Mon. Not. R. Astron. Soc. **356,** 1466–1470 (2005) doi:10.1111/j.1365-2966.2004.08585.x

Correcting systematic effects in a large set of photometric light curves

O. Tamuz,[1]★ T. Mazeh[1] and S. Zucker[2]†

[1] *School of Physics and Astronomy, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[2] *Observatoire de Genève, 51 Ch. des Maillettes, Sauverny CH-1290, Switzerland*

Accepted 2004 November 1. Received 2004 October 13; in original form 2004 August 16

ABSTRACT

We suggest a new algorithm to remove systematic effects in a large set of light curves obtained by a photometric survey. The algorithm can remove systematic effects, such as those associated with atmospheric extinction, detector efficiency, or point spread function changes over the detector. The algorithm works without any prior knowledge of the effects, as long as they linearly appear in many stars of the sample. The approach, which was originally developed to remove atmospheric extinction effects, is based on a lower rank approximation of matrices, an approach which has already been suggested and used in chemometrics, for example. The proposed algorithm is especially useful in cases where the uncertainties of the measurements are unequal. For equal uncertainties, the algorithm reduces to the Principal Component Analysis (PCA) algorithm. We present a simulation to demonstrate the effectiveness of the proposed algorithm and we point out its potential, in the search for transit candidates in particular.

Key words: atmospheric effects – methods: data analysis – methods: statistical – techniques: photometric – surveys – planetary systems.

1 INTRODUCTION

The advent of large, high signal-to-noise ratio (S/N) CCDs for the use of astronomical studies has driven many photometric projects that have already produced unprecedented large sets of accurate stellar light curves for various astronomical studies. An example of such a project is the Optical Gravitational Lensing Experiment (OGLE) search for transit candidates, which has already yielded significant results (e.g. Udalski et al. 2002). Searching for low-amplitude variables, such as planetary transits, involves finding a faint signal in noisy data. It is therefore of prime interest to remove any systematic effects hiding in the data.

Systematic observational effects may be associated, for example, with the varying atmospheric conditions, the variability of the detector efficiency or point spread function (PSF) changes. However, these effects might vary from star to star, depending on the stellar colour or the position of the star on the CCD, a dependence which is not always known. Therefore, the removal of such effects might not be trivial.

We present here an algorithm to remove some of the systematic effects in a large set of light curves, without any a priori knowledge of the different observational features that might affect the measurements. The algorithm finds the systematics and their manifestation in the individual stars, as long as these effects appear in many light curves.

We started the development of our algorithm in an attempt to correct for the atmospheric extinction, with an approach similar to that of Kruszewski & Semeniuk (2003). We derived the best-fitting airmasses of the different images and the extinction coefficients of the different stars, without having any information on the stellar colours. However, the result is a general algorithm that deals with linear systematic effects. It turned out that such an algorithm had already been proposed by Gabriel & Zamir (1979), who had applied it to data from disciplines other than astronomy, chemometrics, for example. In some restricted cases, when one can ignore the different uncertainties of the data points, this algorithm reduces to the well-known Principal Component Analysis (PCA; Murtagh & Heck 1987, chapter 2). However, when the uncertainties of the measurements vary substantially, as in many photometric surveys, the PCA performs poorly, as we demonstrate below.

In Section 2 we present the initial, simpler version of our algorithm, which was meant solely to remove the colour-dependent atmospheric extinction. In Section 3 we put the algorithm in a broader context, and show how the algorithm can remove linear systematic effects, and can even treat several unknown effects. In Section 4 we present a simulation to demonstrate the effectiveness of our algorithm. We discuss some of the algorithm properties and potential developments in Section 5.

2 CORRECTION FOR ATMOSPHERIC EXTINCTION

The colour-dependent atmospheric extinction is an obvious observational effect that contaminates ground-based photometric

★E-mail: omert@wise.tau.ac.il
†Present address: Faculty of Physics, Weizmann Institute of Science, Rehovot 76100, Israel.

measurements. This effect depends on the stellar colours, which are not always completely known. This is especially true for photometric surveys when only one filter is used and no explicit colour information is available. In this section we describe how we find the best stellar extinction coefficients to account for the atmospheric absorption, together with the most suitable airmasses assigned for each image.

Consider a set of N light curves, each of which consists of M measurements. We define the residual of each observation, r_{ij}, to be the average-subtracted stellar magnitude, i.e. the stellar magnitude after subtracting the average magnitude of the individual star.

Let $\{a_j;\ j = 1, \ldots, M\}$ be the airmass at which the jth image was observed. We can then define the effective extinction coefficient c_i of star i to be the slope of the best linear fit for the residuals of this star – $\{r_{ij};\ j = 1, \ldots, M\}$ as a function of the corresponding airmasses – $\{a_j;\ j = 1, \ldots, M\}$. We aim to remove the product $c_i a_j$ from each r_{ij}. In fact, we search for the best c_i that minimizes the expression

$$S_i^2 = \sum_j \frac{(r_{ij} - c_i a_j)^2}{\sigma_{ij}^2}, \tag{1}$$

where σ_{ij} is the uncertainty of the measurement of star i in the image j.

Assuming the airmasses are known, a simple differentiation and equating to zero yields an estimate for the extinction coefficient:

$$c_i = \frac{\sum_j \left(r_{ij} a_j / \sigma_{ij}^2\right)}{\sum_j \left(a_j^2 / \sigma_{ij}^2\right)}. \tag{2}$$

Note that the derivation of each c_i is independent of all the other c_i, but does depend on all the $\{a_j\}$.

The problem can now be turned around. Because atmospheric extinction might depend not only on the airmass but also on weather conditions, we can ask ourselves what is the most suitable 'airmass' of each image, given the known coefficient of every star. Thus, we can look for the a_j that minimizes

$$S_j^2 = \sum_i \frac{(r_{ij} - c_i a_j)^2}{\sigma_{ij}^2}, \tag{3}$$

given the previously calculated set of $\{c_i\}$. The value of the effective 'airmass' is then

$$a_j^{(1)} = \frac{\sum_i \left(r_{ij} c_i / \sigma_{ij}^2\right)}{\sum_i \left(c_j^2 / \sigma_{ij}^2\right)}. \tag{4}$$

We can now recalculate the best-fitting coefficients, $c_i^{(1)}$, and continue iteratively. We thus have an iterative process which in essence searches for the two sets – $\{\bar{c}_i\}$ and $\{\bar{a}_j\}$ – that best account for the atmospheric extinction.

We have performed many simulations that have shown that this iterative process converged to the same $\{\bar{a}_j\}$ and $\{\bar{c}_i\}$, no matter what initial values were used. Therefore, we suggest that the proposed algorithm can find the most suitable airmass of each image and the extinction coefficient of each star. As the next section shows, these airmasses and coefficients may have no relation to actual airmass and colour.

3 GENERALIZATION

The algorithm presented in the previous section is in fact a search to find the best two sets of $\{c_i;\ i = 1, N\}$ and $\{a_j;\ j = 1, M\}$ that minimize the global expression

$$S^2 = \sum_{ij} \frac{(r_{ij} - c_i a_j)^2}{\sigma_{ij}^2}. \tag{5}$$

Therefore, although the alternating 'criss-cross' iteration process (Gabriel & Zamir 1979) started with the actual airmasses of the different images, the values of the final set of parameters $\{\bar{a}_j\}$ and $\{\bar{c}_i\}$ are not necessarily related to the true airmass and extinction coefficient. They are merely the variables by which the global sum of residuals, S^2, varies linearly most significantly. They could represent any strong systematic effect that might be associated, for example, with time, temperature or position on the CCD. This algorithm finds the systematic effect as long as the global minimum of S^2 is achieved.

Now, suppose the data include a few different systematic effects, with different $\{c_i\}$ and $\{a_j\}$. We can easily generalize the algorithm to treat such a case. To do that, we denote by $\{^{(1)}\bar{c}_i\}$ and $\{^{(1)}\bar{a}_j\}$ the first set of parameters found in the data. We then remove this effect and denote the new residuals by

$$^{(1)}r_{ij} = r_{ij} - {}^{(1)}\bar{c}_i\,{}^{(1)}\bar{a}_j. \tag{6}$$

We can then proceed and search for the next linear effect, hidden in $\{^{(1)}r_{ij}\}$. We use the same procedure to find now the $^{(2)}c_i$ and $^{(2)}a_j$ that minimize

$$^{(1)}S^2 = \sum_{ij} \frac{\left(^{(1)}r_{ij} - {}^{(2)}c_i\,{}^{(2)}a_j\right)^2}{\sigma_{ij}^2}. \tag{7}$$

This process can be applied repeatedly, until it finds no significant linear effects in the residuals. The algorithm finds any linear systematic effect that can be presented as $c_i a_j$ for the jth measurement of the ith star.

After developing our algorithm, we found that such an approach had already been proposed by Gabriel & Zamir (1979) as a lower-rank approximation to data matrices. They applied the algorithm to data from other disciplines, such as climate statistics and chemometrics, and discussed its convergence properties. Very similar algorithms were developed and applied for signal and image processing (e.g. Lu, Pei & Wang 1997). If all measurements have the same uncertainties, the algorithm will reduce to the conventional PCA that can be applied through the singular value decomposition (SVD) technique (Press et al. 1992, chapter 2). However, when the uncertainties of the measurements are substantially different, the PCA becomes less effective at finding and removing systematic effects. This can lead to the removal of true variability, and can leave some actual systematic effects in the data.

4 SIMULATION

To demonstrate the power of the algorithm, we present here one out of the many simulations we ran. In this specific example we simulated light curves of 3000 stars in 1000 images. All stars were set to have constant magnitudes with normally distributed noise of various amplitudes. We added three systematic effects that depended on airmass, the CCD position and lunar phase. Finally, we added transit-like light curves to three stars.

To simulate a realistic set of light curves, we assigned different noise levels to different stars, as if we had bright (high S/N) and faint (low S/N) objects. The rms ranges between 0.01 and 1 mag, with an average value of 0.3 mag. We assigned to each measurement an uncertainty which depended on the stellar standard deviation. To avoid an unrealistic case where all measurements of a star have

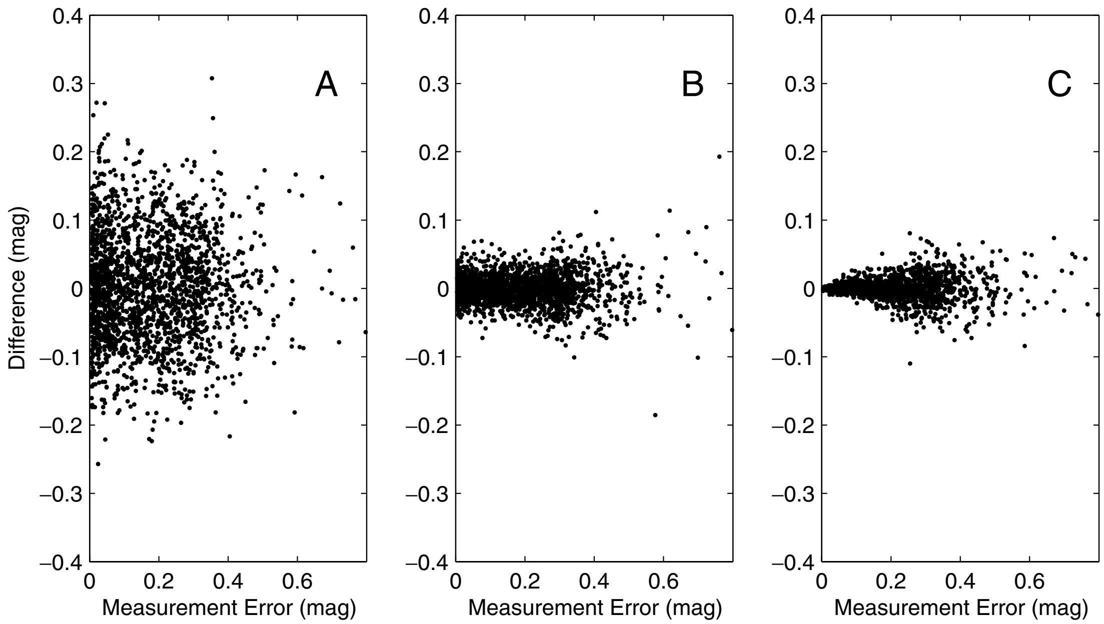

Figure 1. The differences between the derived measurements and their original values (without the systematic effects). Panel A presents the measurements before any correction was applied, panel B after applying the PCA and panel C after applying the proposed algorithm. Data are plotted as a function of the stellar rms; 2000 randomly chosen measurements are plotted.

the same uncertainty, we randomly varied the uncertainty of each measurement by 10 per cent of the stellar rms

We selected three light curves with 0.01 mag rms and added to them planetary transit-like signals. The transit periods were 2.7183, 3.1415 and 1.4142 d, with depths of 15, 20 and 25 mmag, respectively.

We added three systematic effects that depended on airmass (linear and quadratic effects), on the CCD *X*-position and on the lunar phase. Thus,

$$r_{ij} = c_i(am)_j + d_i(am)_j^2 + x_i b_j + f_i \sin(\omega_{\mathrm{lunar}} t_j) + \text{Poisson noise [+Transit]} \qquad (8)$$

where

(i) the observation times $\{t_j\}$ were set to the times of the first 1000 images of the OGLE Carina field survey, available from the OGLE website;[1]

(ii) the airmasses $\{(am)_j\}$ were calculated using these times and the OGLE Carina survey parameters;

(iii) the positions on the CCD $\{x_i\}$ were randomly drawn from a uniform distribution, between 0 and 2047; and

(iv) the coefficients $\{c_i\}$, $\{d_i\}$, $\{b_j\}$ and $\{f_i\}$ were randomly drawn from normal distributions of zero mean. The standard deviations of these distributions were chosen so that the four systematic effects produced rms variabilities of 0.06, 0.04, 0.01 and 0.008, respectively.

We applied our algorithm to the simulated artificial survey four successive times, to eliminate four different linear effects (see Section 5 for a discussion of the number of effects to subtract). For comparison, we applied the PCA subtraction to the same data set the same number of times.

The efficacy of the algorithm is demonstrated in two figures. Fig. 1 presents randomly selected 2000 measurements, before and after the two algorithms were applied. Panel A shows the difference between the magnitudes before and after the systematic effects were added. This difference, which is actually the exact amount added by the systematic effects, is plotted as a function of the rms of each star. We see that the typical systematic error is about 0.1–0.2 mag. For the ‘faint’ stars, with rms of about 0.4–0.7 mag, the systematic error is relatively small. However, for the ‘bright’ stars in the sample, with inherent rms smaller than, say, 0.05 mag, the additional systematic error is relatively large. This added noise can seriously hamper the ability to detect small effects such as planetary transits.

Panel B shows the systematic effects left after the PCA was applied four times, again as a function of the stellar rms. Had the PCA worked perfectly, all differences would have been nullified, and all points in the diagram would have been concentrated on the horizontal line that goes through zero. We can see that the PCA managed to correct all the systematic effects larger than 0.05 mag, but failed to perform for smaller systematic errors.

Panel C of Fig. 1 presents the same 2000 measurements, this time after applying our algorithm four times. We see that the ability of the algorithm to remove the systematic error depends strongly on the stellar inherent rms. The algorithm performs substantially better when the stellar rms is small. For those stars, the advantage of our algorithm over the PCA approach seems clear. In fact, all natural candidates for transit detection are exactly those stars.

The ability to detect transits is depicted in Fig. 2, where we focus on the three stars with simulated transits. Each column presents the stellar folded light curve before and after the successive iterations were applied. The data were folded with the transit period, and were plotted around the mid-transit phase. While initially the systematic errors completely masked the transits, the three of them gradually surfaced as more iterations were applied.

5 DISCUSSION

The proposed algorithm reduces to the PCA approach for the case of equal uncertainties. It is therefore suggested to explore the features

[1] See http://siruis.astrouw.edu.pl/∼ogle.

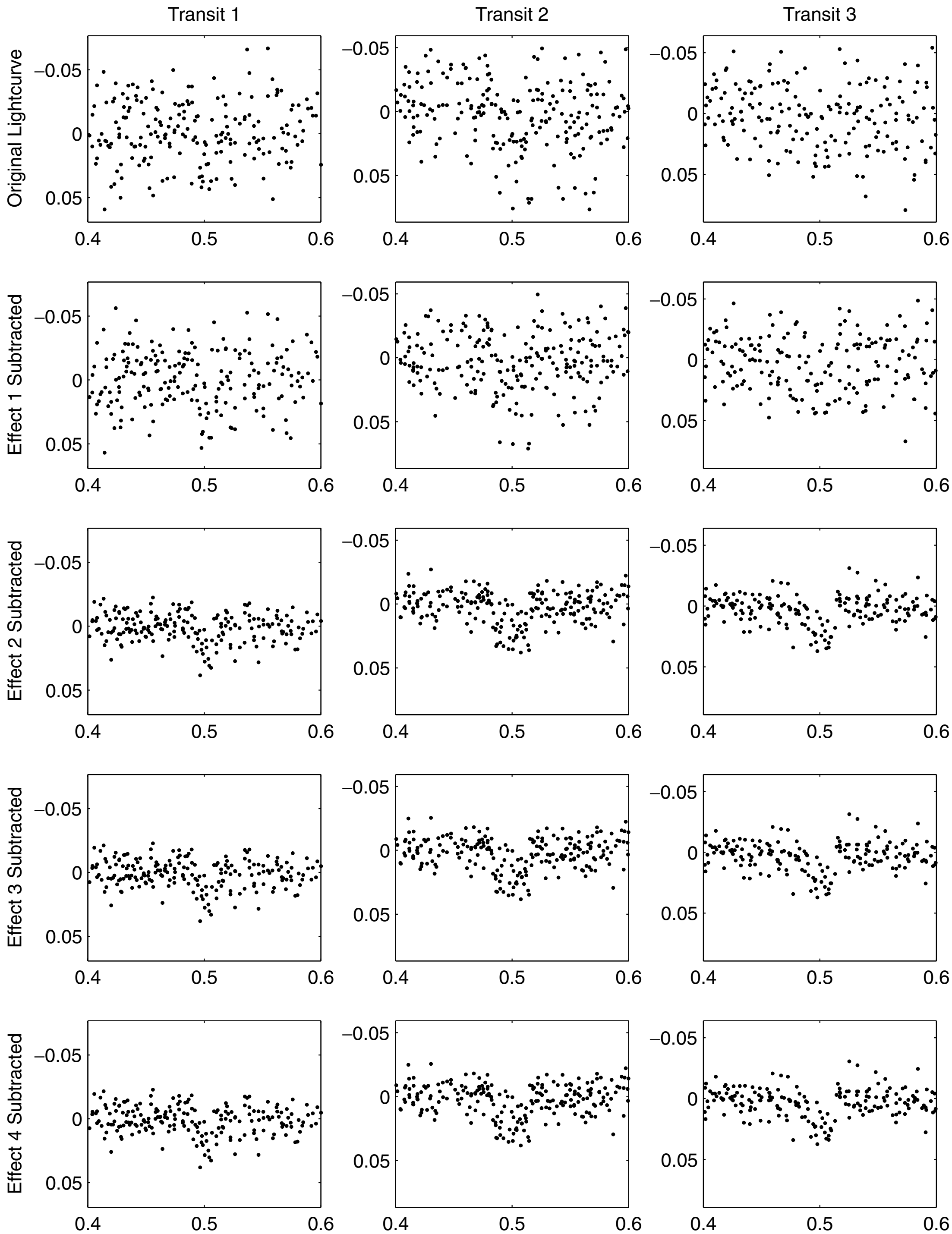

Figure 2. Folded light curves of the three transits planted in the data before and after the few first iterations. Each column depicts the light curves of one transit. The top row shows the data uncorrected, while the following rows show it after successive applications of the algorithm. The light curves are folded on the transit periods and show the points which lie within 0.1 phases of the middle of the transit.

of our algorithm by analogy with the corresponding features of the PCA.

In case of equal uncertainties, the vectors $\{a_j;\ j = 1, \ldots, M\}$ are the eigenvectors of the covariance matrix $\mathbf{R}^{\mathrm{T}}\mathbf{R}$, where $\mathbf{R}$ is the measurement matrix $\{r_{ij};\ i = 1, \ldots, N,\ j = 1, \ldots, M\}$. Because $\mathbf{R}^{\mathrm{T}}\mathbf{R}$ is symmetric, $\{a_j;\ j = 1, \ldots M\}$ constitute an orthogonal set of vectors. The first few a_j are therefore an orthogonal base that spans the vector subspace of the significant systematic effects. Thus, it may very well happen that the strongest effect the PCA derives is a linear combination of some effects we know from prior physical insight, such as a certain combination of the airmass and the X position on the CCD chip. Conversely, it may so happen that two effects about which we have some insight, such as the Y position and the lunar phase, span a vector subspace which includes much of the power of a third effect, say the airmass. In this case, the PCA derives only two significant effects, contrary to our prior intuition.

We suggest that our algorithm exhibits similar behaviour. It is true that in the general case of unequal uncertainties the orthogonality of $\{a_j\}$ is not guaranteed, but the same qualitative behaviour probably persists.

The recent large photometric surveys and the planned photometric space missions (e.g. *CoRoT*, *Kepler*) will face not only the problem of systematic effects, but also the problem of long-term stellar variability. It turns out that the proposed solution can potentially remove some of this variability. In this case, the various $\{a_j\}$ would assume the values of some function, f, of the timing of the jth image: $a_j = f(t_j)$. For equal uncertainties, the space of possible time variability can be spanned by an orthogonal basis of functions (e.g. trigonometric functions in the case of evenly spaced time sampling). From the PCA point of view, these basis functions (such as $\{^{(k)}a_j = \cos(\omega_{(k)}t_j); k = k_1, k_2 \ldots\}$) can be thought of as systematic effects. The contribution of each basis function to the individual stars is reflected through the stellar coefficients, $^{(k)}c_i$. Removing these effects amounts to removing part of the power of the long-term variability. Once again, the general case probably shows similar behaviour.

In general, the main use of the PCA has been to reduce the dimensionality of the data by finding only the significant factors. Thus, an important question in the PCA (Murtagh & Heck 1987, chapter 2) is the number of significant factors to retain. In the PCA, it is easy to solve for the complete set of effects (all eigenvectors of the covariance matrix), and then to decide about the significant factors. In the general case of unequal uncertainties, we can proceed in two alternative ways. One way is to solve simultaneously for an assumed number of effects. The other alternative is to solve for the effects in sequential stages. In each stage, we subtract the effects found in previous stages before solving for a new effect. The two alternatives are equivalent in the PCA case (equal uncertainties), but in the general case they lead to different solutions. Moreover, subtracting the effect in each stage opens up the possibility of subtracting the effect not globally, but only from the stars which are most affected by it. We plan to further explore these issues in order to gain more insight into the features of the solution.

We are currently applying the algorithm presented here to parts of the OGLE III data set. We have already found a few intriguing new planetary transit candidates, and we are still evaluating the statistical significance of these findings. It would be of great interest to apply our algorithm to space mission data, such as *Hubble Space Telescope* photometry, to find out how large the systematic effects hidden in the data are. As we have demonstrated, the advantages of our algorithm are most pronounced in a data set of high S/N measurements with substantially varying uncertainties. Data from space missions exactly fit this description.

ACKNOWLEDGMENTS

This work was supported by the Israeli Science Foundation through grant no. 03/233. SZ wishes to acknowledge support by the European RTN 'The Origin of Planetary Systems' (PLANETS, contract number HPRN-CT-2002-00308) in the form of a fellowship.

REFERENCES

Gabriel K. R., Zamir S., 1979, Technometrics, 21, 489
Kruszewski A., Semeniuk I., 2003, Acta Astron., 53, 241
Lu W. S., Pei S. C., Wang P. H., 1997, IEEE Trans. Circ. Syst. I, 44, 650
Murtagh F., Heck A., 1987, Astrophys. Space Sci. Library, Multivariate Data Analysis. Reidel, Dordrecht.
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in C. The Art of Scientific Computing, 2nd edn. Cambridge Univ. Press, Cambridge
Udalski A. et al., 2002, Acta Astron., 52, 1

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