Mon. Not. R. Astron. Soc. **397,** 558–568 (2009) doi:10.1111/j.1365-2966.2009.14967.x

Detrending time series for astronomical variability surveys

Dae-Won Kim,[1,2,3]★ Pavlos Protopapas,[1,2] Charles Alcock,[1] Yong-Ik Byun[3] and Federica B. Bianco[1,4]

[1]*Harvard Smithsonian Center for Astrophysics, Cambridge, MA 01238, USA*
[2]*Initiative in Innovative Computing, Harvard University, Cambridge, MA 02138, USA*
[3]*Department of Astronomy, Yonsei University, Seoul 120-749, Korea*
[4]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*

Accepted 2009 April 23. Received 2009 April 23; in original form 2008 December 13

ABSTRACT

We present a detrending algorithm for the removal of trends in time series. Trends in time series could be caused by various systematic and random noise sources such as cloud passages, changes of airmass, telescope vibration, CCD noise or defects of photometry. Those trends undermine the intrinsic signals of stars and should be removed. We determine the trends from subsets of stars that are highly correlated among themselves. These subsets are selected based on a hierarchical tree clustering algorithm. A bottom-up merging algorithm based on the departure from normal distribution in the correlation is developed to identify subsets, which we call clusters. After identification of clusters, we determine a trend per cluster by weighted sum of normalized light curves. We then use quadratic programming to detrend all individual light curves based on these determined trends. Experimental results with synthetic light curves containing artificial trends and events are presented. Results from other detrending methods are also compared. The developed algorithm can be applied to time series for trend removal in both narrow and wide field astronomy.

Key words: methods: data analysis – methods: miscellaneous – methods: statistical – surveys.

1 INTRODUCTION

Small-aperture telescopes have detected a large number of exoplanet transits (Alonso et al. 2004; Bakos et al. 2004; McCullough et al. 2005; Bakos et al. 2007; Burke et al. 2008; Pál et al. 2008; Pollacco et al. 2008). A large number of variable stars have also been detected by surveys that use such telescopes (Schmidt 1991; Akerlof et al. 2000; Pojmanski 2005; Schmidt et al. 2007; Pigulski & Pojmański 2008; Szczygiel & Fabrycky 2007). A weakness in these surveys is that the signal-to-noise ratio (S/N) is lower than the S/N obtained by larger-aperture telescopes. The low S/N can be attributed not only to the small-aperture size but also to noise in CCD images such as non-uniform illumination, or to local weather changes throughout the field (especially in the case of wide field surveys). To improve the S/N and thus improve the detectability of variability, these noise sources should be minimized.

Some of these noise sources are strongly correlated between light curves of different stars. For example, if a star appears fainter, other stars near it may appear fainter at the same time. We call such coherent changes through parts of the field *trends*. These trends could be caused by local weather patterns such as thin cloud passages or airmass changes (Howell & Jacoby 1986; Kjeldsen & Frandsen 1992; Gilliland & Brown 1988) throughout the night. The conventional approach for trend removal is differential photometry with a *reasonable* selection of template stars near the star of interest (Young et al. 1991; Everett & Howell 2001). With the help of modern CCDs, it is not hard to select a sufficient number of bright stars as a template set. However, the detrended results are then sensitive to the selection of template stars. If the template stars contain intrinsic variables, the determined trends will be different from the true trends. Therefore, excluding such intrinsically variable stars from template stars is essential. Furthermore, because there is no guarantee that trends are the same for all stars throughout the entire field, the template selection method should be able to handle localization of trends in large fields of wide field surveys.

In this paper, we propose a new detrending method, photometric detrending algorithm (PDT), which incorporates a systematic template selection algorithm that can solve the problems mentioned above and consequently shows superior detrended results. Experiments with simulated light curves show that PDT correctly reproduces localization.

We present details of PDT in Section 2. In Section 3, we show detrended results for synthetic light curves containing artificially added trends and events. In addition, comparison results with the trend filtering algorithm (TFA) (Kovács, Bakos & Noyes 2005) are also presented. In Section 4, we show two examples of astronomical data sets and their detrended results. We outline

★E-mail: dakim@cfa.harvard.edu

future work in Section 5. We summarize our conclusions in Section 6.

2 ALGORITHM

2.1 Outline of the PDT

One of the most widely used methods for the selection of template stars is the method that chooses as a template set a sufficient number of bright stars that are not saturated, not overlapping and not at the edge of the field. Some of these bright stars could have intrinsic variability (e.g. variable or flare stars). If we avoid those stars in the selection of template set, the detrended results will be improved. Ideally, standard stars such as Landolt standard stars (Landolt 1992) could be useful as template set. However, there are not many standard stars in the field (even in the wide field surveys). Thus, one needs to choose template stars from the field, where a few per cent of stars are varying (see e.g. Paczynski & Pojmanski 2000; Everett et al. 2002).

If the light curve of a star manifests a trend without being intrinsically variable, then the light curve should be highly correlated with many other light curves of stars in that field. If a star has both trend and intrinsic variability, the light curve of the star would not be as highly correlated with other light curves. Therefore, a light curve which has strong correlation with many other light curves is a good template candidate. Our approach to the selection of template stars is to choose highly correlated subsets of stars using the similarity matrix C, in which the elements C_{ij} are the Pearson correlation values between light curves of star i and star j.

The Pearson correlation values can be calculated by the following equation:

$$C_{ij} = \frac{1}{n-1} \frac{\sum_{t=1}^{n} [L_i(t) L_j(t)] - n \overline{L}_i \overline{L}_j}{\sigma_i \sigma_j}, \quad (1)$$

where $L_i(t)$ is the flux of star i at time t, n is the total number of measurements, $\overline{L}_i$ is the mean flux of $L_i(t)$ and σ_i is the standard deviation of $L_i(t)$. The number of measurements n for every light curve should be the same.

Using the similarity matrix and a hierarchical tree clustering algorithm explained below, we can extract multiple subsets of template stars; each subset is relatively highly correlated within itself but not with any other subsets. We call the subsets *clusters*. For each extracted cluster, we determine one representative trend light curve by the weighted sum of all light curves from that cluster. To remove the trends from all light curves, we minimize the residuals between each light curve and the determined trends by minimizing the rms r_i,

$$r_i = \sqrt{\frac{1}{n} \sum_t \left[L_i(t) - \lambda_i - \sum_k^m \beta_{ik} T_k(t) \right]^2}, \quad (2)$$

where n is the total number of measurements, $T_k(t)$ are the determined trends for cluster k, m is the total number of clusters, β_{ik} and λ_i are free parameters to be calculated for each light curve. For more details about the minimization process, see Section 2.3.

Sometimes such minimization approaches remove not only trends but also the intrinsic signals because one can adjust the free parameters such that the summed trends resemble the signals. This side effect is more significant when there are more free parameters to be adjusted. Therefore, PDT, which identifies one representative trend per cluster and thus has a small number of free parameters, is better suited for detrending light curves, especially where the rms contribution from the intrinsic signal is significant. This contrasts with TFA or similar methods that assign one free parameter per template star per individual light curve.

In the following sections, we explain how we use the similarity matrix to choose the clusters and how we detrend light curves using the selected clusters.

2.2 Selection of clusters of light curves

First, we summarize some traditional clustering algorithms and their shortcomings in Section 2.2.1. We then explain a selection method for choosing clusters of light curves using a hierarchical tree clustering algorithm, which is more suitable than other clustering algorithms. The selection method consists of two processes. The first step is the construction of a hierarchical tree according to the similarity matrix, explained in detail in Section 2.2.2. The second step is the extraction of clusters from the constructed hierarchical tree using the normality test explained in Section 2.2.3.

2.2.1 Clustering algorithms

In order to extract clusters of template stars, we first group stars using a clustering algorithm based on the similarity matrix. Clustering algorithms are useful for grouping large data according to their similarities (Jain, Murty & Flynn 1999).

We have examined several clustering algorithms, such as density-based clustering (Ester et al. 1996), K-mean (Hartigan & Wong 1979), K-medoids (also known as Partitioning around Medoids or Clustering Large Applications based on Randomized Search (CLARANS), Ng & Han 1994) (hereafter K-methods) and a hierarchical tree clustering algorithm (Jain et al. 1999). These algorithms first define distances between each element (light curves in our case) and then group elements that are similar to each other based on the distance. For all our testing, we used a distance matrix in which the elements are defined as

$$D_{ij} \equiv 1 - C_{ij}, \quad (3)$$

where C_{ij} are the Pearson correlation values between two elements i and j, as shown in equation (1). More correlated, or more similar elements have shorter distances between them.

In choosing template sets, it is important while grouping elements that every element in the same cluster is similar to the others in the cluster. However, some of the clustering algorithms (Hartigan & Wong 1979; Ng & Han 1994; Ester et al. 1996) group into cluster elements which are not pairwise similar. This is a critical disadvantage because we would like to identify only stars that are strongly correlated to one another. Fig. 1 conceptually illustrates the problem. The x- and y-axes indicate the distances between pairs of elements, where closer elements are more similar. By means of these clustering algorithms, one can easily identify the two clusters, C_1 and C_2, in Fig. 1. Yet, some elements in the cluster C_1 are not close to other elements in the same cluster because the cluster C_1 is stretched along the diagonal direction. For example, the bottom left elements are far from the top right elements, even though they are in the same cluster. With the exception of the hierarchical tree clustering algorithm, the clustering methods mentioned above suffer from these disadvantages.

Note that the term 'cluster' in this paper is not used in the conventional way, where C_1 in Fig. 1 would be considered as a cluster. In the rest of the paper, we will be using the term 'cluster' to designate 'zone of influence' which means a group of strongly correlated

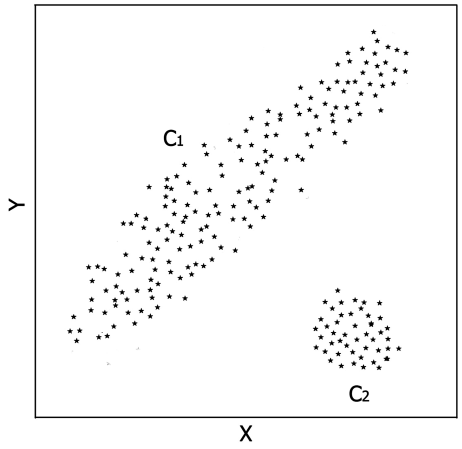

Figure 1. Conceptual illustration of the problem with most clustering methods applied to detrending. Using most algorithms, two clusters in the figure can be easily identified. Even though some of the elements in cluster C_1 are far from each other, they are identified as one cluster. The x- and y-axes indicate the distances between pairs of elements.

elements. In this concept, C_1 would be split into several smaller subgroups.

2.2.2 *Hierarchical tree clustering algorithm*

A hierarchical clustering algorithm is substantially different from density-based clustering or K-methods. It constructs a hierarchical tree by linking all elements together under the same root according to pre-defined distances (see equation 3). During the construction, it does not need to estimate initial parameters such as the minimum number of elements (as in density-based clustering algorithms) or the total number of clusters (as in K-methods). This is an advantage of the hierarchical algorithm.

The constructed hierarchical tree is traditionally represented by a dendrogram as shown in Fig. 2. We use the pre-defined distance matrix in order to link elements and generate the dendrogram. At each stage of linkage, the algorithm joins the two closest nodes into a new set. The 'node' can consist of either a single element or previously connected multiple elements. This process continues until all elements belong under the same root. During this linkage process, we need to define the distances between two nodes as well. There exist several methods to calculate the distance between nodes (Jain et al. 1999). Among these methods, we use the complete-linkage method

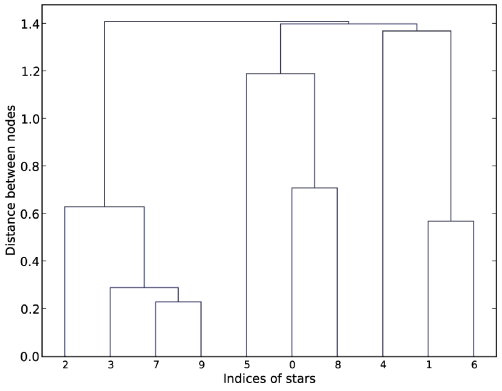

Figure 2. An example of a dendrogram. The x-axis is the index of the star, and the y-axis is the distance between nodes.

to construct the tree. In the complete-linkage method, the distance between two nodes is defined as the longest distance among the pairwise distances between the elements (as defined in equation 3) of the two nodes. Therefore, the distance between any two elements in two nodes is always smaller or equal to the distance between two nodes. The complete-linkage method was chosen because it produces more tightly bound clusters and hierarchies than other methods such as the single-linkage method (Jain et al. 1999).

Fig. 2 shows an example of a dendrogram of a hierarchical tree constructed by the complete-linkage method. We plot only 10 elements in Fig. 2 as an example. The x-axis is the index of each star and the y-axis is the distance between nodes. The height of the horizontal lines in the dendrogram represents the distance between two nodes linked together. We used the PyCluster library (de Hoon et al. 2004) to generate the hierarchical tree and the hcluster library to draw the dendrogram.

Traditionally, hierarchical algorithms do not produce clusters, unlike the other clustering algorithms that group elements into resulting clusters. This is a conventional feature of hierarchical algorithms (Daniels & Giraud-Carrier 2006) and it means that users must decide which elements in the tree should be grouped into resulting clusters. This is equivalent to defining the number of clusters in K-means clustering or defining the connectable distance in the density based clustering. To solve this problem, we propose an extension to the hierarchical algorithm, shown in the following section, that can extract the resulting clusters from the tree without the need of pre-defining such parameters.

2.2.3 *Agglomerative merging algorithm for selection of clusters*

With the constructed dendrogram in hand, we can link every star according to the distance matrix. We now need to extract subsets of stars that are highly correlated among themselves for a template set and to exclude outliers such as intrinsic variables that can be harmful for detrending. Furthermore, if there exist multiple and different trends in data, we should be able to separate them as well. The traditional method to achieve this is to set a certain distance value and extract subsets such that the farthest distance between elements in the subset is smaller than the set distance (e.g. subset [3, 7, 9] will be extracted given a set distance = 0.4 in Fig. 2). On the other hand, it is not easy to choose a set distance, especially for different data sets, for example with data observed under different weather conditions, different dates or with different telescopes. As we mentioned in previous section, this is a conventional feature of the hierarchical tree algorithm in extracting relevant and representative clusters.

To alleviate this problem, we developed an agglomerative merging algorithm (bottom-up merging algorithm) to identify the clusters in the constructed tree, based on the assumption of normal distribution (Kim & Shevlyakov 2008).

First, we note that distances between correlated light curves follow a skewed distribution in contrast to the distribution of distances of uncorrelated light curves that is known to follow a normal distribution.

Second, if one applies Fisher's transformation (Fisher 1915),

$$C'_{ij} = \frac{1}{2} \log \frac{1 + C_{ij}}{1 - C_{ij}} , \quad (4)$$

to the correlation values C_{ij}, the resulting transformed C's are approximately normally distributed (Anderson 1996).

Now we can claim that if a single cluster comprises correlated light curves and does not contain outliers, the transformed distances between the light curves in the cluster are normally distributed.

We then extract subsets by merging the two closest nodes that have the shortest distance in the tree (see details of the process below). We repeat the merging processes and test the normality at every merging step to decide whether to stop the merging processes. To test normality, we use the Anderson–Darling test (Anderson & Darling 1952; Stephens 1974) which tests the null hypothesis that a data set comes from the normal distribution. In other words, the test can statistically quantify how far the data set departs from the normal distribution. Based on the test, one can derive the p-value that indicates the level of significance of the departures from normal distribution (D'Agostino & Stephens 1986). If the subset fails the normality test, it is inferred that there exist outliers in the subset or the subset consists of two or more different trends. Therefore, we stop the merging process below the level where the normality test fails. If we repeat this process for extracting subsets in the hierarchical tree, we can finally obtain multiple clusters of trends without outliers.

Realistically, there is a mixture of various noise sources including Poisson noise and trends, and thus the distances between light curves in a cluster might not be perfectly normally distributed even after Fisher's transformation. Also, because the correlation coefficients in a given cluster are not totally independent (e.g. C_{12} and C_{13} are not totally independent of C_{23}), and because we repeat the p-value testing on the same subset multiple times, the p-value should be considered as a tuning parameter (threshold) instead of its strict statistical definition. Nevertheless, using the normality test, we can extract strongly correlated elements that are placed in the central part (peak) of the distribution. Note that only strongly correlated elements are important to determine trends.

We describe the details of the agglomerative merging algorithm here:

(i) Select initial cluster seeds to be all nodes which consist of only two elements in the constructed tree (e.g. [7, 9], [0, 8] and [1, 6] in Fig. 2).

(ii) Define $C_{\rm seed}$ to be the node that has the shortest distance between two elements among selected cluster seeds from step (i).

(iii) Merge $C_{\rm seed}$ with its next linked node in the tree and call it $C_{\rm merge}$. If the number of elements in $C_{\rm merge}$ is smaller than 5, keep merging with the next linked node. This is because if the number of elements in $C_{\rm merge}$ is too small, the normality test would not be reliable.

(iv) Apply the Anderson–Darling test to the distance list of $C_{\rm merge}$ and derive the p-value.[1] The distance list is the list of all distances between members of $C_{\rm merge}$. For instance, if the indices of members in $C_{\rm merge}$ are [1, 2, 3], the distance list is $[D_{12}, D_{13}, D_{23}]$ where D_{ij} is the distance matrix we defined in equation 3. We apply the Fisher's transformation before we apply the normality test as we mentioned above.

(v) If the calculated p-value is bigger than 0.1, set $C_{\rm merge}$ as new $C_{\rm seed}$ and go to step (iii). Otherwise, stop the merging process and go to step (vi).

(vi) Identify $C_{\rm seed}$ as cluster candidate. Go to step (ii) and choose the next closest pair. Keep these processes until there remain no initial cluster seeds.

(vii) Remove duplicated clusters from the candidates list derived at step (vi). The duplication can happen when there exist multiple seeds in one cluster that can yield identical clusters. Note that as long as the initial cluster seeds defined in step (i) are the same, the resulting clusters are the same no matter which cluster seed we start from.

(viii) Remove clusters whose number of elements are smaller than 10. We need a sufficient number of elements (light curves) to cancel out the uncorrelated noise in the light curves while determining master trends (see Section 2.3).

(ix) Define the list of clusters from step (viii) as C_k, where k is the index of each cluster.

The clusters identified by the algorithm above are used to determine trends which we explain in the following section.

While testing the merging algorithm, we observed that if we constrain the initial seeds at step (i), we can improve our algorithm by (a) decreasing CPU processing time and (b) removing relatively contaminated clusters by other noise sources such as Poisson noise. We explain the details below.

If we select the initial seeds to be just the pairs of elements whose distances are smaller than the average value of the distance matrix ($\overline{D}$) at step (i), we obtain a smaller number of seeds that are more highly correlated. $\overline{D}$ is given by

$$\overline{D} = \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D_{ij}\,, \tag{5}$$

where N is the total number of light curves. The benefit is that we speed up the algorithm by reducing the number of iterative processes that mainly consist of merging nodes and testing for normality. As we explained above, we repeat the merging and the normality test for every initial seed. Therefore, if there are fewer initial seeds, there are fewer iterative processes, thus reducing the CPU processing time. Moreover, we can remove pairs of faint stars in advance from initial seeds. Faint stars suffer from noise more than bright stars, therefore, the clusters derived with pairs of faint stars are less suited to determine trends than the clusters derived with pairs of bright stars. Note that if we use a looser constraint ($\gg \overline{D}$) and thus have too many seeds, the number of weakly correlated clusters and the computational cost will increase. On the contrary, if we use a tighter constraint ($\ll \overline{D}$) and thus fewer initial seeds, we may miss real clusters. We empirically found that any cutting values from $\overline{D}/10$ to $\overline{D}$ give reasonable results. Within this range, the overall characteristics of the determined trends using the resulting clusters were almost identical.

In addition, it is known that the square root of the variance of correlation coefficients are generally

$$\sigma = \frac{1 - C_{ij}^2}{\sqrt{n}}\,, \tag{6}$$

where C_{ij} is the correlation value between two variables and n is the total number of measurements (Bowley 1928; Hotelling 1953; Ghosh 1966). If the light curves consist of random fluctuations (e.g. pure Poisson noise), $C_{ij} \simeq 0$. Thus, equation 6 changes to

$$\sigma \simeq \frac{1}{\sqrt{n}}. \tag{7}$$

We remove all initial seeds from step (i) whose distances are larger than $1 - 3 * \sigma$ because resulting clusters using these initial seeds would contain light curves of mainly random fluctuation that are not correlated with other light curves. Note that this criterion is different from the one above. For example, this occurs when there is a set of light curves of random fluctuations. In that case, $\overline{D}$ is $\sim$1 and several initial seeds whose distances are smaller than 1 would pass the $\overline{D}$ criteria.

We also tested another threshold cut which constrains elements in each cluster to be highly correlated. If a distance between any

[1] We use R statistical packages and RPy library to calculate the p-value.

two elements in a given subset is bigger than $\overline{D}$, we stopped the merging process even if the subset was not rejected by the normality test. Nevertheless, we empirically found that resulting clusters and detrended light curves are not affected by this threshold.

2.3 Determination and removal of trends

With the extracted single or multiple clusters, we next determine the trends for each cluster (hereafter, master-trends), from the weighted sum of the cluster members as

$$T_k(t) = \frac{\sum_{i=1}^{N_k} w_i \, f_i(t)}{\sum_{i=1}^{N_k} w_i},$$
$$f_i(t) = \frac{L_i(t) - \overline{L}_i}{\overline{L}_i},$$
$$w_i = \frac{1}{\sigma_{f_i}^2}, \quad (8)$$

where σ_{f_i} is the standard deviation of f_i, N_k is the total number of template stars in the cluster C_k, t is the time index with the total n measurements, $L_i(t)$ is the light curve of ith template star and $\overline{L}_i$ is the mean value of $L_i(t)$.

This master-trend set, $T_k(t)$, is used to detrend the individual light curves. Each master-trend well represents the characteristic of each cluster because all the light curves in each cluster are selected to be strongly correlated. Note that we determine just one master light curve per cluster.

After we determine the master-trends, we remove the trends from each individual light curve. First we normalize each light curve $L_i(t)$ as

$$\hat{L}_i(t) = \frac{L_i(t) - \overline{L}_i}{\overline{L}_i} \, . \quad (9)$$

We then assume that each light curve, $\hat{L}_i(t)$, is a linear combination of the determined master-trends, $T_k(t)$, and noise, $\epsilon_i(t)$,

$$\hat{L}_i(t) = \sum_{k=1}^{m} \beta_{ik} \, T_k(t) + \epsilon_i(t), \quad (10)$$

where is are the indices of individual light curves to be detrended, ks are the indices of master-trends, m is the total number of master-trends and β_{ik} are free parameters to be determined by means of minimization of $\sum_t \epsilon_i(t)^2$ (equivalent to minimizing r_i^2 in equation 2).

During the minimization of $\sum_t \epsilon_i(t)^2$, there is one more complication we have to consider. Let us assume that there exists a single trend where flux increases monotonically and an intrinsic variable star where flux decreases monotonically. Even though the direction of the trend is different from that of the variable star, the minimization method will eventually reduce the intrinsic signal because the free parameters can take negative values and thus minimize $\sum_t \epsilon_i(t)^2$. To eliminate this undesirable effect, we constraint the free parameters β_{ik}, to be always bigger than or equal to zero using quadratic programming (Goldfarb & Idnani 1983).[2]

[2] Quadratic programming is a mathematical optimization method which minimizes (or maximizes) a quadratic function of several unknown parameters which is subordinate to linear constraints on the parameters. We use R statistical packages to implement quadratic programming.

3 TEST WITH SYNTHETIC LIGHT CURVES

We present here the results from several simulations we performed. First, we describe the method by which we parametrized trends and how we built the simulation (see Section 3.1). Next, we present detrended results of artificially inserted transits and eclipsing binaries using PDT (Section 3.2) and comparison results with TFA (Section 3.3). Finally, we show simulations and results from other unique configurations (Section 3.4).

3.1 Data description

We generated ∼500 artificial light curves, each having different flux and 360 one-minute-exposures. During this simulation, we set x-coordinates of stars as altitude and y-coordinates as azimuth. CCD size was set to 2048 × 2048. The magnitudes of the stars were chosen from the USNO B1.0 catalogue (Monet et al. 2003) within a particular patch of the sky (3 deg^2) [$4^h48^m00^s, 20°46'20''$] and ranged from ∼6 to ∼13 mag. Poisson noise was added to light curves with standard deviation values (σ) set to vary from 0.001 to 0.02 mag. Although there exist other possible sources contributing to the noise budget such as CCD overscan, bias, dark, flat-field, scintillation, incorrect sky subtraction, etc. (Gilliland & Brown 1988), we did not include those noise sources because trends are predominantly due to weather. Bias and other such error sources are usually stable during a night observation and not a major source of trends.

We added three transit signals (Mandel & Agol 2002) into three different light curves with $\sigma = 0.01$ mag. Transit depths were 0.015, 0.020 and 0.025 mag with 60 min duration (one-sixth of total observation duration). We placed the transit signals at the central part of the light curve. We also added two eclipsing binaries into two light curves with $\sigma = 0.01$ mag. The remaining stars were set to have no intrinsic variability.

To add trends, we artificially generated four types of trends:

(i) *First order atmospheric extinction.* This is the typical extinction that linearly depends on the airmass: $a \, M_i(t)$, where a is the extinction coefficient, which is ∼0.16 for the V and ∼0.1 for the R band (Stalin et al. 2008), $M_i(t)$ is the airmass of i th star and is given by

$$M_i(t) = \sec[z_i(t)],$$
$$z_i(t) = 90° - [(c + d\,t) + e\,\hat{y}_i], \quad (11)$$

where c is the starting altitude of the field ($c = 45°$), d is the change of altitude per minute ($d = 0.25°$ min^{-1}), e is the field of view ($e = 3$ deg^2), $\hat{y}_i = y_i / D_y$ is the y-position of i th star normalized by y-size D_y of CCD plane and t is the observational time in minutes. To change airmass with time, we changed the altitude of the field from −45°, passing through 90° to 45°.

(ii) *Position-dependent and time-dependent trend.* We model this type of extinction to imitate stationary 'clouds' and thus to depend on the azimuthal position of the star and observation time as: $b\,\hat{t}\,\hat{x}_i$, where b is the maximum depth (for this simulation we use $b = 0.01$ mag), $\hat{t} = t/t_{\rm TOTAL}$ is the normalized time over the total observation duration ($t_{\rm TOTAL}$), $\hat{x}_i = x_i / D_x$ is the x-position of ith star normalized by D_x, the x-size of CCD plane. Such position-dependent and time-dependent trends can be caused by thin cloud, moonlight or occasionally by CCD noise.

(iii) *Localized trend.* This is an artificially CCD-localized trend that has a simple linear time dependence, $\zeta_i(t)$, given by

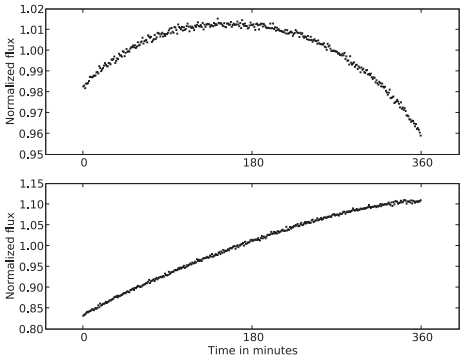

Figure 3. Two sample synthetic light curves of two bright stars which contain trends. The light curve at the top panel consists of the first order atmospheric extinction, position-dependent and time-dependent trend, and Poisson noise. The light curve at the bottom panel contains an additional trend that is artificially CCD-localized.

$$\zeta_i(t) = f\hat{t}\,,$$
$$f \quad = \begin{cases} 0.25 & x_i > 1500 \ \text{ and } \ y_i < 500 \\ 0 & \text{otherwise,} \end{cases} \qquad (12)$$

where x_i is the x-position of ith star, y_i is the y-position of ith star and $\hat{t}$ is the normalized time as explained above. Such localized trends can be caused by non-uniform clouds or the non-uniform illumination structure of CCD images.

(iv) *The second order atmospheric extinction.* This is the other atmospheric extinction related to the star colour, $wrCM_i(t)$, where w is proportional to the square of the optical bandwidth, C is a colour index and r is the difference between the extinction coefficients in the corresponding bands (Young et al. 1991). The coefficients w and r are constant and same for all stars in the field. Even if the airmass changes for two stars are same during observation (e.g. two stars at same altitude), trends could be different due to the differences in colours (typically a few milli-magnitude differences in light curves, Young et al. 1991). We will ignore this term until Section 3.4.1.

Fig. 3 shows two distinctive trends build on two bright stars. The top panel is a light curve of a bright star which consists of the first trend, the second trend [(i), (ii)] and Poisson noise. The bottom panel is a light curve of another bright star which consists of the first trend, the second trend, the third trend [(i), (ii), (iii)] and Poisson noise.

3.2 Identification of clusters of template stars

We applied PDT to test its ability to properly identify the inserted artificial trends. Using PDT, we identified four different clusters in the data set as shown in Fig. 4. The x- and y-axis of Fig. 4 are the x- and y-coordinates of the template stars on the CCD plane. Different symbols indicate different clusters. Each cluster is well separated along the y-axis due to the artificially inserted airmass [(i), $a\,M_i(t)$]. Also, the clusters show a slope along the field due to the second trend [(ii), $b\ \hat{t}\ \hat{x}_i$]. Finally, PDT exactly identified a cluster of localized trend [(iii), $\zeta_i(t)$, marked as circles in Fig. 4]. As the results clearly indicate, PDT can identify and group light curves according to their similarity, even though multiple trends are mixed together and the trends are contaminated by other noise sources such as Poisson noise.

The identified clusters do not contain any stars which are intrinsically variable (three transits and two eclipsing binaries). This shows that our clustering algorithm is also effectively excluding such unwanted outliers.

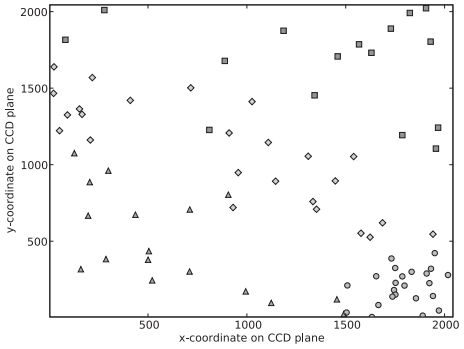

Figure 4. Positions of the identified four clusters in the artificially generated data set. $x(y)$-axis is the $x(y)$-coordinates of stars on the field. Four different shapes mean four different clusters.

3.3 Detrending results and comparison with TFA

Here, we compare our results to TFA. TFA is one of the particularly successful detrending methods (Kovács, Bakos & Noyes 2005; Tamuz, Mazeh & Zucker 2005) and it is used by exo-planet searches such as HATNet (Bakos et al. 2004). It is, therefore, a good comparison algorithm for our detrending algorithm. TFA uses a large number of bright stars as a template set while excluding the light curve being detrended. TFA does not eliminate potentially dangerous stars, such as the stars which have intrinsic variability, from the template set, and it assigns one free parameter per template light curve. In contrast, our algorithm can automatically exclude such intrinsically variable stars and assign one free parameter per cluster of template light curves.

First, we present the detrended results of three transit signals. The top panel of Fig. 5 shows the raw light curves before any detrending treatments. Each column shows three different transits with different depth (0.015, 0.020 and 0.025 mag from left to right). TFA results are shown in the middle panel of Fig. 5, while PDT results are shown in the bottom panel. We used 60 bright stars as a template set for TFA. We excluded the three transit light curves from the template set because in realistic scenarios, it is uncommon for there to be three transit events occurring in the same field and same epoch. However, we did not exclude the two eclipsing binaries from the template set for TFA because variable stars such as eclipsing binaries are common in the field. As the middle panel shows, TFA suppressed each transit signal more than PDT. The suppression was mainly caused by the presence of the eclipsing binaries in the template set. Because TFA tries to minimize the residual between the target light curve to be detrended and a linear combination of light curves from the template set that might contain intrinsic variables such as the two eclipsing binaries in this simulation, it occasionally suppresses the intrinsic signals of the target light curve by removing any similar signals between the target light curve and the template set. In contrast, results from PDT, which can select template sets that do not contain the three transits or the two eclipsing binaries, show less significant signal depression and clearer transit signals than TFA results (see bottom panel of Fig. 5).

Note that one of the eclipsing binaries was phased to the transits to show signal depression effect of TFA. Such coincidences are not common, but we cannot ignore the probability, especially in the case of wide field surveys which simultaneously monitor more than

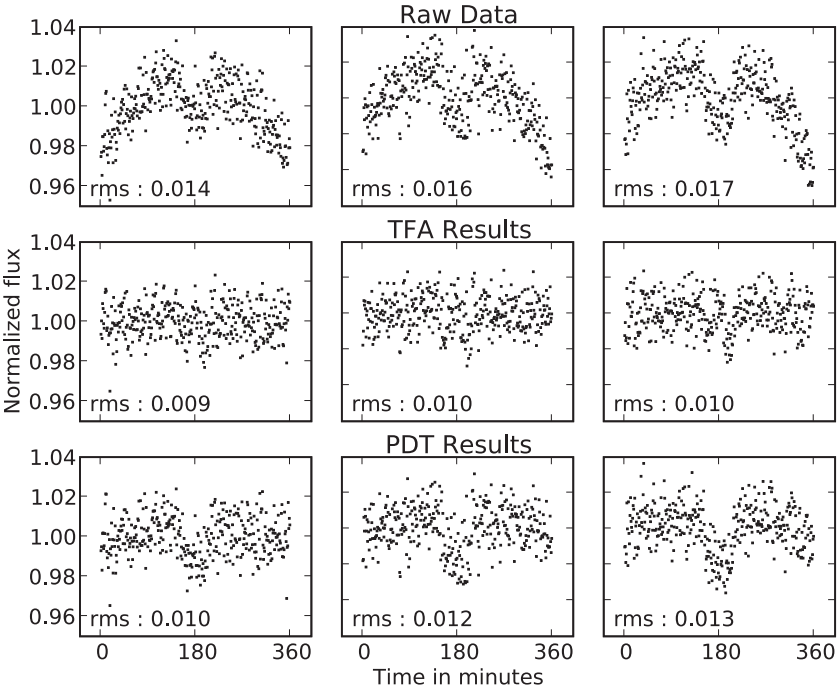

Figure 5. Detrended results of the simulated three transit events. Each column represents each different transit depth of 0.15, 0.20 and 0.25 mag from left to right. The top panel is raw light curves. The middle panel is TFA results, and the bottom panel is PDT results. We indicate the rms of each light curve as well.

Table 1. χ^2 values of each transit and χ^2 ratio of TFA to PDT.

Transit depth	TFA	PDT	χ^2 ratio
0.015	1.08	1.03	1.05
0.020	1.49	1.39	1.07
0.025	1.69	1.50	1.13

several hundred stars. If we exclude the eclipsing binary from the template set, the detrended results using TFA are almost identical to the results using PDT.

We performed χ^2 tests comparing detrended results and original transit signals for all three transits to check how successfully both detrending algorithms regenerated the intrinsic signals. Table 1 shows individual χ^2 values of each transit and χ^2 ratios of TFA to PDT. The χ^2 ratio is defined as $\chi^2_{\rm TFA}/\chi^2_{\rm PDT}$. Therefore, if the χ^2 ratio is bigger than one, it means that PDT results are more similar to the original transit signals than TFA results. As Table 1 shows, all three χ^2 ratios are slightly bigger than one.

If the rms contribution from intrinsic signal is significant, such as the two eclipsing binaries in this simulation, any method which minimizes the rms to detrend light curves will dilute the intrinsic signals. This is the critical problem of the rms minimization algorithm and cannot be perfectly overcome as long as we use the minimization approach. One solution to reduce this side effect is to decrease the number of free parameters (see Section 2) and constraint the free parameters to be bigger than or equal to zero (see Section 2.3). PDT, by construction, has fewer parameters than TFA because we determine one master-trend per cluster. Also, PDT can constraint the free parameters using quadratic programming.

Fig. 6 shows the detrended results of the two eclipsing binaries by both TFA and PDT. The top left panel is the raw light curve of one eclipsing binary affected by all three trends including localized trend [trend (iii)] and thus the average flux of the light curve is increasing along time. The top right panel is the raw light curve of another eclipsing binary affected by only two trends [trend (i) and trend (ii)] and thus it does not show increase of flux because the intrinsic signal is relatively bigger than the two trends. The middle panel is the TFA results and the bottom panel is the PDT results. In both cases, TFA not only removed the trends but also diluted the intrinsic signals. In contrast, PDT removed only the trends and successfully regenerated the intrinsic signals of two binaries.

In addition, we indicate the rms of the detrended light curves in each panel of Figs 5 and 6. The rms values are always smaller in TFA results than in PDT results because TFA has more adjustable free parameters than PDT has and TFA can set free parameters to be any values including negative values. However, as the detrended results show, smaller rms do not always mean better detrended results, especially when intrinsic signals contribute mainly to rms of light curves such as the two eclipsing binaries.

Note the TFA has a *reconstruction* phase which can greatly improve S/N of periodic signals with the initial guess of the signal models (Kovács et al. 2005; Kovacs & Bakos 2008). Nevertheless, PDT is designed to regenerate any types of intrinsic signals whether they are periodic or not.

3.4 Second order extinction and other considerations

3.4.1 Second order extinction test

We now turn our attention to the second order atmospheric extinction related to colours of stars [(iv) at Section 3.1]. After performing several simulations with realistic parameters (e.g. different field of view from 0.1 to 5 deg^2, different bands such as B and V, different observation durations from 1–6 h, etc.) that contain both first and second order extinction, we found that PDT cannot separate clusters according to star colour. The reason is that both extinctions depend

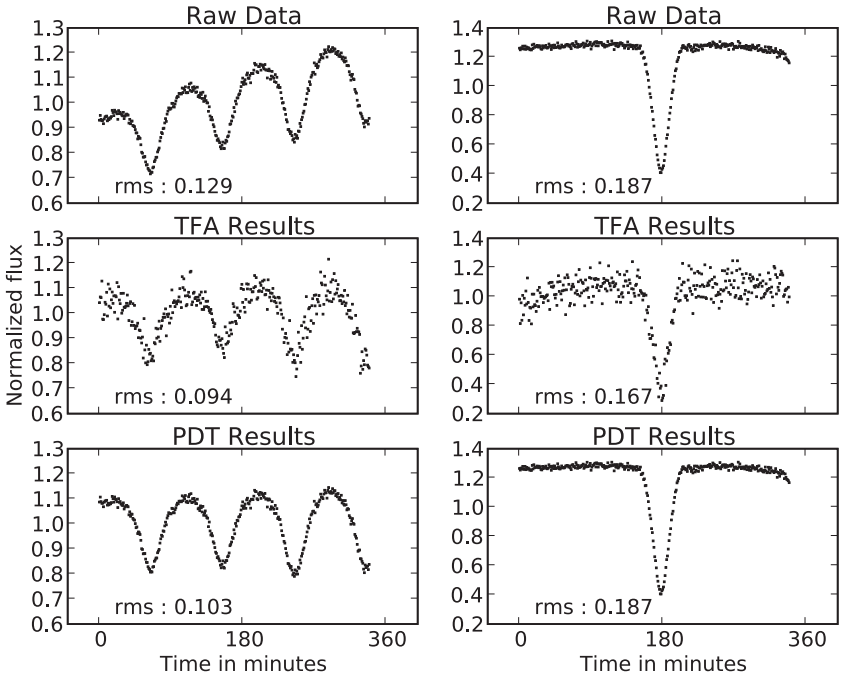

Figure 6. Detrended results of the simulated two eclipsing binaries. The top panel is raw light curves. The middle panel is TFA results, and the bottom panel is PDT results. We indicate the rms of each light curve as well.

Table 2. Mean and standard deviation values (σ) of colours of stars in resulting six clusters.

Cluster	Mean	σ
C_1	1.40	0.11
C_2	1.39	0.17
C_3	1.18	0.27
C_4	0.87	0.15
C_5	0.79	0.09
C_6	−0.25	0.05

linearly on airmass, and the first order extinction is much larger than the second order extinction when using realistic values for the coefficients. Therefore, PDT identifies clusters that mainly depend on the first order extinction.

It is worth mentioning that PDT can identify clusters based on colours if we isolate only the second order extinction. We performed another simulation to test this:

(i) Generate ∼500 light curves that contain only the second order extinction and Poisson noise. We extracted $B - R$ colours of stars from USNO B1.0 catalogue within a particular patch of the sky (3 deg^2) [$4^h48^m00^s, 20^\circ46'20''$], which is the same field of view as in the previous simulation shown in Section 3.1.

(ii) Apply PDT to the light curves and identify clusters.

Table 2 shows the mean and standard deviation values (σ) of the colours of stars in clusters identified by PDT. Although some of the clusters (e.g. C_1 and C_2) could be regarded as clusters of the same trend because they have similar mean colour values, PDT did a good job of separating bluish cluster (C_6) from reddish (C_1 to C_5) clusters.

3.4.2 Pure Poisson noise case

We tested both PDT and TFA with ∼500 synthetic light curves with pure Poisson noise but no trends. We also added three artificial transit events into three individual light curves. We used 60 bright stars as a template set for TFA. Even though there were no trends, TFA still detrended the light curves by using the template stars, and it eventually suppressed the intrinsic signals of the transits. By contrast, PDT did not identify any clusters because we exclude clusters that consist of only Poisson noise (see Section 2.2.3). Consequently, it did not detrend light curves and thus did not suppress any intrinsic signals.

Note that Poisson noise is not always the dominant noise source in light curves. The example referred here shows that if there are none-strongly correlated elements (trends) in data set, then PDT will not detrend the data set.

4 TEST WITH ASTRONOMICAL DATA SETS

We present now two examples of astronomical data sets. One is from The Taiwan-American Occultation Survey (TAOS) (Lehner et al. 2009), and the other is from an occultation survey using Megacam on the 6.5-m Multi-Mirror Telescope (MMT) (Bianco et al. 2009). Both examples show multiple trends that are well localized on the CCD plane. Such localization of trends could be caused by various noise sources such as airmass, cloud passages, noise of CCD images, telescope vibration, defects of photometry and so on. These localizations often happen with wide field observations.

4.1 An example of TAOS data set

The scientific goal of TAOS (Zhang et al. 2008) is to detect km-sized Kuiper Belt Objects (Luu & Jewitt 2002) at a distance of Neptune or beyond. TAOS data usually suffer from low S/N and systematic trends due to the small telescope size (four 50-cm telescopes), noise

of CCD images, defects of photometry and unstable local weather (e.g. cloud passages). The field of view of the TAOS telescopes is 3 deg^2 and the sampling rate is 5 Hz. We chose one sample set of light curves generated by the TAOS photometry pipeline (Zhang et al., in preparation) and detrended the light curves using PDT. The total observation time of the light curves was 1.5 h.

Fig. 7 shows the determined master-trends and examples of detrended light curves. The top left panel shows the position of stars in identified clusters on the CCD plane. Different shapes indicate different clusters. The clusters are localized on the CCD plane due to unstable local weather, noise of CCD images and defects of photometry. The bottom two panels show example light curves of two non-variable stars. The upper light curves of the two bottom panels are before detrending and the lower light curves are after detrending. As the results show, PDT removed trends from both light curves.

The top left panel of Fig. 8 shows the position of stars in identified clusters. Different shapes indicate different clusters. The top right panel shows the determined master-trends. We magnified a part of the light curves (∼5 s) to clearly show the trends. The bottom two panels show two example light curves of non-variable stars before and after detrending.

As the figure shows, two clusters marked as circles and triangles are localized on the CCD plane. In our analysis, we found that often the clusters were divided along the horizontal half divide (e.g. clusters marked as circles and triangles in Fig. 8), and that can be attributed to details of the readout mode, but we also found cases where the clustering that crossed over the horizontal divide (e.g. a cluster marked as squares in 8). The trends are likely due to a combination of weather patterns, photometry and the way the CCD was read out (Bianco et al. 2009).

4.2 An example of Megacam data set

We also applied PDT to a data set obtained using Megacam (McLeod et al. 1998) at the MMT at Mount Hopkins, Arizona. Megacam is a mosaic CCD which consists of 36 chips. The size of each CCD is 2 K by 4 K and the field of view is 24 × 24 $arcmin^2$. Megacam was used in *continuous-readout* mode achieving 200 Hz sampling rate in order to detect stellar occultations caused by Kuiper Belt Objects (Bianco et al. 2009). Due to the high sampling rate, telescope vibrations, defects of photometry and the readout technique, these Megacam data show strong trends. The total observation time of the selected data set was 15 min.

5 NOTES AND FUTURE WORK

A weakness of PDT is that it cannot remove trends that are manifested in just a few light curves and are not highly correlated. For example, moving asteroids or satellites could result in an increase and decrease of the estimated flux of a few background stars in the neighbourhood of the track. These trends are out of phase throughout the light curves because the asteroids or satellites are moving across the field. For these reasons, strongly trended light curves are not highly correlated and thus PDT cannot group them into clusters. We are planning to handle this phase-shift of trends in a future version of PDT.

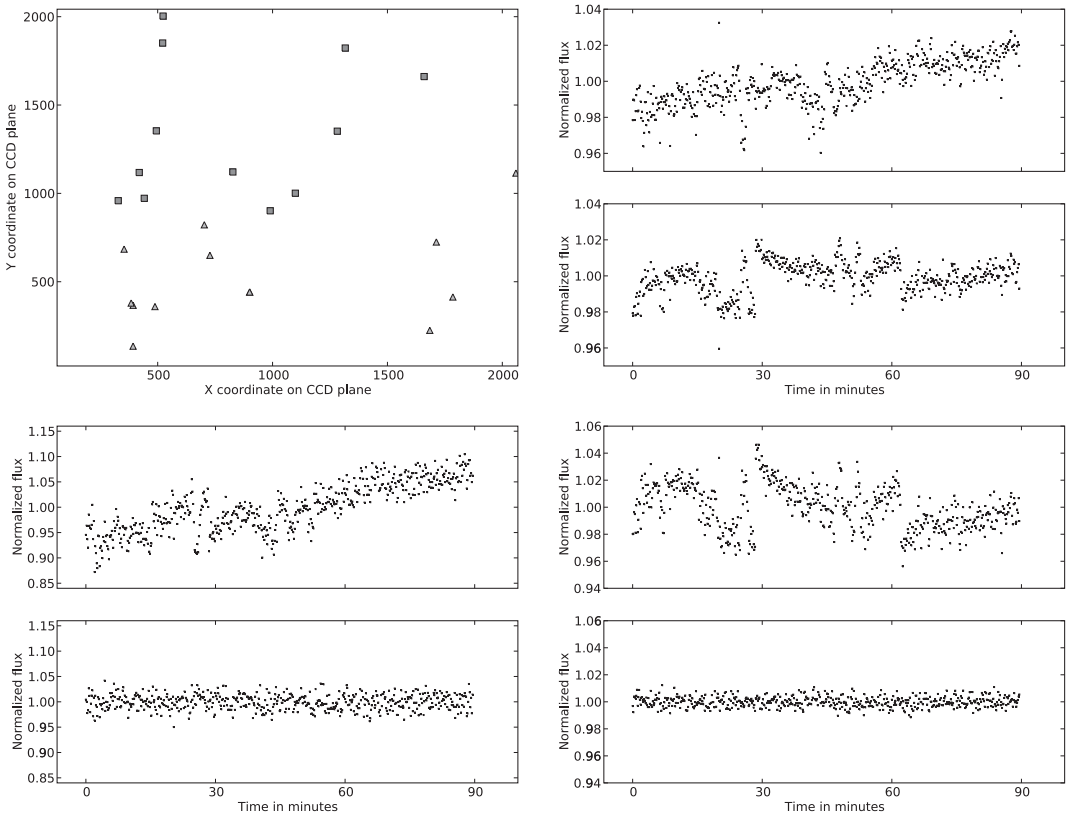

Figure 7. An example of TAOS data set. Top left : position of stars in identified two clusters. $x(y)$-axis is the $x(y)$-coordinate of stars on the CCD plane. Different shapes indicate different clusters. Top right : determined two master-trends. Bottom left and bottom right : two example light curves before and after detrending. Upper panels are before detrending and lower panels are after detrending.

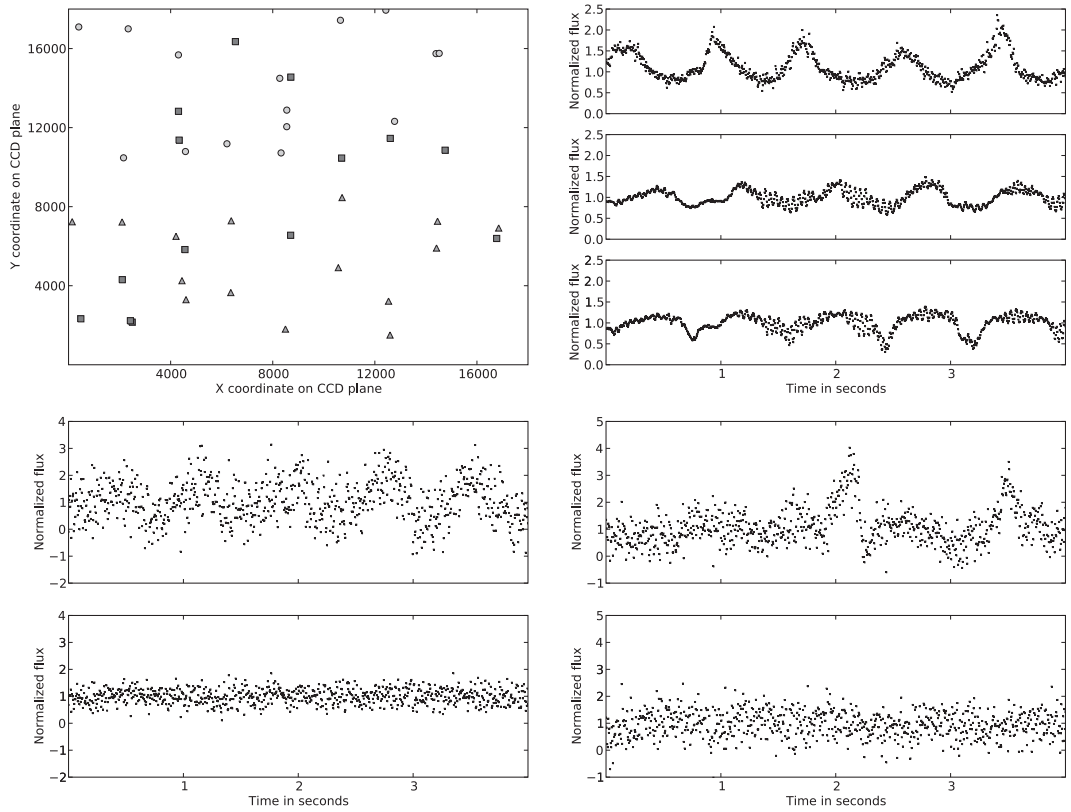

Figure 8. An example of Megacam data set. Top left: position of stars in identified three clusters. $x(y)$-axis is the $x(y)$-coordinate of stars on the CCD plane. Different shapes indicate different clusters. Top right: determined three master-trends. Bottom left and bottom right: two example light curves before and after detrending. Upper panels are before detrending and lower panels are after detrending.

We are also applying PDT to astronomical data sets, e.g. TAOS and MMT, in order to detect various transient events such as Kuiper belt object occultations, flare stars, micro-lensing events and exo-planet transits.

6 CONCLUSION

In this paper, we presented the PDT, a new detrending algorithm. We first determined the trends by constructing a hierarchical tree based on the similarity matrix. Elements of the similarity matrix are the Pearson correlation values of all pairs of light curves. After that, a bottom-up merging algorithm was applied to the constructed tree in order to identify subsets of light curves that we call clusters. At each step of the merging process, we tested the normality of the subsets and determined where to stop. By means of the normality test, we could select reliable clusters of trends. For each cluster, we determined one representative master-trend by weighted sum of the normalized light curves. This procedure greatly constrained the number of free parameters to be calculated, and thus showed less significant signal depression than other detrending algorithms such as TFA. Finally, in order to remove the trends from individual light curves, we used quadratic programming to minimize the residual between each target light curve and the determined master-trends. Note that PDT is designed to remove only the fluctuations that are common among stars. If the fluctuations are unique to an individual star, the fluctuations will be preserved.

We performed several simulations of synthetic light curves with different initial parameters such as total duration of observation, transit duration, field of view, exposure time, etc., to test PDT and showed some of the simulation results in this paper.

First, we tested PDT with ∼500 synthetic light curves that contain the first order atmospheric extinction (airmass), artificial trends, Poisson noise and events (three transits and two eclipsing binaries). We applied PDT to these synthetic light curves in order to determine trends and to regenerate the inserted events. PDT successfully identified multiple clusters of different trends which were the mixture of different trends and noise. These identified clusters well represented the overall characteristic of the trends through the field. We compared detrended results of PDT with one another detrending algorithm (TFA). PDT results were an improvement over TFA results, especially when (a) the data set contains intrinsic variables that would be included in template set of TFA or (b) the rms contribution from the intrinsic signals is significant.

We also tested PDT with ∼500 synthetic light curves that contain colour dependent second order extinction and Poisson noise. Trends appearing in light curves can be slightly different due to differences in colour. We found that PDT can identify clusters according to colour. However, in realistic scenarios, it is not easy to isolate only the second order extinction because, even if one correctly removes the first order extinction for all stars, there exist other various noise sources which dilute trends caused by the second order extinction.

In the case of data set of random fluctuations (e.g. pure Poisson noise), which does not have any trends, we do not need to detrend the data set. PDT can distinguish the light curves of random fluctuations using the characteristics of the distribution of correlation coefficients. Therefore, PDT does not detrend these light curves and thus preserves any intrinsic signals.

Examples of two astronomical data sets are also presented. They show multiple trends in the field caused by various noise sources such as airmass, cloud passages, telescope vibration, defects of photometry and so on. PDT performed well and removed trends that appeared in the data sets.

In this paper, we show the simulation results of wide field data only. However, PDT can be applied to narrow field data as well if there are enough stars in the field ($\sim$ a few hundreds). In addition, PDT is useful to extract global trends that can represent the overall characteristics of a data set. The extracted trends can give a general idea of how much the data are contaminated by the trends.

The software package of PDT is provided at http://timemachine.iic.harvard.edu.

ACKNOWLEDGMENTS

This work is supported by the Korea Research Foundation. Y.-I. Byun also acknowledges the grant of KRF-2007-C00020. We thank R. Reid, R. Dave, G. Wachman and D. Preston at the Harvard, Initiative in Innovative Computing (IIC), and A. W. Blocker at the Harvard University Department of Statistics for comments and suggestions on this paper. We also thank the Harvard-Smithsonian Center for Astrophysics and IIC for providing computing facilities and research space. The simulations and the detrending of data sets in this paper were run on the Odyssey cluster supported by the FAS Research Computing Group at the Harvard.

REFERENCES

Akerlof C. et al., 2000, AJ, 119, 1901
Alonso R. et al., 2004, ApJ, 613, L153
Anderson T. W., 1996, Stat. Sci., 11, 20
Anderson T. W., Darling D. A., 1952, Ann. Math. Stat., 23, 193
Bakos G., Noyes R. W., Kovács G., Stanek K. Z., Sasselov D. D., Domsa I., 2004, PASP, 116, 266
Bakos G. Á. et al., 2007, ApJ, 656, 552
Bianco F. B., Protopapas P., McLeod B. A., Alcock C. R., Holman M. J., Lehner M. J., 2009, preprint (astro-ph/0903.3036)
Bowley A. L., 1928, J. Am. Stat. Assoc., 23, 31
Burke C. J. et al., 2008, ApJ, 686, 1331
D'Agostino R. B., Stephens M. A., 1986, Goodness-of-Fit Techniques. Marcel Dekker, Inc., New York, NY, USA, p. 107
Daniels K., Giraud-Carrier C., 2006, in ICMLA '06: Proc. of the 5th Int. Conf. on Machine Learning and Applications. IEEE Computer Society, Washington, DC, p. 270
de Hoon M. J. L., Imoto S., Nolan J., Miyano S., 2004, Bioinformatics, 20, 1453
Ester M., Kriegel H. P., Sander J., Xu X., 1996, in Simoudis E., Han J., Fayyad U., eds, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, Menlo Park, p. 226
Everett M. E., Howell S. B., 2001, PASP, 113, 1428
Everett M. E., Howell S. B., van Belle G. T., Ciardi D. R., 2002, PASP, 114, 656
Fisher R. A., 1915, Biometrika, 10, 507
Ghosh B. K., 1966, Biometrika, 53, 258
Gilliland R. L., Brown T. M., 1988, PASP, 100, 754
Goldfarb D., Idnani A., 1983, Math. Program., 27, 1
Hartigan J. A., Wong M. A., 1979, Appl. Stat., 28, 100
Hotelling H., 1953, J. R Statist. Soc. B, 15, 193
Howell S. B., Jacoby G. H., 1986, PASP, 98, 802
Jain A. K., Murty M. N., Flynn P. J., 1999, ACM Comput. Surv., 31, 264
Kim K., Shevlyakov G., 2008, IEEE Signal Process. Mag., 25, 2, 102
Kjeldsen H., Frandsen S., 1992, PASP, 104, 413
Kovács G., Bakos G. A., 2008, Commun. Asteroseismol., 157, 82
Kovács G., Bakos G., Noyes R. W., 2005, MNRAS, 356, 557
Landolt A. U., 1992, AJ, 104, 340
Lehner M. J. et al., 2009, PASP, 121, 138
Luu J. X., Jewitt D. C., 2002, ARA&A, 40, 63
Mandel K., Agol E., 2002, ApJ, 580, L171
McCullough P. R., Stys J. E., Valenti J. A., Fleming S. W., Janes K. A., Heasley J. N., 2005, PASP, 117, 783
McLeod B. A., Gauron T. M., Geary J. C., Ordway M. P., Roll J. B., 1998, in D'Odorico S., ed., Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 3355, Megacam: Paving the Focal Plane of the MMT with Silicon. p. 477
Monet D. G. et al., 2003, AJ, 125, 984
Ng R. T., Han J., 1994, in Bocca J., Jarke M., Zaniolo C., eds, 20th International Conference on Very Large Data Bases, 1994 September 12–15, Santiago, Chile Proceedings, Efficient and Effective Clustering Methods for Spatial Data Mining. Morgan Kaufmann Publishers, Los Altos, CA 94022, USA, p. 144
Paczynski B., Pojmanski G., 2000, BAAS, 32, 687
Pál A. et al., 2008, ApJ, 680, 1450
Pigulski A., Pojmański G., 2008, A&A, 477, 907
Pojmanski G., 2005, Acta Astron., 52, 347
Pollacco D. et al., 2008, MNRAS, 385, 1576
Schmidt E. G., 1991, AJ, 102, 1766
Schmidt E. G., Langan S., Rogalla D., Thacker-Lynn L., 2007, AJ, 133, 665
Stalin C. S., Hegde M., Sahu D. K., Parihar P. S., Anupama G. C., Bhatt B. C., Prabhu T. P., 2008, Bull. Astron. Soc. India, 36, 111
Stephens M. A., 1974, J. Am. Stat. Assoc., 69, 730
Szczygiel D. M., Fabrycky D. C., 2007, MNRAS, 377, 1263
Tamuz O., Mazeh T., Zucker S., 2005, MNRAS, 356, 1466
Young A. T. et al., 1991, PASP, 103, 221
Zhang Z.-W. et al., 2008, ApJ, 685, L157

This paper has been typeset from a TEX/LATEX file prepared by the author.