Assignment 2 (ML for TS) - MVA 2021/2022
refrence : http://www.laurentoudre.fr/ast.html

February19, 2022

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

1 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

Thanks to the central-limit theorem, one can write for $X_i \sim X$ i.d.d. of finite variance:

$$\sqrt{n}\frac{\frac{X_1+\cdots+X_n}{n} - \mathbb{E}(X)}{\sigma(X)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

Hence $\bar{X_n} := \frac{X_1+\cdots+X_n}{n}$ converges (in distribution) towards $\mathbb{E}(X)$ at $\frac{\sigma(X)}{\sqrt{n}} = O(\frac{1}{\sqrt{n}})$ rate.

For a wide-sense stationary process, one can write:

$$
\begin{aligned}
\mathbb{E}[(\overline{Y}_n - \mu)^2] &= \mathbb{E}[(\frac{\sum_{i=1}^{n}(Y_i - \mu)}{n})^2] \\
&= \frac{1}{n^2}\sum_{i,j}\mathbb{E}[(Y_i - \mu)(Y_j - \mu)] \\
&= \frac{1}{n^2}\sum_{i,j}\gamma(|i-j|) \\
&= \frac{2}{n^2}\sum_{k=0}^{n-1}\sum_{j-i=k}\gamma(|i-j|) \\
&= \frac{2}{n^2}\sum_{k=0}^{n-1}\gamma(k)\sum_{j-i=k}1 \\
&= \frac{2}{n^2}\sum_{k=0}^{n-1}(n-k)\gamma(k) \\
&\leq \frac{2}{n}\sum_{k=0}^{n-1}\gamma(k) \\
&\underset{n\to+\infty}{\to} 0
\end{aligned}
$$

Since $\sum_{k\geq 0}\gamma(k)$ converges absolutely. Hence:

$$\overline{Y}_n \overset{L_2}{\to} \mu$$

Using Bienaymé-Tchebychev inequality, one can write for every $\epsilon > 0$:

$$\mathbb{P}(|\bar{Y_n} - \mu| > \epsilon) \leq \frac{\mathbb{E}((\bar{Y_n} - \mu)^2)}{\epsilon^2} = \frac{2}{n\epsilon^2}\sum_{k=0}^{n-1}\gamma(k) \leq \frac{2}{n\epsilon^2}\sum_{k\geq 0}\gamma(k) \underset{n\to+\infty}{\to} 0$$

Hence, on the one hand:

$$\overline{Y}_n \overset{\mathbb{P}}{\to} \mu$$

And on the other hand we have:

$$|\bar{Y_n} - \mu| \leq \frac{1}{\sqrt{n}}\sqrt{\frac{2\sum_{k\geq 0}\gamma(k)}{\alpha}}$$

With at least $1-\alpha$ probability, meaning that $\bar{Y_n} - \mu = O(\frac{1}{\sqrt{n}})$ with at least $1-\alpha$ probability, which is the same rate of convergence of the i.i.d. case.

2 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

First case: ϕ_1 and ϕ_2 such that the characteristic polynomial has a complex root of norm 1.

An easy way to enforce ϕ_1 and ϕ_2 in such a configuration is by setting:

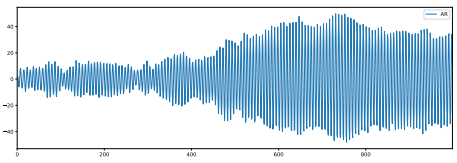
$$1 - \phi_1 x - \phi_2 x^2 = (x - e^{i\frac{2\pi}{T}})(x - e^{-i\frac{2\pi}{T}}) = x^2 - 2cos(\frac{2\pi}{T})x + 1$$

Hence: $\phi_1 = 2cos(\frac{2\pi}{T})$ and $\phi_2 = -1$.

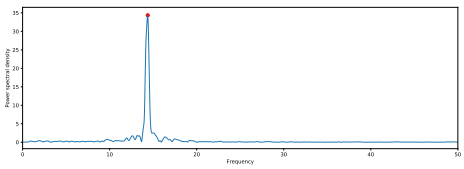
By setting $T = 7$ for example, we get a main period that is equal to:

$$T_{\text{main}} = 6.96 \simeq 7 = T$$

And a the following signal and periodogram shapes:



Signal



Periodogram

Figure 1: First AR(2) process

We observe, then, a periodogram that has almost a pure spike on the chosen period $T \simeq T_{\text{main}}$, whereas a signal that is quasi-periodic and bounded.

Second case: ϕ_1 and ϕ_2 such that the characteristic polynomial has two complex conjugate roots of norm $r = 0.99$ and phase $\theta = 2\pi/3$

Same as before, it is possible to enforce ϕ_1 and ϕ_2 in such a configuration by setting:

$$\begin{aligned} 1 - \phi_1 x - \phi_2 x^2 &= \frac{1}{0.99^2}(x - 0.99e^{i\frac{2\pi}{3}})(x - 0.99e^{-i\frac{2\pi}{3}}) \\ &= \frac{1}{0.99^2}(x^2 - 2 \times 0.99 \times cos(\frac{2\pi}{3})x + 0.99^2) \\ &= \frac{1}{0.99^2}x^2 - \frac{2}{0.99}cos(\frac{2\pi}{3})x + 1 \end{aligned}$$

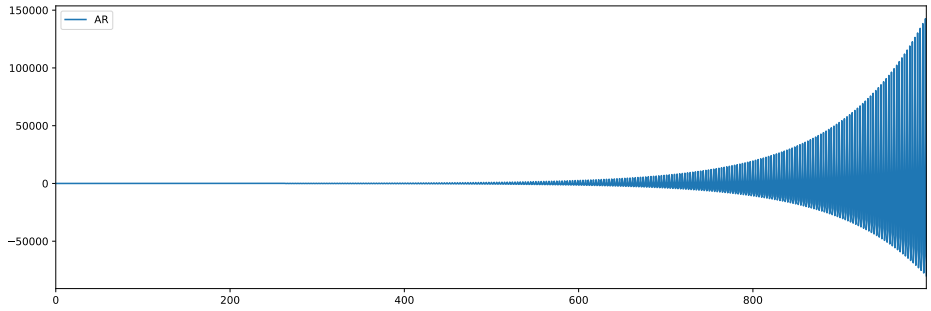
Hence: $\phi_1 = \frac{2}{0.99}cos(\frac{2\pi}{3})$ and $\phi_2 = \frac{-1}{0.99^2}$.

Doing so, we get a two main periods that are equal to:

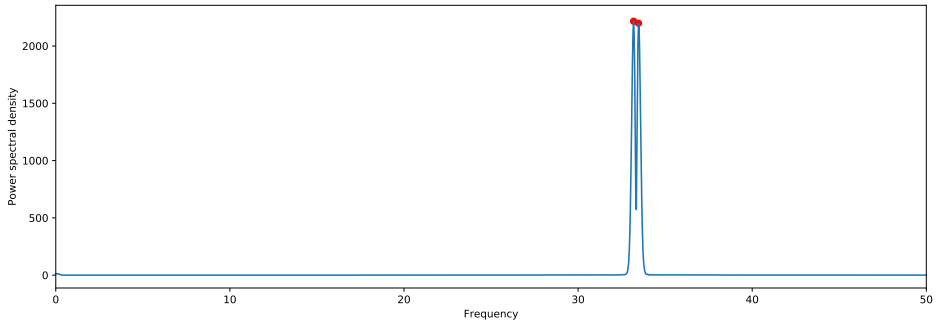
$$T^1_{\text{main}} = 2.99 \simeq 3$$

$$T^2_{\text{main}} = 3.01 \simeq 3$$

And a the following signal and periodogram shapes:



Signal



Periodogram

Figure 2: Second AR(2) process

We observe, then, a periodogram that has two tightly separated spikes, and a signal that is aperiodic and unbounded.

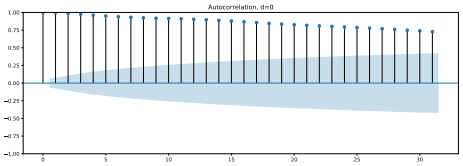
Question 3 *Removing the trend by differencing*

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.

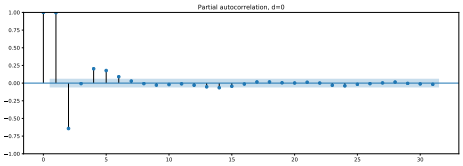
In addition, you should also check visually that after differencing, the autocorrelation decays rapidly to 0 and that no strong negative correlation has appeared at lag 1 (e.g. below -0.5). Box and Jenkins recommend to look at differences of degree 0, 1 or 2 and correlations of lags below 20. (Note that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

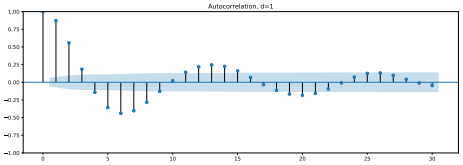
Answer 3



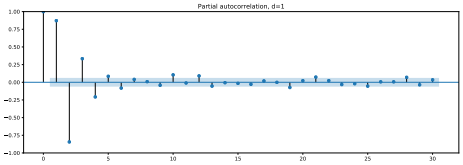
Autocorrelation ($d = 0$), P-value for the ADF test: 0.580048



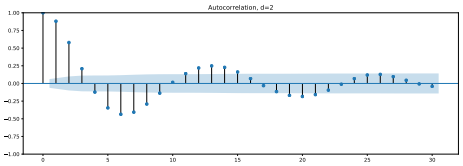
Partial autocorrelation ($d = 0$)



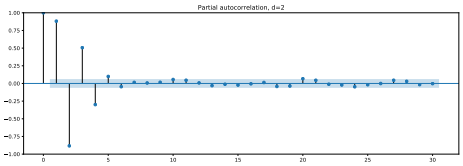
Autocorrelation ($d = 1$), P-value for the ADF test: 0.000000



Partial autocorrelation ($d = 1$)



Autocorrelation ($d = 2$), P-value for the ADF test: 0.000000



Partial autocorrelation ($d = 2$)

Figure 3: Correlograms of the differenced signals

Regarding the figures above and Box and Jenkins recommendations, the most probable value for d is 1. If the serie was to be modelled using an $AR(p)$ model, a good value for p would be 4.

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 isn't stationary since:

$$\mathbb{E}(Y_t^2) = \mathbb{E}(Y_{t-1}^2 + 2\varepsilon_t Y_{t-1} + \varepsilon_t^2) = \mathbb{E}(Y_{t-1}^2) + 1 = \cdots = t$$

  Implying that $\mathbb{E}(Y_t^2)$ depends on t.

- Regarding ΔY, Y is ARIMA(p, d, q) since:

$$\Delta Y_t = Y_t - Y_{t-1} = \varepsilon_t \Rightarrow Y \sim \text{ARIMA}(0, 1, 0)$$

  Hence:

$$p = 0, d = 1, q = 0$$

- Regarding $\Delta^2 Y$, Y is ARIMA(p, d, q) since:

$$\Delta^2 Y_t = \Delta \varepsilon_t \Rightarrow Y \sim \text{ARIMA}(0, 2, 1)$$

  Hence:

$$p = 0, d = 2, q = 1$$

- The simplicity of an $ARIMA(p, d, q)$ can be assessed by evaluating $p + d + q$. Using such a rule, $ARIMA(0, 1, 0)$ is simpler than $ARIMA(0, 2, 1)$, therefore over-differencing leads to unnecessary complexified models.

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

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

for a given maximum lag $k_{\max}$. A naive procedure to test $H_0 : \gamma_n(k) = 0$ for all $k = 1, \dots, k_{\max}$ 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 63.8%, which is much higher than the chosen confidence level for the naive test. This happens because the naive test is a point-wise one, meaning that the confidence level holds for each k individually:

$$\mathbb{P}(\sqrt{n-k}\gamma_n(k) \notin [-1.96, 1.96] | \mathcal{H}_0) = 5\%$$

Yet, we don't have necessarily:

$$\mathbb{P}(\bigcup_{k=1}^{kmax} (\sqrt{n-k}\gamma_n(k) \notin [-1.96, 1.96]) | \mathcal{H}_0) \leq 5\%$$

Besides, using the larger-numbers theorem, we have:

$$\frac{1}{Nmax} \sum_{i=1}^{Nexp} I_{(\cup_{k=1}^{kmax} (\sqrt{n-k}\gamma_n(k) \notin [-1.96,1.96]))} \underset{Nexp \to +\infty}{\to} \mathbb{P}(\bigcup_{k=1}^{kmax} (\sqrt{n-k}\gamma_n(k) \notin [-1.96, 1.96]) | \mathcal{H}_0) \gg 5\%$$

Where the LHS term is exactly the proportion of rejected null hypothesis.

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 hypothesis is 5.4%, which isn't, as desired, much higher than the chosen confidence level for the test. In fact, by increasing the number of launched experiences to 10000 for example, we get a proportion of rejected null hypothesis which is equal to 5.0%, exactly the chosen confidence level, in accordance with theory.

3 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 ($k = 0, \ldots, L-1$) by

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

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

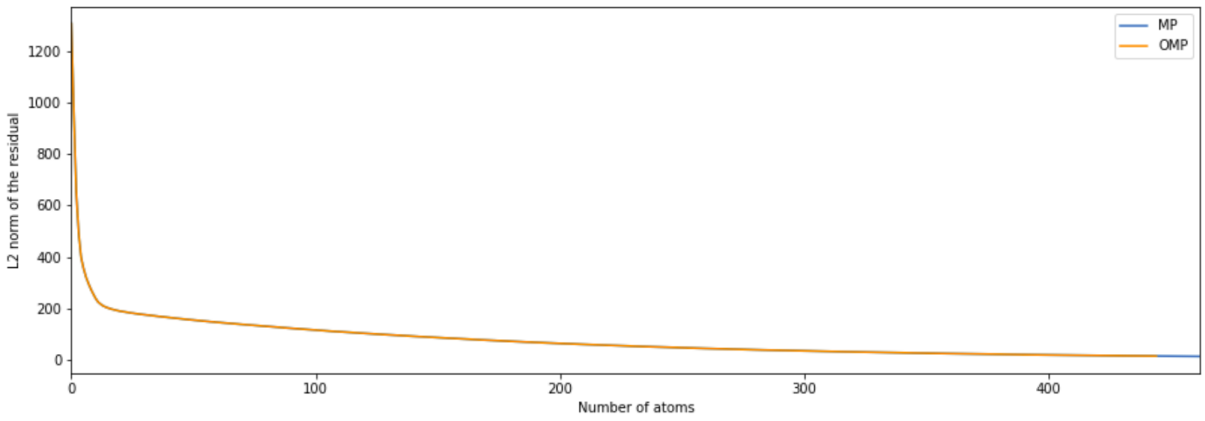
For the signal provided in the notebook, learn a sparse representation with MDCT atoms. The dictionary is defined as the concatenation of all shifted MDCDT atoms for scales L in $[32, 64, 128, 256, 512, 1024]$.

- 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

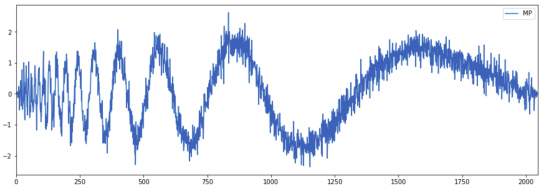
MP OMP

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

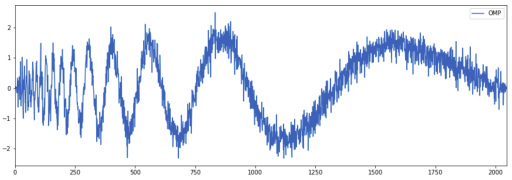


MP

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

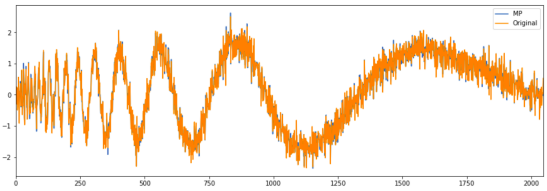


MP

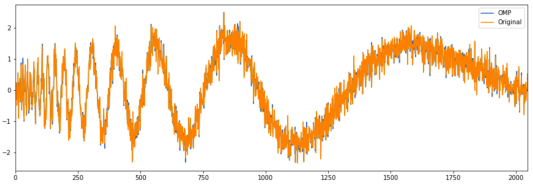


OMP

Figure 6: Chosen reconstruction for MP and OMP



MP/Original



OMP/Origina

Figure 7: Comparing MP and OMP to the original

We notice that *MP* and *OMP* provides almost the same results, however we would like to make few remarks:

- *OMP* converges faster than *MP*.

- For both algorithms, the lowest number of atoms needed to have a residual whose norm is below a threshold 13, is 400 atoms.

- The main difference between *MP* and *OMP* is the fact that *OMP* keeps the residual vector R orthogonal to all atoms that were previously selected during the previous iterations.