# Estimation Laboratory 2

Nathan Magnan, Mathieu Roule

Novembre 2020

## 1 Introduction

**Question 1 :**

For a given data set stored in vector $\underline{y}$, the power mean is given by eq. (1) :

$$P_y = \frac{\underline{y}^T \cdot \underline{y}}{N} \tag{1}$$

The definition for the power mean of the noise is much the same, but simplifies under the hypothesis of an independent, identically distributed, centered noise :

$$\begin{aligned} P_\epsilon &= \frac{\underline{\epsilon}^T \cdot \underline{\epsilon}}{N} \\ &= \left\langle \epsilon_0^2 \right\rangle \\ &= \sigma^2 \end{aligned} \tag{2}$$

Hence, we can expect a Signal to Noise Ratio of 20dB if $\sigma = \frac{1}{10}\sqrt{\frac{\underline{y}^T \cdot \underline{y}}{N}}$.

## 2 Spectral analysis with the Fourier transform

### 2.1 Estimation for known frequencies

**Question 1 :**

We denote as $\underline{\epsilon}$ the noise vector of size $N$, as $\underline{c}$ the complex amplitudes vector of size $2K+1$, and as $\underline{y}$ the data vector of size $N$. We also define the matrix $\underline{\underline{W}}$ of size $N \times (2K+1)$ whose coefficients are :

$$W_{n,k} = e^{2j\pi f_k t_n} \tag{3}$$

From all these definitions, it is clear that :

$$\underline{y} = \underline{\underline{W}}\,\underline{c} + \underline{\epsilon} \tag{4}$$

**Question 2 :**

We are considering a linear problem with additive, independent, centered, Gaussian noise, hence the Maximum Likelihood Estimator is simply :

$$\overset{\circ}{\underline{c}}_{ML}(\underline{y}) = (\underline{\underline{W}}^\dagger \underline{\underline{W}})^{-1} \underline{\underline{W}}^\dagger \underline{y} \tag{5}$$

**Question 3 :**

We implement in MatLab the preceeding equations, and add a method to retrieve the real amplitudes and phase angles from the complex amplitudes :

$$\begin{aligned} \overset{\circ}{\underline{A}}_{ML} &= 2|\overset{\circ}{\underline{c}}_{ML}| \\ \overset{\circ}{\underline{\phi}}_{ML} &= angle(\overset{\circ}{\underline{c}}_{ML}) \end{aligned} \tag{6}$$

We get the results in Tab. 2.1, from which we find that the estimation of the real amplitudes are very good (the relative error is lower than the 1 % noise in the data). However, the estimates of the phases are good, but not as much, with a relative error up to 7 times higher than the noise in the data. Still, the estimates yield a 10 dB Signal-to-Noise Ratio in the worst case, which is acceptable.

| Parameter | Value | Estimation | Relative error |
|---|---|---|---|
| $A_0$ | 0 | 0.008 | N/A |
| $A_1$ | 0.25 | 0.253 | 1.2 % |
| $\phi_1$ | 0.393 | 0.364 | 7.0 % |
| $A_2$ | 0.75 | 0.748 | 0.2 % |
| $\phi_2$ | 0.996 | 0.975 | 2.1 % |
| $A_3$ | 1 | 0.994 | 0.6 % |
| $\phi_3$ | 0.493 | 0.487 | 1.2 % |
| $A_4$ | 0.75 | 0.752 | 0.2 % |
| $\phi_4$ | 0.281 | 0.290 | 3.2 % |
| $A_5$ | 1 | 1.001 | 0.1 % |
| $\phi_5$ | 0.596 | 0.591 | 0.8 % |

Table 1: Table of comparison between original and estimated parameters

## 2.2 Estimation for unknown frequencies

**Question 1 :**

In the irregular sampling case, we cannot simply use the Shannon-Nyquist criteria, because there is no "sampling frequency". We have to study the shape of the spectra to deduce for which frequencies it is reliable, and where the sampling lost information.

The highest frequency in our data in around $35\,days^{-1}$, and we chose to draw a spectrum up to about 3 times that frequency, hence $\nu_{max.} = 100\,days^{-1}$. The smallest difference between 2 frequencies in our data is $0.3\,days^{-1}$, and we chose to draw a spectrum with a frequency comb 10 times thinner than that, hence $M = 3000$.

**Question 2 :**

We plot the time and frequency representations of the single-frequency signal in fig. 1, in both the noisy and non-noisy case.

- For the non-noisy signal, we read easily that the amplitude is close to 0.25 on the time representation, and that the frequency is close to $31\,days^{-1}$ on the frequency representation. We could also read that the amplitude is close to $2 \times 0.12 = 0.24$ on the frequency representation. The true values where $\nu = 31.012\,days^{-1}$ and $A = 0.25$. Hence the parameters are well estimated visually.

- For the noisy signal, we can still estimate the frequency near $31\,days^{-1}$ from the frequency representation. However, we cannot estimate the amplitude from the time representation anymore, we have to rely on the frequency representation to estimate an amplitude around 0.24. But on the whole, the estimated parameters seem as good as in the non-noisy case.

- We did not find a reliable way to estimate the phase of the signal from the plots. We tried to look at the phase of the Fourier transform at the frequency we just found, but the phase of the Fourier transform is too noisy. We tried to look at the time delays of the maximums in the time representation, and to link them to the phase, but even in the non noisy case the imprecision on the frequency makes this task impossible. The only solution left is to use the frequency we found and do the analysis of sec. 2.1 again.

- In the frequency domain, we observe a background noise around 0.02, hence we will not be able to detect a frequency with an amplitude lower than this. We also observe high secondary peaks about $1.5\,days^{-1}$ on both side of the main peak. We believe they are due to the gaps in the sampling, since we see on the time representation that these gaps are periodic with a period close to 1 day. They will clearly be a big issue when we will want to estimate the frequencies in a multiple-frequency signal.

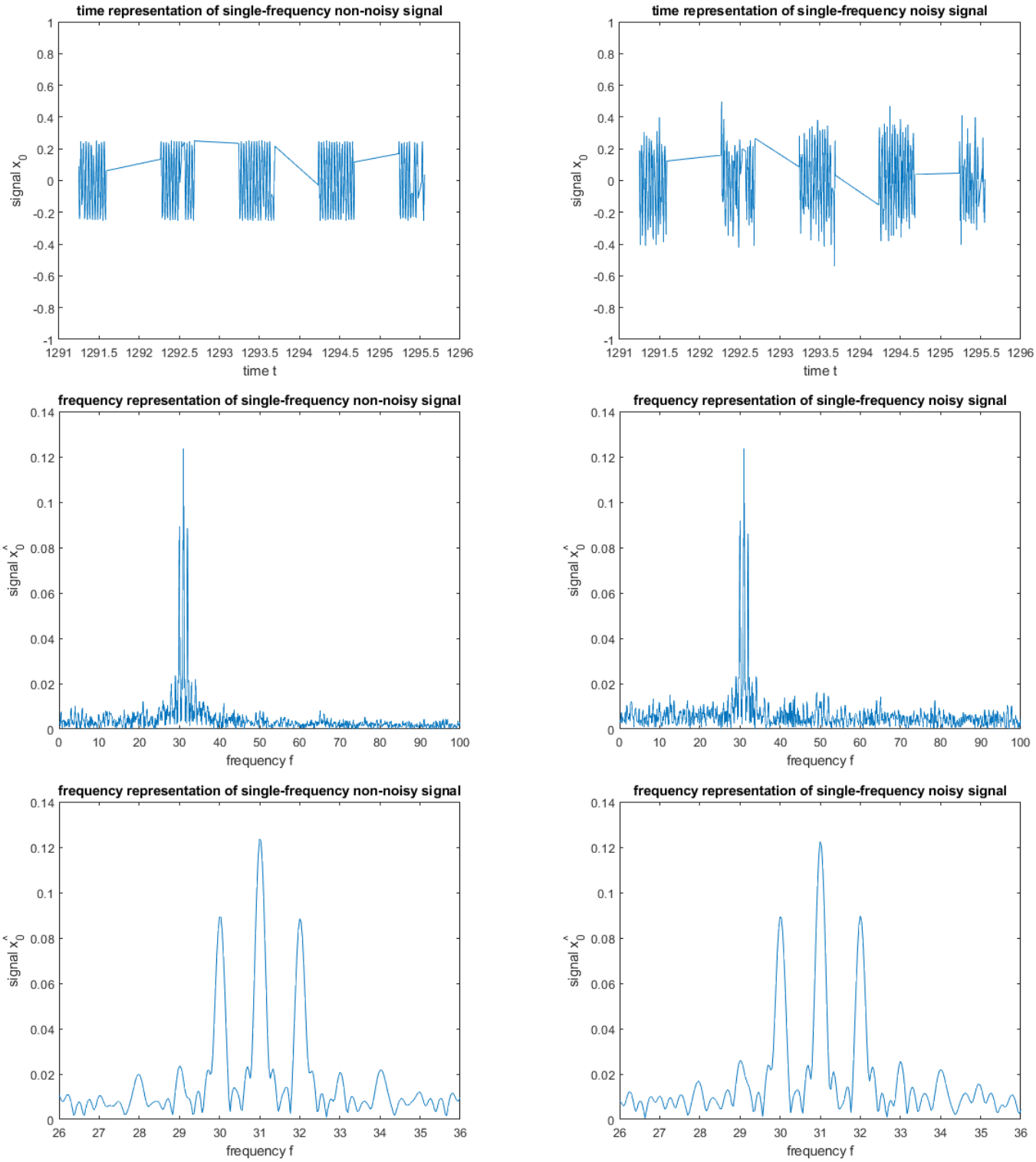


Figure 1: Time and frequency representation of the irregularly sampled, single-frequency signal.

- We observe that there is no Shannon-Nyquist mirror peak due to sampling. This was expected in the irregular sampling case.

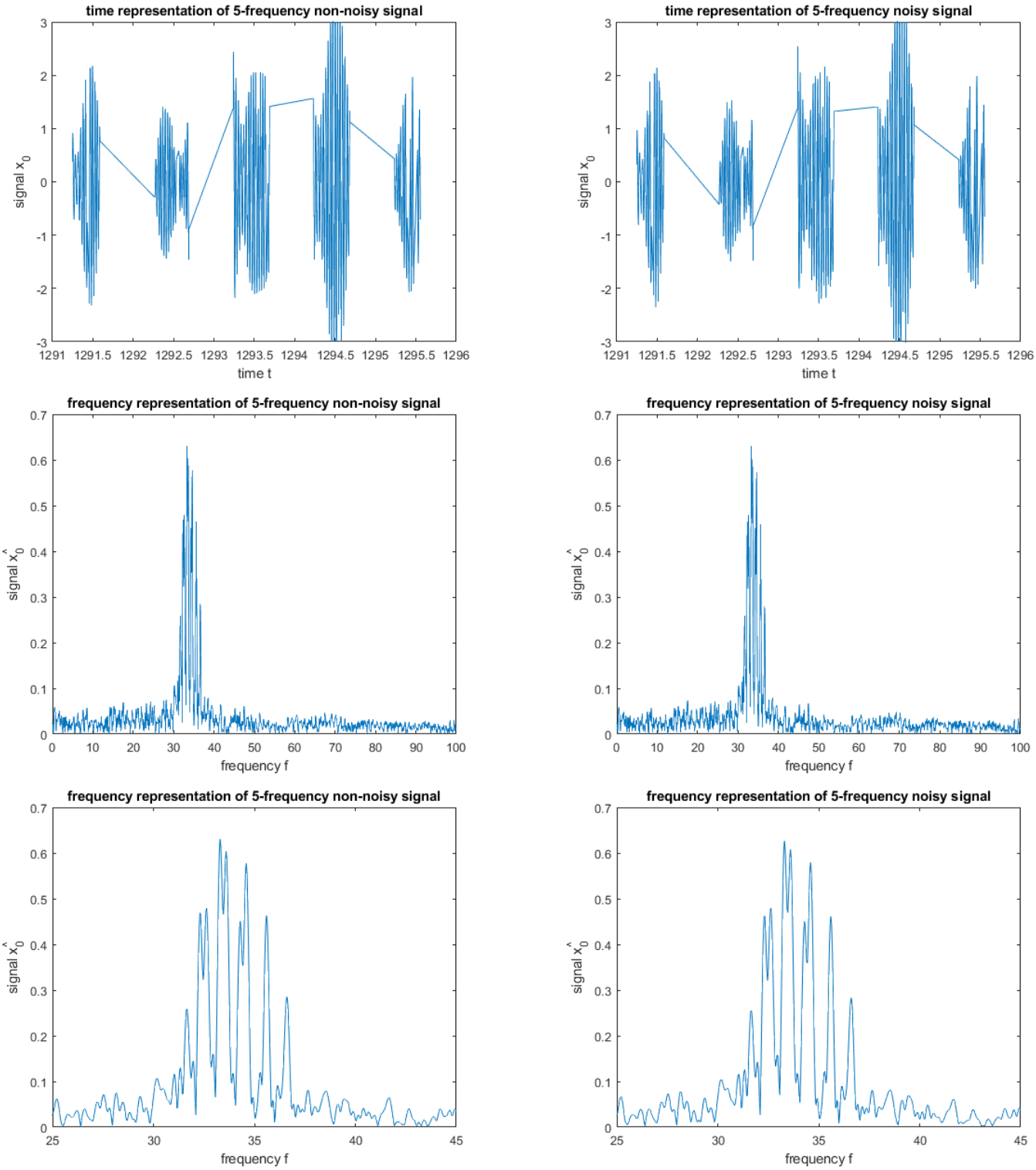


Figure 2: Time and frequency representation of the irregularly sampled, multiple-frequencies signal.

**Question 3 :**

We plot the time and frequency representations of the multiple-frequencies signal in fig. 2, in both the noisy and non-noisy case.

In both cases, we find very similar results : nothing can be deduced from the time representation due to the frequencies mix. And in the frequency representation there is a peak between $30\,days^{-1}$ and $40\,days^{-1}$, but when we zoom we find that this peak is divided in many secondary peaks that do not correspond to any of the expected frequencies. In particular, there is no clear peak at the position of the smallest frequency at $31.012\,days^{-1}$.

What happens is that because of the secondary peaks observed earlier, we cannot distinguish the five frequencies from each other, or from the effects of the sampling. Hence we cannot deduce anything from these spectral representations.

**Question 4 :**

We find the spectral window from fig. 3 :

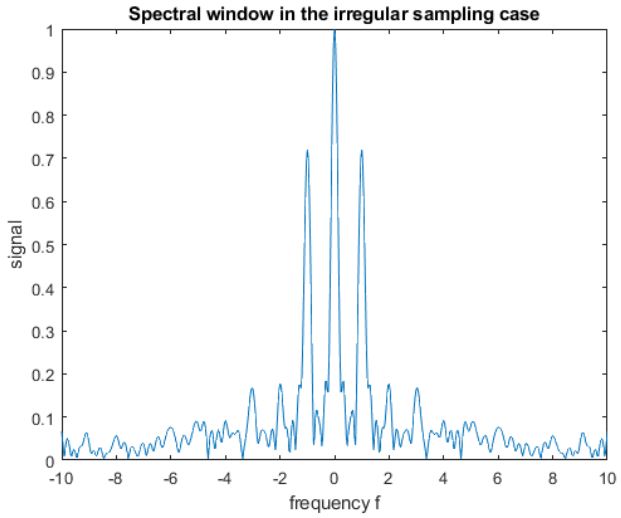


Figure 3: Spectral window for irregular sampling times. We observe symmetric secondary peaks at around 70% amplitude and about $1\,days^{-1}$ around the main peak.

We observe secondary peaks on both sides of the main peak. This is what we predicted in question II.2.2 : The irregular sampling leave an important footprint on the frequency representation. This explains the secondary peaks we observed in question II.2.2, and it also explain why we could not differentiate the frequencies in the multiple-frequency signal : the signal's difference in frequencies was close the secondary peaks' distance to the main peak.

# 3 Sparse representation with greedy algorithms

In this part, we have set $\nu_{max}$ to $50\,days^{-1}$ and M to 5000.

## 3.1 *Matching Pursuit* (MP) algorithm

**Question 2 :**

For `x_0`, the *Matching Pursuit* algorithm detect 2 frequencies : $31.02 days^{-1}$ and its opposite $-31.02 days^{-1}$. This was easily predictable since $cos(\theta) = \frac{1}{2}(e^{i\theta} + e^{-i\theta})$. The small error is due to the sampling of the frequency grid and the result is really satisfying. The amplitude has been nearly equally divided between the positive frequency and its opposite one (0.1272 and 0.1279 respectively). The sum of the amplitude is a little bit higher than the original one (0.25) but it is also pretty satisfying - relative error of 2%-.

For `x`, if we consider only the different frequencies in absolute value, we have 28 different frequencies between $30.67\,days^{-1}$ and $41.7\,days^{-1}$. with some which are really near to one another. To get a better idea, of the distribution of this frequencies and their amplitudes compared to the 5 original ones, we have plotted the estimated spectrum in figure (4). The 5 true frequencies are plotted in red and the 28 estimated one are in blue. We note that the estimated frequencies are

not completely false, they are in the good range of values $(30-40) days^{-1}$ and 4 of them are quite correctly estimated. But the second one ($32.675\, days^{-1}$) is badly recovered, and even if most of the false frequencies have small amplitudes some of them could be wrongly interpreted as characteristic frequencies (especially $31.67\, days^{-1}$). Furthermore, the amplitudes have a mean 25% relative error.

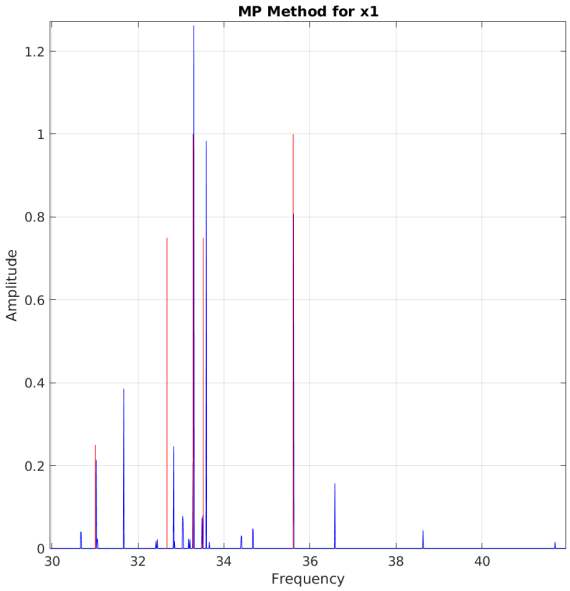


Figure 4: Spectrum of the data `x` estimated by *Matching Pursuit* algorithm - in blue - compared to the true one - in red -

**Question 3 :**

| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Detected frequency | -31.02 | 31.02 | -33.30 | 33.30 | 33.59 | -33.59 | -35.62 | 35.62 | -31.67 | 31.67 |
| Order | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Detected frequency | -32.83 | 32.83 | -33.28 | 31.03 | -31.03 | 33.28 | 33.51 | 36.58 | -33.04 | -33.49 |
| Order | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Detected frequency | -36.58 | 33.05 | -34.67 | 34.68 | -30.68 | 30.67 | 34.41 | -34.40 | -33.18 | 31.06 |
| Order | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | |
| Detected frequency | -38.63 | 38.63 | 32.45 | 33.21 | -31.05 | 32.85 | -32.42 | -33.66 | 41.70 | |

Table 2: Detected frequencies during the iterations of the *Matching Pursuit* algorithm

The table (2) summarizes the detected frequencies along the execution of the *Matching Pursuit* algorithm. In green, we can see the frequencies which have been appropriately detected, in the logical order : the negative one first due to the maximum research implementation - we change the estimated index if and only if we find a strictly better index ; since the negative and the positive frequencies should give the same result and since we search increasingly from $f = -50.00$ to $f = 50.00$, we expect the negative frequency to be chosen before the positive one. In orange, the frequencies near from true frequencies and detected in an unexpected order. In red, the false detection.

Over the 8 first iteration, the *Matching Pursuit* algorithm gives satisfying results, identifying the 4 out of 5 frequencies, in the order of increasing frequency (only the 2nd one is missing - and will be missing until the end), but then return false detection mixed with good detection in wrong order and after the 20th iteration it only returns wrong frequencies.

We could have stopped the algorithm when the frequencies became detected in an unexpected order (at iteration 14 for example). It could be a interesting stopping criterion (it could be coded as : if there is no 2 common absolutes values over the last 3 detected - then break).

## 3.2 *Orthogonal Matching Pursuit* (OMP) algorithm

**Question 2 :**

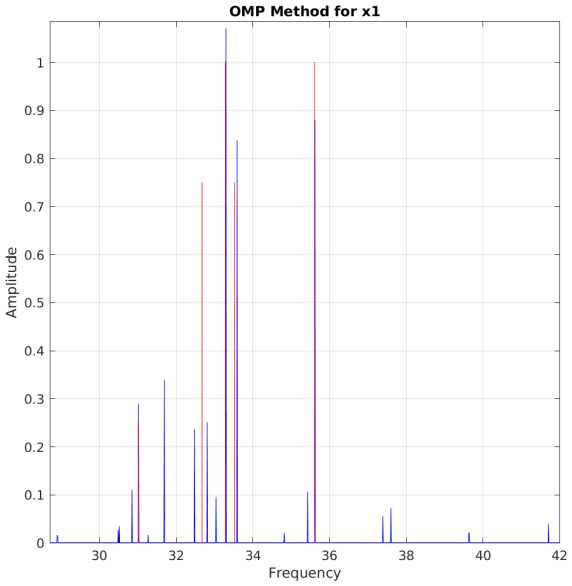


Figure 5: Spectrum of the data `x` estimated by *Orthogonal Matching Pursuit* algorithm - in blue - compared to the true one - in red -

For `x_0`, we obtain the same result (same frequencies) with *Orthogonal Matching Pursuit* as with *Matching Pursuit*. The amplitude is equally distributed between the positive frequency and its opposite (which is quite better than the *Matching Pursuit*), the relative error is a little bit higher (2.3%) but still is very good.

For `x`, we have 21 different frequencies (in absolute values) which is better than the results obtained using *Matching Pursuit* algorithm. But this frequencies are ranged between 28.9 $days^{-1}$ and 41.71 $days^{-1}$ which is more spread than with *Matching Pursuit*. We have plotted in figure (5) the estimated spectrum vs the true one as for *Matching Pursuit* before. We can see that there is less noise around the third (33.283 $days^{-1}$ )and fourth (33.521 $days^{-1}$) frequencies, but a little bit more around the first one (31.012 $days^{-1}$). The second one is still not satisfyingly approached but there is now two pics around it and there still is an issue at 31.69 $days^{-1}$ frequency (was 31.67 $days^{-1}$ for *Matching Pursuit*). Finally, the amplitude estimation are better with a mean 11.5% relative error on the 4 out of 5 well recovered frequencies.

**Question 3 :**

As in question (3.1), we have summarized the iteration in table (3) with the same color coding.

| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Detected frequency | -31.02 | 31.02 | -33.30 | 33.30 | -33.59 | 33.59 | -35.62 | 35.62 | -31.69 | 31.69 |
| Order | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Detected frequency | 31.02 | -31.02 | -32.81 | 32.81 | -32.48 | 32.48 | -33.04 | 33.04 | -30.85 | 30.85 |
| Order | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Detected frequency | -35.43 | 35.43 | -37.60 | 37.60 | 37.39 | -37.39 | 39.63 | 41.71 | -30.52 | -39.64 |
| Order | 31 | 32 | 33 | 34 | 35 | 36 | | | | |
| Detected frequency | -41.71 | 30.49 | 34.82 | 28.90 | -31.27 | -28.91 | | | | |

Table 3: Detected frequencies during the iterations of the *Orthogonal Matching Pursuit* algorithm

As for the *Matching Pursuit*, the first 8 iterations are good (even better since there is no negative/positive frequency order issue) but then the frequency $f = 31.69 days^{-1}$ is wrongly detected, and there is an sign inversion in iterations 11 and 12, and until the end only wrong frequencies are detected. A stopping criterion as the one suggested for *Matching Pursuit* will not work for *Orthogonal Matching Pursuit*. But we could have stopped at the first sign inversion.

### 3.3 *Orthogonal Least Squares* (OLS) algorithm

**Question 1 :**

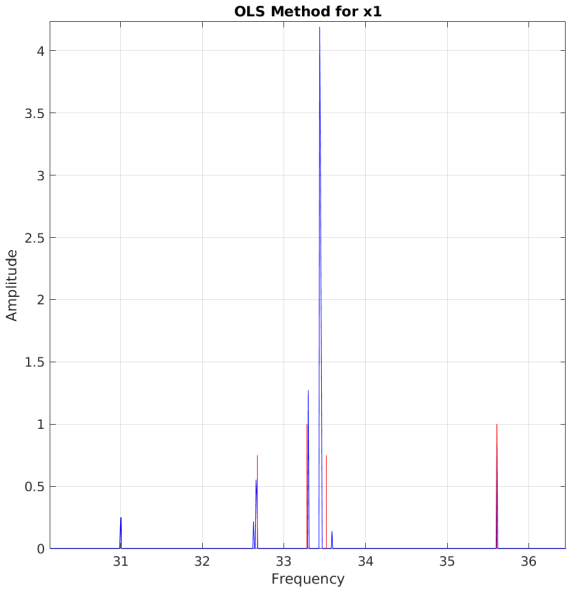


Figure 6: Spectrum of the data `x` estimated by *Orthogonal Least Square* algorithm - in blue - compared to the true one - in red -

For `x_0`, we obtain the exact same results very satisfying as with *Orthogonal Matching Pursuit*.
For `x`, we have 10 different frequencies (in absolute values) which is really better than the results obtained using *Matching Pursuit* or *Orthogonal Matching Pursuit* algorithms. These frequencies are ranged between $31.0\,days^{-1}$ and $35.61\,days^{-1}$ which is a truly good result. There is still one poorly estimated frequency but now it is the fourth one ($33.521\,days^{-1}$) and no more the second one ($32.675\,days^{-1}$). Furthermore, close to this fourth frequency there is an estimated pic spreading on 3 frequencies of the grid (33.44 , 33.45 and $33.46\,days^{-1}$) with an amplitude of nearly 4.2, 3 and 1.5 respectively. This is very surprising compared to the quality of the estimation apart from this error and it could be truly misleading in a study were the true state is unknown. The results have been summarized in tab (4).

| True frequency | True amplitude | Estimated frequencies | Estimated amplitudes |
|---|---|---|---|
| 31.012 | 0.25 | 31.00 | 0.2514 |
| 32.675 | 0.75 | 32.63 | 0.2177 |
| | | 32.66 | 0.5536 |
| | | 32.67 | 0.4171 |
| 33.283 | 1 | 33.30 | 1.2722 |
| 33.521 | 0.75 | 33.44 | 4.1929 |
| | | 33.45 | 2.9772 |
| | | 33.46 | 1.5354 |
| | | 33.59 | 0.1407 |
| 35.609 | 1 | 35.61 | 0.9991 |

Table 4: Spectrum of the data `x` estimated by *Orthogonal Least Squares* vs true spectrum

**Question 2 :**

For this method, we do not have an *a priori* on the order negative/positive frequency, but we note that there is a change at iteration 11 (negative/positive → positive/negative). The number of iterations/detected frequencies have been considerably reduced. The first detected frequency is no more the lowest one ($31.012 days^{-1}$). The order of detection seems to be guided by the amplitude more than by frequency (the frequencies with the highest amplitudes are detected before the ones with small amplitudes) but it is not really clear.

| Order | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Detected frequency | -33.30 | 33.30 | -33.59 | 33.59 | -35.61 | 35.61 | -32.63 | 32.63 | -33.44 | 33.44 |
| Order | 11 | 12 | 13 | 14 | 15 | 16 | | | | |
| Detected frequency | 31.00 | -31.00 | 32.67 | -32.66 | 33.46 | -33.45 | | | | |

Table 5: Detected frequencies during the iterations of the *Orthogonal Least Square* algorithm

**Remark :** This analysis have been done on data noised accordingly to the given data set `y`. By setting the noise correctly (accordingly to the create data), we obtain similar results with an improvement for the *Orthogonal Least Square* method. Indeed, the pic of amplitude decrease to about 2.2 at its maximum.

# 4 Sparse representation with convex relaxation

**Question 1 :**

The amplitude in the data is of order 0.25. Hence, in the Fourier domain, the lowest spectral amplitude will be $0.125 \approx 0.1$. We want a Signal-to-Noise Ratio of 20dB, Hence the maximal spectral amplitudes of the residuals should be lower than $10^{-3}$. Since at convergence we can prove that the maximal spectral amplitude of the residuals is lower or equal to $\lambda/N$, with $N \approx 500$, we chose $\lambda = 0.5$.

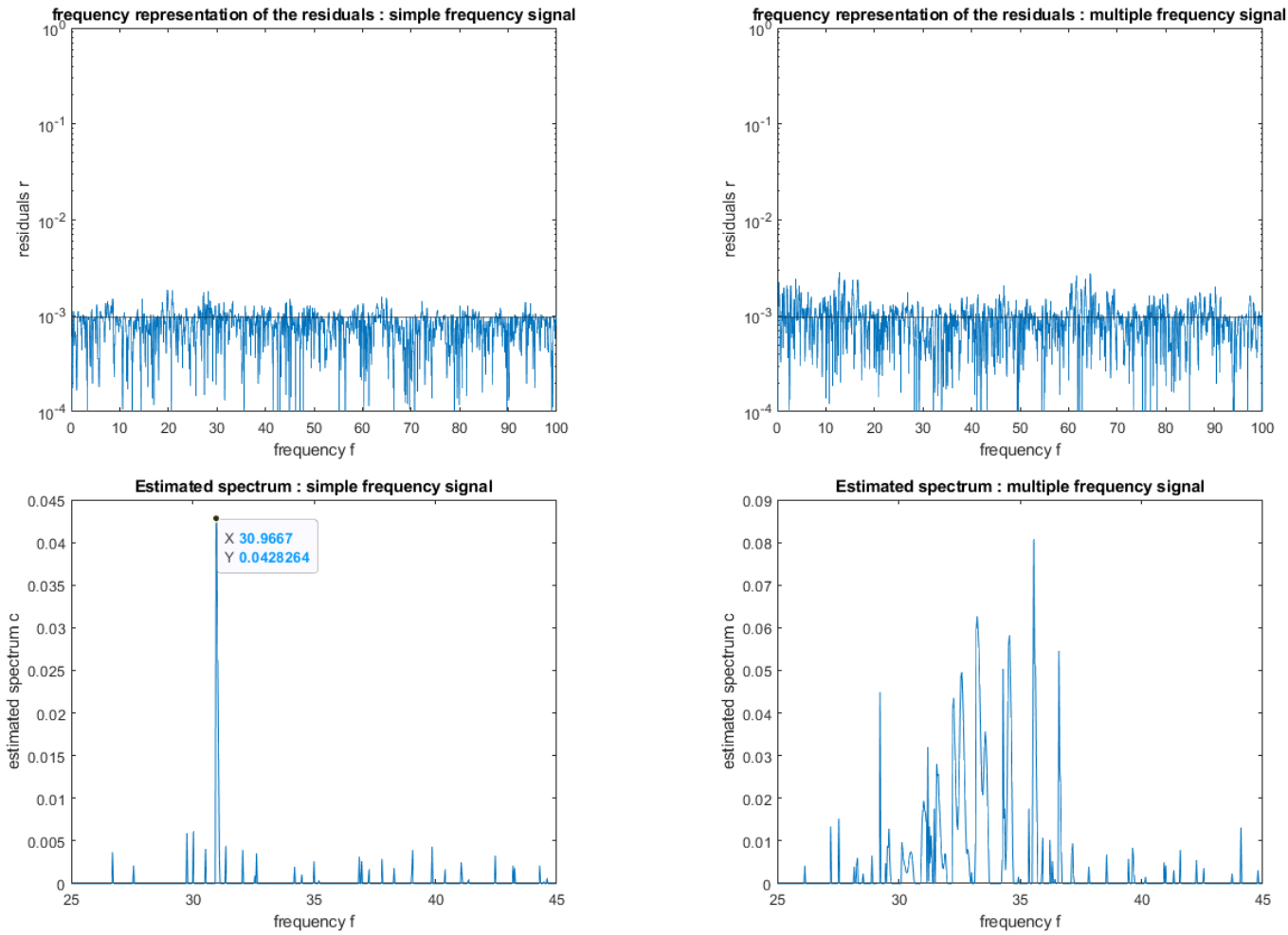


Figure 7: Spectrum of the residuals and estimated spectrum with the convex relaxation method, in the simple-frequency case and the multiple-frequency case. We notice that the residuals are low enough for the signal-to-noise ratio to be at 20dB. But they are slightly above the black line which represents the convergence level, so we do not have pure convergence, only approximate convergence.

For this value of $\lambda$, we find that we need 100 iterations of the algorithm to reach convergence. More precisely, as shown on fig. 7, we would need a little bit more than 100 iterations for pure convergence, but the residuals are already acceptable for 100 iterations so we keep this value and reduce the computation time.

**Question 2 :**

With the simple frequency signal, we see that the estimated spectrum exhibits a single sharp peak at $31 \pm 0.2\, days^{-1}$. The expected value being $31.012\, days^{-1}$, the estimated frequency is good. However, the estimated amplitude of $2 \times 0.04 = 0.08$ is far from the expected value of 0.25. Hence we will have to estimate the amplitude (and the phase) with the method from sec. 2.1.
Let's also notice that this issue was expected, is is caused by the relaxation : once the right frequency is detected, changing the height of the peak will not produce a big difference is the norm L2 of the residuals, because the norm L2 is integral and the peak is very sharp. But the height of the peak will have a big effect on the norm L1 of the estimated spectrum. Hence the algorithm reduces the height of the peak to lower the L1 norm of the spectrum.

With the multiple frequency signal, the picture is shadier. We see 6 or 7 peaks in the estimated spectrum, several of them being double peaks. By decreasing order of amplitude, the frequencies are $35.5\, days^{-1}$ , $33.2\, days^{-1}$, $34.6\, days^{-1}$, $36.6\, days^{-1}$, $32.3\, days^{-1}$, $29.2\, days^{-1}$ and $31.2\, days^{-1}$. We expected only 5 frequencies (all single-peaked), at different places. And, once again, wildly different amplitudes.

Therefore we try to regularize our problem even more and increase the relaxation parameter $\lambda$ to 3, and we get the results of fig. 8 for 100 iterations :

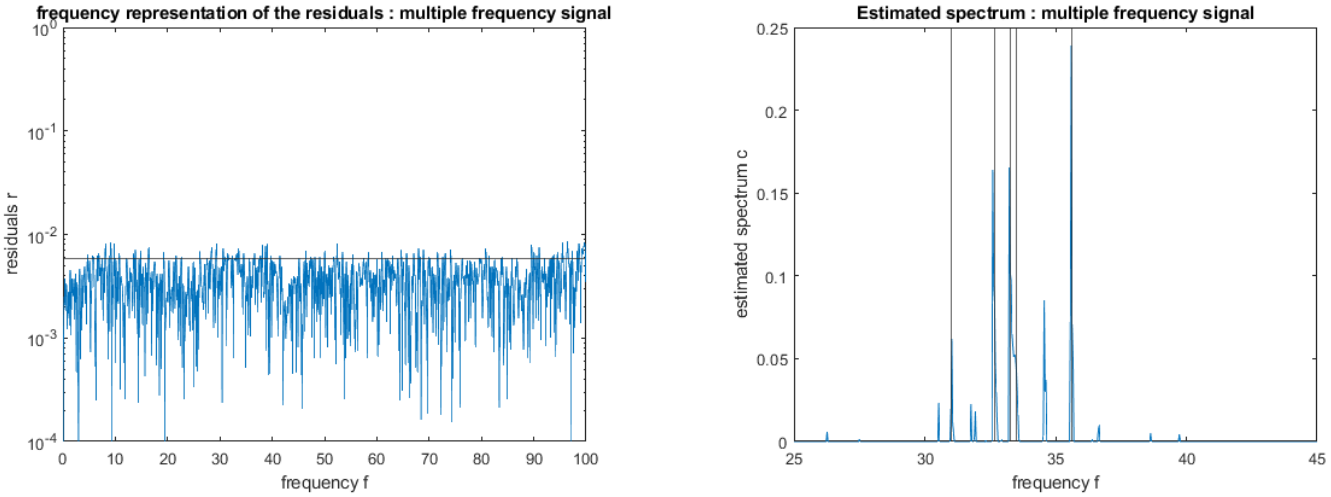


Figure 8: Spectrum of the residuals and estimated spectrum with the convex relaxation method, in the multiple-frequency case, for $\lambda = 3$. The vertical black lines represent the expected frequencies.

From the residuals, we see that we are still close to convergence of the optimisation algorithm. And from the estimated spectrum, we see that for an higher value of $\lambda$ the predicted frequencies are better : we see 5 or 6 peaks, one of them being a double-peak. By decreasing order of magnitude the frequencies are $35.6\, days^{-1}$, $32.6\, days^{-1}$, $33.2\, days^{-1}$, $34.6\, days^{-1}$, $33.5\, days^{-1}$ and $31.0\, days^{-1}$. This means that we have detected all the frequencies of the input signal !
But we also detected a $6^{th}$, foreign, frequency.

**Question 3 :**

First of all, we did not observe the double-peaks invoked in the subject, so we did not have to choose a frequency from each pair.

We run the algorithm of sec. 2.1 with the 6 frequencies found just above, and we get the amplitudes stored in Tab. 4. We find that the estimated amplitudes of the correct frequencies are in the correct order of magnitude, but they can be off the mark by up to 26%. And the estimated amplitude of the foreign frequency is quite high, higher several true frequencies.

| Frequency | Amplitude | Frequency | Amplitude | Frequency | Amplitude |
|---|---|---|---|---|---|
| $0\,days^{-1}$ | 0.013 | $33.2\,days^{-1}$ | 0.83 | 35.6 | 0.61 |
| $31.0\,days^{-1}$ | 0.22 | $33.5\,days^{-1}$ | 0.94 | | |
| $32.6\,days^{-1}$ | 0.94 | $34.6\,days^{-1}$ | 0.80 | | |

Table 6: Maximum Likelihood amplitudes, assuming the frequencies given by the convex relaxation method.

Therefore the results of the convex relaxation estimation method are better than those of the preceding methods, but still not quite sufficient for most astrophysical usages.

# 5 Conclusion

In this laboratory, we have experienced different solutions to extract the characteristic frequencies from a irregularly sampled data set. This irregularity in the sampling is common, for instance, for observations than could be done only during the night (or only during the day). As such, specialized data analysis tools have been developed to tackle this problem. One important point is that there is no Shannon-Nyquist maximal sampling period.

When the frequencies are known, we can reconstruct the amplitudes and phases easily with a maximum likelihood estimator, but we observe that phases are harder to estimate / more sensitive to noise than the amplitudes.

Estimating the spectrum with unknown frequencies is a true challenge. The Least Square gives an idea of the global location of the frequencies but fail to identify clearly the frequencies, due to mixed secondary peaks induced by the spectral window of the irregular sampling.

First, we introduced 3 iterative methods : *Matching Pursuit*, *Orthogonal Matching Pursuit* and *Orthogonal Least Square* to tackle this issue. The first method, *Matching Pursuit*, gave interesting results but too much frequencies were selected as if the algorithm was stopping too late but 4 out of 5 frequencies have been correctly detected (even if the amplitudes are not so good). Improving this method into the *Orthogonal Matching Pursuit*, the results are better since less wrong frequencies are detected and the amplitudes of the good ones begin to fit better with the original model. Finally, the *Orthogonal Least Square*, gave even better results. Nevertheless, this methods are not perfect at all, and it could be difficult to interpret the results without the original model to compare with but it still gives a first approximation not completely incoherent of the spectrum in a irregular sampling case which could be used as an *a priori* for other methods.

Finally we tried convex relaxation, and obtained better results than with the former methods. But they are still not quite satisfying : some estimated frequencies are not present in the original spectrum, the amplitudes are estimated only 25% accuracy, and it is impossible to estimate the phases.

In practice, our advice would be either use the convex relaxation method but to be careful with any affirmation stemming from the results. Or to use several methods independently, and to consider only the frequencies found by several or all methods. Or to consider whether we have a prior on the expected frequencies (which could possibly be constructed with the methods we have seen), in which case a Bayesianist approach would very likely give better results.