Johnston et al. *Computational Astrophysics and Cosmology* (2019) 6:4
https://doi.org/10.1186/s40668-019-0031-2

Computational Astrophysics and Cosmology
a SpringerOpen Journal

RESEARCH **Open Access**

A detection metric designed for O'Connell effect eclipsing binaries

Kyle B. Johnston[1,5*], Rana Haber[2], Saida M. Caballero-Nieves[1], Adrian M. Peter[3], Véronique Petit[4] and Matt Knote[1]

Abstract

We present the construction of a novel time-domain signature extraction methodology and the development of a supporting supervised pattern detection algorithm. We focus on the targeted identification of eclipsing binaries that demonstrate a feature known as the O'Connell effect. Our proposed methodology maps stellar variable observations to a new representation known as distribution fields (DFs). Given this novel representation, we develop a metric learning technique directly on the DF space that is 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. The result is a conservative selection of 124 potential targets of interest out of the Villanova Eclipsing Binary Catalog. Our framework demonstrates favorable performance on Kepler eclipsing binary data, taking a crucial step in preparing the way for large-scale data volumes from next-generation telescopes such as LSST and SKA.

Keywords: Machine learning; Astroinformatics; Eclipsing binaries

1 Introduction

With the rise of large-scale surveys, such as Kepler, the Transiting Exoplanet Survey Satellite (TESS), the Kilodegree Extremely Little Telescope (KELT), the Square Kilometre Array, and the Large Synoptic Survey Telescope (LSST), a fundamental working knowledge of statistical data analysis and data management to reasonably process astronomical data is necessary. The ability to mine these data sets for new and interesting astronomical information opens a number of scientific windows that were once closed by poor sampling, in terms of both number of stars (targets) and depth of observations (number of samples).

This article focuses on the development of a novel, modular time-domain signature extraction methodology and its supporting supervised pattern detection algorithm for variable star detection. The design could apply to any number of variable star types that exhibit consistent periodicity (cyclostationary) in their flux; examples include most Cepheid-type stars (RR Lyr, SX Phe, Gamma Dor, etc...) as well as other eclipsing binary types. Nonperiodic variables would require a different feature space (Johnston and Peter 2017), but the underlying detection scheme could still be relevant. Herein we present the design's utility, by its targeting of eclipsing binaries that demonstrate a feature known as the O'Connell effect.

We have selected O'Connell effect eclipsing binaries (OEEBs) to demonstrate initially our detector design. We highlight OEEBs here because they compose a subclass of a specific type of variable star (eclipsing binaries). Subclass detection provides an extra layer of complexity for our detector to try to handle. We demonstrate our detector design on Kepler eclipsing binary data from the Villanova catalog, allowing us to train and test against different subclasses in the same parent variable class type. We train our detector design on Kepler eclipsing binary data and apply the detector to a different survey—the Lincoln Near-Earth Asteroid Research asteroid survey (LINEAR, Stokes et al. 2000)—to demonstrate the algorithm's ability to discriminate and detect our targeted subclass given not just the parent class but other classes as well.

Classifying variable stars relies on proper selection of feature spaces of interest and a classification framework that can support the linear separation of those features.

*Correspondence: kyjohnst2000@my.fit.edu
[1] Aerospace, Physics and Space Sciences Department, Florida Institute of Technology, Melbourne, USA
[5] Defense Group Melbourne, Perspecta Inc., Melbourne, USA
Full list of author information is available at the end of the article

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Selected features should quantify the telltale signature of the variability—the structure and information content. Prior studies to develop both features and classifiers include expert selected feature efforts (Debosscher 2009; Sesar et al. 2011; Richards et al. 2012; Graham et al. 2013; Armstrong et al. 2016; Mahabal et al. 2017; Hinners et al. 2018), automated feature selection efforts (McWhirter et al. 2017; Naul et al. 2018), and unsupervised methods for feature extraction (Valenzuela and Pichara 2018; Modak et al. 2018). The astroinformatics community-standard features include quantification of statistics associated with the time-domain photometric data, Fourier decomposition of the data, and color information in both the optical and IR domains (Nun et al. 2015; Miller et al. 2015). The number of individual features commonly used is upward of 60 and growing (Richards et al. 2011) as the number of variable star types increases, and as a result of further refinement of classification definitions (Samus' et al. 2017). We seek here to develop a novel feature space that captures the signature of interest for the targeted variable star type.

The detection framework here maps time-domain stellar variable observations to an alternate distribution field (DF) representation (Sevilla-Lara and Learned-Miller 2012) and then develops a metric learning approach to identify OEEBs. Based on the matrix-valued DF feature, we adopt a metric learning framework to directly learn a distance metric (Bellet et al. 2015) on the space of DFs. We can then utilize the learned metric as a measure of similarity to detect new OEEBs based on their closeness to other OEEBs. We present our metric learning approach as a competitive push–pull optimization, where DFs corresponding to known 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.

This article is structured as follows. First, we review the targeted stellar variable type, discussing the type signatures expected. Second, we review the data used in our training, testing, and discovery process as part of our demonstration of design. Next, we outline the novel proposed pipeline for OEEB detection; this review includes the feature space used, the designed detector/classifier, and the associated implementation of an anomaly detection algorithm (Chandola et al. 2009). Then, we apply the algorithm, trained on the expertly selected/labeled Villanova Eclipsing Binary catalog OEEB targets, to the rest of the catalog with the purpose of identifying new OEEB stars. We present the results of the discovery process using a mix of clustering and derived statistics. We apply the Villanova Eclipsing Binary catalog trained classifier, without additional training, to the LINEAR data set. We provide results of this cross-application, i.e., the set of discovered OEEBs. For comparison, we detail two competing approaches. We develop training and testing strategies for our metric learning framework, and finally, we conclude with a summary of our findings and directions for future research.

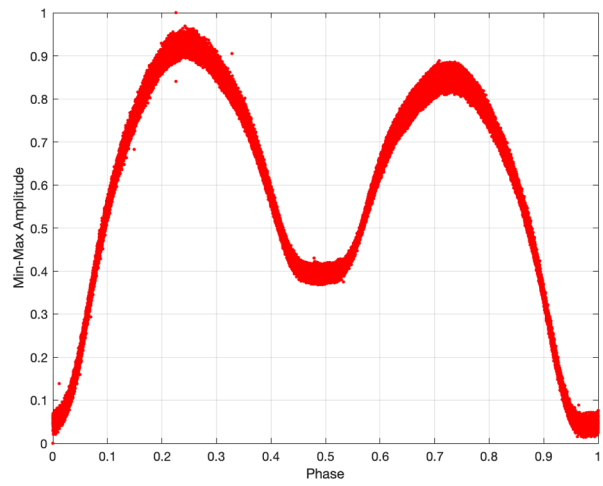

Figure 1 An example phased light curve of an eclipsing binary with the O'Connell effect (KIC: 10861842). The light curve has been phased such that the global minimum (cooler in front of hotter) is at lag 0 and the secondary minimum (hotter in front of cooler) is at approximately lag 0.5. The side-by-side binary orientations are at approximately 0.25 and 0.75. Note that the maxima, corresponding to the side-by-side orientations, have different values

2 Eclipsing binaries with 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 (see Fig. 1). This maxima asymmetry is unexpected, as it suggests an orientation dependency in the brightness of the system. Similarly, the consistency of the asymmetric over many orbits is also surprising, as it suggests that the maxima asymmetry has a dependence on the rotation of the binary system. The cause of the O'Connell effect is not fully understood and additional data and modeling are necessary for further investigation (McCartney 1999; Knote 2019). Our focus in this work is in the application of an automated detector to OEEBs to identify systems of interest for future work.

2.1 Signatures and theories

Several theories propose to explain the effect, including starspots, gas stream impact, and circumstellar matter (McCartney 1999). The work by Wilsey and 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 a consistent decrease in brightness of the star when viewed as a point source. While magnetic surface activity will cause both flares (brightening) and spots

(darkening), flares tend to be transient, whereas spots tend to have longer-term effects on the observed binary flux. Thus, between the two, starspots are the favored hypothesis for causing long-term consistent asymmetry; often binary simulations (such as the Wilson–Devinney code) can be used to model O'Connell effect binaries via including an often large starspot (Zboril and Djurasevic 2006).
- 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 proposes 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 in orbit could result in attenuation, i.e., the difference in maximum magnitude of the phased light curve results from dimming and not brightening.

In the study (McCartney 1999), the authors limited the sample to only six star systems: GSC 03751-00178, V573 Lyrae, V1038 Herculis, ZZ Pegasus, V1901 Cygni, and UV Monocerotis. Researchers have used standard eclipsing binary simulations (Wilson and Devinney 1971) to demonstrate the proposed explanations for each light curve instance and estimate the parameters associated with the physics of the system. Wilsey and Beaky (2009) noted other cases of the O'Connell effect in binaries, which have since been described physically; in some cases, the effect varied over time, whereas in other cases, the effect was consistent over years of observation and over many orbits. The effect has been found in both overcontact, semidetached, and near-contact systems.

While one of the key visual differentiators of the O'Connell effect is $\Delta m_{\max}$, this heuristic feature alone could not be used as a general mean for detection, as the targets trained on or applied to are not guaranteed to be (a) eclipsing binaries and (b) periodic. One of the goals we are attempting to highlight is the transformation of expert qualitative target selection into quantitative machine learning methods.

2.2 Characterization of OEEB

We develop a detection methodology for a specific target of interest—OEEB—defined as an eclipsing binary where the light curve (LC) maxima are consistently at different amplitudes over the span of observation. Beyond differences in maxima, and a number of published examples, little is defined as a requirement for identifying the O'Connell effect (Wilsey and Beaky 2009; Knote 2019).

McCartney (1999) provide some basic indicators/measurements of interest in relation to OEEB binaries: the O'Connell effect ratio (OER), the difference in maximum amplitudes (Δm), the difference in minimum amplitudes, and the light curve asymmetry (LCA). The metrics are based on the smoothed phased light curves. The OER is calculated as Equation (1):

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

where the min-max amplitude (i.e. normalized flux) measurements for each star are grouped into phase bins (n = 500), where 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 we present here, $I_1 = 0$. The difference in max amplitude is calculated as Equation (2):

$$\Delta m = \max_{t<0.5}\big(f(t)_N\big) - \max_{t\geq 0.5}\big(f(t)_N\big), \tag{2}$$

where we have estimated the maximum in each half of the phased light curve. The LCA is calculated as Equation (3):

$$\mathrm{LCA} = \sqrt{\sum_{i=1}^{n/2}\frac{(I_i - I_{(n+1-i)})^2}{I_i^2}}. \tag{3}$$

As opposed to the measurement of OER, LCA measures the deviance from symmetry of the two peaks. Defining descriptive metrics or functional relationships (i.e., bounds of distribution) requires a larger sample than is presently available. An increased number of identified targets of interest is required to provide the sample size needed for a complete statistical description of the O'Connell effect. The quantification of these functional statistics allows for the improved understanding of not just the standard definition of the targeted variable star but also the population distribution as a whole. These estimates allow for empirical statements to be made regarding the differences in light curve shapes among the variable star types investigated. The determination of an empirically observed distribution, however, requires a significant sample to generate meaningful descriptive statistics for the various metrics.

In this effort, we highlight the functional shape of the phased light curve as our defining feature of OEEB stars. The prior metrics identified are selected or reduced measures of this functional shape. We propose here that, as opposed to training a detector on the preceding indicators, we use the functional shape of the phased light curve by way of the distribution field to construct our automated system.

3 Variable star data

As a demonstration of design, we apply the proposed algorithm to a set of predefined, expertly labeled eclipsing

binary light curves. We focus on two surveys of interest: first, the Kepler Villanova Eclipsing Binary catalog, from which we derive our initial training data as well as our initial discovery (unlabeled) data, and second, the Lincoln Near-Earth Asteroid Research, which we treat as unlabeled data.

3.1 Kepler Villanova eclipsing catalog

Leveraging the Kepler pipeline already in place, and using the data from the Villanova Eclipsing Binary catalog (Kirk et al. 2016), this study focuses on a set of predetermined eclipsing binaries identified from the Kepler catalog. From this catalog, we developed an initial, expertly derived, labeled data set of proposed targets "of interest" identified as OEEB. Likewise, we generated a set of targets identified as "not of interest" based on our expert definitions, i.e., intuitive inference.

We have labeled our two populations "of interest" and "not of interest" to represent those targets in the initial training set that a user has found to either be interesting to their research (identified OEEBs) or otherwise. The labeling "of interest" here is specifically used, as we are dependent on the expert selections.

Using the Eclipsing Binary catalog (Kirk et al. 2016), we identified a set of 30 targets "of interest" (see Table 1) and 121 targets of "not of interest" (see Table 2) via expert analysis—by-eye selection based on researchers' interests. Specific target identification is listed in a supplementary digital file at the project repository.[a] We use this set of 151 light curves for training and testing.

3.1.1 *Light curve/feature space*

Prior to feature space processing, the raw observed photometric time domain data are conditioned and processed. Operations include long-term trend removal, artifact removal, initial light curve phasing, and initial eclipsing binary identification; we performed these actions prior to the effort demonstrated here, by the Eclipsing Binary catalog (our work uses all 2875 long-cadence light curves available as of the date of publication as training/testing data, or as unlabeled data to search for new OEEBs). The functional shape of the phased light curve is selected as the feature to be used in the machine learning process, i.e., detection of targets of interest. While the data have been conditioned already by the Kepler pipeline, added steps are taken to allow for similarity estimation between phased curves. Friedman's SUPERSMOOTHER algorithm (Friedman 1984; VanderPlas and Ivezić 2015) is used to generate a smooth 1-D functional curve from the phased light curve data. The smoothed curves are transformed via min-max scaling (4):

$$f(\phi)_N = \frac{f(\phi) - \min(f(\phi))}{\max(f(\phi)) - \min(f(\phi))}, \tag{4}$$

where $f(\phi)$ is the smoothed phased light curve, f is the amplitude from the database source, ϕ is the phase where $\phi \in [0, 1]$, and $f(\phi)_N$ is the min-max scaled amplitude (i.e. normalized flux). Note that we will use the terms $f(\phi)_N$ and min-max amplitude interchangeably throughout this article. We use the minimum of the smoothed phased light curve as a registration marker, and both the smoothed and

Table 1 Collection of KIC of interest (30 total)

10123627	11924311	5123176	8696327	11410485	7696778	7516345	9654476
10815379	2449084	5282464	8822555	7259917	6223646	4350454	9777987
10861842	2858322	5283839	9164694	7433513	9717924	5820209	7584739
11127048	4241946	5357682	9290838	8394040	7199183		

Table 2 Collection of KIC not of interest (121 total)

10007533	10544976	11404698	5560831	7119757	5685072	7335517	5881838
10024144	10711646	11442348	12470530	3954227	10084115	10736223	11444780
10095469	10794878	11652545	2570289	4037163	10216186	10802917	12004834
10253421	10880490	12108333	3127873	4168013	10257903	10920314	12109845
10275747	11076176	12157987	3344427	4544587	10383620	11230837	12216706
10485137	11395117	12218858	3730067	4651526	9007918	8196180	7367833
9151972	8248812	7376500	6191574	4672934	9179806	8285349	7506446
9205993	8294484	7518816	6283224	4999357	9366988	8298344	7671594
9394601	8314879	7707742	6387887	5307780	9532219	8481574	7879399
9639491	8608490	7950964	6431545	5535061	9700154	8690104	8074045
9713664	8758161	8087799	6633929	5606644	9715925	8804824	8097553
9784230	8846978	8155368	7284688	5785551	9837083	8949316	8166095
9935311	8957887	8182360	7339345	5956776	9953894	3339563	4474193
12400729	3832382	12553806	4036687	2996347	4077442	3557421	4554004
6024572	4660997	6213131	4937217	6370361	5296877	6390205	5374999
6467389							

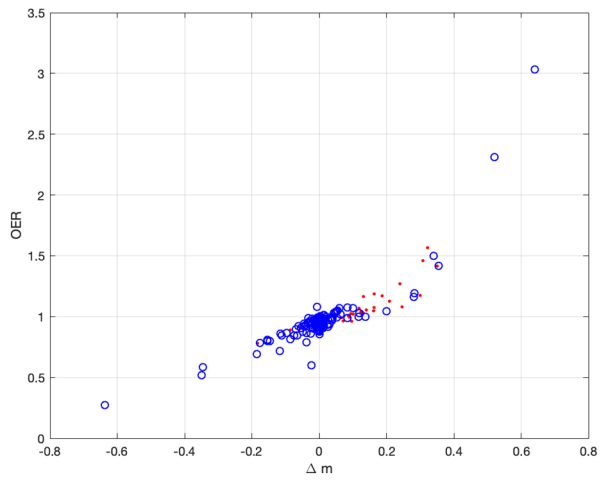

Figure 2 The features Δm and OER for the training data; red dots are the targets are "Of Interest", and blue circles are "Not of Interest"

unsmoothed light curves are aligned such that lag/phase zero corresponds to minimum amplitude (eclipse minima; see McCartney 1999).

3.1.2 Training/testing data

The labeled training data are provided as part of the supplementary digital project repository. We include the SOI and NON-SOI Kepler identifiers here (KIC).

Additionally, we plot the total training and testing dataset in the Δm and OER feature space, to demonstrate the separability of our classes ("Of Interest" vs. "Not of Interest"), see Fig. 2. The values presented here were generated based on phased, unsmoothed, data.

As is apparent from Fig. 2, the data does not exactly separate based on either of the select heuristic measures into the selected categories. For a baseline estimate of performance, we use a simple 1-NN classification algorithm with our selected two heuristic Δm and OER values, using a randomized 50/50 split of our initial Kepler training data (see Sect. 5.1), resulting in a misclassification rate of 23%. This error rate drops to 14% if we use k-NN (with a $k = 5$)

The training data is based on expert requests, with the explicit request that the detector find new observations that are similar to those "of interest" and dissimilar to those identified as "not of interest". These targets were to have a $|\Delta m|$ greater than some threshold, was not to have multiple periods, and was to have a consistent structure in the phased domain. Our objective was to construct a procedure that could find other light curves that fit these user constraints.

3.2 Data untrained

The 2000+ eclipsing binaries left in the Kepler Eclipsing Binary catalog are left as unlabeled targets. We use our described detector to "discover" targets of interest, i.e., OEEB. The full set of Kepler data is accessible via the Villanova Eclipsing Binary website (http://keplerebs.villanova.edu/).

Table 3 Distribution of LINEAR data across classes

Type	Count	Percentage
Algol	287	5.6
Contact Binary	1805	35.6
Delta Scuti	68	1.3
RRab	2189	43.0
RRc	737	14.5

For analyzing the proposed algorithm design, the LINEAR data set is also leveraged as an unknown "unlabeled" data set ripe for OEEB discovery (Sesar et al. 2011; Palaversa et al. 2013). From the starting sample of 7194 LINEAR variables, we used a clean sample of 6146 time series data sets for detection. Stellar class type is limited further to the top five most populous classes—RR Lyr (ab), RR Lyr (c), Delta Scuti / SX Phe, Contact Binaries, and Algol-Like Stars with two minima—resulting in a set of 5,086 observations.

Unlike the Kepler Eclipsing Binary catalog, the LINEAR data set contains targets other than (but does include) eclipsing binaries; the data set we used (Johnston and Peter 2017) includes Algols (287), Contact Binaries (1805), Delta Scuti (68), and RR Lyr (ab-2189, c-737). The light curves are much more poorly sampled; this uncertainty in the functional shape results from lower SNR (ground survey) and poor sampling. The distribution of stellar classes is presented in Table 3.

The full data sets used at the time of this publication from the Kepler and LINEAR surveys are available from the associated public repository.[b]

4 PMML classification algorithm

Relying on previous designs in astroinformatics to develop a supervised detection algorithm (Johnston and Oluseyi 2017), we propose a design that tailors the requirements specifically toward detecting OEEB-type variable stars.

4.1 Prior research

Many prior studies on time-domain variable star classification (Debosscher 2009; Barclay et al. 2011; Blomme et al. 2011; Dubath et al. 2011; Pichara et al. 2012; Pichara and Protopapas 2013; Graham et al. 2013; Angeloni et al. 2014; Masci et al. 2014) rely on periodicity domain feature space reductions. Debosscher (2009) and Templeton (2004) review a number of feature spaces and a number of efforts to reduce the time-domain data, most of which implement Fourier techniques, primarily the Lomb–Scargle (L-S) method (Lomb 1976; Scargle 1982), to estimate the primary periodicity (Eyer and Blake 2005; Deb and Singh

2009; Richards et al. 2012; Park and Cho 2013; Ngeow et al. 2013).

The studies on classification of time-domain variable stars often further reduce the folded time-domain data into features that provide maximal-linear separability of classes. These efforts include expert selected feature efforts (Debosscher 2009; Sesar et al. 2011; Richards et al. 2012; Graham et al. 2013; Armstrong et al. 2016; Mahabal et al. 2017; Hinners et al. 2018), automated feature selection efforts (McWhirter et al. 2017; Naul et al. 2018), and unsupervised methods for feature extraction (Valenzuela and Pichara 2018; Modak et al. 2018). The astroinformatics community-standard features include quantification of statistics associated with the time-domain photometric data, Fourier decomposition of the data, and color information in both the optical and IR domains (Nun et al. 2015; Miller et al. 2015). The number of individual features commonly used is upward of 60 and growing (Richards et al. 2011) as the number of variable star types increases and as a result of further refinement of classification definitions (Samus' et al. 2017). Curiously, aside from efforts to construct a classification algorithm from the time-domain data directly (McWhirter et al. 2017), few efforts in astroinformatics have looked at features beyond those described here—mostly Fourier domain transformations or time domain statistics. Considering the depth of possibility for time-domain transformations (Fu 2011; Grabocka et al. 2012; Cassisi et al. 2012; Fulcher et al. 2013), it is surprising that the community has focused on just a few transforms. Additionally, there has been recent work in exo-planet detection using whole phased waveform data (smoothed), in combination with neural network classification (Pearson et al. 2017), and with local linear embedding (Thompson et al. 2015). Similar to the design proposed here, these methods use the classifier to optimize the feature space (the phased waveform) for the purposes of detection.

Here we propose an implementation that simplifies the traditional design: limiting ourselves to a one versus all approach (Johnston and Oluseyi 2017) targeting a variable type of interest; limiting ourselves to a singular feature space—the distribution field of the phased light curve—based on (Helfer et al. 2015) as a representation of the functional shape; and introducing a classification/detection scheme that is based on similarity with transparent results (Bellet et al. 2015) that can be further extended, allowing for the inclusion of an anomaly detection algorithm.

4.2 Distribution field

As stated, this analysis focuses on detecting OEEB systems based on their light curve shape. The OEEB signature has a cyclostationary signal, a functional shape that repeats with a consistent frequency. The signature can be isolated using a process of period finding, folding, and phasing (Graham et al. 2013); the Villanova catalog provides the estimated "best period." 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.

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 and Peter 2017; Park and Cho 2013). This analysis will leverage the distribution field transform to generate a feature space that can be operated on; a distribution field (DF) is defined as (Helfer et al. 2015; Sevilla-Lara and Learned-Miller 2012) Equation (5):

$$\mathrm{DF}_{ij} = \frac{\sum_k^N [y_j < f(x_i \le \phi_k \le x_{i+1})_N < y_{j-1}]}{\sum_k^N [y_j < f(\phi_k)_N < y_{j-1}]}, \tag{5}$$

where N is the number of samples in the phased data, and [ ] is the Iverson bracket (Iverson 1962), given as

$$[P] = \begin{cases} 1 & P = \text{true} \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

and y_j and x_i are the corresponding normalized amplitude and phase bins, respectively, where $x_i = 0, 1/n_x, 2/n_x, \ldots, 1$, $y_i = 0, 1/n_y, 2/n_y, \ldots, 1$, n_x is the number of time bins, and n_y is the number of amplitude bins. The result is a right stochastic matrix, i.e., the rows sum to 1. Bin number, n_x and n_y, is optimized by cross-validation as part of the classification training process. Smoothed phased data—generated from SUPERSMOOTHER—are provided to the DF algorithm.

We found this implementation to produce a more consistent classification process. We found that the min-max scaling normalization—normalized flux—if applied by itself without smoothing, when outliers are present can produce final patterns that focus more on the outlier than the general functionality of the light curve. Likewise, we found that using the unsmoothed data in the DF algorithm resulted in a classification that was too dependent on the scatter of the phased light curve. Although at first glance, that would not appear to be an issue, this implementation resulted in light curve resolution having a large impact on the classification performance—in fact, a higher impact than the shape itself. An example of this transformation is given in Fig. 3.

Though the DF exhibits properties that a detection algorithm can use to identify specific variable stars of interest, it alone is not sufficient for our ultimate goal of automated detection. Rather than vectorizing the DF matrix

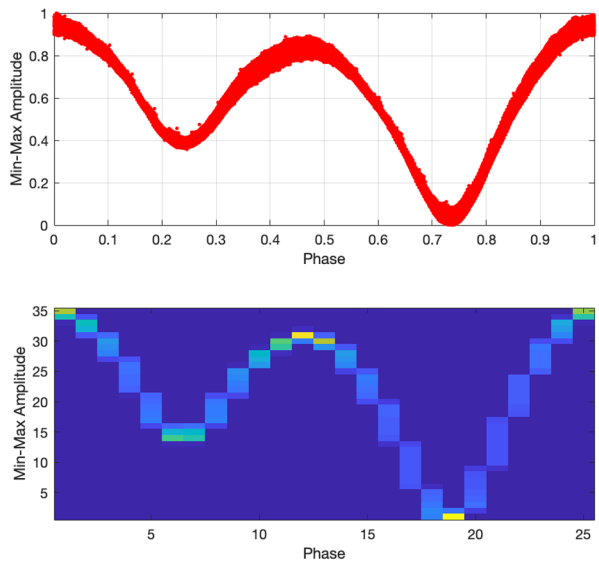

Figure 3 An example phased light curve (top) and the transformed distribution field (bottom) of an Eclipsing Binary with the O'Connell effect (KIC: 7516345)

and treating it as a feature vector for standard classification techniques, we treat the DF as the matrix-valued feature that it is (Helfer et al. 2015). This allows for the retention of row and column dependence information that would normally be lost in the vectorization process (Ding and Dennis Cook 2018).

4.3 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 focuses on defining 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 selected feature space of DF matrices, the distance between two matrices X and Y (Bellet et al. 2015; Helfer et al. 2015) is defined as Equation (7):

$$d(X,Y) = \|X-Y\|_M^2 = \mathrm{tr}\big\{(X-Y)^T M (X-Y)\big\}. \tag{7}$$

M is the metric that we will be attempting to optimize, where $M \succeq 0$ (positive semi-definite). The PMML procedure outlined in (Helfer et al. 2015) is similar to the metric learning methodology LMNN (Weinberger et al. 2009), save for its implementation on matrix-variate data as opposed to vector-variate data. We summarize it here. The developed objective function is given in Equation (8):

$$E = \frac{1-\lambda}{N_c - 1}\sum_{i,j}\big\|\mathrm{DF}_c^i - \mathrm{DF}_c^j\big\|_M^2 - \frac{\lambda}{N-N_c}\sum_{i,k}\big\|\mathrm{DF}_c^i - \mathrm{DF}_c^k\big\|_M^2 + \frac{\gamma}{2}\|M\|_F^2, \tag{8}$$

where N_c is the number of training data in class c; λ and γ are variables to control the importance of push versus pull and regularization, respectively. Bellet et al. (2015) define the triplet $\{\mathrm{DF}_c^i, \mathrm{DF}_c^j, \mathrm{DF}_c^k\}$ as the relationship between similar and dissimilar observations, i.e., DF_c^i is similar to DF_c^j and dissimilar to DF_c^k. Note, the summation over i and j is the summation over similar observations in the training data, and the summation i and k is the summation over dissimilar observations.

There are three basic components: a pull term, which is small when the distance between similar observations is small; a push term, which is small when the distance between dissimilar observations is larger; and a regularization term, which is small when the Frobenius norm ($\|M\|_F^2 = \sqrt{\mathrm{Tr}(MM^H)}$) of M is small. Thus the algorithm attempts to bring similar distribution fields closer together, while pushing dissimilar ones farther apart, while attempting to minimize the complexity of the metric M. The regularizer on the metric M guards against overfitting and consequently enhances the algorithm's ability to generalize, i.e., allow for operations across data sets. This regularization strategy is similar 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 (8) is quadratic in the unknown metric M; hence it is possible to obtain the following closed-form solution to the minimization of Equation (8) as:

$$M = \frac{\lambda}{\gamma(N-N_c)}\sum_{i,k}\big(\mathrm{DF}_c^i - \mathrm{DF}_c^k\big)\big(\mathrm{DF}_c^i - \mathrm{DF}_c^k\big)^T - \frac{1-\lambda}{\gamma(N_c-1)}\sum_{i,j}\big(\mathrm{DF}_c^i - \mathrm{DF}_c^j\big)\big(\mathrm{DF}_c^i - \mathrm{DF}_c^j\big)^T. \tag{9}$$

Equation (9) does not guarantee that M is positive semi-definite (PSD). To ensure this property, we can apply the following straightforward projection step after calculating M to ensure the requirement of $M \succeq 0$:

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: $M = U^T \Lambda_+ U$.

If M is not PSD, then the distance axioms are not held up, and therefore our similarity that we are using $d(X, Y) = \|X - Y\|_M^2$ would not be a true distance, specifically if M is not symmetric then $d(x_i, x_j) \neq d(x_j, x_i)$. 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-nearest neighbor (k-NN) classifier.

4.4 k-NN classifier

The traditional k-NN algorithm is a nonparametric classification method; it uses a voting scheme based on an initial training set to determine the estimated label (Altman 1992). For a given new observation, the L_2 Euclidean distance is found between the new observation and all points in the training set. The distances are sorted, and the k closest training sample labels are used to determine the new observed sample estimated label (majority rule). Cross-validation is used to find an optimal k value, where k is any integer greater than zero.

The k-NN algorithm estimates a classification label based on the closest samples provided in training. For our implementation, the distance between a new pattern DF^i and each pattern in the training set is found, using the optimized metric instead of the standard identity metric that would have been used in L_2 Euclidean distance. The new pattern is classified depending on the majority of the closest k class labels. The distance between patterns is in Equation (7), using the learned metric M.

5 Results of the classification

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 like LSST.

5.1 Training on Kepler data

Optimized feature dimensions and parameters used in the classification process were estimated via five-fold cross-validation (Duda et al. 2012). Our procedure is as follows: the original set of 151 was split into two groups, 76 for training and 75 for testing. The training dataset is partitioned into 5 groups of equal sizes (14), with no replication. To evaluate the optimal DF dimensions we loop over a range of both x and y resolutions. We also include a loop for number of k-Neighbors. Within these loops we evaluate the performance of three classifiers that have limited input parameters to train using our partitioned data. This cycling over the partitioned data is the 5-fold cross-validation, within this loop we cycle over each partition, leaving it out, using the other four for training, and comparing the classifier trained on the four against the one left out.

The resulting error average over the five cycles is used as the performance estimate for the selection of x/y/k resolution. After the resulting analysis, we have three error estimates, per x/y/k resolution pairing. We select the "optimal" x/y resolution pairing based on a minimization of the PMML classifier, for all classifiers trained (the optimal resolutions for the other classifier methods resulted in roughly the same resolution at the PMML one). These optimal resolutions are then used to generate the DF features from all of the training data. This training data is then used to train the classifier selected, and those trained classifiers are then applied to the testing data that was originally set aside. The resulting misclassification rates are provided in the Table 6.

We make the following general notes/caveats about the process used:

- The selection of partitions is a random process (there is a random selection algorithm in the partitioning algorithm). Therefore the resulting errors produced as part of the cross-validation process aren't guaranteed to be the same on subsequent runs (i.e. different partitions selection = different error rates).
- Likewise, one of the alternative classification methodologies requires the use of clustering, so similar to (3) there is some inherent randomness associated with the process that may result in different results from run to run.

The minimization of 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 (i.e. k number of neighbors). Some parameters are more sensitive than others; often this insensitivity is related to the loss function or the feature space, or the data themselves. For example, the γ and λ values weakly affected the optimization, while the bin sizes and k-values had a stronger effect (γ and λ values tested both spanned the range 0.0 to 1.0, k-Values were tested for odd values between 1 and 13).

The cross-validation process was then reduced to optimizing the n_x, n_y, and k values; n_x and n_y values were tested over values 20–40 in steps of 5. The set of optimized parameters is given as γ =1.0, λ = 0.75, n_x = 25, n_y = 35, and k = 3. Given the optimization of these floating variables in all three algorithms, the testing data are then applied to the optimal designs (testing results provided in Sect. 5.2.2).

5.2 Application on unlabeled Kepler data

The algorithm is applied to the Villanova Eclipsing Binary catalog entries that were not identified as either "Of Interest" or "Not of Interest," i.e., unlabeled for the purposes of our targeted goal. The trained and tested data sets are 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 to apply the system to the entire Villanova Eclipsing Binary data (2875 curves).

5.2.1 Design considerations

On the basis of the results demonstrated in Johnston and Oluseyi (2017), the algorithm additionally conditions the detection process based on a maximal distance allowed between a new unlabeled point and the training data set in the feature space of interest.

This anomaly detection algorithm is based on the optimized metric; a maximum distance between data points is based on the training data set, and we use a fraction (0.75) of that maximum distance as a limit to determine "known" versus "unknown." The value of the fraction was initially determined via trial and error, based on our experiences with the data set and the goal of minimizing false alarms (which were visually apparent). This further restricts the algorithm to classifying those targets that exist only in "known space." The k-NN algorithm generates a distance dependent on the optimized metric; by restricting the distances allowed, we can leverage the algorithm to generate the equivalent of an anomaly detection algorithm.

The resulting paired algorithm (detector + distance limit) will produce estimates of "interesting" versus "not interesting," given new—unlabeled—data. Our algorithm currently will not produce confidence estimates associated with the label. Confidence associated with detection can be a touchy subject, both for the scientists developing the tools and for the scientists using them. Here we have focused on implementing a k-NN algorithm with optimized metric (i.e., metric learning); posterior probabilities of classification can be estimated based on k-NN output (Duda et al. 2012) and can be found as $(k_c/(n * \text{volume}))$; linking these posterior probability estimates to "how confident am I that this is what I think this is" is not often the best choice of description.

Confidence in our detections will be a function of the original PPML classification algorithm performance, the training set used and the confidence in the labeling process, and the anomaly detection algorithm we implemented. Even $(k_c/(n * \text{volume}))$ would not be a completely accurate description in our scenario. Some researchers (Dalitz 2009) have worked on linking "confidence" in k-NN classifiers with distance between the points. Our introduction of an anomaly detection algorithm into the design thus allows a developer/user the ability to limit the false alarm rate by introducing a maximum acceptable distance thus allowing some control in the confidence of the result; see Johnston and Oluseyi (2017) for more information.

5.2.2 Results

Once we remove the discovered targets that were also in the initial training data, the result is a conservative selection of 124 potential targets of interest listed in a supplementary digital file at the project repository.[c] We here present an initial exploratory data analysis performed on the phased light curve data. At a high level, the mean and standard deviation of the discovered curves are presented in Fig. 4.

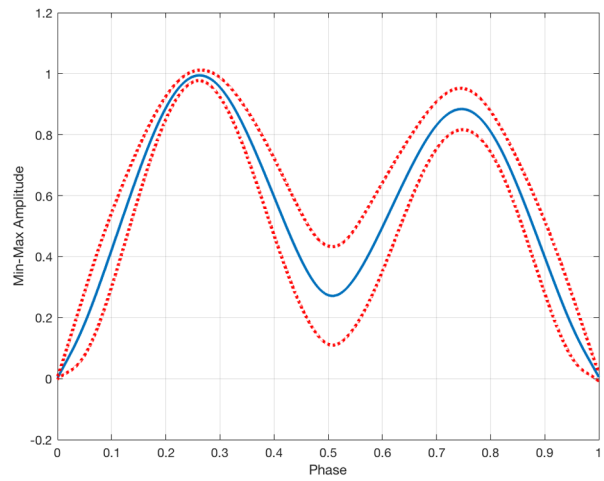

Figure 4 The mean (solid) and a $1-\sigma$ standard deviation (dashed) of the distribution of O'Connell effect Eclipsing Binary phased light curves discovered via the proposed detector out of the Kepler Eclipsing Binary catalog

A more in-depth analysis as to the meaning of the distribution functional shapes is left for future study. Such an effort would include additional observations (spectroscopic and photometric additions would be helpful) as well as analysis using binary simulator code such as Wilson–Devinney (Prša and Zwitter 2005). 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 maximum following the primary eclipse is greater than the second maximum following the secondary eclipse.
2. The light curve relative functional shape from the primary eclipse (minima) to primary maxima is fairly consistent across all discovered systems.
3. The difference in relative amplitude between the two maxima does not appear to be consistent, nor is the difference in relative amplitude between the minima.

We perform additional exploratory data analysis on the discovered group via subgrouping partitioning with unsupervised clustering. The k-means clustering algorithm with matrix-variate distances presented as part of the comparative methodologies is applied to the discovered data set (their DF feature space). This clustering is presented to provide more detail on the discovered group morphological shapes. The associated 1-D curve generated by the SUPERSMOOTHER algorithm is presented with respect to their respective clusters (clusters 1–8) in Fig. 5.

The clusters generated were initialized with random starts, thus additional iterations can potentially result in different groupings. The calculated metric values and the

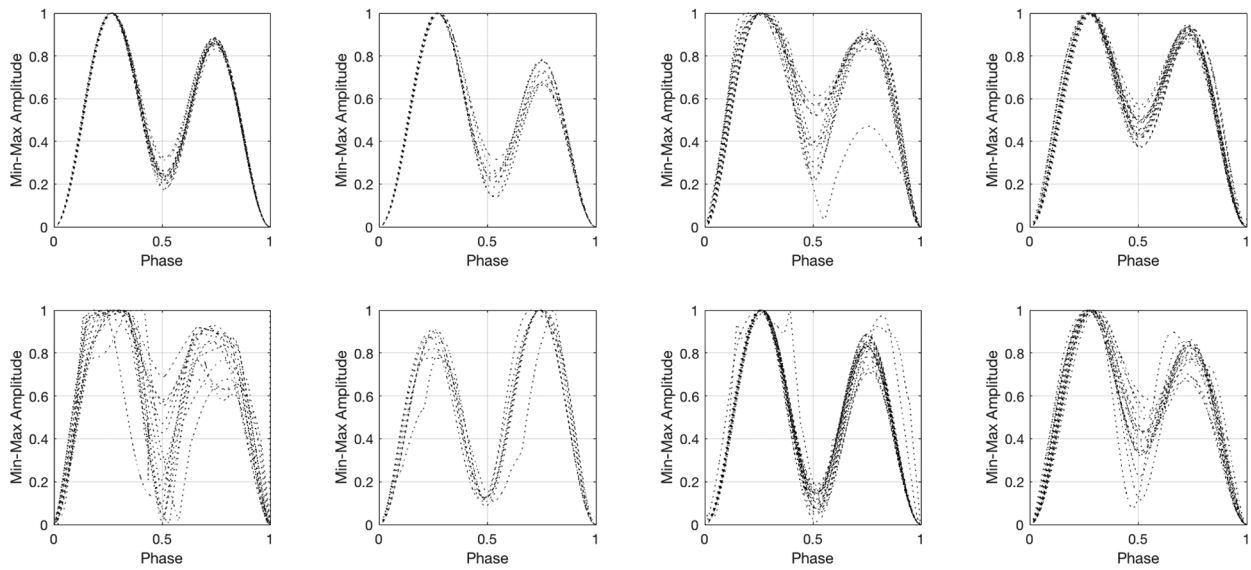

Figure 5 The phased light curves of the discovered OEEB data from Kepler, clustered via k-mean applied to the DF feature space. Cluster number used is based on trial and error, and the unsupervised classification has been implemented here only to highlight morphological similarities. The top four plots represent clusters 1–4 (left to right), and the bottom four plots represent clusters 5–8 (left to right)

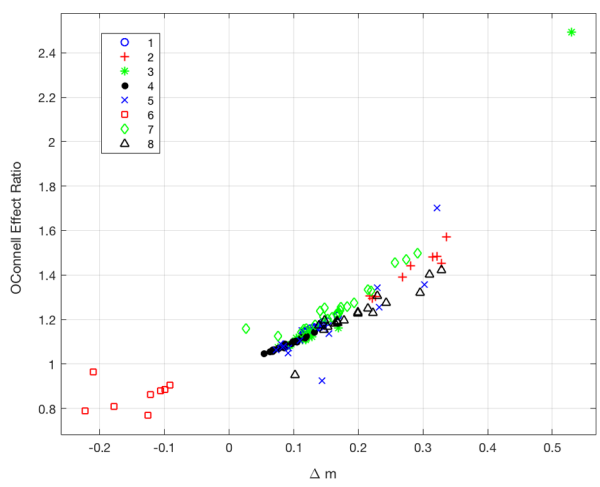

Figure 6 OER versus Δm for discovered Kepler O'Connell effect Eclipsing Binaries. This relationship between OER and Δm was also demonstrated in (McCartney 1999)

clusters numbers for each star are presented in the supplementary digital file. A plot of the measured metrics as well as estimated values of period and temperature (as reported by the Villanova Kepler Eclipsing Binary database), are given with respect to the cluster assigned by k-means.[d] Following Fig. 4.6 in McCartney (1999), plot of OER versus Δm is isolated and presented in Fig. 6.

We note, that based on our methodology, the selection of initial training data (i.e. the expert selected dataset) will have a strong effect on the performance of the detector and the targets discovered. Biases in the initial training data—limits of Δm for example—will be reflected in the the discovered dataset. Between our selection of classifier, which is rooted in leveraging similarity with respect to the initial selected training set, and the anomaly detection algorithm we have supplemented our design with, this design effectively ensures that the targets discovered will be limited to morphological features similar to the initial training data provided by the expert. If there were an OEEB with a radically different shape than those stars used in the initial training data, this methodology would likely not find those stars. An increase in missed detection rate has thus been traded for a decrease in false alarm rate, a move that we feel is necessary given the goals of this design and the amount of potential data that could be fed to this algorithm.

5.2.3 Subgroup analysis

The linear relationship between OER and Δm reported in (McCartney 1999) is apparent in the discovered Kepler data as well. The data set 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 that is extreme. This is comparable to the reported range in (McCartney 1999) of OER $\sim (0.8, 1.2)$ and $\Delta m \sim (-0.1, 0.05)$—a similar OER range, but our Kepler data span a much larger Δm domain, likely resulting from our additional application of min-max amplitude scaling (i.e., normalized flux). The gap in Δm between −0.08 and 0.02 is caused by the bias in our training sample and algorithm goal: we only include O'Connell effect binaries with a user-discernible Δm.

Table 4 Metric measurements from the discovered O'Connell effect eclipsing binaries from the Kepler data set

Cluster	Δm	$\sigma_{\Delta m}/\Delta m$	OER	$\sigma_{\mathrm{OER}}/\mathrm{OER}$	LCA	$\sigma_{\mathrm{LCA}}/\mathrm{LCA}$	#
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

Note. Metrics are based on (McCartney 1999) proposed values of interest.

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 4. All of the clusters have a positive mean Δm, save for cluster 6. The morphological consistency within a cluster is visually apparent in Fig. 5 but also in the relative error of LCA, with clusters 5 and 7 being the least consistent. The next step will include applications to other surveys.

5.3 LINEAR catalog

We further demonstrate the algorithm with an application to a separate independent survey. Machine learning methods have been applied to classifying 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 include 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 apply our Kepler trained detector to the data to generate "discovered" targets of interest.

The discovered targets are aligned, and the smoothed light curves are presented in Fig. 7. Note that the LINEAR IDs are presented in Table 5 and as a supplementary digital file at the project repository.[e]

Application of our Kepler trained detector to LINEAR data results in 24 "discovered" OEEBs. These include four targets with a negative O'Connell effect. Similar to the Kepler discovered data set, we plot OER/Δm features using lower-resolution phased binnings (n = 20) and see that the distribution and relationship from McCartney (1999) hold here as well (see Fig. 8).

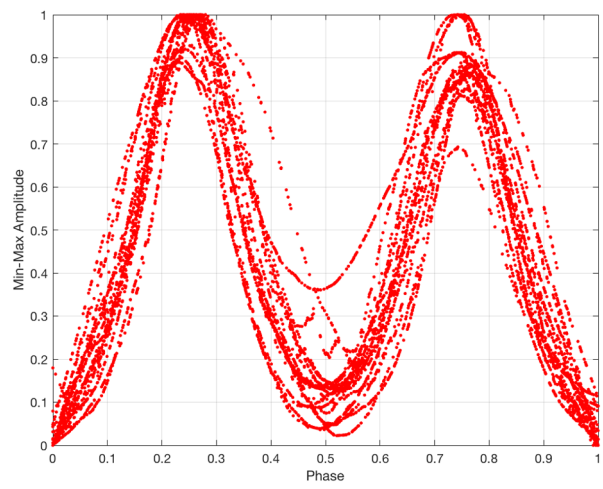

Figure 7 Distribution of phased–smoothed light curves from the set of discovered LINEAR targets that demonstrate the OEEB signature. LINEAR targets were discovered using the Kepler trained detector

6 Discussion on the methodology

6.1 Comparative studies

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 that have trained on this particular data set, targeted this particular variable type, used this feature space, or used this classifier. As we discussed earlier, classifiers that implement matrix-variate features directly are few and far between and almost always not off the shelf. We have developed here two hybrid designs—off-the-shelf classifiers mixed with feature space transform—to provide context and comparison.

These two additional classification methodologies implement more traditional and well-understood features and classifiers: k-NN using L^2 distance applied to the phased light curves (Method A) and k-means representation with quadratic discriminant analysis (QDA) (Method B). 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 data set 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). Method B

Table 5 Discovered OEEBs from LINEAR

13824707	19752221	257977	458198	7087932	4306725	23202141	15522736′
1490274	21895776	2941388	4919865	8085095	4320508	23205293	17074729′
1541626	22588921	346722	4958189	8629192	6946624′		

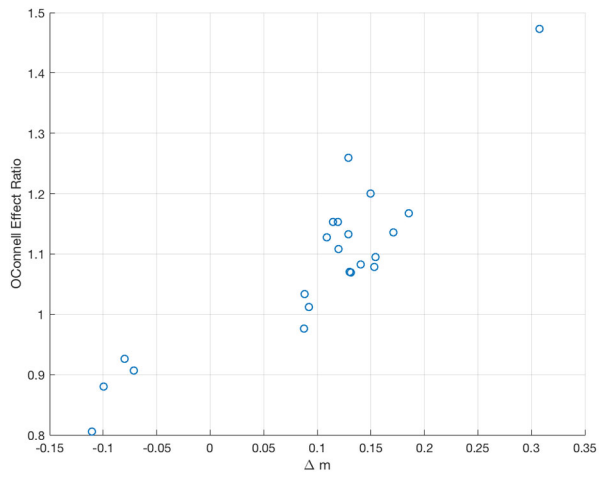

Figure 8 OER versus Δm for the discovered OEEB in the LINEAR data set. This relationship between OER and Δm was also demonstrated in (McCartney 1999) and is similar to the distribution found in Fig. 6

borrows from Park et al. (2003), transforming the matrix-variate data into vector-variate data via estimation of distances between our training set and a smaller set of exemplar means DFs that were generated via unsupervised learning. Distances were found using the Frobenius norm of the difference between the two matrices.

Whereas 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. An alternative to this implementation is a secondary transformation into a lower-dimensional feature space. Following the work of Park et al. (2003), we implement a matrix distance k-means algorithm (e.g., k-means with a Frobenius norm) to generate estimates of clusters in the DF space. Observations are transformed by finding the Euclidean distance between each training point and each of the k-mean matrices discovered. The resulting set of k-distances is treated as the input pattern, allowing the use of the standard QDA algorithm (Duda et al. 2012). The performances of both the proposed methodology and the two comparative methodologies are presented in Table 6. The algorithms are available as open source code, along with our novel implementation, at the project repository.

Table 6 Comparison of performance estimates across the proposed classifiers (based on testing data)

	PPML	Method A	Method B
Error rate	12.5%	15.6%	12.7%

We present the performance of the main novel feature space/classification pairing as well as the two additional implementations that rely on more standard methods. Here we have evaluated performance based on misclassification rate, i.e., 1-accuracy given by Fawcett (2006) as 1 – correct/total. The method we propose has a marginally better misclassification rate (Table 6) and has the added benefit of (1) not requiring unsupervised clustering, which can be inconsistent, and (2) providing nearest neighbor estimates allowing for demonstration of direct comparison. 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 training and testing data are randomized, performance estimates may vary. Of course, increases in training data will result in increased confidence in performance results.

We have not included computational times as part of this analysis, as they tend to be dependent on the system operated on. We can anecdotally discuss that, on the system implemented as part of this research (MacBook Pro, 2.5 GHz Intel i7, 8 GB RAM), the training optimization of our proposed feature extraction and PPML classification total took less than 5–10 min to run—variation depending on whatever else was running in the background. Use of the classifiers on unlabeled data resulted in a classification in fractions of seconds per star. However, we should note that this algorithm will speed up if it is implemented on a parallel processing system, as much of the time taken in the training process resulted from linear algebra operations that can be parallelized.

6.2 Strength of the tools

The DF representation maps deterministic, functional stellar variable observations to a stochastic matrix, with the rows summing to unity. The inherently probabilistic nature of DFs provides a robust way to model interclass variability and handle irregular sampling rates associated with stellar observations. Because the DF feature is indifferent to sampling density so long as all points along the

functional shape are represented, the trained detection algorithm we generate and demonstrate in this article can be trained on Kepler data but directly applied to the LINEAR data, as shown in Sect. 5.3.

The algorithm, including comparison methodologies, designed feature space transformations, classifiers, utilities, and so on, is publicly available at the project repository (see OCD, Johnston et al. 2019);[f] all code was developed in MATLAB and was run on MATLAB 9.3.0.713579 (R2017b). The operations included here can be executed either via calling individual functions or using the script provided (ImplementDetector.m). Likewise, a Java version of all of the individual computational functions has been generated (see JVarStar, Johnston et al. 2019) and is included in the project repository.[g]

6.3 Perspectives

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 and based on prior experience, this design can be replicated for other targets that do not demonstrate a cyclostationary signal (i.e., impulsive, nonstationary, etc.) and even to targets of interest that are not time variable in nature but have a consistent 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—and thus the dominant period allowing a phasing of the light curve—to perform the feature extraction and thus the classification. The DF process here allows for a direct transformation into a singular feature space that focuses on functional shape.

For other variable stars, a multiview 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 radio) would be required in the process of properly defining the variable star type. The solution to this multiview problem is neither straightforward nor well understood (Akaho 2006). Multiple options have been explored to resolve this problem: combination of classifiers, canonical correlation analysis, postprobability blending, and multimetric classification. The computational needs of the algorithm have only been roughly studied, and a more thorough review is necessary in the context of the algorithm proposed and the needs of the astronomy community. The k-NN algorithm dependence on pairwise difference, while one of its strong suits is also one of the more computationally demanding parts of the algorithm. Functionality such as $k-d$ trees as well as other feature space partitioning methods have been shown to reduce the computational requirements.

7 Conclusion

The method we have 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. We compare our methodology to two other methods that have more standard components. The methodology proposed (DF + Push-Pull Metric Learning) is comparable to or outperforms the other two methods. As a demonstration, the design is applied to Kepler eclipsing binary data and LINEAR data. Furthermore, the increase in the number of systems, and the presentation of the data, allows us to make additional observations about the distribution of curves and trends within the population. Future work will involve the analysis of these statistical distributions, as well as an inference as to their physical meaning.

The new OEEB systems we 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 like LSST. Although the effort here targets OEEB as a demonstration, it need not be limited to those particular targets. We could use the DF feature space along with the push–pull metric learning classifier to construct a detector for any variable stars with periodic variability. Furthermore, any variable star (e.g., supernova, RR Lyr, Cepheids, eclipsing binaries) can be targeted using this classification scheme, given the appropriate feature space transformation allowing for quantitative evaluation of similarity. This design is directly applicable to exoplanet 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).

Acknowledgements

The authors are grateful to the anonymous reviewers for their detailed and expansive recommendations on the organization of this article. The authors are grateful for the valuable astrophysics insight provided by C. Fletcher and T. Doyle. Inspiration provided by C. Role. Initial editing and review provided by H. Monteith. The authors would like to additionally acknowledge the Kepler Eclipsing Binary team, without whom much of the training data would not exist.

Funding

Research was partially supported by Perspecta, Inc. This material is based upon work supported by the NASA Florida Space Grant under 2018 Dissertation and Thesis Improvement Fellowship No. 202379. The LINEAR program is sponsored by the National Aeronautics and Space Administration (NRA Nos. NNH09ZDA001N, 09-NEOO09-0010) and the U.S. Air Force under Air Force contract FA8721-05-C-0002.

Abbreviations

TESS, Transiting Exoplanet Survey; KELT, Kilodegree Extremely Little Telescope; LSST, Large Synoptic Survey Telescope; DF, distribution field; OEEBs, O'Connell

effect eclipsing binaries; LINEAR, Lincoln Near-Earth Asteroid Research; MATLAB, Matrix Laboratory; QDA, quadratic discriminate analysis; UCR, University of California Riverside; OER,O'Connell effect ratio; LCA, light curve asymmetry; SDSS, Slone Digital Sky Survey.

Availability of data and materials
The software developed for this article, as well as the reduced data results, is available at https://GitHub.com/kjohnston82/OCDetector. The links to the raw training data are also provided at the public repository.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The detector code was developed by KBJ and RH. Initial targets of interest were identified by MK. Training data and testing data were procured and refined by KBJ. KBJ prepared this manuscript and performed the tests detailed within. All authors read and approved the final manuscript.

Author details
[1]Aerospace, Physics and Space Sciences Department, Florida Institute of Technology, Melbourne, USA. [2]Mathematical Sciences Department, Florida Institute of Technology, Melbourne, USA. [3]Computer Engineering and Sciences Department, Florida Institute of Technology, Melbourne, USA. [4]Physics and Astronomy Department, University of Delaware, Newark, USA. [5]Defense Group Melbourne, Perspecta Inc., Melbourne, USA.

Endnotes
[a] https://github.com/kjohnston82/OCDetector/supplement/KeplerTraining.xlsx
[b] https://github.com/kjohnston82/OCDetector
[c] https://github.com/kjohnston82/OCDetector/supplement/AnalysisOfClusters.xlsx
[d] https://github.com/kjohnston82/OCDetector/Documentation/Figures/ReducedFeaturesKeplerAll_Temp.png
[e] https://github.com/kjohnston82/OCDetector/supplement/LINEARDiscovered.xlsx
[f] https://github.com/kjohnston82/OCDetector
[g] https://github.com/kjohnston82/VariableStarAnalysis

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 2 January 2019 Accepted: 25 October 2019
Published online: 08 November 2019

References
Akaho, S.: A kernel method for canonical correlation analysis. arXiv preprint (2006). arXiv:cs/0609071
Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992)
Angeloni, R., Contreras Ramos, R., Catelan, M., et al.: The VVV templates project towards an automated classification of VVV light-curves. I. building a database of stellar variability in the near-infrared. Astron. Astrophys. **567**, A100 (2014). https://doi.org/10.1051/0004-6361/201423904
Armstrong, D.J., Kirk, J., Lam, K.W.F., et al.: K2 variable catalogue—II. Machine learning classification of variable stars and eclipsing binaries in K2 fields 0–4. Mon. Not. R. Astron. Soc. **456**, 2260–2272 (2016). https://doi.org/10.1093/mnras/stv2836
Barclay, T., Ramsay, G., Hakala, P., et al.: Stellar variability on time-scales of minutes: results from the first 5 yr of the rapid temporal survey. Mon. Not. R. Astron. Soc. **413**, 2696–2708 (2011). https://doi.org/10.1111/j.1365-2966.2011.18345.x
Bellet, A., Habrard, A., Sebban, M.: Metric learning. Synth. Lect. Artif. Intell. Mach. Learn. **9**(1), 1–151 (2015)
Blomme, J., Sarro, L.M., O'Donovan, F.T., et al.: Improved methodology for the automated classification of periodic variable stars. Mon. Not. R. Astron. Soc. **418**, 96–106 (2011). https://doi.org/10.1111/j.1365-2966.2011.19466.x
Cassisi, C., Montalto, P., Aliotta, M., et al.: Similarity measures and dimensionality reduction techniques for time series data mining. In: Advances in Data Mining Knowledge Discovery and Applications. InTech (2012)
Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 15 (2009)
Chen, Y., Keogh, E., Hu, B., et al.: The UCR Time Series Classification Archive (2015)
Dalitz, C.: Reject options and confidence measures for knn classifiers. Schriftenreihe des Fachbereichs Elektrotechnik und Informatik Hochschule Niederrhein **8**, 16–38 (2009)
Deb, S., Singh, H.P.: Light curve analysis of variable stars using Fourier decomposition and principal component analysis. Astron. Astrophys. **507**, 1729–1737 (2009). https://doi.org/10.1051/0004-6361/200912851
Debosscher, J.: Automated classification of variable stars: application to the OGLE and CoRoT databases. PhD thesis, Institute of Astronomy, Katholieke Universiteit Leuven, Belgium (2009)
Ding, S., Dennis Cook, R.: Matrix variate regressions and envelope models. J. R. Stat. Soc., Ser. B, Stat. Methodol. **80**(2), 387–408 (2018)
Dubath, P., Rimoldini, L., Süveges, M., et al.: Random forest automated supervised classification of Hipparcos periodic variable stars. Mon. Not. R. Astron. Soc. **414**, 2602–2617 (2011). https://doi.org/10.1111/j.1365-2966.2011.18575.x
Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2012)
Eyer, L., Blake, C.: Automated classification of variable stars for all-sky automated survey 1–2 data. Mon. Not. R. Astron. Soc. **358**, 30–38 (2005). https://doi.org/10.1111/j.1365-2966.2005.08651.x
Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)
Friedman, J.H.: A variable span smoother. Technical report, Stanford Univ CA lab for computational statistics (1984)
Fu, T.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
Fulcher, B.D., Little, M.A., Jones, N.S.: Highly comparative time-series analysis: the empirical structure of time series and their methods. J. R. Soc. Interface **10**(83), 20130048 (2013)
Gagniuc, P.A.: Markov Chains: From Theory to Implementation and Experimentation. Wiley, New York (2017)
Grabocka, J., Nanopoulos, A., Schmidt-Thieme, L.: Invariant time-series classification. In: Machine Learning and Knowledge Discovery in Databases, pp. 725–740. Springer, Berlin (2012)
Graham, M.J., Djorgovski, S.G., Mahabal, A.A., Donalek, C., Drake, A.J.: Machine-assisted discovery of relationships in astronomy. Mon. Not. R. Astron. Soc. **431**, 2371–2384 (2013). https://doi.org/10.1093/mnras/stt329
Graham, M.J., Drake, A.J., Djorgovski, S.G., et al.: A comparison of period finding algorithms. Mon. Not. R. Astron. Soc. **434**, 3423–3444 (2013). https://doi.org/10.1093/mnras/stt1264
Haber, R., Rangarajan, A., Peter, A.M.: Discriminative interpolation for classification of functional data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 20–36. Springer (2015)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, vol. 2. Springer, Berlin (2009)
Helfer, E., Smith, B., Haber, R., Peter, A.: Statistical analysis of functional data. Technical report, Florida Institute of Technology, (2015)
Hinners, T.A., Tat, K., Thorp, R.: Machine learning techniques for stellar light curve classification. Astron. J. **156**, 7 (2018). https://doi.org/10.3847/1538-3881/aac16d
Iverson, K.E.: A programming language. In: Proceedings of the May 1–3, 1962, Spring Joint Computer Conference, pp. 345–351. ACM (1962)
Johnston, K.B., Caballero-Nieves, S.M., Peter, A.M., Petit, V., Haber, R.: JVarStar: variable star analysis library. Astrophysics Source Code Library (2019)
Johnston, K.B., Haber, R.: OCD: O'Connell Effect Detector using push-pull learning. Astrophysics Source Code Library (2019)
Johnston, K.B., Oluseyi, H.M.: Generation of a supervised classification algorithm for time-series variable stars with an application to the LINEAR dataset. New Astron. **52**, 35–47 (2017). https://doi.org/10.1016/j.newast.2016.10.004
Johnston, K.B., Peter, A.M.: Variable star signature classification using slotted symbolic Markov modeling. New Astron. **50**, 1–11 (2017). https://doi.org/10.1016/j.newast.2016.06.001
Kirk, B., Conroy, K., Prša, A., et al.: Kepler eclipsing binary stars. VII. The catalog of eclipsing binaries found in the entire Kepler data set. Astron. J. **151**, 68 (2016). https://doi.org/10.3847/0004-6256/151/3/68

Knote, M.F., Kaitchuck, R.H., Berrington, R.C.: Observations and preliminary modeling of the light curves of eclipsing binary systems NSVS 7322420 and NSVS 5726288. J. Am. Assoc. Var. Star Obs. **47** (2019)
Lomb, N.R.: Least-squares frequency analysis of unequally spaced data. Astrophys. Space Sci. **39**, 447–462 (1976). https://doi.org/10.1007/BF00648343
Mahabal, A., Sheth, K., Gieseke, F., et al.: Deep-learnt classification of light curves. In: Computational Intelligence (SSCI), 2017 IEEE Symposium Series on, pp. 1–8. IEEE (2017)
Masci, F.J., Hoffman, D.I., Grillmair, C.J., Cutri, R.M.: Automated classification of periodic variable stars detected by the wide-field infrared survey explorer. Astron. J. **148**, 21 (2014). https://doi.org/10.1088/0004-6256/148/1/21
McCartney, S.A.: A 2-D model for the O'Connell effect in W Ursae Majoris systems. PhD thesis, The University Of Oklahoma (1999)
McWhirter, P.R., Steele, I.A., Al-Jumeily, D., Hussain, A., Vellasco, M.M.B.R.: The classification of periodic light curves from non-survey optimized observational data through automated extraction of phase-based visual features. In: Neural Networks (IJCNN), 2017 International Joint Conference on, pp. 3058–3065. IEEE (2017)
Miller, A.A., Bloom, J.S., Richards, J.W., et al.: A machine-learning method to infer fundamental stellar parameters from photometric light curves. Astrophys. J. **798**, 122 (2015). https://doi.org/10.1088/0004-637X/798/2/122
Modak, S., Chattopadhyay, T., Chattopadhyay, A.K.: Unsupervised classification of eclipsing binary light curves through k-medoids clustering. arXiv e-prints (2018). arxiv:1801.09406
Naul, B., Bloom, J.S., Pérez, F., van der Walt, S.: A recurrent neural network for classification of unevenly sampled variable stars. Nat. Astron. **2**, 151–155 (2018). https://doi.org/10.1038/s41550-017-0321-z
Ngeow, C., Lucchini, S., Kanbur, S., et al.: Preliminary analysis of ULPC light curves using Fourier decomposition technique. In: Space Science and Communication (IconSpace), 2013 IEEE International Conference on, pp. 7–12. IEEE (2013)
Nun, I., Protopapas, P., Sim, B., Zhu, M., et al.: FATS: feature analysis for time series. arXiv e-prints (2015). arxiv:1506.00010
O'Connell, D.J.K.: The so-called periastron effect in eclipsing binaries (summary). Mon. Not. R. Astron. Soc. **111**, 642 (1951). https://doi.org/10.1093/mnras/111.6.642
Palaversa, L., Ivezić, Ž., Eyer, L., et al.: Exploring the variable sky with LINEAR. III. Classification of periodic light curves. Astron. J. **146**, 101 (2013). https://doi.org/10.1088/0004-6256/146/4/101
Park, H., Jeon, M., Rosen, J.B.: Lower dimensional representation of text data based on centroids and least squares. BIT Numer. Math. **43**(2), 427–448 (2003)
Park, M.J., Cho, S.S.: Functional data classification of variable stars. Commun. Stat. Appl. Meth. **20**(4), 271–281 (2013)
Pearson, K.A., Palafox, L., Griffith, C.A.: Searching for exoplanets using artificial intelligence. Mon. Not. R. Astron. Soc. **474**(1), 478–491 (2017)
Pichara, K., Protopapas, P.: Automatic classification of variable stars in catalogs with missing data, Astrophys. J. **777**, 83 (2013). https://doi.org/10.1088/0004-637X/777/2/83
Pichara, K., Protopapas, P., Kim, D.-W., et al.: An improved quasar detection method in EROS-2 and MACHO LMC data sets. Mon. Not. R. Astron. Soc. **427**, 1284–1297 (2012). https://doi.org/10.1111/j.1365-2966.2012.22061.x
Prša, A., Zwitter, T.: A computational guide to physics of eclipsing binaries. I. Demonstrations and perspectives. Astrophys. J. **628**, 426–438 (2005). https://doi.org/10.1086/430591
Richards, J.W., Starr, D.L., Butler, N.R., et al.: On machine-learned classification of variable stars with sparse and noisy time-series data. Astrophys. J. **733**, 10 (2011). https://doi.org/10.1088/0004-637X/733/1/10
Richards, J.W., Starr, D.L., Miller, A.A., et al.: Construction of a calibrated probabilistic classification catalog: application to 50k variable sources in the all-sky automated survey. Astrophys. J. Suppl. Ser. **203**, 32 (2012). https://doi.org/10.1088/0067-0049/203/2/32
Samus', N.N., Kazarovets, E.V., Durlevich, O.V., et al.: General catalogue of variable stars: version GCVS 5.1. Astron. Rep. **61**, 80–88 (2017). https://doi.org/10.1134/S1063772917010085
Scargle, J.D.: Studies in astronomical time series analysis. II—statistical aspects of spectral analysis of unevenly spaced data. Astrophys. J. **263**, 835–853 (1982). https://doi.org/10.1086/160554
Sesar, B., Stuart, J.S., Ivezić, Ž., et al.: Exploring the variable sky with LINEAR. I. Photometric recalibration with the sloan digital sky survey. Astron. J. **142**, 190 (2011). https://doi.org/10.1088/0004-6256/142/6/190
Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1910–1917. IEEE (2012)
Stokes, G.H., Evans, J.B., Viggh, H.E.M., Shelly, F.C., Pearce, E.C.: Lincoln near-Earth asteroid program (LINEAR). Icarus **148**, 21–28 (2000). https://doi.org/10.1006/icar.2000.6493
Templeton, M.: Time-series analysis of variable star data. J. Am. Assoc. Var. Star Obs. **32**, 41–54 (2004)
Thompson, S.E., Mullally, F., Coughlin, J., Christiansen, J.L., Henze, C.E., Haas, M.R., Burke, C.J.: A machine learning technique to identify transit shaped signals. Astrophys. J. **812**(1), 46 (2015)
Valenzuela, L., Pichara, K.: Unsupervised classification of variable stars. Mon. Not. R. Astron. Soc. **474**, 3259–3272 (2018). https://doi.org/10.1093/mnras/stx2913
VanderPlas, J.T., Ivezić, Ž.: Periodograms for multiband astronomical time series. Astrophys. J. **812**(1), 18 (2015). https://doi.org/10.1088/0004-637X/812/1/18
Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009) http://dl.acm.org/citation.cfm?id=1577069.1577078. ISSN 1532-4435
Wilsey, N.J., Beaky, M.M.: Revisiting the O'Connell effect in eclipsing binary systems. Soc. Astron. Sci. Ann. Symp. **28**, 107 (2009)
Wilson, R.E., Devinney, E.J.: Realization of accurate close-binary light curves: application to MR cygni. Astrophys. J. **166**, 605 (1971). https://doi.org/10.1086/150986
Zboril, M., Djurasevic, G.: Progress report on the monitoring active late-type stars in 2005/2006 and the analysis of V523 cas. Serb. Astron. J. **173**, 89 (2006). https://doi.org/10.2298/SAJ0673089Z

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com