Randomization Inference of Periodicity in Unequally Spaced Time Series with Application to Exoplanet Detection

Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business

Joint work with Jacob Bean (U Chicago, Astrophysics)

Periodic extinctions? (Raup and Sepkoski, 1986)

Introduction

The estimation of periodicity is a fundamental task in science; e.g., astrophysics/astronomy, paleontology, biology, climate science.

The problem is deceptively simple, however. Standard methods require

- equal or i.i.d. spacings between observation times, and that
- common estimators —e.g., periodogram peaks— are consistent and asymptotically normal.

In practice, these conditions are unrealistic: observation times exhibit patterns while common estimators can substantially deviate from normality.

It is unclear how inference should proceed in such settings.

Motivation

Our work is motivated by the analysis of radial velocity data in exoplanet detection.

High-resolution observatories have made ground-breaking exoplanet detections, including “51 Pegasi b” (Mayor and Queloz, 1995) awarded the 2019 Nobel Prize in Physics.

Recently, a potential discovery was announced in our immediate stellar neighborhood of α Centauri (Anglada-Escude et.al., 2016). This is astonishing as it suggests that exoplanets may be ubiquitous.

Despite these successes, the underlying statistical methods need improvement (or complete overhaul). False discoveries are not uncommon!

Our contributions

Methodological contribution: We develop a set identification method to infer hidden periodicity, θ^*. The idea is to construct a confidence set, $\Theta_{1-\alpha}$, with correct finite-sample coverage:

$$\mathrm{pr}(\theta^* \in \Theta_{1-\alpha}) \geq 1 - \alpha. \quad (1)$$

This construction does not require normality, not even consistency, of the underlying statistic. It also does not require normality or i.i.d. errors, and can seamlessly work with equally or unequally spaced data.

Practical contribution: Our method gives sharp inference on the confirmed exoplanets in our sample. However, it raises doubts for other recent —yet unconfirmed— exoplanets.

Finally, we suggest ways to improve the observation design for sharper inference, which could help with future discoveries.

Potential downside: our method is (quite) computationally intensive. But it can be parallelized.

Outline

1. Background. Detection and estimation of periodicity. Challenges.
2. Main method (parametric).
3. Application: Exoplanet detections.
4. Improving observation designs.

5. (if time): Nonparametric method. Details
6. (if time): Implications for statistical inference. Covid-19 application. Details

Table of Contents

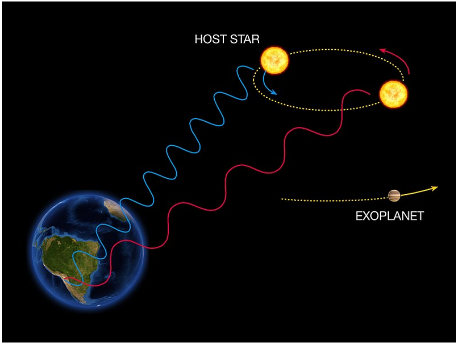
Exoplanet detection in practice

An planet orbiting a star affects the star's emitted light (Doppler effect).

On Earth, we observe regular changes in the star's spectrum.

From these changes we infer the star's radial velocity.

Oscillations in the radial velocity are then attributed to the presence of an exoplanet.



Two main steps in this process:

- Detection of periodicity.
- Estimation of periodicity (if detection was successful).

Data and observation design

Our data are (T^n, Y^n) comprised of

$$\begin{aligned} T^n &= (t_1, \dots, t_n) \quad \text{observation times} \\ Y^n &= (y_1, \dots, y_n) \quad \text{radial velocity measurements.} \end{aligned}$$

The differences $t_i - t_{i-1}$ are the spacings between observation times.

The distribution pr(T^n) on T^n is the observation design and biases towards summer, night, etc. Thus, observations usually exhibit deterministic patterns (e.g., 1-day periodicities).

In earlier work, the spacings are assumed either equal or unequal but i.i.d.

We make a more mild assumption:

$$T^n \perp\!\!\!\perp Y^n. \tag{A1}$$

Detecting periodicity — The periodogram

Standard methods with equal spacings are based on the periodogram (Schuster, 1898). Extension to unequal spacings by Lomb (1976) and Scargle (1982).

Suppose that the following harmonic model is ground-truth:

$$y_i = \psi_1^* + \psi_2^* \cos(2\pi t_i/\theta^*) + \psi_3^* \sin(2\pi t_i/\theta^*) + \varepsilon(t_i) \equiv \underbrace{y^{\mathrm{p}}(t_i;\theta^*,\psi^*)}_{\text{periodic component}} + \underbrace{\varepsilon(t_i)}_{\text{error component}} .$$

Here, $\theta^* \in \Theta$ is the unknown period and $(\psi_1^*, \psi_2^*, \psi_3^*) \equiv \psi^* \in \Psi$ are nuisance parameters.

Then, the generalized Lomb–Scragle (LS) periodogram is defined as:

$$A_n(\theta) = \frac{L_{0n} - L_n(\theta, \hat{\psi}_\theta)}{L_{0n}}, \;\; A_n : \Theta \to \mathbb{R},$$

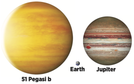
where

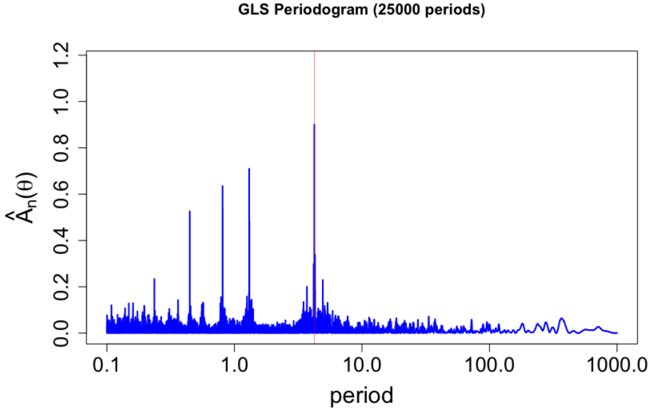
$$L_n(\theta,\psi) = \sum_{i=1}^{n} [y_i - y^{\mathrm{p}}(t_i|\theta,\psi)]^2/\sigma_i^2. \qquad \text{(squared loss / normal likelihood)}$$

$$\hat{\psi}_\theta = \arg\min_{\psi\in\Psi} L_n(\theta,\psi) \qquad \text{(cf. profile likelihood)}$$

$$L_{0n} = \sum_{i=1}^{n} (y_i - \bar{y})^2/\sigma_i^2. \qquad \text{(baseline fit).}$$

Illustration: periodogram from 51 Pegasi b





Fourier power spectrum over periods (1/frequency). Peaks and aliases visible.

But likelihood is non-smooth and multimodal ⇒ Problems for inference (coming up).

Detecting periodicity — Periodogram peak

Main method developed by Fisher (1929). Power refined by (Siegel, 1980; Bolviken, 1983; Chiu, 1989), and extended to more general hypotheses (Juditsky et al., 2015) and sparse alternatives (Cai et al., 2016).

Most methods rely on the periodogram peak, $\hat{\theta}_n = \arg\max_{\theta \in \Theta} A_n(\theta)$.

Idea is to reject the null of no periodicity when the peak exceeds a threshold ("false alarm probability"). See also (Baluev, 2008, 2013; Delisle et al., 2020; Nemec and Nemec, 1985) for adaptations in astronomy.

Under normality assumptions, each $A_n(\theta)$ is associated to a χ^2_2, and so the distribution of $\hat{\theta}_n$ (under the null) can be approximated via extreme value theory.

Detection of periodicity is generally robust and poses no major challenges.

Estimating periodicity

Estimation of periodicity is more involved, however. A common mistake in practice is to interpret detection of periodicity with θ^* being "near $\hat{\theta}_n$".

This implicitly relies on standard asymptotics of the form $\sqrt{n}(\hat{\theta}_n - \theta^*) \to N$ ...

However, in the harmonic model the typical CLT assumptions are implausible:

- Likelihood is irregular, non-smooth and multimodal $\Rightarrow$ Sampling distribution of $\hat{\theta}_n$ may substantially deviate from normal!
- Observation times are not entirely random $\Rightarrow$ Consistency is not guaranteed.
- Other pernicious effects from "hyperparameters" such as the granularity of Θ.

Bayesian methods could resolve these issues? Many reasons why not.. Details

Example 1: Synthetic data

Let $t_i = i + 0.05U_i$, $i = 1, \ldots, 100$, and $y_i = 1.5\cos(2\pi t_i/\sqrt{2}) + \varepsilon_i$, where $U_i \sim$ Unif$[-1, 1]$ and $\varepsilon_i \sim N(0, 1)$ i.i.d. So, $\theta^* = \sqrt{2} \approx 1.414$.

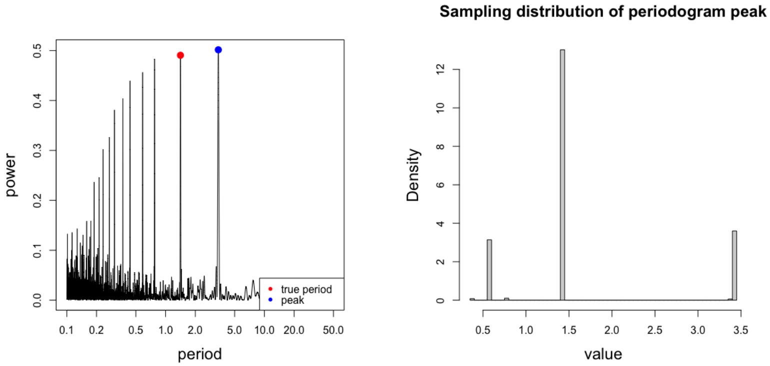


Figure: Left: Periodogram from one problematic dataset. Right: Sampling distribution of the periodogram peak from the same model over 1,000 replications.

Example 2: Real data from α Centauri B

Take (T^n, Y^n) from (Dumusque et.al., 2012). Sample assuming that $\theta^* = \hat{\theta}_n$.

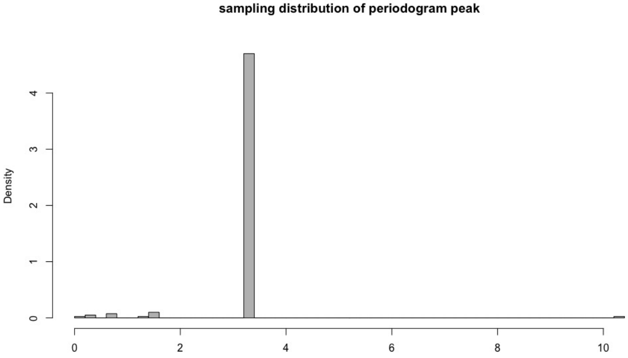


Figure: Sampling distribution of periodogram peak on a grid of $|\Theta| = 10,000$ periods.

Example 2: Real data from α Centauri B (different Θ)

Take (T^n, Y^n) from (Dumusque et.al., 2012). Sample assuming that $\theta^* = \hat{\theta}_n$.

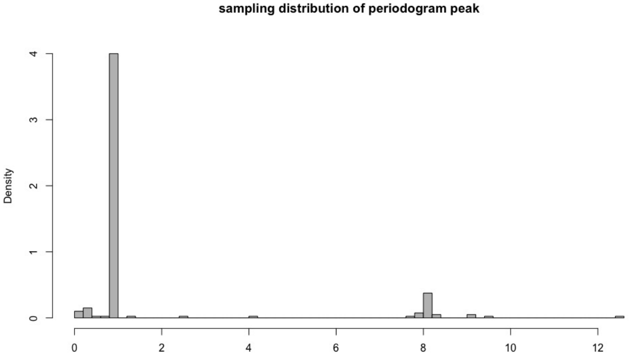


Figure: Sampling distribution of periodogram peak on a grid of $|\Theta| = 2,000$ periods.

Table of Contents

Main method: A global null

We propose to do inference conditional on T^n (and Θ). Start with the following "global null":

$$H_0^{\text{full}}:\ \theta^* = \theta_0, \psi^* = \psi_0.$$

The null implies exact values for the periodic component:

$$Y_0^{n,\text{p}} = [y^{\text{p}}(t_1;\theta_0,\psi_0),\ldots,y^{\text{p}}(t_n;\theta_0,\psi_0)].$$

and the errors

$$\varepsilon^n = Y^n - Y_0^{n,\text{p}},\ \text{where}\ \varepsilon^n = [\varepsilon(t_1),\ldots,\varepsilon(t_n)].$$

Thus, we can test H_0^{full} based on general assumptions on the errors via randomization tests. Background

Error invariance

Our inference will rely on certain invariance assumptions on the errors.

Specifically, for any observation times $T^n = \{t_1, \ldots, t_n\}$, with n finite, there exists an algebraic group $\mathcal{G}^n$ of $n \times n$ matrices such that

$$\mathrm{g} \cdot \varepsilon^n \overset{\mathrm{d}}{=} \varepsilon^n \mid T^n \quad (\mathrm{g} \in \mathcal{G}^n). \tag{A2}$$

To keep things simple, we assume that $\mathcal{G}^n = [\pm]^{n \times n}$, the set of $n \times n$ diagonal matrices with ± 1 in the diagonal.

As such, our inference works with any symmetric distribution of independent errors beyond just Gaussian that is frequently assumed in practice.

This formulation follows the framework of randomization tests (Lehmann and Romano, 2006) where testing is based on structural rather than analytical assumptions.

Example of "structured inference". Details

Testing the global null, H_0^{full}

Define a test statistic $S_n = s_n(Y, T)$ and let $s_{\text{obs}} = s_n(Y^n, T^n)$ denote the observed value in the sample (e.g., periodogram peak).

To construct the null distribution of S_n, we generate data as follows:

$$Y^{n,(i)} = Y_0^{n,\text{p}} + G^{(i)} \cdot (Y^n - Y_0^{n,\text{p}}), \text{ with } G^{(i)} \sim \text{Unif}(\mathcal{G}^n).$$

Then, a p-value for H_0^{full} is:

$$\text{pval}(\theta_0, \psi_0) = E\{s_n(Y^{n,(i)}, T^n) \geq s_{\text{obs}}\}, \tag{2}$$

where the expectation is with respect to $G^{(i)}$ while Y^n, T^n are fixed.

Testing the global null, H_0^{full}

Define a test statistic $S_n = s_n(Y, T)$ and let $s_{\text{obs}} = s_n(Y^n, T^n)$ denote the observed value in the sample (e.g., periodogram peak).

To construct the null distribution of S_n, we generate data as follows:

$$Y^{n,(i)} = Y_0^{n,\text{p}} + G^{(i)} \cdot (Y^n - Y_0^{n,\text{p}}), \text{ with } G^{(i)} \sim \text{Unif}(\mathcal{G}^n).$$

Then, a p-value for H_0^{full} is:

$$\text{pval}(\theta_0, \psi_0) = E\{s_n(Y^{n,(i)}, T^n) \geq s_{\text{obs}}\}, \tag{2}$$

where the expectation is with respect to $G^{(i)}$ while Y^n, T^n are fixed.

Theorem

Suppose that Assumptions (A1)-(A2) hold. Then, the p-value in (2) is exact in finite samples under H_0^{full}, that is, for any finite $n > 0$,

$$\text{pr}\left\{\text{pval}(\theta_0, \psi_0) \leq \alpha \mid H_0^{\text{full}}\right\} = \alpha.$$

A confidence set for θ^*

However, ψ^* is usually a nuisance parameter. We may only want to test for θ^*:

$$H_0 : \theta^* = \theta_0. \tag{3}$$

We can reject H_0 (in a conservative way) by checking $\max_{\psi \in \Psi} \text{pval}(\theta, \psi) \leq \alpha$.

This test can also be inverted, in principle, to build a confidence set for θ^*:

$$\Theta_{1-\alpha} = \left\{ \theta \in \Theta : \max_{\psi \in \Psi} \text{pval}(\theta, \psi) > \alpha \right\}. \tag{4}$$

A confidence set for θ^*

However, ψ^* is usually a nuisance parameter. We may only want to test for θ^*:

$$H_0 : \theta^* = \theta_0. \tag{3}$$

We can reject H_0 (in a conservative way) by checking $\max_{\psi \in \Psi} \text{pval}(\theta, \psi) \leq \alpha$.

This test can also be inverted, in principle, to build a confidence set for θ^*:

$$\Theta_{1-\alpha} = \left\{ \theta \in \Theta : \max_{\psi \in \Psi} \text{pval}(\theta, \psi) > \alpha \right\}. \tag{4}$$

Theorem

Suppose that Assumptions (A1)-(A2) hold. Then, $\Theta_{1-\alpha}$ is a finite-sample valid $100(1-\alpha)\%$ confidence set for θ^*; i.e., for any finite $n > 0$,

$$\text{pr}(\theta^* \in \Theta_{1-\alpha}) \geq 1 - \alpha.$$

An approximate confidence set for θ^*

Maximizing over Ψ may be expensive. To test H_0 efficiently we can just plug in $\hat{\psi}_{\theta_0}$ and use the following p-value:

$$\widehat{\text{pval}}(\theta_0) = E\{s_n(\hat{Y}^{n,(i)}, T^n) \geq s_{\text{obs}}\}, \tag{5}$$

where

$$\hat{Y}^{n,(i)} = \hat{Y}_0^{\text{p}} + G^{(i)} \cdot (Y^n - \hat{Y}_0^{\text{p}}), \text{ and } \hat{Y}_0^{\text{p}} = [y^{\text{p}}(t_1; \theta_0, \hat{\psi}_{\theta_0}), \ldots, y^{\text{p}}(t_n; \theta_0, \hat{\psi}_{\theta_0})]. \tag{6}$$

The following construction for the confidence set of θ^* is valid asymptotically:

$$\hat{\Theta}_{1-\alpha} = \left\{\theta \in \Theta : \widehat{\text{pval}}(\theta) > \alpha\right\}. \tag{7}$$

Theorem

Suppose that Assumptions (A1)-(A2) hold, and that $\hat{\psi}_{\theta_0} \xrightarrow{\text{p}} \psi^*$ under H_0. Then, $\hat{\Theta}_{1-\alpha}$ is an asymptotically valid $100(1-\alpha)\%$ confidence set for θ^*; i.e., as n increases

$$\text{pr}(\theta^* \in \hat{\Theta}_{1-\alpha}) \geq 1 - \alpha + o_P(1).$$

Concrete procedure

1. Choose a grid of possible period values, Θ, that contains θ^* w.p. 1.
   Set $\hat{\Theta}_{1-\alpha} \leftarrow \emptyset$. Pick a test statistic, s_n.

2. Obtain data (Y^n, T^n), possibly after removing known stellar signals, e.g., rotational periods, magnetic cycles, etc. (Feigelson and Babu, 2012).

3. For all $\theta_0 \in \Theta$ do:
   (i) Estimate the nuisance parameters, $\hat{\psi}_{\theta_0} = \arg\min_{\psi \in \Psi} L_n(\theta_0, \psi)$, through weighted least squares.
   (ii) Calculate the observed value, s_{obs}, of the test statistic.
   (iii) With fixed T^n, sample new data, $Y^{n,(i)}$, where $i = 1, \ldots, R$ for some fixed R, by flipping the signs of residuals.
   (iv) Using the samples from 3(iii), calculate the p-value (5), and if it exceeds α then include θ_0 in the confidence set; i.e., set $\hat{\Theta}_{1-\alpha} \leftarrow \hat{\Theta}_{1-\alpha} \cup \{\theta_0\}$ if $\widehat{\text{pval}}(\theta_0) > \alpha$.

4. Return $\hat{\Theta}_{1-\alpha}$ as the $100(1-\alpha)\%$ confidence set of θ^*.

Discussion

Advantages

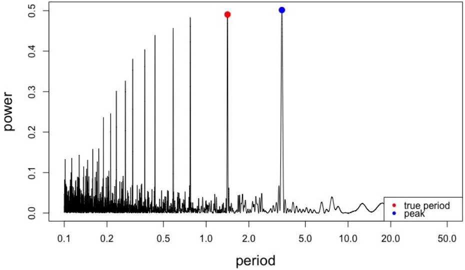
- The confidence set $\Theta_{1-\alpha}$ is valid in finite samples. The confidence set $\hat{\Theta}_{1-\alpha}$ is approximately so.
- No assumption is made for the test statistic. Not necessary to be "well-behaved" (e.g., consistent or normal).
- No assumption on the observation design or spacings.
- Inference conditional on hyperparameters (e.g., Θ).

Challenges

- Choice of test statistic. Details
- Computational challenges (procedure requires computation over entire Θ). Details

Example 1: Synthetic data — What does our method produce?

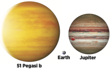
Let $t_i = i + 0.05U_i$, $i = 1, \ldots, 100$, and $y_i = 1.5\cos(2\pi t_i/\sqrt{2}) + \varepsilon_i$, where $U_i \sim \text{Unif}[-1, 1]$ and $\varepsilon_i \sim N(0, 1)$ i.i.d. So, $\theta^* = \sqrt{2} \approx 1.414$.

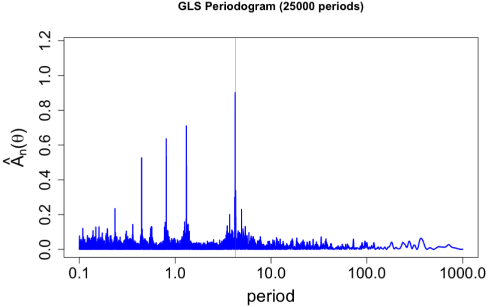


θ_0	p-value	$\hat{\Theta}_{0.95}$	$\hat{\Theta}_{0.99}$
0.1752	0.00	no	no
0.1890	0.00	no	no
0.2124	0.00	no	no
0.2330	0.00	no	no
0.2696	0.00	no	no
0.3036	0.00	no	no
0.3693	0.00	no	no
0.4362	0.00	no	no
0.5857	0.03	no	yes
0.7737	0.17	yes	yes
1.4130	0.48	yes	yes
3.4175	1.00	yes	yes

Table of Contents

51 Pegasi b (Mayor and Queloz, 1995)

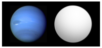


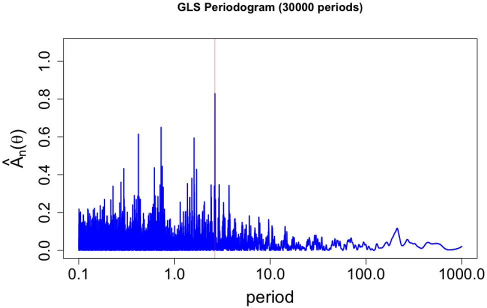


θ_0	p-value	$\hat{\Theta}_{0.95}$	$\hat{\Theta}_{0.99}$
0.3085	0.00	no	no
0.5662	0.00	no	no
0.8069	0.00	no	no
0.8089	0.00	no	no
0.8295	0.00	no	no
1.3047	0.00	no	no
1.3095	0.00	no	no
3.7033	0.00	no	no
4.1807	0.00	no	no
4.2311	1.00	yes	yes
4.2821	0.00	no	no
4.9331	0.00	no	no

Left: Periodogram of radial velocity on exoplanet "51Pegb". Here, $\Theta = \{0.1, \ldots, 1000\}$ is split uniformly in the log-space so that $|\Theta| = 25,000$. Right: Inference of periodicity of 51Pegb based on Procedure 1. The table shows the p-values for the hypothesis $H_0 : \theta^* = \theta_0$ for values of θ_0 that correspond to high peaks of the periodogram shown on the left.

We see that there are no identification issues as the 4.23-day signal is the only one accepted in the confidence sets.

Gliese 436 b (Butler et al., 2004) 

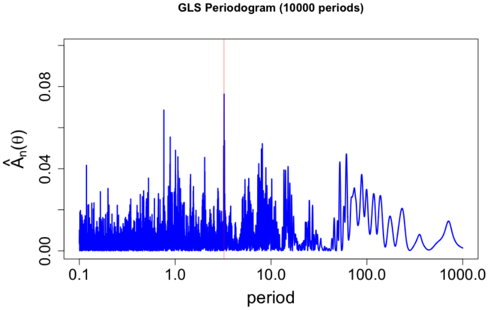


θ_0	p-value	$\hat{\Theta}_{0.95}$	$\hat{\Theta}_{0.99}$
0.4200	0.00	no	no
0.6155	0.00	no	no
0.7067	0.00	no	no
0.7438	0.00	no	no
1.3641	0.00	no	no
1.5187	0.00	no	no
1.6013	0.00	no	no
1.6086	0.00	no	no
1.7008	0.00	no	no
2.4103	0.00	no	no
2.6441	1.00	yes	yes
3.7092	0.0000	no	no

Left: Periodogram of radial velocity on exoplanet GJ436b. Here, $\Theta = \{0.1, \ldots, 1000\}$ is split uniformly in the log-space so that $|\Theta| = 30,000$. Right: Inference of periodicity based on Procedure 1. The table shows the p-values for the hypothesis $H_0 : \theta^* = \theta_0$ for values of θ_0 that correspond to high peaks of the periodogram shown on the left.

We see that there are no identification issues as the 2.64-day signal is the only one accepted in the confidence sets.

α Centauri B (Dumusque et.al., 2012)

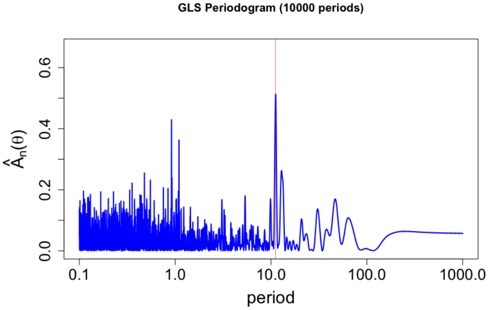


θ_0	p-value	$\hat{\Theta}_{0.95}$	$\hat{\Theta}_{0.99}$
0.7622	0.0705	yes	yes
0.8882	0.0271	no	yes
1.0086	0.0174	no	yes
1.0678	0.0079	no	no
2.0292	0.0122	no	yes
3.2074	0.0163	no	yes
3.2371	1.0000	yes	yes
3.2670	0.0178	no	yes
7.9394	0.0116	no	yes
8.1169	0.0175	no	yes
52.2242	0.0121	no	yes
61.1334	0.0226	no	yes

Left: Periodogram of radial velocity on candidate exoplanet orbiting α Centauri B. Here, $\Theta = \{0.1, \ldots, 1000\}$ is split uniformly in the log-space, so that $|\Theta| = 10,000$. Right: The table shows the p-values for the hypothesis $H_0 : \theta^* = \theta_0$ for values of θ_0 that correspond to high peaks of the periodogram shown on the left.

We see that there are severe identification issues as several signals other than the periodogram peak are accepted in the confidence sets.

Proxima Centauri (Anglada-Escude et.al., 2016)



θ_0	p-value	$\hat{\Theta}_{0.95}$	$\hat{\Theta}_{0.99}$
0.1106	0.0007	no	no
0.3355	0.0022	no	no
0.3552	0.0055	no	no
0.4778	0.0025	no	no
0.5512	0.0047	no	no
0.7532	0.0052	no	no
0.8412	0.0059	no	no
0.9164	0.0173	no	yes
0.9266	0.0005	no	no
1.0957	0.0080	no	no
11.1739	1.0000	yes	yes
12.8769	0.0006	no	no

Left: Periodogram of radial velocity on candidate exoplanet Proxima Centauri b. Here, $\Theta = \{0.1, \ldots, 1000\}$ split regularly in the log-space, so that $|\Theta| = 10,000$. Right: The table shows the p-values for the hypothesis $H_0 : \theta^* = \theta_0$ for values of θ_0 that correspond to high peaks of the periodogram shown on the left.

We see that there are no severe identification issues. The detection appears to be robust except for a nuisance signal at 0.9164 days.

Table of Contents

Observation designs

The importance of observation times in identifying a periodic signal is well understood (Feigelson and Babu, 2012; VanderPlas, 2018; Ivezic et al., 2014).

Surprisingly, there is little (to none) work in the statistical aspects of careful observation design.

Our method makes a contribution to this problem as well. The idea is simply to synthesize data under alternative designs, and then pick the design that yields "ϵ-identification"; i.e., $\hat{\Theta}_{1-\alpha}$ only contains values ϵ-away to a candidate signal θ_*^{cand}.

We address two questions:

1. How much to randomize observation times for ϵ-identification?
2. How many more observations to make for ϵ-identification?

(Candidate) Exoplanet	Design (A): randomness needed for identification (best δ)	Design (A): $\pm$ hrs.	Design (B): #additional obs. needed for identification (best $n' - n$)
51 Pegasi b	0	0	0
Gliese 436 b	0	0	0
α Proxima B	0.18	4.32	137
Proxima Centauri	0.06	1.44	17

Table: Observation designs (A) and (B) to achieve identification in the exoplanet applications. Design (A) introduces randomness in the observation times, while design (B) introduces additional observations.

We see that 51Pegb and GJ436b require no improvement in the observation times.

For α Centauri B: We need an additional variation of ± 0.18 days around the actual observation times (i.e., ± 4.32 hrs./observation). Alternatively, we need 137 new observations with a random variation of ± 15 mins./observation.

For Proxima Centauri: We need an additional variation of ± 0.06 days (i.e., ± 1.44 hrs./observation) on the actual observation times. Alternatively, we only need an 17 additional observations with a random variation of ± 15 mins./observation.

Concluding remarks

- We developed a method of set identification for hidden periodicity in unequally spaced time series. Structured inference approach. Details
- This approach is more appropriate than standard methods of statistical inference because common estimators, such as the periodogram peak, are not well-behaved and may even be inconsistent.
- We validated empirically our method in examples from exoplanet detection using radial velocity data. Inference appears not to be conservative. It also conclusively raises red flags for some recent high-profile detections.
- Our method suggests ways to improve the observation designs, either by randomizing observation times or just adding new observations. These designs could help in scheduling observation times for future discoveries.

Thank You.

Toulis, P. and Bean, J. (2021). Randomization Inference of Periodicity in Unequally Spaced Time Series with Application to Exoplanet Detection (working paper)

Toulis, P. (2020). Estimation of Covid-19 prevalence from serology tests: A partial identification approach. Journal of Econometrics, 220(1), pp. 193-213.

Bayesian methods?

We might expect that a Bayesian approach could address these issues.

However, a Bayesian approach also faces problems.

(i) Prior specification: uniform priors give preference to parameter regions that not only have high likelihood but are also wide. This sweeps the identification problem "under the rug"; see also (Hall and Yin, 2003, Section 1).

(ii) Posterior summarization is challenging when the likelihood is multimodal and non-smooth. Also affected by hyperparameters (e.g., Θ.)

(iii) Model selection: Bayes factors may strongly depend on features that are esoteric to the specified models. See also (Gelman and Yao, 2020, Sections 3 and 6).

Go back

Structured inference

Suppose we want to estimate parameter $\theta^* \in \Theta$ through a statistic S.

Typical asymptotic approach for inference is to derive a law $\sqrt{n}(S - \theta^*) \to \ldots$ and then pivot to CIs. Relies on asymptotics and usually normality.

However, we can do finite-sample valid inference if we know that

$$\mathrm{g}S \overset{\mathrm{d}}{=} S,$$

for some transformation g, via inversion of randomization tests.

The simplest case is when we have access to $f(S \mid \theta)$, the distribution of S. Then, we can build a finite-sample valid confidence set for θ^* (cf. Neyman construction):

Construct 95% confidence set:

$$\widehat{\Theta} = \left\{ \theta \in [0,1]^3 : \sum_{s \in \mathbb{S}} \mathbb{I}\{f(s|\theta) \leq f(s_{\mathrm{obs}}|\theta)\} f(s|\theta) > 0.05 \right\}.$$

In words: "accept all θ for which there is at least 5% of the density mass of $f(S|\theta)$ below $f(s_{\mathrm{obs}}|\theta)$". Outline or Global null

Comparison with standard methods

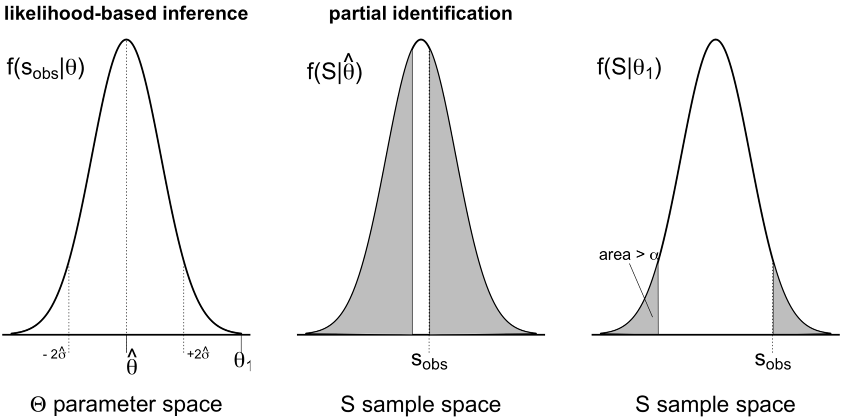
For standard methods:

- Focus is on $f(s_{\text{obs}}|\theta)$ as a function of θ (likelihood-centric).
- Inference "happens around the mode", $\hat{\theta} = \arg\max_\theta f(s_{\text{obs}}|\theta)$. Tails of likelihood are ignored.
- The "hope" is that $\hat{\theta}$ is near θ_0. Asymptotics and approximations are necessary.
- Many problems (usually undetected) when #samples is small, likelihood is multimodal, nonsmooth, modes are not separable, etc. (think of exoplanet detection!).

For structured inference methods:

- Focus is on $f(S|\theta)$ as a function of S or on invariances $gS \stackrel{\text{d}}{=} S$.
- Inference "happens everywhere" in the parameter space. The likelihood value of $f(s_{\text{obs}}|\theta)$ only matters relatively to other values $f(S|\theta)$.
- No asymptotics or approximations are necessary.
- Finite sample guarantee: Works even when #samples is small, likelihood is multimodal, nonsmooth etc.
- Downside: requires computation over entire Θ and possible over $\mathbb{S}$ (sample space).

Illustrative comparison



Go back or Covid-19 application

Covid-19 serology model

We have two calibration studies and one main study:

observed values

$$S_c^- = \#\text{positives in calibration study out of 401 true negatives} \qquad s_c^- = 2;$$
$$S_c^+ = \#\text{positives in calibration study out of 197 true positives} \qquad s_c^+ = 178;$$
$$S_m = \#\text{positives in main study out of 3,330 trials} \qquad s_m = 50.$$

Assume:

$$\begin{aligned} \text{pr(positive result}|\text{actual negative)} &= p \quad \text{[false positive rate]} \\ \text{pr(positive result}|\text{actual positive)} &= q \quad \text{[true positive rate]} \\ \frac{\text{\# actual positives in main study}}{3,330} &= \pi \quad \text{[prevalence]}. \end{aligned} \tag{8}$$

Parameter $\theta = (p, q, \pi) =\in [0,1]^3$, and statistic $S = (S_c^-, S_c^+, S_m) \in \mathbb{S}$.

Key observation: We can calculate the density, $f(S|\theta)$, of the statistic exactly.

Covid-19 serology model

Setup: $\theta = (p, q, \pi) =$ (FPR, TPR, prevalence), data $S = (S_c^-, S_c^+, S_m)$.

Density of data statistic.

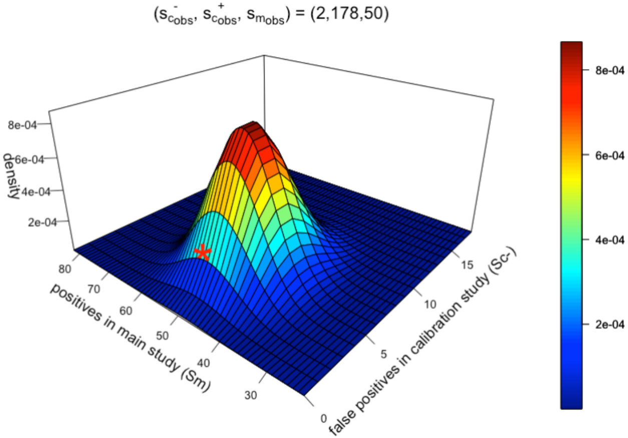
$$f(S|\theta) = \underbrace{\text{Bin}(S_c^-; 401, p)}_{\text{FP in calibration}} \cdot \underbrace{\text{Bin}(S_c^+; 197, q)}_{\text{TP in calibration}} \cdot \underbrace{\sum_i \text{Bin}(i; N_\pi, q) \cdot \text{Bin}(S_m - i; N - N_\pi, p)}_{\text{prob of } S_m \text{ positives out of } N_\pi \text{ actual positives in main study}},$$

where $N_\pi = 3300\pi =$ #actual positives in main study.

- In the sample, we observe $s_{\text{obs}} = (2, 178, 50)$. How to do inference on θ?

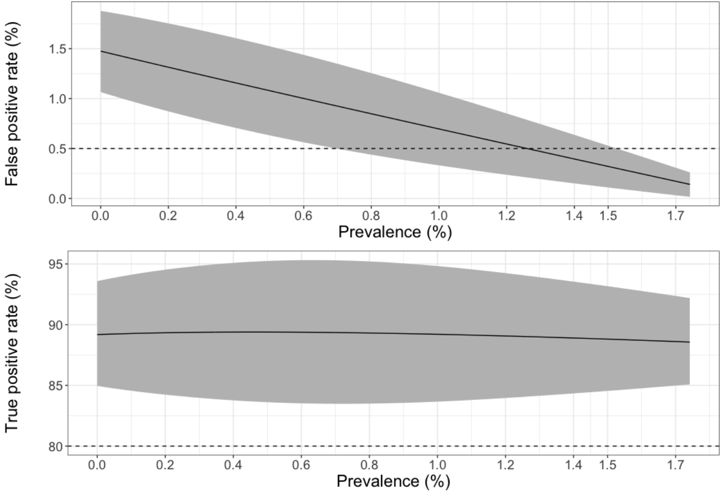
Illustration

Suppose $\theta_0 = (p, q, \pi) = (1.5\%, 100\%, 0\%)$. Then, $f(S|\theta_0)$ looks as follows:



- We have to decide: Is θ_0 plausible?

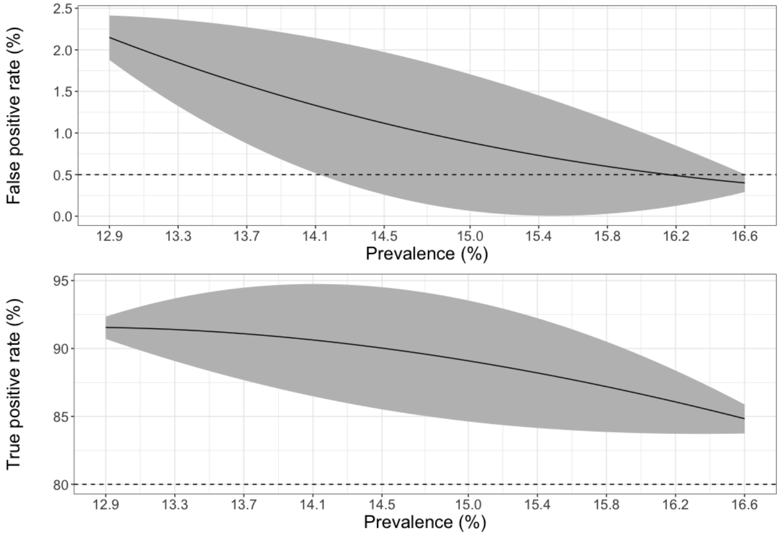
Application: Santa Clara study



Visualization of (p, q, π) in $\widehat{\Theta}$; dashed lines = empirical estimates of FPR, TPR;

Results: $\pi = 0\%$ is included; but [0.7-1.5%] is arguably more plausible.

Application: New York study



Results: Clear evidence for high prevalence. Go back

Discussion: Choice of test statistic

Procedure 1 is valid for any choice of the test statistic, s_n.

However, power depends on how sensitive s_n is in detecting violations of the null hypothesis.

We choose $s_n(Y^n, T^n) = A_n(\hat{\theta}_n) - A_n(\theta_0)$, the difference between periodogram values at the global peak peak and the null, θ_0.

Fisher's classical statistic is $s_n = \max_{\theta \in \Theta} \hat{A}_n(\theta)/\bar{A}_n$, where $\bar{A}_n = |\Theta|^{-1} \sum_\theta A_n(\theta)$.

Improvements using a trimmed mean in place of $\bar{A}_n$ have also been suggested (Bolviken, 1983; Siegel, 1980; Damsleth and Spjotvoll, 1982). See also (McSweeney, 2006) for numerical comparisons. Go back

Discussion: Computation

The complexity of our method is, prima facie, $O(|\Theta|^2 \cdot R \cdot C)$, where C = time for weighted least-squares.

e.g., for $|\Theta| = 10^4$, $R = 10^3$, and $C = 50\mu s$ an analysis on a conventional laptop of a time series with 200 observation times takes a total of 1,388 hrs. of wall clock time (approx. 58 days).

However, several reductions of computation time are possible.

1. Procedure 1 can be fully parallelized in step 3; e.g., with 100 nodes the wall clock time thus drops to 14 hrs.

2. Again in step 3, there is no need to consider all values in Θ but only a proportion; e.g., consider local peaks that are at least 20% as high as the global peak. This leads to a complexity $O(\gamma|\Theta|^2 \cdot R \cdot C)$ with $\gamma \sim 0.1\%$-3%.

As such, the computation in the above example drops dramatically to approximately 30 mins. of wall clock time. Indeed, in our application, get up to $R = 100,000$ and still finish all analyses in a few hours using a cluster with 400 nodes. Go back

Randomization Tests (Lehman and Romano, 2005)

Let $D \in \mathbb{R}^n$ be the data, and $\mathcal{G}^n$ a group of $\mathbb{R}^n \times \mathbb{R}^n$ transformations.
We are testing some H_0 under which:

$$D \overset{d}{=} \mathrm{g}D, \text{ for all } \mathrm{g} \in \mathcal{G}^n.$$

Define a test statistic $T_n = t_n(D)$ and $\mathsf{T}_D = \{t_n(\mathrm{g}D) : \mathrm{g} \in \mathcal{G}^n\}$. Then,

$$T_n \mid \mathsf{T}_D = \text{ Uniform.}$$

To test H_0, we could take the p-value of T_n wrt to T_D.

$*$ This test is (i) exact in finite samples and (ii) works for any choice of T_n.

Go back

Non-parametric approach (1/2)

Define

$$\Pi(T^n;\theta) = \{\pi \in \mathsf{S}_n : \pi(t_i) \equiv t_i (\text{mod } \theta),\ i = 1, \ldots, n\}.$$

In words, $\Pi(T^n;\theta)$ is the set of permutations of $(t_1, \ldots, t_n)$ such that any time t_i is mapped only to an observation time that is equivalent to t_i modulo θ.

We wish to test the following nonparametric null hypothesis of periodicity θ_0:

$$H_0^{\text{np}} : y^{\text{p}}(t') = y^{\text{p}}(t), \text{ for all } t', t \text{ such that } t' \equiv t (\text{mod } \theta_0). \quad (9)$$

To test H_0^{np} we can adapt Procedure 1 as follows.

1. For all $r = 1, \ldots, R$ do:
   - (i) Sample $\pi \sim \text{Unif}\big(\Pi(T^n;\theta_0)\big)$.
   - (ii) Generate synthetic outcome data $Y^{n,(r)} = \pi \cdot Y^n$ obtained by permuting the data Y^n according to π while observation times, T^n, are fixed.
2. Using the samples from 2(ii), calculate the p-value, say $\text{pval}(\theta_0)$, as in (5), and reject if the p-value is less than α.

Theorem

Suppose that Assumptions (A1) and (A2) hold with $\mathcal{G}^n = \Pi(T^n;\theta_0)$. Then, the p-value from Procedure 2 is exact under H_0^{np} conditionally on the observation times, that is,

$$\text{pr}\left\{\text{pval}(\theta_0) \leq \alpha \mid H_0^{\text{np}}, T^n\right\} = \alpha.$$

Non-parametric approach (2/2)

An alternative approach would be to use the nonparametric estimators of θ^* developed by (Hall et al., 2000); (Hall and Li, 2006); (Hall, 2008) together with a variation of Procedure 1 or Procedure 2.

Both these procedures do not require regularity conditions on the observation times but only a consistent estimator for the periodic component, y^{p}. We leave these directions for future work. Go back