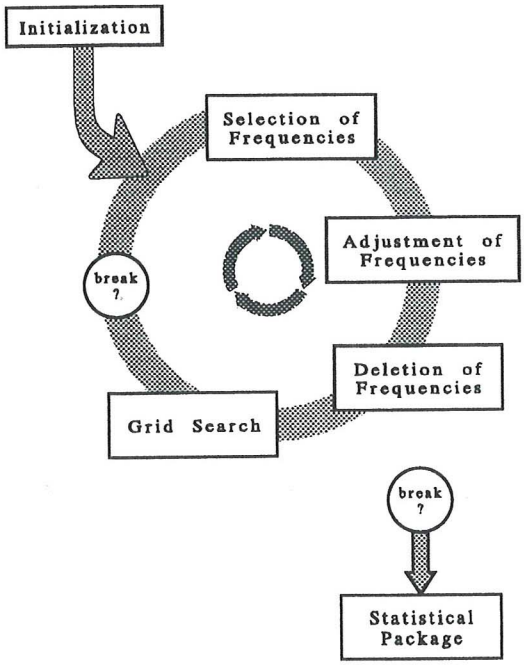
Automatic Analysis for Time-Series with Large Gaps

Mössner Martin, Pfleiderer Jörg,[1]
Department of Astronomy, Leopold-Franzens-University,
Technikerstraße 25, A–6020 Innsbruck, Austria
email: martin.moessner@uibk.ac.at

Abstract

We discuss a new algorithm for time-series analysis with large gaps. Our algorithm uses special techniques for selection and deletion of frequencies. Frequency improvement is done by multidimensional nonlinear optimization combined with grid searching. The results are checked by statistical tests. As a demonstration, we give an evaluation of an ill-conditioned test-example and a re-evaluation of the periods of the δ-Scuti star θ^2 Tauri.

1 Introduction



A set of data $\boldsymbol{m}(t) = (m_j(t_j))$, $j = 1, \ldots, M$, such as time-dependent brightness data, is to be approximated by $\boldsymbol{s} = (s_j)$:

$$s_j = A_0 + \sum_{i=1}^{N} A_i \cos(2\pi f_i t_j - \phi_i). \quad (1)$$

The amplitudes A_i, frequencies f_i, phases ϕ_i, as well as the number N of frequencies considered are unknown. The data and, perhaps, the times contain observational errors. We minimize

$$||\boldsymbol{m} - \boldsymbol{s}||_2 = ||\boldsymbol{r}||_2 = min\,! \quad (2)$$

in the least squares sense. In this paper, we show how to choose N and estimate the sets $\boldsymbol{A}$, $\boldsymbol{f}$, and $\boldsymbol{\phi}$. The quality of the fit is checked by statistical means.
The diagram illustrates the main algorithm. The individual steps will be described in the following sections.

2 Selection of Frequencies

The power function $p(f)$ is the criterion for the selection of frequencies:

$$p(f) = \sqrt{\frac{\left(\sum_{j=1}^{M} \cos(2\pi f t_j)\, m_j\right)^2}{\sum_{j=1}^{M} \cos^2(2\pi f t_j)} + \frac{\left(\sum_{j=1}^{M} \sin(2\pi f t_j)\, m_j\right)^2}{\sum_{j=1}^{M} \sin^2(2\pi f t_j)}} \quad (3)$$

Proceedings of 5^{th} ESO/ST-ECF Data Analysis Workshop, April 26-27, 1993
[1]This work was supported by the Austrian Research Foundation under grant P8568-PHY and by the Austrian Academy of Sciences (Space Research / National Programs)

- Because $p(f)$ is a very rapidly oscillating function, high resolution is necessary for a clear image. Optimal resolution is given by $\Delta f = 0.1\,\frac{1}{T}$ with $T = t_M - t_1$. That means one has to compute $p(f)$ for about $n_p = 10\,T\,(f_{max} - f_{min})$ values. In the case of θ^2 Tauri we have $T \approx 1700$ d and $f_{max} = 100\ \mathrm{d}^{-1}$ and therefore $n_p = 1.7 \cdot 10^6$. At this place it is crucial to take a proper compromise between required accuracy and computational costs.
- In order to overcome local oscillations we smooth $p(f)$ by replacing it by the maximum within overlapping intervals.

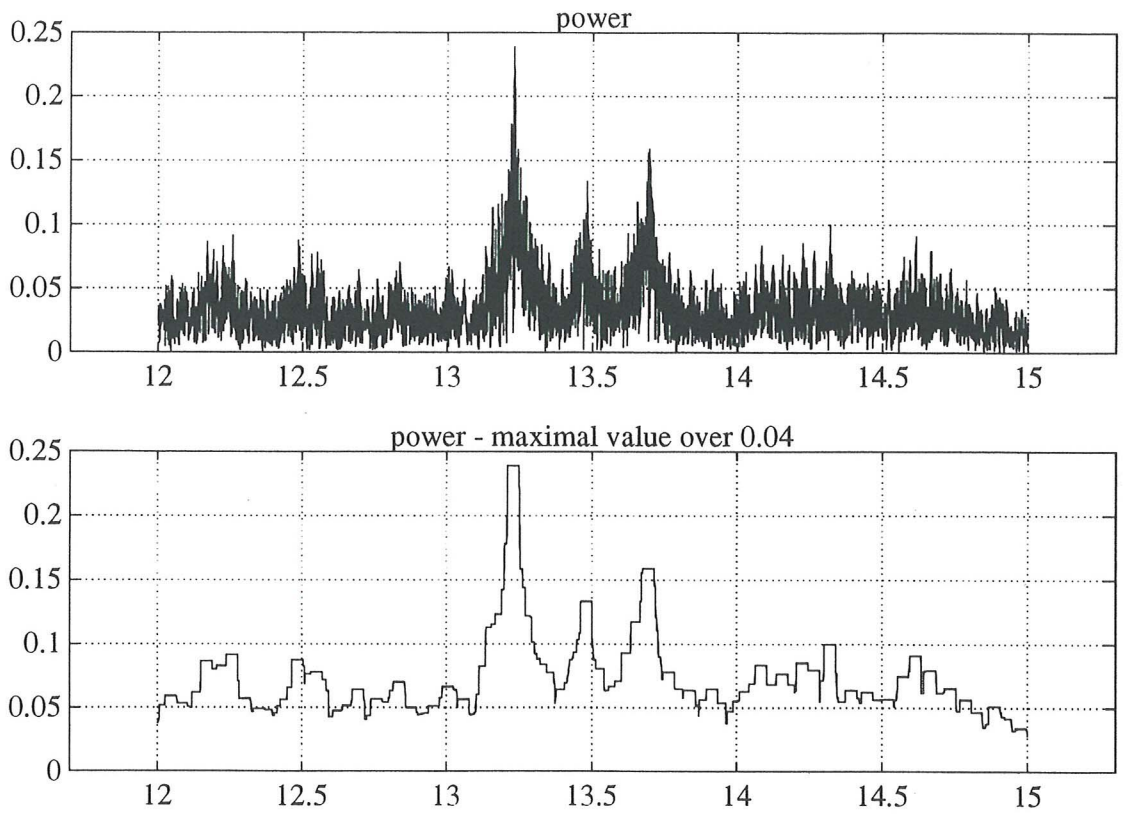


Figure 1: frequency versus (smoothed) power for θ^2 Tauri

- The frequency vector used as starting value for the minimization routine is appended by up to n_{add} values of frequencies, corresponding to maxima of the smoothed power function.
- If in the course of the algorithm more than n_{max} frequencies are proposed, only those n_{max} frequencies are accepted that correspond to the strongest contribution in power.

3 Adjustment of Frequencies – Nonlinear Minimization

Given a start frequency vector $\boldsymbol{f} = (f_1, \ldots, f_N)$ we solve the following overdetermined system of nonlinear equations:

$$\boldsymbol{g}(\boldsymbol{f}) = \left(a_0(\boldsymbol{f}) + \sum_{i=1}^{N} a_i(\boldsymbol{f}) \cos(2\pi\, f_i\, t_j) + \sum_{i=1}^{N} b_i(\boldsymbol{f}) \sin(2\pi\, f_i\, t_j) \right)_j \approx \boldsymbol{m} \tag{4}$$

For the amplitudes $a_0(\boldsymbol{f})$, $a_i(\boldsymbol{f})$, $b_i(\boldsymbol{f})$ we take the best fit at the current frequency vector (see below). Because $M > N$, the code tries to minimize

$$|| \boldsymbol{g}(\boldsymbol{f}) - \boldsymbol{m} ||_2 = min! \tag{5}$$

by varying $\boldsymbol{f}$, in the least squares sense. For this we use the damped Gauss-Newton code **nlscon** (see Deuflhard [3, 4], coding Novak and Weimann [10]). For testing purposes we used the lecture program **newton** (Hairer [8]), which works quite well.

4 Determination of Amplitudes – Linear Minimization

The minimization routine needs a function which computes the amplitudes $\boldsymbol{a}$:

$$\boldsymbol{A} \cdot \boldsymbol{a} \approx \boldsymbol{m}$$

$$\boldsymbol{A} = \begin{pmatrix} 1 & \cos(2\pi f_1 t_1) & \cdots & \cos(2\pi f_N t_1) & \sin(2\pi f_1 t_1) & \cdots & \sin(2\pi f_N t_1) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cos(2\pi f_1 t_M) & \cdots & \cos(2\pi f_N t_M) & \sin(2\pi f_1 t_M) & \cdots & \sin(2\pi f_N t_M) \end{pmatrix} \tag{6}$$

$$\boldsymbol{a} = (a_0,\ a_1,\ \cdots,\ a_N,\ b_1,\ \cdots,\ b_N)^t \qquad \boldsymbol{m} = (m_1,\ \cdots,\ m_M)^t$$

If Equation (6) has full rank, a unique $\boldsymbol{a}$ exists which minimizes $||\boldsymbol{Aa} - \boldsymbol{m}||_2$. This solution is called the least squares solution (see Golub and Van Loan [7]).

The least squares solution assumes an exact matrix $\boldsymbol{A}$ and tries to get the best fit for the data $\boldsymbol{m}$, which are assumed to contain errors. However, in our situation time data are given with a precision of 10^{-5} d. We search for frequencies of the order of 15 d^{-1}. This leads to errors in $\boldsymbol{A}$ of size 10^{-3}. On the other side, brightness data have errors which are bounded by one to two thousands of a magnitude. Hence, both sides of Equation (6) show errors of the same order of magnitude. Therefore we use the method of total least squares (see Golub and Van Loan [6], Van Huffel and Vandewalle [11, 12]).

Astronomical time series have usually pronounced gaps between groups of data (daily, monthly, and seasonal gaps). The linear system (6) is, in such cases, ill-conditioned. Therefore, one has to choose carefully the numerical algorithms. For least squares we use singular value decomposition [7] and for total least squares we use an implementation by Mössner which is based on the paper of Golub and Van Loan [6] and the Lapack library [5].

5 Deletion of Frequencies

In practical examples, it is possible that the strongest power-amplitude is caused by aliasing effects (see test example). Nevertheless, if the correct frequencies are chosen, amplitudes for aliased frequencies become small and can therefore be deleted. We use the following criteria:

- If there is a closely spaced pair of frequencies, a special routine is called. This routine inserts new frequencies and checks whether a better fit exists. If the old pair was a resonance phenomenon, it will be deleted in a later stage.
- Frequencies which are *very* closely spaced are substituted by their mean value.
- Frequencies with small amplitudes are deleted.
- Take only the strongest $n_{max} - 3$ signals into the next iteration. This enables the code to try new frequencies, even if the frequency vector was already full.
- After a certain number of iterations, a special deletion routine is called. This simulates a restart with better initial data.

6 Grid Search

The main break criterion for our algorithm gives the norm of the residual vector of Equation (4) or (6), respectively. However, the norm of the residuals shows strong oscillations as function of any component of the frequency vector $\boldsymbol{f}$. Therefore, most optimization techniques tend to catch one of the local minima, instead of the absolute minimum. The typical frequency for these oscillations is well known ($\Delta f = \frac{1}{2T}$). Therefore, we do a grid search in each component of $\boldsymbol{f}$ (with grid size $0.25/T$) at the end of each iteration of the main loop.

7 Statistical Package

After successful exit or after i_{max} iterations, iteration stops and a statistical package is called:

- Tests are performed on the randomness of the residual vector.
- Confidence intervals are computed for the obtained parameters.
- A test whether all frequencies are necessary is in preparation.

With the Student's t distribution with ν degrees of freedom $A(t|\nu)$ (see Abramowitz and Stegun [1]) and its β-fractil $t_{\beta,\nu}$ defined by $A(t_{\beta,\nu}|\nu) = \beta$, we have the probability (see Mendenhall *et al.* [9]):

$$\begin{aligned} P\Big(|p_{i,w} - p_i| < \delta_i\Big) &= 1 - \beta \\ \delta_i &= S\,\sqrt{c_{ii}}\; t_{1-\beta/2,M-3N-1} \\ S = \frac{||\boldsymbol{r}||_2}{\sqrt{M-3N-1}} \qquad & c_{ii} = \Big((\boldsymbol{J}^t\boldsymbol{J})^{-1}\Big)_{ii} \end{aligned} \tag{7}$$

Here $p_{i,w}$ is the true value of the computed parameter p_i (= A_i, f_i, or ϕ_i). β is the level of significance of the confidence interval and $\boldsymbol{J}$ denotes the Jacobian of the minimization function $\boldsymbol{g}(\boldsymbol{f})$. Altogether, we have confidence intervals ($p_i \pm \delta_i(\beta)$) related to the error probability β.

8 A Test Example

For testing purposes we created data sets for the following signal

$$\begin{aligned} s_j \;=\; & 1 + 4\,\cos(2\pi\,6.5\,t_j) + 6\,\cos(2\pi\,8.6\,t_j + 1) + \\ & 5\,\cos(2\pi\,9.3\,t_j - 2) + 3\,\cos(2\pi\,18.0\,t_j - 3) + r_j \end{aligned} \tag{8}$$

We selected the time distribution:

$$\begin{aligned} & t_j = \frac{0.2\,i}{n_{step}} + d + 300\,y \qquad \text{with} \qquad j = i + 1 + n_{step}\cdot(d + n_{day}\cdot y) \\ & i = 0,\ldots,n_{step}-1 \qquad d = 0,\ldots,n_{day}-1 \qquad y = 0,\ldots,n_{year}-1 \end{aligned} \tag{9}$$

Our test-example uses $n_{step} = 20$, $n_{day} = 30$, and $n_{year} = 2$. Errors were simulated by a uniform random generator. Given two uniform random numbers U_1 and U_2, one gets, by the transformation $G = \sigma\sqrt{-2\ln U_1}\,\cos(2\pi U_2) + \mu$ (see Abramowitz and Stegun [1]) a Gaussian random number G with mean μ and variance σ. The test example contains Gaussian errors with mean $\mu = 0$ and variance $\sigma = 0.2$. Errors in the time data were simulated by adding, to t_j, a uniformly distributed random number of range $\pm 10^{-4}$.

The power of the given example has aliasing maxima at 5.5, 7.5, 8.3, 9.6, 10.3, 10.6. As starting vector for the frequencies we took a single wrong frequency $\boldsymbol{f} = (3.0)$. The above mentioned

parameters were chosen as follows: $n_p = 3000$, $f_{min} = 5$, $f_{max} = 20$, $n_{add} = 7$, $n_{max} = 12$ and $i_{max} = 10$. The code stopped after 5 iterations, because the residuals approached the statistical limit. The result for the error probability $\beta = 0.001$ is

	frequency		amplitude		phase	
	0.000000		1.015	(±0.038)		
3	6.500022	(±0.000029)	3.990	(±0.054)	0.020	(±0.021)
1	8.599976	(±0.000019)	5.985	(±0.054)	0.976	(±0.016)
2	9.300000	(±0.000023)	4.997	(±0.054)	-1.997	(±0.016)
4	17.999928	(±0.000040)	2.952	(±0.054)	2.822	(±0.028)

9 Re-evaluation of the δ-Scuti star θ^2 Tauri

The examined data are due to Breger [2]. The 2806 measurements are sampled over five years and contain large gaps (see table below). The strongest five frequencies are well known and have been published by Breger [2]:

days	number of data
17-28	248
316-322	248
347-358	112
625-738	1145
1355-1359	80
1738-1759	1190

results of Breger	
frequency	amplitude
13.229653	0.0066
13.480733	0.0026
13.693597	0.0045
14.317637	0.0027
14.614537	0.0012

In the following we indicate the ten strongest signals as obtained by our analysis. The confidence intervals correspond to an error probability of $\beta = 0.001$. The given phases (radians) refer to the epoch JD 2445017.0. Particularly the weak signals must be checked in further campaigns. Of special interest is the isolated weak signal at 26.18 which cannot be explained by aliasing effects.

	frequency		amplitude		phase	
	0.000000		-0.00003	(±0.00016)		
7	12.172308	(±0.000088)	0.00075	(±0.00025)	0.61	(±0.68)
1	13.229653	(±0.000010)	0.00654	(±0.00026)	2.60	(±0.08)
4	13.480717	(±0.000027)	0.00246	(±0.00025)	-1.86	(±0.21)
6	13.647071	(±0.000084)	0.00085	(±0.00025)	-2.55	(±0.65)
2	13.693601	(±0.000017)	0.00428	(±0.00024)	1.33	(±0.12)
9	13.827642	(±0.000104)	0.00063	(±0.00024)	-2.66	(±0.81)
3	14.317639	(±0.000033)	0.00258	(±0.00032)	0.81	(±0.22)
8	14.323414	(±0.000140)	0.00063	(±0.00033)	3.05	(±0.94)
5	14.613782	(±0.000060)	0.00111	(±0.00025)	0.03	(±0.47)
10	26.189602	(±0.000126)	0.00050	(±0.00023)	-0.34	(±0.97)

References

[1] Abramowitz, M., Stegun, I.A.: 1972, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Dover Publications, Inc; New York.

[2] Breger M., Garrido R., Huang Lin, Jiang Shi-yang, Guo Zi-He, Frueh M., Paparo M.: 1989, *Astron. Astrophys.* **214**, 209.

[3] Deuflhard P.: 1974, *Numer. Math.* **22**, 289.

[4] Deuflhard P.: 1993, *Newton Techniques for Highly Nonlinear Problems - Theory and Algorithms.* Academic Press Inc. (in press)

[5] Dongarra J. *et al.*: 1992, *LAPACK Users' Guide.* SIAM, Philadelphia.

[6] Golub G.H., Van Loan C.F.: 1980, *SIAM J. Numer. Anal.* **17**, 883-93.

[7] Golub G.H., Van Loan C.F.: 1990, *Matrix Computations.* 2nd Edition, The Johns Hopkins University Press, Baltimore and London.

[8] Hairer E.: 1991, (Private communication, transmission of lecture-programs). Université de Genève, Dept. de mathématiques, Genève, Switzerland.

[9] Mendenhall W., Wackerly D.D., Schaeffer R.L.: 1989, *Mathematical Statistics with Applications.* Fourth Eddition. PWS-Kent Publishing Company, Boston.

[10] Nowak U., Weimann L.: 1991, *A Family of Newton Codes for Systems of Highly Nonlinear Equations - Algorithm, Implementation, Application.* Konrad Zuse Zentrum für Informationstechnik Berlin (ZIB), Technical Report TR 91-10.

[11] Van Huffel, S., Vandewalle, J.: 1989, *Numer. Math.* **55**, 431-49.

[12] Van Huffel, S., Vandewalle, J.: 1989, *SIAM J. Matrix Anal. Appl. 10*, **3**, 294.