Semi-supervised standardized detection of periodic signals with application to exoplanet detection

Sophia Sulis[1], **David Mary**[2], **Lionel Bigot**[2]

[1] Université Aix Marseille, CNRS, CNES, LAM, Marseille, France, Email: sophia.sulis@lam.fr
[2] Université Côte dAzur, Observatoire de la Côte dAzur, CNRS, France

Abstract

We propose a numerical methodology for detecting periodicities in unknown colored noise and for evaluating the 'significance levels' (p-values) of the test statistics. The procedure assumes and leverages the existence of a set of time series obtained under the null hypothesis (a null training sample, NTS) and possibly complementary side information. The test statistic is computed from a standardized periodogram, which is a pointwise division of the periodogram of the series under test to an averaged periodogram obtained from the NTS. The procedure provides accurate p-values estimation through a dedicated Monte Carlo procedure. While the methodology is general, our application is here exoplanet detection. The proposed methods are benchmarked on astrophysical data.

Introduction

The detection of periodic signals is an old topic in which there are persistent problems that still hold, such as the control of the significance levels of detection tests when the noise is colored and the time series is irregularly sampled. Indeed, the irregular sampling creates dependencies between the periodograms ordinates which complicate the statistical interpretation of the periodogram peaks values and often make analytical derivations of the significance levels (P-values) out of reach. Additionally, the unknown statistics of the colored noise complicate the picture. In the field of exoplanets, we can find many examples of false planet detections driven by significance levels with poor meaning.

Main Objective

Our objective is to provide a detection method that allows to exploit a NTS (if available) and to estimate accurately the resulting P-values for correlated noise models and irregularly sampled time series.

Composite hypothesis testing problem

Consider an irregularly sampled data time series $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$ and the composite hypothesis testing problem:

$$\begin{cases} \mathcal{H}_0 \ : \boldsymbol{x} = \boldsymbol{d}|\mathcal{M}_d(\boldsymbol{\theta}_d) + \boldsymbol{n}, \\ \mathcal{H}_1 \ : \boldsymbol{x} = \mathbf{s}|\mathcal{M}_s(\boldsymbol{\theta}_s) + \boldsymbol{d}|\mathcal{M}_d(\boldsymbol{\theta}_d) + \boldsymbol{n}, \end{cases} \ \boldsymbol{n} \sim \mathcal{N}(0, \Sigma) \qquad (1)$$

with $\boldsymbol{n}$ a zero mean, Gaussian stochastic noise component of unknown covariance matrix; $\boldsymbol{d}$ a nuisance signal (unknown mean under $\mathcal{H}_0$) that is generated from some model $\mathcal{M}_d$ with parameters $\boldsymbol{\theta}_d$; and $\boldsymbol{s}$ the unknown deterministic (quasi-)periodic signal from some model $\mathcal{M}_s$.

We consider the semi-supervised situations where part or all of the following side information is available: 1) For $\boldsymbol{d}$, the model $\mathcal{M}_d(\boldsymbol{\theta}_d)$ is available (but the corresponding parameter $\boldsymbol{\theta}_d$ vector is indeed unknown). 2) For the stochastic part $\boldsymbol{n}$, a set of $L \ll N$ time series of the noise (the NTS) is available:

$$\mathcal{T}_L := \{\boldsymbol{n}^{(i)}\}, i = 1, \ldots, L, \quad \boldsymbol{n}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \qquad (2)$$

To cope with unknown noise correlations, our approach is based on periodogram standardization $\widetilde{\boldsymbol{p}}(\nu_k|\boldsymbol{x}, \mathcal{T}_L)$ [3] by an averaged periodogram $\overline{\boldsymbol{p}}_L(\nu_k|\mathcal{T}_L)$ based on the NTS. The p-value of the observed test statistic $t_M := \max_k \widetilde{\boldsymbol{p}}$ is

$$\mathsf{v}(t_M) := \Pr(T_M > t_M \mid \mathcal{H}_0) = 1 - \Phi_{T_M}(t_M) \qquad (3)$$

with Φ_{T_M} the unknown CDF of the test T_M.

Semi-supervised standardized detection

Our semi-supervised standardized detection (3SD) procedure is composed of two algorithms. One to compute the test statistic t_M (Algorithm 1), and one for the computation of its p-values (Algorithm 2). We report the reader to the paper ID 2871 for the detailed description of these algorithms.

Algorithm 1: Considered standardized detection procedure. The procedure is semi-supervised if side information $\mathcal{T}_L$ or $\mathcal{M}_d$ is available.

Inputs : $\boldsymbol{x}$: Times series under test
(P,T): selected couple (periodogram, test)
$\boldsymbol{\Omega}$: considered set of frequencies
$\mathcal{T}_L$ and/or $\mathcal{M}_d$

Output: Test statistic $t(\boldsymbol{x})$

1 **if** $\mathcal{M}_d \neq \emptyset$ **then**
2 Estimate $\widehat{\boldsymbol{\theta}}_d$
3 $\boldsymbol{x} \leftarrow \boldsymbol{x} - \widehat{\boldsymbol{d}}|\mathcal{M}_d(\widehat{\boldsymbol{\theta}}_d)$
4 **end**
5 $\boldsymbol{p}(\boldsymbol{x}) \leftarrow$ Apply P to $\boldsymbol{x}$
6 **if** $\mathcal{T}_L \neq \emptyset$ **then**
7 Compute $\overline{\boldsymbol{p}}_L(\mathcal{T}_L)$ as in (3)
8 **else**
9 $\widehat{\sigma^2} \leftarrow$ Estimate var($\boldsymbol{x}$)
10 $\overline{\boldsymbol{p}}_L \leftarrow \widehat{\sigma^2}\mathbf{1}$
11 **end**
12 Compute $\widetilde{\boldsymbol{p}}$ as in (4)
13 $t(\boldsymbol{x}) \leftarrow$ Apply T to $\widetilde{\boldsymbol{p}}$

Application to exoplanet detection

Fig 2 shows the Lomb-Scargle periodogram of radial velocity data of the star α Centauri B around which a debated Earth-mass planet detection has been claimed at the period of 3.2 days [1]. The results of our Algorithm 2 with a given noise model $\mathcal{M}_d$ are displayed with the solid lines on the left panel: accounting for estimation errors in the nuisance parameters $\boldsymbol{d}$ leads to a p-value of about 1.5% of the largest periodogram peak (against 0.02%, see dashed lines). There is some uncertainty in the noise model, however: a similar though different model (say, $\mathcal{M}'_d$), based on Gaussian processes was proposed in [2]. It is therefore interesting to investigate how Algorithm 2 run with model $\mathcal{M}_d$ reacts when the data in reality undergo the slightly different model $\mathcal{M}'_d$. To this end, Algorithm 2 was run with $\mathcal{M}'_d$, except in row 21, which used model $\mathcal{M}_d$. The results are shown in the right panel. The p-value of the 3.2d peak, now 63.43%, indicates that this peak's height is in fact quite ordinary under the null hypothesis with model $\mathcal{M}'_d$, showing that the 3.2d detection is not robust under such model errors.

Algorithm 2: Monte Carlo procedure for estimating the p-value of the result of Algorithm 1 along with confidence intervals.

Inputs : $\boldsymbol{x}$: Times series under test
(P,T): selected couple (periodogram, test)
$\boldsymbol{\Omega}$: considered set of frequencies
b, B: Monte Carlo sample size
π: parameters' prior distribution
if $\mathcal{T}_L \neq \emptyset$ **then**
$\mathcal{M}_n$: parametric model for $\boldsymbol{n}$
$\widehat{\boldsymbol{\theta}}_n|\mathcal{M}_n, \mathcal{T}_L$: estimated parameters
else
$\widehat{\sigma^2}_w$: estimated variance of WGN
$\boldsymbol{\Delta}_w$: scale parameter for prior π on $\widehat{\sigma^2}_w$
end
if $\mathcal{M}_d \neq \emptyset$ **then**
$\widehat{\boldsymbol{\theta}}_d|\mathcal{M}_d$: estimated parameters
$\boldsymbol{\Delta}_d$: scale parameters for prior π on $\widehat{\boldsymbol{\theta}}_d$
end

Output: $\widehat{\mathsf{v}}(t)$ and 90% confidence interval

1 **for** $i = 1, \ldots, B$ **do**
2 **if** $\mathcal{T}_L \neq \emptyset$ **then**
3 $\mathcal{T}_L^{(i)} \leftarrow$ Generate from $\mathcal{M}_n(\widehat{\boldsymbol{\theta}}_n)$
4 $\widehat{\widehat{\boldsymbol{\theta}}}_n^{(i)} \leftarrow$ Estimate from $\mathcal{T}_L^{(i)}|\mathcal{M}_n$
5 **end**
6 **for** $j = 1, \ldots, b$ **do**
7 **if** $\mathcal{T}_L \neq \emptyset$ **then**
8 $\mathcal{T}_L^{(i,j)} \leftarrow \{\boldsymbol{n}^{(i,j,\ell)}|\widehat{\widehat{\boldsymbol{\theta}}}_n^{(i)}\}_{\ell=1,\ldots,L}$
9 $\boldsymbol{x}^{(i,j)} \leftarrow \boldsymbol{n}^{(i,j,L+1)}|\widehat{\widehat{\boldsymbol{\theta}}}_n^{(i)}$
10 **else**
11 $\epsilon_w \leftarrow$ Generate from $\pi(0, \boldsymbol{\Delta}_w)$
12 $\widehat{\widehat{\sigma^2}}_w^{(i,j)} \leftarrow \widehat{\sigma^2}_w + \epsilon_w$
13 $\boldsymbol{w}^{(i,j)} \sim \{\mathcal{N}(\mathbf{0}, \widehat{\widehat{\sigma^2}}_w^{(i,j)})$
14 $\boldsymbol{x}^{(i,j)} \leftarrow \boldsymbol{w}^{(i,j)}$
15 **end**
16 **if** $\mathcal{M}_d \neq \emptyset$ **then**
17 $\boldsymbol{\epsilon}^{(i,j)} \leftarrow$ Generate from $\pi(\mathbf{0}, \boldsymbol{\Delta}_d)$
18 $\widehat{\boldsymbol{d}}^{(i,j)} \leftarrow$ Generate $\mathcal{M}_d(\widehat{\boldsymbol{\theta}}_d + \boldsymbol{\epsilon}^{(i,j)})$
19 $\boldsymbol{x}^{(i,j)} \leftarrow \boldsymbol{x}^{(i,j)} + \widehat{\boldsymbol{d}}^{(i,j)}$
20 **end**
21 $t^{(i,j)} =$ Algorithm 1$\left(\boldsymbol{x}^{(i,j)}, (\mathsf{P}, \mathsf{T}), \boldsymbol{\Omega}, \mathcal{T}_L^{(i,j)}, \mathcal{M}_d\right)$
22 **end**
23 $\widehat{\Phi}_T^{(i)} \leftarrow$ Estimate CDF from the $\{t^{(i,j)}\}_{j=1,\ldots,b}$
24 $\widehat{\mathsf{v}}^{(i)}(t) \leftarrow 1 - \widehat{\Phi}_T^{(i)}(t)$
25 **end**
26 $\widehat{\mathsf{v}}(t) \leftarrow \frac{1}{B}\sum_{i=1}^{B} \widehat{\mathsf{v}}^{(i)}(t)$
27 90% confidence interval $\leftarrow \{\widehat{\mathsf{v}}^{(i)}(t)\}_{i=1,\ldots,B}$

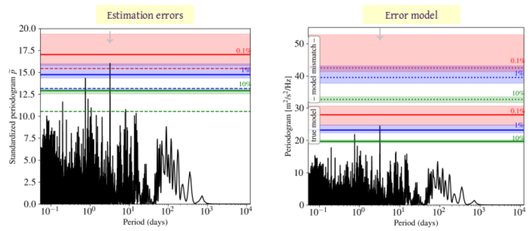

Figure 1: Application to exoplanet detection. Black: data (standardized) Lomb-Scargle periodograms. Prior predictive p-values levels 0.1, 1 and 10% (solid) with their 90% confidence intervals (shade) estimated assuming model $\mathcal{M}_d$ is true in Algorithm 2. **Left:** P-values computed assuming the noise is a WGN and there is no estimation error or error model (dashed). **Right:** P-value levels computed by Algorithm 2 if model $\mathcal{M}'_d \neq \mathcal{M}_d$ is true (dotted).

Conclusions

In conclusion, we have proposed a semi-supervised detection procedure and a MC method allowing to evaluate the resulting p-values and the impact of some model error. This procedure is designed to leverage ancillary data such as a NTS, that is used for periodogram standardization. The method is quite versatile in the periodograms, tests and noise models that can be plugged-in. We have illustrated and validated the procedure for exoplanet detection but it can be easily adapted to other periodicity detection problems.

References

[1] X. Dumusque et al. An Earth-mass planet orbiting α Centauri B. *Nature*, 491(7423):207–211, 2012.

[2] V. Rajpaul et al. Ghost in the time series: no planet for Alpha Cen B. *MNRAS*, 456(1):L6–L10, 2016.

[3] S. Sulis, D. Mary, and L. Bigot. A Study of Periodograms Standardized Using Training Datasets and Application to Exoplanet Detection. *IEEE Trans. on Signal Processing*, 65:2136–2150, 2017.

Acknowledgements

S. Sulis acknowledges support from CNES. D. Mary acknowledges support from the GDR ISIS through the *Projet exploratoire TASTY*. MHD stellar computations have been done on the "Mesocentre SIGAMM" machine, hosted by *Observatoire de la Côte d'Azur*.