THE ASTROPHYSICAL JOURNAL, 460:L107–L110, 1996 April 1

FAST AND STATISTICALLY OPTIMAL PERIOD SEARCH IN UNEVEN SAMPLED OBSERVATIONS

A. SCHWARZENBERG-CZERNY
Astronomical Observatory of Adam Mickiewicz University, ul. Słoneczna 36, 60-286 Poznań, Poland,
and Copernicus Astronomical Centre, ul. Bartycka 18, 00-716 Warsaw, Poland
Received 1995 December 6; accepted 1996 January 18

ABSTRACT

The classical methods for searching for a periodicity in *uneven sampled* observations suffer from a poor match of the model and true signals and/or use of a statistic with poor properties. We present a new method employing periodic orthogonal polynomials to fit the observations and the analysis of variance (ANOVA) statistic to evaluate the quality of the fit. The orthogonal polynomials constitute a flexible and numerically efficient model of the observations. Among all popular statistics, ANOVA has optimum detection properties as the uniformly most powerful test. Our recurrence algorithm for expansion of the observations into the orthogonal polynomials is fast and numerically stable. The expansion is equivalent to an expansion into Fourier series. Aside from its use of an inefficient statistic, the Lomb-Scargle power spectrum can be considered a special case of our method. Tests of our new method on simulated and real light curves of nonsinusoidal pulsators demonstrate its excellent performance. In particular, dramatic improvements are gained in detection sensitivity and in the damping of alias periods.

Subject headings: binaries: eclipsing — methods: data analysis — methods: statistical — pulsars: general — stars: oscillations — X-rays: stars

1. INTRODUCTION

Popularity of the discrete Fourier transform (DFT) based power spectrum in period analysis stems from its simplicity, clear interpretation and computational efficiency. There are two problems in application of DFT to searching for periodicities, however. First, it is well known that DFT uses a sinusoid to model the observations and hence is unsuitable for the nonsinusoidal signals. Second, it is less well known that in practical applications neither power spectrum nor its modification by Lomb (1976) and Scargle (1983) obey the theoretical $\chi^2(2)$ probability distribution. The departure from the χ^2 distribution occurs in the normalization of the periodogram by an empirical variance (§ 3). In § 2 we present a new method which does not suffers from these problems. In §§ 4 and 5 we compare the performance of the new and classical methods on simulated and observed time series, respectively.

2. METHODS

The correspondence between Fourier series and complex polynomials was noted long before the advent of electronic computers. Polynomials useful for the present application are periodic and orthogonal on a discrete set of uneven observations. We are unaware of any recent discussion of the numerical properties of these polynomials. Hence, we present here the details of the algorithm. A Fourier series of N harmonics $F^{(N)}(t)$ corresponds to the complex polynomial of order $2N$, namely $\Psi_{2N}(z) = z^N F^{(N)}(t)$. Writing explicitly, one obtains

$$\begin{aligned} z^N \sum_{n=0}^{N} (a_n \cos n\omega t + b_n \sin n\omega t) \\ = \frac{1}{2} \sum_{n=0}^{N} [(a_n - ib_n) z^{N+n} + (a_n + ib_n) z^{N-n}], \end{aligned} \quad (1)$$

where $a_0 = b_0$. All the arguments $z = e^{i\omega t}$ lay on the unit circle $|z| = 1$. For a given trial frequency ω, times of observation t_k map onto points on the unit circle: $z_k = e^{i\omega t_k}$, $k = 1, \ldots, K$. The complex polynomials form Hilbert space with the scalar product defined by the Stjeltjes integral over the unit circle $(\Phi, \Psi) = (1/2\pi) \int_{-\pi}^{\pi} \Phi(z)\overline{\Psi(z)}\, d\mu(t)$. The integral depends on the weight function $\mu(t)$ defined on the circle. A natural weight function μ for the case of discrete observations is the step function with the steps at the phases ωt_k. Then the scalar product reduces to the weighted sum:

$$(\Phi, \Psi) = \sum_{k=1}^{K} g_k \Phi(z_k) \overline{\Psi(z_k)}. \quad (2)$$

Thus, observations with uneven weights $g_k \sim 1/\mathrm{Var}\,\{X_k\}$ are accounted for here in a natural way.

For a given base $\Phi_N(z) = \Sigma_{n=0}^{N} a_n^{(N)} z^n$, $N = 0, 1, \ldots,$ the expansion of a polynomial Ψ_N is unique:

$$\Psi_N(z) = \sum_{n=0}^{N} c_n \Phi_n(z). \quad (3)$$

In the polynomial Hilbert space, there exists a unique orthonormal base, with real and positive leading coefficients $0 < a_n^{(n)} \in R$. It satisfies the following conditions of orthonormality:

$$(\Phi_n, \Phi_m) = \delta_{n,m}. \quad (4)$$

A convenient way to generate Φ_N is by means of the recurrence relation (Grenander & Szegö 1958):

$$\tilde{\Phi}_0(z) = 1, \quad (5)$$

$$\tilde{\Phi}_{n+1}(z) = z\tilde{\Phi}_n - \alpha_n z^n \overline{\tilde{\Phi}_n(z)}, \quad (6)$$

$$\Phi_n(z) = \frac{\tilde{\Phi}_n(z)}{\sqrt{(\tilde{\Phi}_n, \tilde{\Phi}_n)}}. \quad (7)$$

The formulae for the coefficients α_n and c_n follow from equations (6) and (3) multiplied by Φ_n:

$$\alpha_n = \frac{(z\tilde{\Phi}_n, \tilde{\Phi}_n)}{(z^n\bar{\tilde{\Phi}}_n, \tilde{\Phi}_n)}, \quad (8)$$

$$c_n = \frac{(\Psi, \tilde{\Phi}_n)}{\sqrt{(\tilde{\Phi}_n, \tilde{\Phi}_n)}}. \quad (9)$$

Equation (9) corresponds to a generalized Fourier integral. The recurrence solution for $\tilde{\Phi}$ and c_n proceeds in the following steps: (1) set $n = -1$, $\tilde{\Phi}_{-1} = 1/z$ and $\alpha_{-1} = 0$, (2) increment n by 1, (3) compute equation (6), (4) add the increments to the scalar products (eqs. [8] and [9]) for all observations, (5) compute α_n and c_n, and finally (6) jump back to step 2. In this way, in the z space one fits $\Psi = z^N X$ with the polynomial Ψ_{2N} and in the real space X with the corresponding Fourier series $F^{(N)}$. Although the fit is obtained by a fast recurrence process, the result is identical to the least-squares solution of the overdetermined system $\sum_{n=0}^{N} \Phi_n(z_k)c_n = \Psi(z_k)$, $k = 1, \ldots, K$ (Eadie et al. 1982; Schwarzenberg-Czerny 1995).

The new algorithm requires $\mathbb{O}(WNK)$ operations, i.e., much less than the $\mathbb{O}(WN^2K)$ required by the least-squares solution. Here W is number of frequencies ω. Our algorithm is only a factor N less efficient than the ordinary, non-FFT, DFT power spectrum for uneven sampling. Our orthogonal expansion is advantageous by virtue of its numerical stability and statistical independence. The variance of the parameters attains a minimum in the orthogonal case. It is true that the phase-folding and binning may be considered an expansion in terms of the periodic orthogonal step functions (e.g., Schwarzenberg-Czerny 1989). However, for most applications, step functions are less suitable base functions than the smooth functions corresponding to Φ_n.

Parseval's theorem corresponding to Fourier integral in equation (9) has the following form:

$$(2N + 1)\,\widehat{\mathrm{Var}}\,\{F^{(N)}\} \equiv \sum_{k=0}^{K} (F^{(N)})^2 = \sum_{n=0}^{2N} |c_n|^2. \quad (10)$$

A simple proof follows from equations (1) and (3), and from the identity $(F^{(N)}, F^{(N)}) = (\Psi_{2N}, \Psi_{2N}) = \sum_{n=0}^{2N} (\Psi_{2N}, \Phi_n)(\Phi_n, \Psi_{2N})$. The orthogonal expansion discussed above yields the variance of the fitted Fourier series without explicit use of the sine and cosine amplitudes a_n and b_n. Note that because of orthogonality, c_n, $n = 0, 1, \ldots$ are statistically independent: a change of c_m in equation (3) does not affect the estimate of c_n in equation (9) for $m \neq n$. For $N = 1$ our variance estimate $|c_0^{(1)}|^2 + |c_2^{(1)}|^2$ reduces to Lomb-Scargle statistic: it constitutes the sum of two statistically independent variables, is quadratic in X, and is invariant against time shift (Lomb 1976). A time shift rotates the whole complex plane without affecting the absolute values $|c_n|$. For the reasons discussed in § 3 we do not recommend direct use of the variance (eq. [10]) for period searching.

We recommend instead use of the analysis of variance (ANOVA) statistic. Let observations X consist of the sum of the signal F and of the noise E: $X_k = E_k + E_k$. Their variance estimates satisfy an algebraic identity $(X, X) = (\Psi, \Psi) = (\Psi_{2N}, \Psi_{2N}) + \sum_{n=2N+1}^{K-1} |c_n|^2$ equivalent to $(K - 1)\,\widehat{\mathrm{Var}}\,\{X\} = 2N \times \widehat{\mathrm{Var}}\,\{F\} + (K - 2N - 1)\,\widehat{\mathrm{Var}}\,\{E\}$ (Scheffe 1959). The relevant ANOVA statistic $\Theta \equiv \widehat{\mathrm{Var}}\,\{F\}/\widehat{\mathrm{Var}}\,\{E\}$ has the following form:

$$\Theta(\omega) = \frac{(K - 2N - 1)\sum_{n=0}^{2N} |c_n|^2}{(2N)[(X, X) - \sum_{n=0}^{2N} |c_n|^2]}. \quad (11)$$

Note that it suffices to compute the total variance $(K - 1)\,\widehat{\mathrm{Var}}\,\{X\} = (X, X)$ in equation (11) only once at the beginning. It is assumed in equation (11) that the average value was subtracted from the data X. This is recommended by numerical considerations. If no subtraction has been made, replace $2N$ in the denominator (and *not* in the numerator) by $2N + 1$.

3. STATISTICAL CONSIDERATIONS

For the quantiative evaluation of detection significance one uses the probability distribution of the periodogram for the hypothesis H_0. The hypothesis H_0 states that observations contain only pure white noise. The relevant statistical procedure is hypothesis testing. The statistics $\widehat{\mathrm{Var}}\,\{F\}$ and $\widehat{\mathrm{Var}}\,\{E\}$ are independent by virtue of Fisher's Lemma. So, the AOV periodogram, even for small samples, has Fisher's F probability distribution with $2N$ and $K - 2N - 1$ degrees of freedom $F(2N, K - 2N - 1)$ (Scheffe 1959). The classical statistics (power spectrum, PDM, χ^2 for the residuals, χ^2 for the signal) all have to be normalized by the empirical variance $\widehat{\mathrm{Var}}\,\{X\}$. Because of this normalization their true distribution is no longer the theoretical distribution. Although the normalized statistics are ratios of variances, they do not obey Fisher's F distribution since $\widehat{\mathrm{Var}}\,\{X\}$ is *correlated* to the periodogram itself. Fisher's F distribution holds only for ratios of independent variances. The string-length statistic (Dvoretsky 1983) obeys no known distribution for the same reason. For most of these statistics, the tail of the distribution used for testing may converge to its theoretical distribution very slowly at best, and diverge at worst (Schwarzenberg-Czerny 1989). All classical periodograms rely on some quadratic functions of the observations, i.e., on some variance estimates. Among these statistics, for a fixed number of the parameters, the AOV is the uniformly most powerful test, i.e., it is the most sensitive for detection (Scheffe 1959). This is amply illustrated by our examples in §§ 4 and 5.

A period search method assumes that data may contain a given model of a signal, called by statisticians an alternative hypothesis H. A good model fits data well with a minimum number of the parameters, i.e., with minimal loss of degrees of freedom. Such a model yields decreased $\widehat{\mathrm{Var}}\,\{E\}$ and increased significance or sensitivity of the detection in equation (11). A poor model, inflexible and with too few parameters loses sensitivity by growth of $\widehat{\mathrm{Var}}\,\{E\}$. The pure sinusoid used in the power spectrum is a poor model for a nonsinusoidal signal. The step functions implicitly involved in all phase-binning statistics are poor models for a smooth signal. For other reasons, the loss of degrees of freedom occurs for correlated observations. Correlated observations are equivalent to a smaller number of independent observations (Schwarzenberg-Czerny 1991). For a model with an excess number of parameters, the residuals $(2N + 1)\,\widehat{\mathrm{Var}}\,\{E\}$ do not decrease quickly with N, i.e., the surplus c_n's are small. Averaging out these small c's in the numerator of equation (11) results in a smaller significance of detection. This problem affects the string-length methods. A related problem, concerning the "bandwidth penalty," is discussed below. In practice one rarely tests the simple alternative hypothesis H_1: "The data contain the signal of assumed shape and of frequency ω_1, suspected from other data." More commonly the test is of the composite hypothesis H_{N_f} for unknown ω. It corresponds to many, say N_f, simultaneous tests of the simple hypotheses/models at once, where $1 \leq N_f \leq \min(W, K)$. One pays the "bandwidth penalty" for the corresponding waste of the degrees of freedom for the parameters of all these simple models. The bandwidth penalty is best expressed in terms of the confidence levels $P_{N_f} = 1 - (1 - P_1)^{N_f}$, where P_1 and P_{N_f} are the confidence levels for H_1 and H_{N_f} detections, respectively. Unfortunately, N_f has to be guessed at, or estimated by simulations (Horne & Baliunas 1986). For further interesting discussions of this issue, see the Eadie et al. (1982) monograph, §§ 10.5.5 and 11.5.1.

The present Fourier model enables optimum tuning of the number of the parameters to the data. Additionally, our method is conservative in the sense that for a given number of parameters, the fitted series has minimum variance between the observed points (see Grenander & Szegö 1958). For a pure sinusoidal signal our 1 harmonic periodogram corresponds to the Lomb-Scargle periodogram transformed using equation (11). The computational overhead associated with the transformation is negligible. Still, for this case our method retains an advantage in its optimum statistic.

4. TESTS BY SIMULATIONS

We employed simulations to compare the performances in detection of the periodic signals of our new periodogram and of the power spectrum. To simulate an observed time series, 1000 times of observations t_k were drawn from the standard normal distribution, half of them were then shifted by +10. For these moments of time we computed the values of the signal. The signal consists of the periodic narrow pulse of the height S/N plus unit-variance Gaussian white noise E. The signal was calculated according to the formula $X_k = (\mathrm{S/N}) \exp(-\sin^2 60\pi t_k) + E_k$. A family of time series was simulated for signal-to-noise ratios S/N in the range from 16 to 1/8. Each of the time series was analyzed using our new periodogram for eight harmonics, and the Lomb-Scargle power spectrum, over the frequency range from 0 to 100 cycles day^{-1} with oversampling by a factor of 5. The confidence levels P_1 for the detection of the pulsations at 60 cycles day^{-1} were computed from $\chi^2(2)$ and $F(16,983)$ distributions, respectively. Both were corrected for the estimated $N_f = 600$ independent frequencies (Horne & Baliunas 1986). The results, converted to the normal distribution σ confidence levels, are presented in Figure 1. Note that for the simulation, our new periodogram yields steeper rise of the significance with S/N and unique detections for weaker signals compared to the power spectrum.

5. APPLICATIONS

Walker (1994) obtained high-quality (S/N $\geq$ 30) CCD light curves of the pulsating variables in M68. Although the observations span over 90 days, they are clumped into just 3 weeks. Hence, the aliasing at 1 and 1/70 cycle day^{-1} poses a severe problem in the analysis (see Fig. 6 of Walker 1994). Our method applied to the observations of the single mode RRab pulsating stars yielded periodograms free of most spurious features. The periodograms identify the true period quite unambiguously (Fig. 2). The only remaining spurious features

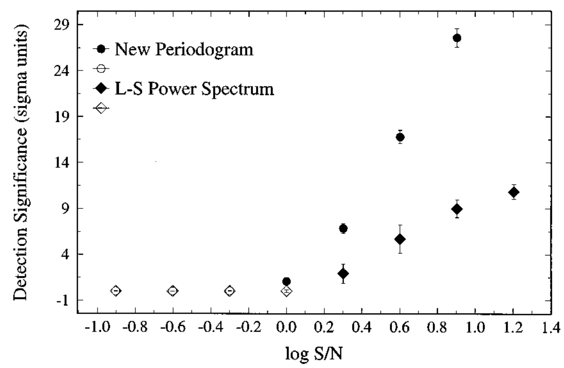

FIG. 1.—Confidence levels for the detection of a simulated narrow pulse signal (see text) plotted against log amplitude S/N ratio, with the Monte Carlo error bars. Filled symbols indicate the unique detections as the highest peaks in the 0 to 100 cycles day^{-1} range. Note that compared to the Lomb-Scargle power spectrum, our new periodogram yields unique detections for smaller S/N. Our detection significance level rises more steeply with S/N.

are the subharmonics of the main period. An alias period enables fit of the light curve with an amplitude comparable to that for the true period. Hence, the alias in the power spectrum may be nearly as high as for the true period. Still, for good observations the residuals and the denominator in equation (11) for the alias period are large compared to these for the true period. The consequence is the damping of the alias in our ANOVA spectrum. The better the fit for the true period, i.e., the larger S/N and the more harmonics are used, the more pronounced is the damping of aliases. For the smooth light curves, e.g., for V15 and V18, the best results were obtained using up to 4 harmonics. For the light curve of V22 featuring a sharp peak, up to 8 harmonics were useful. For the data at hand, use of higher harmonics is counterproductive because of the degrees of freedom economy (§ 3).

For the bimodal pulsating stars (RRc), e.g., V21 or V7 our mathematical model is no better than the pure sinusoid in the power spectrum. Clearly, the residuals are dominated by the other mode. Use of the harmonics higher than 2 never improves the periodogram for these stars. For S/N $\ll$ 1 another oscillation has no effect on the detection of the primary oscillation, except for its contribution to the general noise level. The detection for S/N $\ll$ 1 depends on the sampling of observations. For the well-sampled observations, covering many cycles of both periods, the peaks of the periodogram converge to the values of $\Theta(\omega_j) = (K - 2N - 1)\Pi_j/2N \, \Sigma_{i \neq j}$

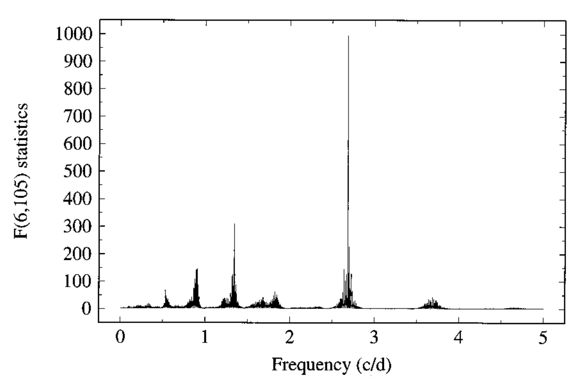

FIG. 2.—Our new periodogram for the single mode pulsating star V15 of the type RRab (see Fig. 6 of Walker 1994). The periodogram was calculated using 3 harmonics expansion. Note the damping of aliases. The only remaining spurious features are the subharmonics of the main period.

Π_i, where Π_i is the total power in all the harmonics of ith mode. The convergence is asymptotic for $S/N \to \infty$ and $N \to \infty$. For an incomplete sampling some insight can be gained by simulations. For Walker's (1994) sampling, S/N = 30, and frequencies and amplitudes of a typical bimodal RRc star, we obtained peak values of Θ_i, $i = 0$, 1 that differed by up to a factor of 3 from their asymptotic values, depending on the initial phases of the modes. The difficulties encountered by our method for these stars affect all the classical methods relying on the single period models, too. As not much overhead is involved, there is little harm in using our method for this case also. The generalization of the present methods for the multiple frequencies is beyond the scope of the present paper.

6. CONCLUSIONS

The statistical advantages of our new period search method are the following: (1) a tunable and flexible Fourier model for oscillations, (2) orthogonal variables, and (3) the optimum (uniformly most powerful) ANOVA statistic which guarantees the best sensitivity. An extra bonus comes with the known statistical properties for the small samples. The model is ideal for pulsating and/or eclipsing X-ray sources and variable stars. The ANOVA statistic takes full advantage of good S/N by making aliases less of a problem. Our orthogonal expansion algorithm is efficient and numerically stable. In fact, for 1 harmonic our method reduces to the non-FFT Lomb-Scargle periodogram, except for our better statistic (eq. [11]).

The author is grateful to Paweł Moskalik and Janusz Kałużny for stimulating and friendly discussions. Thanks are due to Steve Kawaler and the anonymous referee for comments that improved the presentation of the Letter. The present research was supported by KBN grant 2 P304 002 005.

REFERENCES

Dvoretsky, M. M. 1983, MNRAS, 203, 917
Eadie, W. T., Drijard, D., James, F. E., Roos, M., & Sadoulet, B. 1982, Statistical Methods in Experimental Physics (Amsterdam: North-Holland)
Grenander, U., & Szegö, G. 1958, Toeplitz Forms and Their Applications (Berkeley: Univ. California Press)
Horne, J. H., & Baliunas, S. 1986, ApJ, 302, 757
Lomb, N. R. 1976, Ap&SS, 39, 447
Scargle, J. H. 1982, ApJ, 263, 835
Scheffe, H. 1959, The Analysis of Variance (New York: Wiley)
Schwarzenberg-Czerny, A. 1989, MNRAS, 241, 153
———. 1991, MNRAS, 253, 198
———. 1995, A&AS, 110, 1
Walker, A. R. 1994, AJ, 108, 555