Preprint typeset using LaTeX style AASTeX6 v. 1.0

A NOVEL DETECTION METRIC DESIGNED FOR IDENTIFICATION OF O'CONNELL EFFECT ECLIPSING BINARIES

K.B. Johnston
Physics and Space Sciences Dept., Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL, US

R. Haber
Mathematical Sciences Dept., Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL, US

S.M. Caballero-Nieves
Physics and Space Sciences Dept., Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL, US

A.M. Peter
Engineering Systems Dept., Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL, US

V. Petit
Physics and Astronomy Dept., University of Delaware, Newark, DE, USA

M. Knote
Physics and Space Sciences Dept., Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL, US

Abstract

With the advent of digital astronomy, new benefits and new challenges have been presented to the modern day astronomer. Here we focus on the construction and application of a novel time-domain signature extraction methodology and the development of a supporting supervised pattern detection algorithm for the targeted identification of eclipsing binaries which demonstrate a feature known as the O'Connell Effect. Our proposed methodology maps stellar variable observations (time-domain data) to a new representation known as Distribution Fields (DF), whose properties enable us to efficiently handle issues such as irregular sampling and multiple values per time instance. Given this novel representation, we develop a metric learning technique directly on the DF space capable of specifically identifying our stars of interest. The metric is tuned on a set of labeled eclipsing binary data from the Kepler survey, targeting particular systems exhibiting the O'Connell Effect. Our framework demonstrates favorable performance on Kepler EB data, taking a crucial step to prepare the way for large-scale data volumes from next generation telescopes such as LSST.

Keywords: binaries: eclipsing – methods: data analysis – methods: statistical

1. INTRODUCTION

With the rise of large scale surveys such as the Kepler, the Transiting Exoplanet Survey Satellite (TESS), 2MASS, the Kilodegree Extremely Little Telescope (KELT), the Large Synoptic Survey Telescope (LSST), Pan-STARRS, and the like, a fundamental working knowledge of statistical data analysis and data management for the reasonable processing of astronomical data is necessary. The ability to mine these datasets for new and interesting astronomical

kyjohnst2000@my.fit.edu

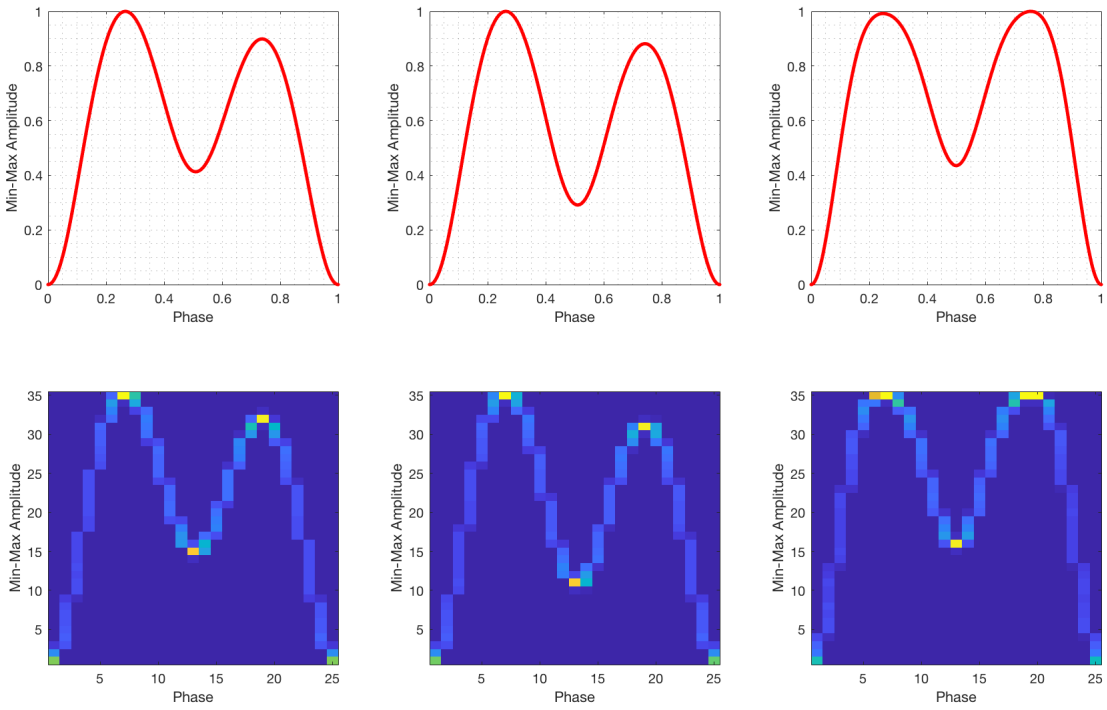

Figure 1. Example of DF feature space representation, phased light curves are presented on the top row, DF features are presented on the bottom row. The left column and center columns are targets of interest (OEEB), the right column is a example of a non-OEEB binary.

information opens a number of scientific windows that were once limited by poor sampling; both in terms of number of stars (targets) and depth of observations (number of samples). While this paper will focus on the construction of a framework for target searching and identification, in a larger context this work is advocating for efforts that are tailored to the needs of specific research, focused on automated analysis algorithms like supervised classification.

This paper focuses on the construction and application of a novel time-domain signature extraction methodology and the development of a supporting supervised pattern detection algorithm for the targeted identification of eclipsing binaries that demonstrate a feature known as the O'Connell Effect. The O'Connell Effect (O'Connell 1951) is defined for eclipsing binaries as an asymmetry in the maxima of the phased light curve. This asymmetry is unexpected, as it suggests an orientation dependency in the brightness of the system. Similarly, the consistancy of the asymmetric over many orbits is also surprising, as it suggests the cause of the asymmetry is dependent on the rotation of the binary system. While the cause of the O'Connell Effect is not fully understood, a number of explanations have been made, and additional data and modeling is necessary for further investigation (McCartney 1999).

To support the discovery of new eclipsing binaries which demonstrate the O'Connell Effect we present a novel detection framework that maps time-domain stellar variable observations to an alternate Distribution Field (DF) representation, and then, develops a metric learning approach to identify O'Connell Effect Eclipsing Binaries (OEEBs). The DF representation (Sevilla-Lara & Learned-Miller 2012) maps deterministic, functional stellar variable observations to a stochastic matrix, with the rows summing to unity. The inherently probabilistic nature of DFs provide a robust way to mitigate inter-class class variability and simultaneously handle irregular sampling rates and multi-valued observations associated with stellar observations. Figure 1 illustrates how two OEBB observations that have noticeable differences in their time-domain form can produce similar DFs; thus improving their chances of being detected as belonging to the same class. It also shows how observations from a non-OEEB maps to different looking DFs.

Though the DF naturally exhibits some discriminative properties, it alone is not sufficient for our ultimate goal of detection. Rather than vectorizing the DF matrix and treating it as a feature vector for standard classification techniques, we treat the DF as the matrix-valued feature that it is. This allows for the retention of row and column dependence information that would normally be lost in the vectorization process (Ding & Dennis Cook 2018). Based on the matrix-valued DF feature, we adopt a metric learning framework to directly learn a distance metric on the space of DFs. The learned metric can then be utilized as a measure of similarity to detect OEEB based on their closeness to other OEEBs. Our metric learning approach is presented as a competitive push-pull optimization, where DFs corresponding

to OEEBs influence the learned metric to measure them as being nearer in the DF space. Simultaneously, DFs corresponding to non-OEEBs are pushed away and result in large measured distances under the learned metric.

The prior efforts employing machine learning to classify variable stars are reviewed in Section 1.1, including a discussion on the target variable star class. We outline the novel proposed pipeline for OEEBs detection in Sections 2.1 to 2.3, and detail two competing approaches that will be used for comparison in Section 2.4. The train and testing strategies for our metric learning framework are developed in Section 3.1. This is followed by an experimental analysis where the performance of our method is evaluated. Assessments include the LINEAR dataset, where we demonstrate the ability to extract new targets of interest. Finally, we conclude with a summary of our findings and directions for future research.

1.1. *Prior Methodologies*

The idea of constructing a supervised classification algorithm for stellar classification is not unique to this paper (Dubath et al. 2011). Methods pursued include the construction of a detector to determine variability (Barclay et al. 2011), the design of random forests for the detection of photometric redshifts in spectra (Carliles et al. 2010), the detection of transient events (Djorgovski et al. 2012) and the development of machine-assisted discovery of astronomical parameter relationships (Graham et al. 2013a). Debosscher (2009) explored several classification techniques for the supervised classification of variable stars, quantitatively comparing the performance in terms of computational speed and performance. Likewise, other efforts have focused on comparing speed and robustness of various methods (Blomme et al. 2011; Pichara et al. 2012; Pichara & Protopapas 2013). These methods span both different classifiers and different spectral regimes, including infrared (IR) (Angeloni et al. 2014; Masci et al. 2014), radio frequency (RF) (Rebbapragada et al. 2011) and optical (Richards et al. 2012). Methods for automated supervised classification include procedures such as: direct parametric analysis (Udalski et al. 1999), fully automated neural networking (Pojmanski 2000, 2002) and Bayesian classification (Eyer & Blake 2005).

Many of these efforts have, with the advent of massive surveys such as the LSST, focused mostly on multi-class classification. The costs and benefits of multi-class verses two-class classification have been demonstrated in Rifkin & Klautau (2004); Beygelzimer et al. (2005). Likewise, the effect that a lack of training data, a greatly imbalanced class set, and an incomplete or ill-defined class space has on multi-class classification has also been reviewed (Johnston & Oluseyi 2017). Methods for addressing these concerns were also addressed by Johnston & Oluseyi (2017), who proposed anomaly detection, and one vs. all classification (detection) as measures for improving pipeline performance and mitigating the risks associated with variable star classification.

The Fourier domain feature space developed in Debosscher (2009), has been used in a number of variable star classification papers (many of the ones addressed above), and has been shown to be useful in generic multi-classification. Johnston & Peter (2017) showed that shape based features can be just as effective in variable star identification. This paper focuses on the construction and application of a novel time-domain signature extraction methodology based on the phased light curve functional shape, and the development of a supporting supervised pattern classification algorithm for the identification of variable stars. Given the reduction of a survey of stars into a standard feature space, the problem of using prior patterns to identify new observed patterns can be addressed via classification algorithms. These methods have two large advantages over manual-classification procedures: the rate at which new data is processed is dependent only on the computational processing power available and the performance of a supervised classification algorithm is quantifiable and consistent.

1.2. *Eclipsing Binaries with the O'Connell Effect*

This work focuses on a particular set of eclipsing binaries whose light curves exhibit the O'Connell Effect (OEEB, O'Connell 1951), an asymmetry of the maxima. This effect suggests that the total brightness of the system in their un-eclipsed state is dependent on the orientation of the binary system, specifically if the primary is coming out of or going into eclipse. Several theories explain the effect have been proposed, including: starspots, gas stream impact, and circumstellar matter (McCartney 1999). The work by Wilsey & Beaky (2009) outlines each of these theories, and demonstrates how the observed effects are generated by the underlying physics.

Starspots result from chromospheric activity, causing flares in brightness. To explain the observed effect, however, these flares would need to be both consistent in brightness as well as in position (long term stability). Gas stream impact results from matter transferring between stars (smaller to larger) through the L1 point and onto a specific position on the larger star, resulting in a consistent brightening on the leading/trailing side of the secondary/primary. The circumstellar matter theory attempts to describe the increase in brightness via free falling matter being swept up, resulting in energy loss and heating, again causing an increase in amplitude. Alternatively, circumstellar matter

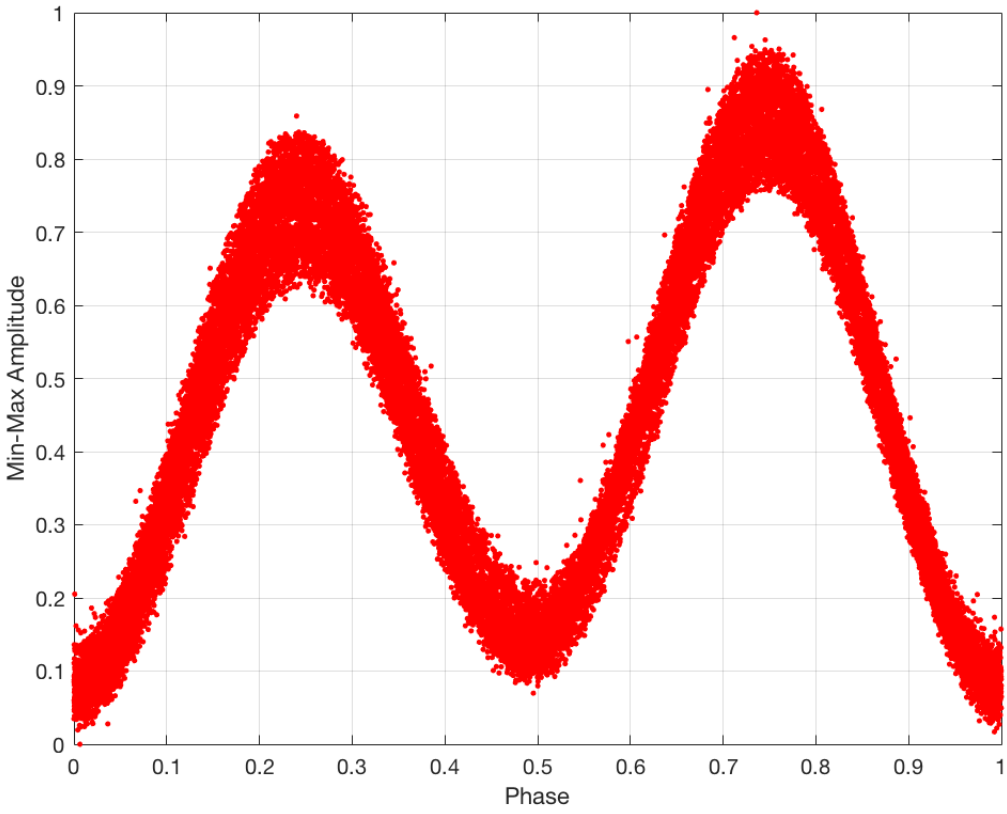

Figure 2. An example phased light curve of a Eclipsing Binary with the O'Connell Effect. Note, the light curve has been phased such that: the global minimum (cooler in front of hotter) is at lag 0 and the secondary minimum (hotter in front of cooler) is at approximately lag 0.5. The side-by-side on binary orientations are at approximately 0.25 and 0.75.

in orbit could also result in attenuation, i.e. the difference in maximum magnitude of the phased light curve results from a dimming and not brightening.

Even in the McCartney (1999) comprehensive study, the sample was limited to only six star systems: GSC 03751-00178, V573 Lyrae, V1038 Herculis, ZZ Pegasus, V1901 Cygni, and UV Monocerotis; the data of two of these are presented in the paper. Standard eclipsing binary simulations were used (Wilson & Devinney 1971) to demonstrate the spot proposed for each light curve instance, and estimate the parameter associated with the physics of the system. Wilsey & Beaky (2009) noted other cases of O'Connell Effect in binaries, which have since been described physically; in some cases the effect varied over time, while in other the effect was consistent over years of observation and over many orbits. The effect has been found in both over-contact, semi-detached, and near-contact systems.

An increased number of identified targets of interest is required to provide the sample size needed for a complete statistical analysis of the O'Connell Effect. The new OEEB systems discovered by the method of automated detection proposed here can be used to further investigate their frequency of occurrence, provide constraints on existing light curve models, and provide parameters to look for these systems in future large-scale variability surveys such as LSST. While the effort here targets OEEB as a demonstration, it need not be limited to those particular targets. Any variable star (e.g. supernova, RR Lyr, Cepheids, Eclipsing Binaries, etc.) can be targeted, given the appropriate feature space transformation allowing for quantitative evaluation of similarity. This design could be directly applicable to exo-planet discovery; either via light curve detection (e.g. to detect eclipsing exo-planets), or via machine learning applied to other means (e.g. spectral analysis).

1.3. *Proposed Methodology*

Proposed is a detection methodology for a specific target of interest, an O'Connell Effect Eclipsing Binary (OEEB), defined as an eclipsing binary where the light curve (LC) maxima are consistently over the span of observation at different amplitudes. Beyond differences in maxima, and a number of published examples, little is defined as a requirement for identifying "the O'Connell Effect" (Wilsey & Beaky 2009; Matthew F. Knote submitted). The bounding or defining of descriptive statistics or functional relationships such as: delta max amplitude (Δm_{max}), delta

min amplitude (Δm_{min}), expected period, etc., requires a larger sample than is presently available.

This effort proposes using functional shape as the target of interest indicator; furthermore, the quantification of functional statistics allow for the improved understanding of not just the standard definition of the variable, but also the population distribution as a whole. These estimates allow for empirical statements to be made regarding the feature distributions of OEEB type light curves. The determination of an empirically observed distribution, however, requires a significant sample from which to make the descriptive statistics for various measurements. To generate these empirical estimates a large dataset is required, the challenge is how to consistently select similar targets of interest in an automated fashion.

2. METHODS AND DESIGN

Relying on previous designs in astroinformatics to develop a supervised detection algorithm (Johnston & Oluseyi 2017), a design is proposed that tailors the requirements specifically towards the detection of OEEB type variable stars. Leveraging the Kepler pipeline already in place, and using the data from the Villanova EB catalog (Kirk et al. 2016), this study focuses on a set of pre-determined EBs identified from the Kepler catalog. From this catalog, an initial labeled dataset of proposed targets of interest is generated and identified as OEEB; likewise a set of targets identified as "not interesting" based on expert definition, i.e. intuitive inference, is also generated. Next, a feature space is selected that will be sensitive to the signature features of interest. Finally, a classification methodology that allows for the tailored feature space and is able to produce values of similarity is selected; the similarity measure is critical for an estimate of confidence when applying the classifier to new unlabeled dataset (Bellet et al. 2015), or in the implementation of an anomaly detection algorithm (Chandola et al. 2009). The algorithm, including comparison methodologies, designed feature space transformations, classifiers, utilities, etc. is publicly available at the project repository via GitHub[1]; all code was developed in MATLAB and was run on MATLAB 9.3.0.713579 (R2017b).

2.1. *Signal Conditioning and Signal Processing*

Prior to feature space processing, the raw observed photometric time domain data is conditioned and processed. This effort includes long term trend removal, artifact removal, initial light curve phasing and initial Eclipsing Binary identification. The is performed prior to the effort by the Eclipsing Binary catalog (this survey will be using all 2875 long-cadence light curves available as of the date of publication). The functional shape of the phased light curve has been selected as the feature to be used in the machine learning process, i.e. detection of targets of interest. While the data has been conditioned already by the Kepler pipeline, added steps are taken to allow for similarity estimation between phased curves. Friedman's SUPERSMOOTHER algorithm (Friedman & Silverman 1989) is used to generate a smooth 1-D functional curve from the phased light curve data. The smoothed curves are normalized via the Min-Max scaling Equation 1:

$$f(t)' = \frac{f(t) - \min(f(t))}{\max\left(f(t)\right) - \min(f(t))} \tag{1}$$

where $f(t)$is the phased light curve, f is the raw amplitude from the database source and t is the phase where $t \in [0, 1]$, and $f(t)'$is the min-max scaled amplitude. The min of the smoothed phased light curve is used as a registration marker and both the smoothed and unsmoothed light curved are aligned such that lag/phase zero corresponds to min amplitude (minima eclipse, see McCartney 1999). Stored with the original aligned phased light curve data is the smoothed aligned curve, as well as a regularly sampled smoothed curve generated by interpolating the smooth, unique, 1-D functional curve generated from the SUPERSMOOTHER algorithm. This set of three datasets is used at different points in this analysis.

2.2. *Distribution Fields for 1D Signal Classification*

As stated, this analysis focuses on the detection of OEEB systems based on their light curve shape. The OEEB signature has a cyclostationary signal, a functional shape that repeats with some frequency. The signature can be isolated using a process of period finding, folding, and phasing (Graham et al. 2013b). The proposed feature space transformation will focus on the quantification or representation of this phased functional shape. This particular implementation design makes the most intuitive sense, as visual inspection of the phased light curve is the way experts identify these unique sources. A transformation of the phased light curve into a feature space that is machine understandable is required for machine learning.

[1] https://GitHub.com/kjohnston82/OCDetector

As discussed, prior research on time domain data identification has varied between generating machine learned features (Gagniuc 2017), implementing generic features (Masci et al. 2014; Palaversa et al. 2013; Richards et al. 2012; Debosscher 2009), and looking at shape or functional based features (Haber et al. 2015; Johnston & Peter 2017; Park & Cho 2013). This analysis will leverage the distribution field transform to generate a feature space that can be operated on a distribution field (DF), and an array of probability distributions, where probability at each element is defined as (Helfer et al. 2015; Sevilla-Lara & Learned-Miller 2012) Equation 2:

$$DF_{ij} = \frac{\sum \left[y_j < f\left(x_i \leq t \leq x_{i+1} \right) < y_{j-1} \right]}{\sum \left[y_j < f\left(t \right) < y_{j-1} \right]} \tag{2}$$

Where, [ ] is the Iverson Bracket (Iverson 1962), and y_j and x_i are the corresponding normalized amplitude and phased time bins, respectively. The result is a 2-D histogram that is a right stochastic matrix, i.e. the rows sum to one. Bin number (in either x or y direction), is optimized by cross-validation as part of the classification training process. The smoothed phased data (generated from SUPERSMOOTHER), is provided to the DF algorithm. This implementation was found to produce a more consistent classification process; it was found that the Min-Max scaling normalization, when outliers are present, can produce final patterns that focus more on the outlier than the general functionality of the light curve. In the original Helfer et al. (2015) proposal for DF learning, the signal was assumed to be a 1-D waveform (functional) input. In the 1-D case, convolution can be used to achieve a representation that takes into account the feature values of the surrounding points, that is the case here as well and convolution is included as part of this analysis. The transformation outlined here of smoothing and mapping to the DF feature space provides a domain in which: (1) the phased light curve (functional shape) is represented, (2) variation in amplitude is provided for, (3) variation in phase is provided for, and (4) some amount of dimensionality reduction as compared to the original light curve or the smoothed dataset is implemented.

2.3. *Push-Pull Matrix Metric Learning*

At its core, the proposed detector is based on the definition of similarity, and more formally a definition of distance. Consider the example triplet "x is more similar to y than to z", i.e. the distance between x and y in the feature space of interest is smaller than the distance between x and z. The field of metric learning concerns itself with the definition of this distance in a given feature space to optimize a given goal, most commonly the reduction of error rate associated with the classification process. Given the feature space selected of DF matrices, the distance between two matrices X and Y Bellet et al. (2015) is defined as Equation 3:

$$d(X,Y) = \|X - Y\|_M^2 = tr\left\{ (X-Y)^T M (X-Y) \right\} \tag{3}$$

where $M \in \mathbb{R}^{d \times d}$ and $M \succeq 0$ (positive semi-definite). The procedure outlined in Helfer et al. (2015) is similar to the metric learning methodology LMNN (Weinberger et al. 2006), and is summerized as following follows: the developed objective function is given in Equation 4:

$$E = \frac{1-\lambda}{N_c - 1} \sum_{i,j} \left\| DF_c^i - DF_c^j \right\|_M^2 - \frac{\lambda}{N - N_c} \sum_{i,k} \left\| DF_c^i - DF_c^k \right\|_M^2 + \frac{\gamma}{2} \|M\|_F^2 \tag{4}$$

where the triplet $\left\{ DF_c^i, DF_c^j, DF_c^k \right\}$ i.e. DF_c^i is similar to DF_c^j and dissimilar to DF_c^k. Clearly there are three basic components: a "Pull term" which is small when the distances between similar observations is small, a "Push term" which is small when the distances between dissimilar observation is larger, and a regularization term which is small when the Forbinus norm of M is small. Thus, the algorithm attempts to bring similar distribution fields closer together, while pushing dissimilar ones further apart, while attempting to minimize the complexity of the metric M. The regularizer on the metric M guards against over-fitting and consequently enhances algorithm's ability to generalize better---a strategy similar in spirit to popular regression techniques like lasso and ridge (Hastie et al. 2009).

Additional parameters λ and γ weight the importance of the push-pull terms and metric regularizer, respectively. These free parameters are typically tuned via standard cross-validation techniques on the training data. The objective function represented by Equation 4 is quadratic in the unknown metric M; hence, it is possible to obtain the following closed-form solution to M Equation 5:

$$M = \frac{\lambda}{\gamma (N - N_c)} \sum_{i,k} \left(DF_c^i - DF_c^k \right) \left(DF_c^i - DF_c^k \right)^T - \frac{1-\lambda}{\gamma (N_c - 1)} \sum_{i,k} \left(DF_c^i - DF_c^j \right) \left(DF_c^i - DF_c^j \right)^T \tag{5}$$

The update in Equation 5 does not guarantee that M is positive semi-definite (PSD). To ensure this property, we can apply the following straightforward projection step after the calculation of M:

1. Perform Eigen decomposition: $M = U^T \Lambda U$
2. Generate $\Lambda_+ = \max(0, \Lambda)$, i.e. select positive eigenvalues
3. Reconstruct the metric M: $M = U^T \Lambda_+ U$

This projected metric is used in the classification algorithm. The metric learned from this Push-Pull methodology is used in conjunction with a standard k-NN classifier. The k-NN algorithm estimates a classification label based on the "closest" samples provided in training (Altman 1992), where $\{x_n\}$ is a set of training data n big. The distance between a new pattern x_i and each pattern in the training set is found. The new pattern is classified depending on the majority of the closest k class labels. Here, the distance between patterns is given in Equation 3, using the learned metric M.

2.4. *Comparative Methodologies*

The pairing of DF feature space and Push-Pull matrix metric learning represents a novel design, thus it is difficult to draw conclusions about performance of the design as there are no similar studies which have: trained on this particular dataset, targeted this particular variable type, used this feature space, or used this classifier. Presented then are two additional classification methodologies which implement more traditional and well understood features and classifiers : k-NN applied to Phased Light Curves (Method A) and k-Means representation with Quadratic Discriminant Analysis (QDA) (Method B). These are shown to allow for comparison of design, likewise the associated testing errors are shown for comparison of performance. Method A is similar to the UCR (Chen et al. 2015) Time Series data baseline algorithm, reported as part of the database. Provided here is a direct k-NN classification algorithm applied directly to the smoothed, aligned, regularly sampled phased light curve. This regular sampling is generated via interpolation of the smoothed dataset, and is required because of the nature of the Nearest Neighbor algorithm requiring one-to-one distance. Standard procedures can then be followed Hastie et al. (2009).

While Method A uses neither the DF feature representation nor the Metric learning methodology, Method B uses DF feature space but not the Metric learning methodology. This presents a problem however, as most standard out of the box classification methods require a vector input. Indeed many methodologies, even when faced with a matrix input, choose to vectorize the matrix allowing an alignment of requirements. An alternative to this implementation is a secondary transformation into a lower dimensional feature space. Following the work of Park et al. (2003), a matrix distance k-Means algorithm is implemented to, unsupervised, generated estimates of "clusters" in the DF space. The observations are transformed by finding the Euclidean distance ($M = \mathbb{I}$) between each training set and each of the k-mean matrices "discovered". The resulting set of k-distances are treated as the input patterns, allowing the use of the standard QDA algorithm Duda et al. (2012). Note, the introduction of k-Means results in variation in implementation depending on the means found for a given run and for Method B a mean error over 50 runs and across the cross-validation process is presented. The performance of both the proposed methodology and the two comparative methodologies is presented in Table 2. Both algorithms are available as open source code, along with our novel implementation, at the project repository.

3. RESULTS

As a demonstration of design, the proposed algorithm is applied to a set of pre-defined eclipsing binary light curves. Using the Eclipsing Binary Catalog (Kirk et al. 2016), a set of 30 targets of interest and 121 target of non-interest were identified via by hand expert analysis. This set of 151 light curves is used for training and testing.

3.1. *Training, Cross-Validation and Testing*

While the initial 151 light curves are consistent for the implementation, the algorithm provided for randomized selection of training data, testing data, and cross-validation segmentation. The algorithm implements 5-fold cross-validation (Duda et al. 2012), the function **Generate_5FoldCrossVal** generates index lists the algorithm will use in the cross-validation process. The algorithmic details are likely beyond the scope of this paper, however as implemented the algorithm splits each class in the labeled data in half, with one half used in training and the other in testing. The training data is further subdivided into five partitions; as the algorithm is trained these partitioned are used by the function **PullTrainingAndCrossFromStruct** to generate a training set using four of the five partitions and a cross-validation set using the fifth. The cross-validation process is the rotation of which partition is used for training and

which as the cross-validation set. The misclassification error for each of the five iterations is averaged together to generate a best estimate. The minimization of this misclassification rate is used to optimize floating parameters in the design such as the number of x-bins, the number of y-bins, and k-values. Some parameters are more sensitive than others; often this insensitivity is related to the loss function or the feature space, or event the data itself (variability of similarity). For example, it was found that the γ and λ values weakly affected the optimization, while the bin sizes and the k-values had a stronger effect (to some degree). The set of optimized parameters is given in Table 1.

Table 1. The DF Push-Pull optimized parameters based on expert labeled data and cross-validation

Parameter	Value
γ	1.0
λ	0.75
$y-bins$	35
$x-bins$	25
k	3

Given the optimization of these floating variables in all three algorithms, the testing data is then applied to the optimal designs; note this is data that was not included in the initial training of the algorithms.

Table 2. A comparison of performance estimates across the proposed classifiers

	PPM	Method A	Method B
Misclass. Rate	12.5%	15.6%	12.7%

The performance of the main novel feature space/classification pairing as well as the two additional implementations that rely on more standard methods is presented. The method proposed has a marginally better misclassification rate (Table 2), and has the added benefit of (1) not requiring unsupervised clustering which can be inconsistent and (2) provides nearest neighbor estimates allowing for demonstration of direct comparison. Note, these performance estimate values are dependent on the initial selected training and testing data. They have been averaged and optimized via cross-validation, however with so little initial training data and with the selection process for which data are training and which are testing randomized, performance estimates may vary. Of course increases in training data will result in increased confidence in performance results.

3.2. *Application to Unlabeled Data*

The data are now applied those data in the initial Eclipsing Binary catalog that were not identified as either "Of Interest" or "Not Of Interest". The trained and tested dataset is combined into a single training set for application, the primary method (push-pull metric classification) is used to optimize a metric based on the optimal parameters found during cross-validation, and apply the system to the entire Kepler Eclipsing Binary data (2875 curves). Based on the results demonstrated in Johnston & Oluseyi (2017), the algorithm additionally conditions the detection process based on a maximal distance allowed between a new unlabeled point and the training dataset in the feature space of interest. This further restricts the algorithm to classify those targets which exist only in "known space". The k-NN algorithm generates a distance, dependent on the optimized metric; this limitation is the equivalent of generating an anomaly detection algorithm. Once the discovered targets that were also in the initial training data were removed, the result is a conservative selection of 124 potential targets of interest listed in the supplementary digital file KeplerTraining.xlsx at the project repository[2]. An initial exploratory data analysis performed on the phased light curve data is presented. At a high-level, the mean and standard deviation of the discovered curves is presented in Figure 3.

[2] https://github.com/kjohnston82/OCDetector/supplement/KeplerTraining.xlsx

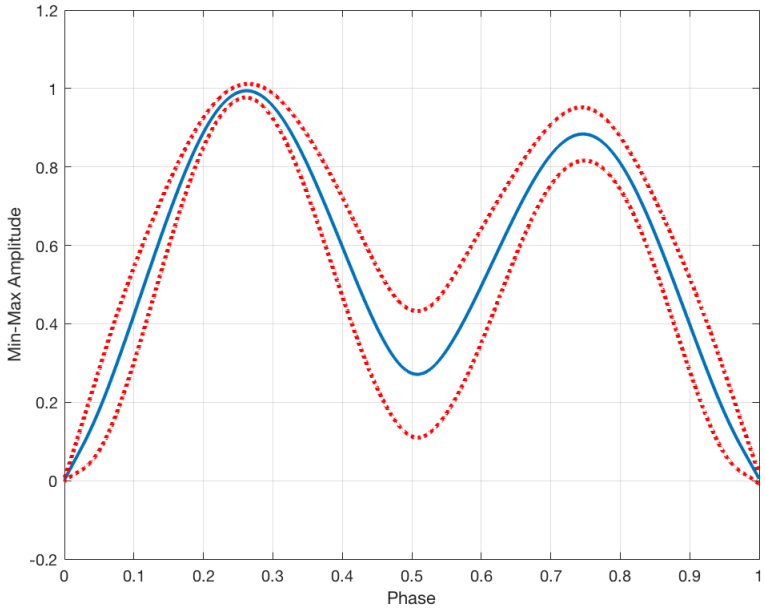

Figure 3. The mean and standard deviation of the distribution of O'Connell Effect Eclipsing Binary phased light curves discovered via the proposed detector out of the Kepler Eclipsing Binary Catalog.

While a more in depth analysis as to the meaning of the distribution functional shapes is left for future study, it is noted that in general there are some morphological consistencies across the discovered targets.

1. In the majority of the discovered OEEB systems the first maxima is greater than the second.
2. The light-curve relative functional shape from the primary minima to primary maxima is fairly consistent across all discovered systems.
3. The difference in amplitude between the two maxima does not appear to be consistent, nor is the difference in amplitude between the minima.

The discovered group is partitioned into sub-groupings via unsupervised clustering. The k-Means algorithm presented as part of the comparative methodologies is applied to the discovered dataset with a predefined cluster number. Clusters are numbered one through eight, the resulting 1-D curve generated by the supersmoother algorithm are presented in their respective clusters in Figure 4, the top four plots represent clusters 1 to 4 (left to right) and the bottom four plots represent clusters 5 to 8 (left to right).

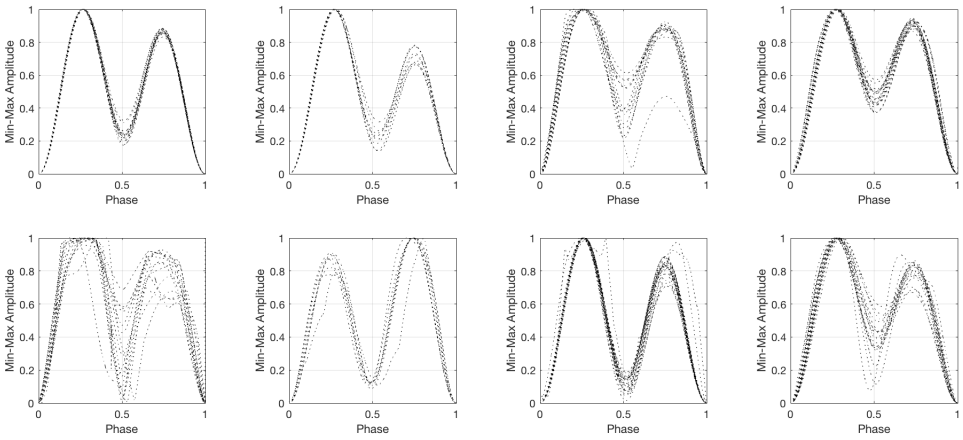

Figure 4. The phased light curves of the discovered OEEB data from Kepler, clustered via k-Mean applied to the DF feature space.

The clusters generated were initialized with random starts, thus additional iterations can potentially result in different groupings. For further analysis of the sub-clusters, a set of the four metrics identified by McCartney (1999) as descriptors of the O'Connell Effect are presented in graphical and tabular form: the O'Connell Effect Ratio (OER), the difference in maximum amplitudes (Δm) , the difference in the minimum amplitudes, and the light curve asymmetry (LCA). The metrics are based on the smoothed phased light curve curves. The O'Connell Effect Ratio (OER) is estimated as Equation 6:

$$OER = \frac{\sum_{i=1}^{n/2} (I_i - I_1)}{\sum_{i=n/2+1}^{n} (I_i - I_1)} \tag{6}$$

where the Min-Max amplitude measurements for each star are grouped into phase bins ($n = 500$) the mean amplitude in each bin is I_i. An $OER > 1$ corresponds to the front half of the light curve having more total flux, note that for the procedure presented here $I_1 = 0$. The difference in max amplitude is estimated as Equation 7:

$$\Delta m = \max_{t<0.5} (f(t)') - \max_{t\geq 0.5} (f(t)') \tag{7}$$

where we have estimated the max in each half of the phased light curve. The Light Curve Asymmetry is estimated as the Equation 8:

$$LCA = \sqrt{\sum_{i=1}^{n/2} \frac{\left(I_i - I_{(n+1-i)}\right)^2}{I_i^2}} \tag{8}$$

As opposed to the measurement of OER, LCA measures the deviance from symmetry of the two peaks. The metrics estimated, as well as the cluster the stars were partitioned into, is presented in a supplementary digital file AnalysisOfClusters.xlsx at the project repository[3]. A plot of the measured metrics as well as estimated values of period and temperature (as reported by the Kepler database), with respect to the cluster assigned by k-Means, are presented in Figure 5.

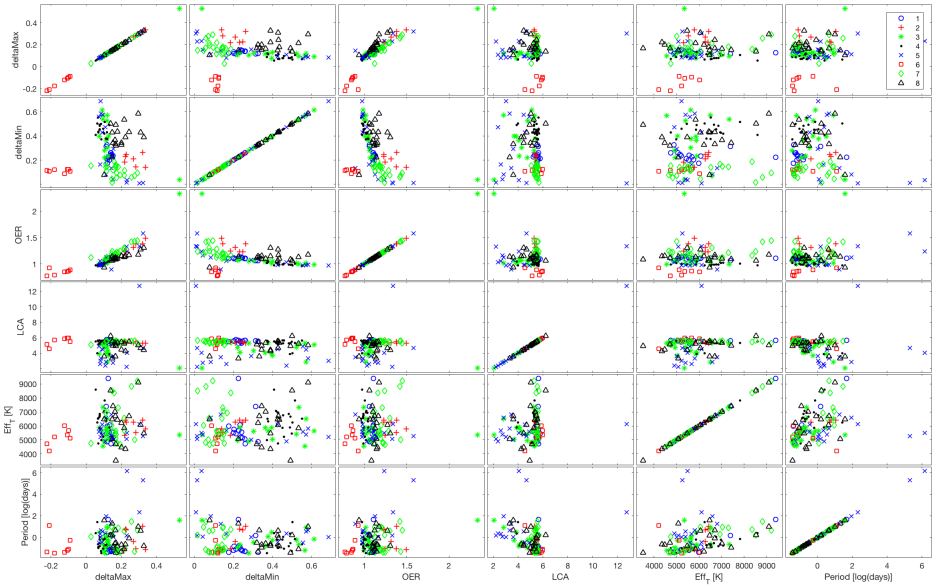

Figure 5. Covariance Plot of OER, LCA, maximum amplitude difference, minimum amplitude difference, Period, and Effective Temperature for the discovered OEEB in the Kepler Eclipsing Binary Dataset.

Following figure 4.6 in McCartney (1999), plot of OER vs. Δm is isolated and presented in Figure 6:

[3] https://github.com/kjohnston82/OCDetector/supplement/AnalysisOfClusters.xlsx

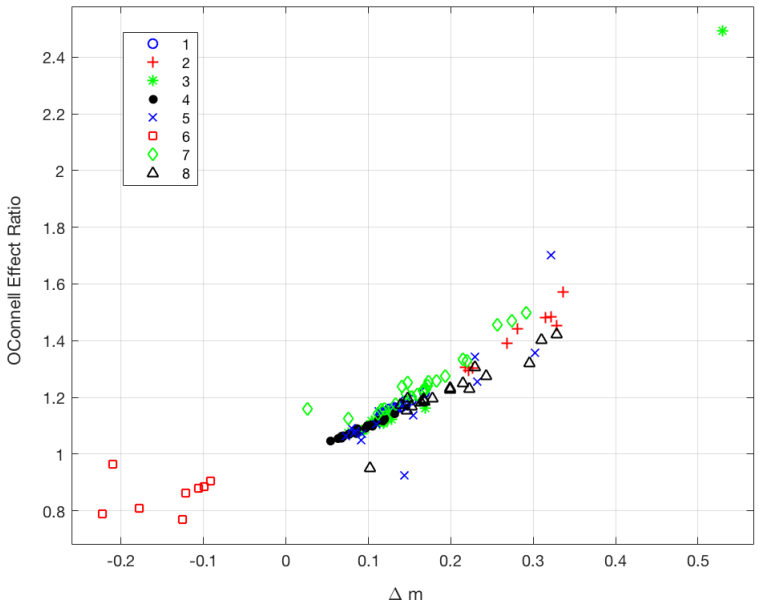

Figure 6. OER vs. Δm for discovered Kepler O'Connell Effect Eclipsing Binaries.

the linear relationship between OER and Δm reported in McCartney (1999) is apparent in the discovered Kepler data as well. The dataset here extends from $OER \sim (0.7, 1.8)$ and $\Delta m \sim (-0.3, 0.4)$, not including the one sample from cluster 3 which is extreme. This is comparable to the reported range in McCartney (1999) of $OER \sim (0.8, 1.2)$ and $\Delta m \sim (-0.1, 0.05)$, similar OER range but our Kepler data spans a much larger Δm domain, likely resulting from our additional application of Min-Max amplitude scaling. The gap in Δm between -0.08 and 0.02 is caused by the bias in our training sample and algorithm goal, only O'Connell Effect Binaries with a user discernible Δm. The clusters identified by the k-Mean algorithm applied to the DF feature space roughly correspond to groupings in the OER/Δm feature space (clustering along the diagonal). The individual cluster statistics (mean and relative error) with respect to the metrics measured here, are given in Table 3:

Table 3. Example O'Connell Effect metric measurements

Cluster	Δm	$\sigma_{\Delta m}/\Delta m$	OER	σ_{OER}/OER	LCA	σ_{LCA}/LCA	#
1	0.13	0.11	1.16	0.02	7.62	0.25	17
2	0.28	0.17	1.41	0.07	8.92	0.16	9
3	0.14	0.78	1.20	0.30	7.13	0.25	15
4	0.09	0.24	1.08	0.02	6.95	0.23	22
5	0.15	0.55	1.17	0.16	8.54	0.58	15
6	-0.14	-0.36	0.86	0.08	8.36	0.19	8
7	0.17	0.36	1.25	0.08	9.41	0.82	24
8	0.20	0.31	1.22	0.08	8.03	0.36	19

All of the clusters have a positive mean Δm save for cluster 6. The morphological consistency within a cluster is visually apparent in figure 4, but also apparent in the relative error of LCA with cluster 5 and 7 being the least consistent. The data discovered will be dependent on the input training data, and as such similar trends in the expert labeled data can be observed. The next step will include applications to other surveys.

3.3. *Cross-Survey Application*

Further demonstration of the algorithm is presented with an application to a separate independent survey. Machine learning methods have been applied to the classification of variable stars observed by the LINEAR survey (Sesar et al. 2011), and while these methods have focused on leveraging Fourier domain coefficients and photometric measurements

$\{u, g, r, i, z\}$ from SDSS, the data also includes best estimates of period as all of the variable stars trained on had cyclostationary signatures. It is then trivial to extract the phased light curve for each star, and since the DF feature is indifferent to sampling density so long as all points along the functional shape are represented, the trained detection algorithm generated and demonstrated above can be directly applied to the LINEAR data. The result is the identification of the potential targets of interest, all initially identified as eclipsing binaries. The discovered target aligned, smoothed, light curves are presented in Figure 7, note the LINEAR IDs are presented in the supplementary digital file LINEARDiscovered.xlsx at the project repository[4]:

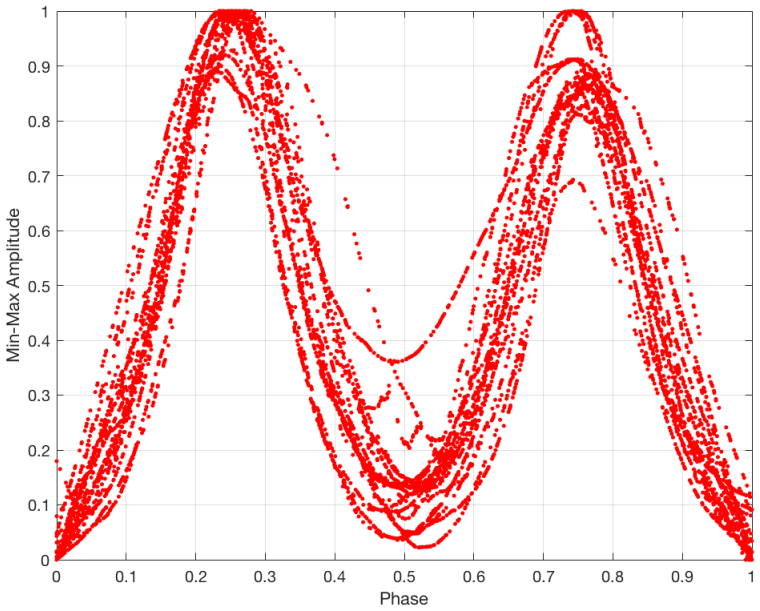

Figure 7. The set of discovered LINEAR targets that demonstrate the OEEB signature.

Unlike the Kepler Eclipsing Binary Catalog, the LINEAR dataset contains targets other than (but does include) eclipsing binaries; the light curves are also much more poorly sampled. Thus, the poor uncertainty in the functional shape results from lower SNR (ground survey) and poor sampling. Similar to the Kepler discovered dataset, we plot $OER/\Delta m$ features using lower resolution phased binnings ($n = 20$), and see that the distribution and relationship from McCartney (1999) hold here as well (see Figure 8):

[4] https://github.com/kjohnston82/OCDetector/supplement/LINEARDiscovered.xlsx

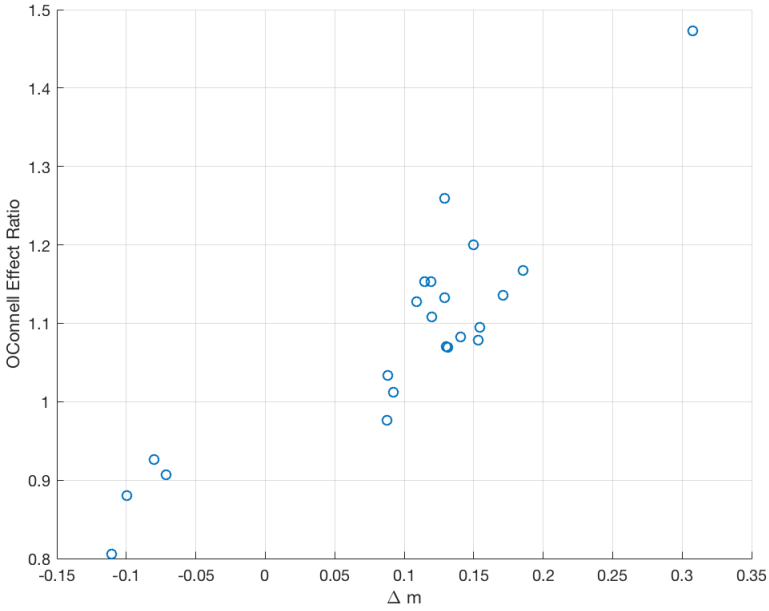

Figure 8. OER vs. Δm for the discovered OEEB in the LINEAR dataset.

4. DISCUSSION

This design is modular enough to be applied, as is, to other types of stars and star systems that are cyclostationary in nature. With a change in feature space, specifically one that is tailored to the target signatures of interest, this design can be replicated for other targets that do not demonstrate a cyclostationary single (i.e. impulsive, non-stationary, etc.), and even to targets of interest which are not time-variable in nature but have a consistant observable signature (e.g. spectrum, photometry, image point-spread function etc.). One of the advantages of attempting to identify the O'Connell Effect Eclipsing Binary is that one only needs the phased light curve to make a classification. The Distribution Field process here allows for a direct transformation into a singular feature space that focuses on functional shape. For other variable stars, a multi-view approach might be necessary; either descriptions of the light curve signal across multiple transformations (e.g. Wavelet and DF), or across representations (e.g. polarimetry and photometry), or across frequency regimes (e.g. optical and RF) would be required in the process of properly defining the variable star type. The solution to this multi-view problem is neither straightforward nor is it well understood (Akaho 2006). Multiple options have been explored to resolve this problem: combination of classifiers, canonical correlation analysis, post-probability blending, and multi-metric classification. The computational needs of the algorithm have only been roughly studied, and a more through review is necessary in the context of the algorithm proposed and the needs of the astronomy community. The k-NN algorithm dependence on pair wise difference, while one of its' strong suits, is also one of the more computationally demanding parts of the algorithm. However, functionality such as $k-d$ trees as well as other feature space partitioning method have been shown to reduce the computational requirements.

The method outlined here has demonstrated the ability to detect targets of interest given a training set consisting of expertly labeled light curve training data. The procedure presents two new functionalities: the Distribution Field, a shape based feature space and the Push-Pull Matrix Metric Learning algorithm, a metric learning algorithm derived from LMNN that allows for matrix-variate similarity comparisons. A comparison to less novel more standard methods was demonstrated on a Kepler Eclipsing Binary sub-dataset that was labelled by an expert in the field of O'Connell Effect binary star systems. The performance of the three methods is presented, the methodology proposed (DF + Push-Pull Metric Learning) is comparable or out performs the other methods. As a demonstration, the design is applied to Kepler Eclipsing Binary data and LINEAR data; discovered systems are found in each dataset and reported. Furthermore, the increase in the number of systems, and the presentation of the data, allow us to make additional observations about the distribution of curves and trends within the population. Future work will involve the analysis of these statistical distribution, as well as inference as to their physical meaning.

The authors are grateful for the valuable machine learning discussion with S. Wiechecki-Vergara. The authors are grateful the valuable astrophysics insight provided by C. Fletcher and T. Doyle. Initial editing and review provided by

G. Langhenry. Research was partially supported by Vencore, Inc. The LINEAR program is sponsored by the National Aeronautics and Space Administration (NRA Nos. NNH09ZDA001N, 09-NEOO09-0010) and the United States Air Force under Air Force Contract FA8721-05-C-0002. The authors would like to additionally acknowledge the Kepler Eclipsing Binary team without whom much of the training data would not exist.

REFERENCES

Akaho, S. 2006, arXiv preprint cs/0609071
Altman, N. S. 1992, The American Statistician, 46, 175
Angeloni, R., Ramos, R. C., Catelan, M., et al. 2014, Astronomy & Astrophysics, 567, A100
Barclay, T., Ramsay, G., Hakala, P., et al. 2011, Monthly Notices of the Royal Astronomical Society, 413, 2696
Bellet, A., Habrard, A., & Sebban, M. 2015, Synthesis Lectures on Artificial Intelligence and Machine Learning, 9, 1
Beygelzimer, A., Langford, J., & Zadrozny, B. 2005, in AAAI, 720–725
Blomme, J., Sarro, L., O'Donovan, F., et al. 2011, Monthly Notices of the Royal Astronomical Society, 418, 96
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, The Astrophysical Journal, 712, 511
Chandola, V., Banerjee, A., & Kumar, V. 2009, ACM Computing Surveys (CSUR), 41, 15
Chen, Y., Keogh, E., Hu, B., et al. 2015, The UCR Time Series Classification Archive, , , www.cs.ucr.edu/~eamonn/time_series_data/
Debosscher, J. 2009, status: published
Ding, S., & Dennis Cook, R. 2018, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80, 387
Djorgovski, S. G., Mahabal, A., Donalek, C., et al. 2012, in E-Science (e-Science), 2012 IEEE 8th International Conference on, IEEE, 1–8
Dubath, P., Rimoldini, L., Süveges, M., et al. 2011, Monthly Notices of the Royal Astronomical Society, 414, 2602
Duda, R. O., Hart, P. E., & Stork, D. G. 2012, Pattern classification (John Wiley & Sons)
Eyer, L., & Blake, C. 2005, Monthly Notices of the Royal Astronomical Society, 358, 30
Friedman, J. H., & Silverman, B. W. 1989, Technometrics, 31, 3
Gagniuc, P. A. 2017, Markov Chains: From Theory to Implementation and Experimentation (John Wiley & Sons)
Graham, M. J., Djorgovski, S., Mahabal, A. A., Donalek, C., & Drake, A. J. 2013a, Monthly Notices of the Royal Astronomical Society, 431, 2371
Graham, M. J., Drake, A. J., Djorgovski, S., et al. 2013b, Monthly Notices of the Royal Astronomical Society, 434, 3423
Haber, R., Rangarajan, A., & Peter, A. M. 2015, in Machine Learning and Knowledge Discovery in Databases (Cham: Springer International Publishing), 20–36
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The elements of statistical learning (Springer)
Helfer, E., Smith, B. Haber, R., & Peter, A. 2015, Statistical Analysis of Functional Data, Technical Report TR-2015-05,, Tech. rep., Florida Institute of Technology
Iverson, K. E. 1962, in Proceedings of the May 1-3, 1962, spring joint computer conference, ACM, 345–351
Johnston, K. B., & Oluseyi, H. M. 2017, New Astronomy, 52, 35
Johnston, K. B., & Peter, A. M. 2017, New Astronomy, 50, 1
Kirk, B., Conroy, K., Prša, A., et al. 2016, The Astronomical Journal, 151, 68
Masci, F. J., Hoffman, D. I., Grillmair, C. J., & Cutri, R. M. 2014, The Astronomical Journal, 148, 21
Matthew F. Knote, Ronald H. Katchuck, R. C. B. submitted
McCartney, S. 1999, PhD thesis, University of Oklahoma Graduate College
O'Connell, D. 1951, Monthly Notices of the Royal Astronomical Society, 111, 642
Palaversa, L., Ivezić, Ž., Eyer, L., et al. 2013, The Astronomical Journal, 146, 101
Park, H., Jeon, M., & Rosen, J. B. 2003, BIT Numerical mathematics, 43, 427
Park, M. J., & Cho, S. S. 2013, CSAM (Communications for Statistical Applications and Methods), 20, 271
Pichara, K., & Protopapas, P. 2013, The Astrophysical Journal, 777, 83
Pichara, K., Protopapas, P., Kim, D.-W., Marquette, J.-B., & Tisserand, P. 2012, Monthly Notices of the Royal Astronomical Society, 427, 1284
Pojmanski, G. 2000, Acta Astronomica, 50, 177
—. 2002, Acta Astronomica, 52, 397
Rebbapragada, U., Lo, K., Wagstaff, K. L., Murphy, T., & Thompson, D. R. 2011, Proceedings of the International Astronomical Union, 7, 397
Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, The Astrophysical Journal Supplement Series, 203, 32
Rifkin, R., & Klautau, A. 2004, The Journal of Machine Learning Research, 5, 101
Sesar, B., Stuart, J. S., Ivezi, Ž., et al. 2011, The Astronomical Journal, 142, 190
Sevilla-Lara, L., & Learned-Miller, E. 2012, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 1910–1917
Udalski, A., Soszynski, I., Szymanski, M., et al. 1999, Acta Astronomica, 49, 437
Weinberger, K. Q., Blitzer, J., & Saul, L. K. 2006, in Advances in neural information processing systems, 1473–1480
Wilsey, N. J., & Beaky, M. M. 2009, in Society for Astronomical Sciences Annual Symposium, Vol. 28, 107
Wilson, R. E., & Devinney, E. J. 1971, The Astrophysical Journal, 166, 605