Assignment 1 (ML for TS) - MVA 2023/2024

Basile Terver terverbasile@gmail.com
Léa Khalil leakhalil@yahoo.fr

November 7, 2023

1 Introduction

Objective. This assignment has three parts: questions about the convolutional dictionary learning, the spectral features and a data study using the DTW.

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 Tuesday 7$^{\text{th}}$ November 23:59 PM.
- Rename your report and notebook as follows:
`FirstnameLastname1_FirstnameLastname2.pdf` and
`FirstnameLastname1_FirstnameLastname2.ipynb`.
For instance, `LaurentOudre_CharlesTruong.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: docs.google.com/forms/d/e/1FAIpQLSdTwJEyc6QIoYTknjk12kJMtcKllFvPlWLk5LbyugW0YO7K6Q/viewform?usp=sf_link.

2 Convolution dictionary learning

Question 1

Consider the following Lasso regression:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 \quad + \quad \lambda \|\beta\|_1 \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ the design matrix, $\beta \in \mathbb{R}^p$ the vector of regressors and $\lambda > 0$ the smoothing parameter.

Show that there exists $\lambda_{\max}$ such that the minimizer of (1) is $\mathbf{0}_p$ (p-dimensional vector of zeros) for any $\lambda > \lambda_{\max}$.

Answer 1

The KKT conditions yield that $\hat{\beta}_\lambda$ is a solution for the minimization problem (1) if and only if

$$-X^\top(y - X\hat{\beta}_\lambda) = \lambda \hat{s}_\lambda, \tag{2}$$

where $\hat{s}_{\lambda,j}$ is an unknown quantity equal to $sgn(\hat{\beta}_{\lambda,j})$ if $\hat{\beta}_{\lambda,j} \neq 0$ and some value lying in $[1,1]$ otherwise, in other words $\hat{s}_{\lambda,j}$ is a subgradient for the absolute value function.

Suppose $\hat{\beta}_\lambda = 0$ and inject it into (2), this yields $-X^\top y = \lambda \hat{s}_\lambda$, and in particular

$$\|X^\top y\|_\infty = \lambda \|\hat{s}_\lambda\|_\infty.$$

If $\|\hat{s}_\lambda\|_\infty \neq 1$, then λ could decrease (with $\|\hat{s}_\lambda\|_\infty$ increased to maintain equality) and the Lasso estimate would still be $\hat{\beta}_\lambda = 0$. Therefore, since $\|\hat{s}_\lambda\|_\infty \leq 1$, if we denote $\lambda_{\max}$ the smallest value of λ such that $\hat{\beta}_\lambda = 0$, we get that

$$\|X^\top y\|_\infty = \lambda_{\max} \cdot 1. \tag{3}$$

As a conclusion, we get

$$\lambda_{\max} = \|X^\top y\|_\infty. \tag{4}$$

Question 2

For a univariate signal $\mathbf{x} \in \mathbb{R}^n$ with n samples, the convolutional dictionary learning task amounts to solving the following optimization problem:

$$\min_{(\mathbf{d}_k)_k, (\mathbf{z}_k)_k \|\mathbf{d}_k\|_2^2 \leq 1} \quad \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 \quad + \quad \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \tag{5}$$

where $\mathbf{d}_k \in \mathbb{R}^L$ are the K dictionary atoms (patterns), $\mathbf{z}_k \in \mathbb{R}^{N-L+1}$ are activations signals, and $\lambda > 0$ is the smoothing parameter.

Show that

- for a fixed dictionary, the sparse coding problem is a lasso regression (explicit the response vector and the design matrix);
- for a fixed dictionary, there exists $\lambda_{\max}$ (which depends on the dictionary) such that the sparse codes are only 0 for any $\lambda > \lambda_{\max}$.

Answer 2

Let us suppose $L << N$, in particular $L < N - L + 1$.

The above problem with fixed $(\mathbf{d}_k)_k$ is a Lasso regression with response vector $\mathbf{x}$. Let us denote $\mathbf{Z} := (\mathbf{z}_1^\top, \ldots, \mathbf{z}_K^\top)^\top \in \mathbb{R}^{K(N-L+1)}$. The dimension of the parameter to learn is $p := K(N - L + 1)$.

Each activation signal $\mathbf{z}_k$ associated to a given atom is a sum of dirac delta (more precisely linear combination of indicator) functions. By the sifting property of the Dirac delta function, we get, in

the framework of discrete convolution: $\mathbf{z}_k * \mathbf{d}_k[i] = \sum_{j=\max(1,i-N+L-1)}^{\min(i,L)} \mathbf{z}_k^{i+1-j}\mathbf{d}_k^j$ for $i \in \{1,\dots,N\}$. Indeed, for $i > N-L+1$, we want to consider valid values of $\mathbf{z}_k$, that is $1 \leq i-j \leq N-L+1$, that is $j \geq i-N+L-1$.

Let us vectorize as $\mathbf{Z} := (\mathbf{z}_1^\top, \dots, \mathbf{z}_K^\top)^\top \in \mathbb{R}^{K(N-L+1)}$ and

$$\mathbf{H} := (\mathbf{H}_1, \dots, \mathbf{H}_k) := (\mathbf{h}_1^1, \dots, \mathbf{h}_1^{N-L+1}, \mathbf{h}_2^1, \dots, \mathbf{h}_K^{N-L+1}) \in \mathbb{R}^{N \times K(N-L+1)} \tag{6}$$

where $\mathbf{h}_k^n$ corresponds to activation of atom k at time n. We want, for each $i \in \{1,\dots,N\}$, to have

$$\sum_{k=1}^K \mathbf{z}_k * \mathbf{d}_k[i] = \sum_{k=1}^K \sum_{j=\max(1,i-N+L-1)}^{\min(i,L)} \mathbf{z}_k^{i+1-j}\mathbf{d}_k^j = \sum_{k=1}^K \sum_{l=1}^{N-L+1} \mathbf{h}_k^l[i]\mathbf{z}_k^l$$

So we want $\sum_{l=1}^{N-L+1} \mathbf{z}_k^l \mathbf{h}_k^l[i] = \sum_{j=\max(1,i-N+L-1)}^{\min(i,L)} \mathbf{z}_k^{i+1-j}\mathbf{d}_k^j = \sum_{j=\max(1,i-L+1)}^{\min(i,N-L+1)} \mathbf{z}_k^j \mathbf{d}_k^{i+1-j}$.

Therefore, we must have $\mathbf{h}_k^l[i] = \mathbf{d}_k^{i+1-l}\mathbb{1}_{\max(1,i-L+1)\leq l \leq \min(i,N-L+1)}$, for each $i \in \{1,\dots,N\}, l \in \{1,\dots,N-L+1\}, k \in \{1,\dots,K\}$.

Since we supposed $L << N$, in the general regime we have $L \leq i \leq N-L+1$, thus, for most values of i, we have $\mathbf{h}_k^l[i] = \mathbf{d}_k^{i+1-l}\mathbb{1}_{\max(1,i-L+1)\leq l \leq \min(i,N-L+1)} = \mathbf{d}_k^{i+1-l}\mathbb{1}_{i-L+1\leq l \leq i}$.

In the regime $i < L$, we have $\mathbf{h}_k^l[i] = \mathbf{d}_k^{i+1-l}\mathbb{1}_{\max(1,i-L+1)\leq l \leq \min(i,N-L+1)} = \mathbf{d}_k^{i+1-l}\mathbb{1}_{1\leq l \leq i}$.

Eventually, in the regime $N \geq i > N-L+1 > L$, we have $\mathbf{h}_k^l[i] = \mathbf{d}_k^{i+1-l}\mathbb{1}_{\max(1,i-L+1)\leq l \leq \min(i,N-L+1)} = \mathbf{d}_k^{i+1-l}\mathbb{1}_{i-L+1\leq l \leq N-L+1}$.

Therefore we can write

$$\mathbf{H}_k := \begin{pmatrix}
\mathbf{d}_k^1 & 0 & \dots & & & \dots & \dots & \dots & 0 \\
\vdots & \ddots & & & & & & & \\
\mathbf{d}_k^L & \dots & \mathbf{d}_k^1 & & & & & & \\
\vdots & \ddots & & \ddots & & & & & \\
\vdots & & \ddots & & \ddots & & & & \\
0 & \dots & 0 & \mathbf{d}_k^L & \dots & \mathbf{d}_k^1 & 0 & \dots & 0 \\
\vdots & & & & \ddots & & \ddots & & \vdots \\
\vdots & & & & & \ddots & & \ddots & 0 \\
\vdots & & & & & & \ddots & & \mathbf{d}_k^1 \\
\vdots & & & & & & & \ddots & \vdots \\
0 & \dots & & & & & & 0 & \mathbf{d}_k^L
\end{pmatrix} \in \mathbb{R}^{N \times (N-L+1)}. \tag{7}$$

The Lasso regression is

$$\min_{\mathbf{Z} \in \mathbb{R}^{K(N-L+1)}} \|\mathbf{x} - \mathbf{H}\mathbf{Z}\|_2^2 + \lambda \|\mathbf{Z}\|_1. \tag{8}$$

Since, for a fixed dictionary, the convolutional sparse coding problem (8) is a Lasso regression, Question 1 gives that for $\lambda > \lambda_{\max}$, we have $\mathbf{Z} = 0$, with

$$\lambda_{\max} = \|\mathbf{H}^\top \mathbf{x}\|_\infty. \tag{9}$$

3 Spectral feature

Let X_n ($n = 0, \dots, N-1$) be a weakly stationary random process with zero mean and autocovariance function $\gamma(\tau) := \mathbb{E}(X_n X_{n+\tau})$. Assume the autocovariances are absolutely summable, i.e. $\sum_{\tau \in \mathbb{Z}} |\gamma(\tau)| < \infty$, and square summable, i.e. $\sum_{\tau \in \mathbb{Z}} \gamma^2(\tau) < \infty$. Denote by f_s the sampling frequency, meaning that the index n corresponds to the time instant n/f_s and for simplicity, let N be even.

The *power spectrum* S of the stationary random process X is defined as the Fourier transform of the autocovariance function:

$$S(f) := \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2\pi f \tau / f_s}. \tag{10}$$

The power spectrum describes the distribution of power in the frequency space. Intuitively, large values of $S(f)$ indicates that the signal contains a sine wave at the frequency f. There are many estimation procedures to determine this important quantity, which can then be used in a machine learning pipeline. In the following, we discuss about the large sample properties of simple estimation procedures, and the relationship between the power spectrum and the autocorrelation.

(Hint: use the many results on quadratic forms of Gaussian random variables to limit the amount of calculations.)

Question 3

In this question, let X_n ($n = 0, \dots, N-1$) be a Gaussian white noise.

- Calculate the associated autocovariance function and power spectrum. (By analogy with the light, this process is called "white" because of the particular form of its power spectrum.)

Answer 3

Let us calculate the autocovariance function of this Gaussian white noise. We know that the random variable has finite variance which we will denote $sigma^2$. In addition, by definition of a Gaussian white noise, we know that each time step in independent from the others.

$$\gamma(\tau) := \mathbb{E}(X_n X_{n+\tau}) = \sigma^2 \delta_0(\tau)$$

Let us now compute the power spectrum of this signal.

$$S(f) := \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2i\pi f \tau / f_s} = \gamma(0) = \sigma^2$$

This process is called white because, like for white light it presents a particular spectrum which is a constant of the frequency.

Question 4

A natural estimator for the autocorrelation function is the sample autocovariance

$$\hat{\gamma}(\tau) := (1/N) \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \tag{11}$$

for $\tau = 0, 1, \dots, N-1$ and $\hat{\gamma}(\tau) := \hat{\gamma}(-\tau)$ for $\tau = -(N-1), \dots, -1$.

- Show that $\hat{\gamma}(\tau)$ is a biased estimator of $\gamma(\tau)$ but asymptotically unbiased. What would be a simple way to de-bias this estimator?

Answer 4

Let us rewrite the estimator for the autocorrelation function as follows:

$$\hat{\gamma}(\tau) := \frac{1}{N}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} = \frac{N-\tau}{N}\frac{1}{N-\tau}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}$$

We know that the random variable $X_n X_n + \tau$ has finite expectation. $\mathbb{E}(\hat{\gamma}(\tau)) = \frac{N-\tau}{N}\mathbb{E}(X_n X_{n+\tau}) = \frac{N-\tau}{N}\gamma(\tau))$ $\lim_{N\to+\infty} \mathbb{E}(\hat{\gamma}(\tau)) = \gamma(\tau)$ This shows that the estimator is asymptotically unbiased.

If we want to work with an unbiased estimator of this quantity, we can instead use:

$$\hat{\gamma_u}(\tau) := \frac{1}{N-\tau}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}$$

Question 5

Define the discrete Fourier transform of the random process $\{X_n\}_n$ by

$$J(f) := (1/\sqrt{N})\sum_{n=0}^{N-1} X_n e^{-2\pi i f n/f_s} \tag{12}$$

The *periodogram* is the collection of values $|J(f_0)|^2, |J(f_1)|^2, \ldots, |J(f_{N/2})|^2$ where $f_k = f_s k/N$. (They can be efficiently computed using the Fast Fourier Transform.)

- Write $|J(f_k)|^2$ as a function of the sample autocovariances.
- For a frequency f, define $f^{(N)}$ the closest Fourier frequency f_k to f. Show that $|J(f^{(N)})|^2$ is an asymptotically unbiased estimator of $S(f)$ for $f > 0$.

Answer 5

$$\begin{aligned}
|J(f_k)|^2 &= \frac{1}{N}\sum_{n=0}^{N-1} X_n e^{-2\pi i k n/N}\sum_{m=0}^{N-1} X_m e^{2\pi i k m/N} \\
&= \frac{1}{N}\sum_{n=0}^{N-1} X_n^2 + \frac{1}{N}\sum_{n=0}^{N-1}\sum_{m=n+1}^{N-1} X_n X_m 2\cos(2k\pi\frac{m-n}{N}) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} X_n^2 + \frac{1}{N}\sum_{n=0}^{N-1}\sum_{t=1}^{N-1-n} X_n X_{n+t} 2\cos(2k\pi\frac{t}{N}) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} X_n^2 + \frac{1}{N}\sum_{t=1}^{N-1} 2\cos(2k\pi\frac{t}{N})\sum_{n=0}^{N-1-t} X_n X_{n+t} \\
&= \hat{\gamma}(0) + 2\sum_{t=1}^{N-1}\cos(2k\pi\frac{t}{N})\hat{\gamma}(t)
\end{aligned}$$

Let f be a frequency such that $f \leq f_{\frac{N}{2}}$. By definition, for a given length of the sequence N, there exists $k \leq \frac{N}{2}$ such that $f^{(N)} = f_k$ (f necessarily belongs to an interval of the form $[kf_s/N; (k+1)f_s/N]$ and this k is defined as the $argmin|f - f_s k/N|$. This means that $|f - f^{(N)}| \leq \frac{f_s}{N}$. (Indeed f belongs to an interval of length $\frac{f_s}{N}$). This proves that $\lim_{N\to+\infty} f^{(N)} = f$ In addition, we can rewrite $|J(f^{(N)})|^2$ as follows by using the parity of the sample autocorrelation function:

$$|J(f^{(N)})|^2 = \sum_{t=-N+1}^{N-1} e^{-2\pi i k t/N} \hat{\gamma}(t) = \sum_{t=-N+1}^{N-1} e^{-2\pi i f^{(N)}/f_s} \hat{\gamma}(t)$$

In addition, we have that:

$$\mathbb{E}(|J(f^{(N)})|^2) = \sum_{t=-N+1}^{N-1} e^{-2\pi i f^{(N)}/f_s} \frac{N-t}{N} \gamma(t)$$

For any $t \in \mathbb{R}$, $\lim_{N\to+\infty} e^{-2\pi i f^{(N)}/f_s} \frac{N-t}{N} \gamma(t) \mathbb{1}_{t\in[-N+1,N-1]} = e^{-2\pi i f/f_s} \gamma(t)$. In addition, we know that $|e^{-2\pi i f^{(N)}/f_s} \frac{N-t}{N} \gamma(t) \mathbb{1}_{t\in[-N+1,N-1]}| \leq |\gamma(t)|$ and that the autocovariance function is absolutely summable.
We can now apply the dominated convergence theorem: $\lim_{N\to+\infty} \sum_{t=-N+1}^{N-1} e^{-2\pi i f^{(N)}/f_s} \frac{N-t}{N} \gamma(t) = \sum_{t=-\infty}^{+\infty} lim_{N\to+\infty} e^{-2\pi i f^{(N)}/f_s} \frac{N-t}{N} \gamma(t) = \sum_{t=-\infty}^{+\infty} e^{-2\pi i f/f_s} \gamma(t) = S(f)$
We conclude that:

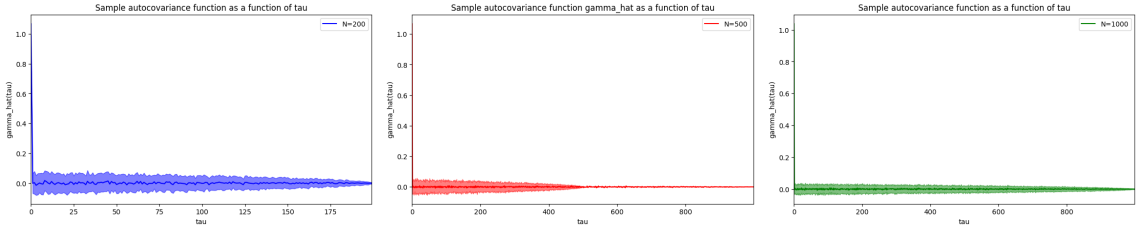
$$\lim_{N\to+\infty} |J(f^{(N)})|^2 = S(f)$$

Question 6

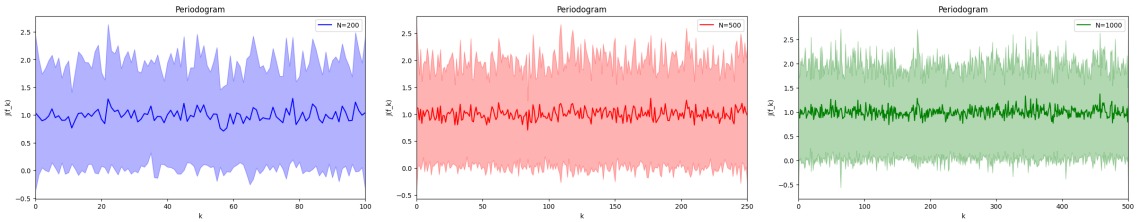
In this question, let X_n $(n = 0, \ldots, N-1)$ be a Gaussian white noise with variance $\sigma^2 = 1$ and set the sampling frequency to $f_s = 1$ Hz

- For $N \in \{200, 500, 1000\}$, compute the *sample autocovariances* ($\hat{\gamma}(\tau)$ vs τ) for 100 simulations of X. Plot the average value as well as the average $\pm$ the standard deviation. What do you observe?
- For $N \in \{200, 500, 1000\}$, compute the *periodogram* ($|J(f_k)|^2$ vs f_k) for 100 simulations of X. Plot the average value as well as the average $\pm$ the standard deviation. What do you observe?

Add your plots to Figure 1.



Autocovariance ($N = 200$) Autocovariance ($N = 500$) Autocovariance ($N = 1000$)



Periodogram ($N = 200$) Periodogram ($N = 500$) Periodogram ($N = 1000$)

Figure 1: Autocovariances and periodograms of a Gaussian white noise (see Question 6).

Answer 6

We have plotted on this Figure the Autocovariance function as well as the periodogram for $N = 200$, $N = 500$ and $N = 1000$ when X is a Gaussian white noise with unit variance.
We observe that the autocovariance sample function is indeed a good estimator for te autocovariance. Our results are consistent with the calculations made in Question 3: the variance is equal to 1 and $\gamma(\tau) = 0$ for $\tau \geq 0$. We also see that the standard deviation of the estimator decreases when N increases, and that it also decreases when t increases for a given value of N. The autocovariance sample function is more precise for larger values of t.
Let us now look at the periodograms. They are constant and equal to the variance, which is again consistent with our theoretical calculations. However, we notice that their standard deviation is rather big compared to that of the plots of the previous rows. We also notice that this standard deviation does not diminish with t for a given N, and does not seem to decrease when the length of acquisition increases.

Question 7

We want to show that the estimator $\hat{\gamma}(\tau)$ is consistent, i.e. it converges in probability when the number N of samples grows to ∞ to the true value $\gamma(\tau)$. In this question, assume that X is a wide-sense stationary *Gaussian* process.

- Show that for $\tau > 0$

$$\mathrm{var}(\hat{\gamma}(\tau)) = (1/N) \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(1 - \frac{\tau + |n|}{N}\right) \left[\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau)\right]. \qquad (13)$$

  (Hint: if $\{Y_1, Y_2, Y_3, Y_4\}$ are four centered jointly Gaussian variables, then $\mathbb{E}[Y_1Y_2Y_3Y_4] = \mathbb{E}[Y_1Y_2]\mathbb{E}[Y_3Y_4] + \mathbb{E}[Y_1Y_3]\mathbb{E}[Y_2Y_4] + \mathbb{E}[Y_1Y_4]\mathbb{E}[Y_2Y_3]$.)

- Conclude that $\hat{\gamma}(\tau)$ is consistent.

Answer 7

$$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\left(\mathbb{E}\left(\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} X_n X_{n+\tau} X_m X_{m+\tau}\right) - \left(\sum_{n=0}^{n=N-\tau-1}\mathbb{E}(X_n X_{n+\tau})\right)^2\right)$$

First of all, we can easily compute the term on the left by exploiting the wide-sense stationarity of the Gaussian process:

$$\left(\sum_{n=0}^{n=N-\tau-1}\mathbb{E}(X_n X_{n+\tau})\right)^2 = (N-\tau)^2\gamma(\tau)^2$$

Let us now develop the expression on the right. In order to do so, we will use the expression suggested in the Question:

$$\mathbb{E}\left(\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} X_n X_{n+\tau} X_m X_{m+\tau}\right) = \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_n X_m)\mathbb{E}(X_{n+\tau}X_{m+\tau}) + \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_n X_{n+\tau})\mathbb{E}(X_m X_{m+\tau}) + \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_n X_{m+\tau})\mathbb{E}(X_{n+\tau}X_m)$$

We will now deal with each of the terms Because of the stationarity of the signal, we can again write:

$$\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_n X_{n+\tau})\mathbb{E}(X_m X_{m+\tau}) = (N-\tau)^2\gamma(\tau)^2$$

$$\begin{aligned}
\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_n X_m)\mathbb{E}(X_{n+\tau}X_{m+\tau}) &= \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_n X_m)^2 \\
&= \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\gamma(n-m)^2 \\
&= \sum_{n=0}^{N-\tau-1}\sum_{k=-N+\tau+1+n}^{N-\tau-n-1}\gamma(k)^2 \\
&= \sum_{k=-N+\tau+1}^{N-\tau-1}\sum_{n=0}^{N-\tau-|k|-1}\gamma(k)^2 \\
&= \sum_{k=-N+\tau+1}^{N-\tau-1}(N-\tau-|k|)\gamma(k)^2
\end{aligned}$$

$$\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\mathbb{E}(X_nX_{m+\tau})\mathbb{E}(X_{n+\tau}X_m) = \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\gamma(m+\tau-n)\gamma(m-n-\tau)$$
$$= \sum_{n=0}^{N-\tau-1}\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}\gamma(k+\tau)\gamma(k-\tau)$$
$$= \sum_{k=-N+\tau+1+n}^{N-\tau-1-n}\sum_{n=0}^{N-\tau-|k|-1}\gamma(k+\tau)\gamma(k-\tau)$$
$$= \sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(N-\tau-|k|)\gamma(k+\tau)\gamma(k-\tau)$$

We can now write the final expression for the variance:

$$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(N-\tau-|k|)(\gamma(k+\tau)\gamma(k-\tau)+\gamma(k)^2)$$
$$= \frac{1}{N}\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(1-\frac{\tau+|k|}{N})(\gamma(k+\tau)\gamma(k-\tau)+\gamma(k)^2)$$

We know that the autocovariance function is such that it is square summable. For $N \in \mathbb{N}, n \in [-N+\tau+1, N-\tau-1]$, $(1-\frac{\tau+|n|}{N})\gamma(n)^2 \leq \gamma(n)^2$ Since the $gamma(\tau)$ series is square summable, we deduce that the sum of $(1-\frac{\tau+|n|}{N})\gamma(n)^2$ converges when N goes to infinity, so $\lim_{N\to+\infty}\frac{1}{N}\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(1-\frac{\tau+|k|}{N})\gamma(n)^2 = 0$ Let us now take a look at the second sum. The Cauchy-Schwarz inequality gives us:

$$\left|\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(1-\frac{\tau+|k|}{N})\gamma(k+\tau)\gamma(k-\tau)\right| \leq \left(\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(1-\frac{\tau+|k|}{N})\gamma\left(k+\tau\right)^2\right)^{1/2}$$
$$\left(\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(1-\frac{\tau+|k|}{N})\gamma\left(k-\tau\right)^2\right)^{1/2} \leq M$$

Where M is a positive real. Indeed, we reuse the previous argument that the sum of the $(1-\frac{\tau+|k|}{N})\gamma\left(k-\tau\right)^2$ is convergent. This means that the second sum is bounded and hence that: $\lim_{N\to+\infty}\frac{1}{N}\sum_{k=-N+\tau+1+n}^{N-\tau-1-n}(1-\frac{\tau+|k|}{N})\gamma(k+\tau)\gamma(k-\tau) = 0$ Finally, this means that$\lim_{N\to+\infty}\text{var}(\hat{\gamma}(\tau)) = 0$

Let us now show that the $\hat{\gamma}(\tau)$ estimator is consistent, meaning that it converges to $\gamma(\tau)$ in probability.
Let $\epsilon, \eta \geq 0$.

$$|\hat{\gamma}_N(\tau)-\gamma(\tau)| > \epsilon \equiv |\hat{\gamma}_N(\tau)-\mathbb{E}\hat{\gamma}_N(\tau)+\mathbb{E}\hat{\gamma}_N(\tau)-\gamma(\tau)| > \epsilon$$

$$\mathbb{P}(\hat{\gamma}_N(\tau)-\gamma(\tau)| > \epsilon) \leq \mathbb{P}(|\hat{\gamma}_N(\tau)-\mathbb{E}\hat{\gamma}_N(\tau)| > \epsilon - |\mathbb{E}\hat{\gamma}_N(\tau)-\gamma(\tau)|)$$

We know that $\hat{\gamma}_N(\tau)$ is asymptotically unbiased, so there is $N_1 \in \mathbb{N}$ such that for $N \geq N_1$, $|\mathbb{E}\hat{\gamma}_N(\tau)-\gamma(\tau)|) \leq \frac{\epsilon}{2}$ Let us now use the Bienaymé-Tchebychev inequation:

$$\mathbb{P}(|\hat{\gamma}_N(\tau)-\mathbb{E}\hat{\gamma}_N(\tau)| \geq \frac{\epsilon}{2}) \leq \frac{4\text{var}(\hat{\gamma_N}(\tau))}{\epsilon^2}$$

We have seen that $\lim_{N\to+\infty} \text{var}(\hat{\gamma_N}(\tau)) = 0$ so there is $N_2 \in \mathbb{N}$ such that for $N \geq N_2$, $\mathbb{P}|\hat{\gamma}_N(\tau) - \mathbb{E}\hat{\gamma}_N(\tau)| \geq \frac{\epsilon}{2} \leq \frac{\eta\epsilon^2}{4}$ Finally, for $N \geq \max(N_1, N_2)$, $\mathbb{P}(|\hat{\gamma}_N(\tau) - \gamma(\tau)| > \epsilon) \leq \eta$ This proves that

$$\lim_{N\to+\infty} \mathbb{P}(|\hat{\gamma}_N(\tau) - \gamma(\tau)| > \epsilon) = 0$$

Contrary to the correlogram, the periodogram is not consistent. It is one of the most well-known estimators that is asymptotically unbiased but not consistent. In the following question, this is proven for a Gaussian white noise but this holds for more general stationary processes.

Question 8

Assume that X is a Gaussian white noise (variance σ^2) and let $A(f) := \sum_{n=0}^{N-1} X_n \cos(-2\pi fn/f_s$ and $B(f) := \sum_{n=0}^{N-1} X_n \sin(-2\pi fn/f_s$. Observe that $J(f) = (1/\sqrt{N})(A(f) + iB(f))$.

- Derive the mean and variance of $A(f)$ and $B(f)$ for $f = f_0, f_1, \ldots, f_{N/2}$ where $f_k = f_s k/N$.
- What is the distribution of the periodogram values $|J(f_0)|^2, |J(f_1)|^2, \ldots, |J(f_{N/2})|^2$.
- What is the variance of the $|J(f_k)|^2$? Conclude that the periodogram is not consistent.
- Explain the erratic behavior of the periodogram in Question 6 by looking at the covariance between the $|J(f_k)|^2$.

Answer 8

X is a Gaussian white noise so it is a centered process. This means that $\mathbb{E}(A(f)) = 0$ and $\mathbb{E}(B(f)) = 0$ for all f. In addition, X is a Gaussian white noise which means that $forn, m \leq N - 1n \neq m \to \mathbb{E}(X_n X_m) = 0$, so:

$$\text{var}(A(f)) = \sum_{n=0}^{N-1} \text{var}(X_n) \cos(-2\pi fn/f_s)^2 = \sigma^2 \sum_{n=0}^{N-1} \cos(-2\pi fn/f_s)^2$$

Let the frequence f be such that there is $k \in [0, N-2]$ such that $f = f_k = f_s k/N$

$$\text{var}(A(f_k)) = \sigma^2 \sum_{n=0}^{N-1} \cos(-2\pi nk/N)^2 = \sigma^2 \sum_{n=0}^{N-1} (1 + \cos(-4\pi nk/N))/2$$

We can also write:

$$\text{var}(B(f_k)) = \sigma^2 \sum_{n=0}^{N-1} \sin(-2\pi nk/N)^2 = \sigma^2 \sum_{n=0}^{N-1} (1 - \sin(-4\pi nk/N))/2$$

Let us here assume that k is different from 0 and $N/2$ (otherwise we could not write the following without dividing by 0). $\sum_{n=0}^{N-1} e^{-4i\pi nk/N} = \frac{1-e^{-4i\pi k}}{1-e^{-4i\pi k/N}} = 0$ By taking the real and imaginary value of the previous sum, we show that:

$$\sum_{n=0}^{N-1} \sin(-4\pi nk/N) = \sum_{n=0}^{N-1} \cos(-4\pi nk/N) = 0$$

If $k = 0$:$\sum_{n=0}^{N-1} \sin(-4\pi nk/N) = 0$ and if $k = N/2$: $\sum_{n=0}^{N-1} \sin(-\pi k) = 0$

$$\text{var}(A(f)) = \frac{N}{2}\sigma^2$$

$$\text{var}(B(f)) = \frac{N}{2}\sigma^2$$

Let k be in $[0, N/2]$, $k \neq 0, N/2$ $|J(f_k)|^2 = \frac{1}{N^2}(A(f_k)^2 + B(f_k)^2)$ The $(X_n)_{n \leq N-1}$ random variables are Gaussian and independent, hence $A(f)$ and $B(f)$ which are linear combinations of these variables are also centered Gaussian variables with variance $\frac{N}{2}\sigma^2$. Let us now prove that $A(f)$ and $B(f)$ are independent. Since they are Gaussian variables, all we have to do is to show that they are decorrelated, or that $\mathbb{E}(A(f)B(f)) = 0$ in this case.

$$\mathbb{E}(A(f)B(f)) = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} \mathbb{E}(X_n X_m)\cos(-2\pi fn/f_s)\sin(-2\pi fm/f_s) = \frac{\sigma^2}{2}\sum_{n=0}^{N-1}\sin(-4\pi fn/f_s) = 0$$

Finally, the random variable $|J(f_k)|^2$ follows a chi-squared law of dimension 2.
If $k = 0$, then $A(f_0) = \sum_{n=0}^{N-1} X_n$ and $B(f_0) = 0$. $\text{var}(A(f_0)) = N\sigma^2$ and $\text{var}(B(f_0)) = 0$. This means that $|J(f_0)|^2 = \frac{1}{N}A(f_0)^2$. $A(f_0)$ is a centered Gaussian variable with variance $N\sigma^2$ so $|J(f_0)|^2$ follows a chi-squared law of dimension 1.
If $k = \frac{N}{2}$, $A(f_{N/2}) = \sum_{n=0}^{N_1}(-1)^n X_n$ and $B(f_{N/2}) = 0$. $\text{var}(A(f_{N/2})) = N\sigma^2$ and $\text{var}(B(f_{N/2})) = 0$. This means that $|J(f_{N/2})|^2 = \frac{1}{N}A(f_{N/2})^2$. $A(f_{N/2})$ is a centered Gaussian variable with variance $N\sigma^2$ so $|J(f_{N/2})|^2$ follows a chi-squared law of dimension 1.

Let $k \in [0, \frac{N}{2}]$, $k \neq 0, N/2$.
We know that $A(f_k)$ and $B(f_k)$ are independent so:

$$\text{var}|J(f_k)|^2 = \frac{1}{N^2}(\text{var}(A(f_k))^2 + \text{var}(B(f_k)^2)) = \sigma^4$$

For $k = 0$ or $k = N/2$,

$$\text{var}|J(f_k)|^2 = 2\sigma^4$$

This means that the variance of the periodogram does not converge to 0. This explains that the periodogram is not consistant.
Let us now look at the covariances between the $|J(f_k)|^2$. Let $k, l \in [0, \frac{N}{2}]$, $k, l \neq 0, N/2$.

$$\text{Cov}(|J(f_k)|^2, |J(f_l)|^2) = \mathbb{E}(|J(f_k)|^2|J(f_l)|^2) - \sigma^4$$

We can show that $\text{Cov}(|J(f_k)|^2, |J(f_l)|^2) = 0$. The variables are uncorrelated which explains the behaviour of the periodogram.

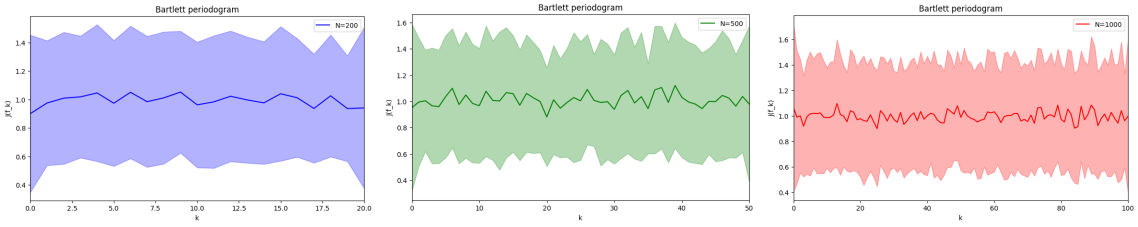
Question 9

As seen in the previous question, the problem with the periodogram is the fact that its variance does not decrease with the sample size. A simple procedure to obtain a consistent estimate is to divide the signal in K sections of equal durations, compute a periodogram on each section and average them. Provided the sections are independent, this has the effect of dividing the variance by K. This procedure is known as Bartlett's procedure.

- Rerun the experiment of Question 6, but replace the periodogram by Barlett's estimate (set $K = 5$). What do you observe.

Add your plots to Figure 2.

Answer 9



Periodogram ($N = 200$) Periodogram ($N = 500$) Periodogram ($N = 1000$)

Figure 2: Barlett's periodograms of a Gaussian white noise (see Question 9).

4 Data study

4.1 General information

Context. The study of human gait is a central problem in medical research with far-reaching consequences in the public health domain. This complex mechanism can be altered by a wide range of pathologies (such as Parkinson's disease, arthritis, stroke,...), often resulting in a significant loss of autonomy and an increased risk of fall. Understanding the influence of such medical disorders on a subject's gait would greatly facilitate early detection and prevention of those possibly harmful situations. To address these issues, clinical and bio-mechanical researchers have worked to objectively quantify gait characteristics.

Among the gait features that have proved their relevance in a medical context, several are linked to the notion of step (step duration, variation in step length, etc.), which can be seen as the core atom of the locomotion process. Many algorithms have therefore been developed to automatically (or semi-automatically) detect gait events (such as heel-strikes, heel-off, etc.) from accelerometer and gyrometer signals.

Data. Data are described in the associated notebook.

4.2 Step classification with the dynamic time warping (DTW) distance

Task. The objective is to classify footsteps then walk signals between healthy and non-healthy.

Performance metric. The performance of this binary classification task is measured by the F-score.

Question 10

Combine the DTW and a k-neighbors classifier to classify each step. Find the optimal number of neighbors with 5-fold cross-validation and report the optimal number of neighbors and the associated F-score. Comment briefly.

Answer 10

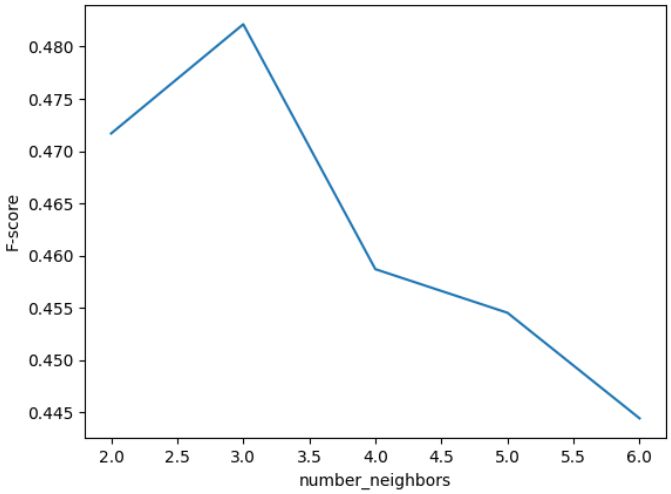


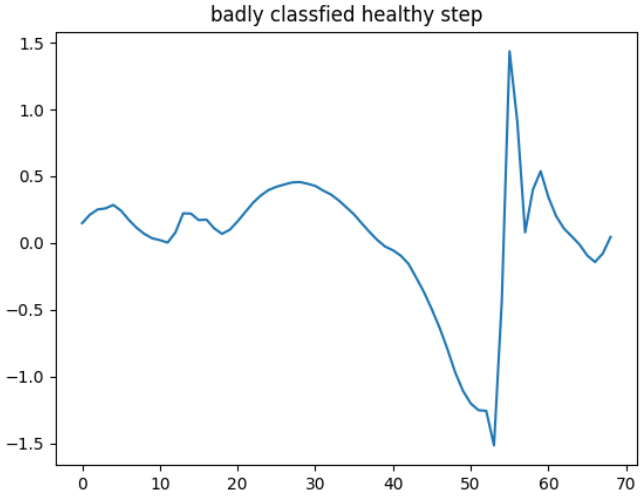
Figure 3: F-scores for some values of the number of neighbors.

We do cross-validation on the following calues of the number of neighbors $[2, 3, 4, 5, 6]$. The best hyperparameter is clearly 3.

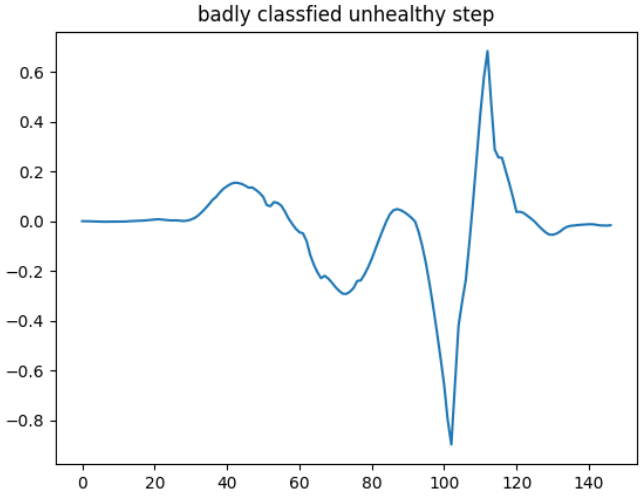
Question 11

Display on Figure 4 a badly classified step from each class (healthy/non-healthy).

Answer 11



Badly classified healthy step



Badly classified non-healthy step

Figure 4: Examples of badly classified steps (see Question 11).