Lomb-Scargle Periodogram
COM-500 Statistical signal and data processing through applications : Mini-project report

Toussain Cardot
IC - EPFL
toussain.cardot@epfl.ch

Samuel Dubuis
IC - EPFL
s.dubuis@epfl.ch

Ivan Snozzi
IC - EPFL
ivan.snozzi@epfl.ch

Abstract

In this mini-project we discuss the Lomb-Scargle periodogram, an advanced method to the standard periodogram. We will generate simulated (unevenly distributed) data and test each one at first, then do the same with real captured data. Finally, we will discuss and compare the two.

I. Introduction

Often in statistical signal processing, the goal of frequency estimation is the process of estimating the complex frequency components of a signal in the presence of noise given assumptions about the number of components [1].

However, the periodogram is a method based on a non parametric model since the only assumption made is that the signal is WSS (wide-sense stationary). It has a high variance and its spectral resolution is determined by the number of samples used to compute it.

The principle of this last one is that the power spectral density of the process sampled, which represents the average distribution of energy over the frequency domain, is also the Fourier transform of its auto-correlation function. It has its pros, however it is badly considered for multiple reasons that we will explain in more details. This will be the first method we will describe.

The second one will be the Lomb-Scargle periodogram. This method is a well-known algorithm for detecting and characterizing periodicity in unevenly sampled time-series [2]. It allows us to make an efficient computation of a Fourier-like power spectrum estimator even with very unpleasant and unevenly sampled data. It is very clean and capable of detecting periodic features, motivated by and closely related to the Fourier analysis and least-square methods, as well as Bayesian probability.

The second part of this report will be dedicated to its understanding, analysis and comparison to the standard periodogram.

II. Procedure

To experiment with both algorithms, we have created data, evenly and unevenly sampled. The results and data were all created with Python as a coding language.

The evenly sampled data is created as such

$$X_{\text{even}}[n] = \sum_{n=1}^{4} \sin(2\pi f n t) \tag{1}$$

with $t = 1000$ samples from 0 to 10 evenly spaced and $f = 10$. We then added noise by extracting random samples from a normal Gaussian distribution, defined here as $W[n]$ with noise power of 0.5.

$$Y_{\text{even}}[n] = X_{\text{even}}[n] + W[n] \tag{2}$$

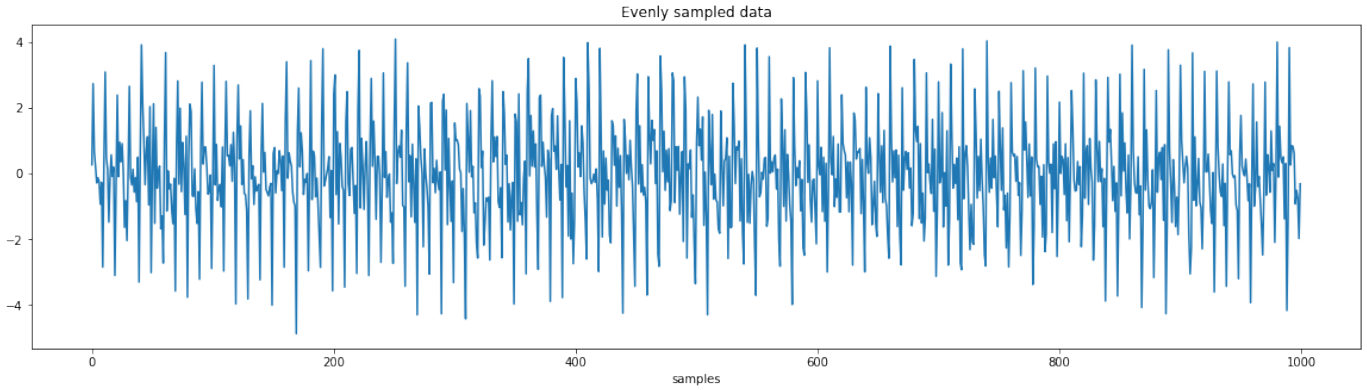

Fig. 1: Representation of the evenly sampled data created $Y_{\text{even}}[n]$

Finally, the unevenly sampled data was created by removing random samples from $X_{\text{even}}[n]$. We chose to drop around 50% of the original data. A random permutation is made to shuffle indices and only 50% are then kept into the new data.

$$Y_{\text{uneven}}[n] = \text{random } 50\% \text{ of } Y_{\text{even}}[n] \tag{3}$$

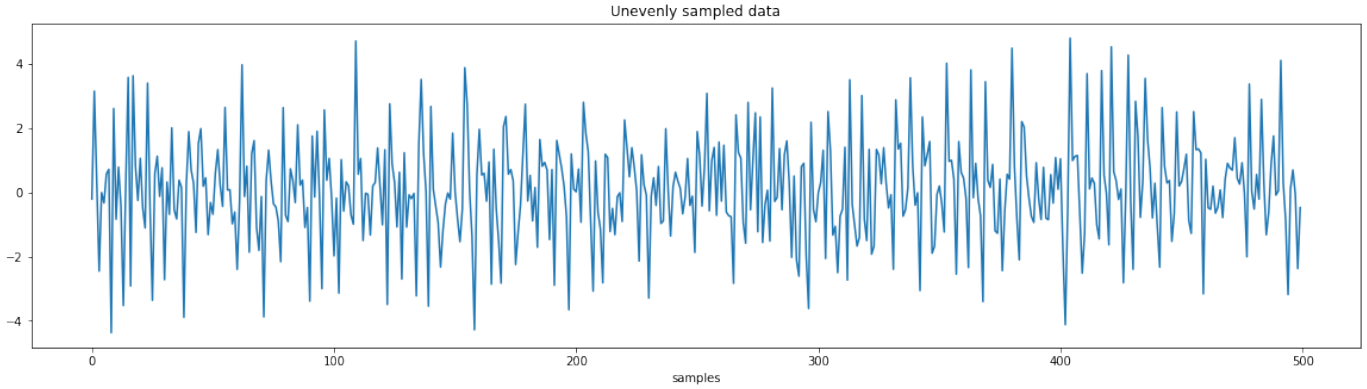

Fig. 2: Representation of the unevenly sampled data created $Y_{\text{uneven}}[n]$

The last data we generated was interpolated data, from the unevenly sampled data. It was interpolated based on a third order (cubic) spline fitting on the one dimension data. We'll call it $Y_{\text{interp}}[n]$.

We also test the selected algorithms on real data that we were given : an ECG signal.

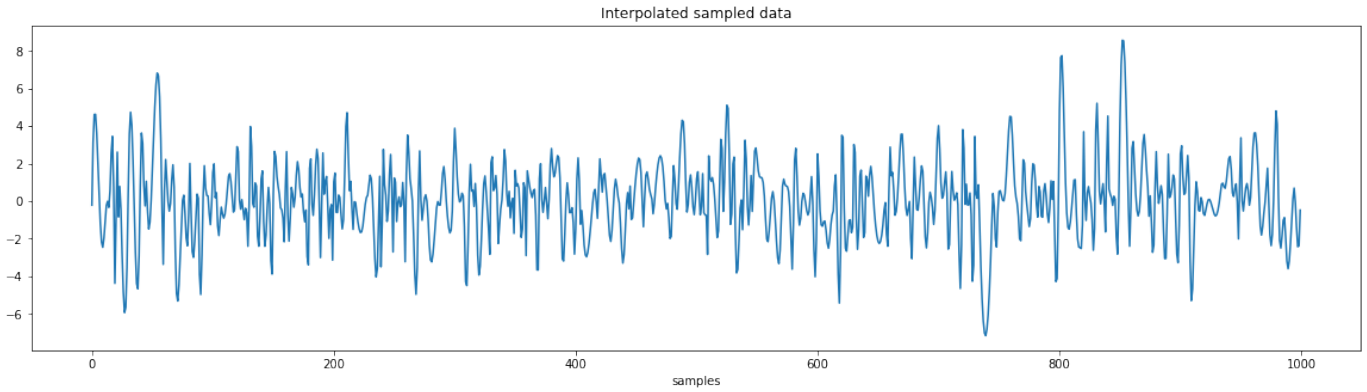

Fig. 3: Representation of the interpolated data created $Y_{\text{interp}}[n]$

A. *Periodogram*

Following [3] and [4], we can explain the process of the classical periodogram.

We define the process $X[n]$ to be WSS and to have been observed over a finite temporal window. The periodogram is a non-parametric estimate of the power spectral density (PSD) of a wide-sense stationary random process, described as such

$$P_X^N(\omega) = \frac{1}{N}\left|\sum_{n=1}^{N} x[n]e^{-j\omega n}\right|^2 = \frac{1}{N}\left|\hat{x}_N(\omega)\right|^2 \tag{4}$$

which is the discrete time Fourier transform (DTFT) of the observed samples

However, the PSD $P_X(\omega)$ represents the average distribution of energy over the frequency of the domain. But the PSD is defined as the Fourier transform of the auto-correlation $R_X[k]$ of the process $X[n]$.

How are the PSD $P_X(\omega)$ and the DTFT of these observations related ?

$$P_X^N(\omega) = \frac{1}{N}\left|\sum_{n=1}^{N} x[n]e^{-j\omega n}\right|^2 \tag{5}$$

$$= \frac{1}{N}\sum_{m=1}^{N}\sum_{l=1}^{N} x[m]x[l]^* e^{-j\omega(m-l)} \tag{6}$$

$$= \frac{1}{N}\sum_{m=1}^{N}\left(\sum_{l<m} x[m]x[l]e^{-j\omega(m-l)} + \sum_{l\geq m} x[m]x[l]e^{-j\omega(m-l)}\right) \tag{7}$$

$$= \frac{1}{N}\sum_{k=-N+1}^{N-1}\sum_{l=1}^{N-|k|} x[l+k]x[l]e^{-j\omega k} \tag{8}$$

Equation 5 and 6 is the periodogram related to the correlation estimator. In equation 7, the observations were assumed to be real and in the final equation 8 we make a change of variable $m-l=k$ and compute the rest.

Knowing that the biased estimator of the correlation is

$$\frac{1}{N}\sum_{l=1}^{N-|k|} x[l+k]x[l] = \hat{R}_X[k] \tag{9}$$

We end up with

$$P_X^N(\omega) = \sum_{k=-N+1}^{N-1} \hat{R}_X[k]e^{-j\omega k} \tag{10}$$

meaning that the periodogram is the DTFT of the correlation estimator $\hat{R}_X[k]$.

Using the periodogram as the Fourier transform of the correlation also shows us that the periodogram is the biased estimator of the PSD.

$$E[P_X^N(\omega)] = E\left[\frac{1}{N}\sum_{k=-N+1}^{N-1}\sum_{l=1}^{N-|k|} X[l+k]X[l]e^{-j\omega k}\right] \tag{11}$$

$$= \frac{1}{N}\sum_{k=-N+1}^{N-1} e^{-j\omega k}\sum_{l=1}^{N-|k|} E[X[l+k]X[l]] \tag{12}$$

$$= \frac{1}{N}\sum_{k=-N+1}^{N-1} e^{-j\omega k}\sum_{l=1}^{N-|k|} R_X[k] \tag{13}$$

$$= \sum_{k=-N+1}^{N-1} \frac{N-|k|}{N} R_X[k]e^{-j\omega k} \tag{14}$$

$$\neq \sum_{k=-N+1}^{N-1} R_X[k]e^{-j\omega k} \tag{15}$$

But with the use of a triangular window, we can modify the expression and see the periodogram as an unbiased estimator of the PSD.

$$E[P_X^N(\omega)] = \sum_{k=-N+1}^{N-1} \frac{N-|k|}{N} R_X[k]e^{-j\omega k} \tag{16}$$

$$= \sum_{k=-\infty}^{\infty} BW[k]R_X[k]e^{-j\omega k} \tag{17}$$

with

$$BW[k] = \begin{cases} \frac{N-|k|}{N}, & \text{if } k = -N+1, \ldots, N-1 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

which is the triangular window with Fourier transform as such :

$$\widehat{BW}(\omega) = \frac{1}{N}\left(\frac{\sin\frac{\omega N}{2}}{\sin\frac{\omega}{2}}\right)^2 \tag{19}$$

As such

$$E[P_X^N(\omega)] = FT(BW[k]R_X[k]) \tag{20}$$

$$= \frac{1}{N}\left(\frac{\sin\frac{\omega N}{2}}{\sin\frac{\omega}{2}}\right)^2 * \sum_{k=-\infty}^{\infty} R_X[k]e^{-j\omega k} \tag{21}$$

That last expressions shows that the periodogram can be seen as unbiased, with mean

$$\sum_{k=-\infty}^{\infty} R_X[k]e^{-j\omega k}$$

convoluted with the square of the "pseudo" sinc function

$$\left(\frac{\sin\frac{\omega N}{2}}{\sin\frac{\omega}{2}}\right)^2$$

In average the Periodogram behaves like the Fourier transform of the correlation convoluted with the square of the "pseudo" sinc function.

B. Lomb-Scargle Periodogram

Following now Jacob VanderPlas and his very detailed paper [2] about the comprehension and background of the Lomb-Scargle periodogram, we can describe it. Scargle first modified the standard periodogram formula to first find a time delay τ such that this pair of sinusoids would be mutually orthogonal at sample times n, and also adjusted for the potentially unequal powers of these two basis functions, to obtain a better estimate of the power at a frequency [5], which made his modified periodogram method exactly equivalent to Lomb's least-squares method.

Considering the original periodogram, we can develop it

$$P_X^N(\omega) = \frac{1}{N}\left|\sum_{n=1}^{N} x[n]e^{-j\omega n}\right|^2 \tag{22}$$

$$= \frac{1}{N}\left[\left(\sum_n x[n]\cos(\omega n)\right)^2 + \left(\sum_n x[n]\sin(\omega n)\right)^2\right] \tag{23}$$

It has nice useful properties which however don't hold if the sampling is non-uniform. Scargle creates a generalized form of the periodogram

$$P_X^N(\omega) = \frac{A^2}{2}\left(\sum_n x[n]\cos(\omega[n-\tau])\right)^2 + \frac{B^2}{2}\left(\sum_n x[n]\sin(\omega[n-\tau])\right)^2 \tag{24}$$

where A, B and τ are arbitrary functions related to the function f defining ω. We can chose specific forms for these such that we get the original periodogram if the samples are evenly sampled, this periodogram is insensitive to time-shift and computable analytically.

These lead to this new form of periodogram

$$P_X^N(\omega) = \frac{1}{2}\left[\frac{\left(\sum_n x[n]\cos(\omega[n-\tau])\right)^2}{\sum_n \cos^2(\omega[n-\tau])} + \frac{\left(\sum_n x[n]\sin(\omega[n-\tau])\right)^2}{\sum_n \sin^2(\omega[n-\tau])}\right] \tag{25}$$

where τ is

$$\tau = \frac{1}{2\omega} \tan^{-1} \left(\frac{\sum_n \sin(2\omega n)}{\sum_n \cos(2\omega n)} \right) \tag{26}$$

The main difference with the classical one is that the denominator are now $\sum_n \cos^2(2\omega n)$ and $\sum_n \sin^2(2\omega n)$ instead of $N/2$, the expected values of each quantities. A very interesting point to the Lomb-Scargle periodogram is that its results are nearly identical to the fitting of a model of simple sinusoids to unevenly sampled data at each frequency. The τ function value serves now as a shift to orthogonalize the equation used for fitting.

III. Results

A. *Classical periodogram*

The first results obtained are the ones of the classical periodogram applied to the generated and real data. To better understand the effects of nonuniform sampling on the classical periodogram we tested it over some simple functions. Figures 4 and 5 show two *sinc* functions: one which has been sampled uniformly and the other one non-uniformly. The non-uniformly sampled function periodogram appear clearly more noisy than the other one. For this results, half of the samples have been dropped randomly to simulate unevenly sampled data.

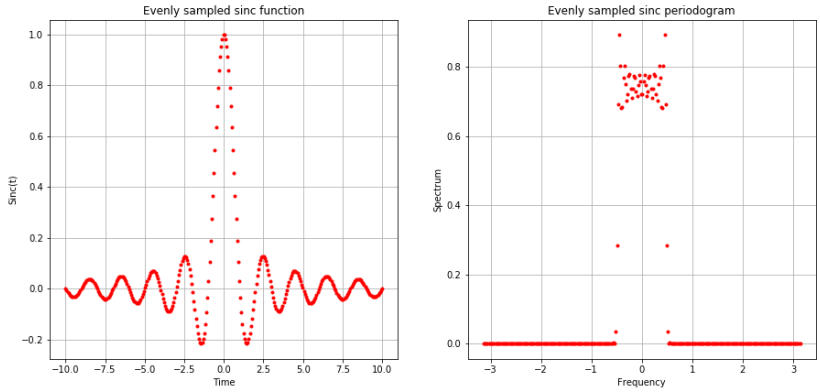

Fig. 4: Simple sinc function evenly sampled and corresponding periodogram

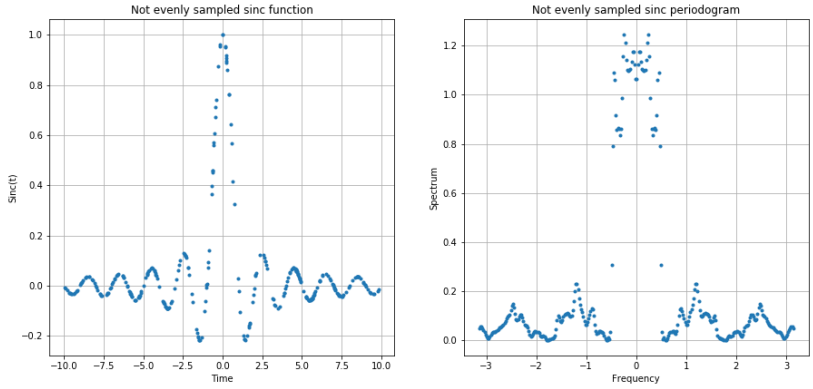

Fig. 5: Simple sinc function not evenly sampled and corresponding periodogram

We also applied the procedure of dropping samples and re-interpolating on 5 seconds of ECG signal. Figures 6 and 7 show us the application of the classical periodogram to this signal, as well as applied to the harmonic generated one.

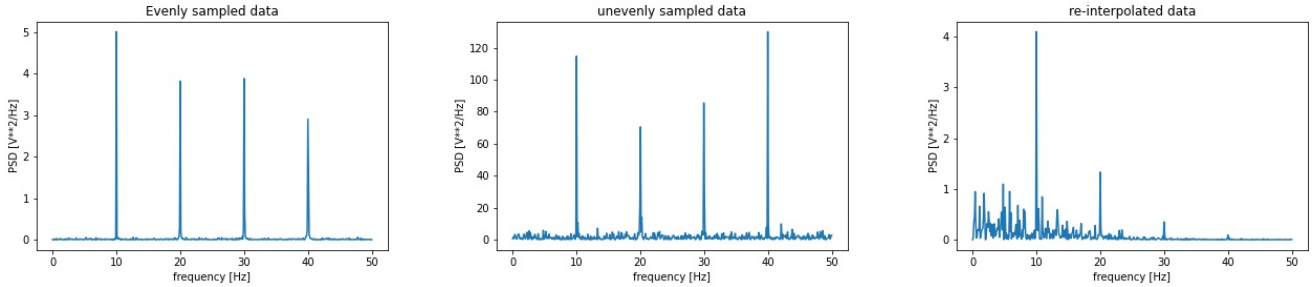

Fig. 6: Periodogram applied to evenly and unevenly sampled harmonic signal, and to the re-interpolated signal

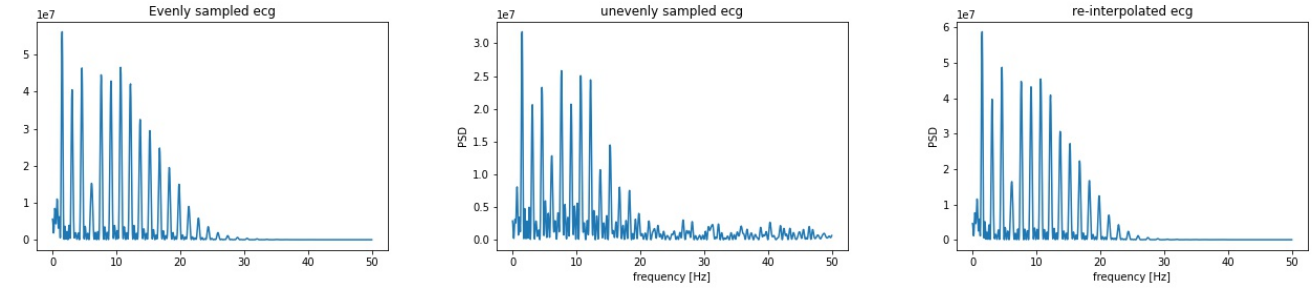

Fig. 7: Periodogram applied to 5 seconds of evenly and unevenly sampled ECG, and to the re-interpolated signal

B. Lomb-Scargle periodogram

Now we display the results, in figure 8, obtained when applying the Lomb-Scargle periodogram to the sum of sinusoid and to the ECG. The procedures remain the same.

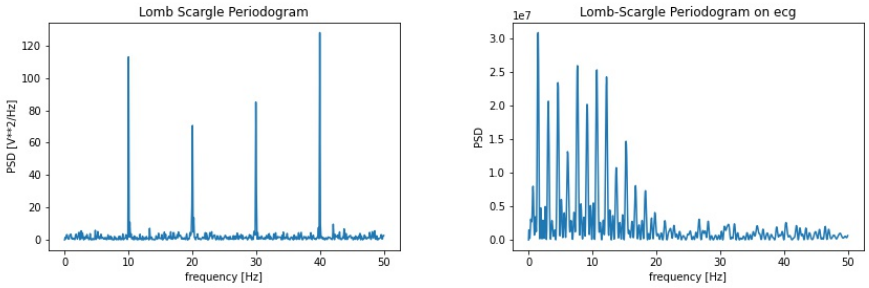

Fig. 8: Lomb-Scargle periodogram applied to the unevenly harmonic signal and ECG

IV. Discussion

After having implemented each algorithm and tested them independently on data, we can make a comparison and explain observations that we made. As we have seen in the procedure subsection, where we explain the theoretical part of both algorithms, they are nearly identical with the main exception being that the Lomb-Scargle periodogram is a generalized form [6]. With also the help of [7], we can describe the Lomb-Scargle periodogram a least-squares spectrum for fitting a sinusoid. With a complex signal, here the ECG, the Lomb-Scargle periodogram achieves the same performances as the classical one.

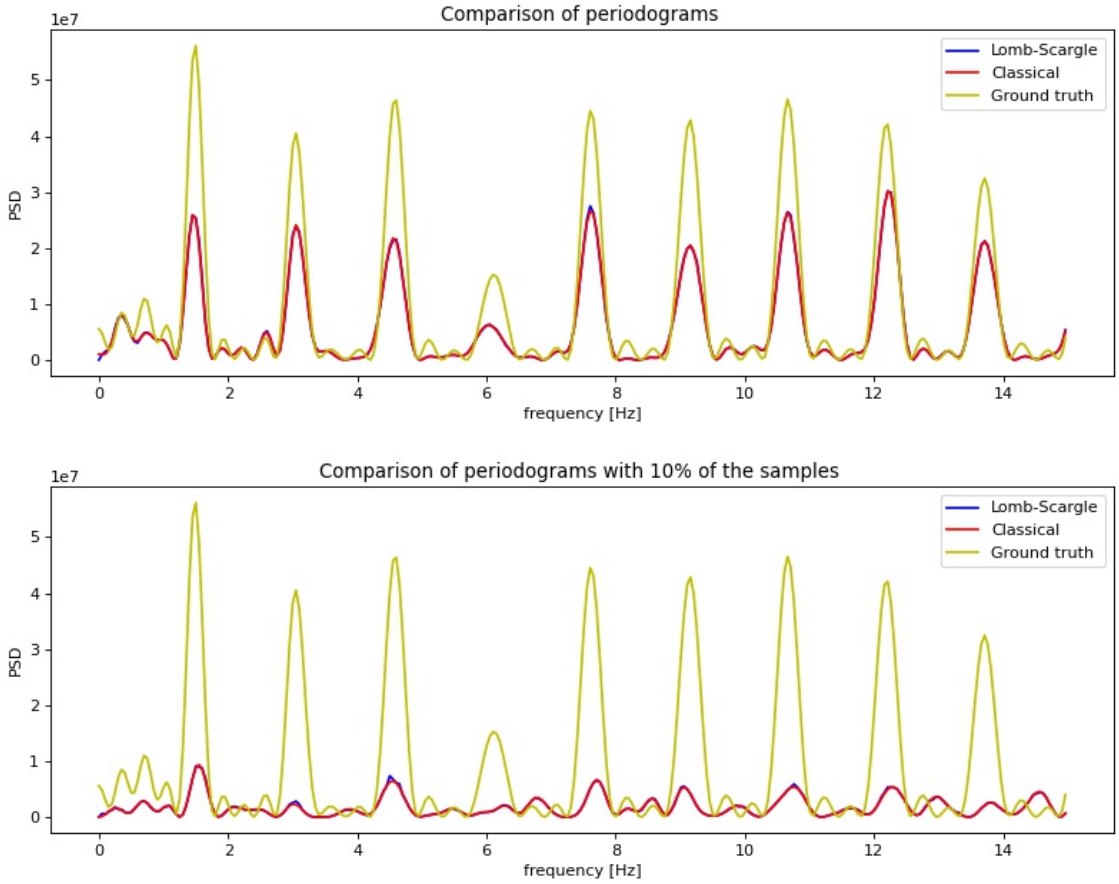

Fig. 9: Lomb-Scargle and classical periodogram applied to unevenly sampled ECG, then compared with the spectrum of the entire ECG. In the second plot, only 10% of the samples are kept

We can see in the plots shown in the result section that with the harmonic signal, applying the classical periodogram without re-interpolating the signal works better, the peak frequencies stay the same but their magnitude changes. The PSD of the re-interpolated signal shows a low pass effect on the signal, which makes the higher frequency peaks less visible.

When applying the same procedure on the five seconds of ECG signal, the PSD of the re-interpolated signal is quite similar to the original one. The PSD of the unevenly sampled signal is quite different. The peak frequencies are not the same and dropping samples seems to create higher frequencies in the periodogram.

Considering the Lomb-Scargle periodogram, it seems to have less of a low-pass effect on the generated data as we can see in the plots. When comparing the classical and the Lomb-Scargle periodogram, we can see that they achieve the same performance, even when a lot of samples are thrown away.

We also compared the mean squared errors of the two methods over (1). We computed the PSD of (1) analytically and used it as ground truth for the MSE. For this case we didn't drop the data but uniformly sampled the data for the unevenly sampled data. There are three main things we can note from Figure[10]. The first thing is that the MSE decrease when we increase the sampling frequency: the more samples of the same signal we have, the more both methods are accurate. We can also note how on average the evenly sampled signal has a lower MSE, but this difference get smaller the more samples we have. Last

thing to note is how the classical and lomb-scargle periodograms MSEs are almost identical for both the evenly and unevenly sampled data.

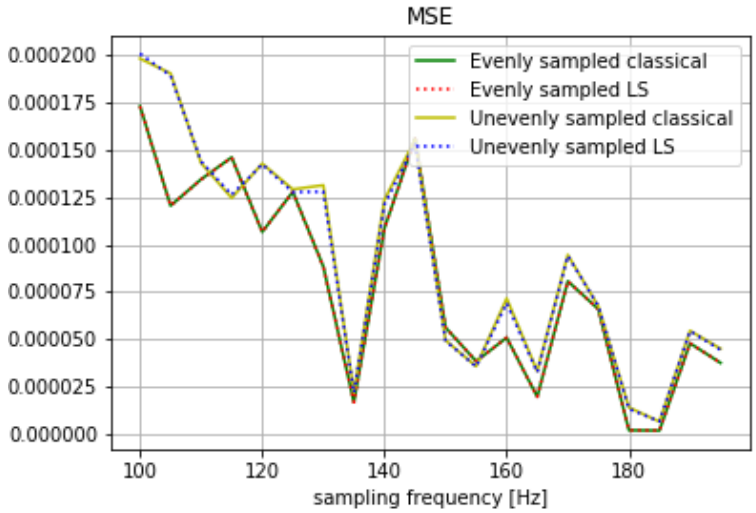

Fig. 10: Mean squared errors over (1)

V. Conclusion

To conclude, it'd seem that depending of the need of the analysis, one periodogram or the other will achieve slightly better.

The classical one is probably easier to implement even though nowadays pre-made functions already exists for anything. With evenly sampled, basic data, the classical periodogram is a bit quicker.

However, despite some comments from [6], the Lomb-Scargle periodogram has a very interesting mathematical aspect and does perform well when used for unevenly sampled, and, or, time-shifted signals. It shines by its generalization form.

In this project we have focused on understanding and testing both algorithms. Another interesting approach would be to test them both on longer signals, and also try to change up the sampling. We only generated unevenly sampled signals by dropping samples, generating random timestamps or dropping entire chunks of the signals could better approach real life situations.

It's finally worth looking at the last figure, Figure 9, to see that both the classical and the Lomb-Scargle periodogram get nearly the exact same results, independent of the sampling, quantity and randomness.

REFERENCES

[1] M. Hayes, *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[2] J. T. VanderPlas, "Understanding the Lomb-Scargle Periodogram," *ArXiv e-prints*, Mar. 2017.

[3] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[4] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 360–371, 2006.

[5] J. D. Scargle, "Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data.," , vol. 263, pp. 835–853, Dec. 1982.

[6] R. Vio and P. Andreani, "A critical comparison of the lomb-scargle and the classical periodograms," 2020.

[7] M. Zechmeister and M. Kürster, "The generalised lomb-scargle periodogram-a new formalism for the floating-mean and keplerian periodograms," *Astronomy & Astrophysics*, vol. 496, no. 2, pp. 577–584, 2009.