Assignment 2 (ML for TS) - MVA 2021/2022

Eya Ghamgui eya.ghamgui@telecom-paris.fr
Siwar Mhadhbi siwar.mhadhbi@telecom-paris.fr

June 28, 2022

1 Introduction

Objective. The goal is to better understand the properties of ARIMA processes, and do signal denoising with sparse coding.

2 General questions

A time series $\{y_t\}_t$ is a single realisation of a random process $\{Y_t\}_t$ defined on the probability space $(\Omega, \mathcal{F}, P)$, i.e. $y_t = Y_t(w)$ for a given $w \in \Omega$. In classical statistics, several independent realisations are often needed to obtain a "good" estimate (meaning consistent) of the parameters of the process. However, thanks to a stationarity hypothesis and a "short-memory" hypothesis, it is still possible to make "good" estimates. The following question illustrates this fact.

Question 1

An estimator $\hat{\theta}_n$ is consistent if it converges in probability when the number n of samples grows to ∞ to the true value $\theta \in \mathbb{R}$ of a parameter, i.e. $\hat{\theta}_n \xrightarrow{\mathcal{D}} \theta$.

- Recall the rate of convergence of the sample mean for i.i.d. random variables with finite variance.
- Let $\{Y_t\}_{t\geq 1}$ a wide-sense stationary process such that $\sum_k |\gamma(k)| < +\infty$. Show that the sample mean $\bar{Y}_n = (Y_1 + \cdots + Y_n)/n$ is consistent and enjoys the same rate of convergence as the i.i.d. case. (Hint: bound $\mathbb{E}[(\bar{Y}_n - \mu)^2]$ with the $\gamma(k)$ and recall that convergence in L_2 implies convergence in probability.)

Answer 1

Part 1:

Let $Y_1, Y_2, ..., Y_n$ be n i.i.d. random variables with overall mean μ and finite variance σ^2.

Let $\bar{Y}_n = \frac{Y_1+Y_2+...+Y_n}{n}$ be the sample mean. Then, according to the Central Limit Theorem (CLT), $\sqrt{n}(\frac{\bar{X}_n-\mu}{\sigma}) \underset{n\to\infty}{\longrightarrow} N(0,1)$, and the rate of convergence of the sample mean is of order $\frac{1}{\sqrt{n}}$.

Part 2:

- To show that the sample mean $\bar{Y}_n$ is consistent, we show that $\bar{Y}_n$ converges in probability to the mean μ. As we know that the convergence in L_2 implies the convergence in probability, it is sufficient to show that $\bar{Y}_n$ converges in L_2 to μ. To do so, we show that $\mathbf{E}[(\bar{Y}_n - \mu)^2] \underset{n\to\infty}{\longrightarrow} 0$.

$\mathbf{E}[(\bar{Y}_n - \mu)^2] = \mathbf{E}[\bar{Y}_n^2 - 2\mu\bar{Y}_n + \mu^2] = \mathbf{E}[\bar{Y}_n^2 - 2\mu\bar{Y}_n + \mu^2] = \mathbf{E}[\bar{Y}_n^2] - 2\mu\mathbf{E}[\bar{Y}_n] + \mu^2$

- $\{Y_t\}_{t>1}$: wide-sense stationary process

$$\implies \text{ we have stationarity of } \begin{cases} \text{order1 : } \forall n,\ \mathbf{E}[Y_n] = \mu \implies \mathbf{E}[\bar{Y}_n] = \mu \quad (\star) \\ \\ \text{order2 : } \mathbf{E}[Y_i Y_j] = \gamma(|i-j|) \qquad\qquad (\star\star) \end{cases}$$

$\mathbf{E}[(\bar{Y}_n - \mu)^2] = \mathbf{E}[\bar{Y}_n^2] - \mu^2$ (from($\star$)) $\implies \mathbf{E}[(\bar{Y}_n - \mu)^2] \leq \mathbf{E}[\bar{Y}_n^2]$

$$\bar{Y}_n^2 = \frac{1}{n^2}(\sum_{i=1}^n Y_i)(\sum_{j=1}^n Y_j) = \frac{1}{n^2}(\sum_{i=1}^n \sum_{j=1}^n Y_i Y_j) \implies \mathbf{E}[\bar{Y}_n^2] \;=\; \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[Y_i Y_j]$$

$$= \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \gamma(|i-j|) \text{ (from}(\star\star))$$

Which can be written as : $\frac{1}{n^2}\sum_{k=0}^n \alpha_k \gamma(k)$ with α_k the multiplicity of the term k that replaces $|i-j|$ for $(i,j) \in [1,n]$.

$$\implies \mathbf{E}[\bar{Y}_n^2] \leq \frac{1}{n^2}\sum_{k=0}^n \underbrace{max_k \ \alpha_k}_{=n} |\gamma(k)| = \frac{1}{n}\sum_{k=0}^n |\gamma(k)|$$

$$\implies (\mathbf{E}[\bar{Y}_n - \mu])^2 \leq \mathbf{E}[(\bar{Y}_n - \mu)^2] \leq \frac{1}{n}\sum_{k=0}^n |\gamma(k)|$$

- When we tend $n \to \infty$: $\begin{cases} \frac{1}{n} \xrightarrow[n\to\infty]{} 0 \\ \\ \sum_{k=0}^{\infty} |\gamma(k)| < \infty \end{cases}$:

  $\mathbf{E}[(\bar{Y}_n - \mu)^2] \xrightarrow[n\to\infty]{} 0$, hence the convergence in L_2, and therefore in probability.

- Moreover, $(\mathbf{E}[\bar{Y}_n - \mu])^2 \leq \frac{1}{n}\sum_{k=0}^n |\gamma(k)| \implies \mathbf{E}[\bar{Y}_n - \mu] \leq \frac{1}{\sqrt{n}}(\sum_{k=0}^n |\gamma(k)|)^{\frac{1}{2}}$.

  Thus, the rate of convergence of the sample mean $\bar{Y}_n$ is equal to $\frac{1}{\sqrt{n}}$ which is the same as the i.i.d case.

3 ARIMA process

Question 2 *Characteristic polynomial*

Let $\{Y_t\}_{t\geq 1}$ be an AR(2) process, i.e.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \tag{1}$$

with $\phi_1, \phi_2 \in \mathbb{R}$. The associated characteristic polynomial is $1 - \phi_1 x - \phi_2 x^2$. Properties on the roots of this polynomial drive the behaviour of this process.

- Choose ϕ_1 and ϕ_2 such that the characteristic polynomial has a complex root of norm 1. Simulate the process Y (with $n = 1000$) and display the signal and the periodogram. What do you observe?
- Choose ϕ_1 and ϕ_2 such that the characteristic polynomial has two complex conjugate roots of norm $r = 0.99$ and phase $\theta = 2\pi/3$. Simulate the process Y (with $n = 1000$) and display the signal and the periodogram. What do you observe?

Answer 2

- In the first part of the question, we chose the roots of the characteristic polynomial as $\{-j, j\}$ such that the characteristic polynomial is written as: $(x-j)(x+j) = x^2 + 1$. Thus, we found that $\phi_1 = 0$ and $\phi_2 = -1$.

  We simulated the process Y with $n = 1000$. The following figures show the simulated signal and its corresponding periodogram. We can clearly notice that this signal is non-stationary. Its absolute amplitude increases over time. Moreover, its periodogram shows a main peak around the frequency $2.5Hz$ and some other peaks around this frequency having almost null power spectral densities.

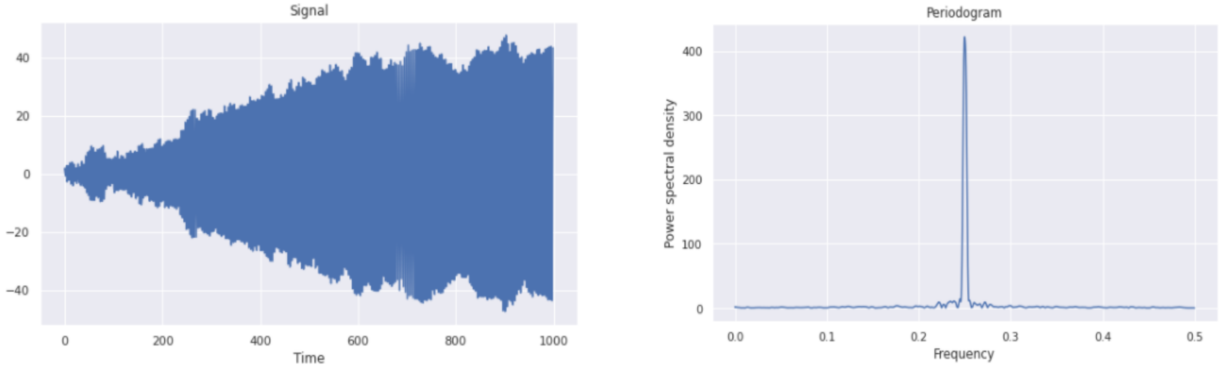


Figure 1: First AR(2) process

- In the second part of the question, we chose the roots of the characteristic polynomial as $\{0.99\ e^{-\frac{2\pi}{3}j}, 0.99\ e^{\frac{2\pi}{3}j}\}$. We found that $\phi_1 = -0.98999999999995$ and $\phi_2 = -0.9800999999999997$.

  The graph on the left corresponds to the simulated signal Y with $n = 1000$. This signal is now sationary. Indeed, the norms of the roots of the characteristic polynomial are different from 1 explaining the stationarity of the signal. The periodogram also reveals a main peak around the frequency $0.33Hz$ with a lower amplitude compared to the previous signal. Around this frequency there are several other peaks. This explains the flat looking series, showing no trend or periodic fluctuations.

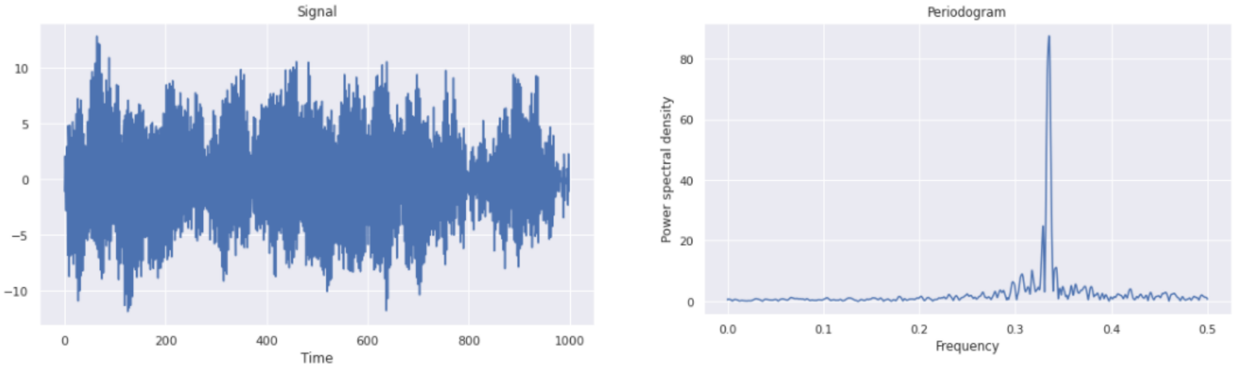


Figure 2: Second AR(2) process

Question 3 *Removing the trend by differencing*

The first step of the Box-Jenkins methodology consists in removing long-memory trends using differencing. To find the correct degree of differencing, the augmented Dickey-Fuller test is often used. The null hypothesis of the augmented Dickey-Fuller test is the presence of a unit root , and the alternative hypothesis of the absence of a unit root.

In addition, you should also check visually that after differencing, the autocorrelation decays rapidly to 0 and that no strong negative correlation has appeared at lag 1 (e.g. below -0.5). Box and Jenkins recommend to look at differences of degree 0, 1 or 2 and correlations of lags below 20. (Note that differencing $d = 0$ is simply returning the original signal.)

- For the signal provided in the notebook, and for degree 0, 1 and 2, display the correlogram and the p-value of the augmented Dickey-Fuller test. Conclude.

Answer 3

The differencing equations are expressed as follows:

- If $d = 0 : y_t = Y_t$
- If $d = 1 : y_t = Y_t - Y_{t-1}$
- If $d = 2 : y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

From the following plots of the autocorrelation and partial autocorrelation of the signal when $d = 0$, we can notice that the first 20 autocorrelations are very high, which means that the signal contains a specific trend. In addition, the P-value of the the augmented Dickey-Fuller test is equal to 0.58. This means that we cannot reject the null hypothesis (H_0 : unit root is present in an autoregressive time series model). Thus, we can conclude that this signal is non-stationary.

For degrees 1 and 2, we can notice that the autocorrelation plot shows a sinusoidal shape. Moreover, the corresponding P-values of the augmented Dickey-Fuller test are very small (lower than 10^{-10}). Thus, we can reject the null hypothesis (H_0). Therefore, this signal is a stationary signal. We can say that, by using the differencing method, we are able to remove the trend from the signal and thus construct a stationary one. From the partial autocorrelation plot, we can notice that the order of the signal is 4 in the case of $d = 1$ and 11 in the case of $d = 2$. This result indicates that the use of the differencing of degree 1 is better in this case to eliminate the trend and obtain a stationary signal.

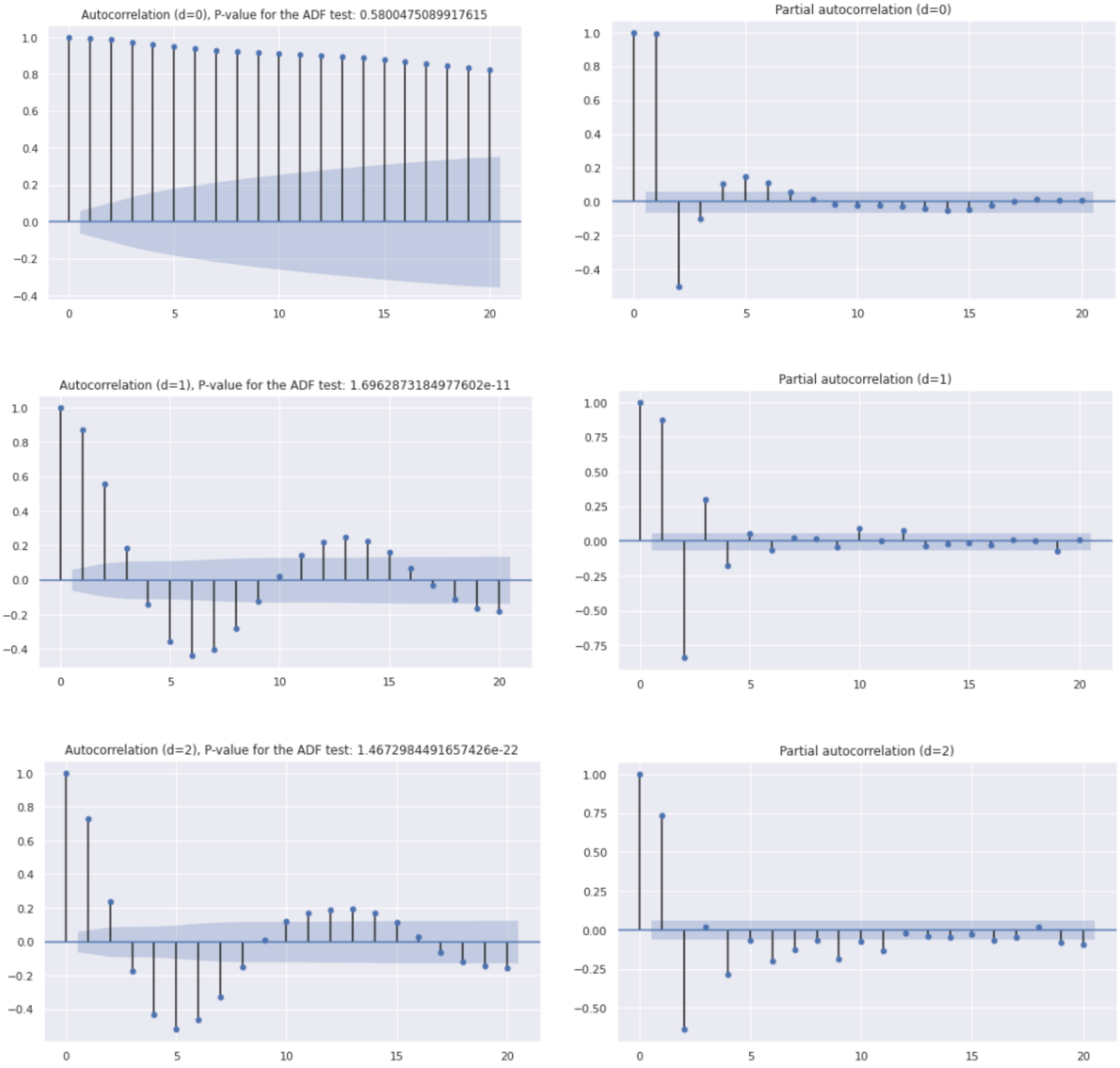


Figure 3: Correlograms of the differenced signals

Question 4 *Over-differencing*

Box and Jenkins warn about over-differencing, because it introduces unwanted correlations between samples. The following example illustrates this observation. Consider the process $Y_t = Y_{t-1} + \varepsilon_t$ where ε_t is a Gaussian white noise and let Δ denote the differencing operator.

- Is Y stationary?
- By looking at ΔY, show that Y is ARIMA(p, d, q) (specify the p, d and q).
- By looking at $\Delta^2 Y$, show that Y is ARIMA(p, d, q) (specify the p, d and q).
- Which of the two previous model is simpler?

Answer 4

- The associated characteristic polynomial for the process $Y_t = Y_{t-1} + \varepsilon_t$ is $1 - x$. This characteristic polynomial has a unit root. Thus, Y is non-stationary.

  We also performed for this task the augmented Dickey-Fuller test on a simulated process in the attached notebook. We obtained a P-value of the order of 0.6 which is greater than the significance level of 0.05. This implies that the null Hypothesis of the presence of a unit root cannot be rejected and therefore Y is considered as non-stationary.

- The ARIMA(p,d,q) process can be written as follows :

$$x^{(d)}[n] = -\sum_{i=1}^{p} a_i x^{(d)}[n-i] + b[n] + \sum_{j=1}^{q} m_j b[n-j]$$

- The first difference of the series is the difference between Y and itself lagged by one period:
  $\Delta Y = Y_t - Y_{t-1} = \varepsilon_t \implies Y$ is an ARIMA model with: $\begin{cases} p = & 0 \\ d = & 1 \\ q = & 0 \end{cases} \implies Y :$ ARIMA(0,1,0).

- The second difference of the series is the first difference of the first difference, i.e. the change-in-the-change of Y at period t :

  $\Delta^2 Y = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = \varepsilon_t - \varepsilon_{t-1} \implies Y$ is an ARIMA model with: $\begin{cases} p = & 0 \\ d = & 2 \\ q = & 1 \end{cases} \implies Y :$ ARIMA(0,2,1).

- Since the Y process is not stationary, the first ARIMA(0,1,0) model is the simplest. Indeed, according to this model, the time series is purely predicted as a stochastic model with a time dependency based only on the previous time point $t-1$. Thus, it is quite easy to understand and to compute. This model is interpreted as a "random walk" model. However, the second model ARIMA(0,2,1) requires a larger number of observations, not only of the previous time point $t-1$ but also of the one before, the time point $t-2$. It is therefore more complex.

Question 5 *Model diagnotic*

The last step of the Box-Jenkins methodology consists in checking if the residuals are uncorrelated. Denote by $\hat{\rho}_n$ the sample autocorrelation of lag k with n samples. For a i.i.d. process $\{Y_t\}_t$, the sample correlation vector converges to a standard multivariate Gaussian variable

$$[\hat{\rho}_n(1), \ldots, \hat{\rho}_n(k_{\max})]' / \sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, Id) \tag{2}$$

for a given maximum lag $k_{\max}$. A naive procedure to test $H_0 : \gamma_n(k) = 0$ for all $k = 1, \ldots, k_{\max}$ vs the alternative $H_1 : \gamma_n(k) = 0$ for at least one lag k is to check if $\gamma_n(k)/\sqrt{n}$ is within the interval $[-1.96, 1.96]$ (at level 5%). However, this procedure suffers from the multiple testing issue (see Question 5).

Simulate a Gaussian white noise ($n = 500$) and compute the $k_{\max} = 20$ first sample autocorrelations. Implement the naive procedure to test if the residual are uncorrelated. Repeat the experiment 500 times and report the proportion of rejected null hypotheses at level 5%. What do you observe?

Answer 5

In this question, we started by simulating a white Gaussian noise with $(n = 500)$. After that, we implemented the naive procedure to test whether the residuals are uncorrelated or not and we repeated this experiment 500 times. We found that **the proportion of rejected null hypotheses is equal to 0.61**.

A white noise process is a sequence of independent and identically distributed random variables (i.i.d). That is, the Gaussian white noise hypothesis imposes that the residuals are uncorrelated. However, when performing the naive procedure testing, we found a very high rejection rate that contradicts the assumption. Thus, we can say that this test is weak.

The following plot corresponds to an example of the first 20 autocorrelations of a white Gaussian noise. We notice that there is only one value not in the interval: $[-\frac{1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}}]$ (at the 5% level). In this case, we reject the hypothesis (H_0 : the correlations in the population from which the sample is taken are zero, and therefore any correlation observed in the data is the result of the randomness of the sampling process) and we accept the hypothesis (H_1) despite the fact that the signal is white Gaussian noise, but this value which is outside but not so far from the interval leads to reject the hypothesis that the residuals are independently distributed.

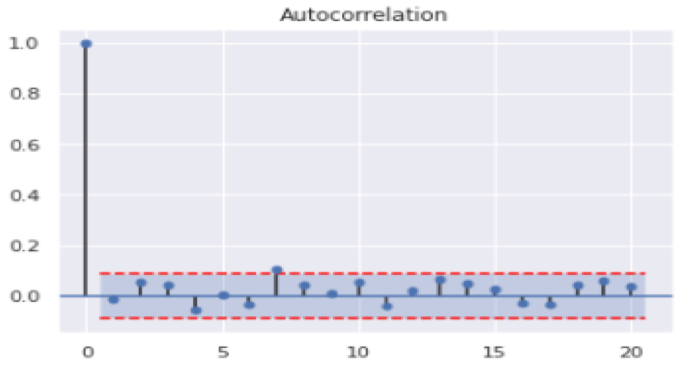


Figure 4: Autocorrelation plot of Gaussian white noise

Question 6 *Model diagnotic (continued)*

The Ljung-Box test is a better alternative. It relies on the statistic

$$n(n+2)\sum_{k=1}^{k_{\max}} \hat{\rho}_T(k)^2/(n-k) \tag{3}$$

which follows a χ^2 distribution with $k_{\max}$ degrees of freedom under the null.

Simulate a Gaussian white noise ($n = 500$) and compute the $k_{\max} = 20$ first sample autocorrelations. Implement the Ljung-Box procedure to test if the residuals are uncorrelated. Repeat the experiment 500 times and report the proportion of rejected null hypotheses at level 5%. Is this proportion in accordance with theory?

Answer 6

The **Ljung-Box test** uses the following hypotheses:

- H_0: The residuals are independently distributed.
- H_1: The residuals are not independently distributed, they exhibit serial correlation.

In this question, we proceed exactly as in the previous one, the only difference is the testing method. Here, we use the chi-square distribution with k_{max} degrees of freedom to calculate the P-value. We reject the null hypothesis in the case when the P-value of the test is less than the significance level 0.05.

We found that the rejection ratio of (H_0) among 500 experiments is equal to 0.056. In this case, this ratio is very low which implies that the residuals are independently distributed. This result is in concordance with the white Gaussian noise assumption.

4 Sparse coding

The modulated discrete cosine transform (MDCT) is a signal transformation often used in sound processing applications (for instance to encode a MP3 file). A MDCT atom $\phi_{L,k}$ is defined for a length 2L and a frequency localisation k ($k = 0, \dots, L-1$) by

$$\forall u = 0, \dots, 2L-1, \quad \phi_{L,k}[u] = w_L[u]\sqrt{\frac{2}{L}}\cos[\frac{\pi}{L}\left(u + \frac{L+1}{2}\right)(k + \frac{1}{2})] \tag{4}$$

where w_L is a modulating window given by

$$w_L[u] = \sin\left[\frac{\pi}{2L}\left(u + \frac{1}{2}\right)\right]. \tag{5}$$

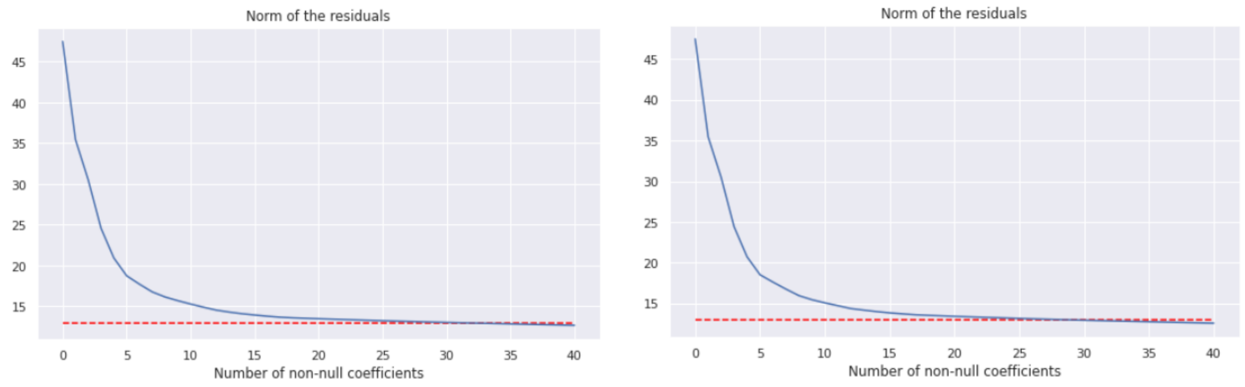
Question 7

For the signal provided in the notebook, learn a sparse representation with MDCT atoms. The dictionary is defined as the concatenation of all shifted MDCDT atoms for scales L in $[32, 64, 128, 256, 512, 1024]$.

- For the sparse coding, implement two different but related algorithms: the Matching Pursuit (MP) and the Orthogonal Matching Pursuit (OMP).
- Display on the same graph the norm of the successive residuals for both algorithms. Does one converge faster than the other?
- For both algorithms, what is the lowest number of atoms needed to have a residual whose norm is below a threshold, say 13? Display the associated reconstructions.

Answer 7

- We display the norm of the successive residuals for the two algorithms separately in Figure 5 and on the same graph in Figure 6. We observe that for both algorithms, the error decreases monotonically and the two curves overlap almost perfectly. However, by zooming in, we can notice that the OMP algorithm converges slightly faster.
- The MP algorithm needs 31 atoms to have a residual whose norm is lower than a threshold of 13, while the OMP algorithm needs 29 atoms. This also highlights the fact that the OMP algorithm converges slightly faster than the MP algorithm.
- We display the associated reconstructions in Figure 7. We can clearly observe that both algorithms manage to reconstruct the signal well. However, we can say that the OMP gives better performance since its signal-to-noise ratio is equal to 10.54 dB which is slightly lower than 10.64 dB, the signal-to-noise ratio of the MP algorithm.



MP OMP

Figure 5: Norms of the successive residuals for MP and OMP

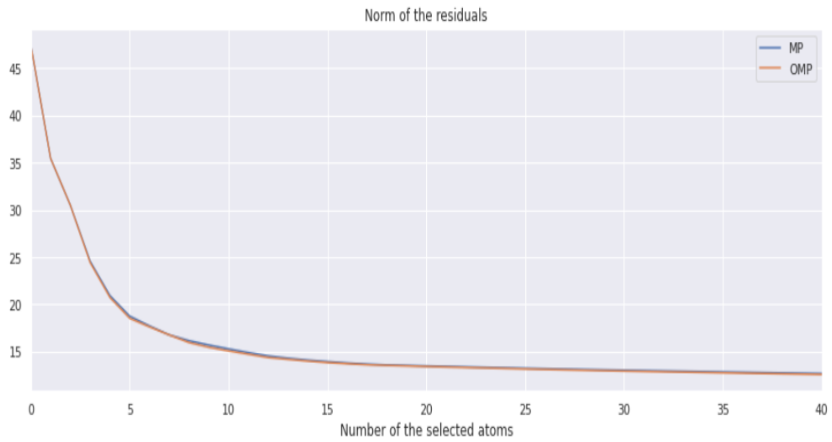
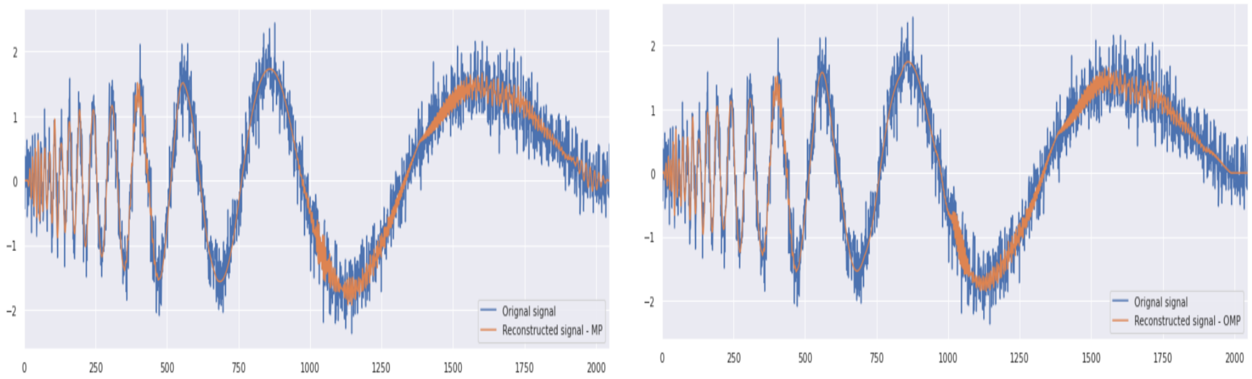


Figure 6: Norms of the successive residuals for MP and OMP displayed on the same graph



MP OMP

Figure 7: Chosen reconstruction for MP and OMP