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

MARENGO Matteo matteo.marengo@ens-paris-saclay.fr
ROBERT Hugo robert.hugo@ens-paris-saclay.fr

December 17, 2023

Part I

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 7th 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.

Part II

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

The first term $\frac{1}{2}\|y - X\beta\|_2^2$ is the residual sum of squares that measures the fit of the model to the data. The second term $\lambda\|\beta\|_1$ is the L_1 penalty on the parameters.

Step 1: Define the Optimality Conditions for the Lasso regression

The optimality conditions for the Lasso objective function can be derived by setting the gradient of the objective function with respect to β equal to zero. The gradient of the Lasso regression is:

$$\nabla_\beta \left(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right) = -X^T(y - X\beta) + \lambda \operatorname{sgn}(\beta)$$

To understand the computation totally it is fully explained in Appendix 1.

When we set the gradient equal to zero, we obtain the optimality conditions for the Lasso problem:

$$-X^T(y - X\beta) + \lambda \operatorname{sgn}(\beta) = 0$$

Step 2: Introduce $\beta = \mathbf{0}_p$ in the optimality conditions

If $\beta = \mathbf{0}_p$, the optimality conditions become:

$$-X^T y + \lambda \operatorname{sgn}(\mathbf{0}_p) = 0$$

Since the sign of a zero vector is undefined, we need to consider that the subdifferential of the L_1 norm at $\beta = \mathbf{0}_p$ is the set of all vectors z such that $-1 \leq z_i \leq 1$ for all i [Ref : S. Boyd and L. Vandenberghe, Notes for EE364b, Stanford University]. Therefore, the optimality conditions for $\beta = \mathbf{0}_p$ is written as:

$$-X^T y = -\lambda \, [-1, 1]^p$$

This means that each element of the vector $X^T y$ must be in the range $[-\lambda, \lambda]$.

Step 3: Find $\lambda_{\max}$

To find $\lambda_{\max}$, we want to find the smallest value of λ such that the vector $\mathbf{0}_p$ is a solution to the Lasso regression. This result is obtained when the maximum absolute value of the elements in $X^T y$ is less than or equal to λ:

$$\max_i |(X^T y)_i| \leq \lambda$$

Therefore:

$$\lambda_{\max} = \max_i |(X^T y)_i| = \|X^T y\|_\infty$$

For any $\lambda > \lambda_{\max}$, the vector $\mathbf{0}_p$ is a solution to the Lasso regression, because each element of $X^T y$ will be between $[-\lambda, \lambda]$, and as explained in the second step it satisfies the optimality conditions for $\beta = \mathbf{0}_p$.

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

1) Sparse coding as lasso regression

For a fixed dictionary, we have the set of dictionary atoms $(\mathbf{d}_k)_k$ as constant. In this case, our goal is to find the activation signals $(\mathbf{z}_k)_k$ that minimize the following objective function:

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

First, let's consider the **convolution operation**. We can then rewrite it (more details given in Appendix 2) as:

$$\left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{D}_k \mathbf{z}_k \right\|_2^2$$

where $\mathbf{D}_k$ is the matrix constructed from the dictionary atom $\mathbf{d}_k$ as described in the Appendix 2. Then, the matrices $\mathbf{D}_k$ and the activation signals matrices $\mathbf{z}_k$ follow this organisation:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 & \dots & \mathbf{D}_K \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_K \end{bmatrix}$$

It allows us to rewrite the first term of the objective function as (we find indeed the same results that were presented in the slides):

$$\|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2$$

Similarly, we can rewrite the second term of the objective function as:

$$\lambda \|\mathbf{z}\|_1$$

Therefore, the sparse coding problem can be formulated as a lasso regression problem:

$$\min_{\mathbf{z}} \quad \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2^2 \quad + \quad \lambda \|\mathbf{z}\|_1$$

2) Existence of $\lambda_{\max}$

For a fixed dictionary, we want to show that there exists a maximum value of λ; $\lambda_{\max}$, so the solution for any $\lambda > \lambda_{\max}$ to the optimization problem isis $\mathbf{z}_k = 0$ for all k. Given the lasso regression formulation and using question 1, there is the existence of a $\lambda_{\max}$, such that for any $\lambda > \lambda_{\max}$, the solution to the sparse coding problem is $\mathbf{z}_k = 0$ for all k, where:

$$\lambda_{\max} = \left\| 2\mathbf{D}^T \mathbf{x} \right\|_\infty$$

Part III

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:

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

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

Step 1: Autocovariance function

For a discrete-time stochastic process X_n, the autocovariance function is defined as:

$$R(\tau) = \mathrm{E}[(X_n - \mu_n)(X_{n+\tau} - \mu_{n+\tau})]$$

where μ_n is the mean of X_n, and τ is the lag. For a Gaussian white noise process, we have $\mu_n = 0$ for all n, and the process is uncorrelated at different time points. Therefore, the autocovariance function becomes:

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

where σ^2 is the variance of the process, further details are given in Appendix 3.

Step 2: Power Spectrum

The power spectrum $S(f)$ of a discrete-time process X_n is the Fourier transform of its autocovariance function $R(\tau)$:

$$S(f) = \sum_{\tau=-\infty}^{\infty} R(\tau) e^{-j2\pi f\tau}$$

Given the autocovariance function we calculated earlier, we can calculate the power spectrum:

$$S(f) = \sum_{k=-\infty}^{\infty} \begin{cases} \sigma^2 & \text{if } k = 0, \\ 0 & \text{if } k \neq 0, \end{cases} e^{-j2\pi fk} = \sigma^2$$

To conclude, the power spectrum of a Gaussian white noise process is constant and equal to its variance σ^2 across all frequencies f. This is why the process is called "white" noise, as it contains all frequencies with equal intensity, similar to white light which contains all colors of the spectrum with equal intensity.

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

Bias of $\hat{\gamma}(\tau)$:

The bias of an estimator $\hat{\theta}$ of a parameter θ is defined as:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Therefore, the bias of $\hat{\gamma}(\tau)$ is (calculations detailed in Appendix 4):

$$\text{Bias}(\hat{\gamma}(\tau)) = \frac{N-\tau}{N}\gamma(\tau) - \gamma(\tau) = -\frac{\tau}{N}\gamma(\tau)$$

Since $-\frac{\tau}{N}\gamma(\tau) \neq 0$ in general, $\hat{\gamma}(\tau)$ is a biased estimator of $\gamma(\tau)$.

Step 2: Asymptotic unbiasedness of $\hat{\gamma}(\tau)$:

To be asymptotically unbiased means that the bias of the estimator goes to zero as the sample size N goes to infinity. In this case, as $N \to \infty$, we have:

$$\text{Bias}(\hat{\gamma}(\tau)) = -\frac{\tau}{N}\gamma(\tau) \to 0$$

Therefore, $\hat{\gamma}(\tau)$ is **asymptotically unbiased**.

Step 3: A simple way to de-bias the estimator:

A simple way to de-bias the estimator $\hat{\gamma}(\tau)$ is to multiply it by a **correction factor** $\frac{N}{N-\tau}$ to counteract the bias:

$$\tilde{\gamma}(\tau) = \frac{N}{N-\tau}\hat{\gamma}(\tau)$$

This will make $\tilde{\gamma}(\tau)$ an unbiased estimator of $\gamma(\tau)$, as we can verify:

$$\mathbb{E}(\tilde{\gamma}(\tau)) = \frac{N}{N-\tau}\mathbb{E}(\hat{\gamma}(\tau)) = \frac{N}{N-\tau}\left(\frac{N-\tau}{N}\right)\gamma(\tau) = \gamma(\tau)$$

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 $|J(f_0)|^2, |J(f_1)|^2, \ldots, |J(f_{N/2})|^2$ 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

Step 1: Write $|J(f_k)|^2$ as a function of the sample autocovariances

The discrete Fourier transform at frequency f_k is defined as:

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

We want to find the expression for $|J(f_k)|^2$, details are given in Appendix 5:

$$|J(f_k)|^2 = \hat{\gamma}(0) + \sum_{\tau=1}^{N-1}\left(\hat{\gamma}(\tau)e^{-\frac{2\pi i k\tau}{N}} + \hat{\gamma}(-\tau)e^{\frac{2\pi i k\tau}{N}}\right)$$

Show that $|J(f^{(N)})|^2$ is an asymptotically unbiased estimator of $S(f)$ for $f > 0$

The relation between $J(f^{(N)})$ and the autocovariances is given by:

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

Now, we want to show that as $N \to \infty$, $|J(f^{(N)})|^2$ converges in expectation to $S(f)$, which is given by:

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

Taking the expectation of $|J(f^{(N)})|^2$, we have:

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

Since $\hat{\gamma}(\tau)$ is an asymptotically unbiased estimator of $\gamma(\tau)$, we have $\mathbb{E}[\hat{\gamma}(\tau)] \to \gamma(\tau)$ as $N \to \infty$. Therefore, as $N \to \infty$, the above expression converges to:

$$\sum_{\tau=-(N-1)}^{N-1} \gamma(\tau) e^{-2\pi i f^{(N)} \tau / f_s}$$

Now, we can observe that the above expression converges to $S(f)$ as $N \to \infty$. $f^{(N)}$ tends to $f > 0$ and as $N \to \infty$, the sum approaches the infinite sum that defines $S(f)$:

$$\sum_{\tau=-(N-1)}^{N-1} \gamma(\tau) e^{-2\pi i f^{(N)} \tau / f_s} \to \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-2\pi i f \tau / f_s} = S(f)$$

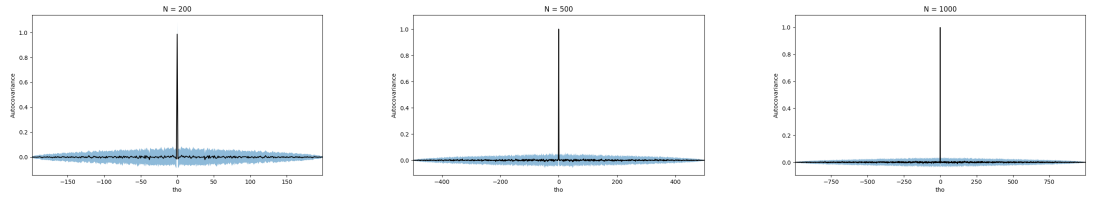
This shows that $|J(f^{(N)})|^2$ is an asymptotically unbiased estimator of $S(f)$ for $f > 0$.

Question 6

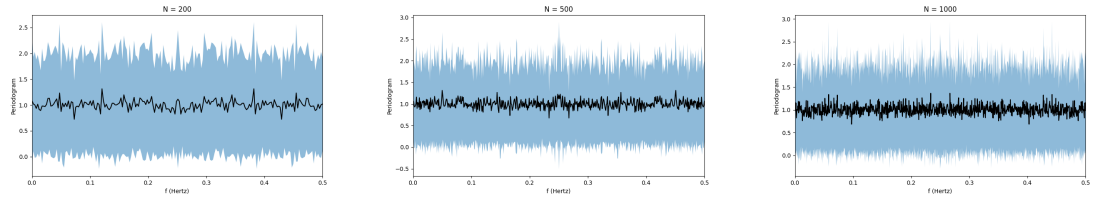
In this question, let X_n $(n = 0, \ldots, N-1)$ be a Gaussian white noise with variance $\sigma^2 = 1$ and set the sampling frequency to $f_s = 1$ Hz

- 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 $\pm$ 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

Autocovariances

As shown in the figures above, and as we said in question 3, the standard deviation of the autocovariance function decreases with the sample size N. Furthermore, the autocovariance function equals

$$\begin{cases} \sigma^2 & \text{if } \tau = 0 \\ 0 & \text{if } \tau \neq 0 \end{cases} \text{ with } \sigma = 1$$

This result shows that for a lag of 0, the correlation is strong, i.e. there is a strong association between the points in the time series. Then, as the lag is increased, this correlation strongly decreases, which shows that the influence of past/future values has no impact on the present value. This result is characteristic of Gaussian white noise.

Periodogram

As shown in the figures above, the periodogram is a constant according to f_s which is logical for white Gaussian noise. Furthermore, the variance does not decrease with the sample size N.

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 $\tau > 0$

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

(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

Step 1: Compute the variance of the estimator $\hat{\gamma}(\tau)$

We start by writing down the definition of the variance:

$$\text{var}(\hat{\gamma}(\tau)) = \mathbb{E}[(\hat{\gamma}(\tau) - \gamma(\tau))^2].$$

After some calculations developed in Appendix 6 we obtain :

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

Step 2: Show that $\hat{\gamma}(\tau)$ is consistent

To show that $\hat{\gamma}(\tau)$ is a consistent estimator of $\gamma(\tau)$, we need to show that as $N \to \infty$, $\hat{\gamma}(\tau)$ converges in probability to $\gamma(\tau)$. This is equivalent to show that:

$$\lim_{N\to\infty} \text{var}(\hat{\gamma}(\tau)) = 0.$$

Indeed, we can use the Bienaymé-Tchebychev inequality:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

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

Given the expression for $\text{var}(\hat{\gamma}(\tau))$ derived in Step 1, and under the assumption that the autocovariances $\gamma(n)$ are absolutely summable, we have:

$$\lim_{N\to\infty} \text{var}(\hat{\gamma}(\tau)) = 0.$$

This shows that $\hat{\gamma}(\tau)$ is a consistent estimator of $\gamma(\tau)$ as it converges in probability.

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 $J(f) = (1/N)(A(f) + iB(f))$.

- 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

1. **Mean and variance of $A(f)$ and $B(f)$:**

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

Similarly, for $B(f)$:

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

The variance of $A(f)$ is:

$$\begin{aligned}\text{var}(A(f)) &= \mathbb{E}(A(f)^2) - (\mathbb{E}(A(f)))^2 \\ &= \mathbb{E}\left((\sum_{n=0}^{N-1} X_n \cos\left(-\frac{2\pi fn}{f_s}\right))^2\right) \\ &= \sum_{n=0}^{N-1} \mathbb{E}(X_n^2) \cos^2\left(-\frac{2\pi fn}{f_s}\right) \quad \text{as Xn and Xm are not correlated} \\ &= \sigma^2 \sum_{n=0}^{N-1} \cos^2\left(-\frac{2\pi fn}{f_s}\right) \\ &= \frac{N\sigma^2}{2}.\end{aligned}$$

Similarly, the variance of $B(f)$ is:

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

2. **Distribution of the periodogram values:**

   Since X is Gaussian, and $A(f)$ and $B(f)$ are linear combinations of X, they are also Gaussian. Therefore, the periodogram values $|J(f_k)|^2$ are distributed as the sum of the squares of two independent normal random variables, which follows a chi-squared distribution with 2 degrees of freedom.

3. **Variance of $|J(f_k)|^2$:**

   The variance of $|J(f_k)|^2$ is the variance of a chi-squared distribution with 2 degrees of freedom, which is $\text{Var}(|J(f_k)|^2) = 2 \cdot k \cdot \frac{\sigma^2 \cdot N}{2}$. As k=2, $\text{Var}(|J(f_k)|^2) = 2 \cdot \sigma^2 \cdot N$

   The variance of $|J(f_k)|^2$ diverges when N is infinite, it explains that the periodogram is not consistent.

4. **Explanation of the erratic behavior of the periodogram:**

   The erratic behavior of the periodogram can be explained by looking at the covariance between the $|J(f_k)|^2$. Since the X_n are independent, the $|J(f_k)|^2$ are also independent for different f_k so the covariance is equal to zero.

   Indeed:

   $$\text{Cov}(|J(f_k)|^2, |J(f_j)|^2) = \frac{1}{N^4} \cdot \text{Cov}(A(f_k)^2 + B(f_k)^2, A(f_j)^2 + B(f_j)^2)$$

   With the bilinearity of the covariance :

   $$\text{Cov}(|J(f_k)|^2, |J(f_j)|^2) = \frac{1}{N^4} \cdot (\text{Cov}(A(f_k)^2, A(f_j)^2) + \text{Cov}(B(f_k)^2, A(f_j)^2) + \text{Cov}(B(f_k)^2, B(f_j)^2) + \text{Cov}(A(f_k)^2, B(f_j)^2))$$

   As there are only different frequencies:

   $$\text{Cov}(|J(f_k)|^2, |J(f_j)|^2) = 0$$

   It then means that the periodogram values can fluctuate widely from one frequency to another this is why there is an erratic behavior.

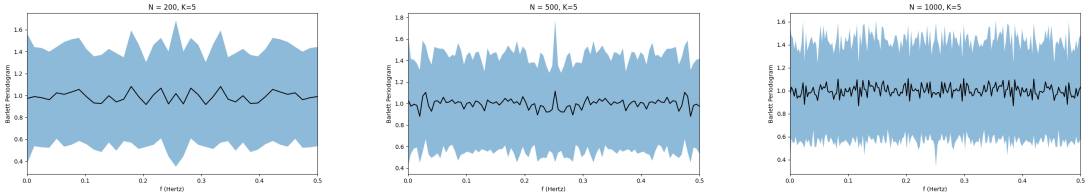
Question 9

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).

As explained before, the variance does not decrease with the sample size N. However, we can see that the variance has been divided by $K = 5$ compared to the results in question 6. To ensure that our periodogram is consistent, it would be interesting to define K according to N (for example, $K = N/constant$) so that the variance decreases with N.

Part IV

Data study

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.

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.

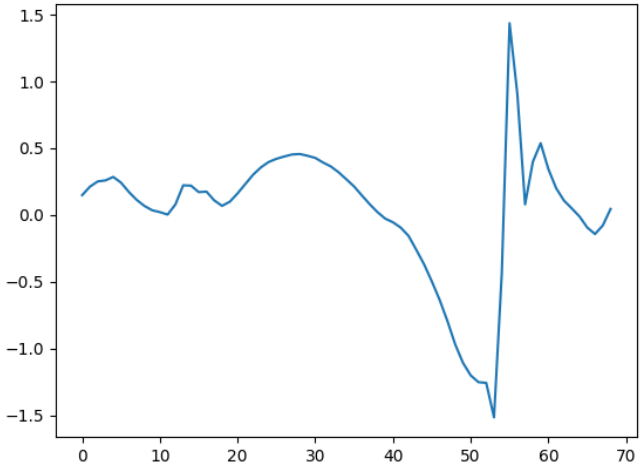
Answer 10

Because the time series does not have the same length, we can not create a KNN model which computes the DTW between times series. So we precompute the distance matrix for X_train and X_test before passing it to the KNN model. The best results are obtained for several neighbours of 1. The associated f1 score is 0.398 based on the prediction with X_test, which is not satisfactory. These bad results can be explained by the use of a k-neighbors classifier model which may not be suited to the study of time series.

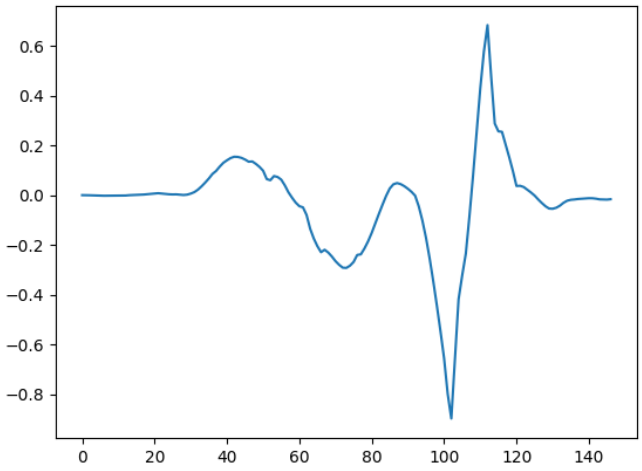
Question 11

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).

The first graph is a badly classified healthy step as we do not observe an erratic behavior when the signal decreases with is the case with the non-healthy step.

Part V

Appendix

1 Question 1

- **First term:** $\frac{1}{2}\|y - X\beta\|_2^2$

  Let's consider $r = y - X\beta$, therefore the first term can be written as $\frac{1}{2}\|r\|_2^2$. To find the gradient of this term with respect to β, we first compute the derivative of r with respect to β:

$$\frac{\partial r}{\partial \beta} = \frac{\partial (y - X\beta)}{\partial \beta} = -X$$

  Next, we use the chain rule. The gradient of $\frac{1}{2}\|r\|_2^2$ with respect to β is

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

  Since $r = y - X\beta$, this can be rewritten as

$$-X^T(y - X\beta)$$

- **Second term:** $\lambda \|\beta\|_1$

  Here, $\text{sgn}(\beta)$ is the sign of β, which is a vector with the same dimension as β, where each element is the sign of the corresponding element in β.

  This explains how we obtained the derivation of this second term :

  The ℓ_1 norm of a vector β is defined as

$$\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|,$$

  where p is the number of elements in β, and β_i is the i-th element of β.

  To find the derivative of the ℓ_1 norm with respect to β, we need to consider each element of β separately. The derivative of the absolute value function $|x|$ with respect to x is given by

$$\frac{d|x|}{dx} = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ \text{undefined} & \text{if } x = 0. \end{cases}$$

  Therefore, the derivative of $\|\beta\|_1$ with respect to β_i is the sign of β_i:

$$\frac{\partial \|\beta\|_1}{\partial \beta_i} = \text{sgn}(\beta_i).$$

The gradient of $\|\beta\|_1$ with respect to β is then a vector whose i-th element is $\text{sgn}(\beta_i)$:

$$\nabla_\beta \|\beta\|_1 = [\text{sgn}(\beta_1), \text{sgn}(\beta_2), \ldots, \text{sgn}(\beta_p)].$$

We define the sign function $\text{sgn}(x)$ as:

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ \text{undefined} & \text{if } x = 0. \end{cases}$$

For the undefined case, the issue is solved in the next steps of the question.

2 Question 2

The convolution can be represented as a matrix-vector multiplication, where the matrix is build from the dictionary atom $\mathbf{d}_k$ in such a way that each row corresponds to a shifted version of $\mathbf{d}_k$. The result of the convolution is then equivalent to the product of this matrix with the activation signal $\mathbf{z}_k$.

To understand this result step by step, we can consider the convolution product for $i \in [1, N]$

$$(\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k)_i = \sum_{k=1}^{K} \sum_{j=1}^{N} z_k(j) d_k(i-j)$$

3 Question 3

Variance, denoted by σ^2, is defined as the expected value of the squared deviation of a random variable from its mean. For a random variable X, this is written as:

$$\sigma^2 = \mathbb{E}[(X_n - \mathbb{E}[X_n])^2]$$

As the mean $\mathbb{E}[X_n]$ is zero, which is often the case for a white noise process, the variance simplifies to:

$$\sigma^2 = \mathbb{E}[X_n^2]$$

That is, the variance is simply the expected value of X_n squared when the mean is zero.

So, when we talk about a Gaussian white noise process X_n, which by definition has a mean of zero and some constant variance σ^2, the autocovariance at lag zero, $\gamma(0)$, is calculated as:

$$\gamma(0) = \mathbb{E}[X_n X_n] = \mathbb{E}[X_n^2]$$

Since the mean of X_n is zero, the above expression for $\gamma(0)$ is equivalent to the variance σ^2 of X_n. Hence, we can conclude that:

$$\gamma(0) = \sigma^2$$

4 Question 4

For the sample autocovariance $\hat{\gamma}(\tau)$, we have:

$$\begin{aligned}
\mathbb{E}(\hat{\gamma}(\tau)) &= \mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}\right) \\
&= \frac{1}{N}\sum_{n=0}^{N-\tau-1} \mathbb{E}(X_n X_{n+\tau}) \ \text{(Linearity of the expectation)} \\
&= \frac{1}{N}\sum_{n=0}^{N-\tau-1} \gamma(\tau) \\
&= \frac{N-\tau}{N}\gamma(\tau)
\end{aligned}$$

5 Question 5

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

Now we can express this as a function of the sample autocovariances $\hat{\gamma}(\tau)$. Notice that when $n = m$, $e^{-\frac{2\pi i k(n-m)}{N}} = 1$, and when $n \neq m$, the exponential term corresponds to a phase shift. We can rewrite the expression as:

$$|J(f_k)|^2 = \hat{\gamma}(0) + \sum_{\tau=1}^{N-1}\left(\hat{\gamma}(\tau)e^{-\frac{2\pi i k\tau}{N}} + \hat{\gamma}(-\tau)e^{\frac{2\pi i k\tau}{N}}\right)$$

6 Question 7

$$\text{var}(\hat{\gamma}(\tau)) = \mathbb{E}[(\hat{\gamma}(\tau) - \gamma(\tau))^2].$$

Now, we can expand $\hat{\gamma}(\tau)$:

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

To compute the expectation, we use the hint given in the question. Let $Y_1 = X_n$, $Y_2 = X_{n+\tau}$, $Y_3 = X_m$, and $Y_4 = X_{m+\tau}$. We then have:

$$\mathbb{E}[Y_1 Y_2 Y_3 Y_4] = \mathbb{E}[Y_1 Y_2]\mathbb{E}[Y_3 Y_4] + \mathbb{E}[Y_1 Y_3]\mathbb{E}[Y_2 Y_4] + \mathbb{E}[Y_1 Y_4]\mathbb{E}[Y_2 Y_3].$$

Now, we can plug in the values of Y_1, Y_2, Y_3, Y_4 to obtain the expression for the variance of $\hat{\gamma}(\tau)$:

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