2009 SURF
Developing a Systematic Detrending Algorithm for Photometric Time Series Data

Giri Gopalan*
Mentor: Dr. Peter Plavchan†

October 31, 2009

Abstract

Detecting transiting planets from light curves requires high precision data. Hence, it is important to intelligently filter systematic trends affecting wide field surveys. We apply an implementation of the Trend Filtering Algorithm (TFA) to the 2MASS calibration catalog and selected Palomar Transient Factory (PTF) photometric time series data. TFA is successful at reducing the overall dispersion of light curves. However it can adversely filter intrinsic variables and increase "instantaneous" dispersion. To rectify these issues, we modify TFA by including measurement uncertainties in its computation, including ancillary data correlated with noise, and algorithmically optimizing the selection of a template set. We develop a package of MATLAB software which implements the original version of TFA along with these modifications.

1 Introduction

Recent technological advancements in astronomy, including the ability to store massive amounts of observational data, have allowed astronomers to explore the time domain in detail. In particular, these advancements have allowed for the production of detailed photometric time series data, or "light curves", which monitor the brightness of a stellar object on time-scales varying from minutes to decades. Analysis of light curves has the potential to reveal exoplanets, planets beyond our solar system which orbit other stars. If an exoplanet transits a star, or in other words moves across the face of the star with respect to our line of sight, this produces a reduction in brightness of the star. This phenomena can be seen in the light curve exhibited in **Figure 1**. To date, approximately 50 exoplanets have been detected by analyzing photometric time series data to look for transits.[10]

Photometric time series data obtained from wide field surveys is affected by systematic sources of noise. Such noise stems from a variety of sources such as varying atmospheric conditions or uncorrected instrumental effects. The ability to intelligently filter out such systematic noise is crucial to detecting transiting planets from photometric time series data. The Trend Filtering Algorithm (TFA) [1] attempts to detrend systematic noise in light curves. The algorithm leverages the fact that wide field surveys generate light curves that are affected by similar systematics. Our overall research efforts have been focused on implementing this algorithm on existing photometric time series data sets and investigating methods to improve its performance.

In § 2, we briefly discuss the 2MASS and PTF data we have applied TFA to. In § 3, we analyze in detail the original version of TFA as well as various modifications to TFA we have implemented. In § 4, we present the results of the detrending on the aforementioned data sets. In § 5 we discuss these results and conclude.

*California Institute of Technology. *email:* giri@caltech.edu
†NASA Exoplanet Science Institute. *email:* plavchan@ipac.caltech.edu

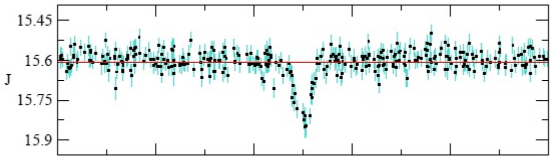

Figure 1: Light curve for a potential transit candidate [5]. The horizontal axis is phase while the vertical axis is near infrared J-band magnitude. The period is .88320 days.

2 Data Selection

We apply TFA and various improvements of the algorithm to two sets of data, the Two Micron All Sky Survey (2MASS) and Palomar Transient Factory (PTF) data. In this section we discuss the production of these data sets in detail.

2.1 2MASS Data

The Two Micron All Sky Survey imaged the entire sky in three near-infrared bands between 1997 and 2001. Photometric calibration for 2MASS was accomplished using nightly observations of selected calibration fields around the sky. The 35 2MASS calibration fields are distributed at approximately 2 hr intervals in right ascension near declinations of 0° and ±30°. Each 2MASS calibration field covers a region 8.50' wide (in right ascension) by 60' long (in declination).

Each night during 2MASS operations, the survey telescopes were directed at one of the calibration fields once per hour. During each visit, six consecutive scans of the field were made in alternating declination directions in approximately 10 minutes of elapsed real time (a "scan group"). Each scan in the set of six was offset from the preceding scan by approximately 5" in R.A. to avoid systematic pixel effects. The calibration fields were observed using the same "freeze-frame" scanning strategy used for the main survey that yielded a net 7.8 s exposure on the sky per scan. Over the course of the 2MASS survey, between 562 and 3692 independent observations were made of each of the 35 calibration fields.

The raw imaging data from each scan of a 2MASS calibration field were reduced using the same automated data processing system used to process the survey observation data. The reduction process detected and extracted source positions and photometry for all objects in the images from each scan. Measurements of the standard stars in each field were used to determine the nightly photometric zero-point solutions as a function of time, and seasonal atmospheric coefficients. All source extractions from all scans were loaded into the 2MASS Calibration Point SourceWorking Database (Cal-PSWDB). This database contains over 191 million source extractions derived from 73,230 scans of the 35 calibration fields. Further descriptions of the Cal-PSWDB and its properties can be found in Cutri et al. (2006).

2.2 PTF Data

Preliminary test data were taken during instrument commissioning time in Feb/March 09, with the purpose of using them to build a differential photometry pipeline. Two data sets were taken. The first was an overlapped tiling of the entire Orion region, R band, 60s exposures. The second was a single-field time-series data set, R band, 30s exposures, approximately 1min cadence, over three nights. The aim of the program is to observe a field where the gas disks of young stars are actively dissipating (approximately 5–10Myr old), leaving behind any newly formed planets. The field just south of Orions belt was chosen for the time-series pilot study.

3 Analysis

In § 3.1, we begin by briefly summarizing the methodology of TFA [1] as well as the derivation of the matrix formulation of the algorithm. In the remaining sections, we analyze various methods of modifying TFA to improve its performance. In § 3.2 we discuss how one may include measurement uncertainties and recast this modification into matrix algebra. In § 3.3, we discuss two clustering based approaches to optimizing the selection of a template set, the first method from Kim, et.al [8]. In the final section we present a method of including external parameters correlated with noise known as External Parameter Decorrelation, as in Bakos et. al [7].

3.1 Formulation of TFA

We begin with the mathematical formulation of TFA [1]. Let $Y(i)$ be a lightcurve which is to be detrended, and assume it is zero averaged. TFA assumes that a filter function $F(i)$ may be constructed as a linear combination of "template" light curves $X_j(i)$ which are selected from the field surveyed, and are zero averaged as well. Implicitly, the template set contains all information of the systematic trends that the algorithm is privy to. To summarize we have the following relations:

$$\begin{aligned} F(i) &= \sum_{j=1}^{j=m} c_j X_j(i) \qquad (1) \\ Y^*(i) &= Y(i) - F(i) \qquad (2) \end{aligned}$$

where $Y^*(i)$ is the filtered light curve. Thus the formulation of a filter function is equivalent to the solution of a particular set of constants c_j. TFA solves for these constants by minimizing the following sum of residuals:

$$min \sum_{i=1}^{i=n} (Y(i) - F(i))^2 \qquad (3)$$

This minimization problem can be recast in terms of matrix algebra. Assume there exist m lightcurves in the template set, and assume each lightcurve consists of n brightness measurements. Then we may arrange the template set into an m by n matrix where each row is a lightcurve. Further, consider placing the constants $\{c_j\}$ into a length m row vector. Then we have the relation:

$$F = c * T \qquad (4)$$

Where F is the filter function viewed as a length n row vector. Using this new notation, our minimization problem now becomes:

$$min||Y - cT||_2 \qquad (5)$$

In other words, a particular choice of c corresponds to a point in the subspace spanned by the rows of T, and we seek a choice of c such that the residual distance $Y - cT$ is minimized. A basic proposition from linear algebra dictates that this choice of c will cause the vector $Y - cT$ to be normal to the hyperplane spanned by the rowspace of T. Hence we seek a c^* such that the vector $Y - c^*T$ is normal to all the rows of T. Since normality is defined by a 0 inner product, we get the following relation:

$$T * (Y - cT)^T = 0 \qquad (6)$$

Hence solving for c:

$$c = (TT^T)^{-1} * TY^T \qquad (7)$$

The salient point of this derivation is that a filter function can be calculated via basic matrix operations.

3.2 Including Measurement Uncertainties

A main drawback of TFA is that is does not rely on individual measurement uncertainties, and hence a highly uncertain measurement is treated equally to a very certain measurement. To rectify this, one may modify the key minimization problem TFA employs to weight certain measurements more than uncertain measurements.

Specifically, we now minimize the following weighted sum of residuals:

$$min \sum_{i=1}^{i=n} w_i (Y(i) - F(i))^2 \tag{8}$$

where $w_i = \sigma_i^{-2}$. As before, we may recast this problem in terms of matrices. If we define the diagonal matrix S such that $S_{ii} = \sigma_i^{-1}$, then we are minimizing the following:

$$min||(Y - cT)S||_2 \tag{9}$$

Since S is a linear map this is equivalent to the problem:

$$min||(YS - cTS)||_2 \tag{10}$$

This is of the same form as the original TFA problem and hence the solution is given by:

$$c = (GG^T)^{-1} * GH^T \tag{11}$$

Where

$$G = TS \tag{12}$$

and

$$H = YS \tag{13}$$

It should be noted that the involvement of measurement uncertainties is more computationally costly than the unmodified version of TFA. With the original algorithm, one may form a template set for an entire field of stars and and use the same template for each star filtered. However, if one corrects by measurement uncertainties the template set must be multiplied by S for each star filtered, where S is dependent on the particular star. Weighting by measurement uncertainties is mentioned in the thesis of A. Pal. [6].

3.3 Optimizing the Template Set

A criticial component of TFA is the selection of a template set, since it implicitly contains all information about systematic noise. Ideally, one would like to minimize the number of template stars while maximizing information about systematic noise. While a large number of template stars yields a large reduction in dispersion, this also yields a greater tendency to over filter intrinsic variables. This is because more free parameters in the minimization problem allows for more freedom to fit any particular light curve and potentially over filter noise and intrinsic variations. We investigate two methods of optimizing the selection of a template set, both based on clustering algorithms.

3.3.1 Agglomerative Heiarchial Clustering

Kim, et al. [8] proposes an algorithm which attempts to select a small number of template stars that well represent systematic trends. The algorithm is, in essence, an implementation of Agglomerative Heiarchial Clustering. The algorithm aims to partition stars into subsets whose light curves correlate highly with each other. The logic behind this approach is that each cluster represents a particular sort of systematic trend. Once partitioning is complete, one may extract a template star from each cluster by performing a weighted average, where weighting is done by the inverse of variance.

There are three main steps involved in this algorithm. First a distance matrix is computed for the light curves. Then, one computes a binary tree using this distance matrix. Finally, one uses the binary tree to determine clusters via a merging algorithm. We detail these steps below:

- **STEP 1**: *Compute the Distance Matrix.* First, we compute the Pearson correlation between all light curves. We store the information in a distance matrix D where $d_{ij} = 1 - c_{ij}$, where c_{ij} is the correlation between light curves i and j.

- **STEP 2**: *Compute the Binary Tree.* We then compute a binary tree using the distance matrix. Specifically, the leaves of the tree are the individual light curves. We then combine the closest two nodes under one parent, where the distance between two nodes a and b is the maximum distance between any two light curves in the nodes. We iterate this linking procedure until all light curves have been merged.
- **STEP 3**: *Determine Clusters via Merging.* Using the binary tree, one can determine clusters in the following manner. Initially we set clusters to be nodes in the tree with at least two children. We then consider merging the closest nodes to form a larger cluster. Call this potential cluster *Cmerge.* If *Cmerge* contains light curves which are correlated (i.e *Cmerge* is a good representation of a particular systematic trend) then the distribution of distances between any two light curves in *Cmerge* should follow a normal distribution. Hence, one applies an Anderson-Darling normality test to the list of distances. If the test produces a p-level below .10 (i.e, we have reason to believe the distances do not come from a normal distribution) then we stop the merging procedure, as we have evidence that the light curves are not all correlated with each other. In this fashion, one may partition all light curves into subsets which are all correlated which each other. Once the clusters are formed, one takes a weighted average of light curves in the cluster to produce a template trend.

3.3.2 KMEANS Clustering

An alternate approach to clustering is the KMEANS algorithm. If one assumes that all light curves are elements of $\mathbb{R}^n$ where n is the number of brightness measurement for each light curve, then one may formulate a notion of Euclidean distance between two light curves. Using this notion of distance, KMEANS partitions a set of light curves into k subsets where the elements of each subset are close to each other. One begins the algorithm by assigning k random points in $\mathbb{R}^n$ as centers. Next, one assigns each light curve to the center it is closest to, and in this process we partition the set into k subsets. Then, we recalculate k centers by choosing the average of each cluster as its center. We then iterate this process of reassigning light curves to clusters and recalculating centers until no new assignments have been made. A subtlety involved is the initial choice of centers, and a particularly efficient method for doing so is given by KMEANS++ [9]. Note that one must ensure light curves are zero averaged to ensure that Euclidean distance is a feasible metric.

3.4 External Parameter Decorrelation

While TFA assumes one has no apriori information regarding systematic noise, it is feasible that certain external parameters, such as seeing or position, correlate with noise. Bakos, et al. [7] suggest a method to involve external parameters which correlate with systematics. In essence, the formulation is the same as TFA, except coefficients are now chosen for the parameters via the same minimization problem.

$$min\sum_{i=1}^{i=n} w_i(Y(i) - \sum_{j=1}^{j=m} c_j X_j(i) - \sum_{j=m+1}^{j=l} c_j P_j(i))^2 \tag{14}$$

Again, we may recast this formulation into matrix algebra; we simply add additional rows to the template matrix T, where each row contains the external parameter values for each time index.

4 Results

We have written a package of MATLAB software which implements the Trend Filtering Algorithm as well as the aforementioned modifications. What follows are a series of quantitative assesments of the algorithm, using both the unmodified version and the various improvements.

4.1 Assessing the Unmodified Version of TFA

4.1.1 TFA Reduces Dispersion

By visualizing dispersion versus apparent magnitude for both the 2MASS and PTF data, we have determined that TFA reduces the overall dispersion of the lightcurves. These graphs are depicted below in **Figures 2 and 3**. Specifically, the algorithm reduced the dispersion most substantially for the PTF data.

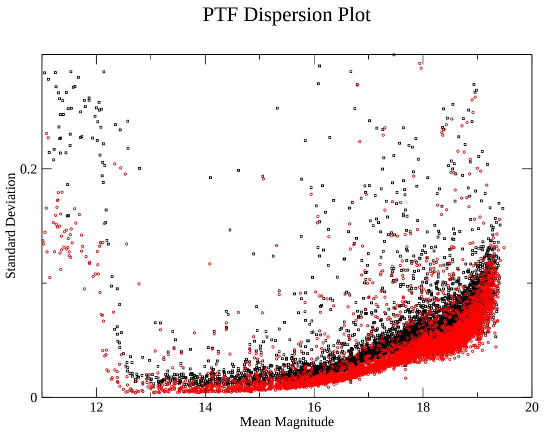

Figure 2: PTF dispersion: Black points indicate light curves before detrending, and red indicates after detrending. Mean magnitudes are of I band.

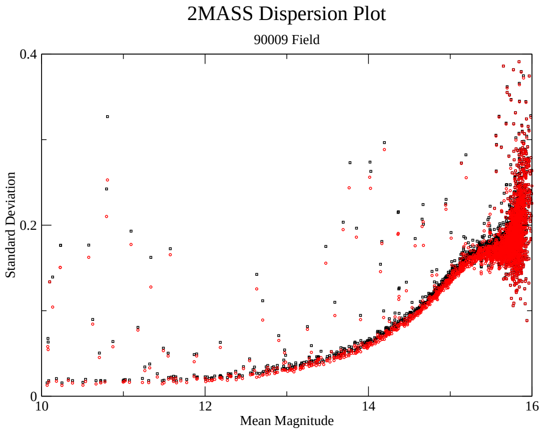

Figure 3: 2MASS dispersion: Black points indicate light curves before detrending, and red indicates after detrending. Mean magnitudes are K band.

4.1.2 TFA Adversely Filters Intrinsic Variables

Evidence suggests that TFA adversely filters intrinsic variable signals. For example, consider **Figure 4**, which presents the original and filtered light curve for a PTF variable candidate. In this example it is clear that any intrinsic variability has been flattened by the filtering algorithm. This light curve also demonstrates that TFA may be increasing its “instantaneous dispersion”. In other words, while the overall standard deviation may be reduced, the standard deviation for points taken within a small time frame has increased, which is not a desired result.

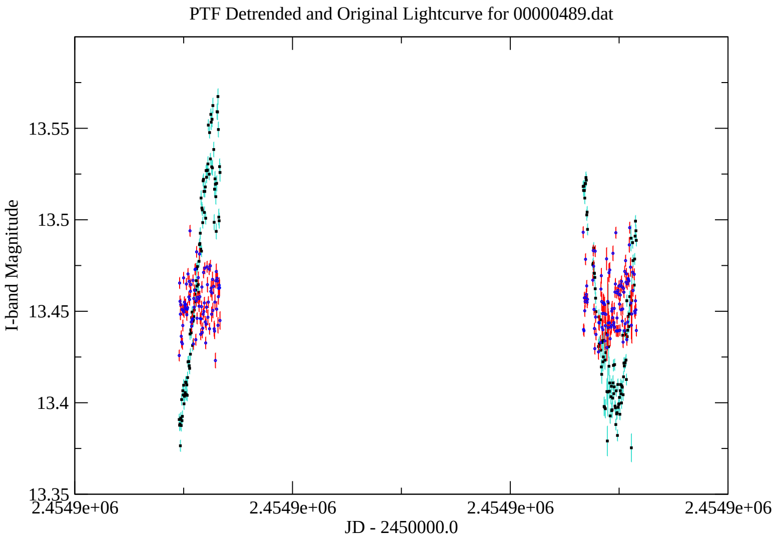

Figure 4: Detrended PTF variable candidate using the unmodified version of TFA. Note that the algorithm appears to filter intrinsic variability of the light curve, which is not desired.

4.2 Assessing the Modified Version of TFA

4.2.1 Modified TFA Reduces Dispersion

It is evident that the modifications made to TFA do not mitigate its power to reduce the dispersion of light curves. The dispersion plots presented in **Figures 5 and 6** are qualitatively identical to dispersion plots of the unmodified version of TFA. In addition to these dispersion plots, we have generated histograms which illustrate the distribution of dispersion improvements. In particular, **Figure 7** present histograms of dispersion improvements. These histograms illustrate both absolute and relative dispersion improvements. Absolute improvement is defined as the difference between the dispersion of light curves post and pre TFA. Relative dispersion improvement is defined as the absolute improvement divided by the post TFA dispersion. From these histograms, we can derive quantitative measure of the modified TFA's performance. On average, dispersion was reduced by 30 percent for the PTF data set and 1.5 percent for the 2MASS data set. The lack of significant improvement to most 2MASS light curves indicates that the data reduction for 2MASS and callibration is thorough in removing most systematics. However, significant improvements can be obtained in special cases, such as extended or confused sources.

4.2.2 Modified TFA Does Not Adversely Filter Intrinsic Variables

Evidence suggests that the modifed version of TFA no longer adversely filters intrinsic variables. For example, consider the detrended PTF light curve in **Figure 8** which was previously over filtered by the unmodified version of TFA. The potential variable is essentially unchanged. In addition, the problem of increased instantaneous dispersion is also alleviated, although it still persists to a slight extent. This success can be most likely attributed to the intelligent selection of a few template trends. The utilization of many template curves allows for many free parameters in the key minimization problem that TFA employs. In turn, this allows any arbitrary light curve to be fit by the trend curves.

4.2.3 Template Optimization

In **Figure 9** we provide a sample of five trends produced by the Heiarchial Agglomerative Clustering algorithm. This is meant to serve as an illustrative example of the template optimization scheme.

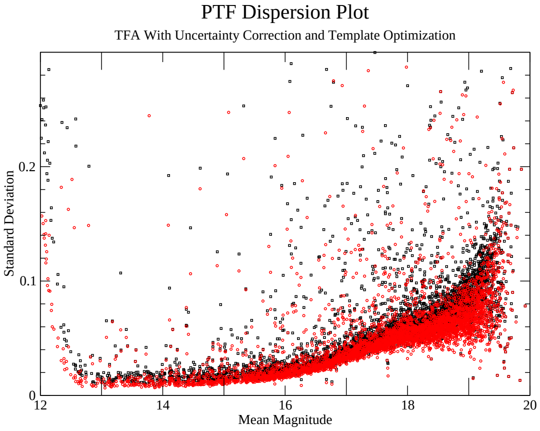

Figure 5: PTF Dispersion Plot, using the modified version of TFA which included measurement uncertainties and template optimization. Mean magnitudes are I band.

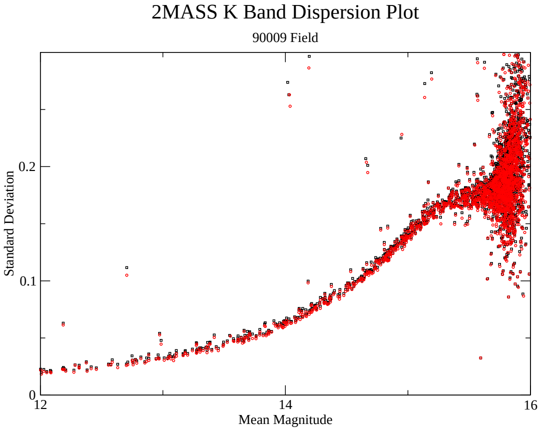

Figure 6: 2MASS Dispersion Plot, using the modified version of TFA which included measurement uncertainties and template optimization. Mean magnitudes are K band.

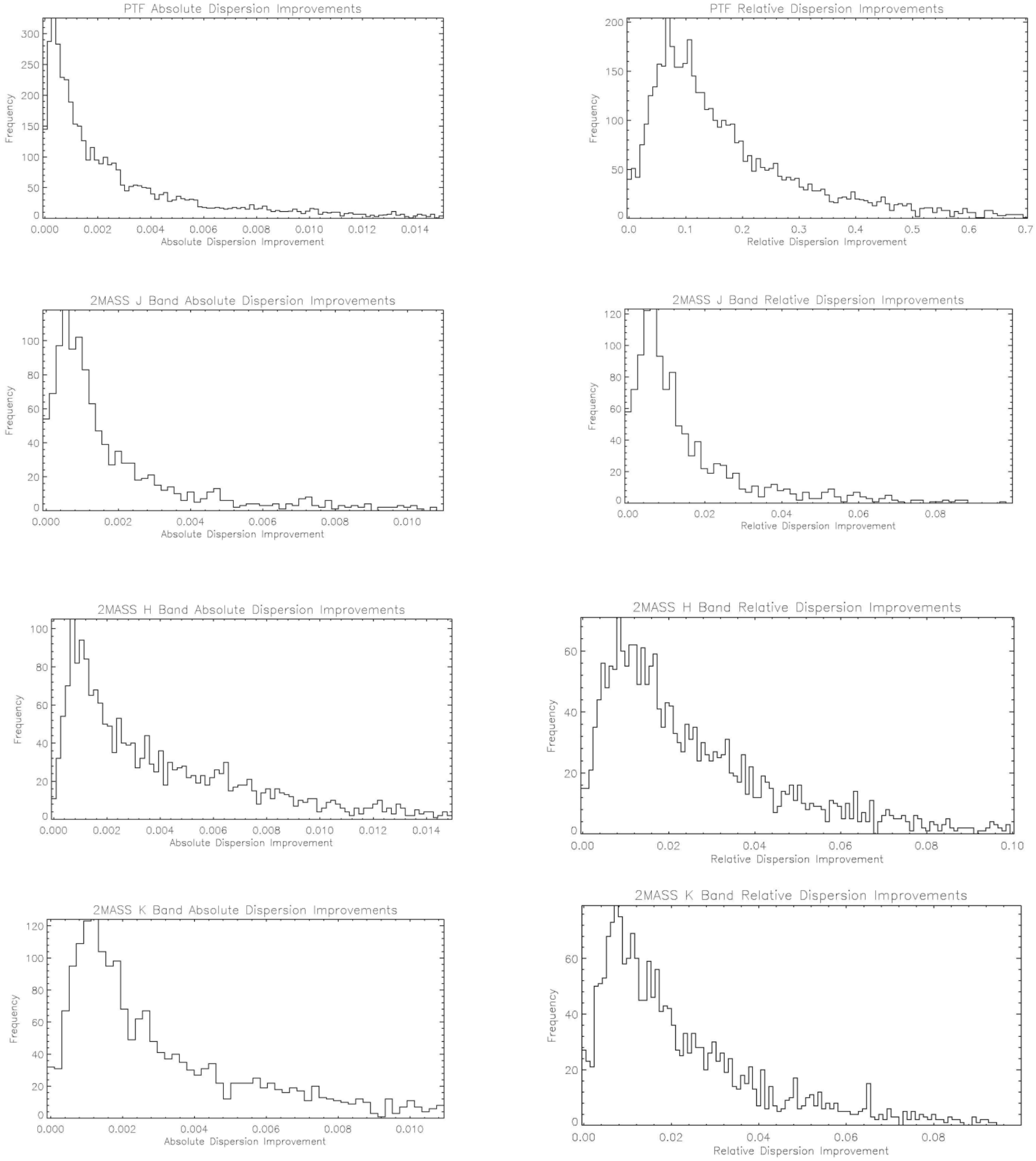

Figure 7: Histograms illustrating the distribution of absolute and relative dispersion improvements using the modified version of TFA. Histograms on the left are absolute improvements while histograms on the right are relative improvements.

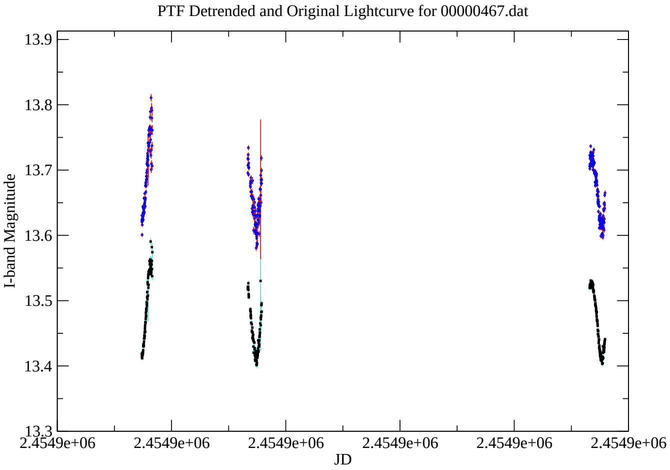

Figure 8: Example of a PTF variable which remains largely unfiltered using the modified version of TFA. Note that the detrended version is shifted .2 magnitude above the original light curve.

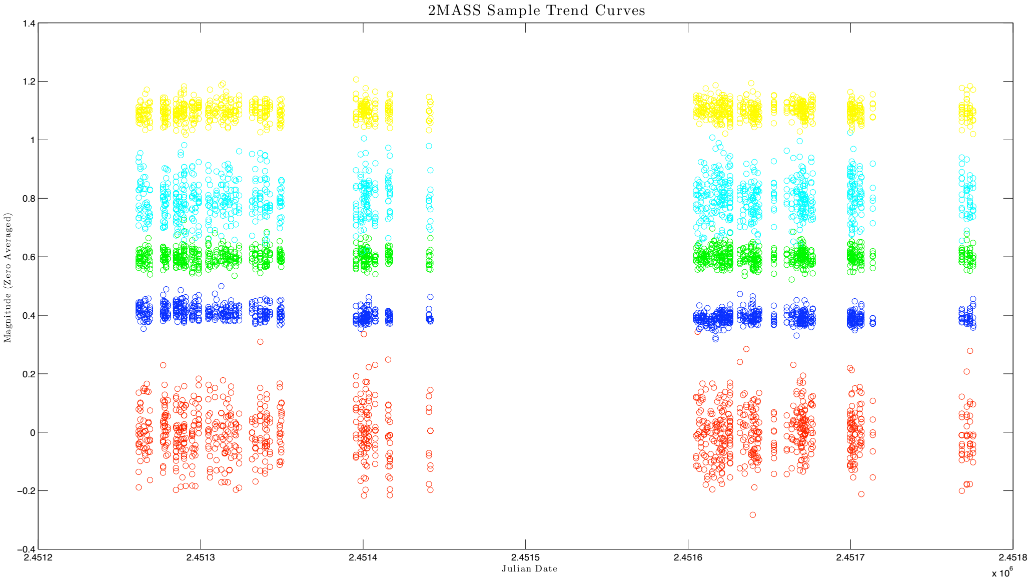

Figure 9: A set of 5 sample 2MASS trends produced using the Agglomerative Heiarchial Clustering approach.

5 Conclusion

After implementing TFA and applying it to the 2MASS calibration data and PTF pilot data we conclude that the algorithm substantially reduces the dispersion of light curves, most notably for the PTF data set. However, it is apparent that TFA may over filter intrinsic variables and increase the instantaneous dispersion of light curves. By modifying TFA to include measurement uncertainties, include ancillary data correlated with noise, and optimally select a template set using clustering algorithms, we believe that these effects can be mitigated. Preliminary implementation and application of these improvements evidences that the over filtering of intrinsic variables is alleviated, without substantially losing the power of the algorithm to reduce overall dispersion.

6 References:

[**1**] Kovacs, G., Bakos, G., Noyes, R. (2005). A Trend Filtering Algorithm for Wide-Field Variability Surveys. MNRAS. Vol. 356, 557-567.
[**2**] Kovacs, G., Bakos, G. (2008). Application of the Trend Filtering Algorithm in the Search for Multiperiodic Signals. Comm. In Asteroseismology. Vol. 157.
[**3**] Tamuz, O., Mazeh, T.., Zucker, S.(2005). The Sys-Rem Detrending Algorithm. MNRAS. Vol. 356, 1446.
[**4**] Plavchan, P., Gee, A H., Stapelfeldt, K., Becker, A. (2008). The Peculiar Periodic YSO WL 4 in p Ophiuchus. The Astrophysical Journal, Volume 684, Issue 1, pp. L37-L40
[**5**] Plavchan, P., Jura, M., Kirkpatrick, J., Cutri, R., Gallagher, S. (2008) Near-Infrared Variability in the 2MASS Calibration Fields: A Search for Planetary Transit Candidates. The Astrophysical Journal Supplement Series, Volume 175, Issue 1, pp. 191-228.
[**6**] Pal, Andras (2009). Tools for discovering and characterizing extrasolar planets. arXiv:0906.3486v1.
[**7**] Bakos, G., et al., 2007, APJ, 670, 826.
[**8**] Kim, D., et al. (2009). De-Trending Time Series for Astronomical Variability Surveys. arXiv:0812.1010v3.
[**9**] Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035.
[**10**] NStED Webpage: `http://nsted.ipac.caltech.edu/`
[**11**] PTF Webpage: `http://www.astro.caltech.edu/ptf/`