Assignment 1 (ML for TS) - MVA

Léa Bohbot lea.bohbot@polytechnique.edu
Grégoire Béchade gregoire.bechade@gmail.com

October 25, 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

$$\lambda_{\max} = \dots \tag{2}$$

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:

$$\min_{(\mathbf{d}_k)_k,(\mathbf{z}_k)_k\|\mathbf{d}_k\|_2^2\leq 1} \quad \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 \quad + \quad \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \tag{3}$$

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

$$\lambda_{\max} = \dots \tag{4}$$

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:

$$S(f) := \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2\pi f \tau / f_s}. \tag{5}$$

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

$\gamma(\tau) = \mathbb{E}(X_n X_{n+\tau})$
We have that $X_n \sim \mathcal{N}(0, \sigma^2)$. Therefore, if $\tau = 0$, we have directly that $\gamma(0) = \mathbb{E}(X_n^2) = \sigma^2$. If $\tau \neq 0$, we have that $\gamma(\tau) = \mathbb{E}(X_n X_{n+\tau}) = \mathbb{E}(X_n)\mathbb{E}(X_{n+\tau}) = 0$, as X_n is a white noise.

Finally, $\gamma(\tau) = \mathbb{1}_{\{\tau=0\}}\sigma^2$.

$$\begin{aligned} S(f) &= \sum_{\tau=-\infty}^{+\infty} \gamma(\tau) e^{-2\pi f \tau / f_s} && (6) \\ &= \gamma(0) e^{-2\pi f \times 0 / f_s} && (7) \\ &= \sigma^2 && (8) \\ & && (9) \end{aligned}$$

Question 4

A natural estimator for the autocorrelation function is the sample autocovariance

$$\hat{\gamma}(\tau) := (1/N) \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \qquad (10)$$

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

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

Answer 4

$$\begin{aligned} \mathbb{E}(\hat{\gamma}(\tau)) &= \mathbb{E}(\frac{1}{N} \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau}) && (11) \\ &= \frac{1}{N} \sum_{n=0}^{N-\tau-1} \mathbb{E}(X_n X_{n+\tau}) && (12) \\ &= \frac{N-\tau}{N} \gamma(\tau) && (13) \\ &\neq \gamma(\tau) && (14) \end{aligned}$$

This estimator is therefore **biased**. However, $\lim_{N\to\infty} \mathbb{E}(\hat{\gamma}(\tau)) = \gamma(\tau)$, so it is **asymptotically unbiased**. A simple way to de-bias this estimator is to multiply it by $\frac{N}{N-\tau}$.

Question 5

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

$$J(f) := (1/\sqrt{N}) \sum_{n=0}^{N-1} X_n e^{-2\pi i f n / f_s} \tag{15}$$

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 $f > 0$.

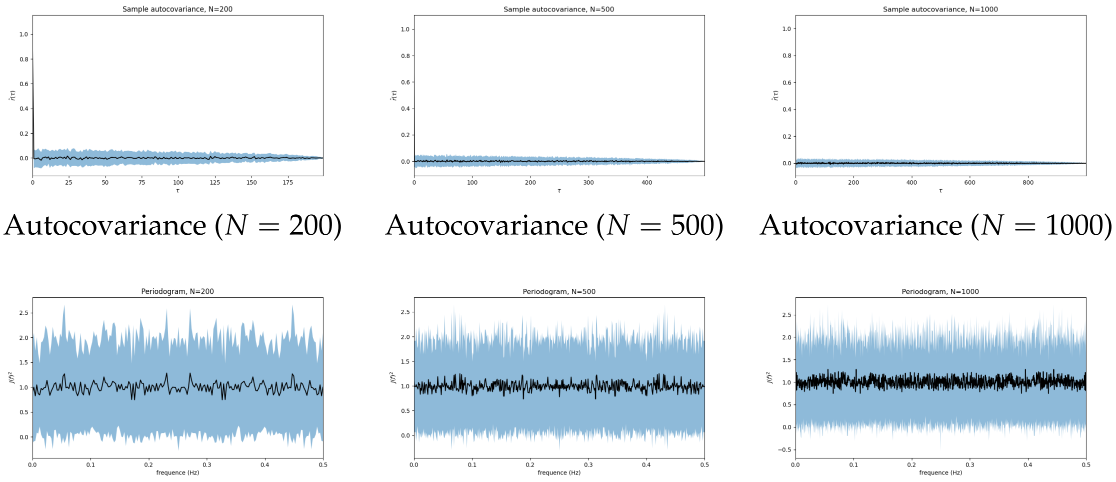
Answer 5

Question 6

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

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

Add your plots to Figure 1.



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

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

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

Answer 6

We observe that the autocovariance function converges towards the expected value ($\mathbb{1}_{\{0\}}$). The fact that the standard deviation decreases with N is expected, as a bigger N means more samples and therefore a more accurate estimation of the autocovariance function. However, the fact that the standard deviation decreases when τ increases is surprising: when τ increases, the number of samples in the set from which the standard deviation is computed decreases, which should lead to a bigger variance. This decrease comes from the fact that the $\hat{\gamma}$ is a sum of $N - \tau - 1$ divided by N, leading to smaller values when τ gets bigger. We would probably observe a different behaviour (with a standard deviation increasing with τ) with the unbiaised estimator.

Increasing the length of the samples does not have any affect on the periodogram.

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 $\tau > 0$

$$\operatorname{var}(\hat{\gamma}(\tau)) = (1/N) \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(1 - \frac{\tau + |n|}{N}\right) \left[\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau)\right]. \tag{16}$$

  (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

$Var(\hat{\gamma}(\tau)) = \mathbb{E}(\hat{\gamma}(\tau)^2) - \mathbb{E}(\hat{\gamma}(\tau))^2$

$\mathbb{E}((\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau})^2) = \mathbb{E}(\sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \sum_{m=0}^{N-\tau-1} X_m X_{m+\tau})$
$= \sum_{n=0}^{N-\tau-1} \sum_{m=0}^{N-\tau-1} \mathbb{E}(X_n X_{n+\tau} X_m X_{m+\tau})$
As the X_n are jointly Gaussian variables, we have :

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

However, we have that $\mathbb{E}(\hat{\gamma})^2 = (\frac{N-\tau}{N}\gamma(\tau))^2$, which leads to

$Var(\hat{\gamma}(\tau)) = \frac{1}{N^2} \times \sum_{n=0}^{N-\tau-1} \sum_{m=0}^{N-\tau-1} \gamma(n-m)^2 + \gamma(\tau+m-n)\gamma(\tau+n-m)$
.

Therefore, it is relevant to **sum on the difference between** n **and** m. $n - m$ takes values in $[-(N - \tau - 1), N - \tau - 1]$, and a value n' in $[-(N - \tau - 1), N - \tau - 1]$ is taken $N - \tau - |n'|$ times in the double sum.

We finally get the result expected:

$$Var(\hat{\gamma}(\tau)) = \frac{1}{N} \times \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(\frac{N-\tau-|n|}{N} \right) \left(\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau) \right).$$

To prove that $\hat{\gamma}$ is consistent, we need to show that $Var(\hat{\gamma}(\tau)) \xrightarrow[n \to \infty]{} \mathbf{0}$, which comes directly from the absolute and square sommabilities of the covariances. Let $\epsilon \in \mathbb{R}_+^*$.

$\mathbb{P}(|\hat{\gamma}(\tau) - \gamma(\tau)| > \epsilon) \leq \frac{Var(\hat{\gamma}(\tau))}{\epsilon^2} \xrightarrow[n \to \infty]{} 0$. (With **Bienaymé-Tchebychev inequality**).

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 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 $A(f) := \sum_{n=0}^{N-1} X_n \cos(-2\pi f n / f_s$ and $B(f) := \sum_{n=0}^{N-1} X_n \sin(-2\pi f n / f_s$. Observe that $J(f) = (1/N)(A(f) + iB(f))$.

- 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

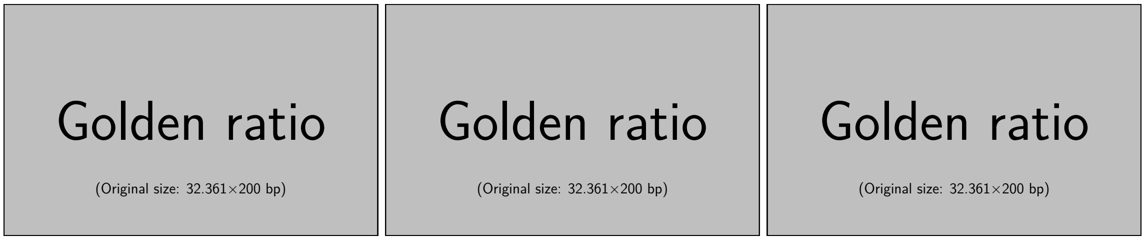
Question 9

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.

Answer 9


Golden ratio
(Original size: 32.361×200 bp)
Golden ratio
(Original size: 32.361×200 bp)
Golden ratio
(Original size: 32.361×200 bp)

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 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.

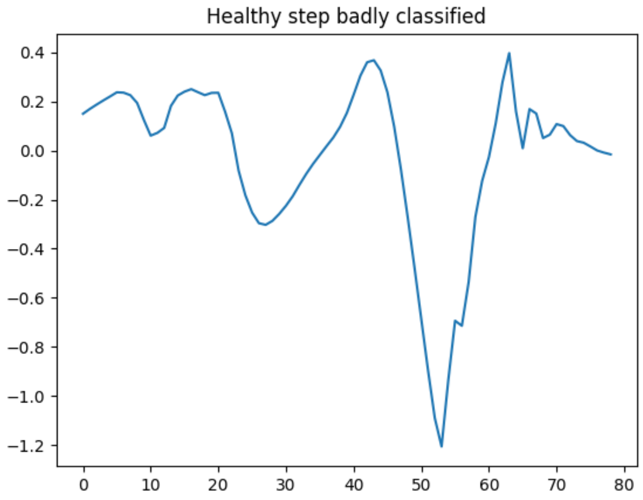
Answer 10

We have splitted the train set in 5 patches and have trained a KNN on 4 of them and tested on the last one. We have then computed the F1-score for each of these 5 situations and averaged the results for each number of neighbors. The optimal number of neighbourghs is **k=3**, with a **F1-score of 0.58**. We have then trained the model on the whole train set and tested it on the test set, obtaining a **F1-score of 0.57.** The huge difference between those two numbers can be explained by the fact that the train set is balanced, while the test set has a proportion of 0.82 unhealthy steps.

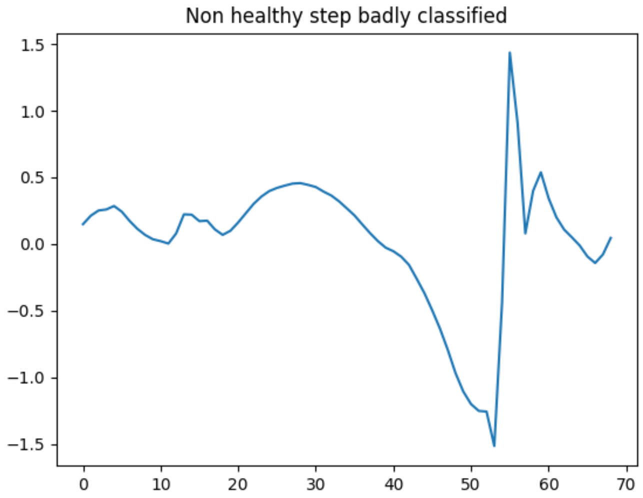
Question 11

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).