Mon. Not. R. Astron. Soc. **301,** 831–840 (1998)

The distribution of empirical periodograms: Lomb–Scargle and PDM spectra

A. Schwarzenberg-Czerny[1,2]
[1]*Astronomical Observatory of Adam Mickiewicz University, ul. Słoneczna 36, 60-286 Poznań, Poland*
[2]*Copernicus Astronomical Center, ul. Bartycka 18, 00-716 Warsaw, Poland*

Accepted 1998 August 7. Received 1998 May 14; in original form 1997 March 4

ABSTRACT

The theoretical probability distributions of periodograms are derived for the assumed variance of noise. In practice, however, the variance is estimated from data and hence it is a random variable itself. The *empirical periodograms*, i.e. the periodograms normalized using the estimated variance, therefore follow a distribution different from that predicted by theory. We demonstrate that in general many empirical periodograms follow the beta distribution. In particular, as an example we consider a Lomb & Scargle (L–S) modified power spectrum with an exponential theoretical distribution. We derive its easy-to-use analytical empirical distribution. We demonstrate that the difference between the tails of the empirical and theoretical distributions is large enough to have a profound effect on the statistical significance of signal detections. The difference persists despite generally good asymptotic convergence of the distributions near their centres. Hence we argue that even for well-behaved statistics (e.g. L–S) one has to use our new empirical beta distributions rather than the theoretical ones. Our conclusions are illustrated by a realistic example. In the example we demonstrate a significant difference between the theoretical and empirical distributions. Additionally, we provide an example of conversion between analysis of variance (AOV), power-spectrum, PDM and χ^2 periodograms.

Key words: methods: data analysis – methods: statistical – binaries: eclipsing – stars: oscillations – pulsars: general – X-rays: stars.

1 INTRODUCTION

Astronomical observations are affected by factors such as seasons and time allocation, which result in uneven sampling. Thus the vast body of statistical literature on the analysis of evenly sampled time series does not apply. Often, in the analysis of astronomical time series, one has to start from basic statistical principles. The statistical principles involved in period detection and fitting curves to experimental data are similar (e.g. Lomb 1976). The key elements of the relevant statistical procedure are

(i) the mathematical model $\xi_{\|}$ of the observations ξ;
(ii) the statistic Θ, i.e. a measure of the data and model consistency;
(iii) the probability distribution $P(\Theta)$ of the statistic Θ for the pure noise signal ξ, i.e. for the *null hypothesis* H_0; and
(iv) the hypothesis testing theory (e.g. Eadie et al. 1971; Lupton 1993).

The general aim of the present paper is to discuss in a uniform way the statistical properties of the general class of periodograms based on the linear orthogonal models (i) and the χ^2 statistics Θ (ii). The models possess a distinct advantage in that their distributions (iii) are known (e.g. Eadie et al. 1971; Bickel & Doksum 1977). Since the relevant material on the algebraic and statistical properties of these models is scattered in the literature, we provide a summary in Section 2 and Appendix A. The class of the L^2-norm statistics, often called the χ^2 statistics, is broad enough to contain most of the known periodograms and narrow enough to obtain their explicit probability distributions. Note that in general the distributions differ from the χ^2 distribution itself. In Section 3 we apply the hypothesis testing theory (iv) to the individual periodograms and discuss their properties.

Our specific aim is to discuss the difference between the distributions of the theoretical and empirical periodograms. The theoretical distributions are derived for a given distribution of the noise (i.e. for the null hypothesis H_0). In practice, one has to estimate the variance of the noise from the same data that are used for the periodogram itself. This interrelation is bound to affect the distribution. It is often argued that the empirical distributions converge to the theoretical distributions for sufficiently large samples. However, Schwarzenberg-Czerny (1989) and Davies (1990) have presented an example in which no such convergence is achieved at all. In Section 4 we consider in the detail an example of a Lomb (1976) & Scargle (1982) (L–S) periodogram. We derive

its empirical distribution and discuss its consequences for the significance of detections. Section 5 contains an example explaining how our ideas work when applied to realistic observations of a pulsating variable star.

The reader should keep in mind that additional effects *must* be taken into account in any practical period analysis: the effect of the correlation in the observations on their effective number (e.g. Schwarzenberg-Czerny 1991) and the effective number of independent frequencies analysed (multiple trial or bandwidth penalty, e.g. Horne & Baliunas 1986). The effects constitute self-contained problems and hence are best discussed elsewhere. We consider the opinion that all statistical problems related to the periodograms can be solved by Monte Carlo simulations to be over-optimistic. The simulations have problems of their own, related chiefly to the untested effects of the discrete random number generators and periodogram algorithms on the tails of the continuous distributions. Experiments often demonstrate difficulties in the reproduction of theoretical single-trial distributions by simulations. Hence the analytical single-trial probabilities discussed here are essential for the verification of Monte Carlo simulations.

2 STATISTICAL PRINCIPLES

Our considerations in this paper concern the statistical evaluation of the fit of a time series by a series of orthogonal functions (Appendix A). The classic statistic measuring the consistency of the observations and model is based on the L^2 vector norm induced by the scalar product.

2.1 The L^2 vector norms

The real periodic functions have particularly simple representations after their mapping into the complex vector space $\mathcal{H}$ of the functions on the unit circle $C : |z| = 1$. For a given trial frequency ω a time series of n observations sampled unevenly at times t_k, where $k = 1, \cdots, n$ corresponds to a function ξ defined on a discrete set of points $z_k = \mathrm{e}^{\omega t_k}$. For a given set of t_k the observed and modelled series ξ and $\xi_\parallel$ [point (i) in Section 1] correspond to the vectors of n components $\xi_k \equiv \xi(z_k)$ in $\mathcal{H}$. All possible values of the discrete model and observation time series define the n-dimensional subspace $\mathcal{H}_n$ of the space $\mathcal{H}$. In fact $\mathcal{H}$ constitutes the Hilbert space, with a scalar product defined by the Stjeltjes integral over C: $(\xi, \psi) = (1/2\pi) \oint_{C:|z|=1} \xi(z)\overline{\psi(z)}\mathrm{d}\mu(z)$. The integral depends on the weight function $\mu(z)$ and on the parameter ω, i.e. frequency. The natural weight function for the case of the discrete time series in $\mathcal{H}_n$ is the step function with discontinuities at the phases of the observations $z_k = \mathrm{e}^{\mathrm{i}\omega t_k}$. The scalar product then reduces to the weighted sum

$$< \xi, \psi > \equiv \xi^\dagger \circ \psi = \sum_{k=1}^{n} \mu_k \xi(z_k)\overline{\psi(z_k)}, \quad \text{where} \tag{1}$$

$$\mu_k = \mathrm{Var}^{-1}\{\xi_k\} \tag{2}$$

are weights of the individual observations and † denotes the matrix complex conjugate transposition (Hermitian) operator.

2.2 Fisher's lemma

In the discussion of the statistical properties in the rest of the paper we assume that the hypothesis H_0 holds, i.e. that all observations are pure Gaussian white noise. The covariance matrix of ξ in the observation coordinates (orthonormal basis E) is proportional to the identity matrix $E\{\xi \circ \xi^\dagger\} \equiv \mathrm{Cov}\{\xi\} = \sigma^2\mathbf{I}$. In changing to the model coordinates (orthonormal basis Φ), the vector ξ is multiplied by the unitary matrix $\Phi^\dagger$ (equation A8). The components $< \phi^{(l)}, \xi >$ of the vector $\Phi^\dagger \circ \xi$ in the model space are linear combinations of the observations, the independent normal random variables. Hence they are normal themselves. This is usually demonstrated by using the characteristic function of the normal distribution, i.e. the Fourier transform $\mathcal{F}$ of the distribution $f_X(\tau) \equiv E\{\exp \mathrm{i}\tau x\} = \exp(-\tau^2\sigma^2/2)$ and noting that for the independent random variables X and Y the characteristic function $f_{X+Y} = f_X f_Y$. Since, for the normal distribution, $f_X f_Y$ retains the form of the normal distribution f_X, and since the original distribution may be uniquely recovered from f_{X+Y} by $\mathcal{F}^{-1}$, the demonstration is complete (Eadie et al. 1971). In the new base the covariance matrix is again diagonal ($< \Phi^{(l)}, \xi >$ are uncorrelated):

$$E\{(\Phi^\dagger \circ \xi) \circ (\Phi^\dagger \circ \xi)\}^\dagger = \Phi^\dagger \sigma^2 \mathbf{I} \circ \Phi = \sigma^2 \mathbf{I}. \tag{3}$$

For the $N(0, \sigma)$ normal random variables the covariance matrix contains full information about their statistical properties. Thus the lack of correlation is equivalent to the independence of the $< \phi^{(l)}, \xi >$s in this special case. Hence, $||\xi_\parallel||^2 = \sum_{l=1}^{r} | < \phi^{(l)}, \xi > |^2$ and $||\xi_\perp||^2 = \sum_{l=r+1}^{n} | < \phi^{(l)}, \xi > |^2$ are independent since all their terms are independent. This conclusion constitutes Fisher's lemma (e.g. Fisz 1963). Fisher's lemma holds for H_0 and for the orthonormal base Φ. The sum

$$||\xi||^2 \equiv ||\xi_\parallel||^2 + ||\xi_\perp||^2 \tag{4}$$

is correlated with each of the terms, of course. The sum does not depend on the model, since the norm is invariant of the unitary transformations (equation A10). $||\xi||^2 = < \xi, \xi >$. As the L^2 norms are sums of the squared independent Gaussian variables $< \Phi^{(l)}, \xi >$, they all have the following χ^2 distributions:

Statistic	Distribution	Expectation
$\|\|\xi\|\|^2$	$\chi^2(d)$	d
$\|\|\xi_\parallel\|\|^2$	$\chi^2(d_\parallel)$	$d_\parallel$
$\|\|\xi_\perp\|\|^2$	$\chi^2(d_\perp)$	$d_\perp$

(5)

where $d = d_\parallel + d_\perp$ and $d = n$, $d_\parallel = r$ and $d_\perp = n - r$ are the respective number of degrees of freedom. For numerical reasons it is advisable to subtract the average value from the observations. Then $d = n - 1$, $d_\parallel = r - 1$ and $d_\perp = n - r$.

2.3 Empirical statistics

In practice, we often deal with the measurements ξ_k, for which standard deviation σ is not known a priori. Then by virtue of equations (2) and (3) the norms and scalar products suffer from σ^2 scalefactor indeterminacy. The indeterminacy propagates to the *empirical* χ^2 statistics generated by the norms $||\xi||^2$, $||\xi_\parallel||^2$ and $||\xi_\perp||^2$. The indeterminacy renders the norms unsuitable for direct statistical evaluation. To remove the indeterminacy, one takes the ratios of norms. Normalization by the mean variance $||\xi||^2$ estimated a posteriori from the data constitutes an example of such a ratio. These ratios of two random variables may not retain the distribution of the numerator. Hence our first conclusion is that *the normalized empirical statistics do not follow, in general, their theoretical distributions* computed for the known variances. Specifically, most of the empirically computed χ^2 statistics, i.e. the L^2 norms in astronomy and physics, *do not* follow the χ^2 distribution.

By virtue of Fisher's lemma, $||\xi_\parallel||^2$ and $||\xi_\perp||^2$ are independent; hence their ratio has a Fisher–Snedecor distribution. The ratios

$||\xi_\parallel||^2/||\xi||^2$ and $||\xi_\perp||^2/||\xi||^2$, where $||\xi||^2 = ||\xi_\parallel||^2 + ||\xi_\perp||^2$, have beta distributions

$$\begin{array}{lll} \text{Statistic} & \text{Distribution} & \text{Expectation} \\ \Theta_o \equiv \dfrac{d_\perp ||\xi_\parallel||^2}{d_\parallel ||\xi_\perp||^2} & F(d_\parallel, d_\perp) & \dfrac{d_\perp}{d_\perp - 2} \to 1 \\ \Theta_\parallel \equiv \dfrac{||\xi_\parallel||^2}{||\xi||^2} & I_{\Theta_\parallel}\left(\dfrac{d_\parallel}{2}, \dfrac{d_\perp}{2}\right) & 0 \le \dfrac{d_\parallel}{d} \le 1 \\ \Theta_\perp \equiv \dfrac{||\xi_\perp||^2}{||\xi||^2} & I_{\Theta_\perp}\left(\dfrac{d_\perp}{2}, \dfrac{d_\parallel}{2}\right) & 0 \le \dfrac{d_\perp}{d} \le 1 \end{array} \tag{6}$$

where $I_x(a, b)$ denotes the incomplete beta function (e.g. Bickel & Doksum 1977; Abramovitz & Stegun 1971). Note that, by design, the left-hand side of equation (A13) does not depend on ω, so all Θ statistics depend on ω via a single function, e.g. $||\xi_\parallel||^2$. Hence a one-to-one correspondence exists between their corresponding periodograms and their probability distributions (Abramovitz & Stegun 1971; Bickel & Doksum 1977 equation 1.3.6; Schwarzenberg-Czerny 1989 equation 12; Davies, 1990 equation 12; Schwarzenberg-Czerny 1997). In particular, equation (26.6.2) and (26.5.2) of Abramovitz & Stegun (1971) read in the present notation

$$1 - F(\Theta_o; d_\parallel, d_\perp) = I_{\Theta_\perp}\left(\frac{d_\perp}{2}, \frac{d_\parallel}{2}\right) = I_{\Theta_\parallel}\left(\frac{d_\parallel}{2}, \frac{d_\perp}{2}\right), \tag{7}$$

where

$$\Theta_\perp = 1 - \Theta_\parallel = \frac{d_\perp}{d_\perp + d_\parallel \Theta_o}. \tag{8}$$

Each of the Θ periodograms and the corresponding distributions may be converted to another by means of the change of variables corresponding to equation (8). Hence the statistical conclusions do not depend on which of the Θ statistics and corresponding distribution are used. *The consistency of the statistics and of the distribution* in use is of crucial importance for the correctness of the analysis, however.

Note that the old Fisher statistic (Fisz 1963; Eadie et al. 1971)

$$Z = \frac{1}{2} \ln \Theta_o \tag{9}$$

may be particularly suitable for plotting in periodograms, because its distribution is closer to the normal distribution than all Θ statistics, hence it is more intuitive to an observer:

$$E\{Z\} = (d_\parallel^{-1} - d_\perp^{-1})/2 - (d_\parallel^{-2} - d_\perp^{-2})/6 \to 0$$

and

$$V\{Z\} = (d_\parallel^{-1} + d_\perp^{-1})/2 + (d_\parallel^{-2} + d_\perp^{-2})/2 + (d_\parallel^{-3} + d_\perp^{-3})/3 > 0.$$

2.4 Corrections to probabilities

Any probabilities derived from the periodograms *must* be corrected for the correlation of the residuals and the effective number of frequencies searched. These corrections may alter the probabilities (and variances) by many orders of magnitude. The corrections apply in a similar way to all types of periodograms. However, since the corrections constitute a well-defined separate topic, we sketch the main ideas only briefly here, and for the details we refer readers to the literature quoted in the following two subsections.

2.4.1 *Correlation effects*

The standard least-squares error and variance estimators are based on the implicit assumption *that the residuals are white noise*, i.e. that their values for the consecutive measurements *are not correlated*. The violation of the assumption manifests itself, more often than not, by long sequences of residuals of the same sign. One cause of the correlation is the severe oversampling of slow intrinsic random variations, e.g. in the fast photometry of the flickering in cataclysmic binary stars. Another cause is poor matching of the observations by the model, e.g. in fitting a sinusoid to narrow-pulse (or narrow-eclipse) light curves. Power-spectrum analysis corresponds to the latter case (Lomb 1976). Let on average k consecutive residuals be correlated, or, equivalently, they change sign n/k times. Then statistically correct results are recovered by substitution of each k observations by their average value. In the corresponding number of degrees of freedom of the probability distributions, one should substitute the true number of observations n by the *effective number of observations* n/k (e.g. Schwarzenberg-Czerny 1991). This decreases the significance of any detection and increases the variance of the fitted parameters by a factor of k. Note that the effective number of observations used in the analysis depends both on the observation mode and on the statistical model (method) in use.

2.4.2 *Multiple frequencies*

The formulae discussed so far are concerned with the single-frequency (i.e. single-trial) probability. Since the realistic periodograms cover multiple frequencies, their statistical evaluation *requires significant correction* for multiple trials (also called the bandwidth penalty). Generally, values of the periodogram at different frequencies are not independent and the bandwidth penalty is no simple function of the number of computed frequencies. Two strategies can be used.

(i) Guess the number of independent trials m and compute the multiple-trial probability from the single-trial probability,

$$P_m(z) = P_1(z)^m. \tag{10}$$

(ii) Use Monte Carlo simulation of observations to compute the multiple-frequency histogram of the periodogram $P_m(z)$.

The Monte Carlo simulations rely on rare events of low probability, for which neither the accuracy of random number generators nor the accuracy of periodogram algorithms is well tested. Experiments often demonstrate difficulties in reproduction of the theoretical single-trial distributions by Monte Carlo simulations. Hence *the single-trial analytical probability distributions are indispensable* in any strategy for the bandwidth correction. In this paper we concentrate on the single-trial probabilities. The bandwidth penalty issue is not specific to the type of the periodogram, and the reader is encouraged to refer to van der Klis (1989) and Horne & Baliunas (1986) for more details.

3 APPLICATIONS

Phase folding and binning of observations into phase histograms corresponds, in fact, to the fitting of orthogonal step functions. The methods involving $r = 2$ and 3 smooth orthogonal functions were considered by Lomb (1976) and Ferraz-Mello (1981), respectively. Until recently, any attempts to extend their technique to larger orthogonal sets were frustrated by the inefficient Gramm–Schmidt algorithm, scaling with $\mathcal{O}(r^3)$. The Gramm–Schmidt orthogonalization is as laborious as the direct least-squares fitting of the non-orthogonal functions. This inefficiency was removed by the new efficient $\mathcal{O}(r^1)$ algorithm for the projection on to trigonometric orthogonal polynomials (Schwarzenberg-Czerny 1996). We shall

classify the periodograms according to the underlying orthogonal models. The following models are encountered in time series analysis (TSA): ordinary orthogonal harmonics in fast Fourier transforms (FFTs), orthogonal step functions, orthogonal combinations of ordinary harmonics and combinations of splines. Wavelets constitute a prospective example. The FFT even-sampling case falls outside the scope of the present paper (see e.g. Press et al. 1992).

3.1 Step function model

The methods relying on phase folding and binning data into histograms rely implicitly on least-squares fitting of step functions:

$$\Psi^l(\omega t) = \begin{cases} \frac{1}{n_l} & \text{if } l-1 \le r\frac{\omega t}{2\pi} \bmod r < l, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where n_l is the number of observations in the lth bin. Obviously, the step functions are orthogonal with respect to each other. The classical methods belonging here are those of Lafler & Kinman (1965), PDM and Whittaker & Robinson (W-R, 1926; Stellingwerf 1978). Except for PDM, these periodograms use the $\Theta_\parallel$ statistic, so they follow the beta distribution $I[(r-1)/2,(n-r)/2]$. Specifically, the Lafler–Kinman periodogram corresponds to narrow binning, $r = n/2$, with double coverage. The PDM periodogram uses the $\Theta_\perp$ statistic and follows the $I[(n-r)/2,(r-1)/2]$ distribution (Schwarzenberg-Czerny 1997). Originally it was claimed that these periodograms follow either χ^2 or F distributions. In fact, only the analysis of variance (AOV/ANOVA) periodograms using the Θ_o statistic follow the F distribution. In greater detail the ANOVA periodogram and its $F(r-1,n-r)$ distribution were investigated by Schwarzenberg-Czerny (1989) and Davies (1990). The computational advantage of the phase folding and binning is the fixed and rather small number of calculations per observation, with no regard for r. In that respect the phase folding may be compared with the methods of fitting the functions with a compact support, e.g. splines (Akerlof et al. 1994). Despite the non-orthogonality of the splines, their covariance matrix has a fixed number of non-vanishing subdiagonals so that the computational overhead in the least-squares solution does not grow with r, depending on the width of the footprint instead. The spline method, as well as the phase folding, is sensitive to the even-phase coverage or else the χ^2 distribution may be violated (equation 5). The methods discussed in the next section are largely insensitive to the phase coverage.

3.2 Fourier harmonics

The Fourier harmonics $e^{2\pi i l}$, $l = 0,\cdots,[r/2]$, are used for the FFT power spectrum for even sampling (e.g. Press et al. 1992). On the FFT grid they are orthogonal. The sine and cosine functions ($r = 2$) are used in calculation of the discrete power spectrum (DPS) for uneven sampling [a discrete Fourier transform (DFT), e.g. Deeming 1975]. However, for uneven sampling the harmonics are no longer orthogonal. Then the theory of Section 2 does not apply and the statistical properties of the DFT remain unknown. Note, that in this case equation (A5) does not provide the least-squares fit. See Foster (1995) for the counterexample demonstrating that not all power is detected by a DPS. Lomb (1976) and Scargle (1982) (L–S) demonstrated that a shift in phase suffices to orthogonalize the sine and cosine curves. They employed $\Theta_\parallel$ statistics for $r = 2$ and claimed an exponential distribution. We postpone any detailed discussion of the L–S statistics to Section 4. Ferraz-Mello (1981), Grison (1994) and Foster (1995) attempted to extend the L–S method by adding harmonics of order 0, 2, 3 and 4 to the set of model functions and obtaining their orthogonal combinations by carrying out Gramm–Schmidt orthogonalization. They employed the $\Theta_\parallel$ statistic for $r > 2$ parameters, claiming that its distribution is $\chi^2(r)$. Because of the necessity to perform the Gramm–Schmidt orthogonalization for each frequency, their procedure is computationally inefficient. As we have demonstrated in Section 2, the L–S and Forster periodograms in fact follow the beta distribution $I(r,n-r)$. Schwarzenberg-Czerny (1996) invented an efficient method for fitting data with a multiple harmonic Fourier series. He employed the ANOVA Θ_o statistic and identified its exact Fisher–Snedecor distribution $F(r-1,n-r)$. The method employs recurrence formulae for efficient calculation of the orthogonal combinations of the harmonics, so that the number of computations scales with $\mathcal{O}(r^1)$, with very little overheads, and hence it is ideally suited for the analysis of the non-sinusoidal oscillations.

In terms of statistical and computational efficiency, this may be the best method known for the detection of non-sinusoidal oscillations with a single frequency in unevenly sampled data. In application to the light curves from mass photometric surveys, the precomputed orthogonal vectors may be reused, with considerable computational saving. All the methods discussed in this section are insensitive to phase coverage. Poor coverage prevents light-curve restoration but does not prevent the methods from determining whether any systematic trends exist in adequately covered phases.

4 EXAMPLE: LOMB–SCARGLE PERIODOGRAM

Lomb (1976) has demonstrated in a particularly clear way the relation between the periodogram and the model fitting. He and Scargle (1982) both use $r = 2$ orthogonal functions $\sin(\omega t-\tau)$ and $\cos(\omega t-\tau)$, with the $\Theta_\perp$ and $\Theta_\parallel$ statistics, respectively. Hence their periodograms have the beta distributions $Q_1(x) = 1-P_1(x) = 1-I_x(2/2,n/2) = I_{1-x}(n/2,1)$ (Section 2.3). For the given arguments, the beta integral is elementary, hence

$$Q_1(z) \equiv 1-P_1(z) = \left(1-\frac{2z}{n}\right)^{\frac{n}{2}} \to e^{-z}, \tag{12}$$

where $z = nx/2$. Lomb and Scargle considered only the theoretical cases of fixed $||\xi||^2$ and random $||\xi||^2_\parallel$ (equation A13). Thus for $\Theta_\parallel = d||\xi||^2_\parallel/d_\parallel||\xi||^2$ they obtained the theoretical exponential distribution $\chi^2(2)/2$, corresponding to

$$Q_1(z) \equiv 1-P_1(z) = e^{-z}. \tag{13}$$

Equation (12) demonstrates that for the empirical $||\xi||^2$ estimated from the data, the distribution approaches the exponential one only asymptotically for $n \to \infty$. The convergence to the asymptotic distribution is fast near the centre of the distribution. However, in the tail used in the estimation of the detection significance this is generally not true, as amply illustrated in Fig. 1. In the figure we plot the $Q_n(z)$ from equation (13) against that computed from equation (12). Both probabilities were converted from the single-trial values Q_1 to the m-trial values Q_m using equation (10), where we have chosen the maximum possible value for m to be equal to the number of observations. Large differences between the beta distribution and its exponential approximation are evident for the number of observations $m \le 100$ and a detection significance in excess of 3σ. In this way we demonstrate that the exponential distribution does not approximate the probability distribution of the L–S spectrum adequately for practical purposes.

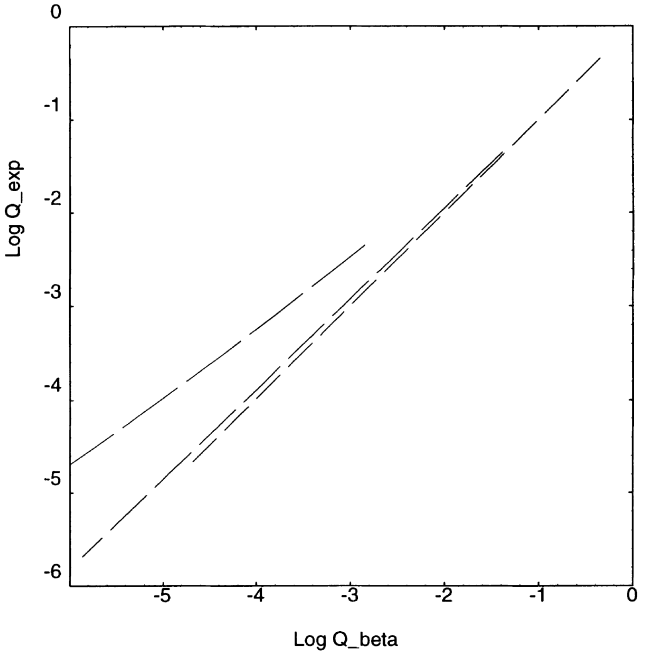

Figure 1. The asymptotic exponential cumulative probability $Q_m(z) = 1 - (1 - \mathrm{e}^{-z})^m$ of the L–S periodogram z against the exact beta distribution probability $Q_m(z) = 1 - [1 - (1 - 2z/n)^{n/2}]^m$. The curves are plotted for number of observations $m = 100, 1000$ and 10 000. The probabilities are corrected for the multiple trials, assuming the number of effective frequencies equals m. This approximates the case of reasonably uniform sampling and a sufficiently broad-band periodogram.

We are in a position to resolve the dispute over the normalization of the L–S spectra (Koen 1990; Horne & Baliunas 1986 and references therein). Both procedures are valid – normalization by the total variance and by the residual variance – and correspond to using $\Theta_\parallel$ and Θ_o statistics, respectively. The difference affects the distribution: in the former case the periodogram follows the beta distribution $I_{\Theta\parallel}[1, (n-3)/2]$, in the latter the Fisher–Snedecor distribution $F(\Theta_\mathrm{o}; 2, n-3)$.

It is assumed implicitly in the derivation of the theoretical distribution of the L–S periodogram that no constant shift of signal is present. Usually, the value of the constant term is a priori unknown, and it is estimated from the data, e.g. as the data average value. For even sampling, the constant is orthogonal with respect to sine and cosine functions and the assumption is of no consequence for the results. However, for uneven sampling, the constant term is not orthogonal with respect to L–S sine and cosine functions (Ferraz-Mello 1981). The L–S procedure involves explicit projection of observations on to directions of sine, cosine functions and implicit projection on to constant function, introduced by subtraction of the mean value. However, for uneven sampling the sine, cosine and constant function are not all orthogonal. If so, the model obtained from the orthogonal projections does not correspond to the least-squares solution (cf. Appendix A3). Such a model yields a biased estimate of amplitude, generally correlated with the constant term and not obeying the exponential distribution $\chi^2(2)$. Foster (1995) discussed an example severely affected by these problems. A fully orthogonal procedure involving sinusoid and constant functions derived by Ferraz-Mello (1981) constitutes a special case of our multiharmonic periodogram (Schwarzenberg-Czerny 1996). Summarizing, for uneven sampling, the power spectrum yields a biased and non-optimal estimate of amplitude A. Although the L–S spectrum is better, as it involves a pair of orthogonal functions, it is also neither unbiased nor optimal. We mean optimality in the sense of least-squares residuals, and bias means that, despite improving signal-to-noise ratio (S/N), estimates do not converge to the true value of A.

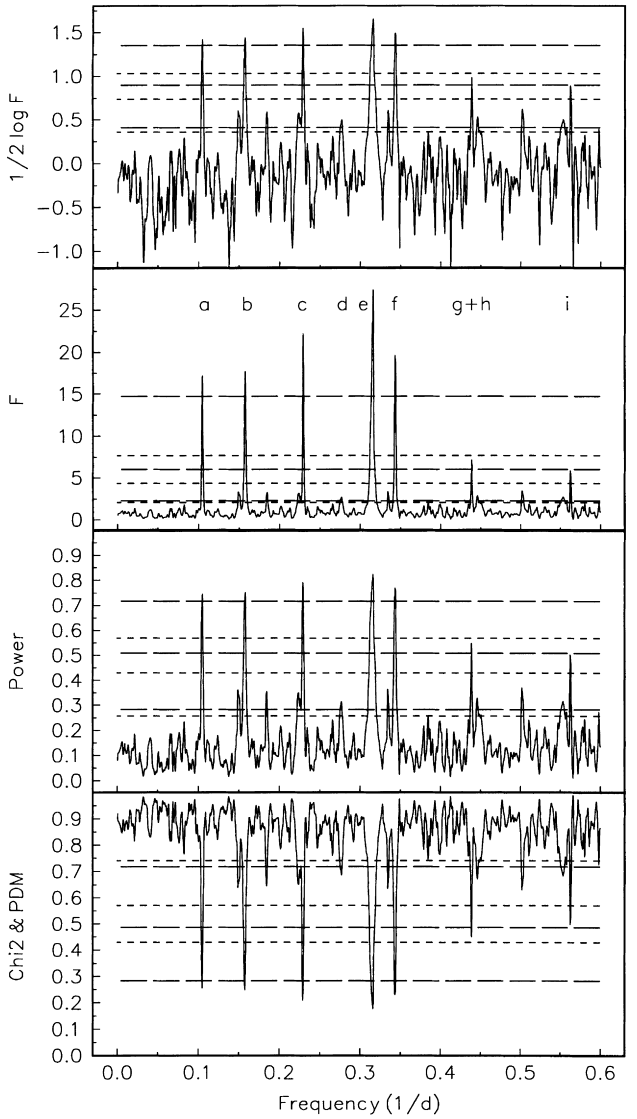

Figure 2. Four entirely equivalent periodograms (cf. equation 7) computed for the same set of 49 observations of BK Cen (Leotta-Janin 1967). All periodograms correspond to the same model function, with Fourier series of three harmonics, but to different fit quality statistics: $Z = (1/2)\ln\Theta_\mathrm{o}$, Θ_o, $\Theta_\parallel$ and $\Theta_\perp$ (equation 6). Their probability distributions are Snedecor $Z(d_\parallel, d_\perp)$ (Eadie et al. 1971), $F(d_\parallel, d_\perp)$, beta$(d_\parallel, d_\perp)$ and beta$(d_\perp, d_\parallel)$ (Abramovitz & Stegun 1971). In terms of the statistic in use, the last three periodograms correspond to the AOV (Schwarzenberg-Czerny 1989, 1996), modified power-spectrum (Lomb 1976; Scargle 1982), PDM and χ^2 (Stellingwerf 1978; Lomb 1976) periodograms. Note, however, that the periodograms in these papers differ from ours in the vertical units and model functions in use (a sinusoid and step function). Critical values of the statistics corresponding to Gaussian 2, 4 and 6σ significance levels for empirical and theoretical distributions are plotted with long and short-dashed lines, respectively.

5 EXAMPLE: OSCILLATIONS OF BK CEN

For a more realistic example, we reanalyse $n_\mathrm{o} = 49$ photometric

observations of the double-mode Cepheid BK Cen made by van Houten (Leotta-Janin 1967). Thus our results are directly comparable with those obtained using PDM and AOV periodograms for the same data, phase folded and binned (Stellingwerf 1978; Schwarzenberg-Czerny 1989). For each trial frequency we fitted data with Fourier series of $n_\mathrm{h} = 3$ harmonics, using our orthogonal projection method (Schwarzenberg-Czerny 1996). There were $d_\parallel = 2n_\mathrm{h} + 1 = 7$ and $d_\perp = n_\mathrm{o} - 1 - d_\parallel = 41$ degrees of freedom in our model and residuals, respectively. In the process we computed the values of three types of empirical fit quality statistics, Θ_o, $\Theta_\parallel$ and $\Theta_\perp$. Their probability distributions are $F(d_\parallel, d_\perp)$, beta$(d_\parallel, d_\perp)$ and beta$(d_\perp, d_\parallel)$ (equation 6). Additionally, we computed values of the Fisher statistic $Z = (1/2)\ln\Theta_\mathrm{o}$. We were motivated by the proximity of the Fisher Z distribution to the familiar Gaussian distribution. In this way we obtained the four equivalent periodograms plotted in Fig. 2. Additionally, we indicated with dashed lines the critical values of the statistics corresponding to Gaussian 2, 4 and 6σ significance level.

The values in these periodograms are uniquely related by equation (7). Except for different model functions (Fourier series instead of a sinusoid or a step function), the probability distribution of our Θ_o periodogram corresponds to the AOV periodogram, the distribution of $\Theta_\parallel$ to the Lomb (1976)–Scargle (1982) modified power spectrum and the distribution of $\Theta_\perp$ to both PDM and χ^2 periodograms of Lomb (1976). The differences are restricted to a different number of parameters, and the corresponding number of degrees of freedom, and to the units of the vertical axis. Normally, the modified power spectrum $\Theta_{\mathrm{L-S}}$, PDM Θ_{PDM} and χ^2 *per degree of freedom* Θ_{χ^2} periodograms are plotted in units such that their expected values for a pure noise signal (i.e. for hypothesis H_0) are 1. Our $\Theta_\parallel$ and $\Theta_\perp$ are plotted in units of the beta distribution, so that their values fit the range (0, 1). To convert our vertical units to more familiar units, one has to divide by expected values of the corresponding beta distributions:

$$\Theta_{\mathrm{L-S}} = \Theta_\parallel \frac{d_\parallel + d_\perp}{d_\parallel}, \tag{14}$$

$$\Theta_{\mathrm{PDM}} \equiv \Theta_{\chi^2} = \Theta_\perp \frac{d_\parallel + d_\perp}{d_\perp}. \tag{15}$$

Note that in equations (14) and (15) the same model functions and number of parameters are assumed on both sides. As simple consequence of the preceding considerations one concludes that all empirical power spectra, PDM and χ^2 periodograms have hard limits, corresponding to $\mathrm{S/N} = \|x_\parallel\|^2/\|x_\perp\|^2 \to \infty$. In particular, for the modified power spectrum, $\max \Theta_{\mathrm{L-S}} = (2 + d_\perp)/2$. The theoretical distribution derived by Lomb (1976) and Scargle (1982) is incorrect in that it does not account for this limit. A similar limit value applies to the ordinary power spectrum. However, its exact value is difficult to compute as it depends to some extent on sampling. Critical values corresponding to many σ significance levels cluster near these limits. Thus, for high S/N, the significance of two aliases of nearly equal height may differ by many σ, while for low S/N they would remain equally significant. Scant attention was paid in the past to the existence of this limit in empirical power spectra and to the associated statistical peculiarities. Observers seem to be totally unaware of these effects. Neither limits nor clustering occur in F and Fisher $Z = (1/2)\ln F$ periodograms. The Fisher periodogram may suit the intuition of observers particularly well as its distribution is very close to Gaussian, and its significance level grows nearly proportionally to the distance from its null expected value (Eadie et al. 1971). To our knowledge, Fig. 2 constitutes the first vivid demonstration of one-to-one correspondence between analysis of variance (AOV, Schwarzenberg-Czerny 1989), power spectrum (Lomb 1976; Scargle 1982), PDM (Stellingwerf 1978) and χ^2 (Lomb 1976) periodogram statistics. In the example we always use a Fourier series model. In the original papers, step functions and a sine function were used instead.

Let us return to the differences between theoretical and empirical probability distributions. Theoretical distributions correspond to assuming that the value of $\|x_\perp\|^2$ is known exactly and corresponds to its expected value $d_\perp \mathrm{Var}\{x\}$. In this extreme case, the $d_\parallel F$ statistics follow the distribution of its numerator, i.e. the χ^2 distribution. This corresponds to a known limit of the F distribution (Eadie et al. 1971):

$$\lim_{d_\perp \to \infty} d_\parallel F(d_\parallel, d_\perp) = \chi^2(d_\perp). \tag{16}$$

Although similar limits exist for the beta distribution, care must be taken over the units in order to preserve expected values that are consistent with periodogram statistics. A simple practical procedure is first to compute critical values for the distribution in equation (16) and then to compute from equation (7) the corresponding beta critical values. The critical values computed in this way for theoretical distributions and for the same significance levels as for empirical distributions are plotted in Fig. 2 with short-dashed lines. Inspection of the figure reveals that, for a small but still realistic number of observations, theoretical and empirical distributions differ dramatically.

6 BAYESIAN DETECTION CRITERION

The Bayesian detection criterion is derived by following a different approach to the decision theory from that used in the rest of this paper. Our derivation follows the methods and notation used by Helstrom (1960). Appendix A2 introduces the relevant concepts. H_0 states that observations x_i are Gaussian white noise, so that their expected value and variance are $E\{x_i\} = 0$, $\mathrm{Var}\{x_i\} = \sigma^2$ and their combined probability density is

$$p_0(u)\mathrm{d}^n u = \mathrm{d}^n u (2\pi)^{-n/2} \exp -||u||^2/2, \tag{17}$$

where $u_i = x_i/\sigma$ are normalized observations and $||u||^2 = \sum_{i=1}^n u_i^2$. Similarly, for H_1, $E\{x_i\} = \sigma S_i(\theta)$, $\mathrm{Var}\{x_i\} = \sigma^2$ and the combined probability density is

$$p_1(u)\mathrm{d}^n u = \mathrm{d}^n u (2\pi)^{-n/2} \exp -||u - S||^2/2, \tag{18}$$

where $S(\theta)$ and θ correspond to an assumed signal shape and its parameter(s), respectively. In this form H_0 has no parameters, i.e. it is a simple hypothesis, and H_1 is composite. The prior probabilities of H_0 and H_1 are respectively ζ and $1 - \zeta$ and the prior probability density of their parameters is $z(\theta)\mathrm{d}\theta$. One searches for a criterion dividing the whole space of possible observations x into two excluding regions R_0 and R_1 corresponding to decisions favouring H_0 and H_1. For a cost matrix

$$(C_{ik}) = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix}, \tag{19}$$

the mean cost of a choice between H_0 and H_1 is

$$\begin{aligned} \overline{C} = {} & \zeta\big[C_{00}(1 - Q_0) + C_{10}Q_0\big] \\ & + (1 - \zeta)\big[C_{01}Q_1 + C_{11}(1 - Q_1)\big], \end{aligned} \tag{20}$$

where $Q_0 = \int_{\mathrm{R}_1} p_0 \mathrm{d}^n u$ and $Q_1 = \int_{\mathrm{R}_0} z p_1 \mathrm{d}^r\theta \mathrm{d}^n u$. The mean cost of choices is minimized for R_0 and R_1 defined so that $\Lambda(x) < \Lambda_0$ holds for $x \in \mathrm{R}_0$ and the reverse inequality holds for R_1. The relevant merit function, called likelihood, and its critical value are

respectively (Helstrom 1960)

$$\Lambda(x) = \frac{\int z p_1 \mathrm{d}^r\theta}{p_0}, \tag{21}$$

$$\Lambda_0 = \frac{\zeta}{1-\zeta}\frac{C_{10}-C_{00}}{C_{01}-C_{11}}. \tag{22}$$

For simplicity we assumed $\partial C/\partial\theta = 0$, but relaxation of this condition is straightforward. Substituting equation (17) and (18) one obtains an equation of a dividing surface:

$$\int z \exp\left(-||u-S||^2/2 + ||u||^2/2\right)\mathrm{d}^r\theta = \Lambda_0. \tag{23}$$

From equation (4), it follows that the maximum value of the exponent in equation (23), $||u_\perp||^2$, is attained for such $\theta^{(\mathrm{o})}$ that $S(\theta^{(\mathrm{o})})$ constitutes a least-squares fit of u. In a volume $\Delta\theta$ around $\theta^{(\mathrm{o})}$ the exponent is large, and it may be approximated by its Taylor expansion:

$$\frac{1}{2}||u_\perp||^2 - \frac{1}{2}(\theta-\theta^{(\mathrm{o})})^T\left(\partial^2||u-S||^2/\partial\theta_i\partial\theta_j\right)(\theta-\theta^{(\mathrm{o})}), \tag{24}$$

i.e. the integrand in equation (23) reduces to a multidimensional Gaussian bell. Note that, by virtue of the Fisher lemma, $||u_\perp||$ does not depend on θ and may be carried in front of the integral sign. Linear orthogonal models are also useful in Bayesian statistics. Namely, for these models the Hessian matrix in equation (24) has the diagonal form $2||\partial S/\partial\theta_i||^2\delta_{ij} = 2||\phi^{(i)}||^2\delta_{ij}$, corresponding to a half-width of each Gaussian cross-section of $1/\sqrt{2}||\phi^{(i)}||$. For base functions $\phi^{(i)}$ of fixed amplitude norms, $||\phi^{(i)}||^2$ scales with number of observations n and the width of each cross-section scales with $\sqrt{1/n}$. Hence for large n and for an orthogonal model the integral in equation (23) may be calculated by the steepest descent method, yielding

$$||u_\perp||^2 \equiv \chi^2 = 2\ln\frac{\Lambda_0}{z(\theta^{(\mathrm{o})})\Pi_{i=1}^r 2\sqrt{\pi}||\phi^{(i)}||}, \tag{25}$$

where $z_\mathrm{o} \equiv z(\theta^{(\mathrm{o})})$. For model parameters scaled so that $E\{\theta\} \sim \mathrm{Var}\{\theta\} \sim 1$, the probability density $z_\mathrm{o} \sim 1$. In reasonably designed experiments, $\zeta \sim (1-\zeta) \sim 1$, hence $\Lambda_0 \sim 1$. Since for large sample size, n, all norms scale as $\sqrt{n}$, the dominant term on the right-hand side of equation (25) is $r\ln n$. Therefore, as already stated, an uncertainty of the prior probabilities collected in Λ_0/z_o plays little role for large samples.

Inspection of equation (25) obtained for the asymptotic case of a large sample reveals that in this case the Bayesian detection criterion reduces to the classical Neyman–Pearson (NP) χ^2 detection criterion with a particular choice of significance level (cf. Appendix B). In other words, a NP detection criterion corresponds to a Bayesian detection criterion with particular costs and prior probabilities. This statement represents a special case of a general result in decision theory (Wald 1950; Fourgeaud & Fuchs 1967). This demonstrates that, for sufficiently large samples, the results derived in our paper for NP criteria remain valid for the corresponding Bayesian criteria. One way to apply our results in Bayesian practice is by using equation (25) to translate between Λ_0 and the significance level in Bayesian and NP criteria, respectively.

7 DISCUSSION

7.1 Empirical periodograms and the beta distribution

The theoretical distributions of the periodogram statistics are derived for a given distribution of noise. In practice, one has to estimate the variance of the noise from the same data that are used for the periodogram itself. It is often argued that the empirical distributions converge to the theoretical distributions for sufficiently large samples. However, Schwarzenberg-Czerny (1989) and Davies (1990) presented an example in which no such convergence is achieved at all. Here, we demonstrate that, even for well-behaved cases, the convergence of the distributions is not fast enough to justify the use of the asymptotic distribution. One has to use the exact distributions instead. We have demonstrated, that a large class of the empirical periodograms, i.e. the periodograms normalized by empirical data variance, follows the beta distribution. The difference between the empirical and asymptotic distributions is particularly pronounced in the extremes of their tails, which are usually used for hypothesis testing, i.e. for evaluation of the significance of a detection. The situation is aggravated by the multiple-trial (bandwidth) penalty pushing the significant detections further away into the tail. *The beta distribution in fact applies to a large class of the so-called χ^2 statistics normalized by the empirical variance*, often encountered in experimental physics and observational astronomy.

7.2 Equivalence of periodograms

We demonstrated that for the same underlying model the three classes of the periodograms in use, corresponding to the Θ_o, $\Theta_\parallel$ and $\Theta_\perp$ statistics, are entirely equivalent to each other. Thus, they yield identical statistical results, provided that their correct probability distributions, correspondingly $F(r-1, n-r)$, $I_x[(r-1)/2, (n-r)/2]$ and $I_x[(n-r)/2, (r-1)/2]$, are used. If no average value was subtracted from the data then $F(r, n-r)$, $I_x[r/2, (n-r)/2]$ and $I_x[(n-r)/2, r/2]$ apply, respectively. From the statistical point of view, the choice between χ^2, ANOVA or PDM periodograms reduces just to a matter of taste, provided that the matching distribution is used and the underlying model remains the same. Thus users may obtain *temporal* spectra (periodograms) with 'emission' or 'absorption' lines corresponding to periodicities, depending on their preferences. Conversion between their respective statistics is effected by means of equation (7) (cf. Abramovitz & Stegun 1971, equations 26.5.2 and 26.6.2). To our knowledge, Fig. 2 constitutes the first vivid example in astronomical literature of the application of these classical formulae to the conversion of periodograms.

7.3 The role of the models

For a general class of the periodograms based on different orthogonal models, we presented a uniform picture of statistical properties. The class is broad enough to contain most known periodograms and narrow enough to obtain their detailed statistical properties. Our considerations have demonstrated that the probability distributions of periodograms for the H_0 hypothesis do not depend on the specific model used (i.e. on the type of orthogonal function). The change in the type of the periodogram (i.e. in the type of the statistic) reduces to the mere change of variables in the distribution, with no statistical consequences, as long as the periodograms belong to Θ type. The distribution does depend on the number of parameters (i.e. the number of functions) of the model. For a given sampling, the periodograms with the same number of parameters fitted have the same distribution for H_0 valid. In this way we obtain a uniform picture of the statistical properties of the most popular types of periodograms. The uniform picture does not yet allow the comparison of the relative sensitivity of periodograms, since the

sensitivity depends on distributions for which H_0 is invalid, not considered yet. The uniform picture constitutes a step in the right direction, however.

It is in order to warn the reader here that at least one dark cloud casts a shadow on the uniform picture: all the periodograms discussed here are as sensitive to the Gaussian distribution of the noise as χ^2 is.

ACKNOWLEDGMENTS

Thanks are due to an anonymous referee for suggestions that led to a broadening of the scope and improvement in the quality of presentation of this paper. This research was supported by KBN grant 2 P03C 001 12. I would like to thank Professor Gerard Vauclair and members of Observatoire du Midi-Pyrenees for their hospitality, and the Polish/French PAN/CNES exchange scheme for support of my visit, which enabled me to complete this work.

REFERENCES

Abramovitz M., Stegun I., 1971, Handbook of Mathematical Functions. Dover, New York
Akerlof C. et al., 1994, ApJ, 436, 787
Bickel P.J., Doksum K.A., 1977, Mathematical Statistics. Holden-Day, San Francisco
Birkhoff G., Mac Lane S., 1954, A Survey of Modern Algebra. Macmillan, New York
Davies S.R., 1990, MNRAS, 244, 93
Deeming T. J., 1975, Ap&SS, 36, 137
Eadie W.T., Drijard D., James F.E., Roos M., Sadoulet B., 1971, Statistical Methods in Experimental Physics. North-Holland, Amsterdam
Ferraz-Mello S., 1981, AJ, 86, 619
Fisz M., 1963, Probability Theory and Mathematical Statistics. Wiley, New York, p. 536 ff
Fourgeaud C., Fuch A., 1967, Statistique. Dunod, Paris
Foster G., 1995, AJ, 109, 1889
Grison P., 1994, A&A, 289, 404
Horne J.H., Baliunas S.L., 1986, ApJ, 302, 757
Helstrom C.W., 1960, Statistical Theory of Signal Detection. Pergamon Press, London
Koen C., 1990, ApJ, 348, 700
Lafler J., Kinman T.D., 1965, ApJS, 11, 216
Leotta-Janin C., 1967, Bull. Astron. Inst. Neth., 19, 169
Lomb N.R., 1976, Ap&SS, 39, 447
Lupton R., 1993, Statistics in theory and practice. Princeton Univ. Press, Princeton, NJ
Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., 1992, Numerical Recipes. Cambridge Univ. Press, Cambridge
Scargle J.H., 1982, ApJ, 263, 835
Schwarzenberg-Czerny A., 1989, MNRAS, 241, 153 (Paper I)
Schwarzenberg-Czerny A., 1991, MNRAS, 253, 198
Schwarzenberg-Czerny A., 1996, ApJ, 460, L107
Schwarzenberg-Czerny A., 1997, ApJ, 489, 941
Stellingwerf R.F., 1978, ApJ, 224, 953 (S78)
van der Klis M., 1989, in Ögelman H., van den Heuvel E., eds, Timing Neutron Stars. Am. Inst. Phys., New York, p. 27
Wald A., 1950, Statistical decision function. Wiley, New York
Whittaker E.T., Robinson G., 1926, The Calculus of Observations. Blackie & Son, London

APPENDIX A: THE ALGEBRAIC PROPERTIES

A1 Scalar product and vector norm

Given a scalar product, the standard linear algebra of the Hilbert vector spaces applies (e.g. Birkhoff & Mac Lane 1954, chapter IX). The product obeys the usual linearity and Hermitian symmetry laws (equations A1 and A2) and it induces the natural vector norm in $\mathcal{H}_n$ (equations A3 and A4):

$$< a\xi + \eta, \psi >= a < \xi, \psi > + < \eta, \psi >, \tag{A1}$$

$$\overline{< \eta, \xi >} = < \xi, \eta >, \tag{A2}$$

$$||\xi||^2 \equiv < \xi, \xi > \geq, 0 \tag{A3}$$

$$||\xi||^2 = 0 \leftrightarrow \xi = 0. \tag{A4}$$

A2 Orthogonal bases

A set Φ of n vectors $\phi^{(l)}$, $l = 1, \ , n$, constitutes a base of $\mathcal{H}_n$ if and only if none of them is parallel: $| < \phi^{(l)}, \phi^{(k)} > | < ||\phi^{(l)}|| \ ||\phi^{(k)}||$ for $l \neq k$. Any vector belonging to $\mathcal{H}_n$ may be uniquely expanded in terms of the base vectors (equation A5):

$$\xi = \sum^{n} c_l \phi^{(l)}. \tag{A5}$$

The base is called orthogonal if all its vectors are pairwise orthogonal, $< \phi^{(l)}, \phi^{(k)} >= 0$. By a simple rescaling the orthogonal base may be converted to the orthonormal base (equation A6). The family E of the time series $\epsilon^{(l)}$, vanishing everywhere except at a single observation z_l, constitutes an example of the Cartesian orthonormal base (cf. equations A10 and 1). The base E may be called the observations base. Let us now consider the transformation of the vector components by the change from one base, say E, to another base, say Φ. Calculation of the coefficients of the vector components c_l is particularly simple for the orthonormal new base Φ. It corresponds to the orthogonal projection (equations A7 and A8). For this reason we restrict our attention in the present paper to the orthonormal bases alone:

$$< \phi^{(k)}, \phi^{(l)} > = \delta_{kl}, \tag{A6}$$

$$c_l = < \phi^{(l)}, \xi > \equiv \phi^{(l)\dagger} \circ \xi. \tag{A7}$$

The proof follows from equation (A5) multiplied by $\phi^{(k)}$ and from equation (A6). Substituting equation (A7) into equation (A5), and using equation (A1), one obtains in the matrix notation equation (A8), where Φ denotes the matrix formed from the columns of the components of the new base vectors $\phi^{(l)}$ in the old base. The form of equation (A8) demonstrates that by the change of base the vector components transform linearly according to the matrix $\Phi^{\dagger}$. By reordering the brackets in equation (A8) we obtain the equation for the identity matrix $\mathbf{I}$, namely $\xi = (\Phi \circ \Phi^{\dagger}) \circ \xi$. Hence the transformation matrix Φ satisfies equation (A9). Such a matrix is called *unitary*. Thus in an arbitrary orthonormal base the scalar product and norm always reduce to a simple scalar product of the components c_l (equation A10). From equations (A8) and (A9) one obtains equation (A10), demonstrating that the values of the scalar product and of the norm are conserved with no regard for the orthonormal base in which it is calculated:

$$\xi = \Phi \circ (\Phi^{\dagger} \circ \xi), \tag{A8}$$

$$\Phi \circ \Phi^{\dagger} = \mathbf{I}, \tag{A9}$$

$$< \xi, \eta > = (\xi^{\dagger} \circ \Phi) \circ (\Phi^{\dagger} \circ \eta), \tag{A10}$$

where † indicates the transposed complex conjugate matrix.

A3 Orthogonality of model and residuals

The analysis of the observations is performed by fitting them with a model. A linear combination of r model functions constitutes a

convenient kind of model. These functions span a vector subspace $\mathcal{M}_r$ of $\mathcal{H}_n$. By the Gramm–Schmidt procedure these functions may always be transformed to the orthonormal model base $\phi^{(l)}$, $l = 1, \cdots, r$, of $\mathcal{M}_r$. By the same procedure the base may be supplemented by another $n - r$ orthogonal functions $\phi^{(l)}$, $l = r + 1, \cdots, n$, to the orthonormal base of $\mathcal{H}_n$. The additional functions span the subspace $\mathcal{R}_{n-r}$, such that $\mathcal{H}_n = \mathcal{M}_r \times \mathcal{R}_{n-r}$. The latter equality means that any vector $\xi \in \mathcal{H}_n$ corresponds to a pair of vectors $\xi_\parallel \in \mathcal{M}_r$ and $\xi_\perp \in \mathcal{R}_{n-r}$ and vice versa. The spaces $\mathcal{M}_r$, $\mathcal{R}_{n-r}$ and $\mathcal{H}_n$ are called here the model, residuals and observations spaces. By grouping together r and $n - r$ terms in equations (A8) and (A10) we obtain

$$\xi = \xi_\parallel + \xi_\perp, \quad \text{where} \tag{A11}$$

$$< \xi_\parallel, \xi_\perp > = 0, \quad \xi_\parallel \in \mathcal{M}_r, \quad \xi_\perp \in \mathcal{R}_{n-r}, \tag{A12}$$

$$||\xi||^2 = ||\xi_\parallel||^2 + ||\xi_\perp||^2. \tag{A13}$$

The vector $\xi_\parallel$ constitutes the least-squares fit of ξ in $\mathcal{M}_r$ in terms of the norm statistic $||\xi - \xi_\parallel||$ [point (ii) in Section 1]. Indeed, for $\tilde{\xi}_\parallel \in \mathcal{M}_r$ the minimum of $||\xi - \tilde{\xi}_\parallel||^2 = ||\xi_\parallel - \tilde{\xi}_\parallel||^2 + ||\xi_\perp||^2$ is obtained for $\xi_\parallel = \tilde{\xi}_\parallel$. We have exploited here equation (A12). The advantage in using the orthogonal models in the statistics stems from the simplicity of the least-squares solution. Note that, for a special type of FFT base and sampling, equation (A8) corresponds to the forward and inverse Fourier transforms (FFTs). Then equation (A10) for (ξ, ξ) corresponds to Parseval's theorem.

APPENDIX B: DECISION THEORY

B1 Statistical hypotheses

Two alternative outcomes of signal detection are possible. Either the signal consists of a pure noise or it contains noise *and* a deterministic signal carrying some information. An observer has to decide which case is true from incomplete evidence in his observations x. Statisticians say that one has to choose between the null and alternative hypotheses, H_0 and H_1, respectively. By choosing between H_0 and H_1 one divides the space of all possible observations x into two excluding regions R_0 and R_1. A general equation of a surface dividing R_0 from R_1 takes the following form:

$$\Theta(x) = \Theta_c. \tag{B1}$$

An equation of the form (B1) is called a decision criterion or decision rule, a function of random observations $\Theta(x)$ is called a statistic and a constant Θ_c is called a critical value of the statistics. Generally, identification of the regions R_0^∞ and R_1^∞ actually corresponding to H_0 and H_1 requires an infinite number of observations. The object of decision theory is to provide decision rules for which R_0 and R_1 are optimum approximations of R_0^∞ and R_1^∞ in a certain sense. The sense of optimality and corresponding Θ depend on the assumed *prior* knowledge about a model deterministic signal in H_1, our prejudice against H_0 and H_1 and the costs of error decisions. Specifically, in period searches Θ is a function of both pulse frequency ω and the profile. The plot of Θ against ω for *a given set of observations* is called a periodogram. The presence in a periodogram of frequencies for which $\Theta(\omega) > \Theta_c$ marks the detection of a periodic signal with a frequency ω. This is a kind of game statisticians call hypothesis testing (e.g. Eadie et al. 1971). The precondition for hypothesis testing is knowledge of the probability distribution of Θ for an assumed distribution of x.

B2 Prior knowledge

The probabilities of H_0 or H_1 being true are ζ or $1 - \zeta$. Often hypotheses depend on unknown parameters θ, such as noise amplitude, signal-to-noise ratio and signal frequency. Let these parameters have probability distribution $z(\theta)$. Often these probabilities may be estimated *prior* to the current experiment, e.g. from past surveys.

The costs in our example may be quantified, e.g. as wasted fractions of the total available photometric and spectroscopic time, C_{01} and C_{10} respectively. The whole cost matrix, C, may appear as follows:

Reality	Decision	
	H_0	H_1
H_0	0	C_{01}
H_1	C_{10}	0

(B2)

B3 Decision rules

Decision rules differ in their use of prior knowledge. In the Neyman–Pearson (NP) rule no explicit use of costs and prior probabilities is made. Instead, one sets the probability of false alarms, $1 - \alpha$, at a small fixed level

$$Q_0 \equiv \int_{R_1} p_0 \mathrm{d}x = 1 - \alpha. \tag{B3}$$

On the one hand, if H_1 is rather improbable, $\zeta \approx 1$, then the commonest errors are false alarms and the NP rule is reasonable, because it ensures that the total probability of errors does not exceed $1 - \alpha$. On the other hand, this is not a reasonable criterion if (i) a small fixed fraction of false alarms is still too costly or (ii) the success rate is high, $\zeta \ll 1$, and/or false alarms are inexpensive compared with misses. In view of the above, the fact that the NP rule is often used in practice with good results may not prove that the NP rule is always optimum and free of prior assumptions. Acception or rejection of this statement pertains to the dispute between Bayesian and anti-Bayesian statisticians.

If information about costs and prior probabilities C, ζ and z is available it may be more reasonable in case (ii) to minimize the mean cost of a decision,

$$\overline{C} = \zeta\big[C_{00}(1 - Q_o) + C_{10}Q_o\big] + (1 - \zeta)\big[C_{01}Q_1 + C_{11}(1 - Q_1)\big], \tag{B4}$$

where $Q_1 \equiv \int_{R_0} z p_1 \mathrm{d}\theta \mathrm{d}x$. The corresponding decision rule is called Bayesian. Use of costs and prior probabilities constitutes a characteristic feature of the Bayesian procedures.

Critics point out that for entirely new experiments no prior estimates of ζ and z are valid. Indeed, the aim of many experiments is to determine the underlying ζ and z from the distribution of observations $p(x; \theta)$. Parameters of these distributions, θ, are estimated in the process. Bayesian advocates point out that the NP criteria involve severe bias against H_1, specified by the confidence level, while experiments are often set up in a way biased in favour of H_1. When unknown parameters θ are involved, Bayesian criteria depend on their *prior* probability distribution $z(\theta)$. Bayesian statisticians point out that the role of the uncertainties in prior probabilities diminishes for large samples. It is also true, that for large samples NP, Bayesian and minimax criteria are all similar. This is illustrated in Section 6. Anti-Bayesian critics point out that

any advantage in the Bayesian use of prior probabilities diminishes accordingly.

In the example discussed in Section B2, bias is introduced because only likely candidates are surveyed photometrically. The Bayesian criterion takes this bias ζ into account, yielding the minimum average cost of decisions. If costs are known but no prior probabilities are available, it may be reasonable in case (i) to minimize the maximum possible cost of a decision. This is achieved by selecting the ζ_0 for which the average cost of a decision is maximum, $\partial\overline{C}/\partial\zeta = 0$, and following the corresponding Bayesian rule. This is called a minimax rule.

In fact, the NP and minimax criteria constitute the Bayesian criteria for particular cost and prior probabilities. All these criteria may be formulated using the same statistic $\Theta(x)$ but different critical values Θ_c (e.g. Helstrom 1960). All these criteria are valid as long as their underlying assumptions are satisfied. The question of which criterion performs best depends largely on circumstances.

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