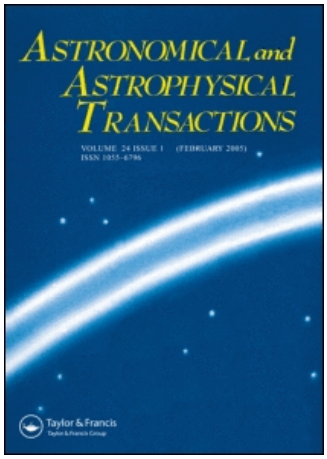

Astronomical & Astrophysical Transactions
The Journal of the Eurasian Astronomical Society
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713453505

Principal component analysis - an efficient tool for variable stars diagnostics

Z. Mikulášek [ab]
[a] Institute of Theoretical Physics and Astrophysics, Masaryk University, Brno, Czech Republic
[b] Observatory and Planetarium of J.Palisa, Ostrava, Czech Republic

Online Publication Date: 01 February 2007
To cite this Article: Mikulášek, Z. (2007) 'Principal component analysis - an efficient tool for variable stars diagnostics', Astronomical & Astrophysical Transactions, 26:1, 63 - 70
To link to this article: DOI: 10.1080/10556790701343850
URL: http://dx.doi.org/10.1080/10556790701343850

Astronomical and Astrophysical Transactions
Vol. 26, Nos. 1–3, February–June 2007, 63–70

Principal component analysis – an efficient tool for variable stars diagnostics

Z. MIKULÁŠEK*†‡

†Institute of Theoretical Physics and Astrophysics, Masaryk University, Kotlářská 2, Brno, Czech Republic
‡Observatory and Planetarium of J.Palisa, Ostrava, Czech Republic

(Received 19 February 2007)

We present two diagnostic methods based on the ideas of principal component analysis and demonstrate their efficiency for sophisticated processing of multicolour photometric observations of variable objects.

Keywords: Variable stars; Light curves; Principal component analysis; Least-squares method; Robust regression

1. Introduction

In the last few decades there has been a much larger volume of an increased common access to high-quality observational data on variable stars; however, the standard methods used for data processing and interpretation have lagged behind this progress. One of the steps taken to overcome this lack of progress is the consequent application of principal component analysis (PCA) combined with robust regression (RR), factor analysis, wavelet analysis and other sophisticated approaches to the treatment of observations of variable stars.

The commonly used method for the treatment of astrophysical data is the simple (unweighted) least-squares method (LSM). As these data usually suffer from outliers and very different degrees of quality, the method yields questionable and misleading results. RR as an adequate alternative of the standard LSM is used only seldomly; LSM weights, if introduced at all, are often used incorrectly.

2. Standard and weighted principal component analysis

PCA is one of the oldest and most elaborate methods employed to treat statistical data. PCA can be used to simplify a data set without loss of information. It is a linear transformation that

*Email: mikulas@physics.muni.cz

Astronomical and Astrophysical Transactions
ISSN 1055-6796 print/ISSN 1476-3540 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10556790701343850

chooses a new coordinate system such that the greatest variance corresponds to the first axis, then the residuals to the second axis, etc. PCA is simple and straightforward; it does not need any model. It reduces the number of uncorrelated parameters necessary for the description of a data set, helps to reveal hidden relationships and effectively suppresses noise. For more details see, for example, [1] or [2].

At present, it is profusely used particularly in image techniques, politics, criminal science, sociology and other human sciences; however, in astronomy it is almost unknown. We shall demonstrate how to apply standard PCA on routine tasks in the processing of the observations of variable stars.

Let us assume that we have p photometric measurements obtained in q photometric colours, which we can arrange in the form of p row vectors with q components: $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p\}$, $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iq}]$, or into the $p \times q$ matrix $\mathbf{Y}$. Each measurement can be then described as a point in q-dimensional (q-D) space; all p observations represent the 'cloud' of points, whose global characteristics we shall study by means of standard PCA.

If we want to use PCA as effectively as possible, we must linearly transform the components of these data vectors into new variables $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_p\}$:

$$z_{ij} = \frac{y_{ij} - \overline{y_j}}{\overline{s_j}}, \tag{1}$$

where $\overline{y_j}$ is the mean value of the jth components (jth colour), $\overline{s_j}$ is the estimate of the mean (typical) error (uncertainty) in the jth component. The purpose of this transformation is to identify the middle of the data cloud of observations with the origin of the new system of coordinates and to equalize all coordinates among them. PCA here implicitly hypothesizes that at least the ratios between the measurements in various colours are roughly constant. 'Error boxes' of particular measurements in q-D space should have the form of spheres of unit radius.

Standard PCA can be easily extended to weighted principal component analysis (WPCA), introducing the weights of individual data vectors. The weight is put inversely proportional to the square of ε_i: $w_i \propto \varepsilon_i^{-2}$, where ε_i is the expected uncertainty of a component of the ith data vector $\boldsymbol{z}_i$. Let $\boldsymbol{w} = [w_1, w_2, \ldots, w_p]$ be a vector describing the weights of individual data vectors; the diagonal matrix of weights, $\mathbf{W}$, of size $p \times p$ is defined as $\mathbf{W} = \mathrm{diag}(\boldsymbol{w})$. In our q-D representation it corresponds to when error spheres of individual sets of multicolour measurements are permitted to have various effective radii, proportional to ε_i. Standard PCA is then the special case of WPCA with equal weights: $\mathbf{W} \approx \mathbf{I}_p$.

The above-mentioned PCA linear transformation of a vector $\boldsymbol{z}$ to a smoothed vector $\boldsymbol{z}_\mathrm{s}$ using the smoothing $q \times q$ **matrix $\hat{\mathbf{A}}$ can be** written as

$$\boldsymbol{z}_\mathrm{s} = \boldsymbol{z}\hat{\mathbf{A}} = \boldsymbol{z}(\mathbf{A}\mathbf{A}^\mathrm{T}), \quad \boldsymbol{y}_\mathrm{s} = [z_{\mathrm{s}1}\overline{s_1} + \overline{y_1}, \ldots, z_{\mathrm{s}q}\overline{s_q} + \overline{y_q}], \tag{2}$$

where $\mathbf{A}$ is the $q \times r$ matrix consisting of r columns of normalized eigenvectors of the symmetric definite $q \times q$ matrix $\mathbf{Z}^\mathrm{T}\mathbf{W}\mathbf{Z}$, where $\mathbf{Z}$ is the $q \times p$ data matrix: $\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q]$. As follows from the definition, each eigenvector $\boldsymbol{a}_i$ together with the corresponding eigenvalue λ_i obeys the relation

$$(\mathbf{Z}^\mathrm{T}\mathbf{W}\mathbf{Z})\boldsymbol{a}_i = \boldsymbol{a}_i\lambda_i. \tag{3}$$

It can be proved that, for the $q \times q$ matrix $\mathbf{Z}^\mathrm{T}\mathbf{W}\mathbf{Z}$, just q eigenvalues and q normalized eigenvectors exist. The total set of q eigenvectors forms an orthonormal vector base. Let us order the eigenvectors according to their eigenvalues from the largest to the smallest in the sequence $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q\}$. Now we take the first r ($r \leq q$) eigenvectors and connect to give the

matrix $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r]$. Their eigenvalues then create the diagonal of the $r \times r$ diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}([\lambda_1, \ldots, \lambda_r])$:

$$(\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})\mathbf{A} = \mathbf{A}\mathbf{\Lambda}; \quad \boldsymbol{a}_i \cdot \boldsymbol{a}_j = \delta_{ij} \Longrightarrow \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}, \tag{4}$$

where δ_{ij} is the discrete version of the Kronecker delta function, $\mathbf{I}$ is the $r \times r$ identity matrix. The vectors $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_r\}$ contained in $\mathbf{A}$ represent the orthonormal vector base of the r-D subspace placed in the q-D space. The arranged set of scalar products of a vector z and vectors $\{\boldsymbol{a}_i\} : \{k_1, k_2, \ldots, k_r\}$, where $k_i = \boldsymbol{z} \cdot \boldsymbol{a}_i$, defines a vector $\boldsymbol{k}$:

$$\boldsymbol{k} = z\mathbf{A}; \quad z_{\mathrm{s}} = \boldsymbol{k}\mathbf{A}^{\mathrm{T}} = z\mathbf{A}\mathbf{A}^{\mathrm{T}}, \quad \mathbf{Z}_{\mathrm{s}} = \mathbf{K}\mathbf{A}^{\mathrm{T}} = \mathbf{Z}\mathbf{A}\mathbf{A}^{\mathrm{T}}. \tag{5}$$

We can introduce the $q \times r$ matrix $\mathbf{K} : \mathbf{K} = [\boldsymbol{k}_1, \ldots, \boldsymbol{k}_q]$. Assuming equation (4), we can write

$$\mathbf{K} = \mathbf{Z}\mathbf{A} \Longrightarrow \mathbf{A}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})\mathbf{A} = \mathbf{K}^{\mathrm{T}}\mathbf{W}\mathbf{K} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{\Lambda} = \mathbf{\Lambda}. \tag{6}$$

Equation (6) shows us that the eigenvalues correspond to the sum of the weighted variance of the projections of all vectors z_i and provides the reason why we should confine ourselves to only those components with eigenvalues that are sufficiently large; others do not contain any true information; they represent only a noise and therefore could be trimmed.

The application of PCA and WPCA should help us to find the number of parameters essential for the description of variability (the number of mechanisms of variability in action); it enables us to examine the relative quality of observations in multicolour measurements. Although we do not know the s_j value of individual colours exactly, we could improve them very quickly using an iterative circle. The convergence of this process is fairly good, because the results are as a rule insensitive to the s_j used.

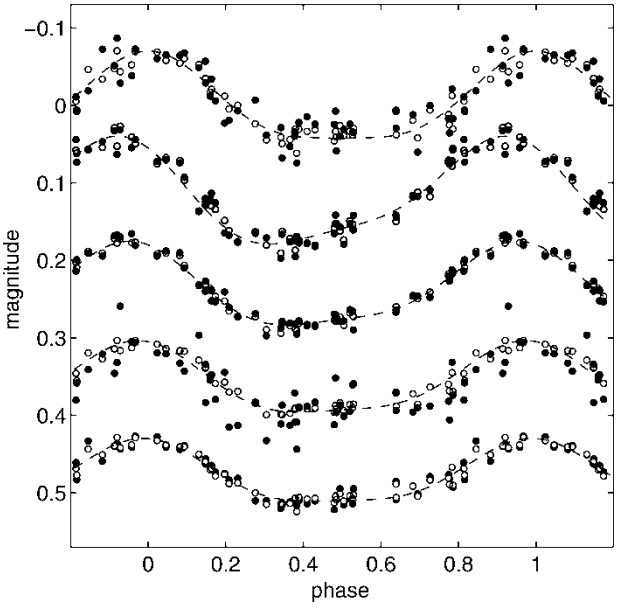

Figure 1. Smoothing of 'observational' data (full circles) by the standard PCA method. Smoothed points are denoted by open circles; the dashed curves represent the original (without noise) light curves in five synthetic colours.

The above-mentioned methods help us mainly in the preliminary processing of observational data, when we want to reach an orientation in the nature of the variability of studied objects, possible relationships between measured quantities and their quality. All this information can be gained without using any physical model and time-dependent smoothing, which can strongly influence finding *a priori* unexpected types of variability (rapid variations, trends, etc.).

We demonstrate the PCA treatment of artificial photometric data (50 observations in five colours), simulating the light variability of a rotating, chemically peculiar (CP) star with two differently coloured photometric spots. The 'observed' points and smoothed points with suppressed noise for $r = 2$ for individual colours are displayed in the phase diagram in figure 1. The treatment does not consider information on the phase.

PCA methods, similarly to LSMs, suffer from outliers, which are quite common in astrophysical data. The introduction of weights into PCA enables us to eliminate their influence by means of an iterative process that adjusts the individual weights of entering data (see, for example, the appendix of [3]).

3. Advanced principal component analysis

The extent of applicability of standard PCA and WPCA methods is rather limited as they are demanding with regard to the completeness and homogeneity of input data. These confinements were obviously one decisive reason why PCA techniques remain beyond the scope of the majority of observing astronomers.

Since 2000 we have developed a qualitatively new method that synthesizes WPCA and RR. We shall call it advanced principal component analysis (APCA). The versatility of APCA proves it to be quite broad. It has been used several times (see, for example, [4–7]); however, it has not been fully described until now. We shall briefly present only the method, without its derivation and strict mathematical proof of lemmas or statements.

3.1 *Vector description of light curves*

Let the course of a light curve be described by means of a preselected model the parameters of which are determined by standard regression methods, such as the LSM with weights or its modification that eliminates the influence of outliers. It is advisable to use a linear model so that the course of a light curve in a certain colour c, denoted $m_c(t)$, would be described by a linear combination of the ensemble of q so-called *elementary* functions $f_1(t), f_2(t), \ldots, f_q(t)$ which define the time dependence by a column vector $\boldsymbol{f}(t) = [f_1(t), f_2(t), \ldots, f_q(t)]$ of length q by the relation

$$\Delta m_c(t) = m_c(t) - \overline{m_c} = \sum_{i=1}^{q} y_{ci} f_i(t) = \boldsymbol{y}_c \cdot \boldsymbol{f}(t), \tag{7}$$

where y_{ci} are the components of a row vector $\boldsymbol{y}_c$, and $\overline{m_c}$ is the mean magnitude in the colour c. The components of the vector are found from the observational data by standard regression procedures (weighted LSM and RR).

We should be very particular about the choice of the base of the elementary functions $f_1(t), f_2(t), \ldots, f_q(t)$. The functions should be selected so that they enable us to express all studied light curves of the object with sufficient accuracy. It is advisable for many reasons (avoiding problems with multicollinearity, and equality of the uncertainty in the components of

the vector $\boldsymbol{y}_c$) to opt for elementary functions so that they would form the base quasiorthonormal of the set of data, which means that

$$\overline{f_i^2} \approx \overline{f_j^2} \approx \overline{f^2}, \quad \overline{f_i f_j} \ll \overline{f^2} \quad \text{for } i \neq j. \tag{8}$$

In the case when the set of elementary functions does not obey the conditions given above, it is trivial to transform the system into the orthonormal system by means of the standard Gram–Schmidt orthonormalization procedure.

Assuming that the observational data are distributed more or less uniformly over the observational interval, it is recommended that Legendre polynomials are used for the orthonormal in the interval $\langle -1; 1\rangle$. If the object is periodically variable, then the condition of quasiorthogonality fulfils any combination of the harmonic polynomials $\sin(2k\pi t/P)$ and $\cos(2k\pi t/P)$, $k = 1, 2, \ldots; \overline{f^2} = 1/2$.

If the functions of the linear regression model are quasiorthonormal, then the uncertainties in particular components ε_c of the vector $\boldsymbol{y}_c$ are the same:

$$\varepsilon_c \approx \frac{s_c}{(N_c \overline{f^2})^{1/2}}, \quad w_c \propto \varepsilon_c^{-2} \propto \frac{N_c}{s_c^2}, \tag{9}$$

where s_c is the standard deviation of the light curve fit and N_c is the number of observations in the particular colour used for the light curve fit. The weight of the corresponding vector of light curve in the c colour then will be proportional to ε_c^{-2}.

The whole set of vectors describing the light curves in all p colours can be arranged into the $p \times q$ matrix $\mathbf{Y}$, with the weights described by the $p \times p$ diagonal matrix $\mathbf{W}$.

3.2 *Advanced principal component analysis; reducing free parameters; use of advanced principal component analysis*

Let us allow the variable part Δm_c of the light curves in all colours to be sufficiently accurately approximated by a linear combination of only r ($r < q$) normalized orthogonal (*principal*) functions $\varphi_j(t)$ determined by the linear combination of all q elementary functions $f_i(t)$ with the coefficients forming the $q \times r$ matrix $\mathbf{B}$:

$$\varphi_j(t) = \sum_{i=1}^{q} b_{ij} f_i(t) = \boldsymbol{b}_j \cdot \boldsymbol{f}(t), \tag{10}$$

$$\Delta m_c(t) = \sum_{j=1}^{r} k_{cj} \varphi_j(t) = \sum_{j=1}^{r} k_{cj} \boldsymbol{b}_j \cdot \boldsymbol{f}(t) = \boldsymbol{k}_c \mathbf{B}^{\mathrm{T}} \boldsymbol{f}(t) = \boldsymbol{y}_{sc}\, \boldsymbol{f}(t), \tag{11}$$

$$\boldsymbol{y}_{sc} = \boldsymbol{k}_c \mathbf{B}^{\mathrm{T}}, \tag{12}$$

where $\boldsymbol{b}_j$ is the normalized vector of the jth principal function and jth column of the matrix $\mathbf{B}$. The row vector $\boldsymbol{k}_c$ ($1 \times r$) represents the semiamplitude components of the light curve in colour c versus r principal functions $\{\varphi_1(t), \ldots, \varphi_r(t)\}$ and the $1 \times q$ vector $\boldsymbol{y}_{sc}$ contains parameters of the APCA smoothed light curve in the colour c.

Further we shall assume that the vector base $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r\}$ is orthonormal; then

$$\boldsymbol{b}_i \boldsymbol{b}_j = \delta_{ij}, \quad \mathbf{B}^{\mathrm{T}} \mathbf{B} = \mathbf{I}. \tag{13}$$

Minimizing the scalar quantity $S(\mathbf{B}, \boldsymbol{k}_c)$, which is defined as the the sum of weighted variances of the differences, $\Delta \boldsymbol{y}_c = \boldsymbol{y}_c - \boldsymbol{y}_{sc}$, and given by

$$S(\mathbf{B}, \boldsymbol{k}_c) = \sum_{c=1}^{p} \Delta \boldsymbol{y}_c \Delta \boldsymbol{y}_c^{\mathrm{T}} w_c = \sum_{c=1}^{p} (\boldsymbol{y}_c \boldsymbol{y}_c^{\mathrm{T}} - \boldsymbol{k}_c \boldsymbol{k}_c^{\mathrm{T}}) w_c; \quad \text{grad}\, S = 0, \tag{14}$$

we arrive after some algebra at the following conclusions:

$$\mathbf{K} = \mathbf{YB}, \quad (\mathbf{Y}^{\mathrm{T}}\mathbf{WY})\mathbf{B} = \mathbf{B}(\mathbf{K}^{\mathrm{T}}\mathbf{WK}) = \mathbf{B}\boldsymbol{\Lambda}. \tag{15}$$

$$\mathbf{Y}_{\mathrm{s}} = \mathbf{Y}\hat{\mathbf{B}} = \mathbf{Y}(\mathbf{BB}^{\mathrm{T}}). \tag{16}$$

The results in equations (15) and (16) are formally identical with those in equations (4)–(6); so we can conclude that the matrix $\mathbf{B}$ contains r column vectors which are eigenvectors of the matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{WY}$ corresponding to its first r largest eigenvalues. The smoothing matrix $\hat{\mathbf{B}}$ is defined identically as $\hat{\mathbf{A}}$.

Nevertheless, we have to emphasize that APCA is not identical with standard PCA or WPCA. APCA and PCA give very similar but not the same results; the smoothing matrix $\hat{\mathbf{B}}$ is not the duplicate of $\hat{\mathbf{A}}$! The main reason is that data treated by PCA have been centred on their mean, while in the case of APCA we handle the obtained data directly without any centring. The difference is formulated using the basic suggestion of APCA (equation (10)), which seems to us physically more valid than the postulates of PCA. The correctness of the APCA method has been verified using several relevant statistical tests and trials with simulated data.

We demonstrate the use of APCA on synthetic photometric data, simulating the light variability of the same model of the rotating CP star. Figure 2 displays the phase diagram of multicolour light variations; 'observed' points are indicated by full circles, and the light curves fitted by the standard LSM technique are indicated by dashed curves. The light curves found by APCA are plotted as full curves; they are indistinguishable from the curves fitted by the standard LSM

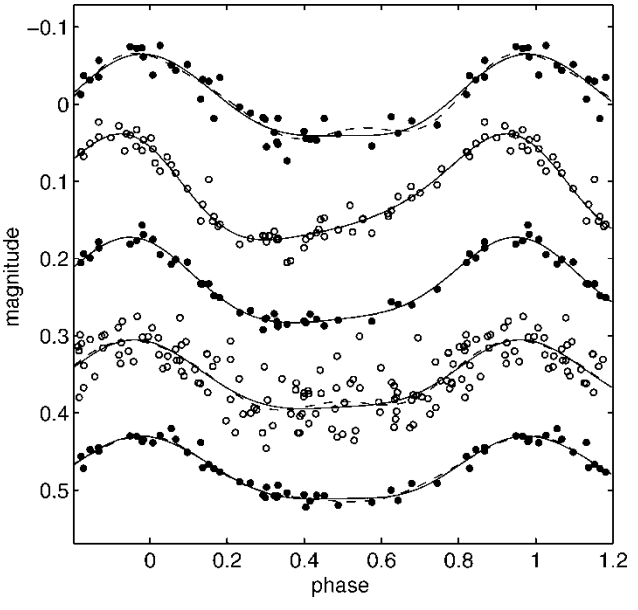

Figure 2. Fitting of 'multicolour observational' data by the APCA method (full curves). The dashed curves represent the fitting by the LSM.

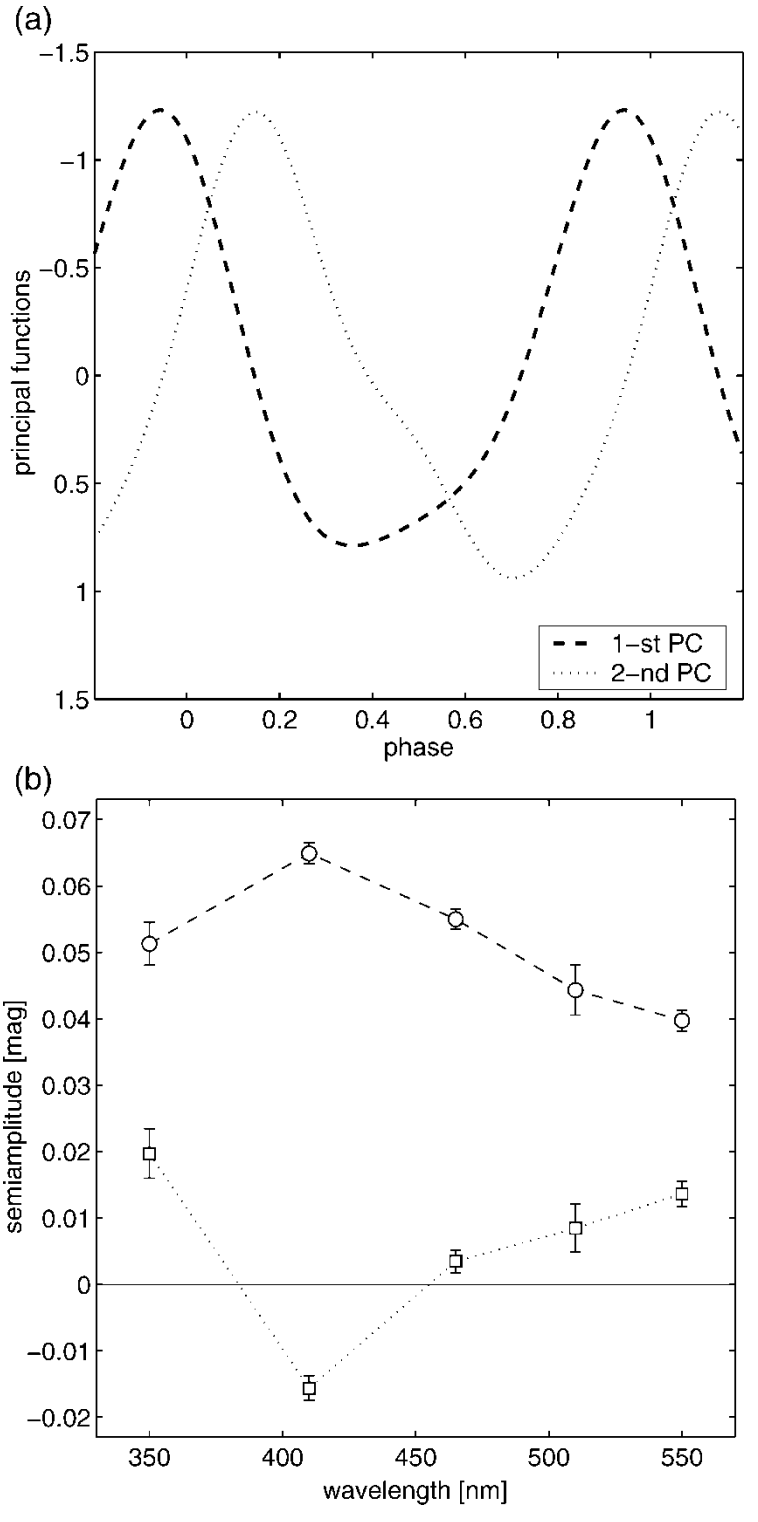

Figure 3. (a) The curves for the first two principal functions. (b) The dependence of the semiamplitude on wavelength for both principal components: dashed curves, first principal component; dotted curves, second principal component.

technique. The first two principal curves are displayed in figure 3(a); the dependences of the semiamplitudes on the wavelength for both principal light curves are plotted in figure 3(b). From the diagram for a real object we can obtain information about the light variability.

APCA can be used for reliable prediction of the multicolour behaviour of an object; the method is very suitable for the quantification and classification of light curves [5, 7], for multicolour $O-C$ measurements [6] and for improvement in the light ephemeris [5]. APCA seems to be a very efficient tool for the analysis of spectral variations and radial velocity measurements [4].

4. Conclusions

PCA and APCA prove to be universal, relatively simple methods with an extremely versatile range of uses in the processing and interpretation of astronomical data (both photometric and spectroscopic). The efficiency and applicability of the PCA increase when we combine it with other sophisticated methods of data treatment, *e.g.* RR, weighted LSM or wavelet analysis.

Acknowledgements

The author is highly indebted to Dr Miloslav Zejda and Dr Jan Janík for careful reading of the manuscript and valuable comments and suggestions. This investigation was supported by the Grant Agency of the Czech Republic, grants 205/04/2063 and 205/06/0217.

References

[1] J.E. Jackson, *A User's Guide to Principal Components*, Wiley, New York (2003).
[2] I.T. Jolliffe, *Principal Component Analysis*, 2nd edition, Springer, Berlin (2004).
[3] Z. Mikulášek, J. Žižňovský, J. Zverko *et al.*, Contrib. Astron. Obs. Skalnaté Pleso **33** 29 (2003).
[4] D. Korčáková, Z. Mikulášek, A. Kawka *et al.*, Inf. Bull. Variable Stars No. 5605 1 (2005).
[5] Z. Mikulášek and T. Gráf, Astrophys. Space Sci. **296** 157 (2005).
[6] Z. Mikulášek, J. Krtička, J. Zverko *et al.*, in *Proceedings of the Conference on Active-OB Stars: Laboratories for Stellar and Circumstellar Physics*, ASP Conference Series, volume 361, edited by S. Stefl, S. Owocki and A. Okazaki, Sapporo, Japan, 2005 (Astronomical Society of the Pacific, San Francisco, California, 2007), p. 466.
[7] Z. Mikulášek, J. Zverko, J. Žižňovský *et al.*, in *Proceedings of the Symposium on the A-Star Puzzle*, IAU Symposium, volume 224, edited by J. Zverko, J. Žižňovský, S.J. Adelman and W.W. Weiss, Poprad, Slovakia, 2005 (Cambridge University Press, Cambridge, 2005), p. 657.

Zdenek Mikulášek