STOCHASTIC DATA IN ASTRONOMY: FOURIER ANALYSIS OF HIGHLY NONUNI-FORM TIME SERIES

A. F. Kholtygin and A. B. Shneiwais   UDC: 524.35: 539.186

The features of the CLEAN algorithm for Fourier analysis of time series with data separated by long pauses are analyzed in detail. Estimates are obtained for the limits of variability of the parameters of harmonic components that can be determined on a specified time grid. This analysis are used to search for harmonic components in the spectral line profile variations of the star λ Ori A (O8III) obtained in 2001 with the 1 m telescope at the Special Astrophysical Observatory of the Russian Academy of Sciences.

Keywords: *stars: early type: line profiles: variability*

1. Introduction

The most efficient way to study the structure of astronomical objects is to make spectral observations over a sufficiently long period. During these observations a large number of spectra are obtained for different times. Sets of spectra of this sort are usually called the dynamic spectrum of the object being studied. The measured flux at frequency ν_j is a random function which takes the random value $F(t_i, \nu_j)$ at a given time t_i. Thus, when analyzing the data obtained over the entire observation period we examine a set of random processes.

This analysis yields information on a specific random process and reveals the constancy or variability of the measured quantity. If the quantity being studied turns out to be variable, then we can ask about the type of variability: is it regular or the changes are random? It is important to keep in mind the statistical nature of the answer to this question, which depends on the accepted level of significance; for one level of significance an answer can be obtained, while for another, not.

When studying random functions (the set of values of $F(t_i, \nu_j)$ at a given time t_i) the problem of clarifying the degree of dependence of the values of this function at different points may arise: are these values independent or does some correlation exist among them? In an analysis of an entire set of time varying random functions (the set of random

Astronomical Institute, St. Petersburg State University, Russia, e-mail: afk@theor1.astro.spbu.ru

processes $F(t_i, \nu_j)$ for fixed values of ν_j) one may formulate the problem of establishing the presence or absence of a dependence of the random processes for different values of the frequency variable ν_j on one another.

Astronomical observations and, in particular, spectral observations, have a number of features which are not typical of observations made under earthbound conditions, so special approaches are required to analyze them. One such feature is their uniqueness and nonreproducibility. Thus, for example, when analyzing astronomical time series we are dealing with a single segment of a series; this makes it difficult to obtain statistical estimates. Most astronomical observations are made at night and the very possibility of making them depends on suitable conditions, the state of the detection apparatus, etc. For this reason, data obtained through astronomical observations are in the form of nonuniform (unevenly spaced) time series.

This article is the first of a series of articles devoted to the analysis of sets of random processes obtained from astronomical observations. The main instrument for this analysis is the methods of mathematical statistics, Fourier and wavelet analysis techniques, and factor and correlation analysis, all of which are fully discussed in the literature [1-5]. We, on the other hand, shall emphasize those details of the analysis which are associated with the above specifics of astronomical observations. In the first article of this series we discuss Fourier analysis techniques for nonuniform time series obtained in the course of spectral observations of stars in early spectral classes.

2. Statement of the problem

Suppose that a series of observations at time t_i yields n values of a random function $F(t_i, x_j)$ with $i=1, ..., m$,, where m is the number of observations that have been made, and $j=1, ..., n$, where x_j is a parameter, such as the average wavelength of a spectral interval within which the radiant flux from the object under study varies. In the case, for example, of spectral observations the $F(t_i, x_j)$ are the measured radiation fluxes at time t_i,[1] arriving from the object within specified spectral intervals.

The measurements are one of the realizations of a two-dimensional random quantity $F(t,x)$ which represent a random process for each value of x. In this article we shall assume that the random quantities $F(t, x_k)$ and $F(t, x_l)$ with $k \neq l$ are independent for arbitrary times t. We shall also assume that the step size for the time grid is the same for all x.

Thus, an analysis of the two-dimensional random quantity $F(t,x)$ reduces to studying n independent time series $F_j(t) \equiv F(t, x_j)$, where $j=1, ..., n$. In the following we shall assume that all the time series are centered, i.e.,

$$\sum_{i=1}^{m} F_j(t_i) = 0, \tag{1}$$

and, also, that any linear trend (if it was present) has been eliminated from the time series. $G_j(t)$ denotes the values of $F_j(t)$ after centering and elimination of any linear trend. For convenience we shall omit the subscript i on the variable t. Let us assume that the function $G_j(t)$ can be represented as a set of harmonic components and white noise. Then

---

[1] In general, observations are made over some finite time interval Δt; however, an observation can usually be attributed to a certain instant of time $\bar{t}$, which might, for example, be taken to be the middle of the interval Δt.

$$G(t) = \sum_{k=1}^{L} A_k \cos(2\pi\nu_k\, t + \varphi_k) + \sigma_N N \,, \tag{2}$$

where A_k, ν_k, and φ_k, respectively, are the amplitudes, frequencies, and phases of the harmonic components, L is the number of harmonics, N is a normally distributed random quantity with zero mathematical expectation and unit dispersion, and σ_N is the standard deviation of the noise component. Let $A_{max} = \max(A_k, k = 1, ..., L)$ be the maximum amplitude of the harmonic components.

We write $\sigma_n = A_{max}/U$, where U is a characteristic of the degree to which the noise component contributes to the experiment signal. $U \gg 1$ corresponds to a small contribution from the noise component.

Two problems can be formulated here. The first is to find the harmonic components at a specified level of significance $\alpha = 1 - q$ $(q \ll 1)$ in real series of observations. The second is the same as the first, but for model time series on a temporal grid determined by the real series with model harmonic components and with the parameters for the harmonic components obtained by solving the first problem. In this case the solution of the model problem should confirm the reliability with which the harmonic components have been found from the real series at the specified level of significance.

3. Fourier analysis of model line profiles

One of the purposes of this article is to construct an optimal algorithm for finding the harmonic components in the experimental signal. The most effective way to detect harmonic components from an initial signal when the time series are nonuniform is the CLEAN algorithm [14]. In this paper we shall use a modified version of this algorithm described in Ref. 3. The Cleanest algorithm [7,8] is also currently in use, but it offers no significant advantages for analyzing highly nonuniform time series.

We shall state the problem in the following way: during an analysis by the CLEAN method of a real time series $G^{obs}(t)$ on a temporal grid determined by the times the observations to be analyzed are made, let the harmonic components ν_k, A_k, and φ_k, with $k = 1, ..., L$, be found. Equation (2) is then used on the same temporal grid to construct a model time series that includes harmonic components with the frequencies, amplitudes, and phases $(\nu_k, A_k, \varphi_k, \quad k = 1, ..., L)$ determined above, which, as we assume can actually be present in the observed time series.

A detailed analysis of the model series is carried out for different values of L and a determination is made of whether the given components can be isolated from the model series for specific values of ν, A, φ or whether the found values of the parameters belong to false peaks in the Fourier spectrum.

In addition, we also solve the problem of how close the parameters ν_k^*, A_k^*, and φ_k^* obtained as a result of the Fourier analysis of the model series are to the corresponding values ν_k, A_k, and φ_k.

3.1. Constructing model time series. Constructing model time series requires selecting a temporal grid, a set of parameters ν_k, A_k, and φ_k in Eq. (2), and a ratio U. In analyzing observations that have already been made, these parameters are determined by the real time grid and parameter set determined by analyzing the observations. For planning future observations with the aim of choosing an optimal strategy for making them, it is appropriate to choose parameters

close to their typical values for real observations.

Based on a study [6] of spectra of several bright O-supergiants obtained on the 1 m telescope at the Special Astrophysical Observatory (SAO), we shall assume that the average exposure time t_{exp} is 10-15 minutes (0.007-0.01 day). The total observation time T_{obs} for the chosen object depends on the time of year, the weather conditions, and its declination and culmination time. For a suitable choice of observation date, T_{obs} is 4-10 hours. Profile variations with a characteristic time of 2-6 hours [6] are studied using observations taken over a number N_{night} of observation nights.

We shall assume initially that the observations are organized in an ideal fashion; that is, on each of the N_{night} nights the observations begin at the same time, so the number $N_{obs} = T_{obs}/t_{exp}$ of observations per night is strictly constant when t_{exp} is held constant for the entire observation period. Thus, on each of the N_{night} nights, N_{obs} observations are made, followed by $N_{gap} = (1 - T_{obs})/t_{exp}$ gaps (for times measured in days).

We shall consider the time of each individual measurement to be the midpoint of the exposure and assume that the time of the midpoint of the first exposure on the first observational night corresponds to time $T = 0$, so the time of observation No. i on observational night j is given by $t_k = (j-1) + (i-1)t_{exp}$, where $k = (j-1)N_{obs} + i$.

In reality, from 15 to 60 spectra of a star can be obtained over a single observation night with a characteristic ratio $\mho = A_{max}/\sigma_N$ for the variable components of the lineshapes in the spectrum of the observed stars of $\mho = 3-6$. The number of observations and their starting time can, in general, vary from night to night. For a more realistic description of the temporal grid for the observations, we shall use the formula

$$t_k = (j-1) + (i-1)t_{exp} + \Delta T_j , \qquad (3)$$

where ΔT_j is the shift in the time of the observations during night No. j relative to the starting time of the observations during the first night.

As an illustration, Fig. 1 shows several temporal grids corresponding to different observation strategies. The time t_{exp} between successive observations has been taken to be 13.1 minutes (≈ 0.0091 day), a typical exposure time for spectral
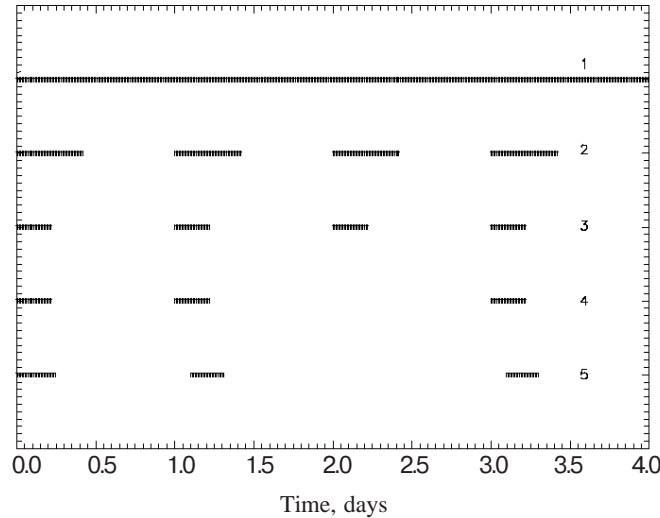

Fig. 1. Temporal grid for the model series. The time is measured in days and the times at which values of the series are obtained are indicated by arrows. The labels correspond to the model sequences 1-5 discussed in Section 3.1.

observations of bright supergiants on the 1 m telescope at the SAO [6].

The upper sequence of time markers (No. 1 in Fig. 1) corresponds to hypothetical around-the-clock observations (e.g., on an extraterrestrial observatory). Such observations might, in principle, be organized on the earth through cooperative observations at several telescopes located at different latitudes; however, organizing observations of this sort is extremely cumbersome and such observations have thus far only been made for about ten objects. (See Ref. 10, for example.) There is also the problem of reducing the observations made on different instruments to a unified system. Time marker sequences 2 and 3 in Fig. 1 correspond to ideally organized earthbound observations. Here we have assumed that observations were made over four days with constant t_{exp}. In real observations, t_{exp} is the sum of the exposure time, as such, plus the array readout time, which we assume to be constant throughout the observation period. For sequence No. 2 of time markers it was assumed that the observations were made for 10 hours, which is, in principle, possible for bright, high-latitude objects, such as αCam during the winter. Here 47 observations can be made during a night or 188 observations over the 4 observation nights.

Time-marker sequence No. 3 corresponds to 6-hour observation periods in each day. Here 25 observations are made per night and 100, over the entire observation period.

In order to account for realistic, practical observations, when bad weather causes the loss of an observation night, we have constructed time sequence No. 4 assuming that no observations were made on the third night.

And, finally, the last time sequence, No. 5, corresponds to real observations of the star λOri made on the 1 m telescope at the SAO in December 2001. Then, 28 observations were made on the first night, 24 on the second, and 23 on the fourth. No observations were made on the third night. The observations on the second and fourth nights began 0.019 and 0.024 days later, respectively, than on the first night.

The frequencies of the periodic components determined by analyzing lineshapes in the spectra of some stars in spectral class O and in early subclasses of spectral class B lie in the range of 0.1-6 d^{-1} with periods $P = 4^{h} \div 10^{d}$ [11,9,10,6,13]. Thus, our problem involves choosing a method for determining the harmonic components of time series with these frequencies specified on the temporal grids shown in Fig. 1.

3.2. Choice of optimal values for the parameters. When determining the optimal values for the parameters ν, A, and φ of the harmonic components of the signals, it is important to make a correct choice for the parameter χ which determines the density of the sample of values for the frequencies used for calculating the refined Fourier spectrum of the signal being analyzed. Larger values of χ correspond to a denser frequency grid. The difference between neighboring values on the grid is $\delta\nu \sim 1/\chi$.

Our analysis shows that for both uniform and nonuniform time series with $A/N > 1$ the parameters ν, A, and φ determined using the CLEAN algorithm depend significantly on the choice of χ. The effect of improvements in the accuracy of these parameters when choosing an optimal value of φ is illustrated in Table 1.

It was assumed that the observations were made over 3.3 days with a step size of 0.00909 days, for a total of 364 values of the test function. The test series on the given time grid is assumed to be a cosine curve with frequencies ranging from 0.2 to 50 with the ratio $\upsilon = 5$, A=1 and $\varphi = 0.0$. The table shows clearly that a suitable choice of χ yields much more accurate values for the parameters than a constant value of $\chi = 5$.

Choice of the optimum value of *g*. The parameter g determines the extent to which a found harmonic is deducted from the "dirty" periodogram. $g = 1$ means that in each step of cleaning up the periodogram the found harmonic is subtracted entirely. For time grids that are close to uniform, choosing $g < 1$ makes it possible to improve the quality

TABLE 1. The Differences Between the Exact Values of the Parameters ν, A, and φ and those determined by the CLEAN method for $\chi = 5$ and an optimal choice of the parameter χ.

ν	$\|\nu - \nu_{calc}\|$		$\|A - A_{calc}\|$		$\|\varphi - \varphi_{calc}\|$	
	$\chi = 5$	$\chi_{optimal}$	$\chi = 5$	$\chi_{optimal}$	$\chi = 5$	$\chi_{optimal}$
0.21	0.032	0.032	0.075	0.075	0.47	0.47
0.51	0.035	0.003	0.029	0.016	0.39	0.05
0.71	0.017	0.003	0.018	0.005	0.17	0.04
1.11	0.010	0.000	0.002	0.002	0.09	0.04
2.20	0.020	0.000	0.008	0.001	0.20	0.03
3.30	0.03	0.000	0.022	0.000	0.33	0.04
4.30	0.00	0.00	0.000	0.000	0.01	0.01

of cleanup of the periodogram [11]. Our calculations showed that for the highly nonuniform series considered in this article, the optimum choice is $g = 1$.

Choice of the optimum value of X_q. X_q determines the threshold for detection of a signal in the noise with a probability of $\alpha = 1 - q$ for uniform and nonuniform time series. To obtain the value of this parameter we simulated white noise for a given time series and studied the maximum counting statistics for the periodogram.

Let us assume that $N \gg 1$ white noise simulations have been carried out and that for m of them the maximum count exceeds X_q. In that case we assume $\alpha = 1 - m/N$.

4. Degree of reliability of the harmonic components determined for the test signal

In the case of highly nonuniform time series the CLEAN algorithm for refining the periodograms does not provide complete confidence that the resultant periodic component of the time series being analyzed is actually present in the series and not a false component. In addition, even if a harmonic component found by analyzing a specific series actually is present in the series, the accuracy with which its parameters are determined may be low.

In order to establish the degree of reliability of the presence of a found harmonic component in the test time series and to estimate the errors in determining its parameters, we shall use the following method. Let us assume that applying the CLEAN algorithm to the time series reveals component with parameters ν, A, and φ. We then construct a sequence of model time series with that fixed value of A and values of ν and φ within the intervals $\nu_{min} \leq \nu_{max}$ and $\varphi_{min} \leq \varphi_{max}$ chosen so as to encompass all the possible values of ν and φ for the given process. In particular, φ is chosen to be within the interval $0 \leq \varphi \leq 2\pi$.

For all values on the (ν, φ) grid a model time series is analyzed using the CLEAN algorithm and the parameters ν^*, A^*, and φ^* are found. then the errors $\Delta\nu = |\nu - \nu^*|$, $\Delta A = |A - A^*|$, and $\Delta\varphi = |\varphi - \varphi^*|$ these model series parameters

are determined.

An *accuracy criterion* for the determination of a given parameter is selected that reduces to choosing the maximum possible deviation of the exact and found value for each of the parameters for the process being analyzed. That is, it is assumed that if the errors in determining the parameters of the found harmonic component ν^*, A^*, and φ^* did not exceed the chosen maximum deviations, then this component is present in the process being analyzed. If, on the other hand, the error in determining the magnitude of even one of the parameters exceeds the maximum possible error, then it is assumed that this harmonic cannot be recovered on the given time grid with sufficient accuracy.

Our preliminary analysis [6] showed that the frequencies of the possible periodic components of the lineshape variations for the time grid shown in Fig. 1 lie within the interval of 0.3-2.0 d^{-1}.

Let us consider three accuracy criteria:

Criterion A: $\left|\nu - \nu^*\right| < 0.1$, $\left|A - A^*\right| < 0.1$, and $\left|\varphi - \varphi^*\right| < 0.1$;

Criterion B: $\left|\nu - \nu^*\right| < 0.2$, $\left|A - A^*\right| < 0.2$, and $\left|\varphi - \varphi^*\right| < 0.2$; and,

Criterion C: $\left|\nu - \nu^*\right| < 0.5$, $\left|A - A^*\right| < 0.5$, and arbitrary $\left|\varphi - \varphi^*\right|$.

The time is taken to be measured in days, frequencies in d^{-1}, and phase φ in radians. The amplitudes of all the model line profiles were assumed to be 1.

Criteria A and B make it possible to evaluate the frequency, amplitude, and phase of an unknown periodic process with a fairly high degree of reliability, while criterion C only indicates that a periodic process is present in a given time series but additional observations are required for a reliable determination of the characteristics of this process.

Using these accuracy criteria for the parameters ν, A and φ, we now introduce the reliability function $R(\nu, \varphi, K)$, defined as

$$R(\nu, \varphi, K) = \begin{cases} 1, \text{criterion } K \text{ is satisfied,} \\ 0, \text{criterion } K \text{ is satisfied,} \end{cases}$$

The value of the reliability function $R(\nu, \varphi, K)$ clearly depends on the choice of accuracy criterion K. The degree of reliability with which an unknown harmonic component is isolated from a time series can be conveniently illustrated with the aid of reliability plots which show the values of $R(\nu, \varphi, K)$ as functions of ν and φ.

Figure 2 is an example of plots of this type for the time grids shown in Fig. 1 and the accuracy criteria A, B, and C. In this figure the ranges of ν and φ within which the reliability function $R(\nu, \varphi, K) = 1$ are indicated in black. The parameters of the model series can be recovered with the specified accuracy within these regions. The figure shows that when the weakest accuracy criterion C is used, the parameters of an harmonic component specified on all the temporal grids can be recovered in the regions with $\nu \geq 2$ for all possible values of φ and for all the time sequences considered here.

At the same time, for a series with large gaps (the middle and right hand reliability plots in the lower series) in the region with $\nu \leq 2$ the possibility of determining ν and φ depends on their specific values. For example, for time sequence 3 there are regions of frequency ν (in particular, for $\nu = 2$) within which the parameters of a harmonic component that is actually present in the time series cannot be recovered for any values of φ.

When the stricter accuracy criteria A and B (upper and middle rows of reliability plots in Fig. 2) are used, the region where the reliability function $R(\nu, \varphi, K)$ equals zero is considerably larger. Even in the case where observations are made
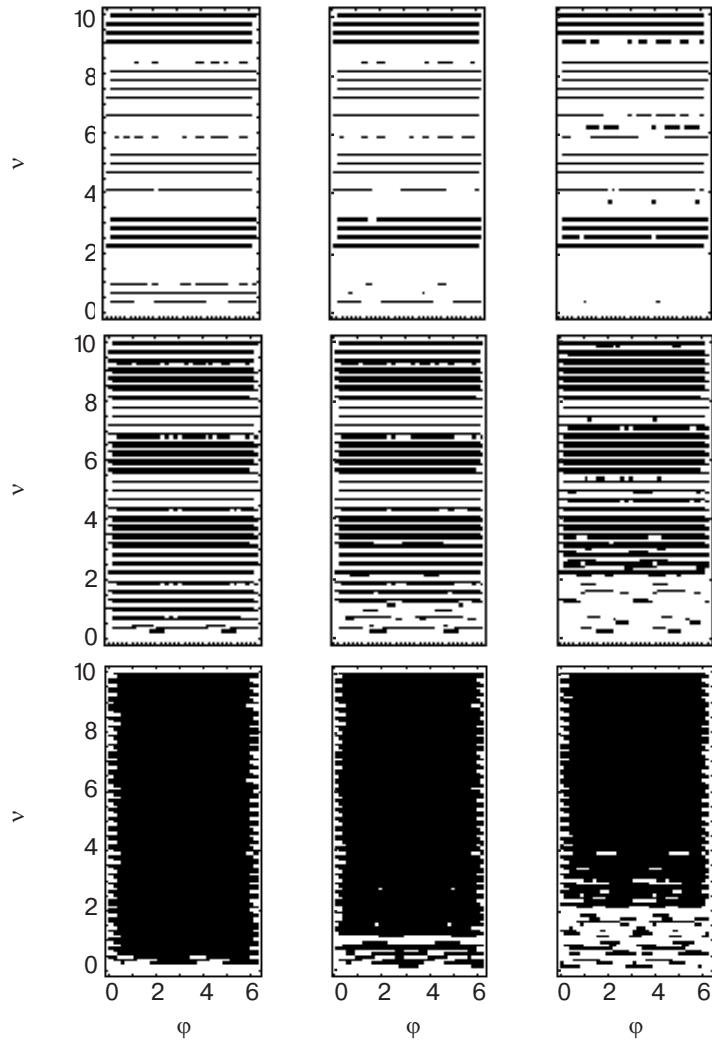
74

Fig. 2. Plots of the reliability with which the parameters ν and φ are determined for accuracy criteria A (upper row), B (middle), and C (lower). In each of these rows, the leftmost plot corresponds to a time sequence 1 (around-the-clock observations for four days), the center plot, to sequence 2 (observations for 10 hours on each of nights), and the rightmost plot, to sequence 4 (observations with the third night omitted). $\upsilon = 6$.

without any gaps, for reliability criterion A (leftmost plot in the upper row of Fig. 2) there is a very large range of values of ν and φ within which an harmonic component of the series is not recoverable in principle with a specified degree of accuracy.

Figure 3 shows more detailed reliability plots for a harmonic signal with the parameters $\nu \in [0,2]$ and $\varphi \in [0,2\pi]$.

This figure shows clearly that for the choice of accuracy criterion A, *recovery of components of a periodic signal that is present in the analyzed time series on a temporal grid with large gaps is possible only for a very narrow region of the parameters ν and φ.*
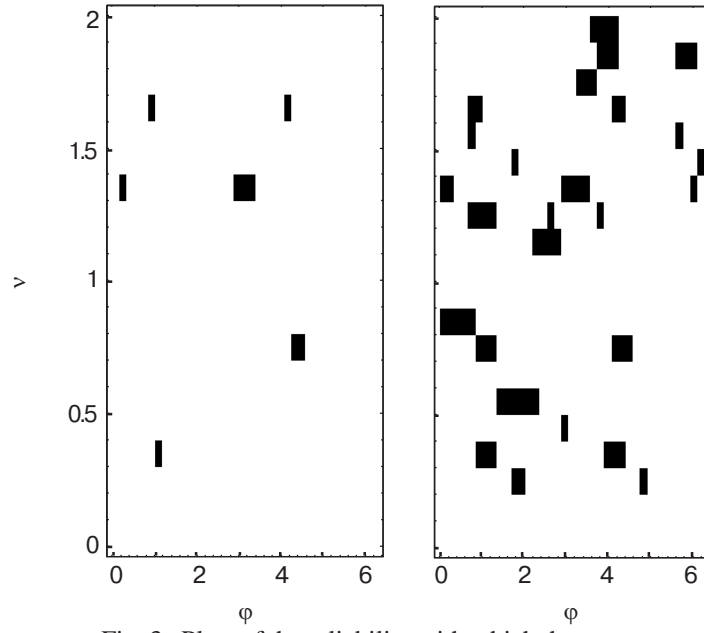
Fig. 3. Plots of the reliability with which the parameters ν and φ are determined for a real time sequence (sequence 5 of Fig. 1). The plot on the left shows the values for reliability criterion A, and on the right, for criterion B.

Going from accuracy criterion A to criterion B increases the region over which the parameters of the harmonic components can be recovered reliably, but even in this case for most of the possible values of ν and φ an harmonic component that is certainly present in the analyzed series cannot be recovered in principle.

5. Search for regular components in the variation of lineshapes from the star λ Ori A

In order to test the method for optimal search for harmonic components in a signal discussed above, we have analyzed the variations in the lineshapes in the spectrum of λ Ori A. The observations of λ Ori A were made on three nights over the period November 29- December 4, 2001, using the 1 m telescope at the SAO.

The observations were made using the CEGS Coude echelle-grating spectrometer on the 1 m telescope at the SAO. A Wright Instruments CCD with a detector size of 1242×1152 pixels. A spectral resolution R = 45000 (0.08 Å/pixel in the Hα region) was obtained in the range λ = 4000-8000 Å for an entrance slit width of 2".

The procedure for processing the CCD images and obtaining the difference spectra is described elsewhere [6]. The wavelength scales were determined taking into account the corrections for the earth's rotation and for its revolution about the sun. A total of 75 spectra were obtained in 2001. The Fourier spectra of the variations in the difference profiles in the spectra of λ Ori A obtained by the above described technique for Fourier analysis of nonuniform time series are shown in Fig. 4.

Taking the results of the preceding analysis and using the reliability plots we have obtained, we can discern the presence of harmonic components with frequencies $\nu_1 \approx 0.5\,d^{-1}$, $\nu_2 \approx 0.75\,d^{-1}$, and $\nu_3 \approx 1.3\,d^{-1}$ in the lineshape
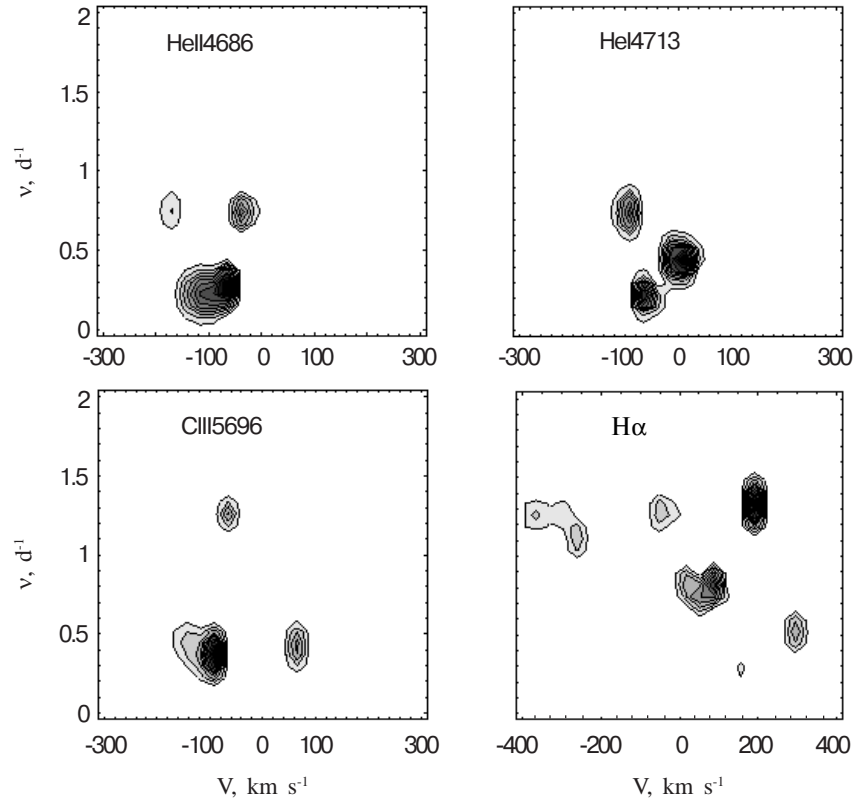
Fig. 4. Fourier spectra of the variations in the difference profiles of the HeII λ4686 Å, HeI λ4713 Å, CIII λ5696 Å, and Hα lines in the frequency range $\nu = 0 - 2\,d^{-1}$. Only those amplitudes of the Fourier spectra corresponding to a significance 1- $q > 0.999$ for the presence of the given variable component are shown. Darker values in these plots correspond to higher amplitudes.

variations for the spectrum of this star; however, we cannot precisely establish the localization of these components within specific wavelength intervals (Doppler velocities), since the analysis of the preceding section shows that, for most phases of a harmonic component, its parameters cannot be obtained for the given temporal grid.

Since the phase of a harmonic component of the variations in the lineshapes varies along the line profile itself [11,12], we may assume that the harmonic components found here are present in other parts of the profile, as well. Our analysis shows that a larger number (200-300) of spectra of this star will be required to identify these components.

6. Conclusions

The preceding analysis yields the following conclusions:
- For nonuniform temporal grids there is a set of frequencies which can be determined (for a given signal/noise

ratio) based on a Fourier analysis of a time series on this grid. Frequencies outside this set cannot be detected.

- For a given temporal grid and a given signal/noise ratio for a harmonic component with a definite frequency ν there is an interval of phases within which the CLEAN algorithm (or modifications of it) can be used to find the parameters A and φ of the component ν with a given accuracy. Outside this phase interval, this problem cannot be solved.

- There is a range of frequencies ν and phases φ within which the parameters A and φ of the harmonic components cannot be found with sufficiently high accuracy (criteria A and B) even for uniform time series.

REFERENCES

1. Z. Brandt, Statistical Methods for Analysis of Observations [Russian translation], Mir, Moscow (1975), p. 87.

2. V. V. Vityazev, Analysis of Uniform Time Series [in Russian], Izd. SPbGU (2001).

3. V. V. Vityazev, Analysis of Nonuniform Time Series [in Russian], Izd. SPbGU (2001), p. 68.

4. J. Dobesi, Ten Lectures on Wavelets [Russian translation], Moscow (2001).

5. Yu. N. Tyurin and A. A. Makarov, Computer Analysis of Data [in Russian], Infra-M, Moscow (2003), p. 544.

6. A. F. Kholtygin, D. N. Monin, A. E. Surkov, and S. N. Fabrika, *Pis'ma Astron. zh*. 29, 208 (2003).

7. J. Foster, *Astron. J*. 109, 1889 (1995).

8. J. Foster, *Astron. J*. 111, 541 (1996).

9. J. A. de Jong, H. F. Henrichs, S. Schrijvers et al., *Astron. Astrophys*. 345, 172 (1999).

10. J. A. de Jong, H. F. Henrichs, L. Kaper et al., *Astron. Astrophys*. 368, 601 (2001).

11. L. Kaper, H. F. Henrichs, A. W. Fullerton et al., *Astron. Astrophys*. 327, 281 (1997).

12. L. Kaper, H. F. Henrichs, J. S. Nichols et al., *Astron. Astrophys*. 344, 231 (1999).

13. C. Neiner, A. M. Hubert, M. Floquet et al., *Astron. Astrophys*. 388, 899 (2002).

14. D. H. Roberts, J. Lehar, and J. W. Dreher, *Astron. J*. 93, 968 (1987).