Phase Estimation from Noisy Data with Gaps

Yitong HUANG * & Clark BOWMAN †& Olivia WALCH ‡& Daniel FORGER†§¶
*Department of Mathematics, Dartmouth College, NH, USA
†Department of Mathematics, University of Michigan, MI, USA
‡Department of Neurology, University of Michigan, MI, USA
§Department of Computational Medicine and Bioinformatics, University of Michigan, MI, USA
¶Michigan Institute for Data Science, University of Michigan, MI, USA
E-mail : forger@umich.edu

***Abstract*—Determining the phase of a rhythm embedded in a time series is a key step in understanding many oscillatory systems. While existing approaches such as harmonic regression and cross-correlation are effective even when some data are missing, we show that they can produce biased estimates of phase when missing data are consecutive (i.e., there is a gap). We propose a simple modification of the least-squares approach, Gap Orthogonalized Accelerated Least Squares (GOALS), which addresses this issue with a negligible increase in computational expense. We test GOALS against other approaches on a synthetic dataset and on a real-world dataset of activity recorded by an Apple Watch, showing in both cases that GOALS is effective at recovering phase estimates from noisy data with gaps.**

I. Introduction

Phase estimation of oscillatory data becomes more difficult when data are noisy or when data points are missing. Often missing data are random, or can be approximated by a random process [6], [15]. However, in many applications, missing data are continuous and form a gap, such as when a sensor breaks or when measurements can only be made at certain times. Here, we review available approaches to phase estimation with noisy data and find that gaps can significantly bias phase estimates; the goal of this paper is to correct this bias.

There are many proposed gap-filling methods for analyzing noisy oscillatory data with missing values, such as cubic splines and empirical mode decomposition [16], [14]. Instead, we propose a modification of the ordinary least squares method, Gap Orthogonalized Accelerated Least Squares (GOALS), which avoids biased phase estimates using an orthogonal basis over existing data. GOALS is comparable in computation time to ordinary least squares and significantly faster than iterative nonlinear least squares.

Rigorously, suppose data $d(t)$ are modeled as the sum of a signal $q(t,\phi)$ and noise $n(t)$:

$$d(t) = q(t,\phi) + n(t),$$

where $q(t,\phi)$ is a periodic function in time t with unknown phase ϕ. We consider the scenario in which the period is known (at least approximately) and the data correspond to roughly one period (over which the phase is assumed constant), i.e., $t \in \Omega \equiv (t_0, t_0 + \frac{2\pi}{\omega})$ in terms of the known frequency ω. When there is a gap in the data, data exist only for $t \in \Omega_g \subset \Omega$.

For notational simplicity, we also define the inner products

$$\langle x(t), y(t)\rangle \equiv \int_{\Omega} x(t')y(t')dt'$$

and

$$\langle x(t), y(t)\rangle_g \equiv \int_{\Omega_g} x(t')y(t')dt'.$$

II. Conventional methods

To illustrate the potential bias caused by gaps in noisy data, we consider several common methods for determining the phase of a rhythm. One approach is to determine when the rhythm peaks or crosses a threshold [1], [9]; since we do not know *a priori* whether peaks or threshold crossings occur during gaps in data, this approach cannot be used here. We will thus consider (1) harmonic regression [1], [2], [3], [7], [8] and (2) cross-correlation [4], [12], [13].

A. Harmonic regression (cosinor analysis)

Since the frequency ω is known, a basic sinusoidal model for the signal is given by

$$q(t,\phi) = a_1 \sin(\omega t) + a_2 \cos(\omega t),$$

where a_1 and a_2 are determined from the phase ϕ. Following the ordinary least squares approach, we seek to minimize the integrated square error

$$\|n(t)\|_2^2 = \|d(t) - a_1 \sin(\omega t) - a_2 \cos(\omega t)\|_2^2,$$

which is minimized when $\pi a_1 = \langle d(t), \sin(\omega t)\rangle$ and $\pi a_2 = \langle d(t), \cos(\omega t)\rangle$, yielding a phase $\phi = \arg(a_1 + ia_2)$.

When a gap exists, inner products must be computed over Ω_g, i.e. $\pi a_1 = \langle d(t), \sin(\omega t)\rangle_g$ and $\pi a_2 = \langle d(t), \cos(\omega t)\rangle_g$. Since $\langle \sin(\omega t), \cos(\omega t)\rangle_g \neq 0$ in general, the resulting phase estimate may not minimize $\|n(t)\|_2^2$.

B. Cross-correlation

Let $s(t)$ be a function with period $2\pi/\omega$ which models oscillations in the observed data $d(t)$ and define the modeled signal as $q(t,\phi) = s(t-\phi)$. Under this assumption, the integrated square error $\|n(t)\|_2^2$ is minimized when $\langle d(t), s(t-\phi)\rangle$ is maximized, i.e., the function $s(t)$ is shifted to best agree with the observed data. Again, when a gap exists, inner products must be computed over Ω_g; the phase ϕ which maximizes $\langle d(t), s(t-\phi)\rangle_g$ may not minimize $\|n(t)\|_2^2$.

III. Gap Orthogonalized Accelerated Least Squares (GOALS)

To account for the loss of orthogonality over Ω_g, we replace the basis functions $\cos(\omega t), \sin(\omega t)$ with an orthogonalized set of basis functions. Specifically, suppose data are observed at times $t_i \in \Omega_g$ for $i = 1, \ldots, N$ and define the matrix A as

$$A = \begin{bmatrix} 1 & \sin(\omega t_1) & \cos(\omega t_1) & \cdots \\ 1 & \sin(\omega t_2) & \cos(\omega t_2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & \sin(\omega t_N) & \cos(\omega t_N) & \cdots \end{bmatrix},$$

i.e., the matrix whose columns are the basis functions over Ω evaluated at t_i (including the constant function, $\sin(\omega t)$ and $\cos(\omega t)$, and any desired higher harmonics); the QR decomposition $A = QR$ then yields an orthonormal matrix Q whose columns are orthogonal over Ω_g.

Let $x(t)$, $y(t)$ refer to the second and third columns of Q, i.e., the columns corresponding to the newly orthogonalized basis of the first harmonic. Following a similar procedure to the least-squares minimization in Section II-A, the new phase estimate is given as

$$\phi = \arg\left(\frac{\langle x(t), d(t)\rangle_g}{\langle x(t), \sin(\omega t)\rangle_g} + i\frac{\langle y(t), d(t)\rangle_g}{\langle y(t), \cos(\omega t)\rangle_g}\right).$$

Phase estimates for higher harmonics can be obtained by repeating the above procedure for any remaining columns of the matrix Q.

IV. Examples and Results

We first illustrate the effect of consecutive gaps using synthetic data. Let

$$f(t, \phi) = -\sin\left(\frac{2\pi}{100}(t - \phi)\right)$$

define a sinusoidal signal with known period 100 and phase ϕ. For times $t_i = 1, \ldots, 100$, we generate noisy data as

$$d_i = f(t_i, \phi) + \beta_i,$$

where β_i are *i.i.d.* samples from $N(\alpha, 1)$, i.e., a standard normal with mean α and unit variance. To simulate a gap, data $d_{30}, \ldots, d_{65}$ are discarded; to simulate random data loss, 36 data d_i are discarded equally likely at random. Remaining data are used to find a phase estimate ϕ_{est}.

For each of $\alpha = 0, 0.5, 0.75, 1.0$ (i.e., varying degrees of bias) and for both consecutive and random data loss, algorithms were tested by repeating the above procedure 1000 times with values of ϕ chosen uniformly at random from 0 to 100. By plotting ϕ_{est} against ϕ, we observe the ability of each method to accurately recover phase; unbiased methods should roughly follow the diagonal $\phi_{est} = \phi$. Although here we change α which changes the mean level, similar effects could also be seen by adding in higher harmonics.

Figure 1 shows results for harmonic regression and cross-correlation (using $f(t, 0)$ as the oscillatory model $s(t)$) in the case of random data loss. Regardless of the bias α, phase estimates ϕ_{est} loosely follow the true values ϕ, confirming existing results that both methods can effectively determine the signal phase when data are distributed evenly across one period of oscillation.

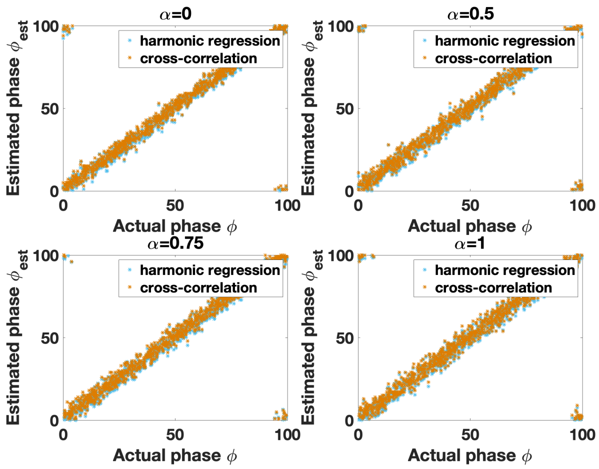

Fig. 1. Phase estimates ϕ_{est} compared to true phase ϕ for harmonic regression and cross-correlation on a synthetic dataset with missing data. For each of 1000 trials, a uniformly random phase ϕ was used to generate data d_i from a 100-period sinusoidal signal $f(t, \phi)$ and i.i.d. noise β_i sampled from a Gaussian with mean α and variance 1; one third of the data were then removed uniformly at random, with harmonic regression and cross-correlation used to estimate phase from the remaining data. Results are shown for four values of the noise bias α. Regardless of α, both methods recover phase to good accuracy (i.e., along the diagonal $\phi_{est} = \phi$).

When data are removed consecutively, both harmonic regression and cross-correlation can provide biased phase estimates (Figure 2). The most significant deviation from the diagonal, i.e., bias in phase estimation, occurs when the signal peaks during the gap ($\phi \sim 72.5$), with more severe effects when the noise bias α is greater. The error can be quantified in terms of the mean squared error (MSE):

$$\text{MSE} = \frac{1}{1000}\sum_{j=1}^{1000}\left(\phi^j - \phi^j_{est}\right)^2,$$

where ϕ^j, ϕ^j_{est} refer the true and estimated phase of the j^{th} trial. The MSE for harmonic regression and cross-correlation appear as a function of the noise bias α in the leftmost plot of Figure 3.

Figure 4 shows phase estimates using GOALS in the case of consecutively removed data. Regardless of the noise bias α, phases are recovered to good accuracy; the center plot of Figure 3 shows the MSE to be independent of α. This suggests that GOALS may be useful for phase estimation in cases where gaps in data result in a biased estimate of the signal mean. GOALS estimates for larger gap sizes appear in Figure 5; re-orthogonalization appears useful up to gaps of roughly two-thirds of a period.

The results have so far shown that harmonic regression and cross-correlation can yield biased estimates of phase when the

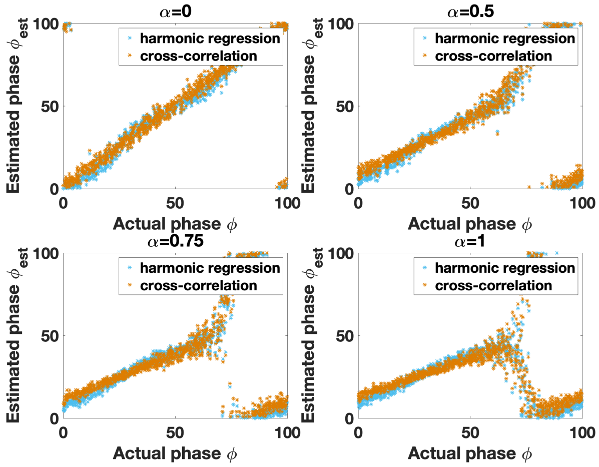

Fig. 2. Repeating the analysis of Fig. 1 when data are removed consecutively, i.e., $d_{30}, \ldots, d_{65}$ are always discarded. Phase estimates are significantly noisier than in Fig. 1 and are heavily biased when the signal peaks during a gap ($\phi \sim 72.5$). Both cross-correlation and harmonic regression are more significantly affected when the noise bias α is high.

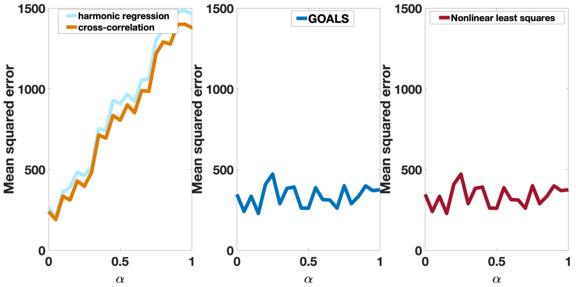

Fig. 3. Mean squared error (MSE) for (left) harmonic regression and cross-correlation, (center) Gap Orthogonalized Accelerated Least Squares (GOALS), and (right) nonlinear least squares when estimating phase in the context of Figure 2, i.e., when data are removed consecutively. MSE is calculated over a range of values for the noise bias α; while GOALS and nonlinear least squares perform equally well regardless of α, both harmonic regression and cross-correlation decrease significantly in accuracy as the noise bias approaches the magnitude of the signal ($\alpha = 1$).

noise bias α is significant; in addition to re-orthogonalization (i.e., GOALS), this concern can also be addressed by incorporating a vertical offset in, e.g., an iterative nonlinear method [10], [11]. The rightmost plot of Figure 3 shows the MSE of such a nonlinear approach to also be independent of α.

Figure 6 illustrates computation times required to estimate phase from 1000 trials of the synthetic dataset using harmonic regression, GOALS, and the Levenberg-Marquardt iterative nonlinear approach implemented in MATLAB on a 2016 Macbook Pro (2.9 GHz Intel Core i7). Larger datasets (up to 10^6 noisy samples) were also considered. GOALS is only marginally more expensive than ordinary least squares, and both are significantly cheaper than an iterative approach; all three methods approach a similar computational complexity with respect to number of data ($\sim N^{1.1209}$, $\sim N^{1.2672}$, $\sim N^{1.1877}$, respectively).

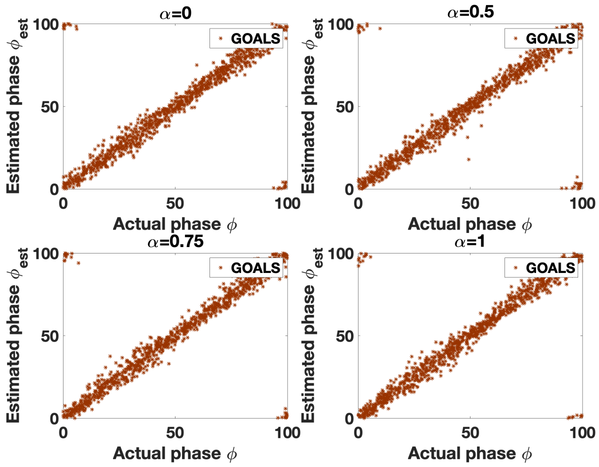

Fig. 4. Phase estimates for the synthetic dataset with gaps (Fig. 2) using Gap Orthogonalized Accelerated Least Squares (GOALS). Regardless of the noise bias α, estimated phases ϕ_{est} are in good agreement with the underlying signal phases ϕ, suggesting that the re-orthogonalized basis of GOALS can effectively avoid estimation bias due to a gap in data.

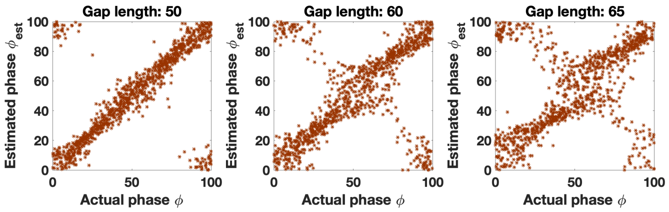

Fig. 5. GOALS phase estimates for the synthetic dataset when the gap is lengthened (up to 50, 60, 65 of 100 data consecutively removed, resp.). Noise bias α is fixed at 0.75. Phase estimates become noisier as the gap becomes larger, becoming heavily biased when roughly two-thirds of data are discarded.

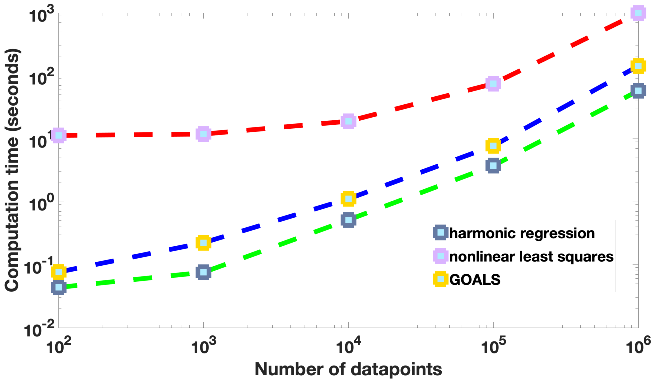

Fig. 6. Log-log comparison of computation time for harmonic regression (green), GOALS (blue), and nonlinear least squares (red) as a function of the size N of the synthetic dataset (the noisy data with gaps from Figs. 2 – 4). GOALS is comparable in expense to harmonic regression but significantly less expensive than nonlinear least squares, with a more pronounced difference for small N. The asymptotic slopes give estimates of the complexity ($\sim N^{1.1209}$, $\sim N^{1.2672}$, $\sim N^{1.1877}$, respectively). Algorithms were implemented in MATLAB on a 2016 Macbook Pro (2.9 GHz Intel Core i7).

We next apply the algorithms to a real-world dataset: roughly 800 days of motion data collected from the Apple Watch of one of the authors. Circadian phase has long been a subject of interest in the biological literature; the ability of wearables to record data such as activity, heart rate, and sleep has generated a renewed interest in understanding the relationship between circadian phase and measurable data [5]. A major difficulty in this context is the presence of regular gaps in data when wearables are removed (especially during sleep) – though we do not attempt to relate results to circadian phase, the Apple Watch data thus provide an example of real-world phase estimation from a noisy signal with gaps.

The Apple Watch dataset features "steps" data for one-hour periods throughout each day, but with roughly eight hours of data missing when the watch is removed during sleep or to recharge. The resulting dataset thus contains 788 days with ~ 16 data points each (one for each hour of observed activity); since humans are typically more active at certain times of day, the underlying signal can be thought of as oscillatory with a 24-hour period.

Daily phase estimates from cross-correlation (center) and GOALS (right) appear in Figure 7. Since a sine wave is a poor approximation for this signal, cross-correlation used a signal $s(t)$ estimated from data: for each hour of the day, the recorded activity during that hour was averaged over the entire dataset, with the resulting averages then combined to form an oscillatory model $s(t)$ tracking average motion throughout the day. Phase estimates from cross-correlation were dominated by the timing of the gap in data and were largely unaffected by day-to-day variation in activity; in contrast, GOALS recovered roughly the same phase on average (i.e., activity tends to be highest in the middle of the day) but was much more effective at following the large variation in daily activity.

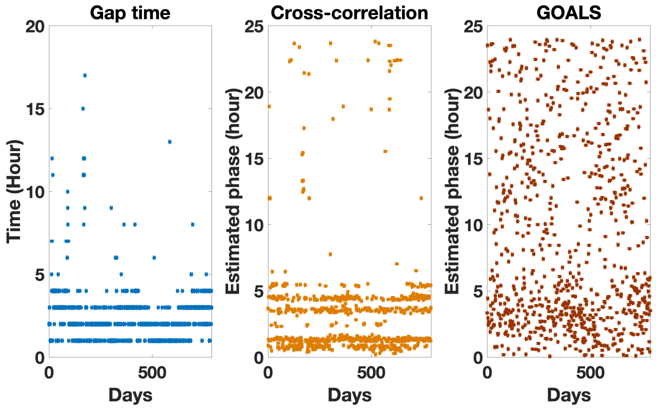

Fig. 7. Estimated phase in activity using cross-correlation (center) and GOALS (right) for a 788-day dataset of steps data recorded from an Apple Watch. Cross-correlation used as its oscillatory model $s(t)$ the average activity for each hour in the dataset. Since the watch was reguarly removed during sleep and to recharge, significant gaps exist in the recorded data (left). Phases estimated by cross-correlation follow closely the timing of data gaps and were not majorly affected by daily variations in recorded steps, while GOALS was significantly more effective at tracking variations in day-to-day activity (i.e., was less biased toward the timing of gaps).

V. Conclusion

Harmonic regression and cross-correlation are commonly used in phase estimation from noisy time series; we have shown that these methods can yield substantially biased estimates of phase when there exist large gaps in observed data. We instead propose a slight modification of the least-squares approach using a new orthogonal basis obtained from QR factorization. GOALS is computationally efficient (comparable to ordinary least squares and significantly faster than iterative nonlinear approaches) and avoids biased phase estimates due to gaps; due to its ease of implementation, it is a simple and practical solution for the estimation of phase from noisy time series with gaps.

VI. Acknowledgments

This work was supported by NSF DMS award #1714094.

References

[1] J. Broadway, J. Arendt & S. Folkard *Bright light phase shifts the human melatonin rhythm during the Antarctic winter.* Neuroscience letters **79** (1987), 185–189.

[2] G. Cornelissen *Cosinor-based rhythmometry.* Theor Biol Med Model **11** (2014), 16.

[3] M. Drennan, D. Kripke, & J. Gillin *Bright light can delay human temperature rhythm independent of sleep.* American Journal of Physiology-Regulatory, Integrative and Comparative Physiology **257** (1989), R136–R141.

[4] S. Folkard, D. Minors, & J. Waterhouse *Demasking the Temperature Rhythm after Simulated Time Zone Transitions.* Journal of biological rhythms **6** (1991), 81–91.

[5] D. Forger *Biological Clocks, Rhythms and Oscillations: The Theory of Biological Timekeeping.* The MIT Press (2017) .

[6] J. Kim, & J. Curry *The Treatment of Missing Data in Multivariate Analysis.* Sociological Methods and Research **6** (1977), 215–240.

[7] E. Klerman, Y. Lee, C. Czeisler, & R. Kronauer *Linear demasking techniques are unreliable for estimating the circadian phase of ambulatory temperature data.* Journal of biological rhythms **14** (1999), 260–274.

[8] E. Klerman, W. Wang, A. Phillips, & M. Bianchi *Statistics for Sleep and Biological Rhythms Research: longitudinal analysis of biological rhythms data.* Journal of biological rhythms **32** (2017), 18–25.

[9] H. Klerman, M. Hilaire, R. Kronauer, J. Gooley, C. Gronfier, J. Hull, S. Lockley, N. Santhi, W. Wang, & E. Klerman *Analysis method and experimental conditions affect computed circadian phase from melatonin data.* PloS one **7** (2012), e33836.

[10] K. Levenberg *A Method for the Solution of Certain Problems in Least-Squares.* Quarterly Applied Mathematics **2** (1944), 164–168.

[11] D. Marquardt *An Algorithm for Least-squares Estimation of Nonlinear Parameters.* SIAM Journal Applied Mathematics **11** (1963), 431–441.

[12] J. Mills, D. Minors, & J. Waterhouse *Adaptation to abrupt time shifts of the oscillator(s) controlling human circadian rhythms.* The Journal of physiology **285** (1978), 455–470.

[13] D. Minors, S. David,& J. Waterhouse *The use of constant routines in unmasking the endogenous component of human circadian rhythms.* Chronobiology international **1** (1984), 205–216.

[14] A. Moghtaderi, P. Borgnat , & P. Flandrin *Gap-filling by the empirical mode decomposition.* 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2012), 3821–3824.

[15] T. Orchard, & M. Woodbury *A missing information principle: theory and applications.* Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability **1** (1972), 697–715.

[16] G. Papadopoulos,& D. Kugiumtzis *Estimation of connectivity measures in gappy time series.* Physica A: Statistical Mechanics and its Applications **436** (2015), 387–398.