Removing systematic noise from lightcurves

A discussion of past approaches and potential future directions

Giri Gopalan[1] Dr. Peter Plavchan[2]

[1]Department of Statistics
Harvard University

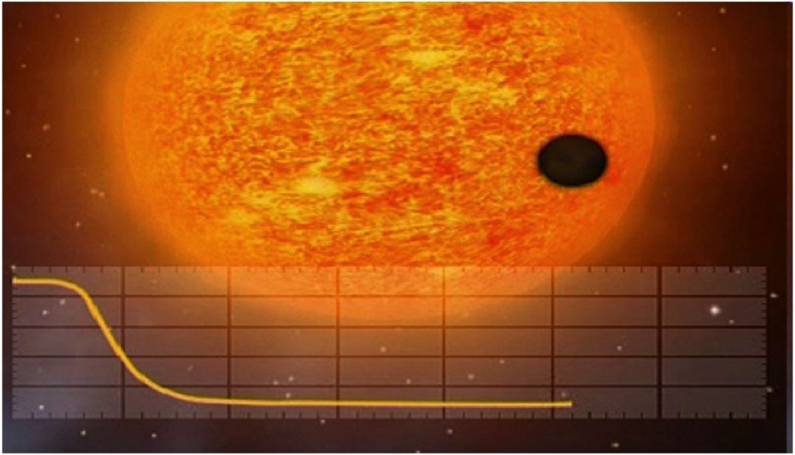
[2]Caltech
NASA IPAC

Stat 310, Feb 11th, 2014

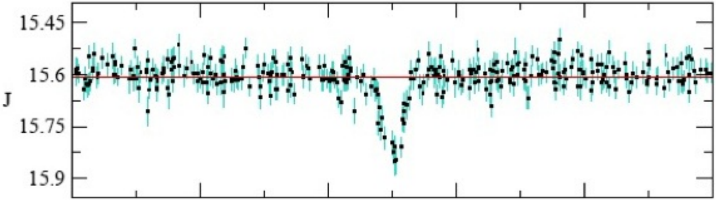
Outline

- **Motivation**: Lightcurves, exoplanets, systematic noise
- **"Trend Filtering Algorithm"**: Concise mathematical and computational foundation and extensions
- **Applying TFA and the Perils of Overfitting: 2MASS and PTF Data**
- **Future Directions**: A more principled Bayesian approach, modeling the frequency domain with wavelets.

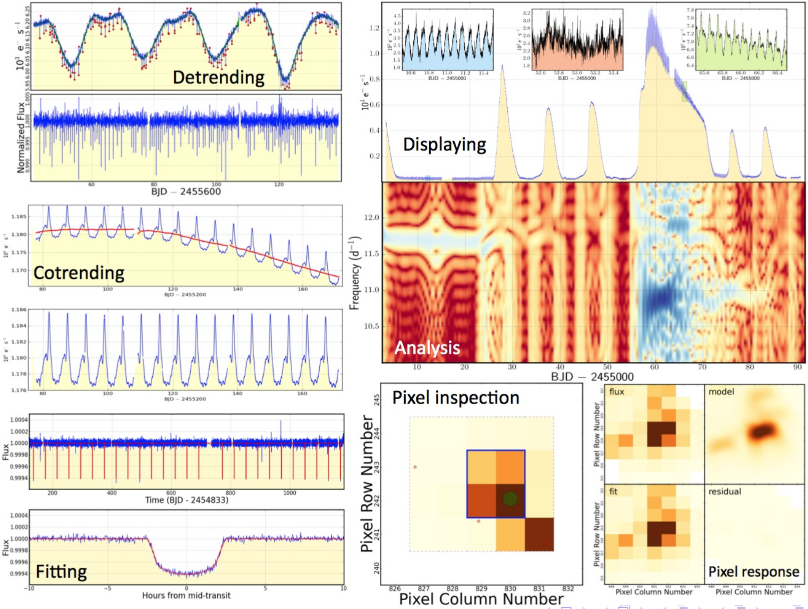
Motivation: transiting exoplanets



Motivation: transiting exoplanets



Motivation: systematic trends



TFA (Kovacs and Bakos): formulation with linear algebra

- Let $Y \in \mathbb{R}^{nx1}$ be an unfiltered lightcurve.
- Let $T \in \mathbb{R}^{kxn}$ be a "template" set where each row represents a "systematic" trend. Presumably k is much smaller than the number of total lightcurves. It is also reasonable to assume $k \leq n$.
- By assumption, the total systematic noise affecting lightcurve Y is a **linear** combination of the rows of T, namely $F = T^t c$ for some $c \in \mathbb{R}^k$.
- The filtered lightcurve is then $Y - T^t c \in \mathbb{R}^n$.
- How do we find c?

TFA (Kovacs and Bakos): formulation with linear algebra

- The original literature algorithm simply uses a least squares estimate: $\underset{c}{\mathrm{argmin}}||Y - T^t c||_2$.
- This has a well known result: $c = (TT^t)^{-1}TY$. (If $rank(TT^t) < k$, we can use the pseudoinverse.)
- So we are simply (orthogonally) projecting the lightcurve onto the vector space spanned by the template trends to determine the noise.

TFA: incorporating measurement uncertainties and ancillary information

- Say we have additional information thought to correlate with noise e.g seeing conditions.
- TFA simply corrects for this by including extra rows to T.
- Say we have a variance estimate for the brightness measurement at a particular time point.
- TFA simply corrects for this by weighting by the inverse of this measurement.
- i.e We now solve the minimization problem: $\underset{c}{\operatorname{argmin}}||(Y - T^t c)S^{-1}||_2$ where T has additional rows for the ancillary information and $S \in \mathbb{R}^n$ is a diagonal matrix with $S_{ii} = \hat{\sigma}_i$.

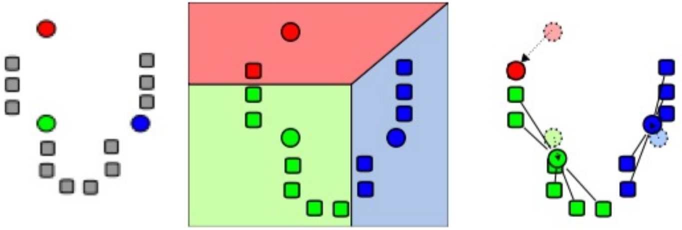
TFA: choosing a template set

- Crux of the approach: determining T, the template set of systematic trends.
- Bakos and Kovacs suggested simply using a cutoff of the standard deviation of a lightcurve as a criterion for including it in the template set which resulted in approximately 50 template lightcurves. In addition to being somewhat ad-hoc, using this approach leads to overfitting, as will be illustrated shortly.
- The approach suggested in Kim et al. was to use unsupervised learning to extract the systematic trends from the data set. We tried two different methods: KMEANS and hierarchical clustering.

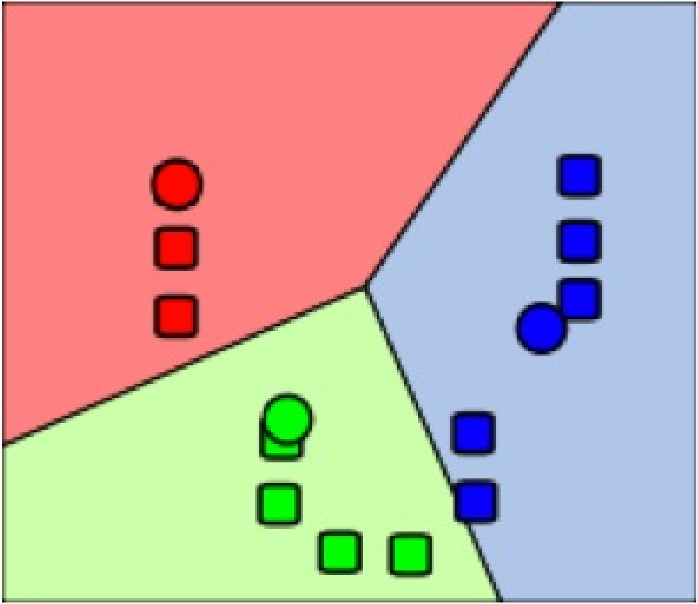
TFA: KMEANS clustering

- Initialize K random points in $\mathbb{R}^n$ as centers.
- Assign each light curve to the cluster it is closest to.
- Terminate when no new assignments are made.
- (Often sold as an instance of the EM algorithm in fitting a mixture of multivariate-normals, but this is not exactly true: there are data sets where the two methods will give you different results.)

TFA: KMEANS clustering



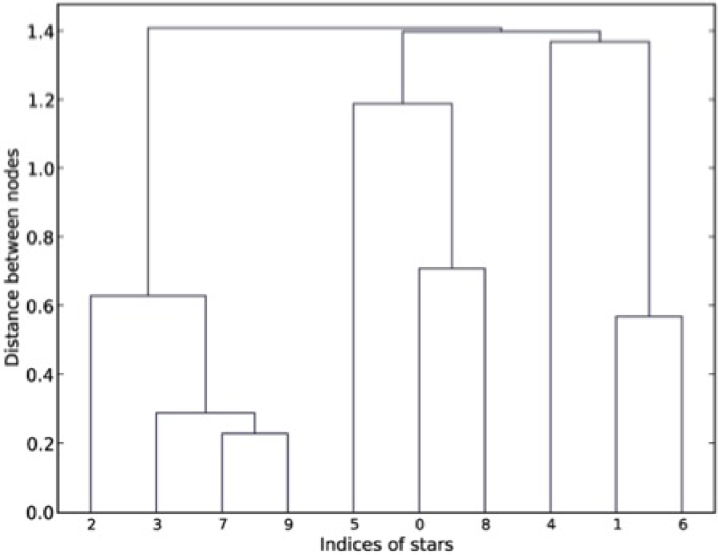
TFA: KMEANS clustering



TFA: Hierarchical Tree Clustering

- Compute a distance matrix for the lightcurves
- Compute a binary tree using the distance matrix
- Use the binary tree to determine clusters via a merging algorithm

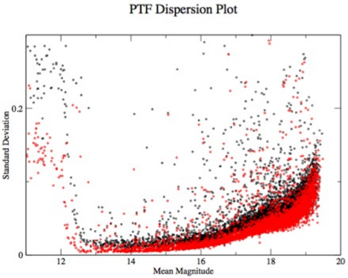
TFA: Hierarchical Tree Clustering

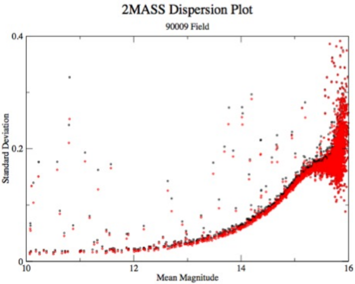


Forming the Clusters Given a Tree

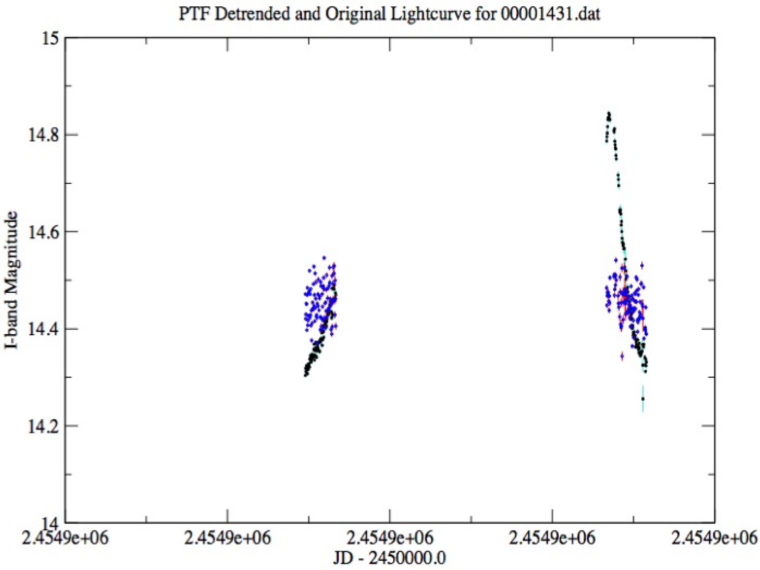
- Set initial clusters to be the singletons.
- Consider merging two closest nodes under one cluster.
- If the distribution of distances in this node is normally distributed, we have reason to believe all stars are correlated. (Seems ad-hoc...)

TFA Results on 2MASS and PTF

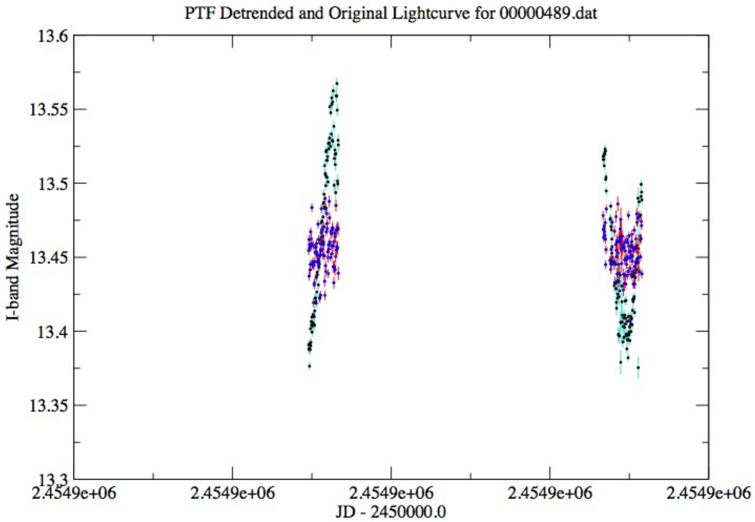




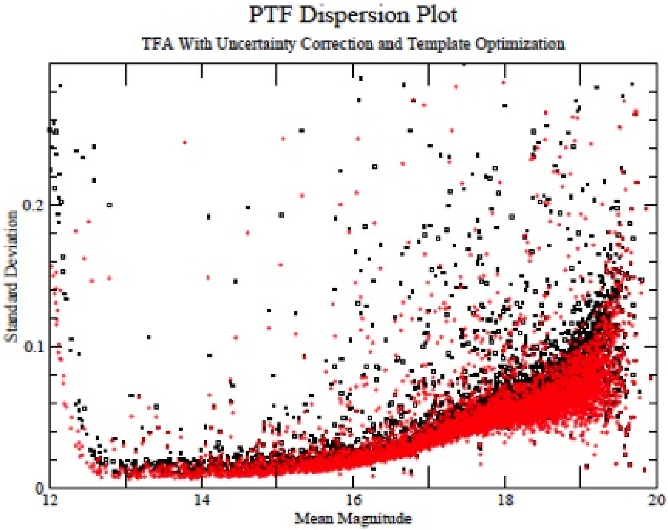
Overfitting



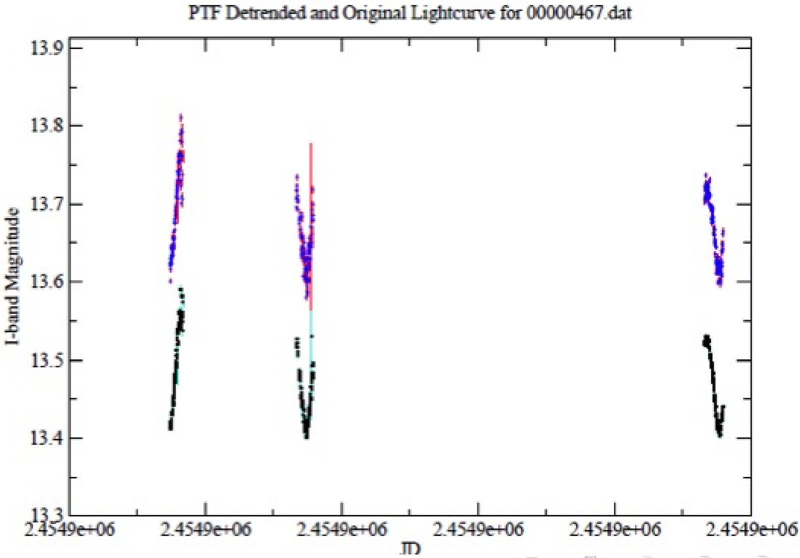
Overfitting



Using clustering to reduce the size of the template set



Using clustering to reduce the size of the template set



Future Directions

- Are we killing signal and therefore missing planets in the data?
- The selection of template trends is messy: we are mixing signal and noise!
- The method is not statistically principled; can we write down one cohesive probabilistic/statistical model for the generation of a widefield survey?
- Can we eliminate overfitting by regularization or wisely chosen priors for c? (The typical approach is to use L_1 penalization).
- Can we model systematic trends directly? What are their properties in the frequency domain?
- Finally, can we put these approaches together into a cohesive hieararchical Bayesian model?

Future Directions: regularization

- One way to eliminate overfitting is to shrink the coefficients of c.
- The non-Bayesian (LASSO) way: $\underset{c}{\operatorname{argmin}} ||Y - T^t c||_2^2 + \lambda ||c||_1$
- The Bayesian way: choose informative priors peaked at zero for c.

Future Directions: exploring the frequency domain

- Due to periodicity, systematic trends may be better represented in the frequency domain than the time domain.
- One possibility: use the orthoronormal, complete eigenbasis supplied by Sturm-Louiville operators: Fourier Transform!
- Another possibility: use a wavelet basis.
- Ideal situation: combine both: seperate trends with a strong frequency component and time components.

Future Directions: wavelets for systematics

- An alternative approach is to utilize a wavelet basis.
- Advantages:certain wavelets (due to Ingrid Daubchies) have compact support!
- Avoid Gibbs' phenomenon
- Potentially sparse in the frequency domain
- Localize in both frequency and wavelet domain.
- Contrast to Fourier Transform where there is a tradeoff in the time and frequency domain: e.g the eigenbasis of Momentum in QM corresponds to the Fourier Transform, so the uncertainty principle in QM is a restatement of this property of the FT.
- If some of the trends are strongly periodic, we'd probably want to use the FT.

A Bayesian Model For A Wide-Field Survey

- Draw K according to a poisson distribution.
- Draw K ''systematic" trends $T \in \mathbb{R}^{kxn}$ as a mixture of Multivariate Normals in the frequency domaIn. We may want to impose structure on the covariance matrices (e.g, treat them as Gaussian Processes).
- For each lightcurve, draw a vector of coefficients $c_i \in \mathbb{R}^k$ with a strongly informative prior peaked at 0 in each component.
- The observed signal is then $Y_i = T^t c_i + S_i$ For S_i the "true" signal.

A Bayesian Model For A Wide-Field Survey

- We need to incorporate chip position. (Some systematics are thought to be dependent on the position.)
- Computational considerations: TFA requires only matrix inversion. Presumably we will use a MCMC scheme to fit this model: how will the computational complexity compare?

References

- Kovacs, G., Bakos, G., Noyes, R. (2005). A Trend Filtering Algorithm for Wide-Field Variability Surveys. MNRAS. Vol. 356, 557-567.
- Kovacs, G., Bakos, G. (2008). Application of the Trend Filtering Algorithm in the Search for Multiperiodic Signals. Comm. In Asteroseismology. Vol. 157.
- Kim, D., et al. (2009). De-Trending Time Series for Astronomical Variability Surveys. arXiv:0812.1010v3.

Thanks

- Caltech SURF
- Dr. Peter Plavchan
- Xiao-Li
- Astrostats group