Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **440,** 2099–2114 (2014) doi:10.1093/mnras/stu372

# Extreme-value modelling for the significance assessment of periodogram peaks

M. Süveges★

*ISDC Data Centre for Astrophysics, Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, CH-1290 Versoix, Switzerland*

Accepted 2014 February 25. Received 2014 February 24; in original form 2012 December 1

**ABSTRACT**

I propose a new procedure to estimate the false alarm probability, the measure of significance for peaks of periodograms. The key element of the new procedure is the use of generalized extreme-value distributions, the limiting distribution for maxima of variables from most continuous distributions. This technique allows reliable extrapolation to the very high probability levels required by multiple hypothesis testing, and enables the derivation of confidence intervals for the estimated levels. The estimates are stable against deviations from distributional assumptions, which are otherwise usually made either about the observations themselves or about the theoretical univariate distribution of the periodogram. The quality and the performance of the procedure are demonstrated on simulations and on two multimode variable stars from Sloan Digital Sky Survey Stripe 82.

**Key words:** methods: data analysis – methods: statistical – stars: variables: general.

## 1 INTRODUCTION

Several research fields of astronomy rely crucially on the analysis of periodic phenomena. For instance, asteroseismology depends on a reliable identification of excited oscillation modes in variable stars, in order to match theoretical stellar models to observed data (Smolec & Moskalik 2007; Grigahcène et al. 2010; Antoci et al. 2011; Balona & Dziembowski 2011, to pick only a few). As another example, extrasolar planetary systems are found by periodic modulations in the light or the radial velocity of their central stars (Cumming, Marcy & Butler 1999; Udry et al. 2007; Mayor et al. 2009; Dawson & Fabrycky 2010). Thus, the detection and analysis of periodic signals in astronomical time series receives much attention in the literature (for a concise summary of the theoretical background, see Schwarzenberg-Czerny 1998).

The general principles of the detection are those of statistical hypothesis testing and model selection. A null hypothesis $H_0$ of no periodic signal in the observed time series is tested against the alternative of the presence of a periodic deterministic component. $H_0$ supposes most often that the observations are white noise (usually Gaussian) around a constant mean, though other error distributions, errors with non-trivial time series structure and non-periodic stochastic processes like random walks are also possible. The white noise null hypothesis is made plausible by pre-processing the data: pre-selection of candidates (e.g. quasar/star separation in a deep survey to filter out cases where random walks may be a good alternative model), and a careful assessment whether independence or uncorrelatedness is a sufficiently good approximation for the errors. The alternative hypothesis is formalized by a series of more complex models $\mathcal{M}_f$ indexed by the frequency $f$, consisting of a periodic signal at $f$ and noise. Both the null model $\mathcal{M}_0$ and the collection $\mathcal{M}_f$ are fitted to the data, and a test statistic $\theta(f)$ is computed for all models, which quantifies the improvement yielded by $\mathcal{M}_f$ over $\mathcal{M}_0$. This collection of test statistics as a function of $f$ is called the periodogram. The frequency at which the largest improvement is achieved is accepted as the most likely frequency of a potential periodic component. Then the decision, whether the object shows a periodic oscillation or not, is based on the significance assessment of the model improvement: the probability is computed that the found periodogram maximum or a higher value is produced under the null hypothesis. This probability is termed the false alarm probability (FAP).

This assessment is a multiple testing situation: as we do not know the frequency of the sought oscillation in advance, we compute the test statistic often at hundreds of thousands of frequencies. Thus, the single-value distribution $F$ (the marginal distribution) of the individual test statistics is not directly applicable to compute the FAP. Instead, we must find the distribution $G$ of the maximum of a large set of test statistics.

The idea which is most commonly used in astronomy to obtain $G$ and the FAP is based on elementary probability calculations for the maxima of $M$ independent variables with common distribution function $F$, yielding the formula $G(z) = F(z)^M$ (Scargle 1982; Horne & Baliunas 1986; Schwarzenberg-Czerny 1998). This formula has the great merits of being simple and, once $M$ fixed, easily applicable even for large surveys in an automated way.

However, there are numerous issues with it, both on the theoretical side and in practice. Theoretically, the characteristics of astronomical time series and period search methods imply that the formula is an ad hoc statistical model rather than a well-founded approximation. On the practical side, the form $F(z)^M$ requires an estimate of $M$, which is time-consuming to produce, and needs a good

★ E-mail: maria.suveges@unige.ch

knowledge about the marginal distribution $F$ of the periodogram. A bootstrap-based method to estimate $M$ was given in Paltani (2004) and in Schwarzenberg-Czerny (2012). A reliable, but less CPU-greedy FAP estimate, with sound statistical foundations and with a possibility to assess the uncertainty in the FAP level estimate would be desirable.

This paper proposes to avoid the shortcomings of the formula $F(z)^M$ by the application of extreme-value statistics. Statistical theory proved that a suitable regularization of the observed maxima leads to a simple three-parameter limiting distribution, called the generalized extreme-value family (GEV; Fisher & Tippett 1928; Gnedenko 1943). It provides a regular approximation to the distribution of maxima of random variables, similarly to the central limit theorem for the sum or the mean of a collection of random variables. The validity range of this limit theorem is very broad: it encompasses practically all continuous distributions and dependent variables too, under some mild conditions (Leadbetter 1974). Regardless of the underlying marginal distribution $F$ of the variables, it has one common parametric form, estimable with standard methods, and provides a stable, mathematically well-founded model for extrapolation to the levels required by the FAP. Instead of the formula $F(z)^M$, its use is standard practice in most sciences or industries that are concerned with risk estimation of rare events like hurricanes, floods, droughts, internet traffic failures, or market crashes (for a range of applications, see e.g. Finkenstadt & Rootzén 2001). The use of extreme-value theory for periodograms in astronomy was first proposed by Baluev (2008), providing an upper bound on FAP under some conditions on the distribution of the periodogram and under the assumption of weak aliasing and spectral leakage. A recent further development is an extension for multifrequency cases (Baluev 2013).

The procedure proposed here fits a GEV model to the tail of the distribution of the periodogram under the null hypothesis that the time series is white noise. First, we construct a large number of noise time series corresponding to the null hypothesis. In the next step, we compute part of the periodograms of the constructed time series, and find the maxima of these partial periodograms. Finally, the parameters of the GEV distribution of these maxima are estimated, and the obtained GEV is used to extrapolate to the desired high levels. Since similarly to the Baluev (2008), Paltani (2004) and Schwarzenberg-Czerny (2012) methods, the new GEV-based procedure is only approximate in the case of strong aliasing, various well-established statistical diagnostics are proposed to check the quality of its approximation.

In Section 2, we first give a brief summary about frequency analysis and its particularities in astronomy, and discuss in detail the consequent problems in the statistical hypothesis testing. Then, we present the basics of extreme-value theory and inference. Section 3.1 describes the proposed procedure, and explains the arguments motivating the applied techniques. Section 4 shows the performance of the procedure on simulations, while Section 5 applies the methods to two variable stars from Stripe 82 of the Sloan Digital Sky Survey (SDSS) with an RR Lyrae-like primary frequency. Section 6 gives a summary of the results.

## 2 STATISTICAL BACKGROUND

### 2.1 Frequency analysis

Suppose we have an observed time series $X_1, \ldots, X_N$ with $N$ points, e.g. magnitudes or radial velocities of a star, measured at irregular times $t_1, \ldots, t_N$. The exposure time and other practical factors limit the precision of the epochs, and there can be found a greatest common divisor $\delta t$ of the time differences $t_i - t_j, i, j = 1, \ldots, N$, such that we can regard the measurements as taken on a dense time grid of resolution $\delta t$, consisting of epochs $0, 1\delta t, 2\delta t, \ldots, T\delta t$ (Eyer & Bartholdi 1999). The observed sequence can then be regarded as a time series taken on a regular grid, with a scarce minority of known observations at $t_1, \ldots, t_N$ and with an overwhelming majority of missing values at other epochs of the grid. In the sequel, the term 'irregularly or unevenly sampled time series' will refer to such an observational sequence.

Periodic signals in an evenly sampled time series $X_1, \ldots, X_T$ are usually detected by the means of the periodogram, an estimate of the spectral density of the time series, on a fixed frequency grid $\mathcal{F}$ between 0 and $f_{\max}$, where $f_{\max} \leq f_{\rm Nyquist} = 1/2\delta t$ (Eyer & Bartholdi 1999). The classical periodogram is defined as

$$I_{T,X}(f) = \frac{1}{T\delta t}\left|\sum_{j=1}^{T} \mathrm{e}^{-\mathrm{i}2\pi f j \delta t} X_j\right|^2 . \quad (1)$$

Many effects can make the identification of a periodicity very difficult. Aliases and other spurious peaks can appear due to the presence of frequencies higher than the Nyquist frequency and to quasi-periodicities in the observation times, and spectral leakage, due to the finite time span of the time sampling window, broadens the spectral features.

For evenly observed signals, the fundamental method to estimate the periodogram at the Fourier frequencies is the fast Fourier transform. For the irregularly sampled case, many methods can be found in the astronomical literature: the Deeming method (Deeming 1975) that can be regarded as the direct generalization of equation (1) to arbitrary times; PDM-Jurkevich (Jurkevich 1971; Stellingwerf 1978; Dupuy & Hoffman 1985); string length (Clarke 2002); SuperSmoother (Friedman 1984; Reimann 1994); Keplerian periodograms (Cumming 2004); Lomb–Scargle and its extension, the generalized Lomb–Scargle method (Lomb 1976; Ferraz-Mello 1981; Scargle 1982; Zechmeister & Kürster 2009); the CLEAN algorithm of Foster (1995); the FASTCHI2 of Palmer (2009); and for photon arrival time series, methods based on Rayleigh's and Kuiper's tests (Paltani 2004). The extremum (most often the maximum) of the estimated periodogram indicates the most likely frequency of the object, and statistical hypothesis testing is used to decide whether the periodic component is significant or not.

This hinges on the knowledge of the distribution $G$ of the maximum. Most arguments for its derivation rely directly or indirectly on the relationship $F(z)^M$, which follows from independence assumptions (Scargle 1982; Horne & Baliunas 1986). If there is a set of independent random variables $Z_1, \ldots, Z_M$ with common distribution function $F$, then the distribution of their maximum can be derived as

$$\begin{aligned}&\Pr(\max\{Z_1, \ldots, Z_M\} \leq z)\\ &\quad = \Pr(Z_1 \leq z, \ldots, Z_M \leq z)\\ &\quad = \Pr(Z_1 \leq z) \times \cdots \times \Pr(Z_M \leq z) = F(z)^M, \end{aligned} \quad (2)$$

where the second equality is true only if the variables $Z_1, \ldots, Z_M$ are independent; otherwise, the joint probability cannot be decomposed into a product of the marginal probabilities.

So far, only a few easy-to-use alternatives were proposed to this formula, and most procedures for FAP estimation rely on it (though there are bootstrap-based alternatives as in Paltani 2004 and Schwarzenberg-Czerny 2012, or procedures using empirically

derived reference distributions as in Koen & Eyer 2002). However, there are several drawbacks.

(i) *No orthogonal set of frequencies exists for an irregular sparse time sampling*. The irregular sampling introduces non-vanishing correlations between sine functions of different frequencies. This entails the loss of any independent frequency systems even in a Gaussian case. Thus, an appropriate derivation of the distribution of the maximum should use the joint multivariate distribution of the periodogram.

(ii) *This joint multivariate distribution is degenerate*. We use $N$ observations to compute $n \gg N$ test statistic values $\theta(f_i)$ $(i = 1, \ldots, n)$. Any mapping $h : \mathbb{R}^N \longrightarrow \mathbb{R}^n$, defined on the space of the observed random variables $X_1, \ldots, X_N$ to compute $n$ values $Z_1, \ldots, Z_n$, produces degenerate joint probability distributions if $n > N$ ($N/2$ for periodograms, as we lose the phase information during the computation). The testing situation is therefore not just mildly, but radically different from the basic assumptions of equation (2).

(iii) *The marginal distribution $F(z)$ is only approximately known*. Theoretically derived approximate marginal distributions for the periodogram usually rely on the orthogonality of the basis functions. However, for many methods this does not hold; an example is the inclusion of a constant besides $\sin 2\pi f t$ and $\cos 2\pi f t$ in the generalized Lomb–Scargle method [though orthogonalization is possible for example by the Gram–Schmidt procedure (Ferraz-Mello 1981) or the decomposition into Szegő polynomials of order 0, 1 and 2 (Schwarzenberg-Czerny 1996)]. Theoretical approximations may be anyway rough in cases when there are only a small number of observations and strong departures from normality in their tail. Fig. 1 illustrates such a case: the standardized Gaussian quantile–quantile plot (in the right-hand panel) of the observed sequence shows deviations from the straight line representing a true Gaussian distribution, especially in the high end.

(iv) *It is necessary to estimate $M$*. $M$ does not correspond to the count of any interpretable or identifiable variable in a simplified model of the testing situation, so its estimation is usually based on fits to simulations from normally distributed white noise at the observational epochs. The estimated value of $M$ is most often larger than the number of observations (Horne & Baliunas 1986; Frescura, Engelbrecht & Frank 2008; Schwarzenberg-Czerny 2012), which is impossible in the context of an independent model. This hints at the ad hoc nature of the approximation $F(z)^M$: there is no straightforward simplification to an equivalent independent set of variables.

(v) *The formula $F(z)^M$ is sensitive to changes in both $M$ and $F$*. Extremely high quantiles of $F$ (of the order of $F(z) = 0.9999$) must be precisely estimated in order to obtain even moderate FAP levels: with $M = 25$ (an unusually low value) and FAP $= 0.01$, we need $z$ such that $F(z) = 0.9996$, since FAP $= 1 - F(z)^M = 1 - 0.9996^{25}$. The form $F(z)^M$, with $M$ in the exponent of a highly uncertain tail distribution, can easily result in an erroneous estimation.

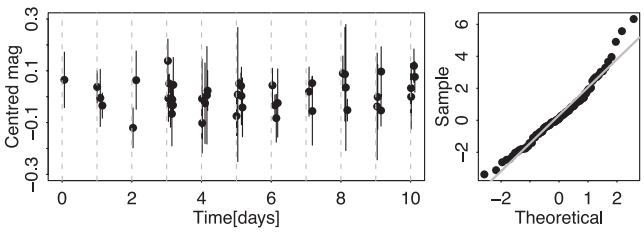


**Figure 1.** A simulated time series (sinusoidal signal and independent Gaussian errors with SNR $= 0.5$ and frequency 3.379 865 $\mathrm{d}^{-1}$) and its quantile–quantile plot. The signal is undetectable in this case.

(vi) *Moreover, the distribution $F(z)^M$ itself becomes degenerate when $M$ increases*. This contradicts a fundamental paradigm in statistical inference, namely, that the distribution of the test statistic should tend to a stable well-defined distribution when the number of observations $N$ increase, and in general, the quality of the estimates should improve. However, in the analysis of astronomical periodograms the increase of $N$ usually causes also $M$ to increase. For a marginal distribution with a finite endpoint, such as the beta distribution, this leads to $F(z)^M$ tending to a step function with an extremely steep rising part just before its endpoint, and in order to give a good probability estimate, we need an increasingly high precision in $z$. When the endpoint is infinite, such as for the Fisher–Snedecor and the exponential distributions, $F(z)^M \to 0$ everywhere on the real line, and $z$ values corresponding to the necessary near-one probability levels run out to infinity. Stabilization is necessary, and this is what is obtained by extreme-value theory.

### 2.2 Generalized extreme-value distributions

Extreme-value theory deals with the statistical analysis of low-probability events. Its development has been motivated by the need of estimating the probability and the level of rare, but possibly catastrophic events, such as the hurricane Katrina in 2005; the three-day rainstorm on 1999 December 14–16 in the coastal areas of Venezuela, which caused around 30 000 deaths; or financial market crashes that can have serious impact on economic stability. The theory has been summarized in several books (e.g. Leadbetter et al. 1983; Resnick 1987; Embrechts, Klüppelberg & Mikosch 1997; Coles 2001; Beirlant et al. 2004; de Haan & Ferreira 2006). Its cornerstone theorem yields a family of asymptotically valid limiting distributions for the (normalized) maxima $Z_{\max,\,n}$ of a large number of random variables $Z_1, \ldots, Z_n$, identically distributed according to a continuous distribution $F(z)$, when $n \longrightarrow \infty$ (the normalization constants are irrelevant for estimation, as they are merged into the parameters of the distribution). The theorem is valid for the maxima of not only independent, but dependent variables too, if the dependence decays between increasingly separated extremely large variables when $n \longrightarrow \infty$. The extremal types theorem (Fisher & Tippett 1928; Gnedenko 1943) states the form of this limiting family, called the generalized extreme-value distribution (GEV):

$$\begin{aligned} G(z) &= \Pr\{Z_{\max,n} \le z\} \\ &= \exp\left\{-\left(1 + \xi\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right\}, \\ &\quad \xi \in \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0, \end{aligned} \tag{3}$$

where $z$ is such that $1 + \xi(z - \mu)/\sigma > 0$. The exact formulation of the extremal types theorem can be found in the references given above.

The GEV family constitutes the limiting family for the maxima of nearly all continuous distributions. The parameter $\xi$, called shape, is related to the tail decay of the underlying distribution of $F(z)$, and divides the GEV family into three well-separated subfamilies. Negative shape parameters provide distributions of maxima of variables with a finite upper boundary, for example the uniform or the beta distribution. In this case, $z < \mu - \sigma/\xi$, and there is zero probability to obtain maximum values higher than this limit. With a positive shape parameter, there is no upper limit, the probabilities in the right tail of the GEV decay slowly as a power law, and there is considerable chance for the maximum to reach very large values. This is the limit distribution of maxima from heavy-tailed

laws like the Student's $t$, or the Cauchy distribution, which has the same mathematical form as a Lorentzian profile. The case $\xi = 0$, called the Gumbel distribution, separates these two distinct types of behaviour. It is defined as the limit function when $\xi \to 0$, and has the form

$$G(z) = \exp\left\{-\exp\left(-\frac{z-\mu}{\sigma}\right)\right\}$$

with $z \in \mathbb{R}$. Its tail decreases exponentially, giving much lower probabilities to observe very high maxima than the case $\xi > 0$, but these probabilities are nowhere 0, differently from the case of a negative shape parameter. This is the limit distribution of maxima of variables from distributions like the normal (Gaussian), lognormal, gamma, exponential or the chi-squared.

### 2.3 Inference and diagnostics for extremes

The estimation of a GEV model for a time series of length $n$ starts usually with the division of the series into blocks of $k$ observations (for example, a block can be a year for a sequence of daily temperature measurements spanning several decades). From each block, we select the maximal value (in the example, the maximum temperature of each year, say $z_1, \ldots, z_m$), and we fit the GEV model to the maxima for example by maximum likelihood (see e.g. Coles 2001) or probability-weighted moment method (Hosking, Wallis & Wood 1985). The maximum likelihood procedure, which straightforwardly yields also error estimates on the parameters, is described in the appendix. About GEV modelling and inference for extremes, Coles (2001) gives an excellent practical summary in its Chapters 2 and 3.

The selection of $k$, the number of observations in a block raises a question of bias-variance trade-off. Since the GEV is a limit distribution when $k \to \infty$, $k$ must be large enough to provide a good extreme-value approximation, otherwise the model will be poor, and yields biased parameter estimation and extrapolation. But if $k$ is too large, we can obtain only a few blocks in the series and therefore too few maxima. This implies a large variance of the parameter estimates. The choice of the block size is governed by pragmatic considerations, as in the case of annual maxima for climatic time series. In cases where the choice is not straightforward, several block size may be used to perform the analysis. Then, the quality of each model is checked by diagnostic plots, and the results using the smallest block size that gives sufficiently good model diagnostics can be accepted as final results.

Model diagnostics must be applied also for another reason. The GEV family is a valid limiting distribution for dependent data too, under some conditions on the strength of the dependence in a limit where the number of observations tend to infinity. However, the number of observations in astronomical time series is finite, often quite small. Together with observational gaps, this results in a strongly dependent, even degenerate periodogram, which is far from the requirements of the asymptotic limit. For this reason, the GEV model proposed here must be checked for model quality by the means of diagnostic plots.

There are two important and easy-to-use types of these. The first is the quantile–quantile plot, generally used in statistics to check adequacy of a fitted model instead of histograms, since histograms are sensitive to bin choice, and in particular, not adapted to detect discrepancies in the tails of the fitted distributions, exactly where FAP estimates are expected to be precise. Let $z_1, \ldots, z_m$ be a collection of block maxima, $z_{(1)}, \ldots, z_{(m)}$ the ordered sample in increasing order, and $\hat{G}(z)$ the estimated GEV distribution. The quantile–quantile plot consists of the points

$$\left\{\hat{G}^{-1}\left(\frac{i}{m+1}\right), z_{(i)}\right\}, \qquad i = 1, \ldots, m,$$

where $\hat{G}^{-1}$ is the inverse function of $\hat{G}$ taken at the estimated parameter values $\hat{\xi}$, $\hat{\sigma}$ and $\hat{\mu}$. If the model is good, the points should closely follow a straight line with intersect 0 and slope 1 without strong systematic deviations. This plot is a direct visual comparison of the empirical distribution function (EDF) of the data and the fitted model. The EDF is defined such that it takes the value $\frac{i}{m+1}$ at the points $\{z_{(i)}\}$; in order to avoid having exactly 0 or 1 outside the data range, which could cause problems, we apply a small correction by using $m+1$ instead of $m$ in the denominator. The same probability levels $\frac{i}{m+1}$ are taken by the fitted model at $\hat{G}^{-1}(\frac{i}{m+1})$. If the model is good, the EDF and the estimated distribution function $\hat{G}(z)$ should be close to each other. Thus, the values $\hat{G}^{-1}(\frac{i}{m+1})$ and $z_{(i)}$ should also be nearly equal, because they are the quantiles belonging to the same probability levels $\frac{i}{m+1}$ in the two distributions.

Fig. 4 provides examples for diagnostic quantile–quantile plots. The top row shows two quantile–quantile plots resulting from GEV fits. Both are acceptable fits: the grey dots representing the observed values versus the model-predicted quantiles form approximately a straight line, and no strong systematic deviations from the fit can be seen, though there is some scatter at the right end of the distribution. Another example of a general quantile–quantile plot is the right-hand panel of Fig. 1, using standard Gaussian quantiles instead of the GEV, corresponding to the supposed distribution of the sample. On that plot, non-Gaussianity can be detected as systematic deviation from the straight line.

Whether the scatter of the right end of the distribution in Fig. 4 is just due to natural fluctuations or to an invalid model can be better judged by the return level plot. It shows the return level function $\zeta_p$ (the inverse of the fitted distribution taken at probability $1-p$, see the appendix) against transformed probability values $\log[-\log(1-p)]$, that is, the points

$$\left\{\log[-\log(1-p)], \zeta_p\right\},$$

where the return level $\zeta_p$ is given by equation (A2) or (A3) for any $p$, using the fitted GEV parameters. The transformation $\log[-\log(1-p)]$ is such that the Gumbel distribution becomes a straight line, a heavy-tailed GEV curves upwards above it, and a finite-tailed one remains below it, monotonically increasing, but eventually tending to a horizontal line. Confidence bands can be calculated as described in Appendix A4. The observed points, plotted as the pairs

$$\left\{\log\left[-\log\left(1-\frac{i}{m+1}\right)\right], z_{(i)}\right\}, \qquad i = 1, \ldots, m.$$

usually show some scatter around the line, especially at the high end. The confidence bands help to see how far from the fitted model they are. Despite some scatter, a model is still acceptable if the points remain within the confidence bands.

As an illustration, the return level plots corresponding to the quantile–quantile plots of Fig. 4 are given in the bottom row. They are constructed using the same collection of maxima and the same model fit. The solid line on both graphs represents the fit, namely the return levels $\zeta_p$ given by equation (A2) against the transformed probabilities $\log[-\log(1-p)]$. The distribution plotted on the left-hand side is finite-tailed, the one on the right is very close to a Gumbel distribution. The confidence bands give an immediate visual impression about the uncertainty of the estimated return levels.

In both cases, the fitted model is acceptable, since all the points remain well within the confidence bands.

## 3 ESTIMATION OF THE FAP

### 3.1 Procedure

1. *Bootstrap of the original time series*. In order to generate noise sequences under $H_0$, we resample the original observations $X_1, \ldots, X_N$ with replacement and with equal probabilities $R$ times, using the same observational epochs (called non-parametric bootstrap in statistics). Thus, we create $R$ repetitions of a white noise series with approximately the same marginal distribution as the original time series. Simulated Gaussian white noise (parametric bootstrap) can be used if the error bars on the observations have no uncertainty, a Gaussian distribution of the errors can be regarded as a good assumption, and outliers are rare. In general, non-parametric bootstrap provides more prudent FAP estimates than the parametric simulations: it does not make use of any preliminary distributional assumptions, and thus also includes uncertainty about the correct model itself.

2. *Maxima of partial periodograms*. From the frequency grid $\mathcal{F}$, we randomly select $L$ non-overlapping frequency intervals of $K$ consecutive frequencies, where $K$ is the oversampling factor, and $L$ is chosen large enough to provide a good extreme-value approximation. This is equivalent to a random draw of $L$ central frequencies $f_{j_1}^{(r)}, \ldots, f_{j_L}^{(r)}$ with equal probabilities, and, around each, taking a frequency interval containing $K$ consecutive grid frequencies. The periodogram must be calculated only at these $KL$ frequencies, and only the maximum of each partial periodogram is needed. The output of this step is thus a sample of $R$ maxima of partial periodograms of white noise sequences, which have distributions similar to the original observations.

3. *GEV modelling of the partial maxima*. Fit an extreme-value model $G(z; \xi, \sigma, \mu)$ to the $R$ maxima, maximizing the log-likelihood equation (A1) to obtain estimates $\hat{\xi}, \hat{\sigma}$ and $\hat{\mu}$ for its parameters. Compute the inverse of the observed information matrix for their uncertainty estimates. Use diagnostic plots to check the quality of the fit.

4. *Extrapolation for the complete periodogram*. Use the estimated $\hat{G}(z; \hat{\xi}, \hat{\sigma}, \hat{\mu})$ to find levels corresponding to the desired FAP values. Give confidence intervals of these levels.

### 3.2 Reasoning behind the steps

1. *Bootstrap*. The parametric and the non-parametric bootstrap, both proposed here, test two different zero hypotheses. $H_0$ in the case of parametric bootstrap is that the observed sequence is a white noise with the specified distribution (usually Gaussian), whose parameters, the error bars, are known quantities, not estimates. In the case of the non-parametric bootstrap, the assumption on the distribution of the errors is relaxed, and $H_0$ is simply the hypothesis that the observed sequence is a white noise, with no restriction on its distribution. In the situation of Fig. 1, it is clear that if this sequence is indeed white noise, then it cannot come from a Gaussian distribution. Generating white noise from a Gaussian distribution would produce on average lower periodogram maxima than another, non-Gaussian distribution that is closer to the observed one, since the observed upper tail seems to be heavier than a Gaussian distribution. Thus, in the case of unknown or unreliable noise distribution, Gaussian simulations would overestimate the significance of the maximum of the periodogram of the observations. Both versions of bootstrap simulate white noise, and are not applicable when the errors are correlated; since the correlation structure induces distortions in the periodogram, modified bootstrap procedures together with more sophisticated extreme-value methods must be developed.

2. *Maxima of partial periodograms*. Two different aims motivate the particular way of the procedure to select 'blocks', subsets from periodograms of which we take the maxima. The principal goal is to decrease the computational load due to a bootstrap. At the same time, the reduced frequency set should reflect the fundamental characteristics of a full periodogram: the long-range dependence manifest in the non-vanishing correlations between sinusoids of two distant frequencies, and the short-range dependencies due to the spectral leakage. The random selection of central frequencies throughout the whole range $(0, f_{\max}]$ provides a sample which is representative of the long-range dependence, whereas taking intervals of width equal to the oversampling factor $K$ accounts for the effects of spectral leakage. Maxima of such partial periodograms carry information on both kinds of dependency, and the GEV fit of step 3. will reflect these, but a careful assessment of model quality is required to enable extrapolation.

3. *GEV modelling*. The fitting can be done in several ways, for example by the method of maximum likelihood (Smith 1985; Coles 2001). The parameter estimates, if the true $\xi > -0.5$, follow an asymptotic multivariate normal distribution (similar to the usual asymptotics of regular likelihood estimates), so inference is straightforward for the estimated model. Due to the strong dependence and the degeneracy present in the periodogram, and to check whether the number of bootstrap repetitions $R$ and the size of the partial periodograms $KL$ provide a sufficiently good extreme-value approximation, it is necessary to assess model quality by the diagnostic plots described in Section 2.3.

4. *Extrapolation*. The model can be used for extrapolation only if the diagnostics have proven the model acceptable. Extrapolation consists of finding quantile levels corresponding to the FAP probability of the full periodogram, based on the extreme-value modelling of the partial periodogram. A level $\zeta_\alpha$ corresponding to FAP $= \alpha$ in the full periodogram is exceeded only once in $1/\alpha$ complete periodograms (in numbers, FAP $= 0.01$ means that out of a hundred periodograms, we can expect on average only one where the maximum exceeds $\zeta_\alpha$). As the complete periodogram contains $n/(KL)$ times more test frequencies than the subsets used for estimation, this corresponds to one exceedance in $n/(\alpha KL)$ partial periodograms. Using equation (A2), we can then compute the desired level as $\zeta_\alpha = \hat{G}^{-1}(1 - [\alpha K L]/n)$, and can add also confidence intervals based on the methods given in Appendix A4.

Though using only the maxima of a reduced frequency set for the estimation of the GEV parameters alleviates the problems due to the time-requirements of the bootstrap, the reduction implies that we need to extrapolate in order to give return levels of maxima of the complete periodogram. If the used frequency set is much smaller than the complete periodogram, the extrapolation must reach to levels far beyond those used in the fit, and the estimated return levels will have a large uncertainty. A frequency set size closer to the complete periodogram provides more reliable extrapolation. Thus, there is a trade-off between the computational load and the necessary range of extrapolation. The choice of $R$ and $L$ can also be checked by the above mentioned diagnostic plots: for bad model fits, increasing $L$ is a possible remedy, that is, using larger partial periodograms, and if the observed points scatter too much around the fitted line, this may be improved by increasing $R$, that is, using a larger number of bootstrapped noise sequences.

## 4 RESULTS

### 4.1 Evenly sampled simulations

The new GEV-based method was compared to the $F(z)^M$ formula in a case where the good performance of the latter is expected: on a signal evenly sampled at 25 points, with time-varying independent Gaussian noise, according to the model

$$Y_i = A\sin(2\pi f t_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_i^2).$$

The standard deviations $\sigma_i$ were random, based on a gamma-distributed variable at each point, yielding an average error $\bar{\sigma}$ of 0.05 mag. With a signal amplitude equal to 0.01 mag, the signal-to-noise ratio (SNR) was therefore approximately SNR $= A/\bar{\sigma} = 0.2$.

Two setups were used. In the first, the deviation added to the sine-wave at point $i$ was generated from $\mathcal{N}(0, \sigma_i^2)$, representing an ideal case with all assumptions about errors satisfied, correctly known error bars, and the marginal distribution $F(z)$ of the periodogram as close to the theoretical one as possible. In the second, the deviations at five epochs were generated from a different Gaussian with several (2–6) times larger standard deviation, so the error bars attached to the time series were at these places false. This setup tested the effect of misspecified errors in the time series. For both setups, one signal and one series of error bars were generated, and the random measurement error generation was repeated fur times, yielding twice 4 different realizations of time series with the same signal and the same stochastic characteristics.

For each simulation, we computed the weighted GLS periodogram on a frequency grid with oversampling factor 16, and assessed the significance of its peak with four procedures:

(i) the combination of non-parametric bootstrap and GEV fitting as proposed in Section 3.1;
(ii) a variant of the latter, replacing the non-parametric bootstrap of the observations with white noise generated from $\mathcal{N}(0, \sigma_i^2)$ using the error bars $\sigma_i$ at each epoch;
(iii) the formula $F(z)^{\hat{M}}$, where $\hat{M}$ was estimated by the procedure of Schwarzenberg-Czerny (2012) and Paltani (2004), based on the peaks of the complete periodograms of 1000 simulated Gaussian white noise sequences;
(iv) the formula $F(z)^M$ supposing $M$ to be the (known) number of independent Fourier frequencies in the searched frequency interval: $M = 12$. $F(z)$ in the latter two procedures was taken to be the theoretical marginal distribution of the periodogram, Beta(1,11) (Schwarzenberg-Czerny 1998).

The periodograms of the eight simulations are shown in Figs 2 (employing the correct error distribution) and 3 (using the partly misspecified errors). With such a weak signal, only two show a maximum at the correct signal frequency, in panels (a) and (b) of Fig. 2. The estimated GEV quantile levels and their confidence intervals, corresponding to FAP $= 0.05$, are added as horizontal lines, black showing the estimate based on non-parametric bootstrap, grey, the estimate based on parametric simulations.

Table 1 compares the FAP estimates of the maxima of the eight simulations with the four different methods. The most conservative FAP estimate is from the non-parametric bootstrap, which does not make any assumptions about the distribution of the time series under $H_0$. Leaving the white noise distribution unconstrained makes it harder to claim there is a signal: the fluctuation in the observation at time $t_i$ cannot be compared to a fully known Gaussian of zero mean and variance $\sigma_i^2$. It must be allowed to have come from a distribution with possibly fatter tails that produces large deviations more easily.

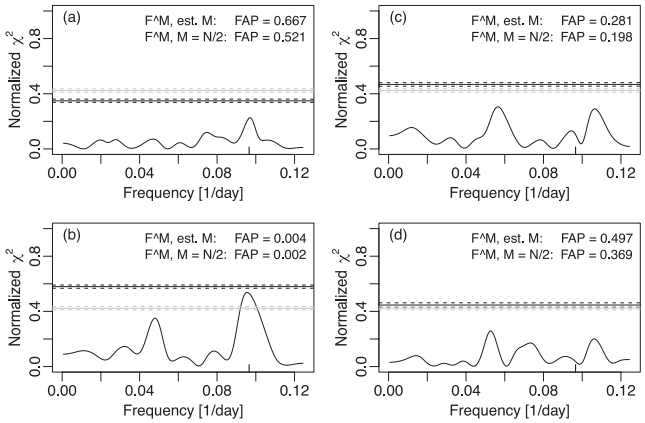


**Figure 2.** GLS periodograms of four evenly sampled sine signals with SNR $= 0.2$ and with known noise distribution. Black horizontal lines: return level with its 95 per cent confidence interval corresponding to FAP $= 0.05$ based on non-parametric bootstrap and GEV fits; grey horizontal lines: return level with its 95 per cent confidence interval corresponding to FAP $= 0.05$, based on Gaussian parametric simulation and GEV fits. The tickmark on the frequency axis indicates the frequency of the true signal.

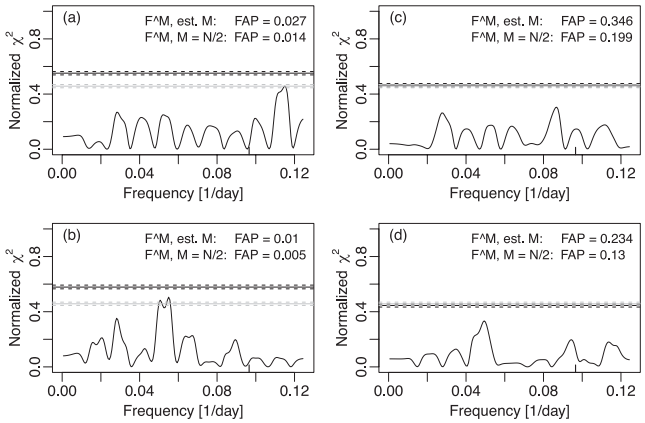


**Figure 3.** GLS periodograms of four evenly sampled sine signals with SNR $\approx 0.2$ and with contaminated noise distribution. Black horizontal lines: return level with its 95 per cent confidence interval corresponding to FAP $= 0.05$ based on non-parametric bootstrap and GEV fits; grey horizontal lines: return level with its 95 per cent confidence interval corresponding to FAP $= 0.05$, based on Gaussian parametric simulation from the assumed incorrect error distribution and GEV fits. The tickmark on the frequency axis indicates the frequency of the true signal.

A stronger constraint on the noise due to extra knowledge, through the Gaussian white noise simulations and the use of the beta margin for the periodogram, results in a less strict periodicity detection criterion; the simulated weak signal becomes detectable when we use our precise knowledge about the error distribution (Fig. 2b).

However, the inclusion of this extra knowledge, when it is incorrect, leads to false identification of periodicities, as can be seen in Fig. 3. Generating five observations out of 25 with higher measurement errors produces strong peaks in the periodogram, which can be falsely estimated significant, if we rely too much on the error bars – mistaken in this case. This happens for two of the four simulations, plotted in panels (a) and (b) of Fig. 3. The use of non-parametric bootstrap avoids this mistake: none of these false peaks are declared significant. This means that the combination of

**Table 1.** Comparison of the different FAP estimates for the four simulations with correct error bars and for the four others with wrong error bars.

| | Estimate of FAP | | | | |
|---|---|---|---|---|---|
| $\frac{\chi^2-\chi_0^2}{\chi_0^2}$ | Non-par. GEV | Par. GEV | $F^{\hat{M}}$ ($\hat{M}$ est.) | $F^{M}$ ($M=N/2$) | Figure |
| | | | Correct errors | | |
| 0.2263 | 0.3472 | 0.6641 | 0.6672 | 0.5209 | Fig. 2(a) |
| 0.5375 | 0.0924 | 0.0069 | 0.0037 | 0.0025 | Fig. 2(b) |
| 0.3052 | 0.4685 | 0.287 | 0.281 | 0.198 | Fig. 2(c) |
| 0.2579 | 0.5368 | 0.4974 | 0.4972 | 0.3687 | Fig. 2(d) |
| | | | Incorrect errors | | |
| 0.4581 | 0.1785 | 0.049 | 0.0269 | 0.0141 | Fig. 3(a) |
| 0.5053 | 0.13 | 0.0234 | 0.0099 | 0.0052 | Fig. 3(b) |
| 0.3048 | 0.4113 | 0.3716 | 0.3464 | 0.199 | Fig. 3(c) |
| 0.3335 | 0.2446 | 0.2708 | 0.234 | 0.1298 | Fig. 3(d) |

non-parametric bootstrap and the GEV provides a reliable, though conservative, FAP estimate in cases when the error bars on the observations may be not well estimated.

Table 1 highlights also the fact that the formula $F(z)^M$ is not connected to any simplified independent model. The time series simulations presented here are evenly sampled, which means that we have an exact, independent Fourier frequency system of 12 frequencies in the tested interval. Nevertheless, using $M = N/2$ clearly underestimates the FAP, and thus overestimates the significance of the found periodicities. Testing a frequency grid denser than the Fourier grid is equivalent to the inclusion of new, functionally dependent test statistics among the random independent variables at the Fourier frequencies. This distorts the distribution of the maximum. What happens is somewhat similar to the difference between examining the maximum of a set of independent, identically distributed positive random variables $\{X_1, \ldots, X_N\}$, or that of an enlarged set $\{X_1, \ldots, X_N, \sum_{i=1}^{N} X_i\}$. The distribution of the maxima of the former or the latter set is clearly not the same: the first is $F(z)^N$, while the second is the distribution of the sum of the variables. The formula $F(z)^M$ is thus only an ad hoc approximation to the true distribution.

Table 1 shows also that the FAP estimates based on parametric simulations + GEV and those based on the formula $F(z)^{\hat{M}}$ are very close, though the GEV estimate is slightly more conservative, especially when misspecified errors are used for the simulations. The similarity of the two estimates using parametric simulations in the evenly sampled case, which is closest to the validity domain of the formula $F(z)^{\hat{M}}$, lends support to the validity and applicability of the GEV approximation. The estimation of $\hat{M}$ requires a large number of simulations and the computation of the complete periodogram for each repetition. The same result, together with an uncertainty estimate, can be produced with a much lower number of repetitions and only partial periodogram calculations. Moreover, the combination of non-parametric bootstrap and the GEV works also when the error bars are not reliable, and prevents the identification of many false periodicities.

### 4.2 Irregularly sampled simulations

The performance of the procedure in a situation closer to practice was assessed using two light-curve patterns, a sine-wave of the form $g(t) = A_{\sin,i} \sin(f_{\sin} t)$ and a broken-line model for detached eclipsing binaries. For the sinusoid, the light-curve parameters were $f_{\sin} = 3.379\,865\,\mathrm{d}^{-1}$ and three different amplitudes $A_{\sin,1} = 0.15$, $A_{\sin,2} = 0.05$ and $A_{\sin,3} = 0.025$ mag. For the eclipsing binary, $f_{\mathrm{ecl}} = 0.424\,3146\,\mathrm{d}^{-1}$ was used, with the depth of the primary minimum equal to $A_{\mathrm{ecl},i} = 2$, 1.33, 1, 0.67, 0.33 and 0.167 mag, and the ratio between the depth of the secondary and the primary minima fixed to 0.375.

To both variability patterns, we added random Gaussian noise with time-varying variance, yielding the model $Y_i = g(t_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. The noise generation was similar to that of Section 4.1, using the correct error bars, such that the approximate SNR of the three sine-wave simulations were $\mathrm{SNR}_{\sin,i} = A_{\sin,i}/\bar{\sigma} =$ 3, 1 and 0.5.

Epochs of observations were chosen on a time grid of 0.005 d (about 10 min) with a total span of 25 d. Imitating random nighttime observations during 4 h each night, two different sequences of epochs were randomly uniformly selected from 25 nightly 4-h periods. One consisted of $N = 100$ observations, the other, of $N = 25$. The input data for the period search were the two sequences of epochs, the nine noisy light curves each sampled at both cadences as observed values and the random gamma variables as error bars, yielding in total 18 different time series.

The time grid parameters led to an upper frequency detection limit $f_{\max} = 100\,\mathrm{d}^{-1}$, a Fourier frequency set of $\mathcal{F}_F = \{0.04, 0.08, \ldots, 100\}\,\mathrm{d}^{-1}$ and a corresponding peak width of 0.04 $\mathrm{d}^{-1}$ due to leakage. An oversampling factor $K = 16$ was used, providing a test frequency grid $\mathcal{F} = \{0.0025, 0.005, \ldots, 100\}\,\mathrm{d}^{-1}$ with $n = 40\,000$ test frequencies. The generalized Lomb–Scargle method was performed on all 18 time series using these test frequencies, once without weighting, and once with weights defined by the normalized inverse squared error bars. The grey spikes in Fig. 7 show the resulting periodograms from the non-weighted version for the six sinusoidal light curves with Gaussian errors.

The procedure presented in Section 3.1 was performed on the 18 light curves. In order to check the stability and the variance of the estimates as a function of the number of bootstrap repetitions $R$ and the number $L$ of the test frequency intervals, we applied all pairwise combinations of $L = 50$, 100, 200, 300, 400, 500 and $R = 200$, 400, 600, 800, 1000 for each of the 18 simulated light curves. We calculated also the complete periodogram for 2000 independent bootstrap noise sequences for each of the simulated light curves, in order to compare the model-based and the empirical high quantiles, corresponding to selected FAP levels.

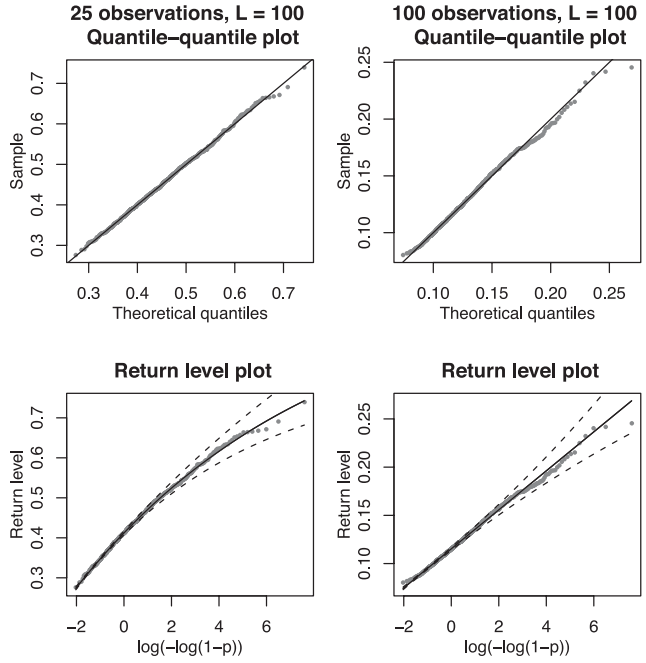


**Figure 4.** Upper panels: quantile–quantile plots for two GEV fits for the maxima of 1000 noise sequences generated by bootstrap, for 25 and for 100 night-time observations (left- and right-hand panel, respectively). In both cases, the bootstrapped original time series was the sinusoidal simulation with SNR = 1, and $L = 100$ random frequency intervals were used. Bottom row: the return level plots for the same data as on the panels above. Dots correspond to the observed values against the transformed empirical probabilities $\log[-\log(1-\frac{i}{m+1})]$, solid lines to the fitted model, and dashed lines to a 95 per cent confidence interval for the fitted return level curve based on asymptotic normality of the maximum likelihood estimator.

### 4.2.1 *Model fits and diagnostics*

Each combination of number of repetitions and number of random frequency intervals $(R, L)$ on each simulation yielded a sample of $R$ maxima of partial periodograms of size $KL$, which were then fitted with the GEV model. The quality of all models was checked by the return level and quantile–quantile plots described in Section 2.3. In general, the models are acceptable and can be used for extrapolation, though some small deviations could be found in many plots, as shown in Fig. 4. The model quality on average is somewhat worse for the shorter time series with $N = 25$ than for time series with $N = 100$, but no other systematic difference can be seen with respect to $R$ or $L$. The high end of the quantile–quantile plots often deviates slightly downward, implying that the observed periodogram maxima are stochastically smaller than the model estimates. The same effect can be remarked also on the return level plots: the theoretical curve is above the points corresponding to the observations. This leads to a conservative error in the significance assessment: the model-based estimated quantiles will be a little higher, and an observed peak will be found somewhat less significant. The converse error is very rare, which implies a lower risk of false periodicity identifications.

### 4.2.2 *Plausibility of quantile estimates and stability*

After the inspection of the diagnostic plots, all GEV models obtained for the 18 simulations with all $(R, L)$ combinations were used to estimate quantile levels $\hat{\zeta}_\alpha$ belonging to fixed FAP levels $\alpha = 0.01$ and 0.005. In order to check their plausibility, we created 2000 bootstrapped white noise sequences from each of the 18 simulated time series. The complete periodograms of every repetitions were computed, and the maxima of these selected. For a FAP equal to $\alpha$, the proportion $\hat{\alpha}$ of the peaks exceeding $\hat{\zeta}_\alpha$ was calculated, and compared to $\alpha$.

Figs 5 and 6 show the results for the six sine-wave simulations with Gaussian noise, $1-\hat{\alpha}$ as a function of $L$ for the short and the long time series, respectively. The agreement between the theoretical $1-\alpha$ and the empirical $1-\hat{\alpha}$ is visibly much better for the non-weighted method than for the weighted for the short time series. An improvement with decreasing SNR, when the deviations from normality become smaller, is visible in Fig. 5.

Small test frequency set sizes ($L = 50$ or 100) cause instability of the empirical exceedance proportions $1-\hat{\alpha}$, especially for the short time series. With $N = 25$ and for the weighted period search version, $L > 200$ seems necessary. Above this, $\hat{\alpha}$ is stable, and approximates well the FAP level $\alpha$ even in the case of short time series. For the

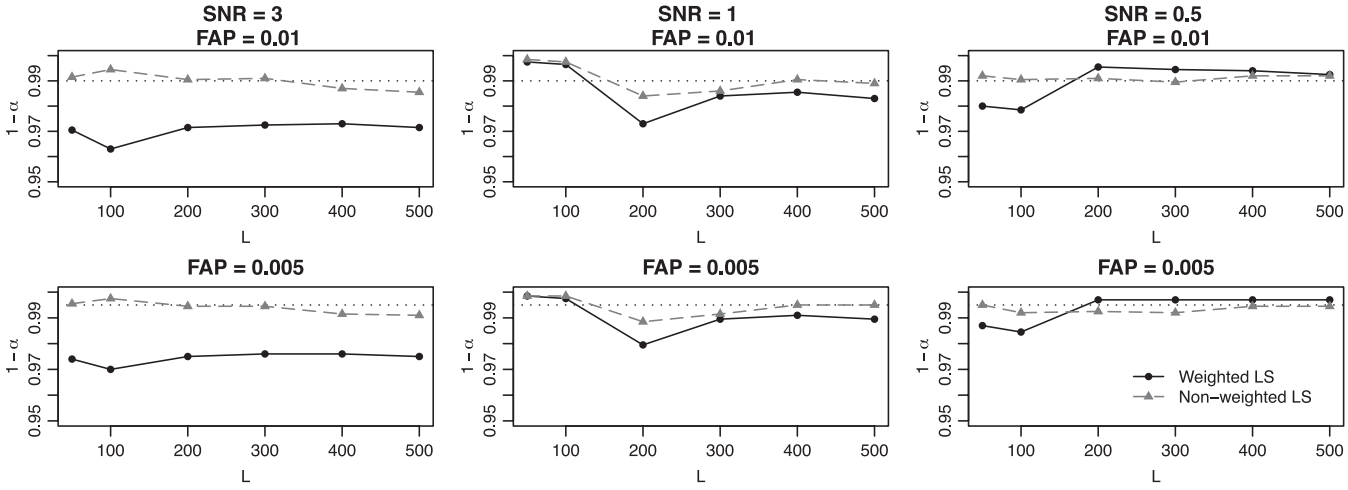


**Figure 5.** The proportion of full periodogram maxima in 2000 that exceed the estimated high quantiles for FAP = 0.01 (top row) and FAP = 0.005 (bottom row), as a function of $L$, for time series length $N = 25$, with $R = 1000$. Dashed grey line with triangles show the results using non-weighted generalized Lomb–Scargle, solid black lines with dots is using the weighted version. The dotted lines denote the FAP levels. The left-hand panels refer to SNR = 3, the middle panels, to SNR = 1, and the right-hand panels, to SNR = 0.5.

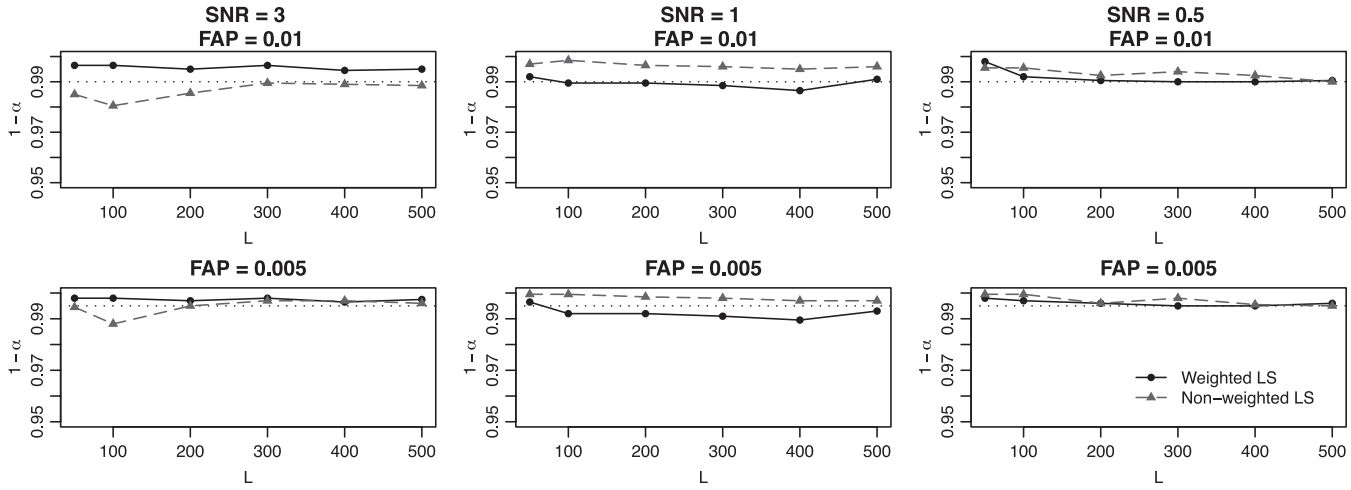


**Figure 6.** The proportion of full periodogram maxima in 2000 that exceed the estimated high quantiles as a function of $L$, for time series length $N = 100$. The symbols and the SNR-FAP combinations are the same as in Fig. 5.

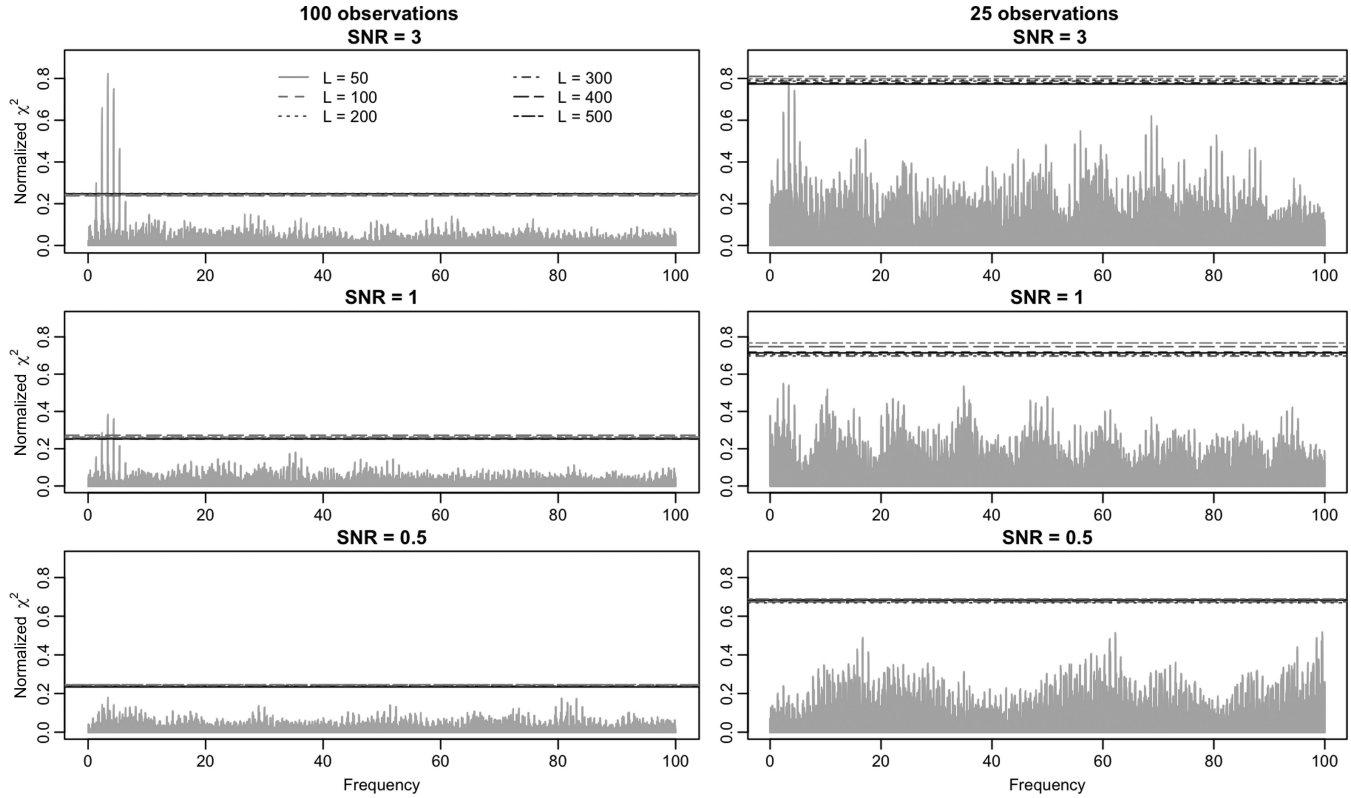


**Figure 7.** Stability of the estimated 0.99 quantile as a function of the number of test frequencies for the sinusoidal simulations with Gaussian errors, $N = 100$ (left-hand column) and $N = 25$ (right-hand column). The grey spikes show the periodograms of the simulated noisy sine-wave signals, the horizontal lines are the quantile levels estimated from fitted GEV models. The different types of the lines correspond to different numbers of test frequency intervals $L$, and are the same for all panels, shown in the legend in the top-left panel. For all plots, $R = 1000$ and the period search method is the non-weighted generalized Lomb–Scargle.

longer time series with $N = 100$, the stability is remarkable, and the approximation is in general good.

The plausibility of the estimates can be easily seen if they are plotted against the periodograms of the time series. In Fig. 7, the decrease of the signal below detection level can be clearly observed, as the SNR decreases. The left-hand panels show the time series of the sinusoidal light curve with Gaussian noise with $N = 100$ observations, the right-hand panels the same light curve–noise combinations with $N = 25$. The levels predicted by the procedure are presented as horizontal lines with different line types for different $L$ values. These levels reflect well the judgment based on the aspect of the whole periodogram. In the cases of SNR = 3 and 1 with $N = 100$ observations, the presence of a signal is obvious because of the absence of other comparable peaks, and accordingly, the plotted FAP = 0.01 levels pass well below the peaks. In the case of the weakest signal with $N = 100$, the periodogram exhibits

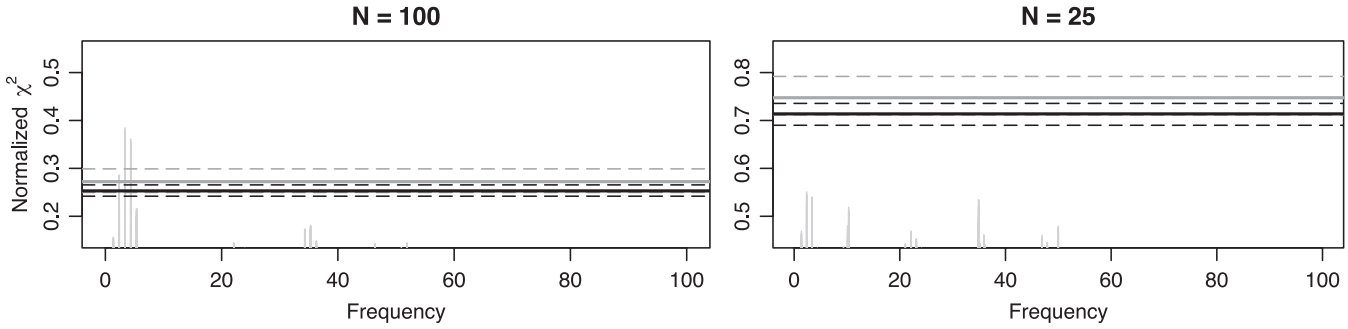


**Figure 8.** 95 per cent confidence intervals of the estimated 0.99 quantile for the SNR = 1 case of the sinusoidal simulations with Gaussian errors. The plots are enlarged versions of the two panels of the middle row of Fig. 7. For visibility, quantiles for only two values of $L$ are plotted ($L = 100$ with grey, $L = 500$ with black); the thick solid lines correspond to the quantile estimates, the dashed ones to 95 per cent confidence intervals from bootstrap.

another group of peaks of comparable height beside the correct one. The signal is undetectable based on the noisy data. The quantiles, in agreement with the impression of non-significance, appear here well above the peaks, and indicates that neither of the peaks is significant.

With less data ($N = 25$), the same plausibility can be observed. So scarce sampling makes even a signal of SNR = 3 barely detectable. A careful judgment says that there is very likely a signal, but the maximum of the periodogram does not exceed very much the secondary peak, and there may be a weak probability that a noise sequence produced this periodogram. Accordingly, the estimated quantile for FAP = 0.01 hovers around the highest peak: its statistical significance is around 0.01, that is, there is ∼1 per cent probability that a noise sequence produces such a peak. For the two other cases with smaller SNR, the signal is not detectable: for SNR = 1, there are many other peaks with comparable size, and for SNR = 0.5, the signal is lost, and the spectrum is dominated by peaks solely due to noise.

The plausibility for the eclipsing binary-like time series is very similar, with the only difference that the period search method finds the double of the frequency, not the correct one. The generalized Lomb–Scargle method does so very often for eclipsing binaries, due to their two more or less equally spaced minima in the light curve. The significance of a peak present in the periodogram is assessed in the same way and with the same plausibility of the results for the double frequency. When applied to real data, a further step of plotting the folded light curve can decide whether the analysed star is an eclipsing binary or not, and whether the correct or the double frequency was found.

The stability with respect to the number of test frequency intervals $L$ can be observed in Fig. 7, too. In the left-hand panels with $N = 100$, all estimated quantile levels corresponding to different numbers $L$ of test frequency intervals are almost indistinguishably close together. Due to the scarce data, the estimated levels are more scattered in the right-hand panels showing the $N = 25$ cases, but the agreement is still quite good, and conveys reasonable judgments about the significances. The agreement between estimates with different $L$ values can be even better appreciated in Fig. 8. This shows an enlarged picture of the estimated quantiles of the simulation with SNR = 1 and $R = 1000$ (middle panels of Fig. 7), together with bootstrap-based confidence intervals, offering a better opportunity to judge their stability. For the sake of visibility, only $L = 100$ and $L = 500$ are plotted. The overlap of confidence intervals confirms the stability of the quantile estimates in a broad range of frequency set sizes.

#### *4.2.3 Stability with respect to R*

The weak dependence of the estimated high quantiles on the number $R$ of bootstrap repetitions can be seen in Fig. 9. The upper row shows the estimated 0.99-quantiles for the complete periodogram of the sinusoidal signals with Gaussian noise with 100 observations, the bottom row shows them for $N = 25$. In the former case, $R$ as low as 200 can be combined with a number $L$ of test frequency intervals as low as 100 giving stable and reliable estimates, though the confidence intervals (here, those originating from bootstrap) are larger with such small values than with $L = 500$ or $R = 1000$. For a scarcely sampled time series, the estimates are more varying according to $L$ or $R$, but over the whole range, they agree with each other within the confidence intervals. The stability of the estimates with respect to $R$ and $L$ allows a strong reduction of computational time, preserving at the same time the quality of the significance assessment.

## 5 CANDIDATE MULTIMODE RR LYRAE STARS FROM SDSS

### 5.1 Data

We give two examples of the use of the proposed procedure, namely, two variable stars from the SDSS Stripe 82[1] (Ivezić et al. 2007). The SDSS provides five-band ($u$, $g$, $r$, $i$ and $z$) photometry of more than 11 000 deg$^2$ of the sky. A region along the celestial equator, called Stripe 82, was imaged repeatedly during a 10 yr long period, providing five simultaneous time series of up to ∼100 data points per object until Data Release 7 (Abazajian et al. 2009). RR Lyrae and high-amplitude $\delta$ Scuti stars were identified in this data set by Sesar et al. (2010) and Süveges et al. (2012). The classification using principal component analysis presented in the latter yielded a number of candidate RR Lyrae stars. Among these, there were many rejected because of a noisy light curve or a non-significant primary frequency. We reconsider two of these stars here, a double-mode candidate and one with a surprisingly high secondary frequency.

All the periodograms were computed by the non-weighted generalized Lomb–Scargle procedure between [0, 6] d$^{-1}$ ([0, 40] d$^{-1}$ for the secondary periodogram for the second candidate), using a resolution of 0.0001 d$^{-1}$. This grid choice implied for both stars an oversampling factor $K \approx 13$, which was checked by plotting the spectral

[1] http://www.astro.washington.edu/users/ivezic/sdss/catalogs/S82variables.html

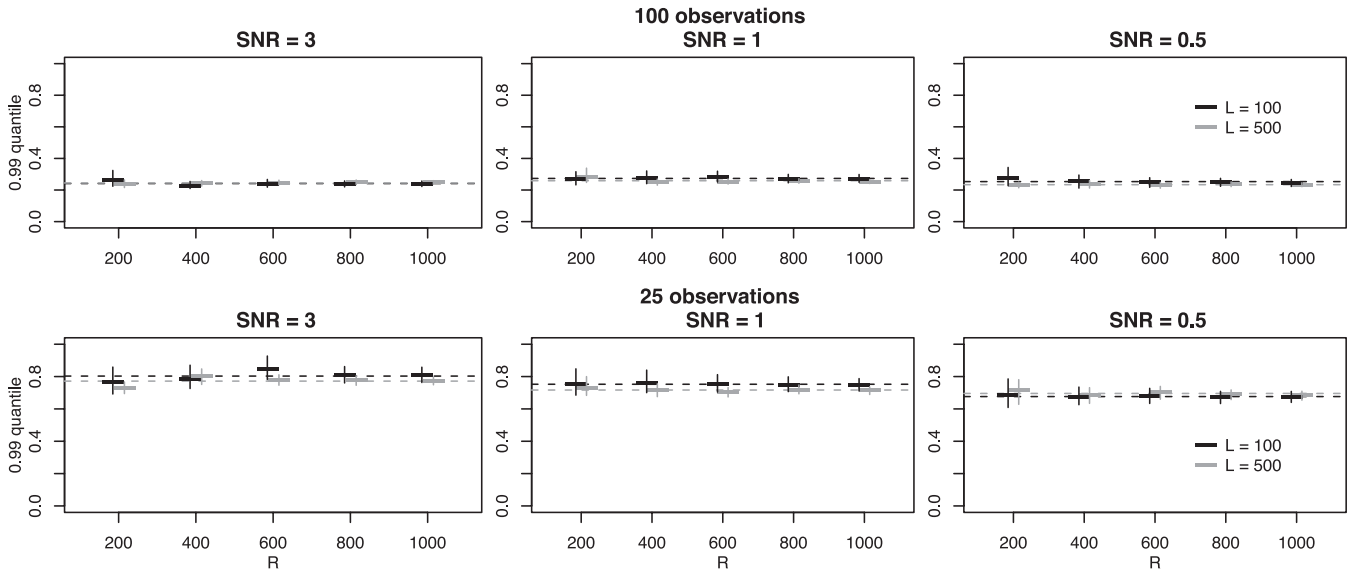


**Figure 9.** Stability of the estimated quantile corresponding to FAP = 0.01 as a function of the number of bootstrap repetitions, for number of test frequency intervals $L = 100$ (black) and 500 (grey). 95 per cent confidence levels based on non-parametric bootstrap are plotted as vertical bars. The plotting positions are slightly shifted horizontally only in order to avoid overlapping marks.

window. Pre-whitening was performed by fitting cyclic cubic splines to the folded light curve, because this is able to remove all harmonics of that frequency simultaneously, at the price of a milder decrease of degrees of freedom than a Fourier harmonic series. A range of smoothing parameters was tried for all light curves, among them one based on generalized cross-validation (GCV; see e.g. Ruppert, Wand & Carroll 2003). For most, the GCV criterion produced a reasonably smooth light curve. In the case of overfitting, the smoothing parameter was adjusted manually. The procedure proposed in Section 3.1 was performed for all observed and pre-whitened light curves, using $R = 500$, combined with $L = 200$ for periodograms with $[0, 6]$ d$^{-1}$ and $L = 800$ for $[0, 40]$ d$^{-1}$, respectively. Return levels for the complete spectrum, corresponding to FAP = 0.05 and 0.01, that is, the 0.95- and 0.99-quantiles were then calculated from a GEV model fitted to the 500 bootstrap periodogram maxima, and the decision about the significance of the periodicity was obtained by comparing the peak in the observed sequence to the estimated level.

### 5.2 A double-mode candidate

The spectral window and primary and residual $g$-band periodograms of the star 538812 (J231332.19-010746.2) are shown in the left-hand panels of Fig. 10. The spectral window exhibits slowly decaying daily alias peaks. Its enlarged version in the top-right panel shows in addition strong yearly aliasing and an oversampling factor $K = 13$. The star has a weakly significant ($0.01 < \mathrm{FAP} < 0.05$) primary peak at the frequency 2.755 272 d$^{-1}$ in $g$. Period search on the other bands supports the existence of a signal at this period: the periodogram maximum in $u$ falls at a yearly alias of this frequency, and there is a prominent peak at the location of this peak in the other three bands as well, though those are not maxima. There are apparently several other periodic signals in the other bands, suggesting a multiperiodic nature. Pre-whitening all bands with the spline smoother yields a very significant peak in the residual spectrum at 2.051 07 d$^{-1}$, providing the fundamental frequency $f_0$ of the star. The proportion between the two frequencies is 0.7444, which corresponds perfectly to the Petersen diagram of double-mode RR Lyrae stars. Performing pre-whitening with the same technique on the other bands yields similarly significant periodicities in $r$ and $i$, but no evident signal in either $u$ or $z$, probably due to the higher noise levels in these bands and to the faintness ($\sim$20 mag in $g$) of the star. The removal of the second cycle does not yield any further significant new frequency, apart from an alias of $f_0 + f_1$, and since the residual degrees of freedom decrease by a value around 10 or more with each pre-whitening, the data set size ($N = 56$) does not allow any further meaningful investigations.

The very weak significance of the primary peak is due to the strong secondary frequency, which causes large residual noise in the folded light curve, and hence a small relative decrease of $\chi^2$ in the primary periodogram. The found frequencies, the position of this star on the Petersen diagram and on the $(u - g, g - r)$ and period–amplitude diagram, together with the results of a principal component analysis presented in Süveges et al. (2012), confirms its double-mode pulsating nature despite the only weakly significant frequency in the primary spectrum.

### 5.3 A candidate with a high-frequency secondary mode

Star 4477012 (J203120.88-001125.3) also was selected as an RR Lyrae candidate in Süveges et al. (2012). Its main frequency found there ($f_0 = 2.660\,207$ d$^{-1}$), its amplitude, its principal components characteristics and the strong variation in the folded $g - i$ colour light curve suggest an RR Lyrae-like pulsating nature. The variable is on the bluest border of the RRab region of the $(u - g, g - r)$ diagram, slightly off from both RRc and RRab regions on the colour–log(period) diagram, and far from the location of both subtypes on the period–amplitude diagram. A double-mode nature would put the star to an admissible position on the colour–log(period) plot, but this is excluded by the analysis of the residual periodogram of the $r$-band observations presented in the second row of Fig. 11: there is no indication of a correct secondary frequency. Instead, we find a weak peak at a high frequency, $f_1 = 13.770\,38$ d$^{-1}$. It attains the level corresponding to FAP = 0.01 only in $r$, but the maxima

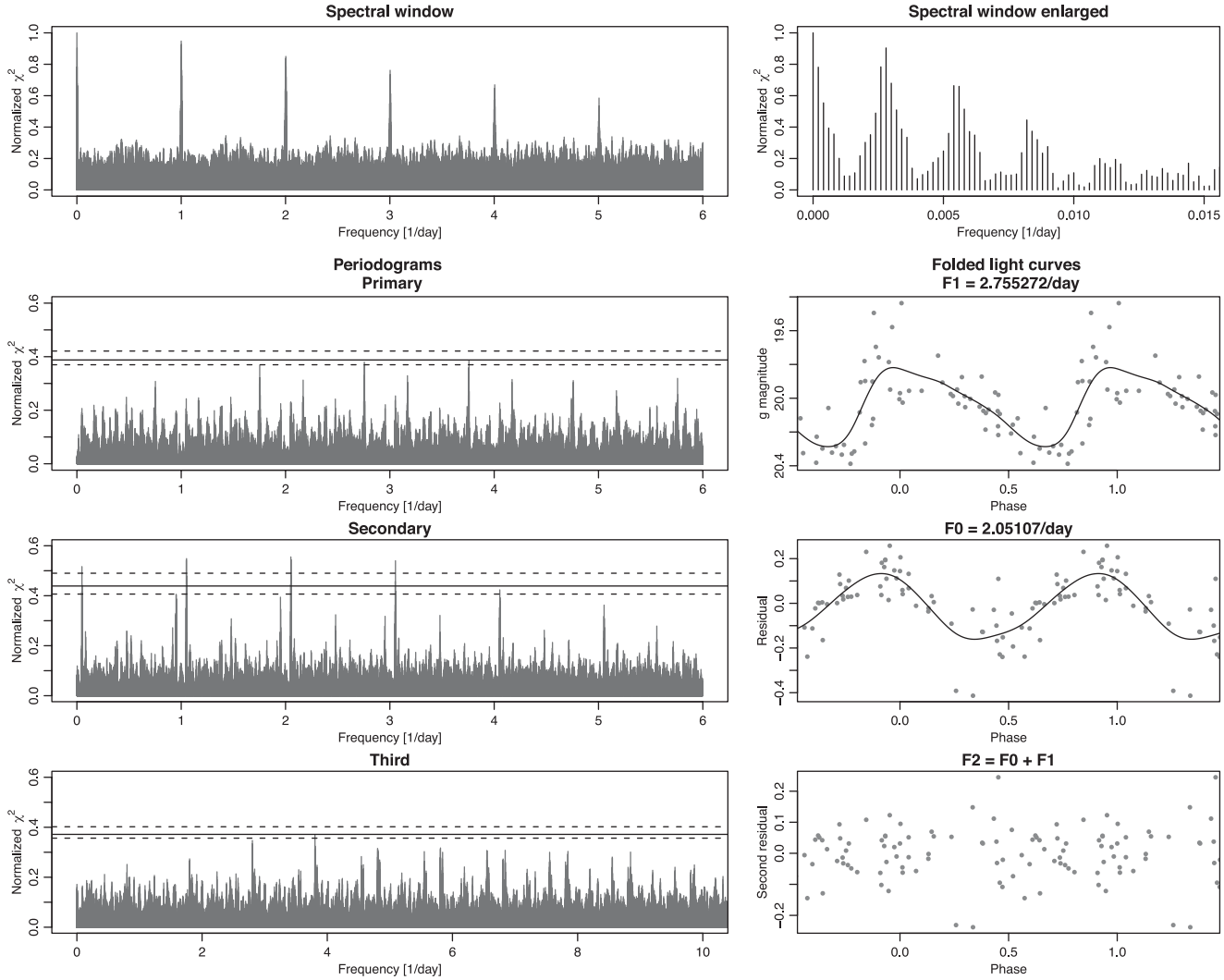


**Figure 10.** Spectral window (top left), enlarged spectral window around 0 (top right), periodograms of $g$-band observations and of residual time series after two successive pre-whitening (left-hand panels, second to fourth row) and the corresponding folded light curves (right-hand panels, second to fourth row) for star 538812 (J231332.19-010746.2). On the periodograms, the FAP levels of 0.01 are plotted as solid black lines, together with their 95 per cent bootstrap confidence intervals.

of the residual periodograms in all other bands falls exactly at the same frequency. This simultaneous occurrence supports that the found periodicity is a real oscillation, either of the star, or of some external origin.

The right-hand panel of the second row of Fig. 11 shows that in the folded residual curve the largest and the smallest residual fall exactly at the minimum and the maximum of the fitted sinusoid. This is so in all bands. The found frequency is thus determined by the separation of these two extremal observations. When omitting them, the primary frequency, now at $f_0 = 3.660\,207\ \mathrm{d}^{-1}$, becomes far more significant, as illustrated in the left-hand panel in the third row of Fig. 11. From the residual periodogram, presented in the bottom row on the left, any significant peaks disappear, and the five filters show no longer marked coinciding patterns.

The decision, whether the signal of $f_1 = 13.770\,38\ \mathrm{d}^{-1}$ exists or not, cannot be taken on purely statistical grounds. It depends on the judgment whether the two influential observations are true or erroneous, and whether we find it plausible that only two observations out of 44 carry most of the information on an existing oscillation. These two observations are of good quality, are not outliers in any of the bands, and their error bars in all bands are rather small. Moreover, in Fig. 12, which shows the observed ($u$, $g$, $r$) and ($r$, $i$, $z$) magnitudes of the observations in 3 dimensions plotted against each other, the light-coloured larger dots representing these points fit perfectly into the joint multivariate distribution of the remaining data. Hence, rejecting them seems unreasonable. On the other hand, accepting them and thus accepting the existence of a secondary period of such a high frequency leads to difficulties in the interpretation of the frequencies and in the class determination for the star. The confirmation or the rejection of this frequency could be obtained only by more data on this object.

## 6 DISCUSSION

This paper offers an alternative to the most commonly used method of astronomy to estimate the FAP in periodograms, the formula $F(z)^M$. The proposed procedure is intended to avoid the shortcomings of the latter: the lack of interpretation of $M$, the sensitivity of the formula both to the tail of $F$ and to $M$, the invalid assumption

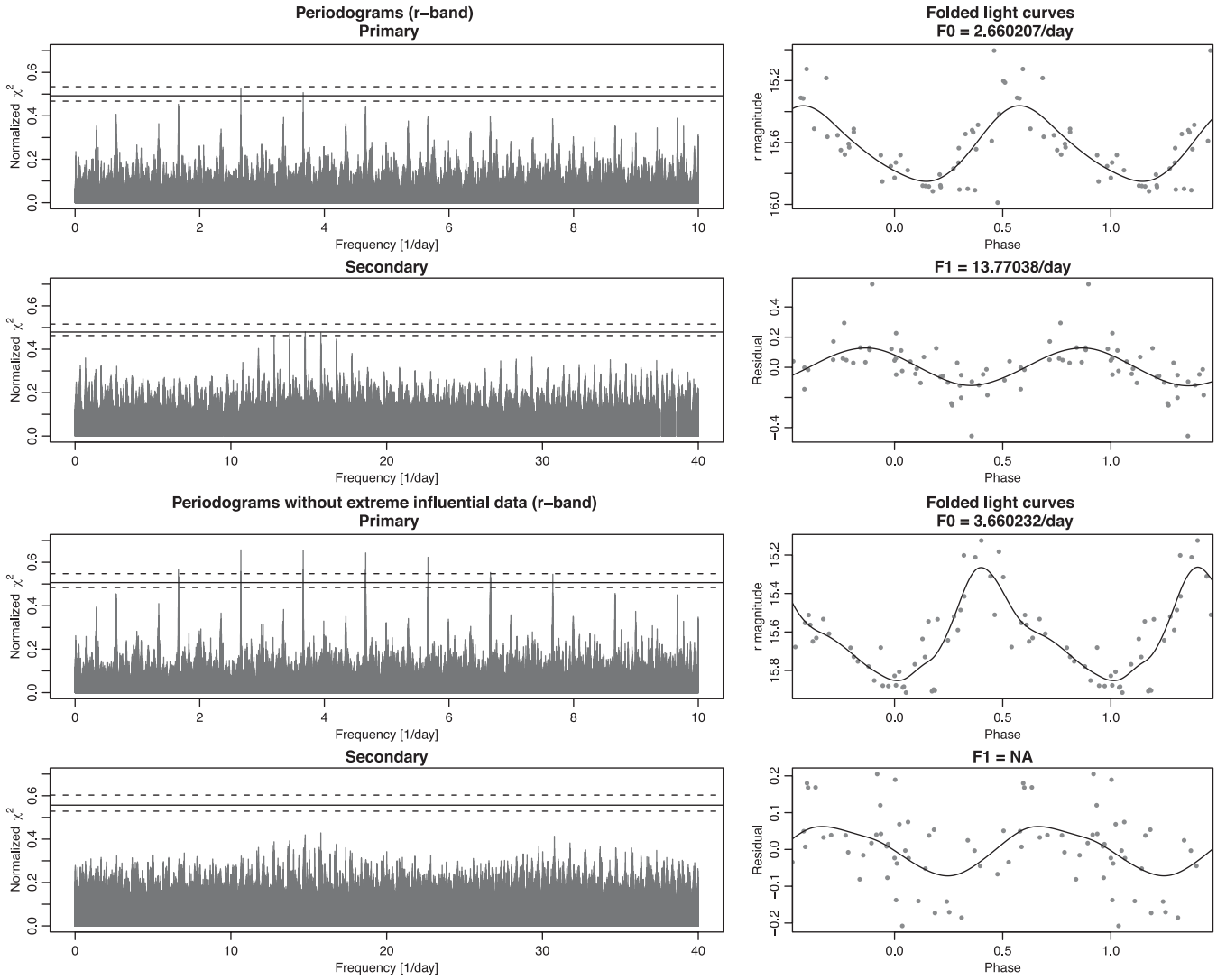


**Figure 11.** Periodograms of *r*-band observations and of residual time series after one pre-whitening (left-hand panels) and the corresponding folded light curves (right-hand panels) of star 4477012 (J203120.88-001125.3). The top two rows show the periodograms and the folded light curves including the two extremal observations, the bottom rows show the same plots when omitting these from the analysis. On the periodograms, the FAP levels of 0.01 are plotted as solid black lines, together with their 95 per cent bootstrap confidence intervals.

of independence in its derivation, its degeneracy when $M$ increases, and the lack of inference for the estimated FAP or quantile levels.

The proposed procedure combines a bootstrap of the original time series with the use of the GEV distribution instead of $F$. The bootstrap is used to produce white noise samples with similar empirical marginal behaviour as the observations, corresponding to the null hypothesis of white noise without imposing additional distributional assumptions. For these bootstrap repetitions, partial periodograms are computed, and the GEV is used to estimate the distribution of their maxima. The GEV is a three-parameter family of distributions, which describes probability levels of maxima from almost all continuous distributions. Thus, it becomes unnecessary to know either the error distribution of the observations or the single-value distribution of the periodogram. The combination of the non-parametric bootstrap and the GEV distribution lifts the sensitivity of the FAP on eventual misspecification of the periodogram distribution $F$. Moreover, this ensures the applicability of the method to many types of periodograms, without regard to their particular single-value distribution.

The formula $F(z)^M$ is heavily impacted also by the effective number of independent frequencies, $M$. Small changes in its value change the estimated quantiles or probabilities. There are no clear theoretical arguments that could help to compute its value for a given time series, or to assess whether an estimated value is reasonable or not. The proposed procedure avoids such issues. Moreover, standard asymptotic normal theory can be used to give inference about the fitted model, and to obtain confidence intervals for the parameters of the GEV and for levels corresponding to fixed FAP values in the periodogram.

The use of the GEV also relieves the computational load due to the bootstrap, since it implies that estimates can be based on maxima of partial periodograms. The quality of the fitted GEV model, which can be compromised by the frequently occurring strong aliasing in astronomical periodograms, can be checked by diagnostic plots, yielding information whether the model is good enough to use for extrapolation.

The subset of frequencies at which we calculate the partial periodograms is selected in a specific way, in order to account for

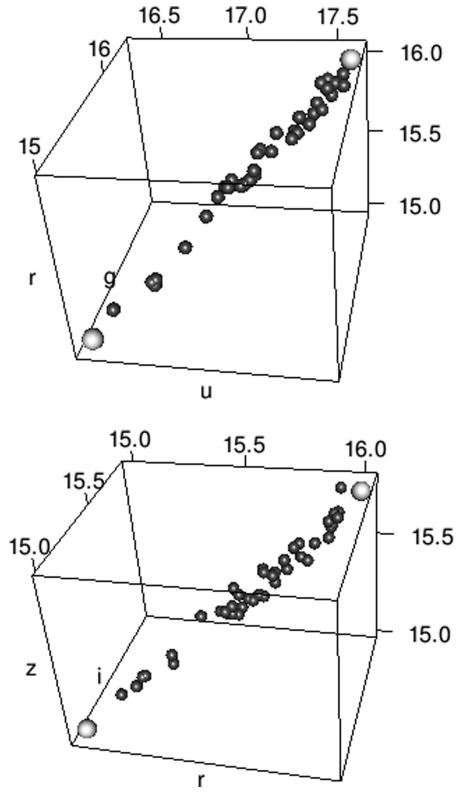


**Figure 12.** Multivariate distribution of the five-band observations in the *u*-, *g*- and *r*-bands (top) and in the *r*-, *i*- and *z*-bands (bottom). The larger light grey blobs on both plots represent the observations with crucial influence on the secondary periodogram in Fig. 11.

the dependence of the periodogram. Intervals of contiguous frequencies are randomly chosen across the tested frequency range $(0, f_{\max}]$, each of length at least equal to the oversampling factor. This ensures that they sample both the long-range correlations across the whole frequency range and the extremely strong local dependence due to spectral leakage.

In tests on simulations of evenly sampled weak sinusoidal signals with known Gaussian errors, the combination of the GEV model with parametric bootstrap gives equivalent results to the $F^M$ method, requiring only the fraction of its computational time. In the case of imprecise error estimates, possible outliers or contamination of the distribution of the measurement errors, the GEV combined with non-parametric bootstrap produces FAP levels less prone to false detections than either the GEV model with parametric bootstrap or the $F^M$ formula.

The confidence intervals on the critical levels corresponding to specific FAP values provide information on the uncertainty of the decision. The levels estimated by GEV are stable in a large range of both number of bootstrap repetitions and frequency subset sizes. A setup that provides good estimates and reasonably narrow confidence intervals on the quantile levels, but computationally is not too heavy, takes a few times the time of the calculation of the complete periodogram (in the simulations and the data examples of this paper, less than 10 times).

The procedure was also applied to two variable stars with unknown class from SDSS Stripe 82, which had marginally significant primary periods in the RR Lyrae range. One of these proved to be a double-mode RR Lyrae, since after pre-whitening, a highly significant secondary peak emerged at a frequency that corresponds to the Petersen diagram of the double-mode RR Lyrae stars. The other star remains an unclear case. A barely significant secondary period was found at a high frequency. The weak significance is due to the fact that this period is mainly determined by only two observations. However, these observations are very unlikely to be erroneous measurements, so the presence or absence of this high-frequency variation in this star cannot be decided purely on statistical grounds.

In summary, the $F^M$ formula is very simple to implement, and does not need optimization procedures. Its drawbacks are the absence of interpretation for its independent frequency system, its reliance on the knowledge of the marginal distribution of the periodogram, the lack of regular statistical inference, and the large CPU requirements. The new procedure is based on the GEV family, which has a similar role of general limit distribution for maxima as the Gaussian distribution holds for sums and averages. Uncertainty estimates can be given by regular statistical inference, and the quality of the model can be validated by diagnostic plots. CPU requirements are significantly lower due to the good extrapolation capacities of the model. Its drawback is the need for an optimization. Both the $F^M$- and the GEV-based procedure deal only with white noise as null hypothesis, and both are approximate in a finite-length, aliased case. In summary, the new procedure is an interesting alternative to obtain significance of periodicities detected in astronomical periodograms.

## ACKNOWLEDGEMENTS

The author thanks R. Gaál for interesting discussions, and P. Dubath and L. Rimoldini for many valuable comments on the manuscript. The work was partly supported by the Swiss National Science Foundation grant PMPDP2_129178.

## REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Antoci V. et al., 2011, Nature, 477, 570
Balona L. A., Dziembowski W. A., 2011, MNRAS, 417, 591
Baluev R. V., 2008, MNRAS, 385, 1279
Baluev R. V., 2013, MNRAS, 436, 807
Beirlant J., Goegebeur Y., Segers J., Teugels J., 2004, Statistics of Extremes. Wiley, New York
Brockwell P. J., Davis R. A., 2006, Time Series: Theory and Methods, 2nd edn. Springer-Verlag, Berlin
Clarke D., 2002, A&A, 386, 763
Coles S. G., 2001, An Introduction to Statistical Modelling of Extreme Values. Springer-Verlag, London
Cumming A., 2004, MNRAS, 354, 1165
Cumming A., Marcy G. W., Butler R. P., 1999, ApJ, 526, 890
Dawson R. I., Fabrycky D. C., 2010, ApJ, 722, 937
de Haan L., Ferreira A., 2006, Extreme Value Theory: An Introduction. Springer-Verlag, Berlin
Deeming T. J., 1975, Ap&SS, 36, 137
Dupuy D. L., Hoffman G. A., 1985, Int. Amat.-Prof. Photoelectr. Photometry Commun., 20, 1
Embrechts P., Klüppelberg C., Mikosch T., 1997, Modelling Extremal Events for Insurance and Finance. Springer-Verlag, Berlin
Eyer L., Bartholdi P., 1999, A&AS, 135, 1
Ferraz-Mello S., 1981, AJ, 86, 619
Finkenstadt B., Rootzén H., 2001, Extreme Values in Finance, Telecommunications and the Environment. Chapman and Hall, London
Fisher R. A., Tippett L. H. C., 1928, Proc. Camb. Phil. Soc., 24, 180
Foster G., 1995, AJ, 109, 1889

Frescura F. A. M., Engelbrecht C. A., Frank B. S., 2008, MNRAS, 388, 1693
Friedman J. H., 1984, LCS Technical Report 5, A Variable Span Smoother. Stanford Univ., Stanford, CA
Gnedenko B. V., 1943, Ann. Math., 44, 423
Grigahcène A. et al., 2010, ApJ, 713, L192
Horne J. H., Baliunas S. L., 1986, ApJ, 302, 757
Hosking J. R. M., Wallis J. R., Wood E. F., 1985, Technometrics, 27, 251
Ivezić Ž. et al., 2007, AJ, 134, 973
Jurkevich I., 1971, Ap&SS, 13, 154
Koen C., Eyer L., 2002, MNRAS, 331, 45
Leadbetter M. R., 1974, Z. Wahrscheinlichkeitstheor. Verwandte Geb., 28, 289
Leadbetter M. R., Lindgren G., Rootzén H., 1983, Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag, New York
Lomb N. R., 1976, Ap&SS, 39, 447
Mayor M. et al., 2009, A&A, 507, 487
Nelder J. A., Mead R., 1965, Comp. J., 7, 308
Palmer D. M., 2009, ApJ, 695, 496
Paltani S., 2004, A&A, 420, 789
R Development Core Team 2010, R Foundation for Statistical Computing, Vienna, Austria, available at http://www.R-project.org
Reimann J. D., 1994, PhD thesis, Univ. California, Berkeley
Resnick S., 1987, Extreme Values, Regular Variation and Point Processes. Springer-Verlag, New York
Ruppert D., Wand M. P., Carroll R. J., 2003, Nonparametric Regression. Cambridge Univ. Press, Cambridge
Scargle J. D., 1982, ApJ, 263, 835
Schwarzenberg-Czerny A., 1996, ApJ, 460, L107
Schwarzenberg-Czerny A., 1998, MNRAS, 301, 831
Schwarzenberg-Czerny A., 2012, in Griffin E., Hanisch R., Seaman R., eds, Proc. IAU Symp. 285, New Horizons in Time-Domain Astronomy. Cambridge Univ. Press, Cambridge, p. 81
Sesar B. et al., 2010, ApJ, 708, 717
Smith R. L., 1985, Biometrika, 72, 67
Smolec R., Moskalik P., 2007, MNRAS, 377, 645
Stellingwerf R. F., 1978, ApJ, 224, 953
Süveges M. et al., 2012, MNRAS, 424, 2528
Udry S. et al., 2007, A&A, 469, L43
Zechmeister M., Kürster M., 2009, A&A, 496, 577

## APPENDIX A: FORMULAE FOR THE ESTIMATION OF GEV MODELS

### A1 Log-likelihood

Let $z_1, \ldots, z_R$ denote a set of maxima of $R$ blocks of random variables (for example, $R$ subsampled periodograms or $R$ years of daily temperature measurements). The log-likelihood for a GEV model is obtained from the cumulative distribution function (3) in the usual way. We first differentiate it with respect to $z$ in order to obtain the corresponding density function, then substitute the set of maxima $z_1, \ldots, z_R$ into the density and take their product to find the likelihood function. Finally, taking the logarithm of the resulting function, the log-likelihood assumes the following form:

$$\ell(\xi, \sigma, \mu) = -R \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{R} \log \left(1 + \xi \frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{R} \left(1 + \xi \frac{z_i - \mu}{\sigma}\right)^{-1/\xi}, \quad \text{(A1)}$$

under the constraint that $1 + \xi(z_i - \mu)/\sigma > 0$ for $i = 1, \ldots, R$. This function is maximized with respect to its arguments $\xi, \sigma$ and $\mu$ to obtain the best-fitting estimates for the parameters. Optimization procedures for it are implemented in several packages of the statistical software R, such as ISMEV, EVIR, EVD, EXTREMES or EVDBAYES; in this paper, I used the GEV procedure of EVIR. Optimization by GEV was performed by the simplex method (Nelder & Mead 1965), which uses only function values, and is very robust.

### A2 Return levels

The estimated GEV distribution can then be used to obtain return levels, probabilities of very high levels or distributions of maxima of longer sequences. The return level $\zeta_p$ associated with the return period $1/p$ is the level that is expected to be exceeded once in every $1/p$ blocks of the same length $k$. This is simply the $1 - p$ quantile of the GEV distribution:

$$\zeta_p = G^{-1}(1 - p) = \mu - \frac{\sigma}{\xi} \left[1 - \{-\log(1 - p)\}^{-\xi}\right], \quad \text{(A2)}$$

or in the case of the Gumbel distribution,

$$\zeta_p = G^{-1}(1 - p) = \mu - \sigma \log \{-\log(1 - p)\}, \quad \text{(A3)}$$

where $G^{-1}$ denotes the inverse function of $G$. For example, if we analyse the behaviour of heatwaves, we might fit a GEV to the annual maxima of a long sequence of daily temperature measurements. In that case, the $\zeta_{1/20}$ return level would be the temperature which is exceeded only once in every 20 years.

### A3 Errors on the estimates: the variance-covariance matrix

Smith (1985) showed that the estimates obtained by maximizing the likelihood (A1) are asymptotically multivariate normal as long as $\xi > -0.5$. Thus, an approximate variance-covariance matrix and standard errors can be straightforwardly derived for the estimates according to likelihood theory. We first compute the negated second derivative matrix

$$\mathbf{I}(\xi, \sigma, \mu) = \begin{bmatrix} -\frac{\partial^2 \ell}{\partial \xi^2} & -\frac{\partial^2 \ell}{\partial \xi \partial \sigma} & -\frac{\partial^2 \ell}{\partial \xi \partial \mu} \\ -\frac{\partial^2 \ell}{\partial \xi \partial \sigma} & -\frac{\partial^2 \ell}{\partial \sigma^2} & -\frac{\partial^2 \ell}{\partial \sigma \partial \mu} \\ -\frac{\partial^2 \ell}{\partial \xi \partial \mu} & -\frac{\partial^2 \ell}{\partial \sigma \partial \mu} & -\frac{\partial^2 \ell}{\partial \mu^2} \end{bmatrix}.$$

Evaluation of this matrix at the estimated parameter values and observed maxima $z_1, \ldots, z_R$, that is, at the maximum of the log-likelihood, yields the observed information matrix $I_O(\hat{\xi}, \hat{\sigma}, \hat{\mu})$. The approximate variance-covariance matrix $\mathbf{V}$ of the estimates can be obtained by inverting the observed information matrix:

$$\mathbf{V} = I_O(\hat{\xi}, \hat{\sigma}, \hat{\mu})^{-1}. \quad \text{(A4)}$$

The diagonal elements give the variance of the corresponding parameter estimate, off-diagonal elements provide the covariance between two estimated parameters. It follows then that a 95 per cent confidence interval for any of the parameters (denoting the vector $(\hat{\xi}, \hat{\sigma}, \hat{\mu})$ by $\theta$) can be given as $\theta_i \pm 1.96\sqrt{V_{ii}}$. The matrix $\mathbf{V}$ is straightforward to estimate numerically, as a by-product of the optimization, though it can be calculated also theoretically.

### A4 Delta method and the errors on the return levels

Confidence intervals for $\zeta_\alpha$ can be obtained by several methods: by parametric or non-parametric bootstrap, by the delta method using the variance-covariance matrix of the GEV parameters, or by profile likelihood. The delta method makes use of the fact that

if a random variable $\theta$ follows a multivariate normal distribution with the true parameter value $\theta_0$ as its mean and $V_\theta$ as its variance-covariance matrix, then a scalar function $g(\theta)$ of it also follows a normal distribution, with $g(\theta_0)$ as its mean and $V_g = (\nabla g)^T V_\theta (\nabla g)$ as its variance-covariance matrix $\nabla g$ is the gradient vector of $g(\theta)$.

Letting $\theta = (\hat{\xi}, \hat{\sigma}, \hat{\mu})$, the maximum likelihood estimator of the GEV model, the return level is a function of $\theta$: $z_p = z_p(\hat{\xi}, \hat{\sigma}, \hat{\mu}) = z_p(\theta)$, according to equation (A2) or (A3). From likelihood theory, we know that $\theta$ follows a multivariate normal distribution. Therefore, $z_p = z_p(\theta)$ is a normal variable with variance $V_{z_p}$. To obtain an uncertainty estimate for the return level, the gradient vector $\nabla z_p$ can be calculated from the return level formula (A2) as

$$\nabla z_p = \left\{ \frac{\partial z_p}{\partial \xi}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \mu} \right\} = \left\{ \sigma \frac{1 - y_p^{-\xi} - \xi y_p^{-\xi} \log y_p}{\xi^2}, -\frac{1 - y_p^{-\xi}}{\xi}, 1 \right\}, \quad \text{(A5)}$$

or in the case of the Gumbel distribution from equation (A3) as

$$\nabla z_p = \left\{ \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \mu} \right\} = \left\{ -\log y_p, 1 \right\}, \quad \text{(A6)}$$

where $y_p = -\log(1 - p)$.

The variance-covariance matrix $V_{z_p}$ of the return level is then $V_{z_p} = (\nabla z_p)^T V (\nabla z_p)$, where $\mathbf{V}$ is the variance-covariance matrix (A4) of the parameters. A 95 per cent confidence interval for $z_p$ can be derived then as $z_p \pm 1.96\sqrt{V_{z_p}}$.

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