Analysis of Variance (AoV)

for period search

based on Alex Schwarzenberg-Czerny, 1989, MNRAS, 241, 153

Milena Ratajczak
Statistics Journal Club, 31.03.2022

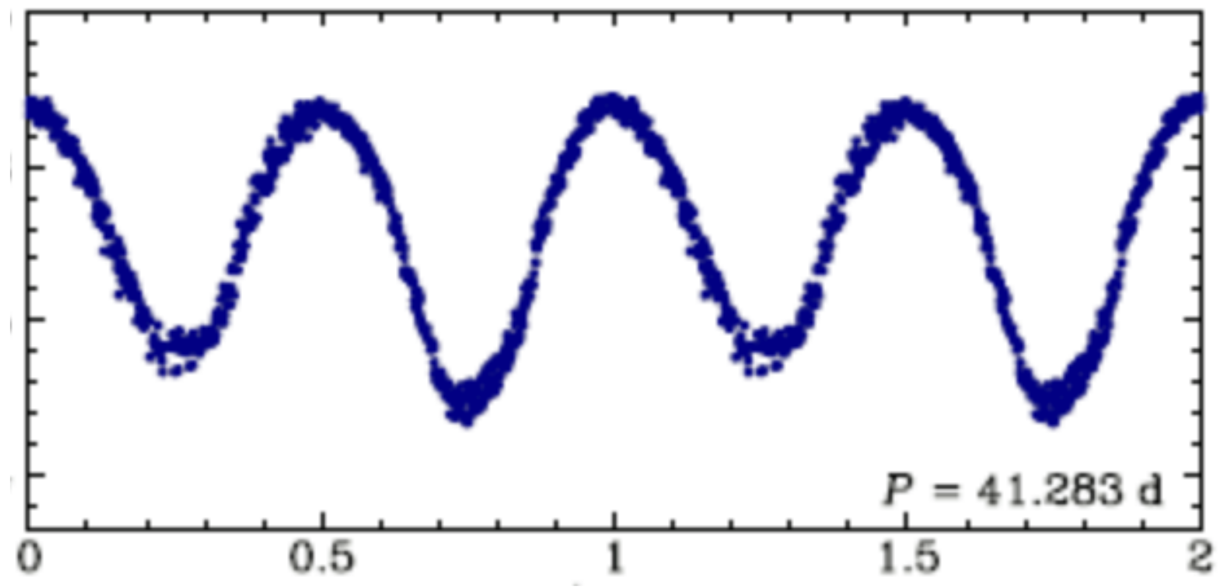
Searching for Periodic Signals in Time Series Data

- Least Squares Sine Fitting
- Discrete Fourier Transform
- Lomb-Scargle Periodogram
- Pre-whitening of Data
- Phase Dispersion Minimisation
- Analysis of Variance

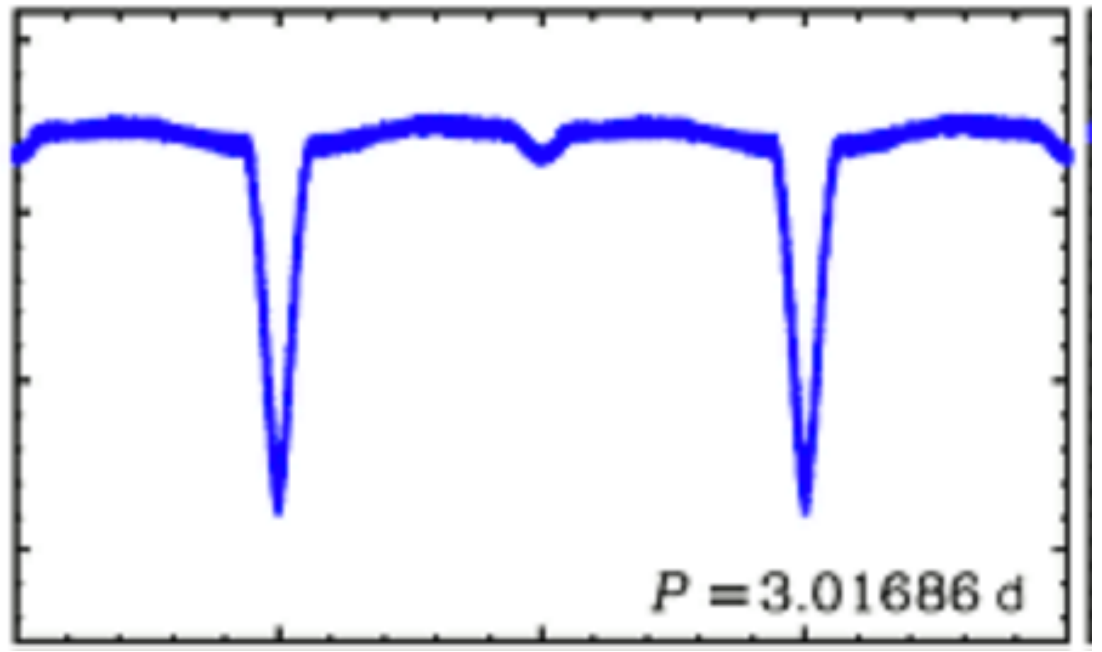
Fourier Transform

- simplicity
- clear interpretation
- computional efficiency

good for sinusoidal data



not good for searching for narrow sharp pulses



Observations

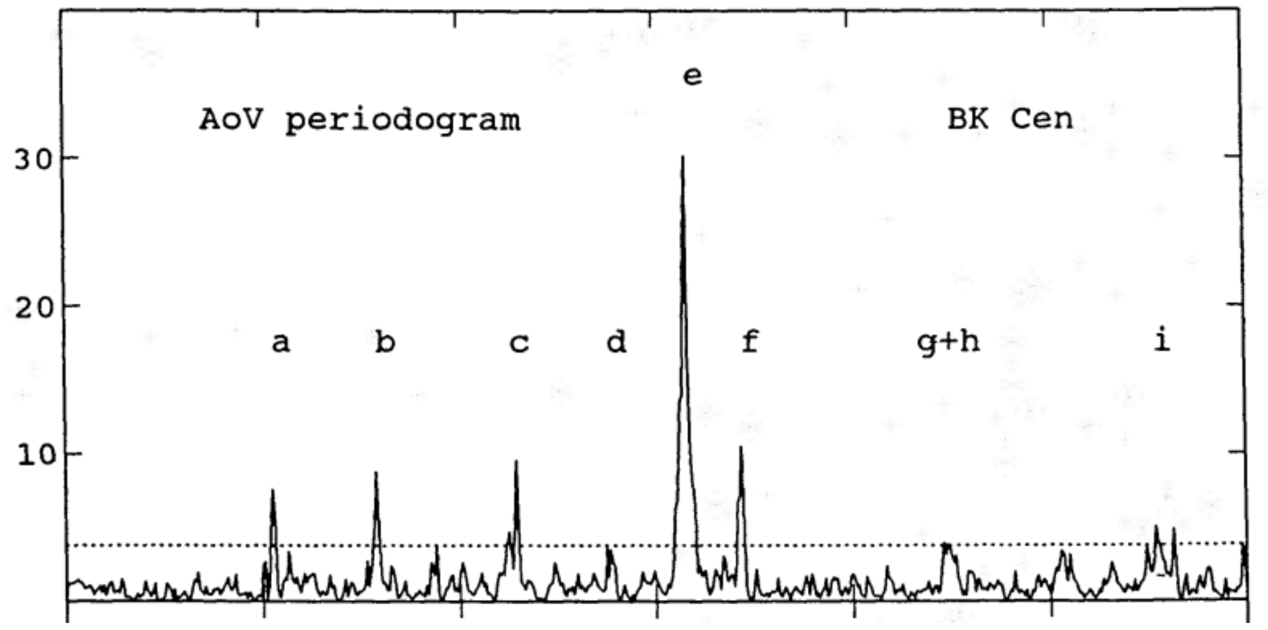
- non-uniformly distributed in time
- with large gaps (eg. full moon, seasons)

Standard procedure → folding data with a trail period & grouping into phase bins

→ several methods → AoV

Methods features

- For each method a test statistics is needed
- It is a function of all observations and a trail period
- Since observations are random - test statistics is random too
- Period is not random - it is a parameter of the statistics
- For each period and for all available observations a test statistics gives a number
- The plot of values of the statistics for a range of periods - periodogram
- Features in periodogram (realated to oscillations in observations) - lines
- Similar lines may arise due to the noise in the data
- We need a criterion for statistical significance of the lines
- Most desirable are cases in which probability distribution of statistics is known



AoV and related test statistics

n - total number of observations
$\bar{x}$ - average of observations
r - number of bins
n_i - number of observations in ith bin
$\bar{x}_i$ - average of observations in ith bin
$x_{ij} = x(t_{ij})$ - jth individual observation in the ith bin obtained at time t_{ij}

We assume $\bar{x} = 0$

We define three statistics s_1^2, s_2^2, s_0^2 as

$$(r-1)s_1^2 = \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2,$$ → „between bins variance"

$$(n-r)s_2^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$ → „within bin variance"

$$(n-1)s_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2.$$ → „standard variance"

They satisfy an algebraic identity

$$(n-1)s_0^2 = (r-1)s_1^2 + (n-r)s_2^2$$

A pure noise signal

Assumption: observations are **Gaussian white noise** with **zero mean** and **variance** chosen for unit of power (null hypothesis H_0)

$$E[x_{ij}]=0, \qquad \mathrm{Var}[x_{ij}]=\sigma^2\equiv 1, \qquad \mathrm{Cov}[x_{ij}x_{kl}]=\delta_{ik}\delta_{jl}.$$

Properties of three statistics

$$E[s_1^2]=E[s_2^2]=E[s_0^2]=\sigma^2\equiv 1$$

$$\mathrm{Var}[s_1^2]=\frac{2}{r-1}, \qquad \mathrm{Var}[s_2^2]=\frac{2}{n-r}, \qquad \mathrm{Var}[s_0^2]=\frac{2}{n-1}$$

- s^2 statistics are unbiased estimates of σ^2
- we don't know the estimated value of σ
- in order to verify H_0 we have to use ratios instead of individual values of s^2

$\Theta_1=s_1^2/s_0^2$, $\Theta_2=s_2^2/s_0^2$ and $\Theta_{\mathrm{AoV}}=s_1^2/s_2^2$ → standard AoV test statistics

A pure noise signal

$$\Theta_{\rm AoV} = s_1^2 / s_2^2$$

ratio of two independent random variables → it has Fisher-Snedecor F-distribution with *r-1* and *n-r* degrees of freedom

$$E[\Theta_{\rm AoV}] = \frac{n-r}{n-r-2}$$

$$\mathrm{Var}[\Theta_{\rm AoV}] = \frac{2(n-r)^2(n-3)}{(r-1)(n-r-2)^2(n-r-4)}$$

s_0^2 is not independent of s_1^2 and s_2^2, so we don't know probability functions of Θ_1 and Θ_2

A pure noise signal

$$\Theta_1 = s_1^2/s_0^2,\ \Theta_2 = s_2^2/s_0^2 \text{ and } \Theta_{\mathrm{AoV}} = s_1^2/s_2^2$$

Let's investigate probability functions of Θ_1 and Θ_2 considering an asymptotic case

$n \to \infty, \quad r \to \infty \quad$ and $\quad r/n \to 0,$ → number of observations and bins goes to infinity, but slow

Assumption: all bins contain an equal number of data n_i

$$E[\Theta_{\mathrm{AoV}}] = E[\Theta_1] = E[\Theta_2] = 1$$

$$\mathrm{Var}[\Theta_{\mathrm{AoV}}] = \frac{2(n-1)}{(n-r)(r-1)}$$ → it has F-distribution

$$\mathrm{Var}[\Theta_1] = \frac{2(n-r)}{(n-1)(r-1)}$$

$$\mathrm{Var}[\Theta_2] = \frac{2(r-1)}{(n-r)(n-1)}.$$ → it does not have F-distribution

A periodic signal with noise

Assumption: observations contain the sum of a **white noise *a*** and a **periodic signal *f*** (hypothesis H_1)

$$x_{ij} = a_{ij} + Af_{ij}$$

A - amplitude of periodic signal of period P

f - arbitrary periodic function with period P_0 normalized to unit amplitude

We assume $\bar{f} = 0$

Expected values of statistics s_1^2, s_2^2, s_0^2 →

$$E[s_1^2] = 1 + A^2 F_1^2$$

$$E[s_2^2] = 1 + A^2 F_2^2$$

$$E[s_0^2] = 1 + A^2 F_0^2,$$

Coefficients:

$$F_1^2 = \frac{1}{r-1} \sum_{i=1}^{r} n_i (\bar{f}_i - \bar{f})^2$$

$$F_2^2 = \frac{1}{n-r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (f_{ij} - \bar{f}_i)^2$$

$$F_0^2 = \frac{1}{n-r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (f_{ij} - \bar{f})^2$$

Depend on the shape of the signal, but don't depend on its amplitude

A periodic signal with noise

Let's consider asymptotic case

$$n \to \infty, \qquad r \to \infty \quad \text{and} \quad r/n \to 0,$$

$$E[\Theta_{\text{AoV}}] = 1 + A^2(F_1^2 - F_2^2) + O(A^4)$$
$$E[\Theta_1] = 1 + A^2(F_1^2 - F_0^2) + O(A^4)$$
$$E[\Theta_2] = 1 + A^2(F_2^2 - F_0^2) + O(A^4).$$

elements dependent on A^4

➔

we can eliminate F_1

$$E[\Theta_{\text{AoV}}] = 1 + A^2(F_0^2 - F_2^2)\frac{n-1}{r-1} + O(A^4)$$
$$E[\Theta_1] = 1 + A^2(F_0^2 - F_2^2)\frac{n-r}{r-1} + O(A^4)$$
$$E[\Theta_2] = 1 - A^2(F_0^2 - F_2^2) + O(A^4).$$

A periodic signal with noise

We shall compare the power *1-β* of various *Θ* statistics for testing randomness of observations

$$1-\beta=\frac{1}{2}\left\{1-\mathrm{erf}\left[\frac{\Theta^{\mathrm{cr}}(\alpha)-E[\Theta;A]}{\sqrt{2\,\mathrm{Var}[\Theta;A]}}\right]\right\}$$

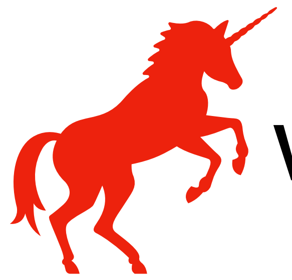
where $\Theta^{\mathrm{cr}}(\alpha)$ is the critical value of Θ statistics for confidence level α

→ case: signal

$$\alpha=\frac{1}{2}\left\{1-\mathrm{erf}\left[\frac{\Theta^{\mathrm{cr}}(\alpha)-E[\Theta;0]}{\sqrt{2\,\mathrm{Var}[\Theta;0]}}\right]\right\}.$$

→ case: only noise

Value above which we can detect periodicity (below which signal is indistinguishable from noise)

After  we obtain expression for power of the test

$$1-\beta=\tfrac{1}{2}\{1-\mathrm{erf}\,[\mathrm{erf}^{-1}(1-2\alpha)-A^2S]\}$$

$$S=(F_0^2-F_2^2)\sqrt{\frac{(n-r)(n-1)}{8(r-1)}}$$

for small amplitudes A

$$S=\left|\frac{dE[\Theta;0]}{dA^2}\right|\frac{1}{\sqrt{2\,\mathrm{Var}[\Theta;0]}}$$

Example 1 - BK Cen

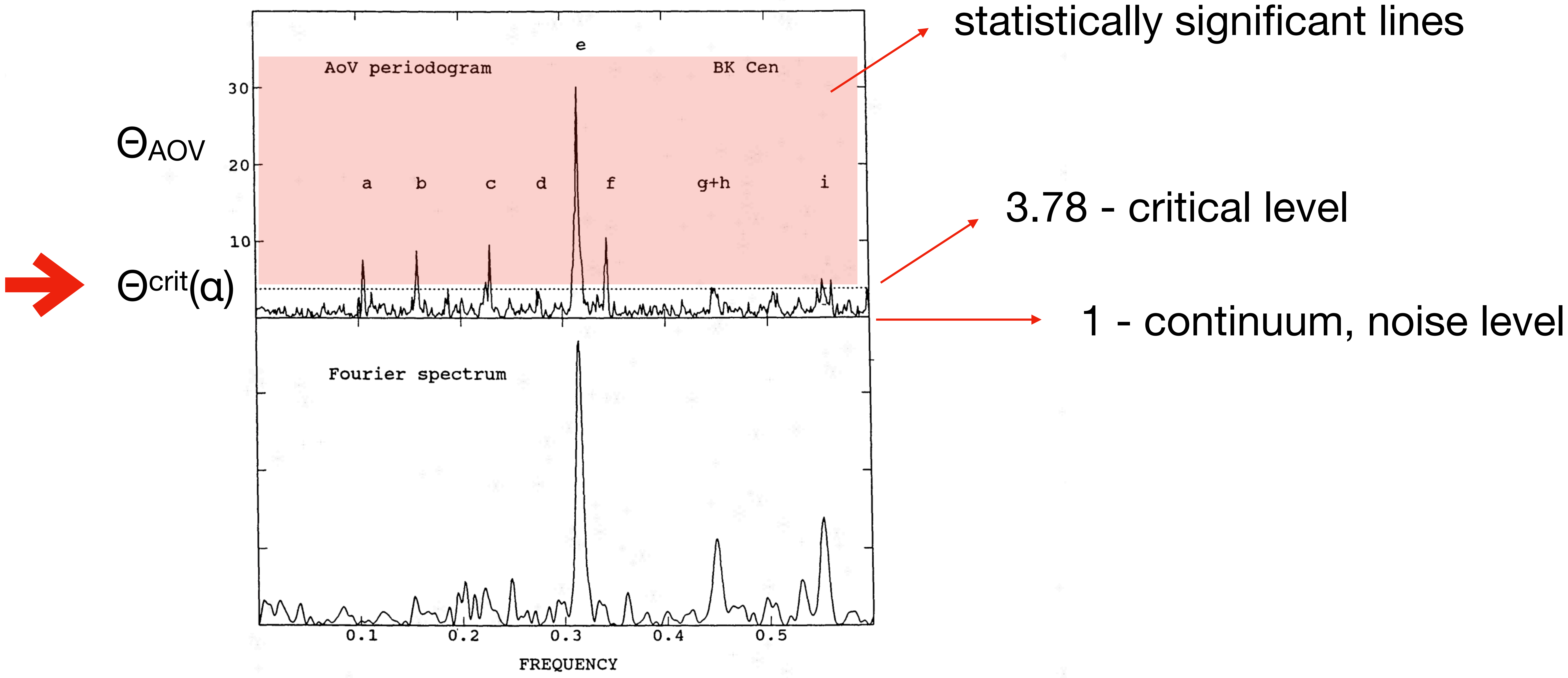


Figure 1. The analysis of variance (AoV) and Fourier periodograms for the photometric observations of the double-mode Cepheid BK Cen (Leotta-Janin 1967). Frequency is in cycles per day and power in arbitrary units. The annotation of the lines follows the PDM analysis by Stellingwerf (1978). The dotted line indicates the critical value of the AoV statistic for the significance level 0.05, uncorrected for bandwidth (see text). Comparison with the PDM results indicates the superior sensitivity of the AoV criterion.

Example 2 - V603 Aql

Assumption:

Θ_{AOV}

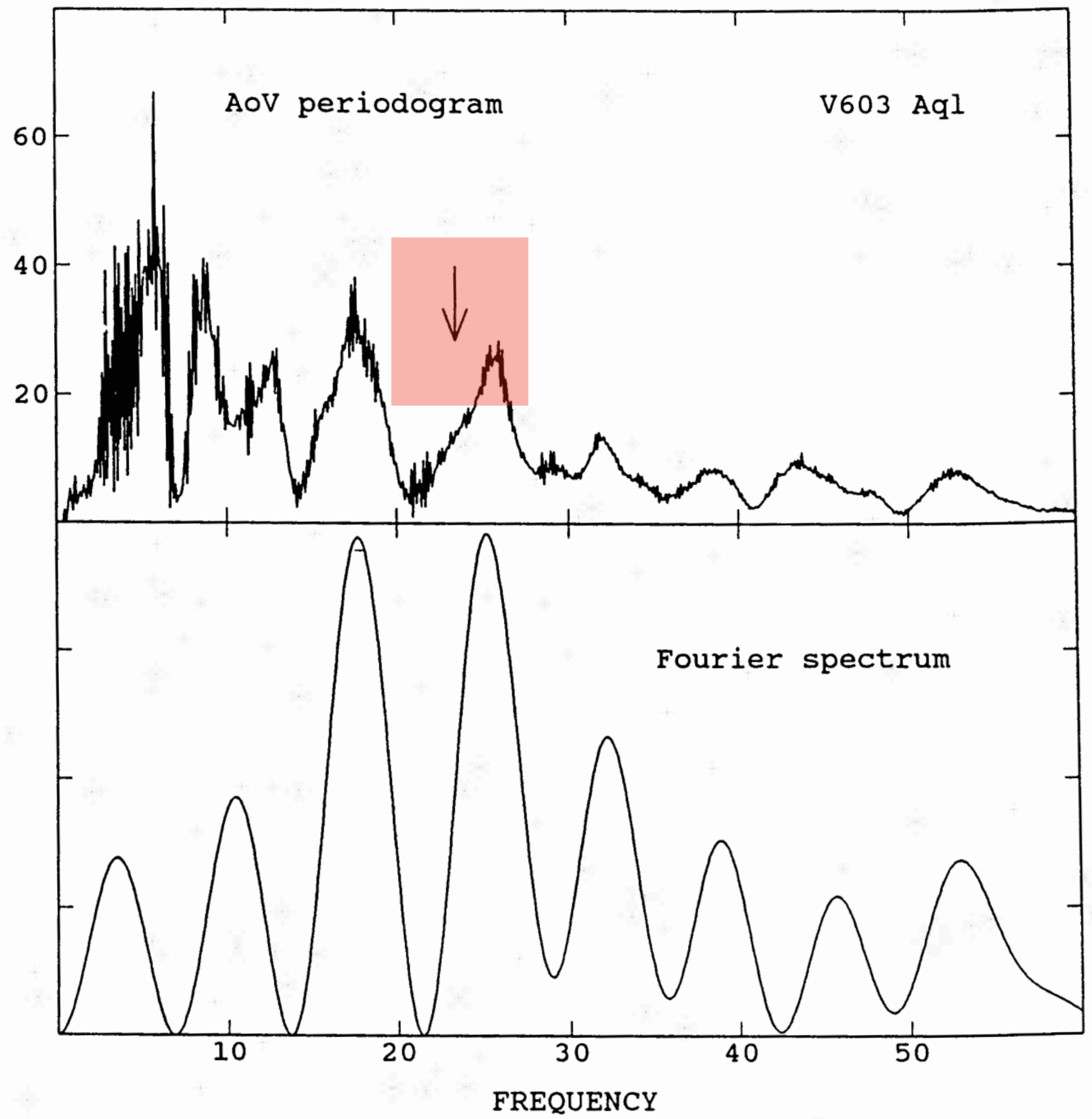


Figure 2. The same as Fig. 1 for one of nine runs of observations of Nova V603 Aquilae obtained by Udalski & ;chwarzenberg-Czerny (1989). The run lasted 3.9 h and the main 3.5-h period was removed from the data. An ırrow indicates the 61.4-min oscillation period discovered by these authors in optical and X-ray data. Since in he AoV periodograms no harmonic artefacts are produced, the first high-frequency feature (at 25 cycles d^{-1}) ught to be real, indicating detection of the 61-min oscillation in this single run (see text for details).

→ 5 - continuum level

flickering of several minutes & 20 s separation of observations → random signals are correlated

→ we cannot use F-significance criterion ✗

Recipe

- assume number of bins r
- assume trail period P
- calculate s_1, s_2
- calculate Θ_{AOV}
- repeat for various P
- assume the shape of signal f
- calculate S for given f and r
- assume confidence level α
- calculate $1-\beta$
- calculate $\Theta^{crit}(\alpha)$
- check if the signal line is above $\Theta^{crit}(\alpha)$

Conclusions

- AoV is recommended for detection **sharp periodic signals**
- Application of AoV requires **folding** and **binning** with a trail period
- Its **probability distribution** function is **known** for any numbers of observations (even for small samples)
- Modifications of the method: multi-harmonic AoV (used in TATRY code)