Functional Data Analysis
MSc 2024 – 2025
Homework assignment
Due by ***TBC***

Instructions: For this home assignment, you work individually. You may use Markdown (or the equivalent in Python) for your report, or you can print it in PDF. The comments on the results should appear in your document. You may use implemented functions (for instance for kernel smoothing), but you have to briefly describe their role. The comments in the report are not expected to be long, but should provide evidence that you master the notions. **The report (one file!) has to be uploaded on Moodle, before 18:00, on ***TBC*****

Introduction. The notation below are the same as on the slides. The topic is study of a test for the mean function. More precisely, you have to compute a test statistic like[1]

$$T_{norm,N} := N\|\widehat{\mu} - \mu\|^2, \tag{1}$$

(where $\widehat{\mu}$ is some estimator of the mean function) and investigate its behavior :

- under the null hypothesis $H_0 : \mu = \mu_0$ for some given mean function μ_0 (e.g., the null function);
- under alternatives (departures from the null).

The test size is denoted α. Using simulated data, you investigate the accuracy of the level α, and the power of the test. The data have to be generated using the computer, but mimicking real data features: the curves observed with error, on a random design set with only few points.

Purpose. The test using $T_{norm,N}$, presented in the lectures with $\widehat{\mu}$ the empirical mean function, is based on asymptotic arguments and is designed for the ideal case where the curves are observed everywhere without error. The purpose of this homework is to evaluate the consequences of applying this test with data which are not like in the theory.

Data. A first task is to write a code for generating artificial (simulated) functional data. The characteristics of the simulated data are given below. Let

- N be the sample size (number of curves in the sample);
- K_i the number of design (domain) points on the curve X_i – it could be fixed (all the K_i are equal), or generated from a Poisson distribution (then the K_i are not all equal); the latter may be more challenging for the code, a fixed K will be accepted;
- σ^2 be the noise variance;
- for each $1 \leq i \leq N$, let T_{ik}, $1 \leq k \leq K_i$ denote the design (domain) points on the curve X_i – they are independent draws from a continuous random variable T taking values in $[0, 1]$, for instance the uniform.

[1] Recall, $\mu(t) = \mathbb{E}[X(t)]$, $t \in [0, 1]$.

The data points are

$$(Y_{ik}, T_{ik}), \qquad 1 \le k \le K_i, 1 \le i \le N,$$

where the T_{ik}'s are independent draws from T, independent of the curves X_i,

$$Y_{ik} = X_i(T_{ik}) + \varepsilon_{ik}, \qquad \mathbb{E}(\varepsilon_{ik}) = 0,\ \mathbb{E}(\varepsilon_{ik}^2) = \sigma^2, \tag{2}$$

and the ε_{ik} are independent draws from a noise ε with $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon^2) = \sigma^2$. The curves X_i are independent sample paths of some process X. The noise is independent of the T_{ik}'s and the curves X_i.

To simulate a sample of functional data, that means the data points (Y_{ik}, T_{ik}), go through the following steps: for each $1 \le i \le N$,

- generate K_i (if the K_i are all equal, there is nothing to do, the K_i is given by your choice);
- generate the T_{ik}'s, $1 \le k \le K_i$;
- use the KL decomposition for a Brownian motion on $[0,1]$, truncated at the first J terms to generate the $X_i(T_{ik})$'s;
- add the mean values $\mu(T_{ik})$;
- generate the errors ε_{ik} from some zero-mean distribution with variance σ^2, for instance the Gaussian;
- compute the Y_{ik} corresponding to the T_{ik}'s according to (2)

Provide plots with few ideal X_i and the data points (Y_{ik}, T_{ik}). See also the figure below for illustration.

Curves reconstruction. In theory, the curves have to be reconstructed for each $t \in [0,1]$, and next averaged to get the mean function estimator. In practice, is not possible to reconstruct the curves X_i in any point t, a refined grid of points will be used instead.

- Consider some (large) L, and an equidistant grid of $L+1$ points on $[0,1]$, that is $t_l = l/L$, $0 \le l \le L$. For each i, you have to compute $\widehat{X}_i(t_l)$ that are estimates of $X_i(t_l)$, for all $t_0, t_1, \dots, t_L$. Two types of estimates will be asked :
  - estimate the $X_i(t_l)$'s by linearly interpolating the Y_{ik};
  - estimate the $X_i(t_l)$'s by smoothing, either kernel smoothing (Nadaraya-Watson) or splines;

Provide plots with few ideal X_i, the data points (Y_{ik}, T_{ik}), and the estimates $\widehat{X}_i(t_l)$.

Compute the test statistic. With at hand the estimates $\widehat{X}_i(\cdot)$, construct

$$\widehat{\mu}(t_l) = \frac{1}{N} \sum_{i=1}^{N} \widehat{X}_i(t_l), \qquad 0 \le l \le L.$$

Next, numerically approximate $T_{norm,N}$ defined in (1) (by a Riemann sum or the trapezoidal rule).

Compute the critical values for the test. As stated in the lectures, under the null hypothesis $H_0 : \mu = \mu_0$, the ideal test statistics $T_{norm,N}$ (computed with the empirical mean function estimator based on the true curves X_i), behaves as the random variable $\sum_{j\ge1} \lambda_j Z_j$, where $\lambda_1, \lambda_2, \cdots$ are the eigenvalues of the covariance operator of X, and Z_j are i.i.d. chi-squared random variables. In our simulation setup, the X_i are sample paths of the Brownian motion, thus the values λ_j are well known.

To get the critical values of the test, we need the quantiles of the random variable $\sum_{j\ge1} \lambda_j Z_j$. For this, proceed as follows :

- Truncate the series $\sum_{j\geq 1}\lambda_j Z_j$ at, say, $M = 250$ terms;
- compute numerically by Monte-Carlo the critical values of the test as the quantile[2] $q_{1-\alpha}$ of the random variable $\sum_{j=1}^{M}\lambda_j Z_j$; provide quantile value you get.

Perform the test many times. For a functional data sample, use the instructions above and compute $T_{norm,N}$ under the null hypothesis. Save the value of the indicator function $\mathbf{1}\{T_{norm,N} \geq q_{1-\alpha}\}$. If the indicator is equal to 1, reject the null hypothesis, otherwise do not reject.

The experiments can be repeated, say, R times. At the end, compute the empirical mean of the R indicators $\mathbf{1}\{T_{norm,N} \geq q_{1-\alpha}\}$, which will provide an estimate of the rejection probability for the test.

If $\mu(\cdot)$ used to generate the Y_{ik} is equal to $\mu_0(\cdot)$ (that means if H_0 holds true), the estimate of the rejection probability is expected to be close to the nominal level α. If $\mu \neq \mu_0$, the estimate of the rejection probability is expected to be close to 1, at least when N (the sample size, the number of curves X_i) increases.

The simulation experiment is expected to reveal to which extent these theoretical properties are realistic when the functional data are discretely observed, with error.

Simulation setup. For your study, consider the following values :

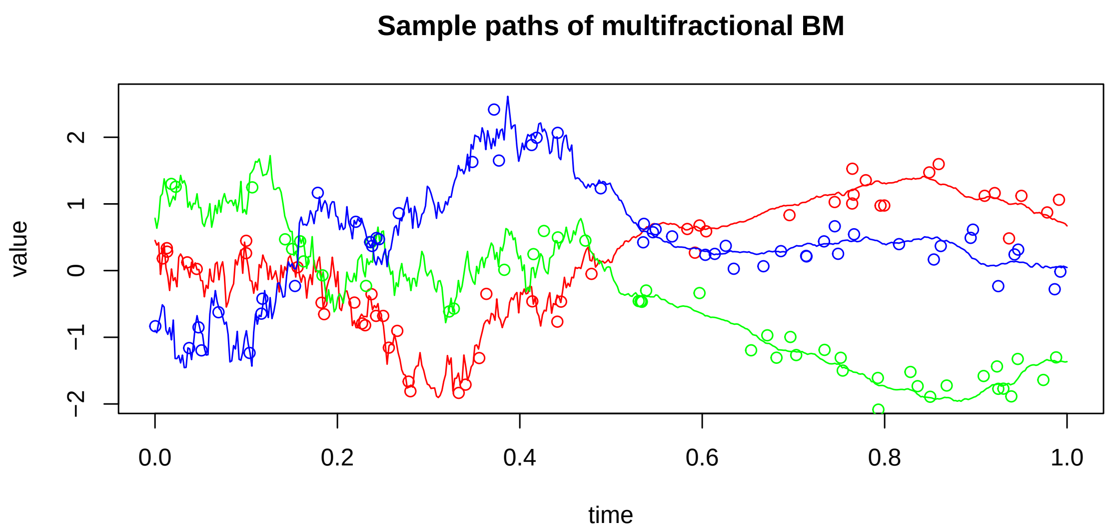
- μ_0 is the null function, and μ_0 is a non-null function built from simple functions (polynomial, sine/cosine,...); consider $\mu_* \neq \mu$, and define alternative hypotheses like $\mu = \gamma\mu_0 + (1-\gamma)\mu_*$, with, say, $\gamma \in \{0.5, 1\}$;
- $R = 200$ (number of replications of the experiment);
- $N \in \{100, 200\}$ (functional data sample size);
- $K = \{10, 100\}$ (number of design points T_{ik} on each curve; if you decide to draw K from a Poisson variable, set the parameter of the Poisson distribution to K);
- $J = 300$ (number of terms in KL decomposition used to generate curves X_i looking like realizations of a Brownian motion);
- $L = 200$ (the size of the grid on which the estimators $\widehat{X}_i(t_l)$ of $X_i(t_l)$ are computed; this was also the number in the first home assignment where the values $X_i(t_l)$ were observed and there was no need for estimation);
- $\sigma^2 \in \{0, 0.5\}$ (the variance of the noise; $\sigma^2 = 0$ means no noise);
- $\alpha = 0.05$ (the test size).

Finally, report and comment the results. The results are the rejection proportions for the different choices of the simulation setup. Is the level of the test accurate? Does it matter the μ_0 is null or not? Is the test powerful? All the results can be simply presented in tables with

- the columns H_0, $H_1(\gamma = 0.5, \mu_0 = \cdots, \mu_* = \cdots)$, $H_1(\gamma = 1, \mu_0 = \cdots, \mu_* = \cdots)$, for the different choices of μ_0 and μ_*;
- the lines $(N = xxx, K = yyy, \sigma^2 = zzz)$, for the different choices of N, K and σ^2.

[2]By definition, here $q_{1-\alpha}$ is the real number such that $\mathbb{P}\left(\sum_{j=1}^{M}\lambda_j Z_j > q_{1-\alpha}\right) = \alpha$.

The picture below present data generated according to the steps described above. The continuous lines are the true curves[3] (here, three curves are represented), the circles are the data points (Y_{ik}, T_{ik}).



[3]In the figure, the true curves X_i, plotted with continuous lines, are generated using another type of process than the Brownian motion X_i, but this does not matter, it is just for illustration purposes. The variable $T \in [0, 1]$ is labeled 'time' in the figure.