Assignment 1 (ML for TS) - MVA 2022/2023

Zakaria El Founoun zakaria.el-founoun@student-cs.fr

May 21, 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 Wednesday 1st February 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: dropbox.com/request/8uHP2WLfYTS1Js8LNkP6.

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

Using the sub-gradient of $\|.\|$.

We have $\quad X^T(y - X\hat{\beta}) = \lambda z \quad$ with : $z_j = \begin{cases} \{+1\} & if \quad \hat{\beta}_j > 0 \\ \{-1\} & if \quad \hat{\beta}_j < 0 \\ [-1,1] & if \quad \hat{\beta}_j = 0 \end{cases}$

Thus $\quad \hat{\beta} = 0 \iff X^T y = \lambda z$
That means $\quad \lambda \geq \left\|X^T y\right\|_\infty$ Now if $\quad \lambda \geq \left\|X^T y\right\|_\infty$
we have $\quad X^T X\hat{\beta} = -\lambda z + X^T y$
Then $\quad \hat{\beta}^T X^T X\hat{\beta} = \hat{\beta}^T(-\lambda z + X^T y)$
However $\quad (\lambda z - X^T y)j$ have the sign of $\hat{\beta}j$
So $\quad \hat{\beta}^T X^T X\hat{\beta} \leq 0$ and $\hat{\beta} = 0$ This means :

$$\lambda_{\max} = \left\|X^T y\right\|_\infty \tag{2}$$

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

We use the matrix formulation for this problem using Parseval theorem and the convolution and Fourier transformation properties :
min $\|x - DZ\|_2 + \lambda \|Z\|_1$
as Seen in Q1 :

$$\lambda_{\max} = \left\|D^T x\right\|_\infty \tag{4}$$

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 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:

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

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.

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

For a white noise of variance σ^2 we have $\gamma(\tau) = \sigma^2 \mathbf{1}_{\tau=0}$

Thus we have: $S(f) = \sigma^2$ which is independant of f.

Question 4

A natural estimator for the autocorrelation function is the sample autocovariance

$$\hat{\gamma}(\tau) := (1/N) \sum_{n=0}^{N-\tau-1} X_n X_{n+\tau} \tag{6}$$

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

We have: $\mathbb{E}(\hat{\gamma}(\tau)) := (1/N) \sum_{n=0}^{N-\tau-1} \mathbb{E}(X_n X_{n+\tau}) = = \frac{N-\tau}{N} \gamma(\tau)$ when N goes to infinity $\frac{N-\tau}{N}$ goes to 1 this the estimator is asymptotically unbiased.

An unbiased estimator would be $\frac{N}{N-\tau}\hat{\gamma}(\tau)$

Question 5

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

$$J(f) := (1/\sqrt{N}) \sum_{n=0}^{N-1} X_n e^{-2\pi i f n/f_s} \tag{7}$$

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

we have : $|J(f_k)|^2 = \frac{1}{N}\sum_{n,m=0}^{N-1} X_n X_m e^{\frac{-2i\pi k(n-m)}{N}}$
Now we introduce the variable $\tau = n - m$, Thus the inequality becomes:
$|J(f_k)|^2 = \frac{1}{N}\sum_{\tau=-(N-1)}^{N-1}\sum_{m=0}^{N-\tau-1} X_m X_{m+\tau} e^{\frac{-2i\pi k\tau}{N}}$
Thus: $|J(f_k)|^2 = \frac{1}{N}\sum_{\tau=-(N-1)}^{N-1} \hat{\gamma}(\tau) e^{\frac{-2i\pi k\tau}{N}}$

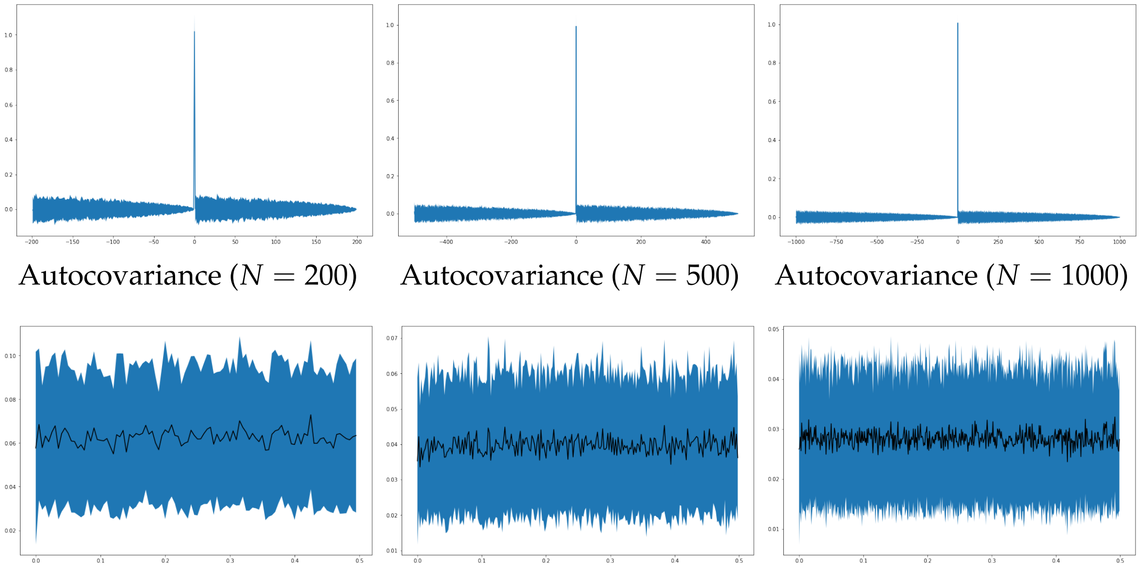
But since $\hat{\gamma}(-\tau) = \hat{\gamma}(\tau)$ we will have
$|J(f_k)|^2 = \frac{1}{N}(\hat{\gamma}(0) + 2\sum_{\tau=1}^{\tau N-1} \hat{\gamma}(\tau) cos(\frac{2\pi k\tau}{N}))$

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

- We can notice for the sample autocovariances that the plot converge to a Dirac at 0. And the variance decrease with the sample size.
- We can notice that the variance of the periodogram do 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 $\tau > 0$

$$\operatorname{var}(\hat{\gamma}(\tau)) = (1/N) \sum_{n=-(N-\tau-1)}^{n=N-\tau-1} \left(1 - \frac{\tau + |n|}{N}\right) \left[\gamma^2(n) + \gamma(n-\tau)\gamma(n+\tau)\right]. \tag{8}$$

  (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

we have $\operatorname{var}(\hat{\gamma}(\tau)) = E(\hat{\gamma}(\tau)^2) - E(\hat{\gamma}(\tau))^2$
And $E(\hat{\gamma}(\tau)) = \frac{N-\tau}{N}\gamma(\tau)$
$\operatorname{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{n_1=0}^{N-\tau-1}\sum_{n_2=0}^{N-\tau-1} E(X_{n_1}X_{n_1+\tau}X_{n_2}X_{n_2+\tau}) - E(\hat{\gamma}(\tau))^2$
Using the hint and that $E(X_nX_{n+\tau}) = \gamma(\tau)$:

$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{n_1=0}^{N-\tau-1}\sum_{n_2=0}^{N-\tau-1}(\gamma(\tau)^2 + \gamma(n_2-n_1)^2 + \gamma(n_2-n_1+\tau)\gamma(n_2-n_1-\tau)) - E(\hat{\gamma}(\tau))^2$
So $\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{n_1=0}^{N-\tau-1}\sum_{n_2=0}^{N-\tau-1}(\gamma(n_2-n_1)^2 + \gamma(n_2-n_1+\tau)\gamma(n_2-n_1-\tau))$
we change n_2 to $n = n_2 - n_1$, and we switch the sum to have :
$\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N^2}\sum_{n=-(N-\tau-1)}^{N-\tau-1}(N-\tau-|n|)(\gamma(n)^2 + \gamma(n+\tau)\gamma(n-\tau))$
Finally : $\text{var}(\hat{\gamma}(\tau)) = \frac{1}{N}\sum_{n=-(N-\tau-1)}^{N-\tau-1}(1-\frac{\tau-|n|}{N})(\gamma(n)^2 + \gamma(n+\tau)\gamma(n-\tau))$

We have $\sum_{\tau\in\mathbb{Z}}\gamma^2(\tau) < \infty$ and $|\sum_{n=-(N-\tau-1)}^{N-\tau-1}\gamma(n+\tau)\gamma(n-\tau)| \leq \|\gamma\|_2^2$
so var $(\hat{\gamma}(\tau)) \xrightarrow[N\to\infty]{} 0$
That Means $\hat{\gamma}$ is consistent

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 $f = f_0, f_1, \dots, f_{N/2}$ 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

- $\mathbb{E}(A(f_k)) = \mathbb{E}(B(f_k)) = 0$
- $var(A(f_k)) = \sum_{n=0}^{N-1}\sigma^2 cos^2(\frac{2\pi nk}{N})$ using the equality: $cos^2(a) = \frac{1+cos(2a)}{2}$ we have $var(A(f_k)) = \frac{\sigma^2 N}{2}$ using the same we find that : $var(B(f_k)) = \frac{\sigma^2 N}{2}$
- Since (X_n) is a white noise, $A(f_k)$ and $B(f_k)$ are Gaussian independent variables hence $A^2(f_k) + B^2(f_k)$ is a distributed according to a $\frac{\sigma\sqrt{N}}{\sqrt{2}}\chi^2$ with 2 degree of freedom.
- Thus we have: $var(|J(f_k)|^2) = \frac{1}{N}2 * 2 * \frac{\sigma^2 N}{2} = 2 * \sigma^2$ Hence the variance do not decrease with the sample size, thus the periodogram is not consistant.

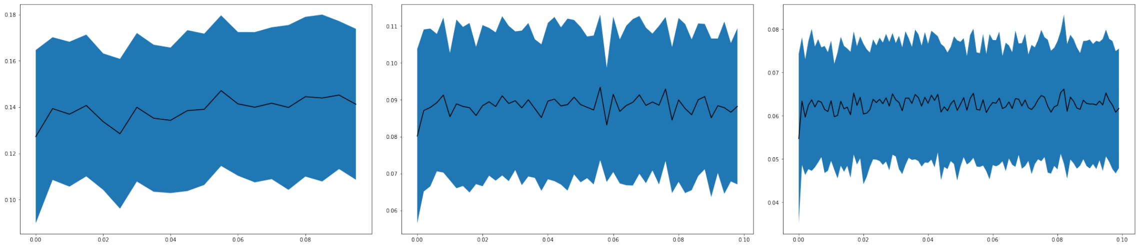
Question 9

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).

- We can notice that now the variance is decreasing.

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.

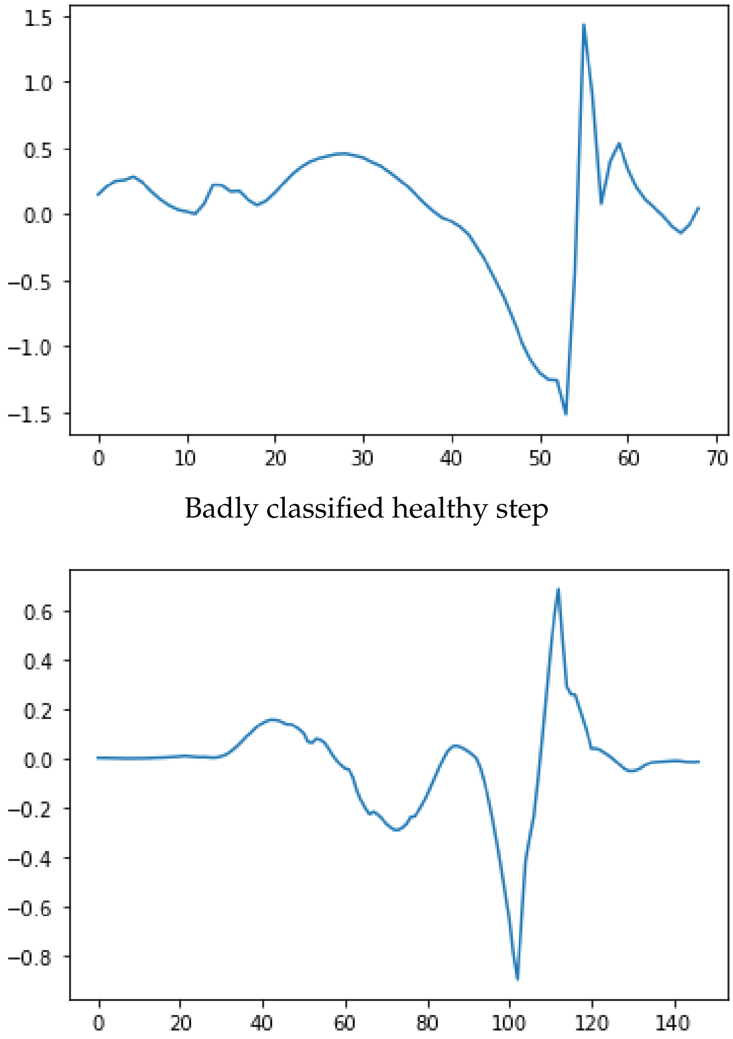
Answer 10

The best k found with 5-fold cross validation is 1 and the best average F-score is 0.61. The model did not perform very will, and more sophisticated methods are required to get better results.

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).