A&A 481, 571–574 (2008)
DOI: 10.1051/0004-6361:20077559

Astronomy & Astrophysics

A heuristic derivation of the uncertainty for frequency determination in time series data

T. Kallinger, P. Reegen, and W. W. Weiss

Institute for Astronomy (IfA), University of Vienna, Türkenschanzstrasse 17, 1180 Vienna, Austria
e-mail: kallinger@astro.univie.ac.at

Received 28 March 2007 / Accepted 22 December 2007

ABSTRACT

Context. Several approaches to estimating frequency, phase, and amplitude errors in time-series analyse have been reported in the literature, but they are either time-consuming to compute, grossly overestimating the error, or are based on empirically determined criteria.
Aims. A simple, but realistic estimate of the frequency uncertainty in time-series analyses is our goal here.
Methods. Synthetic data sets with mono- and multi-periodic harmonic signals and with randomly distributed amplitude, frequency, and phase were generated and white noise added. We tried to recover the input parameters with classical Fourier techniques and investigated the error as a function of the relative level of noise, signal, and frequency difference.
Results. We present simple formulas for the upper limit of the amplitude, frequency, and phase uncertainties in time-serie analyses. We also demonstrate the possibility of detecting frequencies that are separated by less than the classical frequency resolution and of finding that the realistic frequency error is at least 4 times smaller than the classical frequency resolution.

Key words. methods: data analysis – methods: statistical – techniques: miscellaneous

1. Motivation

In the frequency analysis of time series, a realistic estimate of the amplitude, phase, and frequency uncertainties can be useful. A few examples are:

- The comparison of frequencies derived for simultaneously observed stars allows identifying the instrumental signal, if the frequencies occur in different data sets but within the frequency uncertainty.
- One needs to know the observed frequency errors to assess the quality of a fit of models to the observations.
- For mode identifications based on amplitude ratios or phase differences from multi-color photometry, one also needs a reliable estimate for the frequency error.

A combination of Fourier and least-square fitting algorithms (like *SigSpec* by Reegen 2007; *Period04* by Lenz & Breger 2005; or *CAPER* by Walker et al. 2005) is a frequently used method for determining frequencies, amplitudes, and phases of harmonic signals. For a time series consisting of a perfect sine wave and white noise, the frequency error is determined by the total time base of the data set and the signal-to-noise ratio (S/R) of the corresponding amplitude in the Fourier spectrum. Montgomery & O'Donoghue (1999) define the amplitude, phase, and frequency errors as

$$\sigma(a)_{\rm Montgomery} = \sqrt{\frac{2}{N}}\sigma(m), \tag{1}$$

$$\sigma(\phi)_{\rm Montgomery} = \sqrt{\frac{2}{N}}\frac{\sigma(m)}{a}, \tag{2}$$

$$\sigma(f)_{\rm Montgomery} = \sqrt{\frac{6}{N}}\frac{1}{\pi T}\frac{\sigma(m)}{a}, \tag{3}$$

based on an analytical solution for the one-sigma error of a least-square sinusoidal fit with an rms of $\sigma(m)$. The total number of data points, the total time base of the observations, the signal amplitude, phase, and frequency are N, T, a, ϕ, and f, respectively. Hence the last term in Eqs. (2) and (3) represents S/R^{-1} in the time domain. We want to mention that the time domain S/R in these relations is not equal to the commonly used S/R in the amplitude spectrum (peak amplitude divided by the average amplitude in a given frequency range), and it scales to the time domain S/R by a factor of $\approx\sqrt{\pi/N}$. Reegen (2007) shows that this scaling cannot be applied uniquely to the full frequency range and that systematic effects have to be taken into account if an exact description of frequency-domain errors is needed.

However, in reality an intrinsic signal is superposed not only by white noise (e.g. due to photon statistics) but also by correlated noise (e.g. atmospheric scintillation for ground-based data) or non-Gaussian distributed noise (e.g., introduced by the data reduction). Even the star itself can contribute correlated noise, for example due to granulation. All these noise sources increase the real frequency uncertainty, which leads to the unsatisfying situation that several empirical parameters can be found in the literature that tune the frequency error to personal experience.

People quite often use the Rayleigh frequency resolution (T^{-1}), defined by the total time base of the data set, which is a dramatic overestimation of the real uncertainty in most cases. To access the uncertainties of the fitting parameters for the time series analysis, it turned out to be an appropriate way to perform simulations with the actually analyzed data set, as done by Monte Carlo simulations in *Period04* or by bootstrap simulations in *CAPER* (see Rowe et al. 2006, for details). This approach has the disadvantage that the simulations can be very time-consuming especially if the data sets are big and/or include plenty of signal components.

2. Mono-periodic signal

To quantify the effect of white noise on the frequency determination of a coherent mono-periodic signal, a numerical simulation was performed for 42 597 synthetic data sets. Each data set consists of 10 000 data points uniformly distributed over 10 days and includes two components: a single sinusoidal signal with random (uniformly distributed) frequency, amplitude, and phase, and Gaussian distributed scatter with a random (uniformly distributed) amplitude (FWHM of the Gaussian random-number generator). All input parameters are independent of each other.

2.1. Frequency error

The routine *SigSpec*[1] (Reegen 2007) was used for the frequency analysis. It is an automatic program to detect periodic signals in data sets, and it relies on an exact analytical solution for the probability that a given DFT (discrete Fourier transform; Deeming 1975) amplitude is generated by white noise. Its main advantage over commonly used S/R estimates is its appropriately incorporated frequency *and* phase angle in Fourier space and time-domain sampling, hence using all available information instead of mean amplitude alone. The *SigSpec* spectral significance is defined as the logarithm of the inverse false-alarm probability that a DFT peak of a given amplitude arise from pure noise in a non-equidistantly spaced data set.

On average, an S/R of 4 corresponds to a spectral significance value of 5.46. This means that an amplitude of four times the noise level would appear by chance at a given frequency in one out of $10^{5.46}$ cases, assuming white noise.

Figure 1 shows the absolute deviation – scaled to the data set length – between the input frequency and the *SigSpec* frequency as a function of the spectral significance. Given are average values and the $+4\sigma$ (and -1σ) distribution in bins of spectral significance. Not surprisingly, there is a clear dependency of the frequency error on the significance (or S/R). Obviously, the real frequency error quite often (≈30%) exceeds the frequency error given by Eq. (3). However, we could heuristically define a frequency error criterion (top panel of Fig. 1) as

$$\sigma(f)_{\rm Ka} = \frac{1}{T \cdot \sqrt{sig}} \approx \frac{\pi \cdot \log e}{4 \cdot T \cdot S/R}, \qquad (4)$$

representing a good approximation for the upper limit of the frequency uncertainty and showing that the frequency uncertainty is less than the frequency resolution T^{-1}, at least by a factor of $\sqrt{sig}$. Only 4 out of 42 597 simulations result in a frequency error exceeding the thus defined upper frequency error limit. Aware that a simulation need not reflect the reality, we added the frequency error of real observations into Fig. 1 derived from the comparison of ground based data with long-term high-precision space photometry (MOST) of the same stars. We have to mention that plotting the frequency error as a function of the signal frequency (or phase) reveals no correlation between these quantities. To be independent of the spectral significance, synthetic data sets with a fixed S/R have been used.

The deviation from a linear relation at high significances in the log-log scale of Fig. 1 is due to a distortion of the significance scale, which is explained in Fig. 2, where the S/R in the amplitude spectrum is plotted versus the spectral significance for frequencies determined from the synthetic data sets. For spectral significances below about 300, the significance is roughly

[1] **Sig**nificance **Spec**trum, http://www.astro.univie.ac.at/SigSpec/

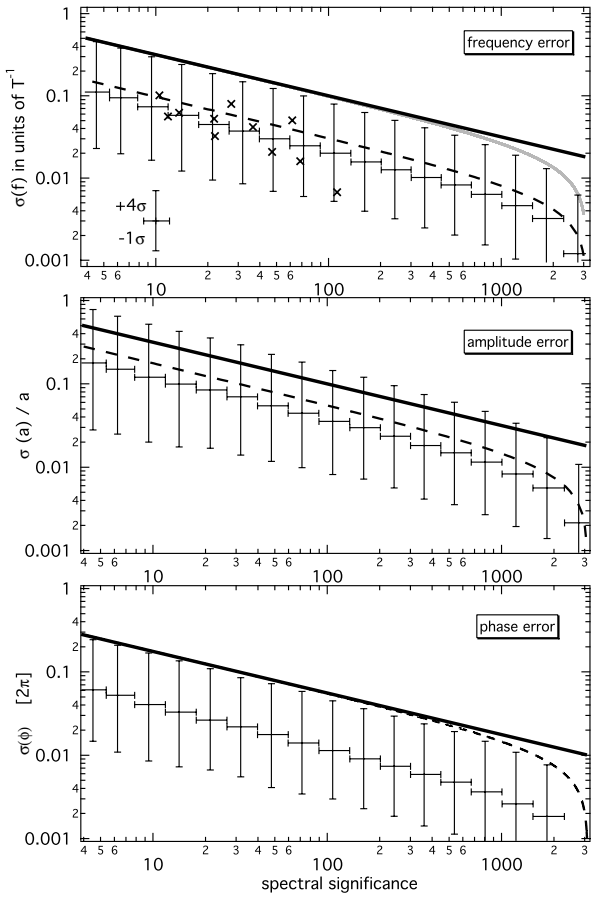

Fig. 1. *Top*: frequency error $\sigma(f)$ normalized to the Rayleigh frequency resolution given by the data set length T versus the spectral significance. Given are average values in bins (represented by the horizontal bars) as a result of a numerical simulation of 42 597 synthetic data sets including a single sinusoidal signal with random frequency, amplitude, phase, and white noise added. Vertical bars indicate the $+4\sigma$ (and -1σ) distribution, illustrating that the heuristically determined frequency error criterion (solid black line) represents a good approximation of the upper limit of the frequency uncertainty, which is at least by a factor significance$^{1/2}$ smaller than the frequency resolution T^{-1}. Cross symbols correspond to frequency errors derived from the comparison of real ground-based data with high-precession space photometry of the same stars. For an explanation of the grey line see the penultimate paragraph of Sect. 2.1. *Middle*: relative amplitude error versus the spectral significance. The solid line indicates the upper limit for the relative amplitude error given in this work. *Bottom*: phase error (in units of 2π) versus the spectral significance. The solid line shows the "Montgomery phase error" converted to spectral significance. *All panels*: the dashed lines represent the analytically determined one-sigma error of a sinusoidal least-square fit (Montgomery & O'Donoghue 1999).

equal to $(\pi \cdot \log e)/4$ times the $(S/R)^2$ in the amplitude spectrum (Reegen 2007). Only for extremely significant signals does one have to take into account that the noise calculation for the S/R and the spectral significance are different. Whereas the S/R is based on the average amplitude in a Fourier spectrum *after* prewithening the signal (corresponds to the rms residual), the spectral significance is based on the rms scatter of the time series *including* the signal. In other words, a pure signal without noise has an infinite S/R but still a finite spectral significance (see Reegen 2007, for details). The grey line in Fig. 1 takes this effect into account.

To explain the difference between the upper frequency error limit and the "Montgomery frequency error", we interpret the

latter to be the statistically expected value for the frequency uncertainty corresponding to the average values in the spectral significance bins of our simulation. We also have to mention that the frequency error distribution of our simulation (for fixed spectral significance) is neither Gaussian nor symmetric, which makes it very difficult to define an analytical average value and scatter for the frequency uncertainty.

2.2. Amplitude error

Whereas the absolute amplitude error only depends on the time series rms scatter (see Eq. (1)), the relative amplitude error $\frac{\sigma(a)}{a}$ should be correlated with the signal's spectral significance (or S/R). The middle panel of Fig. 1 shows the relative amplitude error (deviation between the input amplitude and the *SigSpec* amplitude relative to the *SigSpec* amplitude) versus the spectral significance of our simulated white-noise data sets. The dashed line indicates the relative amplitude error based on the absolute "Montgomery amplitude error" representing the statistically expected value. According to our upper limit for the frequency uncertainty, we could again define an upper limit for the amplitude error of a sinusoidal least-square fit as,

$$\frac{\sigma(a)_{\rm Ka}}{a} = \frac{1}{\sqrt{sig}} \approx \frac{2}{\sqrt{\pi \cdot \log e}} \cdot \frac{1}{S/R}, \tag{5}$$

the solid line in the middle panel of Fig. 1. However, the upper limit of the amplitude error is not defined as well as for the frequency error. But still, ≈98% of the determined amplitude errors are smaller than the given limit.

2.3. Phase error

The bottom panel of Fig. 1 illustrates the absolute deviation between the input phase and the *SigSpec* phase versus the spectral significance of the 42 597 synthetic data sets. Again, the dashed line indicates the phase error for a sinusoidal least-square fit according to Eq. (2). Unlike the "Montgomery frequency error" corresponding to the statistically expected value for the frequency uncertainty, the "Montgomery phase error" is consistent with an upper limit for the real phase error. All but 4 numerically determined phase errors are below the given limit. Equation (2) based on the time-domain S/R, is converted to spectral significances (and frequency-domain S/R) as follows,

$$\sigma(\phi) = \sqrt{\frac{\log e}{2 \cdot sig}} \approx \sqrt{\frac{2}{\pi}} \cdot \frac{1}{S/R}, \tag{6}$$

which is indicated by a solid and a dashed line in the bottom panel of Fig. 1.

3. Multi-periodic signal

Usually the smallest frequency separation of two independent signals in a data set that can be determined separately is called frequency resolution.

For two signals with comparable amplitudes, a frequency separation corresponding to the Rayleigh frequency resolution (T^{-1}) results in a local minimum between the two peaks in the amplitude spectrum. Closer frequencies produce an asymmetric peak, whereas the peak maximum is roughly at the amplitude-weighted mean of the frequencies. After prewhitening the signal (corresponding to the subtraction of a scaled

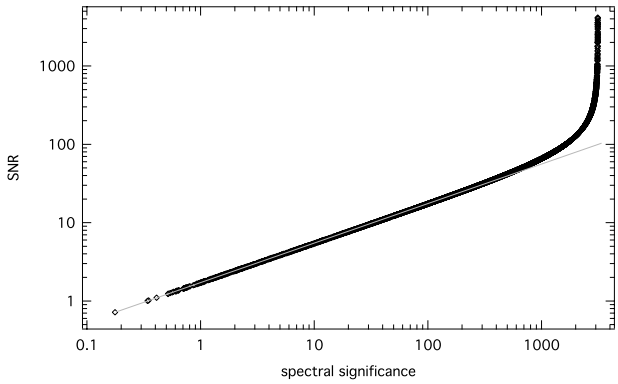

Fig. 2. Amplitude spectrum signal-to-noise ratio (S/R) versus spectral significance for frequencies determined from 42 597 synthetic data sets. The deviation from the linear relation (gray line in the log-log plot) at high significances is due to different noise estimates for S/R and spectral significance.

spectral window at the given frequency, Roberts et al. 1987), some signal will still be left in the amplitude spectrum. In other words, it should be possible to determine frequency, amplitude, and phase of signals separated in frequency by less than the frequency resolution, so the uncertainties of these parameters should be less than given by the Raleigh criterion.

To quantify this uncertainty, a numerical simulation was performed for ~50 000 synthetic data sets now including two signals with random frequency, amplitude, and phase for the first component. The second signal has a frequency randomly separated from the first one between 0 and 5 times the Rayleigh frequency resolution (T^{-1}), a random amplitude between 0.1 and 1 times the amplitude of the first one, and a random phase. Gaussian-distributed scatter with a random amplitude was added to the synthetic data.

Figure 3 shows the average absolute frequency error in bins of the spectral significance of the stronger signal for different ranges of the frequency separation Δf (in units of the Rayleigh frequency resolution) of the two input signals. The presence of a second signal separated by lower than the Rayleigh frequency resolution limits the frequency uncertainty of the stronger signal to $(4 \cdot T)^{-1}$ (see Fig. 3) if the spectral significance exceeds a value of 16. This is where both criteria give the same frequency error. We have to note that this limit is again purely heuristically determined. For a second signal, separated by more than 3 times the Rayleigh frequency resolution, the frequency uncertainty of the stronger signal is limited by the frequency error criterion for a mono-periodic signal given by Eq. (4) (see bottom panel in Fig. 3). There seems to be a smooth transition for $1 < \Delta f < 3$ (middle panel). Remarkably, only 13 out of ~50 000 (≈0.026%) numerically determined frequency errors do not satisfy the following criterion.

If a second signal is present within about three times the Rayleigh frequency resolution and spectral significance >16, the upper limit for the frequency error is

$$\sigma(f)_{\rm Ka} = \frac{1}{4T}. \tag{7}$$

In all other cases the frequency error is smaller than

$$\sigma(f)_{\rm Ka} = \frac{1}{T \cdot \sqrt{sig}}, \tag{8}$$

corresponding to Eq. (4).

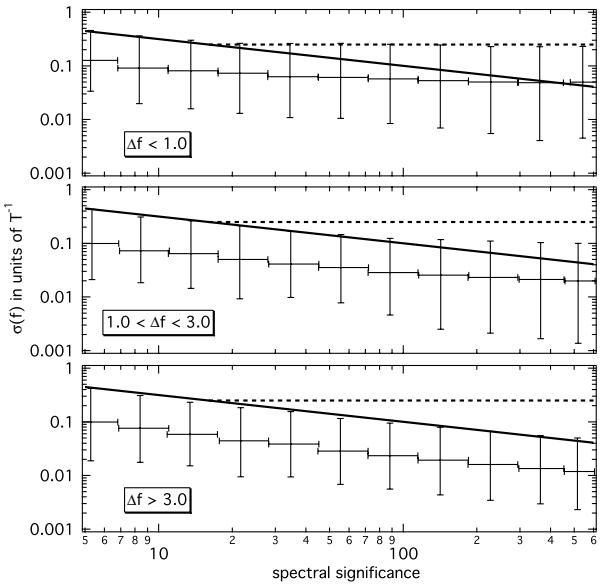

Fig. 3. Same as top panel of Fig. 1 now including two sinusoidal signals illustrating average frequency errors $\sigma(f)$ (normalized to the Rayleigh frequency resolution) of the stronger signal (first detected in the prewhitening sequence) in bins of the spectral significance along with $+4\sigma$ (and -1σ) environments in the bins. The panels refer to different ranges of the frequency separation Δf (in units of the T^{-1}) of the two input signals. The solid line indicates the upper frequency error limit for mono-periodic signals. The dashed line corresponds to the heuristically determined upper frequency error limit for close frequencies and is equal to $(4 \cdot T)^{-1}$.

4. Conclusions

Based on extensive simulations, we have shown that there is an upper limit to the amplitude and frequency error in time-series data analyses. Compared to the statistically *expected* value for the uncertainties given by Montgomery & O'Donoghue (1999), our *upper limits* cover the possible error due to white noise and even leaves room for additional error sources like atmospheric scintillation. A major advantage of calculating amplitude, frequency, and phase errors in terms of spectral significance rather than S/R is that the time-domain noise need not be Gaussian. As pointed out by Reegen (2007), the spectral significance does not depend on the probability distribution associated to the noise, and the only precondition is the uncorrelatedness of consecutive data points. It also must be mentioned that amplitude, frequency, and phase errors derived from spectral significances are only comparable to errors derived from S/R if the time-series is well-sampled (e.g. continuous space observations). Contrary to spectral significance based errors, S/R based error estimations (time-domain as well as frequency-domain) do not take the data sampling into account and can yield as a crude underestimation of the errors for "bad" sampling as is more or less always the case for single-site ground-based observations.

We have shown that the phase error defined by Montgomery & O'Donoghue (1999) is consistent with our simulations. Furthermore, we have shown that the determination of frequency pairs closer than the Rayleigh frequency resolution is possible and that the resulting frequency error is still 4 times smaller than the Rayleigh frequency resolution. However, our simulation does not say anything about the reliability of close frequency pairs in general. It tells us about the frequency uncertainty of a peak if, after prewhitening this peak, a second significant peak is present. It tells us that peaks do not influence each other's frequency determination if they are separated in frequency by 3 times the Rayleigh frequency resolution. For closer peaks, the frequency uncertainty is at least 4 times below the Rayleigh resolution even for peaks within the Rayleigh resolution.

Acknowledgements. This project was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF) within the project *The Core of the HR diagram* (P17580-N02), and the Bundesministerium für Verkehr, Innovation und Technologie (BM.VIT) via the Austrian Agentur für Luft- und Raumfahrt (FFG–ALR).

References

Deeming, T. J. 1975, Ap&SS, 36, 137
Lenz, P., & Breger, M. 2005, CoAst, 146, 53
Montgomery, M. H., & O'Donoghue, D. 1999, DSSN, 13, 28
Reegen, P. 2007, A&A, 467, 1353
Roberts, D. H., Lehar, J., & Dreher, J. W. 1987, AJ, 93, 968
Rowe, J. F., Matthews, J. M., Seager, S., et al. 2006, ApJ, 646, 1241
Walker, G., Kuschnig, R., Matthews, J. M., et al. 2005, ApJ, 635, L77