Assignment 3 (ML for TS) - MVA 2021/2022

Paul-Emile Zafar paul-emile.zafar@telecom-paris.fr
Marine Mercier marine.mer98@gmail.com

March 21, 2022

1 Introduction

Objective. The goal is to present (i) a model selection heuristics to find the number of change-points in a signal 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 18$^{\text{th}}$ March 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: dropbox.com/request/5DKPDBVAJ25hon0ZZsnn.

2 Model selection for change-point detection

Notations. In the following, $\|x\|$ is the Euclidean norm of x if x is a vector and the Frobenius norm if x is a matrix. A set of change-points is denoted by a bold $\boldsymbol{\tau} = \{t_1, t_2, \dots\}$ and $|\boldsymbol{\tau}|$ (the cardinal of $\boldsymbol{\tau}$) is the number of change-points. By convention $t_0 = 0$ and $t_{|\boldsymbol{\tau}|+1} = T$. For a set of change-points $\boldsymbol{\tau}$, $\Pi_{\boldsymbol{\tau}}$ is the orthogonal projection onto the linear subspace of piecewise constant signals with change-points in $\boldsymbol{\tau}$: for a signal $x = \{x_t\}_{t=0}^{T-1}$,

$$(\Pi_{\boldsymbol{\tau}} x)_t = \bar{x}_{t_k..t_{k+1}} \quad \text{if } t_k \leq t < t_{k+1} \tag{1}$$

where $\bar{x}_{t_k..t_{k+1}}$ is the empirical mean of the subsignal $x_{t_k..t_{k+1}} = \{x_t\}_{t_k}^{t_{k+1}-1}$.

Model selection. Assume we observe a $\mathbb{R}^d$-valued signal $y = \{y_t\}_{t=0}^{T-1}$ with T samples that follows the model

$$y_t = f_t + \varepsilon_t \tag{2}$$

where f is a deterministic signal which we want to estimate with piecewise constant signals and ε_t is i.i.d. with mean 0 and covariance $\sigma^2 I_d$.

The ideal choice of $\boldsymbol{\tau}$ minimizes the distance from the true (noiseless) signal f:

$$\boldsymbol{\tau}^\star = \arg\min_{\boldsymbol{\tau}} \frac{1}{T} \|f - \Pi_{\boldsymbol{\tau}} y\|^2 . \tag{3}$$

The estimator $\boldsymbol{\tau}^\star$ is an *oracle* estimator because it relies on the unknown signal f. Several model selection procedures rely on the "unbiased risk estimation heuristics": if $\hat{\boldsymbol{\tau}}$ minimizes a criterion $\mathrm{crit}(\boldsymbol{\tau})$ such that

$$\mathbb{E}\left[\mathrm{crit}(\boldsymbol{\tau})\right] \approx \mathbb{E}\left[\frac{1}{T} \|f - \Pi_{\boldsymbol{\tau}} y\|^2\right] \tag{4}$$

then

$$\frac{1}{T} \|f - \Pi_{\hat{\boldsymbol{\tau}}} y\|^2 \approx \min_{\boldsymbol{\tau}} \frac{1}{T} \|f - \Pi_{\boldsymbol{\tau}} y\|^2 \tag{5}$$

under some conditions. In other words, the estimator $\hat{\boldsymbol{\tau}}$ approximately minimizes the oracle quadratic risk.

Here, we will consider penalized criteria:

$$\mathrm{crit}(\boldsymbol{\tau}) = \frac{1}{T} \|y - \Pi_{\boldsymbol{\tau}} y\|^2 + \mathrm{pen}(\boldsymbol{\tau}) \tag{6}$$

where pen is a penalty function. In addition, let

$$\hat{\boldsymbol{\tau}}_{\mathrm{pen}} := \arg\min_{\boldsymbol{\tau}} \left[\frac{1}{T} \|y - \Pi_{\boldsymbol{\tau}} y\|^2 + \mathrm{pen}(\boldsymbol{\tau})\right] . \tag{7}$$

Question 1 *Ideal penalty*

- Calculate $\mathbb{E}[\|\varepsilon\|^2/T]$, $\mathbb{E}[\|f-\Pi_{\boldsymbol{\tau}} y\|^2/T]$ and $\mathbb{E}[\|y-\Pi_{\boldsymbol{\tau}} y\|^2/T]$.
- What would be an ideal penalty $\mathrm{pen}_{\mathrm{id}}$ such that Equation (4) is verified?

Answer 1

$$
\begin{aligned}
\mathbb{E}[\|\varepsilon\|^2/T] &= \frac{\mathbb{E}[\sum_{t=0}^{T}\varepsilon_t^2]}{T} \\
&= \sum_{t=0}^{T-1}\frac{\mathbb{E}[\varepsilon_t^2]-\mathbb{E}[\varepsilon_t]^2}{T} \\
&= \sum_{t=0}^{T-1}\frac{\sigma^2}{T} \\
&= \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\|f-\Pi_{\boldsymbol{\tau}} y\|^2/T] &= \mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{y_{t_k}})^2]/T \\
&= \mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{f_{t_k}}-\bar{\varepsilon_{t_k}})^2]/T \\
&= \sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(\mathbb{E}[(f_t-\bar{f_{t_k}})^2]+\mathbb{E}[\bar{\varepsilon_{t_k}}^2]-2\mathbb{E}[\bar{\varepsilon_{t_k}}(f_t-\bar{f_{t_k}})])/T \\
&= \mathbb{E}[\|f-\Pi_{\boldsymbol{\tau}} f\|^2/T]+\mathbb{E}[\|\varepsilon\|^2/T]
\end{aligned}
$$

With $2\mathbb{E}[\bar{\varepsilon_{t_k}}(f_t-\bar{f_{t_k}})]) = 2\mathbb{E}[\bar{\varepsilon_{t_k}}]\mathbb{E}[(f_t-\bar{f_{t_k}})] = 2\sum_{t=t_k}^{t_{k+1}-1}\mathbb{E}[\varepsilon_{t_k}]\mathbb{E}[(f_t-\bar{f_{t_k}})] = 0$ since ε_t and $f_t-\bar{f_{t_k}}$ are independant.

$$
\begin{aligned}
\mathbb{E}[\|y-\Pi_{\boldsymbol{\tau}} y\|^2/T] &= \mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(y_t-\bar{y_{t_k}})^2/T] \\
&= \mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t+\varepsilon_t-\bar{y_{t_k}})^2/T] \\
&= \mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{y_{t_k}})^2/T]+\mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}\varepsilon_t^2/T]+2\mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{y_{t_k}})\varepsilon_t/T] \\
&= \mathbb{E}[\|f-\Pi_{\boldsymbol{\tau}} y\|^2/T]+\mathbb{E}[\|\varepsilon\|^2/T]-2(|\tau|+1)\sigma^2/T
\end{aligned}
$$

because

$$
\begin{aligned}
2\mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{y_{t_k}})\varepsilon_t]/T &= 2\mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{f}_{t_k}-\bar{\varepsilon_{t_k}})\varepsilon_t]/T \\
&= 2\mathbb{E}[\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(f_t-\bar{f}_{t_k}-\frac{\sum_{i=t_k,i!=t}^{t_{k+1}-1}\varepsilon_i+\varepsilon_t)}{t_{k+1}-1-t_k})\varepsilon_t]/T \\
&= 2\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(\mathbb{E}[(f_t-\bar{f}_{t_k}-\frac{\sum_{i=t_k,i!=t}^{t_{k+1}-1}\varepsilon_i}{t_{k+1}-1-t_k})\varepsilon_t]-\mathbb{E}[\varepsilon_t^2]/T)/T \\
&= 2\sum_{k=0}^{\tau}\sum_{t=t_k}^{t_{k+1}-1}(\mathbb{E}[(f_t-\bar{f}_{t_k}-\frac{\sum_{i=t_k,i!=t}^{t_{k+1}-1}\varepsilon_i}{t_{k+1}-1-t_k})]\mathbb{E}[\varepsilon_t]-\frac{\mathbb{E}[\varepsilon_t^2]}{t_{k+1}-1-t_k})/T \\
&= -2\sum_{k=0}^{\tau}(t_{k+1}-1-t_k)\frac{\mathbb{E}[\varepsilon_t^2]}{(t_{k+1}-1-t_k)*T} \\
&= -2\sum_{k=0}^{\tau}\sigma^2/T = -2(|\tau|+1)\,\sigma^2/T
\end{aligned}
$$

Therefore we have :

$$
\begin{aligned}
\mathbb{E}[crit(\tau)] &= \mathbb{E}[\|y-\Pi_{\boldsymbol{\tau}}y\|^2/T+pen(\tau)] \\
&= \mathbb{E}[\|f-\Pi_{\boldsymbol{\tau}}y\|^2/T]+\sigma^2-2(|\tau|+1)\sigma^2/T+\mathbb{E}[pen(\tau)]
\end{aligned}
$$

The ideal penalty in order to respect equation (4) is

$$
\mathrm{pen}_{\mathrm{id}}(\boldsymbol{\tau}) = (2\frac{|\tau|+1}{T}-1)\sigma^2 \approx 2\frac{|\tau|+1}{T}\sigma^2
$$

Question 2 *Mallows' C_p*

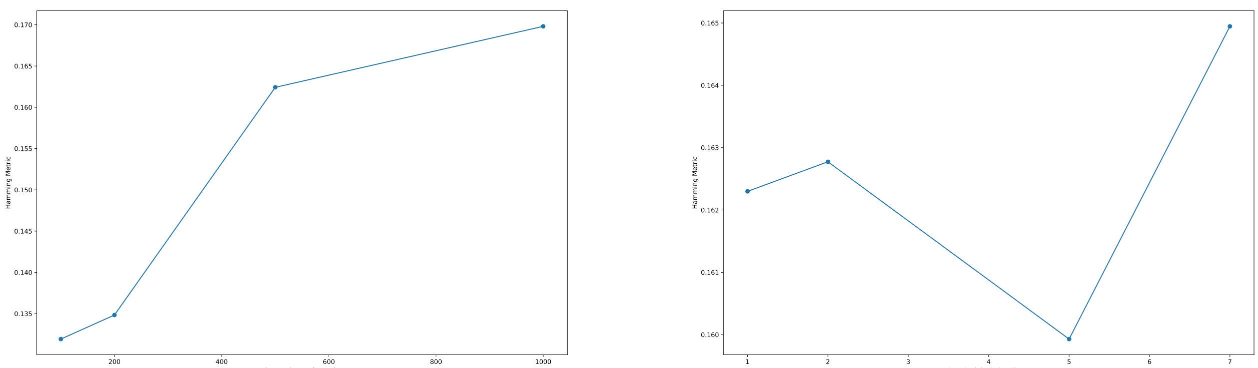
The ideal penalty depends on the unknown value of σ. Pluging an estimator $\hat{\sigma}$ into pen_{id} yields the well-known Mallows' C_p. Use the empirical variance on the first 10% of the signal as an estimator of σ^2.

Simulate two noisy piecewise constant signals with the function `ruptures.pw_constant` (set the dimension to $d = 2$) for each combination of parameters: `n_bkps`$\in \{2, 4, 6, 8, 10\}$, $T \in \{100, 200, 500, 1000\}$ and $\sigma \in \{1, 2, 5, 7\}$.

Using Mallows' C_p,

- for $\sigma = 2$ and $T \in \{100, 200, 500, 1000\}$, compute the Hamming metric between the true segmentation and the estimated segmentation and report the average on Figure 1-a;
- for $T = 500$ and $\sigma \in \{1, 2, 5, 7\}$, compute the Hamming metric between the true segmentation and the estimated segmentation and report the average on Figure 1-b.

Answer 2



(a) Hamming metric vs the number T of samples

(b) Hamming metric vs the standard deviation σ

Figure 1: Performance of Mallows' C_p

Question 3 *Slope heuristics*

The ideal penalty is of shape $\text{pen}(\boldsymbol{\tau}) = Cd|\boldsymbol{\tau}|/T$ where $C > 0$. The slope heuristics is a procedure to infer the best C without knowing σ.

Slope heuristics algorithm.

- Estimate the slope of $\hat{s}$ of $\min_{\boldsymbol{\tau},|\boldsymbol{\tau}|=K} \|\Pi_{\boldsymbol{\tau}} - y\|^2$ as a function of K for K "large enough". Define $\hat{C}_{\text{slope}} := -T\hat{s}$.
- Estimate $\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} \|y - \Pi_{\boldsymbol{\tau}} y\|^2 / T + \hat{C}_{\text{slope}} d|\boldsymbol{\tau}|/T$.

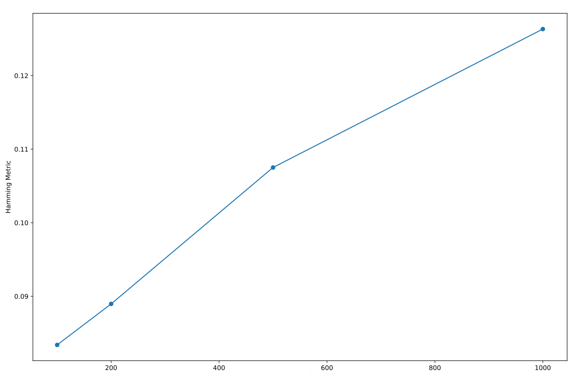
In simulations, "large enough" means for K between 15 and $0.4T$.

Simulate two noisy piecewise constant signals with the function `ruptures.pw_constant` (set the dimension to $d = 2$) for each combination of parameters: `n_bkps`$\in \{2, 4, 6, 8, 10\}$, $T \in \{100, 200, 500, 1000\}$ and $\sigma \in \{1, 2, 5, 7\}$.

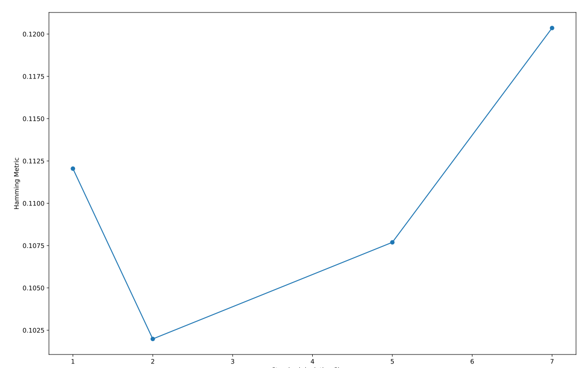
Using the slope heuristics,

- for $\sigma = 2$, $T \in \{100, 200, 500, 1000\}$, compute the average Hamming metric between the true segmentations and the estimated segmentations and report the average on Figure 2-a;
- for $T = 500$ and $\sigma \in \{1, 2, 5, 7\}$, compute the average Hamming metric between the true segmentations and the estimated segmentations and report the average on Figure 2-b.

Answer 3



(a) Hamming metric vs the number T of samples



(b) Hamming metric vs the standard deviation σ

Figure 2: Performance of the slope heuristics

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 $W_f : \{1, \ldots, M\} \times V \longrightarrow \mathbb{R}$:

$$W_f(m, v) := \sum_{l=1}^{n} \hat{g}_m(\lambda_l) \hat{f}_l u_l(v) \tag{8}$$

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

$$\hat{g}_m(\lambda) := \hat{g}^U(\lambda - am) \quad (0 \leq \lambda \leq \lambda_n) \tag{9}$$

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

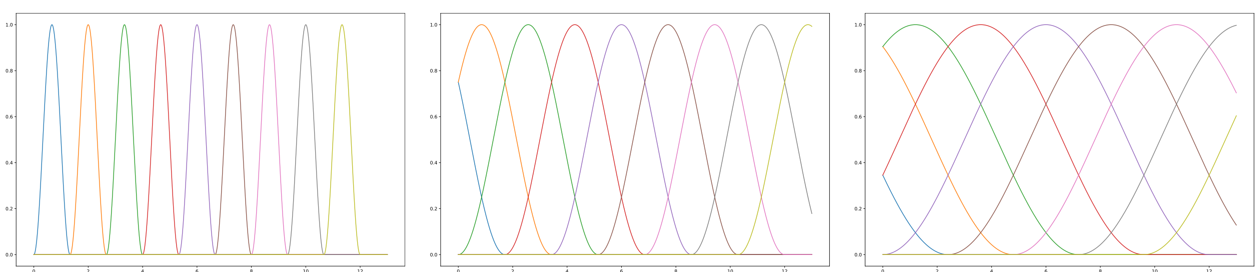
$$\hat{g}^U(\lambda) := \frac{1}{2}\left[1 + \cos\left(2\pi\left(\frac{\lambda}{aR} + \frac{1}{2}\right)\right)\right] \mathbb{1}(-Ra \leq \lambda < 0) \tag{10}$$

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

Question 4

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

Answer 4



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

Figure 3: The SAGW kernels functions

The parameter influence the width of a single cosine half period of the SAGW (as R increases, $1/aR$ decreases), thus the wavelets overlap more and more as R increases.

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

Question 5

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 5

The stations with missing values are Arzal, Batz, Begmeil, Brest-Guipavas, Brignogan, Camaret, Landivisiau, Lannaero, Lanveoc, Ouessant-stiff, Plouay-sa, Ploudalmezeau, Plougonvelin, Quimper, Riec sur Belon, Sizun, St Nazaire-Montoir, Vannes-Meucon

The maximum threshold on the distance matrix is equal to 39.9, thus the minimum threshold on the Gaussian matrix is equal to 0.83.

The signal is the least smooth (when the smoothness is at it's maximum) at 09:00:00 2014/01/10.

The signal is the smoothest at 19:00:00 2014/01/24.

Question 6

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

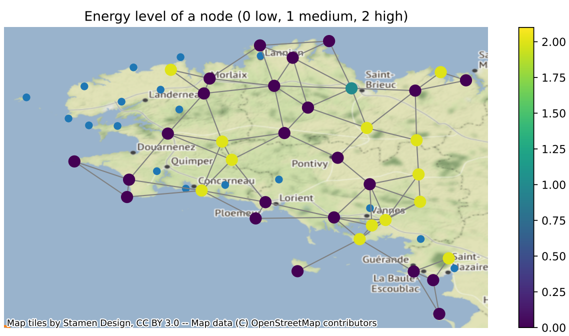
For each node v, the vector $[W_f(1, v), W_f(2, v), \ldots, W_f(M, v)]$ can be used as a vector of features. We can for instance classify nodes into low/medium/high frequency:

- 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.

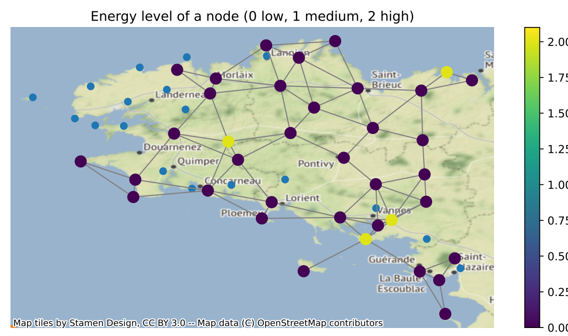
For both signals from the previous question (smoothest and least smooth) as well as the first available timestamp, apply this procedure and display on the map the result (one colour per class).

The legend is: value 2 is a high frequency node (yellow), value 1 is a medium frequency node (blue), value 0 is a low frequency node (purple).

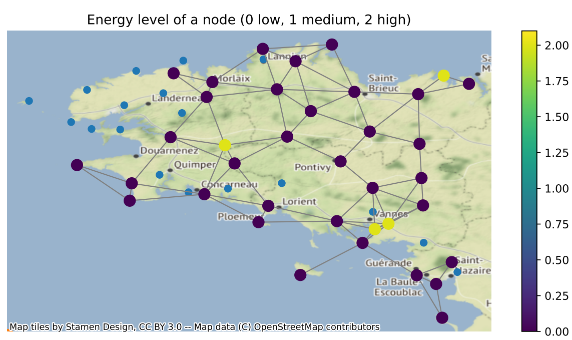
Answer 6



(a) Least smooth signal



(b) Smoothest signal



(c) First available timestamp

Figure 4: Classification of nodes into low/medium/high frequency

Question 7

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 7

Legend of majority classes: **Blue: low frequency, Magenta: medium frequency, Red: High frequency**

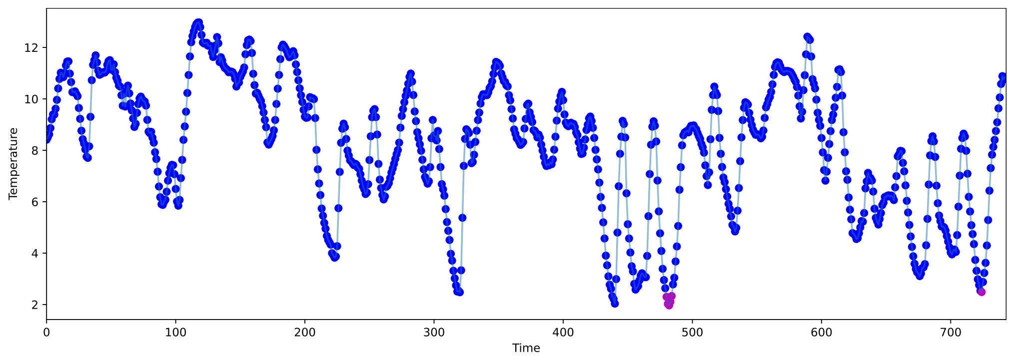


Figure 5: Average temperature. Markers' colours depend on the majority class.

Question 8

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 8

We make a new grah with temporal information from the previous one.

$H = G \times G'$

Let's note D resp W the degree, resp weights matrices of G. (D' and W' are, identically the degree and weights matrices of G'.)
(D'' and W'' are, identically the degree and weights matrices of H'.)
Also $D' = diag(1, 2, 2, ..., 2, 2, 1)$, as the degrees of all nodes are 2, except on the edges.

Because G' is the line graph of size n, we can note that G' is the matrix with ones on the subdiagonal and superdiagonal (every node is connected to the previous and the next one, except the edge nodes).

When making the cartesian product of two graphs, in particular here G and G', we get that:

$$W'' = I_m \otimes W + W' \otimes I_n$$

where $\otimes$ is the Kroenecker product.

When looking at the expression: we can see that

- the first term $(I_n \otimes D)$ gives the weights of G duplicated on diagonal blocks
- the second term $(D' \otimes I_n)$ gives the weights of each node being connected to itself in the previous and next times, meaning subdiagonal and superdiagonal blocks of identity matrices.

We also have that: $D'' = I_m \otimes D + D' \otimes I_n$

Again, when observing the expression, we get back the fact that the resulting degrees of each node is the same as in G, except that we have added 2 (previous + next connection), except for the initial and final graphs that are only added 1 because they connect to one graph only.

As a result, given that $L'' = D'' - W''$, because the Kronecker product is associative, we get that:

$$L'' = I_n \otimes (D - W) + (D' - W') \otimes I_n$$

In conclusion:

$$L'' = Im \otimes L + L' \otimes I_n$$

We note (λ_i) resp (μ_j) the eigenvalues of G, resp G'; and U_i, resp V_j the eigenvectors of L, resp L'. And, the eigenvalues and the eigenvectors the resulting Laplacian are, $\forall i, j, 0 \leq i \leq n, 0 \leq j \leq m$

$$\lambda_i + \lambda_j$$

and the associated eigenvector are

$$V_j \otimes U_i$$

.

Thus, we can compute the eigenvalues and eigenvectors of the spatiotemporal graph without searching the spectrum of the whole Laplacian.

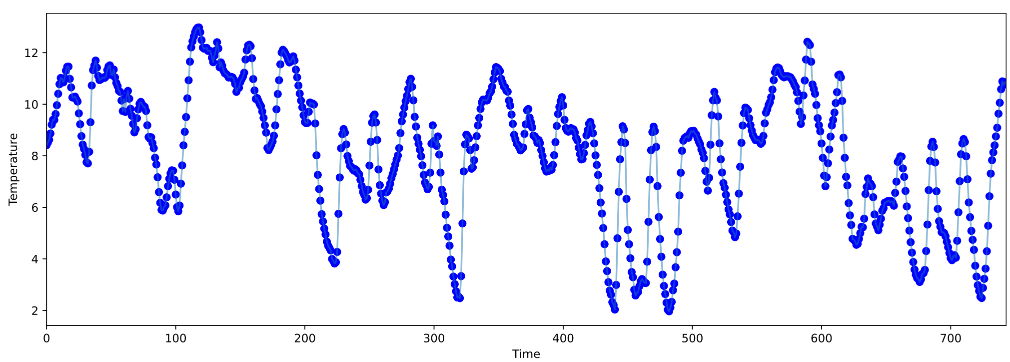


Figure 6: Average temperature. Markers' colours depend on the majority class.