Mon. Not. R. Astron. Soc. **362,** 460–468 (2005) doi:10.1111/j.1365-2966.2005.09305.x

Fast identification of transits from light-curves

Pavlos Protopapas,[1]* Raul Jimenez[2] and Charles Alcock[1]

[1] *Harvard Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA*
[2] *Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*

Accepted 2005 June 16. Received 2005 June 8; in original form 2005 March 2

ABSTRACT

We present an algorithm that allows fast and efficient detection of transits, including planetary transits, from light-curves. The method is based on building an ensemble of fiducial models and compressing the data using the MOPED compression algorithm. We describe the method and demonstrate its efficiency by finding planet-like transits in simulated Panoramic Survey Telescope & Rapid Response System (Pan-STARRS) light-curves. We show that our method is independent of the size of the search space of transit parameters. In large sets of light-curves, we achieve speed-up factors of the order of 10^3 times over an optimized adaptive search in the χ^2 space. We discuss how the algorithm can be used in forthcoming large surveys like Pan-STARRS and the Large Synoptic Survey Telescope (LSST), and how it may be optimized for future space missions like *Kepler* and *COROT* where most of the processing must be done on board.

Key words: techniques: photometric – binaries: eclipsing – planetary systems.

1 INTRODUCTION

If the orbit of a planet around a star is so favourably inclined that $\sin(i) \approx 1$, the planet will transit the disc of the star once per orbit. During the transit, the observed flux from the star is reduced by the ratio of the areas of the planet and the star, typically ~1 per cent for a Jupiter-like planet around a Sun-like star. When this photometric dimming is observed to repeat periodically, a small-radius companion may be inferred to exist. This effect is seen in star HD 209458 (Charbonneau et al. 2000), which was first identified as a planetary system using the radial velocity technique. The added value of the detection of transits is significant: not only is the $\sin(i)$ ambiguity resolved, but the radius of the planet may be inferred, and spectroscopic examination of the object during transit allows the study of the atmosphere of the planet (Charbonneau et al. 2002).

The transit technique to search for planets has some advantages: photometry is less costly in telescope time than spectroscopy, and one knows $\sin(i)$ for all the systems found this way. The major disadvantage is that the yield is comparatively low, since only systems with $\sin(i) \approx 1$ will be detected.

A large number of transit searches for extrasolar planets, both space-based and ground-based, have been completed or are under way (Gilliland et al. 2000; Mochejska et al. 2002; Udalski et al. 2002, 2003; Mallén-Ornelas et al. 2003). Many of these efforts employ small-aperture, wide-field cameras to monitor tens of thousands of nearby, bright stars. Of the surveys using this approach, the only success to date has come from the Trans-Atlantic Exoplanet Survey (TrES), which recently announced the discovery of a planet dubbed TrES-1 (Alonso et al. 2004). The only other success (albeit for fainter stars where follow-up is more difficult) has come from the OGLE survey (Udalski et al. 2002, 2003). The vast majority of transits have been false detections resulting from grazing transits of stellar companions or a blend of an eclipsing binary with a brighter foreground or background star (Torres et al. 2004; Pont et al. 2005). Some progress has been made at differentiating these from planets (Hoekstra, Wu & Udalski 2005). However, the few candidates not eliminated by follow-up studies, in particular OGLE-TR-56b with its 1.2-d orbital period, further challenge our already revised models of planet formation (Konacki et al. 2003).

The detection of a weak, short, periodic transit in noisy light-curves is a challenging task. The large number of light-curves collected makes automation and optimization processes a necessity. This requirement is even stronger in the context of space missions, where much of the processing must be done on board. A number of transit detection algorithms have been implemented in the literature (Doyle et al. 2000; Defaÿ, Deleuil & Barge 2001; Aigrain & Favata 2002; Jenkins, Caldwell & Borucki 2002; Kovács, Zucker & Mazeh 2002; Udalski et al. 2002; Street et al. 2003; Drake & Cook 2004) and there has been some effort to compare their respective performances (Tingley 2003).

Such transit searches are generally performed by comparing light-curves to a family of models with a common set of parameters: the transit period T, the transit duration η, the epoch τ (which is equal to the time t at the start of the first transit) and the transit depth θ. The best set of parameters is identified by finding the model most likely to have given rise to the observed data, i.e. the model with the highest likelihood L. This is exactly the kind of problem the

[*] E-mail: pprotopapas@cfa.harvard.edu

Massively Optimized Parameter Estimation and Data compression algorithm (MOPED) (Heavens, Jimenez & Lahav 2000) was designed to address. In particular, light-curves contain plenty of redundant information: the light between transits. By using MOPED one can weight more the part of the light-curve that is sensitive to the transit, thus constructing one eigenvector for each of the parameters in the transit model. However, for the case of transit detection in light-curves, the MOPED eigenvectors are sensitive to the fiducial model, and thus MOPED incorrectly overweights some data. In this paper we present a solution to this problem by building an ensemble of fiducial models. We find that, for each model in an ensemble of fiducial models, there are many possible solutions. However, only one solution is common to all models in the ensemble of fiducial models: the one with the correct parameter values of the transit. We construct a new statistical measure to determine for the set of fiducial models the correct value of the parameters for the transit. We also show that our algorithm passes the null test, i.e. it correctly identifies a light-curve with no transit. The set of fiducial models can be pre-computed, and we provide a recipe to do this. We show that this needs to be done only once before the search for transits is performed in a set of light-curves.

The speed-up in the analysis is significant. For a simulated light-curve typical of the the Panoramic Survey Telescope & Rapid Response System (Pan-STARRS), we find that our algorithm is 10^3 times faster than an optimized adaptive search in the χ^2 space. The speed-up is due to the fact that, using MOPED, the maximum-likelihood search is performed on four data (the number of parameters) instead of thousands, and that the ensemble of fiducial models can be pre-computed. This achieved increase in speed to compute the likelihood is important for transit analysis since the likelihood surface has *multiple* maxima, of which only one is the desired solution, and therefore the search for this best solution needs to explore the whole likelihood surface.[1]

This paper is organized as follows. In Section 2, we briefly describe MOPED. Section 3 presents the transit model used and how a set of synthetic light-curves were constructed. In Section 4, we describe the extension of MOPED using an ensemble of fiducial models, and we also present how the results should be compared to the null hypothesis. Results are discussed in Section 5, and our conclusions summarized in Section 7. In Section 6, we describe the numerical topics, including a numerical recipe.

2 MOPED

We briefly review the parameter estimation and data compression method MOPED, which was originally described in Heavens et al. (2000). The method is as follows: Given a set of data $\boldsymbol{x}$ (in our case a light-curve) that includes a signal part $\boldsymbol{\mu}$ and noise $\boldsymbol{n}$, i.e.

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{n}, \tag{1}$$

the idea then is to find weighting vectors $\boldsymbol{b}_m$, where m runs from 1 to the number of parameters M, such that

$$y_m \equiv \boldsymbol{b}_m \boldsymbol{x} \tag{2}$$

contain as much information as possible about the parameters (period, duration of the transit, etc.). These *numbers* y_m are then used as the data set in a likelihood analysis, with the consequent increase in speed at finding the best solution. In MOPED, there is one vector associated with each parameter.

In Heavens et al. (2000) an optimal and lossless method was found to calculate $\boldsymbol{b}_m$ for multiple parameters (as is the case with transits). The definition of 'lossless' here is that the Fisher matrix at the maximum-likelihood point is the same whether we use the full data set or the compressed version. The Fisher matrix is defined by

$$\boldsymbol{F}_{\alpha\beta} \equiv -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha\, \partial\theta_\beta} \right\rangle, \tag{3}$$

where the average is over an ensemble with the same parameters (θ_α; θ_β) but different noise. The a posteriori probability for the parameters is the likelihood, which for Gaussian noise is

$$\mathcal{L}(\theta_\alpha) = \frac{1}{(2\pi)^{N/2}\sqrt{\det(\boldsymbol{C})}} \times \exp\left[-\frac{1}{2}\sum_{i,j}(x_i - \mu_i)\boldsymbol{C}^{-1}_{ij}(x_j - \mu_j)\right]. \tag{4}$$

The Fisher matrix gives a good estimate of the errors on the parameters, provided the likelihood surface is well described by a multivariate Gaussian near the peak. The method is strictly lossless in this sense provided that the noise is independent of the parameters, and provided our initial guess of the parameters is correct. This is not exactly true because our initial guess is inevitably wrong. However, the increase in parameter errors is very small in these cases (see Heavens et al. 2000) – MOPED recovers the correct solutions extremely accurately even when the conditions for losslessness are not satisfied. The weights required are

$$\boldsymbol{b}_1 = \frac{\boldsymbol{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}^t_{,1}\boldsymbol{C}^{-1}\boldsymbol{\mu}_{,1}}} \tag{5}$$

and

$$\boldsymbol{b}_m = \frac{\boldsymbol{C}^{-1}\boldsymbol{\mu}_{,m} - \sum_{q=1}^{m-1}\left(\boldsymbol{\mu}^t_{,m}\boldsymbol{b}_q\right)\boldsymbol{b}_q}{\sqrt{\boldsymbol{\mu}^t_{,m}\boldsymbol{C}^{-1}\boldsymbol{\mu}_{,m} - \sum_{q=1}^{m-1}\left(\boldsymbol{\mu}^t_{,m}\boldsymbol{b}_q\right)^2}} \qquad (m > 1), \tag{6}$$

where a comma denotes the partial derivative with respect to the parameter m, and $\boldsymbol{C}$ is the covariance matrix with components $\boldsymbol{C}_{ij} = \langle n_i n_j \rangle$, with i and j running from 1 to the size of the data set. To compute the weight vectors requires an initial guess of the parameters. We term this the fiducial model ($\boldsymbol{q}_{\rm f}$) and we discuss in Section 6 the impact on the MOPED solution of the choice of the fiducial model. For the case of transits, $\boldsymbol{C}$ does not depend on the parameters and therefore the $\boldsymbol{b}_m$ depend only on the fiducial parameters ($\boldsymbol{q}_{\rm f}$). On the other hand, $\boldsymbol{\mu}$ represents the signal part and thus depends on the free parameters, which we denote by $\boldsymbol{q}$.

The data set $\{y_m\}$ is orthonormal: i.e. the y_m are uncorrelated, and of unit variance. The y_m have means

$$\langle y_m \rangle = \boldsymbol{b}_m(\boldsymbol{q}_{\rm f}) \cdot \boldsymbol{\mu}(\boldsymbol{q}). \tag{7}$$

The new likelihood is easy to compute, namely,

$$\ln \mathcal{L}(\theta_\alpha) = \text{constant} - \sum_{m=1}^{M}\frac{(y_m - \langle y_m \rangle)^2}{2} = \text{constant} - \sum_m [\boldsymbol{b}_m(\boldsymbol{q}_{\rm f}) \cdot \boldsymbol{x} - \boldsymbol{b}_m(\boldsymbol{q}_{\rm f}) \cdot \boldsymbol{\mu}(\boldsymbol{q})]^2. \tag{8}$$

Further details are given in Heavens et al. (2000).

It is important to note that, if the covariance matrix is known for a large data set (e.g. a large synoptic survey) or it does not change significantly from light-curve to light-curve, then the $\langle y_m \rangle$ need be

[1] In surveys with high cadence and short observational period (e.g. TrES), the likelihood surface is smooth and methods utilizing smart searches of the likelihood surface are better suited.

computed only *once* for the whole data set, thus massively speeding up the computing of the likelihood.

3 TRANSIT MODEL AND SYNTHETIC LIGHT-CURVES

3.1 Transit model

For the transit analysis, we have constructed a model, $\boldsymbol{\mu}$, that closely represents the shape of a planetary transit light-curve. An obvious and usually chosen approach is to use a square wave: $\mu(t) = -1$ for $-1 < t < 1$ and $\mu(t) = 0$ otherwise. However, in order to allow for softer edges and being analytically differentiable, we used the following function:

$$\mu(t; T, \eta, \theta, \tau) \\ = \text{constant} + \tfrac{1}{2}\theta\left\{2 - \tanh\left[c\left(t' + \tfrac{1}{2}\right)\right] + \tanh\left[c\left(t' - \tfrac{1}{2}\right)\right]\right\}, \tag{9}$$

where t' is given by

$$t'(T, \eta, \tau) = \frac{T \sin[\pi(t - \tau)/T]}{\pi\eta}, \tag{10}$$

T is the period, τ is the epoch, η is the transit duration, θ is the depth of the transit and c is a constant.[2]

Applying the transit model to the MOPED framework, one needs to calculate the $\boldsymbol{b}$ weight vectors (equation 6), which depend on the derivatives of the model $\boldsymbol{\mu}$ [the derivatives of equation (9) with respect to the four parameters T, θ, η and τ]. These derivatives can be analytically calculated and thus are computationally inexpensive since they do not require conditional statements.

3.2 Synthetic light-curves

In order to test our method and estimate the gain in speed, we created a sample of synthetic light-curves by setting the four free parameters to realistic values and generating magnitudes according to equation (9) with Gaussian noise added to simulate real light-curves better. We adjusted the Gaussian noise to achieve desirable signal-to-noise ratio (S/N) values.

We simulated observational sampling patterns from Pan-STARRS (one observation every 10 min, four times a month) and generated magnitudes as described in the equation

$$x(t_i; T, \eta, \theta, \tau) = \mu(t_i; T, \eta, \theta, \tau) + \boldsymbol{n}_i, \tag{11}$$

where t_i are the observational times and $\boldsymbol{n}_i$ is a Gaussian noise obtained from Pan-STARRS photometric accuracy of 0.01 mag. Fig. 1 (top panel) shows a typical synthetic light-curve with period 1.3 d and S/N = 5.

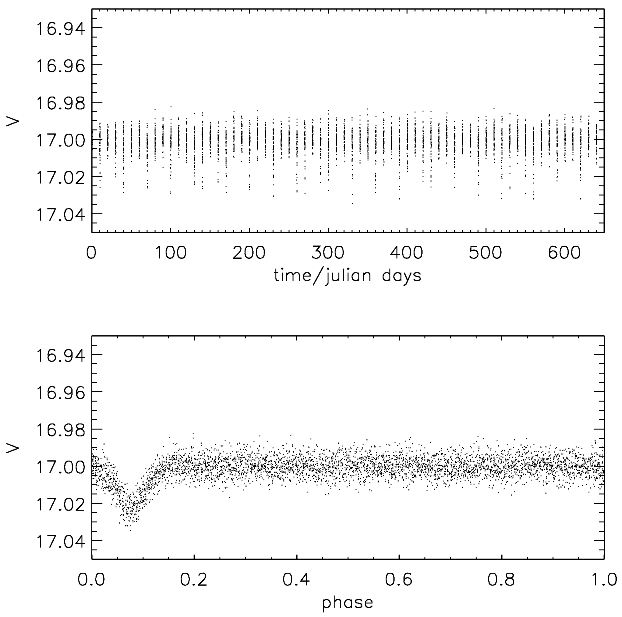

Figure 1. Top: Synthetic light-curve with transit signal: S/N ~ 5, period $T = 1.3$ d, transit duration of $\eta = 0.1$ d, transit depth $\theta \sim 0.01$. Bottom: Same synthetic light-curve folded with the right period $T = 1.3$ d. The folded light-curve has not been used in any of the analysis and it is only shown here for demonstration.

4 EXTENSION TO MOPED USING AN ENSEMBLE OF FIDUCIAL MODELS

Unlike the case of galaxy spectra (Heavens et al. 2000), the fiducial model will weight some data high, very erroneously if the fiducial model is way off from the true model. This is because the derivatives of the fiducial model with respect to the parameters are large near the walls of the box-like shape of the model.

In this section we present an alternative approach to find the best-fitting transit model to a light-curve. The method is based on using an ensemble of randomly chosen fiducial models. For an arbitrary fiducial model, the likelihood function (equation 8) will have several maxima, one of which is guaranteed to be the correct solution. This is the case where the values of the free parameters ($\boldsymbol{q}$) are close to the true one; thus $\boldsymbol{\mu}(\boldsymbol{q})$ in equation (8) is similar to $\boldsymbol{x}$. For a different arbitrary fiducial model, there are also several maxima, but only one will be guaranteed to be a maximum, the true one. Therefore, by using several fiducial models, one can eliminate the spurious maxima and keep the one that is common to all the fiducial models, which is the true one. We combine the MOPED likelihoods for different fiducial models by simply averaging them.[3]

The new measure Y is defined as

$$Y(\boldsymbol{q}) \equiv \frac{1}{N_{\mathrm{f}}} \sum_{\{\boldsymbol{q}_{\mathrm{f}}\}} \ln \mathcal{L}(\boldsymbol{q}; \boldsymbol{q}_{\mathrm{f}}), \tag{12}$$

where $\boldsymbol{q}$ and $\boldsymbol{q}_{\mathrm{f}}$ are the parameter vectors $\{T, \eta, \theta, \tau\}$ and their fiducial values $\{T_{\mathrm{f}}, \eta_{\mathrm{f}}, \theta_{\mathrm{f}}, \tau_{\mathrm{f}}\}$ and N_{f} is the number of fiducial models. The summation is over an ensemble of fiducial models $\{\boldsymbol{q}_{\mathrm{f}}\}$. $\mathcal{L}(\boldsymbol{q}; \boldsymbol{q}_{\mathrm{f}})$ is the MOPED likelihood (equation 8), i.e.

$$\ln \mathcal{L}(\boldsymbol{q}; \boldsymbol{q}_{\mathrm{f}}) = \sum_m [\boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}) \cdot [\boldsymbol{b}_m(\boldsymbol{q}q_{\mathrm{f}}) \cdot \boldsymbol{x} - \boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}) \cdot \boldsymbol{\mu}(\boldsymbol{q})]^2 \tag{13}$$

Fig. 2 shows Y as a function of period T for different size sets of fiducial models for a synthetic light-curve with S/N = 3 and 2000 observations. The top panel shows the value of Y using an ensemble of three fiducial models. As can be seen from the figure, there are

[2] The constant c controls the sharpness of the edges. We used $c = 100$ for all calculations in this work.

[3] This is chosen ad hoc. We have tried other approaches, all of which work similarly well. Averaging turned out to be the functional form in which the error and confidence level of the measurement could be easily and analytically calculated.

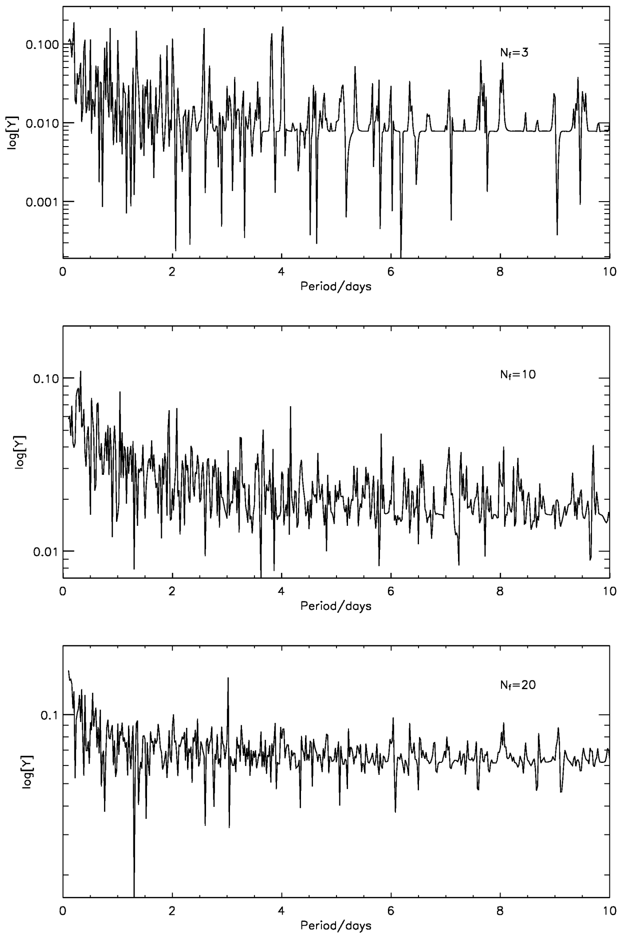

Figure 2. Plots of Y as a function of period T for a set of fiducial models for a synthetic light-curve with S/N = 3 and 2000 observations and $T = 1.3$ d. The top panel shows the value of Y using three randomly selected fiducial models, the middle panel using 10 and the bottom using 20. As the number of fiducial models used increases, the number of minima decreases. At $N_{\mathrm{f}} = 20$ there is only one obvious minimum at $T = 1.3$ d.

more than a few minima. Using an ensemble of 10 fiducial models (middle panel) reduces the number of minima. In the bottom panel we used an ensemble of 20 fiducial models, and there is only one obvious minimum, the true one.

Fig. 3 shows the value of Y as a function of each free parameter for a synthetic light-curve. We set the values of three of the parameters to the 'correct' values (used to construct the light-curve) and we left the fourth free for each panel. Note that the shape of Y as a function of η, θ and τ is smooth; however, the dependence on T is erratic, suggesting that efficient minimization techniques are not applicable.

4.1 Confidence and error analysis

To determine confidently that the minimum found is not spurious, the likelihood of the candidate solution must be compared to the value and distribution of Y derived from a set of light-curves with no transit signal. One can simulate a set of null light-curves and build a distribution by calculating the value of Y for each point in the parameter space for each simulated 'null' light-curve – a very expensive computational task. Alternatively, this null distribution can be analytically derived.

Since $x \sim N(\langle x \rangle, \sigma_x)$ and all other variables are deterministic, then it can be shown that $Y(\boldsymbol{q})$ follows a non-central $\mathcal{X}^2$ distribution $Y(\boldsymbol{q}) \sim \mathcal{X}^2(r, \lambda)$, where r is the number of degrees of freedom and λ is distribution $Y(\boldsymbol{q}) \sim \mathcal{X}^2(r, \lambda)$, where r is the number of degrees of freedom and λ is the

$$\mu = r + \lambda \tag{14}$$

and

$$\sigma^2 = 2(r + 2\lambda), \tag{15}$$

where $r = 4$ and λ is given by

$$\lambda = \frac{\mathrm{E}^2[\mathcal{X}]}{\mathrm{var}[\mathcal{X}]}. \tag{16}$$

The square of the expectation value is

$$E^2[\mathcal{X}] = \sum_m \left[\langle x \rangle B_m(\boldsymbol{q}_{\mathrm{f}}) - D_m(\boldsymbol{q}; \boldsymbol{q}_{\mathrm{f}})\right]^2, \tag{17}$$

where we define

$$B_m(\boldsymbol{q}_{\mathrm{f}}) \equiv \sum_t b^t_m(\boldsymbol{q}_{\mathrm{f}}), \tag{18}$$

and

$$D_m(\boldsymbol{q}; \boldsymbol{q}_{\mathrm{f}}) \equiv \boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}) \cdot \boldsymbol{\mu}(\boldsymbol{q}), \tag{19}$$

the variance is given by

$$\begin{aligned} \mathrm{var}[\mathcal{X}] &= \mathrm{var}\left[\sum_m \boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}) \cdot \boldsymbol{x} - \sum_m \boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}) \cdot \boldsymbol{\mu}(\boldsymbol{q})\right] \\ &= \sum_m |\boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}})|^2 \, \mathrm{var}[x^t] \\ &= \sigma_x^2 \beta_m(\boldsymbol{q}_{\mathrm{f}}) \end{aligned} \tag{20}$$

and where we define $\beta_m(\boldsymbol{q}_{\mathrm{f}})$ to be

$$\beta_m(\boldsymbol{q}_{\mathrm{f}}) \equiv \boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}) \cdot \boldsymbol{b}_m(\boldsymbol{q}_{\mathrm{f}}). \tag{21}$$

Combining the above equations we get

$$\lambda = \frac{\sum_m [\langle x \rangle \; B_m(\boldsymbol{q}_f) - D_m(\boldsymbol{q}; \boldsymbol{q}_f)]^2}{\sigma_x^2 \; \beta_m(\boldsymbol{q}_f)} \tag{22}$$

To compute confidence levels for a particular Y, we integrate a non-central $\mathcal{X}^2$ distribution with non-centrality given by equation (22) from $Y(\boldsymbol{q})$ to infinity. This is done numerically; still, this is a very quick operation. Furthermore, as we will show in Section 6, this will only be performed a few times per light-curve.

Fig. 4 shows the values of $Y(T)$ for the null case (i.e. a light-curve without a transit) both simulated (crosses) and theoretically calculated using the equations above (solid line is the expected value and dotted line is the 80 per cent confidence level). It is clear that the simulated values agree well with the theoretical ones. Note that, because the confidence can be calculated analytically, we do not have to simulate null light-curves and recalculate the Y for each light-curve, thus gaining computational speed.

5 RESULTS

Fig. 3 shows the results of likelihood as a function of each parameter using a typical synthetic light-curve. The above searches were performed only in one parameter at a time; regardless, we successfully recover the true values for the parameters of the transit.

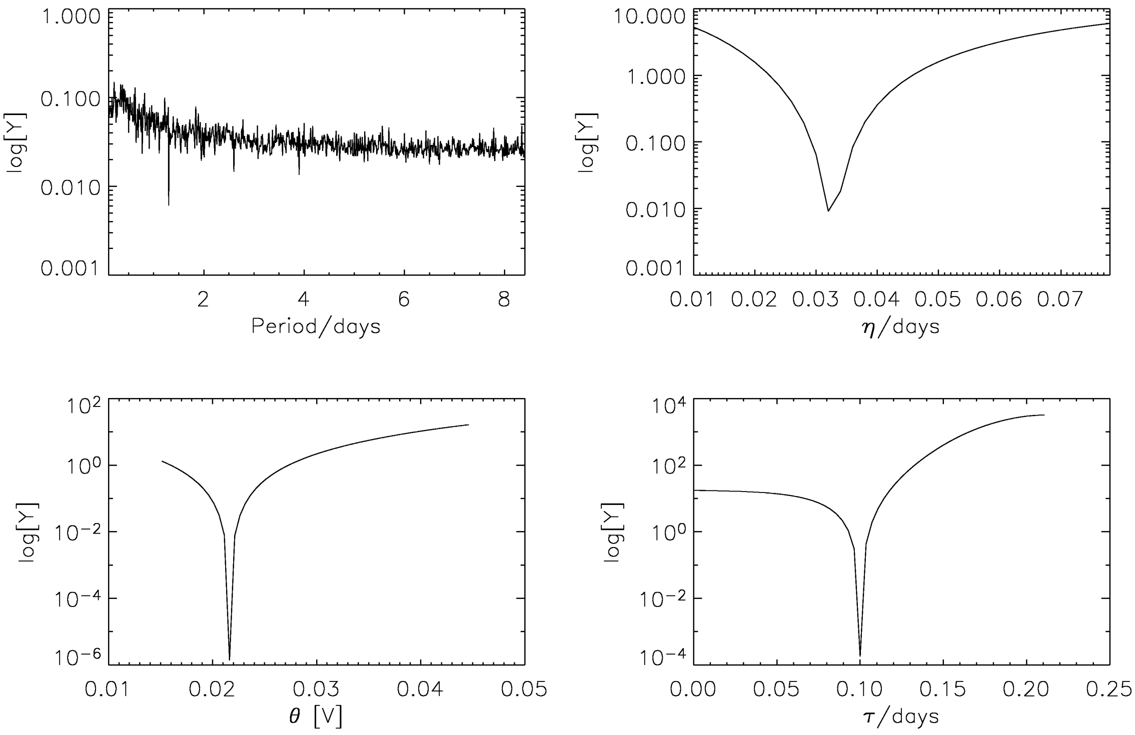

Figure 3. Likelihood as a function of period T (top left), transit duration η (top right), θ (bottom left) and τ (bottom right). In all parameters the correct value is found (we used 20 fiducial models). Note that for T the topology of the likelihood surface is fairly complicated, with many local minima, thus making efficient minimization techniques not applicable.

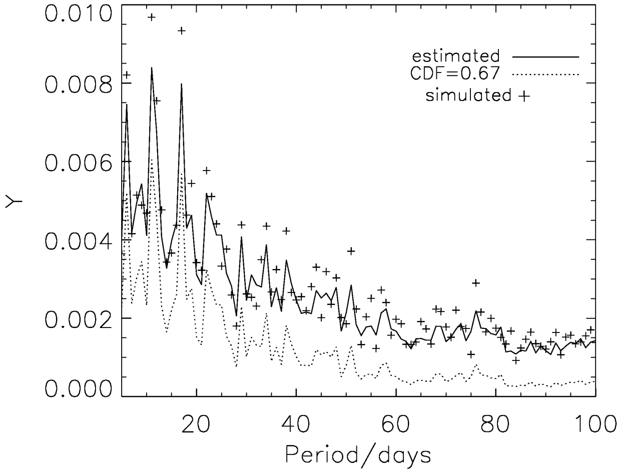

Figure 4. Values of $Y(T)$ for the null case (i.e. a light-curve without a transit) both simulated (crosses) and analytically calculated (see Section 4.1) (solid line is the expected value and dotted line is the 67 per cent confidence level). It is clear that the simulated values agree well with the theoretical ones.

In Fig. 5 we show the value of Y as a function of period for synthetic light-curves with a transit of 1.28 d. The run was done using 40 fiducial models. The different panels show different values of S/N. The dotted line shows the 80 per cent confidence level. For all four cases there is a well-defined minimum at the right period, where the minimum is below the 80 per cent level for S/N as low as 5 and at 71 per cent for S/N = 3.

The more realistic case is to perform the search in the four parameter spaces simultaneously and show that our method successfully recovers the 'correct' values of T, η, θ and τ for a sample of synthetic light-curves. This is shown in Figs 6, 7 and 8, where the 2D projections of the four-dimensional search are presented. The different contours correspond to 50, 65 and 80 per cent confidence levels. It is worth commenting on the 'multiple' maxima in the likelihood. This feature also appears in the one-dimensional search: multiple minima appear at multiples of the true period, but note that the best-fitting model is still the true period only (at the 50 per cent confidence level the other solutions are excluded). This behaviour is expected since, when the period is allowed to be a multiple of the true one, one out of n (n is an integer) transits will fit and therefore will produce a better fit than the null case. These multiple solutions can be easily excluded by keeping the shortest period. This only occurs for T; the other parameters have only one well-defined minimum at the true value.

5.1 Application to multiple light-curves

After having shown that our algorithm works properly on several synthetic light-curves, we now explore the performance of the method for a wide range of values for T, θ, η and τ. In particular, we have simulated light-curves for $0.1\,\mathrm{d} < T < 4\,\mathrm{d}$, $1\,\text{per cent} < \eta < 5\,\text{per cent}$ of the period, $0 < \tau < T$, $R_{\mathrm{planet}}/R_\star \sim 0.1$ and $12 < V < 24$. The observation frequency of the light-curve is similar to that of a Pan-STARRS light-curve. This space parameter and observation frequency should cover the range of light transit observations expected from surveys like Pan-STARRS[4] and the Large Synoptic Survey Telescope (LSST).[5]

[4] http://pan-starrs.ifa.hawaii.edu

[5] http://www.lsst.org

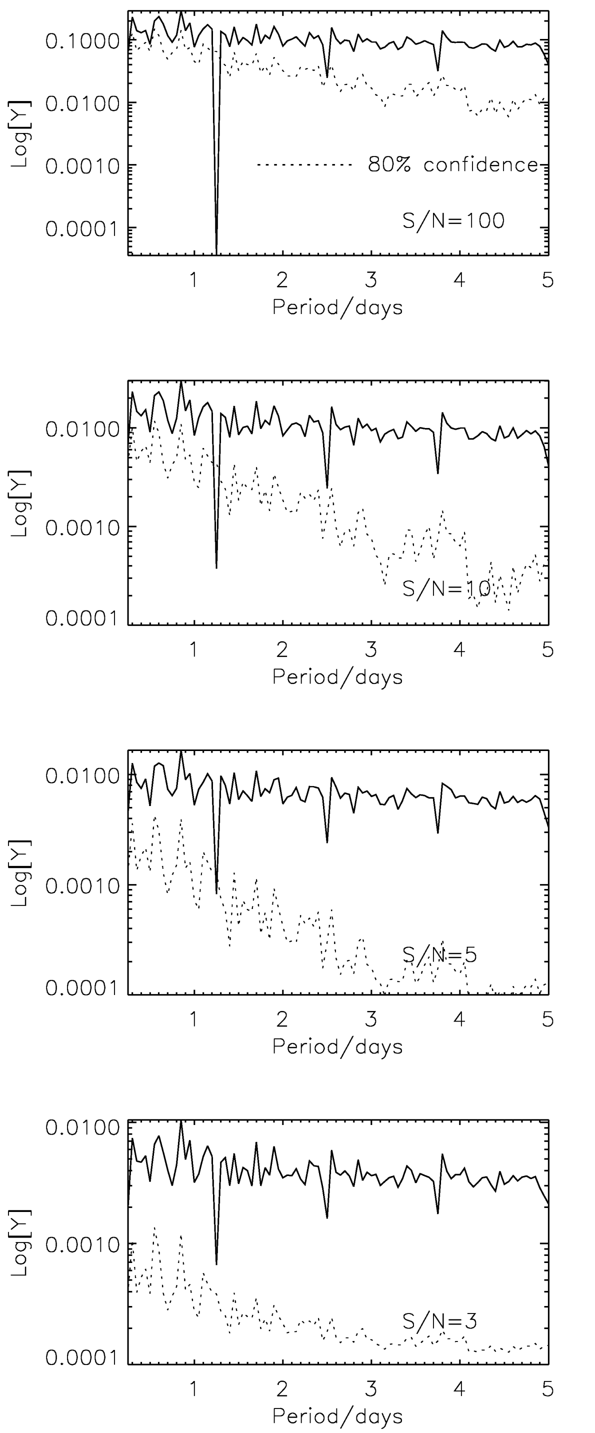

Figure 5. The value of Y as a function of period for a synthetic light-curve with a transit at 1.25 d. The various panels show different values of S/N. Note that there is a well-defined minimum at the right period. The dotted line shows the 80 per cent confidence level. Note that at this level there is only a single minimum at the right period even for S/N as low as 5.

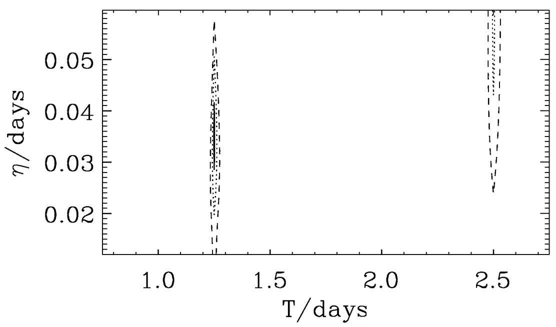

Figure 6. Projection of the four-dimensional likelihood surface on the two-dimensional space η–T. Note that contours close around the right period and that they appear at multiples of the right period as happens for the one-dimensional case (see text for more details).

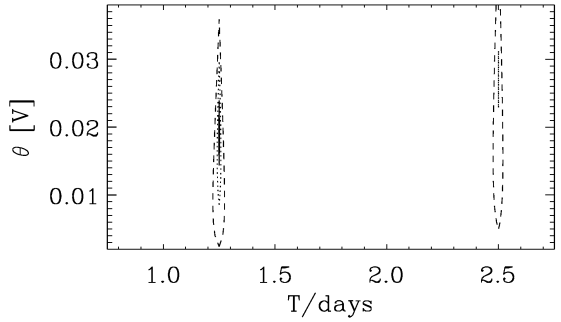

Figure 7. Same as Fig. 6 but for parameters θ and T.

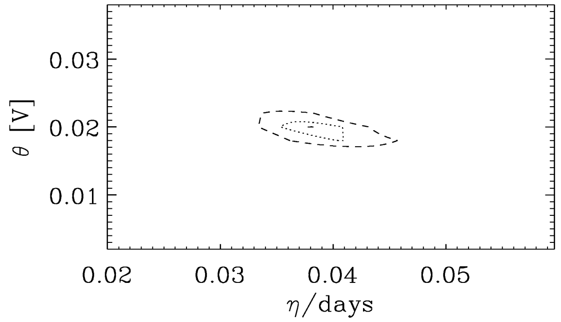

Figure 8. Same as Fig. 6 but for parameters θ and η.

We have simulated 100 light-curves with S/N = 5. For each light-curve we estimated the likelihood $Y(\boldsymbol{q}_{\rm true})$ for the ensemble of fiducial models and then we calculated the confidence that this value is not a spurious detection. Fig. 9 shows the distribution for these confidence values. Two histograms are shown: the dotted one is for curves with a total of 2000 observations (about 1 yr); the thick solid one is for the case when the range is doubled to 4000 measurements (2 yr). For the higher number of observations, there is a significant increase in the confidence of recovering the true period. For this case, most transits (80 per cent) are found with confidence over the null case higher than 70 per cent, i.e. for all stars the recovered period T has a confidence greater than 70 per cent of being the correct one. In about 25 per cent of cases the confidence in recovering the true period is greater than 90 per cent. For the case of 2000 measurements, the success rate is somewhat lower. This is because the error on the estimated parameters depends on the number of observations. For θ, η and τ this depends on the number of observations in the transits. However, for T this depends on the number of transits observed. One can show that the probability of observing a single

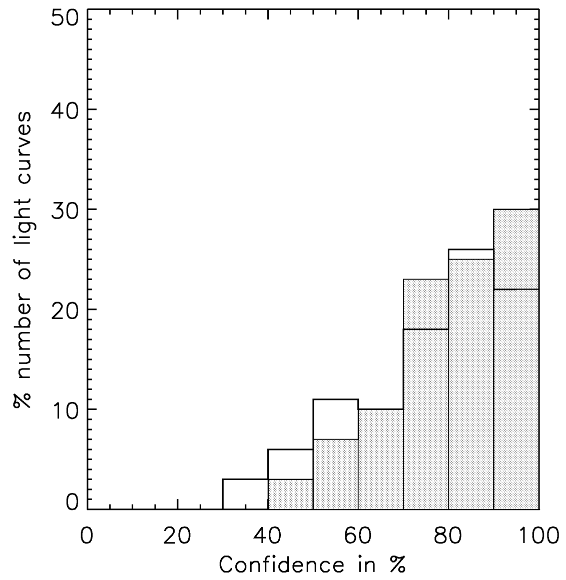

Figure 9. Distribution function for the confidence of finding the true period T for a set of 100 simulated light-curves with S/N = 5 for the range of values of transit parameters described in Section 5.1. Two sets of 100 light-curves have been simulated: a set where the number of measurements per light-curve is 2000 (dotted line) and another with 4000 measurements (solid line). Note that for the 4000 measurements case most T values are recovered with confidence higher than 80 per cent and that for about 30 per cent of the simulated light-curves the confidence in recovering the true period is greater than 90 per cent.

transit is proportional to η/T, though the probability of observing multiple transits is smaller. Furthermore, it also depends on the irregularity of the observational times (the more irregular the times, the better the chance of recovering the signal).

6 NUMERICAL METHOD

The real advantage of the present method lies in the fact that for a set of light-curves most of the work can be done once in calculating the fiducial models. In this section we describe the numerical approach in more detail and we present the numerical gain over the brute-force calculation using the full $\mathcal{X}^2$.

6.1 Calculations of the fiducial models

The second term of equation (13) does not depend on the actual light-curve data $\boldsymbol{x}$. Therefore, $D_m(\boldsymbol{q}_{\rm f}; \boldsymbol{q})$, $b^t_m(\boldsymbol{q}_{\rm f})$ and $qq_{\rm f}; \boldsymbol{q}q)$, $b^t_m(\boldsymbol{q}_{\rm f})$ and stored in files. Thus for each light-curve we only need to calculate $\sum_t b^t_m(\boldsymbol{q}_{\rm f})x^t$, which is independent of the search parameter $\boldsymbol{q}$. This is a major advantage of our method. Before we describe the numerical steps in more detail, we need to address how we choose the fiducial models and how many fiducial models are needed.

6.2 Choice of fiducial models

There are three questions that we need to address about the choice of fiducial models.

(i) *Number of fiducial models.* Since the confidence level can be calculated at each iteration step, the number of fiducial models does not need to be predetermined. If there is only one solution at confidence larger than 70 per cent, that parameter is considered to be the correct value and the iteration is stopped. Yet, for light-curves with low S/N the actual solution may never exceed that threshold. Therefore, we imposed a maximum of 100 fiducial models. Besides, for a typical survey there will be only a few expected light-curves containing a transit signal. Thus for most cases the iteration will be terminated at the 100 fiducial model limit.

(ii) *Choice of fiducial parameters.* It is simplest in finding the true solution to sample the parameter domain uniformly.

(iii) *Choice of search parameters* $\boldsymbol{q}$. Despite the fact that D_m will be calculated only once and it will not contribute to the overall computational burden of finding transits in a set of light-curves, the size of the data base (files) that stores the fiducial model information depends heavily on the choice of the free parameters range and grid size. This is mostly important for

As can be seen from Fig. 3 (top left panel), finding the 'correct' period where there is a non-linear dependence of the model on the period is a most difficult task. This is due to the fact that a small change in the value of period T produces a huge variation at the tail of the light-curve.

Theory suggests that the asymptotic standard deviation of the estimate of the period is of the order of $T^{5/3}$, so the grid should be that small too. We therefore performed the search on a uniform grid in frequency, T^{-1}, rather than on a uniform grid in T. There is also a related question of how fine the search grid should be for η and τ. Since for data folded at period T, the folded observation times are roughly uniform, the average spacing of subsequent folded observations is T/N (N is the total number of observations); thus T/N is a natural choice for the grid size for the η and τ searches. However, the spacing of $\eta/4$ should suffice to pick up the local minima, since the χ^2 surface will be correlated on this scale. For a typical search the total number of searches can be as high as 10^9, which translates to 1 TB of data. This is prohibitive for space missions. In what follows we examine how to reduce the search space further using physical and statistical arguments.

Transit length range

For a given period, one can allow η to take values between 0 and $T/2$. This is a naive estimate based on the fact that the planet spends half of the time in front of the star. The range of η can be further limited using geometrical arguments and Kepler's law. It can be shown that the transit duration is (Sackett 1999)

$$\eta \approx (T/\pi)\sqrt{(R_\star/a)^2 - \cos^2(i)}, \tag{23}$$

where $R_\star$ is the radius of the star, a is the orbit radius of the planet and i is the inclination angle. The maximum value that η can take is when the inclination angle i is zero. Using Kepler's law the ratio of duration over period is

$$\frac{\eta}{T} = \frac{R}{\pi[(T^2GM_\star)/(4\pi^2)]^{1/3}}. \tag{24}$$

For a typical main-sequence star this yields $\eta/T \approx 4$ per cent for periods of 1–2 d. The fraction gets smaller as the period increases, resulting in a gain of a factor of 50 in computational time (compared to the naive approximation $\eta/T \leqslant 1/2$).

Longest period

Equation (24) can be used to determine the longest period that can be recovered from the data. Namely, this is period at which the transit duration over the period is small enough that the probability of observing more than a few occultations is insignificant. It can be shown that for most inclinations the probability per data point of observing an occultation is given by

$$p = \frac{\eta}{T} = \frac{1}{2\pi}\frac{R_\star}{a}. \tag{25}$$

This is basically the probability that an observation taken at random orbital phase falls during a transit.

The probability of observing x occultations during the whole lifetime of the survey is given by a binomial distribution. At the limit where the number of observations is large, the probability distribution becomes a Gaussian distribution

$$P_t(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{26}$$

where p is given in equation (25). The mean value is given by

$$\mu = n_t p, \tag{27}$$

and the standard deviation by

$$\sigma = \sqrt{n_t p(1-p)}, \tag{28}$$

where n_t is the number of complete transits T_{range}/T. The probability of observing at least three transits is therefore given by the integral

$$P(x \geqslant 3) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_3^\infty \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{d}x. \tag{29}$$

For a typical main-sequence star, a planet with a period of 20 d has a probability of observing 10 occultations that is less than a few per cent. Using that as the upper limit to our search reduces the number of iterations by a factor of 5–10.

6.3 Numerical recipe

The steps of the numerical method are described below:

(i) Select a set of fiducial models. The choice of the fiducial parameters spans the domain of the search parameters.

(ii) Calculate $D_m(\boldsymbol{q}_{\text{f}}\,;\boldsymbol{q})$, $b^t_m(\boldsymbol{q}_{\text{f}})$ and $\beta^t_m(\boldsymbol{q}_f)$ and sampling frequency of the free parameters $\boldsymbol{q}$ are according to the physical arguments described above. Save values in a data base (binary files).

(iii) For each light-curve, calculate

$$\sum_t b^t_m(\boldsymbol{q}_{\text{f}})x^t.$$

(iv) Search through the fiducial models for $D_m(\boldsymbol{q}_{\text{f}}\,;\boldsymbol{q})$ with similar values as

$$\sum_t b^t_m(\boldsymbol{q}_{\text{f}})x^t$$

from the previous step. Note that, since the data base is sorted with respect to D, this is a log (N_q) operation, where N_q is the number of free parameter values.

(v) Calculate Y for those parameters such that

$$\sum_t b^t_m(\boldsymbol{q}_{\text{f}})x^t - D_m(\boldsymbol{q}_{\text{f}};\boldsymbol{q})$$

is small.

(vi) Compute the confidence level for the selected $\boldsymbol{q}$ values using equation (22). Note that since the D, B and β values are

(vii) Compute the confidence level for the selected $\boldsymbol{q}$ values using equation (22). Note that since the D, B and β values are precalculated, we only need to compute $\langle x \rangle$.

(viii) If there is only one minimum with confidence level higher than 70 per cent, exit.

(ix) If number of fiducial models is larger than 100, exit.

(x) Go back to step (iii).

6.4 Required number of operations

The brute-force minimization for the likelihood function requires

$$N^{\text{brute}}_{\text{total}} \sim N_{\text{obs}} N_q \tag{30}$$

operations. The number of operations for our method after the fiducial models are computed is

$$N^{\text{MOPED+}}_{\text{total}} \sim N_{\text{obs}} N_{\text{fid}}. \tag{31}$$

For a typical light-curve with low observing frequency like Pan-STARRS in the four-dimensional parameter space, N_q can easily be 10^{10}. This number is large because of the non-linear dependence of the period to the likelihood, and thus $N_T \sim 100\,000$ (see the arguments above).

7 CONCLUSIONS

We have presented a new algorithm for fast and efficient detection of transits in light-curves. Our algorithm produces a major speed-up factor in light transit searches of about eight orders of magnitude, compared to the brute-force method using the full χ^2. This translates into finding a transit on a light-curve with 10^4 observations in well under a second on current desktop computers. We have developed a four-parameter model for the transit of an object and have shown, using synthetic light-curves, that our algorithm is successful at recovering the true parameters of the transit. We have simulated a set of light-curves with the sampling rate and photometric accuracy expected in large synoptic surveys like Pan-STARRS and shown that for a large range in the values of the parameters (T, η, θ, τ) we recover the true values. For surveys like Pan-STARRS and LSST it should be possible to detect transits by Jupiter-like planets and planets several times the size of Earth. Since the expected detection rate of transits in these large surveys is very low, only one transit out of thousands of light-curves, we believe that our method provides a fast and efficient algorithm to detect transits for future surveys.

ACKNOWLEDGMENTS

P. Protopapas wishes to thank Rahul Dave for valuable discussions.

REFERENCES

Aigrain S., Favata F., 2002, A&A, 395, 625
Alonso R. et al., 2004, ApJ, 613, L153
Charbonneau D., Brown T. M., Latham D. W., Mayor M., 2000, ApJ, 529, L45
Charbonneau D., Brown T. M., Noyes R. W., Gilliland R. L., 2002, ApJ, 568, 377
Defaÿ C., Deleuil M., Barge P., 2001, A&A, 365, 330
Doyle L. R. et al., 2000, ApJ, 535, 338
Drake A. J., Cook K. H., 2004, ApJ, 604, 379
Gilliland R. L. et al., 2000, ApJ, 545, L47
Heavens A., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Hoekstra H., Wu Y., Udalski A., 2005, ApJ, 626, 1070
Jenkins J. M., Caldwell D. A., Borucki W. J., 2002, ApJ, 564, 495
Konacki M., Torres G., Sasselov D. D., Jha S., 2003, ApJ, 597, 1076

Kovács G., Zucker S., Mazeh T., 2002, A&A, 391, 369

Mallén-Ornelas G., Seager S., Yee H. K. C., Minniti D., Gladders M. D., Mallén-Fullerton G. M., Brown T. M., 2003, ApJ, 582, 1123

Mochejska B. J., Stanek K. Z., Sasselov D. D., Szentgyorgyi A. H., 2002, AJ, 123, 3460

Pont F., Melo C. H. F., Bouchy F., Udry S., Queloz D., Mayor M., Santos N. C., 2005, A&A, 433, L21

Sackett P. D., 1999, in Mariotti J.-M., Alloin D., eds, Planets Outside the Solar System: Theory and Observations. Kluwer, Boston, p. 189

Street R. A. et al., 2003, MNRAS, 340, 1287

Tingley B., 2003, A&A, 403, 329

Torres G., Konacki M., Sasselov D. D., Jha S., 2004, ApJ, 614, 979

Udalski A. et al., 2002, Acta Astron., 52, 1

Udalski A., Pietrzynski G., Szymanski M., Kubiak M., Zebrun K., Soszynski I., Szewczyk O., Wyrzykowski L., 2003, Acta Astron., 53, 133

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