Name:...

Functional Data Analysis

$\boxed{\textbf{Assignment 3}}$

Eigenfunctions

Assignments constitute part of the examination and must be handed in time. You are asked to hand in the solutions during a week following the week on which the assignment has been discussed in classes. You can submit either an electronic copy or a hard copy of your work. In the latter case, staple your solutions together.

Problem 1 – Basic formulation of FDA model

The set of all square integrable functions is denoted by $\mathcal{L}^2$ or simply H. Let $X, X_1, X_2, ...$ be random functions that are square integrable. We call μ the expectation of a random function X and denote it by EX if $\mu \in H$ and such that $E\langle X, y\rangle = \langle \mu, y\rangle$ for all $y \in H$.

Further, for a random function X, the covariance operator of X is defined by:

$$C(y) = E[\langle X - EX, y\rangle (X - EX)], \qquad y \in H, \tag{1}$$

under the assumption that the right hand side is well-defined.

- Consider a sequence of random numbers $Y_1, \dots, Y_N$ drawn from a certain distribution given by the cdf F. Recall the concept of empirical distribution $\hat{F}$. Argue that $\hat{F}$ is a random square integrable function. What is the expectation of $\hat{F}$?
- Find the covariance operator for the empirical distribution $\hat{F}$.
- Write the covariance operator in the case when the original sample has been drawn from uniform distribution.

Problem 2 – The eigenfunction for the covariance operator

Recall the definition of the covariance matrix for a random vector. This matrix is positive definite.

- Formulate the condition of the positive definiteness for a matrix and suggest its equivalent for the covariance operator for functional data.
- For a matrix, define the concept of eigenvalue/vector. What can you say about the eigenvalues and eigenvector of positive definite matrices.
- What would be an equivalent of eigenvalue/vector pair for the covariance operator?

The eigenfunctions v_i of the covariance operator C allow for the optimal representation of corresponding X. The functional principal components (FPC) are defined as the eigenfunctions of the covariance operator C of X. The representation

$$X = \sum_{i=1}^{\infty} \langle X, v_i\rangle v_i \tag{2}$$

is called the Karhunen-Loeve expansion.

The inner product $\langle X_i, v_j\rangle = \int X_i(t) v_j(t) dt$ is called the jth score of X_j and is interpreted as the weight of the contribution of the FPC v_j to the curve X_j. We often estimate the eigenvalues and eigenfunctions of C, but the interpretation of these quantities as parameters, and their estimation, must be approached with care. The eigenvalues must be identifiable, so we must assume that $\lambda_1 > \lambda_2 > ....$ In practice, we can estimate only the p largest eigenvalues, and assume that $\lambda_1 > \lambda_2 > ... > \lambda_p > \lambda_{p+1}$ which implies that the first p eigenvalues are nonzero. The eigenfunctions v_j are defined by $C(v_j) = \lambda_j v_j$, so if v_j is an eigenfunction, then so is av_j, for any nonzero scalar a (by definition, eigenfunctions are nonzero). The v_j are typically normalized, so that $\|v_j\| = 1$, but this does not determine the sign of v_j. Thus if $\hat{v}_j$ is an estimate computed from the data, we can only hope that $\hat{c}_j\hat{v}_j$ is close to v_j, where

$$\hat{c}_j = sign(\langle \hat{v}_j, v_j \rangle)$$

Note that $\hat{c}_j$ cannot be computed form the data, so it must be ensured that the statistics we want to work with do not depend on $\hat{c}_j$'s.
We define the estimated eigenelements by:

$$\hat{C}_N(\hat{v}_j) = \hat{\lambda}_j \hat{v}_j \quad j = 1, 2, ..., N \tag{3}$$

- Suppose that your data consists from the empirical distribution functions $\hat{F}_i$, $i = 1, \ldots, n$, that are computed for samples of the size 1000 on different occasions for a certain variable over the population (the original data on which the empirical distributions were evaluated are irrelevant here).
- Discuss the eigenfunctions and their empirical estimates for this case.

Problem 3 – Brownian Bridge Recall the concept of Brownian Motion or Wiener process (limit of random walks plus central limit theorem). A Brownian bridge is a continuous-time stochastic process $B(t)$ whose probability distribution is the conditional probability distribution of a Wiener process $W(t)$ subject to the condition that $W(T) = 0$, so that the process is pinned at the origin at both $t = 0$ and $t = T$. More precisely:

$$B_t := (W_t \mid W_T = 0), \quad t \in [0, T]$$

Alternatively,

$$B_T(t) = W(t) - tW(T)/T, \ t \in [0, T],$$

where W is a standard Brownian motion.

1. Find the covariance operator for this process.
2. Do you see any connection with empirical distribution?

In the following figure, we have the graphs of brownian bridge being generated.

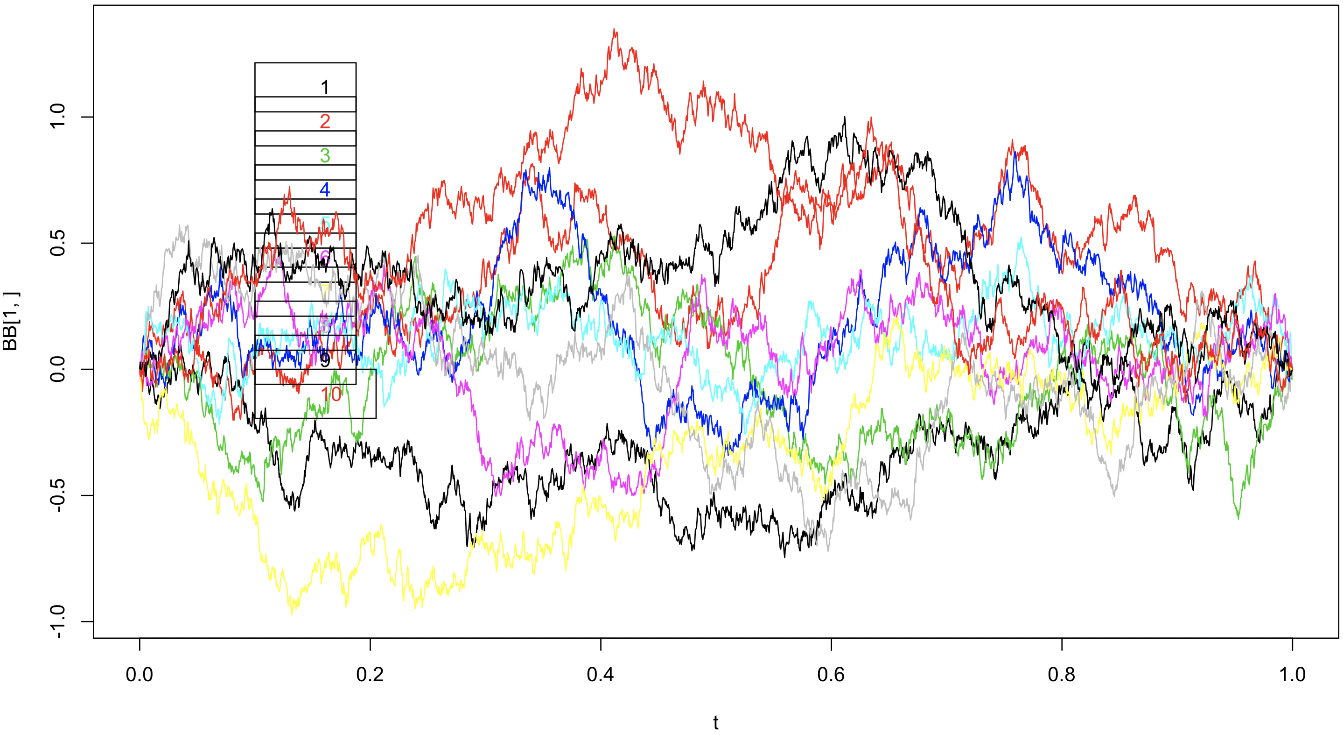

Figure 1: Brownian Motion

Figure 1 has been generated using the code in R described below.

```
1  #Simulation an independent sample of a Brownian bridge
2  #over an equidistant grid
3
4  n=2000 #size of the equidistant one dimensional grid
5  MC=10 #Monte Carlo sample size
6  t=matrix(seq(0,1,by=1/n),nrow=1) #grid
7
8
9  ZZ=matrix(rnorm(n*MC),ncol=n)/sqrt(n) #random noise
10
11 #Simulating Brownian Bridge that starts from zero
12 ZeC=matrix(rep(0,MC),ncol=1)
13 BB=cbind(ZeC,t(apply(ZZ,1,cumsum)))-matrix(apply(ZZ,1,sum),ncol=1)%*%t
14
15 #Ploting trajectories
16 quartz()
17 plot(t,BB[1,],type=’l’,ylim=c(min(BB),max(BB)))
18 legend(0.1,max(BB)-1*0.1*max(BB),1,text.col =1)
19 for(i in 2:MC)
20 {
21   lines(t,BB[i,],type=’l’,col=i)
22   legend(0.1,max(BB)-i*0.1*max(BB),i,text.col =i)
23 }
```

Here, the code is pretty straight forward, we establish the grid first (line 4 to 6), generate a random noise (line 9) and pin it to 0 at time 0 (line 11 to 13). Only one sample has been generated in this case, this can be modified depending on the user's needs.

Now that we can generate the Brownian bridge, we can practice some concepts and their estimates on this example. In fact the Brownian bridge is one of few cases for which the eigenfunctions are known in an analytical form

$$v_k(t) = \sqrt{2}\sin k\pi t$$

$$\lambda_k = \frac{1}{k^2\pi^2}$$

- Verify that the above functions and values are indeed eigenvalue/function pairs.
- Verify that they are orthogonal.
- Write the process in their terms.
- Write the covariance operator in their terms.