Use of PCA for Gaia

Paul Bartholdi - Geneva Observatory

July 11, 2005

Abstract

Principal Component Analysis, PCA, can be applied successfully to the preliminary analysis of the photometric data provided by Gaia, to monitor any variation in the system response, detect micro variability from system noise, reduce the amount of data to a single pseudo filter before searching for periods, and give a spectral view of the variability.

1 Principal Component Analysis

PCA was developed early in the 19^{th} century to analyze data from human sciences. Its use for astronomical data was not frequent. T. J. Deeming wrote one of the first article on the subject. Peter Lucke applied it very successfully to the reddening law for O stars measured in the Geneva Photometric system. Massa wrote three articles on the use of PCA for photometric data analysis.

In this note, we apply PCA specifically to the detection and quantification of photometric variablilty, being intrinsic or instrumental, from the multicolor photometry of the satellite Gaia

1.1 Introduction to PCA

Let s_{ijk} be the measure of the star S_k through the filter f_j at time t_i, and N_k the size of $\{t_i\}$ which we assume is the same for all filters, but not for all stars. n will be the number of filters.

As we analyse all stars individually, we can drop for simplicity the index k.

Then we define $\overline{s}_j = \frac{1}{N}\sum_i s_{ij}$ and $d_{ij} = s_{ij} - \overline{s}_j$ the deviations from the mean for each filters.

The covariance matrix $C_{lm} = \frac{1}{N}\sum_i d_{il}d_{im}$ which is square semi-positive, that is all eigenvalues are real, positive or null.

Let $0 \leq \nu_1 \leq \nu_2 \leq \nu_3 \ldots$ be the eigenvalues of C, and $U = \{u_1, u_2 \ldots\}$ be the corresponding eigenvectors. They form an orthogonal basis for the d.

We will call $e_{ij} = \sum_k u_{jk} d_{ik}$ the projection of d on u_j.

The covariance matrix of e is diagonal, with the ν_i on the diagonal. As we have ordered the ν_i, the e_{in} have the largest variance (ν_n), and the e_{i1} the smallest, possibly null, one. In practice, no ν_i will be null if $N > n$.

2 Applications for Gaia

We foresee four applications of PCA for Gaia, keeping in mind we are interested in all kind of variability in the measurements.

Our analysis will differ from usual PCA in two ways: 1) we use covariance and not correlation matrices, which allows us to keep the original metric of the data, and 2) we will be interested in both small and large eigenvalues, those representing the minimal variability and the largest one.

2.1 Control of intrinsic noise

The smallest eigenvalue(s) represent(s) the variability that is left after all correlated variations have been removed. They can be used to verify the noise budget as a function of magnitude, spectral type, sky coordinates etc.

2.2 Detection of variability

The ratio $R = \nu_n/\nu_1$ is very sensitive to all kind of variabilities, of instrumental or stellar origin. It can be used to select the variable objects (see 4.1 bellow). As a refinement, R could be evaluated as $R = (\nu_n + \nu_{n-1})/(\nu_1 + \nu_2 + \nu_3 + \nu_4)$

2.3 Principal component for period search

As seen above (1.1), the e_{in} is the linear combination of the original data which contains most of the correlated variance from all filters. It will be easier and sufficient to search for periodicity only on the e_{in}. Assuming that the variability is the same on all filters, certainly an optimistic view, then the signal to noise on e is $= \sqrt{n}$ larger than on any d.

2.4 Spectral representation of the variability

If λ_i is the equivalent wavelength of the filter f_i, then the representation of u_{in} as a function of λ_i is a rough spectrum of the variability. This can be used to assign the type of variable to the object.

The sign of the components of u_j is mathematically undefined. A systematic choice will be defined, for example taking the sign such that the component corresponding to the shortest wavelength is positive, or that some gradient is positive etc.

3 Data from Geneva Photometry

The Geneva Seven Color Photometric catalog contains more than 500 000 measurements for about 50 000 objects. About than 150 stars have been measured more than 100 times and have a standard deviation on m_v and on all colors less then .005 . These stars can be considered as constant. This set is used to test the principles given above. We will ignore the variation in magnitude and keep only the six color indices as data (s_{ij}). Later, known variable stars will be added.

In a later analysis, we will consider known and suspected variables, and keep m_v as well as the color indexes.

3.1 Control of intrinsic noise

The histogram of $\sqrt{\nu_1}$ is rather narrow centred on 0.002, indicating that the measurements are intrinsically better than advertised. It does not depends on R. As a control, the analysis of a stronger variable star (HD 114529) gives the same ν_1 while $= \nu_6$ is more than 15 times greater.

3.2 Detection of variability

For most stars the ratio $R = \nu_6/\nu_1$ is close to 3, but a few stars have $R \geq 7$. Finer analysis indicates that they all correspond to variable objects of very low amplitude.

3.3 Principal component for period search

While period search on any color (or magnitude) did not show any significant frequency, the same analysis done on the e_{in} presents clear peaks out of the noise.

3.4 Spectral representation of the variability

For all stars with $R \leq 4$ (probably not variable objects), the representation u_{i1} as a function of λ_i present the same characteristics, which is

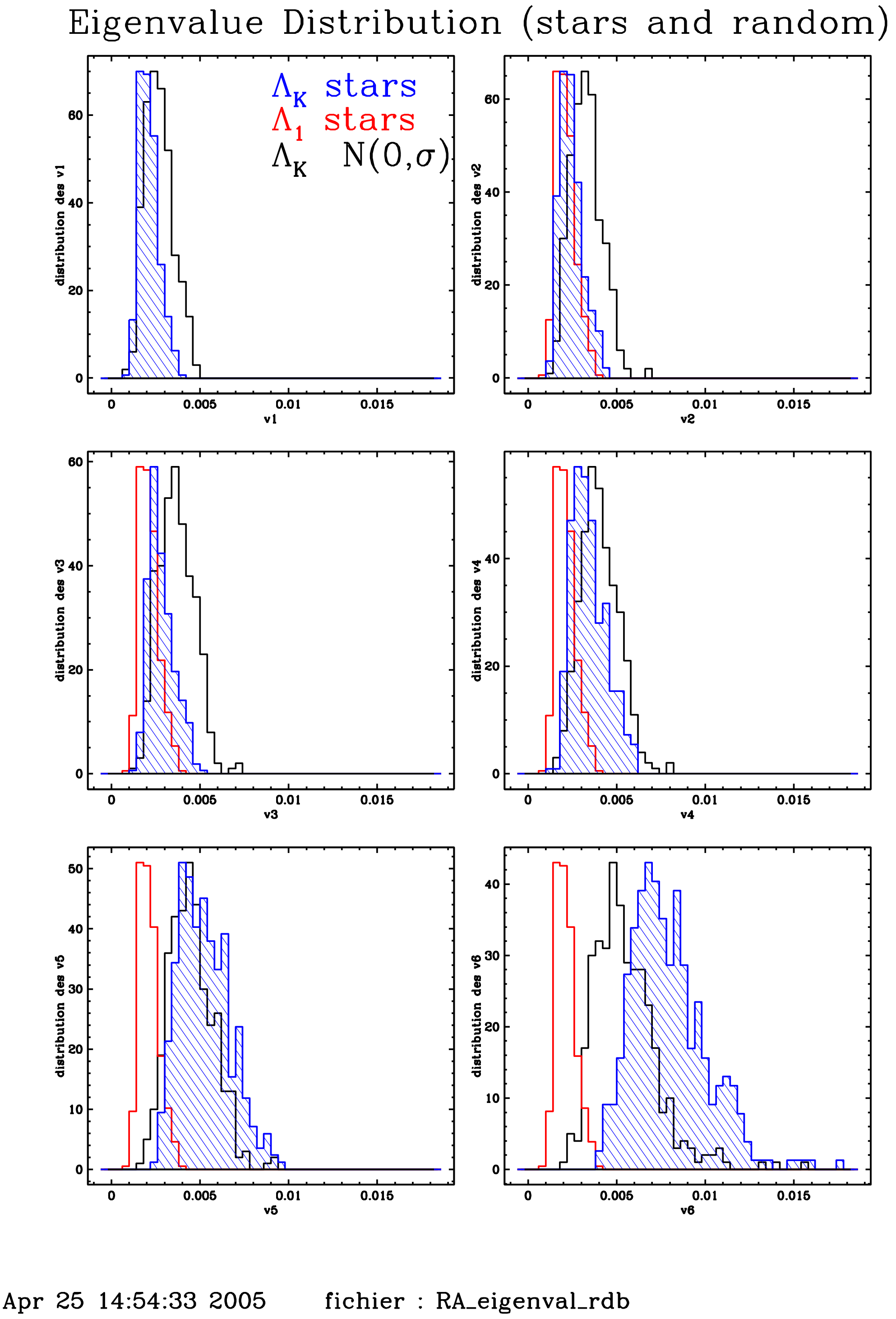

Figure 1: Distribution of Eigenvalues, stars and random measurements

most probably due to residual color effects after the reduction outside the atmosphere.

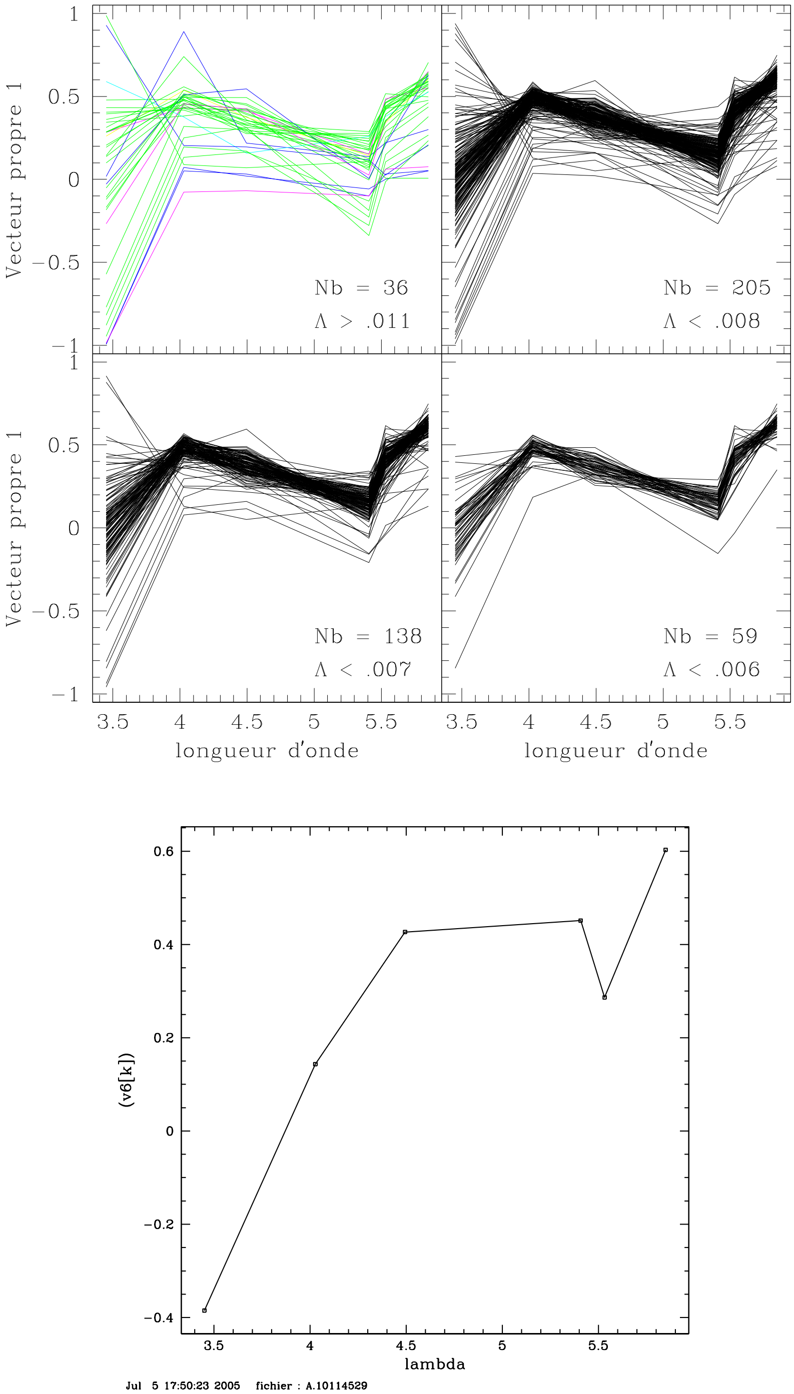

Figure 2: “Spectra” for various R and for HD 114529

For the few stars with $R \geq 7$, the "spectra" are quite different indicating a stellar origin.

3.5 Random simulation

As a comparision template, we simulated observations sith the same variances as the real ones, but with no correlation between filters. The resulting eigenvalues are presented in the figure 3.1 above. The "Spectra" from this simulation are given im figure 3.5.

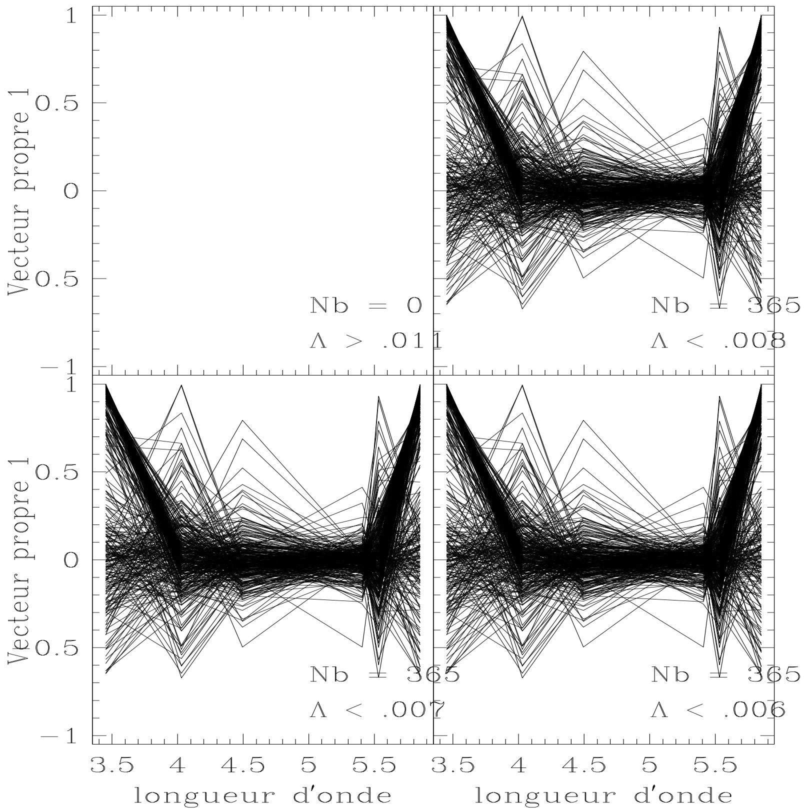

Figure 3: "Spectra" for random measurements

4 Pending work

I see three area were work as to be done:

4.1 R calibration

To be quantitatively useful, the statistics of R should be determined under the following conditions and as a function of N typical for Gaia:

1) The d_j have all the same variance, 2) the d_j have different variances. The figure 4.1 gives the distribution os R for the Geneva Photometry and the random measurements.

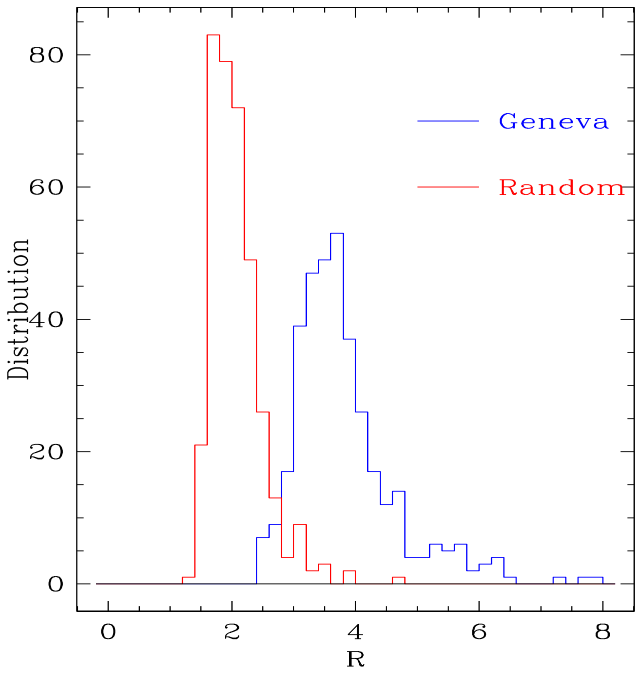

Figure 4: Distribution of R from the Geneva Photometry and from random measurements

4.2 Further Analysis of the Geneva Photometry

The relation between ν_6 and the visual magnitude m_v must be analyzed as the relation with the color index $[B - V]$.

For known variable stars, the number of ν_j significantly above the values for a non variable stars must be established. It probably depends on the type of the variable. For most microvariables, only ν_6 is significant, while for HD 114529 ν_6 and ν_5 are well above normal, but it is the only known variable currently added to our sample.

For the same variable stars, the relation u_i1 as a function of λ_i must be further analyzed to see if it can be used to establish the type of variability.

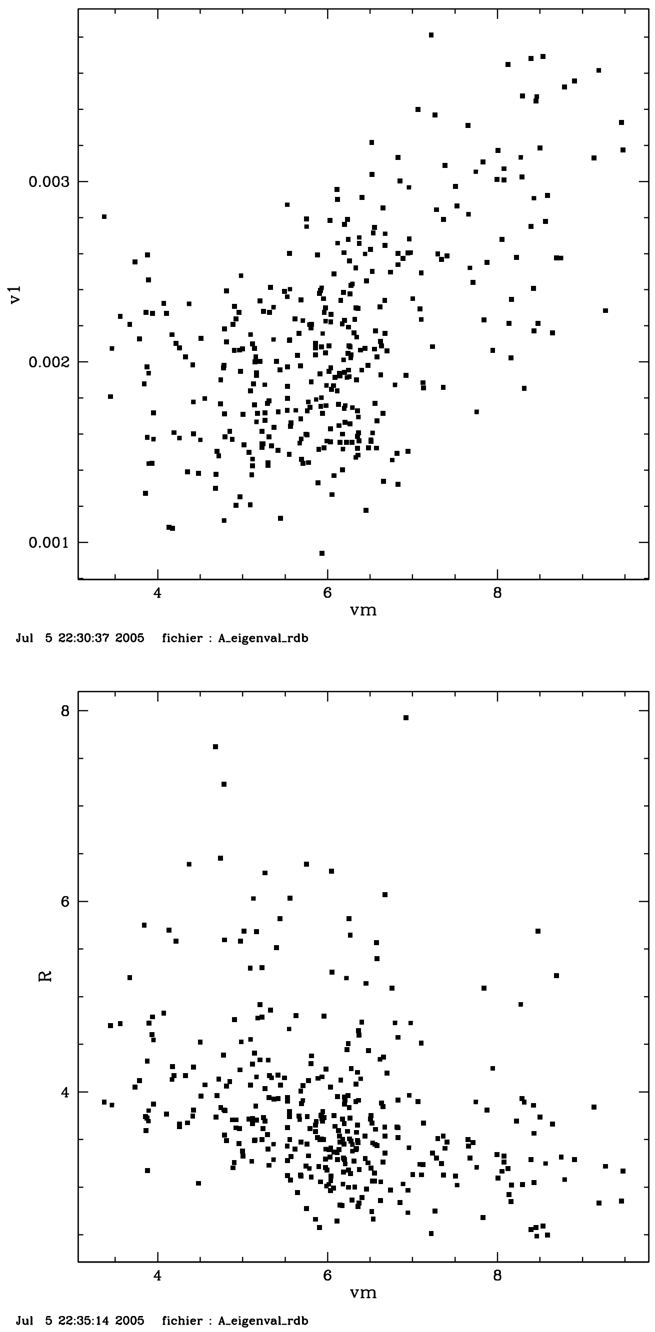

Figure 5: Preliminary relation between R, ν_1 and m_v

4.3 Problems specific for Gaia

Theoretical and/or observational spectra from standard and variable stars should be used with the band pass of the Gaia Photometric System to establish the expected variances for each filter and given spectral type or color index.

The same data for variable stars should provide calibrations for the relation $u_i n$ as a function of λ_i for various variable types.

5 References

Deeming, T. J. Monthly Notices of the Royal Astronomical Society, Vol. 127, p.493.
Lucke, P. B. Astronomy and Astrophysics, vol. 90, no. 3, Oct. 1980, p. 350-354.
Massa, D. Lillie, C. F. Astrophysical Journal, Part 1, vol. 221, May 1, 1978, p. 833-850.
Massa, D. Astronomical Journal, vol. 85, Dec. 1980, p. 1644-1662.