

Principal Component Analysis of RR Lyrae light curves

S. M. Kanbur^{1*} and H. Mariani¹

¹*Department of Astronomy, University of Massachusetts
Amherst, MA 01003, USA*

Received XX XXX 2003 / Accepted XX XXX 2003

ABSTRACT

In this paper, we analyze the structure of RRab star light curves using Principal Component Analysis. We find this is a very efficient way to describe many aspects of RRab light curve structure and: in many cases, a Principal Component fit with 9 parameters can describe a RRab light curve including bumps whereas a 17 parameter Fourier fit is needed. As a consequence we show statistically why the amplitude is also a good summary of the structure of a RR Lyrae light curve. We also use our analysis to derive an empirical relation relating absolute magnitude to light curve structure. In comparing this formula to those derived from exactly the same dataset but using Fourier parameters, we find that the Principal Component Analysis approach has distinct advantages. These advantages are, firstly, that the errors on the coefficients multiplying the fitted parameters in such formulae are much smaller, and secondly, that the correlation between the Principal Components is significantly smaller than the correlation between Fourier amplitudes. These two factors lead to reduced formal errors, in some cases estimated to be a factor of 2, on the eventual fitted value of the absolute magnitude. This technique will prove very useful in the analysis of data from existing and new large scale survey projects concerning variable stars.

Key words: RR Lyraes – Stars: fundamental parameters

1 INTRODUCTION

Kanbur et al (2002), Hendry et al (1999), Tanvir et al (2004) introduced the use of Principal Component Analysis (PCA) in studying Cepheid light curves. They showed that a major advantage of such an approach over the traditional Fourier method is that it is much more efficient: an adequate Fourier description requires, at best, a fourth order fit or 9 parameters, whilst a PCA analysis requires only 3 or 4 parameters with as much as 81% of the variation in light curve structure being explained by the first parameter. Later, Leonard et al (2003) used the PCA approach to create Cepheid light curve templates to estimate periods and mean magnitudes for HST observed Cepheids. The purpose of this paper is to apply the PCA technique to the study of RR Lyrae light curves.

The mathematical formulation and error characteristics of PCA are given in K02 and will only be summarized here.

2 DATA

The data used in this study were kindly supplied by Kovacs (2002 private communication) and used in Kovacs and

Walker (2002, hereafter KW). These data consist of 383 RRab stars with well observed V band light curves in 20 different globular clusters. KW performed a Fourier fit to these data, which, in some cases, is of order 10. Details concerning the data can be found in KW. The data we work with in this paper is this Fourier fit to the magnitudes and we assume that the Fourier parameters published by KW are an accurate fit to the actual light curves. We start with the data in the form used in KW: a list of the mean magnitude, period and Fourier parameters for the V band light curve. The light curve can thus be reconstructed using an expression of the form

$$V = A_0 + \sum_{k=1}^{k=N} A_k \sin(k\omega t + \phi_k), \quad (1)$$

where A_0 is the mean magnitude, $\omega = 2\pi/P$, P the period, A_k, ϕ_k the Fourier parameters given in KW. These light curves are then rephased so that maximum light occurs at phase 0 and then rewritten as

$$V = A_0 + \sum_{k=1}^{k=N} (a_k \cos(k\omega t) + b_k \sin(k\omega t)). \quad (2)$$

The a_k, b_k are the light curve characteristics entering into the PCA analysis (K02). We then solve equation (4) of K02, either after, or before removing an average term from the

* Email: shashi@astro.umass.edu

Fourier coefficients in equation (2). With PCA, the light curve is written as a sum of "elementary" light curves,

$$V(t) = PCA1.L_1(t) + PCA2.L_2(t) + PCA3.L_3(t) + \dots, \quad (3)$$

where $V(t)$ is the magnitude at time t , $PCA1, PCA2..$ etc. are the PCA coefficients and the $L_i(t), i = 1, 2, 3..$ are the elementary light curves at phase or time t . These elementary light curves are not a priori given, but are estimated from the dataset in question. Each star has associated with it a set of coefficients $PCA1, PCA2, \dots$ and these can be plotted against period just as the Fourier parameters in equation (1) are plotted against period. We also note that the PCA results are achieved as a result of the analysis of the *entire* dataset of 383 stars whereas the Fourier method produces results for stars individually. This feature of PCA is particularly useful when performing an ensemble analysis of large numbers of stars obtained from projects such as OGLE, MA-CHO and GAIA.

3 RESULTS

Solving equation (4) of K02 yields the Principal Component scores and the amount of variation carried by each component. What we mean by this is the following: if we carry out an N^{th} order PCA fit, then PCA will assume that all the variation in the dataset is described by 10 components and simply scale the variation carried by each component accordingly. Table 1 shows this latter quantity with and without the average term removed. We see that in the case when we do not remove the average term the first PC explains as much as 97% of the variation in the light curve structure. In the case when we do remove the average term from the Fourier coefficients, the first PCA coefficient explains as much 81 percent of the variation in light curve structure. In either case, the first four components explain more than 99.99% of the variation.

Figures 1 and 2 show some representative light curves from our RRab dataset. In each panel of these two figures, the solid line is the Fourier decomposition of order 15 (that is 31 parameters) used by KW, whilst the dashed line is a PCA generated light curve of order 14 (that is 15 parameters). Straightforward light curves such as the one given in the top and bottom left panels of figures 1 and 2 are easily reproduced by our method. The top left panel of figure 1 provides an example of an RRab light curve with a dip and sharp rise at a phase around 0.8. This is well reproduced by PCA. It could be argued that PCA does not do as well as Fourier in mimicking this feature, for example, in the bottom right panel of figure 2. However, the difference in the peak magnitudes at a phase of around 0.8 is of the order of 0.02mags. It is also important to remember that the PCA method is an ensemble method and analyzes all stars in a dataset simultaneously. With Fourier, it is possible to tailor a decomposition to one particular star. This difference can be seen either as a positive or negative point about either technique. Given this, we contend that PCA does remarkably well in describing the full light curve morphology of RRab stars. On the other hand, the Fourier curve in the bottom left panel of figure 2 at this phase is not as smooth as the PCA curve.

In fact the PCA curves do not change much after about

8 PCA parameters. Even though table 1 implies that the higher order PCA eigenvalues are small, we feel justified in carrying out such a high order PCA fit because its only after about 8 PCA components that the fitted light curve assumes a stable shape. The left panel of figure 3 displays an eighth order PCA fit (9 parameters, dashed line) and a fourth order Fourier fit (9 parameters, solid line). The Fourier curve still has some numerical wiggles whilst the PCA curve is smoother. In addition, the two curves disagree at maximum light. The right panel of figure 3 shows, for the same star, the same order PCA curve as the left panel and an eighth order Fourier fit (17 parameters). Now the two light curves agree very well. Note that in portraying the PCA and Fourier fits of reduced order in this figure, we simply truncated the original representations to the required level.

We suggest that figures 1-3 and table 1 provide strong evidence that PCA is an *efficient* way to describe RRab light curve structure without compromising on what light curve features are captured by this description.

Figures 4-6 display plots of the first three PC scores plotted against log period for our sample. The errors associated with these PCA scores is discussed in section 4 of K02 and given in equation 6 of that section. The orthogonal nature of these scores may well provide insight into the physical processes causing observable features in the light curve structure. A detailed study of these plots, in conjunction with theoretical models, is left for a future paper.

Figure 7 graphs V band amplitude against the first PCA coefficient (after averaging). We see a very tight correlation. Since table 1 implies that PCA1 explains about 81% of the variation in light curve structure, figure 6 shows that the amplitude is a good descriptor of RRab light curve shape. Although the Fourier amplitudes are also correlated with amplitude, with PCA, we can quantify, very easily, the amount of variation described by each PCA component. This has implications for both modeling and observation. On the modeling side, a computer code that can reproduce the observed amplitude at the correct period, will also do a good job of reproducing the light curve structure. On the observational side, this provides insight into why we can use the amplitude, rather than a full blown PCA or Fourier analysis, to study the *general* trends of light curve structure. This is why comparing theoretical and observational RRab light curves on period-amplitude diagrams works reasonably well, though we caution that a careful analysis should consider the finer details of light curve structure.

Figures 6 and 7 display plots of the first two PCA coefficients and Fourier amplitudes, respectively, for our data, plotted against each other. Whilst A_1 and A_2 are correlated with each other, $PCA1$ and $PCA2$ are not, by construction. A similar situation would occur had we plotted A_1 or A_2 against A_3 . This is another advantage of PCA analysis of variable star light curves: the different PCA components are orthogonal to each other. A practical advantage of this feature is outlined in the next section.

4 LIGHT CURVE LUMINOSITY RELATIONS

A major goal of stellar pulsation studies is to find formulae linking global stellar parameters such as luminosity or metallicity to structural light curve properties. If we are in-

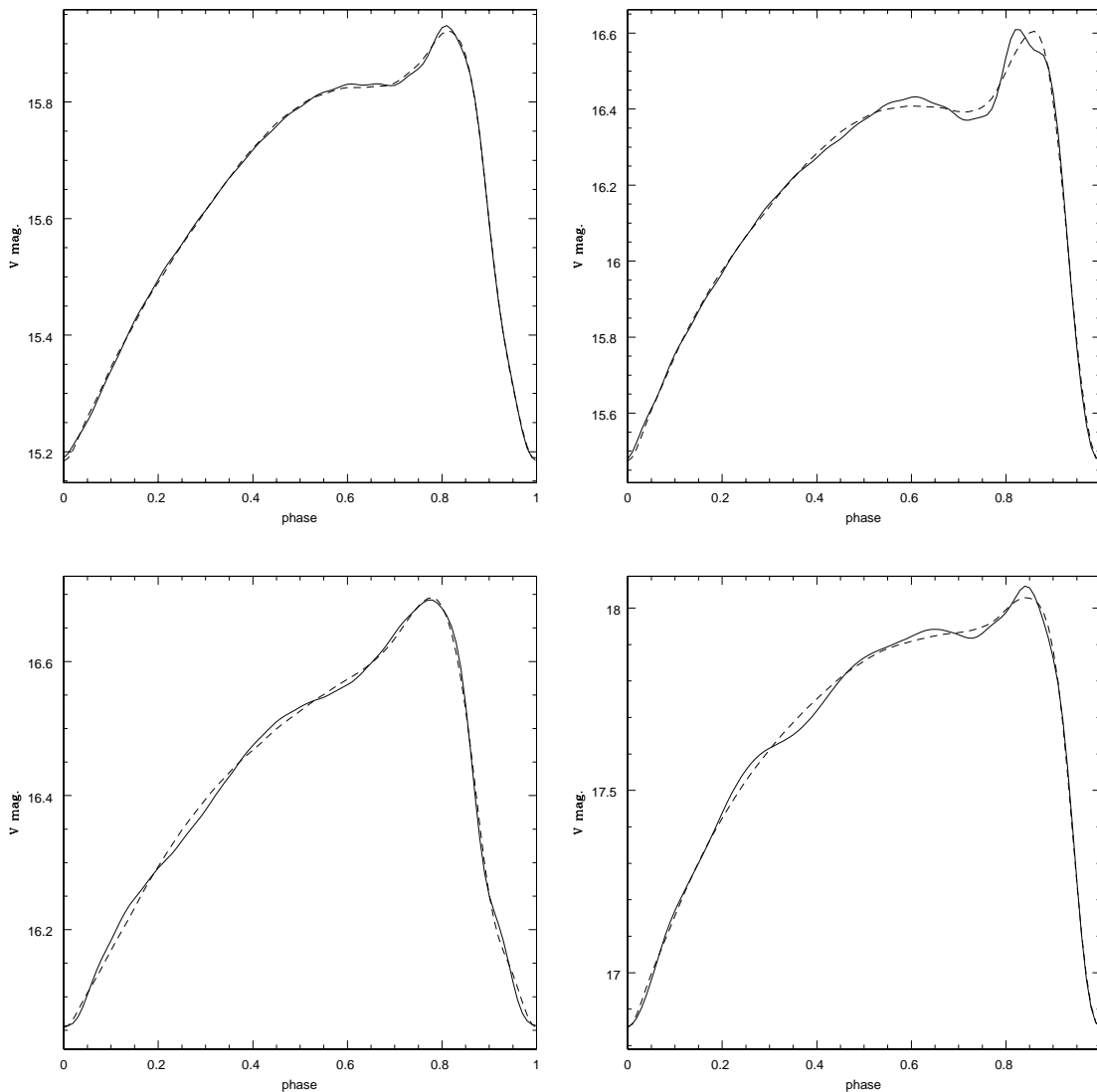


Figure 1. Light curve reproduction using Fourier (solid lines) and PCA (dashed lines) methods

interested in the V band magnitude, then we can write,

$$M_v = f(\text{lightcurvestructure}),$$

where, since we do not know the function f , we try to estimate it empirically. Two different approaches to quantifying light curve structure will, in general, yield different formulations of the function f , but if there does exist a true underlying function f , then both methods should give similar answers for M_v , given the same input data. With a Fourier based method, the function f is related to the Fourier amplitudes and phases, A_k, ϕ_{k1} , usually with a linear relation. With a PCA approach, we use the PCA scores plotted in figures 2-4. Hence a PCA relation, though also linear, will be different. The nature of PCA implies that the error structure in such formulae will be simpler and we quantify this below. Both formulations should, of course, give similar numbers for the final estimated value of the physical parameter in question, in this case, M_v .

KW used the Fourier method and found relations of the

form,

$$M_v = \text{const.} - 1.82 \log P - 0.805A_1, \quad (4)$$

and,

$$M_v = \text{const.} - 1.876 \log P - 1.158A_1 + 0.821A_3. \quad (5)$$

We note that these relations were obtained through an iterative procedure whereby outliers were removed and the relations re-fitted (Kovacs 2004). In this paper, we use the PCA method, but also, we use the entire dataset C mentioned in KW, consisting of 383 stars, and fit the relations just once. We do not remove any outliers. This may be why we obtain slightly different versions of the fit using Fourier parameters than that published in KW. For ease of comparison, we include in table 2 results obtained using both PCA and Fourier parameters. This table gives the name for the relation, the independent variables considered and coefficients together with their standard errors. The value of

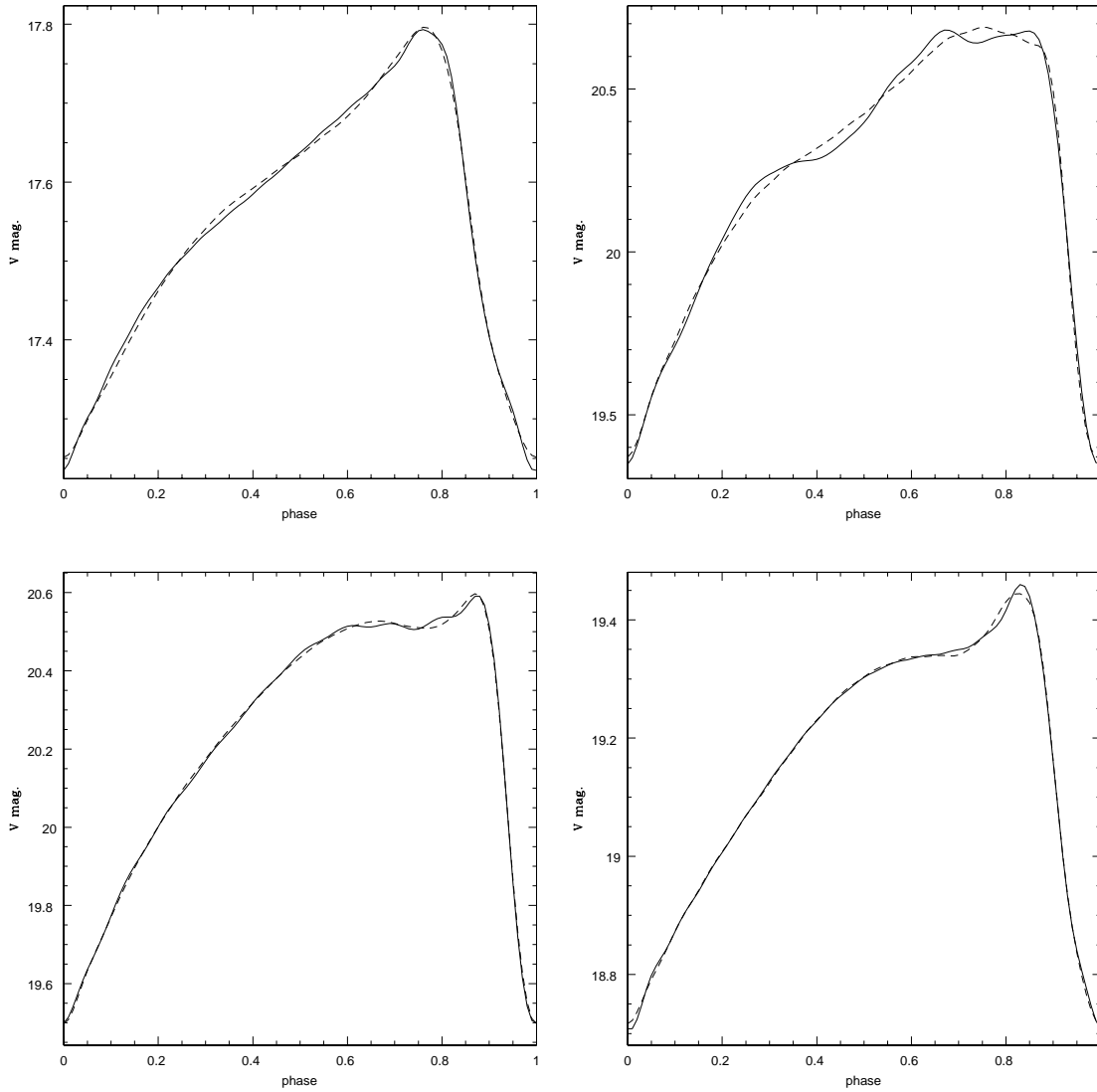


Figure 2. Light curve reproduction using Fourier (solid lines) and PCA (dashed lines) methods

chi-squared in the table is defined as

$$\sum_{k=1}^{k=N} (M_v - \hat{M}_v)^2 / (N - p), \quad (6)$$

where \hat{M}_v is the fitted value of M_v and N, p are the number of stars and parameters respectively in the fit. An examination of this table strongly suggests that

- 1) Similar relations to equations (4) and (5) between M_v and the PCA coefficients exist.
- 2) We can use an F test (Weisberg 1980) to test for the significance of adding a second and then a third PCA parameter to the regression. The F statistic we use is

$$\frac{(RSS_{NH} - RSS_{AH}) / (df_{NH} - df_{AH})}{RSS_{AH} / df_{AH}}, \quad (7)$$

where RSS_{NH}, RSS_{AH} are the residual sum of squares under the null and alternate (NH and AH) hypothesis respectively. Similarly, df_{NH} and df_{AH} are the degrees of freedom

under these two hypotheses. For this problem, the null hypothesis is that the model with the smaller number of parameters is sufficient whilst the alternative hypothesis is that the model with the greater number of parameters is required. Under the assumption of normality of errors, equation (7) is distributed as an $F_{(df_{NH} - df_{AH}), df_{AH}}$, (Weisberg, 1980, p. 88). The large number of stars and an appeal to the central limit theorem suggests that the normality of errors assumption is valid. Applying this F test implies firstly, that adding the first parameter $PCA1$ is a significant addition to $\log P$ and secondly, that adding a second and third parameter, $PCA2$ and $PCA3$ are also highly significant with a p value less than 0.0004. In the case of Fourier parameters, adding the A_1 parameter to $\log P$ is highly significant and adding the A_3 parameter to this is also highly significant. However, a formula involving $(\log P, A_1, A_2)$ has a p value of 0.0058 and a formula involving all 3 Fourier amplitudes and $\log P$ is not a significant addition to a formula involving $(\log P, A_1, A_3)$.

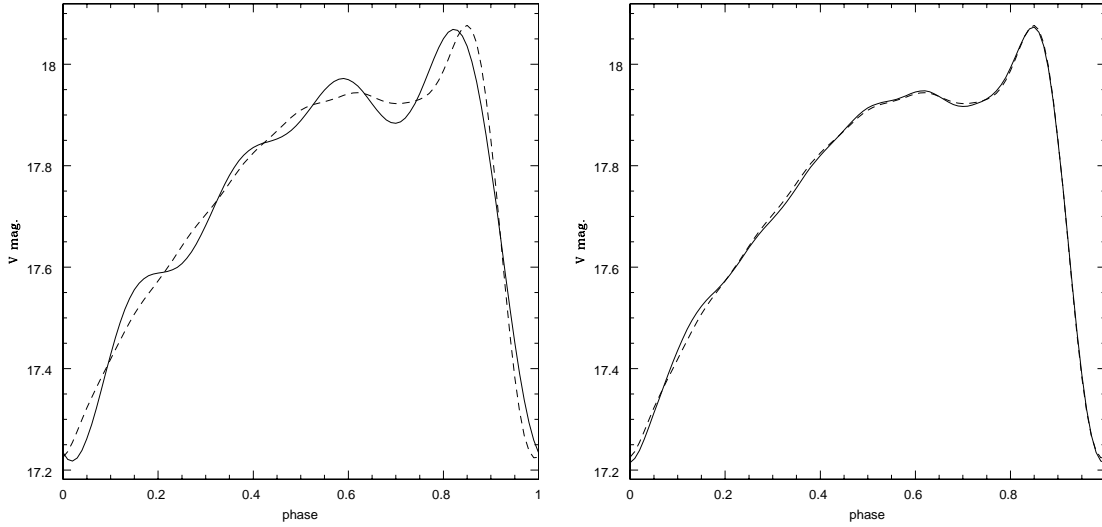


Figure 3. Light curve reproduction using Fourier (solid lines) and PCA (dashed lines) methods. The left panel is a fourth order (9 parameters) Fourier fit and an eight order PCA (9 parameters) fit. The right panel is an eight order (17 parameters) Fourier fit and an eight order PCA (9 parameters) fit. (

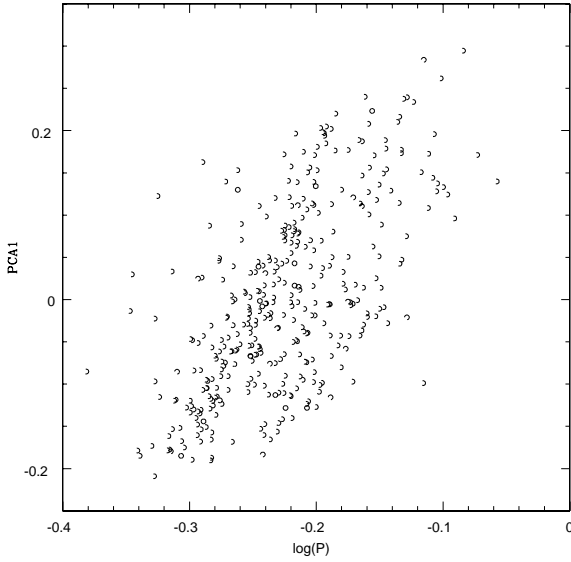


Figure 4. Plot of first Principal Component against log period.

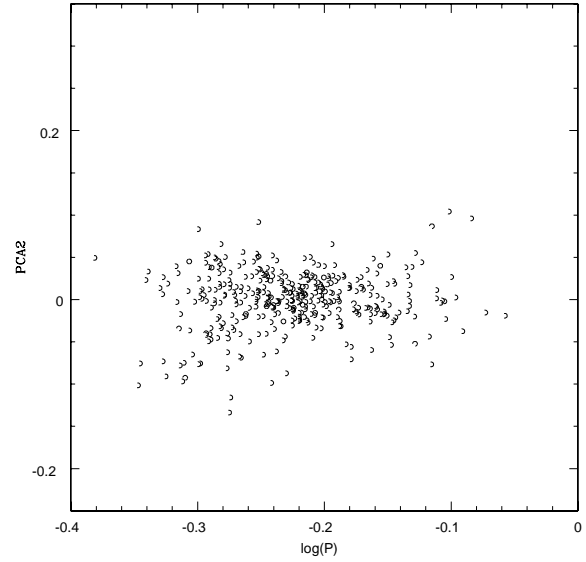


Figure 5. Plot of second Principal Component against log period.

- 3) The standard deviation of the fits given in the last column is generally slightly higher for the PCA case, when considering similar numbers of parameters. This is perhaps caused by the fact that the different PCA components carry orthogonal sets of information.

- 4) The errors on the coefficients in the PCA fits are always significantly smaller. This is an important point when we evaluate the errors on the final fitted value of the absolute magnitude.

- 5) If we write the absolute magnitude as a function of parameters, x_1, x_2, \dots, x_N ,

$$M_v + const. = f(x_1, x_2, \dots, x_N), \quad (8)$$

then the error on the absolute magnitude is given by,

$$\sigma^2(M_v + const.) = \sum_{k=1}^{k=N} \sigma^2(x_k) \left(\frac{\partial f}{\partial x_k} \right)^2 + \sum_{i,j=1, i \neq j}^N \sigma^2(x_i, x_j) \left(\frac{\partial f}{\partial x_i} \right)^2 \left(\frac{\partial f}{\partial x_j} \right)^2. \quad (9)$$

As table 2 indicates, $\sigma^2(x_k)$ is always smaller when the x_k are PCA coefficients rather than Fourier amplitudes. Figure 8 and 9 portray graphs of $PCA1$ vs $PCA2$ and A_1 versus A_2 respectively. We note that $\rho_{i,j} \sigma(x_i) \sigma(x_j) = \sigma^2(x_i, x_j)$. Table 3 presents sample correlation and covariance coefficients

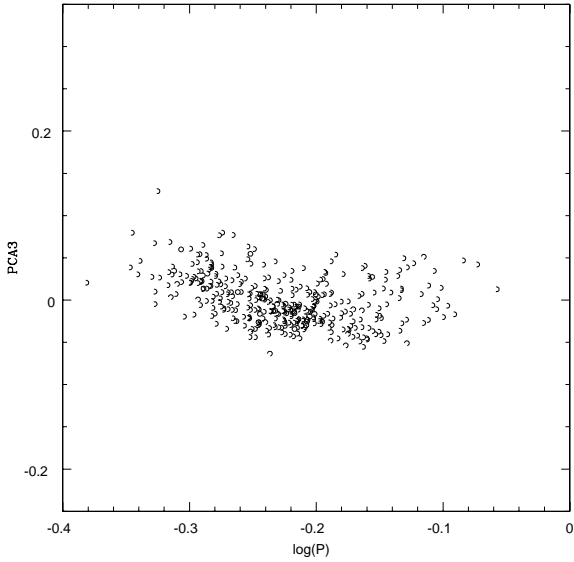


Figure 6. Plot of third Principal Component against log period.

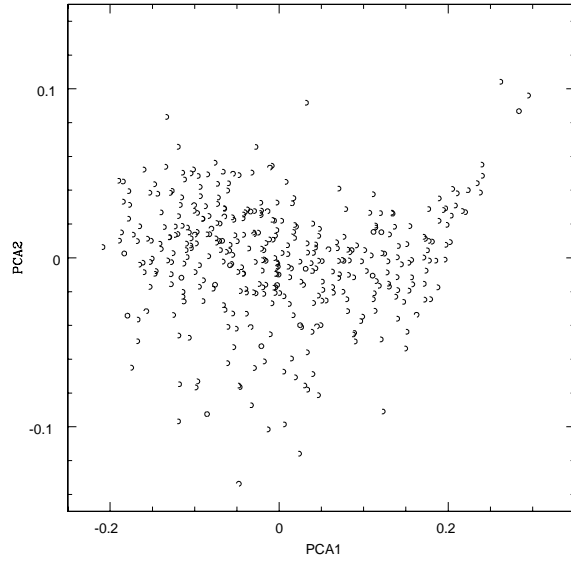


Figure 8. Plot of first Principal Component against second Principal Component.

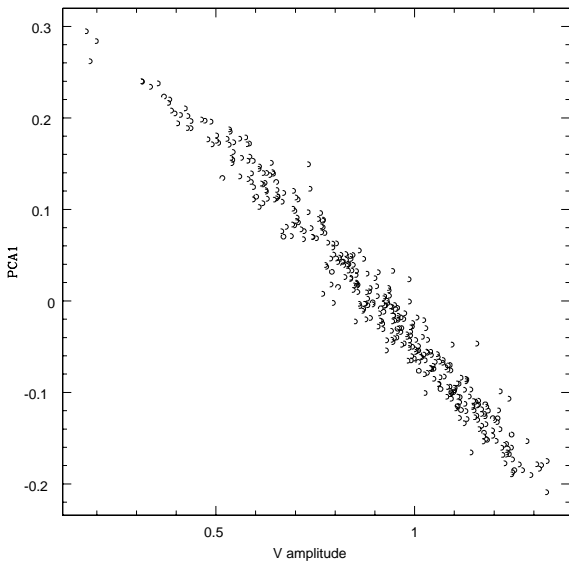


Figure 7. Plot of V band amplitude against the first PCA coefficient.

between the period and PCA parameters and period and Fourier parameters. Table 3, and figures 6 and 7 demonstrate that the correlation coefficient amongst any pair of PCA coefficients is smaller than between any pair of Fourier coefficients. Hence the error on the fitted value of M_v , $\sigma^2(M_v)$, has to be smaller when using a PCA based formula. We can use table 3 and equation (9) to formally calculate the error on $M_v + const$. Table 4 presents these results. The label in the top row of this table refers (P1, F1, etc.) to the appropriate relation in table 2. We see clearly that the PCA formulae do better than their Fourier counterparts with a similar number of parameters. When we consider the $(\log P, PC1, PC2)$ and $(\log P, A1, A3)$ variables, then the "error advantage" using a

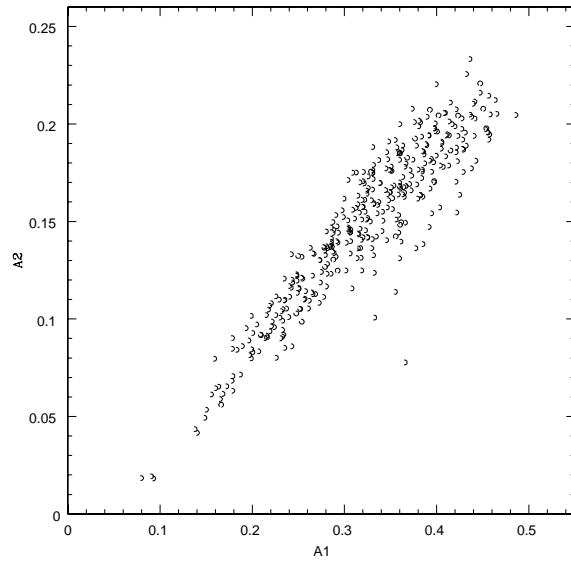


Figure 9. Plot of first Fourier amplitude against second Fourier amplitude.

PCA based method is a factor of two. This occurs not just because the PCA coefficients are orthogonal to each other, but also because the errors on the coefficients in a PCA based formula are significantly smaller than in the Fourier case.

Figure 10 displays a plot of the predicted absolute magnitudes obtained using a two parameter $(\log P, A1, A3)$ Fourier fit and the three parameter $(\log P, PCA1, PCA2, PCA3)$ PCA fit. The two approaches are displaced from each other because we do not consider the constants in this study. Disregarding this, it can be seen that

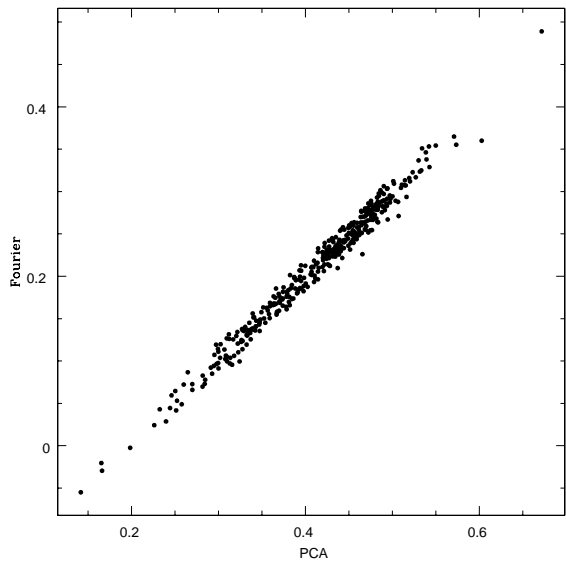


Figure 10. Plot of fitted $M_v + const$ values when using Fourier and PCA methods.

Table 1. Percentage of variation explained by PC components

	PC1	PC2	PC3	PC4	PC5	PC6
without average	81.4	7.8	5.7	2.3	0.74	0.57
with average	96.9	1.9	0.55	0.25	0.07	0.006

the slope of this plot is 1: hence the two methods produce similar relative absolute magnitudes.

5 CONCLUSION

We have shown that the method of PCA can be used to study RR Lyrae light curves. It has distinct advantages over a Fourier approach because

- a) It is a more efficient way to characterize structure since fewer parameters are needed. A typical Fourier fit requires 17 parameters whereas a PCA fit may only need 9.
- b) Using the PCA approach, we see clearly why the amplitude is a good descriptor of RRab light curve shape.
- c) The different PCA components are orthogonal to each other whereas the Fourier amplitudes are highly correlated with each other. This leads to relations linking light curve structure to absolute magnitude using PCA having coefficients with smaller errors and leading to more accurate estimates of absolute magnitudes. This can reduce the formal error, in some cases, by a factor of 2.

In future work we plan to investigate the applicability of this method to light curve structure-metallicity relations, RRc stars and a comparison of observed and theoretical light curves using PCA.

ACKNOWLEDGMENTS

SMK thanks Geza Kovacs for stimulating discussions and for kindly supplying the RRab dataset. SMK thanks D. Iono for help writing the PCA and least squares program and C. Ngeow for help with latex. HM thanks FCRAO for providing a summer internship in 2001 when part of this work was completed.

REFERENCES

- Hendry, M. A., Tanvir, N. A., Kanbur, S. M., 1999, ASP Conf. Series, 167, p. 192
 Kanbur, S., Iono, D., Tanvir, N. & Hendry, M., 2002, MNRAS, 329, 126
 Kovacs, G., Walker, A. R., 2001, A&A, 371, 579
 Kovacs, G., 2004, private communication
 Leonard, D., Kanbur, S.,M., Ngeow, C., Tanvir, N., 2003, ApJ, 594, 247
 Tanvir, N. R., Hendry, M. A., Kanbur, S. M., 2004, in preparation
 Weisberg, S., 1980, *Applied Linear Regression*, John Wiley & Sons, 1st Ed.

Table 2. Light curve luminosity relation using PCA and Fourier methods.

	$\log P$	first	second	third	chisquare
PCA					
P0	-1.134 ± 0.059				0.00321
P1	-1.550 ± 0.082	0.269 ± 0.038			0.00283
P2	-1.609 ± 0.082	0.290 ± 0.038	0.291 ± 0.082		0.00274
P3	-1.744 ± 0.088	0.329 ± 0.039		-0.539 ± 0.107	0.0027
P4	-1.829 ± 0.088	0.359 ± 0.039	0.336 ± 0.079	-0.583 ± 0.105	0.00253
Fourier					
F1	-1.677 ± 0.083	-0.472 ± 0.054			0.00266
F2	-1.700 ± 0.082	-0.726 ± 0.092		0.613 ± 0.179	0.00258
F3	-1.740 ± 0.085	-0.758 ± 0.116	0.536 ± 0.193		0.00261
F4	-1.720 ± 0.085	-0.790 ± 0.117	0.215 ± 0.243	0.490 ± 0.227	0.00258

Table 3. Sample correlation and covariance coefficients between period, PCA and Fourier coefficients

	$\log P, PCA1$	$\log P, PCA2$	$\log P, PCA3$	$PCA1, PCA2$	$PCA2, PCA3$	$PCA1, PCA3$
correlation	0.631	0.099	-0.299	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
covariance	0.0038	0.0002	-0.0006	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
	$\log P, A_1$	$\log P, A_2$	$\log P, A_3$	A_1, A_2	A_2, A_3	A_1, A_3
correlation	-0.655	-0.529	-0.562	0.926	0.931	0.902
covariance	-0.0028	-0.0012	-0.0011	0.0030	0.0013	0.0024

Table 4. Formal error on $M_v + const.$ for PCA and Fourier relations

P1	P2	P3	P4	F1	F2	F3	F4
0.0139	0.0142	0.0216	0.0240	0.0156	0.0313	0.0394	0.0311