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

Paper ID : 2871

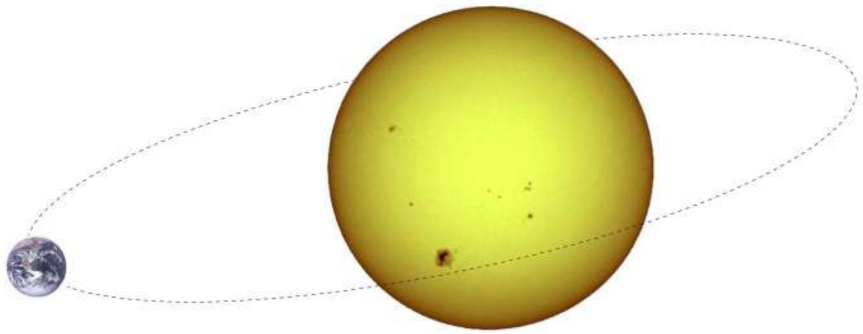
Sophia Sulis
Laboratoire d'Astrophysique de Marseille, Aix-Marseille Université

David Mary & Lionel Bigot
Université Côte d'Azur, Observatoire de la Côte d'Azur

2022 IEEE International Conference on Acoustics, Speech and Signal Processing Conference

Introduction

- Detection of periodic signals
- Irregular temporal sampling →dependencies between periodogram's ordinates
- Noise of partially/unknown distribution
- Problem : compute accurate significance level (p-values) of detection tests
- Work driven by an astrophysical application : detection of small extrasolar planets



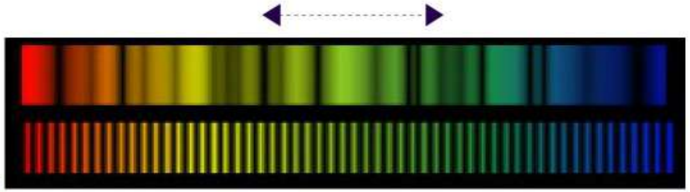
Overview

- Astrophysical context : signal, noise and simulations
- Detection algorithm
- Evaluation of the p-values
- Results

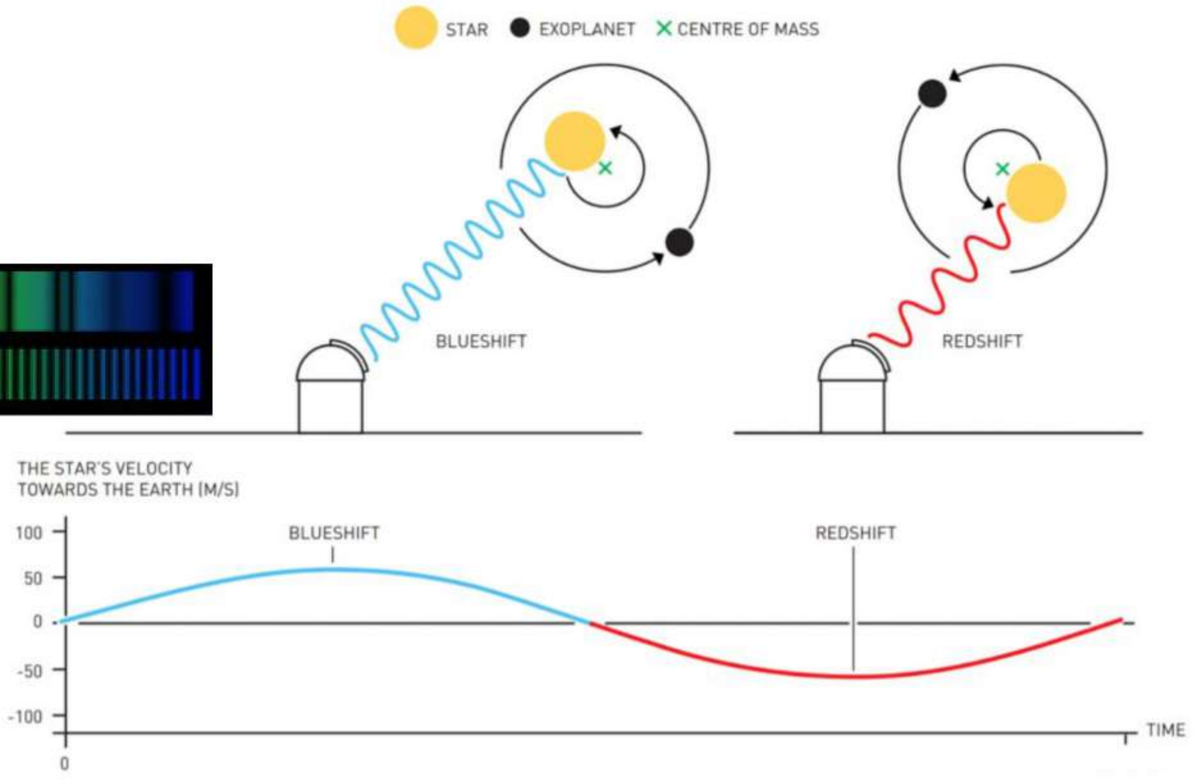
Astrophysical context : exoplanet detection

- The radial velocity technique

Star spectrum



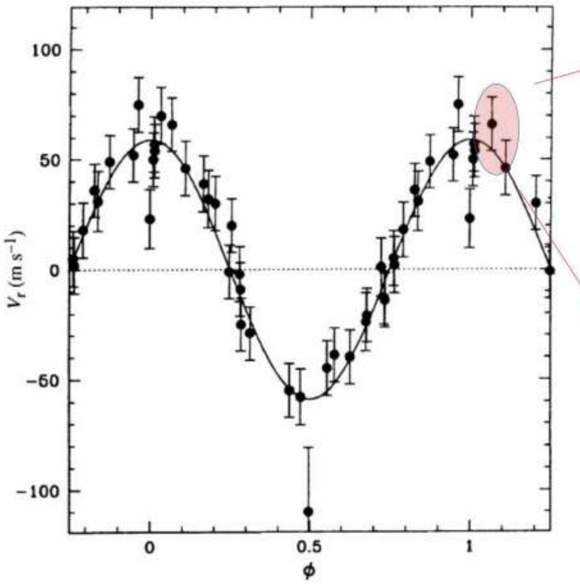
Astro-comb
(wavelength reference)



Astrophysical context : exoplanet detection

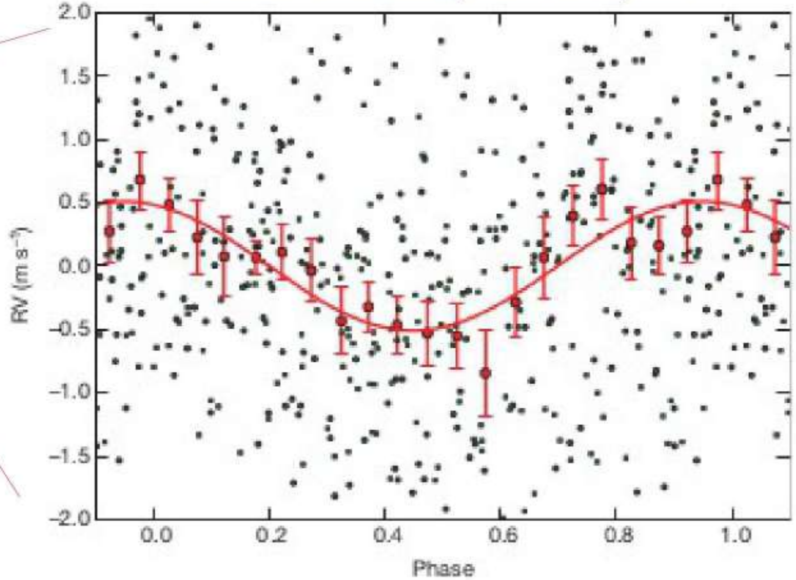
- Signal model : periodic (Keplerian) signature

51 Pegasi b : a Jupiter-mass planet



[Mayor & Queloz, 1995]

aCenBb : a controversy Earth-mass planet

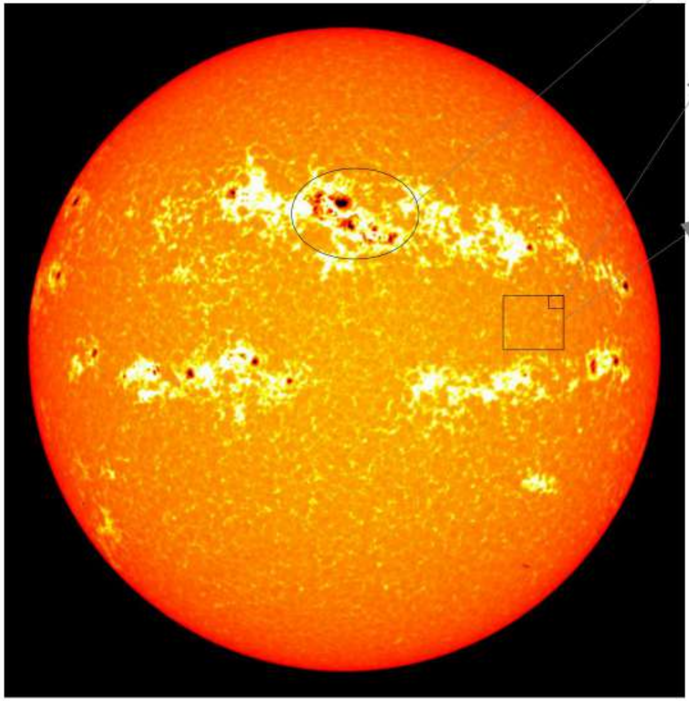


[Dumusque et al, 2012]

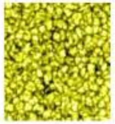
Astrophysical context : exoplanet detection

- Noise : Stellar activity

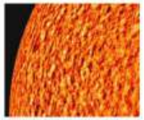
Starspots and faculae



High solar activity - March 28, 2001. Credit : NASA



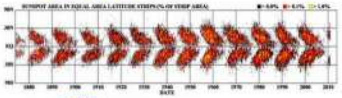
Granulation



Supergranulation

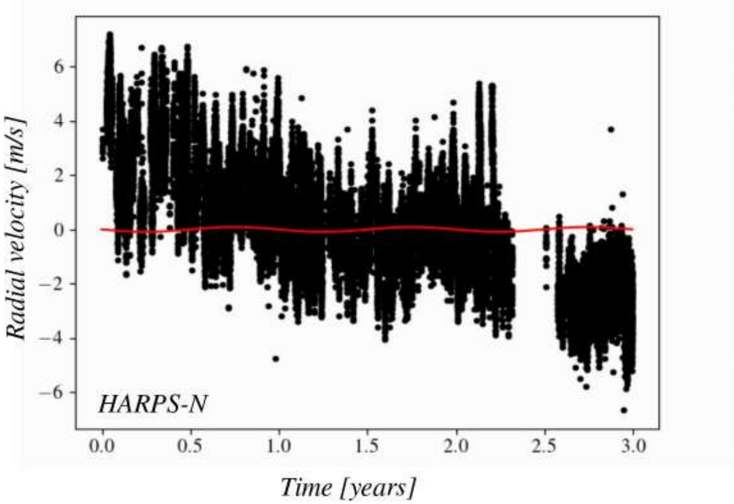
Flares

Stellar pulsations



Stellar cycle

Meridional circulation



Astrophysical context : exoplanet detection

- Noise : null training sample (NTS)

➢ For a target star of known parameters, it is possible to generate realistic realisations of the stochastic fluctuations due to stellar convection (3D magneto-hydrodynamical simulations)

➢ Computationally heavy : only a few training data set available (L<<N)

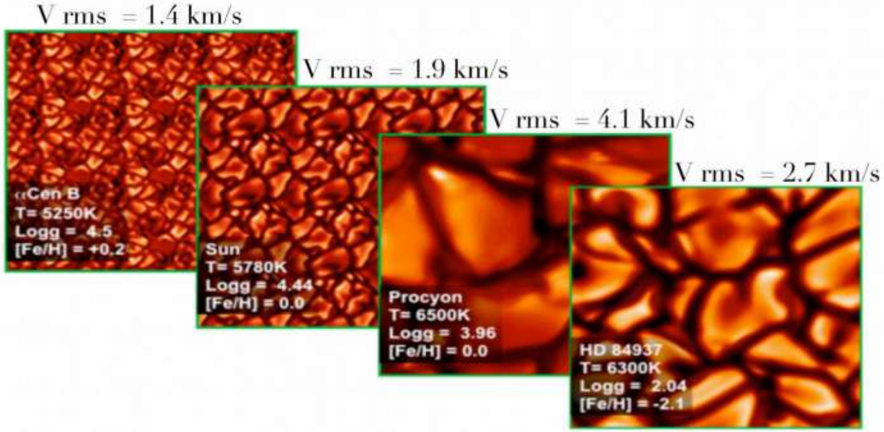
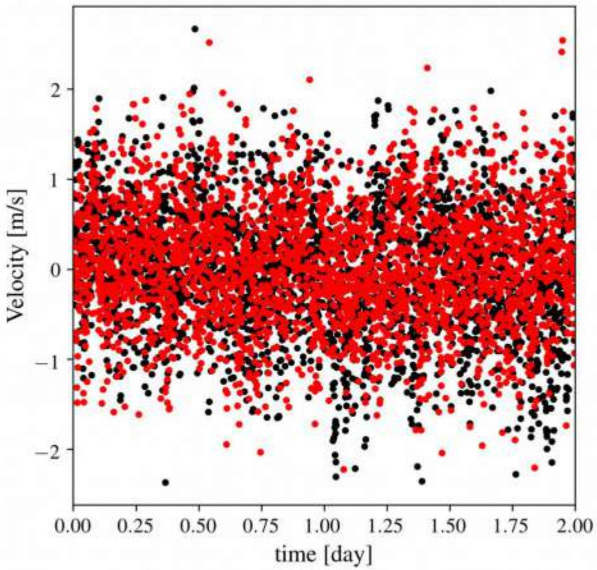
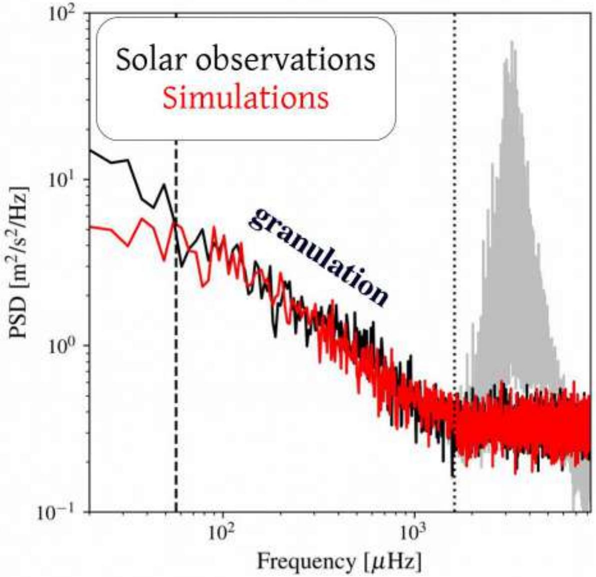


Image credit : L. Bigot





Figures from Sulis et al., (2020)

Astrophysical context : exoplanet detection

- Noise : complementary side information

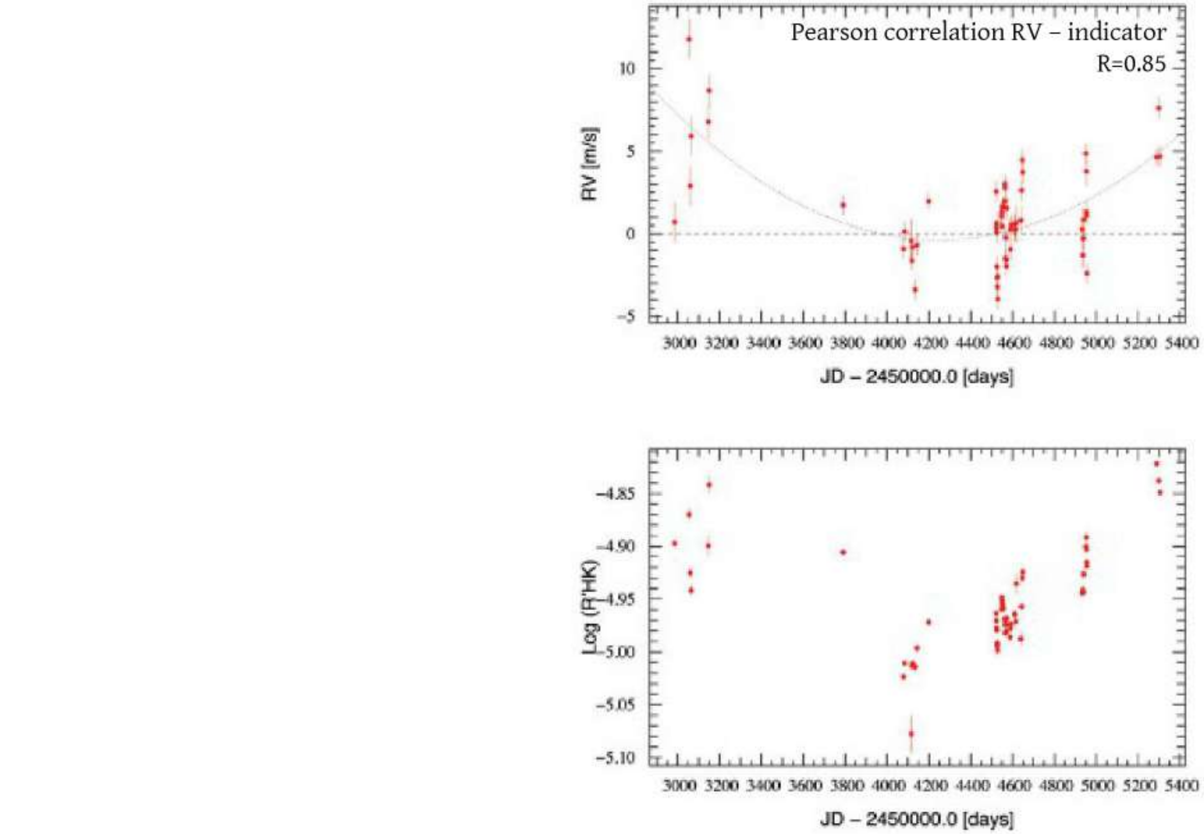


Figure from Dumusque et al. (2010)

Composite hypothesis testing problem

- Given an irregularly sampled data time series $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top$

- We consider the following composite hypothesis testing problem

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

n : NTS available

d : ancillary time series available

n : a zero mean, Gaussian stochastic noise component of unknown covariance matrix $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_c + \sigma_w^2 \boldsymbol{I}$,

d : nuisance signal (stellar magnetic activity, instrumental drift)

s : unknown, deterministic, periodic or quasi-periodic signal

Semi-supervised standardized detection (3SD) procedure

https://github.com/ssulis/3SD

- Consider a null training series $\mathcal{T}_L := \{\boldsymbol{n}^{(i)}\}, i = 1, \ldots, L, \quad \boldsymbol{n}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$
- Built the averaged periodogram $\overline{\boldsymbol{p}}_L(\nu_k|\mathcal{T}_L) := \frac{1}{L}\sum_{\ell=1}^{L} \boldsymbol{p}_\ell(\nu_k|\boldsymbol{x}^{(\ell)}),$
- Define the standardized periodogram as $\widetilde{\boldsymbol{p}}(\nu_k|\boldsymbol{x}, \mathcal{T}_L) := \frac{\boldsymbol{p}(\nu_k)}{\overline{\boldsymbol{p}}_L(\nu_k|\mathcal{T}_L)}$
- A classical detection test statistic ('Max-test') $T_M(\widetilde{\boldsymbol{p}}) := \max_k \widetilde{\boldsymbol{p}}(\nu_k)$ ⟶ observed test statistic t_M
- How `` significant '' is t_M ?

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

unknown ...

Semi-supervised standardized detection (3SD) procedure

https://github.com/ssulis/3SD

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

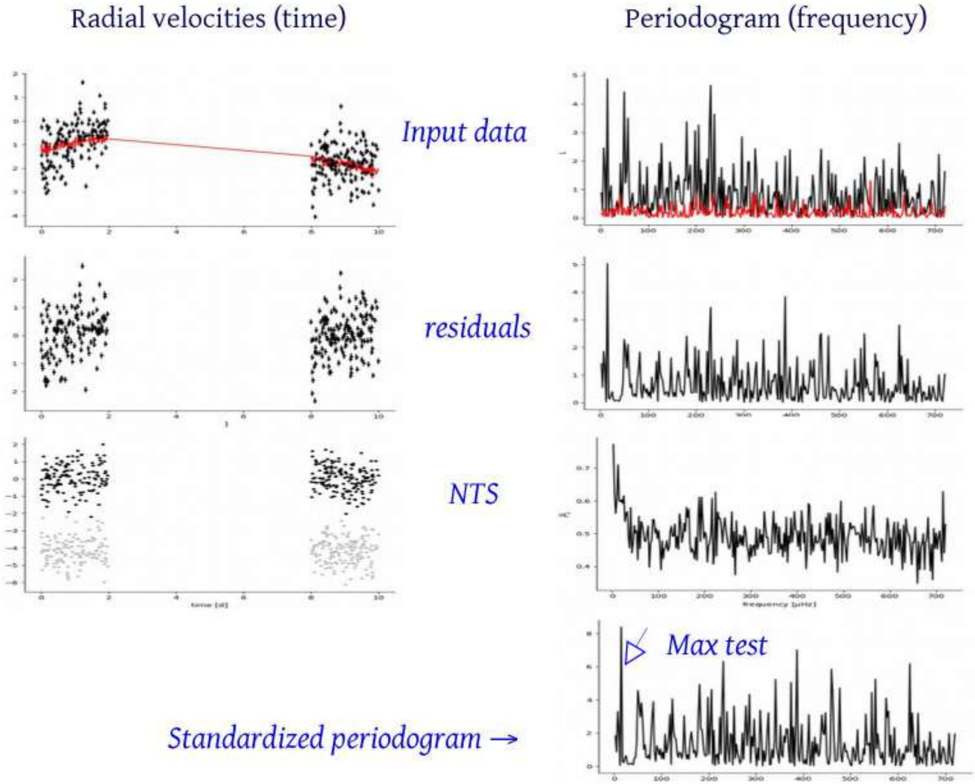
Inputs : $\boldsymbol{x}$: Times series under test
(P,T): selected couple (periodogram, test)
$\boldsymbol{\Omega}$: considered set of frequencies
$\mathcal{T}_L$ and/or $\mathcal{M}_d$

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

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

Detection algorithm

Illustrative example : sinusoid + colored noise



Semi-supervised standardized detection (3SD) procedure

https://github.com/ssulis/3SD

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

for $i = 1, \ldots, B$ **do**
  if $\mathcal{T}_L \neq \emptyset$ **then**
    $\mathcal{T}_L^{(i)} \leftarrow$ Generate from $\mathcal{M}_n(\widehat{\boldsymbol{\theta}}_n)$
    $\widehat{\widehat{\boldsymbol{\theta}}}_n^{(i)} \leftarrow$ Estimate from $\mathcal{T}_L^{(i)}|\mathcal{M}_n$
  end
  for $j = 1, \ldots, b$ **do**
    if $\mathcal{T}_L \neq \emptyset$ **then**
      $\mathcal{T}_L^{(i,j)} \leftarrow \{\boldsymbol{n}^{(i,j,\ell)}|\widehat{\widehat{\boldsymbol{\theta}}}_n^{(i)}\}_{\ell=1,\ldots,L}$
      $\boldsymbol{x}^{(i,j)} \leftarrow \boldsymbol{n}^{(i,j,L+1)}|\widehat{\widehat{\boldsymbol{\theta}}}_n^{(i)}$
    else
      $\epsilon_w \leftarrow$ Generate from $\pi(0, \boldsymbol{\Delta}_w)$
      $\widehat{\widehat{\sigma^2}}_w^{(i,j)} \leftarrow \widehat{\sigma^2}_w + \epsilon_w$
      $\boldsymbol{w}^{(i,j)} \sim \{\mathcal{N}(\boldsymbol{0}, \widehat{\widehat{\sigma^2}}_w^{(i,j)})$
      $\boldsymbol{x}^{(i,j)} \leftarrow \boldsymbol{w}^{(i,j)}$
    end
    if $\mathcal{M}_d \neq \emptyset$ **then**
      $\boldsymbol{\epsilon}^{(i,j)} \leftarrow$ Generate from $\pi(\boldsymbol{0}, \boldsymbol{\Delta}_d)$
      $\widehat{\boldsymbol{d}}^{(i,j)} \leftarrow$ Generate $\mathcal{M}_d(\widehat{\boldsymbol{\theta}}_d + \boldsymbol{\epsilon}^{(i,j)})$
      $\boldsymbol{x}^{(i,j)} \leftarrow \boldsymbol{x}^{(i,j)} + \widehat{\boldsymbol{d}}^{(i,j)}$
    end
    $t^{(i,j)}$ = Algorithm 1$\left(\boldsymbol{x}^{(i,j)}, (\mathsf{P},\mathsf{T}), \boldsymbol{\Omega}, \mathcal{T}_L^{(i,j)}, \mathcal{M}_d\right)$
  end
  $\widehat{\Phi}_T^{(i)} \leftarrow$ Estimate CDF from the $\{t^{(i,j)}\}_{j=1,\ldots,b}$
  $\widehat{\mathsf{v}}^{(i)}(t) \leftarrow 1 - \widehat{\Phi}_T^{(i)}(t)$
end
$\widehat{\mathsf{v}}(t) \leftarrow \frac{1}{B}\sum_{i=1}^{B} \widehat{\mathsf{v}}^{(i)}(t)$
90% confidence interval $\leftarrow \{\widehat{\mathsf{v}}^{(i)}(t)\}_{i=1,\ldots,B}$

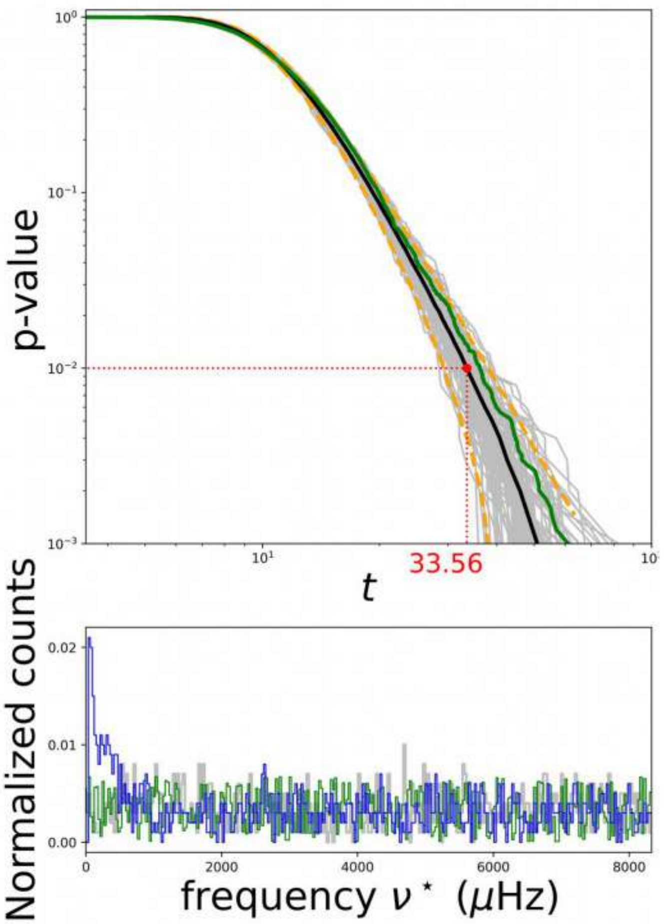
Estimation of p-values

Principle : sample a large number of test statistics that are consistent with the data and the models, by randomly perturbing the parameters within their uncertainties according to some prior distribution.

→ Estimation errors

- Setting :
  - If $\mathcal{T}_L \neq \emptyset$, we use a parametric model $\mathcal{M}_n$ to generate synthetic samples of the NTS.
  - If $\mathcal{T}_L = \emptyset$, an estimate of the noise variance is required
  - If $\mathcal{M}_d \neq \emptyset$, the algorithm requires the model + the parameters estimates

Validation : accuracy of the procedure in evaluating p-values



Error models – diagnostic with the 3SD procedure

- In Algorithm 2 : MC samples are generated under the same noise model $\mathcal{M}_d$
- But ... different models may lead to unconsistent results
- In presence of error model : true p-values are larger than estimated under no error model
- Algorithm 2 : sampling from some model and conducting the detection and p-value estimation procedures with another.

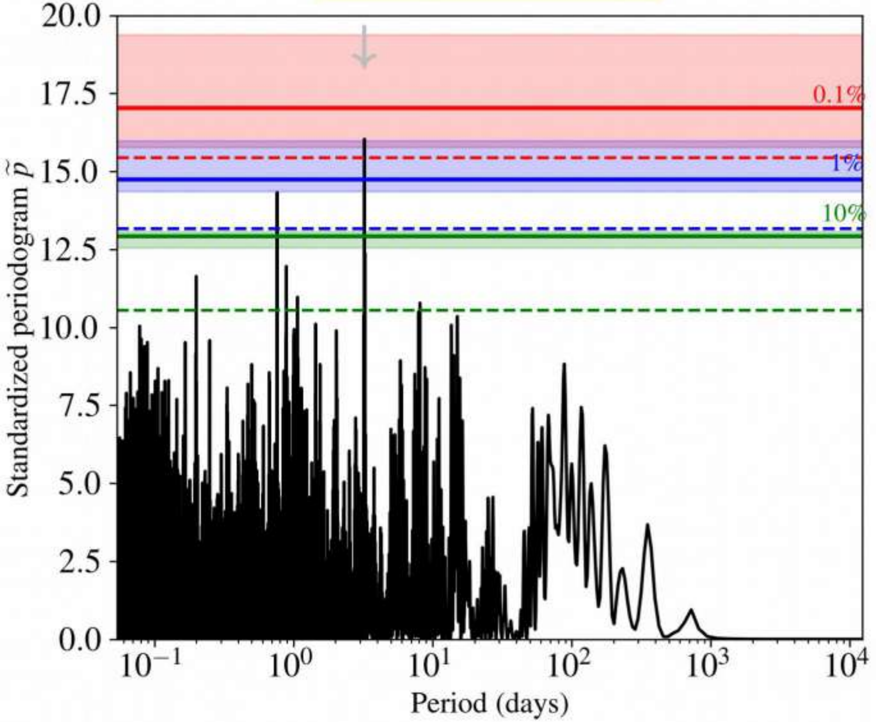
→ Access to the robustness of a specific procedure to error model

Application to exoplanet detection

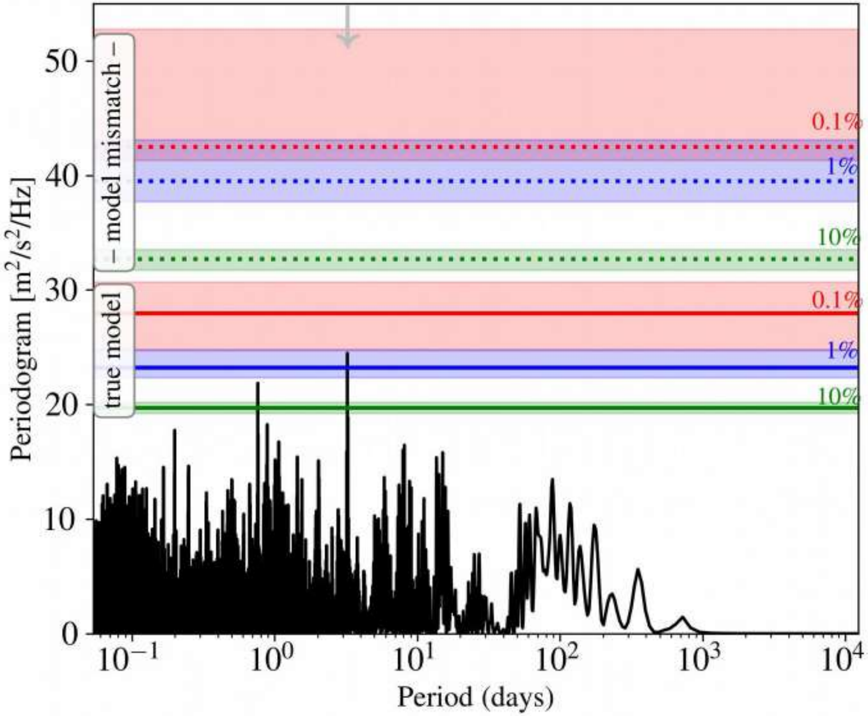
- System αCenB : a controversy planet of 1 Earth mass with 3.2 day orbital period

  Detection paper: Dumusque et al. (2012) Nature
  Controversy papers : Hatzes (2013), Rajpaul et al. (2016), Toulis & Bean (2021)

Estimation errors



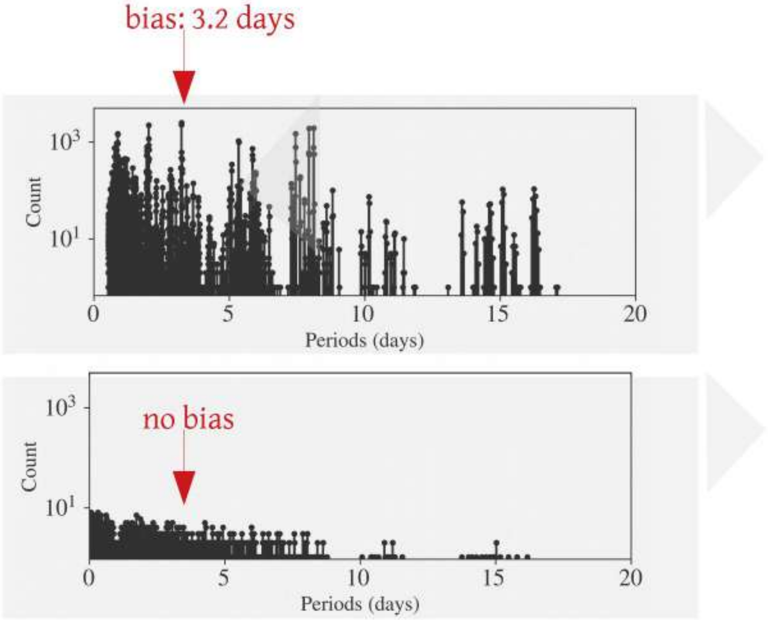
Error model

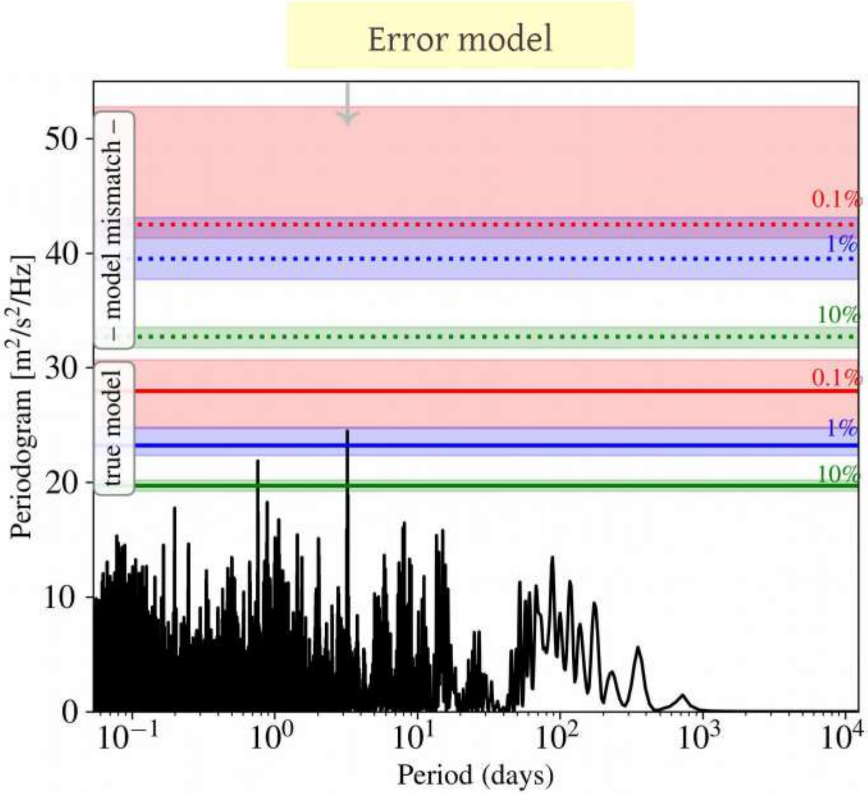


Application to exoplanet detection

- System αCenB : a controversy planet of 1 Earth mass with 3.2 day orbital period

  Detection paper: Dumusque et al. (2012) Nature
  Controversy papers : Hatzes (2013), Rajpaul et al. (2016), Toulis & Bean (2021)





Summary

- Controlling the error rate of detection tests is critical and not easy in many practical applications
- Focus here on detection tests for exoplanet detection
- Noise PSD unknown distribution but i) reliable simulations, ii) ancillary observations available
- We proposed a semi-supervised standardized detection procedure and a MC method to evaluate p-values of detection tests

→ propagate the estimation errors of model parameters on p-values

→ quite versatile (periodograms, detection tests, noise models)

→ available online at : *https://github.com/ssulis/3SD*

→ allow to test the robustness of a given noise model to error model

Thank you ! Questions ? Please, write me an email : sophia.sulis@lam.fr

References

- [Sulis 2017a]

Sophia Sulis. ``Statistical methods using hydrodynamic simulations of stellar atmospheres for detecting exoplanets in radial velocity data''.
PhD thesis (English), Université Côte d'Azur, 2017.

- [Sulis 2017b]

S. Sulis, D. Mary, and L. Bigot, ``A Study of Periodograms Standardized Using Training Datasets and Application to Exoplanet Detection'',
IEEE Transactions on Signal Processing, vol. 65, no. 8, 2136-2150, April, 15 2017. doi: 10.1109/TSP.2017.2652391

- [Sulis 2020a]

S. Sulis, D. Mary, and L. Bigot." 3D magneto-hydrodynamical simulations of stellar convective noise for improved exoplanet detection. I. Case of regularly sampled radial velocity observations. "
Astronomy & Astrophysics, 635, A146, 2020 - doi: 10.1051/0004-6361/201937105

- [Sulis 2017c]

S. Sulis, D. Mary, and L. Bigot. ``A bootstrap method for sinusoid detection in colored noise and uneven sampling. Application to exoplanet detection'',
25th European Signal Processing Conference (EUSIPCO), Kos, 2017, pp. 1095-1099. doi: 10.23919/EUSIPCO.2017.8081377

- [Sulis 2020b]

S. Sulis, D. Mary, and L. Bigot. « Semi-supervised standardized detection of extrasolar planets »,
submitted to Astronomy & Astrophysics (11/2021)