Assignment 3 (ML for TS) - MVA 2022/2023

Théo Di Piazza theo.dipiazza@gmail.com
Stanislas du Ché stanislasduche@gmail.com

April 7, 2023

1 Introduction

Objective. The goal is to implement (i) a signal processing pipeline with a change-point detection method and (ii) wavelets for graph signals.

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 Friday 7th April 11:59 PM.
- Rename your report and notebook as follows:
  `FirstnameLastname1_FirstnameLastname1.pdf` and
  `FirstnameLastname2_FirstnameLastname2.ipynb`.
  For instance, `LaurentOudre_CharlesTruong.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: https://www.dropbox.com/request/rmETjrLAH9Li3pf8JvOt.

2 Dual-tone multi-frequency signaling (DTMF)

In the last tutorial, we started designing an algorithm to infer from a sound signal the sequence of symbols encoded with DTMF.

Question 1

Finalize this procedure–in particular, find the best hyperparameters. Describe in 5 to 10 lines your methodology and the calibration procedure (give the hyperparameter values).

Answer 1

Fixed number of detection points : When the number of points to be detected is known: i.e. 6 segments to be detected for the signal below, it is a question of training the model and then predicting for $6 * 2 = 12$ points because 6 segments to be detected correspond to $6 * 2$ breaks.

Adaptive method : The objective is to detect a succession of horizontal segments of different sizes, spaced discontinuously. When the number of points to be detected is not known in advance or is difficult to estimate, one solution is to define a performance criterion.

Here, 3 to 100 points to be detected are tested. For each point, the detection is performed. For each detected segment, the values for each possible frequency are summed (ordinate axis) and the frequency for which the sum is maximum is chosen, which would correspond to the detected segment. For this given frequency and on this given detected segment: one determines for each pixel if it is higher than **threshold_pixel** (minimum value of the pixel for which one considers a detection); then one calculates the percentage of detected pixel on this segment, this percentage must be higher than **threshold_detected** so that the detection is considered as successful. The same procedure is used for segments where no points are detected, and the average number of pixels detected must be less than **threshold_empty**. **threshold_pixel** is equal to 0.002, slightly higher than the minimum pixel intensity. After several trials with different values, **threshold_detected** and **threshold_empty** are selected at 0.9 and 0.25. As soon as this criterion is validated, we stop and choose the number of points to detect: for example for the signal below, we find that the number of points is 12.

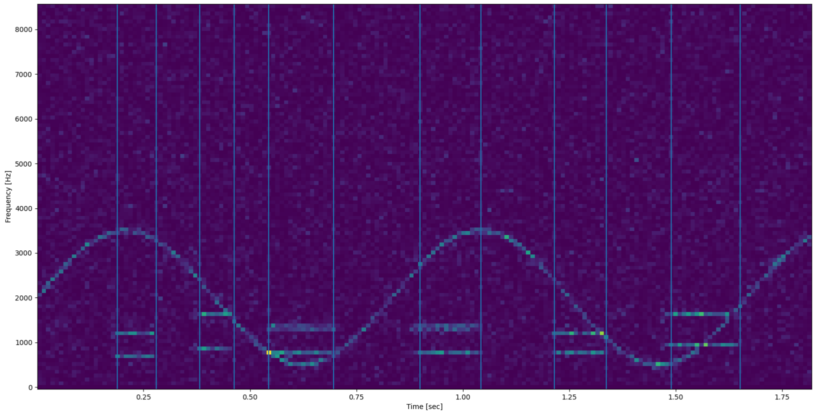


Figure 1: Results for change-point detection from Question 1.

Question 2

What are the two symbolic sequences encoded in the provided signals?

Answer 2

- Sequence 1: B94B68B1
- Sequence 2: CD112639

3 Wavelet transform for graph signals

Let G be a graph defined a set of n nodes V and a set of edges E. A specific node is denoted by v and a specific edge, by e. The eigenvalues and eigenvectors of the graph Laplacian L are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $u_1, u_2, \ldots, u_n$ respectively.

For a signal $f \in \mathbb{R}^n$, the Graph Wavelet Transform (GWT) of f is $W_f : \{1, \ldots, M\} \times V \longrightarrow \mathbb{R}$:

$$W_f(m, v) := \sum_{l=1}^{n} \hat{g}_m(\lambda_l) \hat{f}_l u_l(v) \tag{1}$$

where $\hat{f} = [\hat{f}_1, \ldots, \hat{f}_n]$ is the Fourier transform of f and $\hat{g}_m$ are M kernel functions. The number M of scales is a user-defined parameter and is set to $M := 9$ in the following. Several designs are available for the $\hat{g}_m$; here, we use the Spectrum Adapted Graph Wavelets (SAGW). Formally, each kernel $\hat{g}_m$ is such that

$$\hat{g}_m(\lambda) := \hat{g}^U(\lambda - am) \quad (0 \leq \lambda \leq \lambda_n) \tag{2}$$

where $a := \lambda_n / (M + 1 - R)$,

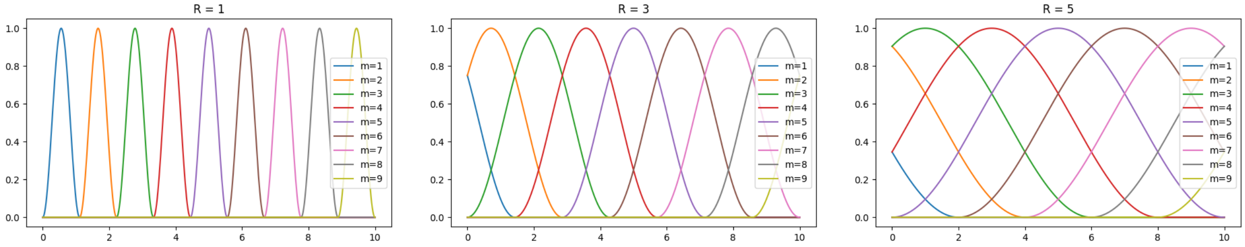
$$\hat{g}^U(\lambda) := \frac{1}{2}\left[1 + \cos\left(2\pi\left(\frac{\lambda}{aR} + \frac{1}{2}\right)\right)\right] \mathbb{1}(-Ra \leq \lambda < 0) \tag{3}$$

and $R > 0$ is defined by the user.

Question 3

Plot the kernel functions $\hat{g}_m$ for $R = 1$, $R = 3$ and $R = 5$ (take $\lambda_n = 12$) on Figure 2. What is the influence of R?

Answer 3



(a) $R = 1$ (b) $R = 3$ (c) $R = 5$

Figure 2: The SAGW kernels functions

R influences the width of the cosine signal: as the value of R increases, the cosine is more spread out. Moreover, by comparing the different kernel functions between them: when R increases, the shifted kernels are more and more overlapped. This effect is generated by the indicator function $\mathbb{1}(-Ra \leq \lambda < 0)$ in equation (3).

We will study the Molene data set (the one we used in the last tutorial). The signal is the temperature.

Question 4

Construct the graph using the distance matrix and exponential smoothing (use the median heuristics for the bandwidth parameter).

- Remove all stations with missing values in the temperature.
- Choose the minimum threshold so that the network is connected and the average degree is at least 3.
- What is the time where the signal is the least smooth?
- What is the time where the signal is the smoothest?

Answer 4

- The stations with missing values are 'BATZ', 'BEG MEIL', 'CAMARET', 'PLOUGONVELIN', 'RIEC SUR BELON', 'ST NAZAIRE-MONTOIR', 'PLOUAY-SA', 'VANNES-MEUCON', 'LANNAERO', 'PLOUDALMEZEAU', 'LANDIVISIAU', 'SIZUN', 'QUIMPER', 'OUESSANT-STIFF', 'LANVEOC', 'ARZAL', 'BREST-GUIPAVAS', 'BRIGNOGAN'.
- The threshold is equal to 0.83.
- The signal is the least smooth at 2014-01-20 06:00:00 !
- The signal is the smoothest at 2014-01-24 19:00:00 !

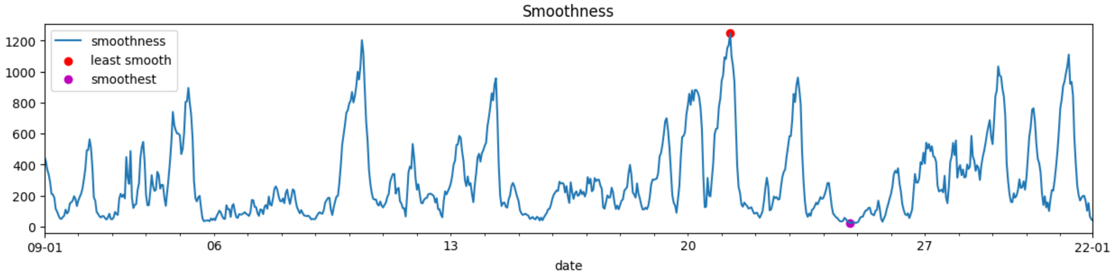


Figure 3: Smoothness VS time.

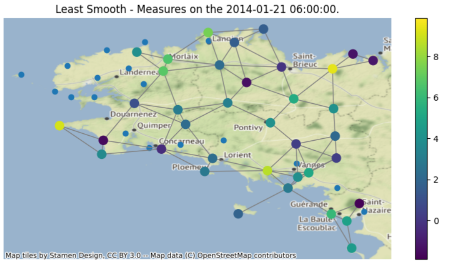


Figure 4: Least smooth

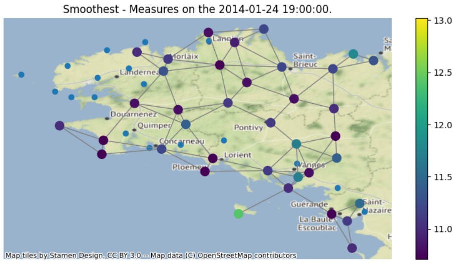


Figure 5: Smoothest

Question 5

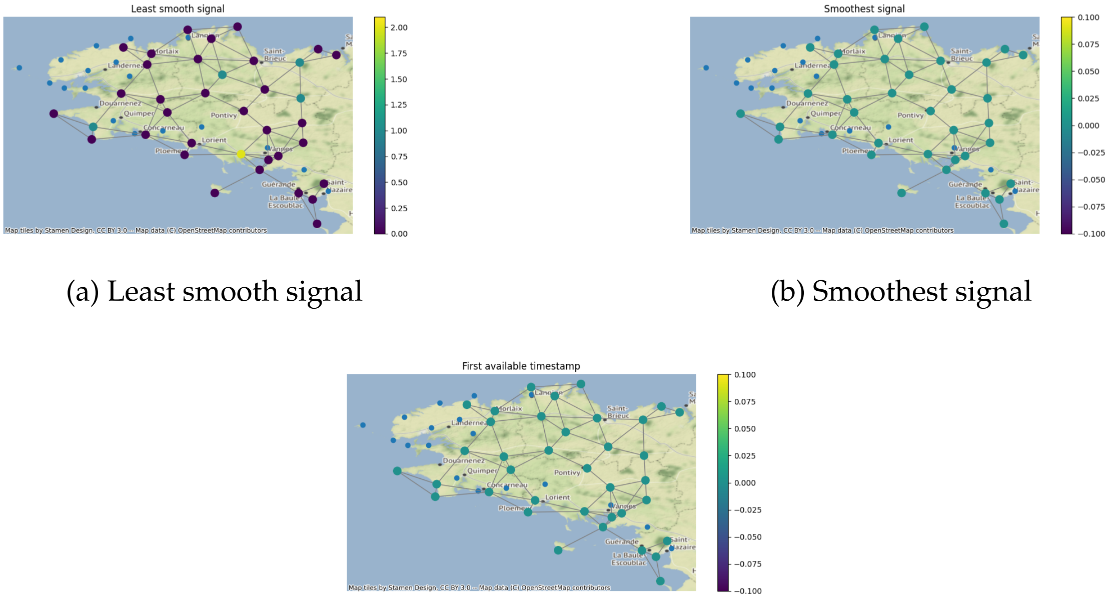
(For the remainder, set $R = 3$ for all wavelet transforms.)

For each node v, the vector $[W_f(1, v), W_f(2, v), \ldots, W_f(M, v)]$ can be used as a vector of features. We can for instance classify nodes into low/medium/high frequency:

- a node is considered low frequency if the scales $m \in \{1, 2, 3\}$ contain most of the energy,
- a node is considered medium frequency if the scales $m \in \{4, 5, 6\}$ contain most of the energy,
- a node is considered high frequency if the scales $m \in \{6, 7, 9\}$ contain most of the energy.

For both signals from the previous question (smoothest and least smooth) as well as the first available timestamp, apply this procedure and display on the map the result (one colour per class).

Answer 5



(a) Least smooth signal

(b) Smoothest signal

(c) First available timestamp

Figure 6: Classification of nodes into low/medium/high frequency

For the smoothest signal and the signal associated with the first timestamp, we see that all nodes are considered as low frequency. For the least smooth signal, some nodes are considered as medium or high frequency.

Question 6

Display the average temperature and for each timestamp, adapt the marker colour to the majority class present in the graph (see notebook for more details).

Answer 6

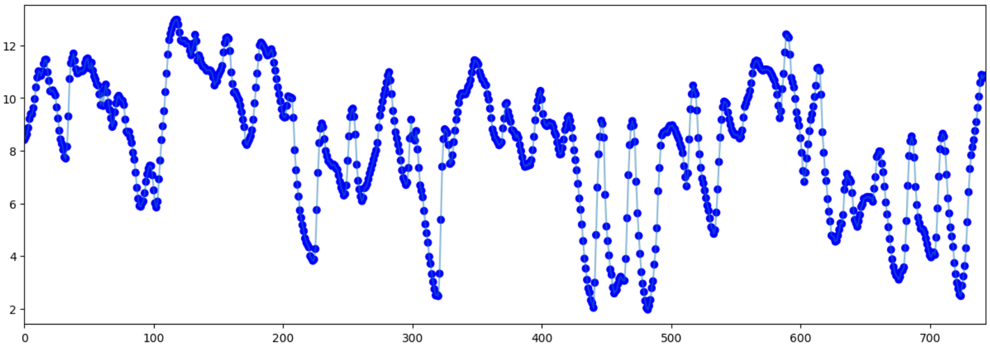


Figure 7: Average temperature. Markers' colours depend on the majority class.

Blue points are associated to low frequency.

Question 7

The previous graph G only uses spatial information. To take into account the temporal dynamic, we construct a larger graph H as follows: a node is now *a station at a particular time* and is connected to neighbouring stations (with respect to G) and to itself at the previous timestamp and the following timestamp. Notice that the new spatio-temporal graph H is the Cartesian product of the spatial graph G and the temporal graph G' (which is simply a line graph, without loop).

- Express the Laplacian of H using the Laplacian of G and G' (use Kronecker products).
- Express the eigenvalues and eigenvectors of the Laplacian of H using the eigenvalues and eigenvectors of the Laplacian of G and G'.
- Compute the wavelet transform of the temperature signal.
- Classify nodes into low/medium/high frequency and display the same figure as in the previous question.

Answer 7

- Let's express the Laplacian of H L_H using the Laplacian of G L_G and the Laplacian of G′ $L_{G'}$ with the Kronecker products. To start, the new spatio-temporal graph H is the Cartesian product of the spatial graph G and the temporal graph G' :

  $$H = G \times G' \implies L_H = L_{G\times G'} = L_G \bigotimes I_n + I_m \bigotimes L_{G'}$$

  where I_n and I_m are the identity matrixes of dim n and m (number of nodes in G and G').

- It comes that :

  $$\begin{aligned} L_H(v_i \bigotimes u_i) &= (L_G \bigotimes I_n + I_m \bigotimes L_{G'})(v_i \bigotimes u_i) \\ &= L_G v_i \bigotimes I_n u_i + v_i I_m \bigotimes u_i L_{G'} \\ &= \lambda_i v_i \bigotimes u_i + \mu_i v_i \bigotimes u_i) \end{aligned}$$

  By noting λ_i eigenvalues of G and μ_i eigenvalues of G', it comes that the eigenvalues of H are :

  $$\{\forall i \in \{1,..,n\} \text{ and } \forall j \in \{1,..,m\} : \lambda_i + \mu_j\}$$

  By noting v_i eigenvectors of G and u_i eigenvectors of G', it comes that the eigenvectors of H are :

  $$\{\forall i \in \{1,..,n\} \text{ and } \forall j \in \{1,..,m\} : v_i \bigotimes u_j\}$$

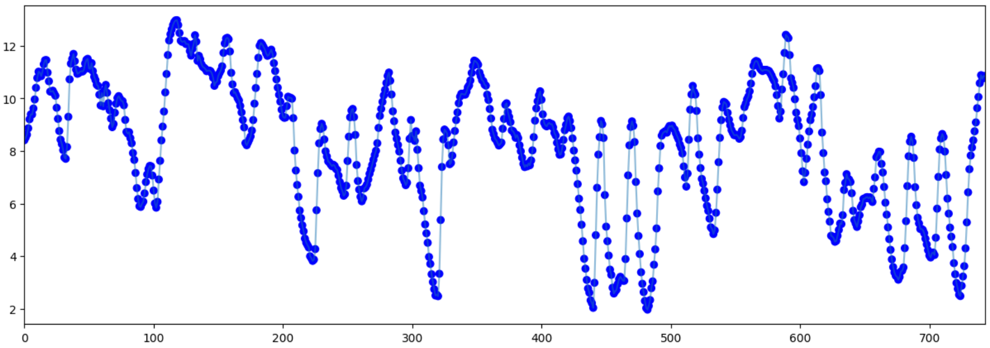


Figure 8: Average temperature. Markers' colours depend on the majority class.

Blue points are associated to low frequency.