Assignment 1 (ML for TS) - MVA

Oliver Jack olijacklu@gmail.com
Paulo Silva paulohenriquecrs@hotmail.com

October 28, 2024

1 Introduction

Objective. This assignment has three parts: questions about convolutional dictionary learning, 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 28th October 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: 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 $\mathbf{0}_p$ (a p-dimensional vector of zeros) for any $\lambda > \lambda_{\max}$.

Answer 1

First of all, as seen in the class of Convex Optimization, we know that Lasso regression is a convex problem. Thus, the KKT conditions are necessary and sufficient to solve our minimization problem. Since we are in the case of an unconstrained problem, this would simply come down to solving:

$$\mathbf{0}_p \in \nabla_\beta \left(\frac{1}{2} \|y - X\beta\|_2^2 \right) + \lambda \partial \|\beta\|_1$$

In this case, because the ℓ_1-norm is not differentiable at $\beta_j = 0$, we need to work with the subgradient of the ℓ_1-norm with respect to β.

The gradient of the least-squares term $\frac{1}{2}\|y - X\beta\|_2^2$ is:

$$\nabla_\beta \left(\frac{1}{2} \|y - X\beta\|_2^2 \right) = -X^T(y - X\beta)$$

The subgradient of the ℓ_1-norm term $\|\beta\|_1$ is:

$$\partial \|\beta_j\|_1 = \begin{cases} 1 & \text{if } \beta_j > 0, \\ -1 & \text{if } \beta_j < 0, \\ [-1, 1] & \text{if } \beta_j = 0. \end{cases}$$

Consider the case when $\beta = \mathbf{0}_p$. The previously mentioned KKT condition would then become:

$$-X^T y + \lambda z = 0$$

where $z_j \in [-1, 1]$ for all $j \in \{1, \dots, p\}$, the subgradient for the ℓ_1-norm

For $\beta = \mathbf{0}_p$ to be a solution, this equation must be satisfied for some $z_j \in [-1, 1]$. Hence, for all $j \in \{1, \dots, p\}$:

$$|X_j^T y| \leq \lambda$$

Since this should hold true for all $j \in \{1, \dots, p\}$, we can define:

$$\lambda_{\max} = \max_{j \in \{1, \dots, p\}} |X_j^T y|.$$

Finally, we can conclude that for any $\lambda > \lambda_{\max}$, the optimal solution to the Lasso regression is $\beta = \mathbf{0}_p$, since the subgradient condition is satisfied at $\beta = \mathbf{0}_p$, meaning that no $\beta_j \neq 0$ can minimize the objective function further.

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

For a fixed dictionary, the problem can be written as:

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

Let us define a new vector $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_k) \in \mathbb{R}^{K(N-L+1)}$ obtained by concatenating the K vectors $\mathbf{z}_k$. Moreover, each $\mathbf{z}_k * \mathbf{d}_k$ can be represented by a matrix multiplication $\mathbf{D}_k \mathbf{z}_k$, where

$$\mathbf{D}_k = \begin{bmatrix} \mathbf{d}_{k,1} & 0 & 0 & \cdots & 0 \\ \mathbf{d}_{k,2} & \mathbf{d}_{k,1} & 0 & \cdots & 0 \\ \mathbf{d}_{k,3} & \mathbf{d}_{k,2} & \mathbf{d}_{k,1} & \cdots & 0 \\ \mathbf{d}_{k,3} & \mathbf{d}_{k,2} & \mathbf{d}_{k,1} & \cdots & 0 \\ \vdots & \mathbf{d}_{k,3} & \mathbf{d}_{k,2} & \ddots & \vdots \\ \mathbf{d}_{k,L} & \vdots & \mathbf{d}_{k,3} & \ddots & 0 \\ 0 & \mathbf{d}_{k,L} & \vdots & \ddots & 0 \\ 0 & 0 & \mathbf{d}_{k,L} & \ddots & \mathbf{d}_{k,1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{d}_{k,L} \end{bmatrix} \in \mathbb{R}^{N \times (N-L+1)}$$

with $\mathbf{d}_{k,i}$ the i^{th} element of the vector $\mathbf{d}_k$.

This matrix satsifies the property:

$$\mathbf{z}_k * \mathbf{d}_k = \mathbf{D}_k \mathbf{z}_k$$

Next, let us define the matrix $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_K \end{bmatrix} \in \mathbb{R}^{N \times K(N-L+1)}$$

Then, by using our previously defined vector $\mathbf{z}_k$, we can represent the sum of convolutions as a single matrix multiplication:

$$\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k = \mathbf{D}\mathbf{z}$$

Finally, our minimization problem can be rewritten under the form:

$$\min_{\mathbf{z} \in \mathbb{R}^{K(N-L+1)}} \quad \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 \quad + \quad \lambda \|\mathbf{z}\|_1$$

This proves that for a fixed dictionary, the sparse coding problem is a Lasso regression with response vector $\mathbf{x}$ and design matrix $\mathbf{D}$.

By combining this result with *Question 1*, we can conclude that there exists a λ_{max} such that the sparse codes are only 0 for any $\lambda > \lambda_{max}$, since our problem can be reduced to a Lasso regression.

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 the sampling frequency by f_s, meaning that the index n corresponds to the time n/f_s. 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:

$$S(f) := \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2\pi f \tau / f_s}. \tag{3}$$

The power spectrum describes the distribution of power in the frequency space. Intuitively, large values of $S(f)$ indicate 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 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 number 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

As X_n is a Gaussian white noise (uncorrelated random variables), we have:

- When $\tau = 0$, $\gamma(\tau) = \mathbb{E}(X_n X_{n+0}) = \mathbb{E}(X_n^2) = \sigma^2$, which is the variance;
- When $\tau \neq 0$, $\gamma(\tau) = \mathbb{E}(X_n X_{n+\tau}) = 0$, since $X_n \neq X_{n+\tau}$ and the white noise process is uncorrelated.

As such, the autocovariance function has the values:

$$\gamma(\tau) = \begin{cases} \sigma^2 & \text{if } \tau = 0, \\ 0 & \text{if } \tau \neq 0. \end{cases}$$

Using this result, we can split the power spectrum equation into two terms, one for $\tau = 0$ and the other for $\tau \neq 0$:

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

As $\gamma(\tau) = 0$ for $\tau \neq 0$, the second term disappears and the power spectrum equation becomes:

$$S(f) = \gamma(0) e^{-2\pi i f \cdot 0 / f_s} = \gamma(0) = \sigma^2$$

Therefore, for all frequencies f, we have $S(f) = \sigma^2$.

Question 4

A natural estimator for the autocorrelation function is the sample autocovariance

$$\hat{\gamma}(\tau) := (1/N) \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \tag{4}$$

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

In order to show that $\hat{\gamma}(\tau)$ is a biased estimator of $\gamma(\tau)$, we need to show that $\text{Bias}(\hat{\gamma}(\tau)) = \mathbb{E}[\hat{\gamma}(\tau)] - \gamma(\tau) \neq 0$, that is, $\mathbb{E}[\hat{\gamma}(\tau)] \neq \gamma(\tau)$. We know that:

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

Therefore, we can calculate the expectation of $\hat{\gamma}(\tau)$:

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

As X_n is weakly stationary, the autocovariance function $\mathbb{E}[X_n X_{n+\tau}]$ does not depend on n, only on τ, which allows us to replace $\mathbb{E}[X_n X_{n+\tau}]$ by $\gamma(\tau)$ and obtain:

$$\mathbb{E}[\hat{\gamma}(\tau)] = \frac{1}{N} \sum_{n=0}^{N-\tau-1} \gamma(\tau) = \frac{N-\tau}{N}\gamma(\tau)$$

Therefore, we can calculate the bias:

$$\text{Bias}(\hat{\gamma}(\tau)) = \mathbb{E}[\hat{\gamma}(\tau)] - \gamma(\tau) = \frac{N-\tau}{N}\gamma(\tau) - \gamma(\tau) = -\frac{\tau}{N}\gamma(\tau)$$

Since $\frac{\tau}{N} \neq 0$ for $\tau > 0$, the estimator $\hat{\gamma}(\tau)$ is biased as the bias is not zero for any nonzero τ.

Now, to show that the estimator $\hat{\gamma}(\tau)$ is asymptotically unbiased, we calculate the bias when $N \to \infty$. As we've shown just above, we know that $\mathbb{E}[\hat{\gamma}(\tau)] = \frac{N-\tau}{N}\gamma(\tau)$. When we apply $N \to \infty$, we obtain:

$$\lim_{N\to\infty} \mathbb{E}[\hat{\gamma}(\tau)] = \lim_{N\to\infty} \frac{N-\tau}{N}\gamma(\tau) = \lim_{N\to\infty} \left(1 - \frac{\tau}{N}\right)\gamma(\tau) = \gamma(\tau)$$

Finally, we can replace this in the bias equation:

$$\lim_{N\to\infty} \text{Bias}(\hat{\gamma}(\tau)) = \lim_{N\to\infty} \mathbb{E}[\hat{\gamma}(\tau)] - \gamma(\tau) = \gamma(\tau) - \gamma(\tau) = 0$$

This shows that the estimator $\hat{\gamma}(\tau)$ is asymptotically unbiased.

Finally, a very simple way to de-bias this estimator $\hat{\gamma}(\tau)$ would be by multiplying it by $\frac{N}{N-\tau}$, since it would cancel the term $\frac{N-\tau}{N}$ in the expectation of the estimator (as $\mathbb{E}[\hat{\gamma}(\tau)] = \frac{N-\tau}{N}\gamma(\tau)$). Therefore, a new unbiased estimator would be:

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

Question 5

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

$$J(f) := (1/\sqrt{N}) \sum_{n=0}^{N-1} X_n e^{-2\pi i f n / f_s} \tag{5}$$

The *periodogram* is the collection of values $|J(f_0)|^2, |J(f_1)|^2, \dots, |J(f_{N/2})|^2$ 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

$$\begin{aligned}
|J(f_k)|^2 &= J(f_k)\overline{J(f_k)} \\
&= \left(\frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} X_n e^{-2\pi i \frac{nk}{N}} \right) \left(\frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} X_m e^{2\pi i \frac{mk}{N}} \right) \\
&= \frac{1}{N} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} X_n X_m e^{(m-n)\frac{2\pi i k}{N}}
\end{aligned}$$

Consider $\tau = m - n \iff m = n + \tau$. Then:

$$|J(f_k)|^2 = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{\tau=-n}^{N-n-1} X_n X_{n+\tau} e^{\tau \frac{2\pi i k}{N}}$$

To change the order of the sums, we first need to find the range of τ. We can see that τ ranges from $-N+1$ (when $n = N-1$) to $N-1$ (when $n = 0$). Meanwhile, for a fixed τ, we know that n has to verify $-n \le \tau$ and $\tau \le N - n - 1$, which equates to $-\tau \le n$ and $n \le N - \tau - 1$. Furthermore, we know that $0 \le n \le N-1$. Thus:

$$\begin{aligned}
|J(f_k)|^2 &= \sum_{\tau=-N+1}^{N-1} e^{\tau \frac{2\pi i k}{N}} \left(\frac{1}{N} \sum_{n=max(0,-\tau)}^{min(N-1, N-\tau-1)} X_n X_{n+\tau} \right) \\
&= \sum_{\tau=0}^{N-1} e^{\tau \frac{2\pi i k}{N}} \left(\frac{1}{N} \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \right) + \sum_{\tau=-N+1}^{-1} e^{\tau \frac{2\pi i k}{N}} \left(\frac{1}{N} \sum_{n=-\tau}^{N-1} X_n X_{n+\tau} \right) \\
&= \sum_{\tau=0}^{N-1} e^{\tau \frac{2\pi i k}{N}} \left(\frac{1}{N} \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \right) + \sum_{\tau=-N+1}^{-1} e^{\tau \frac{2\pi i k}{N}} \left(\frac{1}{N} \sum_{n=0}^{N-\tau-1} X_n X_{n-\tau} \right) \\
&= \sum_{\tau=0}^{N-1} e^{\tau \frac{2\pi i k}{N}} \hat{\gamma}(\tau) + \sum_{\tau=-N+1}^{-1} e^{\tau \frac{2\pi i k}{N}} \hat{\gamma}(-\tau) \\
&= \sum_{\tau=0}^{N-1} e^{\tau \frac{2\pi i k}{N}} \hat{\gamma}(\tau) + \sum_{\tau=1}^{N-1} e^{-\tau \frac{2\pi i k}{N}} \hat{\gamma}(\tau) \\
&= \hat{\gamma}(0) + 2 \sum_{\tau=1}^{N-1} \cos\left(\frac{2\pi k \tau}{N} \right) \hat{\gamma}(\tau)
\end{aligned}$$

For the second part of the question, let us assume that $f^{(N)} = f_k$ for some k. Then we can proceed in the following way:

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

For τ fixed, we have that:

$$\lim_{N\to\infty} \frac{N-\tau}{N} = 1$$

By definition, $f^{(N)}$ is the closest Fourier frequency to f. Thus, there exists k such that:

$$f^{(N)} = \frac{f_s k}{N}$$

As N goes to infinity, the spacing between two successive Fourier frequencies f_k and f_{k+1}, which is $\Delta f = \frac{f_s}{N}$, tends to 0. Therefore:

$$\lim_{N\to\infty} f^{(N)} = f.$$

Hence, by making N tend to infinity:

$$\lim_{N\to\infty} \mathbb{E}[|J(f^{(N)})|^2] = \sum_{\tau=-\infty}^{\infty} e^{-\tau 2\pi i \frac{f}{f_s}} \gamma(\tau) = S(f)$$

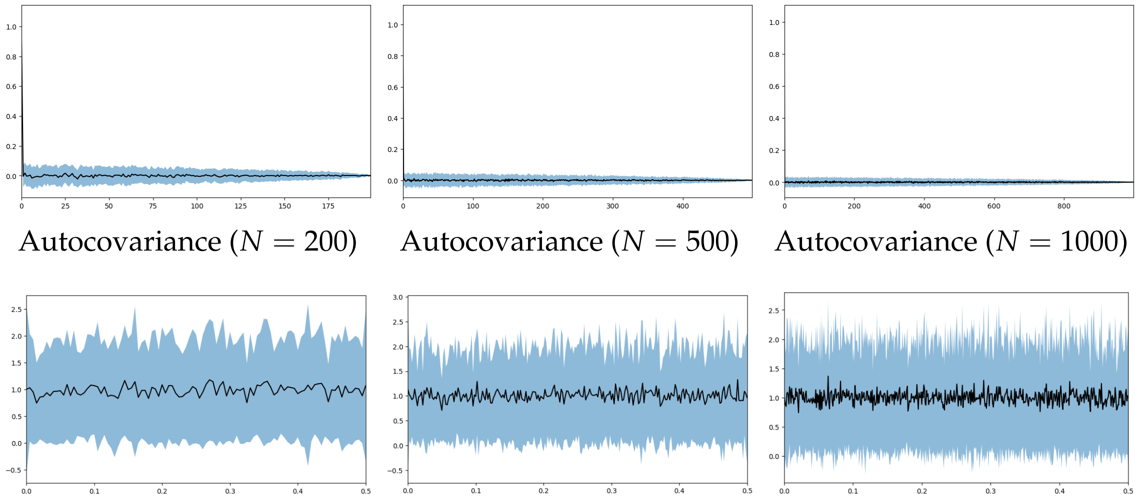
Finally, we have shown that $|J(f^{(N)})|^2$ is an asymptotically unbiased estimator of $S(f)$ for $f > 0$.

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 ±, 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 ±, 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

For Gaussian white noise, the average sample autocovariance $\hat{\gamma}(\tau)$ is has its highest values at $\tau = 0$ and quickly drops to near zero for time lags greater than 0. This is expected since the white noise X_n is weakly stationary, made of uncorrelated random variables. This behavior is similar in all the different values of N tried (200, 500, 1000). The variance decreases as N increases, leading to a tighter spread around zero for larger N, which shows that the estimator is more stable with larger sample sizes.

The periodograms are quite flat across the entire x-axis, which makes sense since Gaussian white noise has equal power across all the frequencies. As N increases, variance remains in the same range (in this case, between 0 and 2), which allows us to conclude that its variance does not decrease with the sample size.

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 $\tau > 0$

$$\operatorname{var}(\hat{\gamma}(\tau)) = (1/N) \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(1 - \frac{\tau + |n|}{N}\right) \left[\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau)\right]. \tag{6}$$

  (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

As shown in *Question 4*, since X_n is a wide-sense stationary Gaussian process, the autocovariance function $\mathbb{E}[X_n X_{n+\tau}]$ does not depend on n, only on τ, which allows us to replace $\mathbb{E}[X_n X_{n+\tau}]$ by $\gamma(\tau)$ and obtain:

$$\mathbb{E}[\hat{\gamma}(\tau)] = \frac{1}{N}\sum_{n=0}^{N-\tau-1}\gamma(\tau) = \frac{N-\tau}{N}\gamma(\tau)$$

Next, let us assume that the $\mu = \mathbb{E}[X_n] = 0$ (if not, it suffices to consider $X_n = Y_n + \mu$ where Y_n is a centered Gaussian process). Then:

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

Since we are in the case where we have four centered jointly Gaussian variables, we can use the *Hint* to decompose the expectation:

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

Finally, we can compute the variance of $\hat{\gamma}(\tau)$:

$$\begin{aligned}
var(\hat{\gamma}(\tau)) &= \mathbb{E}[(\hat{\gamma}(\tau))^2] - \mathbb{E}[\hat{\gamma}(\tau)]^2 \\
&= \frac{(N-\tau)^2}{N^2}\gamma(\tau)^2 + \frac{1}{N}\sum_{k=-(N-\tau-1)}^{N-\tau-1}(1 - \frac{\tau+|k|}{N})(\gamma(k)^2 + \gamma(k-\tau)\gamma(k+\tau)) - \frac{(N-\tau)^2}{N^2}\gamma(\tau)^2 \\
&= \frac{1}{N}\sum_{k=-(N-\tau-1)}^{N-\tau-1}(1 - \frac{\tau+|k|}{N})(\gamma(k)^2 + \gamma(k-\tau)\gamma(k+\tau))
\end{aligned}$$

To prove consistency, let us first show that $var(\hat{\gamma}(\tau))$ converges to 0 as N tends to infinity. First of all, for sufficiently large N, we have that for all $k = -(N-\tau-1), \dots, N-\tau-1$:

$$1 - \frac{\tau+|k|}{N} \leq 1$$

By applying the Cauchy-Schwarz inequality, we obtain:

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

Since the autocovariances are square summable, we know there exists a constant C such that $var(\hat{\gamma}(\tau)) \leq \frac{C}{N}$, which converges to 0 as $N \to \infty$. Thus, $var(\hat{\gamma}(\tau))$ converges to 0 as $N \to \infty$.

By applying Markov's inequality, we get the following result $\forall \epsilon > 0$:

$$\begin{aligned}
\mathbb{P}(|\hat{\gamma}(\tau)-\gamma(\tau)| > \epsilon) &= \mathbb{P}((\hat{\gamma}(\tau)-\gamma(\tau))^2 \geq \epsilon^2) \\
&\leq \frac{\mathbb{E}[(\hat{\gamma}(\tau)-\gamma(\tau))^2]}{\epsilon^2} \\
&= \frac{(\mathbb{E}[\hat{\gamma}(\tau)]-\gamma(\tau))^2+var(\hat{\gamma}(\tau))}{\epsilon^2}
\end{aligned}$$

Since the sample estimator $\hat{\gamma}(\tau)$ is an asymptotically unbiased estimator of $\gamma(\tau)$ and since $var(\hat{\gamma}(\tau) \to 0$ as $N \to \infty$, we can conclude that

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

i.e. $\hat{\gamma}(\tau)$ is consistent.

Question 8

Assume that X is a Gaussian white noise (variance σ^2) and let $A(f) := \sum_{n=0}^{N-1} X_n \cos(-2\pi f n/f_s)$ and $B(f) := \sum_{n=0}^{N-1} X_n \sin(-2\pi f n/f_s)$. 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 $|J(f_0)|^2$, $|J(f_1)|^2$, . . . , $|J(f_{N/2})|^2$.
- 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

Since X is a Gaussian white noise, we have that $\mathbb{E}[X_n] = 0$ and $Cov(X_n, X_m) = 0, \forall n \neq m$. Thus:

$$\mathbb{E}[A(f)] = \sum_{n=0}^{N-1} \mathbb{E}[X_n] \cos\left(\frac{-2\pi f n}{f_s}\right) = 0$$

$$\mathbb{E}[B(f)] = \sum_{n=0}^{N-1} \mathbb{E}[X_n] \sin\left(\frac{-2\pi f n}{f_s}\right) = 0$$

$$var(A(f)) = var\left(\sum_{n=0}^{N-1} X_n \cos\left(\frac{-2\pi f n}{f_s}\right)\right) = \sum_{n=0}^{N-1} var(X_n)\cos^2\left(\frac{-2\pi f n}{f_s}\right) = \sigma^2 \sum_{n=0}^{N-1}\cos^2\left(\frac{-2\pi k n}{N}\right)$$

$$var(B(f)) = var\left(\sum_{n=0}^{N-1} X_n \sin\left(\frac{-2\pi f n}{f_s}\right)\right) = \sum_{n=0}^{N-1} var(X_n)\sin^2\left(\frac{-2\pi f n}{f_s}\right) = \sigma^2 \sum_{n=0}^{N-1}\sin^2\left(\frac{-2\pi k n}{N}\right)$$

Let us distinguish two different cases, when $k = 0$ or $k = N/2$ and if k is any other value.

Case 1: $k = 0$ **or** $k = N/2$

In this case, $\cos^2\left(\frac{-2\pi kn}{N}\right) = 1$ for all n, so $\sum_{n=0}^{N-1}\cos^2\left(\frac{-2\pi kn}{N}\right) = N$. Meanwhile, $\sin^2\left(\frac{-2\pi kn}{N}\right) = 0$ for all n, so $\sum_{n=0}^{N-1}\sin^2\left(\frac{-2\pi kn}{N}\right) = 0$. Hence:

$$var(A(f)) = N\sigma^2, \quad var(B(f)) = 0$$

In this case, we get:

$$|J(f_k)|^2 = \frac{1}{N}A^2(f_k) = \frac{N\sigma^2}{N}\chi_1^2 = \sigma^2\chi_1^2$$

where χ_1^2 is a Chi-squared random variable with 1 degree of freedom. Furthermore:

$$var(|J(f_k)|^2) = var(\sigma^2\chi_1^2) = \sigma^4 var(\chi_1^2) = 2\sigma^4$$

Case 2: $k \neq 0$ **and** $k \neq N/2$

Let us show that $\sum_{n=0}^{N-1}\cos^2\left(\frac{-2\pi kn}{N}\right) = \sum_{n=0}^{N-1}\sin^2\left(\frac{-2\pi kn}{N}\right) = \frac{N}{2}$. Let us set $\theta_n = \frac{-2\pi kn}{N}$ and use the following trigonometric formulas:

$$\cos^2(\theta_n) = \frac{1+\cos(2\theta_n)}{2}, \quad \sin^2(\theta_n) = \frac{1-\cos(2\theta_n)}{2}$$

Then summing over n gives us:

$$\sum_{n=0}^{N-1}\cos^2(\theta_n) = \sum_{n=0}^{N-1}\frac{1+\cos(2\theta_n)}{2} = \frac{1}{2}\sum_{n=0}^{N-1}1 + \frac{1}{2}\sum_{n=0}^{N-1}\cos(2\theta_n) = \frac{N}{2} + \frac{1}{2}\sum_{n=0}^{N-1}\cos(2\theta_n)$$

We know that $2\theta_n$ will increment by an even multiple of $\frac{2\pi}{N}$, i.e. the sequence of angles $2\theta_n$ will wrap around the unit circle uniformly, completing an integer number of full periods.

Due to the periodicity and symmetry of the cosine function, the sum of $\cos(2\theta_n)$ over a complete period is 0. Thus:

$$\sum_{n=0}^{N-1}\cos(2\theta_n) = 0$$

This proves that $\sum_{n=0}^{N-1}\cos^2\left(\frac{-2\pi kn}{N}\right) = \frac{N}{2}$.

Similarly:

$$\sum_{n=0}^{N-1}\sin^2(\theta_n) = \frac{N}{2} - \frac{1}{2}\sum_{n=0}^{N-1}\cos(2\theta_n) = \frac{N}{2}$$

Finally:

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

Next, let us show that $A(f)$ and $B(f)$ are independent. Since they are jointly Gaussian (linear combinations of a set of independent Gaussian variables), this is equivalent to showing that $Cov(A(f), B(f)) = 0$:

$$\begin{aligned}
Cov(A(f), B(f)) &= \mathbb{E}[A(f)B(f)] - \mathbb{E}[A(f))]\mathbb{E}[B(f)] \\
&= \mathbb{E}[A(f)B(f)] \\
&= \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} X_n X_m \cos\left(\frac{-2\pi kn}{N}\right)\sin\left(\frac{-2\pi km}{N}\right)\right] \\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1} \mathbb{E}[X_n X_m] \cos\left(\frac{-2\pi kn}{N}\right)\sin\left(\frac{-2\pi km}{N}\right) \\
&= \sigma^2 \sum_{n=0}^{N-1} \cos\left(\frac{-2\pi kn}{N}\right)\sin\left(\frac{-2\pi kn}{N}\right)
\end{aligned}$$

Here we can use the trigonometric identity $\cos(x)\sin(x) = \frac{1}{2}\sin(2x)$, which gives us:

$$Cov(A(f), B(f)) = \sigma^2 \frac{1}{2} \sum_{n=0}^{N-1} \sin\left(\frac{-4\pi kn}{N}\right)$$

By the same reasoning as before, the sum is equal to 0, which implies that $A(f)$ and $B(f)$ are uncorrelated, hence independent.

Since $\frac{\sqrt{2}A(f)}{\sqrt{N}\sigma}$ and $\frac{\sqrt{2}B(f)}{\sqrt{N}\sigma}$ are two independent standard Gaussian random variables, the distribution of $|J(f_k)|^2$ is given by:

$$|J(f_k)|^2 = \frac{1}{N}(A^2(f_k) + B^2(f_k)) = \frac{N\sigma^2}{2N}(\chi_1^2 + \chi_1^2) = \frac{\sigma^2}{2}\chi_2^2$$

Furthermore:

$$var(|J(f_k)|^2) = var\left(\frac{\sigma^2}{2}\chi_2^2\right) = \frac{\sigma^4}{4} var(\chi_2^2) = \sigma^4$$

Since for all k the variance of the periodogram doesn't decrease to 0 as $N \to \infty$, we can conclude that it is not consistent.

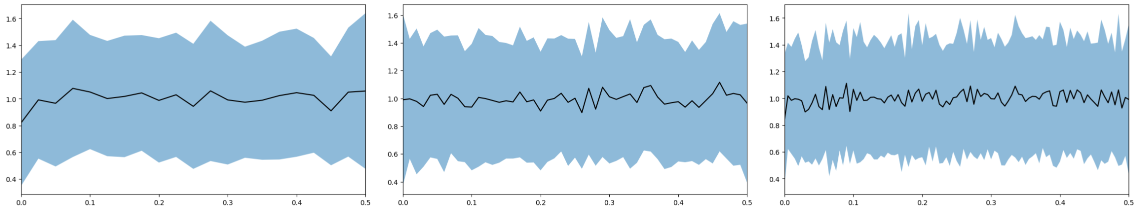
The erratic behaviour of the periodogram can be explained by the fact that the $|J(f_k)^2|$ values are not correlated ($Cov(|J(f_k)|^2, |J(f_{k'})|^2) \neq 0$), which implies that they are not independent. Due to the overlap in frequency components, the resulting dependencies can lead to large fluctuations in 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 into 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.



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

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

Answer 9

In the Bartlett periodograms, the curves are much smoother than the original periodograms, and the variance is reduced because we averaged over multiple sections (K). This effect is especially clear for larger N values (e.g. 1000), where the Bartlett method gives a more stable estimate.

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 falls. 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 and 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

As seen on Figure 3 below, the optimal number of neighbors using the combination of DTW and a k-neighbors classifier was 6, with a mean F1-score of 0.731 for the 5-fold cross validation. There is a stabilization in the F1-score below 0.7 for 8 neighbors and 10 neighbors, indicating that an increase in the number of neighbors will not improve the model. The standard deviation also appears to stabilize around 0.05 with 4 or more neighbors, in contrast to the higher variability observed with only 2 neighbors, where it reaches 0.14.

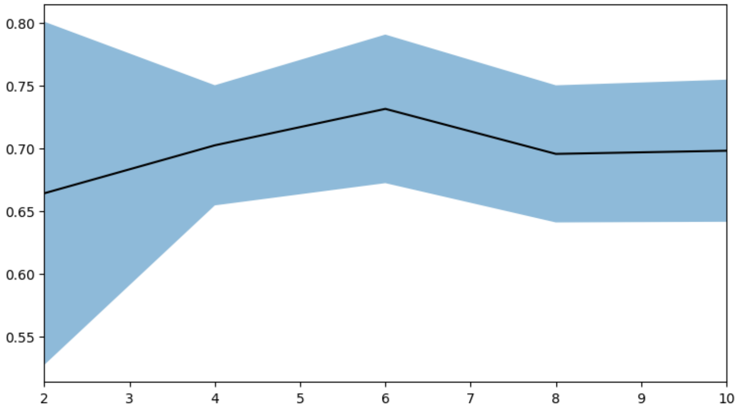
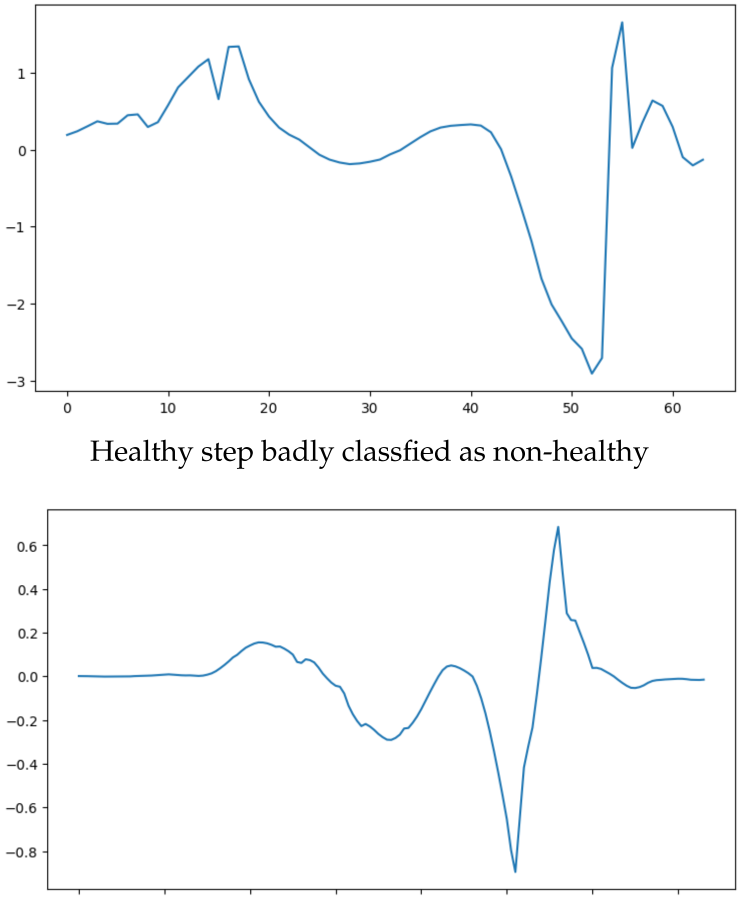


Figure 3: Optimal number of neighbors with 5-fold cross-validation.

Question 11

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

Answer 11



Healthy step badly classfied as non-healthy

Non-healthy step badly classfied as healthy

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