

# Detection of Periodicity Based on Independence Tests – II. Improved Serial Independence Measure

Shay Zucker<sup>\*</sup>

*Dept. of Geosciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel Aviv 6997801, Israel*

7 January 2016

## ABSTRACT

We introduce an improvement to a periodicity metric we have introduced in a previous paper. We improve on the Hoeffding-test periodicity metric, using the Blum–Kiefer–Rosenblatt (BKR) test. Besides a consistent improvement over the Hoeffding-test approach, the BKR approach turns out to perform superbly when applied to very short time series of sawtooth-like shapes. The expected astronomical implications are much more detections of RR-Lyrae stars and Cepheids in sparse photometric databases, and of eccentric Keplerian radial-velocity (RV) curves, such as those of exoplanets in RV surveys.

**Key words:** methods: data analysis – methods: statistical – binaries: spectroscopic – stars: individual: HIP 101453 – stars: variables: RR Lyrae – stars: variables: Cepheids

## 1 INTRODUCTION

In a recent paper (Zucker 2015, ; hereafter Paper I) we have introduced a new non-parametric approach to the detection of periodicities in sparse datasets. The new approach follows the logic of string-length techniques (e.g. Lafler & Kinman 1965; Clarke 2002) and quantifies the dependence between consecutive phase-folded samples, for every trial period. In Paper I we have shown that usually the classical string-length techniques effectively test solely for linear dependence between consecutive samples. On the other hand, the Hoeffding-test approach we have presented there, tests for general dependencies, not necessarily linear. Paper I showed that for two kinds of periodic signals (sawtooth signal and eccentric spectroscopic binary (SB) radial-velocity (RV) curve), our proposed new approach performed better than the conventional techniques.

Inspired by the successful simulations presented in Paper I, we embarked on a wider study to come up with new and improved periodicity metrics, similarly based on dependence measures. The current Letter represents a first step in that direction, a step we have already alluded to in the Discussion of Paper I.

Section 2 describes the details of the modification we propose to the Hoeffding test, Section 3 presents the simulations we have performed in order to test its performance, and in Section 4 we show a test of our new metric on real-life data. In Section 5 we discuss our findings.

## 2 BLUM-KIEFER-ROSENBLATT TEST

The best performing method in Paper I was based on the Hoeffding test. Wassily Hoeffding first proposed it in 1948 (Hoeffding 1948) as a test of independence between two random variables. Essentially it estimates a measure of deviation of the joint empirical distribution function from a distribution that assumes independence. The measure of deviation that Hoeffding used was the so-called Cramér–von Mises criterion for distance between distributions (Cramér 1928; von Mises 1928).

Let us denote by  $G_1$  and  $G_2$  the cumulative distribution functions of the two random variables, and by  $G_{12}$  their joint cumulative distribution function. Then, independence of the two variables would mean  $G_{12} = G_1 G_2$ . Applying Cramér–von Mises criterion for distance between distributions, Hoeffding defined his namesake statistic  $D$  by:

$$D = \int (G_{12} - G_1 G_2)^2 dG_{12} . \quad (1)$$

In estimating  $D$  using the empirical data, we use the empirical distribution functions, determined only by the observed values. This is somewhat reminiscent of the Kolmogorov–Smirnov philosophy, which is popular among astronomers (e.g. Babu & Feigelson 2006). The above definition eventually results in the formulae presented in Paper I.

Blum, Kiefer & Rosenblatt (1961) introduced a new version of the Hoeffding test. They showed that the two tests were equivalent in large samples, but the new test was easier to compute, and also more naturally amenable to generalization to more than two variables. The fundamental definition of their statistic is:

$$B = \int (G_{12} - G_1 G_2)^2 dG_1 dG_2 . \quad (2)$$

While the difference seems to be minute and maybe even insignifi-

\* E-mail: shayz@post.tau.ac.il

cant, the change in the resulting computing formula is not negligible.

Following [Paper I](#), let us denote the phase-folded data by  $x_i$  ( $i = 1, \dots, N$ ) where the index  $i$  reflects the order after the phase folding. In order to make sure our calculations are not affected by the arbitrary zero-phase choice, we also define  $x_{N+1} \equiv x_1$ . Let us further denote by  $R_i$  the phase-folded rank values, such that  $R_i = 1$  means that  $x_i$  is the smallest value. Let us also define the 'bivariate rank'  $c_i$ , as the number of pairs  $(x_j, x_{j+1})$  for which both  $x_j \leq x_i$  and  $x_{j+1} \leq x_{i+1}$ . Note that the existence of the pair  $(x_N, x_{N+1}) = (x_N, x_1)$  accounts for the cyclic wraparound and renders the whole procedure independent of the arbitrary choice of phase.

Now we can define our dependence measure by the formula:

$$B = N^{-4} \sum_{i=1}^N (Nc_i - R_i R_{i+1})^2, \quad (3)$$

which can easily be derived from eq. (5.2) in [Blum et al. \(1961\)](#). The above expression is clearly much simpler than the parallel expressions in eqs. 9–12 in [Paper I](#).

### 3 PERFORMANCE TESTING

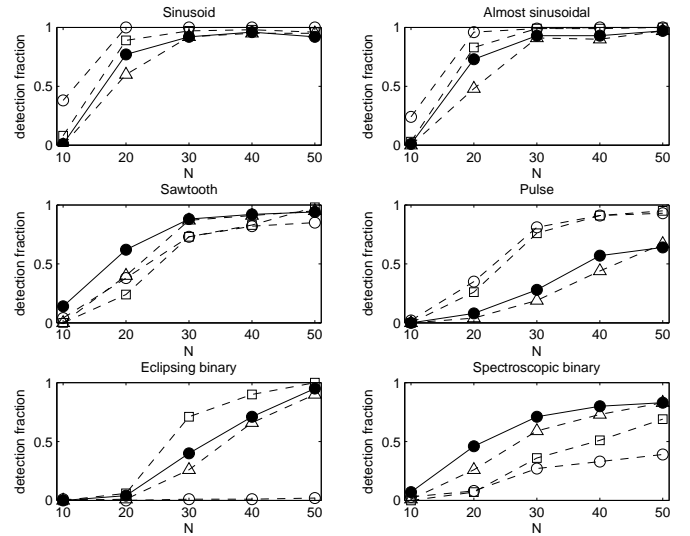
Following the same path as in [Paper I](#), we performed simulations in which we randomly drew a sparse set of sampling times, from a total baseline spanning 1000 time units ('days'). Then we used those times to sample some periodic function with a period of two days, and added white Gaussian noise with a prescribed signal-to-noise ratio (SNR). We tested the same set of six periodic functions we tested in [Paper I](#): sinusoidal, almost sinusoidal, sawtooth, pulse wave, eclipsing-binary light curve and eccentric SB RV curve.

Unlike the approach we followed in [Paper I](#), we have decided this time to use a much simpler and intuitive performance measure: in each tested configuration of signal shape,  $N$  and SNR, we simply counted the number of simulations in which the best score was attained exactly in the correct known period, thus obtaining the 'detection fraction'. We calculated the periodicity metrics for a frequency grid that spanned the range  $10^{-4}$ – $1$  day $^{-1}$ , with steps of  $10^{-4}$  day $^{-1}$ . Thus, counting the fraction of cases with the correct period actually meant a frequency error which was smaller than  $10^{-4}$  day $^{-1}$ .

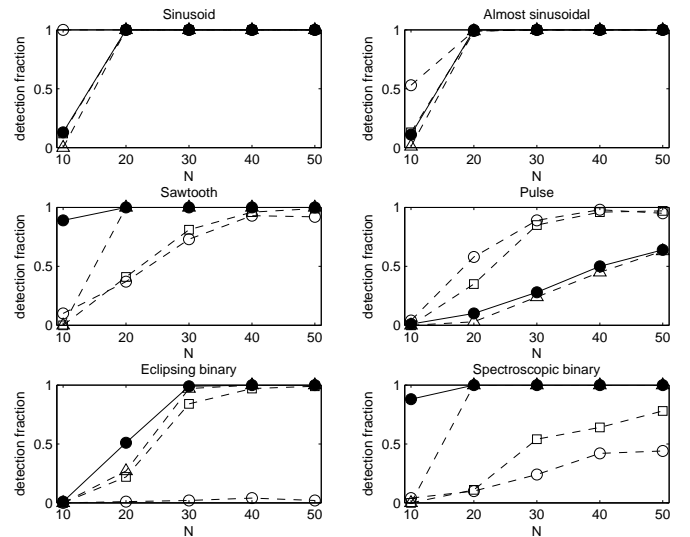
We compared the performance of this Blum-Kiefer-Rosenblatt-test (BKR) periodicity metric to the Hoeffding-test metric we had introduced in [Paper I](#), and also to the two 'traditional' techniques of Generalized Lomb-Scargle ([Lomb 1976](#); [Scargle 1982](#); [Zechmeister & Kürster 2009](#)) and von-Neumann ratio ([von Neumann et al. 1941](#)) as a representative of the string-length techniques ([Clarke 2002](#); [Lafier & Kinman 1965](#)).

[Fig. 1](#) examines the dependence of the detection performance on the number of samples in the time series, for low-SNR. We held the SNR fixed at 3 while we varied the sample size  $N$ . The main feature apparent from examining the Figure is that the new BKR approach constitutes an improvement over the Hoeffding-test approach we introduced in [Paper I](#). In cases where the traditional techniques performed better (e.g., pulse-wave signal shape), they usually performed also better than BKR, and vice versa.

[Fig. 2](#) presents the same test for the case of a high SNR. We fixed the SNR at a value of 100 and repeated the same exercise. In this situation the BKR was again improving over the Hoeffding-test approach, which also meant it performed significantly better than the traditional approaches in the cases of sawtooth-shaped signal



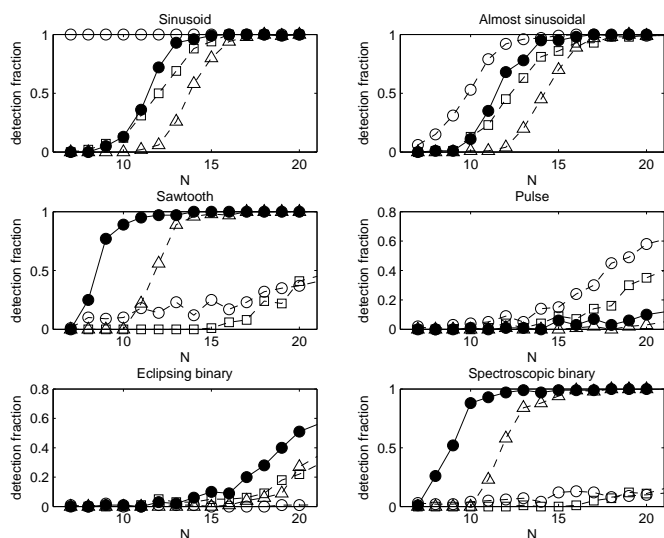
**Figure 1.** Detection fraction as a function of sample size for time series with SNR of 3. The detection fraction is estimated based on 100 simulated light curves. Legend: empty circles, dashed line – Generalized Lomb–Scargle periodogram; empty squares, dashed line – von-Neumann Ratio; empty upward-pointing triangles, dashed line – Hoeffding test; filled circles, solid line – Blum-Kiefer-Rosenblatt test.



**Figure 2.** Detection fraction as a function of sample size for time series with SNR of 100. The detection fraction is estimated based on 100 simulated light curves. Legend: empty circles, dashed line – Generalized Lomb–Scargle periodogram; empty squares, dashed line – von-Neumann Ratio; empty upward-pointing triangles, dashed line – Hoeffding test; filled circles, solid line – Blum-Kiefer-Rosenblatt test.

and eccentric SB RV curve). However, specifically in those two cases, another trend emerged: the BKR method seemed to have an exceedingly better performance for datasets with very few samples ( $N = 10$ ).

In order to make sure this result was not spurious or even erroneous, we have decided to examine this regime more closely. We repeated the simulations with SNR = 100 for a finer grid of  $N$ ,



**Figure 3.** Detection fraction as a function of sample size for time series with SNR of 100, focusing on small  $N$ . The detection fraction is estimated based on 100 simulated light curves. Legend: empty circles, dashed line – Generalized Lomb–Scargle periodogram; empty squares, dashed line – von-Neumann Ratio; empty upward-pointing triangles, dashed line – Hoeffding test; filled circles, solid line – Blum-Kiefer-Rosenblatt test.

namely, for all  $N$  from 7 to 20. Fig. 3, which focuses on this range, shows the very convincing result of a gradual increase of the performance in most cases. Specifically in the cases of the sawtooth signal and the eccentric SB RV curve, The pace of that increase for the BKR method is much faster than that of the other techniques, including the Hoeffding-test approach. It seemed that already when there were 9 (!) samples, the BKR approach had very good chances to detect the periodicity.

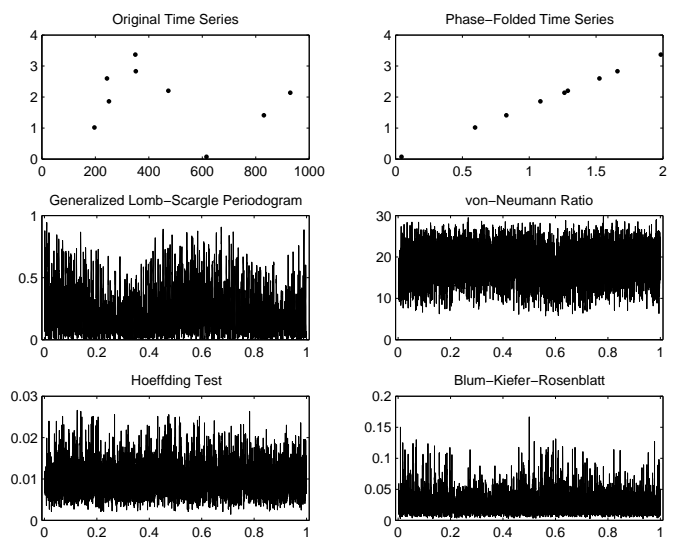
Figs. 4, 5 and 6 show selected concrete examples of cases with a sawtooth signal shape and only 9 samples, in which the performance of BKR was definitely superior over the three other techniques we tested. Those cases were indeed the majority of the simulated cases.

#### 4 REAL-LIFE TEST

We set out to test our new technique in a real-life situation that could potentially emphasize its advantages and allow them to materialize. To this end we chose to apply it to short *Hipparcos* lightcurves. Our sample consisted of *Hipparcos* targets with at most 30 samples in their *Hipparcos* Epoch Photometry Annex entry. We considered only samples which were fully accepted by at least one of *Hipparcos*’ FAST and NDAC consortia (ESA 1997). In total there were 51 lightcurves meeting those criteria.

We scanned the selected lightcurves for periodicity using the BKR test, on a frequency grid ranging from  $10^{-4}$  to  $12 \text{ d}^{-1}$  (following Koen & Eyer (2002)), with steps of  $10^{-4} \text{ d}^{-1}$ . We searched for targets whose peak in the BKR function was significantly prominent. We therefore used the SDE (Signal Detection Efficiency) statistic originally used by Alcock et al. (2000) and Kovács et al. (2002). We chose to single out targets whose SDE statistic was higher than 15. Only one target passed this hurdle – HIP 101453 (also known as CH Aql).

Fig. 7 shows the lightcurve of HIP 101453, as well as its phase



**Figure 4.** An example of the results of applying all the examined periodicity detection methods to a sawtooth simulated time-series, with 9 samples, and a SNR of 100. The upper two panels show the time series and its phase-folded version, using the correct period. The other panels show the periodicity metrics calculated for this time series, with self explanatory titles. Note the poor performance of the first three periodicity metrics, compared to the detection by the BKR technique.

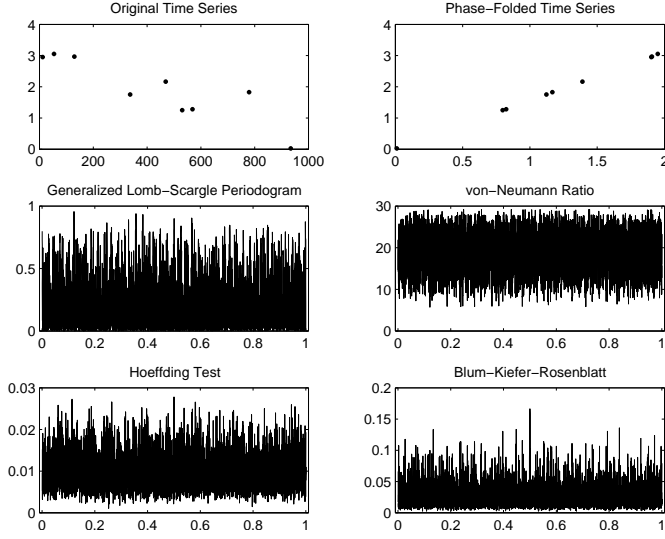
folding using the resulting period, the BKR function and the GLS periodogram we calculated for comparison. The prominence of the BKR peak at a frequency of  $2.5696 \text{ d}^{-1}$  is evident, and indeed the corresponding SDE value is 17.5. The phase-folded lightcurve demonstrates an obvious periodicity. On the other hand, The GLS periodogram shows no hint of a statistically significant periodicity detection.

To complete the picture, this star is a known RR-Lyrae star, with a period of  $0.38918702 \text{ d}$  (Samus et al. 2013), which is consistent with our result –  $0.3891656 \text{ d}$ . The *Hipparcos* catalogue quotes the known period as taken from literature, and adds a comment about the scarcity of the data, which casts doubt about the nature of the periodicity. Therefore, the main catalogue does not quote any period. We show here that by using our new BKR approach, we can assign a high degree of credibility to the periodicity, even with the very scarce *Hipparcos* data.

#### 5 DISCUSSION

This Letter presents an improvement of the serial Hoeffding-test periodicity metric we have presented in Paper I, based on the Blum-Kiefer-Rosenblatt modification of the original Hoeffding test. This new periodicity metric consistently outperforms the Hoeffding-test metric. On top of it, it seems to perform superbly better in sawtooth-like signal shapes, when the number of samples is very small and the SNR is high. This result is in line with the statement of Blum et al. (1961), that their test is asymptotically equivalent to the Hoeffding test for large samples.

This advantage might prove very important for radial-velocity surveys searching for exoplanets. The detection of high-eccentricity Keplerian RV curves is notoriously difficult when based on a small number of samples (Cumming 2004). Another situation in which this feature will prove valuable is the detection of



**Figure 5.** Another example of the results of applying all the examined periodicity detection methods to a sawtooth simulated time-series, with 9 samples, and a SNR of 100. The upper two panels show the time series and its phase-folded version, using the correct period. The other panels show the periodicity metrics calculated for this time series, with self explanatory titles. Note the poor performance of the first three periodicity metrics, compared to the detection by the BKR technique.

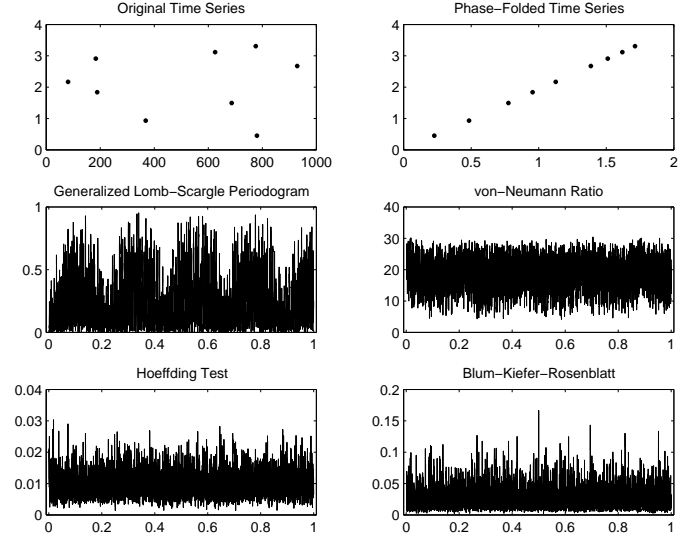
RR-Lyrae stars and Cepheids (whose signal shapes are also essentially sawtooth-like shapes) in sparse datasets such as those of *Hipparcos* and *Gaia* (Eyer et al. 2012). We have provided in Section 4 a demonstration of this potential using *Hipparcos* data, specifically for the case of HIP 101453.

We continue our investigation of harnessing the power of non-parametric independence measures for the purpose of detecting periodicity in extreme circumstances of either poor SNR, small sample sizes, or non-sinusoidal signal shapes. In the meanwhile we also apply our newly developed approaches to existing datasets, both for the sake of testing, but also for detecting previously missed periodic variables. To promote further research and testing of this periodicity metric by the community, we make it available online, in the form of a MATLAB function<sup>1</sup>.

## REFERENCES

- Alcock C. et al. 2000, *ApJ*, 542, 257  
 Babu G. J., Feigelson, E. D. 2006 in Gabriel C., Arviset, C., Ponz D., Enrique S., eds, *ASP Conf. Ser. Vol. 351, Astronomical Data Analysis Software and Systems XV*. Astron. Soc. Pac., San Francisco, p. 127  
 Blum J. R., Kiefer J., Rosenblatt M. 1961, *Ann. Math. Stat.*, 32, 485  
 Clarke D. 2002, *A&A*, 386, 763  
 Cramér H. 1928, *Scand. Actuar. J.*, 11, 13  
 Cumming, A. 2004, *MNRAS*, 354, 1165  
 ESA 1997, *The Hipparcos and Tycho Catalogues*, ESA SP-1200  
 Eyer, L. et al. 2012, *Ap&SS*, 341, 207  
 Hoeffding W. 1948, *Ann. Math. Stat.*, 19, 293  
 Koen, C., & Eyer, L. 2002, *MNRAS*, 331, 45  
 Kovács G., Zucker S., Mazeh T. 2002, *A&A*, 391, 369  
 Lafler J., & Kinman, T. D. 1965, *ApJS*, 11, 216

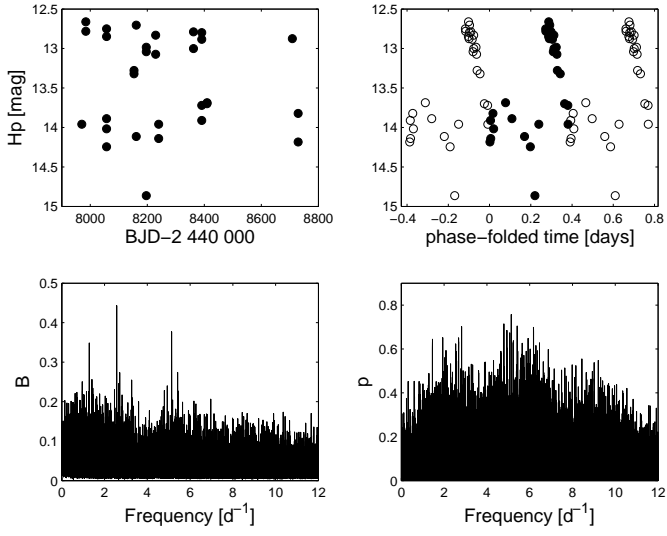
<sup>1</sup> The URL for downloading a MATLAB code to calculate the BKR periodicity metric is <http://www.tau.ac.il/~shayz/BKR.m>



**Figure 6.** Another example of the results of applying all the examined periodicity detection methods to a sawtooth simulated time-series, with 9 samples, and a SNR of 100. The upper two panels show the time series and its phase-folded version, using the correct period. The other panels show the periodicity metrics calculated for this time series, with self explanatory titles. Note the poor performance of the first three periodicity metrics, compared to the detection by the BKR technique.

- Lomb N. R. 1976, *Ap&SS*, 39, 447  
 Samus N. N., Durlевич O. V., Goranskij, V. P., Kazarovets, E. V., Kireeva, N. N., Pastukhova, E. N., Zharova, A. V. 2013, *General Catalog of Variable Stars*, VizieR online data catalog B/gcvs  
 Scargle J. D. 1982, *ApJ*, 263, 835  
 von Mises R. 1928, *Wahrscheinlichkeit, Statistik und Wahrheit*, Julius Springer, Vienna  
 von Neumann J., Kent R. H., Bellinson H. R., Hart B. I. 1941, *Ann. Math. Stat.*, 12, 153  
 Zechmeister, M., & Kürster, M. 2009, *A&A*, 496, 577  
 Zucker S. 2015, *MNRAS*, 449, 2723

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.



**Figure 7.** Detecting the periodicity in the *Hipparcos* lightcurve of HIP 101453. Upper left: the original lightcurve. Upper right: The lightcurve phase-folded using a period of 0.3891656 d. The empty circles represent copies of the original dataset shifted backwards and forwards by one period in order to better visualize the periodicity. Lower left: The BKR periodicity metric. Note the sharp peak at the known frequency. Lower right: Generalized Lomb-Scargle periodogram. Note the absence of any significant peak.