Modeling Lightcurves for Improved Classification of Astronomical Objects

Julian Faraway[1*], Ashish Mahabal[2], Jiayang Sun[3], Xiao-Feng Wang[4], Yi G. Wang[5] and Lingsong Zhang[6]

[1]*Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK*

[2]*Astronomy Department, California Institute of Technology, Pasadena, CA 91125, USA*

[3]*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA*

[4]*Biostatistics Department of Quantitative Health Sciences, Cleveland Clinic Lerner Research Institute, Cleveland, OH 44195, USA*

[5]*Department of Mathematics, Syracuse University, Syracuse, NY 13244, USA*

[6]*Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*

Received 8 September 2014; revised 5 November 2015; accepted 8 January 2016
DOI:10.1002/sam.11305
Published online 1 February 2016 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Many synoptic surveys are observing large parts of the sky multiple times. The resulting time series of light measurements, called lightcurves, provide a wonderful window to the dynamic nature of the Universe. However, there are many significant challenges in analyzing these lightcurves. We describe a modeling-based approach using Gaussian process regression for generating critical measures for the classification of such lightcurves. This method has key advantages over other popular nonparametric regression methods in its ability to deal with censoring, a mixture of sparsely and densely sampled curves, the presence of annual gaps caused by objects not being visible throughout the year from a given position on Earth and known but variable measurement errors. We demonstrate that our approach performs better by showing it has a higher correct classification rate than past methods popular in astronomy. Finally, we provide future directions for use in sky-surveys that are getting even bigger by the day. © 2016 Wiley Periodicals, Inc. Statistical Analysis and Data Mining: The ASA Data Science Journal 9: 1–11, 2016

Keywords: classification, feature selection, Gaussian process regression, irregular sampling, missing data

1. INTRODUCTION

In the last few decades, we have seen advances in imaging technology, and the storage, transfer and processing of data. As a result, astronomy has moved from taking static, sporadic snapshots of the sky to obtaining high-cadence, deep and large images, almost akin to making digital movies of the sky. This, in turn, has resulted in opening up the field of studying the dynamic nature of the Universe, in particular, the cataloging of different types of objects, both within our Galaxy, and all the way to the early Universe. Cataloging goes well beyond stamp-collecting, since it reveals the time scales over which various phenomena occur, directly relating to the physical processes behind the brightness changes in astronomical objects, and allowing us to connect the different families of objects in various ways. A bonus is also the ability to look for connections missing so far, as well as fringe members of different classes.

Much of characterizion or classification for cataloging is done, or at least begins, through the study of variability of objects. Most astronomical objects, be they stars, planets or galaxies, or any of their subclasses, vary in brightness either intrinsically through some physical process such as explosion, merging or infall of matter, or through an extrinsic process such as eclipse or rotation. For a small fraction of objects the variation can happen over a fraction of a second to hundreds of days depending on the

* *Correspondence to:* Julian Faraway (jjf23@bath.ac.uk)

phenomenon. For a majority of objects, the changes are much slower and smaller as the objects evolve through the proverbial astronomical timescales. We can observe large parts of the sky multiple times at different wavelengths, yet these observations are far from continuous, all-sky, or panchromatic. For each part of the sky, and in particular for each object in the part of the sky we image, we get a time series of flux. While all objects vary to an extent, for a vast majority of objects, the variations are non-discernible during the rather sparse sequence (tens to hundreds of epochs) of short exposures (less than a minute) that we have, and over the timescales over which observations occur (a few years). That is precisely the reason, for instance, that when we glance at the night sky we do not find stars suddenly changing their brightness.

This leads to most astronomical objects seeming *non-variable*. When we can discern the variability, e.g. a periodic variation, or a stochastic variation, or even a single sudden jump in brightness, the object could then be called a *variable*. (Due to the somewhat different meanings of the word 'variable' in the two fields, we refer to variables in statistics and *variables* in astronomy.) This functional definition would of course change based on many factors, such as, total interval of observation, type of phenomena involved, etc. An extreme case of a *variable* object is a *transient* - the brightness of which varies by several standard deviations in a short time, of the order of seconds to minutes. It is the study of these types of objects that has really become possible due to wide-field surveys that contain many repeated observations.

In order to understand and classify transients, it is important to understand variability at all levels, including mostly *non-variable* astronomical sources. Past attempts have included analyses for denser lightcurves from Kepler, as in ref. [1,2] or using brighter objects as in ref. [3] and general frameworks based on such approaches as in ref. [4,5] as well as ref. [6–9]. It is important to design measures that can isolate specific classes but which are also derivable based on the available cadence of observations. Our aim is to present new measures based on object lightcurves which help in better discriminating between *variables* and *non-variables*, and among the different transient types. See ref. [10] for an application to larger datasets and ref. [11] for use on specific classes.

Here we use data from the Catalina Real-time Transient Survey (CRTS) [12] which is based on the Catalina Sky Survey (CSS). CSS has been designed to look for near-Earth asteroids. One way to look for asteroids is by looking for the motion of the asteroids with the backdrop of mostly nonmoving stars in the night sky. The cadence used for this is four images taken 10 min apart. Thus, the CSS lightcurves have four points obtained within 30 min. The next such set could be the next night, the next week or even a month later. The sparse and nonuniform nature of the lightcurves presents classification challenges and also allows development of new statistical techniques. The data come from three sites: (i) Mt. Lemon (MLS) - this site covers mainly a narrow region of the sky near the ecliptic, (ii) Siding Spring (SSS) - this site covers the Southern hemisphere, and (iii) Mt. Catalina - this covers the widest area in the Northern hemisphere and also covers part of the Southern hemisphere. About 75% of the sky is covered by CRTS, with parts near the poles and near the plane of our Galaxy excluded. CRTS concerns itself with real-time detection of brightness variations based on the catalog of astronomical sources. Here we have used data from just the Mt. Catalina telescope, and excluded data from MLS or SSS in the current study, but all methods are equally applicable to them as well. Despite the relative sparsity of the CRTS lightcurves, a strength of the survey is its longevity - we have data where the epochs are spread over 10 years and hence there are parts of the sky with several hundred observations, making CRTS one of the richest synoptic datasets.

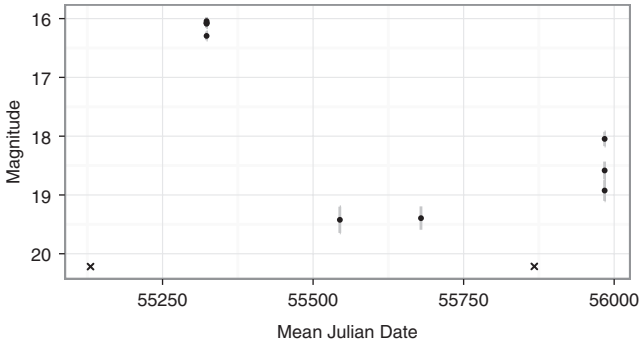

Fig. 1 Fragment of a lightcurve. The solid dots represent recorded observations and are plotted with error bars. The two crosses represent times when an observation was attempted but only an upper limit on the magnitude can be specified.

A small section of about 1 year of light curve observation is shown in Figure 1. The magnitude is the negative logarithm of flux so in keeping with standard practice, we plot the magnitudes on a reversed scale because smaller magnitudes represent brighter objects. There are six time periods where observations are shown in the figure. There are two times where multiple observations were recorded within the typical 'four observations in 30-minutes' CSS sequence, two times when only one reliable observation was made and two other times (marked with a cross) when the usual four observations were made, but there was no reliable detection because the object was insufficiently bright at that time. The magnitudes vary substantially indicating a transient of some type to be determined.

Our method develops informative measures that can represent critical features of the different types of lightcurves,

much like descriptors for an imaging analysis. We require these measures to overcome the irregularities, such as censoring, presence of gaps in observations and irregular sampling of different density. This allows a massive and essential reduction of the data as classification methods cannot be directly applied to the data in the raw lightcurve form. The measures can be computed rapidly and in parallel for classification purposes. The measures can also be used for clustering but this is not explored here.

Our strategies in deriving these critical measures are (i) selecting a collection of relatively balanced, representative lightcurves of various types and scales (Section 2); (ii) exploring the signatures of these lightcurves (also in Section 2); (iii) developing a Gaussian process regression model for the lightcurves with appropriate priors using astronomical information and an empirical Bayesian approach (Section 3); and then (iv) deriving new measures representing characteristics of the lightcurves using the posterior mean regression curve and residuals. These model-based measures complement existing measures. To examine the power of the new measures in classifying lightcurves in comparison to existing measures, five popular classification procedures are used in five schemes of classification problem in Section 4: Linear Discriminant Analysis, Decision Trees, Support Vector Machines, Neural Networks, and Random Forests. The results show that our measures perform better than the existing measures. A discussion for why our approach works better is given in Section 5. Although our modeling approach has been used for an astronomy application, the method could be valuable for other applications involving the classification of sparse, irregularly sampled time series with missing data.

2. DATA

We have selected a sample of lightcurves with which to illustrate our methods. The standard measures available for each object are Right Acension (RA) and Declination (Dec) which provide the position of the object on the sky, Epoch as Julian Date, magnitude (negative logarithm of flux), and an error estimate on the magnitude. The total number of lightcurves considered is 3720. The selection is described below and shown in Table 1.

We started with just the transients detected by CRTS in real-time over about 5 years. These include *Active Galactic Nuclei* (AGNs), *Blazars, Cataclysmic Variables* (CV), *Flare* stars and *Supernovae* (SNe), representing five very different types of lightcurves (e.g. [13]). We also included a set of 15 random locations on the sky covered by the survey and objects within 3' of these locations. These objects are assumed to be *non-variables* because any *variables* in there would have been detected earlier. The transients tend to be fainter than typical objects (by definition - it is easier to catch objects that are not normally seen but brighten and become visible for a short duration). In order to offset that, two classes of brighter *variable* objects were included —CVs from the *Downes* set ([14]) and *RR Lyrae* which are periodic *variables* with a period of approximately 1 day. For the purposes of this article, we have one class called *non-variable* and seven classes which we call *variable*: AGN, Blazars, CV, Flares, SNe, CV Downes, and RR Lyrae with first five also being *transient*. Note that among the labeled types we have considered here, only the *RR Lyrae* are periodic. There are methods for distinguishing periodic objects from nonperiodic ones but these are not addressed in this article.

Table 1. Number of lightcurves for *variable*, specifically transient and *non-variable* objects.

Transients					Bright *variable*		
AGN	Blazar	CV	Flare	SNe	CV Downes	RR-Lyrae	*non-variable*
140	124	461	66	536	376	292	1971

Real-world data are much more assymetric and unbalanced than what we have considered here. If we used a simple random sample, there would be very few transients. Training on samples with enough representatives for all classes would be an immense task. We include sufficient *non-variables* to ensure that the dominant class is represented but not excessively so. In CRTS, the latest catalogs are compared with individual as well as combined catalogs from the archives. The objects that have changed in brightness above a certain threshold (well over a magnitude, i.e. approximately a factor of 2.5) are marked as transients. Only a few are found each night compared to millions of nearly *non-variable* objects.

The number of observations for each lightcurve varied greatly with a maximum of 641. The median length was 52. We excluded lightcurves with fewer than five observations as these cannot reasonably be classified. The earliest date for any of these lightcurves was 53464 Julian Day (JD) and the latest was 56228, i.e. April 5, 2005 to October 28, 2012. We have used 53464 as our zero-point and referred to all dates as number of days beyond this. Our set spans 2764 days. More recent data are analyzed later.

**An examination of the data is helpful in deciding which methods of analysis may be appropriate. In Figure 2, we see four examples of lightcurves. The objects are identified by their catalogue numbers for reference.

Understanding the pattern of measurement is crucial to proper modeling of these curves. The AGN example shows some gaps in an otherwise dense sequence of measurements. No observations were taken during these periods because the orbit of Earth precluded it. In the SN example, there are no observations outside of a narrow

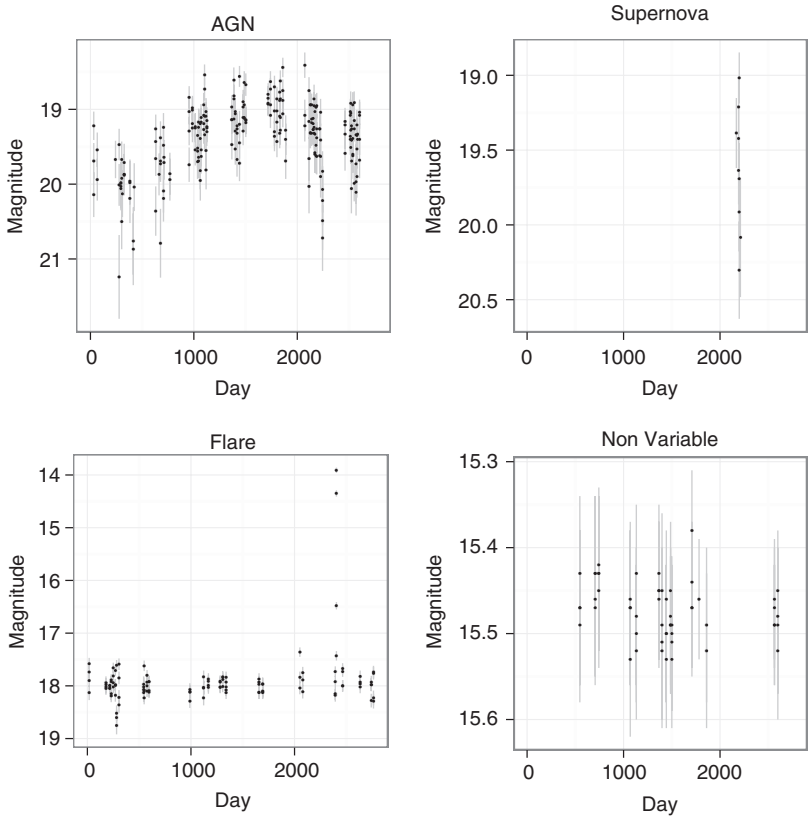

Fig. 2 Examples of four lightcurves: (a) CSS071216:110407-045134, an AGN (b) CSS110405:141104+01115, a Supernova (c) CSS111103:230309+40060, a Flare and (d) 301904800767, a *non-variable*. Note that the range of the vertical axis varies.

range. Observations were attempted at other times but the object was too faint to be observed with a magnitude less than the survey astronomical detection limit of around 20.5. In the flare example, we are fortunate that the spike in brightness was observed as this occurs during a brief period of time. In the *non-variable* example, there are quite long periods with no observations but it seems reasonable to assume that no substantial variations in magnitude occurred during these periods given the nearly constant values of magnitude.

It would be useful to know exactly when observations were attempted for given objects while below detection limit. For the purposes of this analysis, we shall assume that all the objects may be surveyed throughout the period of the study but failures to observe have not been recorded. It will be clear in Section 3 how this information could be incorporated into our methods and that this would improve our results.

3. METHODS

The nature of the data and the requirements of object classification impose some constraints on what methods are practical. The problem could be viewed as one of functional data analysis (see ref. [15]). However, there are several obstacles in pursuing this approach. The lightcurves are very irregular, showing differences both in time and in number of observations. There are methods for processing with such data but there is a more serious obstacle in that there is little sense in which the curves can be registered or aligned. Excepting the rare case where objects are close in the sky and measurements are likely to be correlated due to atmospheric conditions, lightcurves are independent. This prevents us from using the 'borrowing of strength' that registration would allow.

This leads us to another style of analysis based on sample statistics. Judgement is used to devise statistics that measure various characteristics of the observed curves which we believe important in distinguishing them. We prefer that these measures be relatively simple so that they can be applied quickly and automatically for both short and long lightcurves.

About 20 measures are presented in ref. [3] that are mostly derived from previous articles. They found these various measures to be helpful in distinguishing objects. Since these measures have been widely tested, at least for brighter and less sparse data, we use these

as a baseline for our analysis. Our objective is to find additional measures that improve the classification accuracy beyond this set. For ease of reference, we will call this existing set the *Richards measures*. The specific Richards measures we have used from Table 5 of [3] are moment-based measures: `skew`, `kurtosis`, `std` and `beyond1std`, and magnitude-based measures: `amplitude`, `maxslope`, `mad`, `medbuf`, `pairslope` and `rcorbor`, and percentile-based measures: `fpr20`, `fpr35`, `fpr50`, `fpr80`, `peramp` and `pdfp`. We omitted the linear trend measure as this was only large for lightcurves with few observations so it becomes a substitute for a short curve measure. As it happens, including it would not make much difference to the results we present later. We coded these measures from the definitions in ref. [3].

Although the Richards measures encompass a wide variety of measures, they do not use any concept of modeling the curves. The primary innovation of this paper is to use such modeling to generate additional measures. For lightcurve i, we posit a true underlying curve $f_i(t)$ that we would see if we could observe the object continuously without error. However, we are able to observe the object only at times t_{ij} for $j = 1, \ldots n_i$. Note that the times of measurement may be almost the same for objects close in the sky but quite different for objects which are farther apart. We observe only y_{ij} for $j = 1, \ldots n_i$. We assume

$$y_{ij} = f_i(t_{ij}) + \epsilon_{ij}, \quad (1)$$

where the errors ϵ_{ij} are normal with mean zero but will be correlated.

We considered several methods for estimating f but found that Gaussian process regression was the only satisfactory solution compared to standard nonparametric methods like smoothing splines, kernel smoothing or locally weighted smoothing for the following reasons:

1. We have censored data - the lightcurve can fall below the detection limit during the range of observation. Standard methods do not deal with this. They can fit curves where we have data but they will not produce evaluable fitted curves outside this range.

2. Sometimes we have only a handful of observations but for other curves we may have hundreds. Simple parametric methods work well with small datasets while larger datasets require the flexibility of nonparametric methods. But Gaussian process regression works well for both.

3. We know the variance of the measurement error. This information is easily incorporated into the Gaussian process regression but it is not obvious how to take advantage of this information in the standard methods.

3.1. Gaussian Process Regression

Ref. [16] provides a general introduction on Gaussian Process Regression. The introductory article in ref. [17] also has some lightcurve examples. This method requires that we specify a prior for the Gaussian process: $f(x) \sim GP(\psi(x), k(x, x'))$. We use the popular squared covariance kernel:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x - x')^2\right) + \sigma_n^2 \delta(x - x'), \quad (2)$$

where $\delta(x)$ is 0 when $x \neq 0$ and 1 when $x = 0$. Other reasonable choices of kernel are possible but we found the results were similar (see online appendix for details). One advantage of the Gaussian process approach is that an explicit solution for the posterior is available that can be rapidly calculated. We will need to classify large numbers of future lightcurves as the measurements are collected so we need efficient methods.

There are four components of the prior which must be specified:

1. σ_f^2 is the signal variance. A very large fraction of objects to be classified in the future will be *non-variables*. These *non-variables* vary in signal but not very much. For this reason, we set σ_f^2 to be the median observed variance in the *non-variables*.

2. σ_n^2 is the noise variance. Although it is uncommon in other applications, for astronomical data we are often able to estimate the measurement error. In this example, the measurement error varies a little from case to case. For simplicity, we take the mean observed value of the measurement variance for σ_n^2.

3. l is sometimes called the length-scale. It controls the amount of correlation and therefore the amount of smoothness in the resulting posterior fit. We use a value of 140 days as seen in Figure 3. This choice is based on a subjective assessment on how much smoothness should be expected in these curves. Our classification performance is not very sensitive to this choice.

4. $\psi(x)$ is the prior mean. This choice is challenging and requires further discussion below. We take an empirical Bayes approach.

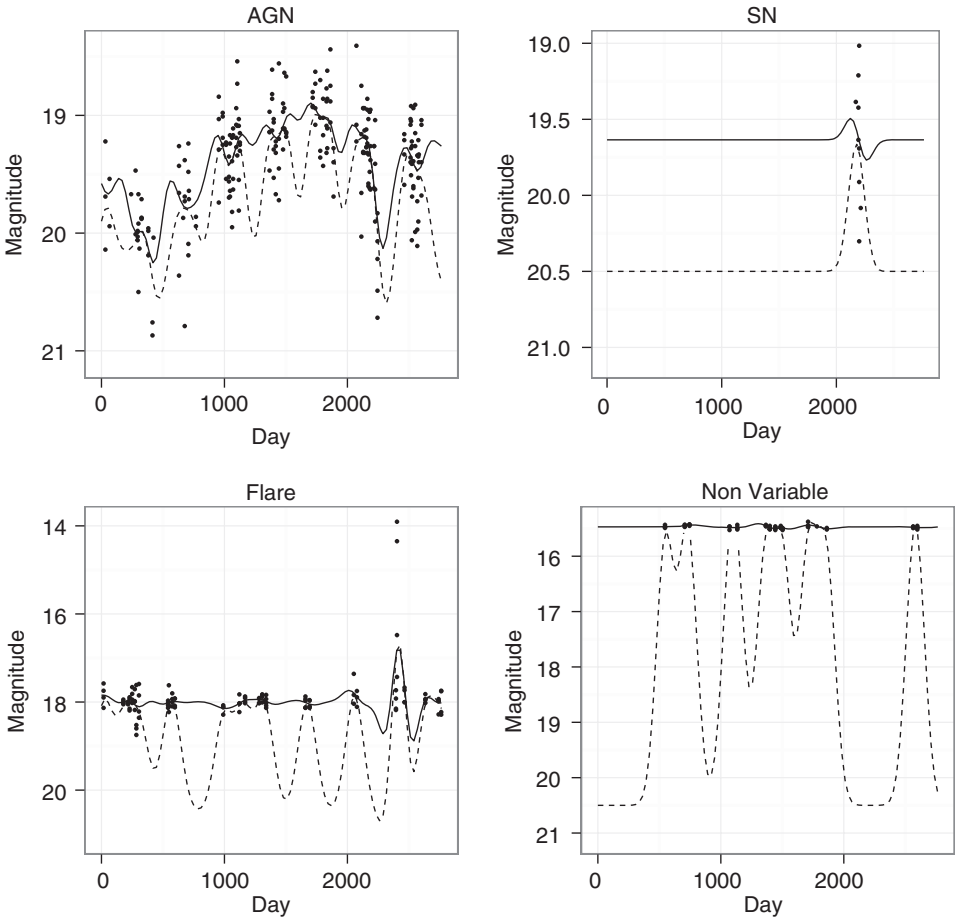

Fig. 3 Gaussian process regression fits to lightcurve data. (a) CSS071216:110407-045134, an AGN (b) CSS110405:141104+01115, a Supernova (c) CSS111103:230309+40060, a Flare and (d) 301904800767, a *non-variable*. The solid line fit derives from a prior mean set at the median magnitude while the dashed line fit corresponds to a prior set at a magnitude of 20.5.

We illustrate the issues in setting the prior mean in Figure 3. What values are expected for the curve in regions where there are no measurements? One answer is that we might expect about the same magnitude as that seen elsewhere for this object. This suggests setting the prior to the median magnitude for the object. This choice can be seen in the solid line fits in Figure 3. This works well enough in three cases but fails for the supernova example because we do not expect this curve to follow a similar magnitude at other times. If it did, it would have been seen. So an alternative approach is to set the prior to the detection limit at a magnitude of 20.5. This gives the dashed line fits as seen in Figure 3. This works well for the supernova case but is problematic for the other three curves. In regions of no measurement, the fitted curve is drawn down toward the detection limit. We can counteract this by increasing the length-scale (i.e. increasing the smoothness in the prior) but this setting tends to attenuate real effects and still does not work well for relatively sparsely measured curves (such as the *non-variable* in this example).

Our solution is to use an adaptive prior. When there is less than 1 year of observations, we use the detection limit, otherwise we use the median magnitude. The choice of a year is large enough that sparse but widely measured curves such as the *non-variable* example in Figure 3 do not use the detection limit. But the choice is small enough that the detection limit is used in cases like the SN example. Using the data to select the prior may make our method at least partly empirical Bayes rather than pure Bayes, but we need to judge the method by its classification performance which is improved by this choice.

3.2. Curve Measures

Given the posterior mean $\hat{f}$, we compute fitted curve measures from $\hat{f}_i$ for curve i computed on an evenly spaced grid of values on the range of observation u_j for $j = 1, \ldots, m = 300$. The italicized word is the name of the variable for future reference:

- *totvar* total variation: $\sum_j |\hat{f}_i(u_{i,j+1}) - \hat{f}_i(u_{ij})|/m$
- *quadvar* quadratic variation: $\sum_j (\hat{f}_i(u_{i,j+1}) - \hat{f}_i(u_{ij}))^2/m$
- *famp* amplitude of fitted function: $\max_t \hat{f}_i - \min_t \hat{f}_i$.
- *fslope* maximum derivative in the fitted curve: $\max_t |\hat{f}_i'|$

We also use the maximum in absolute value of the scaled residuals from the fit, called *outl*, as a measure.

Another feature of this data is the clustering of times of measurement which can occur in groups of up to four observations that are spaced by 10 min within a 30-min period. The Gaussian process regression is not able to model the variation at this finer scale because setting the length-scale l to a much smaller value would result in too rough a fit overall. We need another set of measures to capture the characteristics at this scale of measurement.

We compute the mean within each of these groups of up to four observations as $\breve{f}_{ij}$ and then compute the following measures:

- *lsd*: the log of the standard error, $\breve{\sigma}$, computed using the residuals from these group mean fits.
- *gtvar*: The group total variation $\sum_j |\breve{f}(t_{i,j+1}) - \breve{f}(t_{ij})|/n_i$
- *gscore*: $\sum_j \phi((\breve{f}_{ij} - \overline{f}_i)/\breve{\sigma})/n_i$ where ϕ is the standard normal density, $\overline{f}$ is the mean of the fitted group means.

The last measure is motivated by scoring methods used to judge prediction performance.

There are also some gaps within the Richards measures set of sample curve summary measures. We add the following:

- *shov*: mean of absolute differences of successive observed values: $\sum_j |y_{i,j+1} - y_{ij}|/n_i$
- *maxdiff*: the maximum difference of successive observed values: $\max_j |y_{i,j+1} - y_{ij}|$
- *dscore*: the density score: $\sum_j \phi((y_{ij} - \tilde{f}_i)/s_{ij}))/n_i$ where $\tilde{f}_i$ is the median observed magnitude for curve i and s_{ij} is the observed measurement error at t_{ij}.

There are other measures that may be informative for the current data we are analyzing but may not have predictive value in future examples. We have avoided using such measures. They fall into three categories:

1. Measures based on the number of observations in a lightcurve. Some phenomena, such as supernovae, are not recurrent and subsequent observations may fall below the observable limit. Lightcurves in such cases can be quite short but this is known only in retrospect so this is not usefully predictive. The number of observations does have some impact on the choice of prior and in scaling some of the measures, but we refrain from using this number (or anything closely related) as a direct measure for classification purposes.

2. The classification of an object should be invariant to the addition of a constant to the observed magnitudes. But some biases in the way that our example data was extracted would cause, say, the mean magnitude, to be an effective discriminator among the types. This mean magnitude will not be reproducible in future samples so we do not use this measure or anything related to it.

3. Location in the sky. The method of constructing our example dataset would mean that location would become useful discriminator. As it happens, location would provide some usable information for classification as extra-Galactic objects are more likely to appear away from the Galactic plane but we refrain from using this information here.

There is additional information such as the nearest radio source or the nearest galaxy which could also be useful in classification but we do not use this here. We experimented with a larger set of additional measures (as can be seen in the online appendix) but we have presented only those that appear to have some additional value for classification.

Given this set of measures, we can use any number of classification methods to distinguish objects using lightcurves. We demonstrate the use of our measures using five popular classification methods. We will generally use the default choice of options for the particular implementation in *R*. Our objective is to show that our measures represent an improvement over using the Richards measures alone. It is likely that the classification methods could be better tuned to obtain a better result or that the reader may favor another classification method. But that is not the point of this article. We are not trying to claim one classification method is better than another, just that our measures are better.

The methods we have used are:

LDA Standard linear discriminant analysis method as implemented in ref. [18].

TREE Recursive partitioning as implemented in the `rpart` package by ref. [19].

SVM Support vector machines as implemented in the `kernlab` package by ref. [20].

NN Neural network as implemented in the `nnet` package by ref. [18]

RF Random forest ensemble implemented by the `randomForest` package by ref. [21]

We log-transformed the measures that have extreme skewness in order to improve classification performance. The same transformations were used in all the comparisons below. Without these transformations, both sets of measures would perform less well in general for methods LDA, SVM and NN. The partitioning-based methods, TREE and RF, are invariant to monotone transformations. Explicit details of the implementation can be found in the Appendix.

4. RESULTS

Classification methods usually do not perform as well as expected when applied to new data. When the same data are used to both fit and evaluate a method, the classification rate is inflated. To avoid this problem, we randomly split the data into two-third for training, that is used to develop the classification rule, and one-third for testing, that is to evaluate how well the rule performs. Since we are only interested in the relative performance of the classification measures and methods and because the sample size is relatively large, we present only one random split. In the online appendix, we repeat the calculations for 100 random splits and the results are not qualitatively different.

4.1. Classification Performance

We considered four different types of classification problem with bold labels used for future reference:

All The overall problem of classifying eight types the *non-variables* and the seven *variable* types.

***Variable* or not** Perhaps the first step in any lightcurve classification process will be to determine which objects are *variable*.

***Variable* only** Having separated out the *non-variables*, the next step might be to identify the type of the *variable*. For this problem, we delete the *non-variables* from both the test and training data.

Table 2. Percentage correctly classified in the test set using the Richards measures as the first number of each pair and using our measures in addition to the Richards set as the second number in the pair.

	LDA	TREE	SVM	NN	RF
All	56.7:76.0	58.6:71.9	66.1:80.2	63.3:79.6	67.3:80.5
Variable or not	74.7:90.4	79.5:88.4	81.0:92.0	75.2:91.6	82.5:91.8
Variable only	54.5:70.1	58.9:65.1	64.4:74.3	60.1:72.3	62.9:74.2
Hierarchical	56.4:76.0	60.4:72.7	64.7:79.9	58.8:78.5	65.6:79.8

Hierarchical An alternative approach in classifying all objects directly is to first classify objects into *variable* or not *variable*, then if *variable* to classify among the seven available types.

We show the percentages correctly classified in the test set using the Richards measures and using our measures (which incorporate the Richards measures) in Table 2.

The standard error for the classification rate is just less than 1% which is helpful in judging which differences are notable in these table. Table 2 shows that our measures provide a significant improvement to the Richards measures alone which might be regarded as the previous state of the art. Of course, adding additional measures can only improve the fit of a model, but we are using an independent test set so we can be sure the improvement is more than illusory. There is little to distinguish the hierarchical approach from the one-step method although we would recommend the hierarchical approach on an unbiased sample of lightcurves because these would be dominated by *non-variables*. The percentages are not estimates of how the methods will perform on unidentified objects because we deliberately overloaded the dataset with *variable* types for reasons explained in Section 2. True and false positive rates can be calculated based on assumptions about the frequency of the object types within the Universe.

Table 3 shows the numbers of objects classified in the test set into all eight types compared to their actual types. We present only the random forest results as this was the best performing method. The most noticeable differences between the two sets of measures is that our method results in classifications in all eight types while the older set fails to classify any objects into four of the types. Since the Downes set is just another form of CV, we are not so concerned about a failure to distinguish these two. We can see that Flares are hard to identify.

Variables constitute less than 1% of the population. Hence, even a null method which classified randomly based on prior proportions would achieve around 99% accuracy. Certainly any sensible method will do even better than this and it would take a very large random sample to distinguish different methods. This explains why we have used a more balanced, although non-proportional, representation of the eight types to make an effective comparison of the measures

Table 3. Confusion matrix showing numbers classified for the test set using the Richards measures (first of pair) and our measures (second of pair) with random forest classification of all eight types. The rows are the predicted types while the columns are the actual types. *NV* = *non-variable*.

	Actual types							
Predicted	AGN	Blazar	CV	Downes	Flare	NV	RR-Lyrae	SNe
AGN	**0:31**	0:3	0:0	0:2	0:0	0:2	0:0	0:2
Blazar	0:0	**0:27**	0:3	0:7	0:0	0:0	0:0	0:0
CV	5:2	26:4	**95:93**	53:26	4:0	20:4	3:2	27:14
CV Downes	0:1	0:2	0:15	**0:58**	0:0	0:7	0:5	0:0
Flare	0:0	0:0	0:0	0:3	**0:8**	0:0	0:0	0:0
NT	31:8	7:0	26:9	73:25	16:15	**497:541**	47:1	80:16
RR-Lyrae	0:0	1:1	2:0	3:7	0:0	12:0	**53:95**	3:0
SNe	8:2	5:2	22:25	6:7	4:1	29:4	0:0	**82:160**

and classification methods. Similar strategies are used in case-control studies.

Because these classification methods will be applied to very large numbers of objects, even quite low error rates will result in large numbers of misclassified objects, resulting in wasted resources or missed opportunities. For this reason, the primary classification into *variable* against *non-variable* is particularly important. We can see our proposed measures perform well in this respect, halving the previous error rate, although there remains further room for improvement.

4.2. Testing on Fresh Data

We verified the performance of our measures by applying the methodology to two datasets distinct from the original set used above. One focuses on more recently discovered transients while the other consists of a very large sample of unclassified objects. We assembled 574 transients that have recently been manually classified, consisting of the types AGN, Blazar, CV, Flare and SNe. This is a complete set of CSS transients from 2013 where astronomers using additional auxiliary data were reasonably confident of the nature of the transient. We used all the objects of these five types from the original set of data (updating them to include more recently collected observations to match the lightcurve durations) to train classification rules. We then applied these rules to a test set formed from the recent set of transients. The classification rates from this exercise are shown in Table 4. The addition of our set of measures results in an improved performance over the Richards set alone.

We also assembled a large set of 100 000 lightcurves from CSS. Using the Stetson J (S_J) measure (a weak classifier defined in ref. [22]), we sampled 50 000 objects from $0 < S_J < 0.01$ and labeled these *non-variables*, and 50 000 more from $0.5 < S_J < 1$ which we labeled as *variable*. Any labeling based on a single variable will not be entirely reliable so we have chosen ranges from the extremes of S_J to increase confidence that the labels are correct. See ref. [23] for background. The lightcurves in CSS are not in general labeled so the use of S_J for this purpose, which does not appear in our set of measures, is artificial (and unavoidable) but is sufficient for a comparison of classification performance between sets of measures. Very few *variables* are likely to be transients and most of them will be of types not already considered earlier. We randomly divided the data into two equal parts. One half was used to train classification rules and the other half to test the performance. The classification rates are shown in Table 5 where we see that the addition of our set of measures to the Richards set greatly improves the rates.

Table 4. Percentage of the new set of five transient types correctly classified.

	LDA	TREE	SVM	NN	RF
Richards	60.8	64.0	72.6	71.1	73.1
Ours+Richards	74.6	67.4	78.0	77.2	79.0

Table 5. Percentage of 50 000 *variables* and *non-variables* correctly classified.

	LDA	TREE	SVM	NN	RF
Richards	75.5	79.4	85.3	76.8	85.5
Ours+Richards	97.5	89.2	98.4	98.3	96.8

4.3. Feature Selection

The random forest method provides a means of determining the worth of predictors. Suppose that within a particular node, the proportion classified as type i is given by p_i. The Gini index is defined as $1 - \sum_i p_i^2$ and can be used as a measure of node impurity. It is minimized when only one type is seen and maximized when each type is seen in equal proportion. We can see how much the Gini index averaged across nodes decreases when a measure is removed from the current set. We remove the measure which leads to the least decrease. We refit the model with the reduced set and repeat the process until all measures have gone. The classification rate after each measure is removed is shown in Figure 4 for the problem of classifying all eight types. See also ref. [24] for similar procedures.

There is little difference in classification accuracy between the training and test datasets which is a good indication that we are not over-fitting. We see that this recursive feature selection process removes most of the older nonmodel-based measures without any noticeable loss in classification accuracy. Our fitted curve measures *quadvar*, *famp*, *totvar* and *fslope* are among the most useful classification variables. Hence we can see that deriving

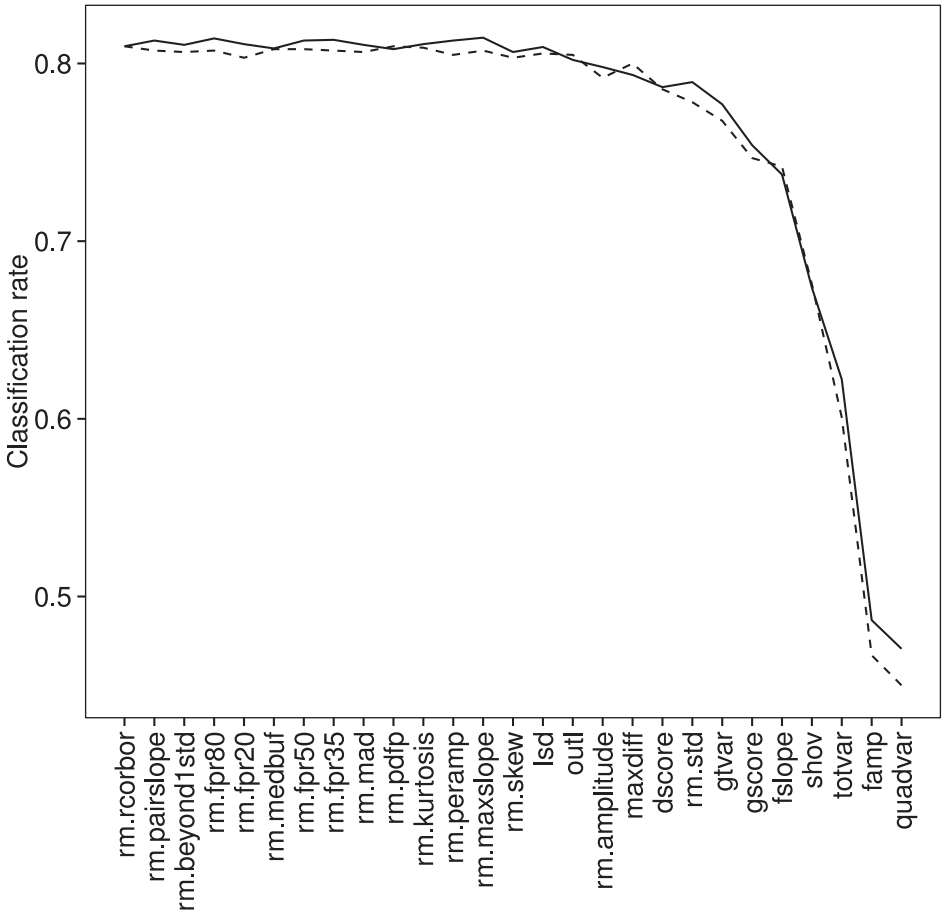

Fig. 4 Recursive feature selection of measures. The classification rates (solid for the training set and dashed for test set) are shown after the named measure is removed from the model. The order in which the measures are removed is determined by the least decrease in the Gini index. Measures from the Richards set have the prefix `rm`.

measures based on our model is a good place to start and not merely a way to supplement existing older measures. The message from the other three classification problems is similar. We do not propose that the older measures be discarded as they are generally simple to compute and may be useful in new situations.

5. DISCUSSION

We have presented two advances. In statistics, we have shown how Gaussian process regression can be adapted and the corresponding priors developed to deal with data of varying sampling density, structures and scales. We are able to deconvolute the underlying curves f_i and measurement noise ϵ_i to distinguish the objects rather than use summary statistics that use samples y_{ij} with mixed up noise and signal. With some further effort we may be able to show that the measures based on our estimated curves, under some regularity conditions, would be consistent for the features of the true underlying curves. An approach based on summary statistics can be biased or inconsistent for these measures. We believe that the modeling is the reason for our success in classification. In astronomy, we have demonstrated a new method of generating measures representing features of lightcurves that are significantly better in classifying objects than previous methods.

There is further scope for improvement in performance by optimizing the classification using routine methods. With more detailed information about when locations were surveyed but no object observed, we can further refine our priors to obtain superior results. The measures we have developed are now being used for several purposes. We can apply the method for new data where the location has only been surveyed for a shorter period of time. The measures can all be scaled appropriately. We have experimented by taking time-wise subsets of this data and have found that although the absolute performance drops with shorter curves, the relative performance over the older set of measures remains. Furthermore, the measures provide

the means to detect objects of unknown type because our measures are sufficiently informative to identify objects that do not belong to known classes by using cluster analysis. This can in turn lead to the discovery of newer classes as well as rarer counterparts of known classes of astronomical objects. CRTS images are available for the entire dataset. When rare or unusual objects are found these can be compared against the images to ensure that no artifacts or spurious features have led to the object being an outlier. The whole process can be automated and human intervention required only at critical junctures like spectroscopic confirmation. Thus, our technique can scale to much larger datasets including the 500 million lightcurves that CRTS now has. By adding a richer and more powerful set of measures, we have increased the potential for interesting discoveries. Some of our measures have already been used in classifying new lightcurves.

ACKNOWLEDGEMENTS

This work is one of results from the Imaging Working Group at the *SAMSI's 2012-13 Program on Statistical and Computational Methodology for Massive Datasets*. We thank SAMSI for bringing us together and for their financial support. The CSS survey is funded by the National Aeronautics and Space Administration under Grant No. NNG05GF22G issued through the Science Mission Directorate Near-Earth Objects Observations Program. The CRTS survey is supported by the U.S. National Science Foundation under grants AST-0909182 and AST-1313422. We are also thankful to the Keck Instutute of Space Studies, the Indo-US Science and Technology Forum (IUSSTF), and part of the work was supported through the Classification grant, IIS-1118041. We thank SG Djorgovski for useful comments and AJ Drake and MJ Graham for help in assembling the 100K dataset.

APPENDIX

Our data, code and detailed results are available as a supplement to be found at
`people.bath.ac.uk/jjf23/modlc`

REFERENCES

[1] J. Blomme, et al, Automated classification of variable stars in the asteroseismology program of the Kepler space mission, Astrophys J Lett 713(2) (2010), L204–L207.

[2] D. R. Ciardi, et al, Characterizing the variability of stars with early-release Kepler Data, Astronom J 141(4) (2011), 108.

[3] J. W. Richards, et al, On Machine-learned classification of variable stars with sparse and noisy time-series data, Astrophys J 733(1) (2011), 23.

[4] A. A. Mahabal, et al, Discovery, classification, and scientific exploration of transient events from the Catalina Real-time Transient Survey, BASI 39 (2011) (2011), 387–408.

[5] S. G. Djorgovski, et al, Flashes in a star stream: automated classification of astronomical transient events. In Proceedings of 2012 IEEE 8th International Conference on E-Science, 2012.

[6] J. S. Bloom, et al, Automating discovery and classification of transients and variable stars in the synoptic survey era, PASP 124 (2012), 1175.

[7] J. W. Richards, et al, Active learning to overcome sample selection bias: application to photometric variable star classification, Ap J 744 (2012), 192.

[8] A. W. Blocker and P. Protopapas, Semi-parametric robust event detection for massive time-domain databases, arxiv 1301.3027 (2013).

[9] M. J. Graham, et al, A novel variability-based method for quasar selection: evidence for a rest-frame 54 d characteristic time-scale, MNRAS 439 (2014), 703.

[10] K. Borne, Scientific data mining in astronomy, arXiv 0911.0505 (2009).

[11] N. Peng, et al, Selecting quasar candidates using a support vector machine classification system, Mon Not R Astron Soc 425 (4) (2012), 2599–2609.

[12] A. J. Drake, et al, First results from the Catalina Real-time Transient Survey, Astrophys J 696 (2009), 870–884.

[13] S. G. D. Djorgovski, et al, The Catalina Real-Time Transient Survey (CRTS), arXiv1102.5004 (2011).

[14] R. A. Downes, et al, A catalog and atlas of cataclysmic variables: the final edition, J Astron Data 11 (2005), 2.

[15] J. Ramsay and B. Silverman, Functional Data Analysis, (2nd ed.), New York, Springer, 2005.

[16] C. Rasmussen and C. Williams, Gaussian Processes for Machine Learning, Cambridge, MIT Press, 2006.

[17] S. Roberts, et al, Gaussian processes for time-series modelling, Phil Trans R Soc A 371 (2013), 1–27.

[18] W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, (4th ed.), New York, Springer, 2002.

[19] T. Therneau, B. Atkinson, and B. Ripley, rpart: Recursive Partitioning, 2013, R package version.

[20] A. Karatzoglou, et al, kernlab—an S4 package for kernel methods in R, J Stat Softw 11 (9) (2004), 1–20.

[21] A. Liaw and M. Wiener, Classification and regression by randomForest. R News 2 (3) (2002), 18–22.

[22] B. Stetson, On the automatic determination of light-curve parameters for cepheid variables, Publ Astron Soc Pac 108 (1996), 851–876.

[23] A. J. Drake, et al, The Catalina surveys periodic variable star catalog, arXiv1405.4290 (2014).

[24] C. Donalek, et al, Feature selection strategies for classifying high dimensional astronomical data sets, CoRRabs/1310. 1976 (2013).