Lecture 4: Two sample inference for Functional Data and Changepoint Detection

Elements of FDA

Autumn 2020

In this section we investigate methods for testing equality of means and covariances for two samples of functional data. An important contribution has been made by Benko et al. [2009] who developed bootstrap procedures for testing the equality of mean functions and their functional principal components. The work of Panaretos et al. [2010] details two-sample testing for covariance functions. Broadly this work follows Horváth and Kokoszka [2012]. We also discuss the problem of detecting changes in mean in streams of independent $L^2([0,1])$-valued data.

1 Two Sample Tests for Equality of Mean Functions

We consider two samples $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_M$. We assume that

$$\begin{aligned} X_i(t) &= \mu(t) + \epsilon_i(t), \quad 1 \leq i \leq N, \\ Y_j(t) &= \mu^*(t) + \epsilon_j^*(t),, \quad 1 \leq j \leq M. \end{aligned}$$

We wish to test the null hypothesis

$$H_0: \quad \mu = \mu^* \text{ in } L^2,$$

against the alternative that H_0 is false. We shall assume that the two samples are independent. More specifically, we make the following assumptions

A1: $\epsilon_1, \ldots, \epsilon_N$ are independent and identically distributed with $\mathbb{E}[\epsilon_i] = 0$ and

A2: $\mathbb{E}\|\epsilon_i\|^4 < \infty$.

A3: Similarly $\epsilon_1^*, \ldots, \epsilon_M^*$ are independent and identically distributed with $\mathbb{E}[\epsilon_i^*] = 0$ and

A4: $\mathbb{E}\|\epsilon_i^*\|^4 < \infty$.

1.1 Approach 1: The Norm Approach

The sample means for each sample are given by

$$\overline{X}_N(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t), \quad \text{and } \overline{Y}_M(t) = \frac{1}{M}\sum_{i=1}^{M} Y_i(t),$$

which are unbiased estimators for $\mu(t)$ and $\mu^*(t)$ respectively, it is natural to reject the null hypothesis if

$$U_{N,M} = \frac{NM}{N+M}\int_0^1 (\overline{X}_N(t) - \overline{Y}_M(t))^2 \, dt$$

We start with establishing the convergence of $U_{N,M}$ under H_0.

Theorem 1.1. *If H_0 holds and the assumptions $A1, \dots, A4$ hold and*

$$\frac{N}{N+M} \to \theta, \text{ for some } 0 \le \theta \le 1,$$

then

$$U_{N,M} \xrightarrow{D} \int_0^1 \Gamma^2(t)\, dt,$$

where $\Gamma(t)$ is a Gaussian process on $[0,1]$ with mean zero and

$$\mathbb{E}[\Gamma(t)\Gamma(s)] = (1-\theta)c(t,s) + \theta c^*(t,s),$$

where $c(t,s) = Cov(X_1(t), X_1(s))$ and $c^*(t,s) = Cov(Y_1(t), Y_1(s))$.

Proof. By the CLT from the previous lecture, there exist two Gaussian processes Γ_1 and Γ_2 with zero means and covariances C and C^* such that

$$\left(N^{-1/2} \sum_{1 \le i \le N} (X_i - \mu), M^{-1/2} \sum_{1 \le j \le M} (Y_i - \mu^*) \right),$$

converges weakly in L^2 to (Γ_1, Γ_2). This proves $\Gamma = (1-\theta)\Gamma_1 + \theta\Gamma_2$. □

According to the KL expansion we can assume that

$$\Gamma(t) = \sum_{k=1}^{\infty} \tau_k^{1/2} N_k \psi_k(t),$$

where the N_k are independent and standard normal random variables $\tau_1 \ge \tau_2 \ge \cdots$ and $\psi_1, \psi_2, \dots$ are the eigenvalues and eigenfunctions of the operator determined by $(1-\theta)c + \theta c^*$. Clearly

$$\int_0^1 \Gamma^2(t)\, dt = \sum_{k=1}^{\infty} \tau_k N_k^2,$$

where the N_k are standard normal random variables $\tau_1 \ge \tau_2 \ge \cdots$ and $\psi_1, \dots,$ are the eigenvalues and eigenfunctions of the operator $\mathcal{C}_\theta$ defined by:

$$\mathcal{C}_\theta f(t) = \int_0^1 ((1-\theta)c(t,s) + \theta c^*(t,s)) f(s)\, ds.$$

Clearly

$$\int_0^1 \Gamma^2(t)\, dt = \sum_{k=1}^{\infty} \tau_k N_k^2,$$

so to provide an approximation for $\int_0^1 \Gamma^2(t)\,dt$ we only need to estimate the $\tau_k's$. This can be done using the eigenvalues of the empirical covariance function

$$\widehat{z}_{N,M} = \frac{M}{M+N}\frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \overline{X}_N(t))(X_i(s) - \overline{X}_N(s))$$
$$+ \frac{N}{M+N}\frac{1}{M}\sum_{j=1}^{M}(Y_j(t) - \overline{Y}_M(t))(Y_j(s) - \overline{Y}_M(s)).$$

The sum $\sum_{k=1}^{d} \widehat{\tau}_k N_k^2$ provides an approximation to the limit if d is sufficiently large.

Theorem 1.2. *Suppose that the assumptions $A1, \ldots, A4$ hold and*

$$\frac{N}{N+M} \to \theta, \text{ for some } 0 \le \theta \le 1,$$

then

$$\int_0^1 (\mu(t) - \mu^*(t))^2\,dt > 0,$$

then $U_{M,N} \xrightarrow{P} \infty$.

1.2 Approach 2: The Spectral Approach

Here we present a second approach based on projection onto finite dimensional subspaces. This allows us to derive a test statistic whose asymptotic distribution is the standard chi–square distribution, which does not depend on any parameters. However, it does involve a truncation level which turns out to be the number of its degrees of freedom. Choosing this trunction is problem dependent, and will contribute to testing errors if chosen poorly.

To this end, we use projections onto the space determned by the leading eigenfunctions of the operator $Z = (1-\theta)C + \theta C^*$. We assume that the eigenvalues of Z satisfy

$$\tau_1 > \tau_2 > \ldots > \tau_d > \tau_{d+1}.$$

The corresponding eigenfunctions are $\phi_1, \ldots, \phi_{d+1}$. We cannot typically have direct access to these, so instead we use the approximate eigenfunctions $\widehat{\phi}_i$, $1 \le i \le d$, and define

$$\widehat{a}_i = \langle \overline{X}_N - \overline{Y}_M, \widehat{\phi} \rangle, 1 \le i \le d,$$

and introduce $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_d)^\top$. We show that under the conditions of Theorem 1.1 the vector $(NM/(N+M))^{1/2}\widehat{a}$ is approximately d-variable normal up to random signs. The asymptotic covariance of $(NM/(N+M))^{1/2}\widehat{a}$ is Q_{ij}, where

$$Q_{ij} = (1-\theta)\mathbb{E}[\langle X_1 - \mu, \phi_i \rangle \langle X_1 - \mu, \phi_j \rangle + \theta \mathbb{E}\langle Y_1 - \mu^*, \phi_i \rangle \langle Y_1 - \mu^*, \phi_j \rangle.$$

It is easy to see that

$$
\begin{aligned}
Q(i,j) &= \int_0^1 \int_0^1 (1-\theta)\mathbb{E}[(X_1(t)-\mu(t))(X_1(s)-\mu(s))]\phi_i(t)\phi_j(s)\,ds\,dt \\
&+ \int_0^1 \int_0^1 (1-\theta)\mathbb{E}[(Y_1(t)-\mu^*)(Y_1(s)-\mu^*(s))]\phi_i(t)\phi_j(s)\,ds\,dt \\
&= \int_0^1 \int_0^1 z(t,s)\phi_i(t)\phi_j(s)\,dt\,ds \\
&= \begin{cases} \tau_i, & \text{if } i=j \\ 0 & \text{if } i \neq j. \end{cases}
\end{aligned}
$$

In the light of this, testing procedures can be based on the statistics

$$
\begin{aligned}
T^{(1)}_{N,M} &= \frac{NM}{N+M}\sum_{k=1}^{d} \frac{\widehat{a}_k^2}{\widehat{\tau}_k} \\
T^{(2)}_{N,M} &= \frac{NM}{N+M}\sum_{k=1}^{d} \widehat{a}_k^2.
\end{aligned}
$$

We have the following theorem, characterising consistency in the large N limit

Theorem 1.3. *If H_0, the assumptions $A1-A4$ hold and*

$$\frac{N}{N+M} \to \theta, \text{ for some } 0 \leq \theta \leq 1.$$

Then

$$T^{(1)}_{N,M} \xrightarrow{d} \chi^2(d),$$

and

$$T^{2}_{N,M} \xrightarrow{d} \sum_{k=1}^{d} \tau_k N_k^2,$$

where $\xi^2(d)$ is a chi-square random variable with d degrees of freedom, and $N_1, N_2, \ldots, N_d$ are independent standard normal random variables.

It is clear that $T^{(2)}_{N,M}$ is a projection version of $U_{N,M}$ we only use the first d terms in L^2 expansion of $\overline{X}_N - \overline{Y}_M$. The statistic $T^{(1)}_{N,M}$ is an asymptotically distribution free modification of $T^{(2)}_{N,M}$ and hence $U_{N,M}$.

The consistency of testing procedures based on $T^{(1)}_{N,M}$ and $T^{(2)}_{N,M}$ can also be easily established.

Theorem 1.4. *Suppose assumptions $A1 - A4$ hold and*

$$\frac{N}{N+M} \to \theta, \text{ for some } 0 \leq \theta \leq 1.$$

and

$$\tau_1 > \tau_2 > \ldots > \tau_d > \tau_{d+1}.$$

If $\mu - \mu^*$ is not orthogonal to the first d components $\phi_1, \ldots, \phi_d$. Then $T^{(1)}_{N,M} \xrightarrow{P} \infty$ and $T^{(2)}_{N,M} \xrightarrow{P} \infty$.

The difference between tests on $U_{N,M}$ and $T^{(1)}_{N,M}$ and $T^{(2)}_{N,M}$ is that the last only see the difference between μ and μ^* in a d-dimensional subspace. If $\mu = \mu^*$ in this subspace, then H_0 will not be rejected. We're not typically worried by this if the first d eigenfunctions explain a large percentage of variance of the difference.

1.3 Simulated Example

In the associated R file `tests_two_sample_mean.R` we apply the two tests for equal means on a simulated dataset. We compare 500 trajectories of a standard Brownian bridge, and compare with 500 trajectories of a Brownian bridge on $[0, 1]$ with a mean of the form

$$\mu(t) = c \sin(\pi(k - 1/2)x),$$

where $c > 0$. See Figure 1 In the file we compute both tests, for increasing c, and demonstrate (roughly) the testing power of both. We note there is no obvious advantage of one over the other in terms of testing power.

2 Two Sample Tests for Equality of Covariance Functions

Consider two samples $X_1, X_2, \ldots, X_N$ and $X_1^*, X_2^*, \ldots, X_M^*$. The functions in each sample are iid mean zero elements of $L^2([0, 1])$ and the two sets are independent. We consider the covariance operators:

$$C(x) = \mathbb{E}[\langle X, x\rangle X], \text{ and } C^*(x) = \mathbb{E}[\langle X^*, X\rangle X^*],$$

where $X_i \sim X$ and $X_j \sim X^*$ for $i = 1, \ldots, N$, $j = 1, \ldots, M$ and where X, X^* are Gaussian random variables taking values in $L^2([0, 1])$. We want to test

$$H_0 : \quad C = C^* \text{ versus } H_A : C \neq C^*.$$

Let $\widehat{C}$ and $\widehat{C}^*$ be the empirical counterparts of C and $C*$. Let $\widehat{R}$ be the "pooled" empirical

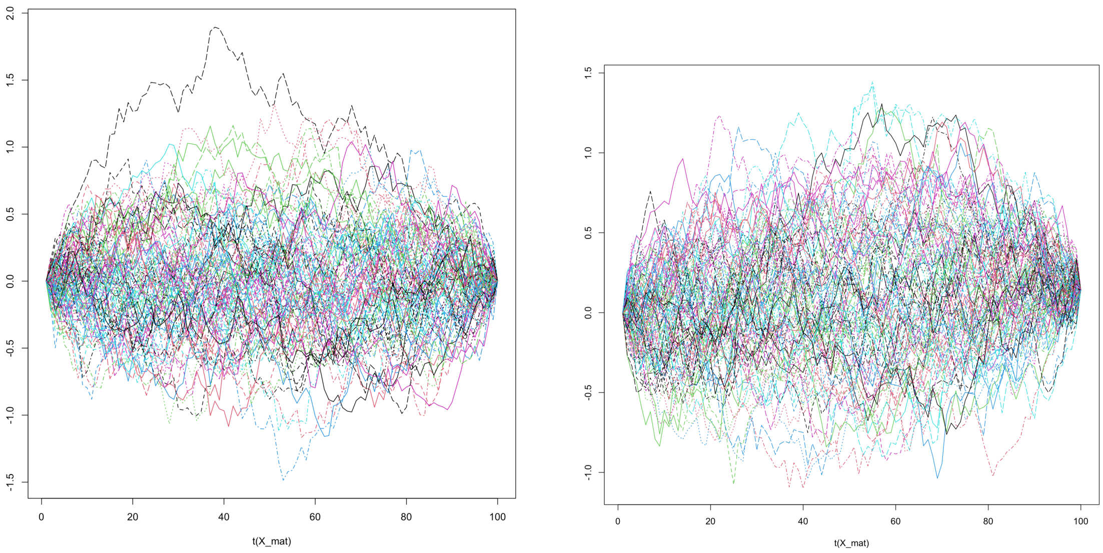


Figure 1: Brownian motion vs mean perturbed Brownian motion

covariance operator, i.e.

$$\begin{aligned}\widehat{R}(x) &= \frac{1}{N+M}\left(\sum_{i=1}^{N}\langle X_i, x\rangle X_i + \sum_{j=1}^{M}\langle X_j^*, x\rangle X_j^*\right) \\ &= \theta\widehat{C}(x) + (1-\theta)\widehat{C}^*(x), \quad x \in L^2,\end{aligned}$$

and $\theta = N/(N+M)$. The operator $\widehat{R}$ has $N+M$ eigenfunctions, which are denoted $\widehat{\phi}_k$. We set

$$\widehat{\lambda}_k = \frac{1}{N}\sum_{n=1}^{N}\langle X_n, \widehat{\phi}_k\rangle^2, \quad \widehat{\lambda}_k^* = \frac{1}{M}\sum_{m=1}^{M}\langle X_m^*, \widehat{\phi}_k\rangle^2.$$

Note that these aren't the eigenvalues, but rather the sample variances of the coefficients of X and X^* with respect to the ONB $\widehat{\phi}_k$ formed by the eigenfunctions $\widehat{R}$. We can now define the following test statistic.

$$\widehat{T} = \frac{N+M}{2}\widehat{\theta}(1-\widehat{\theta})\sum_{i,j=1}^{p}\frac{\langle(\widehat{C}-\widehat{C}^*)\widehat{\phi}_i, \widehat{\phi}_j\rangle^2}{(\widehat{\theta}\widehat{\lambda}_i + (1-\widehat{\theta})\widehat{\lambda}_i^*)(\widehat{\theta}\widehat{\lambda}_j + (1-\widehat{\theta})\widehat{\lambda}_j^*)}$$

Theorem 2.1. *Suppose that X and X^* are Gaussian random variables taking values in $L^2([0,1])$. Suppose that $\widehat{\theta} \to \theta \in (0,1)$ as $N \to \infty$. Then, under the null hypothesis H_0:*

$$\widehat{T} \xrightarrow{d} \chi^2_{p(p+1)/2},$$

as $N, M \to \infty$, where $\chi^2_{p(p+1)/2}$ denotes a chi-square random variable with $p(p+1)/2$ degrees of freedom.

Proof. We shall make use of Theorem 3.7 of Lecture 2 of the notes. Under the fourth moment assumptions on X and X^* we have that

$$N^{1/2}(\widehat{C} - C) \xrightarrow{d} Z_1, \text{ and } M^{1/2}(\widehat{C}^* - C^*) \xrightarrow{d} Z_2,$$

where Z_1 and Z_2 are independent Gaussian elements with the same covariance operator $\mathcal{M}$ under the null hypothesis H_0. By Theorem 1.2 of Lecture 3 we also have that $\widehat{\phi}_i \xrightarrow{P} v_i$, the eigenfunctions of C (and of C^*), where we might have to adjust the sign of the $\widehat{\phi}_i$ to make this work. Combining we obtain

$$W_{N,M}(i,j) = \sqrt{(M+N)\widehat{\theta}(1-\widehat{\theta})}\langle(\widehat{C} - \widehat{C}^*)\widehat{\phi}_i, \widehat{\phi}_j\rangle,$$

so that

$$\widehat{T} = \frac{\sum_{i,j=1}^{p} W_{N,M}^2(i,j)}{2(\widehat{\theta}\widehat{\lambda}_i + (1-\widehat{\theta})\widehat{\lambda}_i^*)(\widehat{\theta}\widehat{\lambda}_j + (1-\widehat{\theta})\widehat{\lambda}_j^*)}.$$

Under H_0 we have

$$W_{N,M}(i,j) = \left\langle \left[(1-\widehat{\theta})^{1/2} N^{1/2}(\widehat{C} - C) - \widehat{\theta}^{1/2} M^{1/2}(\widehat{C}^* - C^*)\right] \widehat{\phi}_i, \widehat{\phi}_j \right\rangle.$$

Therefore in the large sample limit we have

$$W_{N,M}(i,j) \xrightarrow{d} \langle Z v_i, v_j\rangle,$$

where $Z = (1-\theta)^{1/2} Z_1 - \theta^{1/2} Z_2$. Note that the covariance of Z is also equal to $\mathcal{M}$. A straightforward calculation yields that

$$\widehat{T} \xrightarrow{d} \sum_{k=1}^{n} \frac{\langle Z v_k, v_k\rangle^2}{2\lambda_k^2} + \sum_{k<n} \frac{\langle Z v_k, v_n\rangle^2 + \langle Z v_n, v_k\rangle^2}{2\lambda_k \lambda_n}.$$

The "hard part" is now to identify the form of the operator Z, this entails examining the form of the covariance operator $\mathcal{M}$. Details can be found in [Horvath]. Following a calculation we realise that

$$Z = \sqrt{2}\sum_{i=1}^{\infty} \lambda_i \xi_{ii} V_i i + \sum_{i<j} \sqrt{\lambda_i \lambda_j}(V_{ij} + V_{ji}),$$

where ξ_{ij} are standard iid normal. Using the fact that v_i form a basis we see that

$$\langle Z v_k, v_n\rangle = c_{k,n}\xi_{k,n},$$

where

$$c_{k,n} = \begin{cases} \sqrt{2}\lambda_k & \text{if } k = n \\ \sqrt{\lambda_k \lambda_n} & \text{if } k < n \\ \sqrt{\lambda_k \lambda_n} & \text{if } k > n \end{cases}.$$

Combining we obtain that

$$\widehat{T} \xrightarrow{d} \sum_{k=1}^{p} \xi_{kk}^2 + \frac{1}{2}\sum_{k<p} \xi_{kn}^2 + \xi_{nk}^2 \equiv \sum_{k=1}^{p} \xi_{kk}^2 + \sum_{k<p} \xi_{kn}^2 \equiv \chi^2_{p(p+1)/2}.$$

□

3 Detecting Changes in Mean Function

Having presented the two-sample tests for different means and different covariances of functional data, in this section we present a method for detecting changes in the mean of functional observations. At the heart of it lies a significance test for the null hypothesis of a constant functional mean against the alternative of a changing mean. In this case, the change can be not only in the average level of this function, but also in its shape.

It is important to distinguish between a change point problem as is described here and the problem of testing for equal means of the previous sections. In the formal setting, one already knows which population each observation belongs to. In the change point setting, we do not have any partition of the data into several sets with possibly different means. The change can occur at any point, and we want to test if it occurs or not, and if it does, to estimate the point of change.

Such change point methodology is often applied to time series of average annual temperatures at specific locations. Detecting a change point in mean in such a series indicates that the assumption of a constant mean function for the whole series is not tenable. For the daily temperature data, a change in shape may mean, for example, that while the overall annual average stays the same, summers may become warmer and winters colder. Note that, we shall be making the simplifying assumption that observations are independent. This assumption may appear quite strong, but is often approximately satisfied, and allows to focus on the aspect of the methodology directly related to change point detection. Changepoint methods adapted to dependent data have been studied previously, see for example Chapter 16 of Horváth and Kokoszka [2012].

To this end, we assume that the observations $X_i \in L^2([0,1])$ are independent. We wish to test if their mean remains constant in i, that is we wish to test the null hypothesis:

$$H_0 : \quad \mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_N].$$

Note that under H_0, we do not specify the value of the common mean. The proposed test has a particularly good power against the alternative in which the data can be divided into several consecutive segments, and the mean is constant within each segment, but changes from segment to segment.

Under the null hypothesis, each functional observation is given by

$$X_i(t) = \mu(t) + Y_i(t), \quad \mathbb{E}[Y_i] = 0.$$

We shall assume that the mean $\mu \in L^2([0,1])$ and moreover that the errors are IID second-order integrable elements in $L^2([0,1])$, i.e. $\mathbb{E}\|Y_i\|^2 < \infty$. This implies that the covariance function

$$c(t,s) = \mathbb{E}[Y_i(t)Y_i(s)], \quad t,s \in [0,1],$$

is square integrable. Consequently by Mercer's theorem

$$c(t,s) = \sum_{1\le k<\infty} \sum \lambda_k v_k(t) v_k(s),$$

and the Karhunen-Loeve expansion

$$Y_i(t) = \sum_{1\le l<\infty} \zeta_{l,i} v_l(t),$$

where v_k are eigenfunctions of the covariancer operator with kernel $c(t, s)$and $\{\zeta_{l,i}\}$ are independent sequences, such that random variables within the same sequence are uncorrelated, have mean zero and variance λ. This sum converges with probability one.

Recall that the estimated eigenelements are defined by

$$\int \widehat{c}(t,s)\widehat{v}_l \, ds = \widehat{\lambda}_l \widehat{v}_l, \quad l = 1, 2, \ldots,$$

and

$$\widehat{c}(t,s) = \frac{1}{N} \sum_{1 \le i \le N} (X_i(t) - \overline{X}_N(t))(X_i(s) - \overline{X}_N(s)),$$

and

$$\overline{X}_N(t) = \frac{1}{N} \sum_{1 \le i \le N} X_i(t).$$

We make the following assumptions which allow us to appeal to the results in Lecture 3 to obtain consistency of the empirical eigenelements.

Condition 3.1. The eigenvalues λ_l satisfy, for some $d > 0$

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > \lambda_{d+1}.$$

Condition 3.2. The Y_i satisfy $\mathbb{E}\|Y_i\|^4 < \infty$.

By Theorem 1.3 of Lecture 3, we have for each $k \le d$:

$$\lim_{N\to\infty} N\mathbb{E}\|v_k - \widehat{v}_k\|^2 < \infty \text{ and } \limsup_{N\to\infty} N\mathbb{E}|\lambda_k - \widehat{\lambda}_k|^2 < \infty,$$

where we possibly flipped the signs of the v_k's to make these limits true. We first study a test under the alternative of one change point.

Condition 3.3. The observations follow the model

$$X_i(t) = \begin{cases} \mu_1(t) + Y_i(t), & 1 \le i \le k^* \\ \mu_2(t) + Y_i(t), & k^* < i \le N, \end{cases}$$

in which the Y_i satisfy the previous assumptions, the mean functions μ_1 and μ_2 are in $L^2([0,1])$ and

$$k^* = [n\theta] \text{ for some } 0 < \theta < 1.$$

We now specify the test procedure. To this end, denote

$$\widehat{\mu}(t) = \frac{1}{k} \sum_{1 \le i \le k} X_i(t), \quad \widetilde{\mu}(t) = \frac{1}{N-k} \sum_{k < i \le N} X_i(t).$$

If the mean is constant, the difference $\Delta_k(t) = \widehat{\mu}_k(t) - \widetilde{\mu}_k(t)$ is small for $1 \le k < N$ for all $t \in [0,1]$. However, $\Delta_k(t)$ can become large due fluctuations if k is close to 1 or N. It is therefore usual to work with the sequence

$$P_k(t) = \sum_{1 \le i \le k} X_i(t) - \frac{k}{N} \sum_{1 \le i \le N} X_i(t) = \frac{k(N-k)}{N} \left[\widehat{\mu}(t) - \widetilde{\mu}_k(t) \right],$$

in which the variability at the end points is weighted down by a parabolic weight function. If the mean changes at some point, the difference $P_k(t)$ will be large for some values of k and t. Since the observations are in an infinite dimensional domain, we work with the projections of the functions P_k on the principal components of the data.

To this end, onsider thus the scores corresponding to the largest d eigenvalues:

$$\widehat{\xi}_{l,i} = \int (X_i(t) - \overline{X}_N(t)) \widehat{v}_l(t)\, dt, \quad i = 1, 2, \ldots, N, l = 1, 2, \ldots, d.$$

Note that the value of $P_k(t)$ does not change if the $X_i(t)$ are replaced by $X_i(t) - \overline{X}_N(t)$. Consequently, setting $k = [Nx]$, $x \in (0,1)$ we obtain

$$\int \left[\sum_{1 \le i \le Nx} X_i(t) - \frac{[Nx]}{N} \sum_{1 \le i \le N} X_i(t) \right] \widehat{v}_l(t)\, dt = \sum_{1 \le i \le Nx} \widehat{\xi}_{l,i} 0 \frac{[Nx]}{N} \sum_{1 \le i \le N} \widehat{xi}_{l,i}.$$

Thisidentity shows that scores canbe used for testing for shifts in the mean function. This motivates the following test statistic:

$$T_N(x) = \frac{1}{N} \sum_{l=1}^{d} \widehat{\lambda}_l^{-1} \left(\sum_{1 \le i \le Nx} \widehat{\xi}_{l,i} - x \sum_{1 \le i \le N} \widehat{\xi}_{l,i} \right)^2. \tag{1}$$

Theorem 3.1. *Suppose the above conditions hold. Then under* H_0

$$T_N(x) \xrightarrow{d} \sum_{1 \le l \le d} B_l^2(x), \quad 0 \le x \le 1,$$

in the Skorokhod topology of $D[0,1]$*, where* $B_1, \ldots, B_d$ *denote independent standard Brownian Bridges.*

The convergence in the above theorem implies that $U(T_N) \xrightarrow{d} U(\sum_{1 \le l \le d} B_l^2(\cdot))$ for any continuous functional $U : D[0,1] \to \mathbb{R}$. Applying integral or max functionals leads to useful statistics. We focus on the integral of the squared function, known as the Cramer–von–Mises functional, which is known to produce effective tests. To this end, consider the convergence

$$\int_0^1 T_N(x)\,dx \xrightarrow{d} \int_0^1 \sum_{1\leq l\leq d} B_l^2(x)\,dx,$$

which we can rewrite as

$$S_{N,d} = \frac{1}{N^2}\sum_{l=1}^{d} \widehat{\lambda}_l^{-1} \sum_{k=1}^{N} \left(\sum_{1\leq i\leq k} \widehat{\xi}_{l,i} - \frac{k}{N} \sum_{1\leq i\leq N} \widehat{\xi}_{l,i} \right)^2 \xrightarrow{d} \int_0^1 \sum_{1\leq l\leq d} B_l^2(x)\,dx.$$

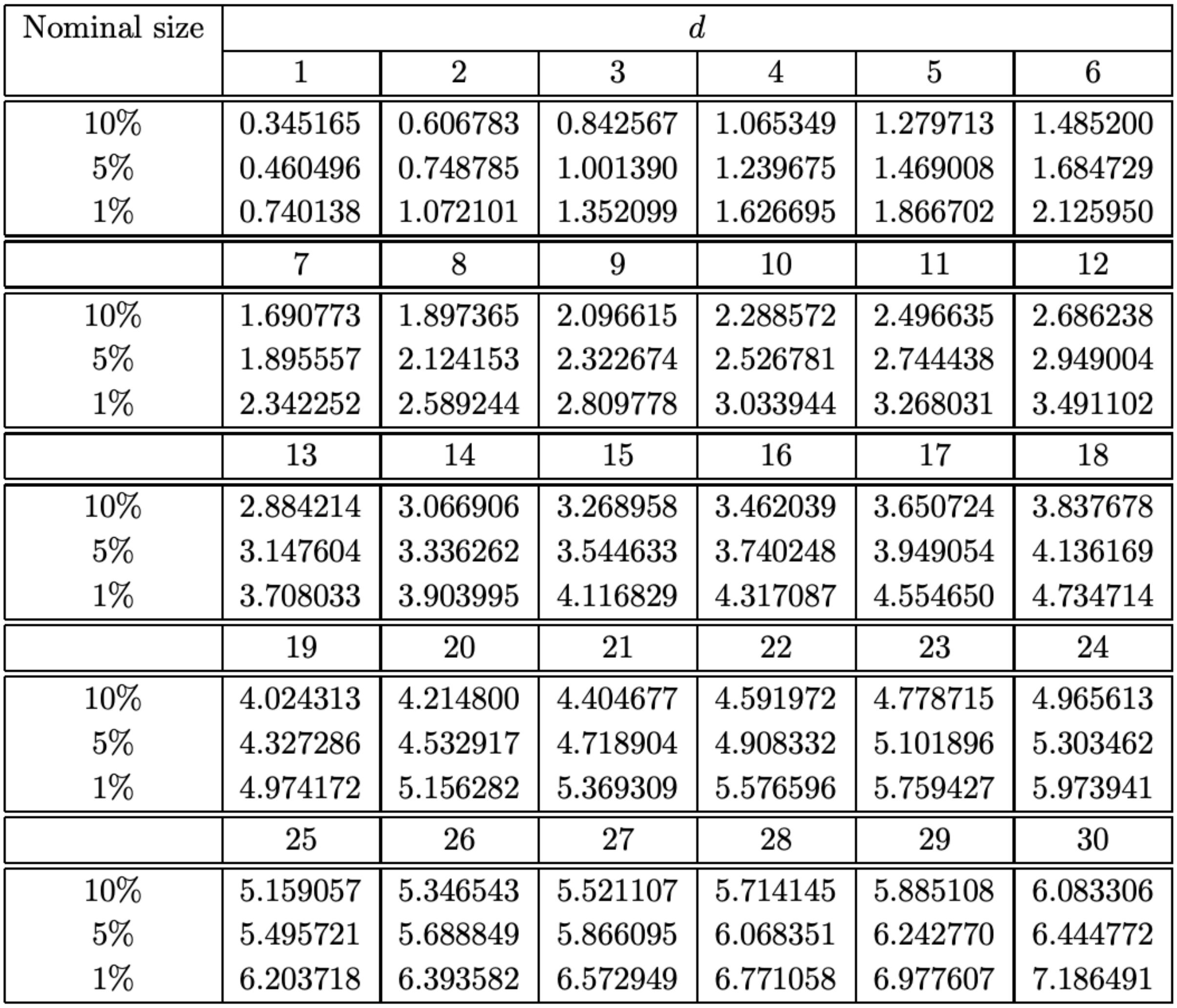
This is a promising approach as the distribution of the random variable

$$K_d = \int_0^1 \sum_{1\leq l\leq d} B_l^2(x)\,dx,$$

was derived by Kiefer [1959], obtained by a series expansion of K_d. Denoting by $c_d(\alpha)$ its $(1-\alpha)th$ quantile, the test rejects H_0 if $S_{N,d} > c_d(\alpha)$. Simulated critical values can be found in Figure 2.

3.1 Code for Equal Mean Two-Sample Tests

```
library(sde)
library(CompQuadForm)

N <- 100
reps <-100
times<-seq(0,1,length=N)


Z = lapply(seq(1,reps),function(x) gen_data(x,c1=0.1))
X_mat = do.call(rbind, Z)
matplot(t(X_mat), type='l')


gen_data<-function(x, N=100, k=1, c1=0){
  mu<-function(x){c1*sqrt(2)*sin((k-1/2)*pi*x)}
  mu_vec<-mu(times)

  y=mu_vec + BBridge(x=0, y=0, t0=0, T=1, N=(N-1))

 return(y)
}

c1s = seq(0,0.1,0.01)
experiments = length(c1s)
norm_pvals = rep(0.0, experiments)
```



Nominal size	*d*					
	1	2	3	4	5	6
10%	0.345165	0.606783	0.842567	1.065349	1.279713	1.485200
5%	0.460496	0.748785	1.001390	1.239675	1.469008	1.684729
1%	0.740138	1.072101	1.352099	1.626695	1.866702	2.125950
	7	8	9	10	11	12
10%	1.690773	1.897365	2.096615	2.288572	2.496635	2.686238
5%	1.895557	2.124153	2.322674	2.526781	2.744438	2.949004
1%	2.342252	2.589244	2.809778	3.033944	3.268031	3.491102
	13	14	15	16	17	18
10%	2.884214	3.066906	3.268958	3.462039	3.650724	3.837678
5%	3.147604	3.336262	3.544633	3.740248	3.949054	4.136169
1%	3.708033	3.903995	4.116829	4.317087	4.554650	4.734714
	19	20	21	22	23	24
10%	4.024313	4.214800	4.404677	4.591972	4.778715	4.965613
5%	4.327286	4.532917	4.718904	4.908332	5.101896	5.303462
1%	4.974172	5.156282	5.369309	5.576596	5.759427	5.973941
	25	26	27	28	29	30
10%	5.159057	5.346543	5.521107	5.714145	5.885108	6.083306
5%	5.495721	5.688849	5.866095	6.068351	6.242770	6.444772
1%	6.203718	6.393582	6.572949	6.771058	6.977607	7.186491

Figure 2: Simulated critical values of the distribution of K_d, obtained from Berkes et al. [2009].

```
PC_pvals= rep(0.0, experiments)

for (i in 1:experiments){

  Z = lapply(seq(1,reps),function(x) gen_data(x,c1=c1s[i]))
  X_mat = do.call(rbind, Z)

  # Estimate functional parameters
  X.f<-Data2fd(times,t(X_mat))
  muhat<-mean.fd(X.f)
  X.pca<-pca.fd(X.f,nharm=20)
  lambda<-X.pca$values
  scores<-X.pca$scores
  v<-X.pca$harmonics


  # Compute tests and p-values
  # PCA test with 3 principal components
  TPC3  <-N*sum(inprod(v[1:3],muhat)^2/lambda[1:3])
  PC_pvals[i]<-pchisq(TPC3,3,lower.tail=FALSE)

  # Norm test
  Tnorm<-N*sum(inprod(v[lambda[1:20]>0],muhat)^2)
  norm_pvals[i]<-imhof(Tnorm,lambda[lambda[1:20]>0])[[1]]
}

plot(c1s, norm_pvals,type='l')
lines(c1s, PC_pvals,col='blue')
abline(h=0.05)
```

References

Michal Benko, Wolfgang Härdle, Alois Kneip, et al. Common functional principal components. *The Annals of Statistics*, 37(1):1–34, 2009.

István Berkes, Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):927–946, 2009.

Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

J Kiefer. K-sample analogues of the kolmogorov-smirnov and cramér-v. mises tests. *The Annals of Mathematical Statistics*, pages 420–447, 1959.

Victor M Panaretos, David Kraus, and John H Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.