Mon. Not. R. Astron. Soc. **363,** 749–762 (2005) doi:10.1111/j.1365-2966.2005.09466.x

Determination of Cepheid parameters by light-curve template fitting

N. R. Tanvir,[1]⋆ M. A. Hendry,[2] A. Watkins,[1] S. M. Kanbur,[3] L. N. Berdnikov[4] and C. C. Ngeow[5]

[1] *Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield, Herts AL10 9AB*
[2] *Department of Physics and Astronomy, University of Glasgow, Glasgow, G12 8QQ*
[3] *Department of Physics, State University of New York, Oswego, NY 13126, USA*
[4] *Sternberg Astronomical Institute, Universitetskij Prospekt 13, Moscow 119992, Russia*
[5] *Department of Astronomy, University of Illinois, 1002 W Green Street, Urbana-Champaign, IL 61801, USA*

Accepted 2005 July 26. Received 2005 July 13; in original form 2005 March 14

ABSTRACT

We describe techniques to characterize the light curves of regular variable stars by applying principal component analysis (PCA) to a training set of high-quality data, and to fit the resulting light-curve templates to sparse and noisy photometry to obtain parameters such as periods, mean magnitudes etc. The PCA approach allows us to efficiently represent the multiband light-curve shapes (LCSs) of each variable, and hence quantitatively describe the average behaviour of the sample as a smoothly varying function of period, and also the range of variation around this average.

In this paper we focus particularly on the utility of such methods for analysing *Hubble Space Telescope* (*HST*) Cepheid photometry, and present simulations which illustrate the advantages of our PCA template-fitting approach. These are: accurate parameter determination, including LCS information; simultaneous fitting to multiple passbands; quantitative error analysis; objective rejection of variables with non-Cepheid-like light curves or those with potential period aliases.

We also use PCA to confirm that Cepheid LCSs are systematically different (at the same period) between the Milky Way and the Large and Small Magellanic Clouds, and consider whether LCS might therefore be used to estimate the mean metallicities of Cepheid samples, thus allowing metallicity corrections to be applied to derived distance estimates.

Key words: Cepheids – stars: variables: other.

1 INTRODUCTION

Many astrophysical investigations rely on the determination of parameters of periodic variable stars. Notably, the use of Cepheid variables as distance indicators requires estimation of periods and (usually) intensity-mean magnitudes in order to establish a period–apparent luminosity relation. With sparse and noisy data this is hard to do reliably. Given the large investment of *Hubble Space Telescope* (*HST*) time in observations of Cepheids in nearby galaxies (e.g. Tanvir et al. 1995; Freedman et al. 2001; Saha et al. 2001), it is particularly important for the techniques employed to be as accurate and efficient as possible.

A number of algorithms have been developed to objectively estimate variable star parameters. Notably the ‘string length’ method of Lafler & Kinman (1965), which essentially minimizes square magnitude differences between successive phased data points, is still frequently used to determine periods. This method works well, especially with precise and well-sampled data, but is likely to be less secure with data ‘at the limit’ – i.e. close to the limiting apparent magnitude of the photometry and/or with sparse phase coverage. To find intensity-mean magnitudes, many authors use the phase-weighted method suggested by Saha & Hoessel (1990), which makes allowance for the non-uniform sampling of the light curve in time. Again, this works well with good data, but is potentially inefficient (in the sense of not making full use of all the data) with sparsely-sampled data.

Most *HST* studies have gone one step further in using the shape of the light curve in the *V* band to predict its form in the *I* band, and hence to allow the *I*-band intensity-mean magnitude to be estimated from only a very few photometric data points. The motivation for this approach is to provide colour information relatively cheaply, which is required to estimate – and then correct for – reddening by dust.

The simplest such recipe (Freedman 1988) uses only prior knowledge of the typical ratio of *V*- to *I*-band amplitude and the typical

⋆E-mail: nrt@star.herts.ac.uk

phase shift between *V* and *I* bands at maximum-light for Cepheids. With this model, a correction can be made to the *I* mean photometry, assuming that it is the same as the correction which would have to be applied to an equivalently undersampled *V* light curve, multiplied by the adopted ratio of amplitudes. Obviously, as with other similar methods, errors are introduced here, both those dependent on the *V* and *I* photometric quality (or lack thereof) and possibly also the accuracy of the prior information. Subsequently a rather more sophisticated algorithm was developed by Labhardt, Sandage & Tammann (1997). This involves predicting and fitting a template light-curve in the *I* band based on the parameters (i.e. the period, phase, amplitude and shape) already determined from the *V*-band data. The strong correlations between the light curves of Cepheids in different bands make this a productive approach.

Fitting template light-curves as a means of estimating Cepheid parameters was first introduced by Stetson (1996) who used templates based on Fourier decomposition of a set of well-observed MW and Large Magellanic Clouds (LMC)/Small Magellanic Clouds (SMC) Cepheids. In his method, initial values of plausible periods are determined by string-length analysis, and then templates fitted with each of these periods as a starting point, and the overall amplitude left as a free parameter (in addition to the period, phase and mean magnitudes). A scoring system is then used to identify the most plausible fit. Stetson argued that the advantage of automated classification of variables and determination of their parameters is not so much that a computer algorithm will necessarily do better than an experienced human analyst, but that the biases and systematics can be more easily studied and characterized.

A further refinement to the Fourier-fitting method was presented in Ngeow et al. (2003), where 'simulated annealing' is used to improve the quality of the Fourier decomposition of sparsely sampled *HST V*-band light curves. This technique restricts the allowed range for the Fourier amplitudes in the minimization procedure, and thus performs substantially better than conventional least-squares fitting on data with significant gaps in phase coverage. *I*-band light curves are reconstructed from the *V* band using interrelations of the Fourier coefficients.

A new approach to Cepheid light-curve template generation was introduced by Tanvir, Ferguson & Shanks (1999; see also Hendry, Tanvir & Kanbur 1999) who used Principal Component Analysis (PCA) to statistically characterize a training set of MW, LMC and SMC Cepheids, and fitted these templates to *V* and *I* data for *HST* observed Cepheids in M96. By fitting well-defined and realistic template curves, several parameters can be determined, together with estimates of their uncertainties. One of the very attractive features of this technique is that photometry in different bands can be handled simultaneously, so that the natural correlations between bands are automatically built into the templates and all of the data is used to determine the parameters. Kanbur et al. (2002) described the PCA method in more detail, used it to investigate variation in Cepheid light-curve structure as a function of period, and described the error properties of the PCA coefficients. PCA template fitting was also successfully applied to *HST*-observed Cepheids in NGC 1637 by Leonard et al. (2003).

In this paper we provide a complete description of the PCA-based method of characterizing light curves, present an updated training set and consider in detail the subsequent template-fitting algorithm which was used in Tanvir et al. (1999) and Leonard et al. (2003). We describe simulations which illustrate the potential of the methods, and discuss future directions. Although we focus on their application to *V*- and *I*-band Cepheid data, these techniques may easily be extended to other passbands and also used to analyse other classes of periodic variable stars. For example, Kanbur & Mariani (2004) consider PCA of photometric data for RRab stars.

The structure of the paper is as follows. In Sections 2 and 3 we present our training set of well-observed MW and LMC Cepheids and describe in more detail how PCA is applied in order to define the template light-curves, and the advantages of this approach over other methods. In Section 4 we discuss an algorithm for fitting templates to noisy data in order to estimate light-curve parameters and their errors. Of course, such a procedure is required however the templates are generated, but in our case the fitting process also returns estimates for the coefficients of the first two principal components. In Section 5 we then go on to generate simulations of poorly sampled Cepheids with noisy photometry, mimicking 'typical' and 'difficult' *HST* data sets, and extract their parameters by template fitting. We consider distance determination and light-curve parameter estimation using both *mean*- and *maximum*-light estimates. This serves to illustrate how our method performs in practice compared to other methods, and also allows us to explore the limits of *HST*-like data sets. In Section 6 we introduce an SMC Cepheid sample, and consider the question of whether light-curve shape (LCS), for either individual Cepheids or averaged over populations, contains other useful information – in particular its potential as an indicator of metallicity. Our conclusions are given in Section 7.

2 PRINCIPAL COMPONENT ANALYSIS OF OUR TRAINING SET

PCA is a widely used statistical tool and has been applied in recent years to a number of astrophysical problems, such as spectral classification, photometric redshift determination and morphological analysis of galaxy surveys (e.g. Li, Kong & Cheng 2001). For a detailed account of the statistical basis of PCA the reader is referred to e.g. Morrison (1967). The central principle behind PCA is easily stated, however: it provides a means of transforming a multidimensional data set consisting of a number of statistically dependent variables into a set of statistically independent variables, which are the principal components. Specifically, the first principal component is determined to be the linear combination of the original variables which accounts for as much of the variability in the data as possible; the second principal component is the linear combination which accounts for as much of the remaining variability as possible – subject to the constraint that it is orthogonal to the first principal component – and so on. In many situations the first few principal components may explain a high proportion of the variability in the data, so that one may substantially reduce the number of variables used to describe the data set with very little loss of information.

Our starting point is a calibrating set of 127 Cepheids, with periods $P > 10$ d and high-quality well-sampled *V*- and *I*-band light curves. We used this 'training set' to establish relationships between multicolour LCS and period. The training set consists of the following.

(i) 61 Galactic Cepheids with photometry from Berdnikov (unpublished data base), Berdnikov & Turner (1995) and Moffett & Barnes (1984).

(ii) 66 LMC Cepheids, covering a wider period range with photometry primarily from the OGLE catalogue of Classical Cepheids (Udalski et al. 1999a; web archive at http://bulge.princeton.edu/~ogle/ogle2/cep_lmc.html; Fourier analysis from Ngeow et al. 2003), but supplemented by data taken from various sources, particularly Moffett et al. (1998).

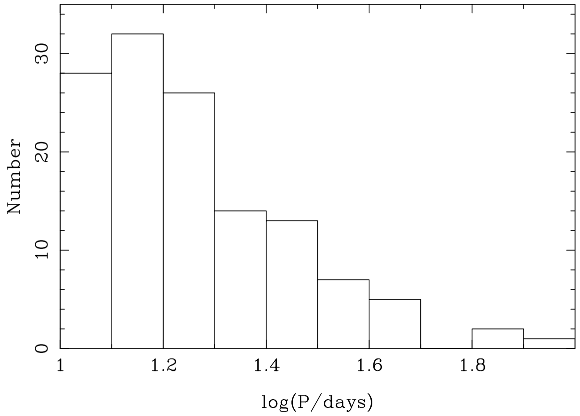

Figure 1. Distribution of log period for our 'training set' of 127 Cepheids.

Fig. 1 shows the distribution of periods in our sample. Note that we do not include any SMC Cepheids in our training set; this reflects the fact that the metallicity of target galaxies for e.g. *HST* Cepheid distance estimation is generally significantly higher than that of the SMC. We do consider SMC Cepheids in Section 6, however, in our discussion of LCS characterized by PCA as a possible diagnostic of metallicity.

To apply PCA to this sample, we first Fourier analyse the photometric data (for those variables not already analysed by Ngeow et al.), up to eighth order, i.e. perform a least-squares fit of the following:

$$m(t) = m_0 + \sum_{k=1}^{k=8} a_k \sin(2\pi kt/P) + b_k \cos(2\pi kt/P). \qquad (1)$$

The coefficients of the Fourier terms constitute a vector consisting of 32 elements for each member of the training set (i.e. eight sine amplitudes and eight cosine amplitudes for both the V and I bands – but note that the phase is shifted such that the first cosine term in V is always zero). Note also that the mean V and I magnitudes, the a_0 terms, of the calibrating Cepheids are not included in the PCA because they are distance dependent).

The mean V and I LCS is established simply by averaging these vectors. PCA is then applied to the whole set of residual vectors (i.e. with the average vector subtracted) in order to determine the most significant variations from the mean LCS. Full numerical details of the analysis are given in Appendix A. Incorporating both V and I data in each vector means that the correlations between the coefficients in each band are automatically encoded in the resulting analysis. This could, of course, be extended to more bands, but we restrict ourselves to V and I here because only those filters have been used in the large majority of *HST* studies.

In practice, the first principal component largely reflects simple variations in amplitude. Subsequent components encode more subtle LCS information, such as 'bumps'. Of course, sets of Fourier amplitudes are not the only vectors which could be used as input to the PCA. One could, for example, work directly with the observed V and I magnitudes for each calibrator, smoothed and interpolated on to a regular grid of phase values. We find, however, that the use of Fourier components as input vectors works very well, naturally incorporating a degree of smoothing of the input data and providing a link with previous approaches to LCS analysis.

Table 1 shows for our calibrating set the proportion of the variance explained by the first few principal components. We can see from this table that one requires only a few components to explain a large proportion: for example, the first three principal components account for 89 per cent of the variance of LCS within the sample. Moreover, because the observed scatter in the data includes the effects of photometric errors and finite sampling on the estimated Fourier coefficients which are input to the PCA, the proportion of the *intrinsic* variation in LCS explained by the first three principal components will, in fact, be even higher than 89 per cent.

Table 1. Proportion of the total variance in the calibrating set explained by the first few principal components. (In PCA the variance associated with each component is equal to the corresponding eigenvalue of the covariance matrix).

Component	Normalized variance	Cumulative variance
1	0.627	0.627
2	0.199	0.827
3	0.064	0.890
4	0.026	0.917

Fig. 2 shows an example of two Cepheids from the OGLE data set with very good phase coverage. This illustrates that excellent light curves are reconstructed from just two PCA terms. As these reconstructed curves incorporate information from the whole training set, they necessarily reflect average Cepheid behaviour, and do not fit perfectly any individual Cepheid. However, this has the advantage that they do not follow noise in the data either, as the Fourier fits in the V band are beginning to do in these examples.

3 DERIVING TEMPLATE LIGHT-CURVES

With the PCA coefficients for each variable in hand, we can plot them as a function of period, as shown for the first four principal components in Fig. 3. This figure reveals some important trends – notably that the behaviour of the first principal component, as expected, is similar to a simple plot of amplitude versus period, with a peak at around $P = 30$ d (see e.g. Schaltenbrand & Tammann 1972). There is clearly also systematic structure in the plots for the coefficients of the second and third principal components. By the fourth component the distribution of coefficients is becoming increasingly dominated by noise, although small but statistically significant correlations of the coefficients with log (P) are seen up to at least eight PCA terms.

We have fitted low-order polynomials through these scatter plots, to define 'typical' values of the PCA coefficients for a Cepheid of given period, and also obtain some estimate of the spread around these typical values. The polynomial fits are shown by the solid curves in each of the panels of Fig. 3, terminating at $\log(P/\text{days}) = 1.8$ where the number of training Cepheids becomes very few.

Before discussing the construction of template light-curves, it is interesting to compare the distribution of coefficients for the Milky Way (MW) and LMC subsamples. For the first and third principal components we can see from Fig. 3 that the distributions appear to be well mixed, with no obvious distinction between the two samples. This is broadly consistent with the results of Kanbur & Ngeow (2004), who obtained period–colour and amplitude–colour relations for a similar sample of MW and LMC OGLE Cepheids. While those authors found strong evidence for a difference in the slope of these relations between long- (i.e. $P > 10$ d) and short-period Cepheids, they did not find a statistically significant difference in the long-period sample slopes between the LMC and MW. For the second principal component, on the other hand, at a given period the LMC coefficients typically appear to be less than those for the MW. The distributions are clearly not disjoint, so that the adoption

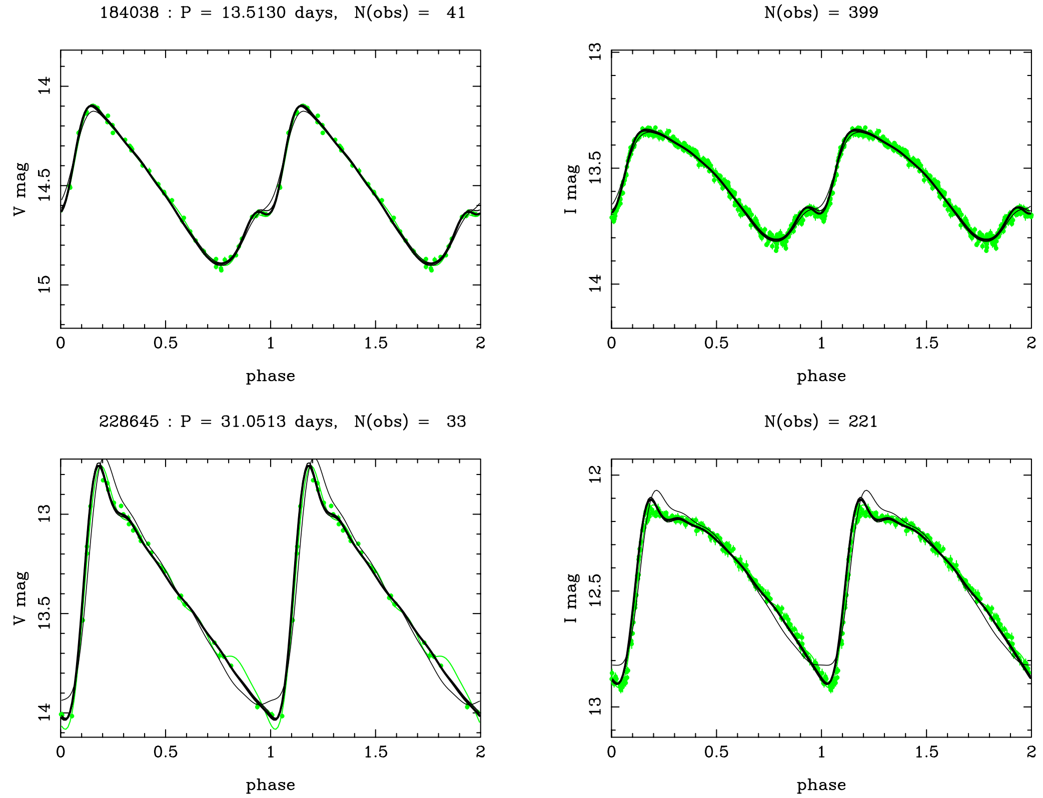

Figure 2. Light-curves for two LMC Cepheids of different periods (OGLE data). The Fourier fits are shown as a grey line (i.e. 32 terms describing both V and I), and the reconstructed PCA curves with one, two, three and four terms are shown in successively thicker, black lines. We emphasize that V and I are simultaneously described by the PCA curves because the analysis is performed on the combined data set. Note that the fits, whilst not perfect, are very good, and that in fact beyond two terms further changes in the PCA curves are almost entirely within the thickness of the line.

of a single polynomial fit to describe the structure in the combined training set is reasonable. Nevertheless, Fig. 3 suggests that the second principal component, at least, might be a useful discriminator between different Cepheid samples. We discuss this point further in Section 6 below.

In order to generate a realistic Cepheid light-curve template all that is now required is to read off the PCA coefficients corresponding to the desired period according to the polynomial fits in Fig. 3, and hence, with knowledge of the PCA vectors and the average light curves, reconstruct a full sequence of Fourier terms. We emphasize again that because V and I Fourier coefficients are both included in each vector, the light curves in each band are reconstructed simultaneously. An extra degree of sophistication can be achieved by considering the PCA coefficients within a range around the polynomial fit corresponding to the scatter in the training set. In this case we find not a single template at a given period, but a whole family of allowable light curves.

Recalling that the primary motivation of the present study is to extract optimal light-curve parameters from observed noisy data sets, the next challenge is to find the best-fitting template light-curves to such a data set when the period and other parameters are a priori unknown. Our solution to this problem is described in more detail in the next section.

4 LIGHT-CURVE PARAMETER ESTIMATION VIA TEMPLATE FITTING

To characterize the Cepheid light-curves of sparsely sampled, noisy data our approach is to find the best-fitting PCA templates, simultaneously in V and I, determined from a search of the plausible parameter space in period, V and I mean magnitude, phase and PCA coefficients. Determining period, phase etc. using both bands is unusual, but makes most efficient use of all the data.

In practice – mainly for computational expediency – only the first two PCA coefficients are allowed to vary. The higher PCA coefficients are set to zero, but as seen already Fig. 2), using more than two PCA terms only modifies the light curves at a subtle level.

The fitting procedure for an individual Cepheid is summarized as follows.

(i) Loop over a large range of trial periods (usually between 10 and 65 d with steps of 0.001 in the log).

(ii) For each trial period, loop over a range of PC1 and PC2 coefficients around their typical values for that period, thus generating template light-curves in V and I for each pair of values of the PC1 and PC2 coefficients.

(iii) For each pair of templates, find the values of phase and intensity-mean magnitudes in V and I which minimize the χ^2 statistic, where χ^2 is defined as the sum of the squared deviations (normalized by the photometric errors) between the observed V and I data points and the magnitudes predicted by the templates. This procedure utilizes the Amoeba algorithm (e.g. Press et al. 1992).

(iv) Move to next pair of PC1 and PC2 coefficients.

(v) Move to next trial period.

At the end of this process, the trial period and V and I light curves with the overall lowest χ^2 is then assumed to provide the best estimates of all the parameters.

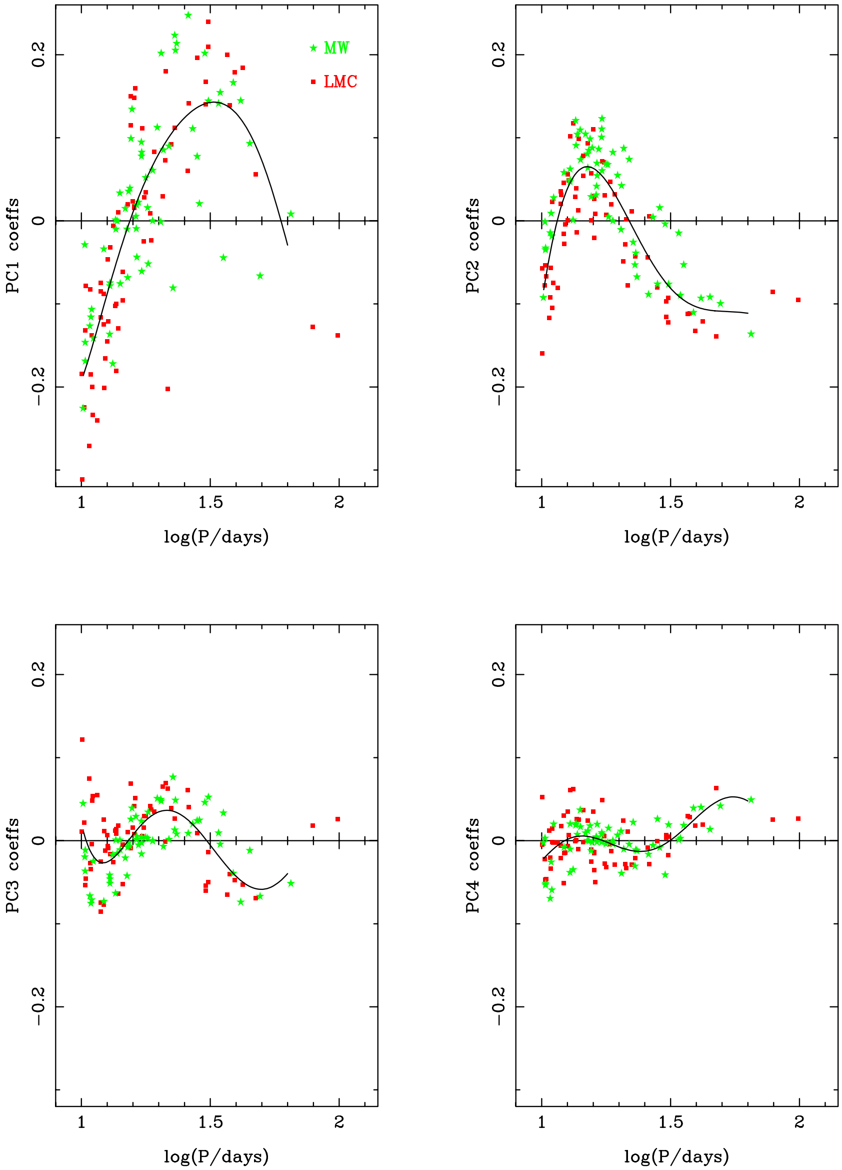

Figure 3. PCA coefficients plotted as a function of log period in days, for our training set of 127 MW (light stars) and LMC Cepheids (darker squares). The four panels illustrate the diminishing strength of successive principal components, and by plotting against log (*P*) also reveal the systematic trends in light-curve amplitude and shape with period. Low-order polynomial fits are overplotted which will be used to obtain typical coefficients at a given period. The small but real difference between the distributions of LMC and MW points is discussed further in the text.

In practice we plot exp $(-\chi^2{}_{\rm red}/2)$, as an indicator of relative likelihood, against test period. This plot reveals: first, the best period (the peak position); secondly, the goodness of the best fit (the height of the peak); thirdly, how well the period is determined (the width of the peak); and, fourthly, whether there are any potential aliases at completely different periods (essentially the height of the second highest peak). As we show later, this information can be very useful in deciding objectively which variables to include and which to exclude in a period–luminosity (PL) analysis.

4.1 Estimating the distance modulus and its error

To estimate the distance modulus of the Cepheid we now ignore the data points and calculate intensity-mean magnitudes, $\langle V \rangle$, $\langle I \rangle$, of the fitted templates themselves. These are compared to the absolute magnitudes calculated for the given period using the calibrating PL relations. The colour excess is used to correct for extinction following the procedure detailed by Tanvir (1997), and hence an estimate of the true, unreddened distance modulus, μ_0, is made.

We can also obtain in a straightforward manner an estimate of the uncertainty on μ_0 by computing the posterior distribution, $p(\mu_0 \,|\, \mathrm{data})$, of μ_0 given the observed *V*- and *I*-band data. Note that this procedure accounts for internal errors due to photometric noise and finite sampling, but not to errors in the original templates (likely to be relatively small in practice) or calibration errors (which must be estimated independently).

We proceed as follows. Let ϕ, P, PC$_1$, PC$_2$ denote respectively the phase constant, period and the first and second PCA components of the underlying light curves. Ignoring the higher order PCA coefficients, and also neglecting for the moment the impact on the LCS of other stellar parameters such as metallicity (see below), we assume that these four parameters, together with the intensity-mean magnitudes, $\langle V\rangle$ and $\langle I\rangle$, completely specify the V- and I-band light curves. Note, moreover, that under this approximation P is uniquely defined by the values of $\langle V\rangle$, $\langle I\rangle$ and μ_0, so that we need not consider the period as an independent parameter.[1] To simplify notation, we denote the remaining light-curve parameters collectively by the column vector $\mathbf{\Lambda}$; i.e.

$$\mathbf{\Lambda} \equiv (\langle V\rangle, \langle I\rangle, \phi, \mathrm{PC1}, \mathrm{PC2})^T . \tag{2}$$

Formally, we may then write

$$p(\mu_0 \mid \mathrm{data}) = \int p(\mu_0, \mathbf{\Lambda} \mid \mathrm{data})\,\mathrm{d}\mathbf{\Lambda} \tag{3}$$

i.e. we marginalize $p(\mu_0, \mathbf{\Lambda} \mid \mathrm{data})$ over the other independent parameters.

To simplify matters we assume that PC_1 and PC_2 are equal to the values determined previously for the globally best-fitting template. In practice, we also assume that $p(\mu_0, \mathbf{\Lambda} \mid \mathrm{data}) = 0$ unless ϕ is equal to its best-fitting estimate for the period determined by the particular values of μ_0, $\langle V\rangle$ and $\langle I\rangle$ This allows equation (3) to be rewritten as

$$p(\mu_0 \mid \mathrm{data}) = \int p(\mu_0, \mathbf{\Lambda} \mid \mathrm{data})\,\mathrm{d}\langle V\rangle\,\mathrm{d}\langle I\rangle, \tag{4}$$

which, in turn, can be approximated by a sum over a series of 'trial' values of $\langle V\rangle$ and $\langle I\rangle$. From Bayes' theorem we may write

$$p(\mu_0, \mathbf{\Lambda} \mid \mathrm{data}) = p(\mathrm{data} \mid \mu_0, \mathbf{\Lambda})p(\mu_0, \mathbf{\Lambda}), \tag{5}$$

where the first term is the likelihood function, expressing the probability of obtaining the observed photometric data, given a set of light-curve parameters and a Cepheid at distance modulus μ_0, and the second term is a prior distribution for those parameters and for the distance modulus. Assuming a flat prior for $p(\mu_0, \mathbf{\Lambda})$, equation (4) may be further reduced to

$$p(\mu_0 \mid \mathrm{data}) \propto \sum_j \sum_k p(\mathrm{data} \mid \mu_0, \langle V\rangle_j, \langle I\rangle_k, \phi, \mathrm{PC}_1, \mathrm{PC}_2) \tag{6}$$

where $\langle V\rangle_j$ and $\langle I\rangle_k$ denote a series of (equally spaced) trial values of the mean V- and I-band magnitudes. One can, if appropriate, easily generalize equation (6) to the case of a non-uniform prior and a non-uniform grid of $\langle V\rangle_j$ and $\langle I\rangle_k$ values.

Assuming that the photometric errors are normally distributed, finally we obtain

$$p(\mu_0 \mid \mathrm{data}) \propto \sum_j \sum_k \exp\left(-\chi^2_{jk}\right), \tag{7}$$

where χ^2_{jk} is the chi-squared obtained from comparing the observed and predicted magnitudes, given values of μ, $\langle V\rangle_j$, $\langle I\rangle_k$, ϕ, PC_1 and PC_2. In fact, we could take as our estimate of μ_0 the value which maximizes the posterior likelihood $p(\mu_0 \mid \mathrm{data})$ – or equivalently the value which minimizes χ^2_{jk} – however, in practice these values differ from those for the individual best-fitting template by only one or two hundredths of a magnitude in distance modulus. Instead we use the fact that $p(\mu_0 \mid \mathrm{data})$ should be properly normalized, to compute σ_μ, the uncertainty in the estimated distance modulus from the width of the resulting likelihood function. Notice that this analysis gives reasonable estimates of the uncertainty providing the estimated period itself is not greatly in error. If, in fact, the best-fitting period is an alias then the true error on the distance modulus is likely to be significantly larger.

The above analysis can be readily extended to include a metallicity dependence in the shape of the light-curve templates and in the PL relations. In this case the period, P, would be determined by the values of μ, $\langle V\rangle$, $\langle I\rangle$ and metallicity, Z, which itself might be estimated along with the other independent parameters by the template-fitting approach. Such an extension in practice seemed inappropriate for the training set considered in this paper, because it contained Cepheids from different metallicity environments (but see Section 6, below). The study of metallicity effects using PCA is straightforward in principle, however, (see Kanbur et al. 2002). Moreover, one could extend the model for the prior distribution, $p(\mu_0, \mathbf{\Lambda})$, of light-curve parameters to include a dependence on other fundamental stellar parameters, such as effective temperature and mass, reflecting one's state of knowledge about, for example, the width of the instability strip, initial mass function and mass luminosity relation for Cepheids (see, for example, Kochanek (1997) for an example of such a prior model).

5 SIMULATIONS OF SPARSE AND NOISY LIGHT CURVES

Another use of our template light-curves is to provide the underlying models for production of artificial Cepheid photometry. In this section we describe simulations designed to resemble the sparse and noisy photometry from typical *HST* Cepheid monitoring campaigns. We then run the PCA template-fitting program on these data sets to obtain maximum likelihood estimates of the parameters for each simulated Cepheid, and hence establish how accurately the input parameters are recovered. We also compare the template-fitting results to those obtained from more traditional parameter estimation methods. Although the simulations are realistic, they are not designed to replicate specific cases of *HST*-observed Cepheids in external galaxies but rather to represent generic examples similar to those found by most *HST* studies. Moreover, we have not compared our template-fitting method with all algorithms which have been used to determine Cepheid light-curve parameters – partly because many such studies have involved some degree of subjectivity, for example in selecting the Cepheids themselves, which is hard to replicate.

We chose to simulate data for Cepheids of $\log(P/\mathrm{days}) = 1.4$ (about 25 d), being typical of the variables observed in the *HST* programs, and in the middle of the range for which the sampling strategy is optimized. Phases were random, and PCA coefficients chosen to be as those for real Cepheids (i.e. based on the polynomial fit to the training set) at the period in question. The sampling of the light curves was carried out using a particular sequence of observations based on that adopted by the *HST* H_0 Key-Project group Specifically this meant 12 epochs of observation in V and four epochs in I. Realistic, magnitude-dependent photometric noise was added to the data points, with one set of 400 simulations with error bars on each point around 0.1–0.2 mag, being representative of 'typical' *HST* data, and another set of 1000 simulations representing 'difficult', low signal-to-noise (S/N) ratio data with error bars more like 0.15–0.3 mag per data point. The latter set approximates the worst-case data which have been obtained in some *HST* studies.

5.1 Results of simulations

To provide a benchmark we first analysed each simulated data set with the methods of period finding via string-length minimization (Lafler & Kinman 1965) and phase-weighted intensity-mean magnitude estimation (Saha & Hoessel 1990). We then analysed the

[1] In other words we assume that the Cepheid lies exactly on the fiducial V and I linear PL relations.

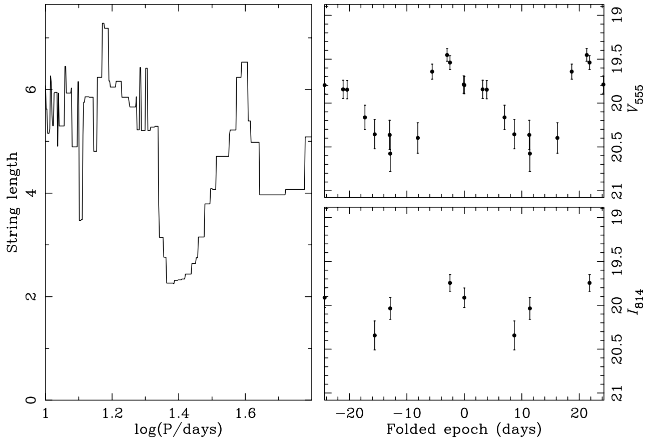

Figure 4. Example of artificial Cepheid photometry simulated to have S/N ratios typical of *HST* data and analysed with the OLD algorithms. The left-hand panel shows the Lafler & Kinman (1965) string-length measure applied just to the *V*-band data points and plotted as a function of trial period – the best period being indicated by the minimum string length. As the input period is $\log(P/\text{days}) = 1.4$, the method obviously works well for this simulation. The worst aliases tend to occur at half the true period (i.e. offset by about 0.3 in the log), but in this case (and in fact for nearly all the simulations with this S/N ratio) the worst alias is not as good a fit as is the correct period. The right-hand panels show the data points folded on this best period. Note that the zero points for the magnitude scales are chosen arbitrarily.

same synthetic data using our PCA template approach described above. For the sake of brevity we henceforth refer to these as the 'OLD' and 'NEW' algorithms respectively, whilst clearly recognizing these particular OLD methods are by no means the only ones used in previous studies. They do, however, have the benefit of being easily and mechanically applied to the data.

We consider first the simulations of 'typical', moderate S/N ratio data. Examples of the simulations and period determination are given in Fig. 4 for the OLD algorithms, and Fig. 5 for the PCA template-fitting algorithm. Both approaches work well in this example, in the sense of correctly identifying the period. The fact that subtle LCS information is largely erased by this degree of sparse sampling and noise addition means that the use of the same templates for both creating and fitting to the simulated data should not significantly bias the results. We checked this expectation by running further noisy simulations but this time starting with the observed light curves of individual Cepheids from the training set which were outliers from the curves in Fig. 3. The results and level of improvement with template fitting were qualitatively similar to those reported above, with the proviso that both algorithms (especially the OLD ones) do somewhat better when the overall Cepheid amplitude is larger rather than smaller. In other words, as one would expect, the outliers with PC1 above the average are easier to find periods for than the outliers with PC1 below the average.

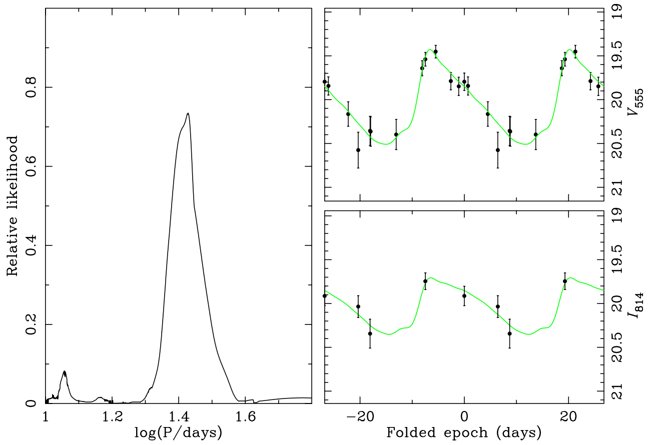

Figure 5. The same simulated data shown in Fig. 4 analysed with the NEW, template-fitting algorithm. In this case we look for a peak in the relative likelihood curve (left-hand panel) to identify the best fitting period (see description in text). The period found by the algorithm is very close to that input in the simulation, $\log(P/\text{days}) = 1.4$, and again a small alias is seen to appear near $P/2$. The right-hand panels again show the data folded on this best period.

In Fig. 6 we summarize the results of all 400 simulations in histograms showing the returned periods and inferred distance moduli. The extinction-corrected distance moduli are calculated using P, $\langle V \rangle$ and $\langle I \rangle$ as described by Tanvir (1997), which is essentially the same method as used by the *HST* H_0 Key-Project group. Neither method is confused by aliases with this S/N ratio data, although the template fitting produces rather more accurate periods and distance moduli. Specifically the 1σ rms scatters for the NEW and OLD methods are 0.16 and 0.23 mag respectively, suggesting that the uncertainties in distance modulus resulting from light-curve parameter estimation for samples of several tens of Cepheids should only be a few hundredths of a magnitude.

For each fit to the simulated data sets we also evaluated the uncertainty in distance modulus as described in Section 4.1. The average turned out to be 0.15 mag, very close to the observed scatter, giving confidence that the errors returned by the template-fitting procedure itself are realistic.

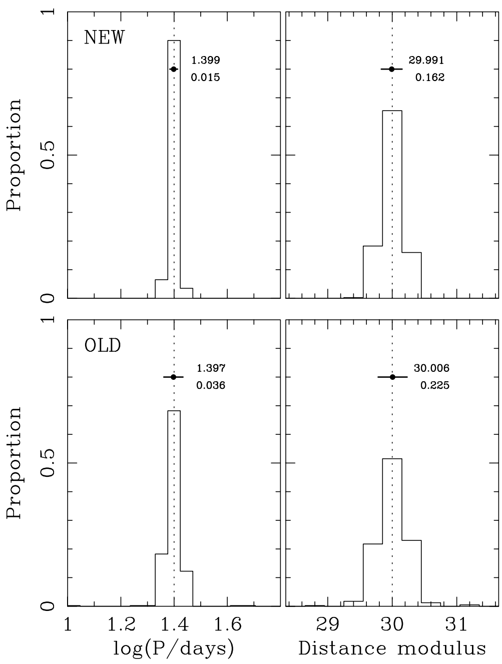

Figure 6. Histograms of the results from 400 simulations with 'typical' *HST* S/N ratio. The input values are indicated by vertical dotted lines, namely $\log(P/\text{days}) = 1.4$ and $\mu_0 = 30$. The latter is chosen arbitrarily as being typical of *HST* studied galaxies. The upper panels are for template fitting and the lower panels using the OLD algorithms. Both period and distance modulus are well determined with this S/N ratio, as indicated by the solid dot and bar which represent the mean and standard deviation of each distribution (numerical values are printed next to the bar). The number of occasions where an alias period is wrongly identified as the true period is negligible.

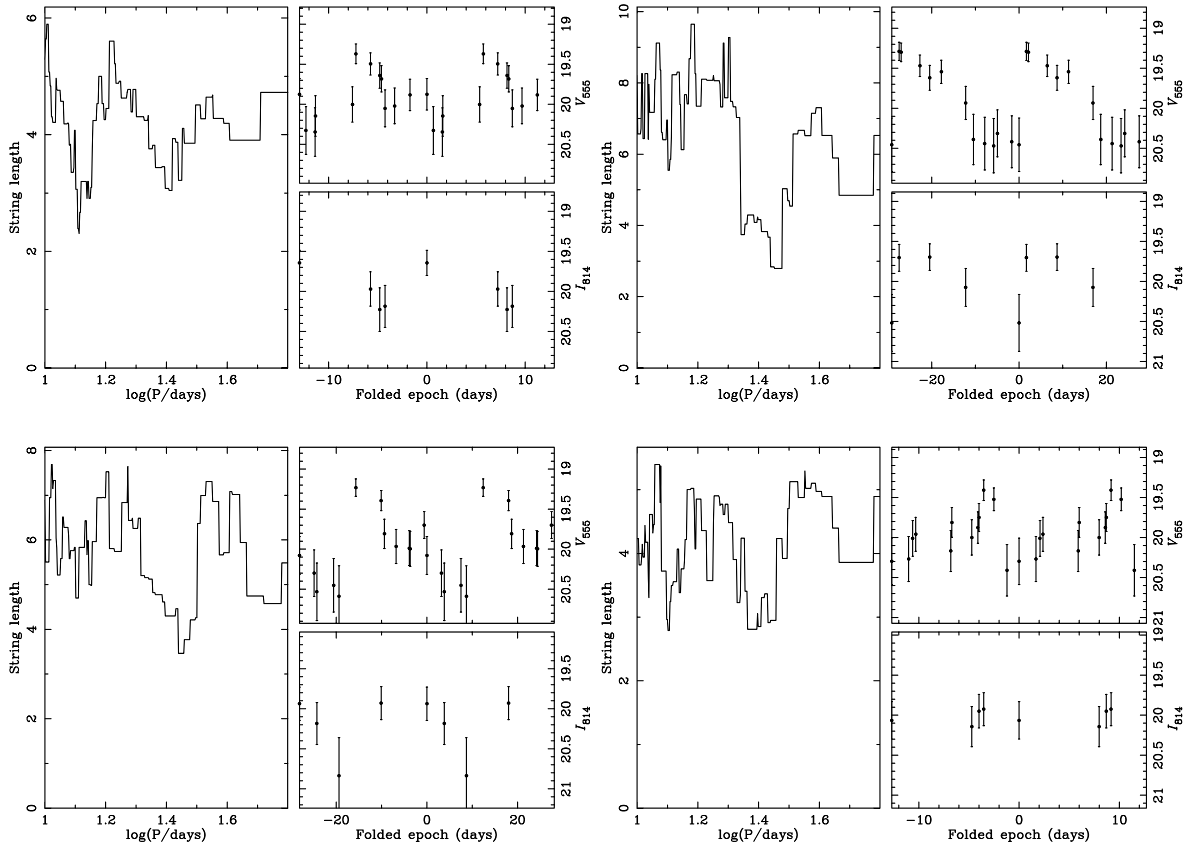

Figure 7. Examples of 'difficult', low S/N ratio simulated data analysed with the OLD algorithms. The panels are similar to those in Fig. 4. The four cases were chosen to illustrate a range in behaviour, with the top left and bottom right variables suffering from bad aliases at half the true period. Note, in the latter case the folded data points do not trace a very Cepheid-like light curve.

The situation with the 'difficult', low S/N ratio data is illustrated in Fig. 7 for the OLD algorithms and Fig. 8 for the NEW template-fitting ones. In this case we show four examples to highlight the fact that now results range from cases where the input parameters are well recovered, to instances where the best fit is actually obtained with an alias period.

This behaviour is summarized in Fig. 9 for 1000 simulated data sets which shows that now both techniques produce the occasional period aliases. In general terms the periods and distance moduli show more scatter than was the case with higher S/N ratio, although overall an accuracy of about 0.4 mag per Cepheid in distance modulus is still reasonably good. Once again the template-fitting performs a little better than the OLD algorithms, but both reveal a slight bias to lower distance moduli, over and above the increased scatter.

However, we must be cautious in interpreting these results for a number of reasons. In practice, because of the low S/N ratio, some of our simulated variables would probably not have been classified as variables in the first place had they appeared in an *HST* study. Furthermore, bad fits would often be rejected as not being sufficiently 'Cepheid-like'. It would not be surprising if such cases of bad template fits also produced the most discrepant distance moduli.

In order to assess these effects, and also make the test more realistic, we clipped the sample of simulations to exclude those for which the degree of scatter of the data points about a constant, non-variable line was such that it would only occur by chance 1 time in 5000. In other words we insisted (as do most Cepheid studies one way or another), that the threshold for treating a star as a variable is high enough that very few non-variable stars ever exceed it by chance. Further we set an upper limit to the acceptable reduced $\chi^2_{\rm red}$ for the template fit of 1.3, which means that 22 per cent of true Cepheids will be lost, but ensures that only those which fold to produce genuine 'Cepheid-like' light curves are retained. Finally, we rejected any variables for which there was an alias period with a $\chi^2_{\rm red} < 1.5$ which was separated by less than 0.1 in log(P) from the highest likelihood peak. This procedure loses a few fits which produced accurate periods, but is particularly good at removing probable aliases.

The results of this whole clipping procedure are shown in Fig. 10. As expected, the scatter in the results of both methods is reduced, but in particular problems with aliases for the template fitting are largely removed, as is the bias in distance modulus determination.

Viewed as a whole the simulations permit the following conclusions.

(i) Template fitting results in a roughly 30 per cent reduction in scatter in estimates of distance modulus for the 'typical' S/N ratio data compared to the OLD methods of string-length minimization to determine periods and phase-weighted averaging to obtain intensity-mean magnitudes.

(ii) There is a tendency to slightly underestimate periods for 'difficult', low S/N ratio data. Again template fitting does somewhat better than string length. This small bias in period also leads to a small bias in distance modulus. In fact, we have also performed simulations for longer period log(P) = 1.7 Cepheids (i.e. around 50 d), although not reported here in detail, and find this underestimation

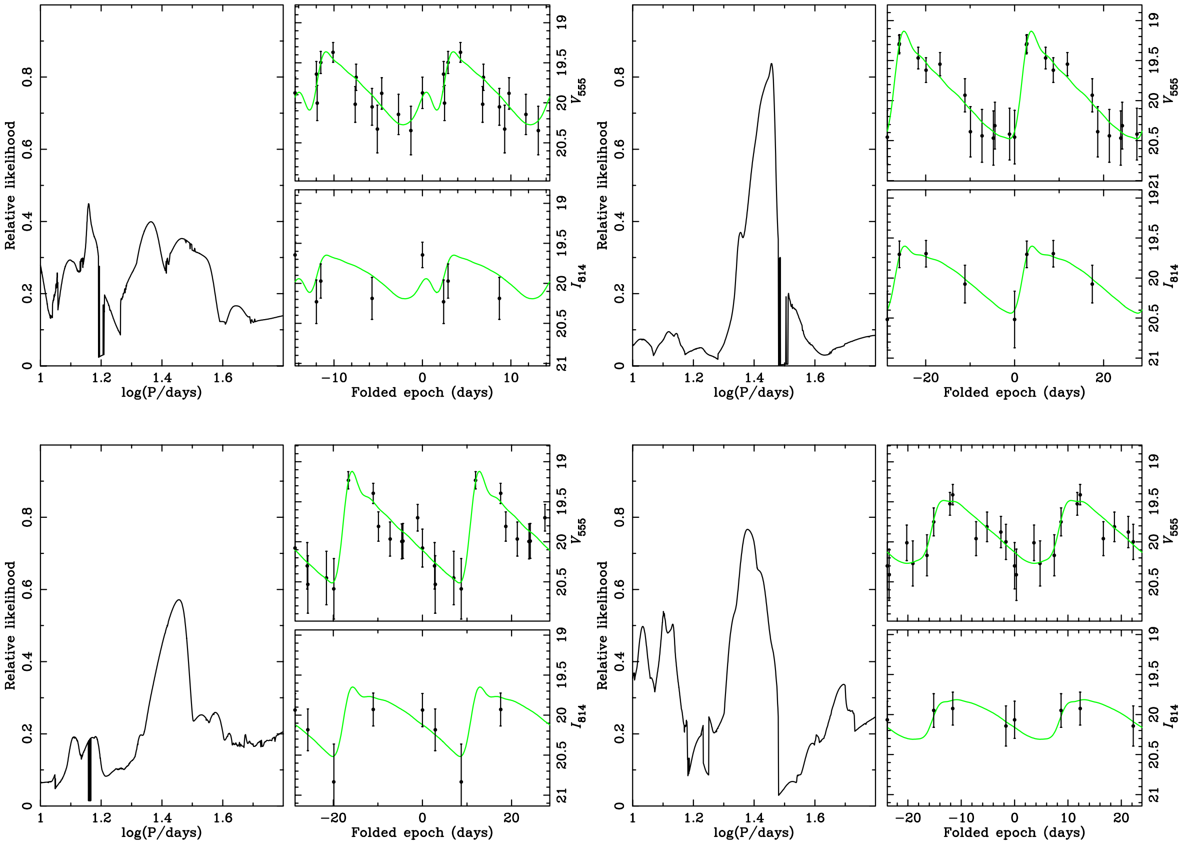

Figure 8. The same simulated data as Fig. 7 analysed with the NEW, template-fitting algorithms. Again the top left case hits the same problem with an alias period providing a better fit that the true period. However, in the bottom-right case this time, the fact that we are fitting a template rather than simply minimizing string-length has correctly identified the period.

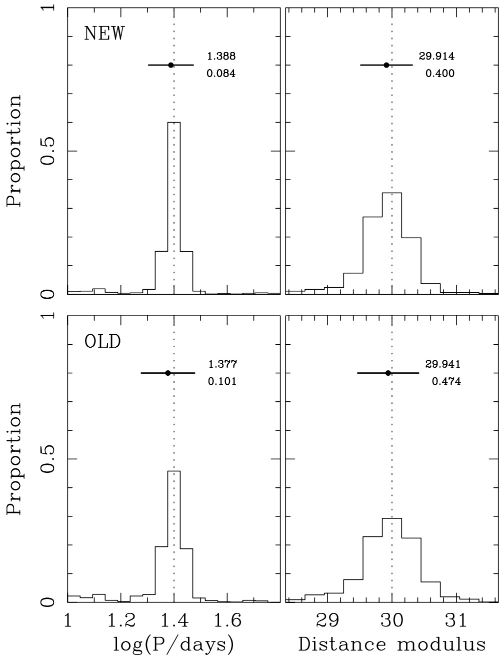

Figure 9. Histograms of the results from 1000 simulations of 'difficult', low S/N data. Compared to Fig. 6 we see that the distribution of returned periods and inferred distance moduli shows more scatter, a small but non-negligible contamination by aliased periods and a small but statistically significant bias toward lower values of μ_0.

becomes a little worse. This is not surprising because the period is now approaching the total length of the observing sequence, and is only a little below the maximum period employed in the trials.

(iii) Both approaches perform respectably even for low S/N ratio data, but with a increasing incidence of period aliases. However, not only does the template fitting do somewhat better, it also computes a goodness-of-fit (χ^2) for the template fit, providing an objective way of selecting the variables for inclusion in the analysis. A sample culled on the basis of only including good template fits reduces the scatter by a factor of around 2 compared to the traditional OLD methods.

(iv) The distance modulus uncertainties calculated as described in Section 4.1 are on average very good. For example, for the 25-day period simulations the average uncertainty calculated for the typical S/N case was 0.15 mag compared to the actual dispersion around the true (input) value of 0.16 mag. For the low S/N ratio data the numbers are 0.28 mag for the estimated errors compared to 0.31 mag as the dispersion for the clipped sample of simulations.

While these are interesting results, and establish the utility of the PCA template-fitting method, we caution that they do not imply significant problems with the results obtained by the various groups reporting *HST* Cepheid observations in the past. For one thing, for the typical S/N ratio photometry, the NEW algorithm only performed a little better than the OLD ones. Furthermore, most recent studies have not simply adopted the OLD methods considered above, but have also either applied 'chi-by-eye' rejection of doubtful variables, and/or performed other variants on the template-fitting scheme. The

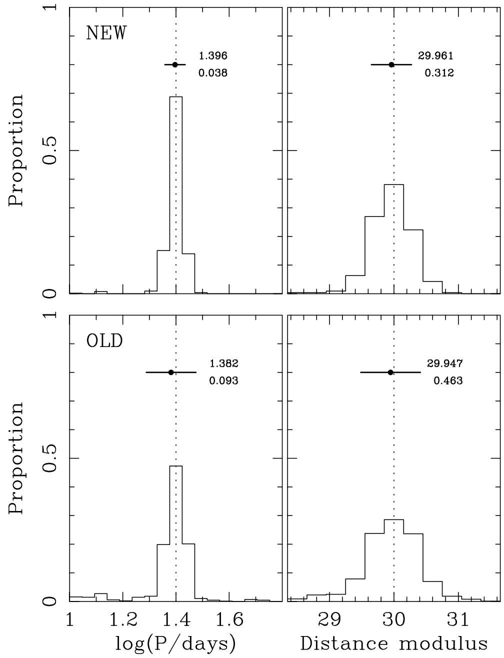

Figure 10. The same information as in Fig. 9 but this time clipped of the low amplitude variables and those with badly fitting light curves. This process mimics what is often done in practice, selecting candidates above some threshold criterion for variability, and rejecting those which do not appear 'Cepheid-like'. In addition, as described in the text, we have removed variables with potential aliases, which is a fairly well-defined, automated procedure (i.e. based on the height of the second highest peak in the relative likelihood plot for the variable) when doing template fitting. These steps obviously reduce scatter, remove many of the aliases, particularly for the template-fitting, and largely remove the small bias. The price which is paid is the loss of 12.5 per cent of the variables used for the OLD algorithm analysis, and about 46 per cent of the variables used for the template fitting.

results of our simulations reinforce the conclusion that such template fitting has many benefits, but we have also shown that our procedure has a more rigorous statistical basis and provides a more efficient means of encoding relevant LCS information than previous methods.

5.2 Estimating maximum light from template fits

Various authors have suggested that Cepheids at maximum light may be as good, if not better, standard candles than Cepheids at intensity mean-light (e.g. Sandage & Tammann 1968; Kanbur & Hendry 1996; Kanbur et al. 2003). Aside from arguments based on intrinsic physical properties, another advantage could be that maximum light may be more precisely determined than mean light if the Cepheid is faint and hence poorly observed through minimum.

However, maximum-light has rarely been used in practice, perhaps partly because many epochs are required to give a decent chance of sampling close to the maximum. One also loses some of the benefit of averaging many observations to reduce noise. Obtaining estimates of maximum-light from template fits may be a way of benefiting from the advantages while not suffering the disadvantages. All the data is used, with appropriate weighting, and reasonable estimates of maximum-light can be obtained without requiring dense sampling over the maximum itself.

We have also tested our ability to estimate maximum-light using the template fits to the simulated data. The resulting histograms are actually so similar to the ones presented for mean-light that we feel they are not worth showing separately (mean-light is very marginally better). Admittedly here the fact that the same templates are used to create the artificial noisy data in the first place and then to fit to it, may make the simulation results appear slightly rosier than reality. But the point is clear: although we find no evidence that maximum-light is superior to mean-light, it is certainly reasonable to use maximum-light with template fitting. Interestingly we also find a strong correlation between the maximum-light and mean-light distance moduli for individual simulated Cepheids, indicating, perhaps unsurprisingly, that they encode very similar information and can not therefore be combined in any way to provide an improved distance indicator.

6 LIGHT-CURVE SHAPE AND METALLICITY

An obvious question which arises from our analysis thus far is whether LCS is sensitive to physical parameters other than just period. It would be particularly useful, for example, if LCS were found to be sensitive to a property – such as metallicity – which is expected to correlate with the absolute magnitude of a Cepheid (see e.g. Caputo et al. 2000, and references therein). Quantifying the (reddening corrected) sensitivity of Cepheid distances to metallicity has proven hard, but estimates have tended to be in the region of 0.2 mag in distance modulus ($\sim$10 per cent in distance) for a factor 10 in metallicity (e.g. Sakai et al. 2004). In most *HST* studies, the Cepheid metallicity is estimated from spectroscopy of the gas phase – sometimes from nebulae in other parts of the galaxy from the Cepheids. Directly constraining the metallicity of the Cepheid sample itself via observations of LCS would have obvious advantages over this approach.

In fact, Paczynski & Pindor (2000) already pointed out that OGLE observed Cepheids in the SMC have systematically lower amplitudes than those in the LMC in the period range $1.1 < \log(P) < 1.4$, which they suggest is likely to be a metallicity effect. Similarly, Kanbur et al. (2002) presented evidence suggesting a difference in the average LCS of SMC and LMC first-overtone Cepheids, based on a principal component analysis of densely sampled *V* and *I* Cepheid light-curves from the OGLE (Udalski et al. 1999a,b) and EROS (Beaulieu et al. 1995) microlensing surveys. Fig. 12 of Kanbur et al. shows, as a function of period, the coefficients of the first and second principal components for OGLE first-overtone Cepheids with periods less than 10 d. A comparison from that figure of the PCA coefficient distribution for LMC and SMC Cepheids shows some clear differences – most notably that the distribution of first principal component coefficients for the SMC Cepheids generally lies below that for the LMC Cepheids. Hence the PCA approach picks out changes in the structure of overtone light curves which correlate with metallicity.

Fig. 11 shows the distribution, as a function of log period, of the first four principal component coefficients of the same MW and LMC Cepheids as Fig. 3 together with a further 52 Cepheids from the SMC. The data for the latter are largely from the OGLE data base (Udalski et al. 1999b) and Moffett et al. (1998). Note that the SMC light curves have been decomposed on to the PCA basis established from the MW/LMC data rather than being included in an expanded training set.

Addition of the SMC data clearly increases the spread at a given period, but most of this appears to be systematic rather than increased random scatter. In particular, the PC1 coefficients are generally lower than those of the MW and LMC, confirming the reduced amplitude noted by Paczynski & Pindor (2000). Particularly prominent are a group of five SMC Cepheids with

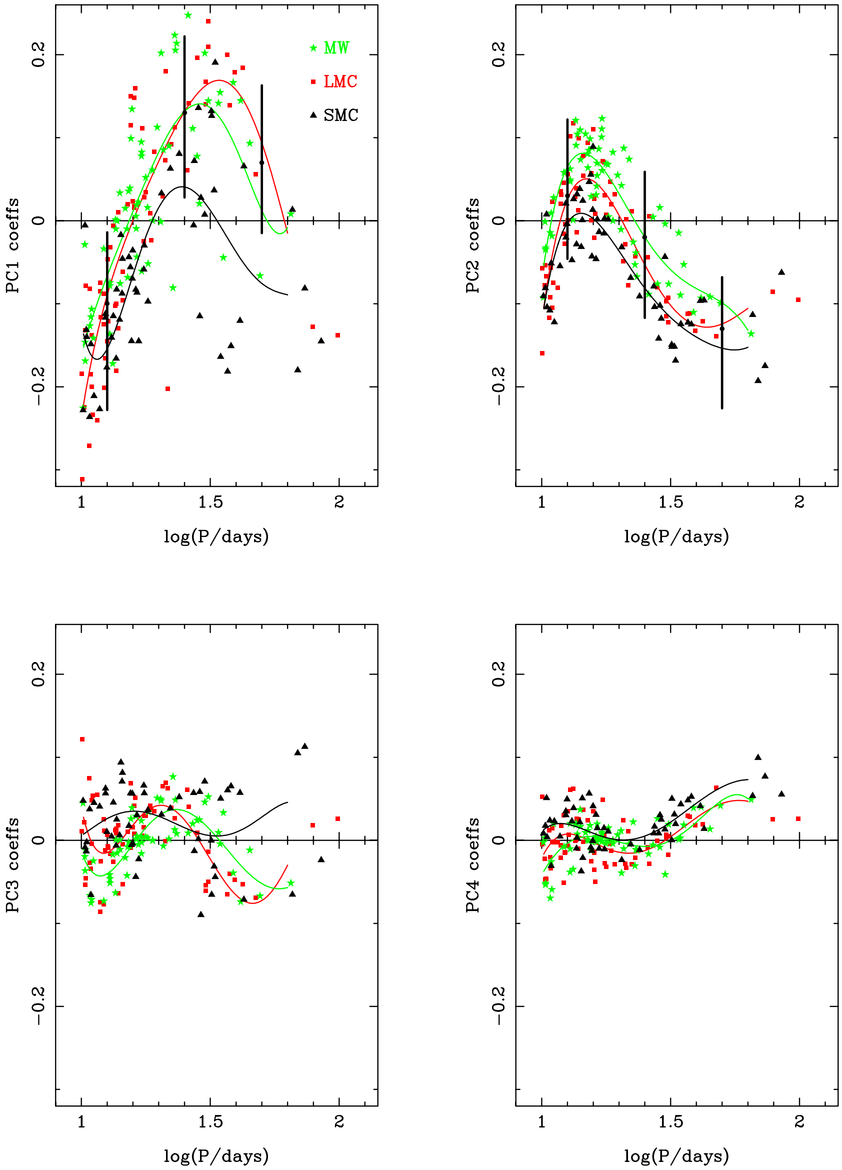

Figure 11. PCA coefficients plotted as a function of log period, for our training set of MW (light stars), LMC (darker squares), extended by the addition of a sample of SMC (dark triangles) Cepheids. The PCA basis used is that of Fig. 3 (i.e. the SMC data have been decomposed on to this basis). Separate low-order polynomial fits are shown for each set of points, illustrating the systematic differences in LCS from galaxy to galaxy, presumably due to metallicity differences. The large error bars show the spread in derived PCA coefficients expected from template fitting to 'typical' noisy data, as estimated from our simulations. This indicates that photometry for individual *HST*-observed Cepheids will not usually be good enough to quantify metallicity, but averaging together a reasonable sample for a particular galaxy might be.

$\log(P/\mathrm{days}) \sim 1.5$, although these Cepheids are *not* significant outliers in the PC2 coefficient distribution.

For the PC2 coefficients, the SMC points are again generally lower, as is most clearly seen by looking at the separate polynomial fits for each galaxy. Of course, given our particular focus on *HST* Cepheid studies, our training set consists only of fundamental mode Cepheids with period, $P > 10$ d. Thus Fig. 11 is not directly comparable to fig. 12 of Kanbur et al. (2002). Unlike the plots we presented for the MW and LMC alone (Fig. 3) we now also see that PC3 and PC4 coefficients for the SMC tend to follow the upper envelopes of distributions.

The metallicities of LMC and SMC Cepheids are thought to be about 50 and 20 per cent of the MW Cepheids, respectively. If we assume that the observed differences in PCA coefficients are due to metallicity then it may provide a route to estimating the average metallicities of other samples of Cepheids directly. The feasibility of this is indicated by the bold vertical bars in the top panels of Fig. 11. These show, at three different values of log P, the 1σ spread in returned PC1 and PC2 coefficients for the simulated 'typical S/N' data. (Given the typical V- and I-band sampling of *HST* Cepheid observations, it would be unrealistic to extract reliable LCS information from the third, or higher, principal components.) It is apparent that for individual Cepheids only very weak constraints can be placed with this quality of data. However, with better data or by averaging a reasonable sample of Cepheids, a useful, direct diagnostic of metallicity may well be achievable. We intend to investigate further the dependence of PCA on metallicity in a future paper.

7 CONCLUSIONS

We have presented in some detail our techniques to characterize Cepheid light-curves using principal component analysis of the Fourier coefficients for a set of well-observed Cepheids in the LMC and MW. We have also described how light-curve parameters can be extracted by fitting these templates to sparse and noisy data, and illustrated the method with extensive simulations.

The advantages of this approach are (i) very realistic light curves as a (smooth) function of period are obtained with only one or two principal components – in fact they are frequently better than the full Fourier fits to the calibrating data because averaging over the full set of Cepheids removes some numerical noise; (ii) multicolour data can be accommodated with a single combined fit which automatically accounts for the correlations between bands; (iii) template fitting to all data (weighted by the errors on each measurement) makes optimal use of the information in determining light-curve parameters; (iv) variables with poor fits (which might be produced by non-Cepheids or those whose photometry is badly affected by crowding) and potential period aliases are easily identified, and hence can be removed from consideration by applying objective, statistical criteria (rather than, for example, by visual inspection, as has often been the case in the past); (v) errors can be estimated in a moderately rigorous way, and Cepheids can be selected on the basis of goodness-of-fit; and (vi) maximum-light is straight forward to estimate and can be used as an alternative to mean-light in the PL relation.

The simulations themselves show that most Cepheids observed in *HST* campaigns should *individually* give distance moduli to about 0.2 mag (with the template fitting doing somewhat better than less sophisticated approaches), indicating that for typical sample sizes (several tens of variables), random errors in the derived Cepheid parameters should only be at the few per cent level for the sample as a whole, and systematics are likely to be the dominant source of uncertainty. Interestingly we have found that even with very poor S/N ratio data, errors can be as little as 0.3 mag for individual Cepheids using template fitting.

Finally we note that these methods can easily be extended to photometry in more than two bands, and to the analysis of other kinds of periodic variable stars.

ACKNOWLEDGMENTS

This work made use of the McMaster Cepheid Database, maintained by Doug Welch.

REFERENCES

Beaulieu J. P. et al., 1995, A&A, 303, 137
Berdnikov L. N., Turner D. G., 1995, Astron. Lett., 21, 717
Caputo F., Marconi M., Musella I., Santolamazza P., 2000, A&A, 359, 1059
Freedman W. L., 1988, ApJ, 326, 691
Freedman W. L. et al., 2001, ApJ, 553, 47
Hendry M. A., Tanvir N. R., Kanbur S. M., 1999, in Egret D., Heck A., eds, ASP Conf. Ser. Vol. 167, Harmonizing Cosmic Distance Scales in a Post-Hipparcos Era. Astron. Soc. Pac., San Francisco, p. 192
Kanbur S. M., Hendry M. A., 1996, A&A, 305, 1
Kanbur S. M., Mariani H., 2004, MNRAS, 355, 1361
Kanbur S. M., Ngeow C. C., 2004, MNRAS, 350, 962
Kanbur S. M., Iono D., Tanvir N. R., Hendry M. A., 2002, MNRAS, 329, 126
Kanbur S. M., Ngeow C., Nikolaev S., Tanvir N. R., Hendry M. A., 2003, A&A, 411, 361
Kochanek C., 1997, ApJ, 491, 13
Labhardt L., Sandage A., Tammann G. A., 1997, A&A, 322, 751
Lafler J., Kinman T. D., 1965, ApJS, 11, 216
Leonard D. C., Kanbur S. M., Ngeow C. C., Tanvir N. R., 2003, ApJ, 594, 247
Li C., Kong X., Cheng F., 2001, Prog. Astron., 19, 9
Moffett T. J., Barnes T. G., 1984, ApJS, 55, 389
Moffett T. J., Gieren W. P., Barnes T. G., Gomez M., 1998, ApJS, 117, 135
Morrison D. F., 1967, Multivariate Statistical Methods. McGraw-Hill, New York, p. 221
Ngeow C. C., Kanbur S. M., Nikolaev S., Tanvir N. R., Hendry M. A., 2003, ApJ, 586, 959
Paczynski B., Pindor B., 2000, ApJ, 533, L103
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes, 2nd edn. Cambridge Univ. Press, Cambridge, p. 292
Saha A., Hoessel J. G., 1990, AJ, 99, 97
Saha A., Sandage A., Tammann G. A., Dolphin A. E., Christensen J., Panagia N., Macchetto F. D., 2001, ApJ, 562, 314
Sakai S., Ferrarese L., Kennicutt R. C., Saha A., 2004, ApJ, 608, 42
Sandage A., Tammann G. A., 1968, ApJ, 151, 531
Schaltenbrand R., Tammann G. A., 1972, A&AS, 4, 265
Stetson P., 1996, PASP, 108, 851
Tanvir N. R., 1997, in Livio M., ed, The Extragalactic Distance Scale, Cambridge University Press, Cambridge, p. 91
Tanvir N. R., Shanks T., Ferguson H. C., Robinson D. R. T., 1995, Nat, 377, 27
Tanvir N. R., Ferguson H. C., Shanks T., 1999, MNRAS, 310, 175
Udalski A., Soszynski I., Szymanski M., Kubiak M., Pietrzynski G., Wozniak P., Zebrun K., 1999a, Acta Astron., 49, 223
Udalski A., Soszynski I., Szymanski M., Kubiak M., Pietrzynski G., Wozniak P., Zebrun K., 1999b, Acta Astron., 49, 437

APPENDIX A: RECONSTRUCTION OF CEPHEID LIGHT-CURVES

Although the primary purpose of this paper is to illustrate the general advantages of template fitting in obtaining Cepheid parameters, some readers may be interested in using the coefficients we have determined, for example to generate template light-curves for their own use.

The light curves for all Cepheids in both V and I are initially decomposed into Fourier terms (equation 1). The principal component analysis allows us to rewrite these light curves in terms of the PC vectors $\boldsymbol{P}_k$ and an average Cepheid light-curve $\boldsymbol{A}$.

$$m(t) = m_0 + \boldsymbol{A} + \sum_{k=1}^{k=32} \gamma_k \boldsymbol{P}_k \tag{A1}$$

Each PC vector (and indeed the 'average' light curve vector) are simply a sum of sine/cosine terms, the coefficients for which are given in Table A1.

$$\boldsymbol{P}_j = \sum_{k=1}^{k=16} \alpha_k \sin(2\pi k t/T) + \beta_k \cos(2\pi k t/T). \tag{A2}$$

In order to generate typical Cepheid light-curves at any given period, the γ_k coefficients can be obtained from the fits to the training-set data shown in Fig. 3.

$$\gamma_k = \sum_k \lambda_k [\log(P) - 1.4]^k, \tag{A3}$$

where period, P, is in days.

The coefficients for these equations are given in Table A2, and similarly in Table A3 for the polynomial fits shown in Fig. 11.

Table A1. The top row in this table gives the coefficients for the average Cepheid light-curve, to be used in conjunction with A2. The subsequent rows are the coefficients for the first 10 PC vectors.

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}
β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
0.395	0.000	−0.018	0.101	−0.022	−0.037	0.030	−0.002	−0.004	0.018	−0.009	−0.004	0.004	−0.005	0.000	0.000
0.245	0.064	−0.015	0.057	−0.011	−0.023	0.019	0.001	−0.004	0.012	−0.007	−0.004	0.003	−0.004	0.001	0.001
0.523	0.000	−0.039	0.469	−0.347	−0.077	0.108	−0.216	0.098	0.103	−0.076	0.034	0.005	−0.042	0.004	0.000
0.335	0.125	−0.059	0.275	−0.220	−0.059	0.071	−0.130	0.057	0.063	−0.048	0.018	0.005	−0.025	0.003	0.002
0.376	0.000	−0.141	−0.127	0.271	−0.319	0.266	0.328	−0.316	0.123	−0.024	−0.220	0.111	0.026	−0.030	0.017
0.233	−0.019	−0.048	−0.091	0.211	−0.183	0.162	0.221	−0.198	0.074	−0.025	−0.144	0.074	0.010	−0.022	0.014
−0.383	0.000	−0.134	−0.081	−0.255	−0.295	0.389	−0.179	0.015	0.330	−0.233	−0.076	0.062	−0.117	0.010	0.028
−0.308	0.012	−0.089	−0.057	−0.141	−0.202	0.225	−0.119	0.001	0.192	−0.165	−0.051	0.041	−0.075	−0.002	0.015
−0.256	0.000	0.138	0.409	−0.224	0.021	0.082	0.150	−0.333	−0.051	0.278	−0.276	0.055	0.260	−0.109	−0.043
−0.198	0.057	0.047	0.227	−0.117	0.022	0.022	0.113	−0.238	−0.096	0.202	−0.174	0.036	0.204	−0.114	−0.037
−0.071	0.000	−0.658	0.036	−0.075	−0.300	−0.148	0.068	0.004	−0.251	0.148	0.159	−0.167	0.067	−0.047	−0.098
−0.041	−0.144	−0.367	0.067	−0.046	−0.183	−0.075	0.004	0.037	−0.169	0.110	0.117	−0.117	0.064	−0.051	−0.080
−0.065	0.000	−0.397	0.033	0.131	0.393	−0.178	−0.127	−0.074	0.200	−0.173	−0.322	0.308	−0.103	−0.103	0.071
0.097	−0.216	−0.239	0.125	−0.020	0.235	−0.154	−0.122	−0.058	0.055	−0.047	−0.220	0.175	−0.021	−0.089	0.017
−0.074	0.000	0.079	−0.320	−0.235	−0.098	−0.008	−0.267	−0.066	−0.062	0.301	−0.105	0.058	0.212	−0.125	−0.035
0.429	−0.479	0.214	−0.166	−0.164	−0.087	0.068	−0.181	0.024	0.043	0.076	−0.077	0.042	−0.007	−0.011	−0.025
0.149	0.000	−0.037	−0.236	−0.173	0.155	−0.006	−0.108	−0.016	0.156	−0.074	−0.039	−0.144	0.035	−0.254	−0.552
−0.029	0.326	−0.062	−0.232	0.024	0.070	−0.052	0.013	−0.099	0.011	0.046	−0.069	−0.057	0.118	−0.224	−0.419
−0.331	0.000	0.142	0.185	0.186	−0.122	−0.103	0.215	−0.100	−0.005	−0.048	0.012	0.014	−0.320	0.060	−0.453
0.301	−0.134	0.066	0.269	−0.062	−0.070	−0.021	−0.042	−0.015	0.116	−0.142	0.101	−0.008	−0.260	0.079	−0.303
−0.037	0.000	0.147	0.110	0.190	−0.253	−0.096	−0.030	0.357	−0.103	−0.129	0.030	0.397	0.011	−0.441	−0.012
−0.030	0.013	0.056	0.001	−0.027	−0.187	0.022	0.127	0.217	−0.164	0.042	−0.012	0.271	0.096	−0.363	−0.026

Table A2. The coefficients determined in equation (A3) from a polynomial fit (shown in Fig. 3) as a function of log period to the first four principal component coefficients, $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, for our training set. The final column gives the rms scatter of the data points around the fits.

	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	Scatter
γ_1	0.124	0.313	−1.201	−0.281	−4.546	−2.476	17.944	0.076
γ_2	−0.035	−0.557	0.717	3.530	−8.867	−0.906	10.359	0.042
γ_3	0.028	−0.232	−1.439	3.233	7.009	−14.092	3.154	0.037
γ_4	−0.013	0.034	0.986	0.745	−7.507	−2.470	14.399	0.024

Table A3. As for Table A2, but in this case describing the polynomial fits displayed in Fig 11.

		λ_0	λ_1	λ_2	λ_3	λ_4	λ_5
γ_1	LMC	0.131	0.505	−1.214	−3.228	−2.477	11.339
	MW	0.132	0.301	−2.291	−3.442	5.852	17.230
	SMC	0.041	−0.054	−2.765	3.913	11.513	−20.621
γ_2	LMC	−0.053	−0.577	0.588	3.480	−6.082	1.647
	MW	−0.020	−0.478	0.739	1.970	−7.774	3.412
	SMC	−0.081	−0.411	0.529	1.050	−5.489	8.000
γ_3	LMC	0.025	−0.367	−1.570	4.613	9.032	−17.867
	MW	0.036	−0.102	−1.510	1.718	6.689	−8.885
	SMC	0.015	−0.136	0.217	2.044	−0.943	−5.493
γ_4	LMC	−0.012	0.102	0.774	−0.901	−3.864	5.015
	MW	−0.007	−0.005	0.664	1.357	−3.699	−3.809
	SMC	0.005	0.113	0.638	−0.680	−2.547	2.894

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