

# Searching for exoplanets in the Kepler public data

Xiaofan Jin

David Glass

December 12, 2014

## Abstract

NASA’s Kepler mission to search for extrasolar planets has collected data from hundreds of thousands of star systems, and has discovered nearly 1000 confirmed exoplanets to date in addition to over 3000 unconfirmed candidates. The mission detects exoplanets using transit photometry, which detects the transit of a planet in front of a star as transient drops in stellar intensity. Raw data is collected in the form of a sequence of stellar images, which are processed into “light-curves” tracking the brightness of a star over time. An algorithm automatically searches for periodic planetary transits in these light curves, but spurious intensity dips and other noise in the data due to non-planetary stellar variability has led to high false-positive rates for detecting transits. As the initial planetary candidates found by this search method require extensive and costly subsequent validation, there is a need to reduce the error rate in exoplanet candidate identification. We present here an algorithm to classify Kepler Objects of Interest (KOIs) as confirmed exoplanets or false positives, using publicly available Kepler data. The algorithm achieves a 93% accuracy, and uses previously generated features extracted from the lightcurve time-series data, as well as newly generated autocorrelation features derived from the time-series specifically during planetary transit.

## 1 Introduction

NASA’s Kepler spacecraft spent over four years collecting data on hundreds of thousands of star systems in search of exoplanets. This data was collected by taking images of a constant patch of space every 30 minutes on a continuous basis for 3 months (i.e. 1 quarter). Every 3 months the spacecraft was recalibrated and the cycle was repeated for 17 separate quarters before the Kepler spacecraft failed. From these images, the pixels corresponding to stars were isolated, and the pixel location and intensity values were recorded over time. Taken together, this generated a series of light-curves for each tracked star, from which exoplanets would be detected using transit photometry by looking for characteristic intensity dips as planets transit in front of the star.

Currently, these light curves are further processed to remove instrument-related artifacts (known as Presearch Data Conditioning or PDC), before being fed into a wavelet-based filter (known as Transiting Planet Search or TPS) that detects the periodic dips characteristic of planetary transits. These automatic algorithms generate a series of exoplanet candidates known as KOIs (Kepler Objects of Interest), and manual follow-up measurements are conducted either to confirm KOIs as exoplanets or false positives. So far,  $\sim 7000$  KOIs have been identified, out of which  $\sim 4000$  have been followed-up on. The results of the follow-up observation has yielded a large false positive rate, as only  $\sim 1000$  exoplanets have been confirmed. Note that some systems contain more than one KOI, and so accounting by stars, Kepler has identified  $\sim 4000$  stars with KOIs, out of which  $\sim 400$  have been confirmed to actually harbour exoplanets. False positives come from a variety of sources, such as intrinsic stellar variability (eg. bursting), random background objects, or from binary star systems, in which one star transits in front of another .

Given the time-consuming and costly nature of follow-up experiments (much of the analysis is done manually), better algorithms are needed to predict from the raw light-curve data whether or not a star system contains exoplanets. For our project, we sought to produce a learning algorithm that can predict whether a KOI is a false positive or a confirmed exoplanet, ideally producing as few false negatives as possible (high sensitivity so as to not miss exoplanets), while reducing the number of false positives that require expensive follow-up studies.

## 2 Data

Light-curve files for the  $\sim 4000$  classified KOIs are publicly available from NASA’s MAST servers, in the form of .fits files (an astronomy standard). These .fits files store time-series data on stellar intensity and pixel location, as well as various other data such as measurement uncertainties and data quality flags. Note that since multiple KOIs can be based around the same star, different KOIs can have identical .fits data (there is only one set of .fits time series for each star). Over the course of a 3 month-long quarter,  $\sim 4500$  data points are gathered in the time-series (one cadence every 30 minutes). In addition to the time-series data, the .fits files are supplemented by previously gathered spectroscopic and other potentially relevant information on the stars of interest (eg. size, temperature, etc), reported as 25 numerical features. Using this set of data, it makes sense to classify stars as harbouring (or not harbouring) exoplanets, rather than classify KOIs themselves.

Note that not all quarters of data are available for all stars and furthermore during each quarter there are certain times when the spacecraft faced technical difficulties for which the time-series data has missing values reported as NaN. Removing these problematic points, our final data consists of just under 2000 stars, roughly 20% of which harbour exoplanets, the rest being false positives. For each of these examples, we are able to retrieve a 25-long feature vector of stellar characteristics, as well as a set of approximately 4500-long time-series data on stellar intensity (both raw and calibrated for Kepler spacecraft artifacts) and pixel location (horizontal and vertical). To correct for the imbalance between positive and negative examples in this dataset, when training we selected a random subset of negative examples equal in size to the positive example set.

Another set of features includes a set of 42 KOI-specific measurements extracted from the time series data using previously published algorithms. These data are again available from the MAST servers, and we did not have to calculate them ourselves. These features include the estimated magnitude and duration of detected transits, estimated radius, temperature, and eccentricity of the exoplanet, and other similar features which are derived by fitting physical models to the time series data. These features are specific to the KOI (rather than to the star), so by using this data, we can directly classify KOIs as exoplanets or false positives. We noted there is some overlap between this NASA-generated KOI-specific data and the star-specific data discussed above. Furthermore, the KOI-specific data had numerous missing values; removing these examples reduced our dataset to 1200 KOI examples, roughly half of which are confirmed exoplanets.

Our initial tests used only star-specific features and sought to binary-classify stars as harbouring exoplanets or not. Our second round of tests used the KOI-specific data provided by NASA (as well as cross-correlation features we generated ourselves from the raw time-series data, see below), and sought to classify KOIs as true exoplanets or false positives.

## 3 Features and Preprocessing

### 3.1 Extracting features for prediction of stars that harbour true exoplanets

#### 3.1.1 Star-specific characteristics

We discovered that there is significant dependence among the 25 variables on intrinsic stellar characteristics (derived directly from the .fits files), so data reduction was performed by using principal components analysis and selecting the top 22 components to use as features.

#### 3.1.2 Star-specific time-series processing

Starting from the raw light-curve and stellar characteristics data, a key challenge of this problem is in extracting informative features with which to predict exoplanet existence. Given that the length of the time-series data is over 4000 data points per example, which exceeds the total number of examples (and outnumbers positive training examples by more than a factor of 10) it is virtually infeasible to directly use the numerical time-series as a feature vector for classification. To gain an intuition for other strategies of feature extraction, we looked in more detail at the raw light curve data. Strikingly, these light curve data often appear to have a periodic nature to them, though simply looking for periodic transits themselves can be very misleading (see Figure 1). Large transits can be seen in both positive and negative data, and a lack of any obvious transits is present as well in example data from both positive and negative data. We hypothesized that in addition to periodicity of the transits (or lack thereof), additional information stored in other frequencies of the light-curves might help distinguish genuine planetary transits from artifacts.

As a basic feature extraction strategy, we attempted to characterize each time series by simple global measurements, including the mean, median, min, max, standard deviation, and inter-quartile range of the light curve as a whole. Next, we tried downsampling the  $\sim 4500$  pixel intensities over time into a  $\sim 250$  long vector. We also tried feature mapping by running a fast Fourier transform (FFT) on the raw light curves, and sampling a  $\sim 250$  dimensional array of the FFT as the new feature vector. The benefit of this method is that it separates out the long-time variation in the star system’s intensity, which could be likely due to intrinsic variation in the star’s luminous output, and also isolates much of the high-frequency noise in the data collection as well. Next, we combined not just the basic light curve data but also measurements on the motion of the star centroid, which gives an estimate of the wobble caused by a potential exoplanet. That dataset is an entire time series on its own, but by measuring the correlations between that dataset (for both  $x$  and  $y$  centroid motion) and the light curves, we came up with the 6 pairwise correlations between each of the three time series to use as features.

### 3.2 Extracting features for prediction of KOIs that are true exoplanets

#### 3.2.1 KOI-specific characteristics

We discovered that there is significant dependence among the 42 KOI-specific variables and the 25 star-specific variables mentioned, so data reduction was performed by combining these into a 67-long vector and using principal components analysis to select the top 30 components as features.

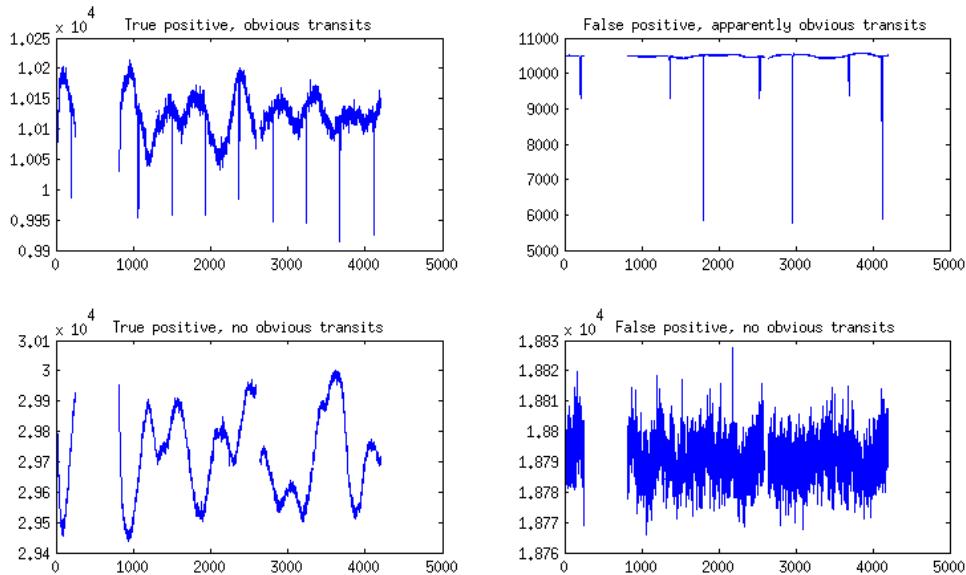


Figure 1: Light curves can be misleading. One can see that the left traces represent exoplanet-harboring stars, whereas the right traces represent non-exoplanet-harboring stars; however clear transits (or lack thereof) can be observed in both cases.

### 3.2.2 KOI-specific time-series processing

A key variable available from the KOI-specific data is the timepoint at which planet transit occurs within the raw timeseries data. Using this, we reduced our  $\sim 4500$  dimensional raw timeseries data to the 15 timepoints closest to the transit event. We then calculated centroid-to-flux cross-correlation using these short time windows, and used the corresponding 29-long cross-correlation vectors as additional features. Compared to calculating correlations with full time-series, we expected this technique to much better capture exoplanet-related wobble. To further reduce noise, we first applied a flatten / detrend algorithm to the timeseries data, using the PyKE data package (software developed by NASA for analysing Kepler data).

## 4 Models and Results

Our problem is one of binary classification. The algorithms we used are Naive Bayes, Logistic regression, and SVM (Gaussian and/or linear kernel, see below for details). The performance of these algorithms was quantified using 5-fold cross validation to generate test error rates.

### 4.1 Initial attempts at classifying stars

In our initial work, we ran a series of classification attempts on the star dataset, ultimately with marginal success. First, we began by exploring the notion of how well the star-specific features of the star system (no light curve data) could classify exoplanets. Given that certain types of stars may be intrinsically more likely than others to harbour exoplanets, the ability to predict exoplanets based on these characteristic features provides a useful baseline. After applying PCA, we find that using these features alone yields 60-65% accuracy in detection, depending on the classification algorithm used. Of the classification algorithm types tested, SVM appears to suffer from overfitting (training error  $\ll$  test error) while Naive Bayes suffers from high variance (high training/test error) so logistic regression emerges as the frontrunner. Following the classification algorithms using just the star features, we re-ran the classification algorithms, but added features extracted from the light-curve time-series in the form of global time-series statistics, downsampled time-series, FFT, and momentum correlations. Using these additional features did not increase algorithm performance with any of the three classification algorithms, as compared to using only the global star features. As before, SVM appears to suffer from overfitting (training error  $\ll$  test error) while Naive Bayes has lost almost all accuracy, and logistic regression test accuracy remains near 65%. A summary of all these techniques tried with the stellar characteristics, light curve, and centroid data is given in Table 4.1. Note we later re-analysed this data using linear kernel SVM, which yielded performance similar to Logistic regression (data not shown).

### 4.2 Classification of KOIs

As our initial attempts at binary-classification of stars did no better than  $\sim 67\%$ , we shifted our focus to the alternate strategy of classifying KOIs directly using the KOI-specific (rather than star-specific) features, applied on the dataset of

Feature mapping	Naive Bayes	SVM (Gaussian kernel)	Logistic Regression
Stellar features only	60.0/57.7	89.5/58.9	71.2/65.9
+Downsampling	51.0/50.3	81.3/60.7	71.6/66.3
+FFT	51.6/50.5	84.6/59.2	71.0/66.0
+Correlations	55.6/55.0	97.9/56.2	71.3/66.0
+Global statistics	55.2/53.7	100.0/51.3	71.2/65.1

Table 1: Summary of performance for each classification method with the feature set listed in the first column. Each table entry includes as percentages the training accuracy / 5-fold cross-validation accuracy.

KOIs (labelled as confirmed exoplanet or false positive).

As before, we began by exploring the notion of how well the characteristic KOI features (no time-series data) could classify exoplanets. After applying PCA, we find that using these features alone yields  $\sim 90\%$  accuracy in detection, depending on the classification algorithm used. Given earlier problems with overfitting using Gaussian-kernel SVM, we tried a linear kernel SVM. Of the classification algorithm types tested, logistic regression performs best with a 93% accuracy, but the performance of SVM and Naive Bayes is not far behind. Training and test accuracies are similar, suggesting that overfitting is not a problem with our current feature set. For details on performance, see Table 2, and the ROC curves shown in Figure 2. The high accuracy suggests that the currently available Kepler features can be directly used to predict exoplanets versus false positives, even without costly manual follow-up. As a side-note, we did run this dataset with Gaussian-kernel SVM, and as before we observed major over-fitting (data not shown).

Model	Accuracy (train/test %s)	F1 score	Area under ROC
SVM (linear kernel)	91.5/90.4	0.908	0.943
Naive Bayes	88.3/88.2	0.887	0.853
Logistic regression	94.3/93.2	0.933	0.968

Table 2: List of accuracies and related characteristics for the classification on stellar and KOI characteristics

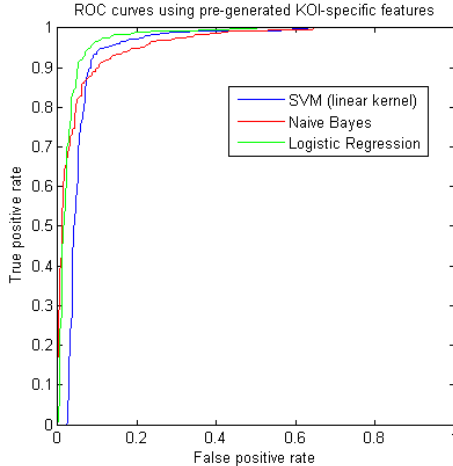


Figure 2: ROC curves for the classification on stellar and KOI characteristics

We further tested whether our features we derived for the earlier attempts could improve the classification attempts here. In particular, we supplemented the stellar and KOI characteristics with the correlations we calculated earlier between the light curves and the centroid movements. We then reran classification with the new feature set (results depicted in Table 3 and in Figure 3). No significant improvement is seen in performance with any of the three classification algorithms, suggesting that little if any predictive information is available from the cross-correlation vectors. Again, logistic regression appears to be the best-performing algorithm, followed by linear kernel SVM.

## 5 Discussion and Future directions

Our initial work for this project attempted to classify stars as harboring or not harboring exoplanets based on measurements from the Kepler spacecraft. This work is applicable to any stars which were measured by the spacecraft, regardless of whether they were picked out by NASA algorithms as Kepler objects of interest (KOIs). Despite several attempts, our best classification algorithm classified stars solely based on their intrinsic characteristics, and not by any time series data (light

Model	Accuracy (train/test %s)	F1 score	Area under ROC
SVM (linear kernel)	90.5/90.4	0.909	0.936
Naive Bayes	80.6/80.8	0.8273	0.7654
Logistic regression	95.1/93.3	0.934	0.960

Table 3: List of accuracies and related characteristics for the classification on stellar and KOI characteristics plus our correlation features

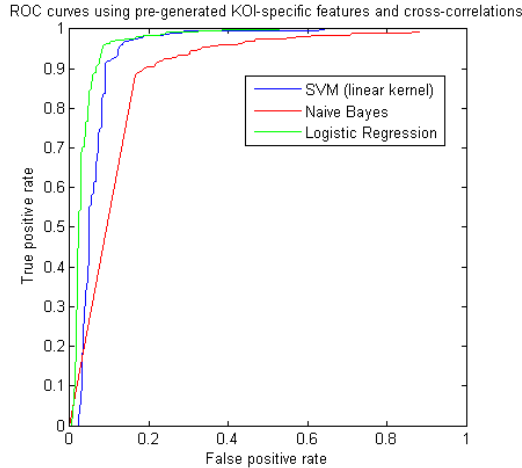


Figure 3: ROC curves for the classification on stellar and KOI characteristics plus our correlation features

curves and centroid data). Interestingly, this still achieved an accuracy of 65%, which indicates that either certain types of stars are more likely to harbor exoplanets (certainly possible given that a neutron star has little in common with a red giant, for example), but could also indicate that NASA’s algorithms for KOI detection has a bias for certain types of stars. Given the first case, we could run our algorithm on unclassified data to obtain an order-of-magnitude estimate (with 66% accuracy) for the number of planet-harboring stars. Such orders of magnitude are of interest to astronomers. Given the second case, NASA’s algorithms could benefit from a correction of such bias.

Our subsequent work for this project attempted to directly classify KOIs as confirmed or false exoplanets based on measurements from the Kepler spacecraft. As such, this work is not applicable to all stars for which Kepler took measurements, only those identified by NASA’s transit detection algorithms. Our classification algorithm can be used as another step in the planet-finding pipeline, where it serves to weed out false positives prior to follow-up measurements. Given that the manual follow-up measurements can be quite costly, there is a real need for algorithms which reduce the large number of false positive KOIs. To demonstrate that ability, we ran our logistic regression classifier with the KOI-specific feature set (ignoring the added correlation features) on 2212 currently unclassified KOIs for which sufficient data was available. Of those, 644 were classified as true exoplanets, and 1568 were weeded out as false positives. A list of the predicted exoplanets and false negatives is given in the appendix; we hope these serve as useful starting points for other planet hunters.

Given the variability in the data and relatively high accuracy achieved on KOI classification using logistic regression, it seems unlikely that a simple classification algorithm will yield significant improvements. Future possible directions could be to find better feature mappings to make use of the time-series data, such as hidden Markov models for transit detection and dynamic time-warping for detecting similarities in time-series data. Further classification algorithms can also be explored, such as random forest and neural networks for deep-learning of time-series data.

## References

- [1] D. Fraquelli and S. E. Thompson *Kepler Archive Manual*. (KDMC-10008-005) 2014
- [2] M. Still and T. Barclay, *PyKE: Reduction and analysis of Kepler Simple Aperture Photometry data*. Astrophysics Source Code Library. Provided by the SAO/NASA Astrophysics Data System 2012
- [3] K. Mandel and E. Agol *Analytic light curves for planetary transit searches*. The Astrophysical Journal. 580, 171-175, 2002
- [4] L. Walkowicz, et al. *Mining the Kepler Data using Machine Learning*. American Astronomical Society. AAS Meeting 223, 2014
- [5] T. Lee, et al. *Feature extraction methods for time series data in SAS Enterprise Miner*. SAS Institute Inc. 2014

