints to be satisfied. The polynomial must connect with the third interval I_3 at the knot location x_4 (connectivity) and the phase space wrapping from 0.5 to -0.5. As the connectivity has constrained the intercept of the 4^{th} polynomial, $a_0^{(4)}$, calculated as before from equation 6 the phase wrapping constraint constrains the first order parameter of the polynomial fit $a_1^{(4)}$ through constraint equation 7 revealing the remaining free parameter $a_2^{(4)}$.

$$P_4(x) = a_0^{(4)} + a_2^{(4)}(x - x_4)(x - x_1) \tag{7}$$

The original PolyFit implementation placed the four knots where the light curve data points crossed the mean magnitude of the light curve, randomly perturbed the knots using a random Gaussian 'kick' and then allowed them to relax into a minimum χ^2 state over a small number of iterations (Paegert et al., 2014). Each iteration must be carefully checked as the phase intervals must have an appropriate number of data points to prevent degeneracy in the polynomial fits. This means that the set of knots x_k must be rejected if the interval I_1 lacks 5 data points, I_2 and I_3 lack 4 data points each and I_4 lacks 3 data points. Finally, to prevent knots from adopting values which place two or more knots too close to each other, the fitting function must have an additional penalty term which disincentives this undesirable outcome. This is accomplished by using a quadratic repulsion term as shown in equation 8 which decreases the performance of a given fit by the square of the distance between each pair of knots with the size of this repulsion defined by an argument ϵ (Prsa et al., 2008).

$$r_{\text{cost}}(x_k; \epsilon) = \epsilon \left[(x_2 - x_1)^{-2} + (x_3 - x_2)^{-2} + (x_4 - x_3)^{-2} + (x_1 + 1 - x_4)^{-2} \right] \tag{8}$$

Due to the noise in the Skycam light curves this approach was not sufficient to produce good models as the nearest cost function minimum was highly dependent on where the noise located the initial state of the knots. To initially limit the noise from the Skycam light curves, the data points are phase binned into 100 mean averaged bins. This reduces the high frequency noise from affecting the PolyFit model and is particularly effective on Skycam due to the large number of observations in many light curves as well as a substantial reduction on computation time as the regression has less data points to compute. Despite this binning operation, the white noise in the light curves was still of sufficient amplitude to produce many local minima in the cost function minimisation procedure. As a result, the fitting procedure was insufficient for reliably determining the optimal PolyFit model for a given light curve.

3.1. Genetic Optimisation for PolyFit

The poor performance of the PolyFit algorithm's original fitting routine was a substantial problem in the use of this method on the Skycam light curves. Fortunately, the genetic algorithm optimisation developed for use in the GRAPE method provide a novel solution to the issues with PolyFit on noisy light curves (McWhirter et al., 2018). Genetic Algorithms are highly capable at the identification of the global optimum of a highly non-linear fitness function with many local optima (Charbonneau, 1995). The fitness function of the PolyFit algorithm when applied to the noisy Skycam light curves exhibits these properties.

The Genetic Algorithm method from GRAPE was modified to identify the optimal knot locations for the set of 4 knots x_k through the computation of the χ^2 fitness function augmented by the repulsion term in equation 8. This fitness function is showed in equation 9.

$$\chi^2(x_k;\epsilon) = \sum_{j=1}^{N} w_j \left(p(x_j) - y_j\right)^2 + r_{\text{cost}}(x_k;\epsilon) \qquad (9)$$

where $p(x_j)$ is the PolyFit interpolated magnitude of phase point x_j, y_j is the magnitude of the phase binned data point j at a phase of x_j, w_j are the weights of each phase bin which are kept at $w_j = 1$ and $r_{\text{cost}}(x_k;\epsilon)$ is the knot repulsion from equation 8 which is a function of a repulsion strength ϵ and the knot positions x_k and N is the number of binned data points in the light curve.

Where GRAPE utilises a one-dimensional feature space, the genetic PolyFit method requires the use of a four-dimensional feature space, the phase locations of the 4 knots. As the 12 polynomial parameters are generated through regularised regression as a function of the 4 knot positions x_k, they do not need to be determined by the genetic algorithm leaving just the 4 knots. The initial population of size N_{pop} is established by a uniform random number generator which creates N_{pop} sets of $x_k = [-0.5, 0.5]$ sorted from -0.5 to 0.5. This population is then encoded into chromosome strings by rescaling the phases to between 0 and 1 (which is simply performed by computing $\hat{x}_k = x_k + 0.5$) followed by the recording of the top 5 decimal places for each of the 4 knots into a concatenated string of 20 base-10 numerals. Similar to the GRAPE method, these chromosomes undergo a genetic update process where knots which minimise the χ^2 fitness function are bred into children and have crossover, mutation and fitness selection operations applied for a set number of generations N_{gen} where the knots have converged to the global optimum.

The arguments of the genetic algorithm are selected through a grid cross-validation procedure on a set of 859 light curves with the limitation that the PolyFit routine must complete in under two seconds. The input arguments were as follows: $N_{\mathrm{pop}} = 100$, $N_{\mathrm{pairups}} = 20$, $N_{\mathrm{gen}} = 100$, $P_{\mathrm{crossover}} = 0.65$, $P_{\mathrm{mutation}} = 0.03$, $P_{\mathrm{fdif}} = 0.6$ and $P_{\mathrm{dfrac}} = 0.7$. For further information on these genetic algorithm arguments we recommend reading the GRAPE paper (McWhirter et al., 2018). We found that this genetic optimisation routine produced more reliable optimal knot locations x_k regardless of the distribution of the initial knot candidates.

There remained one limitation relative to the original expected PolyFit performance on Eclipsing Binaries. The eclipse features are intended to be modelled by two knots at the beginning and end of the eclipse with a highly quadratic polynomial modelling the narrow eclipse dip. Unfortunately, the increased noise in the Skycam light curves resulted in the genetic PolyFit algorithm determining an optimal knot as the base of the eclipse and modelling the two sides of the dip in two separate phase intevals. This allows the PolyFit to use the second knot elsewhere in the phase space usually overfitting on the noise. This defect is shown in figure 4 where the top plot demonstrates the desired PolyFit model and the bottom plot demonstrates a model with a superior χ^2 performance but overfit on noise by placing the knot at the bottom of the peak. Our solution is

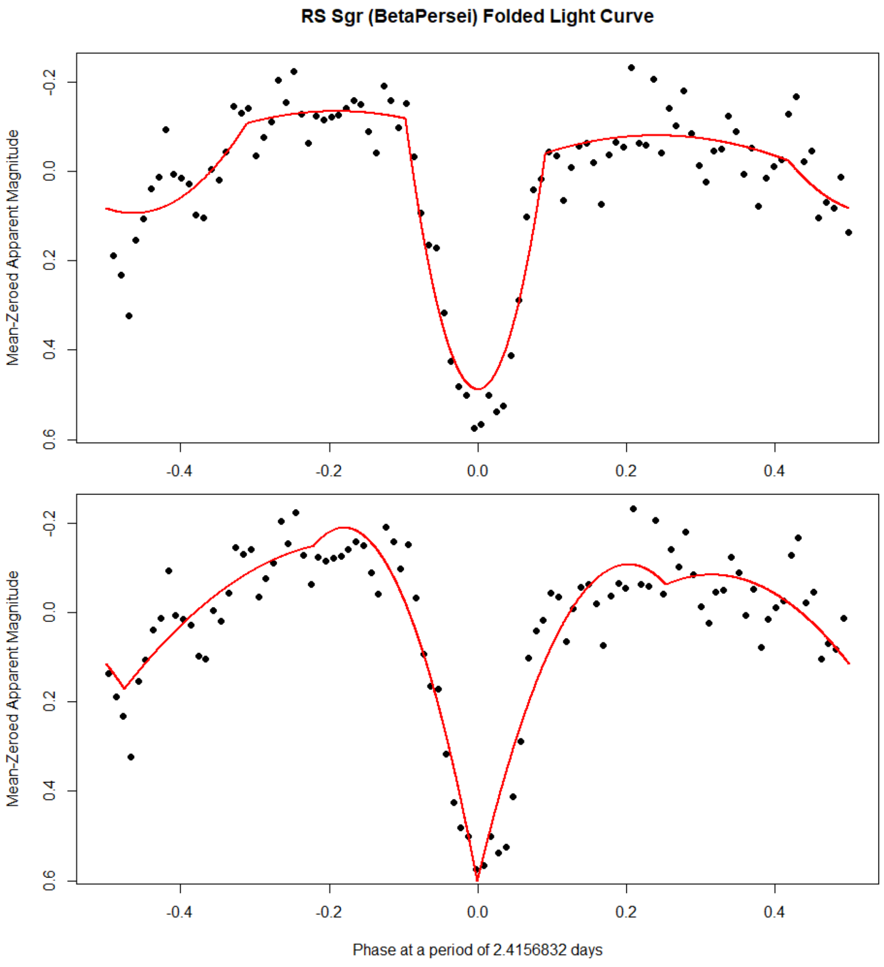


Figure 4: The same Skycam light curve with two different PolyFit optimisations. The top plot demonstrates the desired PolyFit model where the knots are located either side of the eclipse at the beginning and end of the dimming event. The bottom plot demonstrates a PolyFit model with a superior fit according to the fitness function in equation 9. This is not an ideal model as the knot has been located at the base of the eclipse, not the intended location and the second knot has been used to overfit on noise. Whilst the bottom fit minimises the fitness function, it is not the desired model and therefore the fitness function requires an additional corrective penalty term.

to employ a second additional term to the χ^2 fitness function shown in equation 9 which introduces a penalty to selecting knot phase locations x_k where the resulting polynomial fit interpolated magnitude at that phase is far from the median magnitude of the light curve. This penalty incentivises the genetic PolyFit algorithm to place the knots at locations with interpolated magnitudes near the out-of-eclipse magnitude. This penalty is shown in equation 10 and is weighted by an argument δ which defines the relative cost of placing knots far from the median.

$$m_{\text{cost}}(x_k;\delta) = \delta \sum_{j=1}^{4} \left(a_0^{(j)} - \text{median}(y) \right)^2 \tag{10}$$

where $m_{\text{cost}}(x_k;\delta)$ is the new cost term, δ is an argument that defines the strength of the penalty, $a_0^{(j)}$ is the polynomial intercept term for phase interval I_j and median(y) is the median magnitude of the phase binned light curve. Equation 11 defines the final fitness function of the genetic PolyFit algorithm with the two penalty terms.

$$\chi^2(x_k;\epsilon;\delta) = \sum_{j=1}^{N} w_j \left(p(x_j) - y_j\right)^2 + r_{\text{cost}}(\epsilon;x_k) + m_{\text{cost}}(\delta;x_k) \tag{11}$$

This method, whilst designed to correctly fit narrow features such as eclipses, does not adversely affect smooth continuous light curves such as pulsating variables. The $m_{\text{cost}}(x_k;\delta)$ penalty term does apply an increased cost to the knots which may be correctly located far from the median magnitude for these light curve shapes. In this case, the substantial number of data points far from the median magnitude in these phase locations allow the initial χ^2 component to 'outweigh' this penalty term allowing for the correct PolyFit model to be applied.

To verify the performance of our genetic algorithm approach to optimising the PolyFit algorithm, we implement an experiment to compare the knots fit by the genetic algorithm against those fit using a random perturbation approach across a set of 50 initial knot positions. This is applied to the 859 SkycamT light curves shown in table 1 selected for use by a correct period match or

Table 1: Data Summary for the 859 object, 12 class, SkycamT light curve dataset.

Number	Class	Type	Count
1	β Lyrae	Eclipsing Binary	57
2	β Persei	Eclipsing Binary	106
3	Chemically Peculiar	Rotational Variable	18
4	Classical Cepheid	Pulsating Variable	67
5	δ Scuti	Pulsating Variable	14
6	Mira	Pulsating Variable	369
7	RR Lyrae Fundamental Mode	Pulsating Variable	26
8	RR Lyrae Overtone Mode	Pulsating Variable	9
9	RV Tauri	Pulsating Variable	5
10	Semiregular Variable	Pulsating Variable	50
11	Spotted Variable	Rotational Variable	22
12	W Ursae Majoris	Eclipsing Binary	116

submultiple match with the American Association of Variable Star Observers (AAVSO) Variable Star Index (VSI) catalogue period using the GRAPE period estimation method. These light curves are phased around the AAVSO catalogue period in order to generate a set of pulsating, rotational and eclipsing light curve shapes for the PolyFit algorithm to model.

Figure 5 shows the results of this experiment with the two distinct distributions of standard deviation performance for the two optimisation method. For every knot the genetic algorithm produced more consistent knot locations regardless of the initial knot positions with phase standard deviations of 0.01 to 0.1 for most of the 859 light curves. The random perturbation method was substantially less stable with the standard deviation of the final knots of the 50 initial states varying by a standard deviation of 0.1 for most of the light curves with many performing more poorly than this out to 0.2 phase standard deviation. Both these methods require approximately 1.5 seconds of runtime to determine the optimal knots and produce a final PolyFit model. The genetic algorithm based optimisation has a larger range of standard deviations to the random perturbation optimisation. This is likely a result of the quality of the light curve having more effect on the resulting fit as the very noisy light curves can still vary substantially with the genetic generational updates due to the minimal difference between the fitness function values for many potential models and the optimal model. The variability in the genetically optimised PolyFit model is still not ideal; however, it is satisfactory as the possible optimised models are all acceptable for the extraction of classification features for a given light curve.

3.2. PolyFit Principal Component Analysis

Despite the efforts of implementing the δ cost and using the genetic algorithm to optimise the knot locations, there is still substantial variation on the 16 parameters of the PolyFit model. Therefore they cannot reliably be used as features as, at best, the relationship between these parameters for multiple classes is highly non-linear and difficult to learn. Therefore, a different method of representing the important features of the interpolated PolyFit model shape is required which does not rely on the parameter values but simply the magnitudes of the fitted model. Principal Component Analysis has been previously used to learn a set of features from interpolated Fourier models applied to light curves (Deb and Singh, 2009, Tanvir et al., 2005, Yoachim et al., 2009). The extraction of features from fitted models has also been accomplished using other methods such as echo-state-networks, a form of recurrent neural network (Kugler et al., 2016) and local linear embedding (Matijevic, 2012) with positive results. The method used is not dependent on the model used to interpolate the light curves and simply on the distribution of interpolated magnitudes. It is therefore suitable to investigate the performance of such an approach on sets of interpolated magnitudes extracted from SkycamT light curves by the PolyFit algorithm. This method can potentially generate learned features useful for Machine Learning classification in the automated pipeline.

Principal Component Analysis (PCA) is a mathematical method that trans-

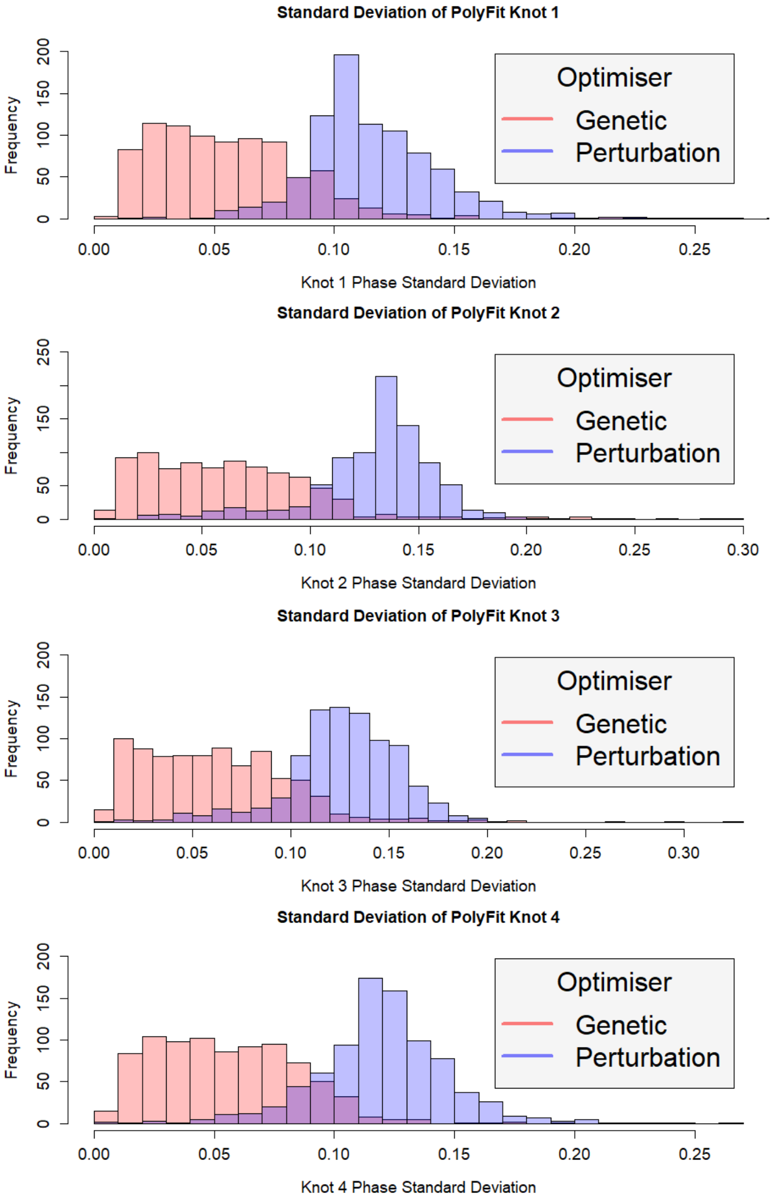


Figure 5: Histograms of the measured standard deviation of the 4 PolyFit knots on the set of 859 SkycamT light curves from table 1. The knots produced by each light curve over 50 initial states are recorded and the standard deviation measured from this set. Each histogram contains two distributions. The red distribution was produced by the knots optimised by the genetic algorithm method and the blue distribution was produced by the knots optimised by the random perturbation method from the original PolyFit algorithm (Prsa et al., 2008).

forms a number of variables or features which may be correlated into a set of uncorrelated variables called principal components (Pearson, 1901, Hotelling, 1933, 1936). The principal components account for the variability of the dataset in order with the first principal component describing a large amount of the initial variance and the second principal component describing much of the remaining variance until the last principal component contains the remaining variance.

Principal components can be calculated from a design matrix X, with columns containing the variables of the dataset and the rows containing the observations. Initially the variables in the design matrix must be scaled so they have comparable values using equation 12.

$$\bar{x_j} = \frac{x_j - \mu_j}{\sigma_j} \quad \text{for } j = 1, 2, \ldots, N \tag{12}$$

where $\bar{x_j}$ is the rescaled j^{th} variable, μ_j is the mean of the j^{th} variable, σ_j is the standard deviation of the j^{th} variable and N is the number of variables in the design matrix X. The covariance matrix of the rescaled design matrix X is then computed using equation 13.

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \left(X_{[i]}\right) \left(X_{[i]}^{\top}\right) \tag{13}$$

where $X_{[i]}$ is the row vector of the design matrix X for the i^{th} observation and m is the total number of observations (rows) in the design matrix. Using the covariance matrix, the eigenvalues and eigenvectors are computed. The sorted eigenvalues from high to low give the principal components in the solution in order of the variance contained in each principal component. The eigenvector associated with the eigenvalues can be used to determine the principal components using equation 14.

$$\text{PCA}_j = \Theta_j^{\top} X \tag{14}$$

where PCA_j is the j^{th} principal component, Θ_j is the eigenvectors associated with the j^{th} sorted eigenvalue and X is the design matrix. This produces uncorrelated principal components and can also be used for dimensionality reduction through the selection of an integer value k where $1 \leq k \leq N$. The choice of

k can be decided through the computation of how much variance is described by the reduced design matrix $\hat{X}$ where $\hat{X} = \Theta^{\top}_{1,\ldots,k}X$. Equation 15 shows the inequality which must be satisfied with the minimum value of k to retain $1 - \epsilon$ of the variance of the design matrix X.

$$\epsilon \geq \frac{\frac{1}{m}\sum_{i=1}^{m} ||x_{[i]} - \hat{x_{[i]}}||^2}{\frac{1}{m}\sum_{i=1}^{m} ||x_{[i]}||^2} \tag{15}$$

where $(1-\epsilon)$ is the required retained variance, $x_{[i]}$ is the row vector of the design matrix X for the i^{th} observation, $\hat{x_{[i]}}$ is the row vector of the reduced design matrix $\hat{X}$ for the i^{th} observation reconstructed using k principal components and m is the total number of observations (rows) in the design matrix.

The PCA method is used as a dimensionality reduction technique to define the interpolated magnitudes of the PolyFit model as a set of features determined by a training set of light curves determined by cross-matching the SkycamT database on the set of AAVSO catalogue objects with period information and with types present in the BigMacc All-Sky Automated Survey (ASAS) light curve classification pipeline (Richards et al., 2012). This produces 6897 light curves across 18 variable star classes shown in table 2. The set of light curves is much bigger in this set as the trained PCA must be capable of modelling any possible PolyFit interpolation on the Skycam database. Many of these light curves are of dubious quality and GRAPE does not agree with the catalogue period for many of the objects as they are not in the 859 light curve dataset. Despite this, the learned PCA components using the larger light curve dataset generalises better than the smaller dataset.

Upon the computation of the PolyFit model for any given light curve, the model is used to interpolate 99 magnitude data points on an evenly distributed grid of phase values from -0.49 to 0.49 with intervals of 0.01. The interpolation does not go to -0.5 or 0.5 due to limitations in the spline fitting code as the edge values could be out of bounds. Of course this is not a limitation of the PolyFit algorithm as the entire phase space is mapped by the polynomial chain $P(x)$ yet in the event of the PolyFit algorithm failing to converge on a light curve the spline algorithm may still produce an acceptable (but likely inferior) fit to collect

Table 2: The class distribution of the STILT 6897 variable light curves used for the PCA training.

Number	Class	Type	Count
1	β Lyrae	Eclipsing Binary	412
2	β Persei	Eclipsing Binary	1518
3	Chemically Peculiar	Rotational Variable	477
4	Classical Cepheid	Pulsating Variable	195
5	δ Scuti	Pulsating Variable	453
6	Ellipsoidal	Non-eclipsing Binary	131
7	Mira	Pulsating Variable	1256
8	Pop II Cepheid	Pulsating Variable	29
9	R Coronae Borealis	Eruptive Variable	3
10	RR Lyrae Dual Mode	Pulsating Variable	3
11	RR Lyrae Fundamental Mode	Pulsating Variable	99
12	RR Lyrae Overtone Mode	Pulsating Variable	60
13	RS Canum Venaticorum	Non-eclipsing Binary	528
14	RV Tauri	Pulsating Variable	36
15	Small Amplitude Red Giant	Pulsating Variable	4
16	S Doradus (Luminous Blue Variable)	Eruptive Variable	1
17	Semiregular Variable	Pulsating Variable	989
18	W Ursae Majoris	Eclipsing Binary	703

some useful information. The interpolated magnitudes are zeroed by having the mean magnitude of the phase-binned data points subtracted from their values. Using the phase-binned mean magnitude instead of the interpolated mean magnitude preserves the interpolated skewness of the light curve. The PolyFit interpolated magnitudes are then used to recalculate the phase zero-point through the identification of the maximum magnitude y_i data point (which corresponds to the minimum brightness). The phase space is then adjusted so that the minimum brightness of the light curve occurs at phase 0.0. For eclipsing binaries this is the primary eclipse and for other variables it simply corresponds with the minimum brightness of the variability. The interpolated magnitudes are then normalised by equation 16 which means that the amplitude of the light curve is not a component of the learned PCA components. The mean zero and scaling operation corresponds to the operation shown by equation 12.

$$\hat{y}_i = \frac{y}{|\max(y) - \min(y)|} \tag{16}$$

The genetic PolyFit algorithm was then applied to the 6897 light curves in table 2 producing a training matrix of 99 columns (the 99 interpolated magnitudes across the phase space) and 6889 rows (6889 light curves generated a PolyFit model, for 8 light curves the PolyFit did not converge and were discarded). Using equations 13 and 14, the PCA operation is computed producing 99 Principal Components sorted by their variance where the first Principal Component contains most of the information of the system. Figure 6 shows the top 10 principal components of the resulting PCA model. The top ten principal components describe 96% of the variance in the PolyFit modelled light curves whereas the remaining 89 principal components only add an additional 4% variance which is likely noise dominated.

The top ten Principal Components (PCs) can be used to closely reconstruct the original PolyFit model as each learned component contains information about a specific transformation of the light curve weighted by the specific value for a given light curve. Figure 7 demonstrates this reconstruction on the eclipsing binary RS Sagitarii. The black line shows the original PolyFit model as

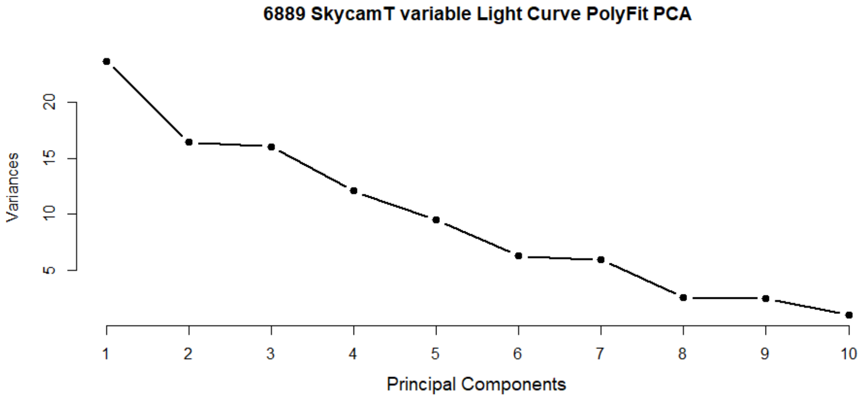


Figure 6: Plot showing the variance described by each principal component determined from the 6889 SkycamT light curves. The first principal component describes 23.7% of the variance of the light curves and the second principal component an additional 16.4%. The first ten principal components describe 96% of the total variance of the light curves. This allows these ten values to describe the shape of the light curves as the remaining 4% is likely noise related.

seen previously in figure 2. Each coloured line represents the reconstruction of the original model by adding on an additional principal component. From this reconstruction it is clear that the first principal component PC1 models the general shape of the light curve determining the range between the minimum and the maximum magnitudes of the light curve but does not model the secondary eclipse. PC2 and PC3 are responsible for modelling the asymmetries between the $[-0.5, 0]$ and $[0, 0.5]$ phase intervals as well as the presence of a secondary eclipse. This is important in the modelling of eccentric eclipsing binaries as well as containing the weighting that distinguishes eclipsing binary light curves from pulsating light curves. PC4 to PC9 are used to reconstruct variations in the smooth continuum of the light curve such as the 'bumps' common to some Cepheid and RR Lyrae type pulsating variables although there is likely a large noise component to these principal components. The final principal component, PC10 appears specifically for modelling the very narrow primary eclipses seen in the longer period β Persei variables. As there are few examples of this type of variable in the set of Skycam light curves, this explains why this important

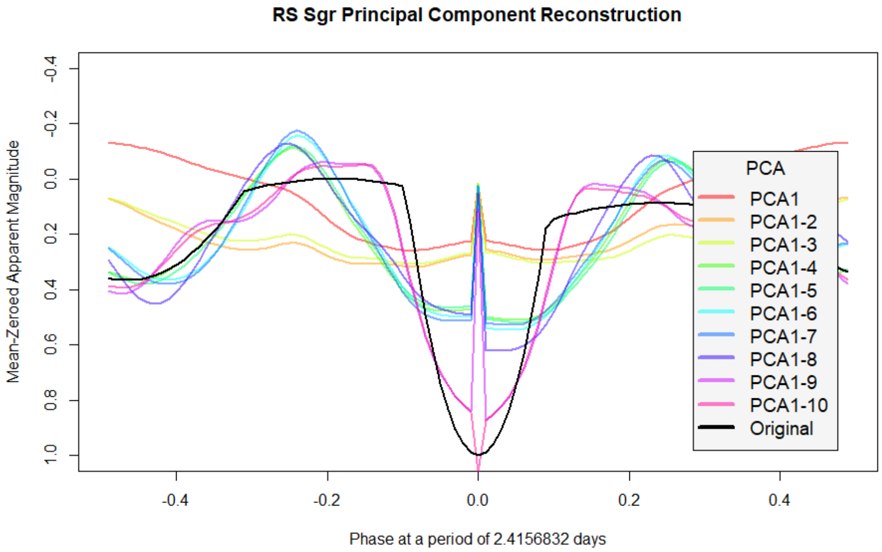


Figure 7: Plot showing the reconstruction of the SkycamT light curve of the eclipsing binary RS Sagitarii from the top ten principal components of the learned PCA model.

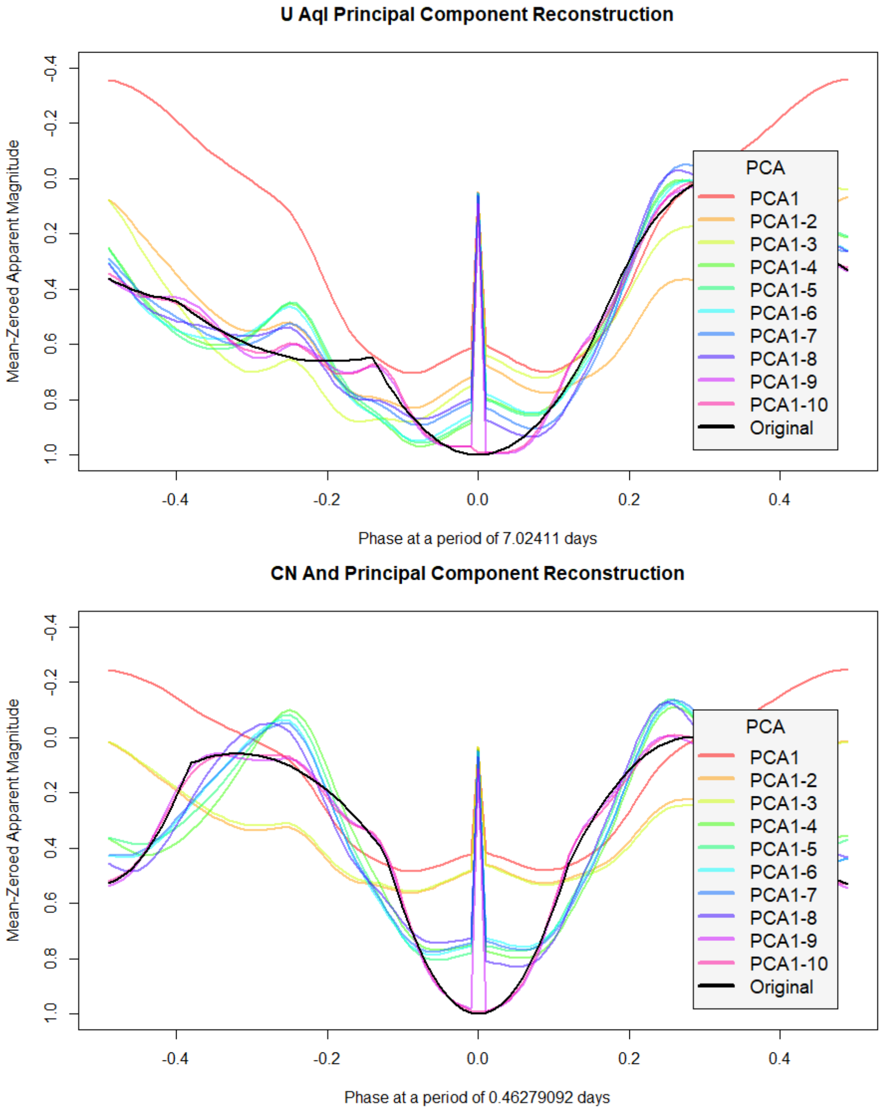
principal component has a lower variance as the variance describes how informative it is to the given training dataset. The most distinguishing feature of the principal components is the narrow spike feature located near phase zero. This is the component of the learned representation that models the width of the eclipses and is weaker for objects lacking a narrow eclipse. This is why PC1 to PC9 contain this spike; it is there to cancel out the narrow eclipse in the majority of light curves which lack this feature. PC10 is used to apply this feature when it is required for a given light curve.

The features utilised from this PCA are the ten weights which, in conjunction with the PCA model, allow the approximate reconstruction of the PolyFit model. The different shapes of light curve are expected to produce different sets of the ten weights and therefore can be used in the classification task. Figure 8 shows the principal component reconstruction for U Aquilae a Classical Cepheid and CN Andromedae, a β Lyrae eclipsing binary. Figure 9 shows this reconstruction for S Pegasi, a Mira-type Long Period Variable and RS Bootis, a fundamental mode RR Lyrae variable. Each variable light curve contains shapes distinctive to each of their classes and therefore they have unique PCA weights which can be used to determine the class of an unknown light curve.

Comparing the PCA features for the set of 859 light curves across 12 classes allow for the identification of features which may be of potential interest to a classification method. The best feature for discriminating the classes is the PC2 feature which is not surprising as PC2 and PC3 are the strongest indicators of the differences between the eclipsing binary 'double dip' light curve to the pulsating variable 'single dip, single peak' light curves. Figure 10 demonstrates the plot of the base-10 logarithm of the Period (the best feature for distinguishing the classes) against the second principal component feature. Whilst there is substantial overlap between the classes, the eclipsing binary classes tend to adopt lower values of P.PCA2 around 1 to 2 whereas the pulsating variables have larger values of P.PCA2 closer to 3 and 4. The feature appears very useful for separating the short-period eclipsing binaries from the RR Lyrae pulsating variables which means it could be of important use in the machine learning classifiers.

3.3. Interpolated Statistical Features

The features extracted directly from the PolyFit interpolation are not the only set of useful information extractable from the PolyFit method. Many of the original statistical features, derived from previous studies (Richards et al., 2011b, Kim and Bailer-Jones, 2016), such as their standard deviation, kurtosis and amplitude have similar inter-class distributions when computed on the Skycam light curves. This is likely a result of the substantial noise component in the Skycam light curves propagating into the features. As a result, the larger noise causes a larger overlap between the features of different variable star classes leading to poor discrimination. The interpolated PolyFit model provides an alternative method to define these statistics by computing them directly from the interpolated data. A number of features are produced to potentially replace the original variability indices and Fourier components as the binned genetic Poly-


U Aql Principal Component Reconstruction

CN And Principal Component Reconstruction

Figure 8: Plot showing the reconstruction of a selection of SkycamT light curves using the learned PCA model. The top light curve is of U Aquilae, a Classical Cepheid pulsating variable. The bottom light curve is of CN Andromedae, a β Lyrae eclipsing binary.

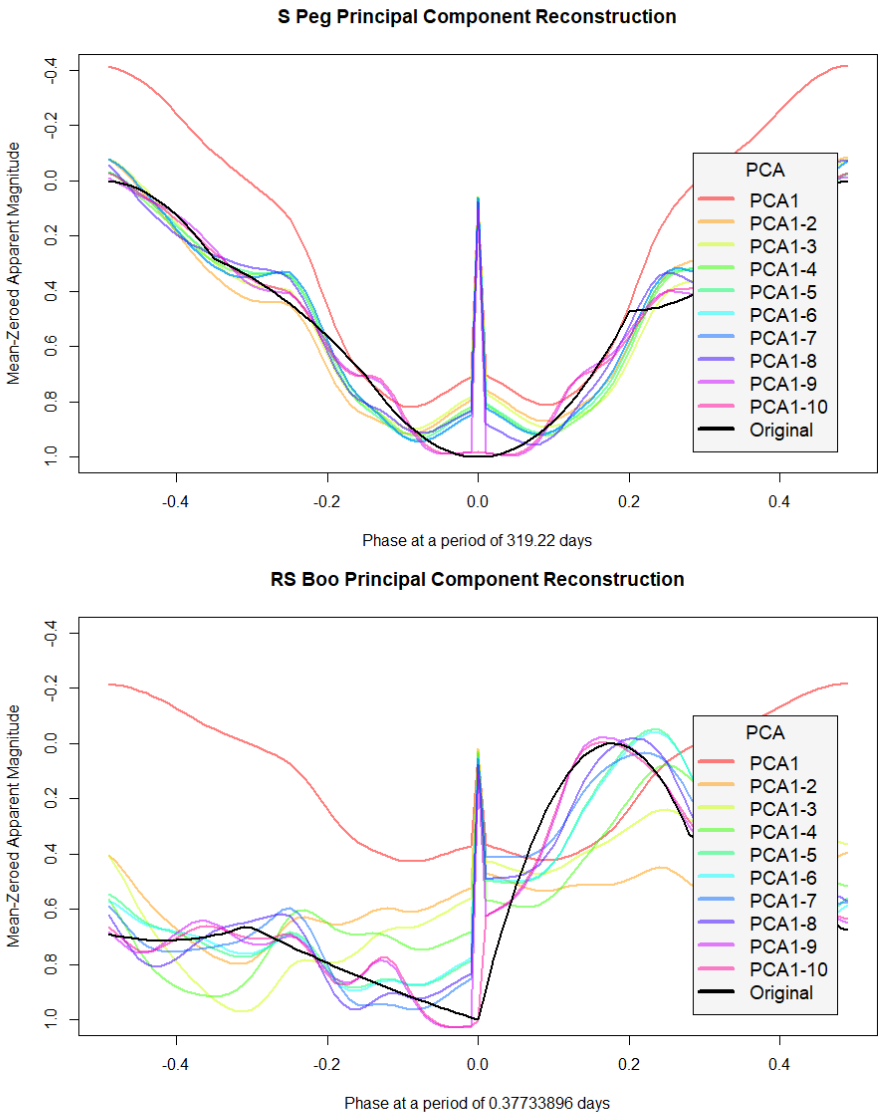


Figure 9: Plot showing the reconstruction of a selection of SkycamT light curves using the learned PCA model. The top light curve is of S Pegasi, a Mira-type Long Period Variable. The bottom light curve is of RS Bootis, a fundamental mode pulsating RR Lyrae variable.

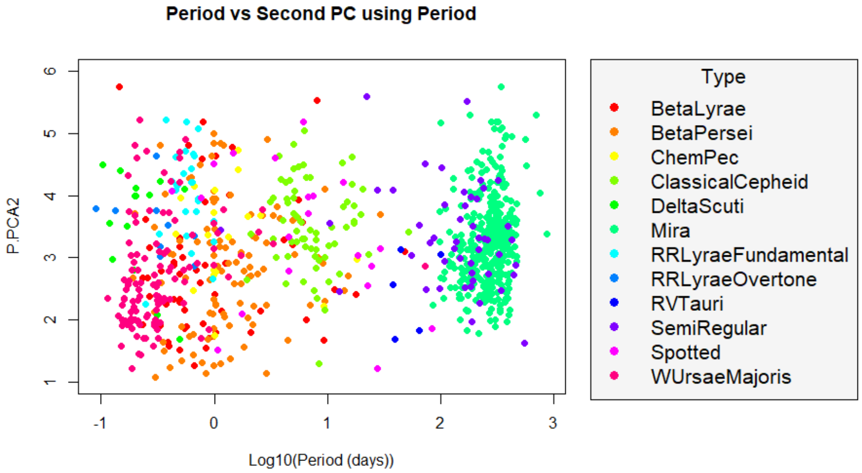


Figure 10: Plot showing the base-10 logarithm of the GRAPE-determined Period feature against the P.PCA2 feature, the second principal component calculated by the PolyFit model and the PCA model at the estimated GRAPE period. The feature partially seperates the pulsating variables and eclipsing binaries with a specific stength at classifying RR Lyrae fundamental mode and overtone mode pulsators from the short-period eclipsing binaries of similar period range.

Fit algorithm has reduced the high frequency noise and fit models which are capable of good quality modelling of any light curve shape. The 99 interpolated magnitude data points at the 99 evenly split phase positions $[x_i, y_i]$ replace the $[\phi_i, m_i]$ data points from the phase-binned light curves in the computation of the following features:

- PolyFit.Phase.Binned.Ratio

  The PolyFit phase binned ratio is a measure of how well sampled a light curve is. It is the ratio of phase bins containing at least one phased data point from the light curve n to the total number of phase bins N_{bins}. It is calculated by equation 17.

  $$\text{PBR}_{\text{PF}} = \frac{n}{N_{\text{bins}}} \tag{17}$$

  This feature is not directly related to the classification of variable star light curves but it is a measure of how well sampled the light curve is when epoch-folded around the estimated period. Poorly sampled light curves have less reliable fitted models.

- PolyFit.Goodness.of.Fit

  The PolyFit Goodness of Fit feature is a measure of how well the PolyFit model matches the phase-binned light curve data points and is defined as the χ^2 value of the fitted model prior to the addition of the penalty terms r_{cost} and m_{cost} as shown in equation 11. This feature is expected to be of moderate usefulness as it indicates light curves with multi-periodic signals as they have significant variance remaining after fitting the dominant period. The feature is limited by the presence of noise which also increases the χ^2 statistic along with poorly selected candidate periods although this would also disrupt the rest of the PolyFit features.

- PolyFit.Interpolated.Amplitude

  The PolyFit Interpolated Amplitude is calculated by computing equation

18 on the interpolated PolyFit magnitude data.

$$a_{\mathrm{PF}} = \frac{|\max(y) - \min(y)|}{2} \tag{18}$$

where a_{PF} is the interpolated amplitude and y is the vector of interpolated magnitudes. This feature is expected to be very important as many variable star types are classified by the amplitude of their light curves and the interpolated amplitude is expected to be less distorted by noise than the other amplitude features.

- PolyFit.Interpolated.StD

The PolyFit Interpolated Standard Deviation is calculated by computing equation 19 on the interpolated PolyFit magnitude data.

$$\sigma_{\mathrm{PF}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu_{\mathrm{PF}})^2} \tag{19}$$

where μ_{PF} is the mean of the y_i magnitudes computed by equation 20.

$$\mu_{\mathrm{PF}} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{20}$$

- PolyFit.Interpolated.Skewness

The PolyFit Interpolated Skewness is calculated by computing equation 21 on the interpolated PolyFit magnitude data.

$$b_{\mathrm{PF}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_{\mathrm{PF}})^3}{\sigma_{\mathrm{PF}}^3} \tag{21}$$

- PolyFit.Interpolated.Small.Kurtosis

The PolyFit Interpolated Small Kurtosis is calculated by computing equation 22 on the interpolated PolyFit magnitude data.

$$k_{\mathrm{PF}} = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{y_i - \mu_{\mathrm{PF}}}{\sigma_{\mathrm{PF}}}\right)^4\right) - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{22}$$

- PolyFit.Interpolated.Beyond.1.StD

  The PolyFit Interpolated Beyond 1 Standard Deviation feature calculates the ratio of interpolated data points $[x_i, y_i]$ which have magnitude values outside of plus or minus the standard deviation of the mean of the interpolated magnitudes. This feature is calculated by equation 23.

$$(\text{beyond}1\sigma)_{\text{PF}} = \frac{n_{>\sigma_{\text{PF}}}}{n} \tag{23}$$

  where $n_{>\sigma_{\text{PF}}} = \sum_{i=1}^{n}$ if $|y_i - \mu_{\text{PF}}| > \sigma_{\text{PF}} = 1, \quad \text{otherwise } 0.$

- PolyFit.Interpolated.Range.Cumulative.Sum

  The PolyFit Interpolated Range of a Cumulative Sum is calculated by first computing the vector of cumulative sums S_{PF} using equation 24 on the interpolated PolyFit magnitude data.

$$S_{\text{PF}} = \frac{1}{n\sigma_{\text{PF}}} \sum_{i=1}^{l} (y_i - \mu_{\text{PF}}) \quad \text{for } l = 1, 2, \ldots, n \tag{24}$$

  The range of the cumulative sum $R(S_{\text{PF}})$ is then determined using equation 25.

$$R(S_{\text{PF}}) = \max(S_{\text{PF}}) - \min(S_{\text{PF}}) \tag{25}$$

4. Experimental Results

Using the 859 SkycamT light curves selected by GRAPE shown in table 1, we compute a PolyFit model for each light curve phased at 2× the GRAPE estimated period and calculate these interpolated features. These distributions of features are shown in a histogram relative to the equivalent feature from the non-interpolated data. Figure 11 (top) demonstrates the Interpolated Amplitude relative to the variability index Amplitude. The Interpolated Amplitude distribution is closer to zero than the Amplitudes for the 859 light curves. This is due to the interpolated model reducing the high frequency noise allowing the fit to more accurately reflect the true amplitude of the variability.

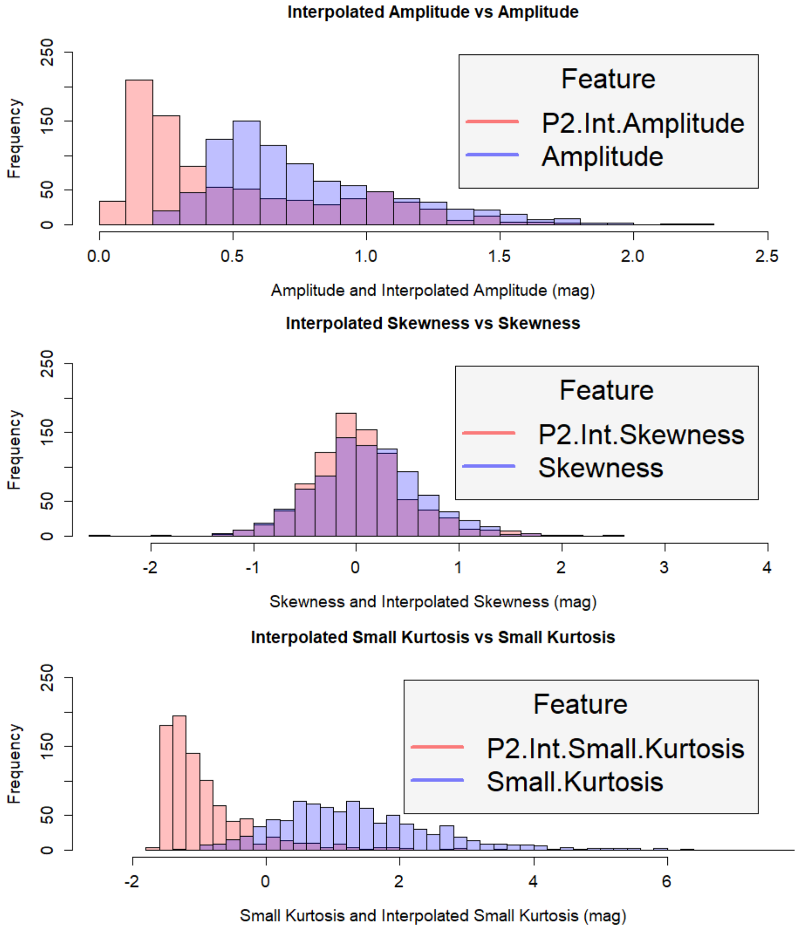


Figure 11: Histogram showing the distribution of the Interpolated features compared to the variability index features for the 859 SkycamT light curves. Many of the features exhibit superior performance due to their resilience to the noise in the data. This explains the interpolated amplitude being closer to zero and the narrower distribution of the interpolated skewness and interpolated small kurtosis. This figure contains Interpolated Amplitude (top), Interpolated Skewness (middle), and the Interpolated Small Kurtosis (bottom).

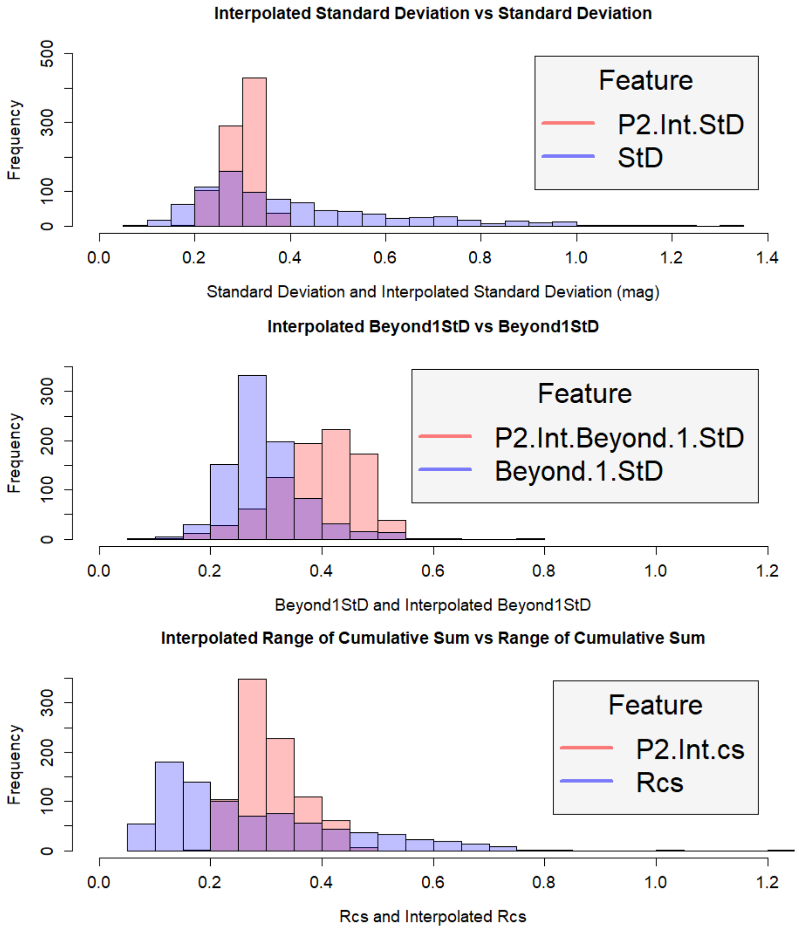


Figure 12: Histogram showing the distribution of the Interpolated features compared to the variability index features for the 859 SkycamT light curves. This figure contains Interpolated Standard Deviation (top), Interpolated Beyond 1 σ (middle), and the Interpolated Range of a Cumulative Sum (bottom).

Figure 11 (middle) shows the Interpolated Skewness feature relative to the variability index Skewness. The distribution of these two features is very similar which suggests that the skewness feature is not distorted by the noise. This is surprising as the skewness feature from the raw light curves did have a wide inter-class distribution. The conclusion that this is due to noise may not be accurate and it is possibly due to a sample bias due to lack of narrow eclipses in our set of β Persei eclipsing binaries. This bias is not due to noise and is a result of poor sampling of this class of object by the Skycam cadence. This is a problem which has been discussed for other surveys such as Kepler and the Large Synoptic Survey Telescope (LSST) (Prsa et al., 2011, Wells et al., 2017, Parvizi et al., 2014, LaCourse et al., 2015).

Figure 11 (bottom) demonstrates the distribution of the Interpolated Small Kurtosis feature relative to the variability index Small Kurtosis. This feature is interesting as the Interpolated Small Kurtosis feature is primarily negative for this set of light curves whereas the small kurtosis feature is positive. The Interpolated Small Kurtosis feature has a narrower distribution which is an effect of the reduced noise component in the interpolated feature which likely causes the difference in the distribution centres. Figure 12 (top) shows the interpolated standard deviation feature relative to the variability index standard deviation. The interpolated standard deviation has a much narrower distribution due to a lower noise component. Figure 12 (middle) demonstrates the interpolated Beyond 1 Standard Deviation feature. This feature adopts higher values as the approximately 0.2-0.25 mag white noise component suppresses the non-interpolated feature in the low amplitude variable classes. Figure 12 (bottom) shows the interpolated Range of a Cumulative Sum feature which also exhibits a narrower distribution relative to the variability index Range of a Cumulative Sum feature due to the noise reduction from the interpolated PolyFit model.

The performance of these features on the 859 SkycamT light curves appear superior to the associated variability indices. This demonstrates the strength of this approach although as the period is an important component of the generation of these models, failure of the period estimation method will disrupt these features substantially more than the variability indices. As the period is the dominant feature in the variable star classification task, the poorer interpolated features are unlikely to be the dominant source of error in the classification of poorly detected variability.

Using the representation learning features, a dataset was generated from the 859 SkycamT light curves from the GRAPE period match operation. A set of 38 features are produced and they are displayed in tables 3 and 4. These features include the GRAPE estimated period, and the PolyFit features, both PCA and interpolated variability indices generated by a PolyFit model phased at the GRAPE estimated period. This process is repeated to produce a second set of PolyFit features at two times the GRAPE estimated period, the 'double period'. The final feature is the ratio of the variances for the period and the double period.

A Random Forest classifier was selected to determine the individual importance and overall performance of these features in the classification task of assigning the 859 SkycamT light curves to the correct variability class out of 12 possible classes. We perform a hyper-parameter optimisation on the three Random Forest arguments. We found that the number of trees in the Random Forest model did not heavily influence the performance of the classification task therefore we kept this value at ntree = 500. The $\mathrm{m_{try}}$ and nodesize parameters are determined using a grid-search from 8 to 18 with intervals of 2 for the $\mathrm{m_{try}}$ parameter and 10 to 30 with intervals of 5 for the nodesize parameter. Figure 13 demonstrates the surface plot generated from the F1 Score of a 5-fold cross-validation with 2 repeats, our figure of merit (FoM) in this experiment as a function of the Random Forest arguments $\mathrm{m_{try}}$ and nodesize. This hyper-parameter optimisation procedure selects the optimal values as $\mathrm{m_{try}} = 14$ and nodesize = 30 for 500 trees in the Random Forest with a 12-class mean F1 score of 0.4729 with a standard deviation of 0.0931. The results of this PolyFit features model were compared to a model trained using a set of the original engineered features. This model uses the same methodology as the PolyFit

Table 3: The first half of the 38 features used in the classification of the 859 SkycamT light curves using the PolyFit algorithm.

Feature	Description
Period	Period (P) estimated by the GRAPE method
P.Binned.Ratio	PolyFit phase binned ratio at P
P.Goodness.of.Fit	χ^2 of PolyFit model at P
P.Int.Std	PolyFit interpolated Standard Deviation at P
P.Int.Skewness	PolyFit interpolated Skewness at P
P.Int.Small.Kurtosis	PolyFit interpolated small Kurtosis at P
P.Int.Amplitude	PolyFit interpolated Amplitude at P
P.Int.Beyond.1.StD	PolyFit interpolated Beyond 1 StD at P
P.Int.cs	PolyFit interpolated Range of a Cumulative Sum at P
P.PCA1	PolyFit interpolated PC1 at P
P.PCA2	PolyFit interpolated PC2 at P
P.PCA3	PolyFit interpolated PC3 at P
P.PCA4	PolyFit interpolated PC4 at P
P.PCA5	PolyFit interpolated PC5 at P
P.PCA6	PolyFit interpolated PC6 at P
P.PCA7	PolyFit interpolated PC7 at P
P.PCA8	PolyFit interpolated PC8 at P
P.PCA9	PolyFit interpolated PC9 at P
P.PCA10	PolyFit interpolated PC10 at P

Table 4: The second half of the 38 features used in the classification of the 859 SkycamT light curves using the PolyFit algorithm.

Feature	Description
P2.Binned.Ratio	PolyFit phase binned ratio at 2P
P2.Goodness.of.Fit	χ^2 of PolyFit model at 2P
P2.Int.Std	PolyFit interpolated Standard Deviation at 2P
P2.Int.Skewness	PolyFit interpolated Skewness at 2P
P2.Int.Small.Kurtosis	PolyFit interpolated small Kurtosis at 2P
P2.Int.Amplitude	PolyFit interpolated Amplitude at 2P
P2.Int.Beyond.1.StD	PolyFit interpolated Beyond 1 StD at 2P
P2.Int.cs	PolyFit interpolated Range of a Cumulative Sum at 2P
P2.PCA1	PolyFit interpolated PC1 at 2P
P2.PCA2	PolyFit interpolated PC2 at 2P
P2.PCA3	PolyFit interpolated PC3 at 2P
P2.PCA4	PolyFit interpolated PC4 at 2P
P2.PCA5	PolyFit interpolated PC5 at 2P
P2.PCA6	PolyFit interpolated PC6 at 2P
P2.PCA7	PolyFit interpolated PC7 at 2P
P2.PCA8	PolyFit interpolated PC8 at 2P
P2.PCA9	PolyFit interpolated PC9 at 2P
P2.PCA10	PolyFit interpolated PC10 at 2P
Period.Double.Ratio	Ratio of the GRAPE statistic at P and 2P

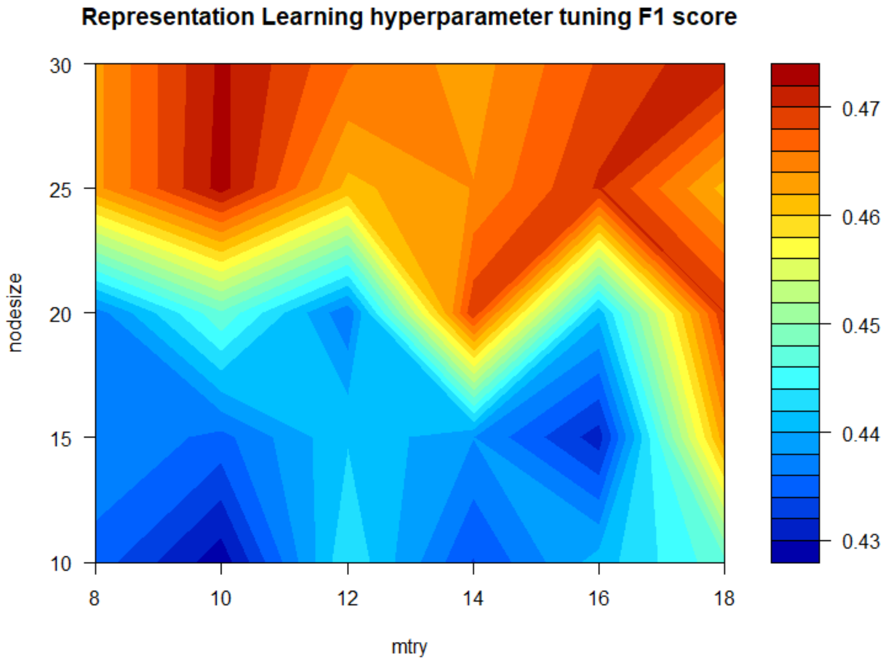


Figure 13: Contour plot of the F1 score performance of the 5-fold cross-validation using a Random Forest classifier with 500 trees on the 859 SkycamT light curves as a function of the m_{try} and nodesize hyperparameters. The optimal hyperparameters are $m_{try} = 14$ and nodesize = 30.

model training and makes use of 25 engineered features selected from previous studies to be similar to the PolyFit features (Richards et al., 2011b, Kim and Bailer-Jones, 2016). The hyperparameter optimisation procedure for the feature engineering model selects the optimal values as $m_{try} = 18$ and nodesize = 10 for 500 trees in the Random Forest with a 12-class mean F1 score of 0.3902 with a standard deviation of 0.0619. The PolyFit model appears to slightly outperform the original features on the set of Skycam variable light curves due to the improvement in the feature extraction.

Figure 14 demonstrates the Receiver Operator Characteristic (ROC) Curve of the PolyFit model trained with the optimal hyperparameters. The poorest performance is found on the spotted stars, overtone mode RR Lyrae variables

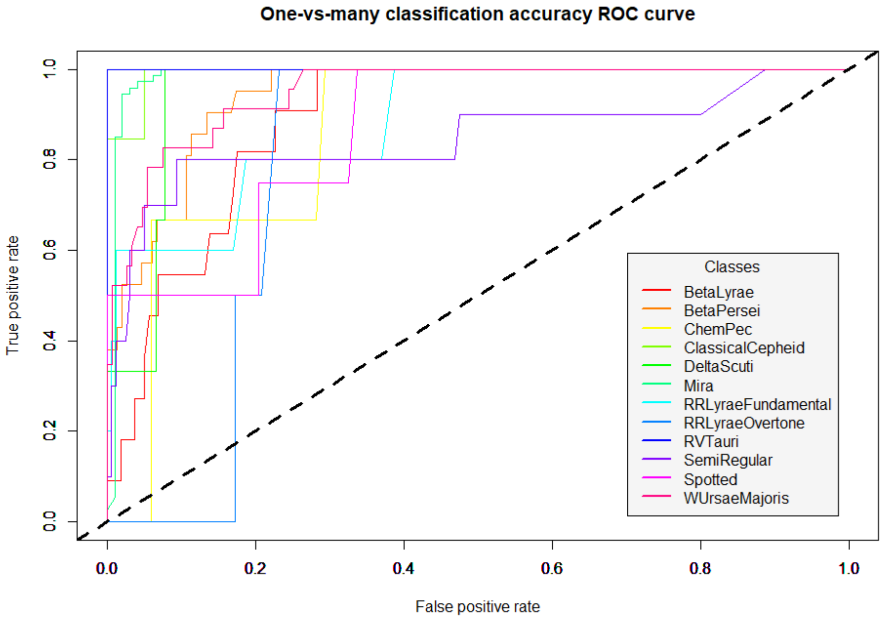


Figure 14: ROC curve of the model trained at the optimal hyperparameters. Many classes exhibit good performance as they obtain high recall of the true light curves of their respective classes whilst minimising the false positives. The poorest performing classes are the spotted stars, overtone mode RR Lyrae variables and the chemically peculiar variables. This failure is likely due to the low amplitude of the variability of these classes resulting in lower signal-to-noise in Skycam.

and the chemically peculiar variables. These three classes exhibit highly sinusoidal variability which benefits the period estimation routine in GRAPE but they also tend to have low amplitudes. Combined with the relatively high noise present in the Skycam data, it results in a poor signal-to-noise for many of these objects which impedes the period estimation and PolyFit interpolation procedures. As the representation learning features are dependent on these steps being performed accurately, the distributions of these learned feature for these classes can be uninformative.

The mean decrease GINI of the Random Forest can be used to display the importance of the features in the trained classification model. Figure 15 demonstrates the importance of the top 20 features in the classification of the 859 light curves with the PolyFit model. The period is the dominant feature as expected by the definition of many variable star types being based on this property of the variability. The interpolated amplitude features are the next most important which again relates to the amplitude being an important part of the definition of the variability classes. The interpolated Range of a Cumulative Sum and interpolated Skewness features are also of higher importance than many other features. The interesting selection is the use of the second principal component of the PolyFit model folded at the period. This was highlighted as a possible discriminator between a number of classes which overlap strongly in the Period and Amplitude feature space. The mean decrease GINI feature can also be determined for a specific class and for the two RR Lyrae variable types and the W Ursae Majoris eclipsing binaries this feature became the second or third most important feature replacing the interpolated amplitude features although period still retained the top position.

These trained models indicate that the PolyFit derived features contain significant knowledge on the shape and distribution of the variable light curves. These features allow the discrimination of the twelve chosen variability classes with reasonably strong performance using the SkycamT light curve database. The features are also limited as they are specifically tuned to detect shaped based information without taking into account the suitableness of the initial epoch-folding operation. This means the features rapidly loose importance and meaning in the event of a light curve being semi-periodic or non-periodic.

The other primary limitation is the dependency on period. If an incorrect period is estimated, the interpolated features will be poor and possibly inferior to the variability indices they were designed to replace. Ultimately, due to the importance of period, an incorrect period estimation is likely to have wider ranging problems than those caused by a poorly generated PolyFit. The PolyFit features have been shown to be powerful on the noisy SkycamT light curves yet they cannot be applied to the light curve classification task alone and should be used in concert with the features discussed in chapter 5. Regardless of the

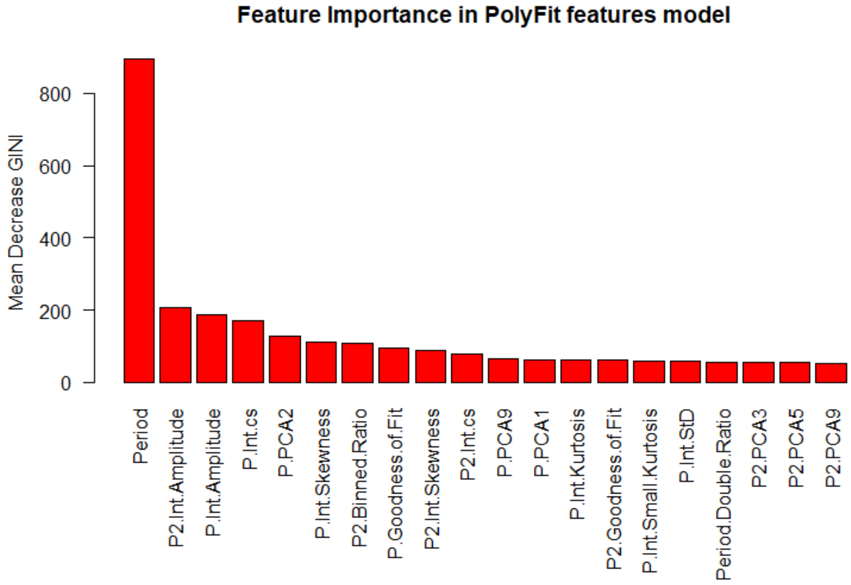


Figure 15: Bar Plot of the Mean Decrease GINI of the top 20 features in the PolyFit feature Random Forest model. As expected, the Period feature is dominant in the classification task followed by the interpolated amplitudes. The second principal component of the PolyFit model folded at the period is also important as was suspected earlier. Other important features include the interpolated Range of a Cumulative Sum feature and the interpolated Skewness.

strengths and weaknesses of this approach, this investigation has shown that representation learning methods can generate new and informative features for learned light curve classification models.

5. Conclusion

We have implemented an unsupervised dimensionality reduction method in the form of a PCA applied to PolyFit interpolated phased light curves to automatically extract variability representative features from a training set of 6897 variable stars of 18 different classes. These features are used to train classification models using the Random Forest algorithm and compared with models produced by the engineered features of previous studies. We find that the new representation learning models slightly outperform the engineered feature models on Skycam data by 0.4729 with a standard deviation of 0.0931 compared to the 0.3902 with a standard deviation of 0.0619 of the feature engineering due to the new features being more resilient to the noise inherent to the Skycam light curves. The mean decrease GINI measure of the feature importance reveals that period is the dominant classification feature. This is not surprising as many variable star classes are defined by their periods. However, it is clear that the PolyFit interpolated features such as the interpolated amplitude were just as capable as the Fourier amplitudes at recording the amplitude of variability of a source. Additionally, representation learned features such as the second principal component are useful in separated certain object classes such as the eclipsing and pulsating short period variables which have a heavily convolved period space.

This improvement is dependent on a good measurement of the periodicity of any candidate light curve shifting alot of the pressure for correct classification onto the GRAPE method. As this method was designed for use on this dataset, we believe it's capability is sufficient for this important task. As the representation learned features are directly learned from a bulk of Skycam light curves, the cadence artifacts related to this survey are automatically incorporated into the classification model. However, it is important to note that due to the sampling limitations, major features of an underlying astrophysical signal may not be sampled. This case is true for eclipsing binaries with long periods as the eclipse feature becomes a very narrow, short duration event (Prsa et al., 2011, Wells et al., 2017, Kochoska et al., 2017). It is possible to introduce an element of manual assistance in the training of these classification models through the use of Active Learning (Richards et al., 2011a). This method was applied to the training of 50,000 variable sources in the All-Sky Automated Survey (ASAS) allowing for a probabilistic analysis of the variability of these light curves (Richards et al., 2012).

Our follow-up work will involve using the GRAPE method and the PolyFit representation learning described in this paper to produce a classification system to run alongside the Liverpool Telescope which will automatically produce candidate sources for future study during normal telescope operation. As noted above, this is a unique system as previous survey telescopes were separate to the follow-up science telescopes. By combining both in a single location, costs can be substantially reduced whilst maintaining scientific output. The success of this method and the development of the classification system it will facilitate will allow for many telescope installations to deploy similar style wide field cameras. As these cameras do not require use of telescope resources, primarily time and cost, they are relatively inexpensive ways of collecting large quantities of astronomy data. Combined with systems such as ours designed to exploit this data, the sky can be monitored for time-domain events routinely without requiring dedicated all sky surveys outside of specific examples such as solar system objects and some extragalactic astronomy. Even a catalogue of variable stars alone can provide interesting objects for follow-up research. Variable stars are probes into fascinating phases of stellar evolution allowing our understanding of these processes to be improved. It also offers a great potential for the expert systems community to get involved in big astronomical survey infrastructure projects as there are possible survey styles currently being ignored due to the lack of software and methods to fully realise them.

There are also a number of further investigations recommended to extend the conclusions of this paper. The PolyFit method has been demonstrated here and elsewhere (Prsa et al., 2008, Paegert et al., 2014, Parvizi et al., 2014) to be a potent method of fitting the folded light curves of pulsating and eclipsing variable stars. There are other interpolation methods which may produce superior fitting models without requiring an input period which can be a substantial source of error. Gaussian Processes are a method which have been successfully used for other astronomical light curves such as transients which can operate over variable timescales without the rigid limitations of a period (Faraway et al., 2014). The only concern is the possibility such a model may overfit noisy light curves such as those from the Skycam data. The dimensionality reduction methods can also be improved with more powerful algorithms. PCA is a powerful method but is limited by its linear nature. By definition it can only produce principal components which are a linear sum of the original features. There are a number of methods which can extend this process into the non-linear regime. The t-distributed stochastic neighbour embedding algorithm is another dimensionality reduction technique originally designed for visualization (van der Maaten and Hinton, 2008). This method makes use of a similarity measure between objects in the original feature space to derive a 'probability of neighbourhood' such that a lower dimension feature space is selected where high probabilities neighbours are located a small distance from eachother in this new feature space. Further to this, the interactions can be modelled directly using a deep learning architecture, such as non-linear PCA and deep Autoencoders which are also capable of determining similar non-linear interactions (Baldi, 2011). By applying these more powerful techniques, the PolyFit model can better model the noisy light curves with a set of features which describe the important properties of the light curves required for classification. Dimensionality reduction is likely an easier problem to address as noisier light curves are not a big limitation as long as the training data size is increased appropriately to prevent overfitting.

Acknowledgment

The Liverpool Telescope is operated on the island of La Palma by Liverpool John Moores University in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias with financial support from the UK Science and Technology Facilities Council. We acknowledge with thanks the variable star observations from the AAVSO International Database contributed by observers worldwide and used in this research.

References

Alcock, C., Allsman, R. A., Alves, D. R., Axelrod, T. S., Becker, A. C., Bennett, D. P., Cook, K. H., Dalal, N., Drake, A. J., and Freeman, K. C. (2000). The MACHO project: Microlensing results from 5.7 years of large magellanic cloud observations. *The Astrophysical Journal*, 542(1):281–307.

Baldi, P. (2011). Autoencoders, unsupervised learning and deep architectures. *UTLW'11 Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop*, 27:37–50.

Benavente, P., Protopapas, P., and Pichara, K. (2017). Automatic survey-invariant classification of variable stars. *The Astrophysical Journal*, 845(2):147.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Bertin, E. and Arnouts, S. (1996). SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, 117(2):393–404.

Charbonneau, P. (1995). Genetic algorithms in astronomy and astrophysics. *The Astrophysical Journal Supplement Series*, 101:309.

Copperwheat, C. M., Steele, I. A., Piascik, A. S., Bersier, D., Bode, M. F., Collins, C. A., Darnley, M. J., Galloway, D. K., Gomboc, A., and Kobayashi, S. (2016). Liverpool telescope follow-up of candidate electromagnetic counterparts during the first run of advanced ligo. *Monthly Notices of the Royal Astronomical Society*, 462(4):3528–3536.

Deb, S. and Singh, H. P. (2009). Light curve analysis of variable stars using fourier decomposition and principal component analysis. *Astronomy & Astrophysics*, 507(3):1729–1737.

Debosscher, J., Sarro, L. M., Aerts, C., Cuypers, J., Vandenbussche, B., Garrido, R., and Solano, E. (2007). Automated supervised classification of variable stars. *Astronomy & Astrophysics*, 475(3):1159–1183.

Eyer, L. and Mowlavi, N. (2008). Variable stars across the observational hr diagram. *Journal of Physics: Conference Series*, 118:012010.

Faraway, J., Mahabal, A., Sun, J., Wang, X., Wang, Y., and Zhang, L. (2014). Modeling light curves for improved classification. *arXiv:1401.3211*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 & 498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Huijse, P., Estevez, P. A., Protopapas, P., Principe, J. C., and Zegers, P. (2014). Computational intelligence challenges and applications on large-scale astronomical time series databases. *IEEE Computational Intelligence Magazine*, 9(3):27–39.

Huijse, P., Estevez, P. A., Protopapas, P., Zegers, P., and Principe, J. C. (2012). An information theoretic algorithm for finding periodicities in stellar light curves. *IEEE Transactions on Signal Processing*, 60(10):5135–5145.

Ivezic, Z. (2014). LSST: from science drivers to reference design and anticipated data products. *arXiv:0805.2366v4*.

Kaiser, N. (2002). Pan-STARRS: A large synoptic survey telescope array. *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 4836:154–164.

Kim, D.-W. and Bailer-Jones, C. A. L. (2016). A package for the automated classification of periodic variable stars. *Astronomy & Astrophysics*, 587.

Kim, D.-W., Protopapas, P., Byun, Y.-I., Alcock, C., Khardon, R., and Trichas, M. (2011). Quasi-stellar object selection algorithm using time variability and machine learning: Selection of 1620 quasi-stellar object candidates from macho large magellanic cloud database. *The Astrophysical Journal*, 735(2):68.

Kochoska, A., Mowlavi, N., Prsa, A., Lecoeur-Tabi, I., Holl, B., Rimoldini, L., Sveges, M., and Eyer, L. (2017). Gaia eclipsing binary and multiple systems. a study of detectability and classification of eclipsing binaries with gaia. *Astronomy & Astrophysics*, 602.

Kugler, S. D., Gianniotis, N., and Polsterer, K. L. (2016). An explorative approach for inspecting Kepler data. *Monthly Notices of the Royal Astronomical Society*, 455(4):4399–4405.

LaCourse, D. M., Jek, K. J., Jacobs, T. L., Winarski, T., Boyajian, T. S., Rappaport, R., Sanchis-Ojeda, R., Conroy, K. E., Nelson, L., Barclay, T., Fischer, D. A., Schmitt, J. R., Wang, J., Stassun, K. G., Pepper, J., Coughlin, J. L., Shporer, A., and A., P. (2015). Kepler eclipsing binary stars - vi. identification of eclipsing binaries in the k2 campaign o data set. *Monthly Notices of the Royal Astronomical Society*, 452(4):3561–3592.

Lang, D., Hogg, D. W., Mierle, K., Blanton, M., and Roweis, S. (2010). Astrometry.net: Blind astrometric calibration of arbitrary astronomical images. *The Astronomical Journal*, 139(5):1782–1800.

Larson, S. (2003). The CSS and SSS NEO surveys. *AAS/Division for Planetary Sciences Meeting Abstracts*, 35:982.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462.

Matijevic, G. (2012). Kepler eclipsing binary stars. III. classification of Kepler eclipsing binary light curves with locally linear embedding. *The Astronomical Journal*, 143:123–128.

Mawson, N. R., Steele, I. A., and Smith, R. J. (2013). STILT: System design and performance. *Astronomische Nachrichten*, 334(7):729–737.

McWhirter, P. R., Steele, I. A., Al-Jumeily, D., Hussain, A., and Vellasco, M. M. B. R. (2017). The classification of periodic light curves from non-survey optimized observational data through automated extraction of phase-based visual features. *2017 International Joint Conference on Neural Networks (IJCNN)*.

McWhirter, P. R., Steele, I. A., Hussian, A., Al-Jumeily, D., and Vellasco, M. M. B. R. (2018). Grape: Genetic routine for astronomical period estimation. *Monthly Notices of the Royal Astronomical Society*, 479(4):5196–5213.

McWhirter, P. R., Wright, S., Steele, I. A., Al-Jumeily, D., Hussain, A., and Fergus, P. (2016). A dynamic, modular intelligent-agent framework for astronomical light curve analysis and classification. *Intelligent Computing Theories and Application Lecture Notes in Computer Science*, pages 820–831.

Mortier, A. and Cameron, A. C. (2017). Stacked bayesian general lomb-scargle periodogram: Identifying stellar activity signals. *Astronomy & Astrophysics*, 601.

Mortier, A., Faria, J. P., Correia, C. M., Santerne, A., and Santos, N. C. (2015). BGLS: A bayesian formalism for the generalised lomb-scargle periodogram. *Astronomy & Astrophysics*, 573.

Neff, J. E., Wells, M. A., Geltz, S. N., and A., B. (2014). Automated variability classification and constant stars in the Kepler database. *18th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun.*

Paegert, M., Stassun, K. G., and Burger, D. M. (2014). The eb factory project. i. a fast, neural-net-based, general purpose light curve classifier optimized for eclipsing binaries. *The Astronomical Journal*, 148(2):31.

Parvizi, M., Paegert, M., and Stassun, K. G. (2014). The eb factory project. II. validation with the Kepler field in preparation for K2 and TESS. *The Astronomical Journal*, 148(6):125.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.

Percy, J. R. (2008). *Variable stars.* Cambridge University Press.

Pichara, K., Protopapas, P., and Len, D. (2016). Meta-classification for variable stars. *The Astrophysical Journal*, 819(1):18.

Protopapas, P., Giammarco, J. M., Faccioli, L., Struble, M. F., Dave, R., and Alcock, C. (2006). Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 369(2):677–696.

Protopapas, P., Huijse, P., Estvez, P. A., Zegers, P., Prncipe, J. C., and Marquette, J.-B. (2015). A novel, fully automated pipeline for period estimation in the eros 2 data set. *The Astrophysical Journal Supplement Series*, 216(2):25.

Prsa, A., Guinan, E. F., Devinney, E. J., Degeorge, M., Bradstreet, D. H., Giammarco, J. M., Alcock, C. R., and Engle, S. G. (2008). Artificial intelligence approach to the determination of physical properties of eclipsing binaries. i. the ebai project. *The Astrophysical Journal*, 687(1):542–565.

Prsa, A., Pepper, J., and Stassun, K. G. (2011). Expected large synoptic survey telescope (LSST) yield of eclipsing binary stars. *The Astronomical Journal*, 142(2):52.

Rahal, Y. R., Afonso, C., Albert, J.-N., Andersen, J., Ansari, R., Aubourg, ., Bareyre, P., Beaulieu, J.-P., Charlot, X., and Couchot, F. (2009). The EROS2 search for microlensing events towards the spiral arms:the complete seven season results. *Astronomy & Astrophysics*, 500(3):1027–1044.

Richards, J. W., Starr, D. L., Brink, H., Miller, A. A., Bloom, J. S., Butler, N. R., James, J. B., Long, J. P., and Rice, J. (2011a). Active learning to overcome sample selection bias: Application to photometric variable star classification. *The Astrophysical Journal*, 744(2):192.

Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., and Rischard, M. (2011b). On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10.

Richards, J. W., Starr, D. L., Miller, A. A., Bloom, J. S., Butler, N. R., Brink, H., and Crellin-Quick, A. (2012). Construction of a calibrated probabilistic classification catalog: Application to 50k variable sources in the all-sky automated survey. *The Astrophysical Journal Supplement Series*, 203(2):32.

Scargle, J. D. (1982). Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835.

Steele, I. A., Smith, R. J., Rees, P. C., Baker, I. P., Bates, S. D., Bode, M. F., Bowman, M. K., Carter, D., Etherton, J., and Ford, M. J. (2004). The liverpool telescope: performance and first results. *Ground-based Telescopes*.

Tanvir, N. R., Hendry, M. A., Watkins, A., Kanbur, S. M., Berdnikov, L. N., and Ngeow, C. C. (2005). Determination of cepheid parameters by light-curve template fitting. *Monthly Notices of the Royal Astronomical Society*, 363(3):749–762.

Udalski, A., Kubiak, M., and Szymanski, M. (1997). Optical gravitational lensing experiment. OGLE-2 – the second phase of the OGLE project. *Acta Astronomica*, 47:319–344.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2608.

Vaughan, S. (2011). Random time series in astronomy. *Philosophical Transactions of the Royal Society*, 371.

Wells, M., Prsa, A., Jones, L., and Yoachim, P. (2017). Initial estimates on the performance of the lsst on the detection of eclipsing binaries. *Publications of the Astronomical Society of the Pacific*, 129(976):065003.

Yoachim, P., Mccommas, L. P., Dalcanton, J. J., and Williams, B. F. (2009). A panoply of cepheid light curve templates. *The Astronomical Journal*, 137(6):4697–4706.

York, D. G., Adelman, J., and Anderson, J. E. (2000). The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579–2000.