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

Bariller Halvard halvard.bariller@ens-paris-saclay.fr
Ellouze Farah farah.ellouze@gmail.com

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:

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

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 $\mathbf{0}_p$ (a p-dimensional vector of zeros) for any $\lambda > \lambda_{\max}$.

Answer 1

The problem (1) is equivalent to the following optimization problem

$$\min_{\beta \in \mathbb{R}^p} f(x) = \frac{1}{2} \|y - X\beta\|_2^2$$

s.t

$$\|\beta\|_1 \leq t$$

This is a convex optimization problem because f is convex and the feasible set is convex. The first order condition holds.

$$L(\beta, \lambda) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda(\Sigma_{i=1}^p |\beta_i| - t)$$

Taking the sub-differential of the Lagrangian with respect to β, we have:

$$\nabla L(\beta, \lambda) = -X^T y + X^T X\beta + \lambda \Sigma_{i=1}^p s(\beta_i) = -X^T(y - X\beta) + \lambda \Sigma_{i=1}^p s_i(\beta_i)$$

where $\mathrm{s}_i(\beta_i) = \begin{cases} 1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ \in [-1, 1] & \text{otherwise} \end{cases}$

$$\nabla L(x, \lambda) = 0 \iff X^T(y - X\beta) = \lambda s(\beta)$$

If the minimizer of (1) is $\mathbf{0}_p$ then

$$\lambda s = X^T y$$

Thus

$$\lambda \|s\|_\infty = \left\|X^T y\right\|_\infty$$

If $\|s\|_\infty \neq 1$, then $\forall \lambda$, $\lambda \|s\|_\infty < \lambda$ and $\left\|X^T y\right\|_\infty < \lambda$ which contradicts the equality obtained if minimizer of (1) is $\mathbf{0}_p$.
Hence, if $\beta = 0$, we have that $\|s\|_\infty = 1$, and the smallest value of λ for $\beta = 0$ is $\lambda \|s\|_\infty = \lambda = \left\|X^T y\right\|_\infty$.
Finally, we have shown that $\forall \lambda \geq \lambda_{max} = \left\|X^T y\right\|_\infty$, the minimizer of (1) is $\mathbf{0}_p$.

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:

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

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

- $\mathbf{d}_k \in \mathbb{R}^L$ are the K dictionary atoms
- $\mathbf{z}_k \in \mathbb{R}^{N-L+1}$ are activations signals
- In this question, we aim to write

$$\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k = \mathbf{DZ} \tag{Eq1}$$

where $\mathbf{D} \in \mathbb{R}^{N*K(N-L+1)}$ and $\mathbf{Z} \in \mathbb{R}^{K(N-L+1)}$, and

$$\lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 = \lambda \|\mathbf{Z}\|_1 \tag{Eq2}$$

Starting with (Eq1),

$$\left(\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right)[i] = \sum_{k=1}^{K} \sum_{j=1}^{N-L+1} (\mathbf{z}_k)[j](\mathbf{d}_k)[i-j] \tag{$*$}$$

for i in $\in [1, N]$

We then define $(\mathbf{Z}_k)_j$ = $\begin{cases} (\mathrm{z}_k)[j] & \text{if } 1 \leq j \leq N-L+1 \\ 0 & \text{otherwise} \end{cases}$

and $(\mathbf{D}_k)_{ij}$ = $\begin{cases} (\mathrm{d}_k)[i-j] & \text{if } 1 \leq i-j \leq L \\ 0 & \text{otherwise} \end{cases}$

$$(\mathbf{D_k Z_k})_i = \sum_{j=1}^{N-L+1} (\mathbf{D}_k)_{ij} (\mathbf{Z}_k)_j$$

We can view the matrix $\mathbf{D}$ as K matrices in $\mathbb{R}^{N*(N-L+1)}$ stacked horizontally; and Z, as K vectors in $\mathbb{R}^{(N-L+1)}$ stacked vertically where $\mathbf{D}_k$ is the k-th matrix of $\mathbf{D}$ and $\mathbf{Z}_k$ is the k-th vector of $\mathbf{Z}$. Hence,

$$(\mathbf{DZ})_i = \sum_{k=1}^{K} \sum_{j=1}^{N-L+1} (\mathbf{D}_k)_{ij} (\mathbf{Z}_k)_j = \left(\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right)[i]$$

Thus, $\left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 = \|\mathbf{x} - \mathbf{DZ}\|_2^2$.

Now, for (Eq2),

$$\begin{aligned}
\lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 &= \lambda \sum_{k=1}^{K} \sum_{j=1}^{N-L+1} |\mathbf{z}_k[j]| \\
&= \lambda \sum_{k=1}^{K} \sum_{j=1}^{N-L+1} |(\mathbf{Z}_k)_j|
\end{aligned}$$

with $(Z_k)_j$ defined as earlier for (Eq1)

$$\begin{aligned}
&= \lambda \sum_{j=1}^{K(N-L+1)} |\mathbf{Z}_j| \\
&= \lambda \|\mathbf{Z}\|_1
\end{aligned}$$

Thus, for a fixed dictionnary, the sparse coding problem is equivalent to the following Lasso problem :

$$\min_{(\mathbf{z}_k)_k} \quad \|\mathbf{x} - \mathbf{D}\mathbf{Z}\|_2^2 \quad + \quad \lambda \|\mathbf{Z}\|_1 \tag{Pb.2}$$

Using the result in **Question 1**,

$$\lambda_{max} = \left\|\mathbf{D}^T\mathbf{x}\right\|_\infty \tag{3}$$

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:

$$S(f) := \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2i\pi f\tau/f_s}. \tag{4}$$

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

By definition of a Gaussian white noise, $(X_n)_n$ is a stochastic process with independant and identically distributed random variables, following a Gaussian distribution such that : $\forall n \in \{0, \ldots, N-1\},\ X_n \sim \mathcal{N}(\mu, \tau^2)$ with finite variance.
As a direct consequence of the independance, we have that:

$$\gamma(\tau) := \mathbb{E}(X_n X_{n+\tau}) = \begin{cases} \sigma^2 & \text{if } \tau = 0 \\ 0 & \text{otherwise} \end{cases}$$

It follows that:

$$S(f) \triangleq \gamma(0) = \sigma^2.$$

The power spectrum of such process is constant over the frequency domain, with the variance as constant value.

Question 4

A natural estimator for the autocorrelation function is the sample autocovariance

$$\hat{\gamma}(\tau) := (1/N) \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \tag{5}$$

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

- 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

We compute the expected value of the sample estimator:

$$\mathbb{E}\left[\hat{\gamma}(\tau)\right] = \mathbb{E}\left[\frac{1}{N} \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}\right] = \frac{1}{N} \sum_{n=0}^{N-\tau-1} \mathbb{E}\left[X_n X_{n+\tau}\right] = \left(1 - \frac{\tau}{N}\right) \gamma(\tau) \xrightarrow[N \to \infty]{} \gamma(\tau).$$

This proves that the sample autocovariance is a biased estimator, but asymptotically unbiased.
An unbiased version would be obtained by taking $\frac{N}{N-\tau}\hat{\gamma}(\tau)$ as estimator.

Question 5

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

$$J(f) := (1/\sqrt{N}) \sum_{n=0}^{N-1} X_n e^{-2\pi i f n / f_s} \tag{6}$$

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

- As a result of the previous definitions, we express the spectrum element $|J(f_k)|^2$ of the discrete Fourier transform as follows:

$$
\begin{aligned}
|J(f_k)|^2 &= J(f_k)\overline{J(f_k)} \\
&= \frac{1}{N}\left(\sum_{n=0}^{N-1} X_n e^{-2\pi i \frac{kn}{N}}\right)\left(\sum_{m=0}^{N-1} X_m e^{2\pi i \frac{km}{N}}\right) \\
&= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} X_n X_m e^{2\pi i \frac{k(m-n)}{N}} \\
&= \frac{1}{N}\left(\sum_{n=0}^{N-1} X_n^2 + \sum_{n=0}^{N-1}\sum_{\substack{m=0\\ m\neq n}}^{N-1} X_n X_m e^{2\pi i \frac{k(m-n)}{N}}\right) \\
&= \hat{\gamma}(0) + \frac{1}{N}\sum_{n=0}^{N-1}\sum_{\substack{m=0\\ m\neq n}}^{N-1} X_n X_m e^{2\pi i \frac{k(m-n)}{N}} \\
&\underset{m-n=\tau}{=} \hat{\gamma}(0) + \frac{1}{N}\left[\sum_{\tau=-N+1}^{-1}\sum_{n=0}^{N+\tau-1} X_n X_{n-\tau} e^{2\pi i \frac{k\tau}{N}} + \sum_{\tau=1}^{N-1}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} e^{2\pi i \frac{k\tau}{N}}\right] \\
&= \hat{\gamma}(0) + \sum_{\tau=-N+1}^{-1} \hat{\gamma}(-\tau) e^{2\pi i \frac{k\tau}{N}} + \sum_{\tau=1}^{N-1} \hat{\gamma}(\tau) e^{2\pi i \frac{k\tau}{N}} \\
&= \hat{\gamma}(0) + \sum_{\tau=-N+1}^{-1} \hat{\gamma}(\tau) e^{2\pi i \frac{k\tau}{N}} + \sum_{\tau=1}^{N-1} \hat{\gamma}(\tau) e^{2\pi i \frac{k\tau}{N}} \\
&= \sum_{\tau=-N+1}^{N-1} \hat{\gamma}(\tau) e^{2\pi i \frac{k\tau}{N}}.
\end{aligned}
$$

  Finally, using the parity of $\hat{\gamma}$, we have that:

$$|J(f_k)|^2 = \sum_{\tau=-N+1}^{N-1} \hat{\gamma}(\tau) e^{-2\pi i \frac{k\tau}{N}}.$$

- We define the frequency resolution $\Delta f = \frac{f_s}{N}$, which corresponds to the space between two observable frequencies. For a frequency $f > 0$, we can then define $f^{(N)}$ as follows:

$$\begin{cases} f^{(N)} = f_{k^*} \\ k^* = \lfloor \frac{f}{\Delta f} + \frac{1}{2} \rfloor \end{cases}$$

  that is the frequency $f^{(N)}$ minimizing $|f - f^{(N)}| = |f - f_{k^*}| = |f - \frac{f_s k^*}{N}|$.

We then have that:

$$\begin{aligned}
\mathbb{E}\left[|J(f_k^*)|^2\right] &= \mathbb{E}\left[\sum_{\tau=-N+1}^{N-1} \hat{\gamma}(\tau) e^{-2\pi i \frac{k^* \tau}{N}}\right] \\
&= \sum_{\tau=-N+1}^{N-1} \mathbb{E}\left[\hat{\gamma}(\tau)\right] e^{-2\pi i \frac{k^* \tau}{N}} \\
&= \sum_{\tau=-N+1}^{N-1} \left(1 - \frac{\tau}{N}\right) \gamma(\tau) e^{-2\pi i \tau \frac{f^{(N)}}{f_s}} \\
&\xrightarrow[N \to \infty]{} \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2\pi i \tau \frac{f^{(N)}}{f_s}} = S(f).
\end{aligned}$$

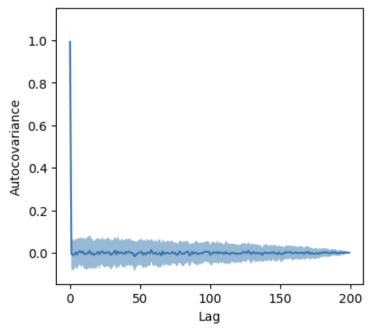
This concludes that for a frequency $f > 0$, $|J(f^{(N)})|^2$ is an asymptotically unbiased estimator of $S(f)$.

Question 6

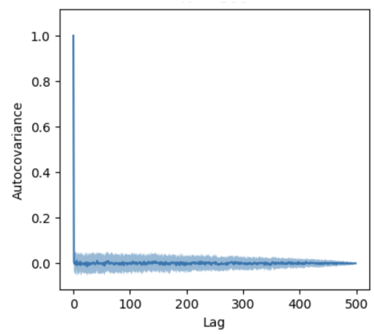
In this question, let X_n ($n = 0, \ldots, N-1$) be a Gaussian white noise with variance $\sigma^2 = 1$ and set the sampling frequency to $f_s = 1$ Hz

- 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?

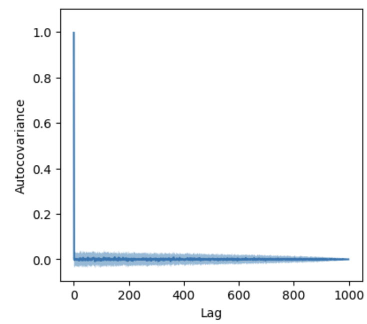
Add your plots to Figure 1.



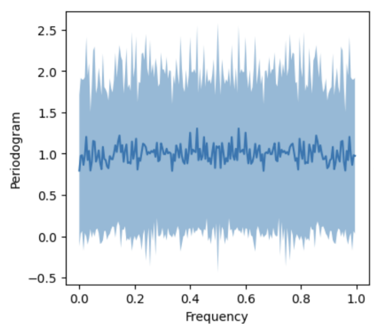
Autocovariance ($N = 200$)



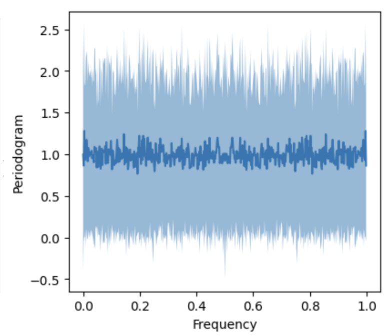
Autocovariance ($N = 500$)



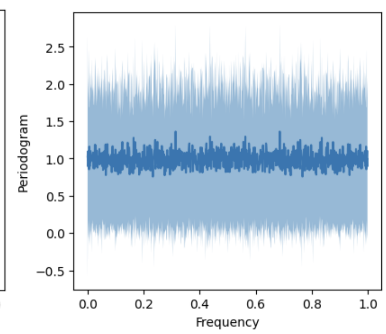
Autocovariance ($N = 1000$)



Periodogram ($N = 200$)



Periodogram ($N = 500$)



Periodogram ($N = 1000$)

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

Answer 6

The obtained plots confirm the conclusions made in question 3.
As expected for Gaussian white noise, we observe that the sample autocovariance takes null values for lag different than 0, confirming that the X_n are independent.
Regarding the periodogram, we notice that the values taken seem uniformly distributed across all frequencies, centered around the variance of the random process.
Finally, on the periodogram plots, we observe that increasing the sample size does not lead to a reduction of the variance (unlike for the sample autocovariance).

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 $\tau > 0$

$$\mathrm{var}(\hat{\gamma}(\tau)) = (1/N) \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(1 - \frac{\tau + |n|}{N}\right) \left[\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau)\right]. \tag{7}$$

(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

Let X be a wide-sense stationary Gaussian process.

- We begin by computing the variance of the sample autocorrelation for $\tau > 0$.

$$\begin{aligned}
\operatorname{var}(\hat{\gamma}(\tau)) &= \mathbb{E}\left[\hat{\gamma}^2(\tau)\right] - \mathbb{E}\left[\hat{\gamma}(\tau)\right]^2 \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}\right)^2\right] - \left(\frac{(N-\tau-1)}{N}\gamma(\tau)\right)^2 \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} X_n X_{n+\tau} X_m X_{m+\tau}\right] - \left(\frac{(N-\tau-1)}{N}\gamma(\tau)\right)^2 \\
&= \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \mathbb{E}\left[X_n X_{n+\tau} X_m X_{m+\tau}\right] - \left(\frac{(N-\tau-1)}{N}\gamma(\tau)\right)^2 \\
&= \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \left(\mathbb{E}\left[X_n X_{n+\tau}\right]\mathbb{E}\left[X_m X_{m+\tau}\right] + \mathbb{E}\left[X_n X_m\right]\mathbb{E}\left[X_{n+\tau} X_{m+\tau}\right]\right. \\
&\quad \left. + \mathbb{E}\left[X_n X_{m+\tau}\right]\mathbb{E}\left[X_m X_{n+\tau}\right]\right) - \left(\frac{(N-\tau-1)}{N}\gamma(\tau)\right)^2 \\
&= \frac{1}{N^2}\left((N-\tau-1)^2\gamma^2(\tau) + \sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \left(\mathbb{E}\left[X_n X_m\right]\mathbb{E}\left[X_{n+\tau} X_{m+\tau}\right]\right.\right. \\
&\quad \left.\left. + \mathbb{E}\left[X_n X_{m+\tau}\right]\mathbb{E}\left[X_m X_{n+\tau}\right]\right)\right) - \left(\frac{(N-\tau-1)}{N}\gamma(\tau)\right)^2 \\
&= \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \left(\mathbb{E}\left[X_n X_m\right]\mathbb{E}\left[X_{n+\tau} X_{m+\tau}\right] + \mathbb{E}\left[X_n X_{m+\tau}\right]\mathbb{E}\left[X_m X_{n+\tau}\right]\right) \\
&= \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1} \left(\gamma^2(m-n) + \gamma(m+\tau-n)\gamma(m-n-\tau)\right).
\end{aligned}$$

Moreover, we have the following:

$$\begin{aligned}
&\sum_{n,m=0}^{N-\tau-1}\gamma^2(m-n) \underset{m-n=k}{=} \left[\gamma^2(0)+\cdots+\gamma^2(N-\tau-1)\right]+\cdots+\left[\gamma^2(-(N-\tau-1)+\cdots+\gamma^2(0)\right] \\
&= (N-\tau)\gamma^2(0) + (N-\tau-1)[\gamma^2(1)+\gamma^2(-1)] + \cdots + [\gamma^2(N-\tau-1)+\gamma^2(-(N-\tau-1))] \\
&= \sum_{k=-(N-\tau-1)}^{N-\tau-1}(N-\tau-|k|)\gamma^2(k).
\end{aligned}$$

Using the aforementioned change of indices, we have:

$$\begin{aligned}
\operatorname{var}(\hat{\gamma}(\tau)) &= \frac{1}{N^2}\sum_{n=0}^{N-\tau-1}\sum_{m=0}^{N-\tau-1}\left(\gamma^2(m-n)+\gamma(m+\tau-n)\gamma(m-n-\tau)\right)\\
&= \frac{1}{N^2}\sum_{k=-(N-\tau-1)}^{N-\tau-1}(N-\tau-|k|)\gamma^2(k)+\gamma(k+\tau)\gamma(k-\tau)\\
&= \frac{1}{N}\sum_{k=-(N-\tau-1)}^{N-\tau-1}(1-\frac{\tau+|k|}{N})\gamma^2(k)+\gamma(k+\tau)\gamma(k-\tau).
\end{aligned}$$

- We recall that we are under the assumptions that the autocovariances are absolutely summable, i.e. $\sum_{\tau\in\mathbb{Z}}|\gamma(\tau)|<\infty$, and square summable, i.e. $\sum_{\tau\in\mathbb{Z}}\gamma^2(\tau)<\infty$.
  The Bienaymé-Chebyshev's inequality gives us the following:

  $$\mathbb{P}(|\hat{\gamma}(\tau)-\mathbb{E}[\hat{\gamma}(\tau)]|>\epsilon)\leq\frac{\operatorname{var}(\hat{\gamma}(\tau))}{\epsilon^2}.$$

  Therefore, using the assumptions and the expression found for the variance, the factor term in $1/N$ when passing to the limit gives us:

  $$\lim_{N\to\infty}\mathbb{P}(|\hat{\gamma}(\tau)-\mathbb{E}[\hat{\gamma}(\tau)]|>\epsilon)=\mathbb{P}(|\hat{\gamma}(\tau)-\gamma(\tau)|>\epsilon)\leq\lim_{N\to\infty}\frac{\operatorname{var}(\hat{\gamma}(\tau))}{\epsilon^2}=0.$$

  This concludes the proof for the consistency of $\hat{\gamma}(\tau)$.

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 $f=f_0,f_1,\dots,f_{N/2}$ where $f_k=f_sk/N$.
- What is the distribution of the periodogram values $|J(f_0)|^2$, $|J(f_1)|^2$, $\dots$, $|J(f_{N/2})|^2$.
- 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

Let $X=(X_1,\dots,X_{N-1}),X_n\sim\mathcal{N}(0,\sigma^2)$, be a Gaussian white noise.

- We compute the mean of $A(f)$ and $B(f)$:
  $$\begin{cases}\mathbb{E}[A(f)]=\mathbb{E}\left[\sum_{n=0}^{N-1}X_n\cos(-2\pi fn/f_s)\right]=\sum_{n=0}^{N-1}\mathbb{E}[X_n]\cos(-2\pi fn/f_s)=0.\\ \mathbb{E}[B(f)]=\mathbb{E}\left[\sum_{n=0}^{N-1}X_n\sin(-2\pi fn/f_s)\right]=\sum_{n=0}^{N-1}\mathbb{E}[X_n]\sin(-2\pi fn/f_s)=0.\end{cases}$$

Recalling the independence of the X_n's in the case of white noise, we compute the variance of $A(f)$ for $k \in \{0, N/2\}$:

$$\begin{aligned}
\text{var}(A(f_k)) &= \text{var}\left(\sum_{n=0}^{N-1} X_n \cos\left(-2\pi \frac{f_k n}{f_s}\right)\right) \\
&= \text{Cov}\left(\sum_{n=0}^{N-1} X_n \cos\left(-2\pi \frac{f_k n}{f_s}\right), \sum_{m=0}^{N-1} X_m \cos\left(-2\pi \frac{f_k m}{f_s}\right)\right) \\
&= \sum_{n,m=0}^{N-1} \cos\left(-2\pi \frac{f_k n}{f_s}\right) \cos\left(-2\pi \frac{f_k m}{f_s}\right) \text{Cov}(X_n, X_m) \\
&= \sum_{n=0}^{N-1} \cos^2\left(-2\pi \frac{f_k n}{f_s}\right) \text{var}(X_n) \\
&= \sigma^2 \sum_{n=0}^{N-1} \cos^2\left(-2\pi \frac{f_k n}{f_s}\right) \\
&= \sigma^2 \sum_{n=0}^{N-1} \frac{1}{2}\left(1 + \cos\left(-4\pi \frac{kn}{N}\right)\right) \\
&= \frac{\sigma^2}{2}\left(N + \sum_{n=0}^{N-1} \cos\left(-4\pi \frac{kn}{N}\right)\right) \\
&= \frac{\sigma^2 N}{2} + \frac{\sigma^2}{2} \sum_{n=0}^{N-1} \cos\left(4\pi \frac{kn}{N}\right).
\end{aligned}$$

In the cases where $k \in \{0, \frac{N}{2}\}$, the cosinus term is constant equal to 1, the second term then becomes equal to $\frac{\sigma^2 N}{2}$ and as a result, $\text{var}(A(f_k)) = \sigma^2 N$.
For all $k \in \{1, \ldots, \frac{N}{2} - 1\}$, we write the sum as follows [Knapp, 2009]:

$$\sum_{n=0}^{N-1} \cos\left(4\pi \frac{kn}{N}\right) = \frac{\sin\left(N \frac{2\pi k}{N}\right)}{\sin\left(\frac{2\pi k}{N}\right)} \cos\left((N-1)\frac{2\pi k}{N}\right)$$

The numerator being equal to 0 with k integer, it thus follows that the sum is equal to 0 all-together.
Therefore, we have the following:

$$\text{var}(A(f_k)) = \begin{cases} \sigma^2 N & \text{if } f \in \{f_0, f_{N/2}\} \\ \frac{\sigma^2 N}{2} & \text{otherwise} \end{cases}$$

With the same reasoning, the variance of $B(f)$ for $k \in \{0, N/2\}$ is:

$$\begin{aligned}
\text{var}(B(f_k)) &= \text{var}\left(\sum_{n=0}^{N-1} X_n \sin\left(-2\pi\frac{f_k n}{f_s}\right)\right) \\
&= \sigma^2 \sum_{n=0}^{N-1} \sin^2\left(-2\pi\frac{f_k n}{f_s}\right) \\
&= \sigma^2 \sum_{n=0}^{N-1} 1 - \cos^2\left(-2\pi\frac{f_k n}{f_s}\right) \\
&= \sigma^2 N - \text{var}(A(f_k)) \\
&= \begin{cases} 0 & \text{if } f \in \{f_0, f_{N/2}\} \\ \frac{\sigma^2 N}{2} & \text{otherwise} \end{cases}
\end{aligned}$$

- As sums of independent Gaussian random variables, $A(f)$ and $B(f)$ are also Gaussian random variables with mean 0 and variance $\frac{\sigma^2 N}{2}$ for $k \in \{1, \ldots, \frac{N}{2} - 1\}$ or variance of respectively $\sigma^2 N$ and 0 for $k \in \{0, \frac{N}{2}\}$.
  Moreover, we have that [Knapp, 2009]:

$$\begin{aligned}
\text{Cov}(A(f), B(f)) &= \sum_{n,m=0}^{N-1} \cos\left(-2\pi\frac{f_k n}{f_s}\right) \sin\left(-2\pi\frac{f_k m}{f_s}\right) \text{Cov}(X_n, X_m) \\
&= \sigma^2 \sum_{n=0}^{N-1} \cos\left(-2\pi\frac{f_k n}{f_s}\right) \sin\left(-2\pi\frac{f_k n}{f_s}\right) \\
&= \sigma^2 \sum_{n=0}^{N-1} \frac{1}{2} \sin\left(-4\pi\frac{f_k n}{f_s}\right) \\
&= \begin{cases} 0 & \text{if } f \in \{f_0, f_{N/2}\} \\ \frac{\sigma^2}{2} \frac{\sin\left(N\frac{2\pi k}{N}\right)}{\sin\left(\frac{2\pi k}{N}\right)} \sin\left((N-1)\frac{2\pi k}{N}\right) = 0 & \text{otherwise} \end{cases}
\end{aligned}$$

  Therefore, $A(f)$ and $B(f)$ are independent Gaussian variables.
  Finally, we have that:

$$\begin{aligned}
|J(f)|^2 &= \left|\frac{1}{\sqrt{N}}\left(A(f) + iB(f)\right)\right|^2 \\
&= \frac{1}{N}\left(A^2(f) + B^2(f)\right)
\end{aligned}$$

  As a sum of independent squared Gaussian random variables, we have that:

$$\begin{cases} \frac{1}{\sigma^2}|J(f_k)|^2 \sim \chi^2(1) & \text{if } f \in \{f_0, f_{N/2}\} \\ \frac{2}{\sigma^2}|J(f_k)|^2 \sim \chi^2(2) & \text{otherwise} \end{cases}$$

- Following from the distribution of the $|J(f_k)|^2$, we have:

$$\begin{cases} \text{var}(\frac{1}{\sigma^2}|J(f_k)|^2) = 2 & \text{if } k \in \{0, N/2\} \\ \text{var}(\frac{2}{\sigma^2}|J(f_k)|^2) = 4 & \text{otherwise} \end{cases} \implies \begin{cases} \text{var}(|J(f_k)|^2) = 2\sigma^4 & \text{if } k \in \{0, N/2\} \\ \text{var}|J(f_k)|^2) = \sigma^4 & \text{otherwise} \end{cases}$$

For each value composing the periodogram, the variance of the corresponding random variable does not depend of the sample size N. Therefore, increasing the sample size won't lead to a decrease of the variance as we had when considering the sample autocovariance estimator. Consequently, despite being asymptotically unbiased, the periodogram is not consistent.

- We consider the covariance between the $|J(f_k)|^2$:

$$\mathrm{Cov}(|J(f_{k_1})|^2, |J(f_{k_2})|^2) = \frac{1}{N^2}\Big[\mathrm{Cov}(A^2(f_{k_1}), A^2(f_{k_2})) + \mathrm{Cov}(A^2(f_{k_1}), B^2(f_{k_2})) + \mathrm{Cov}(B^2(f_{k_1}), A^2(f_{k_2})) + \mathrm{Cov}(B^2(f_{k_1}), B^2(f_{k_2}))\Big].$$

In particular, using the previous results and formulas, we have for $k_1 \neq k_2$:

$$\mathrm{Cov}(A(f_{k_1}), A(f_{k_2})) = \frac{\sigma^2}{2}\left(\sum_{n=0}^{N-1} \cos\left(2\pi\frac{(k_1-k_2)n}{N}\right) + \cos\left(2\pi\frac{(k_1+k_2)n}{N}\right)\right) = 0.$$

Similarly, we can show that all of the previous covariances are equal to zero. Therefore, the $|J(f_k)|$ are independent, and so are the $|J(f_k)|^2$.
The independency of these coefficients on top of of the stochastic effect throughout the sampling of the random process are the reason explaining the erratic behavior when estimating the power spectrum.

Question 9

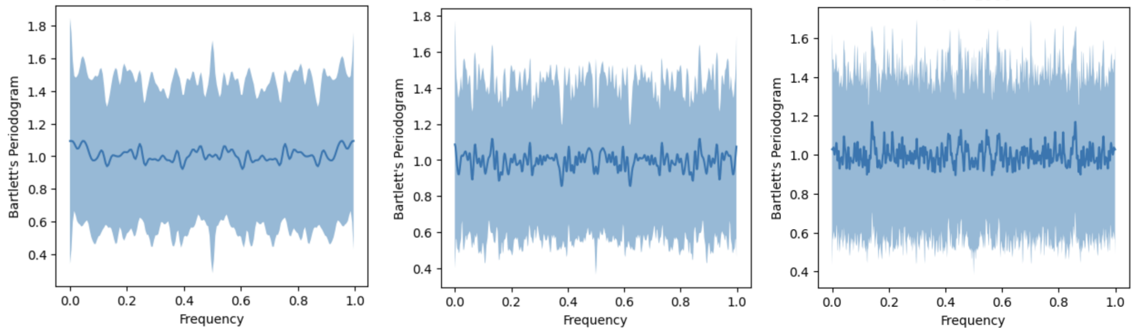
As seen in the previous question, the problem with the periodogram is the fact that its variance does not decrease with the sample size. A simple procedure to obtain a consistent estimate is to divide the signal in K sections of equal durations, compute a periodogram on each section and average them. 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

We notice that as compared to the results in question 6, the variance of the periodogram values obtained with Bartlett's estimate has been reduced by half, going from 1 to ~ 0.5.



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.

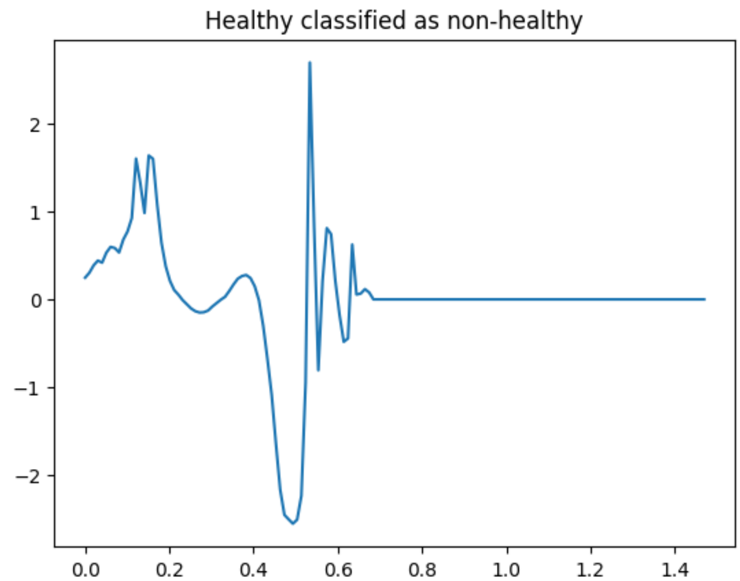
Answer 10

A Grid Search was applied to find the optimal number of neighbors. Using a 5-fold cross validation, the values of k tested ranged from 2 to 10 (the value 1 was not considered because it would create an overfitted model with low bias and high variance). The best value of k returned by the grid search is 5.
A 5-NN model was then created and fitted on the training set. The F-1 score for 5-NN model on the test set is 0.4504 which can be interpreted as the fact that the model makes two false predictions for every true prediction.
We notice that the F-1 score for 5-NN model on the training set is 0.8860, which means that our model overfitted considering the huge gap between the scores. We could consider increasing the number of training examples for future improvement.

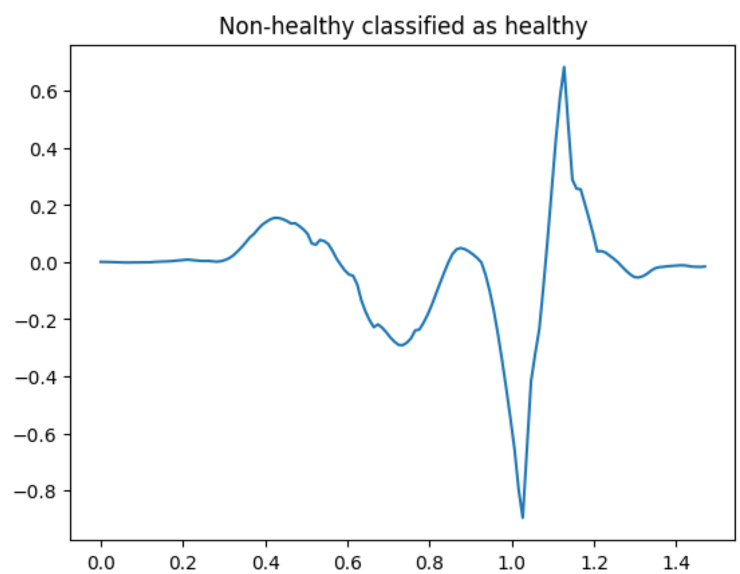
Question 11

Display on Figure 3 a badly classified step from each class (healthy/non-healthy).

Answer 11



Badly classified healthy step



Badly classified non-healthy step

Figure 3: Examples of badly classified steps (see Question 11).

References

Michael P. Knapp. Sines and cosines of angles in arithmetic progression. *Mathematics Magazine*, 82 (5):371–372, December 2009. doi: 10.4169/002557009x478436. URL https://doi.org/10.4169/002557009x478436.