Chapter 2

Transit detection algorithms

The present time is an exciting one for the transit searching community. After a few years of tuning both observation and data analysis techniques, the first ground-based programs (in particular the Optical Gravitational Lensing Experiment, or OGLE) are now yielding detections, albeit in smaller numbers than expected (Konacki et al. 2003a; Bouchy et al. 2004). Is the heralded 'landslide' of extra-solar planets detected via transits really just around the corner? Will the planned space missions really deliver as expected? Their success hangs, at least partially, on our ability to automatically search huge datasets for tiny, repetitive signals deeply buried in noise. The work presented here represents an attempt to contribute to this challenging task.

At the time of starting the present thesis in 2001, there was no widespread agreement as to the best approach for planetary transit detection. However, a small number of papers dedicated to the problem (see e.g. Jenkins et al. 1996; Doyle et al. 2000; Gilliland et al. 2000; Defaÿ et al. 2001) provided a natural starting point from which to develop new methods, or improve on existing ones. By contrast, at the time of writing, the number of published articles devoted to or dealing with this problem has vastly increased (see e.g. Aigrain & Favata 2002; Jenkins et al. 2002; Kovács et al. 2002; Koen & Lombard 2002; Street et al. 2003; Aigrain & Irwin 2004). The references listed here are by no means exhaustive, but provide a good overview of the variety of techniques commonly used today or under investigation for future missions.

The first comparative studies (Tingley 2003a,b, see also Chapter 6) represent an attempt to synthesise and sort through this zoo of methods. One subject on which there is widespread consensus, is that a diversity of available algorithms is a better guarantee of success than a single, uniformly used approach. Not only is it likely that different methods will perform better in different circumstances, but the independent detection of a given event using two algorithms might improve the confidence level in the detection. On the other hand, most published algorithms are based on the same underlying maximum likelihood principles. Clarifying this common case helps to understand the apparently disparate menagerie in terms of a family of closely related siblings, differing mostly in their implementation details – which affect speed, robustness, and adaptability to specific circumstances.

The algorithms presented in this chapter were conceived primarily with upcoming space-based planetary transit search missions in mind, such as COROT, *Eddington* and *Kepler*, but they are equally applicable to data from ground-based programs. Both types of datasets, despite the differences highlighted in Chapter 1, place very similar requirements on the detection algorithms.

The space missions are expected to produce 10 000's of light curves per target field, each containing 100 000's of individual observations, as the time sampling is high and a single field is observed for months or even years. Most small aperture ground-based programs have similar fields of view, or larger fields of view but brighter magnitude limits, leading to similar numbers of stars per target field. While the number of observations of each field is much less than for the space missions (100's or 1000's), is it usual for a number of fields to be observed in a given observing season, thereby leading to similar sized datasets in a given year.

The time sampling of the data from the space missions is expected to be regular, and the duty cycle very high (up to $\geq$95 % for *Eddington* and *Kepler*), though a number of data gaps, e.g. due to telemetry losses, cannot be avoided. In ground-based data the sampling is, of course, very irregular, with interruptions between nights and between observing runs, as well as due to weather or technical problems. In order to keep the algorithms as general as possible, they were designed to be applicable to data with irregular sampling. In some cases regular sampling could allow for small alterations to the code, leading to slight gains in computing time.

To tackle the enormous databases involved, speed and automation are vital. This leads, at least at the detection stage, to a single motto: the simpler the better. This is consistent with maximising the statistical efficiency (see Section 2.1.1.3). The present chapter reflects a learning process in this respect, in that the first algorithm investigated, a Bayesian approach, is more sophisticated, and for detection purposes less effective, than the second, a stripped down version where the use of Bayesian priors was dropped and the assumptions about the shape of the transit signal further simplified. This does not mean that sophisticated methods, incorporating as much prior knowledge about the sought-after signal and the noise characteristics as possible, should be discarded, but they are expected come into their own at the characterisation, rather than detection stage – though that stage is beyond the scope of the present work.

Section 2.1 describes the Bayesian algorithm, together with performance tests which were carried out on simulated light curves affected by photon noise as expected for *Eddington*. Section 2.2 describes the simpler, but derivative, box fitting algorithm.

2.1 A Bayesian, step-function based algorithm

Transit detection algorithms based on a Bayesian approach, already investigated in the context of the COROT mission (Defaÿ et al. 2001), constitute an interesting alternative to more conventional approaches, e.g. matched filters. They maximise the use of whatever information is available on the phenomenon one is trying to detect, and are relatively flexible, allowing the seamless incorporation of new information into the detection process as it becomes available. While a global 'marginalised' statistic can be used for the detection, information is directly available to reconstruct the detected signal if wanted, therefore providing a tool to discriminate between planetary transits and other types of periodic signals (Defaÿ 2001), as well as directly measuring additional parameters such as the planet's radius.

After a brief investigation of the method of Defaÿ (2001), a decision was made (for reasons outlined below) to develop a novel algorithm, based on the GL method of Gregory & Loredo (1992) (hereafter GL92), which was also the starting point for Defaÿ et al. (2001). The GL92 method was developed for the search of periodic variations in emission from pulsars in X-ray data.

The approach of Defaÿ et al. (2001) was based on the expansion of the light curve into a truncated Fourier series. C. Defaÿ kindly provided a coded implementation of this method, but experiments with this code showed that performing the detection in the Fourier domain made the algorithm computationally sensitive to data gaps and discrete sampling rates. The *direct space* approach investigated here is expected to be more robust, though it does not have the advantage of providing a direct means of reconstructing the shape of the detected signal.

The GL method was initially developed for Poisson noise dominated light curves (as is the case for X-ray pulsars) and later extended to the Gaussian noise case (Gregory 1999, hereafter G99). At the flux levels of interest for transit searches, the photon shot noise per detection element (which is Poisson distributed) can be very well represented by Gaussian noise. The original formulation of the GL method made no assumptions about the shape of the variations, the model consisting of a step-function with m even duration, arbitrary level bins. The new algorithm was developed to be as 'general purpose' as possible, but makes intrinsic use of prior knowledge of the expected signal, by allowing one of the bins to have a variable width, to represent the out of transit constant signal level. This formulation also permits the phase of the transits, or epoch (start time) of the first transit, to be identified, a task the original GL

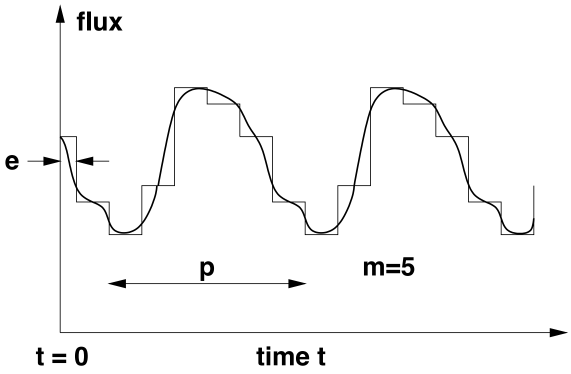

Figure 2.1: Schematic illustration of the type of model used in the GL method. There are m equal duration bins per period p. The (arbitrary) epoch e is defined as the time elapsed between the start of the light curve and the end of the first bin.

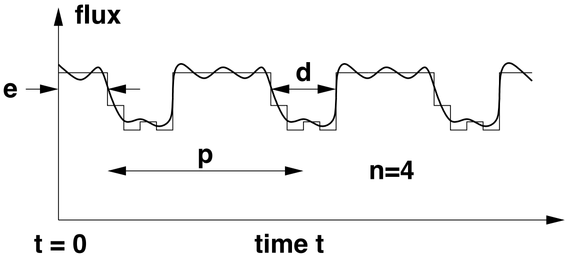

Figure 2.2: Schematic illustration of the type of model used in the modified GL method. There are n in-transit bins of duration d/n plus 1 out-of-transit bin of duration $p - d$ per period p. The epoch e is defined as the time elapsed between the start of the data and the start of the next transit.

method is not suited for (see Section 2.1.1.1). The fitted parameters are the period, duration and phase of the transit. The shape of the transit can then be reconstructed from the phase-folded light curve.

The new algorithm, referred to hereafter as the modified GL method, is derived from the GL method in Section. 2.1.1. Using simulated light curves, the modified GL method is compared to the original in Section 2.1.2, and its performance in the presence of photon noise evaluated in Section 2.1.3. Section 2.1.4 then outlines tests on light curves containing, in addition to simulated photon noise and transits, solar micro-variability.

2.1.1 Derivation of the algorithm

2.1.1.1 The models

In GL92 and G99, the periodic hypothesis is represented by a class of periodic step-function models, resembling a histogram, with m equal duration bins per period (see Figure 2.1). The model level in each bin is allowed to vary, giving these models the ability to describe light curves of unknown shape. A given model $\mathcal{M}_m$, with a given value of m, has $m + 2$ parameters: period p, epoch e (or phase ϕ) and m bin levels $R = \{r_1, r_2, \ldots, r_m\}$.

This *periodic* hypothesis is contrasted with *aperiodic* and *constant* hypotheses, which are special cases of the periodic model, both having a period equal to the total duration of the light curve and the latter having $m = 1$.

In the case of transit searches, some information about the shape of the sought after signal is available. The model should therefore consist of a long flat section followed by a dip. A family of models similar to those of the GL method but where one bin is much longer than the others (see Figure 2.2) is therefore used. There are now $n+1$ bins, n being the number of *in-transit* bins. A given model $\mathcal{M}_n$, with a given value of n, has $n+4$ parameters: period p, transit duration d, epoch e, and $n+1$ bin levels $R = \{r_0, r_1, \ldots, r_n\}$. Models with lower n will incur a lower Occam penalty factor, as emphasised in G99.

Only two hypotheses (transit, denoted by $\mathcal{H}_\mathrm{T}$ and constant, denoted by $\mathcal{H}_\mathrm{C}$) are considered: the constant hypothesis is a special case of the transit hypothesis with a period equal to the full duration of the light curve and $n = 0$, but the adopted model is not suited to aperiodic variations. The special case where the period is equal to the total duration of the light curve corresponds to a single transit, which is considered to be part of $\mathcal{H}_\mathrm{T}$.

2.1.1.2 Bayesian analysis

Just like any signal detection problem, transit searching consists in hypothesis testing. Bayesian analysis provides a framework for the a posteriori incorporation of new or additional information in a detection process. Given the rate at which our knowledge of the characteristics of extra-solar planets is increasing, such an approach has distinct advantages. Already, one can directly incorporate into the detection the fact that transits are rare events due to the low alignment probability. One can also incorporate the fact that this alignment probability is lower for the more distant, i.e. longer period planets, by setting the a priori probability distribution for the period, or period *prior*, to be lower for longer periods.

An excellent discussion of Bayesian detection and parameter estimation is given in GL92. However, given the fact that Bayesian methods are relatively rarely used in the present field, the process is described below in some detail.

The hypothesis under test is that the light curve $X = \{x_1, x_2, \ldots, x_N\}$, contains (periodic), short and shallow dips against an otherwise constant background. This global transit hypothesis will be referred to as $\mathcal{H}_\mathrm{T}$. It is generally relatively straightforward, under given assumptions about the noise characteristics of the data, to compute a *likelihood* for $\mathcal{H}_\mathrm{T}$, that is a measure of the extent to which $\mathcal{H}_\mathrm{T}$ predicts X, $\mathrm{P}(X|\mathcal{H}_\mathrm{T})$ (see Section 2.1.1.3). But the quantity of interest is $\mathrm{P}(\mathcal{H}_\mathrm{T}|X)$, the *posterior probability* for the model. The relationship between the two is governed by Bayes' theorem (Bayes 1764):

$$\mathrm{P}(\mathcal{H}_\mathrm{T}|X) = \frac{\mathrm{P}(\mathcal{H}_\mathrm{T})}{\mathrm{P}(X)} \times \mathrm{P}(X|\mathcal{H}_\mathrm{T}), \tag{2.1}$$

where $\mathrm{P}(\mathcal{H}_\mathrm{T})$ encompasses a priori information about hypothesis and the underlying physical processes, measurement effects, etc., and is consequently known as a *prior*, and $\mathrm{P}(X)$ is a normalisation constant.

In the case of transit searches, the prior for $\mathcal{H}_\mathrm{T}$, $\mathrm{P}(\mathcal{H}_\mathrm{T})$, represents any a priori information as to the probability that the light curve contains transits. In the range of planetary radii already probed by other methods, its value could be deduced from the frequency of planets observed to date, as well as from the alignment probability of the orbit with the line of sight (requiring some assumption about the distribution of orbital inclinations). This would require a relatively complex integration over several parameters, and the possibility that other variations in the light curve might dominate over any transits would still need to be accounted for. At first glance however, the alignment requirement alone suggests that $\mathrm{P}(\mathcal{H}_\mathrm{T}) \leq 0.1$.

$\mathrm{P}(X)$, the normalising factor, is given by:

$$\mathrm{P}(X) = \sum_i \mathrm{P}(\mathcal{H}_i|X) \times \mathrm{P}(\mathcal{H}_i). \tag{2.2}$$

where the $\mathcal{H}_i$ represent each of the hypotheses under test. The need to evaluate $\mathrm{P}(X)$ is circumvented by comparing the hypothesis under consideration, $\mathcal{H}_\mathrm{T}$, to that for another, *null* hypothesis. In the present context, the null hypothesis consists of a constant light curve, and is referred to as $\mathcal{H}_\mathrm{C}$. One computes the *odds ratio*, or ratio of the posterior probabilities for the two hypotheses:

$$\mathcal{O} = \frac{\mathrm{P}(\mathcal{H}_\mathrm{T}|X)}{\mathrm{P}(\mathcal{H}_\mathrm{C}|X)} = \frac{\mathrm{P}(\mathcal{H}_\mathrm{T})}{\mathrm{P}(\mathcal{H}_\mathrm{C})} \times \frac{\mathrm{P}(X|\mathcal{H}_\mathrm{T})}{\mathrm{P}(X|\mathcal{H}_\mathrm{C})}, \tag{2.3}$$

If $\mathcal{O}$ is greater than 1, there is evidence for transits in the light curve. If, as in the example above, $\mathrm{P}(\mathcal{H}_\mathrm{T}) = 0.1$ (and assuming all light curves correspond to either $\mathcal{H}_\mathrm{T}$ or $\mathcal{H}_\mathrm{C}$), then $\mathrm{P}(\mathcal{H}_\mathrm{T})/\mathrm{P}(\mathcal{H}_\mathrm{C}) = 0.11$.

In practice, one does not compute a likelihood for $\mathcal{H}_\mathrm{T}$ as a whole, but rather for a single member $\mathcal{M}_n$ of the class of models represented by $\mathcal{H}_\mathrm{T}$, with a particular value of n, and for a particular set of parameters (period, duration, epoch, shape) represented by the $(n+4)$-element vector V. Likelihoods are computed for each value of V, and combined through a marginalisation process. Given a multi-dimensional array of likelihood values $\mathrm{P}(X|\mathcal{M}_n, V)$, unwanted parameters are marginalised over. One marginalises over each parameter v_i in turn by multiplying the likelihood by a one-dimensional prior for that parameter and integrating over (or summing over all sampled values) of v_i:

$$\mathrm{P}\left(X|\mathcal{M}_n, V'\right) = \sum_{v_i=v_i^{\mathrm{min}}}^{v_i^{\mathrm{max}}} \left\{\mathrm{P}\left(v_i|\mathcal{M}_n\right) \times \mathrm{P}\left(X|\mathcal{M}_n, V\right)\right\}, \tag{2.4}$$

where V' is the $(n+3)$-element vector of all parameters contained in V except v_i. The choice of prior depends on pre-existing knowledge about each parameter. For example, in the case of transits the prior for period might be chosen to reflect the orbital period distribution observed to date. Specific forms of prior can also reflect common sense, for example a logarithmic prior in period or frequency ensures that the results are the same whether one works in period or frequency (see GL92 and references therein). The marginalisation is repeated until no parameters remain (in that respect n can be treated as a parameter), and the left hand side of Equation (2.4) is the global likelihood for $\mathcal{H}_\mathrm{T}$. After computing the likelihood for $\mathcal{H}_\mathrm{C}$, one then obtains the global odds ratio $\mathcal{O}$.

If $\mathcal{O} > 1$, i.e. there is evidence for the transit hypothesis, it is natural to ask which set of transit parameters best describes the data. A transit detector must provide some information about the location of the transits in the light curve to be useful, at least the period and epoch. The posterior probability distribution for each parameter v_i is obtained from the one-dimensional likelihood function, marginalised over all other parameters (including n), using Bayes' theorem:

$$\mathrm{P}\left(v_i|X, \mathcal{H}_\mathrm{T}\right) = \frac{\mathrm{P}\left(v_i|\mathcal{H}_\mathrm{T}\right)}{\mathrm{P}\left(X\right)} \times \mathrm{P}\left(X|\mathcal{H}_\mathrm{T}, v_i\right). \tag{2.5}$$

Given that $\mathrm{P}(X)$ is constant with respect to V, the best trial value of v_i is that which maximises $\mathrm{P}(v_i|X, \mathcal{H}_\mathrm{T}) \times \mathrm{P}(X) = \mathrm{P}(v_i) \times \mathrm{P}(X|\mathcal{H}_\mathrm{T}, v_i)$.

2.1.1.3 Likelihood calculation

The calculations are given below in sufficient detail to allow the reader to reproduce the algorithm, but some lengthy derivations which were taken from G99 have not been reproduced. The original GL method was implemented in parallel to the modified version to provide a benchmark for test purposes, but with the same simplifications as the modified method, that is no noise scale parameter. The large number of parameters in the original GL method is its very weakness: each of them implies a dilution of the likelihood function (whose peak is spread over more dimensions of parameter space). This leads to larger variances in the estimated parameters, and hence to smaller statistical efficiencies.

Each data value x_i, corresponding to time t_i, is decomposed as

$$x_i = r_i + e_i, \tag{2.6}$$

where r_i is the value predicted by the model at time t_i and e_i represents any variations in the data not accounted for by the model. In the present context e_i will contain, apart from true variations not accounted for by the model, a photon shot noise contribution, well approximated by a Gaussian distribution for high photon counts, plus other instrumental and astrophysical noise of unknown distribution. According to the Central Limit Theorem, the most conservative assumption for the distribution of e_i is a Gaussian.

The treatment of errors has been simplified relative to G99: we assume the noise standard deviation for each data point x_i has a known value σ_i. This assumption is justified in the context of transit searches where the noise characteristics should be well determined from the large numbers of simultaneous high-precision light curves, and removes the need for the rather confusing noise scale parameter b of Gregory (1999).

The likelihood for a given value of n (i.e. model $\mathcal{M}_n$) with a given set of $n+4$ parameters (p, d, e, R) is then given by a product of Gaussian probability distributions (assuming the data points are independent):

$$\mathrm{P}\left(X|\mathcal{M}_n, p, d, e, R\right) = \prod_{i=1}^{N} \left\{ \frac{\sigma_i^{-1}}{\sqrt{2\pi}} \times \exp\left[-\frac{(x_i - r_i)^2}{2\sigma_i^2} \right] \right\}, \tag{2.7}$$

where N is the total number of data points.

Before calculating this likelihood one must determine in which bin j of the model a given data point falls:

$$j(t_i) = \left\{ \begin{array}{lll} t_i^{\mathrm{mod}} & : & \text{if } 0 < t_i^{\mathrm{mod}} \leq n \\ 0 & : & \text{otherwise} \end{array} \right. , \tag{2.8}$$

where

$$t_i^{\mathrm{mod}} = \mathrm{int}\left(\frac{(t_i + p - e) \bmod p}{d/n} + 1 \right), \tag{2.9}$$

n is the number of bins per transit, int(y) is the nearest integer lower than or equal to y and a mod b is the remainder of a divided by b.

As a small aside, it is interesting to compare this expression with Equation (6) of Gregory (1999), which corresponds to the original set of models with m equal duration bins:

$$j(t_i) = \mathrm{int}\left\{ 1 + m\left((\omega t + \phi) \bmod 2\pi \right) / 2\pi \right\}, \tag{2.10}$$

where the angular frequency $\omega = 2\pi/p$ and ϕ is a phase parameter representing "*the position of the first bin relative to the start of the data*" and running from 0 to 2π. The definition and range of this parameter are inconsistent, unless one bin can somehow be identified as the first, which is not the case given the unknown shape of the model. As Gregory (1999) employed a uniform prior for ϕ, this inconsistency has no effect on the detection, and the period and shape determinations, which were the quantities of interest in that paper. However it explains some of the results presented in Section 2.1.2.

The likelihood can now be expressed in terms of the $n+1$ bins of the model:

$$P\left(X|\mathcal{M}_n, p, d, e, R\right) = \prod_{j=0}^{n} \left[(2\pi)^{-(n_j/2)} \times \left(\prod_{i=1}^{n_j} \frac{1}{\sigma_i} \right) \times \exp\left(-\frac{\alpha_j}{2}\right) \right], \tag{2.11}$$

where n_j is the number of data points in bin j, and

$$\alpha_j = \sum_{i=1}^{n_j} \frac{\left(x_i - r_j\right)^2}{\sigma_i^2}, \tag{2.12}$$

r_j being the model value in bin j. For all purposes except the determination of the light curve shape inside the transit, the individual r_j's do not matter. It is therefore desirable to marginalise over the r_j's, that is to compute a combined likelihood for all possible values of the r_j's within a range set a priori. According to Bayes' theorem, this is done by multiplying the likelihood by a prior and integrating over the range of r_j:

$$P\left(X|\mathcal{M}_n, p, d, e\right) = \prod_{j=0}^{n} \left[(2\pi)^{-(n_j/2)} \times \left(\prod_{i=1}^{n_j} \frac{1}{\sigma_i} \right) \times \mathcal{R}_j \right], \tag{2.13}$$

where

$$\mathcal{R}_j = \int_{r_{\rm min}}^{r_{\rm max}} dr_j \, P\left(r_j|\mathcal{M}_n\right) \exp\left(-\frac{\alpha_j}{2}\right), \tag{2.14}$$

$r_{\rm min}$ and $r_{\rm max}$ being the minimum and maximum value of the r_j's, respectively. The distinct advantage of step-function models is that, as shown by Gregory, it is possible to perform this marginalisation analytically. Using a uniform prior for the r_j's: $P\left(r_j|\mathcal{M}_n\right) = (\Delta_r)^{-1}$ where $\Delta_r = r_{\rm max} - r_{\rm min}$, and following the derivation of G99, we obtain

$$\begin{aligned} \mathrm{P}\left(X|\mathcal{M}_n,p,d,e\right) &= \left(\tfrac{1}{\sqrt{2\pi}}\right)^{N}\left(\tfrac{1}{\Delta_r}\right)^{(n+1)}\left(\tfrac{\pi}{2}\right)^{\left(\frac{n+1}{2}\right)}\left(\prod_{i=1}^{N}\frac{1}{\sigma_i}\right)\exp\left(-\sum_{j=0}^{n}\frac{\chi^2_{W_j}}{2}\right) \\ &\times \left\{\prod_{j=0}^{n} W_j^{1/2}\left[\mathrm{erfc}(y_{j\mathrm{min}})-\mathrm{erfc}(y_{j\mathrm{max}})\right]\right\}, \end{aligned} \tag{2.15}$$

where the quantities W_j, $\chi^2_{W_j}$, $y_{j\mathrm{min}}$ and $y_{j\mathrm{max}}$ are taken directly from Equations (11) to (16) in G99, and erfc(y) is the complementary error function. In fact, it is shown in Section 2.2.2.1 that the bin levels are not independent parameters, and are fully determined by the data. This fact is exploited by the improved method that was later derived from the one presented in this Section.

2.1.1.4 Choice of priors

Following G99, we use a Jeffreys prior for the period:

$$\mathrm{P}\left(p|\mathcal{M}_n\right) = \frac{1}{p\ln\left(p_{\mathrm{max}}/p_{\mathrm{min}}\right)}, \tag{2.16}$$

where p_{min} and p_{max} are the limits of the period-space explored and $\ln\left(p_{\mathrm{max}}/p_{\mathrm{min}}\right)$ is a normalisation constant to ensure that $\int_{p_{\mathrm{min}}}^{p_{\mathrm{max}}} \mathrm{P}\left(p|\mathcal{M}_n\right)\ dp = 1$.

As pointed out by G99 (see references therein), this prior arises naturally from considerations of invariance with respect to changes in time scale and ensures that an investigator working in terms of p with this prior would obtain the same results as an investigator working in terms of frequency ν with a $1/\nu$ prior. In the present context it also reflects (though in a qualitative rather than quantitative fashion) the lower transit probability for longer periods that arises from geometric alignment considerations.

As already mentioned in Section 2.1.1.3, we use a flat prior for the r_j's, introducing only a normalising factor $1/\Delta r$. Similar priors are also used for epoch and duration.

Provided the results of the individual likelihood calculation (Equation 2.15) are stored, the period and duration priors could be changed at a later date, for example once the observed distributions for these parameters in the case of planetary transits are better known.

Finally, all values of n are considered equally likely a priori, bearing in mind that higher values of n are automatically affected by an implicit Occam's razor penalty factor, as discussed by GL92.

2.1.1.5 Weighting factor to compensate for uneven distribution into the bins

When the number of periods is low such that one bin might be represented four times while another only three times, or if there are gaps in the data which may not be evenly distributed over the bins, GL92 noted that some of their initial assumptions may fail, leading to the appearance of an erroneous trend in the posterior probability for the period.

In an appendix to GL92, a solution to this problem was proposed. A weighting factor s_j is applied to each bin:

$$s_j = \left(\frac{n_j m}{N}\right)^{-n_j}. \tag{2.17}$$

It is important to note that this factor was derived in the context of Poisson statistics.

Despite the low number of periods in our light curves, we found that no weighting factor was required in the benchmark algorithm that reproduced the GL method identically. However, it is clear that the problem is more acute in the modified algorithm. The 'out of transit' bin contains many more data points than the others, and therefore has a much larger effective weight. A weighting factor is required to compensate for this problem. The expression given above for s_j is only appropriate in the photon count context in which it was derived, not in the Gaussian noise case relevant here. A different weighting factor can be heuristically derived by considering Equation (2.12). The contribution of each model level to the likelihood is a χ^2 sum. The variance of a χ^2 distribution is equal to twice the number of degrees of freedom $\mathcal{N}$. In each bin there are n_j data points and $n_{\rm par}$ parameters to adjust (in the modified GL method $npar = n + 4$). As $n_j \gg n_{\rm par}$, $\mathcal{N} = n_j - n_{\rm par} \simeq n_j$. Weighting each bin by a factor $1/n_j$ is therefore equivalent to weighting proportionally to the inverse variance. In practice this is achieved by maintaining the expressions for $\chi^2_{W_j}$, $y_{j\rm min}$ and $y_{j\rm max}$ given in Gregory (1999) in terms of x_i and σ_i, but replacing W_j by W_j/n_j.

This modification was implemented in our algorithm and found to give more robust results.

2.1.1.6 Minimising the computing time

For a given set of parameters, the calculation of the likelihood involves summing over each element in each bin. The time required to compute the likelihood for a given set of p, d, e therefore scales linearly with the number of points in the light curve. It also increases with the number of bins, but this is a slow increase. It does not depend on the individual parameter values.

The overall computing time also depends, of course, on how tightly the parameter space is sampled. It is necessary to minimise the number of trial values for each parameter without missing potentially localised likelihood maxima. Because of the relative sharpness of the peak in the posterior probability for the period, the period increment needs to be kept fairly small (typically once or twice the time step between data points). Attention was therefore concentrated on what increment was suitable in terms of epoch. The results are not significantly worsened by increasing the *posital* phase ϕ_{pos}[1] increment from $1/p$ (i.e. shifting the model by 1 sampling time at each increment) to $d/2np$ (i.e. shifting the model by half the duration of an in-transit bin at each increment). Further increase leads to sharp steps in the posterior probability distribution (analogous to Shannon's sampling theorem).

However, the computing time is inversely proportional to the increment, and a posital phase increment of $d/2np$ is still prohibitively expensive. In practice, steps in the posterior probability distribution that result from a larger increment can be effectively removed by dividing it by the equivalent distribution for an entirely flat light curve with the same duration, sampling and data gaps as the light curve. We call this dividing function the 'window function' [2]. We therefore use a posital phase increment of $d/2p$ and perform the division before analysing the results. As the window function only needs to be calculated once per period and duration, this is much faster than using a smaller increment (see Section 2.1.3).

Note that due to the use of this window function one should not strictly speaking use the word 'posterior probability' when talking about the output of the algorithm. Hereafter, we will refer to 'modified posterior probability' to mean 'posterior probability distribution divided by the window function'. This also implies that the global odds ratios mentioned in Section 2.1.1.2 cannot be used to directly measure the ratio of the probabilities for a periodic model compared to a constant model. Instead, we use bootstrap simulations (see Section 2.1.3.1) to set a threshold value of the detection statistic above which a detection is accepted.

2.1.2 Comparison with the original GL method

In order to establish a reference point and to gain a preliminary estimate of the modified algorithm's performance, some qualitative tests were run on both the original and the modified version.

For this purpose light curves containing transits and photon noise were generated with the parameters of the *Eddington* mission in mind. We describe below a reference light curve simulated with one particular set of parameter values. Each

[1] $\phi_{\mathrm{pos}} = \phi/2\pi$.

[2] This also has the advantage of ironing out any residual effects of the uneven bin duration not removed by the weighting factor.

parameter was then varied in turn over a small but representative range.

All light curves have a sampling time of 15 minutes. The total light curve duration is 4 months. To simulate the transit signal from a planet in a 1 year orbit observed for 3 or 4 years without the computational expense such a duration would imply, a 1 month orbital period was used.

Given the presence of limb darkening in stellar photospheres, planetary transits are not perfectly 'flat bottomed' (nor are they, strictly speaking, truly grey). To simulate transits in a realistic way, we use the Universal Transit Modeller (UTM) software written by H. J. Deeg (Deeg et al. 2001). UTM can simulate light curves from any number of luminous or dark objects, including stars, planets, rings and moons. Circular orbits are assumed, and a linear limb darkening law is adopted for the stars. We used limb darkening coefficients from Van Hamme (1993). The dark objects – planets, rings and moons – are assumed to have zero albedo, i.e. no reflected light is included. This is justified because, in white light, the amount of light reflected by planets is expected to be small compared to any transit signal. Even a large close-in planet (0.05 AU) with the size and albedo of Jupiter ($\sim 30\,\%$[3]) would reflect approximately 8×10^{-6} of the light of its parent star, while it would cause transits of $\sim 1\,\%$ (for a Sun-like star).

The reference light curve corresponds to a $1\,R_{\rm Jup}$ planet orbiting a $1\,R_\odot$ star with $V = 10$, resulting in a transit depth of $\sim 14\sigma$. The chosen orbital distance of $a = 15.3\,R_\odot$, for a 1 month period, results in 4 transits lasting $\sim$ 15 hours each, the ingress and egress lasting approximately 3.3 hours each. The posital phase was set to 0.25.

The level of photon noise in the light curve was computed from the photon counts expected for a G2V star, based on the throughput and aperture of the *Eddington* baseline design as described in Favata & the *Eddington* Science Team (2000), i.e. a collecting area of $0.6\,\mathrm{m}^2$ and a total system throughput of 70 %. Such an instrument would detect $\simeq$ 50 photons per second from a $V = 21.5$ G2V star. The photon noise for each point in the light curves is then simulated by independent draws from a Gaussian distribution, with a standard deviation equal to the square root of the expected photon count per integration for that star.

Both versions of the algorithm were run on the reference light curve described above and the modified posterior probabilities were plotted as a function of period and as a function of phase. The number of bins used was $m = 10$ in the case of the GL method, and $n = 4$ in the case of the modified method. In order to sample the transit as well with the GL method as with the modified method, a much higher value of m

[3] `http://nssdc.gsfc.nasa.gov/planetary/factsheet/jupiterfact.html`

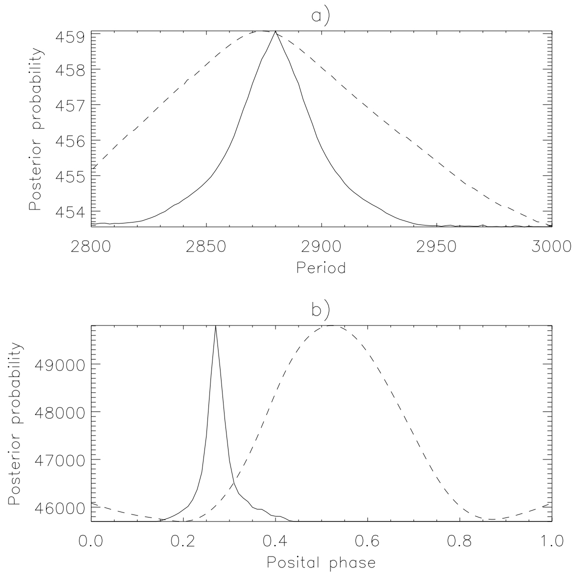

Figure 2.3: Comparison of the GL and modified GL methods for the case of a Jovian planet transiting across a 10^{th} magnitude star, with a period of 30 days (a) and a posital phase of 0.25 (b). Solid line: modified GL method. Dashed line: GL method. Both methods successfully detect the period of the transits although the peak is sharper with the modified method. The GL method is unsuccessful in the phase domain (the GL phase results are folded over the 10 bins). Note that the probabilities are expressed in arbitrary units.

would need to be used, but this would be too computationally expensive. Instead the values of m and n were chosen such that the computing times were similar. The results obtained for this benchmark case are shown in Figure 2.3.

Each of the parameters (be they associated with the light curve or with the model) was then varied over a small range of representative values. These one-off tests on a small parameter space confirmed some expected trends.

- For a given light curve duration the detection is less precise for longer periods as the light curve contains fewer transits.
- As expected, the original GL method is not well suited to detecting the phase, as there is no natural way of labelling one particular bin as the first one. On the other hand the phase is very successfully recovered with the modified version.
- The larger the value of m (GL method), the sharper the detection. However, $m = 10$ appeared sufficient for our purposes.
- Increasing the value of n (modified method) does not necessarily improve the detection ability since one starts to fit the noise inside the transits, which is not periodic. When fitting Gaussian profiles it is standard to require a minimum of 2 bins per FWHM. The shape of the transit is not Gaussian but it is relatively simple, hence we multiplied by a safety factor of 2, leading to $n = 4$ in further calculations. However when dealing with a particular value of d it is advantageous to choose n so that d is a multiple of it to avoid introducing extra noise by splitting individual data points across bin boundaries.
- Although the modified method should in principle allow us to determine the duration of the transit, in practice this is not successful. The program may be fitting a much wider region than the transit itself. In the GL method, as there

are only 10 to 20 bins per period, with p of order several hundred sampling times or more, the bin in which the transit falls is much larger than the transit itself. We have seen that the loss of information this implies does not prevent the detection of the period by the GL method. The modified algorithm is likely to overestimate the transit duration because fitting a region larger than the transit does not significantly reduce the likelihood. For now the duration of the transit was simply marginalised over; once the presence of a transit is asserted and its period known, phase folding should allow a fairly quick determination of the shape and duration;

- For a given set of parameters, with $m = 10$ and $n = 4$, such that both algorithms have similar computing times, the detection peaks are much sharper with the modified version.

2.1.3 Performance evaluation in white Gaussian noise

2.1.3.1 Method

As mentioned in Section 2.1.1.6, the use of a window function to remove the effects of under-sampling in the phase domain, while minimising the computing time, rules out the possibility of direct computation of a global odds ratio, whose value could be used to determine whether or not a given light curve contains transits. Instead, to evaluate the performance of the algorithm, it was run on simulated light curves with similar noise characteristics, some containing transits and some not, and the results were compared.

This method was previously used in a similar context by Doyle et al. (2000). For each set of trial parameters the algorithm was run first on a set of one hundred simulated light curves containing only Gaussian noise and no transits. Subsequently it was run on another set of one hundred simulated light curves containing Jovian-type planetary transits with the characteristics described in Section 2.1.2, with the same level but different realisations of the photon noise, and with uniformly distributed random phases

For each simulation, the modified posterior probabilities were plotted versus period and the value of the maximum was noted. This maximum is our `detection statistic', on the basis of which we wish to determine whether there is a transit or not. We then plot a histogram of the detection statistics measured for all the light curves with transits and one histogram for all the light curves with noise only. In other words, one histogram corresponds to the cases where the transit hypothesis is correct and one to the cases where the null hypothesis is correct. Ideally, the two histograms should be completely separated, with no overlap, and choosing a detection thresh-

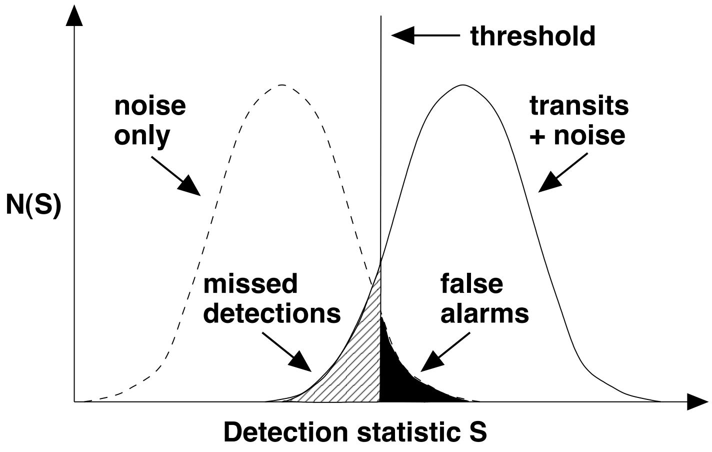

Figure 2.4: Schematic diagram of the performance evaluation method. Solid line: detection statistic distribution for the light curves with transits. Dashed line: idem for the transit-less light curves. Vertical solid line: threshold. Hashed area: missed detections. Filled area: false alarms.

old located between the two histograms would guarantee a 100 % detection rate and a 0% false alarm rate. In practice, for the cases of real interest, close to the detectability limit, the two histograms will overlap. A compromise has to be found by choosing a threshold which minimises a penalty factor designed to take into account both false alarm and missed detection rates. This is illustrated in Figure 2.4.

Depending on the circumstances, it may be more important to minimise the false alarm rate than the missed detection rate. This is the approach followed by Jenkins et al. (2002), on the basis that detections from space experiments are hard to follow-up from the ground. An alternative view is any real transit that is rejected is a loss of valuable scientific information. As long as the false alarm rate is kept to a manageable level, further analysis of the light curves will prune out the false events. We have opted here for an intermediate position, and our penalty factor is simply the sum of the missed detection rate N_{MD} and the false alarm rate N_{FA}:

$$F_{\mathrm{penalty}} = N_{\mathrm{FA}} + N_{\mathrm{MD}}. \tag{2.18}$$

After marginalisation over the other parameters, the detection algorithm yields modified posterior probabilities as a function of period and as a function of phase. The simultaneous use of the two detection statistics S_{per} and S_{ph} (plotting 2–D rather than 1–D distributions) increases the discriminating power of the algorithm, (as long as the two distributions do not have secondary maxima in 2–D space). This is shown when comparing the false alarm and missed detection rates obtained from period and phase information separately and together. The threshold in the 2–D case takes the form of a line: $S_{\mathrm{ph}} = a + b \times S_{\mathrm{per}}$. Here the optimal values of a and b were found by trial and error, although standard discriminant analysis techniques could be used to determine them automatically.

2.1.3.2 A preliminary test case

In Defaÿ (2001), analysis performed on the basis of 200 bootstrap samples for the COROT observations of a star with V = 13 and an Earth-sized planet, containing 6 transits lasting 5 hours each, yielded a probability of true detection of $\simeq$ 0.3. We performed the simulations described in Section 2.1.3.1 for a similar case: an Earth-sized planet orbiting a K5V type star with V = 13, a period of 30 days and a transit duration of 5 hours. The sampling time was 15 minutes. The noise was different from the COROT case, as we concentrate uniquely on the photon noise expected for *Eddington*.

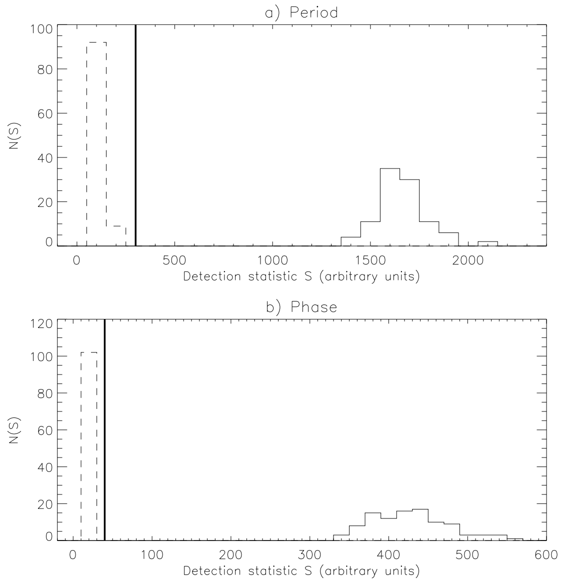

Figure 2.5: Detection statistic distributions for an Earth-sized planet orbiting a V=13 star with a 30 day period (light curve duration 120 days). Solid line: light curves with transits. Dashed line: transit-less light curves. Vertical solid line: threshold. a) Detection statistic obtained from the modified posterior probability distributions as a function of period. b) Idem as a function of phase. Over 100 realisations there were no false alarms and no missed detections.

The results are shown in Figure 2.5 for period and phase separately. As the distributions for the noise only and transit light curves are completely separated, each parameter alone is sufficient to determine a threshold ensuring null false alarm and missed detection rates.

2.1.3.3 Finding the magnitude limit of *Eddington* for Earth-like planets

Given that the key scientific goal of *Eddington* in the field of planet-finding is the detection of habitable planets, the performance of the algorithm was extensively tested for habitable planets at (or close to) the noise limit of *Eddington*. The case of an Earth-sized planet orbiting a K dwarf in a habitable orbit was used as benchmark. The light curve was simulated for a system with the following parameters:

- K5V star ($R_\star$ = 0.8 $R_\odot$) with a range of apparent V–band magnitudes V = 14.0, 14.5 and 15.0;

- Earth-sized planet ($R_\text{p} = R_\oplus$) with an orbital period of 4 months, orbiting the star at a distance of 0.64 A.U. (leading to a transit duration of $\sim$ 10.5 hours);
- light curve duration of 16 months;
- sampling time 1 hour.

An example of a light curve is shown in Figure 2.6. The resulting transit event has a depth $\Delta F/F = 1.4\times 10^{-4}$. For the *Eddington* baseline collecting area a star at $V = 14$ will result in a photon count of 1.8×10^8 per hour, so that the Poisson noise standard deviation will be 1.34×10^4. The S/N of the transit event in each 1 hour bin will thus be 1.88. Following the same reasoning for the $V = 15$ case, the S/N of the transit event in a single one hour bin is 1.19. As there are 4 transits lasting 10 hours each in the light curves considered, the overall transit signal has a S/N of $\sqrt{40} \times 1.19 \simeq 7.5$.

With the results of the simulations, an example of which is shown in Figure 2.7, the analysis described in Section 2.1.3.1 was performed for all three magnitudes, confirming that the combined use of the two statistics improves the results. This is illustrated for the $V = 14.5$ case in Figures 2.8 & 2.9 (for this particular case 1000 rather than 100 runs were computed to improve precision).

As illustrated in Figure 2.10, a mean error rate (the mean of the false alarm and missed detection rates) $< 3\,\%$ can be achieved down to $V = 14.5$. This magnitude is therefore taken as the performance limit for the algorithm for an Earth-sized planet around a K5V-type star. However this analysis is not complete enough to allow a precise determination of the magnitude limit. First the noise treatment is incomplete, photon noise only being considered. Second, one would need more runs per simulation to compute meaningful errors on the false alarm and missed detection rates. Sets of 1000 runs, as was done for the limiting $V = 14.5$ case, should be computed for all cases.

The asymmetric shape of the distributions shown in Figures 2.5, 2.8 & 2.9 implies that, even though the thresholds are chosen to minimise false alarms and missed detections equally, the optimal threshold results in more false alarms than missed detections. This could easily be avoided, if needed, by replacing Equation (2.18) by:

$$F_\text{penalty} = A \times N_\text{FA} + N_\text{MD}. \qquad (2.19)$$

where A is a factor greater than 1. Alternatively one could keep the penalty factor unchanged but set a strict requirement on the maximum acceptable false alarm rate.

As in any unbiased search for periodicity in a time-series, the inclusion of a larger range of periods in the search will lead to a higher chance of finding a spurious (noise-induced) period signal in the data. The simulations used here to assess the

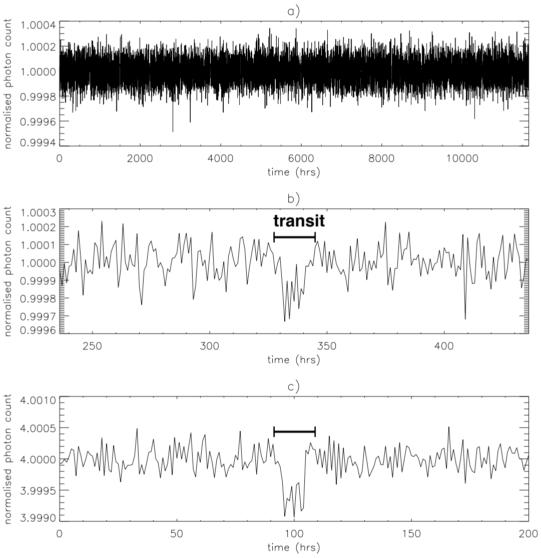

Figure 2.6: Example light curve containing 4 transits of an Earth-sized planet orbiting a K5V star with V=14.5. a) Full light curve. b) Portion around a transit. c) The four transits co-added.

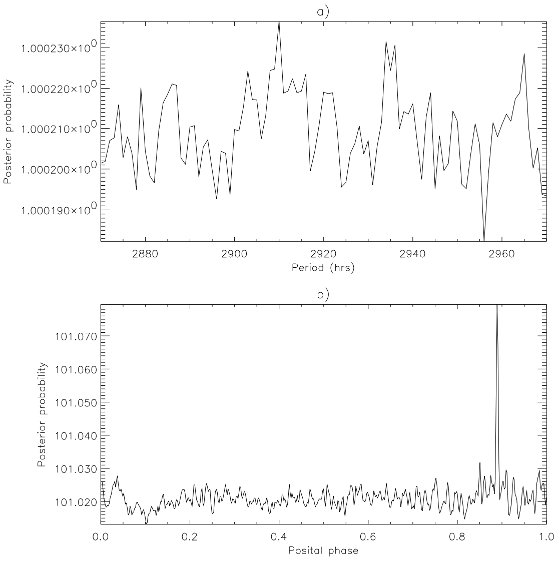

Figure 2.7: Example detection statistic distributions arising from the light curve shown in Figure 2.6 (arbitrary units). a) Period – real value 2912 hours, error -2 hours. b) Phase: real value 0.885, error 0.005.

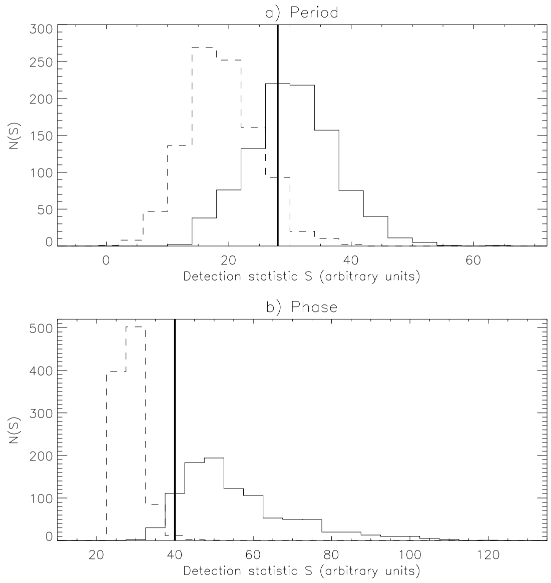

Figure 2.8: Detection statistic distributions for an Earth-sized planet orbiting a V=14.5 star with a 4 month period (light curve duration 16 months). Solid line: light curves with transits. Dashed line: transitless light curves. Vertical solid line: threshold. a) Period: 190 false alarms and 185 missed detections over the 1000 realisations. b) Phase: 27 false alarms and 14 missed detections over the 1000 realisations.

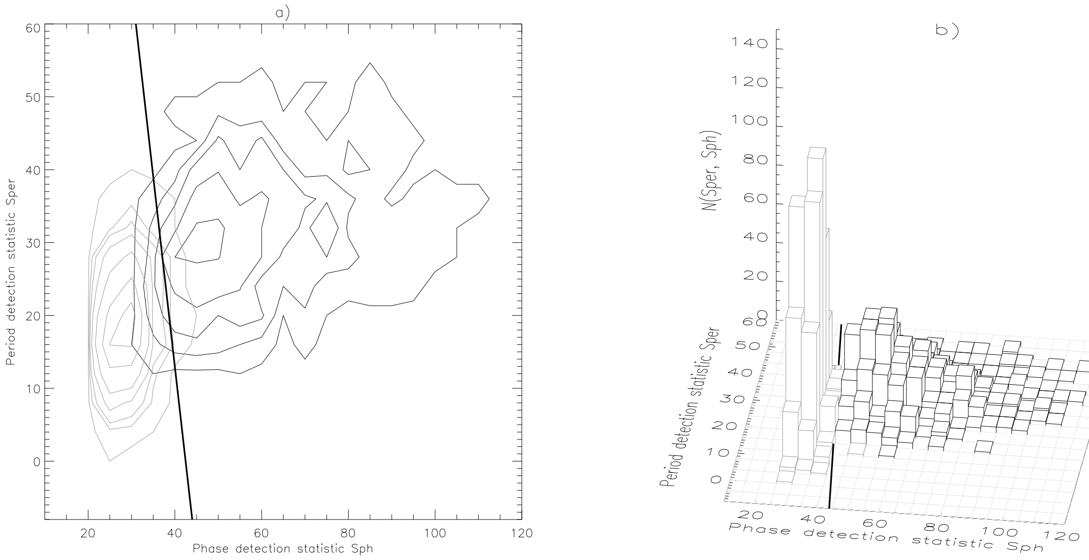

Figure 2.9: a) Contour plot and b) 3–D representation of the two-dimensional (period & phase) detection statistic distribution for an Earth-sized planet orbiting a V=14.5 star with a 4 month period (light curve duration 16 months). Black: lightcurves with transits. Grey: transit-less light curves. Solid line: threshold: $S_{\rm ph} = 42.47 - 1.191 \times S_{\rm per}$, yielding 29 false alarms and 9 missed detections over the 1000 realisations.

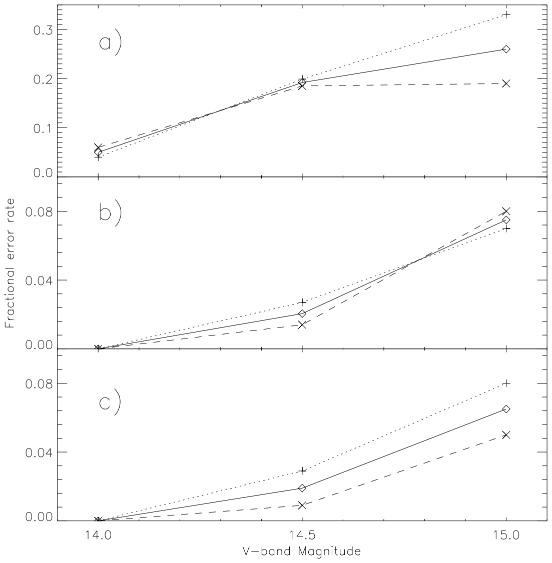

Figure 2.10: Evolution of the algorithm's performance (in terms of fractional error rates) with magnitude (period 4 months, light curve duration 16 months) a) using the period statistic only, b) using the phase statistic only and c) combining the two statistics. Dotted line: false alarm rate. Dashed line: missed detection rate. Solid line: mean error rate.

algorithm's performance are based on a search through a relatively small range of periods. In practice, lacking any a priori knowledge of the possible periodicity of planetary orbits around the star being observed, one will want to test a large range of periods, ranging from a few days (the physical limit of the period of planetary orbits) all the way to the duration of the data set (searching for individual transit events).

2.1.3.4 Data gaps

Any realistic data set will suffer from gaps in the data. While the orbits of both *Eddington* and *Kepler* have been chosen to minimise gaps, 100 % availability is not realistic, and gaps will be present due to e.g. telemetry dropouts, spacecraft momentum dumping manoeuvres, showers of solar protons during large solar flares, etc... For this reason, any realistic algorithm must be robust against the presence of gaps in the data, showing graceful degradation as a function of the fraction of data missing from the time series.

We have therefore tested the algorithm discussed here using simulated light curves with 5 %, 10 % and 20 % data gaps, randomly distributed in the data, i.e. 5 % of the points in the time series are selected randomly with a uniform distribution and removed from the light curve. The gaps will probably not be randomly distributed in reality, but as the typical gap duration is expected to be of order 1 or 2 hours, simulated random gaps can already be used to test the algorithm's robustness. For reasons of computing time, to avoid having to recalculate the 'window function' at each run, the distribution of the data gaps is the same for all runs of a simulation. As the gaps are chosen one by one there are rarely gaps of more than two consecutive time steps, i.e. 2 hours. Note that e.g. the *Eddington* mission is designed to produce light curves with a duty cycle ≥ 90 %, so that the case with 20 % data gaps represents a worst case scenario.

The results are shown in Figure 2.11. There is visibly very little degradation up to 20 % data gaps. When using S_{ph} alone or the two statistics combined there is no perceptible difference. We can therefore say this this algorithm is robust at least for data gaps of the type likely to occur due to e.g. telemetry dropouts, which last only a few hours. One would also expect the algorithm to perform well in the presence of longer gaps: the effect of gaps is to render the number of samples per bin uneven, and this is already the case for this particular method with no gaps at all.

Note that the impact of gaps was tested using a configuration closer to the detectability limit ($V = 14.5$) with the second algorithm, the box-shaped transit finder (see Section 2.2.4).

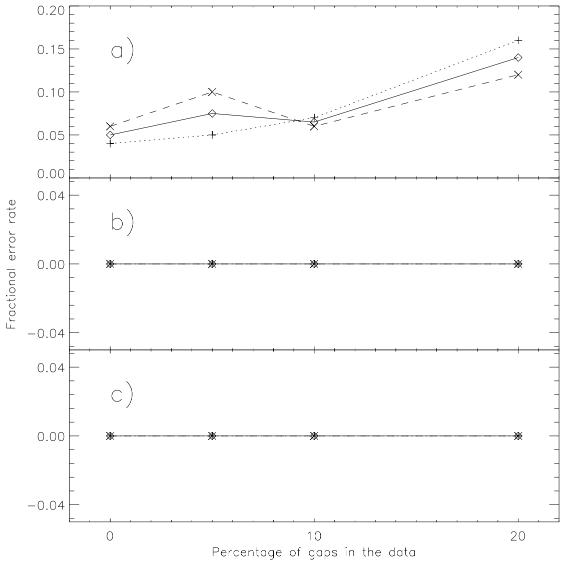

Figure 2.11: Evolution of the algorithm's performance with data gaps (period 4 months, light curve duration 16 months, V=14.0) a) using the period statistic only, b) using the phase statistic only and c) combining the two statistics. Dotted line: false alarm rate. Dashed line: missed detection rate. Solid line: mean error rate.

2.1.3.5 Number of transits in the light curve

The planetary transits detection phase of the *Eddington* mission is planned to last 3 years with a single pointing for the entire duration of that phase. There will therefore be three or four transits in the light curve for a typical habitable planet. However, other missions such as COROT are planned with shorter (5 months) pointings and it is of interest for this type of mission to study the degradation of the algorithm's performance as the number of transits in the light curve reduces. If the algorithm performs well with 2 or less transits, in the context of *Eddington* it may also allow the detection of `cool Jupiters', i.e. Jupiter-sized planets with orbits more similar to those of the gaseous giants in our solar system. This would be of relevance to the question of how typical our solar system is.

Sets of 100 runs with the characteristics specified in Section 2.1.3.3 for a star with $V = 14.5$ were computed for light curve durations of 4, 8, 12, 16 and 20 months, containing between 1 and 5 transits. The results are shown in Figure 2.12. The degradation only becomes significant when less than three transits are present. However, even mono-transits could be detectable for larger planets at that magnitude.

Defaÿ (2001) compared a matched filter approach with a Bayesian method based on the decomposition of the light curve into its Fourier coefficients. Their results suggest that the performance degradation in the low number of transits case is faster for the Bayesian method than for the matched filter. This is because the matched filter makes use of assumptions about the transit shape. It is also shown that when the Bayesian method fails to detect a transit, it can still reconstruct it if the detection is performed using a matched filter. Our algorithm has not been directly compared

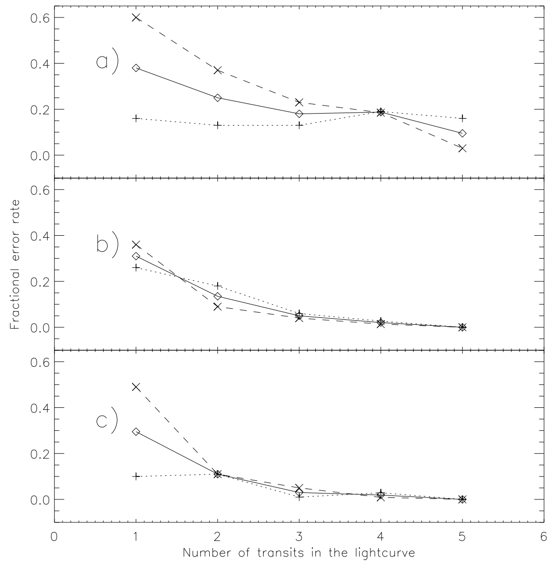

Figure 2.12: Evolution of the algorithm's performance with the number of transits in the light curve (period 4 months, V=14.5) a) using the period statistic only, b) using the phase statistic only and c) combining the two statistics. Dotted line: false alarm rate. Dashed line: missed detection rate. Solid line: mean error rate.

to a matched filter. Its very design is based on the search for a short periodic signal in an otherwise flat light curve, which is itself an assumption about the shape of the signal. The matched filter makes use of more detailed knowledge of the transit shape and is therefore likely to perform better in the low transit number limit. However our algorithm with $n = 1$ may provide already a very good approximation to the relatively simple shape that is a transit, and may therefore perform nearly as well.

2.1.3.6 Differences in the two statistics

The two a posteriori probabilities show a different behaviour. In general the phase statistic is far more discriminatory than the period statistic. This is illustrated by contour plots of the likelihood as a function of trial period and posital phase, as shown in Figure 2.13. The period statistic's lesser effectiveness may be explained in the following way. If the phase is wrong, even if the period is right, it is likely none of the transits will be matched. If the phase is right, whatever the period, at least the first transit will be matched by the model. First we consider the likelihood distribution as a function of phase, normalised over all periods. For an incorrect phase the contribution from the correct period is nil as all transits are generally missed, but for the correct phase all trial periods produce a non-negligible contribution (the correct period of course contributing most). The likelihood distribution as a function of phase is therefore sharply peaked. Then we consider the likelihood distribution as a function of period, normalised over all phases. The contribution from the correct phase is non-negligible whatever the period. When the period is correct, the contribution from

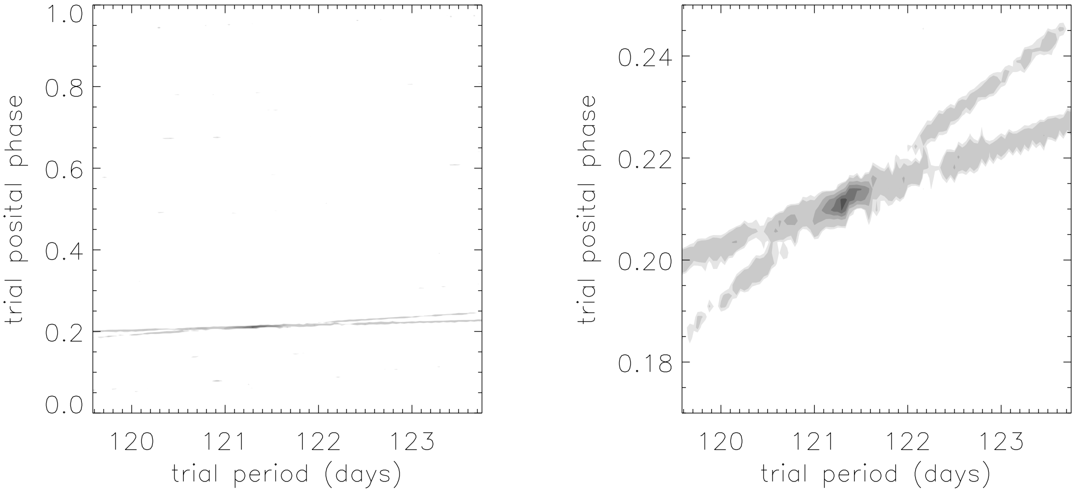

Figure 2.13: Example contour plot of the likelihood as a function of trial period and posital phase for a simulated 16 month light curve containing photon noise as for a V = 14.5 star with 1 hour sampling and transits of a 1 $R_\oplus$ planet with a period of $\sim$ 4 months (121.33 days). Left: full parameter space explored. Right: zooming in on the true posital phase, which is 0.209. The diagonal lines correspond to pairs of trial periods and phases where at least one of the four transits is matched by the model.

the correct phase is washed out by the contributions from all the incorrect phases. The likelihood distribution as a function of period is therefore less sharply peaked. Additionally, the range of periods explored was, for computational reasons, kept relatively small, so that only a small region of parameter space is covered in that direction.

However, the combined use of the two parameters is more successful than the phase statistic alone. The reason for this is illustrated in Figure 2.9: in 2–D space the two distributions are aligned on a diagonal, such that no single value cutoff is optimal in either direction, compared to the line shown. The global odds ratio described in Section 2.1.1.2 could be used for such a purpose. We have noted in Sect. 2.1.1.6 that the global odds ratio for a given light curve cannot be used as an absolute statistic in the context of the present method. It can however be used as relative detection statistic, like $S_{\rm per}$ & $S_{\rm ph}$, combined with bootstrap simulations.

The algorithm described in Section 2.2, which was derived using the lessons from the present one, directly combines phase and period information into a single statistic.

2.1.3.7 Discussion

Efficient data processing is one of the challenges for the upcoming generation of large scale searches for exo-planets through photometric transits. While radial velocity searches concentrate on limited number of stars, transit searches will investigate simultaneously large numbers of stars, and produce large amounts of data (photometric light curves) for each of them. A computationally efficient and robust algorithm for the processing of these data sets is necessary to make transit searches feasible. It is likely that the photometric time series which represent the observational product of the transit searches will be analysed in different stages, using more than a single approach. In particular, a first level of processing (after instrumental effects have been removed) should concentrate on singling out high-probability transit candidates, while efficiently pruning out the large number (more than 90 %, even if all stars have planets, due to the low probability of transit events) of light curves in which no transits are present. In this first stage of analysis the ability to efficiently screen real transits in the data – even at the price of a moderate number of false alarms – is a key requirement for the algorithm. The candidate light curves in which a transit is suspected will then later be subject to a more detailed processing, which can then afford to be computationally less efficient (given it has to operate on a much smaller amount of data).

The modified GL method is able to detect transit events at the limit of the photon noise present in the light curve. It shows a graceful degradation of its performance as a function of different parameters of interest, e.g. the noise level in the data, as well as the presence of data gaps and the number of transits actually observed. Its strong sensitivity to the phase of periodic transits supplies significant additional information to be then used by further steps of processing for e.g. the reconstruction of the transit parameters. Thus, while little used in astronomy, Bayesian algorithms appear to be a powerful tool in the processing of transit data.

However, given the robustness and computational efficiency requirements, simplicity should be the guiding factor for subsequent work on transit detection. Simplifying assumptions will be tried one by one, and those that do not degrade the performance while improving the computing time will be incorporated in the algorithm for future use.

2.1.4 Tests with solar micro-variability

The performance of the modified GL method has been evaluated for simulated data from *Eddington*-like missions, containing simulated transits and photon noise. However, the influence of stellar variability induced by activity is likely to be the main

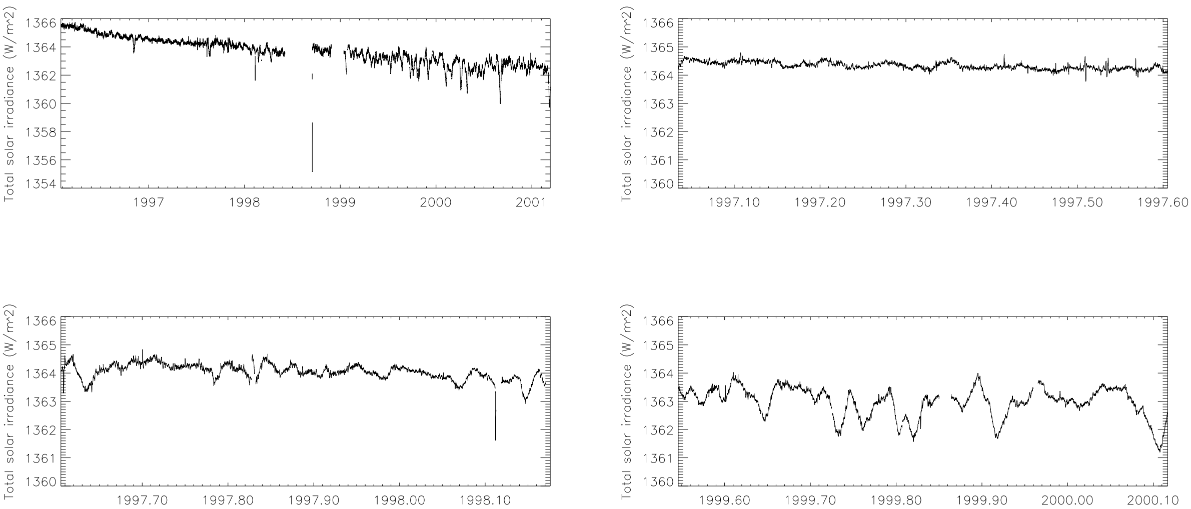

Figure 2.14: PMO6 light curves. Top left: full light curve, January 1996 to September 2001 (no correction for instrumental degradation). Top right: section used as low activity sample. Bottom left: idem for medium activity. Bottom right: idem for high activity.

limitation to *Eddington*'s ability to detect planetary transits. The present section summarises the results of preliminary tests carried out by incorporating in the simulated light curves observed variations in the total irradiance of the Sun, as measured by the PMO6 radiometer, a part of the VIRGO experiment on the SoHO satellite, at solar activity minimum and maximum and at an intermediate phase.

2.1.4.1 Light curves

The transit light curves were simulated using UTM (see Section 2.1.2), with the parameters listed in Table 2.1. The full PMO6 light curve is shown in the top panel of Figure 2.14. This dataset is discussed in more detail in Chapter 3. Three 6 month long segments were selected at low, medium and high activity (bottom 3 panels of Figure 2.14). They were chosen to illustrate a particular feature of the Sun's variability, for example sun-spot signatures or modulation on the time scale of the solar rotational period. Care was taken also to avoid very long data gaps, although there are frequent short gaps (every few days, lasting a few tens of minutes to a few hours), and in the medium and high activity samples, some gaps last a few days. (N.B. This leads in some cases to the absence of one entire transit from the light curve, a situation which was not explored in Section 2.1.3, but with which the algorithm seems to cope reasonably well.) In each set of tests performed, the relevant segment of the PMO6 light curve was rebinned to 1 hour bins, normalised to a mean value of 1,

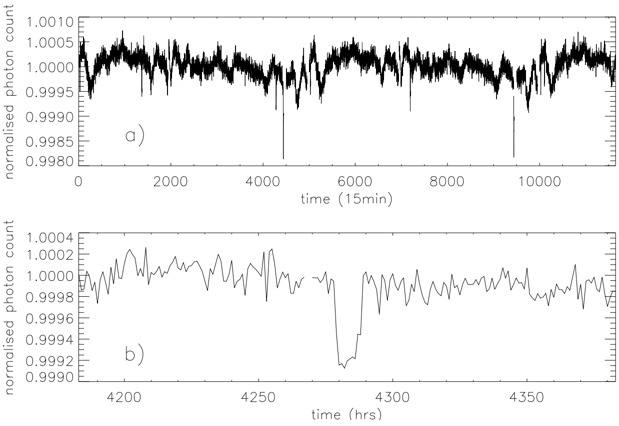

Figure 2.15: Example light curve with medium activity level, $R_\mathrm{p} = 2\,R_\oplus$. a) Full light curve. b) Portion around first transit.

repeated in order to cover the full duration of the simulated transit light curve and the product of stellar and transit light curves was taken. Care was taken to ensure that the 6 month repetition period of the PMO6 light curve was not too close to the transit period (4 months). Photon noise was then added as in the previous simulations according to the photon counts expected from the *Eddington* baseline design.

An example of the light curves produced is shown in Figure 2.15. Although individual transits are (except for the Earth–sized case) easily visible when zooming in on a portion of the curve, the most visible signal in the full curve is activity.

2.1.4.2 Results

The algorithm's performance with these light curves was tested using the method described in Section 2.1.3.1. However, only 10 realisations of the noise were computed in each case, these tests being intended as exploratory rather than systematic. As a result, the threshold analysis described in Section 2.1.3.1 would have been meaningless. Instead, results were recorded in terms of the number, if any, of detections which were not in fact transits but activity-induced features, i.e. the number of times the highest peak in the detection statistic distribution did not correspond to the tran-

Table 2.1: Parameters used for the simulated light curves with solar variability. Transit durations of 5 & 15 hours were also tested in the low activity case with a $2\,R_\oplus$ planet.

Star:		Planet:	
spectral type	K5V	radius	10, 8, 6, 4, 2 & 1 $R_\oplus$
radius	$0.74\,R_\odot$	orbital period	4 months
activity level	low, med, high	orbital distance	0.43 A.U.
V–band magnitude	14.5	transit duration	10 hours

sit's period or phase. For each such case, the corresponding transit-less light curve (with identical noise and stellar variability realisations, but no transits) was used to check that the spurious detection was due to activity and not photon noise.

Table 2.2 gives an overview of the results in terms of the number of spurious detections for each set of 10 light curves with transits. We ascertain whether the detected peak corresponds to the correct transit period or phase by comparing with the distributions generated from transit–less light curves and by requiring that the error in the detected value be less than 2 hours (period) or 0.006 (posital phase).

As can be seen in table 2.2, the performance starts to degrade at 1, 2 and $2\,R_\oplus$ for low, medium and high activity levels respectively. Noticeably, the breakdown is sudden, because it occurs when the transit induced peaks, whose height depends on the transit depth, i.e. on the planet radius, become smaller than the activity-induced peaks (which are of constant height and shape, as the same PMO6 sequence was always used for a given activity level). Although $V = 14.5$ was found to be the limiting magnitude for the reliable detection of an Earth–sized planet in the absence of activity (see Section 2.1.3), variability rather than photon noise is the limiting factor of the algorithm's performance, even in the low activity case. The results obtained here are therefore unlikely to improve much with increasing brightness. As the Sun is also a relatively quiet star for its type, and as other types of stars are likely to be generally more active, the problems outlined by these results are likely to have a serious impact on *Eddington*'s performance and need to be addressed.

2.1.4.3 Example cases

It is helpful to highlight certain characteristics of the detection statistic distributions in a few representative cases in order to suggest ways to address the activity problem.

Figure 2.16 shows the distributions obtained for a planet twice the size of the Earth at the three activity levels. The middle row corresponds to the light curve illustrated in Figure2.15. As activity increases, spurious activity-induced features appear in the distributions, and at high activity the peak corresponding to the actual transit is

R_p	low		medium		high	
$10\,R_\oplus$	0	0	0	0	0	0
$8\,R_\oplus$	0	0	0	0	0	0
$6\,R_\oplus$	0	0	0	0	0	0
$4\,R_\oplus$	0	0	0	0	0	0
$2\,R_\oplus$	0	0	0	10	10	10
$1\,R_\oplus$	10	9	10	8	10	10

Table 2.2: Number of cases where the highest peak in the posterior probability distribution is spurious (activity rather than transit induced) over each set of 10 noise realisations as a function of activity level (columns) and planet radius (rows). The first and second number in each cell relate to the period and phase detection statistic respectively.

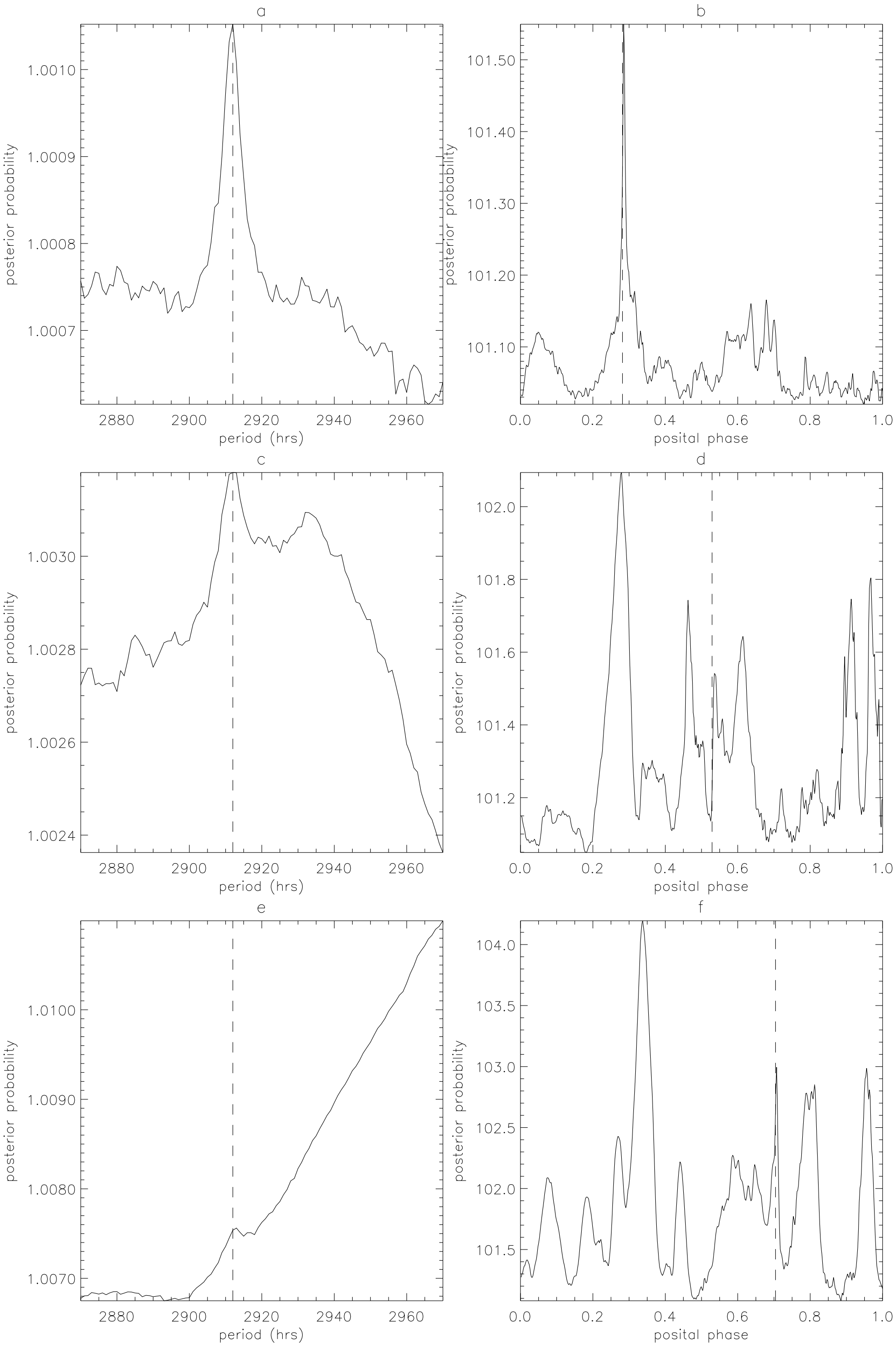

Figure 2.16: Example detection statistic distributions for period (left column) and phase (right column) at low (top row) , medium (middle row) and high (bottom row) activity levels, with R_p =2 $R_\oplus$. In each panel the dashed line indicates the true value of the period or epoch.

no longer detected. The activity-induced features are particularly noticeable in the phase distributions, each peak corresponding to the first of the model light curve's transits being phase-folded onto (say) a sunspot transit. The period of repetition of the PMO6 sequences is well outside the range of periods tested, which avoids the detection of a peak at that period, but there is a discernible trend, which reduces the sensitivity of the period statistic.

Importantly, at the same magnitude, a similar planet was easily detected in the absence of variability. This implies that photon noise is not the limiting factor, hence that the same planet would not be detected around a brighter but active star.

2.1.4.4 Implications

The first step towards understanding the significance of the results presented here is to asses how typical the Sun's variability is, relative to other Sun-like stars and to stars of other spectral types and ages. Very little information on stellar activity on this kind of amplitude and timescale is available at present but there are a number of exploitable datasets either existing or expected within the next few years, which can be used to construct and calibrate a simple model of stellar micro-variability applicable to the variety of planetary transit search stars. A number of pre-processing techniques may be used to remove as much of the micro-variability signal as possible, before applying the modified GL method, or other methods designed primarily for light curves with white Gaussian noise. These topics are discussed in Chapters 3 and 4 respectively.

Given a more qualitative understanding of how micro-variability varies from star to star, and how it will impact transit detection, several aspects of the design of missions such as COROT, *Eddington* & *Kepler* may be affected. These include the choice of target field(s), which should be optimised to contain as many as possible of the least variable stars, or those whose variability can easily be filtered out. The observing strategy, for example the sampling time, may also be affected. Finally, it may be important to make use of additional information besides single bandpass photometry, for example to use the colour signature of the detected events to assess their planetary or stellar origin.

2.2 A stripped-down box-shaped transit finder

In Section 2.1, a dedicated Bayesian transit search algorithm was derived, based on the more general GL method for period finding. Here we develop this algorithm further and attempt to reconcile the apparent diversity of the extant transit algorithms.

Starting afresh from the original GL prescription, appropriate sequential simplifications can be made. We demonstrate that the levels of the step-function bins – which define the shape of the detected event – are not free parameters, their optimal values being fully defined by the data. The use of Bayesian priors can be dropped, given the lack of information currently available on the appropriate form for these priors. Finally, for detection purposes, the model can be simplified to an unequal bin duration square wave with only one out-of-transit and one in-transit value, where the out-of-transit section lasts much longer than the in-transit section.

After a brief aside on the close links between different families of detection methods in white Gaussian noise (Section 2.2.1), the algorithm itself is derived in Section 2.2.2, and its implementation is presented in Section 2.2.3. Its performance in terms of both detection capability and computational requirements is compared to that of the GL method in Section 2.2.4, and the results are discussed in Section 2.2.5.

2.2.1 Likelihood maximisation in Gaussian noise

Transit searches are generally performed by comparing light curves to a family of models with a common set of parameters, differing from each other according to the different values used for these parameters. A variety of methods exists to identify the best set of parameters. The most commonly used, in astronomy, is probably the matched filter, shown by Kay (1998) to be the optimal detector in the presence of white Gaussian noise. In the present section we show how the matched filter is derived from likelihood maximisation, and its equivalence to χ^2 minimisation and to a simple cross-correlation method. This is by no means a new result, but it helps to clarify the very close links that exist between the variety of transit search methods published in the literature, which are based on all of the above approaches, and to which the box-fitting algorithm will later be compared.

If the noise in each data point is assumed to be independently drawn from a Gaussian distribution (an assumption also valid for Poisson noise in the limit of large numbers of photons), the likelihood (or probability that the observed data is the result of adding noise to the model) can be written as the product of independent Gaussian probability distribution functions:

$$\mathcal{L} = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i - r_i)^2}{2\sigma_i^2} \right] \right\}, \tag{2.20}$$

where x_i is the flux (or count, or magnitude) value at time t_i and r_i is the corresponding model value, N is the total number of data points and σ_i the expected error on

x_i. Equation (2.20) can be rewritten as

$$\mathcal{L} = \left(\frac{1}{2\pi}\right)^{(N/2)} \times \prod_{i=1}^{N}\left(\frac{1}{\sigma_i}\right) \times \exp\left(-\frac{\chi^2}{2}\right), \tag{2.21}$$

where

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{(x_i - r_i)^2}{\sigma_i^2}\right], \tag{2.22}$$

so that likelihood maximisation, in the case of Gaussian noise, is equivalent to χ^2 minimisation, since the noise properties σ_i are assumed to be known, i.e. fixed.

Expanding Equation (2.22) gives

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{x_i^2}{\sigma_i^2}\right] + \sum_{i=1}^{N}\left[\frac{r_i^2}{\sigma_i^2}\right] - 2\sum_{i=1}^{N}\left[\frac{(x_i \times r_i)}{\sigma_i^2}\right]. \tag{2.23}$$

The first term is the error-weighted sum of all the data points, and is constant whatever the model. The second term is the equivalent sum for the model. In the case of shallow, short duration transits and provided the errors are relatively constant over the timespan of the observations, this term can also be considered constant. In such circumstances likelihood maximisation, or χ^2 minimisation, is therefore equivalent to maximising the third term, which is a zero-offset cross-correlation between the model and the data, i.e. a generalised matched filter:

$$\mathcal{MF} = \sum_{i=1}^{N}\left[\frac{(x_i \times r_i)}{\sigma_i^2}\right]. \tag{2.24}$$

In the case of transit searches, one of the parameters to be adjusted is the phase of the transit(s), i.e. the time at which the first model transit starts. Different phases are tested by simply introducing an offset between the model and the data, leading to formulation of the problem as a general cross-correlation:

$$\mathcal{CC} = \max\left\{\sum_{i=1}^{N}\left[\frac{(x_i \times r_{i+n})}{\sigma_i^2}\right]\right\}_{n=1}^{N}. \tag{2.25}$$

where r is now defined for a unique reference epoch.

2.2.2 Simplification of the algorithm

2.2.2.1 Optimum χ^2 calculation for a generalised step-function model

We consider here a general periodic step-function model of the type used in the original GL method, characterised by the following parameters: number of bins m, period p, and epoch e (time elapsed between the start of the 1st bin and the start of the light curve) and bin levels $R = \{r_1, r_2, \ldots, r_m\}$. By directly maximising the likelihood, or in this case minimising χ^2, for such a model, it is straightforward to show that whatever the number and relative duration of the bins, the optimal values for the bin levels can be determined directly from the data given m, p and e. If we refer to the contribution from bin j to the overall χ^2 as χ_j^2, and define J as the ensemble of indices falling into bin j, we have

$$\chi_j^2 = \sum_{i \in J} \left[\frac{(x_i - r_j)^2}{\sigma_i^2} \right]. \tag{2.26}$$

The value $\widetilde{r_j}$ of the model level r_j that minimises χ_j^2 is then simply given by the standard inverse variance-weighted mean of the data inside bin j, since by setting $\partial \chi_j^2 / \partial r_j$ to zero we have

$$\frac{\partial \chi_j^2}{\partial r_j} = 2 \sum_{i \in J} \left(\frac{x_i - r_j}{\sigma_i^2} \right) = 0, \tag{2.27}$$

hence

$$\widetilde{r_j} = \overline{x_j} = \left[\sum_{i \in J} \sigma_i^{-2} \right]^{-1} \sum_{i \in J} x_i \sigma_i^{-2}. \tag{2.28}$$

Substituting into Equation (2.26), χ_j^2 now becomes

$$\widetilde{\chi_j^2} = \sum_{i \in J} \left[\frac{\left(x_i - \overline{x_j}\right)^2}{\sigma_i^2} \right], \tag{2.29}$$

where $\widetilde{\chi_j^2}$ denotes the minimised value of χ_j^2 for a given period, epoch and number of bins. The contribution from each of the m bins can be simplified by expanding Equation (2.29):

$$\widetilde{\chi_j^2} = \sum_{i \in J} \left[\frac{x_i^2 - 2 x_i \overline{x_j} + \overline{x_j}^{\,2}}{\sigma_i^2} \right]; \tag{2.30}$$

$$\widetilde{\chi_j^2} = \sum_{i \in J} \frac{x_i^2}{\sigma_i^2} - 2 \overline{x_j} \sum_{i \in J} \frac{x_i}{\sigma_i^2} + \overline{x_j}^{\,2} \sum_{i \in J} \frac{1}{\sigma_i^2}. \tag{2.31}$$

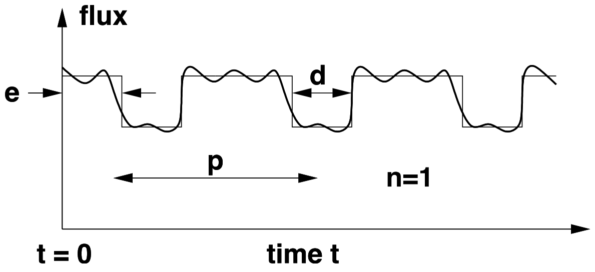

Figure 2.17: Schematic illustration of the type of model used in the box-fitting method. The is a single in-transit bin (cf. Figure 2.2, where the number of in-transit bins is 4) of duration d and one out-of-transit bin of duration p − d where p is the period. The epoch e is defined as the time elapsed between the start of the data and the start of the next transit.

From Equation (2.28) we have

$$\sum_{i \in J} \frac{x_i}{\sigma_i^2} = \overline{x_j} \sum_{i \in J} \frac{1}{\sigma_i^2}, \tag{2.32}$$

so that

$$\widetilde{\chi_j^2} = \sum_{i \in J} \frac{x_i^2}{\sigma_i^2} - \overline{x_j}^{\,2} \sum_{i \in J} \frac{1}{\sigma_i^2}. \tag{2.33}$$

The overall minimised χ^2 over all bins is thus

$$\widetilde{\chi^2} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - \sum_{j=1}^{m} \left[\overline{x_j}^{\,2} \sum_{i \in J} \frac{1}{\sigma_i^2} \right]. \tag{2.34}$$

The first term in Equation (2.34) is entirely independent of the model, and hence stays constant, so that only the second term needs to be calculated for each set of trial parameters.

2.2.2.2 Making use of the known characteristics of planetary transits

The Gregory-Loredo method makes no assumptions about the shape of the variations, and is fairly computationally intensive. However, when trying to detect planetary transits, most of the information is concentrated in a very small portion of the light curve. In Section 2.1, we adapted the Gregory-Loredo method to the planetary transit case by having one long out-of-transit bin (bin 0) and *n* short in-transit bins (see Figure 2.2). The value of *n* used was typically 4. For a given *n*, the parameters defining each candidate model are then *p*, *e*, and the transit duration *d*. The likelihood computation was carried out as described in G99.

This algorithm performed well when tested on simulated data (with photon noise only), but the likelihood calculation was still computationally intensive. A number of improvements were made following the completion of the simulations described in Section 2.1.3:

1. Given the current state of exo-planet research, the use of Bayesian priors is not expected to contribute significantly to the performance of the algorithm at the detection stage. The information available on period and duration distributions is relatively scarce for giant planets, and non-existent for terrestrial planets. The priors used in the modified GL method were generic and mostly identical to those used by G99 for X-ray pulsars, rather than specifically optimised for transit searches.

2. Using the χ^2 rather than the likelihood as a detection statistic, and implementing the calculation as outlined in Section 2.2.2.1, significantly reduces the computational requirements of the detection process.

3. The shape of most planetary transits is sufficiently simple that, for detection purposes (as opposed to detailed parameter estimation), a single in-transit bin, as illustrated in Figure 2.17, provides enough information. A significant advantage of this simplification is that it makes the method far more robust and capable of coping with real data, and all its concomitant problems, with negligible loss in detection efficiency.

4. Once a detection is made, a shape-estimation phase with either a large value of *n*, or by detailed model fitting of the phase folded light curve, can be implemented. As the dependency of transit shapes as a function of the stellar and planetary parameters is relatively well-known, Bayesian priors may have a part to play in this phase. This is, however, outside the scope of the present chapter.

2.2.2.3 χ^2-minimisation with a box shaped transit.

The algorithm used in the present paper evolved from that of Section 2.1 taking into consideration the points listed in Section 2.2.2.2. The model therefore consists of one out-of-transit bin and a single level in-transit bin. (Although this simplification may seem disingenuous, by suitably pre-processing, or adaptively filtering, the signal to remove intrinsic stellar variability, this is a valid approximation to transit detection in practice.) All the data points falling into the out-of-transit bin form the ensemble *O*, while those falling into the in-transit bin form the ensemble *I*. No Bayesian priors are used. Adapting Equation (2.34) to this model gives

$$\widetilde{\chi^2} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - \overline{x_O}^{\,2} \sum_{i \in O} \frac{1}{\sigma_i^2} - \overline{x_I}^{\,2} \sum_{i \in I} \frac{1}{\sigma_i^2}. \tag{2.35}$$

Provided the transits are shallow and of short duration (i.e. the most common case), the ensemble *O* contains the vast majority of the data points, so that $\overline{x_O} \approx \overline{x}$ (where

$\overline{x}$ is the weighted mean of the entire light curve). Substituting this approximation into Equation (2.35):

$$\widetilde{\chi^2} \approx \sum_{i=1}^{N} \left\{ \frac{x_i^2}{\sigma_i^2} - \frac{\overline{x}^2}{\sigma_i^2} \right\} - \overline{x_I}^{\,2} \sum_{i \in I} \frac{1}{\sigma_i^2}. \tag{2.36}$$

The first two terms in Equation (2.36) are constant. The minimisation of χ^2 is therefore achieved by maximising the statistic S^2, given by

$$S^2 = \overline{x_I}^{\,2} \sum_{i \in I} \frac{1}{\sigma_i^2}, \tag{2.37}$$

which can also be expanded as:

$$S^2 = \left(\sum_{i \in I} \frac{x_i}{\sigma_i^2} \right)^2 \left(\sum_{i \in I} \frac{1}{\sigma_i^2} \right)^{-1}. \tag{2.38}$$

If the light curve is robustly 'mean-corrected' prior to running the algorithm, such that x_i is replaced by Δx_i, $\overline{x_I}$ becomes $\overline{\Delta x_I}$, the depth of the model transit. This results in a further simplification where the only free parameters are now the phase, period, and duration of the transit, since the depth is determined given the other three. It is also apparent that S^2 is simply equal to the square of the in-transit signal-to-noise ratio. This is easier to see in the case where $\sigma_i = \sigma$ for all i (a good approximation to the case for space data). Equation (2.38) then becomes

$$S = \mathrm{SNR} = \frac{\sum_{i \in I} \Delta x_i}{n_I^{1/2} \sigma} = \frac{n_I^{1/2} \times \overline{\Delta X}}{\sigma}, \tag{2.39}$$

where n_I is the number of points in I, and $\overline{\Delta X} = \sum_{i \in I} \Delta x_i / n_I$ is the mean of the in-transit points, i.e. the model transit depth (the weighting being unnecessary in that case).

Equation (2.39) is used when the errors are constant, or when no individual error estimates are available for each data point. In the latter case, the Median Absolute Deviation (MAD) of the dataset is used to estimate σ, as this is more robust to outliers than a simple standard error estimate (Hoaglin et al. 1983). For a Gaussian distribution $\sigma_{rms} = 1.48 \times$ MAD and this factor is used throughout to scale the MAD sigmas. If individual error estimates are available, Equation (2.38) provides a more precise estimate of S at the cost of a slight increase in computation time.

If the noise is Gaussian, a theoretical signal-to-noise threshold (i.e. S threshold) can in principle be computed a priori to keep the false alarm rate below a certain value (Jenkins et al. 2002).

2.2.2.4 Detection of triangular or curved eclipses with a box-shaped model

The use of a single class of simple, box-shaped model greatly simplifies the problem, but how does it affect the sensitivity to highly triangular (grazing) eclipses? This can be quantified as follows.

Let $x(t)$ be a light curve which contains a single triangular eclipse of duration d and depth D starting at time e, and has a constant noise level σ. If we analyse this light curve using a matched filter[4] with a model $r(t)$ of similar (triangular) shape, the maximum S_r in the matched filter statistic is obtained when the model parameters d_r, D_r and e_r are equal to d, D and e respectively. The expectation value of S_r is thus

$$\langle S_r \rangle = \int_0^T x(t)\, r(t)\, dt = \int_0^T r^2(t)\, dt = 2 \int_0^{d/2} r^2(t')\, dt'. \tag{2.40}$$

where $t' = t - e$. In the range $0 \leq t' \leq d/2$, $r(t') = -2\, D\, t'/d$, so that

$$\langle S_r \rangle = \frac{8\, D^2}{d^2} \int_0^{d/2} t'^2\, dt' = \frac{d\, D^2}{3}. \tag{2.41}$$

The variance of S_r is given by

$$V(S_r) = \int_0^T \sigma^2\, r^2(t)\, dt = \sigma^2 \int_0^T r^2(t)\, dt = \sigma^2\, \langle S_r \rangle, \tag{2.42}$$

so that the signal-to-noise ratio of S_r is

$$\mathrm{SNR}_r = \frac{\langle S_r \rangle}{\sqrt{V(S_r)}} = \frac{d^{1/2}\, D}{3^{1/2}\, \sigma}. \tag{2.43}$$

Now let us perform the same analysis with a box-shaped dip model $b(t)$ of duration d_b and depth D_b starting a time e_b. The maximum in the detection statistic occurs when the centre of the dip in the model coincides with that of the eclipse in the light curve, i.e. when $e_b + d_b/2 = e + d/2$. One can see immediately that the optimal model duration and depth fulfil $d_b \leq d$ and $D_b \leq D$. The expectation value of the detection statistic for this model, S_b, is

$$\langle S_b \rangle = \int_0^T x(t)\, b(t)\, dt = 2 \int_{d/2 - d_b/2}^{d/2} \frac{2\, D\, t'}{d}\, D_b\, dt. = \frac{2\, D\, D_b}{d} \int_{d/2 - d_b/2}^{d/2} 2\, t'\, dt' \tag{2.44}$$

$$\langle S_b \rangle = \frac{D\, D_b\, d^2}{2\, d} \left[1 - \left(1 - \frac{d_b}{d} \right)^2 \right] = \frac{D\, D_b\, d}{2} \left[2\, \frac{d_b}{d} - \left(\frac{d_b}{d} \right)^2 \right], \tag{2.45}$$

[4] As outlined in Section 2.2.1, a matched filter is equivalent to a χ^2 minimisation approach.

while its variance is

$$V(S_b) = \int_0^T \sigma^2\, b^2(t)\, dt = \sigma^2 \int_0^{d_b} D_b^2\, dt = \sigma^2\, D_b^2 d_b, \tag{2.46}$$

so that

$$\mathrm{SNR}_b = \frac{\langle S_b \rangle}{\sqrt{V(S_b)}} = \frac{D\, d^{1/2}}{2\,\sigma} \left(\frac{d_b}{d}\right)^{-1/2} \left[2\,\frac{d_b}{d} - \left(\frac{d_b}{d}\right)^2\right]. \tag{2.47}$$

Maximising SNR_b with respect to d_b/d yields $d_b/d = 2/3$ and hence

$$\mathrm{SNR}_b = \frac{2^{3/2}\, d^{1/2}\, D}{3^{3/2}\, \sigma}. \tag{2.48}$$

The difference in sensitivity is given by the ratio Equations (2.48) to (2.43):

$$\frac{\mathrm{SNR}_b}{\mathrm{SNR}_r} = \frac{2^{3/2}}{3} \approx 0.9428. \tag{2.49}$$

The loss of sensitivity is thus small: although the best-fit box-shaped model doesn't cover all the eclipse, most of the signal is concentrated in the central part which is covered. The expected loss for curved transits (where limb-darkening is important) is even smaller, as these resemble a box-shape more closely than triangles do.

2.2.2.5 Comparison with other transit search techniques

In following through the steps of the previous sections our prime motives were to modify a general purpose Bayesian periodicity estimation algorithm to make it simpler, faster and more robust. In so doing we have arrived at a very similar formulation to that developed by other authors, though the details of the implementation differ. For example, Kovács et al. (2002) derived and tested a box-fitting method (BLS) similar to the present algorithm on simulated ground based data with white noise, and showed that significant detections followed for in-transit signal-to-noise ratios (our S statistic) greater than 6.

Street et al. (2003) used a transit finding algorithm based on a matched filter technique. After identifying and removing large amplitude variable stars they generated model light curves consisting of a constant out-of-transit level and a single in-transit section. The models were generated for a series of transit durations and phases, and a χ^2-like measure was then used to select the best model (indeed their Equation 3 is essentially a special case of the method derived in Sect. 2.2.1 for single transits).

Udalski et al. (2002b), who made the first direct detections of transiting planetary candidates later to be confirmed with the radial velocity method, also implemented a version of the BLS algorithm and noted that it was much more efficient than their own algorithm based on *"a simple cross-correlation with an error-less transit light curve"* (Udalski et al. 2002a).

In a comparison of several transit finding algorithms, Tingley (2003a) found that matched filters and cross-correlation gave the best results compared with progressively more general methods ranging from BLS, through Deeg's method (Doyle et al. 2000) to Defaÿ's (Defaÿ et al. 2001) Bayesian approach. The fact that matched filters and cross-correlation methods give good results is hardly surprising, and can easily be deduced from the χ^2 minimisation developed in Section 2.2.1. The more general methods suffer from the added complexity of the underlying model, which through the Bayesian view of Occam's Razor, reduces the tightness of the posterior probability distribution of the parameter estimation. What is however surprising, is that the BLS method did not give at least as good a result as the matched filter and cross-correlation methods. We would expect the BLS method to have similar performance to the matched filter as it is mathematically almost identical. In fact, the same author published more recently a revised comparison, in which he implemented the removal of the transit depth as a parameter in the BLS method, along the same lines as advocated here. After additional modifications to make the comparison method more rigorous, this modified BLS – now even closer to the present algorithm – compared well with the matched filter and cross-correlation (Tingley 2003b).

2.2.3 Improvements in the implementation

2.2.3.1 Optimised parameter space coverage

The formulation of the detection statistic presented in Section 2.2.2.3 is fully defined given only the dataset and the start and end times of each model transit. The model parameters are thus the duration d, period p and epoch e (defined for our purposes as the time at the start of the first transit in the dataset).

The range of expected transit durations is relatively small – from a few hours for close-in, rapidly orbiting planets, to almost a day for the most distant planets transiting more than once within the timescale of the planned observations. A simple discrete sampling prescription can therefore be adopted for the duration without leading to large numbers of trial values. One option is to choose the step δd between successive trial durations to be approximately equal to the average time step δt between consecutive data points. This ensures that models with the same period and epoch and neighbouring trial durations differ on average by ~ 1 data point per transit. However, if the observation sampling rate is high – such as the sampling rate of 10 min envisaged for most targets for *Eddington* in planet-finding mode (Favata

2004) – a larger step in duration can be used, provided it is smaller than the shortest significant feature in the transit, namely the ingress and egress, which have typical durations of $\sim$ 30 minutes.

The period sampling prescription is designed to ensure that the error in the phase (or equivalently epoch) of the last model transit in the light curve is smaller than a prescribed value. Capping the error on the period (by using a constant trial period step) is not sufficient, as the error on the epoch of the n^{th} transit will be n times the error on the epoch of the first. This would lead to a larger overall error for shorter periods, where the number of transits in the light curve is large, thus introducing a bias in the distribution of detection statistic with period. This bias is not present if one uses a constant step in trial frequency. Defining the relative frequency $\nu_{\mathrm{r}} = T/p$, T being the total light curve duration, the phase of an event occurring at time t is given by $\phi = 2\pi t/p = 2\pi t \nu_{\mathrm{r}}/T$, so that for the last transit in the light curve $\phi \approx \phi_{\mathrm{max}} = 2\pi\nu_{\mathrm{r}}$. A fixed step in ν_{r} thus leads to a fixed error in ϕ_{max}. By trial and error, a value of 0.05 was found to be suitable for $\delta\nu_{\mathrm{r}}$.

One caveat in the case of space missions with high sampling rates lasting several years, is that the above prescription can lead to very large numbers of trial periods. This implies that the overall algorithm must be extremely efficient. Some steps taken to optimise the efficiency are described below.

The phase, or epoch step interval, is set to the average sampling rate of the data since by so doing one can generate the phase information at no extra computational cost using an efficient search algorithm, detailed below.

2.2.3.2 A weighting scheme to account for non-continuous sampling

A further complication stemming from irregular sampling and from the finite duration of each sample, is that data points nominally corresponding to a time outside a transit may correspond partly to the out-of transit bin and partly to the in-transit bin. To account for this, the indices of points falling either side of the transit boundaries are also stored and included in the calculation of S, but with a weight which is < 1 and is inversely proportional to the interval between the time corresponding to the data point and the start/end time of the transit. This weighting scheme is particularly important for data with irregular sampling where transits might fall, for example, at the end of a night of ground-based observations, or even with space-based observations during a gap in the temporal coverage.

2.2.3.3 Speeding up the algorithm

By far the most time consuming operation in computing S and finding the set of parameters which maximises it, is the identification of the in-transit points, which must be identified for each model d, p and e. If one is dealing with a large number of light curves sharing the same observation times, it is more efficient to process many light curves simultaneously and compute $S(d, p, e)$ for the entire block of light curves for each set of parameters, as follows. For each trial period, the time array is phase-folded. At a given trial duration, the in-transit points are identified for the first trial epoch, by stepping through the folded time array one element at a time until the start time of the transit is reached, and then continuing, storing the corresponding indices, until the end time of the transit is reached. $S(d, p, e)$ is then computed and stored for each light curve. When moving to the next trial epoch, one steps backward through the folded time array from the end time of the old transit (which is stored between successive trial epochs) until the start time of the new transit is found. One then steps forward through the time array, storing the indices, until the end time of the new transit is reached. $S(d, p, e)$ is then computed and stored, and the epoch incremented, and so forth.

This minimises the overall number of calculations needed. As the number of in-transit points is the same for all light curves and σ only needs to be computed once per light curve (in the constant error case), this leaves only the sum of the in-transit points to be computed once per set of parameters and per light curve. The optimum number of light curves to process simultaneously depends on the amount of memory available.

A further speed increase is obtained by noting the redundancy within the computation of S for a range of phase/epoch and period trial values. Breaking down the search to a two-stage process consisting of a single transient event detector (essentially a matched filter stage) followed by a multiplexed period/phase search, removes the inner loop summation of data from the main search and gives a factor of ~ 10 improvement in execution time.

Example run-times computed using a laptop equipped with a 1.2 GHz Pentium IV processor with 512 MB of RAM are as follows. The light curves consisted of 157 680 floating point numbers, i.e. each was $\sim$ 630 KB in size. The trial period and duration ranges were 180 to 400 days and 0.5 to 0.7 days respectively. These ranges are roughly appropriate to search for transits of planets in the habitable zone of a Sun-like star, and correspond to a total number of tested (p, d, e) combinations of $\sim 5 \times 10^7$. After finding the optimal number of light curves to search simultaneously, the runtime per light curve was $\sim$ 4 seconds.

Note that close-in planets with periods below the range included in this sim-

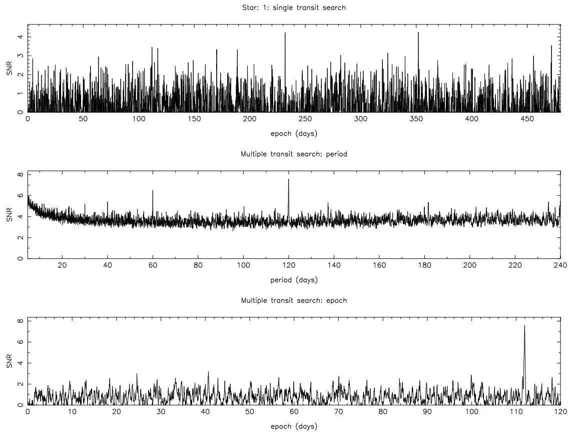

Figure 2.18: Example detection statistic distribution for the $V = 14.5$ case, showing S as a function of trial epoch for the single event search (top), as a function of period for the multiple event search (middle) and as a function of epoch at the best period (bottom). The true epoch and period are 112 and 120 days respectively.

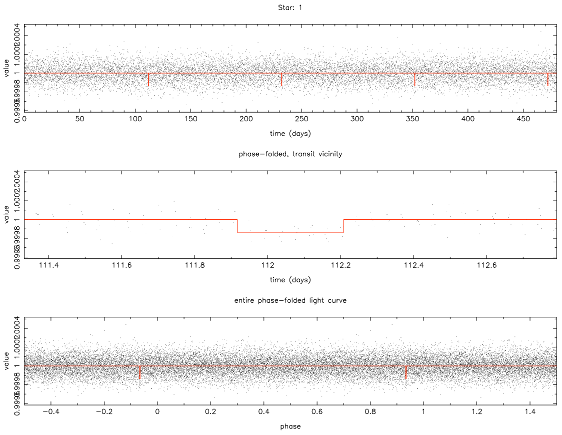

Figure 2.19: Example light curve corresponding to Figure 2.18. Top: full light curve with the positions and depths of the detected transits in red. Middle: phase-folded light curve (portion around transit). Bottom: full phase-folded light curve. Transits occur 112, 232, 352 and 472 days after the start of the light curve

ulation are, of course, of interest, so that lower trial periods (and hence lower trial durations) would also be included when searching for transits in real data, thereby increasing the runtime. As the trial period range is increased, the number of trial periods becomes prohibitively large due to the use of even sampling in frequency space (see Section 2.2.3.1): this leads to very small trial period steps at the low period end of the range if the steps are to be kept reasonable at the high period end of the range. This can be remedied by splitting the required range of trial periods and running the algorithm separately for each period interval. The runtime increases linearly with the number of trial durations.

2.2.4 Performance evaluation

Bootstrap simulations were carried out to evaluate the performance of the box-fitting algorithm in the same manner as for the modified GL method. We present here the evolution of the algorithm's performance as a function of magnitude, designed to verify that the simplifications which led from the modified GL method to the present one did not reduce the performance. Rather than going into more detailed simulations for different star-planet configurations at this stage, the box-fitting method will be tested more extensively in combination with simultaneously developed filtering tools to reduce the impact of stellar micro-variability (see Chapter 5).

To save time, a single transit duration value was used in the bootstrap simu-

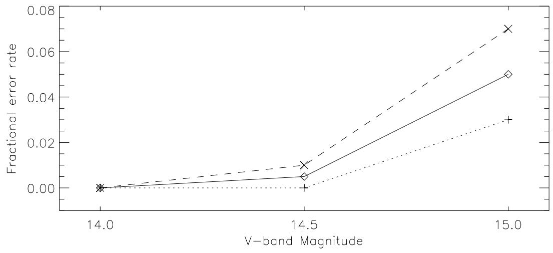

Figure 2.20: Evolution of the box-fitting algorithm's performance (in terms of fractional error rates) with magnitude. Dotted line: false alarm rate. Dashed line: missed detection rate. Solid line: mean error rate.

lations (corresponding roughly to the FWHM of the input transits). Single tests on a given light curve with a range of trial durations did however show that, contrary to the modified GL method, the box-fitting method can be used to provide a rough estimate of the transit duration, which can then be refined using standard least-squares model fitting techniques on the phase-folded light curve.

Simulations were run for *V*-band magnitudes of 14, 14.5 and 15 for a system identical to that used in Section 2.1.3.3: a 1 $R_\oplus$ planet orbiting a K5V star with a period of 4 months, the light curves lasting 16 months with 1 hour sampling. Even though the *Eddington* baseline design at the time these simulations were carried out had evolved from what it was at the time the modified GL method was tested, the photon counts expected with the older design were used to keep the comparison between the two algorithms fair. Figure 2.18 shows an example of the detection statistic distributions obtained for V = 14.5 for the single and multiple event search, while Figure 2.19 shows the corresponding light curve and phase-folded transits.

The results of these simulations are shown in Figure 2.20. Comparing with the bottom panel of Figure 2.10, a small improvement is indeed observed, the mean error rate being slightly lower at all three magnitudes (0, 0.5 and 5 % are V = 14, 14.5 and 15 respectively, compared to 0, 2.5 and 8 % with the modified GL method). Combined with the significant improvement in computing time, this makes the new algorithm decidedly more attractive for the detection stage. Running the modified GL algorithm on each set of 200 light curves (100 with and 100 without transits) took approximately 2 weeks on a Sun Sparc 5, while running the box-fitting method on

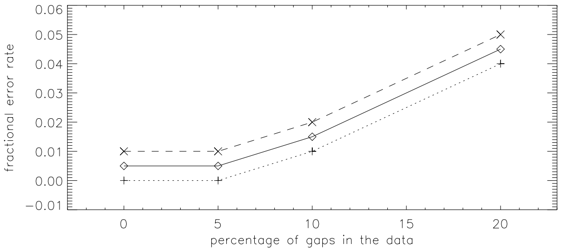

Figure 2.21: Evolution of the box-fitting algorithm's performance (in terms of fractional error rates) with the percentage of gaps in the data for V =14.5. Dotted line: false alarm rate. Dashed line: missed detection rate. Solid line: mean error rate.

the same light curves took just over an hour on a laptop PC with 512 MB of RAM and a 1.2 GHz processor.

Tests were also conducted inserting short, random gaps (as was done for the modified GL method in Section 2.1.3.4) for the $V = 14.5$ case. The results are shown in Figure 2.21, showing a smooth, slow degradation of performance with increasing percentage of gaps. Given that the expected duty cycle for *Eddington* is over 95 %, short gaps of this type should not pose a problem for detection.

2.2.5 Discussion and future work

Through a process of simplification and consolidation, a robust least-squares box-fitting method for transit searching was derived from the modified GL method. Its performance is slightly improved compared to its 'ancestor', while the computational requirements have been vastly reduced.

In white noise, it is capable of reliably detecting periodic transits with a combined signal to noise ratio down to ≥ 6, a limit similar to that found by the authors of its closest relative, the BLS of Kovács et al. (2002). As the *Kepler* mission was designed to produce a combined signal-to-noise ratio ≥ 8 for three transits of an Earth-analogue, the present algorithm should detect such events in *Kepler* data provided most of the stellar variability can be filtered out.

The fact that the detection statistic S is equal to the transit signal-to-noise ratio makes the interpretation of the distributions of the statistic with period and epoch, and thus the initial appraisal of potential candidates, relatively straight forward. The distribution of S with trial epoch at the best trial period (bottom panel of Figure 2.18) is, in simple terms, the convolution of the phase-folded light curve with a top-hat function of width equal to the trial duration, so that the actual duration (and to some extent shape) of the detected event, as well as the presence of other signal (secondary eclipses of a binary for example) can be assessed directly from that distribution. The reduction in sensitivity for triangular or curved eclipses due to the use of a box-shaped model is not expected to exceed 6 %.

The next step is to test this algorithm in a more realistic context, that is in the presence of data gaps and non-Gaussian noise. The two issues are linked: while non-Gaussian noise sources such as stellar micro-variability make a pre-processing stage necessary, the natural approach to filter out this noise involves Fourier domain decomposition. In the presence of data gaps, this is a non-trivial process, which is explored further in Chapter 4.

Following that, the algorithm has been applied blindly to simulated data containing a more complex realistic mix of noise sources, produced by a consortium of members of the COROT Exoplanet Working Group (see Chapter 6).