

The Use of Principal Components Analysis in Analysing Variable Star Data

S. M. Kanbur, D. Iono

*Department of Physics and Astronomy, University of Massachusetts,
Amherst, MA, USA*

N. A. Tanvir

*Department of Physics and Astronomy, University of Central
Hertfordshire, UK*

M. A. Hendry

Department of Physics and Astronomy, University of Glasgow, UK

Abstract. We describe the technique of Principal Components Analysis (PCA) as applied to the analysis of variable star data. It is shown that PCA is an efficient way of describing light curve structure.

1. Introduction

Recent microlensing surveys have produced a vast quantity of variable star data. While Fourier decomposition is an effective way to analyse the nonlinear structure of variable star light and velocity curves, we seek a way of analysing an entire family of light or velocity curves simultaneously. One approach is to use Principal Components Analysis (PCA). PCA seeks to explain variance in data by projecting it onto a set of orthogonal axes so as to maximise the variance of the projections. Let X_{ij} be the j^{th} observed point on the light curve of the i^{th} star, $1 \leq j \leq P$, $1 \leq i \leq N$. Then we can form the correlation matrix

$$S_{jk} = \sum_{i=1}^{i=N} X_{ji} X_{ki} . \quad (1)$$

This measures the correlation between the j^{th} and k^{th} point on the light curve, averaged over all N stars in the sample. It can be shown that in order to maximise the variance between a set of new orthogonal axes u^t it is necessary to solve the eigenvalue equation,

$$Su = \lambda u . \quad (2)$$

This equation results in a set of vectors u^t , the eigenvectors, which are the new axes we seek, and their corresponding eigenvalues, λ^t . We can project each

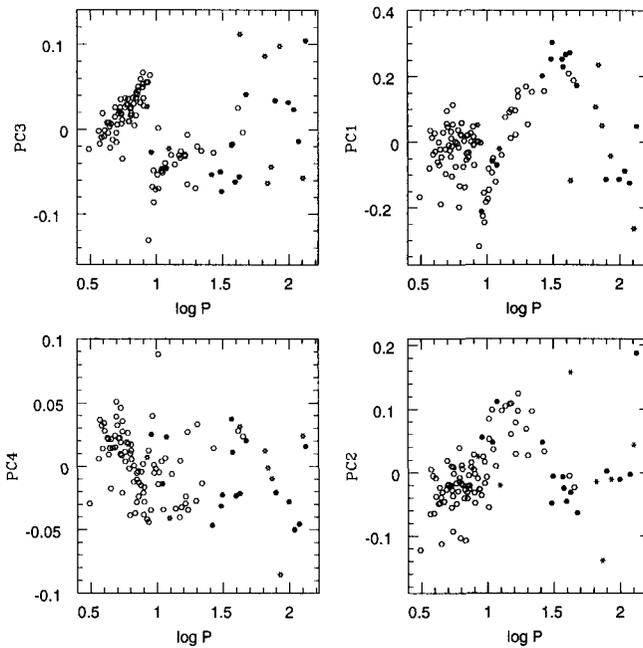


Figure 1. The first four principal components plotted against period.

observed light curve onto a given axis by forming the dot product,

$$PC_t(\text{star } i) = \sum_{j=1}^{j=P} X_{ij}u_j^t \quad (3)$$

In equation (3), we refer to the left hand side as the t^{th} principal component. This is a number for each star. The observed light curves can be reproduced by the formula

$$\sum_{t=1}^{t=P} (PC_t(\text{star } i))u^t \quad (4)$$

Equation (4) shows that an observed light curve is expressed as a linear combination of “basis light curves”, which are the eigenvectors u^t of the correlation matrix S . The quantity λ^{t2} , after suitable normalisation, shows the percentage variance in the sample explained by the t^{th} principal component. A plot of $PC_t(\text{star } i)$ against period indicates how that particular coefficient in equation (4) changes with period.

2. Data and Results

Using data from the Galaxy (Moffett & Barnes 1984) and the Clouds (Berdnikov & Turner 1995), we can perform the PCA analysis described in equations (1) to

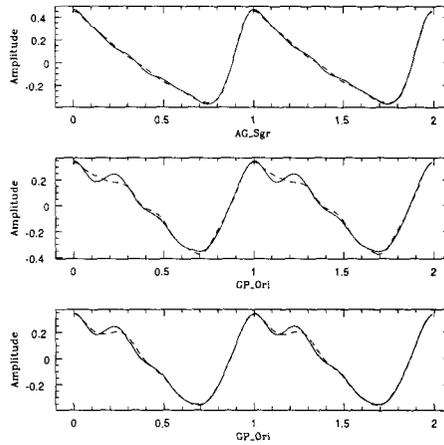


Figure 2. Reproduction of light curves using PCA, solid lines are data and dashed lines are reproductions using PCA.

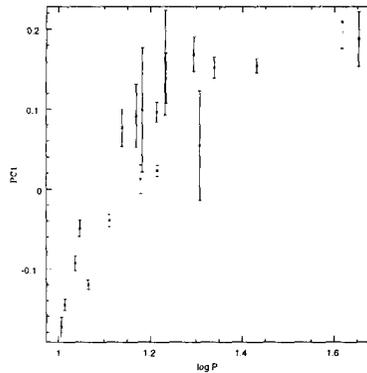


Figure 3. Error estimates for the first principal component.

(4). In our case the X matrix actually consisted of the Fourier coefficients a, b in the following decomposition of the observed light curves,

$$\sum_{j=1}^{j=N} (a_j \cos(k\omega t) + b_j \sin(k\omega t)) \quad , \quad (5)$$

so that $X_{i1} = a_1$, $X_{i2} = b_1$, and so on. Fig. 1 shows plots of $PC_i(\text{stari})$ against period. We see clearly the presence of the $P_2/P_0 = 0.5$ resonance at $P_0 = 10$ days. Further, there is an indication of resonances at longer period and intriguing features, for example, at $\log P = 1.6$. It's interesting to note that the first four principal components explain about 80–90% of the total variance.

Fig. 2 shows how the PCA coefficients in Fig. 1 can be used to reconstruct light curves. The middle and bottom panels show a bump Cepheid where we reconstruct the light curve with 4 and 6 parameters in equation (4), respectively.

Note that bump Cepheid curves can be reproduced with as few as 6 parameters in contrast to a Fourier decomposition where an order 6 fit or greater (12 parameters) is often required.

Fig. 3 shows our estimate of the error on the first PCA coefficient. We use the formulae derived by Petersen (1986) to estimate the error on the coefficients in equation (5). Then we use a convenient attribute of PCA: if the data noise is normally distributed then so is the noise on the principal components,

$$\sigma(\text{PCA coefficients}) = \sigma(X_{ij}).$$

Fig. 3 uses this expression to plot error bars on the first principal component, and indicates that the progressions with period seen in Fig. 1 are real. Similar figures exist for the other PCA components shown in Fig. 1.

3. Conclusions

The technique of Principal Components Analysis offers an efficient way to describe the structure of variable star light curves and is particularly appropriate for large homogeneous datasets (e.g. MACHO, EROS).

References

- Berdnikov, L. N. & Turner, D. G. 1995, *Astron. Lett.*, 21, 717
 Moffett, T. J. & Barnes, T. G. 1984, *ApJS*, 55, 389
 Petersen, J. O. 1986, *A&A*, 170, 59

Discussion

Ivan Andronov: 1. Orthogonal vectors (functions, components) are mostly used to make the error estimates smaller and uncorrelated. Are the principal light curves orthogonal?

2. How is the number of parameters determined, because the total number of principal components should be equal to the number of data?

3. Are the errors of the coefficients estimated?

Shashi Kanbur: 1. The principal light curves are orthogonal.

2. The number of parameters is determined by the solution of $Su = \lambda u$.

3. Yes, the errors are estimated as shown in the diagram.

Géza Kovács: In the figures you showed, none of the light curves exhibited bumps. How does your method of principal components analysis work if you have bump(s) on the light curve?

Shashi Kanbur: The first two eigenvectors have bumps on either side of the maximum. So the coefficient multiplying these eigenvectors changes to reflect bumps passing from one side to the other. The higher order eigenvectors also have bumps on them which could explain other bumps on observed light curves.