Frequentist Estimation and Hypothesis Testing for the Presence of One Sinusoidal Component in Time Series Data

A.E. Charman*

Department of Physics, University of California, Berkeley

(Dated: 12/2/13)

Some frequentist estimators for the amplitude of a sinusoidal component in timer series data of given frequency are suggested.

I. STATEMENT OF THE BASIC PROBLEM

We seek a test for the presence of one sinusoidal component of a postulated frequency ω in one-dimensional time series data, and either known or unknown phase, assuming also at minimum a possible zero-frequency offset, to be treated as a nuisance parameter. We desire a method to both test for the presence of a such a sinusoidal component and eventually a means to estimate its amplitude (and optionally the phase) and the uncertainty in the estimate(s).

Specifically, suppose we make N_k measurements in each of $K \geq 1$ equal-duration time intervals, each assumed to extend over one fundamental period $\tau = \frac{2\pi}{\omega}$, and each assumed to start and end at same time modulo this period. The fundamental period may be, for example, a sidereal or CMB day or year, or a lunar month, etc.

Within each period, the data consist of a discrete set of (actual or simulated) real-valued observations z_{k_j} at times t_{k_j} measured relative to the beginning of the kth time period, where $j = 1, \ldots, N_k$ ranges over the observations in each fundamental period and $k = 1, \ldots, N_k$ ranges over the periods, not necessarily assumed to be contiguous.

We suppose that in the absence of noise the data would fall on the curve

$$z_{k_j} = \sum_{\ell=-\infty}^{+\infty} = A_\ell \, e^{-\ell \omega t_{k_j}} \tag{1}$$

where $A_{-\ell} = A_\ell^*$ since the z_{k_j} are real.

The null hypothesis is that all $A_\ell = 0$ except for A_0. Our first task will be to define an estimator for A_1 whose sampling distribution under the null hypothesis we can estimate by Monte Carlo simulation. As a first step we can then use the estimator to derive an hypothesis test for $A_1 = 0$ versus $A_1 \neq= 0$.

Conversely, we could use the estimated sampling distribution to put a confidence limit on the value of A_1.

Also, we may want to combine estimate estimates from distinct classes of experiments (specifically, with different bias electric fields).

---

*Electronic address: `acharman@physics.berkeley.edu`

II. MOMENT-BASED ESTIMATOR

Assuming any harmonics to be negligibly small compared to the fundamental and the constant offset, we have

$$z_{j_k} = A_1 e^{-\omega t_{k_j}} + A_1^* e^{+\omega t_{j_k}} + A_0 + \eta_{k_j} \tag{2}$$

where $A_0 = A_0^*$ and η_{k_j} is some random noise term, assumed IID, zero-mean, and uncorrelated with any of the expectation values we will subsequently introduce. (These assumptions need not hold for the estimator to still be useful for hypothesis testing). We can then derive an estimator for A_1 by a method of moments, where we replace distributional expectation values with sample averages. Specifically define:

$$N = \sum_k N_k, \tag{3}$$

and then define

$$\langle f \rangle = \frac{1}{N} \sum_k \sum_j f_{k_j} \tag{4}$$

for any sequences f_{k_j}. Assuming a sufficiently large sample, we expect that

$$\langle \eta \rangle \approx 0 \tag{5a}$$

$$\langle \eta \, e^{\pm i\omega t} \rangle \approx \langle \eta \rangle \, \langle e^{\pm i\omega t} \rangle \approx 0, \tag{5b}$$

so we demand

$$\langle z \rangle = A_1 \langle e^{-i\omega t} \rangle + A_1^* \langle e^{+i\omega t} \rangle + A_0 \tag{6a}$$

$$\langle z e^{+i\omega t} \rangle = A_1 + A_1^* \langle e^{+2i\omega t} \rangle + A_0 \langle e^{+i\omega t} \rangle \tag{6b}$$

$$\langle z e^{-i\omega t} \rangle = A_1 \langle e^{-2i\omega t} \rangle + A_1^* + A_0 \langle e^{-i\omega t} \rangle \tag{6c}$$

Eliminating A_0, we have

$$\begin{aligned} C(\omega) = \langle z e^{+i\omega t} \rangle - \langle z \rangle \langle e^{+i\omega t} \rangle &= A_1 \left[1 - \langle e^{-i\omega t} \rangle \langle e^{+i\omega t} \rangle \right] + A_1^* \left[\langle e^{+2i\omega t} \rangle - \langle e^{+i\omega t} \rangle \langle e^{+i\omega t} \rangle \right] \\ &= A_1 \, \Gamma_{+-}(\omega) + A_1^* \, \Gamma_{++}(\omega), \end{aligned} \tag{7}$$

plus the complex conjugate of this equation. Solving for A_1, we find our first estimator:

$$A_1 = \frac{\Gamma_{+-}(\omega) \, C(\omega) - \Gamma_{++} \, C(\omega)^*}{|\Gamma_{+-}(\omega)|^2 - |\Gamma_{++}(\omega)|^2} \tag{8}$$

However, note that while we obtained this estimator by solving equations which were linear in A_1 and A_1^*, the estimator itself is a nonlinear function of the data moments, so it will not be an unbiased estimator in general. Also note that, even though we assumed the higher harmonics were negligible, a Fourier moment at 2ω does appear in our estimator for A_1.

III. QUADRATURE-BASED ESTIMATOR

Assuming the data are a sample from a continuous function

$$z(t) = \sum_{\ell=-\infty}^{+\infty} = A_\ell \, e^{-\ell\omega t} + \eta(t) \tag{9}$$

where the noise is zero-mean and uncorrelated with any of the Fourier components, we have

$$A_1 = \frac{1}{KT} \sum_k \int_{\tau_k}^{\tau_k + T} dt \, z(t) \, e^{+i\omega t}, \tag{10}$$

where τ_k is the the starting time of the kth fundamental period of duration $T = 2\pi/\omega$.

Also assume that within each interval, the measurements have been sorted in chronological order so that

$$t_{k_j} \leq t_{k_j}. \tag{11}$$

We can then estimate each integral by the trapezoidal rule using the sample points:

$$\begin{aligned} A_1 = \tfrac{1}{KT} \sum_k \sum_{j=1}^{N_k-1} \tfrac{1}{2}\big(z_{k_j}\, e^{+i\omega t_{k_j}} + z_{k_{j+1}}\, e^{+i\omega t_{k_{j+1}}}\big)\big(t_{k_{j+1}} - t_{k_j}\big) \\ + \tfrac{1}{KT} \sum_k \tfrac{1}{2}\big(z_{k_1}\, e^{+i\omega t_{k_1}} + z_{k_{N_k}}\, e^{+i\omega t_{k_{N_k}}}\big)\big(T - [t_{k_{N_k}} - t_{k_1}]\big) \end{aligned} \tag{12}$$

where we have assumed periodic boundary conditions at the end of each integral.

This estimator is linear in the measured z_{k_j}. However, for fixed times of observation, this will not be guaranteed to be unbiased, but for a large data set its bias should be less than the previous estimator. However, note that if the data are concentrated in a small subset of the time interval, then the first and last data points carry much weight, which might degrade the accuracy. This suggests a better estimator.

IV. A BETTER INTEGRAL-BASED ESTIMATOR

Given the presumed periodicity, we can always just translate all measurement times into a a single fundamental period $0 \leq t \leq T$. Let us assume that has been done, so that now $K = 1$ but $N_1 = N$. Because statistical accuracy due to averaging several independent integrals improves much more slowly than the numerical accuracy due adding points to one trapezoidal quadrature, we expect this to lead to a more accurate estimator.

To avoid confusion with the previous formula, let us denote the measurements by $\tilde{z}_j$ and the times by $\tilde{t}_j$, for $j = 1, \ldots, N$, where

$$\tilde{t}_j \leq \tilde{t}_{j+1}. \tag{13}$$

So now we take

$$\begin{aligned} A_1 = \tfrac{1}{T} \sum_{j=1}^{N-1} \tfrac{1}{2}\big(\tilde{z}_j\, e^{+i\omega \tilde{t}_j} + \tilde{z}_{j+1}\, e^{+i\omega \tilde{t}_{j+1}}\big)\big(\tilde{t}_{j+1} - \tilde{t}_j\big) \\ + \tfrac{1}{T}\tfrac{1}{2}\big(\tilde{z}_1\, e^{+i\omega \tilde{t}_1} + \tilde{z}_N\, e^{+i\omega \tilde{t}_N}\big)\big(T - [\tilde{t}_N - \tilde{t}_1]\big) \end{aligned} \tag{14}$$

This is probably the best of the three estimators to use, at least for daily signals. For yearly signals accuracy will suffer if the data are concentrated only over a few months.

V. COMBINING RESULTS FROM DISTINCT EXPERIMENTS

I am dubious that we can assume that A_1 simply changes sign when the external electric fields are reversed. There may be other terms, depending, say, on $\boldsymbol{P}\cdot\boldsymbol{B}$ where $\boldsymbol{P}$ is the average momentum of the CMB radiation and $\boldsymbol{B}$ is the magnetic field. So I would advise against combining estimates of A_1 for distinct experimental configurations.

We can however, still combine the data when performing a hypothesis test regarding $|A_1| \neq 0$.

VI. HYPOTHESIS TESTING

The simplest way to combine the data for distinct bias fields is just to use a weighted average of, say, $|A_1|$:

$$S = \frac{N_L|A_{1_L}| + N_L|A_{1_R}|}{N_L + N_R} \tag{15}$$

where N_L and N_R are the number of data points in the two cases. From Monte Carlo Simulations, we can estimate the p-value associated with this statistic S being as large or large than that actually observed, under the null hypothesis.