Mon. Not. R. Astron. Soc. **385,** 1279–1285 (2008) doi:10.1111/j.1365-2966.2008.12689.x

Assessing the statistical significance of periodogram peaks

R. V. Baluev★

Sobolev Astronomical Institute, St Petersburg State University, Universitetskij prospekt 28, Petrodvorets, St Petersburg 198504, Russia

Accepted 2007 November 2. Received 2007 November 2; in original form 2007 June 13

ABSTRACT

The least-squares (or Lomb–Scargle) periodogram is a powerful tool that is routinely used in many branches of astronomy to search for periodicities in observational data. The problem of assessing the statistical significance of candidate periodicities for a number of periodograms is considered. Based on results in extreme value theory, improved analytic estimations of false alarm probabilities are given. These include an upper limit to the false alarm probability (or a lower limit to the significance). The estimations are tested numerically in order to establish regions of their practical applicability.

Key words: methods: data analysis – methods: statistical – surveys.

1 INTRODUCTION

When analysing astronomical time series, it is often necessary to choose between at least two hypotheses, a base one $\mathcal{H}$ and an alternative one $\mathcal{K}$, based on the data array. In the signal detection problem, it is necessary to check whether the observations are consistent with some base model or whether they contain an extra deterministic signal. Under the presence of random errors, such problems can be solved only in a probabilistic sense. It is possible to make two types of mistakes, namely the false retraction of $\mathcal{H}$ (the false alarm) and the false non-retraction of $\mathcal{H}$ (the false non-detection). False alarms are generally believed to be the more dangerous, and hence the problem of the estimation of the false alarm probability (hereafter FAP) associated with a candidate signal is very important. Given some small critical value FAP_* (usually between 10^{-3} and 0.1), we can claim that the candidate signal is statistically significant (if $\mathrm{FAP} < \mathrm{FAP}_*$) or is not (if $\mathrm{FAP} > \mathrm{FAP}_*$).

For the Lomb (1976)–Scargle (1982) periodogram (hereafter the L–S periodogram), the base hypothesis is that the observations incorporate only zero-mean uncorrelated and Gaussian errors (also called the white Gaussian noise). The alternative hypothesis is that a sinuous harmonic is also present. Every single value of the L–S periodogram represents a test statistic for the corresponding problem of hypothesis testing. In routine practical cases, however, the period of a possible signal is not known a priori, and it is necessary to scan many periodogram values within a wide frequency range. In this case, the FAP is provided by the probability distribution of the maximum periodogram value under the base hypothesis (i.e. no signal in the data). Existing methods of calculating this distribution for a continuous frequency range require time-consuming Monte Carlo simulations. The aim of the present paper is to propose analytic approximations that could allow Monte Carlo simulations to be avoided (at least in many practical cases). Such approximations of the distribution of the maximum have already been constructed by mathematicians specializing in the field of extreme values of random processes. In Section 3, these results are adapted for and extended to the specific features of the periodogram analysis of astronomical time series. In Section 4, numerical simulations are used to explore the quality of the analytic results and to show regions of their practical applicability.

2 GENERAL FORMULATIONS

Let us recover the principles of the periodogram analysis in a somewhat more general formulation than is usually used.[1]

Let $x_1, x_2 \ldots x_N$ be observations made at N epochs $t_1, t_2, \ldots t_N$. The errors of x_i are assumed to be independent and Gaussian with standard deviations σ_i. Each value of the periodogram can be recovered as a test statistic that allows one to calculate how likely is the hypothesis that the data contain a signal of a given frequency f. Mathematically, it should be checked whether the observations are fitted well by some base model having only $d_{\mathcal{H}}$ free parameters $\boldsymbol{\theta}_{\mathcal{H}}$, or whether they require an enlarged model of $d_{\mathcal{K}}$ parameters $\boldsymbol{\theta}_{\mathcal{K}} = \{\boldsymbol{\theta}_{\mathcal{H}}, \boldsymbol{\theta}\}$, with $d = d_{\mathcal{K}} - d_{\mathcal{H}}$ parameters $\boldsymbol{\theta}$ of an extra periodic signal. We will assume that, for any fixed frequency, both models are linear, and construct them by means of $d_{\mathcal{H}}$ and $d_{\mathcal{K}}$ base functions forming vectors $\boldsymbol{\varphi}_{\mathcal{H}}(t)$ and $\boldsymbol{\varphi}_{\mathcal{K}}(t, f) = \{\boldsymbol{\varphi}_{\mathcal{H}}(t), \varphi(t, f)\}$. Thus the base model to be fitted is $\mu_{\mathcal{H}}(t, \boldsymbol{\theta}_{\mathcal{H}}) = \boldsymbol{\theta}_{\mathcal{H}} \cdot \boldsymbol{\varphi}_{\mathcal{H}}(t)$, the model of the signal is $\mu(t, \boldsymbol{\theta}, f) = \boldsymbol{\theta} \cdot \boldsymbol{\varphi}\,(t, f)$, and the complete model to be fitted is $\mu_{\mathcal{K}}(t, \boldsymbol{\theta}_{\mathcal{K}}, f) = \boldsymbol{\theta}_{\mathcal{K}} \cdot \boldsymbol{\varphi}_{\mathcal{K}}(t, f) = \mu_{\mathcal{H}}(t, \boldsymbol{\theta}_{\mathcal{H}}) + \mu(t, \boldsymbol{\theta}, f)$. We wish to test whether the hypothesis $\mathcal{H} : \boldsymbol{\theta} = 0$ should be rejected in favour of the alternative $\mathcal{K}(f) : \boldsymbol{\theta} \neq 0$.

For the L–S periodogram, $d_{\mathcal{H}} = 0, d = 2$, and the signal model is given by a harmonic function $\theta_1 \cos \omega t + \theta_2 \sin \omega t$ (here $\omega = 2\pi f$). Schwarzenberg-Czerny (1998a,b) considered cases with

★E-mail: roman@astro.spbu.ru

[1] Much of the mathematical notation used in the present paper is described in Appendix A.

$d_{\mathcal{H}} = 0$ and arbitrary d. Ferraz-Mello (1981) put $d_{\mathcal{H}} = 1$ and added a floating constant term to the harmonic model with $d = 2$, and Cumming, Marcy & Butler (1999) accounted in addition for a possible linear trend ($d_{\mathcal{H}} = 2$).

An optimal statistical test for solving such problems in general is developed in chapter 7 of Lehman (1959). First, the minima are computed (by $\boldsymbol{\theta}_{\mathcal{H},\mathcal{K}}$) of the function $\chi^2 = \langle (x-\mu_{\mathcal{K}})^2 \rangle$ under hypotheses $\mathcal{H}$ and $\mathcal{K}(f)$. This can be done by means of any accessible linear least-squares algorithm (see also Schwarzenberg-Czerny 1998a,b). If the σ_i are known precisely, both minima, $\chi^2_{\mathcal{H}}$ and $\chi^2_{\mathcal{K}}(f)$, can be computed, and the least-squares periodogram can be defined as an advance in χ^2 provided by the transition from $\mathcal{H}$ to $\mathcal{K}(f)$:

$$z(f) = \left[\chi^2_{\mathcal{H}} - \chi^2_{\mathcal{K}}(f)\right]/2. \tag{1}$$

The error variances are often not known precisely and have to be estimated from the time series, explicitly or implicitly. It is usually assumed that $\sigma_i = \kappa \sigma_{\mathrm{meas},i}$, where the 'measured' uncertainties $\sigma_{\mathrm{meas},i}$ determine the weighting pattern of the time series, whereas the coefficient κ is unconstrained. In this case, only the ratio $\chi^2_{\mathcal{H}}/\chi^2_{\mathcal{K}}$ can be computed exactly, and the periodogram (1) has to be modified. We will consider the following modified periodograms:

$$z_1(f) = N_{\mathcal{H}} \frac{\chi^2_{\mathcal{H}} - \chi^2_{\mathcal{K}}(f)}{2\chi^2_{\mathcal{H}}}, \quad z_2(f) = N_{\mathcal{K}} \frac{\chi^2_{\mathcal{H}} - \chi^2_{\mathcal{K}}(f)}{2\chi^2_{\mathcal{K}}(f)}, \quad z_3(f) = \frac{N_{\mathcal{K}}}{2} \ln \frac{\chi^2_{\mathcal{H}}}{\chi^2_{\mathcal{K}}(f)}. \tag{2}$$

Here, $N_{\mathcal{H}} = N - d_{\mathcal{H}}$ and $N_{\mathcal{K}} = N - d_{\mathcal{K}}$ are the numbers of degrees of freedom in $\chi^2_{\mathcal{H}}$ and $\chi^2_{\mathcal{K}}$, respectively. The periodograms $z_1(f)$ and $z_2(f)$ are the well-known normalizations of $z(f)$ by variances of residuals under the respective hypotheses. All periodograms (2) are entirely equivalent because they are unique functions of each other:

$$2z_1/N_{\mathcal{H}} = 1 - \mathrm{e}^{-2z_3/N_{\mathcal{K}}}, \quad 2z_2/N_{\mathcal{K}} = \mathrm{e}^{2z_3/N_{\mathcal{K}}} - 1, \quad (1 - 2z_1/N_{\mathcal{H}})(1 + 2z_2/N_{\mathcal{K}}) = 1. \tag{3}$$

3 FALSE ALARM PROBABILITY

Let us pick any of the periodograms introduced above and denote it by $Z(f)$. If the frequency of a possible signal were known, the FAP could be retrieved as $\mathrm{FAP}_{\mathrm{single}} = 1 - P_{\mathrm{single}}(Z)$, where $P_{\mathrm{single}}(Z)$ is the cumulative distribution function of $Z(f)$ (taken under the base hypothesis). Under the hypothesis $\mathcal{H}$, the statistic $2z$ follows a χ^2-distribution with d degrees of freedom, $2z_2/d$ obeys a Fisher–Snedecor F-distribution with d and $N_{\mathcal{K}}$ degrees of freedom, and $2z_1/N_{\mathcal{H}}$ obeys a beta-distribution with the same numbers of degrees of freedom (Lehman 1959; section 7.1). Using relations (3), the distribution function of z_3 can be easily derived. The corresponding expressions of false alarm probability for $d = 2$ are given in Table 1. Note that the third modified periodogram obeys exactly the same distribution as the basic one if $d = 2$.

Now let us assume that we scan all frequencies from the interval $[0, f_{\max}]$ and look for the maximum value $Z_{\max} = \max_{[0,\,f_{\max}]} Z(f)$. Then the false alarm probability associated with this maximum is $\mathrm{FAP}_{\max} = 1 - P_{\max}(Z_{\max}, f_{\max})$, where $P_{\max}(Z_{\max}, f_{\max})$ denotes the cumulative distribution function of $Z_{\max}$ (under the base hypothesis). The precise expression for the latter distribution is not known even for equally spaced time series. It is always possible to use Monte Carlo simulations to obtain this function, but it is very time-consuming, especially for the most important region of low FAPs (high significances). The function $P_{\max}(Z, f_{\max})$ is often computed

Table 1. FAPs for the Lomb–Scargle periodogram and its modifications ($d = 2$).

$Z(f)$	$\mathrm{FAP}_{\mathrm{single}}(Z)$	$\tau(Z, f_{\max})$, approximately
$z(f)$	e^{-Z}	$W\mathrm{e}^{-Z}\sqrt{Z}$
$z_1(f)$	$\left(1 - \frac{2Z}{N_{\mathcal{H}}}\right)^{\frac{N_{\mathcal{K}}}{2}}$	$\gamma_{\mathcal{H}}\, W \left(1 - \frac{2Z}{N_{\mathcal{H}}}\right)^{\frac{N_{\mathcal{K}}-1}{2}} \sqrt{Z}$
$z_2(f)$	$\left(1 + \frac{2Z}{N_{\mathcal{K}}}\right)^{-\frac{N_{\mathcal{K}}}{2}}$	$\gamma_{\mathcal{K}}\, W \left(1 + \frac{2Z}{N_{\mathcal{K}}}\right)^{-\frac{N_{\mathcal{K}}}{2}} \sqrt{Z}$
$z_3(f)$	e^{-Z}	$\gamma_{\mathcal{K}}\, W \mathrm{e}^{-Z\left(1-\frac{1}{2N_{\mathcal{K}}}\right)} \sqrt{N_{\mathcal{K}} \sinh \frac{Z}{N_{\mathcal{K}}}}$

The factors $\gamma_{\mathcal{H},\mathcal{K}} = \sqrt{\frac{2}{N_{\mathcal{H},\mathcal{K}}}}\, \Gamma(\frac{N_{\mathcal{H}}}{2}) / \Gamma(\frac{N_{\mathcal{H}}-1}{2})$ can be neglected for $N_{\mathcal{H}} \geqslant 10$. If the spectral leakage is low, $\mathrm{FAP}_{\max} \approx \tau(Z, f_{\max})$ for realistic values of parameters. See text for detailed discussion.

(Schwarzenberg-Czerny 1998a,b) as

$$P_{\max}(Z, f_{\max}) \approx P_{\mathrm{single}}(Z)^{N_{\mathrm{ind}}(f_{\max})}, \tag{4}$$

where $N_{\mathrm{ind}}(f_{\max})$ is an effective 'number of independent frequencies' found within $[0, f_{\max}]$. There is no general analytic expression for the quantity N_{ind}, but a common method is to use a short Monte Carlo simulation to assess it and then to extrapolate (4) to low FAPs (Cumming 2004; Horne & Baliunas 1986). However, the multiple-trial formula (4) is only heuristic and is not necessarily precise even for equally spaced observations that do not produce significant aliasing.

A better estimation of $P_{\max}(Z, f_{\max})$ can be obtained using the theory of stochastic processes. The theory of extremes of random processes is developed in depth in the mathematical literature. For our aims, it is worthwhile to mention the series of works by Davies (1977, 1987, 2002). This author considered (in rather general formulations) extreme value distributions for χ^2, F, and beta random processes that can include our periodograms z and $z_{1,2}$ as special cases. The main result of these works is an analytic lower limit to the corresponding extreme value distributions. This result is potentially very useful for astronomical applications, because it yields directly an upper limit to the false alarm probability and a lower limit to the significance of a candidate periodicity. However, the formulae published in the cited papers are not yet ready for use and require some adaptations to specific applications. Moreover, these results can be improved so that they provide not only an upper limit but also a uniform approximation to the false alarm probability, which would be useful in the case of low spectral leakage.

A brief description of these results, adapted for the uneven time series analysis and with details of my extensions, is given in Appendix B. Summarizing them, the 'Davies bound' can be written as

$$\mathrm{FAP}_{\max}(Z, f_{\max}) \leqslant \mathrm{FAP}_{\mathrm{single}}(Z) + \tau(Z, f_{\max}). \tag{5}$$

The function τ will be specified below. If the aliasing effects can be neglected within the frequency band being scanned,[2] and if $f_{\max}$ is large enough, then

$$P_{\max}(Z, f_{\max}) \approx P_{\mathrm{single}}(Z)\, \mathrm{e}^{-\tau(Z, f_{\max})}. \tag{6}$$

The right-hand side in (5) should approach the FAP more closely for large Z (even the asymptotic equality under $Z \to \infty$ is expected,

[2] This means that the spectral window of the time series has no significant peaks in the doubled frequency band $[0, 2f_{\max}]$, except for the main one at $f = 0$.

but not proved strictly yet). In general, the quantity $\tau(Z, f_{\max})$ looks like

$$\tau = \left(\frac{z}{\pi}\right)^{\frac{d-1}{2}} \frac{\mathrm{e}^{-z}}{2\pi} A(f_{\max}) \tag{7}$$

for the basic least-squares periodogram (1) and like

$$\tau = \frac{\gamma}{2\pi} A(f_{\max}) \times \begin{cases} \left(\frac{2z_1}{\pi N_{\mathcal{H}}}\right)^{\frac{d-1}{2}} \left(1 - \frac{2z_1}{N_{\mathcal{H}}}\right)^{\frac{N_{\mathcal{K}}-1}{2}}, \\ \left(\frac{2z_2}{\pi N_{\mathcal{K}}}\right)^{\frac{d-1}{2}} \left(1 + \frac{2z_2}{N_{\mathcal{K}}}\right)^{-\frac{N_{\mathcal{H}}}{2}+1}, \\ \left(\frac{2}{\pi}\sinh\frac{z_3}{N_{\mathcal{K}}}\right)^{\frac{d-1}{2}} \mathrm{e}^{-z_3\left(1+\frac{d-3}{2N_{\mathcal{K}}}\right)} \end{cases} \tag{8}$$

for the modified periodograms (2). Here, the coefficient $\gamma = \Gamma(\frac{N_{\mathcal{H}}}{2})/\,\Gamma(\frac{N_{\mathcal{K}}+1}{2})$. Note that the asymptotic $(2/N_{\mathcal{H},\mathcal{K}})^{(d-1)/2}\gamma \to 1$ holds true for $N \to \infty$. The factor $A(f_{\max})$ depends on the bases $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}_{\mathcal{H}}$, on the time series sampling, and on the weighting pattern. Unfortunately, the general form of $A(f_{\max})$, obtained in Appendix B, is not simple. For now, let us restrict ourselves to the L–S periodograms and neglect the aliasing effects. In the next section we will show that such an approximation for $A(f_{\max})$ works well even for strong aliasing. Of course, it is perfectly possible to calculate $A(f_{\max})$ numerically from the formulae given in the Appendix, such calculations are still much less computationally expensive than the Monte Carlo simulation of $P_{\max}(Z, f_{\max})$. The practicalityof the expressions (5, 6) will be explored numerically in the next section.

To derive $A(f_{\max})$ from the formula (B7), we calculate the eigenvalues of the matrix $\mathbf{M}$, which is defined by the group of equalities (B4). To do this, we have to concretize the functions $\boldsymbol{\varphi}\,(t, f)$. For the usual L–S periodogram, the harmonic base

$$\boldsymbol{\varphi}(t,\, f) = \{\cos\omega t,\, \sin\omega t\} \quad (\omega = 2\pi f) \tag{9}$$

produces the matrices

$$\begin{aligned} \mathbf{Q} &= \frac{1}{2}\begin{pmatrix} 1+\overline{\cos 2\omega t} & \overline{\sin 2\omega t} \\ \overline{\sin 2\omega t} & 1-\overline{\cos 2\omega t} \end{pmatrix}, \\ \mathbf{S} &= \pi\begin{pmatrix} -\overline{t\sin 2\omega t} & \bar{t}+\overline{t\cos 2\omega t} \\ -\bar{t}+\overline{t\cos 2\omega t} & \overline{t\sin 2\omega t} \end{pmatrix}, \\ \mathbf{R} &= 2\pi^2\begin{pmatrix} \overline{t^2}-\overline{t^2\cos 2\omega t} & -\overline{t^2\sin 2\omega t} \\ -\overline{t^2\sin 2\omega t} & \overline{t^2}+\overline{t^2\cos 2\omega t} \end{pmatrix}, \\ \mathbf{M} &= \mathbf{Q}^{-1}(\mathbf{R} - \mathbf{S}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{S}). \end{aligned} \tag{10}$$

If we consider the alias-free case, the terms in (10) containing sine and cosine functions of frequencies $2f \leqslant 2f_{\max}$ are averaged out. Under this approximation, $\mathbf{M} \approx 4\pi^2\mathbb{D}t\,\mathbf{I}$, where $\mathbb{D}t = \overline{t^2} - \bar{t}^2$ is the weighted variance of the observational epochs. Then both eigenvalues required are equal to the constant $4\pi^2\mathbb{D}t$ and $A(f_{\max}) \approx 2\pi^{3/2}\,W$, where $W = f_{\max}T_{\mathrm{eff}}$ is a rescaled frequency bandwidth and $T_{\mathrm{eff}} = \sqrt{4\pi\mathbb{D}t}$ is the effective time series length. If t_i are spanned uniformly and all σ_i are equal, then T_{eff} almost coincides with the actual time series span. Table 1 contains the alias-free approximations of $\tau(Z, f_{\max})$ for all L–S periodograms considered. These expressions and the ones in (5, 6) can be used to obtain the corresponding alias-free approximation of $P_{\max}(Z, f_{\max})$ and its Davies bound. We routinely deal with large values of z and W. In this case, either the factor $P_{\mathrm{single}}(Z)$ in (6) or the term $\mathrm{FAP}_{\mathrm{single}}(Z)$ in (5) can be safely neglected. For instance, for the usual L–S periodogram,

$$P_{\max}(z,\, f_{\max}) \approx (1-\mathrm{e}^{-z})\,\mathrm{e}^{-W\mathrm{e}^{-z}\sqrt{z}} \approx \mathrm{e}^{-W\mathrm{e}^{-z}\sqrt{z}}. \tag{11}$$

Such alias-free approximations are valid if $f_{\max}$ is well resolved ($W \gtrsim 1$) and if the spectral leakage is low. Only the latter assumption is practically significant. If one considers strong spectral leakage, the approximate inequality

$$\mathrm{FAP}_{\max}(z,\, f_{\max}) \lesssim \mathrm{e}^{-z} + W\mathrm{e}^{-z}\sqrt{z} \approx W\mathrm{e}^{-z}\sqrt{z} \tag{12}$$

holds true for the basic L–S periodogram. The relations (11, 12) are equally valid if the base model is not empty but includes a low-order polynomial drift and/or several harmonics of fixed frequencies that can be considered as independent of any frequency within the range being scanned.

For large N, every modified periodogram $z_{1,2,3}$ obeys approximately the same extreme value distribution as the basic one. However, this convergence is not uniform in Z. It is easy to derive from (7, 8) that, for the periodograms $z_{1,2}(f)$, an extra condition $Z \ll \sqrt{N}$ must be satisfied to keep the relative errors of the FAP low. This condition is rarely satisfied in practice. For the periodogram z_3, a corresponding condition $Z \ll N$ is mild and is often satisfied in practical applications. This fact means that we need to consider the third modified periodogram more closely. The log-likelihood function of our Gaussian observations is given by

$$\ln\mathcal{L} = -\chi^2/2 - \sum_{i=1}^{N}\ln\sigma_i + \mathrm{const}. \tag{13}$$

As we adopted $\sigma_i = \kappa\,\sigma_{\mathrm{meas},i}$, this expression can be rewritten as $\ln\mathcal{L} = -\tilde{\chi}^2/(2\kappa^2) - N\ln\kappa + \mathrm{const}$, where $\tilde{\chi}^2$ does not depend on κ. Maximizing $\ln\mathcal{L}$ by κ under the hypotheses $\mathcal{K}$ and $\mathcal{H}$ yields that the logarithm of the ratio of the corresponding likelihood maxima is equal to $(N_{\mathcal{K}}/N)z_3$.

4 NUMERICAL SIMULATIONS

We now test the analytic results introduced above. For this purpose, we will use simulations of time series of N quasi-random data points imitating the white Gaussian noise. The temporal moments t_i cover a segment of length T. The uncertainties σ_i are equal to each other unless otherwise stated. For every simulation discussed below, no fewer than 10^5 Monte Carlo trials were generated (t_i and σ_i were, of course, fixed during every such simulation). This should provide accuracies of simulated FAPs of about 1 per cent for $\mathrm{FAP} = 0.1$ and of about 10 per cent for $\mathrm{FAP} = 10^{-3}$. The simulated tail of $\mathrm{FAP} < 10^{-3}$ often showed unstable deviations comparable with the FAP.

If the time series consists of a large number of equally spaced observations, any aliasing should be negligible. Indeed, in such a case the Davies bound (12) appears very sharp (for $\mathrm{FAP} < 0.1$) and the analytic approximation (11) perfectly follows the simulated distribution (Figs 1 and 2). However, even time series do not allow frequencies higher than the Nyquist frequency $f_{\mathrm{Ny}} = (N-1)/(2T)$ to be searched. An uneven time series allows much higher frequencies to be accessed. However, this access cannot be entirely 'free of charge'. Within a wide frequency range ($f_{\max} \gtrsim f_{\mathrm{Ny}}$), an essential aliasing is normally present purely as a result of random fluctuations of observational moments, even if there is no physical reason for their gapping. Thus we may expect that for $W \gtrsim N$ a significant 'natural' aliasing should take place. According to the numerical results shown in Fig. 1, the alias-free approximation indeed becomes significantly less precise when N decreases, but for large FAP only (larger than a few per cent, say). Even for $W > 100\,N$ the loss of precision remains moderate for important practical values of the FAP. We can use (11) for practical calculations even if W is 10 times larger than N (or even larger, depending on the desired precision).

When a 'physical' spectral leakage is large, the quality of the alias-free approximation depends on the frequency range too. If $f_{\max}$ does not exceed the Nyquist frequency of periodic breaking

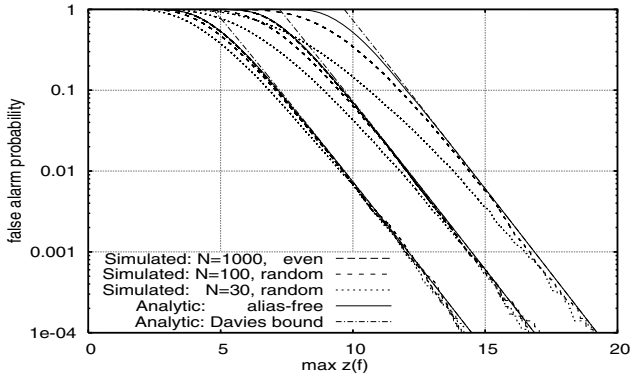

Figure 1. Simulated versus analytic FAP for the Lomb–Scargle periodogram with no forced data gapping. Simulations for 1000 evenly, and 100 and 30 randomly spaced observations, 10^5 Monte Carlo trials, and $f_{\max}\,T = 50, 500, 5000$ (bunches from left to right). In the even cases, the simulated curves almost coincide with their alias-free approximations. As the Nyquist frequency is exceeded, there is no curve for the even case with $f_{\max}\,T = 5000$. All graphs of analytic expressions are plotted for $T_{\rm eff} = T$ (this equality holds true within a few per cent for the three time series used here).

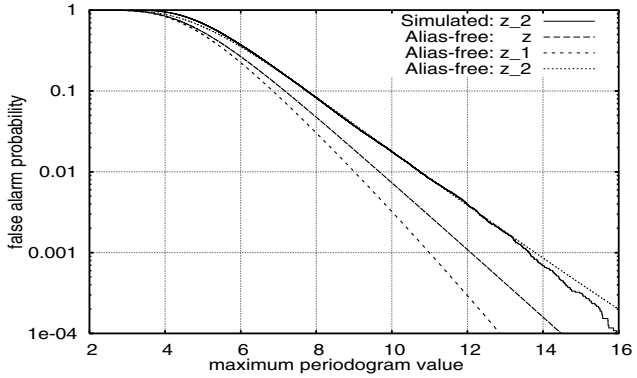

Figure 2. Simulated versus analytic FAP for the modification z_2 of the Lomb–Scargle periodogram: 100 evenly spaced observations, 10^5 Monte Carlo trials, $f_{\max}\,T = 50$. For comparison, the theoretical distribution curves for the periodograms z and z_1 are also plotted (the curve for z_3 almost coincides with that for z and is not shown).

of observations, the interval $[0, f_{\max}]$ is free from aliasing and we can use (11) without significant loss of precision. If the frequency range increases, the model (11) deviates from the real distribution and overestimates the FAP. Such a simulation is shown in Fig. 3. In this example, the frequency of periodic data breaks corresponds to $f_{\max}\,T = 9$ and the respective Nyquist frequency corresponds to half of this value ($f_{\max}T = 4.5$, just before the value $f_{\max}T = 5$ corresponding to the first simulation curve in Fig. 3).

Although errors of the alias-free model may become practically significant for some extremal situations, they are not very large and (more importantly) are not fatal. The significance of a candidate periodicity is underestimated. This underestimation does not favour false alarms. The aliasing may decrease the detectability of low-amplitude signals (if numerical simulations are not used). In this case, the error of the threshold level (i.e. the critical level z_*, corresponding to a given FAP_*) is more important. Examination of Fig. 3 shows that the relative shift $\Delta\, z_*/z_*$ does not exceed 10 per cent for $\mathrm{FAP} < 0.1$. As the amplitude of the corresponding signal scales as $\sqrt{z}$, this translates into only a 5 per cent relative error of the amplitude threshold. Recall that this 5 per cent offset corresponds to a very strong aliasing. Such spectral leakage takes place, for instance, for a sequence of observations that are made over 10 d with only 4.8 night hours in a day, or over 10 yr with only 2.4 observational months in a year.

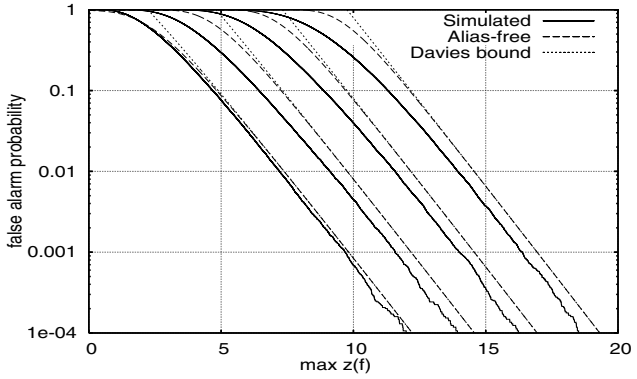

Figure 3. Simulated versus analytic FAP for the Lomb–Scargle periodogram with forced data gapping. $N = 100$ observations were clumped in 10 equal groups, and each group spans (randomly) only a fifth of its natural duration. About 1.7×10^5 Monte Carlo trials were used, $f_{\max}\,T = 5$, 50, 500, 5000 (from left to right).

Note that the multiple-trial formula (4) can work well only in restricted regions. When constructed from a short Monte Carlo simulation, it can fit the centre of the distribution well (i.e. large FAPs), but fails to fit low FAPs. This is the case even for negligible aliasing. The spectral leakage strongly perturbs the distribution centre but only weakly affects its high-significance tail. Hence, any multiple-trial models constructed from short Monte Carlo simulations cannot be extrapolated to the most important region of low FAPs. Such extrapolation overestimates the statistical significance of candidate periodicities, and favours false alarms.

The last pair of Monte Carlo simulations in this paper deals with real astronomical time series. Epochs and standard errors of 153 and 35 radial velocity measurements of the stars 51 Pegasi and 70 Virginis obtained with the ELODIE spectrograph (Naef et al. 2004) are used.[3] These time series are not even. For the star 51 Peg, the effective time series length $T_{\rm eff} \approx 9.0$ yr is close to the actual one, $T \approx 9.2$ yr, but the spectral window (Fig. 4) shows several high peaks, indicating periodic gapping of observations. For the star 70 Vir, the time series has $T_{\rm eff} \approx 8.5$ yr, $T \approx 7.2$ yr, and possesses a more 'noisy' spectral window (Fig. 5), indicating significant natural aliasing. In the first case, the simulated extreme value distributions for the L–S periodogram do not show large deviations from alias-free models (relative error $\Delta(\mathrm{FAP})/\mathrm{FAP} \lesssim 30$ per cent and $\Delta z_*/z_* \lesssim 5$ per cent for $\mathrm{FAP} < 0.1$). In the second case, the simulated FAP may be half of its alias-free approximation, but this may still be tolerated because $\Delta z_*/z_* \lesssim 10$ per cent (again for $\mathrm{FAP} < 0.1$). Note that both time series possess a strong leakage with a 1-d period. Such gapping affects extreme value distributions for $P_{\min} = 1/f_{\max} \leqslant 2$ days only. In the case of 51 Peg, this aliasing could introduce a significant error in FAP for unrealistic frequency ranges (for $P_{\min} = 1/f_{\max} \lesssim 0.1$ d, say). In the case of 70 Vir, the respective deviation is exacerbated by the low number of observations, which leads to rather large errors of FAP even for $P_{\min} = 1$ d. Note

[3] Note that stars 51 Peg and 70 Vir both have a planetary companion (Mayor & Queloz 1995; Marcy & Butler 1996).

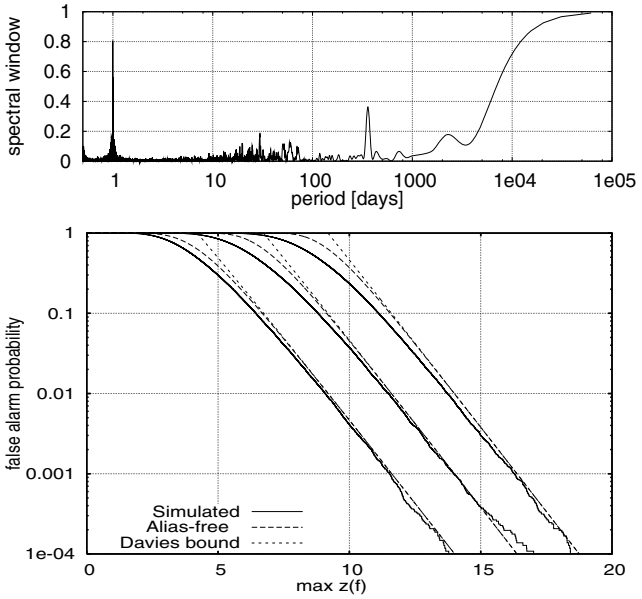

Figure 4. Top: spectral window function for ELODIE radial velocities of 51 Peg. Bottom: simulated and analytic FAPs for this time series for 10^5 Monte Carlo trials, $P_{\min} = 1/f_{\max} = 100$, 10, 1 d (from left to right).

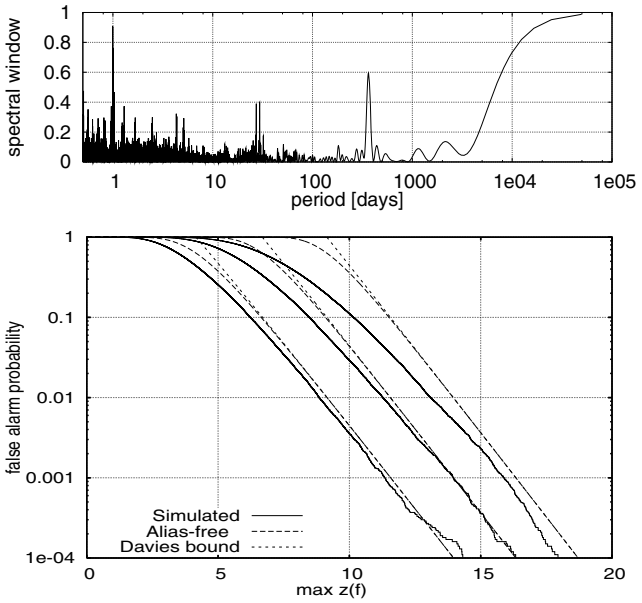

Figure 5. As Fig. 4, but for the star 70 Vir.

also that in both cases the errors of the alias-free approximations decrease significantly when FAP drops to values 10^{-3}–10^{-2}.

Finally, we need to consider the quality of the alias-free approximation for the factor $A(f_{\max})$. Fig. 6 shows a graph of the ratio $A(f_{\max})/(T f_{\max})$ along with graphs of its alias-free approximation and upper Carlson bound (see Appendix B). The observations were spanned in the same way as for Fig. 3. The spectral leakage appears only in small splashes near the Nyquist frequency of the periodic data breaks and near its overtones (i.e. at $f_{\max} T = 4.5, 9.0, 13.5$). For $W > 3$, the function $A(f_{\max})$ is well approximated by the alias-free model regardless of the strong aliasing.

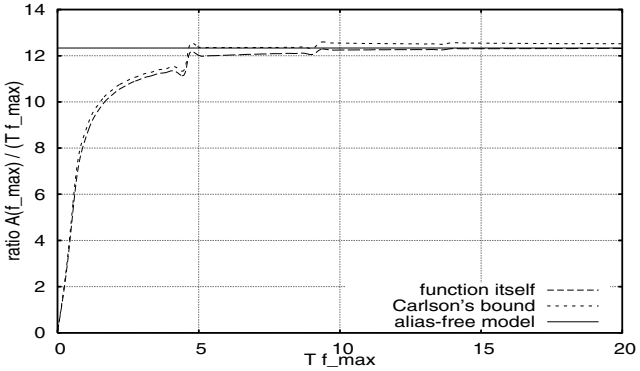

Figure 6. The factor $A(f_{\max})$ and its approximations.

5 CONCLUSIONS

The problem of estimating the statistical significance of periodogram peaks is discussed in this paper. Results published in the field of extreme values of random processes are adapted for and extended to the periodogram analysis of astronomical time series. For the Lomb–Scargle periodogram and its modifications, the corresponding extreme value distributions are given by closed formulae ready for use. If the spectral leakage cannot be neglected, similar expressions provide upper limits to the FAP (or lower limits to the significance).

It is established numerically that the region of validity of these approximations is large and has no sharp boundaries. Even if the aliasing is very strong, the error of the analytic estimation of the FAP does not favour false alarms and thus is not fatal. For strong aliases, use of this analytic approximation slightly decreases the sensitivity to low-amplitude signals. However, the corresponding increase of amplitude thresholds should not exceed several per cent in the worst practical cases (like a strong aliasing enforced by lack of observations).

These results may be very useful in a wide variety of astronomical applications. They will be useful especially for systematic surveys that deal with large amounts of data consisting of many separate time series. Indeed, it would be very difficult or even impossible to perform Monte Carlo simulation for each such time series. By contrast, it is easy to use the simple analytic formulae (11, 12) or their analogues for the modified L–S periodograms. This will eliminate the need for Monte Carlo simulations in cases for which the observed periodogram peak exceeds the adopted threshold and in the opposite cases for which this peak is lower than this threshold by more than, say, 10 per cent. The rare intermediate cases are easy to study by means of Monte Carlo simulations. It is also admissible not to use numerical simulations at all, especially for large data sets ($N \gtrsim 100$). In this case, the number of undetected low-amplitude periodicities may be increased by a negligible quantity.

ACKNOWLEDGMENTS

I would thank Drs V. V. Orlov, K. V. Kholshevnikov, L. P. Ossipkov and the anonymous referee for a critical reading of this paper, fruitful suggestions and linguistic corrections. This work is supported by the Russian Foundation for Basic Research (grants 05-02-17408, 06-02-16795) and by the President Grant NS-4929.2006.2 for the state support of leading scientific schools.

REFERENCES

Azaïs J.-M., Wschebor M., 2002, in Sidoravicius V., ed., Progress in Probability, Vol. 51, In and Out of Equilibrium: Probability with a Physics Flavor. Birkhäuser, Boston, p. 321

Cumming A., 2004, MNRAS, 354, 1165

Cumming A., Marcy G. W., Butler R. P., 1999, ApJ, 526, 890

Davies R. B., 1977, Biometrika, 64, 247

Davies R. B., 1987, Biometrika, 74, 33

Davies R. B., 2002, Biometrika, 89, 484

Ferraz-Mello S., 1981, AJ, 86, 619

Horne J. H., Baliunas S. L., 1986, ApJ, 302, 757

Kratz M. F., 2006, Probability Surveys, 3, 230

Lehman E. L., 1959, Testing Statistical Hypotheses. John Wiley and Sons, New York

Lomb N. R., 1976, Ap&SS, 39, 447

Marcy G. W., Butler R. P., 1996, ApJ, 464, L147

Mayor M., Queloz D., 1995, Nat, 378, 355

Naef D., Mayor M., Beuzit J. L., Perrier C., Queloz D., Sivan J. P., Udry S., 2004, A&A, 414, 351

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

Schwarzenberg-Czerny A., 1998a, MNRAS, 301, 831

Schwarzenberg-Czerny A., 1998b, Baltic Astron., 7, 43

Tee G., 2005, NZ J. Math., 34, 165

APPENDIX A: NOTATION

We introduce the following averaging operations:

$$\langle\phi(t)\rangle = \sum_{i=1}^{N} \phi(t_i)/\sigma_i^2, \qquad \overline{\phi(t)} = \langle\phi(t)\rangle/\langle 1\rangle,$$

with σ_i^2 being the error variance at the observational epoch t_i. The function $\phi(t)$ can be defined at the set of t_i only; that is, it may be a discrete sequence. The quantity $\langle\phi_1(t)\,\phi_2(t)\rangle$ can be treated as a scalar product in Hilbert space (Schwarzenberg-Czerny 1998a).

All vectors are assumed to be column ones by default. The notation $\{x_1, x_2, \ldots\}$ corresponds to a column vector formed by the quantities inside the braces. Similarly, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$ is a vector constituted by elements of the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$.

$\mathbf{I}$ is the identity matrix.

$*^{\mathrm{T}}$ denotes the transpose of a matrix or a vector.

If $\boldsymbol{x}$, $\boldsymbol{y}$ are vectors then $\boldsymbol{x} \otimes \boldsymbol{y} := \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ is a matrix constituted by the pairwise products $x_i y_j$.

$p(x_1, x_2, \ldots)$ is the joint probability density of the random variables x_1, x_2, ..., and $p(x_1 = a_1, x_2 = a_2, \ldots)$ is the same joint probability density, calculated at the point $(a_1, a_2, \ldots)$.

APPENDIX B: RICE METHOD AND PERIODOGRAMS

In the so-called 'Rice method', one considers an integer random variable $N^+(Z_0, f_{\max})$, the number of up-crossings of a given level Z_0 by the random process $Z(f)$ within $[0, f_{\max}]$. The distribution function of the maximum $Z_{\max} = \max_{[0,\,f_{\max}]} Z(f)$ can be represented by the expansion

$$\Pr\{Z_{\max} \leqslant Z_0\} = \Pr\{Z(0) \leqslant Z_0\} \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \nu_j, \tag{B1}$$

where $\nu_0 = 1$ and ν_j are the conditional factorial momenta of $N^+(Z_0, f_{\max})$ under the condition that $Z(0) \leqslant Z_0$. Let $\tilde{\nu}_j$ be the unconditional factorial momenta of N^+. The quantities ν_j and $\tilde{\nu}_j$ are explicitly expressed in terms of the so-called 'Rice formulae'. For instance,

$$\tilde{\nu}_1(Z, f_{\max}) = \int_0^{f_{\max}} \mathrm{d}f \int_0^{\infty} Z' p(Z, Z')\,\mathrm{d}Z', \tag{B2}$$

$$\tilde{\nu}_j(Z, f_{\max}) = \int_{[0,\,f_{\max}]^j} \mathrm{d}f_1 \ldots \mathrm{d}f_j \int_{[0,\infty)^j} Z'_1 \ldots Z'_j \times p(Z_1 = Z, Z'_1; \ldots; Z_j = Z, Z'_j)\,\mathrm{d}Z'_1 \ldots \mathrm{d}Z'_j, \tag{B3}$$

where $p(Z, Z')$ is the joint probability density of Z, and $Z' = \mathrm{d}Z/\mathrm{d}f$, both taken at the same frequency f, and $p(Z_1, Z'_1; \ldots; Z_j, Z'_j)$ is the joint probability density of the pairs $Z_i = Z(f_i)$, $Z'_i = Z'(f_i)$. For details on the Rice method and further references, see the paper by Azaïs & Wschebor (2002).

The expected number of up-crossings plays an important role in what follows. For the sake of convenience, we introduce the synonymous notation $\tau \equiv \tilde{\nu}_1$. Exact analytic expressions for $\tau(Z, f_{\max})$ for the periodograms $z(f)$ and $z_{1,2}(f)$ can be derived from results of Davies (1977, 1987, 2002). Davies dealt with the case for which the weights of measurements are equal to each other; however, his results can be directly extended to the case of unequal weights. The quantity τ not only provides the upper bound (5) on the FAP, but also is expected to yield its asymptotic representation for large z (low FAP) levels. Unfortunately, the asymptotic character of the Davies bound was strictly proved only for restricted families of random processes, such as stationary Gaussian and stationary χ^2 ones. Nevertheless, this asymptotic seems to be non-specific to the distribution of the process values and to the strict stationariness (see references and discussion in the cited works by Davies and in the review by Kratz 2006). Hence, we can expect the asymptotic character of (5) for all of our periodograms. Note that the periodogram $2z(f)$ can be treated as a χ^2 random process, $2z_2/d$ as an F process, and $2z_1/N_{\mathcal{H}}$ as a beta process, according to Davies (2002).

The high-order Rice formulae are significantly more complicated than the first-order one. We will not compute here the high-order Rice terms for our periodograms in the general case. However, the calculations are significantly simplified if the long-distance correlations of the periodogram can be neglected (equivalently, the aliasing is negligible). Indeed, under the approximation stated, the density $p(Z_1, Z'_1; \ldots; Z_j, Z'_j)$ can be factorized as $p(Z_1, Z'_1) \ldots p(Z_j, Z'_j)$ for all frequencies except in the narrow vicinities of the diagonals $f_i = f_j$. This property results in the fact that, if $f_{\max}$ is large enough to be well resolved by the periodogram, the relations $\nu_j \approx \tilde{\nu}_j \approx \tau^j$ hold true. The extreme value distribution of $Z(f)$ is then given by (6). An alternative way to obtain the latter expression is to assume a Poisson distribution for N^+ (Kratz 2006).

The factor $A(f_{\max})$ in equalities (7, 8) determines the dependence on the frequency range (the so-called bandwidth penalty). Before considering it, let us denote $\boldsymbol{\varphi}'_f = \partial\boldsymbol{\varphi}/\partial f$ and define the matrices

$$\begin{aligned}
\mathbf{Q} = \overline{\boldsymbol{\varphi} \otimes \boldsymbol{\varphi}}, \quad \mathbf{S} = \overline{\boldsymbol{\varphi} \otimes \boldsymbol{\varphi}'_f}, \quad \mathbf{R} = \overline{\boldsymbol{\varphi}'_f \otimes \boldsymbol{\varphi}'_f}, \\
\mathbf{Q}_{\mathcal{H}} = \overline{\boldsymbol{\varphi}_{\mathcal{H}} \otimes \boldsymbol{\varphi}}, \qquad \mathbf{S}_{\mathcal{H}} = \overline{\boldsymbol{\varphi}_{\mathcal{H}} \otimes \boldsymbol{\varphi}'_f}, \\
\mathbf{Q}_{\mathcal{H},\mathcal{H}} = \overline{\boldsymbol{\varphi}_{\mathcal{H}} \otimes \boldsymbol{\varphi}_{\mathcal{H}}}, \\
\tilde{\mathbf{Q}} = \mathbf{Q} - \mathbf{Q}_{\mathcal{H}}^{\mathrm{T}} \mathbf{Q}_{\mathcal{H},\mathcal{H}}^{-1} \mathbf{Q}_{\mathcal{H}}, \qquad \tilde{\mathbf{S}} = \mathbf{S} - \mathbf{Q}_{\mathcal{H}}^{\mathrm{T}} \mathbf{Q}_{\mathcal{H},\mathcal{H}}^{-1} \mathbf{S}_{\mathcal{H}}, \\
\tilde{\mathbf{R}} = \mathbf{R} - \mathbf{S}_{\mathcal{H}}^{\mathrm{T}} \mathbf{Q}_{\mathcal{H},\mathcal{H}}^{-1} \mathbf{S}_{\mathcal{H}}, \qquad \mathbf{M} = \tilde{\mathbf{Q}}^{-1} (\tilde{\mathbf{R}} - \tilde{\mathbf{S}}^{\mathrm{T}} \tilde{\mathbf{Q}}^{-1} \tilde{\mathbf{S}}).
\end{aligned} \tag{B4}$$

In general, all these matrices, except for the matrix $\mathbf{Q}_{\mathcal{H},\mathcal{H}}$, depend on the frequency. Note the relations $\mathbf{S}_{\mathcal{H}} = \mathbf{Q}'_{\mathcal{H}}$ and $\mathbf{S}^{\mathrm{T}} + \mathbf{S} = \mathbf{Q}'$. The definitions (B4) look rather bulky, but they are significantly simplified under certain conditions. For example, if the base functions

$\boldsymbol{\varphi}$ for any f are orthogonal to the functions $\boldsymbol{\varphi}_{\mathcal{H}}$, then $\mathbf{Q}_{\mathcal{H}} = \mathbf{S}_{\mathcal{H}} = 0$ and the matrices in (B4) labelled with a tilde are equal to the corresponding matrices without a tilde. Moreover, if the base $\boldsymbol{\varphi}$ is orthonormal for any f, then $\mathbf{Q} = \mathbf{I}$, $\mathbf{S}^{\mathrm{T}} = -\mathbf{S}$ and $\mathbf{M} = \mathbf{R} + \mathbf{S}^2$. The last matrix $\mathbf{M}(f)$ is necessarily positive-definite and possesses the positive eigenvalues $\lambda_i(f)$. In fact, we need only these eigenvalues. They satisfy the characteristic equation $\det(\tilde{\mathbf{R}} - \tilde{\mathbf{S}}^{\mathrm{T}}\tilde{\mathbf{Q}}^{-1}\tilde{\mathbf{S}} - \lambda\tilde{\mathbf{Q}}) = 0$.

The factor $A(f_{\max})$ appears implicitly in the papers by Davies after integration of the mathematical expectation $\mathbb{E}|\boldsymbol{\eta}|$ by f, where the random vector $\boldsymbol{\eta}$ is Gaussian with zero mean and has statistically independent components with $\mathbb{D}\eta_i = \lambda_i(f)$. Davies (1987) gave some exact and approximate integral formulae for this expectation. I present here (in terms of the factor A) a number of new integral representations that may be useful in practice. The first two can be derived easily and are given by

$$A(f_{\max}) = \int_0^{f_{\max}} \mathrm{d}f \oint_{\mathcal{S}_d} m(\boldsymbol{n})\,\mathrm{d}\Omega, \quad m^2(\boldsymbol{n}) = \boldsymbol{n}^{\mathrm{T}}\mathbf{M}\boldsymbol{n},$$

$$A(f_{\max}) = \int_0^{f_{\max}} \mathrm{d}f \int_{\boldsymbol{x}^2<1} \sqrt{\boldsymbol{x}^{\mathrm{T}}\mathbf{M}\boldsymbol{x}}\,\frac{\mathrm{d}\boldsymbol{x}}{|\boldsymbol{x}|^d}, \tag{B5}$$

where $\mathrm{d}\Omega$ denotes an infinitesimal solid angle in $\mathbb{R}^d$, directed by the unit-length integration vector $\boldsymbol{n}$. The integration in the first formula is performed over all possible directions within the whole space solid angle $S_d = 2\pi^{d/2}/\Gamma(d/2)$. Note that the matrix $\mathbf{M}$ can be diagonalized by means of a solid-body rotation of $\boldsymbol{n}$, so that $m^2 = \sum \lambda_i n_i^2$ and the function $A(f_{\max})$ is determined by the eigenvalues λ_i only.

The inner integrals in (B5) are equal to each other. They may be expressed in terms of the (hyper)area of an ellipsoidal (hyper)surface in d dimensions having semi-axes $q_i = \lambda_i^{-1/2}$. Indeed, changing the integration variable in the inner integral in the second of equations (B5) as $\boldsymbol{x} = \mathbf{M}^{1/2}\tilde{\boldsymbol{x}}$, then $\tilde{\boldsymbol{x}} = \tilde{x}\tilde{\boldsymbol{n}}$ ($\tilde{\boldsymbol{n}}^2 = 1$) and integrating by $\tilde{x}$ we obtain

$$\oint_{\mathcal{S}_d} m(\boldsymbol{n})\mathrm{d}\Omega = \Pi_d\sqrt{\det\mathbf{M}}, \quad \Pi_d = \oint_{\mathcal{S}_d} \frac{\sqrt{\tilde{\boldsymbol{n}}^{\mathrm{T}}\mathbf{M}^2\tilde{\boldsymbol{n}}}}{\left(\tilde{\boldsymbol{n}}^{\mathrm{T}}\mathbf{M}\tilde{\boldsymbol{n}}\right)^{\frac{d+1}{2}}}\,\mathrm{d}\tilde{\Omega}. \tag{B6}$$

It can be directly checked that the integrand in the last expression represents an infinitesimal (within $\mathrm{d}\tilde{\Omega}$) area element on the surface $\tilde{\boldsymbol{x}}^{\mathrm{T}}\mathbf{M}\tilde{\boldsymbol{x}} = \tilde{x}^2\tilde{\boldsymbol{n}}^{\mathrm{T}}\mathbf{M}\tilde{\boldsymbol{n}} = 1$, and that Π_d is equal to its total area. It is not hard to show that $\Pi_1 = 2$. The circumference of an ellipse, Π_2, and the usual surface area of an ellipsoid, Π_3, can be expressed by means of elliptic integrals (complete and incomplete, respectively). For $d \geqslant 4$ an Abelian integral can be used to compute Π_d (Tee 2005). There are useful inequalities for Π_d; for example, Carlson's inequality bounds the inner integrals in (B5) by the quantity $S_d\sqrt{(\lambda_1 + \ldots + \lambda_d)/d} = S_d\sqrt{\mathrm{Tr}\mathbf{M}/d}$. Finally,

$$A(f_{\max}) = \int_0^{f_{\max}} \frac{\Pi_d(q_1 \ldots q_d)}{q_1 \ldots q_d}\,\mathrm{d}f \leqslant S_d \int_0^{f_{\max}} \sqrt{\frac{\mathrm{Tr}\mathbf{M}}{d}}\,\mathrm{d}f. \tag{B7}$$

The latter inequality seems to be very sharp in practical situations (Fig. 6). Note also that if every $\lambda_i(f) \equiv \lambda$ then $A(f_{\max}) = S_d\, f_{\max}\sqrt{\lambda}$.

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