THE ASTROPHYSICAL JOURNAL, 489:941–945, 1997 November 10

THE CORRECT PROBABILITY DISTRIBUTION FOR THE PHASE DISPERSION MINIMIZATION PERIODOGRAM

A. SCHWARZENBERG-CZERNY
Astronomical Observatory of Adam Mickiewicz University, ul. Słoneczna 36, 60-286 Poznań, Poland;
Copernicus Astronomical Centre, ul. Bartycka 18, 00-716 Warsaw, Poland; alex@amk.edu.pl

Received 1996 October 15; accepted 1997 June 20

ABSTRACT

The phase dispersion minimization (PDM) statistic is a popular method for searching for nonsinusoidal pulsations. The method involves comparison of the PDM value with the Fisher-Snedecor F distribution to assess the significance of pulsations. Later, it was demonstrated that the PDM statistic does not follow the F distribution and that the sensitivity of the method in its original form, based on the F distribution, is poor. In the present paper we demonstrate that the PDM statistic follows a *beta* distribution, which we designate PDM*. We find that the significance of a given detected pulsation according to the relevant beta distribution is higher than the corresponding significance obtained using the F distribution. In addition, we demonstrate that all methods relying on phase binning and variance estimates related to χ^2 are *equivalent* for a given sampling, binning, and weighting pattern. We conclude by observing that with the invention of the high-performance Fourier series method, based on orthogonal projections, the original motivation for the use of phase binning for nonsinusoidal signals has weakened considerably.

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

1. INTRODUCTION

The statistical evaluation of detected periods resembles the evaluation of a theoretical curve fitted to experimental data (Lomb 1976). The periodogram statistic, Θ, measures the fit for a given pulsation frequency. The probability distribution of Θ is then used to calculate the probability, $P(\Theta > c)$, of obtaining the value of the periodogram higher than the actual observed value, $\Theta = c$, from a hypothetical pure noise signal. An unlikely good fit, corresponding to small P, is interpreted as detection of the corresponding period. Its complement probability, $\alpha = 1 - P(\Theta > c)$, is called the significance level. This is the kind of the game statisticians call hypothesis testing (e.g., Eadie et al. 1971). A precondition of hypothesis testing is a knowledge of the probability distribution of Θ.

The Θ statistics considered in the present paper were derived from methods devised by early observers of spectroscopic binaries. They searched for periods that would either minimize the scatter of their observations around the smooth curve or maximize the amplitude of the curve. Later these methods were combined with phase histograms and χ^2-related statistics. The work of Whittaker & Robinson (1926, hereafter WR), Lafler & Kinman (1965), and Stellingwerf (1978, hereafter S78) was important in that these authors insisted on adopting rigorous statistical machinery for evaluating detected periods. The phase dispersion minimization method (PDM) of Stellingwerf became particularly popular. However, Schwarzenberg-Czerny (1989, hereafter Paper I), and slightly later Davies (1990), noted that the Fisher-Snedecor distribution proposed in S78 for the PDM statistic is applied incorrectly and that it reduces sensitivity. These authors proposed instead using the analysis of variance (ANOVA, AOV) periodogram, giving examples of ways in which an incorrect statistic for PDM reduces sensitivity, and discussing the relation of the PDM, WR, and AOV statistics. No correct probability distribution for the PDM statistic itself has so far been published. In Paper I the asymptotic normal PDM distribution was derived. In § 2 of this paper, we derive the exact analytical probability distribution for the PDM periodogram. Section 3 is devoted to a discussion of its properties. The new statistic is compared with the previously employed F statistic in § 4. We also suggest ways to reinterpret the results of PDM obtained with the F statistic in agreement with the correct beta statistic. As incorrect probability distributions for the PDM periodogram have been used in the astronomical literature for the past two decades, our conclusions are new; for the first time the correct distribution for the PDM statistic is identified. For statisticians, our discussion contains nothing new, as it simply constitutes an application of the classic results of Fisher. Our discussion is particularly relevant for astronomical observations, since it holds for uneven sampling.

The papers cited above are all concerned with single-frequency (i.e., single-trial) probability. Since realistic periodograms cover multiple frequencies, their statistical evaluation *requires significant correction* for multiple trials. This correction is called a bandwidth penalty or bandwidth correction. In this paper we concentrate on single-trial probabilities, specifically for phase-folding periodograms. These single-trial probability are required for the bandwidth correction, either for analytical approximations or for testing Monte Carlo simulations. The bandwidth penalty issue is not specific to the type of periodogram used, and the reader is encouraged to refer to van der Klis (1989) for an introduction and to Horne & Baliunas (1986) for a description of Monte Carlo simulations. As pointed out by this paper's anonymous referee, the analytical formulae fitted to simulations by the latter authors often yield wrong results (e.g., for evenly spaced observations), and generally should not be used.

2. THE DISTRIBUTION OF THE PDM STATISTICS

Let us consider n observations x_{ij} folded in phase into r bins, so that the indices i and j stand for bin number and observation number, respectively. The ranges of the indices are $1 \leq i \leq r$ and $1 \leq j \leq n_i$, where n_i denotes the number of observations in the ith bin, so that $n = \sum_{i=1}^{r} n_i$. We follow the notation of Paper I and introduce the auxiliary statistics s_0^2, s_1^2, and s_2^2, such that

$$d_0 s_0^2 = \sum_{ij} (x_{ij} - \bar{x})^2 \quad (1)$$

$$d_1 s_1^2 = \sum_{i} n_i(\bar{x}_i - \bar{x})^2 \quad (2)$$

$$d_2 s_2^2 = \sum_{ij} (x_{ij} - \bar{x}_i)^2 \quad (3)$$

$$n_i \bar{x}_i = \sum_{j=1}^{n_i} x_{ij} \quad (4)$$

$$n\bar{x} = \sum_{i=1}^{r} n_i \bar{x}_i = \sum_{ij} x_{ij} , \quad (5)$$

where $d_0 = n - 1$, $d_1 = r - 1$, and $d_2 = n - r$ give the corresponding numbers of the degrees of freedom. For arbitrary but not necessarily random x, these statistics satisfy the algebraic relation (Fisz 1963)

$$d_0 s_0^2 = d_1 s_1^2 + d_2 s_2^2 , \quad (6)$$

and therefor not all of the statistic are independent.

As shown in Paper I, the PDM, Whittaker-Robinson (WR, S78), and AOV periodogram statistics can be expressed in terms of s_k^2,

$$\Theta_{\rm PDM} = \frac{s_2^2}{s_0^2} \, \Theta_{\rm WR} = 1 - \frac{s_1^2}{s_0^2} \, \Theta_{\rm AOV} = \frac{s_1^2}{s_2^2} . \quad (7)$$

The one-to-one correspondence of the Θ statistics follows from equation (6):

$$U \equiv \frac{d_2}{d_0} \Theta_{\rm PDM} = \frac{d_2}{d_2 + d_1 \Theta_{\rm AOV}} = \frac{d_2}{d_0} + \frac{d_1}{d_0} \Theta_{\rm WR} . \quad (8)$$

To assess the significance of a period detected using a statistic, one considers the probability distribution of the statistic assuming that observations are purely random noise. The assumption that observations are noise is called the null hypothesis, H_0. We assume that the noise is white, i.e., that its values are mutually independent, and that the distribution is Gaussian, $N(0, 1)$. Then by virtue of Fisher's Lemma $d_1 s_1^2$ and $d_2 s_2^2$ are independent random variables with $\chi^2(d_1)$ and $\chi^2(d_2)$ distributions (Fisz 1963). The ratio $\Theta_{\rm AOV} = s_1^2/s_2^2$ has a Fisher-Snedecor distribution (Paper I) given by

$$P(\Theta_{\rm AOV} > f) = 1 - F(d_1, d_2; f) . \quad (9)$$

The Fisher-Snedecor and beta distributions are closely related (see Theorem 1.2.3 in Bickel & Doksum 1977; eq. 26.6.2 of Abramovitz & Stegun 1971);

$$1 - F(d_1, d_2; \Theta_{\rm AOV}) = I\left(\frac{d_2}{2}, \frac{d_1}{2}; \frac{d_2}{d_2 + d_1 \Theta_{\rm AOV}}\right), \quad (10)$$

where I denotes the regularized incomplete beta function, and I and F are the beta and Fisher-Snedecor cumulative distributions, respectively. The formulae for the calculation of $I_x(a, b)$ and their implementation in the code have been published by Abramovitz & Stegun (1971) and Press et al. (1986), respectively. A comparison of equations (8) and (10) reveals the identity of U and the beta random variable; hence,

$$P\left(\Theta_{\rm PDM^*} < \frac{d_0}{d_2} u\right) = P\left[\Theta_{\rm WR} < \frac{d_0}{d_1}\left(u - \frac{d_2}{d_0}\right)\right] = I\left(\frac{d_2}{2}, \frac{d_1}{2}; u\right) . \quad (11)$$

To avoid confusion, the incorrect Fisher-Snedecor distribution (S78) for the PDM statistic and the new correct beta distribution (eq. [11]), are denoted as P(PDM) and P(PDM*), respectively. The difference affects only the distributions. The definition of the $\Theta_{\rm PDM}$ statistic remains the same for both cases, so we drop the $*$ for Θ. By simply changing the notation of the classical statistical formulae, we obtain the probability distributions of the PDM* and WR periodograms (S78). The distributions are exact for an arbitrary number of bins and observations r and n, respectively, such that $1 < r < n$.

3. PROPERTIES OF THE PDM* DISTRIBUTION

Since $0 < U < 1$, the whole PDM* distribution is contained in the finite interval $0 \leq \Theta_{\rm PDM} \leq d_0/d_2$. From the standard formulae for the beta distribution, rescaled by the factor d_2/d_0 (eq. [11]), we obtain the probability density and the moments of the PDM* distribution for the hypothesis H_0 such that

$$dP^*_{\rm PDM} = \frac{(d_2/d_0)^{d_2/2}\Theta^{d_2/2-1}[1 - (d_2/d_0)\Theta]^{d_1/2-1}\, d\Theta}{B(d_2/2, d_1/2)} , \quad (12)$$

$$\mu'_k = \left(\frac{d_0}{d_2}\right)^k \frac{\Gamma(d_0/2)\Gamma(d_2/2 + k)}{\Gamma(d_2/2)\Gamma(d_0/2 + k)} \quad (13)$$

$$= \prod_{l=0}^{k-1} \frac{1 + 2l/d_2}{1 + 2l/d_0} , \quad (14)$$

where $B(p, \; q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$ is the complete beta function and $\mu'_k = \int \Theta^k \, dP^*$ is obtained by the change of the integration variable to U. The expected value, variance, skew, and kurtosis are given by

$$E^*\{\Theta_{\rm PDM}\} = 1 \quad (15)$$

$$\text{var}^* \, \{\Theta_{\rm PDM}\} = \frac{2(r - 1)}{(n + 1)(n - r)} \quad (16)$$

$$\gamma_1^* = - \frac{2(n - 2r + 1)\sqrt{2(n + 1)}}{(n + 3)\sqrt{(r - 1)(n - r)}} \quad (17)$$

$$\gamma_2^* = \frac{12[n^3 - n^2(5r - 4)]}{(n + 3)(n + 5)(r - 1)(n - r)} + \frac{12[n(5r^2 - 12r + 6) + 7r^2 - 7r + 1]}{(n + 3)(n + 5)(r - 1)(n - r)} . \quad (18)$$

The asymptotic distribution for $1 \ll r \ll n \rightarrow \infty$ is $N[1, 2(r - 1)/n^2]$. This agrees with the result derived in Paper I for PDM*. For the small values of r in use, the asymptotic normality is never reached. Equation (18) demonstrates that no Gaussian limit holds for small r, since the kurtosis

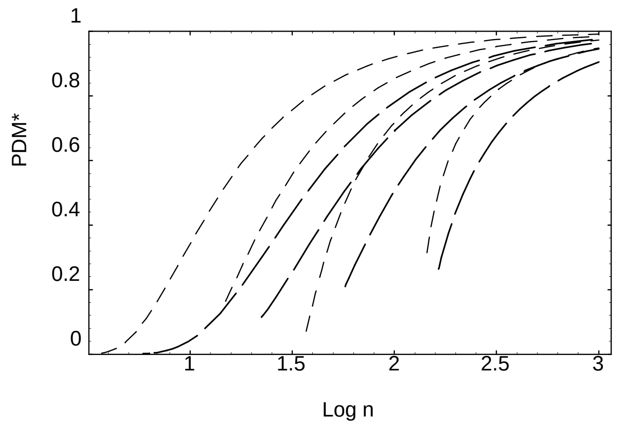

FIG. 1.—Critical values of the $\Theta_{\rm PDM*}$ statistic at 3 σ and 5 σ confidence levels (*short-dashed and long-dashed lines*, respectively) are plotted against the sample sizes n. Curves from left to right correspond to the number of bins, $r = 2$, 8, 32, and 128. The curves for $r = 128$ lie within 5 of their asymptotic limit for $N[1, 2(r-1)/n^2]$.

$\gamma_2 \to 12/(r-1) > 0$. In Figure 1, we plot the critical values of the PDM* statistic for the range of parameters where no normality limit applies. This plot may be used to interpret the significance of features in published PDM periodograms. Depending on the problem, these critical probabilities may or may not require bandwidth correction (see § 1).

By rescaling and shifting the beta distribution in equation (11), we find that the whole WR distribution is contained in the interval $-d_2/d_1 \leq \Theta_{\rm WR} \leq 1$. Its expected value and variance are

$$E\{\Theta_{\rm WR}\} = 0 \tag{19}$$

$$\text{var}\,\{\Theta_{\rm WR}\} = \frac{2(n-1)^2}{(n+1)(n-r)(r-1)} \,. \tag{20}$$

For a given n and r, the WR statistic is equivalent to the PDM* and AOV statistics. Its possible change of sign poses a mild inconvenience. The beta statistic, $d_1\, s_1^2/d_0\, s_0^2$, is more convenient in that respect. However, WR requires fewer calculations than AOV and PDM, and combined with equation (8) it may constitute a more efficient way to compute AOV and PDM.

4. OLD VS. CORRECT PDM DISTRIBUTION

The distribution claimed for PDM was (S78):

$$P(\Theta_{\rm PDM} > c_3) = F(d_2, d_0; c_3) \tag{21}$$

$$E\{\Theta_{\rm PDM}\} = \frac{n-1}{n-3} \tag{22}$$

$$\text{var}\,\{\Theta_{\rm PDM}\} = \frac{2(n-1)^2(2n-r-3)}{(n-r)(n-3)^2(n-5)} \,. \tag{23}$$

For comparison, in Figure 2 we plot this claimed F distribution and the actually valid beta distributions (eqs. [21] and [11], respectively). The families of curves are plotted for four phase bins, $r = 4$, and for a range of sample sizes n.

The F distributions are flat, with positive skew, extending from 0 to ∞, and approach normal for large n. The beta distributions extend over a finite interval of Θ, are concentrated around 1 and retain a negative skew for arbitrarily large n. Their variances diverge even more with growing sample size $n \to \infty$. At this point, the ratio var* $\{\Theta_{\rm PDM}\}$/var $\{\Theta_{\rm PDM}\} \to (r-1)/2n \ll 1$. Clearly, the F distribution (eq. [21]) is incorrect and may not serve even as an approximation of the true PDM* distribution. The most undesirable effect of the F distribution was in decreasing the significance of detected signals. It now seems plausible that the deficiencies attributed in the past to the PDM method were in fact related to an incorrect F distribution.

5. AN EXAMPLE

The inconsistency between the incorrect and correct distributions, PDM and PDM*, can be illustrated by means of a simple simulation. For this purpose we replace the true values of $n = 49$ observations of BK Cen discussed by S78 with simulated Gaussian noise. The PDM periodograms are calculated following the prescription of S78, with $r = 5$ bins and $n_c = 1$ coverage. The whole procedure is repeated 1000 times. In this way, for a given frequency we obtained a sample of 1000 independent values of the $\Theta_{\rm PDM}$ statistic. Estimates of the sample mean and variance were calculated in the usual way for each frequency (cf. eq. (1) of S78). These estimates may be compared with the moments of the PDM* and PDM distributions, equations (15) and (16), (22) and (23), respectively. To facilitate the comparison, the esti-

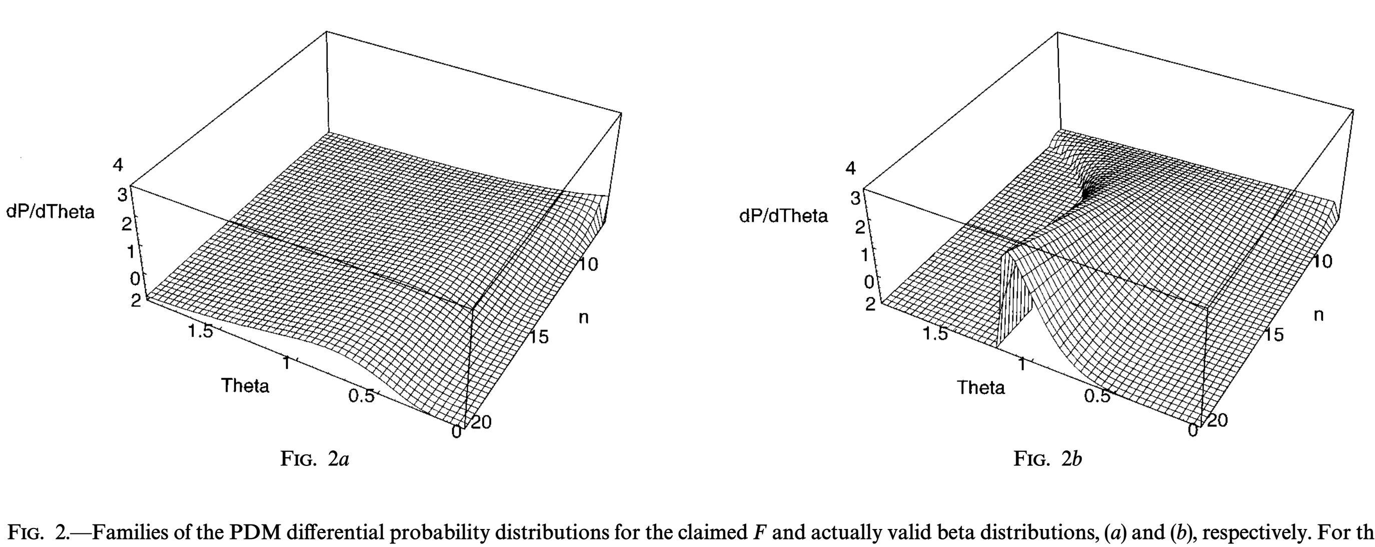

FIG. 2*a*

FIG. 2*b*

FIG. 2.—Families of the PDM differential probability distributions for the claimed F and actually valid beta distributions, (*a*) and (*b*), respectively. For the fixed number of bins, $r = 4$, the families cover a range of sample sizes n.

TABLE 1

Moments of the PDM Statistic

	Sample Moments	
	Mean	Variance
Simulated, $n_c = 1$	1.00191	0.00347
Standard deviation	0.00528	0.00060
PDM* (this paper)	1.00000	0.00327
PDM (S78)	1.04348	0.09596
Simulated, $n_c = 2$	1.00216	0.00242
Standard deviation	0.00510	0.00044

mated sample mean and variances were averaged over frequencies. The average values and standard deviations of individual samples are listed in Table 1 with their corresponding theoretical values for PDM* and PDM distributions. Interpretation of these results requires some care. Owing to power leakage and aliasing, samples corresponding to different frequencies may be correlated, and the usual relation of the individual standard deviation and the standard deviation of an average does not hold in general. Therefore, Table 1 lists the standard deviation for a single-frequency sample, rather than for the frequency average value. Clearly, the distribution of simulated values of the PDM statistic is inconsistent with the PDM distribution of S78. Both mean value and variance of the simulated samples are inconsistent with S78, differing by many standard deviations. The effect of changing from a PDM to a PDM* distribution for this example is comparable to the change from the PDM method to the AOV method for the same sample considered by S78 and in Paper I.

Let us turn our attention to coverages. Coverages are defined by S78 as sets of phase bins differing from each other by their phase offset. The number of bins in each coverage, r, is the same. It was assumed by S78 that the effect of the coverages on the probability is the same as the effect of multiplying the number of phase bins. This assumption is not correct, as an increased number of coverages does not decrease the magnitude of residuals. The coverages correspond to different step functions shifted in phase. Their steps remain coarse for large numbers of coverages. The variance of the PDM statistics is expected to grow with the number of phase bins (eq. [16]). Inspection of Table 1 reveals that the variance of the simulated distribution decreases with n_c, growing from 1 to 2. The decrease is approximately proportional to a factor of $n_c^{-1/2}$. Such a decrease would arise for averaging n_c independent values of $\Theta_{\rm PDM}$. Further simulations reveal that the proportionality breaks down for large n_c as the total number of bins, $n_c r$, approaches the number of observations, n, and Θ statistics computed for different coverages become severely correlated. We conclude by recommending against the use of multiple coverages, as a corresponding probability distribution remains unknown.

6. DISCUSSION

The correct distribution of the PDM* periodogram turns out to be a beta distribution, not an F distribution as originally claimed. The dramatic difference of the two distributions for the same periodogram and parameters are illustrated in Figure 2. The old results of the PDM method using an F statistic are incorrect and should be reanalyzed using a beta statistic. For this purpose, in Figure 1 we plot the critical values of the PDM* statistic for a range of parameters. The plot may be used to interpret the significance of features in published PDM periodograms. Generally, features discovered in the PDM* periodogram are more significant than was previously thought. For the same data and phase binning, the results obtained using the PDM* statistic and the AOV statistic (Paper I) are the same. Depending on the problem, the critical PDM* probabilities may or may not require a bandwidth correction (see § 1).

The PDM*, WR, and AOV methods (S78, Paper I) are fully equivalent. Their differences reduce to simply a change of variables. In the present paper, we demonstrate that these statistics are equally useful in practice by deriving the analytical distributions for PDM* and WR. The distributions are exact for an arbitrary number of bins r and observations n, such that $1 < r < n$. In order to compute each of the statistics, it suffices to compute just one sum, say s_1^2 for each trial frequency, and to use equation (6) with $s_0^2 = \text{const}$. The choice of the PDM*, WR, or AOV statistic is a matter of taste, whether one prefers the "emission" or "absorption" lines in the *temporal spectrum* (introduced by E. Nather). Since AOV, PDM, and WR statistics are all based on the same signal model, namely the step function, their equivalence is understandable. For a given sampling, binning, and weighting pattern, the equivalence in fact extends to *all* methods relying on phase binning and variance estimates resembling χ^2. In view of the equivalence of the statistics, and our discussions in Paper I and in Schwarzenberg-Czerny (1991), the sensitivity of the period detection depends on the consistency of the underlying theoretical signal model with observations, and not on the statistics used in the periodograms. On the one hand, it is important that the model use sufficient bins, harmonics, etc. to match its resolution to the characteristic features in the signal. On the other hand, it is important to avoiding fitting the noise with an excessive number of parameters. The more frequencies analyzed, the more likely is the occurrence of a good fit just by chance, i.e., the less significant the detection. However, a fraction of the frequencies do not count in the game, owing to the effects of oversampling and aliasing (cf. § 1).

Phase binning was originally introduced to raise the sensitivity for narrow pulses, for which the power spectrum is rather insensitive. However, the step function, involved implicitly in phase binning, does not fit well with real, smooth signals, and its use causes uneven phase sensitivity. The Fourier series is a much better statistical model in that sense. The invention of the high-performance Fourier series method, based on the finite-step orthogonal projection (Schwarzenberg-Czerny 1996), has considerably weakened the original cost-saving motivation for using phase binning for nonsinusoidal signals. The algorithm spends as much time per harmonic as the ordinary discrete Fourier transform (DFT) does for the power spectrum.

The author's research was supported by KBN grant 2 P304 002 005.

REFERENCES

Abramovitz, M., & Stegun, I. 1971, Handbook of Mathematical Functions (N.Y.: Dover)

Bickel, P. J., & Doksum, K. A. 1977, Mathematical Statistics (San Francisco: Holden-Day)

Davies, S. R. 1990, MNRAS, 244, 93

Eadie, W. T., Drijard, D., James, F. E., Roos, M., & Sadoulet, B. 1971, Statistical Methods in Experimental Physics (Amsterdam: North-Holland)

Fisz, M. 1963, Probability Theory and Mathematical Statistics (New York: Wiley)

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

Lafler, J., & Kinman, T. D. 1965, ApJS, 11, 216

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes (Cambridge: Cambridge Univ. Press)

Schwarzenberg-Czerny, A. 1989, MNRAS, 241, 153 (Paper I)

———. 1991, MNRAS, 253, 198

———. 1996, ApJ, 460, L107

Stellingwerf, R. F. 1978, ApJ, 224, 953 (S78)

van der Klis, M. 1989, in Timing Neutron Stars, ed. H. Ögelman & E. van den Heuvel (Dordrecht: Kluwer), 27

Whittaker, E. T., & Robinson, G. 1926, The Calculus of Observations (London: Blackie) (WR)