Poster

Application of PCA to light curves of three optical pulsars

G. Naletto[1,2], G. Codogno[1], C. Barbieri[3], E. Verroi[4], M.Barbieri[5] and L. Zampieri[5]

[1] Department of Information Engineering, University of Padova, Via Gradenigo, 6/A, I-35131 Padova PD, Italy;
[2] CNR/IFN/LUXOR, Via Trasea, 7, I-35131 Padova, Italy
[3] presenter, Department of Physics and Astronomy, University of Padova, Vicolo Osservatorio 3, I-35122 Padova PD, Italy; e-mail: cesare.barbieri@unipd.it
[4] University Center of Studies and Activities for Space (CISAS) G. Colombo, University of Padova, Via Venezia 15, I 35131 Padova, Italy
[5] INAF Astronomical Observatory of Padova, Vicolo dell'Osservatorio 5, 35122 Padova, Italy

ABSTRACT

This paper describes a method for determining the optimal period of pulsars and consequently their light curves, by means of the Principal Component Analysis (PCA) applied to the so-called waterfall diagram, which is a bi-dimensional representation of the pulsar photometric data.
To test the method, the PCA was applied to sets of arrival times of photons from the brightest optical pulsars (Crab, B0540-69, Vela) measured with our fast photon counters Aqueye and Iqueye. Comparison with results obtained with the more usual technique of epoch folding shows that the periods determined via the PCA not only agree, but that actually their errors are smaller.

1. Introduction

High time resolution astrophysics (HTRA) investigates celestial objects presenting rapid photometric variability, such as occultations, oscillations in white dwarfs, flickering in cataclysmic variables, rapid variability in neutron stars, X-ray binaries and accreting compact objects, and so on. Among these phenomena, pulsars are probably the most studied ones with HTRA instruments and methods. A characteristic of these objects is their (quasi) periodic signal, produced by a spinning neutron star with an approximately conical beacon-like emission beam. With our instruments Aqueye (Barbieri et al. 2009a,b) applied to the Asiago Copernicus telescope (Italy), and Iqueye (Naletto et al. 2009) applied to the ESO New Technology Telescope (La Silla, Chile), we have investigated three optical pulsars accessible to medium size telescopes, namely PSR B0531+21 (the Crab pulsar, Germanà et al. 2012, Zampieri et al. 2014), PSR B0540-69 in the Large Magellanic Cloud (Gradari et al. 2011), and PSR B0833-45 (the Vela pulsar, Codogno et al. 2011). For the properties of optical pulsars, see e.g. Mignani 2011.

A most important analysis on the pulsar light curves is the optimal determination of the instantaneous period and of its variation with time, because many information can be obtained on emission mechanisms, lifetime and evolution. An important observational quantity is the so called 'braking index' n, consistently found below the canonical value $n = 3$, applicable if braking was due only to radiative losses.
To obtain the best periods and light curves, different techniques have been developed, depending on the available instrumentation, accessible spectral range and signal strength. The usual case for optical pulsars observed with medium class telescopes is acquiring photons in time bins much shorter than the period of the pulsar. Therefore, each time bin will contain very few or even no photons, and actually there will be a null signal in the large majority of cases, with no clear evidence of periodicity. Dedicated analysis tools are requested to identify accurate values of periodic or quasi periodic features.

In this paper we will describe an application of the Principal Component Analysis (PCA) to the optimal determination of periodicity in the Aqueye and Iqueye time stamps from the Crab pulsar, PSR B0540-69 and the Vela pulsar. Section 2 gives a short description of standard techniques used to estimate pulsar periods. Section 3 describes the PCA method. Finally, Sections 4 and 5 explain how to apply the PCA to the waterfall diagram to determine the pulsar optimal period and a confidence interval . It is shown that the PCA periods not only agree with those determined by other techniques, in particular epoch-folding, but actually have slightly better precisions.

2. Common techniques to determine the pulsar period

A first convenient step to identify the presence of periodicity is to calculate the Fast Fourier Transform (FFT) of the signal, as shown in Figure 1.

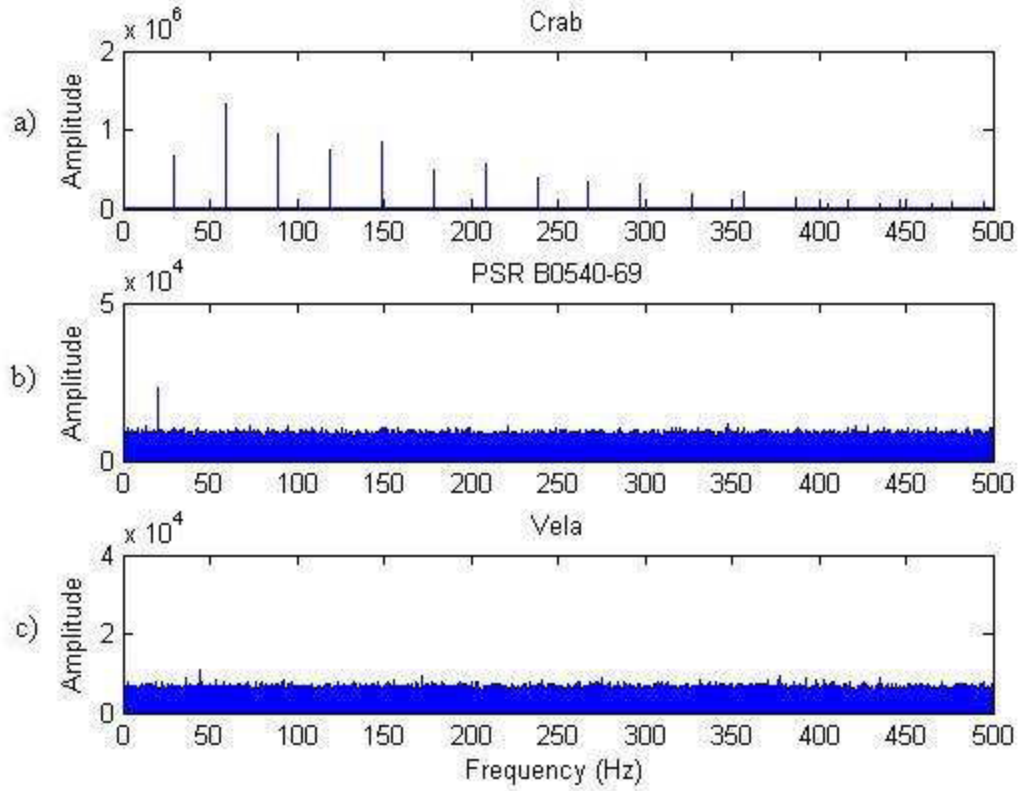

Figure 1 – The frequency spectrum of three optical pulsars, Crab (top), B0540-69 (middle), Vela (bottom) obtained with Iqueye at the NTT. One hour long acquisitions, data binned at 1 ms.

The FFT algorithm has some significant limitations, mainly because the determination of the optimal frequency is hampered by the poor frequency resolution, which depends on the number of input elements and on the ability of the software to deal with large arrays. Therefore, the FFT can be used for the detection of a periodic signal, but in a second step more accurate analyses must be performed. A very consolidated technique is 'epoch folding' (Leahy (1983); Leahy et al. (1983); Leahy (1987); Gregory & Loredo (1992); Larsson (1996)). A reasonable initial period P is assumed, then the detected signal as function of time is divided modulo a trial period P_t close to P. The trial period P_t is divided into N period time bins and the data modulo P_t are co-added into these time bins (often, the phase period is used instead of the time period, with the phase period ranging between 0 and 1). By co-adding (folding) the signal over a large number of periods, assuming the period constant over the time of folding, the statistics per period time bin is largely increased, and a first light curve is generated. At this point, P_t is varied by small amounts, and a whole set of light curves is produced. The optimal light curve is the one which maximizes a suitable merit function. Thus, the problem of finding the best light curve is reduced to the definition of the best merit function. A

fairly common algorithm used to this aim is the calculation of the χ^2 value: the highest χ^2 value indicates the most accurate light curve. Suitable algorithms can be found on dedicated libraries, for example the timing analysis software for X-ray Astronomy Xronos
https://heasarc.gsfc.nasa.gov/docs/xanadu/xronos/xronos.html

Another method is the 'waterfall diagram'. The signal is similarly folded to increase the statistics, but the procedure is different: the total observation time is divided into *M* time intervals, usually with $M \geq 20$, and folding is performed separately over each time interval with a common trial period P_t, so obtaining *M* light curves. Then, each light curve is stacked as a row in a matrix; by associating a colour scale to the light curve signal intensity, this matrix can finally be represented as a colour image. Notwithstanding its simplicity, the accuracy of this method is quite remarkable, and period values as accurate as those obtained by the epoch folding technique can be obtained. An example of such accuracy is shown in Figure 2, obtained from a 120 minutes observation of the Crab pulsar with Iqueye and using *M* = 2000. The time bin in such example was 4×10^{-5} seconds. A very slight curvature of the lines can be noticed. Indeed, the folding period was correct at the centre of the observations interval, but phase shifts occurred as a consequence of the extremely small Crab pulsar spin-down. The residual curvature allows to measure such period variation, namely $dP/dt \approx 4.2 \times 10^{-13}$ s/s at the time of observation (Zampieri et al. 2014), corresponding to a change in the period of only 3 ns from the beginning to the end of the observation, or equivalently to a total phase variation of ~ 0.0096.

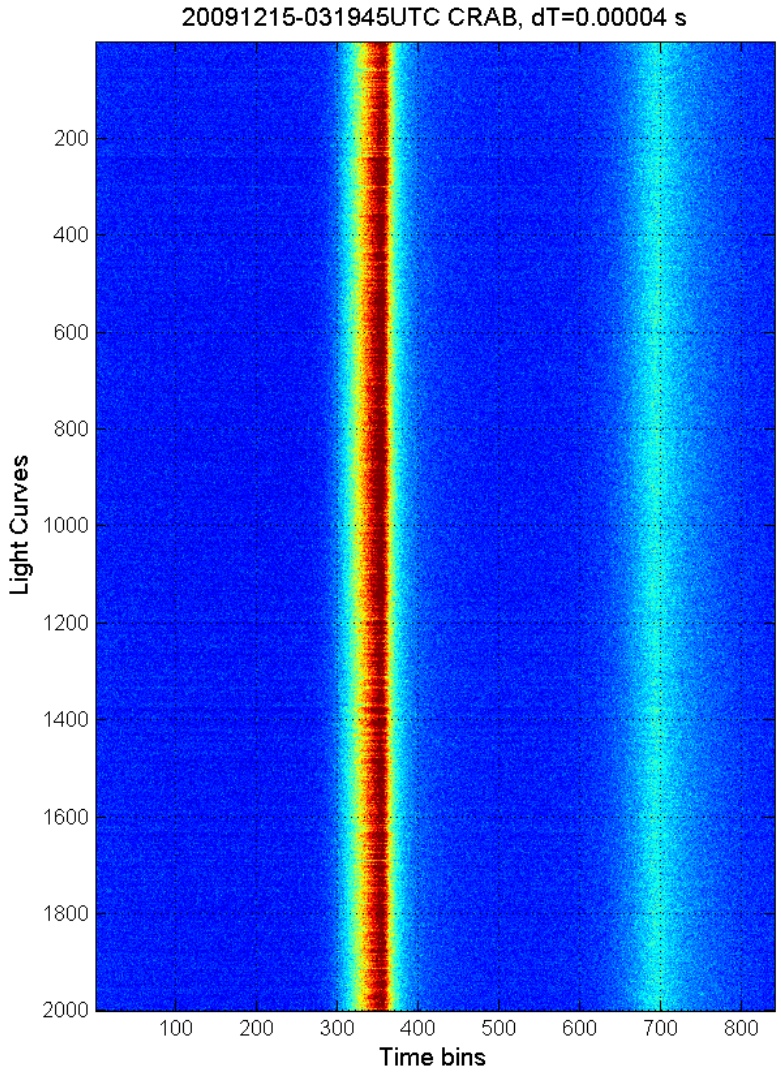

Figure 2 . Waterfall diagram of a two hours Crab pulsar acquisition with Iqueye at NTT on 15 December 2009, binned at 4×10^{-5} s. The slight curvature of the vertical lines is due to the pulsar spin-down during the two hours.

Although extremely accurate and sensitive, the waterfall diagram method can be applied with good results only when the searched periodic signal is high with respect to the noise (i.e. SNR > 5). When the noise is not negligible, the signal in the waterfall image has a very poor contrast making extremely difficult to detect any periodicity. As an example, Figure 3 shows the waterfall diagram for a set of Iqueye data of B0540-69, (where SNR $\ll$ 1),using M = 20 and the nominal folding period. Clearly, no straight vertical feature is evident, at least by eye.

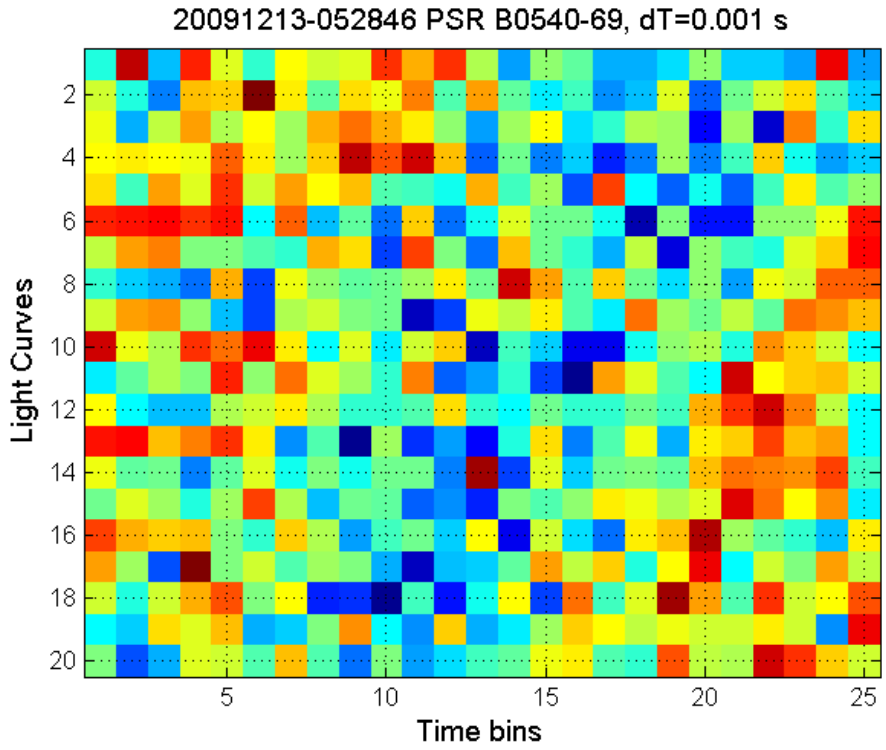

Figure 3. Waterfall diagram of B0540-69 observed for 30 minutes with Iqueye at the NTT and binned at 1x10^{-3} s. No vertical feature is evident, even if the data have been folded in only 20 rows.

The idea at the basis of this paper is to understand if a suitable analysis can extract the 'hidden' information from images like the one shown in Figure 3. We show that the Principal Component Analysis offers such a tool.

3. Principal Component Analysis

The Principal Components Analysis (Pearson (1901)) is a mathematical tool used in statistics, signal processing, meteorology, image compression, computer graphics, and so on. Depending on the application, the technique is subject to small differences in implementation, and so it is often named in different ways, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Proper Orthogonal Decomposition (POD), Karhunen-Loève Transform (KLT, Karhunen 1947, Loève 1978), Hotelling Transform (HT, Hotelling 1933, 1936), and many others.

PCA is a non-parametric method to extract information from data of otherwise difficult interpretation. Practically, PCA provides a way to identify patterns in data: from a geometrical point of view, this corresponds to represent the data in a reference frame able to highlight the data structures. To achieve such representation, the algorithm uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called Principal Components (PCs). The first PC has a variance as high as possible, accounting for as much as possible of the variability in the data, and each following component has in turn the highest possible variance under the constraint that it has to be orthogonal to (uncorrelated with) the preceding components. Figure 4 illustrates the simple

example of a two-dimensional (x,y) data sets having a first principal direction of variation, identified with the u axis and a second important direction orthogonal to u, that is the v axis.

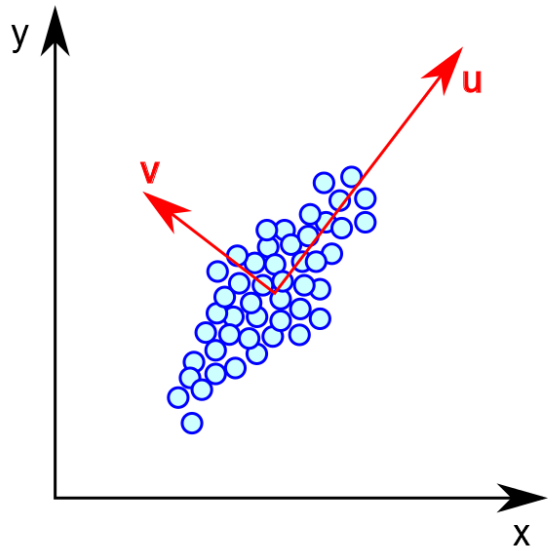

Figure 4. Given a bi-dimensional dataset, the PCA allows to identify the directions (u,v) of orthogonal maximum variances.

The PCA allows to identify these two directions and assign them a priority on the basis of where the data have the largest dispersion. In other words, the data set is arranged by PCA in a new reference frame (u, v) having origin at the centroid of the data and oriented in order to have uncorrelated data, that is their co-variance with respect to the (u, v) coordinates is zero. The directions of these axes are the so-called Principal Components.

The PCA algorithm consists of five steps. The first is to subtract the data mean values, independently on each variable. Then, the covariance matrix **C** is calculated: in the example of Figure 4, **C** will be a 2 × 2 matrix. The third step is the calculation of eigenvectors $\mathbf{e}_i$ and corresponding eigenvalues λ_i of the covariance matrix **C**. This process is equivalent to finding the reference frame in which the covariance matrix is diagonal, that is where the variables are uncorrelated: in this reference frame, the covariances (i.e. the non-diagonal elements of the matrix) are all equal to zero and only the variances (the diagonal elements) remain. The orientation of the axes of this reference frame are provided by the eigenvectors $\{\mathbf{e}_i\}$, which constitute an orthonormal basis, and the value of each eigenvalue λ_i is equal to the variance along the direction of the corresponding eigenvector $\mathbf{e}_i$. In the given example, the eigenvalues are equal to the variance of the data along the two eigenvectors (**u**, **v**), with the eigenvalue associated to the **u** eigenvector being the largest. In the fourth step the eigenvectors $\mathbf{e}_i$ are ordered by the corresponding eigenvalue λ_i from highest to lowest, to put the direction given by the eigenvectors in order of importance. If the ordered eigenvectors are used as columns of a matrix, a so-called feature vector **F** is obtained. This step allows to prioritize the PCs, with the most important ones retaining the largest amount of information, and decide the amount of original information to maintain, simply deciding how many PCs will be used in the following step. It often happens that only the first few PCs have large variances: discarding the less important PCs, the complexity of the system can be largely reduced at the acceptable price of taking out a minor amount of information. With the final fifth step, the new data set is obtained: the feature vector **F** is used as a transformation matrix that takes the data points from the (x,y) reference frame to the (u,v) one by means of the equation:

$$A(u, v) = (A(x, y) - < A(x;, y) >) \cdot \mathbf{F}$$

where A(x, y) is a point in the (x, y) reference frame, <A(x, y) > is the dataset centroid, and A(u, v) is the corresponding point in the (u,v) reference frame. Routines for PCA algorithms can be found in many scientific libraries.

4. Application of PCA to the waterfall diagram of the Crab pulsar

In order to verify the possibility to apply the PCA to the pulsar waterfall diagrams to determine the optimal folding period, tests have been done firstly in the easiest case of the Crab pulsar. We used a dataset of 30 minutes acquired with Aqueye in October 10, 2008 (Germanà et al. (2012)). The procedure was first to produce different $N \times M$ waterfall diagrams $\mathbf{W}_i$ where i is the index associated to each trial folding period T_i. In the test we used N = 336 columns, i.e. 336 time bins in the folding period, and M = 100 rows, i.e. 100 segments in which the whole acquisition was divided. Then we applied the PCA to each waterfall diagram, considering these images as M-dimensional datasets (i.e. a set of 336 100-dimensional hyper-vectors corresponding to the columns of the waterfall). From this analysis, a set of M-component eigenvectors $\mathbf{e}_{i,j}$ and of the corresponding eigenvalues $\lambda_{i,j}$ (j = 1,…, M) was obtained per each waterfall $\mathbf{W}_i$. Looking at the eigenvalues obtained with the waterfall corresponding to the nominal period T*, it could be noticed that the one associated to the first PC was largely dominant over the others, providing a ratio $\Lambda_{i,1} = \lambda_{I,1}/(\Sigma_{j=1,M}\, \lambda_{I,j})$ = 0.985, and that the components of the corresponding eigenvalue had all substantially the same value (≈ 1=√100). This was actually an expected result: when the period is the optimal one, the data are ideally the same per each row of the waterfall, and the components of each hyper-vector are the same. This means that in the 100-dimension space defined by the eigenvector basis, the optimal M-dimensional dataset lays along the hyper-diagonal. In other words, all the points are aligned along the privileged direction where the dataset has the largest variance, which is the value provided by the eigenvalue, while in the other perpendicular directions the variance is almost zero. By analysing the eigenvalues for the various waterfalls $\mathbf{W}_i$, it could be noticed that the ratio $\Lambda_{i,1}$ was increasing getting closer to the optimal folding period. All these considerations suggested to monitor the behaviour of the eigenvalue of the first PC to estimate the best folding period.

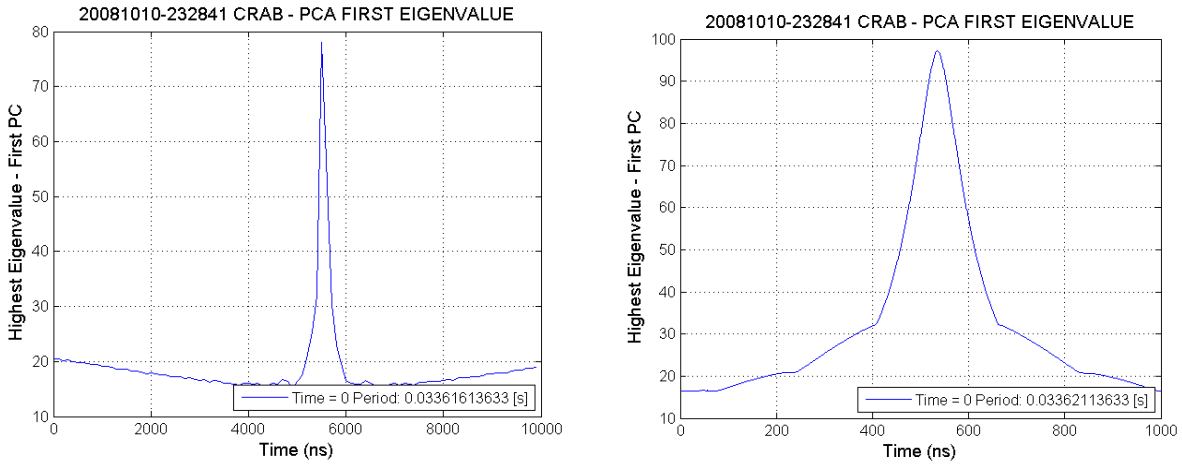

Figure 5. PCA of for a 30 min acquisition of the Crab pulsar with Aqueye. Left: first eigenvalue vs folding period with 100 ns step period increase. Right: First eigenvalue vs folding period with 0.1 ns step period increase.

In Figure 5 the first PC eigenvalue is shown for two different cases: i) on the left, the folding period is changed at 100 ns steps, to make a rough search of the peak region; ii) on the right, the peak search is realized more finely in a smaller region with a step of 0.1 ns. In this case the highest value of the first eigenvalue is found in correspondence of a pulsar period of 33.62167033 ms: comparing this result with the one provided by the Jodrell Bank monthly ephemerides (Lyne et al. 1993, http://www.jb.man.ac.uk/pulsar/crab.html) at the same date of the Aqueye observations, a difference of only ≈ 10 ps is found. As shown in Figure 6, the first component contains essentially all the amount of useful information.

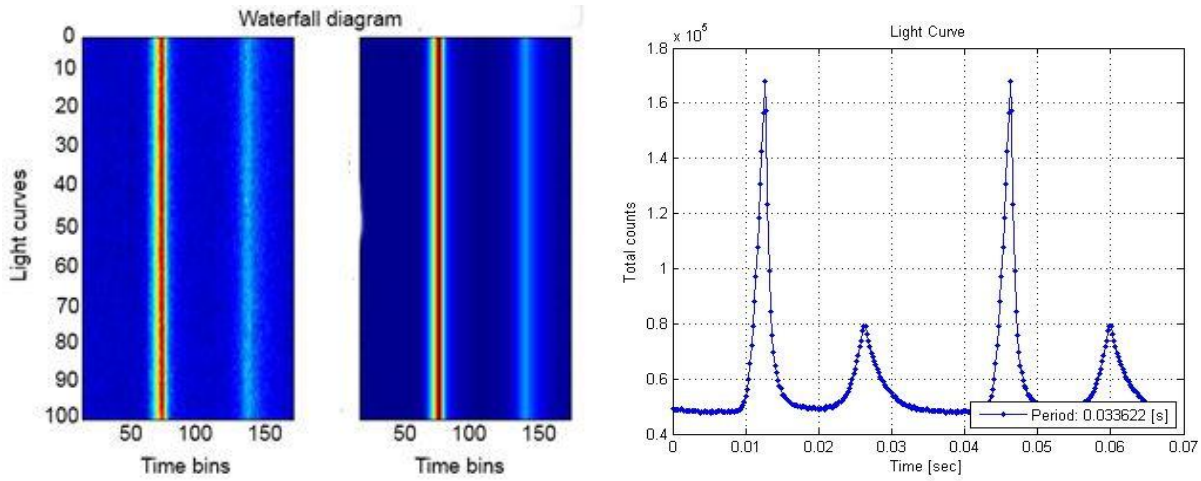

Figure 6. Left: the original waterfall diagram. Centre: the waterfall diagram obtained using only the first PC. Right: the light curve obtained by folding the data with the period found by means of the PCA.

The application of PCA to the waterfall diagram actually allows to make a further check of the goodness of the result. Since ideally the first PC eigenvector should be parallel to the hyper-diagonal in the M-dimensional space, the optimal period is obtained when the scalar product of this eigenvector with the hyper-diagonal unitary vector is maximum and equal to 1. The behaviour of this scalar product in the previously examined case is shown in Figure 7.

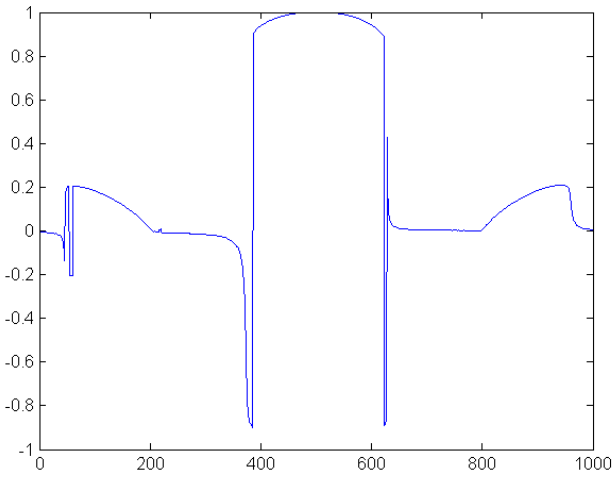

Figure 7. Scalar product of the first PC eigenvalue with the 100-dimension hyper-diagonal by varying the trial period. In this plot, the nominal period is at the centre of the abscissa axis.

This plot not only confirms that the largest scalar product corresponds to the nominal period, but in addition it allows to make a further interesting consideration: the value of the scalar product assumes values close to 1 only in a small range of the abscissa, with a discontinuity at about ±120 ns from the nominal period. Outside this range, the data are not aligned, and the first PC eigenvector is far from being parallel to the hyper-diagonal. Therefore, the behaviour of the scalar product can be used as a significant check on the confidence of the resulting period.

5. Application of PCA to the waterfall diagrams of the two other visible pulsars

The same technique has been applied to the more difficult cases of the other two optical pulsars observed in 2009 with the Iqueye at the NTT: PSR B0540-69 in the Large Magellanic Cloud (Gradari et al. 2011), and PSR B0833-45 in the Vela supernova remnant (Codogno et al., 2011). In these cases, to reduce the noise it has been necessary to take M = 20 only, but the results have been extremely interesting anyway.

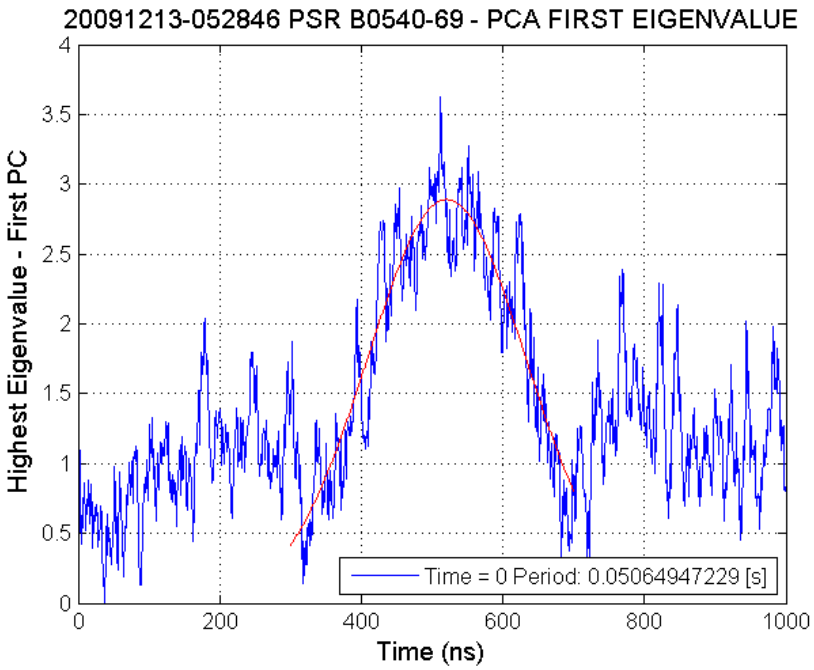

Figure 8. First eigenvalue vs folding period (0.1 ns step period increase) for an 100 min acquisition of B0540-69. The central portion of the data is interpolated with a Gaussian fit, shown in red, to find the optimal peak value.

As seen in Figure 8, which refers to an acquisition of 100 minutes of PSR B0540-69 on Dec. 13, 2009, the behaviour of the first PC eigenvalue as a function of the trial period is very noisy, and no peak is clearly defined. This problem was overcome by means of a standard Gaussian fit (in red): the period corresponding to the first PC eigenvalue peak is P_{PCA} = (0.0506499723 ± 62)·10^{-9} s. To cross check the goodness of this result, we determined the best period on the same dataset also by means of the standard epoch folding technique. The period obtained in this way was P_{EF} = (0.0506499745 ± 81)· 10^{-9} s, just 2 ns longer than the PCA one. Not only the two values are equal within the quoted errors, but in addition the errors associated with the PCA are slightly lower than those of epoch folding. Such errors (2 sigma values) were estimated by the following method: a Monte Carlo analysis was applied to a synthesized signal equivalent to the pulsar observation, noise included. Thousand iterations were performed in each case and we applied either standard epoch folding analysis or waterfall PCA.

The light curve obtained with the PCA best fitting period is compared in Figure 9 with the light curve obtained using the epoch folding technique and reported in Gradari et al. 2011. The latter has been obtained using more than 400 minutes of observations, and is clearly much smoother; however, it is also evident that the main features of the light curve, in particular the double structure of the main peak, are the same.

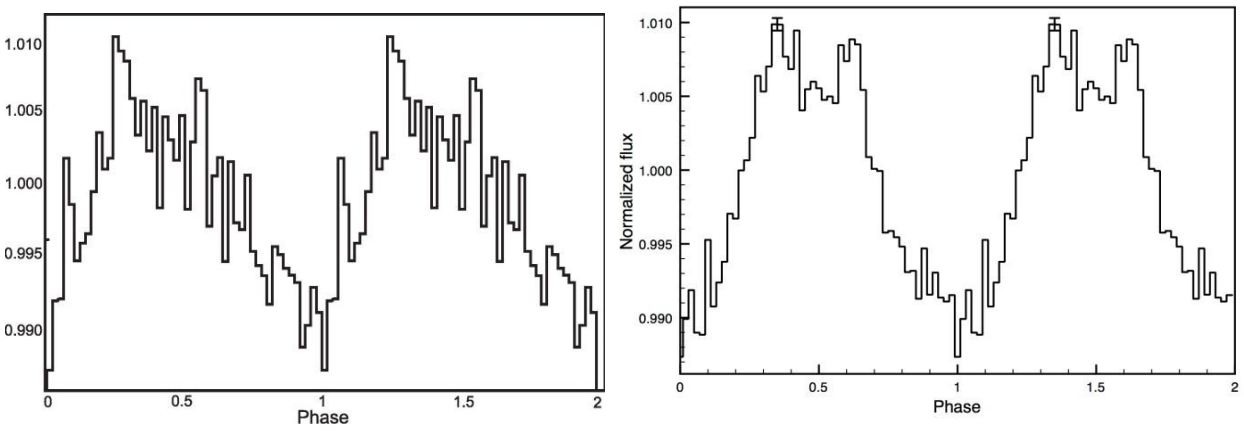

Figure 9. B0540-69. Left: light curve of a 100 minutes long data set obtained by folding the data with the period obtained by the PCA analysis. Right: same pulsar light curve obtained by epoch folding technique (Gradari et al. 2011) with a 428 minutes observation. Two cycles are shown for clarity.

Finally, we show the results obtained in the case of the Vela pulsar, acquired on 18 Dec. 2009. Although this pulsar is much fainter than B0540-69, the function of the first PC eigenvalue versus trial period is less noisy (see Figure 10), probably because the SN remnant around the pulsar is comparatively fainter than that of PSR B0540-69, and the disturbance due to the background is reduced.
As before, a Gaussian fit and Monte Carlo analysis identifies the PCA period and its confidence interval, P_{PCA} = (0.089366994 ± 21)· 10^{-9} s, to be compared with the epoch folding values P_{EF} = (0.089366983 ± 44)·10^{-9} s. The difference between the two periods is ΔP = 11 ns, well within the quoted errors. Also in this case, the error estimated by the Monte Carlo simulation is smaller for the PCA than for the epoch folding technique.

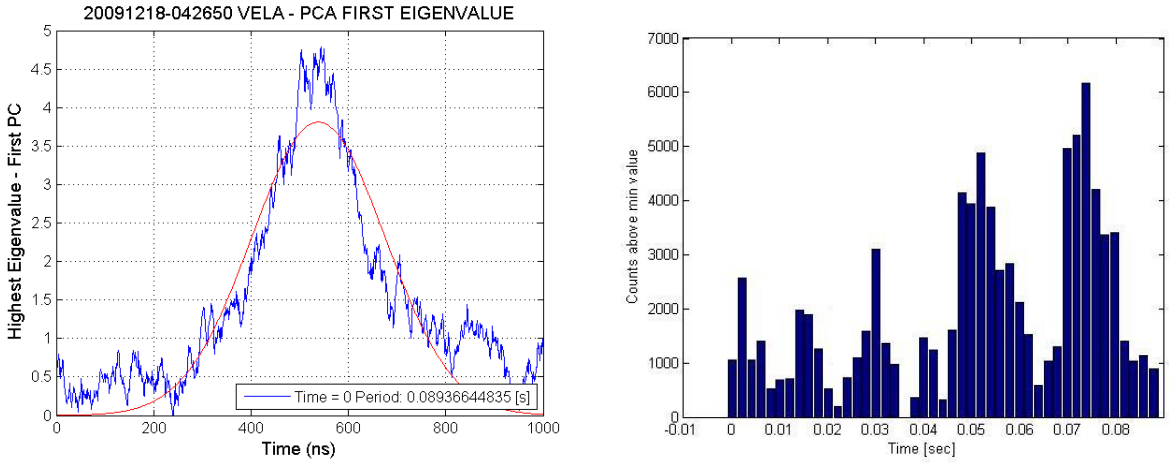

Figure 10. Left: first eigenvalue vs folding period (1 ns step period increase) for an acquisition of the Vela pulsar. The data are interpolated with a Gaussian fit, shown in red, to find the optimal peak value. Right: the very complex Vela pulsar light curve.

The right panel shows the very complex light curve obtained by using the first component of PCA (Codogno et al., 2011). This optical light curve, to our knowledge the first one since 1998 (Gouiffes, 1998), is very complex and of difficult interpretation.

We monitored also in these two cases the behaviour of the scalar product between the first PC eigenvalue and the unity vector parallel to the hyper-diagonal in the M-dimensional space (M = 20 in this case). The obtained results are shown in Figure 11.

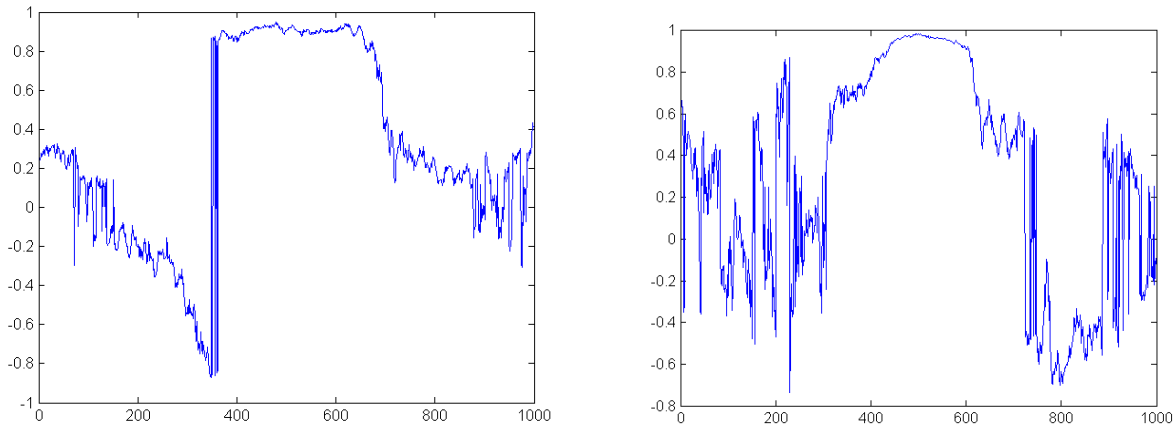

Figure 11. Scalar product of the first PC eigenvector and the hyper-diagonal unity vector as a function of the trial period. The optimal period corresponds to the centre of the abscissa axis. Left: B0540-69 pulsar. Right: Vela pulsar.

It can therefore be concluded that the scalar product allows to locate the optimal period in the limited region where the scalar product has values larger that 0.9. This is a sort of range of confidence which is slightly but significantly better than the one provided by the epoch folding technique.

6. Conclusions

We described in this paper the application of Principal Component Analysis to analyse the light curves of quasi-periodic objects like pulsars, in order to obtain their optimal periods and estimate a the range of confidence.

PCA results have been compared to those of epoch folding. The two methods provide very similar values for the best period. The PCA provides an additional tool to determine a range of confidence on such values. It turns out that the PCA value is affected by a slightly but definitely smaller uncertainty.

We conclude that PCA is a powerful tool for this type of analysis also when the signal is very weak. We applied the technique only to data acquired with our instruments, Aqueye and Iqueye, which work in visible light, but clearly it can be applied to other instruments and in all spectral ranges.

Acknowledgements. This work is based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 082.D-0382 and 084.D-0328(A), and on observations collected at the Copernicus Telescope (Asiago, Italy) of the INAF-Osservatorio Astronomico di Padova.We acknowledge the use of the Crab pulsar radio ephemerides available at the web site of the Jodrell Bank radio observatory (http://www.jb.man.ac.uk/ pulsar/crab.html; Lyne et al. (1993)).
Aqueye and Iqueye have been realized with the support of the Strategic Program Quantum Future of the University of Padova, the Italian Ministry of Research and University MIUR, the Italian Institute of Astrophysics INAF, and the Fondazione Cariparo Padova.

References

Barbieri, C., Naletto, G., Occhipinti, T., et al. 2009a, J. Modern Optics, 56, 261
Barbieri, C., Naletto, G., Verroi, E., et al. 2009b, in Science with the VLT in the ELT Era, Astrophysics and Space Science Proceedings, 249_253
Codogno G., Verroi, E., Naletto, G., Barbieri, C., 2011 *Principles of Multi-wavelength High-Time Resolution Astrophysics*, October 10-15 2011, Pula, Sardinia, Italy
Germaná, C., Zampieri, L., Barbieri, C., et al. 2012, Astron. Astroph., 548, A47
Gouiffes, C. 1998, in Neutron Stars and Pulsars, ed. N. Shibazaki et al. (Tokyo: Univ. Acad. Press), 363

Gradari, S., Barbieri, M., Barbieri, C., et al. 2011, Mon. Not. R. Astron. Soc., 412, 2689
Gregory, P. & Loredo, T. 1992, Astr. J., 398, 146
Hotelling, H. 1933, J. Educational Psychology, 24, 417
Hotelling, H. 1936, Biometrika, 27, 321
Karhunen, K. 1947 *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*. Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys. 37: 1–79.
Larsson, S. 1996, Astron. Astrophys. Suppl. Ser., 117, 197
Leahy, D. 1983, Astr. J., 272, 160
Leahy, D. 1987, Astron. Astrophys., 180, 275
Leahy, D., Elsner, R., & Weisskopf, M. 1983, Astr. J., 272, 256
Loève, M. 1978 *Probability theory*. Vol. II, 4th ed. Graduate Texts in Mathematics 46. Springer-Verlag. ISBN 0-387-90262-7.
Lyne, A., Pritchard, R., & Graham-Smith, F. 1993, Mon. Not. R. Astron. Soc., 265, 1003
Mignani, R. 2011, Adv. Space Res., 47, 1281
Naletto, G., Barbieri, C., Occhipinti, T., et al. 2009, Astr. Astr., 508, 531
Pearson, K. 1901, Philosophical Magazine, 2, 559
Zampieri, L., Čadež, A., Barbieri, C., et al. 2014, Mon. Not. R. Astron Soc., in press