Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **449,** 1098–1105 (2015) doi:10.1093/mnras/stv288

The analysis of indexed astronomical time series – XI. The statistics of oversampled white noise periodograms

Chris Koen*

Department of Statistics, University of the Western Cape, Private Bag X17, Bellville, 7535 Cape, South Africa

Accepted 2015 February 10. Received 2015 February 5; in original form 2014 December 12

ABSTRACT

The distribution of the maxima of periodograms is considered in the case where the time series is made up of regularly sampled, uncorrelated Gaussians. It is pointed out that if there is no oversampling, then for large data sets, the known distribution of maxima tends to a one-parameter Gumbel distribution. Simulations are used to demonstrate that for oversampling by large factors, a two-parameter Gumbel distribution provides a highly accurate representation of the simulation results. As the oversampling approaches the continuous limit, the two-parameter Gumbel distribution takes on a simple form which depends only on the logarithm of the number of data. Subsidiary results are the autocorrelation function of the oversampled periodogram; expressions for the accuracy of simulated percentiles; and the relation between percentiles of the periodogram and the amplitude spectrum.

Key words: methods: statistical – stars: variables: general.

1 INTRODUCTION

The periodogram (power spectrum) and its scaled square-root form, the amplitude spectrum, are standard tools used to identify periodicities in time series. Given a set of observations y_t ($t = 1, 2, \ldots, N$) obtained at regularly spaced time points, the periodogram value at angular frequency $\omega = 2\pi\nu$ is

$$I_y(\omega) = \frac{1}{N}\left\{\left[\sum_t (y_t - \overline{y})\cos\omega t\right]^2 + \left[\sum_t (y_t - \overline{y})\sin\omega t\right]^2\right\} \qquad 0 < \omega < \pi\ , \tag{1}$$

where $\overline{y}$ is the mean of the measurements. (Note that the time interval between successive measurements thus defines the units of time.)

Consider a time series consisting of pure white noise, i.e. $y_t = e_t$, where the e_t are identically and independently Gaussian with variance σ_e^2. It is then easily shown that for given ω, $I_e(\omega)$ is exponentially distributed (e.g. Scargle 1982). Since the exponential distribution is heavy-tailed compared to the Gaussian, chance large values of $I_e(\omega)$ are common, which complicates deciding whether a large peak in a periodogram is due to a signal, or whether it is 'spurious' (i.e. noise-induced). Consequently, testing whether periodograms are consistent with pure noise has given rise to a voluminous literature – see Frescura, Engelbrecht & Frank (2008) for a fairly recent review of the relevant astronomy literature.

If the periodogram is calculated only in the Fourier frequencies

$$\nu_j = \frac{j}{N}\ , \qquad j = 1, 2, \ldots, N/2, \tag{2}$$

then the $N/2$ periodogram ordinates $I(\omega_j)$ are all independent, and the cumulative distribution function (CDF) of the periodogram maximum V is known to be (e.g. Frescura et al. 2008)

$$F_V(v) = [1 - \mathrm{e}^{-v}]^{N/2}\ , \tag{3}$$

provided the $I(\omega)$ have been standardized (i.e. divided by σ_e^2). For large data sets, the standardization is simple, even if σ_e^2 is unknown, since the mean of the periodogram over all frequencies provides an excellent estimate of this parameter. This method of standardization is attractive, since comparison of the highest spectrum peak to the mean spectrum has become popular – see the many references to Breger et al. 1993, who proposed a simple 'significance test' based on this criterion. Alternatively, σ_e^2 can of course be estimated directly from the data.

In astronomical applications periodograms are almost always oversampled, i.e. calculated in a finer grid of frequencies than the Fourier values. The reason is that the spectrum is not fully resolved by the set of Fourier frequencies – features of interest, occurring at frequencies between Fourier values could be missed if spectra are not oversampled. This is particularly important in the field of asterioseismology, where many pulsation frequencies may be present in a spectrum (e.g. Gilliland et al. 2010). Of course, the denser the sampling of the spectrum, the more accurate the determination of the precise values of frequencies of interest will also be.

* E-mail: ckoen@uwc.ac.za

For closely spaced frequencies ν and $\nu + \Delta$ ($\Delta \ll 1$), it is shown in Appendix A that the autocorrelation function of the periodogram is given by

$$\rho(\Delta) \approx \mathrm{sinc}^2(N\Delta) ,$$

i.e. a very steeply decreasing function of the frequency difference Δ, essentially non-zero only over the interval $(-\Delta, \Delta)$. It follows that, for oversampled periodograms, ordinate values at close frequencies will be strongly correlated. This invalidates the use of equation (3), a fact which is well known in the literature on extreme value distributions (EVDs; e.g. Castillo et al. 2005).

The problems engendered by the lack of independence of periodogram ordinates are generally acknowledged in the literature, particularly in the case of irregularly spaced time series, where the issue seems unavoidable (e.g. Frescura et al. 2008).

Deriving appropriate forms for F_V in the case of oversampling is a difficult, if not insurmountable problem, but it is possible to derive very accurate approximation formulae for it in terms of EVDs, as will be demonstrated below. The author is aware of two recent papers in the astronomy literature which argue for the use of EVDs to determine significance levels of periodogram peaks, namely Cuypers (2012) and Süveges (2014). The former author presented a short account of work based on the generalized extreme value (GEV) probability density function (PDF) given by

$$f(x) = \frac{1}{\sigma} \exp\left\{-\left[1 + k\frac{(x-\mu)}{\sigma}\right]^{-1/k}\right\}\left[1 + k\frac{(x-\mu)}{\sigma}\right]^{-(k+1)/k} , \quad (4)$$

where σ, μ and k are, respectively, the scale, location and shape parameters. Data sets consisting of 50 Gaussian white noise measurements were simulated, with a variety of assumed time spacings. The conclusion was that overall the GEV did not provide a better description of the distribution of periodogram maxima than other previously used methods.

The study of Süveges (2014), also based on use of the GEV, was more extensive, and the conclusions drawn were more positive. Contrasting with this work, no special time spacing of the observations was assumed, nor was it required that the error distribution be homogeneous. A bootstrap procedure, followed by extrapolation to smaller tail probabilities, was used to determine large percentiles tailored to the specifics of the particular data set.

The aim here is narrower: to provide accurate formulae which can be used to find percentiles of periodogram maxima ('False Alarm Probabilities'), given the data restrictions stated in the first two paragraphs above. In practice, these would apply to e.g. continuous monitoring from space, or 'high speed' ground-based photometry.

Sturrock & Scargle (2010) approached the oversampling problem in a different way: the authors compared the highest peak in the spectrum to all other *peak* values in the spectrum. It is assumed that the peak heights are all independent. In practice 'peaks' are defined as values such that 'the power is greater than both the powers at adjacent frequencies'. A potential problem is that peaks may thus occur at close frequencies, which may invalidate the independent assumption.

A caution: in this paper, 'log' indicates logarithms taken to the base e.

2 DISTRIBUTION OF THE MAXIMUM OF $I_e(\omega)$.

Denote the upper order statistic (i.e. the maximum) of M independent but identically distributed random variables $v_1, v_2, \ldots, v_M$ by $v_{(M)} \equiv V$. Then it is easy to show that the CDF of V is

$$F_V(V) = [F_v(V)]^M$$

where F_v is the CDF of the v_j. Periodogram ordinates as defined in equation (1) are exponentially distributed with mean value

$$\mathrm{E}I_e(\omega) = \sigma_e^2 .$$

It follows that, to very good approximation $v = I_e(\omega)/\overline{I_e}$ is exponentially distributed with $\mathrm{E}v = 1$, i.e. PDF

$$f_v(v) = \mathrm{e}^{-v}$$

and corresponding CDF

$$F_v(v) = 1 - \mathrm{e}^{-v} .$$

If measurements are regularly spaced in time then periodogram values in the $M = N/2$ Fourier frequencies in equation (2) are independent, and hence the standardized periodogram has the one-parameter EVD

$$F_V(V) = [1 - \mathrm{e}^{-V}]^M . \quad (5)$$

It is well known in theory of EVDs that the exponential distribution is in the 'domain of attraction' of the Gumbel distribution, and hence the distribution of $W = V - \log M$ approaches the Gumbel form as $M \to \infty$ (e.g. Castillo et al. 2005):

$$F_W(W) = \exp\left[-\mathrm{e}^{-W}\right] .$$

The corresponding PDF is

$$f_W(W) = \exp\left[-\mathrm{e}^{-W} - W\right] .$$

It is more convenient to work in terms of the periodogram maximum V than W, i.e. the PDF

$$f_V(V) = \exp\left[-\mathrm{e}^{-(V-\log M)} - (V - \log M)\right] \quad (6)$$

is a useful limiting form of equation (5) as $M \to \infty$.

The Gumbel distribution is a special case of the GEV (4). The two-parameter Gumbel PDF obtained when $k = 0$ is

$$f(x) = \frac{1}{\sigma} \exp\left\{-\left[\frac{(x-\mu)}{\sigma}\right] - \exp\left[-\frac{(x-\mu)}{\sigma}\right]\right\} , \quad (7)$$

i.e. as M increases, equation (5) approaches a two-parameter Gumbel distribution with $\sigma = 1$, $\mu = \log M$.

The distribution (7) will play an important role below in the consideration of the maxima of oversampled periodograms. Parameters of equation (7) can be obtained by the maximum likelihood technique, a methodology which is described in any textbook on statistical inference. Briefly, the likelihood function of K independent random variables $\{x_1, x_2, \ldots, x_K\}$, with common PDF $f(x)$, is

$$L = \prod_{k=1}^{K} f(x_k) .$$

The method consists simply of the maximization of L (or more commonly $\log L$) with respect to any unknown parameters in the PDF $f(x)$ [e.g. μ and σ in the case of equation (7)].

Maximum likelihood estimation of the parameters in equation (7) is easily reduced to a single non-linear equation in σ. First, differentiation of the log likelihood function

$$\mathcal{L} = -m \log \sigma - m\frac{\overline{x} - \mu}{\sigma} - \mathrm{e}^{\mu/\sigma} \sum_k \mathrm{e}^{-x_k/\sigma} \tag{8}$$

with respect to μ leads to

$$\widehat{\mu} = \widehat{\sigma} \left[\log m - \log \sum_k \exp(-x_k/\widehat{\sigma}) \right] . \tag{9}$$

In equations (8) and (9), m denotes the number of periodogram values x_k under consideration. Equation (9) can be substituted into the equation obtained by differentiating equation (8) with respect to σ to find

$$\widehat{\sigma} = \overline{x} - \frac{\sum_k x_k \exp(-x_k/\widehat{\sigma})}{\sum_k \exp(-x_k/\widehat{\sigma})} . \tag{10}$$

In the first instance consideration is now given to the proposition that in the case of oversampling, the distribution of periodogram maxima could be described by the GEV, rather than the one-parameter Gumbel distribution. As the reader will see, the simplified two-parameter Gumbel form of the GEV is sufficient for the task. Since the problem is not necessarily solvable by analytic means, we resort to simulations.

The results below are based on simulated data sets of sizes $N = 10\,000$, 20 000, 50 000, 80 000, 100 000, 130 000. For each N, at least 20 000 Gaussian white noise realizations were generated. Periodograms were calculated, oversampled by rates of $R = 0$, 1, 2, 4, 6, 8, 10, 14, 20, i.e. periodograms were calculated in the frequencies

$$\omega_j = \frac{2\pi j}{(R+1)N}, \quad j = 1, 2, \ldots, (R+1)\frac{N}{2} = m \, . \tag{11}$$

For each periodogram, the value of

$$V = \max_\omega \left[I_\mathrm{e}(\omega)/\overline{I_\mathrm{e}} \right]$$

was noted.

GEV distributions of the form (4) were fitted to the ≥20 000 periodogram maxima corresponding to each of the combinations of N and R. This was accomplished by numerical maximization of the GEV likelihood function. Only 4 (out of 54) values of the shape parameter k differed from zero by more than 1.5σ; the largest (in absolute value) being 0.013. Effectively, this means that the two-parameter Gumbel distribution (7) provides, statistically, as good a description of the extreme values as does the full three-parameter GEV (4).

Equations (9) and (10) were therefore used to fit two Gumbel distribution models to the simulated periodogram maxima: in the first, $\sigma = 1$ was assumed, and equation (9) used to estimate $\widehat{\mu}$. In the second model, both parameters were free. Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) one-sample statistics were used to assess the goodness-of-fit of the two models. Although the model with fixed $\sigma = 1$ was found to be acceptable (by a wide margin) for $R = 0$, 1, it did not fit data with larger oversampling factors well. The second model, on the other hand, always fitted well – p-values of the goodness-of-fit statistics were generally larger than 0.5, the smallest being 0.13 and 0.18 for the KS and AD tests, respectively.

[It should be admitted that a shortcut was taken in these tests: estimated parameters were treated as known. However, since the sample sizes are large (at least 20 000), confidence intervals for the estimated parameters are quite narrow. The implication is that the parameters are very well determined.]

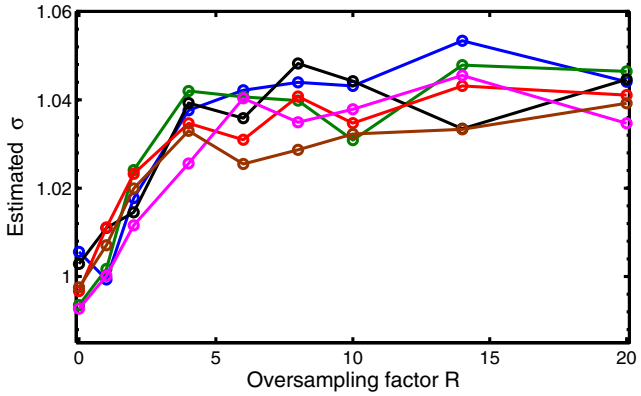

Figure 1. Values of $\widehat{\sigma}$ in the two-parameter Gumbel PDF (7), for different samples sizes N and oversampling factors R. Lines connect results for the same N.

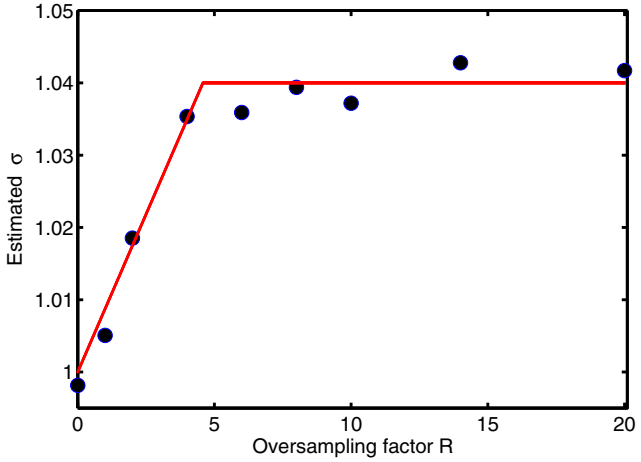

Figure 2. Dots are the means, at each oversampling rate R, of the data plotted in Fig. 1. The lines are a simple representation (equation 12) of the dependence of the mean values on R.

Estimated values of σ are plotted in Fig. 1, with lines connecting $\widehat{\sigma}$ for the same N (but different R). The general impression is that of a steep, approximately linear, rise for $R \leq 6$, followed by a slower increase with increasing R, reaching a plateau near $R \sim 14$. The implication is that the width of the PDF increases slightly with increasing R, but that a 'saturation point' is reached for very well-sampled periodograms. Given the substantial scatter at any given value of R, only the very simple broken line model

$$\widehat{\sigma} = \begin{cases} 1 + BR & R \leq R_0 \\ 1.04 & R \geq R_0 \end{cases}$$

was fitted to the data by least squares. The optimal solution has breakpoint $R_0 = 4.6$ and slope $B = 0.0087$ – see Fig. 2 in which the fit to the means (over N) at various oversampling factors is plotted. The situation can be summarized by saying that $\sigma = 1$ for $R = 0$, 1 (see above); $\sigma \to 1.04$ for large R ($R > 5$); and $\sigma \approx 1 + 0.0087R$ for intermediate values of R.

In summary,

$$\sigma \approx \begin{cases} 1 + 0.0087R & R \leq 4.6 \\ 1.04 & R \geq 4.6. \end{cases} \tag{12}$$

The equivalent of Fig. 1, but for the estimated Gumbel parameter μ, can be seen in Fig. 3. Plotted values have been scaled by

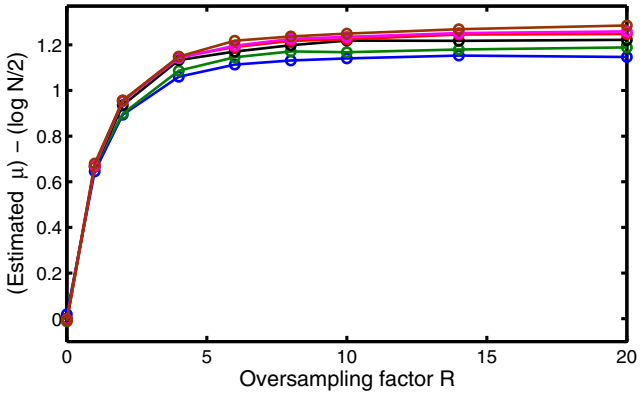

Figure 3. Similar to Fig. 1, but for the parameter $\widehat{\mu}$. Clearly asymptotic values are approached as R is increased, with limiting values increasing slightly as N is increased.

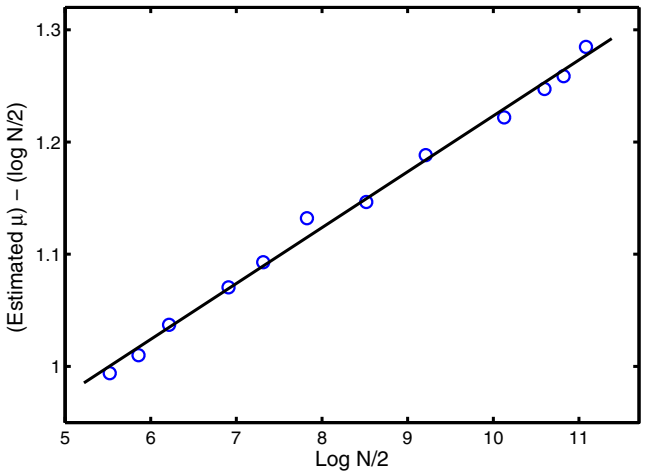

Figure 4. Open circles show values of $\widehat{\mu} - \log N/2$ when oversampling with $R = 20$, for sample sizes ranging from $N = 500$ to $N = 130\,000$. The line is the linear least-squares fit to the points.

subtraction of $\log N/2$ (see equation 9). As expected, the number of independent frequencies $M = N/2$ if there is no oversampling, hence the scaled value of the mean is zero for $R = 0$. As the rate of oversampling increases, $\mu - \log N/2$ approaches an asymptotic limit, which is larger, the larger the N.

The scaled values of the estimated mean parameter μ at $R = 20$ are plotted in Fig. 4, which has been supplemented by results for $N = 500$, 700, 1000, 2000, 3000, 5000. The least-squares linear fit is

$$\widehat{\mu} - \log N/2 = 0.725(0.016) + 0.0498(0.0018) \log N/2 \,, \qquad (13)$$

where quantities in brackets are standard errors. The implication is that the asymptotic Gumbel distribution means can be closely approximated by

$$\mu = 0.032(0.016) + 1.05(0.002) \log N \approx 1.05 \log N \,, \qquad (14)$$

at least for $500 \leq N \leq 130\,000$.

It is possible to derive a fairly accurate general formula for $\widehat{\mu}$, as a function of both the sample size N and the oversampling factor R. Fig. 5 is a modified version of Fig. 3, in which equation (13) has been used to normalize $\widehat{\mu}$:

$$g(N, R) = [\widehat{\mu} - \log N/2]/[0.725 + 0.05 \log N/2] \,.$$

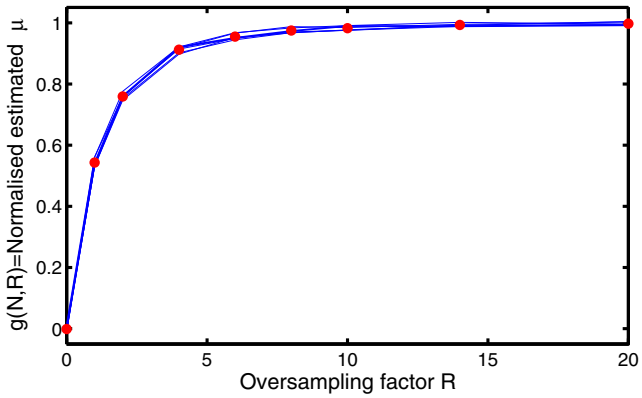

Figure 5. A normalized version of the data in Fig. 3 (see text). The dots represent the mean values at each oversampling rate.

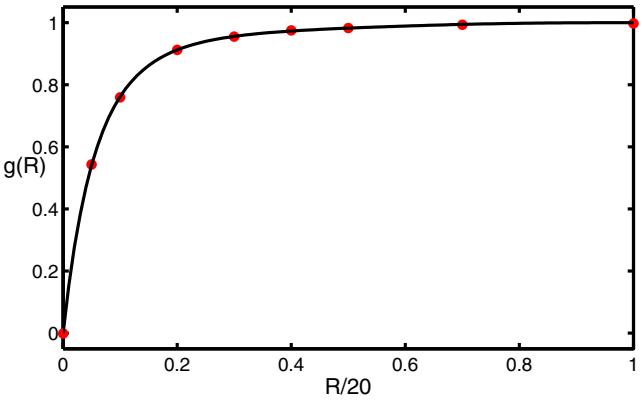

Figure 6. The three-parameter function (15) (solid line) fitted to the averaged data (over N, at each R) in Fig. 5 (dots).

It is clear that the dependence on R is very similar for different N. This dependence can be modelled quite well by

$$g(R) = 1 - \exp\left[-16.92(0.11)(R/20) + 27.9(1.1)(R/20)^2 - 20.3(2.0)(R/20)^3\right] . \qquad (15)$$

The quality of the fit is demonstrated in Fig. 6: it is excellent. It follows that generally

$$\mu \approx \log \frac{N}{2} + (0.725 + 0.05 \log N/2) g(R). \qquad (16)$$

Equations (12), (16) and (15) summarize the results of the simulation experiments. These formulae can provide accurate percentage points for a wide range of sample sizes and periodogram oversampling rates: from

$$p = \Pr(V > x) = 1 - F_V(x) = 1 - \exp\left\{-\exp\left[-\frac{(x-\mu)}{\sigma}\right]\right\} \qquad (17)$$

the percentiles follow as

$$x_{\rm p} = \mu - \sigma \log[-\log(1-p)] \,. \qquad (18)$$

The consistency between percentiles $x_{\rm p}$ calculated from (18), and percentiles $x_{\rm s}$ determined directly from simulation results, can be

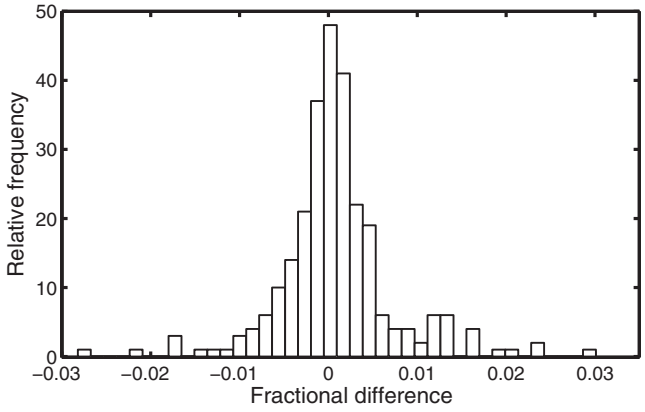

Figure 7. The scaled differences between simulated percentiles and those calculated from equations (12), (15) and (16).

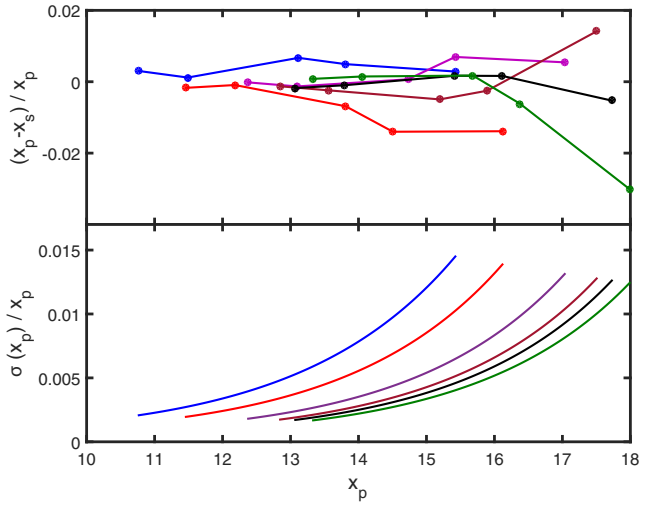

Figure 8. Top panel: fractional differences between the simulated percentiles x_s and the predicted percentiles x_p, for no oversampling. Lines connect values for different p (ranging from 0.1 to 0.001), at fixed N. Bottom panel: theoretical uncertainties in the percentiles, as calculated from order statistics in samples of size $K = 20\,000$ (see Appendix B and Fig. B1). Uncertainties have been scaled by x_p for easy comparison with the graphs in the top panel. The different curves are for the same collection of values of N as in the top panel ($N = 10\,000$, $20\,000$, $50\,000$, $80\,000$, $100\,000$, $130\,000$). For each curve p ranges from 0.1 at the bottom left to 0.001 at the top right.

investigated by studying the fractional differences

$$(x_p - x_s)/x_p \ .$$

Fig. 7 is a histogram of the fractional differences (where x_s are the percentiles extracted from the 20 000 (or more) simulated data sets), for p-values 0.1, 0.05, 0.01, 0.005, 0.001. All fractional differences greater than 1.5 per cent correspond to the 0.1th percentile (i.e. $p = 0.001$); of course, even with 20 000 simulations these values of x_s are poorly determined.

The top panel of Fig. 8 shows fractional differences between the theoretical percentiles x_p for the zero oversampling case, and the corresponding simulated values x_s. Since in this case equation (17) is exact, the information in Fig. 7 can be placed in context. As expected, the percentage differences between x_p and x_s are small for the largest percentiles, and increase with decreasing percentile value. Since x_p is exact, the diagram essentially reflects the uncertainties in x_s, the simulated percentiles. Comparison of Figs 7 and 8 suggests that the differences between x_s and x_p are no worse for $R > 0$ than for $R = 0$.

The accuracy of simulation of the large percentiles, associated with small values of p, is discussed in more detail in Appendix B. The bottom panel of Fig. 8 displays the theoretical uncertainties, scaled by x_p. Comparison of the two parts of the plot shows good agreement.

3 CONCLUSIONS

The simulation results of Section 2 suggest that the maxima of the fully oversampled noise periodogram (divided by $\overline{I_e}$) are distributed as the two-parameter Gumbel distribution (7), with $\sigma \approx 1.04$ and $\mu \approx 1.05 \log N$. Compared to the simple one-parameter form (6), the scale parameter σ is increased from unity to 1.04 (i.e. a 4 per cent increase in the distribution width), and the location parameter μ is shifted to larger values. The latter result, in particular, is hardly surprising, given that the larger R, the periodogram maxima are more fully resolved. The PDF is

$$f_V(V) = (1/1.04) \exp\left\{ -\frac{(V - 1.05 \log N)}{1.04} - \exp\left[-\frac{(V - 1.05 \log N)}{1.04} \right] \right\} . \quad (19)$$

Maxima of partially oversampled spectra are similarly distributed, with σ obtainable from equation (12) and μ given by equation (16). Once μ and σ have been calculated for given N and R, p-values follow from equation (17), or percentiles from equation (18).

Table 1 gives selected percentiles for a few values of N, for the fully oversampled case, and, for purposes of comparison, percentiles calculated from equation (3). The percentiles for the fully oversampled case are, of course, larger. Note for example that for large N, the 0.1 per cent points calculated from equation (3) are similar to the 0.5 per cent points corresponding to equation (19). As expected, use of equation (3) would lead to overestimates of the significance of peaks in oversampled spectra by as much as a factor ∼5.

It is tempting to think that the corresponding percentage points of the scaled amplitude spectrum are simply square roots of the entries in Table 1. However, it is shown in Appendix C that the correct transformation is

$$V_S = \frac{2}{\sqrt{\pi}} \sqrt{V_I} = 1.128 \sqrt{V_I} \ ,$$

where V_S and V_I are, respectively, the amplitude and power spectrum statistics.

As the oversampling factor $R \to \infty$, the periodogram is fully resolved and approaches a continuous stochastic process, a point addressed by Baluev (2008). Although some relevant theory has been available for a considerable time (Sharpe 1978), a full solution using this fact is still lacking.

ACKNOWLEDGEMENTS

A very careful reading of the manuscript by the referee, Dr Mathias Zechmeister, is appreciated. His comments led to an improved version of the paper. This research was partially funded by a South African National Research Foundation grant.

Table 1. Upper percentage points of $V_I = \max[I_e(\omega)]/\overline{I_e}$, for the fully oversampled case (left-hand columns) and for the case of no oversampling (right-hand columns).

	Percentiles, fully oversampled case					Percentiles from equation (3)				
	0.1	0.05	0.01	0.005	0.001	0.1	0.05	0.01	0.005	0.001
500	8.87	9.61	11.31	12.03	13.71	7.77	8.49	10.12	10.82	12.43
750	9.29	10.04	11.74	12.46	14.13	8.18	8.90	10.53	11.22	12.83
1000	9.59	10.34	12.04	12.76	14.44	8.47	9.18	10.81	11.51	13.12
1500	10.02	10.77	12.46	13.19	14.86	8.87	9.59	11.22	11.92	13.53
2000	10.32	11.07	12.77	13.49	15.16	9.16	9.88	11.51	12.20	13.82
3000	10.75	11.50	13.19	13.91	15.59	9.56	10.28	11.91	12.61	14.22
5000	11.28	12.03	13.73	14.45	16.13	10.07	10.79	12.42	13.12	14.73
7500	11.71	12.46	14.15	14.88	16.55	10.48	11.20	12.83	13.53	15.14
10 000	12.01	12.76	14.46	15.18	16.85	10.77	11.49	13.12	13.81	15.42
15 000	12.44	13.19	14.88	15.60	17.28	11.17	11.89	13.52	14.22	15.83
25 000	12.97	13.72	15.42	16.14	17.82	11.68	12.40	14.03	14.73	16.34
50 000	13.70	14.45	16.14	16.87	18.54	12.38	13.10	14.73	15.42	17.03
75 000	14.13	14.88	16.57	17.29	18.97	12.78	13.50	15.13	15.83	17.44
100 000	14.43	15.18	16.87	17.60	19.27	13.07	13.79	15.42	16.12	17.73
130 000	14.70	15.45	17.15	17.87	19.55	13.33	14.05	15.68	16.38	17.99
175 000	15.02	15.77	17.46	18.18	19.86	13.63	14.35	15.98	16.68	18.29
250 000	15.39	16.14	17.83	18.56	20.23	13.99	14.71	16.34	17.03	18.64
400 000	15.88	16.63	18.33	19.05	20.73	14.46	15.18	16.81	17.50	19.11
700 000	16.47	17.22	18.92	19.64	21.32	15.02	15.74	17.37	18.06	19.67
1000 000	16.85	17.60	19.29	20.01	21.69	15.37	16.09	17.72	18.42	20.03

REFERENCES

Ahsanullah M., Nevzprov V. B., Shakil M., 2013, An Introduction to Order Statistics. Atlantis Press, Amsterdam

Anderson T. W., 1971, The Statistical Analysis of Time Series. Wiley, New York

Baluev R. V., 2008, MNRAS, 385, 1279

Breger M. et al., 1993, A&A, 271, 482

Castillo E., Hadi A. S., Balakrishnan N., Sarabia J. M., 2005, Extreme Value and Related Models with Applications in Engineering and Science. Wiley, Hoboken, NJ

Cuypers J., 2012, in Griffin R. E. M., Hanisch R. J., Seaman R., eds, Proc. IAU Symp. 285, New Horizons in Time-Domain Astronomy. Cambridge Univ. Press, Cambridge, p. 299

Frescura F. A. M., Engelbrecht C. A., Frank B. S., 2008, MNRAS, 388, 1693

Gilliland R. L. et al., 2010, PASP, 122, 131

Mosteller F., 1946, Ann. Math. Stat., 17, 377

Scargle J. D., 1982, ApJ, 263, 835

Sharpe K., 1978, Adv. Appl. Probab., 10, 373

Sturrock P. A., Scargle J. D., 2010, ApJ, 718, 527

Süveges M., 2014, MNRAS, 440, 2099

APPENDIX A: AUTOCORRELATION FUNCTION OF THE NOISE SPECTRUM I_e

Trigonometric identities can be used to easily show that equation (1) is equivalent to

$$I_y(\omega) = \frac{1}{N}\sum_j \sum_k (y_j - \overline{y})(y_k - \overline{y})\cos(j-k)\omega\ . \qquad \text{(A1)}$$

Proceeding from equation (A1), the covariance between periodogram values in angular frequencies ω and ψ is

$$\begin{aligned} C_e(\omega,\psi) &= \mathrm{cov}\,[I_e(\omega), I_e(\psi)] \\ &= \mathrm{E}I_e(\omega)I_e(\psi) - \mathrm{E}I_e(\omega)\mathrm{E}I_e(\psi) \\ &= \frac{1}{N^2}\sum_{j,k=1}^{N}\sum_{r,s=1}^{N}\cos(j-k)\omega\,\cos(r-s)\psi\,\mathrm{E}e_j e_k e_r e_s \\ &\quad - \mathrm{E}I_e(\omega)\mathrm{E}I_e(\psi). \end{aligned}$$

For Gaussian e_j,

$$\mathrm{E}e_j e_k e_r e_s \equiv \sigma(j,k)\sigma(r,s) + \sigma(j,r)\sigma(k,s) + \sigma(j,s)\sigma(k,r),$$

where $\sigma^2(j,k) \equiv Ee_j e_k = \sigma_e^2\delta_{jk}$ (e.g. Anderson 1971, p. 444). Furthermore, again from equation (A1),

$$\mathrm{E}I_e(\omega)\mathrm{E}I_e(\psi) = \frac{1}{N^2}\sum_{j,k=1}^{N}\sum_{r,s=1}^{N}\cos(j-k)\omega\,\cos(r-s)\psi\,\mathrm{E}e_j e_k \mathrm{E}e_r e_s$$

$$= \frac{1}{N^2}\sum_{j,k=1}^{N}\sum_{r,s=1}^{N}\cos(j-k)\omega\,\cos(r-s)\psi\,\sigma(j,k)\sigma(r,s).$$

It follows that

$$\begin{aligned} C_e(\omega,\psi) &= \frac{1}{N^2}\sum_{j,k=1}^{N}\sum_{r,s=1}^{N}\cos(j-k)\omega\,\cos(r-s)\psi \\ &\quad \times [\sigma(j,r)\sigma(k,s) + \sigma(j,s)\sigma(k,r)] \\ &= \frac{2}{N^2}\sum_{j,k=1}^{N}\sum_{r,s=1}^{N}\cos(j-k)\omega\,\cos(r-s)\psi\,\sigma(j,r)\sigma(k,s) \\ &= \frac{2\sigma_e^4}{N^2}\sum_{j,k=1}^{N}\cos(j-k)\omega\,\cos(j-k)\psi \\ &= \frac{2\sigma_e^4}{N^2}\Big\{N + 2[(N-1)\cos\omega\cos\psi + (N-2) \end{aligned}$$

$$\times \cos 2\omega \cos 2\psi + \cdots + \cos(N-1)\omega \cos(N-1)\psi] \Big\}$$

$$= \frac{2\sigma_e^4}{N^2}\left[N + 2\sum_{j=1}^{N-1}(N-j)\cos j\omega \cos j\psi\right], \qquad \text{(A2)}$$

which is more convenient for explicit evaluation. For example, for $\omega = \psi$,

$$\sum_{j=1}^{N-1}\cos^2 j\omega = \frac{1}{2}\sum_{j=1}^{N-1}(1+\cos 2j\omega)$$

$$= \frac{N-2}{2} + \frac{1}{2}\cos(N-1)\omega \sin N\omega \operatorname{cosec}\omega$$

$$\sum_{j=1}^{N-1} j\cos^2 j\omega = \frac{1}{2}\sum_{j=1}^{N-1} j(1+\cos 2j\omega)$$

$$= \frac{N(N-1)}{4} + \frac{N-1}{2\sin\omega}\sin(2N-1)\omega$$

$$- \frac{\sin^2(N-1)\omega}{2\sin^2\omega}$$

so that

$$C_e(\omega,\omega) = \frac{(N-1)\sigma_e^4}{N} + \frac{\sigma_e^4}{N^2\sin\omega}[2N\cos(N-1)\omega\sin N\omega$$

$$-(N-1)\sin(2N-1)\omega + \sin^2(N-1)\omega\operatorname{cosec}\omega] \ .$$

For $N \gg 1$, this reduces to

$$\mathrm{var}[I_e(\omega)] \approx \sigma_e^4 \ . \qquad \text{(A3)}$$

Although an exact expression for the case $\psi \neq \omega$ can be derived, the algebra is tedious and the precision not required in the present context of large N. Instead, note from equation (A2) that

$$C_e(\omega,\psi) = 2\sigma_e^4 \sum_{j=1}^{N}\sum_{k=1}^{N}\cos N\left(\frac{j}{N}-\frac{k}{N}\right)\omega$$

$$\times \cos N\left(\frac{j}{N}-\frac{k}{N}\right)\psi \ \frac{1}{N}\frac{1}{N}$$

$$\to 2\sigma_e^4\int_0^1\int_0^1 \cos N(x-y)\omega \cos N(x-y)\psi \ \mathrm{d}x\,\mathrm{d}y$$

$$= \sigma_e^4\int_0^1\int_0^1 [\cos N(x-y)(\omega+\psi)$$

$$+\cos N(x-y)(\omega-\psi)]\,\mathrm{d}x\,\mathrm{d}y$$

$$= \frac{2\sigma_e^4}{N^2}\left[\frac{1-\cos N(\omega+\psi)}{(\omega+\psi)^2} + \frac{1-\cos N(\omega-\psi)}{(\omega-\psi)^2}\right]. \qquad \text{(A4)}$$

Clearly, for large N, the covariances between periodogram ordinates are generally $\sim 1/N^2$, i.e. negligible. In order to study the short-range autocovariance, let $\psi = \omega + 2\pi\Delta$, with $|\Delta| \ll 1$ the *frequency* difference,

$$C_e(\omega,\omega+2\pi\Delta) \approx \frac{2\sigma_e^4}{N^2}\frac{1-\cos 2\pi N\Delta}{(2\pi\Delta)^2}$$

$$= \frac{2\sigma_e^4}{N^2}\frac{2\sin^2\pi N\Delta}{(2\pi\Delta)^2}$$

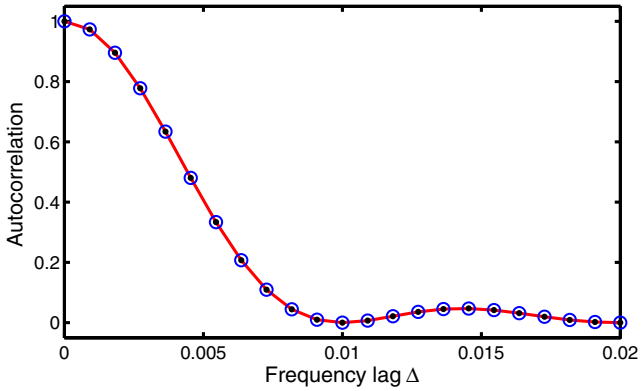

Figure A1. Approximation of the exact autocorrelation function of the noise spectrum, as derived from (A2) (dots), compared to approximations by the integration formula (A4) (open circles) and the sinc^2 function (A5) (solid line), for $N = 100$.

$$= \sigma_e^4\left(\frac{\sin\pi N\Delta}{\pi N\Delta}\right)^2$$

$$= \sigma_e^4 \mathrm{sinc}^2(N\Delta). \qquad \text{(A5)}$$

The autocorrelation function follows from equations (A3) and (A5) as

$$\rho(\Delta) \approx \mathrm{sinc}^2(N\Delta) \ . \qquad \text{(A6)}$$

The levels of approximation of equations (A4) and (A5) to equation (A2) are shown in Fig. A1, for $N = 100$; both approximations are excellent, even for such relatively small N. Clearly, the autocorrelation is effectively zero outside the range $(-1/N, +1/N)$. Interestingly, the autocorrelation function (A6) is exactly zero for frequencies spaced integer multiples of $1/N$ apart – compare this with equation (2).

APPENDIX B: THE ACCURACY OF SIMULATED PERCENTILES

It is assumed that the two-parameter Gumbel CDF

$$F(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \ , \qquad \text{(B1)}$$

with associated PDF (7), is indeed the correct distribution. (This is not critical – it is only required that it provides a reasonably accurate representation of the true distribution of periodogram maxima.) A large sample $\{x_j\}$ of size K is drawn (by computer) from this distribution, and percentiles are estimated from the order statistics. Clearly, the accuracy of the estimated percentiles will be determined by the sampling variability of the order statistic.

Let the percentage of interest be p, with associated percentile x_p. The order statistic found from the sample is

$$\widehat{x}_p = x_{(pK)} \ ,$$

e.g. if $K = 20\,000$, then the upper 1 per cent point is estimated by $x_{(19800)}$. Mosteller (1946) showed that asymptotically, order statistics are normally distributed with variances given by

$$\mathrm{var}(\widehat{x}_p) = \frac{p(1-p)}{Kf^2(x_p)} \ . \qquad \text{(B2)}$$

This clearly demonstrates the substantial increase in variability in the low-probability $[f(x) \ll 1]$ tails of the distribution. From

equation (7),

$$f(x_\mathrm{p}) = \frac{1}{\sigma}\exp\left\{-\exp\left[-\frac{(x_\mathrm{p}-\mu)}{\sigma}\right] - \left[\frac{(x_\mathrm{p}-\mu)}{\sigma}\right]\right\}$$

$$= \frac{1}{\sigma}F(x_\mathrm{p})\exp\left[-\frac{(x_\mathrm{p}-\mu)}{\sigma}\right].$$

Since $p = F(x_\mathrm{p})$, it follows that

$$\log p = -\exp\left[-\frac{(x_\mathrm{p}-\mu)}{\sigma}\right]$$

and hence

$$f(x_\mathrm{p}) = -p\log p/\sigma\ .$$

Substitution into equation (B2) then gives

$$\mathrm{var}(\widehat{x}_\mathrm{p}) = \frac{(1-p)\sigma^2}{Np(\log p)^2}\ . \tag{B3}$$

An alternative approach, which does not rely on asymptotics, proceeds from the exact distribution (e.g. Ahsanullah, Nevzprov & Shakil 2013)

$$G(x_\mathrm{p}) = I\left[Kp, K-Kp+1; F(x_\mathrm{p})\right], \tag{B4}$$

where

$$I(a,b;x) = \frac{1}{B(a,b)}\int_0^x u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$$

is the incomplete beta function, and $B(a, b)$ the standard beta function. The CDF (B4) could be used to obtain confidence intervals for x_p. Instead, we use it to determine, for given p and K, 16, 50 and 84 per cent points in the distribution of $\widehat{x}_\mathrm{p}$: these can be used to obtain two estimates of the usual 'σ' (i.e. the central extent of the distribution which contains 68 per cent of the probability). This simply requires numerical solution of the three equations

$$G(u_0) = 0.5 \qquad G(u_1) = 0.16 \qquad G(u_2) = 0.84$$

followed by

$$\sigma_1 = u_0 - u_1 \qquad \sigma_2 = u_2 - u_0\,. \tag{B5}$$

Of course, if the distribution of the order statistic is symmetrical (which is asymptotically the case, according to the Mosteller 1946 result), then $\sigma_1 = \sigma_2$.

Two sets of results are given in Fig. B1, for $K = 20\,000$ and $K = 40\,000$, respectively. Circles and squares respectively denote σ_1 and σ_2. The latter are virtually indistinguishable from the asymptotic results (lines). The analytical result (B3) is clearly more than adequate for present purposes, particularly for the larger K.

Perhaps the most important point demonstrated by these results is the substantial uncertainty in the 99.9 per cent points obtained from the simulations.

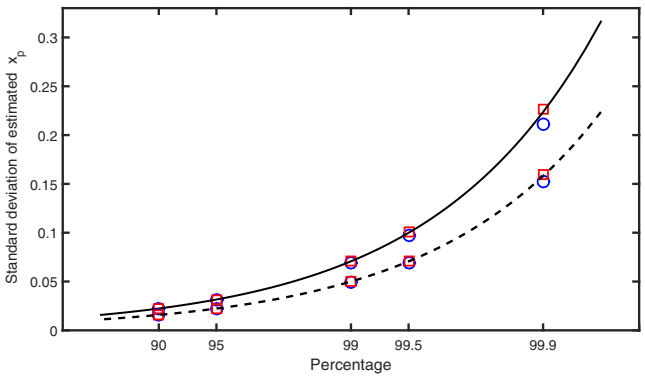

Figure B1. The uncertainty in the percentage points derived from simulation. The solid and broken lines show the asymptotic results equation (B3) for $K = 20\,000$ and 40 000, respectively. Circles and squares, respectively, denote σ_1 and σ_2 in equation (B5).

APPENDIX C: RELATION BETWEEN PERCENTAGE POINTS OF SCALED AMPLITUDE SPECTRA AND SCALED PERIODOGRAMS

The PDF of $I(\omega)$ is exponential:

$$f_I(x) = \frac{1}{\sigma^2}\exp(-x/\sigma^2),$$

where σ^2 is the variance of the time series. The expected value (mean) of the periodogram is

$$\mathrm{E}I(\omega) = \sigma^2\ . \tag{C1}$$

A common definition of the amplitude spectrum is

$$S(\omega) = 2\sqrt{\frac{I(\omega)}{N}}\ . \tag{C2}$$

It is not difficult to show that the PDF of S is then of the Rayleigh form

$$f_S(S) = \frac{N}{2\sigma^2}S\exp\left(-\frac{NS^2}{4\sigma^2}\right)\ .$$

The corresponding expected value is

$$\mathrm{E}S(\omega) = \sigma\sqrt{\frac{\pi}{N}} = \sqrt{\frac{\pi\mathrm{E}I(\omega)}{N}}. \tag{C3}$$

The statistic of interest is the scaled maximum of the amplitude spectrum, i.e.

$$V_S = \max_\omega S(\omega)/\overline{S(\omega)}\ .$$

From equations (C1)– (C3), it follows that

$$V_S = 2\sqrt{\max_\omega I(\omega)}/[\sqrt{N}\ \overline{S(\omega)}]$$

$$\approx 2\sqrt{\max_\omega I(\omega)}/[\sqrt{N}\ \mathrm{E}S]$$

$$= 2\sqrt{\max_\omega I(\omega)/[\pi\mathrm{E}I]}$$

$$\approx \frac{2}{\sqrt{\pi}}\sqrt{V_I}, \tag{C4}$$

where V_I is the scaled maximum of the periodogram:

$$V_I = \frac{\max_\omega I(\omega)}{\overline{I(\omega)}}\ .$$

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