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

Paul-Emile Zafar paul-emile.zafar@telecom-paris.fr
Marine Mercier marine.mer98@gmail.com

February 21, 2022

1 Introduction

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

Warning and advice.

- Use code from the tutorials as well as from other sources. Do not code yourself well-known procedures (e.g. cross validation or k-means), use an existing implementation.
- The associated notebook contains some hints and several helper functions.
- Be concise. Answers are not expected to be longer than a few sentences (omitting calculations).

Instructions.

- Fill in your names and emails at the top of the document.
- Hand in your report (one per pair of students) by Monday 21st February 11:59 PM.
- Rename your report and notebook as follows:
  `FirstnameLastname1_FirstnameLastname1.pdf` and
  `FirstnameLastname2_FirstnameLastname2.ipynb`.
  For instance, `LaurentOudre_CharlesTruong.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: .

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

The central limit theorem states that for i.i.d. random variables and finite variance, the sample mean converges with a rate of $\frac{1}{\sqrt{n}}$.

Let's rewrite $\mathbb{E}[(\bar{Y}_n - \mu)^2]$:

$$\mathbb{E}[(\bar{Y}_n - \mu)^2] = \mathbb{E}[(\frac{1}{n}\sum_{j=1}^{n} Y_i - \mu)^2]$$

$$\mathbb{E}[(\bar{Y}_n - \mu)^2] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j}\mathbb{E}[(Y_i - \mu)(Y_j - \mu)]$$

For wide-sense stationary process, we have that: $\gamma(|i-j|) = \mathbb{E}\left[(Y_i - \mu)\left(Y_j - \mu\right)\right]$

Thus:

$$\mathbb{E}[(\bar{Y}_n - \mu)^2] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(|i-j|)$$

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

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

Thus we get convergence of $\mathbb{E}[(\bar{Y}_n - \mu)^2]$ towards 0, which implies the convergence of $(\bar{Y}_n - \mu)^2$ in probability, and the rate of convergence is still in $\frac{1}{\sqrt{n}}$.

3 ARIMA process

Question 2 *Characteristic polynomial*

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

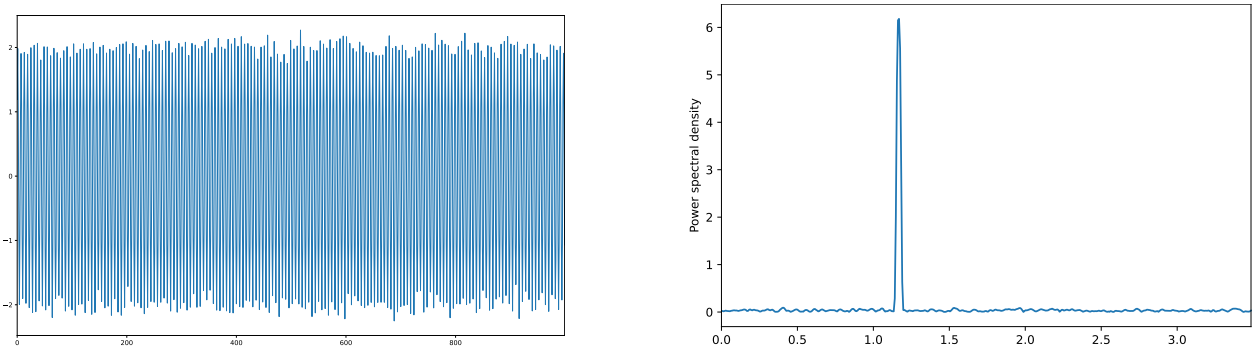
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \tag{1}$$

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?

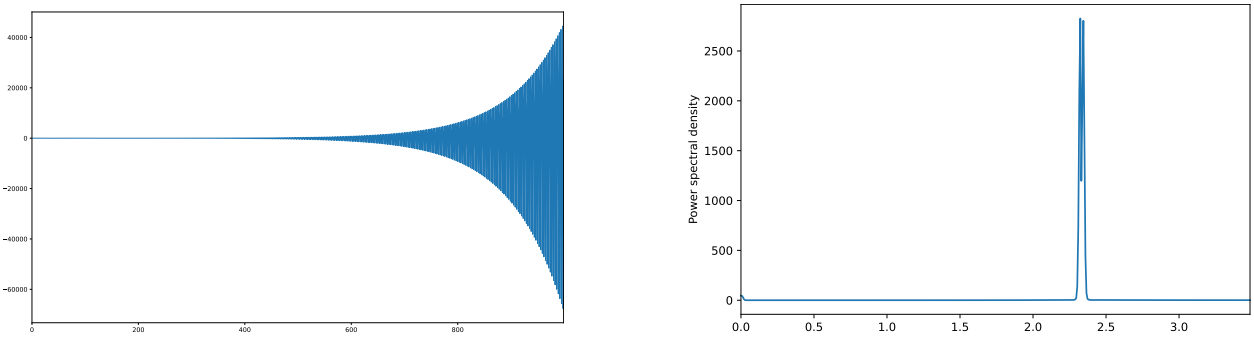
Answer 2

Remark: only the positive frequencies part of the (symmetrical) periodogram have been displayed below).



Signal Periodogram

Figure 1: First AR(2) process



Signal Periodogram

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.

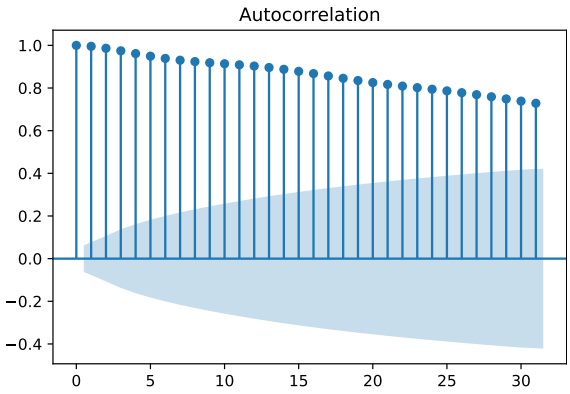
In addition, you should also check visually that after differencing, the autocorrelation decays rapidly to 0 and that no strong negative correlation has appeared at lag 1 (e.g. below -0.5). Box and Jenkins recommend to look at differences of degree 0, 1 or 2 and correlations of lags below 20. (Note that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

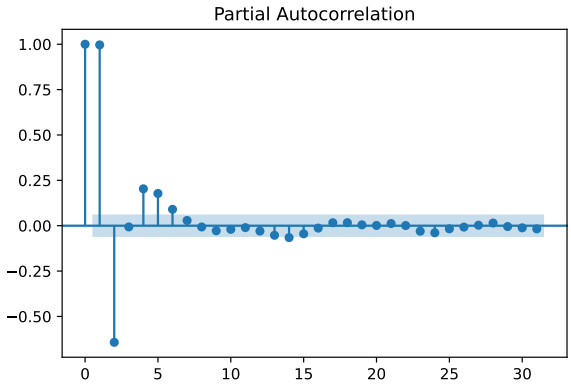
Answer 3

By observing the corelograms, we notice that with no differentiating ($d = 0$), the correlations coefficients remain very important, thus the signal is not stationary, and his is confirmed by the important p-value (≈ 0.60), which does not allow to reject the null hypothesis.

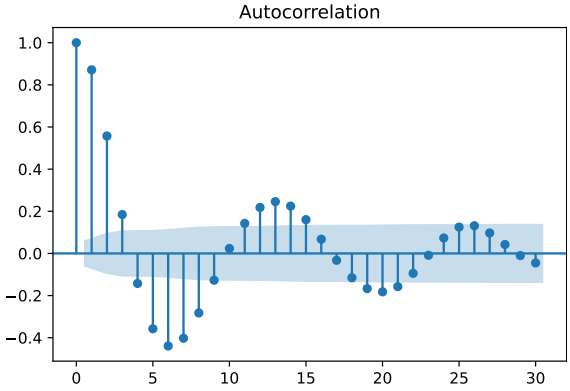
On the other hand, for $d = 1$ and $d = 2$, correlation coefficients drop quickly, and the associated p-values are of the order of 10^{-11} and 10^{-22}, the Dickey-Fuller test thus confirms stationarity.



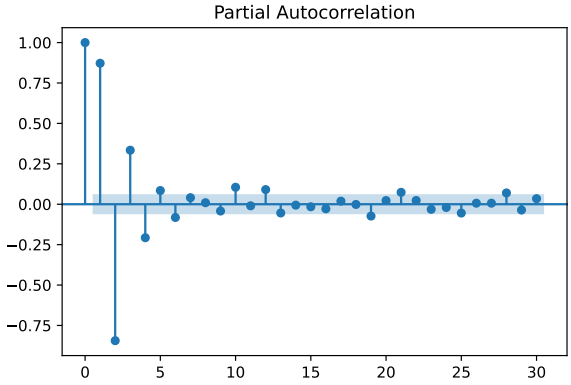
Autocorrelation ($d = 0$), P-value for the ADF test: 0.580



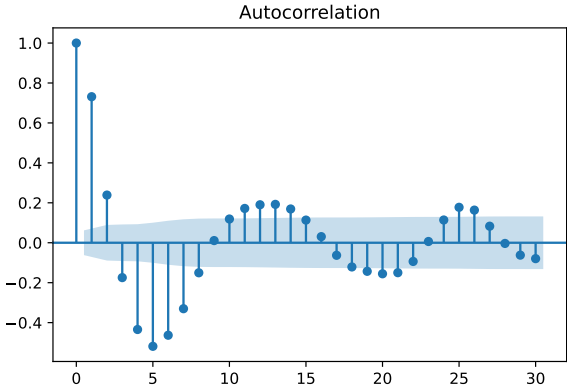
Partial autocorrelation ($d = 0$)



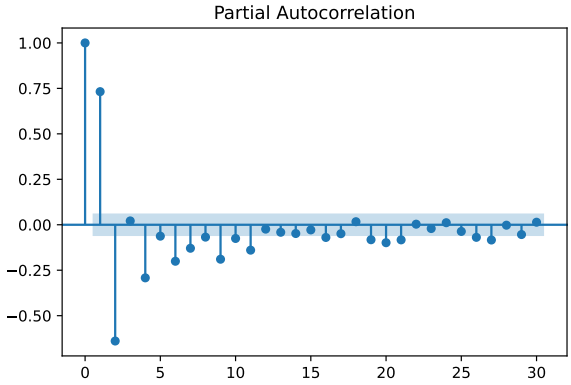
Autocorrelation ($d = 1$), P-value for the ADF test: 1.70×10^{-11}



Partial autocorrelation ($d = 1$)



Autocorrelation ($d = 2$), P-value for the ADF test: 1.46×10^{-22}



Partial autocorrelation ($d = 2$)

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

Y_t is first order stationary : $\forall t > 0, \mathbb{E}[Y_t] = \mathbb{E}[Y_{t-1}] + \mathbb{E}[\varepsilon_t] = \mathbb{E}[Y_{t_1}]$ since Y_{t-1} and ε_t are independent, and by recurrence $\mathbb{E}[Y_t] = \mathbb{E}[Y_0] = Y_0$.

But Y_t is not second order stationary : $\forall t_1 > t_2 > 0, Y_{t_1} = Y_{t_2} + \sum_{k=t_2+1}^{t_1} \varepsilon_k$ then$\mathbb{E}[(Y_{t_1} - Y_0)(Y_{t_2} - Y0)] = \mathbb{E}[(Y_{t_2} - t0)^2] + \mathbb{E}[\sum_{k=t_2+1}^{t_1} \varepsilon_k Y_{t_2} = \mathbb{E}[(Y_{t_2} - t0)^2] = \mathbb{V}(Y_{t_2})$ because Y_{t_2} and ε_k are independant for $k > t_2$. In moreover $\mathbb{V}(Y_{t_2}) = \mathbb{V}(\sum_{k=1}^{t_2} \varepsilon_k) = t_2\sigma_\varepsilon^2$ depends of t_2.

Therefore Y_t is not a stationary process. But $\Delta Y_t = Y_t - Y_{t-1} = \varepsilon_t$ is a stationary process.

The autoregressive integrated moving average model ARIMA(p,d,q) has formula $x^{(d)}[n] = \sum_{i=1}^{p} a_i x^{(d)}[n-1] + b[n] + \sum_{j=1}^{d} m_j b[n-j]$.

$\Delta Y_t = Y_t - Y_{t-1} = Y_t^{(1)} = \varepsilon_t$, therefore Y is ARIMA(0,1,0).

$\Delta^2 Y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = \varepsilon_t - \varepsilon_{t-1} = Y_t^{(1)} - Y_{t-1}^{(1)}$, then $Y_t^{(1)} = Y_{t-1}^{(1)} + \varepsilon_t - \varepsilon_{t-1}$, Y is also ARIMA(1,1,1).

The model ARIMA(0,1,0) is simpler.

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

The proportion of rejected null hypotheses is 61%. Having such a high rate of rejected null hypothesis (supposed to be at 5%) on white iid noise demonstrates that this method in not adequate.

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 proportion of rejected null hypotheses is around 6%. This is completely in accordance with theory, because we are testing the proportion of rejected null hypothesis at 5%, and shows that this is a far better suited test.

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

The orthogonal matching pursuit converges better and faster in terms of number of iterations, but the complexity of the algorithm is increased. Therefore OMP is a bit slower, 54s s for 60 iterations for OMP, while MP took 50s, but this difference of complexity is not a problem.

For MP, the norm of the residual converges to 13.08. The norm of the residual was always above 13, we changed the comparison threshold between the two models to 14. With MP, we need 21 iterations or number of atoms to reach this threshold. With OMP, we need 15 number of atoms to reach the threshold of 14. With 43 number of atoms, the norm of the residual is even under 13. The performance of OMP is indeed better than MP, and the additional complexity is not significant.

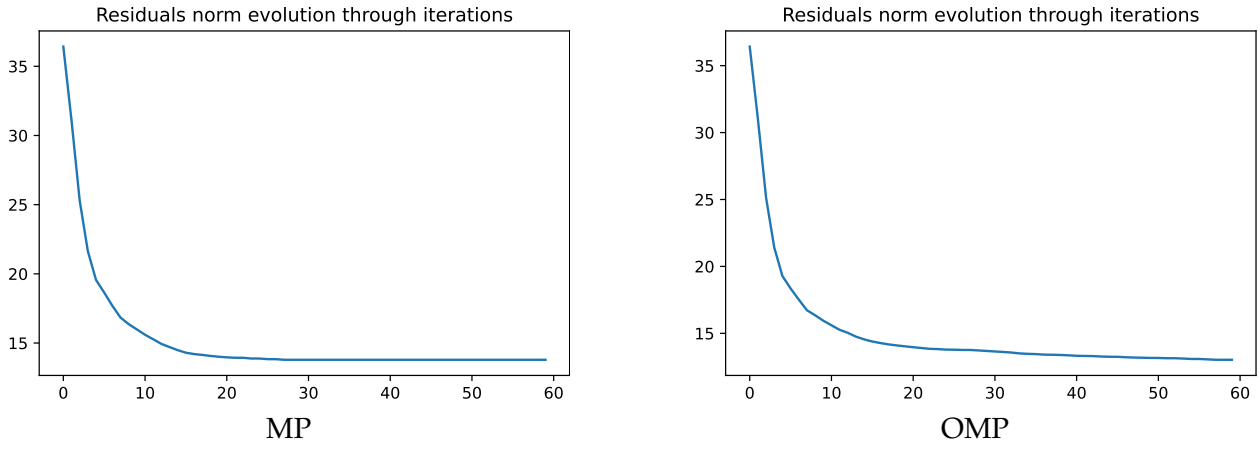


Figure 4: Norms of the successive residuals for MP and OMP

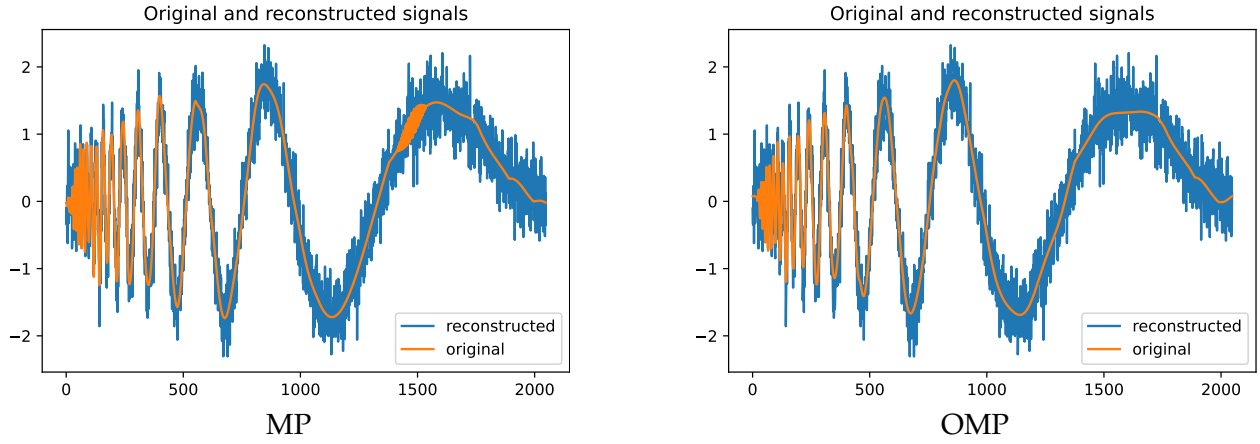


Figure 5: Chosen reconstruction for MP and OMP