Chapter 2: Hilbert Space Model for Functional Data

Elements of Finite Data Analysis

January 2021

Most FDA tools make heavy use of Hilbert space theory, which is essentially a mathematical framework which permits one to treat functions like (infinite dimensional) vectors. In this section we survey some of the aspects of Hilbert space theory which are most relevant to functional data analysis. This is not intended to be a complete survey of Hilbert Space theory. The interested reader is invitated to consult Akhiezer and Glazman [2013] for a more complete exposition, or the monographs of Hsing and Eubank [2015] for an exposition tailored to FDA.

1 Review of Hilbert Space Theory

A vector space is defined over a field of scalars. We shall assume that the vector space is over the field of real numbers. A vector space is a set V whose elements are called vectors, and in which two operations are defined: addition ($x + y$, for $x, y \in V$) and scalar multiplication λx for $\lambda \in \mathbb{R}$, $x \in V$. The following are the axioms of a vector space.

This first set of axioms correspond to V being a commutative group under addition.

1. For every $x, y \in V$, $x + y \in V$ and $x + y = y + x$, $x + (y + z) = (x + y) + z$.
2. The space V contains a unique vector 0 such $x + 0 = x$ for every $x \in V$. For every $x \in V$, there exists a unique vector $-x$ such that $x + (-x) = 0$.

The second set of axioms relate to scalar multiplication

1. For every $a \in \mathbb{R}$ and $x \in V$ there exists $ax \in V$ and
2. $1x = x$.
3. $a(bx) = (ab)x$.
4. $a(x + y) = ax + ay$.
5. $(a + b)x = ax + by$.

All the usual intuitive properties follow from these axioms, e.g. $0x = 0 \in V$, etc. A set $U \subset V$ is called a *subspace* of V if it is a vector space with respect to the operations defined on V. A subset $U \subset V$ is a vector space if and only if, for all $a, b \in \mathbb{R}$, $ax + by \in U$. A set of vectors $x_1, \ldots, x_n$ are said to be linearly independent if

$$\lambda_1 x_1 + \ldots + \lambda_n x_n = 0 \Rightarrow \lambda_i = 0, i = 1, \ldots, n.$$

Otherwise the set of vectors is said to be linearly dependent.

Definition 1.1. A vector space V is called an *inner product space* if for each pair of vectors x, y there is a scalar $\langle x, y\rangle$ called the inner product of x and y which satisfies

1. $\langle x, y\rangle = \langle y, x\rangle$.
2. $\langle ax_1 + bx_2, y\rangle = a\langle x_1, y\rangle + b\langle x_2, y\rangle$;
3. $\langle x, x\rangle \geq 0$ with equalty if and only if $x = 0$.

An inner product defines an inner product norm defined by $\|x\| = \sqrt{\langle x, x\rangle}$.

Proposition 1.1. *The inner product defines a norm $|\cdot\|$ on the inner product space which has the following properties.*

1. $\|ax\| = |a|\|x\|$.
2. $|\langle x, y\rangle| \leq \|x\|\|y\|$.
3. $\|x + y\| \leq \|x\| + \|y\|$.
4. $\|x - y\|$ *defines a distance between x and y.*

The last part of this proposition implies that every inner product space is a normed space, which is in turn a metric space. The distinction between inner product spaces and Hilbert spaces is not algebraic, but topological. Recall from real-analysis the definition of a complete space:

Definition 1.2. A metric space in which every Cauchy sequence has a limit is *complete*.

For example, the unit interval $[0, 1]$ is complete, similarly $\mathbb{R}$ is complete. On the other hand the open interval $(0, 1)$ is not. Indeed, consider $x_n = 1 - 1/n$, $n \geq 1$. Then $|x_n - x_m| = |1/m - 1/n| \leq |1/m| + |1/n| \to 0$ as $m, n \to \infty$, so it is Cauchy, however, $x_n \to 1 \notin (0, 1)$, so that is not complete.

Definition 1.3. A complete inner product space is called a *Hilbert space*.

Example 1.1. The space l^2 is defined to be the set of all sequences $x = (x_1, x_2, \ldots)$ such that $\sum_{i=1}^{\infty} |x_i|^2 < \infty$. It is a vector space with operations $x + y = (x_1 + y_1, x_2 + y_2, \ldots)$ and $\lambda x = (\lambda x_1, \lambda x_2, \ldots)$. The inner product is defined by

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i.$$

The space l^2 is a Hilbert space with the above operations. It is an easy exercise to check that l^2 defines an inner product space. Completeness requires a bit more work. The interested reader can consult Kreyszig [1978] or Akhiezer and Glazman [2013] for a straightforward proof of this.

In this course, we will focus primarily on the space of square integrable real-valued functions defined on an interval (without loss of generality $[0, 1]$.).

Definition 1.4. The space $L^2([0, 1])$ is the collection of Lebesgue measurable real-valued functions x defined on the interval $[0, 1]$ such that

$$\int_0^1 x^2(t)\, dt < \infty.$$

The vector space operations are defined via

$$(x + y)(t) = x(t) + y(t), \quad (\lambda x)(t) = \lambda x(t).$$

The inner product is defined by

$$\langle x, y \rangle = \int_0^1 x(t) y(t)\, dt,$$

which induces the L^2-norm

$$\|x\| = \sqrt{\int_0^1 x(t)^2\, dt}.$$

The space $L^2([0, 1])$ with the above operations defines a Hilbert space. It is an easy exercise to check that L^2 defines an inner product space. Completeness requires a bit more work. The interested reader can consult Kreyszig [1978] or Akhiezer and Glazman [2013] for a straightforward proof of this.

In this space, the Cauchy-Schwartz inequality takes the form:

$$\left| \int_0^1 x(t) y(t)\, dt \right| \leq \left(\int x^2(t)\, dt \right)^{1/2} \left(\int_0^1 y(t)^2\, dt \right)^{1/2}$$

Note that one may define in an almost identical fashion the space $L^2(\mathcal{D})$ where $\mathcal{D}$ is any compact subset of $\mathbb{R}^d$. More generally one can define $L^2(\mathcal{D})$ for any closed subset of $\mathbb{R}^d$.

Example 1.2. The space $L^2(\mathbb{R})$ is the space of all functions on $\mathbb{R}$ that are square integrable. Due to the unboundedness of the domain $\mathbb{R}$ we expect that any function in $L^2(\mathbb{R})$ would be decaying in the tails, to ensure integrability. Indeed, the non-zero constant functions cannot be elements of $L^2(\mathbb{R})$. It is also possible to define weighted spaces, such as $L^p_w(\mathbb{R})$, which consists of all functions f such that $\int f(x)^2 w(x)\,dx < \infty$, for some non-negative weighting function w. One such example is the Gaussian weighted space $L^2_w(\mathbb{R})$ where $w = e^{-w^2/2}$. The introduction of a decaying weighting function, permits larger functions to lie in $L^2_w(\mathbb{R})$, for example, all constant functions will now lie in $L^2_w(\mathbb{R})$. Note that if $w > 0$ almost everywhere then $L^2_w(\mathbb{R})$ is a Hilbert space with inner prduct $\langle f, g\langle_w = \int f(x)g(x)w(x)\,dx$.

As a space of functions, $L^2(\mathcal{D})$ spaces are quite "large", the only condition for admission into this space relates to growth and integrability. In Functional Data Analysis we often want to consider "smooth" functions. In the context of Hilbert spaces, this is best approached through *Sobolev Spaces*.

Definition 1.5. Define $H^K([0,1])$ to be the space of functions in $L^2([0,1])$ whose (weak) derivatives up to order K are also in $L^2([0,1])$. This is a Hilbert space with inner product given by

$$\langle x, y\rangle_{H^k} = \sum_{k=0}^{K} \int D^{(k)}x(t)D^{(k)}y(t)\,dt,$$

where $D^{(k)}x$ denotes the order k weak derivative of the function x.

It is worth mentioning that not all function spaces are Hilbert spaces. In many cases there is not a natural inner product structure. This includes the space of p-integrable functions L^p where $p \neq 2$ as well as spaces of smooth functions. These fall into the category of Banach spaces, i.e. complete, normed vector spaces.

Example 1.3. The space $C[0,1]$ denotes the set of all continuous functions over $[0,1]$. When equipped with the sup-norm: $\|x\| = \sup_{t\in[0,1]} |x(t)|$ the space $C[0,1]$ becomes a Banach space. Note that the set $C[0,1]$ is itself a subspace of $L^2([0,1])$, however it is not complete under the $L^2[0,1]$.

Definition 1.6. We say that $\{e_i, i \in I\}$ is an orthonormal system if

$$\langle e_i, e_j\rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Definition 1.7. A Hilbert space H is said to be *separable* if there is a countable orthonormal system $\{e_1, e_2, \dots, \}$ such that every $x \in H$ admits the expansion

$$x = \sum_{j=1}^{\infty} a_j e_j$$

In this case, $\{e_1, e_2, \dots, \}$ is said to be an *orthonormal basis*.

What does the infinite sum in the above definition mean? We should interpret it in the sense that

$$\lim_{J\to\infty} \|x - \sum_{j=1}^{J} a_j e_j\| = 0.$$

We can characterise the coefficients a_j very explicitly, and moreover, rewrite the norm of x in terms of the coefficients:

Theorem 1.2. *Suppose H is separable and $\{e_i\}$ is a complete orthonormal basis. Then, for any $x \in H$:*

$$x = \sum_{j=1}^{\infty} \langle x, e_j \rangle e_j,$$

and

$$\|x\|^2 = \sum_{j=1}^{\infty} |\langle x, e_j \rangle|^2,$$

known as Parseval's identify.

1.1 Examples of orthonormal bases for $L^2([0,1])$

One of the most commonly used basis is the *Fourier basis*:

$$e_1(t) = 1, e_2(t) = \sqrt{2}\sin(2\pi t), e_3 = \sqrt{2}\cos(2\pi t), \dots.$$

This provides a Fourier basis for $L^2([0,1])$. What's interesting is that the cosine only basis $\{f_0, f_1, \dots\}$:

$$f_0(t) = 1, \quad f_k(t) = \sqrt{2}\cos(\pi k t),$$

forms an orthonormal basis for $L^2([0,1])$. Similarly, the sine-only series $g_k(t) = \sin(k\pi t)$, $k = 1, 2, \dots$ also forms an ONB for $L^2([0,1])$.

Exercise 1.1. Show that the series $\{e_k\}$, $\{f_k\}$, $\{g_k\}$ each form an ONB for $L^2([0,1])$.

Proving completeness is more involved. The reader is invited to consult Hsing and Eubank (2015). The most appropriate choice of basis will depend on the type of function being represented. For example all of the basis elements of $g_k(t)$ are zero at 0 and 1. Given that it is a basis, it is indeed possible to approximate any element $x \in L^2([0,1])$ with a linear combination of elements of g_k however, this is by no means a uniform approximation. All we can say is that

$$\left\| x - \sqrt{2} \sum_{k=1}^{K} \langle x, g_k \rangle g_k \right\| \to 0,$$

as $K \to \infty$. This would suggest that $x(t) = t$ is not efficiently expressable in such a basis, and one might consider using e_k instead.

The same approach can be employed to construct an orthonormal basis on $L^2(\mathcal{D})$ for any compact set $\mathcal{D}$. For function spaces over unbounded domains one can resort to other bases, for example polynomials. More specifically on $L^2_w(\mathbb{R})$ the *Hermite polynomials* given by

$$H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1, H_3(x) = x^3 - 3x, \ldots.$$

One can show that these functions are orthonormal with respect to the weighted norm $\langle \cdot, \cdot \rangle_w$.

1.2 Linear Functionals and Linear Operators on Hilbert Spaces

A function $l : H \to \mathbb{R}$ is called a *linear functional* if it satisfies

$$l(\alpha x + \beta y) = \alpha l(x) + \beta l(y), \text{ for all } \alpha, \beta \in \mathbb{R}, x, y \in H.$$

The linear functional is said to be bounded if there exists a constant $\|l\|$ such that $\|l(x)\| \leq \|l\|\|x\|$. A linear functional l is continuous if and only if it is bounded. One such example of a linear functional is defined by

$$l_g(x) = \langle g, x \rangle, \quad x \in H,$$

for $g \in H$. It is a straightforward exercise to show that l_g is linear and bounded. It turns out that all bounded linear functionals are of this form. This is known as the *Riesz Representation Theorem* Kreyszig [1978].

Theorem 1.3. *If $l : H \to \mathbb{R}$ is linear and continuous, then there is a unique $y \in H$ such that $l(x) = \langle x, y \rangle$.*

The set of bounded linear functions on H under addition, scalar multiplication is a vector space. Indeed, one introduce an inner product making it a Hilbert space, known as the *dual space* of H.

A function $L : H \to H$ is said to be a linear operator on H if

$$L(\alpha x + \beta y) = \alpha L(x) + \beta L(y), \quad \text{for all } \alpha, \beta \in \mathbb{R}, x, y \in H.$$

An operator is bounded if there exists a constant K such that

$$\|Lx\| \leq K\|x\|, \quad \forall x \in H.$$

The smallest such K is known as the operator norm of L, i.e.

$$\|L\| = \sup_{\|x\|=1} \|Lx\|.$$

One can show that a linear operator is bounded if and only if it is continuous. The set of bounded operators is a vector space under addition and scalar multiplication. With the operator norm it forms a Banach space. An operator L is said to be *symmetric* or *self-adjoint* if

$$\langle Lx, y\rangle = \langle x, Ly\rangle, \quad \text{for all } x, y \in H.$$

It's worth mentioning there is a distinction between symmetric and self-adjoint operators, but for bounded operators they are equivalent. An operator is said to be *nonnegative-definite* if

$$\langle Lx, x\rangle \geq 0, \quad \forall x \in H,$$

and *positive definite* if

$$\langle Lx, x\rangle > 0, \quad \forall x \in H,$$

Definition 1.8. Let $L : H \to H$ be a bounded operator. The operator L^* defined by

$$\langle Lx, y\rangle = \langle x, L^* y\rangle$$

is called the adjoint operator of L.

One can show that $\|L^*\| = \|L\|$.

1.3 Hilbert Schmidt and Trace Class Operators

We can further distinguish between operators in terms of how they act on elements of H.

Definition 1.9. A bounded linear operator L is said to be *Hilbert-Schmidt* if for an orthonormal basis $\{e_j : j = 1, \dots, \}$ we have

$$\|L\|_{HS}^2 = \sum_{j=1}^{\infty} \langle Le_j, Le_j\rangle < \infty.$$

This is known as the Hilbert-schmidt norm. One can show that this definition is independent of the choice of basis. The set of Hilbert-Schmidt operators forms a separable Hilbert space with inner product:

$$\langle L_1, L_2 \rangle_{HS} = \sum_{j=1}^{\infty} \langle L_1(e_j), L_2(e_j) \rangle$$

Definition 1.10. A bounded linear operator L is said to be *trace class* if for an orthonormal basis $\{e_j : j = 1, \dots, \}$ we have

$$Tr(L) = \sum_{j=1}^{\infty} \langle e_j, Le_j \rangle < \infty,$$

where this sum is independent of the choice of the orthonormal basis. If the sum does not converge we assume it is ∞.

One can check (exercise) that every trace class operator is Hilbert-Schmidt, though the converse is not true.

Example 1.4 (Integral Operator)**.** Let $H = L^2([0,1])$ and define an operator L by

$$(Lx)(t) = \int k(t,s)x(s)\, ds,$$

where $k : [0,1] \times [0,1] \to \mathbb{R}$ is known as a *kernel*, and satisfies

$$K := \int \int k^2(t,s)\, dt\, ds < \infty.$$

Claim: The operator L is bounded. By Fubini's theorem the function $s \to k(t,s)$ lies in L^2 for almost all t. By the Cauchy Schwartz inequality we have that

$$\left(\int k(t,s)x(s)\, ds \right)^2 \leq \int k(t,s)^2\, ds \int x^2(s)\, ds,$$

and so

$$\|Lx\|^2 \leq \left(\int \int k(t,s)^2\, dt\, ds \right) \|x\|^2.$$

It follows that $\|L\| \leq \sqrt{M}$. We now show that L is Hilbert Schmidt. Let $\{e_j; j \geq 1\}$ be an orthonormal basis, then

$$\begin{aligned} \sum_{j=1}^{\infty} \|Le_j\|^2 &= \int \sum_{j=1}^{\infty} \langle k(t,\cdot), e_j \rangle^2\, dt \\ &= \int \|k(t,\cdot)\|^2\, dt \\ &= M. \end{aligned}$$

It follows that $\|L\|_{HS} = \sqrt{M}$.

1.4 Spectral Theory

In many ways, linear operators behave similarly to linear transformations on finite dimensional vector spaces, and share many common concepts. In this subsection we summarise some results about the spectra of linear operators on Hilbert space.

Definition 1.11. Suppose L is an operator on a Hilbert space H. A number λ is called an *eigenvalue* of L if a nonzero vector x such that

$$Lx = \lambda x.$$

Every such x which satisfied this relationship is called an *eigenfunction*.

Theorem 1.4. *All eigenvalues of non-negative definite operator are non-negative. All eigenvalues of a positive definite operator are positive.*

The proof is a straightforward exercise. The following result is a direct generalisation of the finite dimensional case.

Theorem 1.5. *Eigenvectors corresponding to distinct eigenvalues of a symmetric operator are orthogonal.*

Applying the above result would give us an orthonormal system $e_1, e_2, \ldots$ where e_i is an eigenfunction associated with eigenvalue λ_i. When does this form a basis. The following key result provides us with sufficient conditions for this

Theorem 1.6 (Hilbert-Schmidt Theorem)**.** *Let L be a symmetric, Hilbert-Schmidt operator. Then there is a sequence of nonzero real eigenvalues $\lambda_1, \lambda_2, \ldots$ and an associated orthonormal system $\{e_j; j \geq 1\}$ such that every $x \in H$ can be uniqely written as*

$$x = \sum_{j=1}^{\infty} a_j e_j + v,$$

where v is an element of the nullspace of L, i.e. $Lv = 0$.

Note that in fact, we only need that the operator is compact, but do not define that concept here. An immediate corollary of this is the following, which provides a representation of the operator L.

Theorem 1.7. *Suppose L is symmetric and Hilbert schmidt, then there is an orthonormal basis $\{e_j; j \geq 1\}$ such that for every $x \in H$*

$$Lx = \sum_{j=1}^{\infty} \lambda_j \langle x, e_j \rangle e_j,$$

where λ_j is an eigenvalue associated with the eigenfunction e_j.

Introducing tensor notation $(a \otimes b)x = \langle a, x \rangle b$, we can rewrite the above expression as

$$L = \sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j$$

1.5 Rayleigh quotient and maximum principle for eigenvalues

Let L be a symmetric Hilbert Schmidt operator which is non-negative definite, so that all eigenvalues are non-negative. Assume they are arranged in nondecreasing order $\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots$ and assume $\lambda_1 > \lambda_2$. Suppose we wish to find a unit-norm element $x \in H$ such that $\langle x, Lx \rangle$ is maximum. We seek to maximise

$$\langle Lx, x \rangle = \sum_{j=1}^{\infty} \lambda_j \langle x, e_j \rangle \langle e_j, x \rangle = \sum_{j=1}^{\infty} |\langle x, e_j \rangle|^2,$$

subject to $\sum \langle x, e_j \rangle^2 = 1$. It is straightforward to see that the maximum is attained when all the 'mass' is put on the first eigenvalue λ_1, i.e. choosing $x = \pm e_1$, so that the maximum becomes λ_1. More generally, we can iterate this process. Suppose that $\lambda_2 > \lambda_3$ and that we wish to find a maximiser of $\langle x, Lx \rangle$ subject to x being orthogonal to e_1. Similar to the previous step, one can show that this is λ_2 and moreover the maximimum occurs at $x = e_2$. Repeating we get the following result.

Theorem 1.8. *Suppose L is a symmetric, non-negative definite Hilbert-Schmidt operator with distinct, decreasing eigenvalues λ_j; $j = 1, \ldots$ and associated eigenfunctions e_j; $j = 1, \ldots,$. Then*

$$\sup \left[\langle Lx, x \rangle \, : \, \|x\| = 1, \text{ and } \langle x, e_i \rangle = 0, i = 1, \ldots, j-1 \right] = \lambda_j,$$

where this maximum is attained at $x = e_j$.

2 Random Elements in a Hilbert Space

As hinted at in the previous lecture, a dual perspective on how we think about a stochastic process, say $X(t, \omega)$, $t \in [0, 1]$. The process viewpoint characterises $X(t, \omega)$ as a collection of random

variables, indicated by a continuous index t, such that $X(t,\omega)$ and $X(t',\omega)$ may be correlated. From the point of view of statistical inference, particularly for longitudinal data, it makes sense to view $X(\cdot,\omega)$ as a random variable, but in an appropriate space of functions. The particular choice of function space is a modelling assumption, and has implications on how $X(t,\omega)$ and $X(t',\omega)$ might related. This requires us to understand the implications of having infinite-dimensional random variables (in particular infinitely many degrees of freedom). In the context of FDA we typically assume our curves are lying within a separable Hilbert space $(H,\langle\cdot,\cdot\rangle)$, which is usually, but not always, $L^2([0,1])$.

Some thought is needed how we reason about such random variables. Consider the expectation $\mathbb{E}[X]$. The usual construction starts in the usual fashion in which one constructs intervals on other structures: we start an appropriate definition on the class of "simple functions", i.e. linear combinations of indicator functions and work our way up to a sufficiently rich class of random variables. The resulting expectation is known as the *Bochner integral*. A random variable is (Bochner) integrable if and only if $\mathbb{E}\|X\| < \infty$, see Bosq [2000].

In the context of functional data analysis, it sometimes makes sense to consider weaker forms of integrals. One such approach is based on the observation that we know how to compute the expectation $\mathbb{E}X(t,\cdot)$ for each $t \in [0,1]$ as a real-valued random variable. Based on this idea, we say that X is *Pettis* integrable or weakly integrable if there exists $E \in H$ such that $\mathbb{E}[l(X)] = l(E)$ for all bounded linear functionals l. It can be shown that if the Bochner integral does exist and is finite then it coincides with the weak integral.

If $\mathbb{E}[X] = 0$ then we say that the measure is centered. The *covariance operator* $\mathcal{K}$ is characterised by the identity that

$$\langle u, \mathcal{K}v\rangle = \mathbb{E}\left[\langle u, X - \mathbb{E}[X]\rangle\langle v, X - \mathbb{E}[X]\rangle\right].$$

In tensor form we can write this as

$$\mathcal{K} = \mathbb{E}\left[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])\right]$$

2.1 Gaussian Random Variables on Hilbert Spaces

As with classical multivariate inference, the Gaussian distribution plays an important role in FDA. A random variable X on H is said to be Gaussian if $l(X)$ is Gaussian on $\mathbb{R}$, for all bounded linear functionals l. In this case $l(X) \sim \mathcal{N}(m_l, \sigma_l^2)$ where $m_l = l(\mathbb{E}[X])$ and $\sigma_l^2 = \langle l, \mathcal{K}l\rangle$. The mean m and covariance operator $\mathcal{K}$ are indeed well-defined this definition. Indeed, we have the following result.

Theorem 2.1. *A Gaussian measure on a Hilbert Space H with mean m and covariance $\mathcal{K}$ has characteristic function*

$$\varphi(l) = \exp\left(i\langle l, m\rangle - \frac{1}{2}\langle l, \mathcal{K}l\rangle\right).$$

This implies that m and $\mathcal{K}$ completely characterise the Gaussian measure, hence we are justified

in expressing the distribution as $\mathcal{N}(m, \mathcal{K})$ It is natural to ask what conditions an operator must satisfy in order to be a covariance operator. This is partially answered by the following theorem.

Theorem 2.2. *If $\mathcal{N}(0, \mathcal{K})$ is a Gaussian measure on a Hilbert Space H, then $\mathcal{K}$ is a symmetric, non-negative definite, trace-class operator on H. Furthermore, for any integer p, there is a constant $C = C_p > 0$ such that $X \sim \mathcal{N}(0, \mathcal{K})$*

$$\mathbb{E}\|X\|^{2p} \leq C_p \left(tr(\mathcal{K})\right)^p .$$

Conversely, if $m \in H$ and $\mathcal{K}$ is a symmetric, nonnegative definite trace class operator on H, then there is a Gaussian measure $\mathcal{N}(m, \mathcal{K})$ on H.

2.2 Gaussian Random Variables on Function Spaces

When H is a function space, such as $L^2([0,1])$, then we can view a Gaussian measure $\mathcal{N}(m, \mathcal{K})$ as the *path-measure* associated with a Gaussian process $X(t, \omega)$. In this setting, we can make explicit the connection between the Gaussian distribution on the Hilbert space, and associated Gaussian process.

Lemma 2.3. *Let X be a random variable on $L^2([0,1])$ such that $X(\cdot, \omega)$ which is continuous for almost every $\omega \in \Omega$. Then X is a Gaussian distributed random variable if and only if and all the* finite dimensional distributions *are Gaussian, i.e., every finite dimensional vector $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is a $\mathcal{N}(\mu_k, K_k)$ random variable for some vector μ_k and a symmetric non-negative definite matrix K_k, for all $k \in \mathbb{N}$ and $t_1, t_2, \ldots, t_k \in [0,1]$.*

In this case, we keep in mind that the expected value is a function of t, i.e. $m = m(t)$, $t \in [0,1]$. Moreover the following result characterises the covariance operator in terms of its kernel.

Theorem 2.4. *Any Hilbert-Schmidt operator on $L^2([0,1])$ can be expressed as an integral operator, i.e if $\mathcal{K}$ is a Hilbert-Schmidt operator then there exists $k \in L^2([0,1] \times [0,1])$ such that*

$$(\mathcal{K}x)(t) = \int k(t,s)x(s)\, ds.$$

In fact, under some moderate conditions Mercer's theorem (presented below) provide an explicit expressin for k in terms of the spectrum of $\mathcal{K}$. The distribution of a Gaussian process is uniquely determined by the mean function $m(t) = \mathbb{E}[X(t)]$ and its covariance function $\gamma(s,t) = \mathrm{Cov}(X(s), X(t))$. In general, the smoothness of a Gaussian process is determined by the smoothness of the covariance

function near the diagonal (i.e. where $|t-s|$ is small). Some well-known examples of Gaussian processes which will be discussed throughout this course are the following:

1. **Brownian motion:** $T = [0, \infty)$, $\mu(t) = 0$, and $\gamma(t, s) = \min(t, s)$.
2. **Ornstein Uhlenbeck Process:** $T = [0, \infty)$, $\mu(t) = 0$, $\gamma(t, s) = \exp(-|t - s|)$.
3. **Squared Exponential Process:** $T = [0, \infty)$, $\mu(t) = 0$, $\gamma(t, s) = \exp(-|t - s|^2)$.

Example 2.1 (**Sampling Gaussian Processes**)**.** It is not difficult to simulate Gaussian stochastic processes on a computer. Given a random-number generator that generates $\mathcal{N}(0, 1)$ (pseudo)random numbers, we can sample from a Gaussian stochastic process over a finite set of time-points $0 \leq t_1 \leq \ldots \leq t_N$. Define Σ^N to be the covariance matrix

$$\Sigma^N_{i,j} = \gamma(t_i, t_j).$$

We know that Σ^N is a non-negative definite matrix, and possesses a square root Γ such that $\Sigma^N = \Gamma\Gamma^\top$. Setting $\mu^N = (\mu(t_1), \ldots, \mu(t_N))$ we then have that

$$X^N = \mu^N + \Gamma\mathcal{N}(0, I_{N\times N}),$$

is a sample of the Gaussian process over the time-steps $\{t_1, \ldots, t_n\}$. This scheme provides us with a quick and (relatively) efficient way to generate samples of a Gaussian process, which can be easily implemented in MATLAB, R or Python.

2.3 Translation Invariant Measures on Hilbert spaces and the lack of densities

Much of the methodology we use in classical multivariate statistical analysis is tied to the assumption of the law of a random variable X having a density $\rho(x)$ with respect to the Lebesgue measure on the underlying space. This is contingent on $\mathcal{L}(X)$ being absolutely continuous with respect the Lebesgue measure, allowing one to appeal to the Radon-Nikodym theorem.

A crucial distinction between functional and multivariate data settings is that there is no analogue to the Lebesgue integral in infinite dimensions. More specifically, we have the following result.

Theorem 2.5. *Let H be an infinite-dimensional separable Hilbert space. Then the only locally finite and translation-invariant Borel measure μ on H is the trivial measure assigning zero to all measurable sets. Equivalently, every translation-invariant measure that is not identically zero assigns infinite measure to all open subsets of H.*

Differently from the finite dimensonal setting, to work with densities on function spaces, we must therefore choose a reference measure which exhibits some form of concentration (i.e. not translation invariant). A natural candidate is to choose a Gaussian measure $\mathcal{N}(m, \mathcal{K})$ as a reference measure. When doing so, one must be careful that the absolute continuity assumption, required for the Radon-Nikodym theorem to hold, is very stringent in infinite dimensions. This is best demonstrated in the context of Gaussian distributions by the *Feldman-Hajek theorem*.

Theorem 2.6 (Feldman-Hajek theorem)**.** *Two Gaussian measures $\mu_i = \mathcal{N}(m_i, \mathcal{K}_i)$, $i = 1, 2$ on a Hilbert space H are either* singular *or* equivalent*. They are equivalent if and only if the following three conditions hold:*

1. $E := Im(\mathcal{K}_1^{1/2}) = Im(\mathcal{K}_2^{1/2})$.
2. $m_1 - m_2 \in E$.
3. *The operator $T = (\mathcal{K}_1^{-1/2}\mathcal{K}_2^{1/2})(\mathcal{K}_1^{-1/2}\mathcal{K}_2^{1/2})^* - I$ is Hilbert Schmidt, where I is an identity operator.*

Thus to represent a Gaussian measure μ_1 in terms of a density with respect to Gaussian measure μ_2 one needs to ensure that the conditions for equivalence hold. It is sometime counterintuitive how restrictive this condition is. For example if μ_1 is the distribution for standard Brownian motion $(B_t; t \in [0, 1])$ and μ_2 is the distribution for rescaled Brownian moton $(\sigma B_t; t \in [0, 1])$ with $\sigma \neq 1$, then μ_1 and μ_2 are singular. When a density does exist with respect to a Gaussan measure, then the *Cameron Martin Theorem* provides us with a closed form expression for the density. Related to this is *Girsanov's theorem* which is often employed to infer the coefficents of Stochastic Differential Equations (SDEs) from observations of sample paths. See Da Prato and Zabczyk [2014] for a more complete exposition of these results.

2.4 The Karhunen-Loève Expansion

The Karhunen-Loève (KL) expansion makes rigorous the concept we introduced in the introduction, that a stochastic process X_t can be viewed as a random variable on the space of trajectories/paths on the domain $[0, 1]$. Indeed, via the KL expansion, we can view in a very direct fashion that X_t, $t \in [0, 1]$ can be viewed as a random variable taking values in the Hilbert space $H = L^2([0, 1])$. As we shall see later, the Karhunen-Loeve expansion is merely functional principal component analysis in disguise, and we shall revisit further in Lecture 3.

Let $f \in L^2([0, 1])$ and let $\{e_n\}_{n=1}^{\infty}$ be an orthonormal basis in $L^2([0, 1])$. Then f can be written as

$$f = \sum_{n=1}^{\infty} f_n e_n,$$

where $f_n = \int_\Omega f(x)e_n(x)\,dx$. The convergence of the infinite sum should be interpreted in the

$L^2([0,1])$ sense, i.e.

$$\lim_{N\to\infty}\left\| f-\sum_{n=1}^{N} f_n e_n\right\|_{L^2([0,2]} = 0.$$

Consider a process X_t on $[0,1]$, with finite second moments. Without loss of generality we shall assume that

$$\mathbb{E}[X_t] = m(t) = 0, \quad \text{so that} \quad \gamma(t,s) = Cov(X_t, X_s) = \mathbb{E}(X_t X_s).$$

Assume (for now!) an expansion of the form

$$X_t(\omega) = \sum_{n=1}^{\infty} \zeta_n(\omega) e_n(t), \quad t\in[0,1], \tag{1}$$

where $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis on $L^2([0,1])$. If such a decomposition was possible, the random variables ζ_n would be calculated using the orthonormality of $\{e_n\}$:

$$\langle X_t, e_k\rangle = \int_0^1 X_t e_k(t)\,dt = \int_0^t \sum_{n=1}^{\infty} \zeta_n e_n(t)\,dt = \sum_{n=1}^{\infty}\zeta_n\delta_{n,k} = \zeta_k,$$

where we assumed that we can interchange summation and integration. We will assume that these random variables are uncorrelated such that

$$\mathbb{E}(\zeta_n\zeta_m) = \lambda_n\delta_{n,m},$$

where $\{\lambda_n\}_{n=1}^{\infty}$ are positive numbers yet to be determined.

Assuming that expansion (1) exists for X_t, then we can rewrite the covariance as

$$\begin{aligned}\gamma(t,s) = \mathbb{E}(X_t X_s) =& \mathbb{E}\left(\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\zeta_k e_k(t)\zeta_l e_l(s)\right)\\ =&\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\mathbb{E}(\zeta_k\zeta_l)e_k(t)e_l(s)\\ =&\sum_{k=1}^{\infty}\lambda_k e_k(t)e_k(s).\end{aligned}$$

Consequently, in order for the expansion (1) to be valid, we need

$$\gamma(t,s) = \sum_{k=1}^{\infty}\lambda_k e_k(t)e_k(s). \tag{2}$$

In particular

$$\begin{aligned}\mathcal{K}e_n := \int\gamma(t,s)e_n(s)\,ds =& \int_0^1\sum_{k=1}^{\infty}\lambda_k e_k(t)e_k(s)e_k(s)\,ds\\ =&\sum_{k=1}^{\infty}\lambda_k e_k(t)\int_0^1 e_k(s)e_n(s)\,ds\\ =&\sum_{k=1}^{\infty}\lambda_k e_k(t)\delta_{k,n}\\ =&\lambda_n e_n(t),\end{aligned}$$

so that, $\{\lambda_n, e_n(t)\}_{n=1}^{\infty}$ be necessarily be the set of eigenvalues of the integral operator operator $\mathcal{K}$ with kernel given by $\gamma(t,s)$.

Thus, proving the existence of expansion (1) requires a careful study of the eigenvalue problem for the integral operator $\mathcal{K}$. The results we require are summarized in *Mercer's Theorem*. Recall, that a function $C : [a,b] \times [a,b] \to \mathbb{R}$ is symmetric if $\gamma(t,s) = \gamma(s,t)$ and non-negative definite iff

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i \gamma(t_i, t_j) c_j^* \geq 0,$$

for all $n \in \mathbb{N}$, $\{t_i\}_{i=1}^{N} \in [a,b]$, and $c_1, \ldots c_n \in \mathbb{C}$.

Theorem 2.7 (Mercer's theorem)**.** *Suppose γ is a continuous symmetric non-negative definite kernel. Then there is an orthonormal basis $\{e_n\}_{n\in\mathbb{N}}$ of $L^2[0,1]$ consisting of eigenfunctions of $\mathcal{K}$ such that the corresponding sequence of eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}}$ is nonnegative. The eigenfunctions corresponding to the non-zero eigenvalues are continuous on $[0,1]$ and γ has the representation*

$$\gamma(s,t) = \sum_{n=1}^{\infty} \lambda_n e_n(s) e_n(t),$$

where the convergence is absolute and uniform.

Mercer's theorem applied to $C(t,s)$ guarantees the existence of the $\{e_n\}_{n\in\mathbb{N}}$ and $\{\lambda_n\}_{n\in\mathbb{N}}$. Furthermore, for every $f \in L^2([0,1])$, we can write

$$f = f_0 + \sum_{n=1}^{\infty} f_n e_n(t),$$

for $\mathcal{C} f_0 = 0$, where the convergence of the series is in L^2. Moreover, the expansion (2) is valid, and the series converges absolutely and uniformly in (t,s).

We can now prove the expansion (1) holds.

Theorem 2.8 (Karhunen-Loève)**.** *Let $\{X_t, t \in [0,1]\}$ be an $L^2([0,1])$ process with zero mean and continuous covariance function $\gamma(t,s)$. Let $\{\lambda_n, e_n(t)\}_{n=1}^{\infty}$ be the eigenvalues and eigenfunctions of the operator $\mathcal{K}$ defined by*

$$(\mathcal{K}f)(t) := \int_0^1 \gamma(t,s) f(s)\, ds.$$

Then

$$X_t = \sum_{n=1}^{\infty} \zeta_n e_n(t), \quad t \in [0,1],$$

where

$$\zeta_n = \int_0^1 X_t e_n(t)\, dt, \quad \mathbb{E}\zeta_n = 0, \quad \mathbb{E}(\zeta_n \zeta_m) = \lambda_n \delta_{n,m}.$$

The series converges in L^2 to $X(t)$, uniformly in t.

Proof. The fact that $\mathbb{E}\zeta_n = 0$ follows from the fact that X_t is mean-zero. The orthogonality of the random variables $\{\zeta_n\}_{n=1}^{\infty}$ follows from the orthogonality of the eigenfunctions of $\mathcal{K}$:

$$\begin{aligned}\mathbb{E}(\zeta_n\zeta_m) =& \mathbb{E}\int_0^1\int_0^1 X_tX_se_n(t)e_m(s)\,dtds\\ &\int_0^1\int_0^1 \gamma(t,s)e_n(t)e_m(s)\,dtds\\ &\lambda_n\int_0^1 e_n(s)e_m(s)\,ds = \lambda_n\delta_{n,m}.\end{aligned}$$

Consider now the partial sum $S_N = \sum_{n=1}^{N}\zeta_n e_n(t)$. We have

$$\begin{aligned}\mathbb{E}|X_t - S_N|^2 =& \mathbb{E}X_t^2 + \mathbb{E}S_N^2 - 2\mathbb{E}(X_tS_N)\\ =&\gamma(t,t) + \mathbb{E}\sum_{k,l=1}^{N}\zeta_k\zeta_l e_k(t)e_l(t) - 2\mathbb{E}\left(X_t\sum_{n=1}^{N}\zeta_n e_n(t)\right)\\ =&\gamma(t,t) + \sum_{k=1}^{N}\lambda_k|e_k(t)|^2 - 2\mathbb{E}\sum_{k=1}^{N}\int_0^1 X_tX_se_k(s)e_k(t)\,ds\\ =&\gamma(t,t) - \sum_{k=1}^{N}\lambda_k|e_k(t)|^2 \to 0,\end{aligned}$$

as $N \to 0$, by Mercer's theorem. □

All of the above can be extended to processes on the domain $[0,T]$, for a general $T > 0$ without much effort.

The Karhunen-Loève expansion is straightforward to apply to Gaussian stochastic processes. Let X_t be a Gaussian second-order process with continuous covariance $\gamma(t,s)$. Then the random variables $\{\zeta_k\}_{k=1}^{\infty}$ are Gaussian, being the time-integral of a Gaussian process. Furthermore, since they are Gaussian, $\mathbb{E}[\zeta_k\zeta_l] = \delta_{k,l}$, implies that they are also independent. Hence, for Gaussian processes, the Karhunen-Loève expansion becomes

$$X_t = \sum_{n=1}^{\infty}\sqrt{\lambda_n}\zeta_n e_n(t),$$

where $\{\zeta_n\}_{n=1}^{\infty}$ are independent $\mathcal{N}(0,1)$ random variables.

Example 2.2 (The Karhunen-Loève expansion for Brownian motion)**.** The covariance function of Brownian motion is $C(t,s) = \min(t,s)$. The eigenvalue problem $\mathcal{C}e_n = \lambda_n e_n$ becomes

$$\int_0^1 \min(t,s)e_n(s)\,ds = \lambda e_n(t). \tag{3}$$

Let us assume that $\lambda_n > 0$ (check that 0 is not an eigenvalue). Setting $t = 0$, we obtain $e_n(0) = 0$. Therefore (3) becomes

$$\int_0^t s e_n(s)\, ds + t \int_t^1 e_n(s)\, ds = \lambda_n e_n(t).$$

We differentiate this equation once with respect to t:

$$\int_t^1 e_n(s)\, ds = \lambda_n e_n'(t).$$

Setting $t = 1$ we obtain the second boundary condition $e_n'(1) = 0$. A second differentiation yields

$$-e_n(t) = \lambda_n e''(t).$$

Thus the eigenvalue problem has been reduced to the following Sturm-Liouville problem

$$-e_n(t) = \lambda e_n''(t), \quad e_n(0) = e_n'(1) = 0.$$

Solving, we obtain the following (normalized) eigenfunctions:

$$e_n(t) = \sqrt{2} \sin\left(\frac{1}{2}(2n-1)\pi t\right), \quad \lambda_n = \left(\frac{2}{(2n-1)\pi}\right)^2.$$

Thus the Karhunen-Loève expansion of Brownian motion on $[0, 1]$ is

$$W_t = \sqrt{2} \sum_{n=1}^{\infty} \zeta_n \frac{2}{(2n-1)\pi} \sin\left(\frac{1}{2}(2n-1)\pi t\right).$$

This is known as the Wiener representation of Brownian motion, which provides a Fourier series decomposition for Brownian motion on $[0, 1]$. Clearly, the scaled process

$$\sqrt{c} W\left(\frac{t}{c}\right),$$

is a Brownian motion on $[0, c]$.

Exercise 2.1. Using the Karhunen-Loève (KL) expansion for a mean-zero, L^2-process X_t on $[0, 1]$, show that

$$\mathbb{E} \int_0^1 |X_t|^2\, dt = \text{trace}[\mathcal{K}] = \sum_{n=1}^{\infty} \lambda_n.$$

Exercise 2.2. Calculate the KL expansion for the following:

1. A second-order stochastic process with covariance function $\gamma(t, s) = ts$.

2. The Brownian bridge on $[0,1]$.

Given the Hilbert space $H = L^2([0,1])$, we can define in a rigorous way a Gaussian distribution over H, $\mathcal{N}(0,\mathcal{K})$, having as covariance a self-adjoint trace-class operator $\mathcal{K}$. For example, Brownian motion can described by a Gaussian random variable on $H = L^2([0,1])$, with covariance operator $(-\Delta)^{-1}$.

Exercise 2.3. Show that $\mathcal{L}$ given by

$$\mathcal{K}f = \int_0^1 \min(t,s) f(s)\, ds$$

can be written as $\mathcal{C} = (-\Delta)^{-1}$.

Remark 2.1. Note that the coefficients ζ_n are always uncorrelated but the independence is only guaranteed in the Gaussian case. This is one of the reasons why Gaussianity is so commonly assumed in FDA. In Delaigle et al. [2010] the authors argued that "Particularly in the infinite-dimensional setting of functional data analysis,it seems impossible to use effectively general models for random variables that are uncorrelated but not independent. Such an approach leads to cumbersome methods and does not seem to allow useful insight into theoretical properties". See also the discussion in Cuevas [2014].

3 Estimation of Mean and Covariance Functions

In FDA applications we often observe N curves $X_1, \ldots, X_N$. Vewing each as independent realisations of a random variable X taking values in $L^2([0,1])$. To this end, we shall assume that $X_1, \ldots, X_N \in L^2([0,1])$ and have the same square-integrable distribution.

The mean function $m(t)$ can be estimated by the sample mean function

$$\widehat{m}(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t),$$

and the covariance function by

$$\widehat{\gamma}(s,t) = \frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \widehat{\mu}(t))(X_i(s) - \widehat{\mu}(s)).$$

The associated covariance operator can be estmated by

$$\widehat{\mathcal{K}}x = \int \widehat{\gamma}(s,\cdot)x(s)\,ds = \frac{1}{N}\sum_{i=1}^{N}\langle X_i - \widehat{m}, x\rangle(X_i - \widehat{m}).$$

Remark 3.1. Note that $\widehat{\mathcal{K}}$ maps $L^2([0,1])$ into a finite dimensional subspace spanned by $X_1, X_2, \ldots, X_N$. This illustrates the limitations of statistical inference for functional observations; a finite sample can recover a finite dimensional projection of an infinite dimensional object.

The following result shows us that $\widehat{m}$ is an unbiased, consistent estimator for the mean function m.

Theorem 3.1. *We have that $\mathbb{E}\widehat{m} = m$ and $\mathbb{E}\|\widehat{m} - m\| = O(1/N)$.*

Proof. Clearly,

$$\mathbb{E}\left[\widehat{m}\right] = \frac{1}{N}\sum_{i}\mathbb{E}[X_i] = m,$$

so that the estimator is unbiased. Now consider

$$\begin{aligned}\mathbb{E}\|\widehat{m} - m\|^2 &= N^{-2}\sum_{i,j=1}^{N}\mathbb{E}[\langle X_i - m, X_j - m\rangle] \\ &= N^{-2}\sum_{i=1}^{N}\mathbb{E}\|X_i - m\|^2 \\ &= \frac{1}{N}\mathbb{E}\|X - m\|^2,\end{aligned}$$

where we've used the fact that the samples are independent and identically distrbuted. □

Chebyshev's inequality, this implies that the estimators converge to the corre- sponding population parameters in probability. To develop significance tests, it is useful to find the asymptotic distributions of these estimators. Under the assumptions of the previous theorem, one can show that the mean estimator is asymptotically normal. The results of this section provide central limit theorems for the estimator around their mean. We shall make use of the following central limit theorem for Hilbert space valued random variables (see Bosq [2000] Theorem 2.7)

Theorem 3.2. *Suppose $X_i, i = 1, 2, \ldots$ are IID square-integable random variables with expectation m and covariance operator $\mathcal{K}$, then*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(X_i - m) \xrightarrow{D} Z,$$

where Z is normal with mean zero and covariance $\mathcal{K}$.

Connected to this central limit theorem is a Law of Large numbers for Hilbert-space valued random variables.

Theorem 3.3 (Section 7.2 of Laha)**.** *Suppose $X_i, i = 1, 2, \ldots$ are IID integrable random variables with mean m. Then*

$$\frac{1}{M}\sum_{i=1}^{N} X_i \xrightarrow{a.s.} m, \text{ as } N \to \infty.$$

As with most LLNs and CLT results, keep in mind that the IID assumption can often be relaxed in many ways, permitting weak dependence. This will be revisted in later lectures. We now present the asymptotic normality result for the sample mean estimator.

Theorem 3.4. *Let $X_1, X_2, \ldots,$ be an iid sequence in H with $\mathbb{E}\|X_1\|^2 < \infty$, then*

$$\sqrt{N}\,(\widehat{m} - m) \xrightarrow{d} \mathcal{N}(0, \mathcal{K}) \text{ in } H,$$

where $\mathcal{K}$ is the covariance operator associated with the random variable X_1.

Proof. The proof is a direct application of the CLT for Hilbert space valued random variables. □

Let us now turn to properties of the covariance estimator. Replacing m with $\widehat{m}$ has a neglible effect on the second order statistics. Clearly, $\mathbb{E}\widehat{\gamma}(s,t) = N/(N-1)\gamma(s,t)$, so it is biased, but the bias is asymptotically negligible. For this reason, we shall assume without loss of generality that $m = 0$. One can quickly check that $\widehat{\mathcal{K}}$ is a Hilbert-Schmidt operator. We have the following result.

Theorem 3.5. *If $\mathbb{E}\|X\|^4 < \infty$ and $\mathbb{E}[X] = 0$ then*

$$\|\widehat{C}\|_{HS}^2 \leq \mathbb{E}\|X\|^4.$$

Proof. We have

$$\mathbb{E}\|\widehat{C}\|_{HS}^2 = \sum_{i=1}^{\infty} \mathbb{E}\left[\langle \widehat{C}e_i, \widehat{C}e_i\right].$$

Plugging in the expression for $\widehat{C}$ we see that

$$\mathbb{E}\|\widehat{C}\|_{HS}^2 = N^{-2}\sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{i=1}^{\infty}\mathbb{E}\langle\langle X_n, e_i\rangle X_n, \langle X_m, e_i\rangle X_m\rangle.$$

By Parsevals theorem this implies that

$$\sum_{i=1}^{\infty} \mathbb{E}\langle\langle X_n, e_i\rangle X_n, \langle X_m, e_i\rangle X_m\rangle = \mathbb{E}\langle X_n, X_m\rangle^2,$$

so that

$$\mathbb{E}\|\widehat{C}\|_{HS}^2 = N^{-2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\langle X_n, X_m\rangle^2.$$

Applying Cauchy Schwartz inequality we get

$$\mathbb{E}\|\widehat{C}\|_{HS}^2 \leq \mathbb{E}\|X\|^4.$$

□

Finally we conclude with the consistency of the covariance operator $\widehat{\mathcal{K}}$.

Theorem 3.6. *If* $\mathbb{E}\|X\|^4 < \infty$ *and* $\mathbb{E}[X] = 0$*, then*

$$\mathbb{E}\|\widehat{\mathcal{K}} - \mathcal{K}\|^2 \leq N^{-1}\mathbb{E}\|X\|^4.$$

This is left as an exercise. The interested reader is invited to consult Horváth and Kokoszka [2012]. We now show the asymptotic normality of the estimator.

Theorem 3.7. *Let* $X_1, X_2, \ldots,$ *be an iid sequence in* H *with* $\mathbb{E}\|X_1\|^4 < \infty$*, then*

$$\sqrt{N}\left(\widehat{\mathcal{K}} - \mathcal{K}\right) \xrightarrow{d} \mathcal{N}(0, \mathcal{M}) \text{ in } H,$$

where

$$\mathcal{M} = \mathbb{E}[(X_1 - m)\otimes(X_1 - m) - \mathcal{K}] \otimes [(X_1 - m)\otimes(X_1 - m) - \mathcal{K}]$$

Proof. This proof is left as an exercise to the interested reader. □

References

Naum Il ich Akhiezer and Izrail Markovich Glazman. *Theory of linear operators in Hilbert space*. Courier Corporation, 2013.

Denis Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2000.

Antonio Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge university press, 2014.

Aurore Delaigle, Peter Hall, et al. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, 2010.

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

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.

RG Laha. k rohatgi, vk 1979 probability theory.