Functional Data Analysis

Victor M. Panaretos

Institut de Mathématiques – EPFL

victor.panaretos@epfl.ch

- Lectures Thu 13.15–15.00
- Main reference books (but we might deviate):

  Hsing & Eubank, *Theoretical Foundations of Functional Data Analysis*, Wiley

  Da Prato & Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge
- Webpage: moodle
- Written final exam (cheat sheet allowed)

"Modern science and technology provide statistical problems with observable random variables taking their values in functional spaces."

Jerzy Neyman, 1966.

In a nutshell, this is precisely what this course is about:

the blending of *statistical* and *functional* analysis.

Functional analysis is the study of infinite-dimensional vector spaces, often with additional structures (inner product, norm) with typical examples given by function spaces. The subject also includes the study of linear and non-linear operators on these spaces, as well as measure, integration, probability on infinite dimensions, and also manifolds with local structure modelled by these vector spaces.

`math.stackexchange.com`

Linear Functional Analysis `vs` Nonlinear Functional Analysis
(Linear Algebra `vs` Differential Geometry)

- Multivariate analysis (MVA) was about the statistical analysis of random vectors in finite-dimensional vector spaces
- Functional data analysis (FDA) is about the statistical analysis of random vectors in infinite-dimensional vector spaces
- Matrix algebra played a fundamental role in MVA. Operator theory will play a similarly central role in FDA.
- Measure, integration, and probability obviously play a crucial role in both.

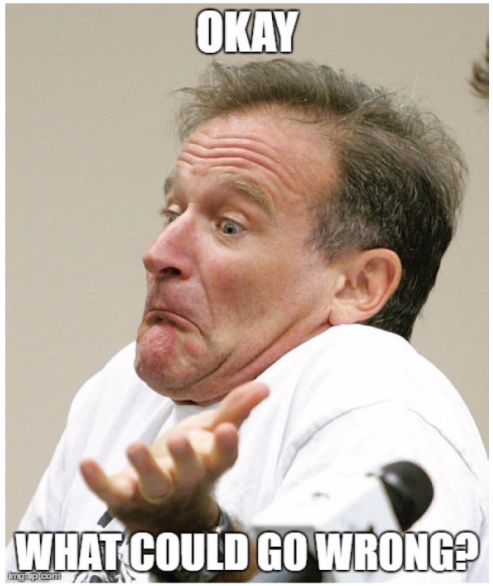
Function spaces are generically infinite dimensional

- Consider any function space that is rich enough to contain (all) polynomials .
- Assume such a space is of dimension $p < \infty$.
- Then, no $p + 1$ elements of said space can be linearly independent.
- But $\sum_{k=0}^{p} \alpha_k x^k = 0 \implies \alpha_k = 0, \forall k$ (differentiate and evaluate at 0).

  (nothing special about polynomials, can work similarly with other systems)

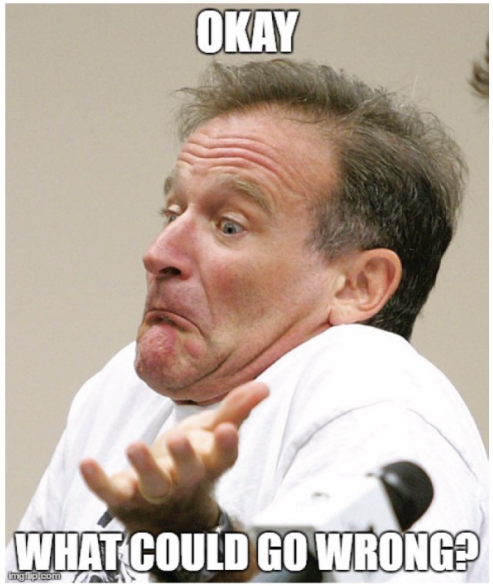
Is this a problem?

- Often no. Many key concepts, structures, and even theorems mimic their Euclidean ($\mathbb{R}^p$, $p < \infty$) counterparts mutis mutandis, so we can be guided by our usual intuition and proceed "by extension".



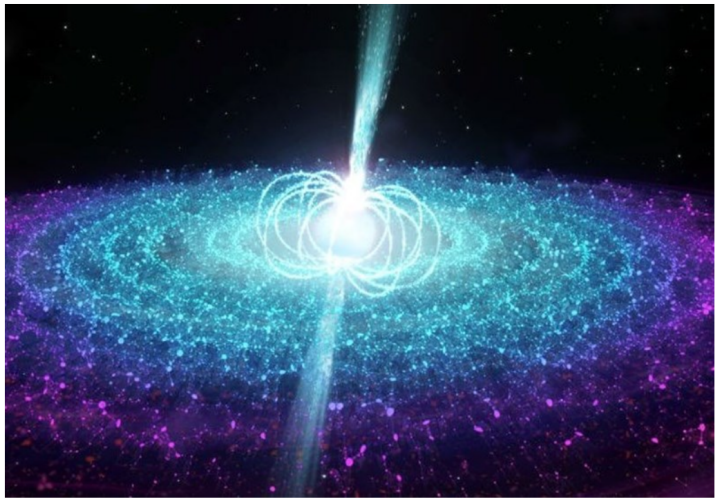
Is this a problem?

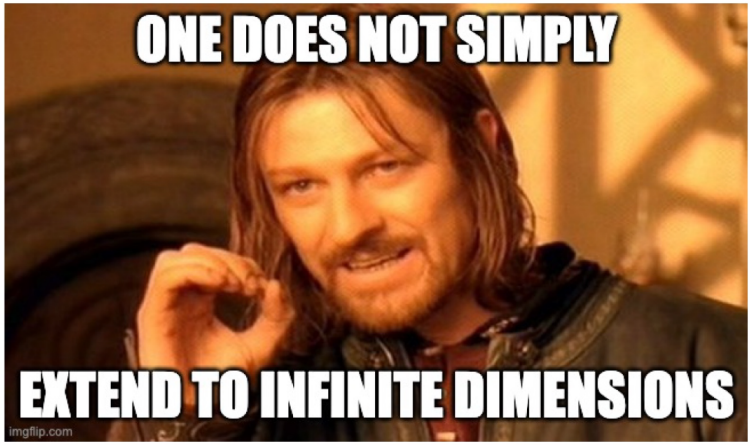
- Often no. Many key concepts, structures, and even theorems mimic their Euclidean ($\mathbb{R}^p$, $p < \infty$) counterparts mutis mutandis, so we can be guided by our usual intuition and proceed "by extension".

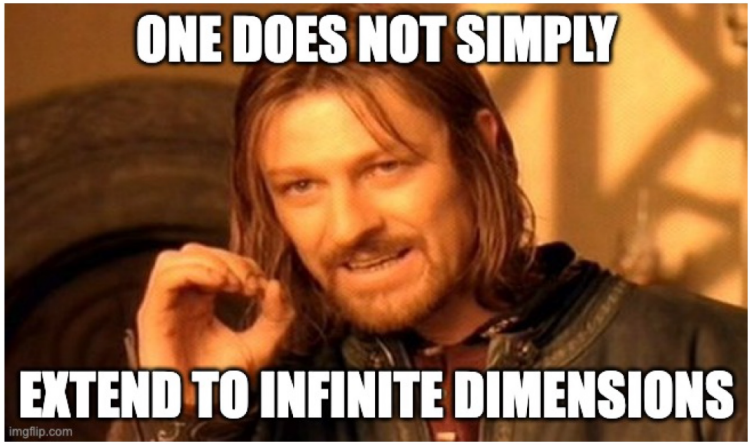


but also

- Often yes. Many concepts and structures behave in strong contrast with our Euclidean intuition. Correspondingly, many "Euclidean theorems" are generically false and we must proceed cautiously.




ONE DOES NOT SIMPLY
EXTEND TO INFINITE DIMENSIONS
imgflip.com



Some strange things about infinite dimensions:

- The unit ball is not compact.
- There exist open subspaces.
- There exist discontinuous linear maps.
- There is no analogue of Lebesgue measure.
- Norms are typically not equivalent.
- Covariances are never invertible.


Toto, I've got a feeling we're not in Kansas anymore

On December 21st, 1807, a truly unforgettable presentation took place in the French Academy of Sciences.

... this presentation was the that of a thesis by the 39 year old mathematician and physicist named Joseph Fourier.

- Fourier, like many contemporary scientists, was working on the problem of heat conduction in metal rods
- This work had led him to consider different ways of representing functions on closed intervals

During the presentation of his thesis he made the following remarkable and at the same time outrageous claim:

any function on a compact interval can be expressed as a trigonometric series

To be more specific, taking $I = [-\pi, \pi]$, Fourier claimed that any $f : I \to \mathbb{R}$, as bizarre as it may be, can be expressed as

$$f(x) = \frac{\alpha_0}{2} + \sum_{n=1}^{\infty} (\alpha_n \cos(nx) + \beta_n \sin(nx))$$

- The precise statement was wrong, but can be made correct by a more assumptions or more nuanced version of "expressed".
- The jury, comprised of Lagrange, Laplace and Legendre, unanimously rejected the thesis!
- The jury had several concerns:
  - Since trigonometric functions are infinitely smooth, their sums should also be
  - Since trigonometric functions are analytic, their local behaviour determines their function globally
  - How could this be consistent with the arbitrarily local behaviour of a general function?
- Fourier's critics made the mistake -quite common for that era- to assume that a property that holds for the elements of a sequence (the partial sums) is also true for the limit of the sequence
- Phrased in more modern language, they thought that subspaces of function spaces (in this case, trigonometric partial sums) are closed (they are not necessarily so, in this case trigonometric series are not closed in $L^2[-\pi, \pi]$).
- So if you ever feel stumped in the functional world, don't be discouraged: some of the greats were also equally stumped!

Still, could this all just be a mathematical indulgence?

- In practice, aren't all data ultimately recorded to finite precision?
- Can't we just set the measured resolution as the dimension p?
- Can't we then translate into an MVA setting and go on with life?

While it certainly appeals to our mathematical tastes, it's also very real:

- We need to make sure our inferences are stable to resolution (blowup).
- Functional nature of data can sometimes be a blessing (perfect testing).
- Fixing a finite resolution can introduce serious bias (graphical models).
- Some forms of statistical variation are functional in essence (phase variation).
- Some sampling regimes are much more naturally functional (sparse/irregular).
- Sometimes there is no a priori discretisation (except machine precision).
- Should exploit the natural ordering and possible (even low degree) regularity

$\vdots$

- Ultimately:

For the same reasons science treats certain objects within the realm of functional analysis, statistics must also treat them within realm of functional data analysis.

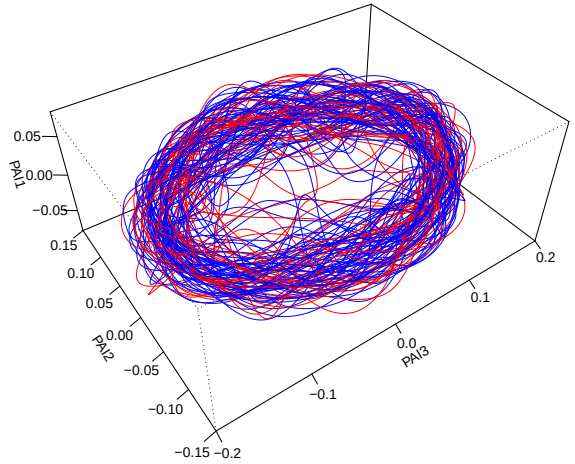
"Currently in the period of dynamic indeterminism in science, there is hardly a serious piece of research which, if treated realistically, does not involve operations on stochastic processes. The time has arrived for the theory of stochastic processes to become an item of usual equipment of every applied statistician."

Jerzy Neyman, 1960.

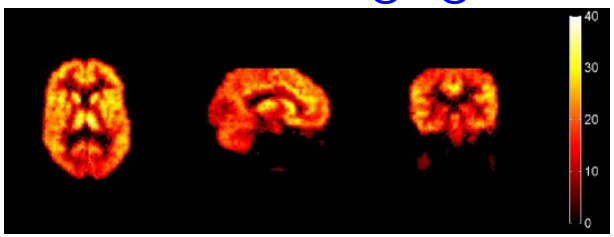
Finance

Handwriting Analysis

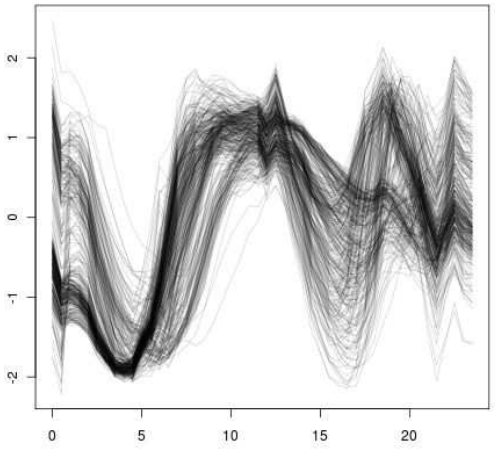
Biophysics



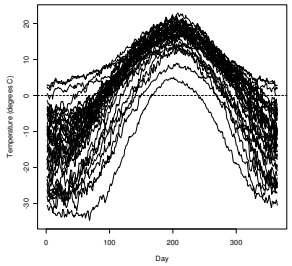
Brain Imaging



Energy



Environmetrics



What do you notice in the functional data plots?

In each scenario we had replicate realisations of a random function.

- This is to be compared with classical inference for random processes.
- The latter treated circumstances with single realisation.
- This required invariance and/or parametric/distributional assumptions.

Replication allows us to be ambitious and seek non-parametric methodology.

This distinguishes FDA from traditional inference for stochastic processes.

> *"If I were actively concerned with the analysis of data from stochastic processes (other than as related to spectra), I believe that I should try to seek out techniques of data processing which were not too closely tied to individual models, which might be likely to be unexpectedly revealing, and which were being pushed by the needs of actual data analysis."*
>
> John W. Tukey, The Future of Data Analysis

Compare the two '60s quotes of Neyman.

- Stochastic process: a collection of jointly distributed real random variables:

$$\{X_t : t \in T\}, \qquad \forall t \in T, \omega \mapsto X_t(\omega) \text{ real-valued measurable map}$$

- Random vector: a random element of some infinite dimensional vector space:

$$X, \qquad \omega \mapsto X(\omega) \text{ vector space-valued measurable map}$$

In finite dimensions (finite T) there is no distinction – a random vector is

- the jointly distributed collection of its coordinates, $X = (X_1, \ldots, X_p)^\top$
- a random element of $\mathbb{R}^p$
- They are related via the evaluation functionals: $X \mapsto \langle X, e_j \rangle = X_j$

But in infinite dimensions the two are not always equivalent

FDA seamlessly blends the two, but we need to be aware.

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

Let $(\mathcal{S}, \mathcal{T})$ be a topological space:

- The closure $\overline{A}$ of a $A \subseteq \mathcal{S}$ is the intersection of all closed supersets of A.
- A set $A \subseteq B \subseteq \mathcal{S}$ is dense in $B \subseteq \mathcal{S}$ if $\overline{A} = B$ in the B-subspace topology.
- A subset $B \subseteq \mathcal{S}$ is separable there is a countable A that is dense in B.
- A is compact if every cover of A has a finite sub-cover.
- A is sequentially compact if every sequence in A has a subsequence that converges in A.

Let $(\mathcal{M}, d)$ be a metric space:

- Denseness is transitive and separability is inherited.
- M is compact if and only if it is sequentially compact.
- $A \subseteq \mathcal{M}$ is relatively compact (or pre-compact) if $\overline{A}$ is compact.
- a sequence $\{x_n\}_{n=1}^{\infty} \subset \mathcal{M}$ is Cauchy, if $\sup_{m,n>N} d(x_m, x_n) \overset{N\to\infty}{\longrightarrow} 0$.
- $\mathcal{M}$ is complete if every Cauchy sequence is convergent.
- $\mathcal{M}$ is totally bounded if, for any $\epsilon > 0$, it admits a finite cover of ϵ-balls.
- (Heine-Borel) if $\mathcal{M}$ is complete, then $A \subset \mathcal{M}$ is totally bounded if and only if it is pre-compact. (hence A is compact $\Leftrightarrow$ A is closed and totally bounded)

Let $(\mathcal{V}, \|\cdot\|)$ be a normed vector space over the reals:

- A subset $\mathcal{V}' \subset \mathcal{V}$ is a subspace if $(\mathcal{V}', \|\cdot\|)$ is itself a normed vector space.
- The span of $A \subset \mathcal{V}$, $\mathrm{span}(A)$ is the smallest subspace containing A (smallest w.r.t. set inclusion order, i.e. the intersection of all subspaces containing A).
- Equiv., $\mathrm{span}(A)$ is the set of all finite linear combinations of A-elements.
- A set $A \subset \mathcal{V}$ is linearly independent if linear combinations of finite subsets thereof vanish only under identically zero coefficients.
- A Hamel basis is a linearly independent set A such that $\mathrm{span}(A) = \mathcal{V}$.
- If $\mathcal{V}$ has a finite Hamel basis, then $\mathcal{V}$ is called finite dimensional, and the dimension of $\mathcal{V}$, $\dim(\mathcal{V})$, is defined as the cardinality of this[1] Hamel basis.
- If $\dim(\mathcal{V}) = p < \infty$ then V is isometrically isomorphic to $\mathbb{R}^p$.

[1] or any other one, all Hamel bases must have the same cardinality.

- The map $d : \mathcal{V} \times \mathcal{V} \to [0, \infty)$ defined as $d(x, y) = \|x - y\|$ is a metric.
- Metric/topological statements made in the context of $\mathcal{V}$ without specifying the metric/topology, are always understood to be w.r.t. $d(x, y) = \|x - y\|$.
- Two norms on $\mathcal{V}$ are called equivalent if they generate the same topology.
- Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on $\mathcal{V}$ are equivalent if and only if

$$c\|x\|_1 \leq \|x\|_2 \leq C\|x\|_1$$

  for all $x \in \mathcal{V}$ and some fixed $c, C < \infty$.
- If $\dim(\mathcal{V}) < \infty$, then all norms on $\mathcal{V}$ are equivalent.
- If $\dim(\mathcal{V}) < \infty$, it is complete and separable regardless of choice of norm.

- A normed vector space $(\mathcal{V}, \|\cdot\|)$ that is complete is called a Banach space.
- If A is a subset of a separable Banach space $\mathcal{B}$, the closure $\overline{A}$ in $\mathcal{B}$ equals the completion of A in $\mathcal{B}$.
- Caution: a Banach space can be non-separable if it's not finite dimensional.
- (Baire) The intersection of countably many dense open sets in $\mathcal{B}$ is dense in $\mathcal{B}$
- Caution: a Banach space has countable Hamel basis **iff** it's finite dimensional.
- Classical Banach spaces that are not finite-dimensional include:
  - $C[0,1]$ with the supremum norm $||f||_\infty = \sup_{x\in[0,1]} |f(x)|$.
  - $L^p[0,1]$, $1 \leq p < \infty$, with the L^p-norm $||f||_p = \left(\int_0^1 f^p(u)du\right)^{1/p}$
  - $L^\infty[0,1]$ with the essential supremum norm

$$||f||_\infty = \operatorname*{ess\,sup}_{x\in[0,1]} |f(x)| := \inf\{u \in [0,1] : |A_u(f)| > 0\}$$

where $A_u(f) = \{x \in [0,1] : f(x) > u\}$ is the (strict) u-superlevel set of f.

- All but the last example are separable.

- One can define L^p-spaces of real functions on more general measure spaces
- For example real random variables $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ with norm $(\mathbb{E}|X|^p)^{1/p}$.
- Important inequalities (used to establish that L^p spaces are Banach):

Hölder: For all $1 \leq p \leq q \leq \infty$ with $p^{-1} + q^{-1} = 1$, we have $\|fg\|_1 \leq \|f\|_p \|g\|_q$

Minkowski: For all $p \geq 1$, $\|f + g\|_p \leq \|f\|_p + \|g\|_p$ (yields triangle inequality)

- Note that we cannot interpret functions in $L^p[0,1]$ pointwise:

$$\|f - g\|_p = 0 \iff |\{x : f(x) \neq g(x)\}| = 0$$

- And yet, $C[0,1]$ is dense in $L^p[0,1]$, for any $p \geq 1$.
- Surprising? And yet coherent:
  - L^p norms "blind" to perturbations over null sets, and
  - continuity cannot tolerate perturbations restricted to null sets[2].

[2] any nonempty open set has positive Lebesgue measure.

- If there exists a symmetric bilinear form $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ on a real normed vector space $(\mathcal{V}, \|\cdot\|)$ such that

$$\|x\|^2 = \langle x, x \rangle, \qquad \forall x \in \mathcal{V},$$

  then $\mathcal{V}$ is called an inner product space with inner product $\langle \cdot, \cdot \rangle$.

- A normed vector space is an inner product space if and only if its norm satisfies the parallelogram law,

$$\|x\|^2 + \|y\|^2 = \frac{\|x - y\|^2 + \|x + y\|^2}{2}$$

  in which case the polarisation identity elicits the underlying inner product,

$$\langle x, y \rangle = \frac{\|x + y\|^2 - \|x - y\|^2}{4}$$

- An inner product is continuous in the product norm topology.

- A Hilbert space is a Banach space with (norm generated by) an inner product.

- Consequently, $L^2[0,1]$ is a Hilbert space but $L^p[0,1]$ with $p \neq 2$ is not.
- We say vectors $x, y \in \mathcal{H}$ are orthogonal, and write $x \perp y$, when $\langle x, y \rangle = 0$.
- An orthonormal system is a set $E \subset \mathcal{H}$ comprised of unit norm vectors that are pairwise orthogonal, i.e. $\|x\| = 1$ and $\langle x, y \rangle = 0$ for all $E \ni x \neq y \in E$.
- An orthonormal system $E \subset \mathcal{H}$ is called complete (in $\mathcal{H}$) or total if

$$\overline{\text{span}(E)} = \mathcal{H}.$$

(equivalently, if $\langle e, x \rangle = 0 \, \forall \, e \in E \iff x = 0$.)

If furthermore E is countable, it is called a (countable) orthonormal basis.

ONBs and Separability

A Hilbert space $\mathcal{H}$ has a countable orthonormal basis if and only if it is separable.

- In the non-separable case, we can still show a complete orthonormal set always exists - provided we use Zorn's lemma.
- Consequence: the unit ball in $\mathcal{H}$ is compact if and only if $\dim(\mathcal{H}) < \infty$.

- Two Hilbert spaces $\{(\mathcal{H}_i, \langle \cdot, \cdot \rangle_i)\}_{i=1}^2$ are isometrically isomorphic if there exists a unitary map $U : \mathcal{H}_1 \to \mathcal{H}_2$ (linear bijection that preserves inner products)

Classification of separable Hilbert spaces

A separable Hilbert space is isometrically isomorphic to either $\mathbb{R}^p$ or ℓ_2, depending on whether it is finite dimensional or not.

- Unitary maps map countable orthonormal bases to countable orthonormal bases, so the isometry manifests via the implied coordinate systems.
- These are also known as Fourier coefficients.

(Generalised) Fourier Series in separable Hilbert spaces

Let $\mathcal{H}$ be a separable Hilbert space with countable orthonormal basis $\{e_n\}$. Then,

1. The sequence $\sum_{n=1}^N \alpha_n e_n$ converges in $\mathcal{H}$ if and only if $\{\alpha_n\} \in \ell_2$.
2. When $x = \lim_{N\to\infty} \sum_{n=1}^N \alpha_n e_n = \sum_{n=1}^\infty \alpha_n e_n$ exists we have $\alpha_n = \langle x, e_n \rangle$.
3. For any $y \in \mathcal{H}$ we have $y = \sum_{n=1}^\infty \langle y, e_n \rangle e_n$

Theorem (Projection Theorem)

Let $\mathcal{S}$ be a linear subspace of $\mathcal{H}$. If $\mathcal{S}$ is closed, then given any $y \in \mathcal{H}$

1. *the functional $f : \mathcal{S} \to [0, \infty)$,*

$$s \mapsto \|s - y\|$$

admits a unique minimiser $\hat{y} \in \mathcal{S}$.

2. *The unique minimiser $\hat{y} \in \mathcal{S}$ is characterised by the condition*

$$\langle y - \hat{y}, s \rangle = 0, \qquad \forall s \in \mathcal{S}.$$

The minimiser is called the projection of y onto $\mathcal{S}$, and the characterisation says that the residual $y - \hat{y}$ should be orthogonal to all elements of $\mathcal{S}$.

This motivates the definition of the orthogonal complement of a subspace $\mathcal{S} \in \mathcal{H}$

$$\mathcal{S}^{\perp} := \{x \in \mathcal{H} : \langle x, s \rangle = 0 \, \forall \, s \in \mathcal{S}\}.$$

So the characterisation (2) now reads that $y - \hat{y} \in \mathcal{S}^{\perp}$.

A consequence of the last theorem is that given a closed subspace $\mathcal{S}$ we can uniquely decompose any $x \in \mathcal{H}$ as $x = x_{\mathcal{S}} + x_{\mathcal{S}^\perp}$,with $x_{\mathcal{S}} \in \mathcal{S}, x_{\mathcal{S}^\perp} \in \mathcal{S}^\perp$. Recalling that a direct sum of two orthogonal subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ (meaning that $x \perp y$ for all $x \in \mathcal{S}_1$ and $y \in \mathcal{S}_2$) is defined as:

$$\mathcal{S}_1 \oplus \mathcal{S}_2 = \{x + y : x \in \mathcal{S}_1, y \in \mathcal{S}_2\}$$

we have essentially shown that:

Theorem (Projection Direct Sum Decomposition)

Given any closed subspace $\mathcal{S}$ we may decompose $\mathcal{H}$ into the direct sum

$$\mathcal{H} = \mathcal{S} \oplus \mathcal{S}^\perp.$$

Finally we consider what happens when we take the orthocomplement of an orthocomplement:

Proposition (Orthocomplements)

Let A be a subset of a Hilbert space $\mathcal{H}$. Then:

- *$A^\perp$ is a closed subspace of $\mathcal{H}$.*
- *$A \subset (A^\perp)^\perp$.*
- *if A is a subspace $(A^\perp)^\perp = \overline{A}$.*

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

When a normed vector space is finite dimensional, it is isomorphic to $\mathbb{R}^p$. Hence integration of vector-valued functions can be defined coordinate-wise.

For a general separable Banach space $\mathcal{B}$, one has to proceed from first principles (and there are more than one notion of integral that can be defined).

- Our goal: given measurable $f : (\Omega, \mathcal{F}, \mu) \to \mathcal{B}$, define the integral $\int f d\mu = \int f(\omega)\mu(d\omega)$ Banach-valued function f.
- Our tool: we know what $\int \|f(\omega)\|\mu(d\omega)$ means, so maybe we use that.
- Recalling that Lebesgue integrals are defined through simple functions, it is reasonable to start with $f : \Omega \to \mathcal{B}$ being a simple function, i.e. of the form

$$f(\omega) = \sum_{i=1}^{k} c_i \mathbf{1}\{\omega \in E_i\}, \quad E_1, \ldots, E_k \in \mathcal{F},\ c_1, ..., c_k \in \mathcal{B}.$$

- When $\mu(E_i) < \infty$ for all i, such f is said to be Bochner integrable, and

$$\int f d\mu \equiv \int_\Omega f(\omega)\mu(d\omega) = \sum_{i=1}^{k} c_i \mu(E_i)$$

  is called its Bochner integral.
- It can be seen that this is well-defined: the integral is invariant to re-parametrising f, since one can always partition the E_i into disjoint sets.

Like with Lebesgue integration, we can then go by approximation:

Bochner Integral

A measurable function $f : (\Omega, \mathcal{F}, \mu) \to \mathcal{B}$ is said to be Bochner integrable if there exists a sequence of simple Bochner integrable functions $\{f_n\}$ such that

$$\lim_{n\to\infty} \int_\Omega \|f(\omega) - f_n(\omega)\| \mu(d\omega) = 0.$$

In this case, the Bochner integral of f is defined as

$$\int f d\mu \equiv \int_\Omega f(\omega)\mu(d\omega) = \lim_{n\to\infty} \int f_n \, d\mu.$$

- The definition implicitly uses completeness: $\| \int f_n \, d\mu \| \leq \int \|f_n\| \, d\mu$ so that

$$\left\| \int f_n \, d\mu - \int f_m \, d\mu \right\| \leq \int \|f_n - f_m\| \, d\mu \leq \int \|f_n - f\| \, d\mu + \int \|f_m - f\| \, d\mu$$

  . Thus, $\{\int f_n \, d\mu\}$ is a Cauchy sequence in $\mathcal{B}$ and hence converges.

- Again, the value of $\int f d\mu$ is independent of the approximating sequence $\{f_n\}$.

(merge any two sequences and use same argument as above)

If f is Bochner integrable, then we can show $\int \|f\| d\mu < \infty$ (and, consequently, $\|\int f d\mu\| \leq \int \|f\| d\mu$). But the converse fails without further assumptions. One avenue is via "fidi approximation":

Theorem

Let $f : (\Omega, \mathcal{F}, \mu) \to \mathcal{B}$ be a measurable and $\int_{\mathcal{B}} \|f(\omega)\| \mu(d\omega) < \infty$. Suppose that

$$\lim_{n \to \infty} \int \|f - g_n\| d\mu = 0$$

for some sequence of Bochner-integrable (not necessarily simple) functions

$$g_n : (\Omega, \mathcal{F}, \mu) \to \mathcal{S}_n \subset \mathcal{B},$$

with $\{\mathcal{S}_n\}$ a sequence of finite dimensional subspaces of $\mathcal{B}$. Then, there exists an approximating sequence of simple functions, and f is Bochner integrable.

In a separable Hilbert space, we do get the converse:

Corollary

Given a separable Hilbert space $\mathcal{H}$, a measurable $f : (\Omega, \mathcal{F}, \mu) \to \mathcal{H}$ is Bochner integrable if and only if $\int_{\Omega} \|f(\omega)\| \mu(d\omega) < \infty$.

Finally, we can use the Lebesgue dominated convergence theorem to get a Bochner dominated convergence theorem:

(Bochner) Dominated Convergence Theorem

Let $\{f_n\}$ be a sequence of Bochner integrable functions valued in $\mathcal{B}$ that converges to some f valued in $\mathcal{B}$. If there exists a non-negative Lebesgue integrable function $g : (\Omega, \mathcal{F}, \mu) \to [0, \infty)$ such that

$$\|f_n(\omega)\| \leq g(\omega), \qquad \text{for } \mu - \text{almost all } \omega \text{ and for all } n \geq 1$$

then f is Bochner integrable, $\lim_{n\to\infty} \int \|f - f_n\| d\mu = 0$, and $\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$.

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

- The study of infinite-dimensional Hilbert spaces was motivated primarily by the need to study function spaces.
- But it is a prior not clear whether (or when) the norm's topology furnishes pointwise information.

Turns out that this question is settled by way of the notion of reproducing kernel:

Definition (Reproducing Kernel)

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space of real-valued functions defined over a set E. A bivariate function K on $E \times E \to \mathbb{R}$ is said to be a reproducing kernel for $\langle \cdot, \cdot \rangle$ on $\mathcal{H}$ if it satisfies the following two properties:

1. For every fixed $t \in E$, the function $K_t \equiv K(\cdot, t) : E \to \mathbb{R}$ belongs itself in $\mathcal{H}$.
2. K_t is reproducing for $\langle \cdot, \cdot \rangle$: for all $f \in \mathcal{H}$ and $t \in E$, one has $f(t) = \langle f, K_t \rangle$.

When a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ possesses reproducing kernel, it is said to be a Reproducing Kernel Hilbert Space (RKHS).

A quick sanity check reveals that these definitions are not empty:

- One can readily check that $\mathbb{R}^p$ with the usual inner product is an RKHS, taking $E = \{1, ..., p\}$ and defining $K(s, t) = \sum_{i=1}^{p} e_i(s) e_j(t)$ where $\{e_i\}$ is the canonical basis.
- Therefore, all finite-dimensional Hilbert spaces are RKHS (by isometric isomorphism).

How can we know whether a kernel is reproducing?

Fortunately, we can characterise reproducing kernels as non-negative definite (a.k.a. positive semidefinite) kernels. Recall that a kernel $K : E \times E \to \mathbb{R}$ is positive-semidefinite (write $K \succeq 0$) if it satisfies:

1. $K(x, y) = K(y, x)$ for all $x, y \in E$.
2. $\sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \alpha_j K(x_i, x_j) \geq 0$ for all $p \in \mathbb{N}$, all $x_1, ..., x_p \in E$, and all $\alpha_1, ..., \alpha_p \in \mathbb{R}$.

A kernel is strictly positive definite (write $K \succ 0$) if it is non-negative definite, and the inequality in (2) is strict unless $\alpha_1 = \ldots \alpha_p = 0$.

Theorem (Moore-Aronsjan)

The reproducing kernel of a given RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is unique and non-negative definite. Conversely, given a non-negative definite kernel K on $E \times E$, there exists a unique RKHS $(\mathcal{H}(K), \langle \cdot, \cdot \rangle_K)$ of functions on E with K as reproducing kernel.

Caution: When we speak of a Hilbert space, we really mean the set of its elements and the associated inner product. The same set of functions can be furnished with two different inner products (meaning all have finite norm) – resulting in two different Hilbert spaces.

So two different positive-definite functions may be reproducing kernels for the same set of functions, but they will be so under different inner products (note that the 'reproducing property' is with respect to a specific inner product).

Proof

If two kernels K_1 and K_2 for $\mathcal{H}$ have the reproducing property for $(\mathcal{H}, \langle \cdot, \cdot \rangle)$,

$$f(t) = \langle f, K_1(\cdot, t) \rangle = \langle f, K_2(\cdot, t) \rangle \ \Rightarrow \langle f, K_1(\cdot, t) - K_2(\cdot, t) \rangle = 0$$

for all $f \in \mathcal{H}$ and all $t \in E$, which implies that $K_1 = K_2$. Since $K_t \in \mathcal{H}$,

$$K(s,t) = \langle K_t, K_s \rangle = \langle K(\cdot, t), K(\cdot, s) \rangle = \langle K(\cdot, s), K(\cdot, t) \rangle = K(t,s).$$

Finally using both the reproducing property and the fact that $K_t \in \mathcal{H}$,

$$\sum_{i=1}^{k} \sum_{j=1}^{k} a_i \, a_j \, K(t_i, t_j) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_i \, a_j \, \langle K_{t_i}, K_{t_j} \rangle = \left\langle \sum_{i=1}^{k} a_i \, K_{t_i}, \sum_{j=1}^{k} a_j \, K_{t_j} \right\rangle \ \geq \ 0.$$

To show that a positive semidefinite-kernel generates a unique RKHS, define

$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{n} a_i \, K_{t_i} \, : \, a_i \in \mathbb{R}, t_i \in E, n \geq 1 \right\}.$$

to be the span $\operatorname{span}\left(\{ K_t \}_{t \in E} \right)$.

Define a symmetric bilinear form in $\mathcal{H}_0$

$$\left\langle \sum_{i=1}^{m} a_i K_{s_i}, \sum_{j=1}^{n} b_j K_{t_j} \right\rangle_0 := \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K(s_i, t_j).$$

We note that, by construction, K is reproducing on $\mathcal{H}_0$ for $\langle \cdot, \cdot \rangle_0$:

$$\langle f, K_t \rangle_0 = f(t), \qquad \forall\, t \in E \;\&\; f \in \mathcal{H}_0.$$

Now, we claim that $\langle \cdot, \cdot \rangle$ is an inner product. Since K is non-negative definite it must be that $\langle f, f \rangle_0 \geq 0$ for $f \in \mathcal{H}_0$. Defining $\|f\|_0 = \langle f, f \rangle_0$, it remains to show that $\|f\|_0 = 0 \implies f = 0$. To show this:

- We can directly verify that $\langle \cdot, \cdot \rangle_0$ satisfies the Cauchy-Schwarz inequality, i.e.,

$$|\langle f, g \rangle_0|^2 \leq \langle f, f \rangle_0 \langle g, g \rangle_0 = \|f\|_0^2 \|g\|_0^2, \qquad \forall\, f, g \in \mathcal{H}_0$$

- Next, since K is reproducing for $\langle \cdot, \cdot \rangle_0$ on $\mathcal{H}_0$, we have

$$|f(t)|^2 = |\langle f, K_t \rangle_0|^2 \leq \langle f, f \rangle_0 \langle K(\cdot, t), K(\cdot, t) \rangle_0 = \langle f, f \rangle_0 K(t, t)$$

  which vanishes if $\langle f, f \rangle_0 = 0$.

To complete our construction (pun intended), we must now complete $\mathcal{H}_0$ with respect to $\langle \cdot, \cdot \rangle_0$. If $\{f_n\}$ is a Cauchy sequence in $\mathcal{H}_0$, then, for any $t \in E$,

$$|f_n(t) - f_m(t)|^2 = |\langle f_n - f_m, K_t \rangle_0|^2 \leq \|f_n - f_m\|_0^2 K(t,t)$$

so $\{f_n(t)\}$ is a Cauchy sequence in $\mathbb{R}$, and hence $f_n(t)$ converges to a real limit. Collecting these limits as t ranges in E, we obtain a real function $f(t)$. We will now need to relate these pointwise limits to limits in the norm.

First, we treat a special case "lemma": if $\{f_n\}$ is Cauchy in $\mathcal{H}_0$ and converges pointwise to zero, then it also does so in the $\|\cdot\|_0$-norm (i.e. $\|f\|_0 \to 0$ as well).

To this aim, note that a Cauchy sequence is necessarily bounded, so let $\|f_n\|_0 < B < \infty$. Furthermore, the Cauchy property implies that for any $\epsilon > 0$ there exists an N such that $\|f_n - f_N\|_0 < \epsilon / B$ for all $n > N$. Now $f_N \in \mathcal{H}_0$ so,

$$f_N(x) = \sum_{i=1}^{p} \alpha_i K_{t_i}(t). \qquad \text{(for some } \{\alpha_i\} \text{ and } \{t_i\}\text{)}$$

Now, we can use Cauchy-Schwarz and the reproducing property to write

$$\|f_n\|_0^2 = \langle f_n - f_N, f_n \rangle_0 + \langle f_N, f_n \rangle_0 \leq \|f_n - f_N\|_0 \|f_n\|_0 + \sum_{i=1}^{p} \alpha_i f_n(t_i) \leq \epsilon + \sum_{i=1}^{p} \alpha_i f_n(t_i)$$

By pointwise convergence of f_n to 0 we can choose n sufficiently large to bound the second term on the RHS by ϵ. In summary: for any $\epsilon > 0$, there exists an n sufficiently large such that $||f_n||_0^2 < 2\epsilon$, so that pointwise convergence to zero implies norm convergence to zero.

Now let $\mathcal{H}$ be the collection of functions on E that are pointwise limits of Cauchy sequences in $\mathcal{H}_0$. For $f, g \in \mathcal{H}$, define

$$\langle f, g \rangle_{\mathcal{H}} = \lim_{n \to \infty} \langle f_n, g_n \rangle_0$$

where $\{f_n\}$ and $\{g_n\}$ Cauchy sequences in $\mathcal{H}_0$ converging pointwise f and g, respectively. To show that this is well-defined we need to:

- Show that the limit exists.
- Show that the limit is invariant to the choice of Cauchy sequences

For existence, using the polarisation identity and the reverse triangle inequality,

$$\left| \langle f_n, g_n \rangle_0 - \langle f_m, g_m \rangle_0 \right| = \frac{1}{4} \left| ||f_n + g_n||_0 - ||f_n - g_n||_0 - ||f_m + g_m||_0 + ||f_m - g_m||_0 \right|$$

$$\leq \frac{1}{4} \left(||(f_n - f_m) + (g_n - g_m)||_0 + ||(f_m - f_n) + (g_n - g_m)||_0 \right) \leq \frac{1}{2} \left(||f_n - f_m||_0 + ||g_n - g_m||_0 \right)$$

so $\langle f_n, g_n \rangle_0$ is Cauchy in $\mathbb{R}$ (since f_n, g_n are Cauchy in $\mathcal{H}_0$) and thus has a limit.

For invariance, take $\{\tilde{f}_n\}$ and $\{\tilde{g}_n\}$ are another pair of Cauchy sequences in $\mathcal{H}_0$ that converge to f and g pointwise, then $\{\tilde{f}_n - f_n\}$ and $\{\tilde{g}_n - g_n\}$ are Cauchy sequences in $\mathcal{H}_0$ converging to zero pointwise. Thus our "lemma" implies that $\|\tilde{f}_n - f_n\|_0 \to 0$ and $\|\tilde{g}_n - f_n\|_0 \to 0$. A second use of the polarisation identity and reverse triangle inequality now gives

$$\left|\langle f_n, g_n\rangle_0 - \langle \tilde{f}_n, \tilde{g}_n\rangle_0\right| = \frac{1}{4}\left|\|f_n + g_n\|_0 - \|f_n - g_n\|_0 - \|\tilde{f}_n + \tilde{g}_n\|_0 + \|\tilde{f}_n - \tilde{g}_n\|_0\right|$$

$$\leq \frac{1}{4}\left(\|(f_n - \tilde{f}_n) + (g_n - \tilde{g}_n)\|_0 + \|(\tilde{f}_n - f_n) + (g_n - \tilde{g}_n)\|_0\right) \leq \frac{1}{2}\left(\|f_n - \tilde{f}_n\|_0 + \|g_n - \tilde{g}_n\|_0\right)$$

So $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a Hilbert space. By direct calculation, we can furthermore verify that it admits K as a reproducing kernel, and so is an RKHS for K.

To show that $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is the only Hilbert space with kernel K, let $(\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$ be another RKHS with kernel K. Then, $\mathcal{H}_0$ is a subspace of $\mathcal{G}$. Hence, $\mathcal{G} = \mathcal{H}_0^{\perp} \oplus \overline{\mathcal{H}_0}$ (in the $\|\cdot\|_{\mathcal{G}}$-sense). So, for any $f \in \mathcal{H}_0^{\perp}$ ($\|\cdot\|_{\mathcal{G}}$-sense):

- on the one hand $f(t) = \langle f, K_t\rangle_{\mathcal{G}}$ since K is reproducing for $(\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$
- on the other hand $\langle f, K_t\rangle_{\mathcal{G}} = 0$ since $K_t \in \mathcal{H}_0$ and $f \in \mathcal{H}_0^{\perp}$.

So $\mathcal{H}_0^{\perp} = \{0\}$. Hence $\mathcal{G} = \overline{\mathcal{H}_0}$ i.e. $\mathcal{G}$ is the completion of $\mathcal{H}_0$. But K reproduces both $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{G}}$, so we must have $\langle\cdot,\cdot\rangle_{\mathcal{H}} = \langle\cdot,\cdot\rangle_{\mathcal{G}} = \langle\cdot,\cdot\rangle_0$ on $\mathcal{H}_0$. It follows that $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}}) = (\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$, as both are completions of $(\mathcal{H}_0, \langle\cdot,\cdot\rangle_0)$. □

Notice that our proof automatically establishes that

$$\mathcal{H}(K) = \overline{\operatorname{span}\left(\{K_t\}_{t\in E}\right)}^{\|\cdot\|_{\mathcal{H}(K)}}$$

i.e. finite linear combinations of K_t's are dense in $\mathcal{H}(K)$. Consequently:

Theorem (Separability)

If E is separable and K is continuous then $(\mathcal{H}(K), \langle\cdot,\cdot\rangle_K)$ is separable.

Proof.

Let $\{t_n\}$ be a countable dense subset of the separable set E. Then, by conituinty of K, the collection of functions $\{\sum_{i=1}^{k} a_i K(\cdot, t_i) : a_i \in \mathbb{Q}, k \geq 1\}$ is a countable dense subset of $\operatorname{span}\left(\{K_t\}_{t\in E}\right)$, which is in turn dense in $\mathcal{H}(K)$. By transitivity of denseness in metric spaces, we conclude that $\mathcal{H}(K)$ is separable. □

Theorem (Continuity)

If K is continuous near $\{(t,t) : t \in E\}$, functions in $\mathcal{H}(K)$ are continuous on E.

Proof.

If $f \in \mathcal{H}(K)$, then $|f(t) - f(s)| = |\langle f, K_t - K_s \rangle| \leq \|f\|_K \, \|K_t - K_s\|_K$. However,

$$\|K_t - K_s\|_K^2 = \langle K_t - K_s, K_t - K_s \rangle_K = K(t,t) + K(s,s) - 2K(t,s),$$

which converges to zero as $s \to t$ by the continuity of K near $\{(t,t) : t \in E\}$.

□

Corollary (continuity of Kernel and continuity near diagonal)

A reproducing kernel $K : E \times E \to \mathbb{R}$ is continuous everywhere if and only if it is continuous near the diagonal $\{(t,t) : t \in E\}$.

RKHS are distinguished among Hilbert spaces, because norm convergence implies pointwise convergence:

Theorem (From Norm to Pointwise Convergence)

Let $\mathcal{H}(K)$ be an RKHS containing functions of E. If $\{f_n\}$ is a sequence of functions in $\mathcal{H}(K)$ such that $\|f_n - f\|_K \to 0$, then $f_n(t) \to f(t)$ as $n \to \infty$ for all $t \in E$. If, furthermore, $\sup_{t \in E} K(t,t) < \infty$, then the convergence is uniform: $\sup_{t \in E} |f_n(t) - f(t)| \to 0$.

Proof.

Imitating a step in previous proofs, (use of reproducing property and Cauchy-Schwarz: $|f_n(t) - f(t)|^2 = |\langle f_n - f, K_t\rangle_K|^2 \leq \|f_n - f\|_K^2 K(t,t)$. □

- It follows that the evaluation functional $e_t : \mathcal{H}(K) \to \mathbb{R}$ given by

$$e_t(f) = f(t) = \langle f, K(\cdot, t)\rangle_K$$

  is a continuous linear map for all $t \in E$ for an RKHS $(\mathcal{H}(K), \langle\cdot,\cdot\rangle_K)$.
- Conversely we can ask, for $\mathcal{H}$ a Hilbert space of functions on E, when are the evaluation functionals continuous? Turns out this is only true for RKHS.

Thus, RKHSs are characterized by the continuity of the evaluation functionals:

Theorem (Evaluation Functionals and RKHS)

Let $\mathcal{H}$ be a Hilbert space of real functions on E. Then, the evaluation functionals are continuous maps, if and only if $\mathcal{H}$ is a RKHS.

The proof makes use of the Riesz representation theorem, and is postponed to that point (Riesz representation itself depends in no way on RKHS).

A natural question is: when are two RKHS topologically equivalent?

We answer this in two steps.

Proposition (Loewner order and RKHS inclusion)

Let $K_1, K_2 \succeq 0$ on $E \times E$. If there exists $C > 0$ s.t. $C^2 K_2 - K_1 \succeq 0$, then:

- $\mathcal{H}(K_1) \subset \mathcal{H}(K_2)$.
- *for any $f \in \mathcal{H}(K_1)$, $\|f\|_{K_1} \leq C\|f\|_{K_2}$.*

So using the above result twice, we obtain:

Proposition (Equivalence of RKHS norms)

If there exist $C, c > 0$ s.t. $C^2 K_2 \succeq K_1 \succeq c^2 K_2 \succeq 0$ on E, then:

- $\mathcal{H}(K_1) = \mathcal{H}(K_2)$.
- *the corresponding norms are equivalent, i.e. for any $f \in \mathcal{H}(K_1) \equiv \mathcal{H}(K_2)$*

$$c\|f\|_{K_2} \leq \|f\|_{K_1} \leq C\|f\|_{K_2}.$$

- This clarifies an earlier comment qualifying uniqueness, in that two *distinct* kernels can generate the same RKHS as a set, but still with different norm.

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

A map $\mathcal{T}$ with domain $\mathrm{dom}(\mathcal{T})$ and range $\mathrm{range}(\mathcal{T})$ is called a linear operator if

1. $\mathrm{dom}(\mathcal{T})$ and $\mathrm{range}(\mathcal{T})$ are vector (sub)spaces over the reals,
2. $\mathcal{T}(\alpha x + y) = \alpha \mathcal{T}(x) + \mathcal{T}(y)$, for all $x, y \in \mathrm{dom}(\mathcal{T})$ and all scalars $\alpha \in \mathbb{R}$

We will be primarily concerned with linear operators whose domain and range are subspaces of (possibly different) Banach spaces, say $(\mathcal{B}_1, \|\cdot\|_1)$ and $(\mathcal{B}_2, \|\cdot\|_2)$.

A linear operator is bounded on its domain if there exists $0 < C < \infty$ such that

$$\|\mathcal{T}x\|_2 \leq C\|x\|_1, \qquad \forall x \in \mathrm{dom}(\mathcal{T}).$$

Often we speak of bounded linear operators acting on a Banach space $\mathcal{B}$ without specifying the domain. It is then implicitly understood that the domain equals $\mathcal{B}$.

Caution: Linear operators may be bounded on a proper subspace of a Banach space, but not over the entire Banach space.

The kernel (or null space) of a linear operator $\mathcal{T}$ is defined as

$$\ker(\mathcal{T}) := \{x \in \mathrm{dom}(\mathcal{T}) : \mathcal{T}x = 0\}.$$

The rank of a linear operator $\mathcal{T}$ is $\mathrm{rank}(\mathcal{T}) := \dim(\mathrm{range}(\mathcal{T}))$.

Theorem (Boundedness and Continuity)

The following three statements are equivalent:

- *The linear operator* $\mathcal{T} : \mathrm{dom}(\mathcal{T}) \to \mathrm{range}(\mathcal{T})$ *is bounded on its domain.*
- *The linear operator* $\mathcal{T} : \mathrm{dom}(\mathcal{T}) \to \mathrm{range}(\mathcal{T})$ *is a continuous.*
- *The linear operator* $\mathcal{T} : \mathrm{dom}(\mathcal{T}) \to \mathrm{range}(\mathcal{T})$ *is Lipschitz continuous.*

Proof.

We will show (1) $\implies$ (3) $\implies$ (2) $\implies$ (1). Assuming (1), we have that

$$\|\mathcal{T}(x+u) - \mathcal{T}x\|_2 = \|\mathcal{T}(u)\|_2 \leq C\|u\|_2$$

which establishes Lipschitz continuity, and hence continuity. Now assume (2). Then there exists $\epsilon > 0$ such that for all $\|u\|_1 < 2\epsilon$ in $\mathrm{dom}(\mathcal{T})$,

$$\|\mathcal{T}(u)\|_2 = \|\mathcal{T}(u) - \mathcal{T}(0)\|_2 \leq 1.$$

Consequently, noting that (obviously) $\epsilon \frac{x}{\|x\|_1} < 2\epsilon$ for all $0 \neq x \in \mathrm{dom}(\mathcal{T})$

$$\|\mathcal{T}(x)\|_2 = \left\| \frac{\|x\|_1}{\epsilon} \mathcal{T}\left(\frac{\epsilon}{\|x\|_1} x \right) \right\|_2 = \frac{\|x\|_1}{\epsilon} \left\| \mathcal{T}\left(\epsilon \frac{x}{\|x\|_1} \right) \right\|_2 \leq \frac{\|x\|_1}{\epsilon} \cdot 1$$

□

- The vector space $\mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ of bounded linear maps from $\mathcal{B}_1$ to $\mathcal{B}_2$ equipped with the operator norm,

$$\left|\left|\left|\mathcal{T}\right|\right|\right|_\infty = \sup_{x \in \mathcal{B}_1 : ||x||_1 = 1} ||\mathcal{T}x||_2, \qquad \mathcal{T} \in \mathcal{B}_1,$$

  is a Banach space. If $\mathcal{B}_1 = \mathcal{B}_2 = \mathcal{B}$, we denote this space by $\mathcal{B}(\mathcal{B})$.

- By definition of boundedness, for any $x \in \mathcal{B}_1$, we have

$$||\mathcal{T}x||_2 \leq \left|\left|\left|\mathcal{T}\right|\right|\right|_\infty ||x||_1.$$

- The Banach space $\mathcal{B}^* := \mathcal{B}(\mathcal{B}, \mathbb{R})$ is called the dual space of $\mathcal{B}$, and its elements are called bounded linear functionals.

- We say that sequence $\{x_n\}$ in a Banach $\mathcal{B}$ converges weakly to $x \in \mathcal{B}$ if the real sequence $\mathcal{T}x_n$ converges to $\mathcal{T}x$ for all $\mathcal{T} \in \mathcal{B}^*$.

A couple of examples:

1. Evaluation functionals on $\mathcal{B}_1 = C[0,1]$: Let $t \in [0,1]$ be an arbitrary point, and define the map $\mathcal{T}_t : f \mapsto f(t)$ from $\mathcal{B}_1 = C[0,1]$ to $\mathcal{B}_2 = \mathbb{R}$. Clearly, $\mathcal{T}_t$ is linear, and $||\mathcal{T}_t f||_2 = |f(t)| \leq ||f||_1 = \sup_{s \in [0,1]} |f(s)|$ with equality holding iff f constant function.
2. Hilbert-Schmidt integral operators on $\mathcal{B}_1 = L_2[0,1]$: Let $\mathcal{B}_1 = L_2[0,1]$ and define the linear map $\mathcal{T}$ by $(\mathcal{T}f)(\cdot) = \int_0^1 K(\cdot, t) f(t) dt$ for $f \in L_2[0,1]$ and $K(\cdot,\cdot) \in L_2([0,1]^2)$. Then, $\mathcal{T}f \in L_2[0,1]$. Further, $||\mathcal{T}f||^2 \leq ||f||^2 \int_0^1 \int_0^1 K^2(s,t) ds dt$, so $\left|\left|\left|\mathcal{T}\right|\right|\right|_\infty \leq ||K||$.

A bounded linear operator $\mathcal{T} \in \mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ is said to be

- 1-to-1 if $\ker(\mathcal{T}) = \{0\}$.
- onto if $\text{range}(\mathcal{T}) = \mathcal{B}_2$.
- bijective if it is 1-to-1 and onto.
- open if $\mathcal{T}(U) \subset \mathcal{B}_2$ is open for all open $U \subset \mathcal{B}_1$.

The identity operator $\mathcal{I} \in \mathcal{B}(\mathcal{B})$ uniquely satisfies $x \mapsto x$ for all $x \in \mathcal{B}$.

If $\mathcal{T} \in \mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ is bijective, it has an inverse: an operator $\mathcal{T}^{-1} : \mathcal{B}_2 \to \mathcal{B}_1$ s.t.:

$$\mathcal{T}^{-1}\mathcal{T} = \mathcal{I}.$$

The inverse $\mathcal{T}^{-1}$ of a bijective operator $\mathcal{T}$ is unique and is invertible with inverse

$$[\mathcal{T}^{-1}]^{-1} = \mathcal{T}$$

When $\mathcal{T}$ is linear, then so is $\mathcal{T}^{-1}$.

- Uniqueness: if $\mathcal{A}$ and $\mathcal{B}$ are both inverses then $(\mathcal{A} - \mathcal{B})\mathcal{T}x = 0$ for all $x \in \mathcal{B}_1$, or equivalently (since $\mathcal{T}$ is onto) we have $(\mathcal{A} - \mathcal{B})y = 0$ for all $y \in \mathcal{B}_2$.
- Inverse: for any $y \in \mathcal{B}_2$, there is a unique $x \in \mathcal{B}_1$ such that $y = \mathcal{T}x$ since $\mathcal{T}$ is a bijection. Hence $\mathcal{T}(\mathcal{T}^{-1})y = \mathcal{T}(\mathcal{T}^{-1})\mathcal{T}x = \mathcal{T}x = y$ showing the existence/form of the inverse.
- Linearity: $\forall\ y_1, y_2 \in \mathcal{B}_2$ there exist $x_1, x_2 \in \mathcal{B}_1$ s.t. $y_i = \mathcal{T}x_i$ & $\mathcal{T}^{-1}y_i = x_i$, $i = 1, 2$. Thus $\mathcal{T}^{-1}(y_1 + cy_2) = \mathcal{T}^{-1}(\mathcal{T}x_1 + c\mathcal{T}x_2) = \mathcal{T}^{-1}(\mathcal{T}(x_1 + cx_2)) = x_1 + cx_2 = \mathcal{T}^{-1}y_1 + c\mathcal{T}^{-1}y_2$.

Theorem (Open Mapping theorem)

If $\mathcal{T} \in \mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ is onto, then it is open.

The proof is a consequence of the Baire property of Banach spaces.

Corollary (Banach Inverse Theorem)

If $\mathcal{T} \in \mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ is bijective, then $\mathcal{T}^{-1} \in \mathcal{B}(\mathcal{B}_2, \mathcal{B}_1)$.

Proof.

When $\mathcal{T}$ is bijective, then $\mathcal{T}^{-1}$ exists and is bijective (and thus onto), so by the open map theorem $(\mathcal{T}^{-1})^{-1}(U) = \mathcal{T}(U)$ is open for all open $U \subset \mathcal{B}_1$. Hence, $\mathcal{T}^{-1}$ is continuous. □

Theorem (Unit Perturbations of the Identity)

For $\mathcal{B}$ a Banach space let $\mathcal{T} \in \mathcal{B}(\mathcal{B})$ with $\left|\left|\left|\mathcal{T}\right|\right|\right|_\infty < 1$. Then $\mathcal{I} - \mathcal{T}$ is bijective and

$$(\mathcal{I} - \mathcal{T})^{-1} = \mathcal{I} + \sum_{j=1}^{\infty} \mathcal{T}^j.$$

For perturations of the form $\mathcal{I} + \mathcal{T}$, the summation would be over alternating signs.

Proof

$(\mathcal{I} - \mathcal{T})x = 0 \implies x = \mathcal{T}x$ and so $||x|| = ||\mathcal{T}x||$. But $||\mathcal{T}x|| \leq \left|\left|\left|\mathcal{T}\right|\right|\right|_\infty ||x|| < ||x||$ since $\left|\left|\left|\mathcal{T}\right|\right|\right|_\infty < 1$. So $x \in \ker(\mathcal{T})$ implies that $||x|| = 0$, and $\mathcal{I} - \mathcal{T}$ is bijective. Since $\left|\left|\left|\mathcal{T}\right|\right|\right|_\infty < 1$, we also have $\sum_{j=1}^{\infty} \left|\left|\left|\mathcal{T}\right|\right|\right|_\infty^j < \infty$. By the triangle inequality, the partial sum sequence $\mathcal{S}_n := \mathcal{I} + \sum_{j=1}^{n} \mathcal{T}^j$ is Cauchy in $\mathcal{B}(\mathcal{B})$ and thus has a limit in $\mathcal{B}(\mathcal{B})$, say $\mathcal{S} = \mathcal{I} + \sum_{j=1}^{\infty} \mathcal{T}^j$. Now $\left|\left|\left|\mathcal{I} - \mathcal{S}_n(\mathcal{I} - \mathcal{T})\right|\right|\right|_\infty = \left|\left|\left|\mathcal{T}^{n+1}\right|\right|\right|_\infty \leq \left|\left|\left|\mathcal{T}\right|\right|\right|_\infty^{n+1} \to 0$. So $\mathcal{S}(\mathcal{I} - \mathcal{T}) = \mathcal{I}$ □

Lemma (Sherman-Morrison-Woodbury identity)

For operators $\mathcal{S}$, $\mathcal{T}$, $\mathcal{U}$ and $\mathcal{V}$ with $\mathcal{S}$ and $\mathcal{T}$ invertible, we have

$$(\mathcal{T} + \mathcal{U}\mathcal{S}^{-1}\mathcal{V})^{-1} = \mathcal{T}^{-1} - \mathcal{T}^{-1}\mathcal{U}(\mathcal{S} + \mathcal{V}\mathcal{T}^{-1}\mathcal{U})^{-1}\mathcal{V}\mathcal{T}^{-1}.$$

(proof by direct verification, multiply RHS by $\mathcal{T} + \mathcal{U}\mathcal{S}^{-1}\mathcal{V}$)

Theorem (Riesz representation)

Any Hilbert space $(\mathcal{H}, \|\cdot\|)$ is isometrically isomorphic to its dual $\mathcal{H}^*$. Said differently, $\mathcal{T} \in \mathcal{B}(\mathcal{H}, \mathbb{R})$ be a bounded linear functional. Then, there is a unique element $r_{\mathcal{T}} \in \mathcal{H}$, called the representer of $\mathcal{T}$, such that

$$\mathcal{T}x = \langle x, r_{\mathcal{T}} \rangle, \qquad \forall\, x \in \mathcal{H}.$$

Furthermore, $\left\|\mathcal{T}\right\|_{\infty} = \|r_{\mathcal{T}}\|$.

Proof

If $\mathcal{T} \equiv 0$, take $r_{\mathcal{T}} = 0$. Otherwise, consider $\ker(\mathcal{T})^{\perp}$. This is a closed subspace of $\mathcal{H}$ by an earlier result. Choose $y \in \ker(\mathcal{T})^{\perp}$ such that $\mathcal{T}y = 1$. Then, $\forall\, x \in \mathcal{H}$,

$$\mathcal{T}(x - (\mathcal{T}x)y) = \mathcal{T}x - \mathcal{T}x\mathcal{T}y = \mathcal{T}x - \mathcal{T}x = 0,$$

i.e., $x - (\mathcal{T}x)y \in \ker(\mathcal{T})$. As $y \in \ker(\mathcal{T})^{\perp}$, we have $\langle y, x - (\mathcal{T}x)y \rangle = 0$, which implies that $\langle x, y \rangle = \mathcal{T}x\langle y, y \rangle = \mathcal{T}x\|y\|^2$. So we may take $r_{\mathcal{T}} = y/\|y\|^2$ and get

$$\mathcal{T}x = \langle r_{\mathcal{T}}, x \rangle \qquad \forall x \in \mathcal{H}.$$

To show uniqueness If we could find another representer $r'_{\mathcal{T}}$, then $\langle x, r_{\mathcal{T}} - r'_{\mathcal{T}} \rangle = 0$ for all $x \in \mathcal{H}$, which implies that $r_{\mathcal{T}} = r'_{\mathcal{T}}$. □

Corollary (Existence and Uniqueness of the Adjoint)

Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be Hilbert spaces. To every $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ corresponds a unique $\mathcal{T}^* \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_1)$ determined by the relation

$$\langle \mathcal{T}x_1, x_2\rangle_2 = \langle x_1, \mathcal{T}^* x_2\rangle_1 \qquad \forall, x_1 \in \mathcal{H}_1, x_2 \in \mathcal{H}_2.$$

- The operator $\mathcal{T}^*$ is called the adjoint of $\mathcal{T}$.
- When case $\mathcal{H}_1 = \mathcal{H}_2$, we say that $\mathcal{T}$ is self-adjoint if $\mathcal{T}^* = \mathcal{T}$.

Proof

The functional $\langle \mathcal{T}x_1, x_2\rangle_2$, seen as a function of x_1 for fixed x_2, is bounded and linear. So, by the Riesz representation, there exists a unique $y \in \mathcal{H}_1$ (depending on x_2) such that $\langle \mathcal{T}x_1, x_2\rangle_2 = \langle x_1, y\rangle_1$. Thus, we take $\mathcal{T}^* x_2 = y$. This definition gives us a linear mapping. To see that it is bounded, first note that for any $x_1 \in \mathcal{H}_1$, we have $\langle \mathcal{T}^* x_2, x_1\rangle_1 = \langle y, x_1\rangle = \langle x_2, \mathcal{T}x_1\rangle$. So,

$$\|\mathcal{T}^* x_2\|_1^2 = \langle \mathcal{T}^* x_2, \mathcal{T}^* x_2\rangle = \langle x_2, \mathcal{T}(\mathcal{T}^* x_2)\rangle \leq |||\mathcal{T}|||_\infty \|\mathcal{T}^* x_2\|_1 \|x_2\|_2,$$

which implies that $\|\mathcal{T}^* x_2\|_1 \leq |||\mathcal{T}|||_\infty \|x_2\|_2$. □

Note that by the Riesz representation, $x_n \to x$ weakly in a Hilbert space $\mathcal{H}$ if and only if $\langle x_n, y\rangle_{\mathcal{H}} \to \langle x, y\rangle$ for all $y \in \mathcal{H}$.

- Therefore, if x_n converges weakly to y in $\mathcal{H}_1$, then $\mathcal{T}x_n \to \mathcal{T}x$ weakly in $\mathcal{H}_2$:

$$\langle \mathcal{T}x_n, z\rangle_2 = \langle x_n, \mathcal{T}^* z\rangle_1 \to \langle x, \mathcal{T}^* z\rangle_1 = \langle \mathcal{T}x, z\rangle_2, \quad \forall, z \in \mathcal{H}_2$$

.

Some specific examples:

1. Matrices: when $\mathcal{H} = \mathbb{R}^d$, any $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ is a $d \times d$ matrix $\mathbf{T}$, and $\mathcal{T}^*$ is the linear transformation associated with the matrix $\mathbf{T}^\mathsf{T}$.

2. Integral operators on $L^2[0,1]$:
$\langle \mathcal{T}f, g\rangle = \int_0^1 \int_0^1 K(s,t)f(t)g(s)dtds = \langle f, \int_0^1 K(s,\cdot)g(s)ds\rangle$. Thus,
$(\mathcal{T}^* g)(\cdot) = \int_0^1 K(s,\cdot)g(s)ds$. So, $\mathcal{T}$ is self-adjoint if K is symmetric.

3. Evaluation maps: Let $\mathcal{H}(K)$ be an RKHS of functions on E and $\mathcal{T}_t$ be the evaluation functional corresponding to a fixed $t \in E$. Then, for $\alpha \in \mathbb{R}$, we have $\alpha\mathcal{T}_t(f) = \alpha f(t) = \alpha\langle f, K(\cdot,t)\rangle = \langle f, \alpha K(\cdot,t)\rangle$. Thus,
$\mathcal{T}_t^*(\alpha) = \alpha K(\cdot,t)$ for $t \in E$. Also, $\left|\left|\left|\mathcal{T}_t\right|\right|\right|_\infty = \left|\left|\left|\mathcal{T}_t^*\right|\right|\right|_\infty = K^{1/2}(t,t)$.

Theorem

For $(\mathcal{H}_i, \langle \cdot, \cdot \rangle_i)$, $i = 1, 2$ two Hilbert spaces, and $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$, we have

$$|||\mathcal{T}|||_\infty = \sup\{|\langle \mathcal{T}f, g \rangle_2| : \|f\|_1 = 1, \|g\|_2 = 1\}.$$

Furthermore, if $\mathcal{H}_1 \equiv \mathcal{H}_2$ and $\mathcal{T}$ is self-adjoint, then

$$|||\mathcal{T}|||_\infty = \sup\{|\langle \mathcal{T}f, f \rangle| : \|f\| = 1\}.$$

Proof

Assume $\mathcal{T} \neq 0$ wlog and denote the above supremum by M. By Cauchy-Schwarz,

$$M = |\langle \mathcal{T}f, g \rangle_2| \leq \|\mathcal{T}f\|_2 \|g\|_2 \leq |||\mathcal{T}|||_\infty \|g\|_2 = |||\mathcal{T}|||_\infty.$$

For the reverse, for any $\|x\| = 1$ with $\mathcal{T}x \neq 0$ (one exists since $\mathcal{T} \neq 0$). Then,

$$M \geq \left| \left\langle \mathcal{T}x, \frac{\mathcal{T}x}{\|\mathcal{T}x\|_2} \right\rangle_2 \right| = \frac{\|\mathcal{T}x\|_2^2}{\|\mathcal{T}x\|_2} = \|\mathcal{T}x\|.$$

which implies that $M \geq |||\mathcal{T}|||_\infty$.

Now consider the self-adjoint, and this time set $M = \sup\{|\langle \mathcal{T} f, f\rangle| : \|f\| = 1\}$.
Just as in the first part of the proof, $\|\|\mathcal{T}\|\|_\infty \geq M$.
For the reverse inequality, note that if $f, g \in \mathcal{H}$ satisfy $\|f\| = \|g\| = 1$, then the polarization identity holds since $\mathcal{T} = \mathcal{T}^*$.

$$4\langle \mathcal{T} f, g\rangle = \langle \mathcal{T}(f+g), (f+g)\rangle - \langle \mathcal{T}(f-g), (f-g)\rangle.$$

Since $|\langle \mathcal{T} h, h\rangle| \leq M\|h\|^2$, it follows that

$$|\langle \mathcal{T} f, g\rangle| \leq M\{\|f+g\|^2 + \|f-g\|^2\}/4 = M\{\|f\|^2 + \|g\|^2\}/2 = M.$$

Thus, the first part of the proof implies that $\|\|\mathcal{T}\|\|_\infty \leq M$. □

Theorem

Let $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Then,

1. $(\mathcal{T}^*)^* = \mathcal{T}$,
2. $\left\|\mathcal{T}^*\right\|_\infty = \left\|\mathcal{T}\right\|_\infty$ and $\left\|\mathcal{T}^*\mathcal{T}\right\|_\infty = \left\|\mathcal{T}\right\|_\infty^2$,
3. $\ker(\mathcal{T}) = (\text{range}(\mathcal{T}^*))^\perp$,
4. $\ker(\mathcal{T}^*\mathcal{T}) = \ker(\mathcal{T})$ and $\overline{\text{range}(\mathcal{T}^*\mathcal{T})} = \overline{\text{range}(\mathcal{T}^*)}$,
5. $\mathcal{H}_1 = \ker(\mathcal{T}) \oplus \overline{\text{range}(\mathcal{T}^*)} = \ker(\mathcal{T}^*\mathcal{T}) \oplus \overline{\text{range}(\mathcal{T}^*\mathcal{T})}$, and
6. $\text{rank}(\mathcal{T}^*) = \text{rank}(\mathcal{T})$.

It's worth reflection on (3), (4), and (5) in the self-adjoint case.

Proof

Part (1) follows from definition. For part (2), let $x_i \in \mathcal{H}_i$ for $i = 1, 2$. Recall from an earlier proof that $\left\|\mathcal{T}^*\right\|_\infty \leq \left\|\mathcal{T}\right\|_\infty$. This used in conjunction with part (1) yields $\left\|\mathcal{T}\right\|_\infty \leq \left\|\mathcal{T}^*\right\|_\infty$ proving $\left\|\mathcal{T}\right\|_\infty = \left\|\mathcal{T}^*\right\|_\infty$. From this identity, and using the definition of operator norm, we can get that $\left\|\mathcal{T}^*\mathcal{T}\right\|_\infty \leq \left\|\mathcal{T}\right\|_\infty^2$. On the other hand, $\|\mathcal{T}x_1\|_2^2 = \langle \mathcal{T}^*\mathcal{T}x_1, x_1\rangle_1 \leq \left\|\mathcal{T}^*\mathcal{T}\right\|_\infty \|x_1\|_1^2$, which implies $\left\|\mathcal{T}\right\|_\infty^2 \leq \left\|\mathcal{T}^*\mathcal{T}\right\|_\infty$.

For part (3), let $x_1 \in \ker(\mathcal{T})$ in which case, $\langle x_1, \mathcal{T}^* x_2 \rangle_1 = 0$ for all $x_2 \in \mathcal{H}_2$. Thus, $x_1 \in (\text{range}(\mathcal{T}^*))^\perp$. Conversely, let $x_1 \in (\text{range}(\mathcal{T}^*))^\perp$. Then, since $\mathcal{T}^*\mathcal{T}x_1 \in \text{range}(\mathcal{T}^*)$, we have $\|\mathcal{T}x_1\|_2^2 = \langle x_1, \mathcal{T}^*\mathcal{T}x_1 \rangle_1 = 0$.

For part (4), if $x_1 \in \ker(\mathcal{T})$, then of course $x_1 \in \ker(\mathcal{T}^*\mathcal{T})$. If $x_1 \in \ker(\mathcal{T}^*\mathcal{T})$, then $0 = \langle x_1, \mathcal{T}^*\mathcal{T}x_1 \rangle_1 = \|\mathcal{T}x_1\|_1^2$, which proves the first identity of part (4). The second identity follows from the first one, part (3), and the fact that for any subspace $\mathcal{M}$, $(\mathcal{M}^\perp)^\perp = \overline{\mathcal{M}}$.

Part (5) follows from parts (3) and (4) and the fact that for any closed subspace $\mathcal{M}$ of $\mathcal{H}_1$, we have $\mathcal{H}_1 = \mathcal{M} \oplus \mathcal{M}^\perp$.

For part (6), first assume that $\text{rank}(\mathcal{T}) < \infty$ so that $\text{range}(\mathcal{T})$ is finite dimensional and hence closed. Applying part (5), we have that for $x \in \mathcal{H}_2$, $\mathcal{T}^* x = \mathcal{T}^* x'$, where x' is the projection of x onto $\text{range}(\mathcal{T})$. So, $\text{range}(\mathcal{T}^*) \subset \mathcal{T}^*(\text{range}(\mathcal{T})) \Rightarrow \text{rank}(\mathcal{T}^*) \leq \text{rank}(\mathcal{T}) < \infty$. Interchanging $\mathcal{T}$ and $\mathcal{T}^*$ yields $\text{rank}(\mathcal{T}) \leq \text{rank}(\mathcal{T}^*)$ implying that $\text{rank}(\mathcal{T}) = \text{rank}(\mathcal{T}^*)$ if one of these ranks is finite. If one of them is infinite, the same argument also shows that the other rank must be infinite. □

Definition

An operator $\mathcal{T}$ on a Hilbert space $\mathcal{H}$ is said to be non-negative definite (or simply non-negative) if it is self-adjoint and $\langle \mathcal{T}x, x \rangle \geq 0$ for all $x \in \mathcal{H}$. It is called positive definite (or just positive) if strict inequality holds for all $x \neq 0$. For two operators $\mathcal{T}_1$ and $\mathcal{T}_2$, we write $\mathcal{T}_1 \preceq \mathcal{T}_2$ (respectively, $\mathcal{T}_1 \prec \mathcal{T}_2$) if $\mathcal{T}_2 - \mathcal{T}_1$ is non-negative (respectively, positive) definite.

For any operator $\mathcal{T}$, we can verify that the operator $\mathcal{T}^*\mathcal{T}$ is non-negative definite.

Theorem

Let $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ for a Hilbert space $\mathcal{H}$. If $\mathcal{T}$ is non-negative (respectively, positive), then there exists a unique non-negative (respectively, positive) operator $\mathcal{S} \in \mathcal{B}(\mathcal{H})$ such that $\mathcal{S}^2 = \mathcal{T}$ and $\mathcal{S}$ commutes with any operator that commutes with $\mathcal{T}$.

We use the notation $\mathcal{T}^{1/2}$ for $\mathcal{S}$, and it is called the square root operator of $\mathcal{T}$.

- Note that the non-negative operator is only assumed bounded.
- For compact operator (to be defined soon) the spectral theorem (to be proven soon) will directly give this result, but compactness is not necessary.

Proof

Assume w.l.o.g. that $\left\|\mathcal{T}\right\|_\infty \leq 1$ so that we can verify that $\left\|\mathcal{I} - \mathcal{T}\right\|_\infty \leq 1$. The proof relies on the fact that the Maclaurin expansion $(1-z)^{1/2} = 1 + \sum_{j=1}^{\infty} c_j z^j$ is absolutely convergent for all $|z| \leq 1$ with all the c_j's being negative.

Consequently, the series $\mathcal{S}_n := \mathcal{I} + \sum_{j=1}^{n} c_j (\mathcal{I} - \mathcal{T})^j$ is Cauchy and must therefore converge to some operator $\mathcal{S} \in \mathcal{B}(\mathcal{H})$. We can directly verify that $\mathcal{S}$ is self-adjoint.

Writing $\mathcal{S} = \mathcal{I} + \sum_{j=1}^{\infty} c_j (\mathcal{I} - \mathcal{T})^j$, we can rearrange terms by absolute convergence to show that $\mathcal{S}^2 = \mathcal{T}$. Now,

$$\langle \mathcal{S}x, x \rangle = 1 + \sum_{j=1}^{\infty} c_j \langle (\mathcal{I} - \mathcal{T})^j x, x \rangle \geq 1 + \sum_{j=1}^{\infty} c_j = 0$$

as $c_j < 0$ and $0 \leq \langle (\mathcal{I} - \mathcal{T})^j x, x \rangle \leq 1$. Further, since $\mathcal{S}_n$ commutes with any operator that commutes with $\mathcal{T}$ this property also holds for the limit $\mathcal{S}$.

We can verify that $\mathcal{S}$ is positive if and only if $\mathcal{T}$ is positive.

To prove uniqueness, suppose there is another operator $\mathcal{V}$ with these properties. Then,

$$\begin{aligned} & (\mathcal{V}-\mathcal{S})\mathcal{V}(\mathcal{V}-\mathcal{S}) + (\mathcal{V}-\mathcal{S})\mathcal{S}(\mathcal{V}-\mathcal{S}) \\ = \quad & (\mathcal{V}^2-\mathcal{S}^2)(\mathcal{V}-\mathcal{S}) \;=\; 0. \end{aligned}$$

As both operators on the left hand side of the last expression are non-negative definite, they must each be identically zero. Thus,

$$(\mathcal{V}-\mathcal{S})\mathcal{V}(\mathcal{V}-\mathcal{S}) - (\mathcal{V}-\mathcal{S})\mathcal{S}(\mathcal{V}-\mathcal{S}) = (\mathcal{V}-\mathcal{S})^3 = 0.$$

So, $(\mathcal{V}-\mathcal{S})^4 = 0$, which implies that for all $x \in \mathcal{H}$,
$\|(\mathcal{V}-\mathcal{S})^2 x\|^2 = \langle (\mathcal{V}-\mathcal{S})^4 x, x\rangle = 0$. Consequently, $(\mathcal{V}-\mathcal{S})^2 = 0$. Applying a similar argument now yields $\mathcal{V}-\mathcal{S} = 0$ and completes the proof.

Recall the projection theorem: If $\mathcal{M}$ is a closed subspace of a Hilbert space $\mathcal{H}$, then for each $x \in \mathcal{H}$, there exists a unique $y \in \mathcal{H}$ such that $\|x - y\| = \inf\{\|x - v\| : v \in \mathcal{M}\}$.

Let $\mathcal{P}_{\mathcal{M}}$ be the map that sends x to its projection onto $\mathcal{M}$. Call it a projection operator.

Theorem

If $\mathcal{M}$ is a closed subspace of $\mathcal{H}$, then $\mathcal{P}_{\mathcal{M}}$ is a self-adjoint operator in $\mathcal{B}(\mathcal{H})$ and satisfies $\mathcal{P}_{\mathcal{M}} = \mathcal{P}_{\mathcal{M}}^2$.

Proof

We first show that $\mathcal{P}_{\mathcal{M}}$ is linear. For $x_1, x_2 \in \mathcal{H}$, $a_1, a_2 \in \mathbb{R}$ and $y \in \mathcal{M}$, we have $\langle a_1 \mathcal{P}_{\mathcal{M}} x_1 + a_2 \mathcal{P}_{\mathcal{M}} x_2, y\rangle = a_1 \langle \mathcal{P}_{\mathcal{M}} x_1, y\rangle + a_2 \langle \mathcal{P}_{\mathcal{M}} x_2, y\rangle = a_1 \langle x_1, y\rangle + a_2 \langle x_2, y\rangle = \langle a_1 x_1 + a_2 x_2, y\rangle$. Thus, $\mathcal{P}_{\mathcal{M}}(a_1 x_1 + a_2 x_2) = a_1 \mathcal{P}_{\mathcal{M}} x_1 + a_2 \mathcal{P}_{\mathcal{M}} x_2$. We can verify the self-adjointness of $\mathcal{P}_{\mathcal{M}}$. Finally for $x \in \mathcal{H}$, $\mathcal{P}_{\mathcal{M}} x \in \mathcal{M}$. The norm minimization feature of projection now has the consequence that $\mathcal{P}_{\mathcal{M}}^2 = \mathcal{P}_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \mathcal{P}_{\mathcal{M}}$.

- Being idempotent, projection operators are non-negative definite.

- Also, $\|\mathcal{P}_{\mathcal{M}} x\| = \|\mathcal{P}^2_{\mathcal{M}} x\| \leq \left\|\mathcal{P}_{\mathcal{M}}\right\|_\infty \|\mathcal{P}_{\mathcal{M}} x\|$, which implies that $\left\|\mathcal{P}_{\mathcal{M}}\right\|_\infty \geq 1$. Note that $\|\mathcal{P}_{\mathcal{M}} x\| \leq \|x\|$, i.e., projection operators are contractions. Combining the two statements, we obtain $\left\|\mathcal{P}_{\mathcal{M}}\right\|_\infty = 1$.

- If $\mathcal{M}$ has dimension one and is spanned by x with $\|x\| = 1$, then $\mathcal{P}(\mathcal{M})$ can be written as $x \otimes x$, where $(x \otimes x)y = \langle x, y\rangle x$ for any $y \in \mathcal{H}$.
  (this is because of the uniqueness of the projection along with the identity $\langle x, y\rangle = \langle (x \otimes x)y, x\rangle$, which implies that $\langle \langle x, y\rangle x, (x - \langle x, y\rangle x)\rangle = 0$.

- $\otimes$ is called the tensor product operator. It can be defined more generally as follows. Let $x_i \in \mathcal{H}_i$ for $i = 1, 2$. Then $x_1 \otimes_1 x_2 : \mathcal{H}_1 \to \mathcal{H}_2$ is defined as $(x_1 \otimes_1 x_2)y = \langle x_1, y\rangle x_2$ for all $y \in \mathcal{H}_1$. If $\mathcal{H}_1 = \mathcal{H}_2$, we use $\otimes$ in place of $\otimes_1$. Further, $\left\|x_1 \otimes x_2\right\|_\infty = \|x_1\|_1 \|x_2\|_2$.

So far our study of operators is very coarse – we mostly looked at:

- Questions of boundendess/continuity.
- Questions of bijectivity and inversion.

None of these speak of their "operational structure". In finite dimensions, this is reflected by the spectral theorem and the SVD.

- Roughly speaking, these tell us how the operator transforms to the unit ball.
- This, in turn, gives us an "X-ray" of the operator's internal structure.

So far we only touched on boundedness – which tells us that there is no direction along which the unit ball is stretched infinitely.

- In finite dimensions the unit ball is compact, and so its image under a (necessarily continuous) linear operator is also compact.
- This fails in infinite dimensional vector spaces (see Riesz's theorem, up next).
- So we will need to elicit some form of compactness of the image, for a similar study – this motivates the notion of compact operator.

Definition

An linear operator $\mathcal{T} : \mathcal{B}_1 \to \mathcal{B}_2$ is said to be compact if for any bounded sequence $\{x_n\} \subset \mathcal{B}_1$, $\{\mathcal{T}x_n\}$ contains a convergent subsequence in $\mathcal{B}_2$.

In other words, if the unit ball in $\mathcal{B}_1$ is mapped to a pre-compact set in $\mathcal{B}_2$.

- Think of what a linear operator does to the unit ball in finite dimensions.
- Exercise: Compactnes $\implies$ boundedness
- Exercise Bounded linear operators with finite rank are compact.
- "Unfortunately" the compactness can fail if the rank is not finite.

Lemma (Riesz's lemma)

Let $\mathcal{X}$ be a normed linear space, $\mathcal{Y}$ be a closed proper subspace of $\mathcal{X}$, and $\alpha \in (0, 1)$. There exists $x \in \mathcal{X}$ with $\|x\| = 1$ such that $\|x - y\| > \alpha$ for all $y \in \mathcal{Y}$.

Proof.

Since $\mathcal{Y} \subsetneq \mathcal{X}$, there is an $x \in X$ such that $x \notin Y$. Since Y is closed, x is at strictly positive distance from any Y, i.e. $\inf_{y \in Y} \|x - y\| = d > 0$. Since $\alpha \in (0, 1)$, there is some $y_0 \in Y$ at distance $\alpha^{-1} d$ from from x (or else the distance of x from Y would be at least $\alpha^{-1} d$ and so certainly not d as stated), i.e. $\|x - y_0\| = \alpha^{-1} d$.
We claim that

$$x_\alpha := \frac{x - y_0}{\|x - y_0\|} = \frac{x - y_0}{\alpha^{-1} d}$$

verifies the theorem's claim. It is obviously a unit vector, and furthermore,

$$\|x_\alpha - y\| = \left\| \frac{x - y_0}{\|x - y_0\|} - \frac{\|x - y_0\|}{\|x - y_0\|} y \right\| = \frac{\|x - (y_0 + \|x - y_0\| y)\|}{\|x - y_0\|} \geq \frac{d}{\alpha^{-1} d} = \alpha,$$

because $y_0 + \|x - y_0\| y \in Y$. Hence, x_α is at distance at least α from Y. □

Corollary (Riesz's Theorem)

The unit ball in a Banach space $\mathcal{B}$ is compact iff it is finite dimensional. In other words, the identity operator is a compact operator iff $\mathcal{B}$ is finite dimensional.

Proof: Exercise.

Corollary

For infinite dimensional Banach spaces $\mathcal{B}_1$ and $\mathcal{B}_2$, let $\mathcal{T} \in \mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ be a bijective operator. Then $\mathcal{T}$ is not a compact operator.

Proof: Exercise

Theorem

1. The closure of the range of a compact operator is separable.
2. If either of two operators is compact, so is their composition.
3. The set of compact operators in $\mathcal{B}(\mathcal{B}_1, \mathcal{B}_2)$ is closed.

Exercise Prove parts (1) and (3) the above theorem.

Theorem

Let $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Then,

1. $\mathcal{T}$ is compact iff there exists a sequence of bounded linear finite rank operators $\mathcal{T}_n$ such that $\left|\left|\left|\mathcal{T} - \mathcal{T}_n\right|\right|\right|_\infty \to 0$ as $n \to \infty$.
2. $\mathcal{T}$ is compact iff $\mathcal{T}^*$ is compact.

Exercise Prove the 'if' part of part (1) of the above theorem, as well as part (2).

- Symmetric matrices admits a spectral decomposition.
- Since compact operators on Hilbert spaces are "limits of matrices", can we hope for a similar decomposition for self-adjoint such operators?

Given a Hilbert space $\mathcal{H}$ let $\mathcal{T} \in \mathcal{B}(\mathcal{H})$. If $\lambda \in \mathbb{R}$ is such that

$$\mathcal{T}e = \lambda e$$

for some non-zero $e \in \mathcal{H}$ (not necessarily unique), we say that:

- λ is an eigenvalue of $\mathcal{T}$.
- e is an eigenvector of of $\mathcal{T}$, associated with the eigenvalue λ.
- $\mathcal{E}_\lambda(\mathcal{T}) := \{x \in \mathcal{H} : \mathcal{T}x = \lambda x\} \equiv \ker(\mathcal{T} - \lambda\mathcal{I})$ is the eigenspace of λ.

The collection of all eigenvalues $\sigma(\mathcal{T})$ is called the spectrum of $\mathcal{T}$.

- Since the kernel of a bounded operator is always a closed subspace of $\mathcal{H}$, $\ker(\mathcal{T} - \alpha\mathcal{I})$ is itself a Hilbert space for any scalar α.
- By definition, $\ker(\mathcal{T} - \alpha\mathcal{I}) \neq \{0\}$ iff α is an eigenvalue. In other words, an eigenspace is by definition non-trivial.

Two ingredients will work for us:

- Self-adjointness will yield orthogonality of eigenspaces.
- Compactness will yield countability and convergence to zero of eigenvalues, on the one hand, and finite-dimensionality of the eigenspaces, on the other.
- Combined they will yield existence of an eigenvalue.

Theorem

For a bounded and self-adjoint operator, eigenspaces corresponding to distinct eigenvalues are orthogonal.

For operators that are only bounded, one can make the weaker statement that eigenvectors corresponding to distinct eigenvalues are linearly independent.

Proof

Let $\mathcal{T}^* = \mathcal{T}$ be bounded, and $\lambda_1 \neq \lambda_2$ be two distinct eigenvalues of $\mathcal{T}$. Take x_1, x_2 to be vectors in the corresponding eigenspaces. Then,

$$\lambda_1 \langle x_1, x_2 \rangle = \langle \lambda_1 x_1, x_2 \rangle = \langle \mathcal{T} x_1, x_2 \rangle = \langle x_1, \mathcal{T}^* x_2 \rangle = \langle x_1, \mathcal{T} x_2 \rangle = \lambda_2 \langle x_1, x_2 \rangle.$$

If $\langle x_1, x_2 \rangle \neq 0$, we could divide by $\langle x_1, x_2 \rangle$ arriving at the contradiction $\lambda_1 = \lambda_2$. □

Now let's bring in compactness...

Theorem

Let $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ be a compact and self-adjoint operator. Then,

1. The subspaces $\{\mathcal{E}_\lambda(\mathcal{T}) : \lambda \neq 0\}$ are all finite-dimensional (hence closed)
2. For any $\epsilon > 0$, the set $\{\lambda \in \sigma(\mathcal{T}) : |\lambda| \geq \epsilon\}$ is finite.
3. The spectrum of $\mathcal{T}$ is countable.

Self-adjointness is actually superfluous, it just makes the proof a bit shorter.

Proof

(1) Suppose $\dim\{\ker(\mathcal{T} - \lambda\mathcal{I})\} = \infty$. Then, by Riesz's lemma, there exists a sequence $\{e_n\}_{n=1}^{\infty} \subset \ker(\mathcal{T} - \lambda\mathcal{I})$ with $\|e_n\| = 1$ for all n and $\|e_m - e_n\| > 1/2$ for all $n \neq m$. So, $\|\mathcal{T}e_m - \mathcal{T}e_n\| = \lambda\|e_m - e_n\| > \lambda/2$ for all $n \neq m$. If $\lambda \neq 0$, this contradicts the assumption that $\mathcal{T}$ is a compact operator.

(2) Suppose the stated set is infinite (possibly uncountably infinite). Then, it contains an infinite sequence $\{\lambda_j\}_{j=1}^{\infty}$ of distinct eigenvalues of $\mathcal{T}$, which by definition satisfies $|\lambda_j| \geq \epsilon > 0$ for all j. Choose unit vectors $e_j \in \mathcal{E}_{\lambda_j}(\mathcal{T})$, and observe that these constitute an infinite orthonormal sequence, since the λ_j are distinct. This for any $n \neq m$,

$$\|\mathcal{T}e_n - \mathcal{T}e_m\|^2 = \|\lambda_n e_n - \lambda_m e_m\|^2 = \|\lambda_n e_n\|^2 + \|\lambda_m e_m\|^2 =\geq 2\epsilon^2$$

which again contradicts compactness of $\mathcal{T}$.

(3) $\sigma(\lambda) = \cup_{n\geq 1}\{\lambda \in \sigma(\mathcal{T}) : |\lambda| \geq n^{-1}\}$ which is a countable union of finite sets. □

Lemma (Existence of an Eigenvalue)

If $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ is compact and self-adjoint, then it always possesses either $\|\mathcal{T}\|_\infty$ or $-\|\mathcal{T}\|_\infty$ as an eigenvalue.

Consequently, a compact and self-adjoint operator is non-zero if and only if it possesses a non-zero eigenvalue.

Proof.

By self-adjointness, $\left\|\mathcal{T}\right\|_\infty = \sup\{|\langle \mathcal{T}x, x\rangle| : \|x\| = 1\}$. So there is a sequence $x_n \in \mathcal{H}$ such that $\|x_n\| = 1$ and $|\langle \mathcal{T}x_n, x_n\rangle| \to \alpha$, where $|\alpha| = \left\|\mathcal{T}\right\|_\infty$. Thus

$$\|\mathcal{T}x_n - \alpha x_n\|^2 = \|\mathcal{T}x_n\|^2 + \|\alpha x_n\|^2 - 2\langle \mathcal{T}x_n, \alpha x_n\rangle \leq \alpha^2 + \alpha^2 - 2\alpha\langle \mathcal{T}x_n, x_n\rangle \to 0.$$

By compactness of $\mathcal{T}$, we can extract a subsequence $\{x_{n_k}\}$ such that

$$\mathcal{T}x_{n_k} \overset{k\to\infty}{\to} x \in \mathcal{H}.$$

which combined with the fact that $\mathcal{T}x_{n_k} - \alpha x_{n_k} \to 0$ yields $\alpha x_{n_k} \to x$, so that

$$\mathcal{T}\alpha x_{n_k} \equiv \alpha \mathcal{T}x_{n_k} \to \mathcal{T}x.$$

by continuity of $\mathcal{T}$. The two blue limits stipulate that $\mathcal{T}x = \alpha x$. □

Theorem (Spectral Theorem for Compact Self-Adjoint Operators)

Let $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ be a compact self-adjoint operator with (countably many) eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \ldots$, indexed by non-increasing magnitude. Then,

1. The eigenspaces of $\mathcal{T}$ corresponding to non-zero eigenvalues yield a direct sum decomposition of $\overline{\text{range}(\mathcal{T})}$,
2. Writing $\mathcal{P}_k$ for the projection onto the eigenspace of an eigenvalue λ_k,

$$\mathcal{T}(x) = \sum_{n \geq 1} \lambda_n \mathcal{P}_n(x), \qquad \forall\, x \in \mathcal{H}.$$

Corollary (Spectral Decomposition)

There exists an ONB $\{e_n\}$ of of $\overline{\text{range}(\mathcal{T})}$ comprised of $\mathcal{T}$-eigenvectors, and

$$\mathcal{T}(x) = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle e_n, \qquad \forall\, x \in \mathcal{H}.$$

If each eigenspace is one-dimensional, the basis is unique.

Notice that the second statement is not merely a corollary – the two statements are actually equivalent (exercise).

Proof (of the corollary, which is equivalent to the theorem).

If we show that $\overline{\text{range}(\mathcal{T})}$ admits a countable orthonormal basis of $\mathcal{T}$-eigenvectors $\{e_j\}$, then it will follows that $\text{range}(\mathcal{T}) \ni \mathcal{T}x = \sum_{j\geq 1} a_j e_j$ for all $x \in \mathcal{H}$ with $a_j = \langle \mathcal{T}x, e_j \rangle = \langle x, \mathcal{T}e_j \rangle = \lambda_j \langle x, e_j \rangle$ for all $j \geq 1$.

To construct such a basis, consider the set $\{u_{i,j}\}$, where for fixed j, $u_{1,j}, u_{2,j}, \ldots$ is an orthonormal basis for the jth eigenspace. There are countably many eigenspaces, they are finite dimensional, and they are mutually orthogonal. So we can order $\{u_{i,j}\}$ into a countable orthonormal set, say $\{e_j\}$.

It remains to show completeness of $\{e_j\}$ in $\overline{\text{range}(\mathcal{T})}$, i.e.

$$\overline{\text{range}(\mathcal{T})} = \overline{\text{span}}\{e_j : j \geq 1\}.$$

For any finite n and $c_1, c_2, \ldots, c_n \in \mathbb{R}$ (not all of which are zero), we have $\sum_{j=1}^{n} c_j e_j \in \overline{\text{range}(\mathcal{T})}$ because each e_j is an eigenvector:

$$e_j = \lambda_j^{-1} \mathcal{T} e_j \in \text{range}(\mathcal{T}).$$

Thus, $\text{span}\{e_j : 1 \leq j \leq n\} \subseteq \overline{\text{range}(\mathcal{T})}$ for all finite n, which in turn implies that $\text{span}\{e_j : j \geq 1\} \subseteq \overline{\text{range}(\mathcal{T})}$. Consequently, $\overline{\text{span}}\{e_j : j \geq 1\} \subseteq \overline{\text{range}(\mathcal{T})}$.

From the projection theorem, we can now write

$$\overline{\text{range}(\mathcal{T})} = \overline{\text{span}}\{e_j : j \geq 1\} \oplus \mathcal{N},$$

where $\mathcal{N}$ containsall elements of $\overline{\text{range}(\mathcal{T})}$ orthogonal to $\overline{\text{span}}\{e_j : j \geq 1\}$. To complete the proof, we must show $\mathcal{N} = \{0\}$.

To this aim, let $\mathcal{T}_{\mathcal{N}}$ be the restriction of $\mathcal{T}$ to $\mathcal{N}$. This restriction is:

1. self-adjoint on $\mathcal{N}$ because $\mathcal{T}$ maps $\mathcal{N}$ to $\mathcal{N}$ (for $x \in \mathcal{N}$ and $y \in \overline{\text{span}}\{e_j : j \geq 1\}$, we have $\langle \mathcal{T}x, y\rangle = \langle x, \mathcal{T}y\rangle = 0$ and $\mathcal{T}x \in \overline{\text{span}}\{e_j : j \geq 1\}^{\perp} = \mathcal{N}$).
2. compact, which is inherited directly from $\mathcal{T}$.

Hence $\mathcal{T}_{\mathcal{N}} \neq 0$ if and only if it posesses a non-zero eigenvalue. But any non-zero eigenvalue of $\mathcal{T}_{\mathcal{N}}$ is also an eigenvalue for the original operator $\mathcal{T}$. And all such non-zero eigenvalues were already captured in the collection $\{\lambda_j\}_{j=1}^{\infty}$, which leaves $\mathcal{T}_{\mathcal{N}}$ being the zero operator as the only option.

It follows that, $\mathcal{N} \subseteq \ker(\mathcal{T})$, and since we also have $\mathcal{N} \subseteq \overline{\text{range}(\mathcal{T})}$, it can only be that $\mathcal{N} = \{0\}$. □

The spectral decomposition paves the way for many important results...

- If $\mathcal{T} \succeq 0$ is compact with spectrum $\{(\lambda_j, e_j)\}_{j=1}^{\infty}$, then

$$\lambda_k = \max_{e \in \mathrm{span}\{e_1, e_2, \ldots, e_{k-1}\}^{\perp}} \frac{\langle \mathcal{T}e, e \rangle}{\|e\|^2}$$

  for all k with $\mathrm{span}\{e_1, e_2, \ldots, e_{k-1}\}$ being the entire $\mathcal{H}$ if $k = 1$.

- So, if we arrange the λ_j's so that $|\lambda_1| \geq |\lambda_2| \geq \ldots$, then

$$|\lambda_k| = \max_{e \in \mathrm{span}\{e_1, e_2, \ldots, e_{k-1}\}^{\perp}} \frac{\|\mathcal{T}e\|}{\|e\|}.$$

- Thus, for a compact, self-adjoint operator $\mathcal{T}$, we have $\|\mathcal{T}\|_{\infty} = |\lambda_1|$.
- If $\mathcal{T}$ is a compact, self-adjoint operator then so is $\mathcal{T}^n$ for all $n \geq 1$. Further, $\mathcal{T}^n = \sum_{j \geq 1} \lambda_j^n e_j \otimes e_j$, i.e., $\{(\lambda_j^n, e_j)\}_{j=1}^{\infty}$ are the eigenpairs of $\mathcal{T}^n$.
- A compact and self-adjoint operator is non-negative iff all of its eigenvalues are non-negative.
- One can define fractional powers of compact and non-negative operators. If $\mathcal{S} := \sum_{j \geq 1} \lambda_j^{1/2} e_j \otimes e_j$, then it is easily verified that $\mathcal{S} = \mathcal{T}^{1/2}$, the square root operator of $\mathcal{T}$. Obviously, $\mathcal{T}^{1/2}$ is compact.

- It is possible to extend the spectral decomposition result to non-self-adjoint compact operators in $\mathcal{B}(\mathcal{H})$ and to compact operators in $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$.
- Let $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ be a compact operator. So, $\mathcal{T}^*\mathcal{T}$ and $\mathcal{T}\mathcal{T}^*$ are compact, non-negative operators in $\mathcal{B}(\mathcal{H}_1)$ and $\mathcal{B}(\mathcal{H}_2)$, respectively.
- They have the same non-zero eigenvalues, say, $\eta_1 \geq \eta_2 \ldots \geq 0$. The eigenfunction of $\mathcal{T}^*\mathcal{T}$ (respectively, $\mathcal{T}\mathcal{T}^*$) associated with η_j is denoted by f_{1j} (respectively, f_{2j}). Further, $f_{2j} = \mathcal{T}f_{1j}/\lambda_j$ and $f_{1j} = \mathcal{T}^*f_{2j}/\lambda_j$ with $\lambda_j = \sqrt{\eta_j}$ for all j. We always take λ_j's to be non-negative.
- The triple $(\lambda_j, f_{1j}, f_{2j})$, $j = 1, 2, \ldots$, is called a singular system of $\mathcal{T}$ with the λ_j's being called the singular values of $\mathcal{T}$, while the f_{1j}'s and the f_{2j}'s are called the right and the left singular functions, respectively.

Theorem (Singular Value Decomposition)

Let $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ be a compact operator. Then, $\mathcal{T} = \sum_{j \geq 1} \lambda_j f_{1j} \otimes_1 f_{2j}$, i.e., $\mathcal{T}x = \sum_{j \geq 1} \lambda_j \langle x, f_{1j} \rangle f_{2j}$ for any $x \in \mathcal{H}_1$ with

(a) $\{\eta_j\} = \{\lambda_j^2\}$ being the non-increasing eigenvalues of $\mathcal{T}^*\mathcal{T}$ and $\mathcal{T}\mathcal{T}^*$,

(b) $\{f_{1j}\}$ being the orthonormal eigenfunctions of $\mathcal{T}^*\mathcal{T} \Rightarrow \{f_{1j}\}_{j=1}^{\infty}$ is an o.n.b. of $\overline{\text{range}(\mathcal{T}^*\mathcal{T})} = \overline{\text{range}(\mathcal{T}^*)} = \ker(\mathcal{T})^{\perp}$, and

(c) $\{f_{2j}\}$ being the orthonormal eigenfunctions of $\mathcal{T}\mathcal{T}^*$ satisfying $\mathcal{T}^* f_{2j} = \lambda_j f_{1j}$ for all j. Also, $\{f_{2j}\}_{j=1}^{\infty}$ is an o.n.b. of $\overline{\text{range}(\mathcal{T}\mathcal{T}^*)} = \overline{\text{range}(\mathcal{T})}$.

- Let $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ be a compact operator. Then, it follows from the previous theorem that $\|\mathcal{T}\|_\infty = \lambda_1$.
- An operator $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is compact iff the SVD holds.
- The singular values converge to zero necessarily $\lambda_n \to 0$.
- A pseudo-inverse can be obtained as follows. For each $i = 1, 2$, let $\{e_{ik}\}_{k=1}^{\infty}$ be an o.n.b. of $\mathcal{H}_i$ (generated from the singular system). Now, if $x = \sum_{k=1}^{\infty} \langle x, e_{1k} \rangle_1 \, e_{1k}$, we have $\mathcal{T}x = \sum_{j=1}^{\infty} \lambda_j \langle x, f_{1j} \rangle f_{2j}$. A condition that characterizes $y = \sum_{k=1}^{\infty} \langle y, e_{2k} \rangle_2 \, e_{2k} \in \text{range}(\mathcal{T})$ is

$$\sum_{j=1}^{\infty} \langle y, f_{2j} \rangle_2^2 / \lambda_j^2 < \infty.$$

This is called Picard's condition and, when it holds, we may set

$$\mathcal{T}^\dagger y = \sum_{j=1}^{\infty} \lambda_j^{-1} \langle y, f_{2j} \rangle_2 f_{1j}.$$

Theorem

The range of an infinite rank compact operator is not closed.

Proof

Let $\mathcal{T}$ have the singular system $\{(\lambda_j, f_{1j}, f_{2j})\}_{j=1}^{\infty}$. Since $\lambda_j \downarrow 0$ as $j \to \infty$, we can choose a subsequence $\{j_k\}$ such that $\lambda_{j_k} < k^{-1}$ for all k. Define $y = \sum_{k=1}^{\infty} \lambda_{j_k} f_{2j_k}$. Clearly, $y \in \overline{\text{range}(\mathcal{T})}$ as $f_{2j_k} = \mathcal{T}(f_{1j_k}/\lambda_{j_k})$. But Picard's condition fails to hold for y. Thus, $y \notin \text{range}(\mathcal{T})$.

- For compact operators, the existence of an approximate solution is guaranteed when Picard's condition holds.
- But the unboundedness of the pseudoinverse makes the solution very unstable.
- The degree of instability depends on how fast the singular values of $\mathcal{T}$ decays to zero: "mildly ill-posed" if $\lambda_j \sim j^{-\alpha}$ for some $\alpha > 1$, and "severely ill-posed" if $\lambda_j \sim exp(-\beta j)$ for some $\beta > 0$.
- To obtain stability (further to existence) in ill-posed linear equations, we must use "regularization" leading to "regularized inverses".

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be separable Hilbert spaces and $\{e_{ij}\}$ be an a countable ONB for $\mathcal{H}_i$. If $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ satisfies

$$\sum_{j=1}^{\infty} \|\mathcal{T} e_{1j}\|_2^2 \equiv \sum_{j=1}^{\infty} \langle \mathcal{T} e_{1j}, e_{2j} \rangle_2^2 < \infty,$$

then $\mathcal{T}$ is called a Hilbert-Schmidt operator.

The collection of Hilbert-Schmidt operators $\mathcal{H}_1 \to \mathcal{H}_2$ is denoted by $\mathcal{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$.

- The infinite sum in the above definition is independent of the choice of the o.n.b., i.e., if the sum converges for some o.n.b., it does for all o.n.b.'s and they all have the same value (exercise).
- In the same vein, if $\mathcal{T}$ if Hilbert-Schmidt, then so is $\mathcal{T}^*$.
- Finally, for some (and thus any) o.n.b.'s $\{e_{1j}\}$ and $\{e_{2k}\}$ of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, we have $\sum_{k=1}^{\infty} \|\mathcal{T}^* e_{2k}\|_2^2 = \sum_{j=1}^{\infty} \|\mathcal{T} e_{1j}\|_2^2$.

- If $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is compact, then it follows from the spectral decompostion that if we choose $\{e_{1j}\}$ as the union of the right singular functions and an o.n.b. of $\ker(\mathcal{T})$, then $\sum_{j=1}^{\infty} \|\mathcal{T} e_{1j}\|_2^2 = \sum_{j=1}^{\infty} \lambda_j^2$. Thus, $\mathcal{T}$ is Hilbert-Schmidt iff its singular values are square summable.
- Clearly, $\mathcal{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$ is a linear space. One can define an inner product on it as follows.

$$\langle \mathcal{T}_1, \mathcal{T}_2 \rangle = \sum_{j=1}^{\infty} \langle \mathcal{T}_1 e_{1j}, \mathcal{T}_2 e_{1j} \rangle_2,$$

where $\{e_{1j}\}$ is an o.n.b. of $\mathcal{H}_1$. The corresponding norm is given by $|||\mathcal{T}||| = \{\sum_{j=1}^{\infty} \|\mathcal{T} e_{1j}\|^2\}^{1/2} = \{\sum_{j=1}^{\infty} \langle \mathcal{T}^* \mathcal{T} e_{1j}, e_{1j} \rangle\}^{1/2}$.
(exercise: check that the inner product is well-defined)

Theorem

The linear space $\mathcal{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$ equipped with the above inner product is a separable Hilbert space. For any choice of o.n.b.'s $\{e_{1j}\}$ and $\{e_{2k}\}$ of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, $\{e_{1j} \otimes_1 e_{2k}\}$ is an o.n.b. of $\mathcal{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$.

Theorem

Any Hilbert-Schmidt operator is compact.

Proof

Let $\mathcal{T}$ be a Hilbert-Schmidt operator, and for each ≥ 1, define $\mathcal{T}_n$ by $\mathcal{T}_n x = \sum_{j=1}^{n} \langle \mathcal{T}x, e_{2j} \rangle_2 \, e_{2j}$, $x \in \mathcal{H}_1$, where $\{e_{2j}\}$ is an o.n.b. of $\mathcal{H}_2$. Clearly, $\mathcal{T}_n$ is a finite rank operator for each n. Thus, it is enough to show that $\left\| \mathcal{T} - \mathcal{T}_n \right\|_\infty \to 0$ as $n \to \infty$. Let $\|x\|_1 \leq 1$. Then, the Cauchy-Schwarz inequality and the fact that $(\mathcal{T} - \mathcal{T}_n)x = \sum_{j>n} \langle \mathcal{T}x, e_{2j} \rangle_2 \, e_{2j}$ implies $\|(\mathcal{T} - \mathcal{T}_n)x\|_2^2 = \sum_{j>n} \langle \mathcal{T}x, e_{2j} \rangle_2^2 = \sum_{j>n} \langle x, \mathcal{T}^* e_{2j} \rangle_2^2 \leq \sum_{j>n} \|\mathcal{T}^* e_{2j}\|_2^2$. The proof is now complete on observing that since $\mathcal{T}^*$ is a Hilbert-Schmidt operator, it follows that $\sum_{j>n} \|\mathcal{T}^* e_{2j}\|_2^2 \to 0$ as $n \to \infty$.

Exercise: Let $(\Omega, \mathcal{F}, \mu)$ be measure space and $\mathcal{T} : \Omega \to \mathcal{B}_{HS}(\mathcal{H}_1, \mathcal{H}_2)$ be measurable with $\int \left\| \mathcal{T} \right\| d\mu < \infty$. Show that $\int \mathcal{T} f d\mu = (\int \mathcal{T} d\mu) f$ for any $f \in \mathcal{H}_1$.

Theorem (Schmidt-Mirsky-Eckart-Young)

Let $\mathcal{T}$ be a Hilbert-Schmidt operator with singular system $\{(\lambda_j, f_{1j}, f_{2j})\}_{j=1}^{\infty}$. Then, for any finite k, we have

$$\left\| \mathcal{T} - \sum_{j=1}^{k} x_j \otimes_1 y_j \right\| \geq \left\| \mathcal{T} - \sum_{j=1}^{k} \lambda_j f_{1j} \otimes_1 f_{2j} \right\|$$

for any set of functions $x_j \in \mathcal{H}_1$, $y_j \in \mathcal{H}_2$, $j = 1, 2, \ldots, k$.

- The above theorem states that the spectral decomposition of a Hilbert-Schmidt (and thus compact) operator provides the best finite dimensional approximation in a Hilbert-Schmidt sense.
- This is core in FDA, as well as its other versions (and its refinements under additional structure).

Proof

It is enough to show that $\left\|\left|\mathcal{T}-\sum_{j=1}^{k} x_j \otimes_1 y_j\right|\right\| \geq \left\|\left|\mathcal{T}\right|\right\|^2 - \sum_{j=1}^{k}\lambda_j^2 = \sum_{j=k+1}^{\infty}\lambda_j^2$. W.l.o.g., we can assume that the (y_j, x_j) are orthonormal. In that case, if $\{e_j\}$ is any o.n.b. of $\mathcal{H}_1$, then

$$\begin{aligned}
&\left\|\left|\mathcal{T}-\sum_{j=1}^{k} x_j \otimes_1 y_j\right|\right\|^2 \\
= &\sum_{l=1}^{\infty}\left\langle\left(\mathcal{T}^*\mathcal{T}+\sum_{j=1}^{k}(x_j-\mathcal{T}^*y_j)\otimes_1(x_j-\mathcal{T}^*y_j)^*\right)e_j, e_j\right\rangle_1 \\
&-\sum_{j=1}^{k}\|\mathcal{T}^*y_j\|_1^2.
\end{aligned}$$

As $(x_j-\mathcal{T}^*y_j)\otimes_1(x_j-\mathcal{T}^*y_j)^*$ is non-negative, the result will follow once we establish that $\sum_{j=1}^{k}\|\mathcal{T}^*y_j\|_1^2 \leq \sum_{j=1}^{k}\lambda_j^2$.

Now, the spectral decomposition of $\mathcal{T}^*$ gives $\mathcal{T}^* y_j = \sum_{l=1}^{\infty} \lambda_l \langle y_j, f_{2l} \rangle_2 f_{1l}$ so

$$\begin{aligned} ||\mathcal{T}^* y_j||_1^2 &= \lambda_k^2 + \left(\sum_{l=1}^{k} \lambda_l^2 \langle y_j, f_{2l} \rangle_2^2 - \lambda_k^2 \sum_{l=1}^{k} \langle y_j, f_{2l} \rangle_2^2 \right) \\ &\quad - \left(\lambda_k^2 \sum_{l=k+1}^{\infty} \langle y_j, f_{2l} \rangle_2^2 - \sum_{l=k+1}^{\infty} \lambda_l^2 \langle y_j, f_{2l} \rangle_2^2 \right) \\ &\quad - \lambda_k^2 \left(1 - \sum_{l=1}^{\infty} \langle y_j, f_{2l} \rangle_2^2 \right). \end{aligned}$$

The last two terms on the RHS are non-positive meaning that

$$\begin{aligned} \sum_{j=1}^{k} ||\mathcal{T}^* y_j||_1^2 &\leq k\lambda_k^2 + \sum_{j=1}^{k} \sum_{l=1}^{k} (\lambda_l^2 - \lambda_k^2) \langle y_j, f_{2l} \rangle_2^2 \\ &= \sum_{l=1}^{k} [\lambda_k^2 + (\lambda_l^2 - \lambda_k^2) \sum_{j=1}^{k} \langle y_j, f_{2l} \rangle_2^2] \leq \sum_{l=1}^{k} \lambda_l^2. \end{aligned}$$

The last inequality follows from the orthonormality of the y_j's and Parseval. □

- Let $\mathcal{T} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. We can then define the square root of the non-negative operator $\mathcal{T}^*\mathcal{T}$, denoted by $(\mathcal{T}^*\mathcal{T})^{1/2}$, on $\mathcal{B}(\mathcal{H}_1)$. Set $|\mathcal{T}| = (\mathcal{T}^*\mathcal{T})^{1/2}$.

Definition

An operator $\mathcal{T}$ is called trace class if for some o.n.b. $\{e_j\}$ of $\mathcal{H}_1$, the quantity $\|\mathcal{T}\|_{tr} := \sum_{j=1}^{\infty} \langle (|\mathcal{T}|\, e_j, e_j \rangle_1$ is finite. In this case, $\|\mathcal{T}\|_{tr}$ is called the trace norm of $\mathcal{T}$.

- Similar argument as in the Hilbert-Schmidt case shows that the infinite sum does not depend on the choice of the o.n.b.
- For any trace class operator $\mathcal{T}$, we have $\|\mathcal{T}\|_{tr} = \||\mathcal{T}|^{1/2}\|^2$.
- Exercise: Show that if $\mathcal{T}$ is trace class, then $\mathcal{T}^*\mathcal{T}$ is compact. Thus, if $\{(\eta_j, f_j)\}_{j=1}^{\infty}$ denote the eigenvalue-eigenfunction pairs of $\mathcal{T}^*\mathcal{T}$, then $\|\mathcal{T}\|_{tr} = \sum_{j=1}^{\infty} \sqrt{\eta_j}$.
- All this simplifies in the case of non-negative operators (symmetrisation/positivation unnecessary), in fact the more natural definition goes through the non-negative case first.
- Hilbert-Schmidt inner product: We can now re-express the HS inner product as

$$\langle \mathcal{T}_1, \mathcal{T}_2 \rangle = \mathrm{tr}\{\mathcal{T}_1^* \mathcal{T}_2\}.$$

(verify this as an exercise)

- $\sum_{j=1}^{\infty} \|\mathcal{T} f_j\|_2^2 = \sum_{j=1}^{\infty} \langle \mathcal{T}^* \mathcal{T} f_j, f_j \rangle_1 = \sum_{j=1}^{\infty} \eta_j \leq \{\sum_{j=1}^{\infty} \sqrt{\eta_j}\}^2 = \|\mathcal{T}\|_{tr}^2$. Consequently, if $\mathcal{T}$ is trace class, then $\mathcal{T}$ is Hilbert-Schmidt. This implies that a trace class operator is compact. Further, $\|\mathcal{T}\| \leq \|\mathcal{T}\|_{tr}$. Also, $\mathcal{T}$ is trace class iff its singular values are summable.
- If $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ is a trace class operator, then its trace is defined as $\text{tr}(\mathcal{T}) = \sum_{j=1}^{\infty} \langle \mathcal{T} e_j, e_j \rangle$, where $\{e_j\}$ is an o.n.b. of $\mathcal{H}$. It can be shown that this infinite series is absolutely convergent and does not depend on the choice of the o.n.b.
- Thus, if $\mathcal{T}$ is self-adjoint, then $\text{tr}(\mathcal{T}) = \sum_{j=1}^{\infty} \lambda_j$, where $\{\lambda_j\}$ is the sequence of eigenvalues of $\mathcal{T}$. If $\mathcal{T}$ is non-negative, then $tr(\mathcal{T}) = \|\mathcal{T}^{1/2}\|^2$.
- Moreover, if $\mathcal{T}_1$ and $\mathcal{T}_2$ are two trace class operators and $a_1, a_2 \in \mathbb{R}$, then $\text{tr}(a_1 \mathcal{T}_1 + a_2 \mathcal{T}_2) = a_1 \text{tr}(\mathcal{T}_1) + a_2 \text{tr}(\mathcal{T}_2)$ and $\text{tr}(\mathcal{T}_2 \mathcal{T}_1) = \text{tr}(\mathcal{T}_1 \mathcal{T}_2)$.
- Exercise: If $\mathcal{T}$ is a non-negative operator such that $\sum_{j=1}^{\infty} \langle \mathcal{T} e_j, e_j \rangle = 0$ for an o.n.b. $\{e_j\}$ of $\mathcal{H}$, then $\mathcal{T} = 0$.

- The following inclusions hold:

  finite rank $\Rightarrow$ trace class $\Rightarrow$ Hilbert-Schmidt $\Rightarrow$ compact $\Rightarrow$ bounded

  and the inclusions are strict if the operators are defined on an infinite dimensional Hilbert space.

- Rank 1 operators. Let $u, v \in \mathcal{H}$ for some Hilbert space $\mathcal{H}$. Define the linear operator $u \otimes_1 v : \mathcal{H} \to \mathcal{H}$ by $(u \otimes_1 v)(\cdot) = \langle u, \cdot \rangle_{\mathcal{H}} v$. Then $u \otimes_1 v$ is trace class, and

$$\left\| u \otimes_1 v \right\|_{\infty} = \left\| u \otimes_1 v \right\| = \left\| u \otimes_1 v \right\|_{tr} = \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}.$$

- For $\mathcal{T} \in \mathcal{B}(\mathcal{H})$ we have:

$$\left\| \mathcal{T} \right\|_{\infty} \leq \left\| \mathcal{T} \right\| \leq \left\| \mathcal{T} \right\|_{tr}.$$

- Let $\mathcal{U}, \mathcal{V} \in \mathcal{B}(\mathcal{H})$ and $\mathcal{T}$ be a trace class operator on $\mathcal{B}(\mathcal{H})$. Then,

$$\left\| \mathcal{U}\mathcal{T}\mathcal{V} \right\|_{tr} \leq \left\| \mathcal{U} \right\|_{\infty} \left\| \mathcal{T} \right\|_{tr} \left\| \mathcal{V} \right\|_{\infty} \quad \& \quad \left\| \mathcal{U}\mathcal{T}\mathcal{V} \right\| \leq \left\| \mathcal{U} \right\|_{\infty} \left\| \mathcal{T} \right\| \left\| \mathcal{V} \right\|_{\infty}.$$

- The above are consequences of Hölder's inequality for Schatten p-norms, of which all the above are special cases ($p = 1$, $p = 2$, $p = \infty$), and defined via

$$\left(\mathrm{tr} \left\{ |\mathcal{T}|^p \right\} \right)^{1/p}$$

  If $\mathcal{T}$ is compact, then

- For a measure space $(E, \mathcal{F}, \mu)$, recall the definition of an integral operator $\mathcal{K}$ associated with a measurable square integrable kernel $K(\cdot, \cdot)$ on $E \times E$.
- We have seen $\mathcal{K} \in \mathcal{B}(L_2(E, \mathcal{F}, \mu))$ and $\|\|\mathcal{K}\|\|_\infty \leq \|K\|_{L^2(E \times E)}$.
- We will be interested in the following context:
  1. E is a compact metric space
  2. $\mathcal{F}$ is the Borel σ-field on E
  3. the support of μ is the whole of E.
  4. K is continuous on $E \times E$ (hence also uniformly continuous).
- Exercise: For every $f \in L_2(E, \mathcal{F}, \mu)$, $(\mathcal{K}f)(\cdot)$ is uniformly continuous on E.

Theorem

Under the above assumptions, $\mathcal{K}$ is compact.

The proof relies on establishing a finite rank approximation of $\mathcal{K}$ by considering a uniform finite dimensional approximation of K by polynomials (Stone-Weierstrass theorem).

- If K is symmetric, then $\mathcal{K}$ is self-adjoint. So, the spectral decomposition yields $\mathcal{K} = \sum_{j=1}^{\infty} \lambda_j \, e_j \otimes e_j$, where $\{(\lambda_j, e_j)\}$ is the eigenvalue-eigenfunction sequence of $\mathcal{K}$.
- The uniform continuity of $(\mathcal{K}f)(\cdot)$ implies that the version of e_j given by $e_j(\cdot) = \lambda_j^{-1} \int_E K(s, \cdot) e_j(s) \mu(ds)$ is uniformly continuous in t. We will always assume this for the e_j's.

Example

Let $E = [0, 1]$ and $K(s, t) = \min(s, t)$. This is the covariance kernel of the Brownian motion on $[0, 1]$. The operator $\mathcal{K}$ is given by

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

The eigenvalues and the eigenfunction of $\mathcal{K}$ are found by solving the following equation in λ and e.

$$\int_0^t s e(s) ds + t \int_t^1 e(s) ds = \lambda e(t) \quad \text{for all } t.$$

Differentiating both sides with respect to t yields

$$te(t) + \int_t^1 e(s)ds - te(t) = \lambda e'(t) \Leftrightarrow \int_t^1 e(s)ds = \lambda e'(t) \quad \text{for all } t.$$

Differentiating the above equation once again yields $e(t) = -\lambda e''(t)$ for all t. Note that if $\lambda = 0$, then $e \equiv 0$ so that the $\ker(\mathcal{K}) = \{0\}$. For $\lambda \neq 0$, a general solution for this differential equation is

$$e(t) = a\sin(t/\sqrt{\lambda}) + b\cos(t/\sqrt{\lambda})$$

with boundary constraints $e(0) = 0$ and $e'(1) = 0$. Thus, $b = 0$ and $a\cos(1/\sqrt{\lambda}) = 0$, which leads to $1/\sqrt{\lambda} = (2j-1)\pi/2$, $j = 1, 2, \ldots$. Thus, the eigenvalues and the associated eigenfunctions of $\mathcal{K}$ are

$$\lambda_j = \frac{1}{\{(j-0.5)\pi\}^2} \quad \text{and} \quad e_j(t) = \sqrt{2}\sin\{(j-0.5)\pi t\}, \quad j \geq 1.$$

Theorem

An integral operator is non-negative (positive) definite iff its integral kernel is non-negative (positive) definite.

A central theorem now is:

Theorem (Mercer)

Let $\mathcal{K} \in \mathcal{B}(L_2(E, \mathcal{F}, \mu))$, where

1. E is a compact metric space
2. $\mathcal{F}$ is the Borel σ-field on E
3. the support of μ is the whole of E.
4. K is continuous on $E \times E$ (hence also uniformly continuous).

If the operator's integral kernel is non-negative definite, $K \succeq 0$, then the spectral decomposition of $\mathcal{K}$ admits a pointwise version,

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j \, e_j(s) e_j(t) \quad \text{for all } s, t$$

where $\{(\lambda_j, e_j)\}$ are the eigenpairs of $\mathcal{K}$, and convergence is absolute and uniform.

Mercer's theorem gives an equivalent representation of the kernel in terms of the eigenvalues and the eigenfunctions of $\mathcal{K}$. There are several important consequences of Mercer's theorem.

Theorem

Under the conditions of Mercer's theorem, $\|\mathcal{K}\|_{tr} = \text{tr}(\mathcal{K}) = \int_E K(t,t)\mu(dt)$ and $\|\mathcal{K}\| = \|K\|_{L_2(E\times E)}$.

Theorem

Let K be a continuous, symmetric and non-negative definite kernel with the eigen-decomposition $K(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t)$ in the sense of Mercer's theorem. Then, for any positive integer r for which $\lambda_r > 0$, we have

$$\min_{\text{rank}(W)=r} \int_{E\times E} \{K(s,t) - W(s,t)\}^2 \mu(ds)\mu(dt) = \sum_{j>r} \lambda_j^2$$

where the minimum is achieved by $W(s,t) = \sum_{j=1}^{r} \lambda_j e_j(s) e_j(t)$.

The proof of Mercer's theorem follows from the following string of result.

Lemma

Under the conditions of Mercer's theorem, we have
(1) $\sum_{j=1}^{\infty} \lambda_j \, e_j^2(t) \leq K(t,t)$ for all t,
(2) $\sum_{j=1}^{\infty} |\lambda_j \, e_j(s) e_j(t)| \leq \{K(t,t)K(s,s)\}^{1/2}$ for all s, t,
(3) $\lim_{n\to\infty} \sup_{s,t} \sum_{j=n+1}^{\infty} |\lambda_j \, e_j(s) e_j(t)| = 0$, and
(4) the function $\sum_{j=1}^{\infty} \lambda_j \, e_j(s) e_j(t)$ is well-defined and uniformly continuous in (s,t) with the sum converging absolutely and uniformly.

Let's reconsider the RKHS associated with a continuous, symmetric and non-negative definite kernel K. It follows from Mercer's theorem that $K(s,t) = \sum_{j=1}^{\infty} \lambda_j \, e_j(s) e_j(t)$ for all s, t. With some work, we can show

$$f \in \mathcal{H}(K) \Leftrightarrow f(t) = \sum_{j=1}^{\infty} \lambda_j \, a_j \, e_j(t), \qquad \text{with} \sum_{j=1}^{\infty} \lambda_j \, a_j^2 < \infty.$$

Equivalently, $f(t) = \sum_{j=1}^{\infty} c_j \, e_j(t)$, where the c_j's satisfy $\sum_{j=1}^{\infty} c_j^2 / \lambda_j < \infty$. This implies that $\sum_{j=1}^{\infty} c_j^2 < \infty$ so that $f \in L_2(E)$. Since the e_j's are orthogonal in the $L_2(E)$ sense, it follows that $c_j = \langle f, e_j \rangle_{L_2(E)}$.

Also, $\|f\|_{\mathcal{H}(K)} = \{\sum_{j=1}^{\infty} c_j^2 / \lambda_j\}^{1/2}$. Hence, for $f \in L^2(E)$, we have

$$f \in \mathcal{H}(K) \Leftrightarrow \sum_{j=1}^{\infty} \langle f, e_j \rangle^2 / \lambda_j < \infty.$$

Since $\mathcal{K}$ is compact with $\{(\lambda_j^{1/2}, e_j)\}_{j=1}^{\infty}$ as its eigenvalue-eigenfunction sequence, it follows that $\mathcal{K}^{1/2}$ is compact with $\{(\lambda_j^{1/2}, e_j)\}_{j=1}^{\infty}$ as its eigenvalue-eigenfunction sequence.

So, it follows from Picard's condition that for $f \in L^2(R)$

$$f \in \mathcal{H}(K) \Leftrightarrow f \in \text{range}(\mathcal{K}^{1/2}) \quad \text{i.e.,} \quad \mathcal{H}(K) = \text{range}(\mathcal{K}^{1/2}).$$

Since $\mathcal{K}^{1/2}$ is compact, $\mathcal{H}(K)$ is not closed in $L_2(E)$.

Furthermore, $\|e_j\|_{\mathcal{H}(K)} = \lambda_j^{-1/2}\|e_j\|_{L_2(E)}$ so that $\|\widetilde{e}_j\|_{\mathcal{H}(K)} = 1$, where $\widetilde{e}_j = \sqrt{\lambda_j}\, e_j$. Also, $\langle \widetilde{e}_i, \widetilde{e}_j \rangle_{\mathcal{H}(K)} = \delta_{ij}$.

Since $\{e_j\}_{j=1}^{\infty}$ forms an o.n.b. of $\overline{\text{range}(\mathcal{K}^{1/2})}$ (equipped with the $L_2(E)$ inner product), it follows that $\{\widetilde{e}_j\}_{j=1}^{\infty}$ is an o.n.b. of $\mathcal{H}(K) = \text{range}(\mathcal{K}^{1/2})$ (equipped with the $\mathcal{H}(K)$ inner product).

Moreover, $K(s,t) = \sum_{j=1}^{\infty} \lambda_j\, e_j(s) e_j(t) = \sum_{j=1}^{\infty} \widetilde{e}_j(s) \widetilde{e}_j(t)$.

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

Henceforth, let $\mathcal{H}$ be a separable Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. What is a natural σ-field of events?

- The usual Borel σ-field $\mathcal{B}(\mathcal{H})$ is given by the smallest σ-field containing all the open subsets of $\mathcal{H}$ and is denoted by.
- But events are more naturally formulated through functionals – we could also define another σ-field on $\mathcal{H}$, namely the smallest one which makes all bounded linear functionals measurable, which by Riesz representation is

$$\mathcal{C} = \sigma\{\langle \cdot, f \rangle^{-1}(B) : B \in \mathcal{B}(\mathbb{R}), f \in \mathcal{H}\}.$$

Theorem

The σ-fields $\mathcal{C}$ and $\mathcal{B}(\mathcal{H})$ are the same.

Consequently, given a measure space $(\Omega, \mathcal{F}, P)$ we have:

Theorem

$X : (\Omega, \mathcal{F}, P) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is measurable iff $\langle X, f \rangle$ is measurable for all $f \in \mathcal{H}$. Furthermore, if X is measurable, its distribution is uniquely determined by the collection of marginal distributions of $\langle X, f \rangle$ as f ranges in $\mathcal{H}$.

A random vector in $\mathcal{H}$ is (a.k.a random element X of $\mathcal{H}$) is a measurable map from a probability space $(\Omega, \mathcal{F}, P)$ to $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$.

Lemma

When $\mathcal{H}$ is separable, X (or more precisely, its distribution) is tight:

given any $\epsilon > 0$, $\exists$ a compact set K such that $\mu(K^c) \equiv \mathbb{P}\{X \notin K_\epsilon\} < \epsilon$

This will follow easily later, when we give a characterization of compact subsets of $\mathcal{H}$ via the notion of flatness (totally bounded $\iff$ flat and bounded).

The argument below is almost as easy, and applies to any separable metric space (so that we have a countable dense subset) that is complete (so that compact = closed + totally bounded).

- For $\{s_n\} \subseteq \mathcal{H}$, let $b_n(\delta) := \{x \in \mathcal{H} : \|x - s_n\| \leq \delta\}$.
- Obviously, $\mathcal{H} = \cup_{k=1}^{\infty} b_k(\delta)$ for any $\delta > 0$, so $\mu\Big(\cup_{k=1}^{n} b_k(\delta)\Big) \uparrow \mu(\mathcal{H}) \equiv 1$.
- So there is $n_m \geq 1$ such that $\mu\Big(\cup_{k=1}^{n_m} b_k(1/m)\Big) \geq 1 - 2^m\epsilon$.
- Now define $K = \cap_{m=1}^{\infty} \cup_{k=1}^{n_m} b_k(1/m)$. We claim it is compact. It is obviously closed.
- For any $\delta > 0$, $K \subset \cup_{k=1}^{n_m} b_k(1/m) \subset \cup_{k=1}^{n_m} b_k(\delta)$ by taking m sufficiently large ($m > 1/\delta$). So we have shown it is totally bounded (admits a finite δ-cover for any $\delta > 0$).
- Finally,

$$\mu(K^c) \leq \mu\left(\cup_{m=1}^{\infty} \cap_{k=1}^{n_m} b_k(1/m))\right) \leq \sum_{m=1}^{\infty} \mu\left(\cap_{k=1}^{n_m} b_k(1/m)\right) < \sum_{m=1}^{\infty} (1 - 2^m\epsilon) = \epsilon.$$

The characteristic functional is defined naturally:

Definition

The characteristic functional of X is given by $\chi(f) = E(\mathrm{e}^{i\langle X, f\rangle})$ for $f \in \mathcal{H}$.

As a direct corollary to our earlier theorem, it determines the distribution of X.

Yet it is not as useful as in finite dimensions:

1. Bochner's theorem is false when $\dim(\mathcal{H}) = \infty$.
2. Lévy's continuity theorem is false when $\dim(\mathcal{H}) = \infty$.

To see # 1, recall Bochner's theorem:

> *a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ with $f(0) = 1$ is non-negative definite if and only if it is the characteristic function of a probability distribution.*

Now take $\chi(x) = \exp(-\|x\|^2/2)$, $x \in \mathcal{H}$, which fits the bill (cts, PSD, $\chi(0) = 1$). If there is a random vector X with c.f. χ, then for any orthonormal $\{e_i, e_j\}$,

$$\mathbb{E}(\mathrm{e}^{i(t_i\langle X, e_i\rangle + t_j\langle X, e_j\rangle)}) = \exp\{-(t_i^2\|e_i\|^2 + t_j^2\|e_j\|^2)/2\} = \mathrm{e}^{-t_i^2/2}\mathrm{e}^{-t_j^2/2} \leftrightarrow N(0, I_{2\times 2})$$

So, for an ONB $\{e_k\}$ of $\mathcal{H}$, we have $\langle X, e_k\rangle \overset{i.i.d.}{\sim} N(0,1)$ which leads to $\|X\|^2 = \sum_{k=1}^{\infty}\langle X, e_k\rangle^2 = \infty$ almost surely...

The Bochner mean $\mathbb{E}[X]$ of a random vector $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ is defined to be its Bochner integral $\int_\Omega X(\omega)\mathbb{P}(d\omega)$, provided it exists.

Recall that for Hilbert spaces, a necessary and sufficient condition for existence of the Bochner integral is the existence of $\int_\Omega \|X(\omega)\|\mathbb{P}(d\omega)$. Hence:

> *The Bochner mean of a $\mathcal{H}$-valued random element X exists if and only if $\mathbb{E}(\|X\|) < \infty$, and in this case $\|\mathbb{E}(X)\| \leq \mathbb{E}(\|X\|)$.*

A different (weaker) approach is to go via linear functionals:

- Suppose that $\mathbb{E}(|\langle X, f\rangle|) < \infty$ for each $f \in \mathcal{H}$.
- Define the linear functional $\phi : \mathcal{H} \to \mathbb{R}$ given by $\phi(f) = \mathbb{E}(\langle X, f\rangle)$.
- It can be shown that ϕ is a bounded.
- By Riesz representation $\exists\ \mu_\phi \in \mathcal{H}$ such that $\langle \mu_\phi, f\rangle = \mathbb{E}(\langle X, f\rangle)\ \forall\ f \in \mathcal{H}$.
- We call μ_ϕ the Gelfand-Pettis mean of X.

> *The Gelfand-Pettis mean of a $\mathcal{H}$-valued random element X exists if and only if $\mathbb{E}(|\langle X, f\rangle|) < \infty$ for each $f \in \mathcal{H}$.*

- When $\mathbb{E}(\|X\|) < \infty$, the two means agree and the Bochner mean μ satisfies

$$\langle \mu, f \rangle = \mathbb{E}(\langle X, f \rangle), \qquad \forall f \in \mathcal{H}.$$

- Exercise: Show by counterexample that existence of a Gelfand-Pettis mean need not imply existence of mean in Bochner sense when $\dim\{\mathcal{H}\} = \infty$.
- Consequently, the Gelfand-Pettis mean is a weaker notion of mean than the Bochner mean.

What happens when the evaluation functionals are continuous?

- Let $\mathcal{H}$ is a RKHS of functions defined over a set E
- Evaluation can be represented as a continuous linear functionals
- Consequently, if the weak mean μ_ϕ exists then it is also a pointwise mean

$$\{\mu_\phi\}(t) = \mathbb{E}[X(t)], \qquad \forall t \in \mathbb{E}.$$

- So if the strong mean exists, it also equals the pointwise mean,

$$\{\mathbb{E}[X]\}(t) = \mathbb{E}[X(t)] \qquad \forall t \in E.$$

- But when $\mathcal{H}$ is not an RKHS (e.g. $\mathcal{H} = L^2$, which strictly speaking is not a function space) the pointwise mean doesn't even have a meaning.

In summary, we can think of different notions of 'mean' when $\dim\mathcal{H}$, which may exist under different settings; but notice that when any **two** exist, they coincide.

The covariance operator of a random vector $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathcal{H}$ in a separable Hilbert space $\mathcal{H}$ is defined as the Bochner integral

$$\mathcal{K} = \mathbb{E}\{(X - \mu) \otimes (X - \mu)\}$$

in the Hilbert-Schmidt sense, provided it exists. Here $\mu = \mathbb{E}(X)$ is the Bochner mean, which exists whenever the covariance exists.

- In effect, we consider the tensor $(X - \mu) \otimes (X - \mu)$ as a random vector in the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}$, and take its Bochner mean.
- So, a necessary and sufficient condition for the existence of covariance operator is $\mathbb{E}(|||(X - \mu) \otimes (X - \mu)|||) < \infty$.
- Since $|||(X - \mu) \otimes (X - \mu)||| = ||X - \mu||^2$, an equivalent necessary and sufficient condition is $\mathbb{E}(||X||^2) < \infty$. This in turn implies that the Bochner mean also exists.
- Note that we can re-write $\mathcal{K} = \mathbb{E}(X \otimes X) - \mu \otimes \mu$.

Exercise: The covariance operator satisfies

$$\langle \mathcal{K} f, g \rangle = \text{cov}(\langle X - \mu, f \rangle, \langle X - \mu, g \rangle), \qquad \forall\, f, g \in \mathcal{H}.$$

Relate this to Gelfand-Pettis integrals, and to pointwise integrals when $\mathcal{H}$ is an RKHS.

Theorem

The covariance $\mathcal{K}$ of a random vector X in a separable Hilbert space such that $\mathbb{E}\|X\|^2 < \infty$ is:

1. self-adjoint
2. non-negative definite
3. trace class (and hence compact) with trace

$$\|\mathcal{K}\|_{tr} = \text{tr}(\mathcal{K}) = \mathbb{E}(\|X - \mu\|^2) = \sum_{j=1}^{\infty} \lambda_j,$$

with $\{\lambda_j\}$ the eigenvalues of $\mathcal{K}$.

Proof

Clearly, $\mathcal{K}$ is self-adjoint and non-negative definite. Let $\{e_j\}$ be an o.n.b. of $\mathcal{H}$. Then, $\|X\|^2 = \sum_{j=1}^{\infty} \langle X, e_j \rangle^2$, so that
$\mathbb{E}(\|X\|^2) = \sum_{j=1}^{\infty} \mathbb{E}(\langle X, e_j \rangle^2) = \sum_{j=1}^{\infty} \langle \mathcal{K} e_j, e_j \rangle = \|\mathcal{K}\|_{tr} = \text{tr}(\mathcal{K}) = \sum_{j=1}^{\infty} \lambda_j.$

So, when $\dim(\mathcal{H}) = \infty$, there is no random element with the identity operator as its covariance operator...

Corollary (Spectral Decomposition)

When the covariance operator $\mathcal{K}$ of a random vector X in a separable Hilbert space exists, it admits the spectral decomposition

$$\mathcal{K} = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j$$

where $\{\phi_j\}$ are eigenfunctions thereof, forming an ONB of $\overline{\text{range}}(\mathcal{K})$, and $\{\lambda_j\}$ are its eigenvalues, which in turn are non-negative, have finite multiplicities, and form a sequence in ℓ_1.

Theorem

In the same context as above,

$$\mathbb{P}\big[X - \mu \in \overline{\text{range}}(\mathcal{K})\big] = 1.$$

Further, with probability one,

$$X - \mu = \textstyle\sum_{j=1}^{\infty} \langle X - \mu, \phi_j \rangle \phi_j \,,$$

where the $\langle X - \mu, \phi_j \rangle$'s are uncorrelated, with zero mean and variance λ_j.

Proof.

For the first part of the theorem, note that $(\text{range}(\mathcal{K}))^{\perp} = \ker(\mathcal{K}^*) = \ker(\mathcal{K})$ since $\mathcal{K}$ is self-adjoint. So, for any $f \in (\text{range}(\mathcal{K}))^{\perp}$, we have

$$\mathbb{E}(\langle X - \mu, f\rangle^2) = \langle \mathcal{K}f, f\rangle = 0.$$

. (alternatively, $\mathcal{K}f \in \text{range}(\mathcal{K})$ so that $\langle \mathcal{K}f, f\rangle = 0$)

Thus, with probability one, $X - \mu \perp (\text{range}(\mathcal{K}))^{\perp}$ i.e.

$$X - \mu \in \left[(\text{range}(\mathcal{K}))^{\perp}\right]^{\perp} = \overline{\text{range}(\mathcal{K})}.$$

The second part follows from the spectral decomposition of $\mathcal{K}$.

The the Fourier expansion of X w.r.t. the eigenvectors of $\mathcal{K}$ is "optimal":

Theorem (Optimal Fourier Truncation)

Let $\{e_j\}$ be an ONB of $\mathcal{H}$. Then, for any $r \geq 1$,
$\mathbb{E}\{\|X - \mu - \sum_{j=1}^{r}\langle X - \mu, e_j\rangle e_j\|^2\} = \mathbb{E}(\|X - \mu\|^2) - \sum_{j=1}^{r}\langle \mathcal{K}e_j, e_j\rangle \geq$
$\mathbb{E}\{\|X - \mu - \sum_{j=1}^{r}\langle X - \mu, \phi_j\rangle \phi_j\|^2\} = \sum_{j=r+1}^{\infty}\lambda_j$
with equality holding when $e_j = \phi_j$ for $j = 1, 2, \ldots, r$. □

Exercise: Prove the last statement.

The cross-covariance operator of two $\mathcal{H}$-valued random elements X_1 and X_2 satisfying $\mathbb{E}(\|X_i\|^2) < \infty$, $i = 1, 2$, is defined as

$$\mathcal{K}_{12} = \mathbb{E}\{(X_1 - \mathbb{E}(X_1)) \otimes (X_2 - \mathbb{E}(X_2))\}.$$

- $|\langle \mathcal{K}_{12} f, g \rangle| \leq \langle \mathcal{K}_1 f, f \rangle^{1/2} \langle \mathcal{K}_2 g, g \rangle^{1/2}$ for any $f, g \in \mathcal{H}$, where $\mathcal{K}_i$ is the covariance operator of X_i for $i = 1, 2$.
- Clearly, $\mathcal{K}_{12}^* = \mathbb{E}\{(X_2 - \mathbb{E}(X_2)) \otimes (X_1 - \mathbb{E}(X_1))\} = \mathcal{K}_{21}$.
- In general this is not a self-adjoint operator, so it will admit an SVD:

$$\mathcal{K}_{12} = \sum_{j \geq 1} \sigma_j \phi_{1j} \otimes \phi_{2j}, \qquad \sigma_j \geq 0,$$

  - $\{\sigma_j^2\}$ being the non-increasing eigenvalues of $\mathcal{K}_{12}^* \mathcal{K}_{12} = \mathcal{K}_{12} \mathcal{K}_{21}$
  - $\{\phi_{1j}\}$ is ONB for $\overline{\text{range}(\mathcal{K}_{12}^*)}$ comprised of $\mathcal{K}_{12}^* \mathcal{K}_{12}$-eigenfunctions
  - $\{\phi_{2j}\}$ is ONB for $\overline{\text{range}(\mathcal{K}_{21}^*)}$ comprised of $\mathcal{K}_{21}^* \mathcal{K}_{21}$-eigenfunctions

Note that we can defined cross-covariance for jointly distributed random vectors $\{X_i\}_{i=1}^2$ being in distinct Hilbert spaces $(\mathcal{H}_i, \langle\cdot,\cdot\rangle_i)$.

- Effectively a single random vector in the product Hilbert space

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \text{ in } \mathcal{H}_1 \times \mathcal{H}_2$$

- So we can think of covariance and cross-covariance operators in "block format" – assuming the means are zero, to alleviate notation,

$$\mathbb{E}\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \otimes \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \mathbb{E}\left[\begin{pmatrix} X_1^2 & X_1 \otimes X_2 \\ X_2 \otimes X_1 & X_2^2 \end{pmatrix}\right] = \begin{pmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{12}^* & \mathcal{K}_{22} \end{pmatrix}$$

- Obviously this generalises to a product $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_p$, yielding a $p \times p$ covariance operator matrix. (check that it is $\succeq 0$ on the product space)

- Example: take a Hilbert space of functions on $E = [0,1]$. You may partition $[0,1]$ into finitely many disjoint intervals and consider the restrictions of a random function on each interval.

Theorem (Baker, 1973)

Let $(\, X_1 \;\; X_2 \,)^\top$ be a random vector in the product Hilbert space $\mathcal{H}_1 \times \mathcal{H}_2$ such that $\mathcal{K}_{11}$ and $\mathcal{K}_{22}$ exist. Then, there exists a unique bounded $\mathcal{R}_{12} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ such that

1. $\mathcal{K}_{12} = \mathcal{K}_{11}^{1/2} \mathcal{R}_{12} \mathcal{K}_{22}^{1/2}$
2. $\left\| \mathcal{R}_{12} \right\|_\infty \leq 1$
3. $\mathcal{R}_{12} = \mathcal{P}_1 \mathcal{R}_{12} \mathcal{P}_2$ where $\mathcal{P}_j$ is the projection onto $\overline{\text{range}}(\mathcal{K}_{jj})$

The operator $\mathcal{R}_{12}$ is the cross-correlation operator of X_1 with X_2

- Note that only cross-correlation operators make sense (there is no intrinsic self correlation of X_1 alone, unless $\mathcal{H}_1$ is itself a product space).
- The theorem allows us to express $\mathcal{R}_{12} = \mathcal{K}_{11}^{+1/2} \mathcal{K}_{12} \mathcal{K}_{22}^{+1/2}$, for $\mathcal{K}_{ii}^{+1/2}$ the pseudoinverse of $\mathcal{K}_{11}^{+1/2}$.
- Accordingly, we can define the correlation operator matrix $\begin{pmatrix} \mathcal{I}_{\mathcal{H}_1} & \mathcal{R}_{12} \\ \mathcal{R}_{12}^* & \mathcal{I}_{\mathcal{H}_2} \end{pmatrix}$

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

Gaussian Vector

A random vector X in $\mathcal{H}$ is said to be a Gaussian if $\langle X, y\rangle$ has a Gaussian distribution on $\mathbb{R}$ for every $y \in \mathcal{H}$.

Note that the Dirac measure δ_α at some $\alpha \in \mathbb{R}$ is considered Gaussian '$N(\alpha, 0)$'

It follows from the definition that a Gaussian measure must be fully determined by a weak (Gelfand-Pettis) mean/covariance pair $(\mu, \mathcal{K})$, implicitly specified via

$$\langle \mu, h\rangle = \mathbb{E}(\langle X, h\rangle) \quad \& \quad \langle h, \mathcal{K}g\rangle = \mathrm{cov}(\langle X, h\rangle\langle X, g\rangle).$$

Theorem (Fernique)

If X is a Gaussian random element then there exists a $t_0 > 0$ such that $\mathbb{E}\{\exp(t_0\|X\|^2)\} < \infty$.

- In particular, a Gaussian vector X satisfies $\mathbb{E}(\|X\|^k) < \infty$ for all $k \geq 1$.
- So Gaussian vectors have "Bochner moments" of all orders.
- Hence the Gaussian measure has (and is determined by) a strong (Bochner) mean and covariance pair, which necessarily coincides with the weak ones.

Theorem (Hajék-Feldman Dichotomy)

Let $N(\mu_1, \mathcal{K}_1)$ and $N(\mu_2, \mathcal{K}_2)$ be two Gaussian measures on a separable Hilbert space $\mathcal{H}$. Then:

- $N(\mu_1, \mathcal{K}_1)$ and $N(\mu_2, \mathcal{K}_2)$ are either singular or equivalent.
- They are equivalent if and only if the following three conditions all hold:
  1. $\mathbf{range}(\mathcal{K}_1^{1/2}) = \mathbf{range}(\mathcal{K}_2^{1/2}) = \mathcal{H}_0$
  2. $\mu_1 - \mu_2 \in \mathcal{H}_0$
  3. $\left(\mathcal{K}_1^{-1/2}\mathcal{K}_2^{1/2}\right)\left(\mathcal{K}_1^{-1/2}\mathcal{K}_2^{1/2}\right)^* - \mathcal{I}$ is Hilbert-Schmidt on $\overline{\mathcal{H}_0}$

In the case where the two measures are shifts of each other (they share the same covariance $\mathcal{K}$), only condition 2 is non-vacuous.

In this case, we can write for N_2-almost all $x \in \mathcal{H}$

$$\frac{dN_1}{dN_2}(x) = \exp\left\{\left\langle \left(\mathcal{K}^+\right)^{1/2}(\mu_1 - \mu_2), \left(\mathcal{K}^+\right)^{1/2}(x - \mu_1)\right\rangle - \frac{1}{2}\left\|\left(\mathcal{K}^+\right)^{1/2}(\mu_1 - \mu_2)\right\|^2\right\}$$

with $\mathcal{K}^+$ the pseudoinverse of $\mathcal{K}$.

- One can easily verify that the characteristic functional of a $N(\mu, \mathcal{K})$ on $\mathcal{H}$ is

$$\chi(f) = \exp\{i\langle f, \mu\rangle - \langle \mathcal{K}f, f\rangle/2\}, \qquad f \in \mathcal{H}.$$

  Conversely, if the characteristic functional of a probability measure on $\mathcal{H}$ is of the above form, then the measure is Gaussian with mean μ and covariance $\mathcal{K}$.

- Exercise: Show that if X is Gaussian, then $\mathbb{P}(X - \mu \in \mathtt{range}(\mathcal{K})) = 0$.

- Exercise: Let X be a $\mathcal{H}$-valued Gaussian random element with mean μ and a strictly positive-definite covariance $\mathcal{K}$. Then, for any proper closed subspace $\mathcal{S}$ of $\mathcal{H}$, we have $\mathbb{P}(X \in \mathcal{S}) = 0$.

- In fact, it can be shown that if X is a Gaussian random element in $\mathcal{H}$, and $\mathcal{S}$ is a Borel measurable subspace of $\mathcal{H}$, then $\mathbb{P}(X \in \mathcal{S}) = 0$ or 1.

- Consequently, the penultimate statement is also true for Borel measurable affine sets.

Theorem (Gaussian Equivalence and Correlation Operators, Bogachev)

Let $(X_1, ..., X_p)$ be a Gaussian random vector in the product Hilbert space $\mathcal{H}_1 \times \ldots \times \mathcal{H}_p$ with (trace-class) covariance operator matrix $\{\mathcal{K}_{ij}\}$. The joint distribution of $(X_1, ..., X_p)$ is equivalent to the product of its marginal distributions if and only if the following two conditions hold:

1. The correlation operator $\mathcal{R}_{ij}$ is Hilbert-Schmidt.
2. The correlation operator matrix $\{\mathcal{R}_{ij}\}$ is strictly positive-definite.

We now obtain an important (IMHO) generalisation:

Theorem (Partition Operator and Cond. Independence, Waghmare/Masak/Panaretos)

In the same (Gaussian) setting, assume that the joint distribution of $(X_1, ..., X_p)$ is equivalent to the product its marginals. Then the correlation operator matrix $\mathrm{R} = \{\mathcal{R}_{ij}\}$ is invertible on $\mathcal{H}_1 \times \ldots \times \mathcal{H}_p$, and the following two are equivalent

1. X_i is conditionally independent of X_j given X_k
2. The (i, j) operator-entry of R^{-1} vanishes.

We can now legitimately call R^{-1} the precision operator matrix.

If we have random vectors (X_1, X_2, X_3) in $\mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3$, we say that X_1 is conditionally independent of X_2 given X_3

$$\text{Law}(X_1, X_2|X_3) = \text{Law}(X_1|X_3)\text{Law}(X_2|X_3).$$

Since we know that the σ-fields $\mathcal{C}$ and $\mathcal{B}(\mathcal{H})$ are the same, we can equivalently interpret this in terms of linear functionals:

$$\text{for all } f_i \in \mathcal{H}_i, \qquad \langle X_1, f_1\rangle_1 \perp\!\!\!\perp \langle X_2, f_2\rangle_2 | \{\langle X_3, h\rangle_3 : h \in \mathcal{H}_3\}.$$

The latter is particularly useful in case $\mathcal{H}_i$ are RKHS.

Warning:

- the equivalence condition may seem natural, but it excludes the case of "partitioning function's domain", unless the function is allowed to have discontinuities at the partition boundaries.
- we need to grapple with continuum issues when considering intrinsic conditional independence (as opposed to extrinsic conditional independence).

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

If we insist on the Hilbert space perspective, then interpreting a random vector as a random function requires us to consider an RKHS ambient space.

But there exists a different notion of a "random function" devoid of any RKHS requirements: a stochastic process

How can the two be aligned? Could it be that we simply have to make assumptions on the law of X?

Let's take it step-by-step:

- When our "data" are functions on E, typically is a compact metric space, we could consider a random function X as a stochastic process $X = \{X(t) : t \in E\}$.
- This immediately makes sense as a "function", compared to the more abstract representation of $X \in \mathcal{H} = L^2(E, \mathcal{B}(E), \nu)$, say.
- For $X = \{X(t) : t \in E\}$ to be a stochastic process, all that is required is for each $X(t)$ to be measurable $(\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$
- This alone won't guarantee that X is measurable as $(\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ (let alone valued in $\mathcal{H}$)

Consider a stochastic process $X = \{X(t) : t \in E\}$, and define its mean function and covariance kernel by

$$m(t) = \mathbb{E}[X(t)], \quad t \in E, \quad \text{and}$$
$$K(t,s) = \text{cov}\{X(t), X(s)\}, \quad t, s \in E,$$

respectively, which exist provided $\mathbb{E}|X(t)|^2 < \infty$ for all $t \in E$.

Processes with well-defined mean function and covariance kernel are called second-order processes.

Note that $K(\cdot,\cdot)$ is bone-fide non-negative definite.

A priori $X(t)$ does not seem to possess any form of continuity – we don't even have a topological context to make sense of continuity.

But we have finite "variances", so such context could be furnished by the Root Mean Square Distance between random variables:

$$\text{RMSE}(X(t), X(s)) = \left(\mathbb{E}|X(t) - X(s)|^2\right)^{1/2} = \|X(t) - X(s)\|_{L^2(\Omega,\mathcal{F},\mathbb{P})}$$

(and recall that $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is in fact a Hilbert space)

A second-order process $X = \{X(t) : t \in E\}$ is called mean-square continuous if

$$\lim_{t_n \to t} \mathbb{E}\left\{[X(t_n) - X(t)]^2\right\} = 0$$

for every $t \in E$ (finer/higher notions of mean square regularity similarly defined)

Since MSE $=$ bias^2 $+$ variance, we expect that mean-square continuity is characterised by the regularity of the mean and covariance kernel – and it *is*:

Theorem (mean-square continuity and mean/covariance covariance)

Let $X = \{X(t) : t \in E\}$ be a second-order process. Then, the following statements are equivalent.

1. X is mean-square continuous.
2. Both $m(\cdot)$ and $K(\cdot, \cdot)$ are continuous.
3. $m(\cdot)$ is continuous and $K(\cdot, \cdot)$ is continuous near the set $\{(t, t) : t \in E\}$.

Proof

(b) $\Rightarrow$ (a): Follows from the definition of mean-square continuity.
(a) $\Rightarrow$ (b): Note that

$$|m(t) - m(s)| \leq \mathbb{E}\{|X(t) - X(s)|\} \leq \mathbb{E}^{1/2}\{|X(t) - X(s)|^2\}.$$

Thus, continuity of $m(\cdot)$ follows from (a).
To prove continuity of $K(\cdot, \cdot)$, we can without loss of generality assume that $m \equiv 0$. Also, for any t, s, t' and s', we have

$$K(t,s) - K(t',s') = \{K(t,s) - K(t,s')\} + \{K(t,s') - K(t',s')\}.$$

By using the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned} |K(t,s) - K(t,s')| &= |Cov(X(t), X(s) - X(s'))| \\ &\leq K^{1/2}(t,t)\mathbb{E}^{1/2}\{|X(s) - X(s')|^2\}, \quad \text{and} \\ |K(t,s') - K(t',s')| &= |Cov(X(t) - X(t'), X(s'))| \\ &\leq K^{1/2}(s',s')\mathbb{E}^{1/2}\{|X(t) - X(t')|^2\}. \end{aligned}$$

So, mean square continuity of X implies continuity of $K(\cdot, \cdot)$.

(b) $\Rightarrow$ (c): Direct.
(c) $\Rightarrow$ (a): Let $m \equiv 0$. The proof follows upon noting that for any $t \in E$ and any sequence $t_n \to t$,

$$\begin{aligned} \mathbb{E}\left\{[X(t_n) - X(t)]^2\right\} &= [K(t_n, t_n) - K(t,t)] - 2[K(t_n, t) - K(t,t)] \\ &\to 0 \quad \text{as } n \to \infty. \end{aligned}$$

Since mean-square continuity implies continuity of K, we can use Mercer's theorem to obtain the immediate corollary:

Corollary (Mercer for Mean Square Continuous Processes)

Let $X = \{X(t) : t \in E\}$ be a mean-square continuous stochastic process with covariance kernel $K(\cdot, \cdot)$. Then, for any finite measure ν supported on E, we have

$$K(t, s) = \sum_{j=1}^{\infty} \lambda_j^{\nu} \phi_j^{\nu}(t) \phi_j^{\nu}(s),$$

where the sum converges absolutely and uniformly on E. Here, the λ_j^{ν}'s and the ϕ_j^{ν}'s are the eigenvalues and the eigenfunctions of the integral operator on $L^2(E, \mathcal{B}(E), \nu)$ with $K(\cdot, \cdot)$.

But can we relate the stochastic process $X(t)$ with the eigenfunctions $\phi_j^\nu(t)$ of K, like we are able to relate a random X in $\mathcal{H} = L^2(E, \mathcal{B}, \nu)$ with the eigenvectors e_j of $\mathcal{K}$?

The answer to this will come in the form of the Karhunen-Loève Expansion.

- We cannot just write $\sum_j \left(\int_E (X(u) - m(u))\phi_j(u)\nu(du)\right)\phi_j(t)$ because even though $\phi_j \in \mathcal{H}$, it's not guaranteed that $\int_E X^2(t)\nu(dt) < \infty$.
- Need to construct a version of "$\langle X - m, \phi_j \rangle$", say $I_X(\phi_j)$
- Consider a sequence of successive refinements of a measurable partition of E, say, $\mathcal{E}_n = \{E_{i,n} : 1 \leq i \leq m(n)\}$ that eventually separates points. Let $t_{i,n} \in E_{i,n}$ for $1 \leq i \leq m(n)$, and $\mathcal{T}_n = \{t_{i,n} : 1 \leq i \leq m(n)\}$.
- For any $f \in \mathcal{H}$, define a "stepwise approximate version" of "$\langle X - m, \phi_j \rangle$" as

$$I_X(f, \mathcal{T}_n) = \sum_{i=1}^{m(n)} \{X(t_{i,n}) - m(t_{i,n})\} \int_{E_i} f(s)\nu(ds) = \int_{E_i} (X_{\mathcal{T}_n}(s) - \mu_{\mathcal{T}_n}(s)) f(s)\nu(ds)$$

with $X_{\mathcal{T}_n}(s) - \mu_{\mathcal{T}_n}(s) = \sum_{i=1}^{m(n)} \{X(t_{i,n}) - m(t_{i,n})\}\mathbf{1}\{s \in E_{i,n}\}$

Such approximation is a reasonable vehicle to get to our expansion, as long as it does not depend on the choice of partition sequence.

Let's calculate $\mathbb{E}\left[I_X(f,\mathcal{T}_n) - I_X(f,\mathcal{T}_n')\right]^2$ for two partitions $(\mathcal{E}_n,\mathcal{T}_n)$ and $(\mathcal{E}_n',\mathcal{T}_n')$,

$$
\begin{aligned}
= & \sum_{i_1=1}^{m(n)}\sum_{i_2=1}^{m(n)} K(t_{i_1,n},t_{i_2,n}) \int_{E_{i_1,n}\times E_{i_2,n}} f(s_1)f(s_2)\nu(ds_1)\nu(ds_2) \\
& + \sum_{j_1=1}^{m'(n)}\sum_{j_2=1}^{m'(n)} K(t_{j_1,n},t_{j_2,n}) \int_{E'_{j_1,n}\times E'_{i_2,n}} f(u_1)f(u_2)\nu(du_1)\nu(du_2) \\
& - 2\sum_{i=1}^{m(n)}\sum_{j=1}^{m'(n)} K(t_{i,n},t_{j,n}) \int_{E_{i,n}\times E'_{j,n}} f(s)f(u)\nu(ds)\nu(du).
\end{aligned}
$$

Since $\{X(t)\}$ is mean square continuous, K is continuous on $E\times E$, uniformly so since E is compact.

Thus, each term (excluding signs and multipliers) can be made arbitrarily close to

$$
\int_{E\times E} K(s,u)f(s)f(u)\nu(ds)\nu(du)
$$

for large values of n.

We now use the completeness of the space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ to conclude that:

1. $\exists$ a random variable $I_X(f) \in L^2(\Omega, \mathcal{F}, P)$ which is the limit of $I_X(f, \mathcal{T}_n)$
2. $I_X(f)$ is independent of the choice of the partition sequence $(\mathcal{E}_n, \mathcal{T}_n)$.

Looking back: we are effectively using mean square continuity to "relate" the Hilbert structures of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $L^2(E, \mathcal{B}(E), \nu)$. Concretely:

Theorem (Loève Isometry)

Let E be compact, $\{X(t)\}_{t \in E}$ be means-square cts, and $f, g \in L^2(E, \mathcal{B}(E), \nu)$.

1. $\mathbb{E}[I_X(f)] = 0$,
2. $\text{cov}(I_X(f), I_X(g)) = \mathbb{E}[I_X(f) I_X(g)] = \int_{E \times E} K(s,u) f(s) g(u) \nu(ds) \nu(du)$
3. $\mathbb{E}[I_X(f)\{X(t) - m(t)\}] = \int_E K(s,t) f(s) \nu(ds) = (\mathcal{K}f)(t)$ for any $t \in E$, where $\mathcal{K}$ is the integral operator on $\mathcal{H}$ associated with $K(\cdot, \cdot)$.
4. If $\{\lambda_j, \phi_j\}$ are the eigenpairs of $\mathcal{K}$, then $\text{cov}\{I_X(\phi_i), I_X(\phi_j)\} = \lambda_j \mathbf{1}\{i = j\}$.

The proof is essentially already done – just use our partitioning approximation.

The isometry is basically in (b): interpreting

- $\mathbb{E}[I_X(f)I_X(g)] = \langle I_X(f), I_X(g)\rangle_{L^2(\Omega,\mathcal{F},\mathbb{P})}$
- $\langle \mathcal{K}^{1/2}f, \mathcal{K}^{1/2}g\rangle_{L^2(E,\mathcal{B}(E),\nu)} = \int_{E\times E} K(s,u)f(s)g(u)\nu(ds)\nu(du)$

we get an isometry $h \mapsto I_X(\mathcal{K}^{-1/2}h)$ (w/ inverse via extending $X(t) \mapsto K(\cdot, t)$) between

$$\text{the RKHS } \mathrm{range}(\mathcal{K}^{1/2}) \subset L^2(E,\mathcal{B}(E),\nu) \text{ and}$$
$$\overline{\mathrm{span}}\{X(t) - m(t) : t \in E\} \subset L^2(\Omega,\mathcal{F},\mathbb{P})$$

the former with its RKHS inner product (i.e. no closure in $L^2(E,\mathcal{B}(E),\nu)$).

So it is two subspaces of $L^2(\Omega,\mathcal{F},\mathbb{P})$ and $L^2(E,\mathcal{B}(E),\nu)$ that are isometric:

- The RKHS of K is clear. Remember that this can be seen as the closure

$$\overline{\mathrm{span}\left(\{K_t\}_{t\in E}\right)}^{\|\cdot\|_{\mathcal{H}(K)}} = \overline{\left\{\sum_{i=1}^{n} a_i K(\cdot, t_i) : a_i \in \mathbb{R}, t_i \in E, n \geq 1\right\}}^{\|\cdot\|_{\mathcal{H}(K)}}$$

  where $K(\cdot, t)$ is in $L^2(E,\mathcal{B}(E),\nu)$ by continuity and compactness of E.
- The closed span of $X(t) - m(t)$ is

$$\overline{\mathrm{span}\left(\{X(t) - m(t)\}_{t\in E}\right)}^{L^2(\Omega,\mathcal{F},\mathbb{P})} = \overline{\left\{\sum_{i=1}^{n} (X(t_i) - m(t_i))a_i : a_i \in \mathbb{R}, t_i \in E, n \geq 1\right\}}^{L^2(\Omega,\mathcal{F},\mathbb{P})}$$

The main result now is:

Theorem (Karhunen-Lòeve expansion)

For E compact and $X = \{X(t) : t \in E\}$ mean-square continuous process, if we define

$$X_n = \sum_{j=1}^{n} I_X(\phi_j)\phi_j,$$

Then,

$$\lim_{n\to\infty} \sup_{t\in E} \mathbb{E}[\{X(t) - m(t) - X_n(t)\}^2] = 0.$$

Proof

Define $X^c = X - m$. The previous theorem implies that

$$E[\{X_n(t) - X^c(t)\}^2] = E[\{X_n(t)\}^2] - 2E[X_n(t)X^c(t)] + E[\{X^c(t)\}^2]$$

$$= \sum_{j=1}^{n} \lambda_j \phi_j^2(t) - 2\sum_{j=1}^{n} \lambda_j \phi_j^2(t) + K(t,t) = K(t,t) - \sum_{j=1}^{n} \lambda_j \phi_j^2(t).$$

We have seen that the right hand side above converges to zero uniformly over $t \in E$, due to "Mercer for mean-square continuous processes".

The two ingredients that go into the KL expansion are:

- The Loève isometry.
- Mercer's theorem for continuous covariances over compact sets.

Here is a more familiar version of the statement (obviously equivalent):

Theorem (Karhunen-Lòeve expansion)

For E compact and $X = \{X(t) : t \in E\}$ mean-square continuous,

$$X(t) = m(t) + \sum_{n=1}^{\infty} \lambda_n^{1/2} \xi_n \phi_n(t),$$

where

- $\{\lambda_n, \phi_n\}$ are the eigenpairs of K
- $\{\xi_n\}$ are uncorrelated random variables of mean 0 and variance 1.

and the series converges in mean square uniformly over t.

And what about optimality of the expansion? For any ONB of $L^2(E, \mathcal{B}(E), \nu)$, say $\{e_j\}_{j\geq 1}$ we can directly check that

$$\lim_{n\to\infty} \int_E E[\{X(t) - m(t) - \widetilde{X}_n(t)\}^2]\nu(dt) = 0$$

where $\widetilde{X}_n = \sum_{j=1}^n I_X(e_j)e_j$. In terms of IMSE, K's eigensystem is best:

Theorem (KL Optimality, IMSE)

In the same context as in the previous theorem,

$$\int_E \mathbb{E}[\{X(t) - m(t) - \widetilde{X}_n(t)\}^2]\nu(dt) \geq \int_E \mathbb{E}[\{X(t) - m(t) - X_n(t)\}^2]\nu(dt)$$

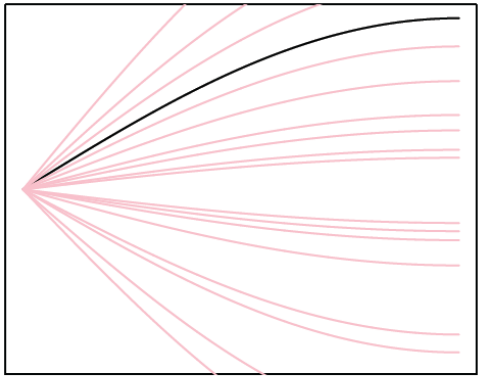
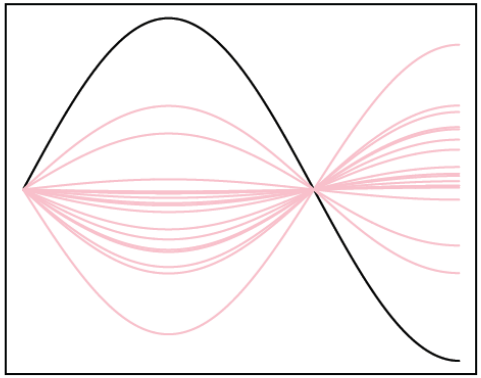
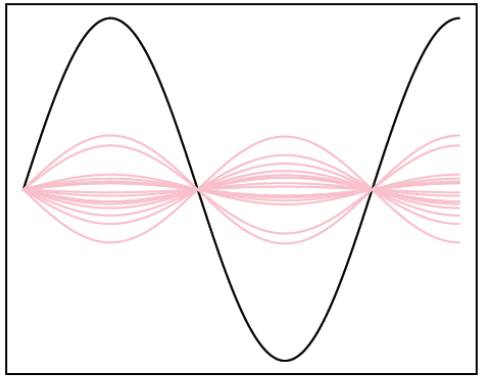
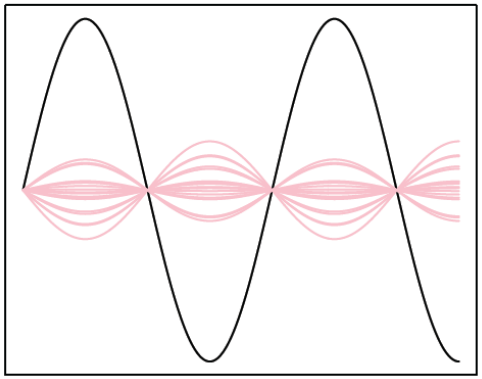
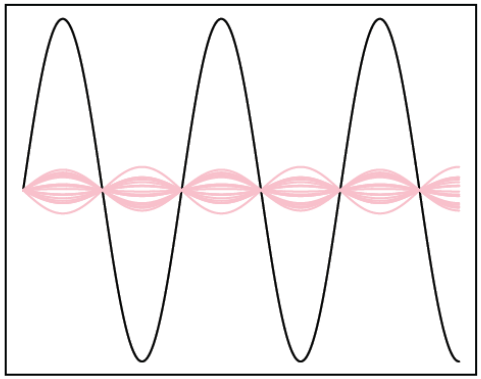
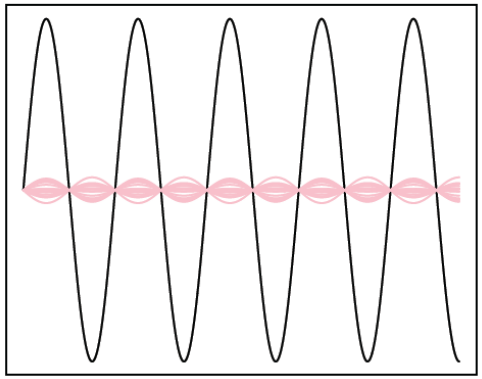
for any choice of ONB $\{e_j\}_{j\geq 1}$ of $L_2(E)$.

Recall that we can find the eigensystem of the kernel $\min\{x, y\}$ explicitly.
:

$$\min\{x, y\} = \sum_{k=1}^{\infty} \frac{2}{\left(k - \frac{1}{2}\right)^2 \pi^2} \sin\left(\left(k - \frac{1}{2}\right) \pi x\right) \sin\left(\left(k - \frac{1}{2}\right) \pi y\right)$$

Thus, when started at 0, BM can be represented as KL expansion:

$$\sum_{k=1}^{\infty} \xi_k \sin\left(\left(k - \frac{1}{2}\right) \pi x\right), \qquad \xi_k \sim N\left(0, \frac{2}{\left(k - \frac{1}{2}\right)^2 \pi^2}\right)$$

(Wish) List of Topics

- The KL expansion obtained is interpretable only in the mean squared sense unlike the almost sure expansion of a random vector X in $L^2(E, \mathcal{B}(E), \nu)$.
- What additional measurability assumption is required to ensure that the collection of pointwise measurable random variables

$$\{X(\omega, t) : t \in E\}$$

is also measurable map

$$\text{from } (\Omega, \mathcal{F}, P) \text{ to } (L^2(E, \mathcal{B}(E), \nu), \mathcal{B}(L^2(E, \mathcal{B}(E), \nu)))?$$

Definition (Joint measurability)

A random element $X : \Omega \times E \to \mathbb{R}$ is said to be jointly measurable if it is measurable with respect to the product σ-field $\mathcal{F} \times \mathcal{B}(E)$.

- Joint measurability implies that:
  - for each $t \in E$, $X(\cdot, t)$ is a random variable $(\Omega, \mathcal{F}) \to \mathbb{R}$
  - for each $\omega \in \Omega$, $X(\omega, \cdot)$ is a measurable function $(E, \mathcal{B}(E)) \to \mathbb{R}$
- So now, we can think of $X(\omega, \cdot)$ as a vector in $L^2(E, \mathcal{B}(E), \nu)$, provided the norm is finite, and ask is it also measurable with respect to the L^2-norm's Borel σ-algebra.

Theorem

Suppose that a stochastic process $\{X(\omega, t) : t \in E\}$ is jointly measurable and that $\int_E X^2(\omega, t)\nu(dt) < \infty$ for all $\omega \in \Omega$. Then, the vector

$$X : \Omega \to L^2(E, \mathcal{B}(E), \nu), \quad \omega \mapsto X(\omega, \cdot)$$

is measurable from $(\Omega, \mathcal{F}, P)$ to $(L^2(E, \mathcal{B}(E), \nu), \mathcal{B}(L^2(E, \mathcal{B}(E), \nu)))$. Equivalently, X is a random vector in $L^2(E, \mathcal{B}(E), \nu)$.

Proof

By joint measurability, the assumption that $\int_E X^2(\omega, t)\nu(dt) < \infty$ for all $\omega \in \Omega$ for each $\omega \in \Omega$ and Fubini's theorem, it follows that for each $f \in L^2(E, \mathcal{B}(E), \nu)$, the map

$$\omega \mapsto \langle X(\omega, \cdot), f \rangle_{L^2(E, \mathcal{B}(E), \nu)}$$

is measurable. Thus, from an earlier theorem (equality of the Borel and bounded linear functional σ-algebras), it follows that the map X is measurable map from $(\Omega, \mathcal{F}, P)$ to $(L^2(E, \mathcal{B}(E), \nu), \mathcal{B}(L^2(E, \mathcal{B}(E), \nu)))$. Equivalently, X is a $L^2(E, \mathcal{B}(E), \nu)$-valued random element. □

- Measurability (joint or not) is often tedious to check. But when we deal with Borel σ-algebras, it suffices to check continuity.
- Is there a similar approach here?

Theorem

Let $X(\cdot) = \{X(\cdot, t) : t \in E\}$ be a stochastic process with continuous sample paths, i.e., $X(\omega, \cdot)$ is a continuous function on E for each $\omega \in \Omega$. Then,

1. X is jointly measurable and hence is a random element in $\mathcal{H}$.
2. The distribution of X is uniquely determined by its finite dimensional marginals, i.e., the distributions of $(X(\cdot, t_1), X(\cdot, t_2), \ldots, X(\cdot, t_n))$ for all $t_1, t_2, \ldots, t_n \in E$ and all $n \geq 1$.

- Gaussian processes vs Gaussian vectors. We can now compare the definitions of a Gaussian *process* (defined via FIDI) to that of a Gaussian *vector* (defined via bounded linear functionals).
- OK, but mean square continuity does not equate to sample path continuity – and it can't, we need a bit more.
- It will turn out that we can translate quantitatively refined continuity (Hölder) in mean square to continuity of sample paths .

Theorem (Kolmogorov's inequality)

Let $X = \{X(t) : t \in E\}$ be a stochastic process, and suppose that there are finite constants $\alpha, \beta, C > 0$ such that

$$\mathbb{E}\{|X(t) - X(s)|^{\alpha}\} \leq C|t - s|^{1+\beta}$$

for all $t, s \in E$. Then, there is a modification of X with continuous sample paths, indeed γ-Hölder paths for all $\gamma \in (0, \beta/\alpha)$.

To see why translates to a Hölder condition at the level of the $L^2(\Omega, \mathcal{F}, \mathbb{P})$ norm/metric, take $\alpha = 2$ and re-write the inequality as

$$\|X(t) - X(s)\|_{L^2(\mathbb{P})} = \left(\mathbb{E}\left\{|X(t) - X(s)|^2\right\}\right)^{1/2} \leq \sqrt{C}|t - s|^{\frac{1+\beta}{2}}, \ \beta, C > 0.$$

- So we need at least $1/2$-Hölder continuity in RMSE ($\beta > 0$) if we are to use this result for sample path continuity.
- Lipschitz continuity in RMSE will yield $1/2$-Hölder paths.
- Can relate this directly to covariance (exercise): If for a centred stochastic process $X = \{X(t) : t \in E\}$ and constants $C > 0, a > 1$, we have

$$K(t,t) + K(s,s) - 2K(s,t) \leq C|t - s|^{a}$$

then for $\gamma < a/2$, X admits a modification with γ-Hölder continuous paths.

- "Modification" refers to the fact that we can't discern deterministic continuity but rather "continuity in law"
- A stochastic process $\{Y(t)\}_{t\in E}$ is said to be a modification of another process $\{X(t)\}_{t\in E}$ if $\mathbb{P}[X(t) = Y(t)] = 1$ for all $t \in E$.
- In this case, $\mathbb{P}[X(t) = Y(t)\ \forall\ t \in S] = 1$ for any countable subset S of E (and thus for a dense such set, if E is separable), but maybe not for $S = E$.
- Modifications have the same FIDI as the original process.

An example: Brownian motion. Let $X = \{X(t) : t \in [0,1]\}$ be a centered Gaussian process with

$$\text{cov}\{X(t), X(s)\} = \min(t,s), \quad s,t \in [0,1].$$

We can check that it has independent increments and that $\mathbb{P}\{X(0) = 0\} = 1$.

- Now take $a = 4$ and verify that $\mathbb{E}\{(X(t) - X(s))^4\} = 3|t-s|^2$.
- So, there is a modification, say, Y of X with continuous sample paths.
- Indeed they are γ-Hölder for all $\gamma < 1/4$.
- In fact, this is true for any $\gamma < 1/2$ because $E\{(X(t) - X(s))^{2k}\} = C_k|t-s|^k$ ($\alpha = 2k$ and $\beta = k-1$) for all $k \geq 2$.

So what happens to the Karhunen-Lòeve expansion when a mean-square continuous process on E is indeed a random element in $L^2(E)$?

Theorem (Mean Square Continuity)

For E compact and $X = \{X(t) : t \in E\}$ jointly measurable and mean-square cts,

1. We have $\int_E X^2(t)\nu(dt) < \infty$, and the process can also be viewed as a random vector in $L^2(E)$.
2. The process mean $m(t) = \mathbb{E}[X(t)]$ satisfies $\int_E m^2(t)\nu(dt) < \infty$ and moreover $m = \mathbb{E}(X)$, in the Bochner sense.
3. The covariance operator $\mathbb{E}\{(X - m) \otimes (X - m)\}$ exists in the Bochner sense and coincides with the integral operator with kernel is $\mathrm{cov}\{X(s), X(t)\}$.
4. for any $f \in L^2(E)$, we have $I_X(f) = \langle X - m\,, f\rangle_{L^2(E)}$.

So, in the setup of the above theorem, the Karhunen-Lòeve expansion holds both:

- in mean-square, uniformly in t
- almost surely, in the L^2-sense
- The random coefficients obtained by the process from the Loève isometry coincide with the eigenbasis linear functionals applied to the random vector

Proof

Note that

$$\mathbb{E}\left\{\int_E X^2(t)\nu(dt)\right\} = \int_E \{K(t,t) + m^2(t)\}\nu(dt),$$

and the RHS is finite, because of mean square continuity. So we have both joint measurability and square integrability, which establishes (1),

Finiteness of the RHS also establishes that $m \in L^2(E)$, so that for any $f \in L^2(E)$,

$$\mathbb{E}(\langle X, f\rangle_{L^2(E)}) = \mathbb{E}\left(\int_E X(t)f(t)\nu(dt)\right) = \int_E m(t)f(t)\nu(dt) = \langle m, f\rangle.$$

Here we used joint measurability and Fubini. It follows that the vector m is the Gelfand-Pettis expectation of X. But we already established that $\mathbb{E}(\|X\|^2_{L^2(E)}) < \infty$ so the Gelfand-Pettis mean is the Bochner mean, proving (2).

Next the fact that $\mathbb{E}(\|X\|^2_{L^2(E)}) < \infty$ means that the Bochner covariance operator is well-defined (in the Hilbert-Schmidt sense) and will be uniquely determined by $\langle \mathbb{E}\{(X-m)\otimes(X-m)\}f, g\rangle_{L^2(E)}$ as f, g range in $L^2(E)$. This is because an ONB $\{e_j\}$ for $L^2(E)$ gives rise to a tensor basis $\{e_i \otimes e_j\}$ for the space of Hilbert-Schmidt operators on $L^2(E)$.

So, using Fubini's theorem and joint measurability, for any $f, g \in L^2(E)$, we have

$$\langle \mathbb{E}\{(X-m)\otimes(X-m)\}f, g\rangle_{L^2(E)} = \mathbb{E}(\langle (X-m), f\rangle\langle (X-m), g\rangle_{L^2(E)})$$
$$= \mathbb{E}\left(\int_{E\times E}\{X(s)-m(s)\}\{X(u)-m(u)\}f(s)g(u)\nu(ds)\nu(du)\right)$$
$$= \int_{E\times E} K(s,u)f(s)g(u)\nu(ds)\nu(du) = \langle \mathcal{K}f, g\rangle_{L^2(E)}.$$

where $\mathcal{K}$ is the integral operator with kernel $K(s,t) = \mathrm{cov}\{X(s), X(t)\}$. It follows that the operators $\mathcal{K}$ and $\mathbb{E}\{(X-m)\otimes(X-m)\}$ coincide, proving (3). For the last part, let $(\mathcal{E}_n, \mathcal{T}_n)$ be a partition as defined earlier. Then, the joint measurability and Fubini's theorem implies that for any $f \in L^2(E)$,

$$E\left\{(I_X(f, \mathcal{E}_n, \mathcal{T}_n) - \langle (X-m), f\rangle)^2\right\}$$
$$= \sum_{i_1=1}^{m(n)}\sum_{i_2=1}^{m(n)} K(t_{i_1}, t_{i_2})\int_{E_{i_1}\times E_{i_2}} f(s_1)f(s_2)\nu(ds_1)\nu(ds_2) + \int_{E\times E} K(s,u)f(s)f(u)\nu(ds)\nu(du)$$
$$- 2\sum_{i=1}^{m(n)}\int_{E_i\times E} K(t_i, u)f(s)f(u)\nu(ds)\nu(du).$$

and the RHS converges to zero as $n \to \infty$. □

One can imagine that things simplify if we take some RKHS instead of $L^2(E, \mathcal{B}(E), \nu)$ – since evaluation maps are continuous. We expect that joint measurability and mean square continuity should be automatic.

Indeed, so it is. Let $\mathcal{H}(R)$ be a RKHS associated with a continuous kernel $R(\cdot, \cdot)$ defined on $E \times E$, with Mercer expansion

$$R(s,t) = \sum_{n \geq 1} \theta_n r_n(s) r_n(t).$$

Theorem

Let X be a random element of $\mathcal{H}$. Then $\{X(t) : t \in E\}$ (i.e., X viewed as a collection of random variables) is a stochastic process. Conversely, if the stochastic process $\{X(t) : t \in E\}$ is such that the sequence $\theta_n^{-1/2} \int_E X(t) r_n(s) \nu(ds)$ is in ℓ_2 for all $\omega \in \Omega$, then it is also a random vector in $\mathcal{H}(R)$.

Proof

Assume first that X is a random vector in $\mathcal{H}(R)$. Then, $X(t) = \langle X, R(\cdot, t) \rangle_{\mathcal{H}(R)} = e_t(X)$ for all $t \in E$. By continuity of the evaluation maps $e_t(\cdot)$'s, it follows that $X(t)$ is a random variable for all $t \in E$.

For the other direction, the condition implies that $\{X(t) : t \in E\}$ is in $\mathcal{H}(R)$ for each $\omega \in \Omega$. Fix a $f \in \mathcal{H}(R)$. Then, there exists a sequence $f_n \in \mathcal{H}(R)$ such that $f_n(\cdot) = \sum_{i=1}^{n} a_i R(\cdot, t_i)$ and $f_n \to f$ in $\|\cdot\|_{\mathcal{H}(R)}$. Now,

$$\langle X(\omega, \cdot), f_n \rangle_{\mathcal{H}(R)} = \sum_{i=1}^{n} a_i X(\omega, t_i).$$

By measurability of $X(t)$ for all $t \in E$, it follows that the mapping $\omega \mapsto \langle X(\omega, \cdot), f_n \rangle_{\mathcal{H}(R)}$ is measurable for all $n \geq 1$. Thus, the mapping

$$\omega \mapsto \langle X(\omega, \cdot), f \rangle_{\mathcal{H}(R)} = \lim_{n \to \infty} \langle X(\omega, \cdot), f_n \rangle_{\mathcal{H}(R)}$$

is measurable. Hence, X is measurable with respect to the bounded linear functional σ-algebra of $\mathcal{H}(R)$, and thus with the Borel σ-algebra. Since it is also of finite norm almost surely it is a bona-fide random vector in $\mathcal{H}(R)$. □

So if we start with a random vector X in the RKHS $\mathcal{H}(R)$, we can get a stochastic process automatically. Is it mean-square continuous? And what is its mean function and covariance kernel?

Theorem

Let X be a random element of $\mathcal{H}(R)$ with Bochner second moment, i.e. $E(\|X\|_R^2) < \infty$. Denote its mean vector and covariance operator by μ and $\mathcal{K}$. Then, $X(t) := \langle X, R_t \rangle_R$ is a mean-square continuous process on E. Furthermore,

- The process pointwise mean satisfies

$$m(t) := \mathbb{E}[X(t)] = \langle \mu, R(\cdot, t) \rangle_R = \mu(t).$$

- The process covariance kernel satisfies

$$K(s,t) \equiv \text{cov}\{X(t), X(s)\} = \langle \mathcal{K}R(\cdot, t), R(\cdot, s) \rangle_R.$$

Moreover, $K \in \mathcal{H}(R) \otimes \mathcal{H}(R)$ with $\|K\|_{\mathcal{H}(R)\otimes\mathcal{H}(R)} \leq \mathbb{E}(\|X - \mu\|_R^2) < \infty$.

Note that $\mathcal{K}$ is NOT the integral operator on $\mathcal{H}(R)$ associated with the kernel $K(\cdot, \cdot)$ - difference from the $L^2(E, \mathcal{B}(E), \nu)$ setup.

Proof

Since $X(t) = \langle X, R(\cdot, t)\rangle_R$, the expressions of $m(t)$ and $K(t,s)$ follow. The continuity of R on $E \times E$ implies the continuity of $m(\cdot)$ and $K(\cdot,\cdot)$, which in turn implies that X is a mean-square continuous process.

Define a new stochastic process $Y = \{Y(t,s) : (t,s) \in E \times E\}$ with $Y(t,s) = [X(t) - m(t)][X(s) - m(s)]$. Note that $Y = (X - m) \otimes (X - m)$ takes values in the tensor product space $\mathcal{H}(R) \otimes \mathcal{H}(R)$, which is also a RKHS associated with the kernel $\tilde{R}(t,s) = R(\cdot,t)R(\cdot,s)$. Thus, Y is a random element of $\mathcal{H}(R) \otimes \mathcal{H}(R)$. Also,

$$\mathbb{E}(\|Y\|_{\mathcal{H}(R)\otimes\mathcal{H}(R)}) = \mathbb{R}(\|(X-m)\otimes(X-m)\|_{\mathcal{H}(R)\otimes\mathcal{H}(R)}) = \mathbb{E}(\|X-m\|^2_{\mathcal{H}(R)}) < \infty.$$

Further, $K(\cdot,\cdot)$ is the mean function (and thus the mean element) of the random element Y, and thus $K \in \mathcal{H} \otimes \mathcal{H}(R)$. So,

$$\|K\|_{\mathcal{H}(R)\otimes\mathcal{H}(R)} = \|\mathbb{R}(Y)\|_{\mathcal{H}(R)\otimes\mathcal{H}(R)} \leq \mathbb{E}(\|Y\|_{\mathcal{H}(R)\otimes\mathcal{H}(R)}) = \mathbb{E}(\|X - m\|^2_{\mathcal{H}(R)}).$$

Theorem (Karhunen-Lòeve expansion, RKHS case)

Let X be a random element in the RKHS $\mathcal{H}(R)$ with $E(\|X\|^2_{\mathcal{H}(R)}) < \infty$. Denote its Bochner mean vector and covariance operator by m and $\mathcal{K}$, respectively. If $\{(\lambda_j, \phi_j)\}$ denote the eigenpairs of $\mathcal{K}$, then

$$\operatorname{cov}\{X(t), X(s)\} = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \phi_j(s),$$

where the sum converges absolutely and uniformly, and

$$\lim_{n\to\infty} \sup_{t\in E} \mathbb{E}\left[\{X(t) - m(t) - X_n(t)\}^2\right] = 0,$$

where $X_n(t) := \sum_{j=1}^{n} \langle X - m, \phi_j \rangle \phi_j(t)$ for all $t \in E$.

Note that we also have $X - m = \sum_{j=1}^{\infty} \langle X - m, \phi_j \rangle \phi_j$ almost surely.

Proof

From the previous theorem, it follows that

$$K(t,s) = \langle \mathcal{K}R(\cdot,t), R(\cdot,s)\rangle_R = \sum_{j=1}^{\infty} \lambda_j \langle \phi_j, R(\cdot,t)\rangle_R \langle \phi_j, R(\cdot,s)\rangle_R = \sum_{j=1}^{\infty} \lambda_j \phi_j(t)\phi_j($$

Since $K(t,t) = \langle \mathcal{K}R(\cdot,t), R(\cdot,t)\rangle_R$, we have

$$\sup_{t\in E} K(t,t) \leq \left\|\mathcal{K}\right\|_{\infty} \sup_{t\in E} R(t,t) < \infty.$$

So, we have

$$\begin{aligned} \sup_{t,s\in E} \sum_{j=1}^{\infty} \lambda_j |\phi_j(t)\phi_j(s)| &\leq \sup_{t,s\in E} \left(\sum_{j=1}^{\infty} \lambda_j \phi_j^2(t) \sum_{j=1}^{\infty} \lambda_j \phi_j^2(s) \right)^{1/2} \\ &= \sup_{t,s\in E} \{K(t,t)K(s,s)\}^{1/2} < \infty. \end{aligned}$$

Finally, the last assertion can be established by following arguments similar to those used in the proof of the Karhunen-Lòeve expansion earlier and noting that the $\langle X - m, \phi_j\rangle$'s are uncorrelated.

- When does a process take values in a RKHS?
- Consider the standard Brownian motion X on $E = [0, 1]$, which has covariance kernel $K(s, t) = \min(s, t)$.
- Observe that

$$K(t, s) = \int_0^1 \mathbf{I}\{u \in [0, t]\}\mathbf{I}\{u \in [0, s]\} du = \int_0^1 (t - u)_+ (s - u)_+ \, du.$$

- Defining

$$\mathcal{H}_1 = \left\{ f : f(t) = \int_0^1 (t - u)_+ g(u) \, du; \, g \in L^2[0, 1] \right\},$$

it follows that any $f \in \mathcal{H}_1$ is absolutely continuous, satisfies $f(0) = 0$, and admits an almost everywhere (weak) derivative, say $\dot{f}$, that lies in $L^2[0, 1]$.

- Further, from our earlier results, it follows that $\mathcal{H}_1$ is a RKHS with kernel $K(t,s) = \min(t,s)$. The inner product of the RKHS is given by

$$\langle f_1, f_2 \rangle_{\mathcal{H}_1} = \langle \dot{f}_1, \dot{f}_2 \rangle_{L^2[0,1]} = \int_0^1 \dot{f}_1(u)\dot{f}_2(u)\,du.$$

- is the space $W_1^0[0,1] = \{f \in W_1[0,1] : f(0) = 0\}$.

- $W_1[0,1]$ is the space of all absolutely continuous functions that are a.s. differentiable with the a.s. derivative $\in L_2[0,1]$.

- However, it is well known that a.s. all the sample paths of the standard Brownian motion are nowhere differentiable. Thus,

$$\mathbb{P}(X \in \mathcal{H}(K)) = 0.$$

Theorem

Suppose that X is a random vector in the RKHS $\mathcal{H}(K)$ with finite Bochner second moment $\mathbb{E}(\|X\|_K^2) < \infty$. If $K(\cdot,\cdot)$ is the coincides with the covariance kernel of X, it must be that $\dim\{\mathcal{H}(K)\} < \infty$.

Proof

Since $\mathbb{E}(\|X\|_K^2) < \infty$, it follows that the covariance operator $\mathcal{K}$ of X is a trace class operator on $\mathcal{H}(K)$, and so it is compact.
At the same time, by the penultimate theorem the reproducing property $K(\cdot,\cdot)$, we have for all $t, s \in E$,

$$K_t(s) = \langle K_s, K_t\rangle_K = K(t,s) = \langle \mathcal{K}K(\cdot,t), K(\cdot,s)\rangle_K = \langle \mathcal{K}K_t, K_s\rangle_K = (\mathcal{K}K_t)(s).$$

This implies that $\mathcal{K}K_t = K_t$ for all $t \in E$, which in turn implies that $\mathcal{K}f = f$ for all $f \in \mathcal{H}(K)$ (by a limiting argument, approximating any such f from within the span of K_t). So, $\mathcal{K}$ is the identity operator on $\mathcal{H}(K)$. However, the identity operator is compact if and only if $\mathcal{H}(K)$ is finite dimensional. □

- Driscoll (1973) – Necessary and sufficient conditions for the sample paths of a Gaussian process to lie in a RKHS.

- Lukíc and Beder (2001) – General case. They showed that for a stochastic process with covariance kernel K,

  - necessary condition for the process to take values in $\mathcal{H}(R)$ is that $\mathcal{H}(K) \subseteq \mathcal{H}(R)$, and the random element in $\mathcal{H}(R)$ associated with the process has a valid covariance operator.
    For the last condition, it is enough to assume that $E(||X||^2_{\mathcal{H}(R)}) < \infty$.

  - The above "subset" condition is also sufficient if $\mathcal{H}(K)$ is a separable Hilbert space, e.g., when K is continuous.

- The condition $\mathcal{H}(K) \subseteq \mathcal{H}(R)$ holds if there exists a constant $c \in (0, \infty)$ such that $cR - K$ is a non-negative definite kernel. Also, in that case, there exists a $b > 0$ such that $||f||_{\mathcal{H}(R)} \leq b||f||_{\mathcal{H}(K)}$ for all $f \in \mathcal{H}(K)$.

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

Denote by $\mathcal{P}(\mathcal{H})$ the set of probability measures on (the Borel σ-algebra) of a separable Hilbert space $\mathcal{H}$.

We say that a collection $\{\nu_n\}_{n\geq 1} \subset \mathcal{P}(\mathcal{H})$ converges weakly to $\nu \in \mathcal{P}(\mathcal{H})$ if

$$\int_{\mathcal{H}} f(x)\nu_n(dx) \overset{n\to\infty}{\to} \int_{\mathcal{H}} f(x)\nu(dx), \qquad \forall\, f \in \mathcal{C}_b(\mathcal{H}),$$

that is, for all continuous, bounded functions $\mathcal{H} \to \mathbb{R}$. In that case, we write

$$\nu_n \overset{w}{\to} \nu.$$

A collection $\{\nu_n\}_{n\geq 1} \subset \mathcal{P}(\mathcal{H})$ is uniformly tight if for every $\epsilon > 0$, there exists a compact subset $C \subset \mathcal{H}$ such that

$$\nu_n(C) \geq 1 - \epsilon, \qquad \forall\, n \geq 1.$$

Theorem (Prokhorov)

Weak convergence is metrisable. Furthermore, a collection $\{\nu_n\}_{n\geq 1} \subset \mathcal{P}(\mathcal{H})$ is sequentially pre-compact in the topology of weak convergence if and only if it is uniformly tight.

In view of Prokhorov's theorem, to prove that $\nu_n \xrightarrow{w} \nu$, we must

1. Show that $\{\nu_n\}$ is uniformly tight
   (not always easy)
2. Show that all subsequences $\{\nu_{n_k}\}$ converge to the same limit, namely, ν.
   (usually easier)

But for (1) we will need a characterisation of compact sets in $\mathcal{H}$. To this aim:

- Define $d(x, A) = \inf_{z \in A} \|x - z\|$ be the distance of $x \in \mathcal{A}$ from $A \subset \mathcal{H}$.
- Define the ϵ-extension of $A \subset \mathcal{H}$ as $A_\epsilon := \{x \in \mathcal{H} : d(x, A) < \epsilon\}$
- Call a set $C \subset \mathcal{H}$ flat if for every $\epsilon > 0$, there exists a finite dimensional subspace $\mathcal{S} \subset \mathcal{H}$ whose ϵ-extension covers C.

We can now state:

Theorem (Compact = Closed + Bounded + Flat)

A subset F of a Hilbert space $\mathcal{H}$ is totally bounded if and only if it is flat and bounded. So, $F \subset \mathcal{H}$ is compact if and only if it is closed, bounded and flat.

Alternatively: $C \subset \mathcal{H}$ is compact iff $\forall\ \epsilon > 0$, $\exists$ a dimension $d \equiv d(\epsilon) < \infty$ such that C can be covered by the ϵ-extension of a d-dimensional closed ball.

Proof

The second claim follows immediately once we establish the first claim.

We start with the $1 \Rightarrow'$ part. Let $C \subset \mathcal{H}$ be totally bounded. Then, C is bounded. Let $\epsilon > 0$, and choose finite cover of ϵ-balls of C, say, $F = \{B_\epsilon(x_1), \ldots\ldots, B_\epsilon(x_{M_\epsilon})\}$. Then, $C \subset \cup_{i=1}^{M_\epsilon} B_\epsilon(x_i)$. But for $\mathcal{S} = \text{span}\{x_1, \ldots, x_{M_\epsilon}\}$ we also have $C \subset \mathcal{S}_\epsilon$.

For the '$\Leftarrow'$ part, let $C \subset \mathcal{H}$ be bounded and flat. Fix $\epsilon > 0$. By flatness, there exists a finite dimensional subspace $\mathcal{S}$ such that $C \subset \mathcal{S}_\epsilon$. By boundedness, there exists $c > 0$ such that $C \subset \{x : \|x\| \leq a\} =: \overline{B}_a(0)$. Since $\mathcal{S}$ is a finite dimensional subspace and hence closed, the set $A := \mathcal{S} \cap \overline{B}_{(a+\epsilon)}(0)$ is compact being a closed and bounded subset of a finite dimensional space. Thus, there exists an ϵ-ball cover, say, $F = \{B_\epsilon(x_1), \ldots\ldots, B_\epsilon(x_{M_\epsilon})\}$ so that $d(y, F) < \epsilon$ for each $y \in A$. Noting that $A^c \cap \mathcal{S} \subset \overline{B}^c_{(a+\epsilon)}(0)$ and $C \subset \overline{B}_a(0)$, it follows that $d(x, A^c \cap \mathcal{S}) \geq \epsilon$ for any $x \in C$. Thus, for any $x \in C$, we have $\epsilon > d(x, \mathcal{S}) = d(x, A)$. So, if $z \in A$ is such that $\|x - z\| = d(x, A)$ (such a point exists because A is compact), we have that $\|x - z\| < \epsilon$, and so, for some $i = 1, 2, \ldots, M_\epsilon$,

$$\|x - x_i\| \leq \|x - z\| + \|z - x_i\| < \epsilon + \epsilon = 2\epsilon.$$

Thus, F forms an ϵ-cover of C, which implies that C is totally bounded. □

[Boundedness+flatness] is equivalent to an "equi-small tail" condition:

Theorem

Let $\{e_k\}$ be an ONB of a separable Hilbert space $\mathcal{H}$. The following are equivalent:

1. $C \subset \mathcal{H}$ is flat and bounded.
2. $\forall\ \epsilon > 0,\ \exists\ d_\epsilon \geq 1$: $C \subset S_\epsilon$, where $S = \mathrm{span}\{e_1, \ldots, e_{d_\epsilon}\}$.
3. $\forall\ \epsilon > 0,\ \exists\ d_\epsilon \geq 1$: $\sup_{x \in C} \left\| x - \sum_{j=1}^{d_\epsilon} \langle x, e_j \rangle e_j \right\|^2 \equiv \sum_{k=d_\epsilon}^{\infty} \langle x, e_k \rangle^2 < \epsilon^2$

Effectively: a closed and bounded is compact if it can be approximated uniformly well through finite dimensional subspaces ($\equiv$ through basis truncation).

In the context of $L^2(E)$, and if we are merely interested in a sufficient condition for compactness (as opposed to a characterisation), the following is very useful:

Theorem (RKHS balls in $L^2(E)$ are pre-compact)

Let $\mathcal{H}(K)$ be the RKHS of functions on E associated with a Mercer kernel $K(\cdot, \cdot)$ on $E \times E$. Then, any $C \subset \mathcal{H}(K)$ that is bounded in the metric of $\mathcal{H}(K)$ is pre-compact in $L_2(E)$ (so $\overline{C}^{\|\cdot\|_{L^2(E)}}$ is compact).

Exercise: Establish this by showing that the unit ball in $\mathcal{H}(K)$ is contained within the image of a bounded ball around zero in $L_2(E)$ under the map $\mathcal{K}^{1/2}$, where $\mathcal{K}$ is the (compact!) integral operator with kernel K.

We will now see that tightness of a sequence of probability measures can be obtained, by showing that they are "mostly flatly supported", and then showing tightness at a finite dimensional level.

A collection $\{\nu_n\}_{n\geq 1} \in \mathcal{P}(\mathcal{H})$ is said to be flatly concentrated if for every $\epsilon > 0$, there exists a finite dimensional subspace $\mathcal{S}$ such that $\nu_n(\mathcal{S}_\epsilon) \geq 1 - \epsilon$ for all $n \geq 1$.

Theorem

A collection $\{\nu_n\}_{n\geq 1} \in \mathcal{P}(\mathcal{H})$ is uniformly tight if for each $\epsilon, \delta > 0$, there exist $y_1, \ldots y_k$ with associated (finite dimensional) span $\mathcal{S} = \text{span}\{y_1, \ldots, y_k\}$, yielding

1. Flat concentration:
$$\inf_{n\geq 1} \nu_n(\mathcal{S}_\epsilon) \geq 1 - \delta.$$
2. Tightness of the associated finite dimensional marginals:
$$\inf_{n\geq 1} \nu_n \left(\{x \in \mathcal{H} : \max_{1\leq j\leq k} |\langle x, y_j \rangle| \leq r\} \right) \geq 1 - \delta$$
for some $r \in (0, \infty)$.

Note that the last display can be interpreted as $\mathbb{P}\{|\langle X_n, y_j \rangle| \leq r\} \geq 1 - \delta$ for all $n \geq 1$, for X_n with distribution ν_n.

Suppose that $\{\nu_n\}_{n\geq 1}$ satisfies (a) above, where ν_n is the distribution of X_n. If the sequence of distributions of $\langle X_n, f\rangle$ converge weakly for all $f \in \mathcal{H}$, then (b) above holds. So it is not so surprising that now we can get:

Theorem (Weak Convergence via Marginals + Flat Concentration)

Let X, X_n be random elements in a separable Hilbert space $\mathcal{H}$. Assume that

1. $\langle X_n, f\rangle \xrightarrow{w} \langle X, f\rangle$ as $n \to \infty$ for all $f \in \mathcal{H}$.
2. For all $\epsilon, \delta > 0$, there exists a finite dimensional subspace $\mathcal{S}$ such that $\mathbb{P}(X_n \in \mathcal{S}_\epsilon) \geq 1 - \delta$ for all $n \geq 1$.

Then, $X_n \xrightarrow{w} X$ as $n \to \infty$.

Proof

The assumptions of the previous theorem are satisfied. Defining $\nu_n = P \circ X_n^{-1}$, it follows that $\{\nu_n\}$ is uniformly tight and hence precompact in the topology of weak convergence. Suppose that there exists two subsequences $\{\nu_{n'}\}$ and $\{\nu_{n''}\}$ of $\{\nu_n\}$ such that $X_{n'} \xrightarrow{w} \tilde{X}_1$ as $n'' \to \infty$ and $X_{n''} \xrightarrow{w} \tilde{X}_2$ as $n'' \to \infty$. By the continuous mapping theorem, $\langle X_{n'}, f\rangle \xrightarrow{w} \langle \tilde{X}_1, f\rangle$ as $n'' \to \infty$ and $\langle X_{n''}, f\rangle \xrightarrow{w} \langle \tilde{X}_2, f\rangle$ as $n'' \to \infty$ for all $f \in \mathcal{H}$. But by the first assumption in the theorem, the distributions of $\langle \tilde{X}_1, f\rangle$, $\langle \tilde{X}_2, f\rangle$ and $\langle X, f\rangle$ agree for all $f \in \mathcal{H}$. Hence, $\tilde{X}_1 \stackrel{d}{=} \tilde{X}_2 \stackrel{d}{=} X$. □

A useful variant is:

Theorem

Let $\{e_j\}_{j\geq 1}$ be an ONB of $\mathcal{H}$. For each $N \geq 1$, denote $\mathcal{P}_N$ to be the orthogonal projection onto $\text{span}\{e_1, e_2, \ldots, e_N\}$. Let $\{X_n\}_{n\geq 1}$ be a sequence of random elements in $\mathcal{H}$. Suppose that there exists a random element $X \in \mathcal{H}$ such that

1. $\mathcal{P}_N X_n \xrightarrow{w} \mathcal{P}_N X$ as $n \to \infty$ for all $N \geq 1$.
2. $\lim_{N\to\infty} \limsup_{n\to\infty} \mathbb{P}(\|X_n - \mathcal{P}_N X_n\| \geq \epsilon) = 0$ for all $\epsilon > 0$.

Then, $X_n \xrightarrow{w} X$ as $n \to \infty$.

In the special case of Gaussian measures we have:

Theorem (Gaussian weak convergence)

A sequence of Gaussian measures $N(\mu_n, \mathcal{K}_n)$ on a separable Hilbert space converges weakly, if and only if the following two both hold true:

1. μ_n converges in the Hilbert norm to some μ.
2. $\mathcal{K}_n^{1/2}$ converges in the Hilbert-Schmidt norm to some $\mathcal{K}^{1/2}$.

When $N(\mu_n, \mathcal{K}_n)$ converges weakly, the limit is itself Gaussian.

We now have all the tools to upgrade the FIDI CLT to a general CLT:

Theorem (CLT in Separable Hilbert Space)

Let $X_1, ..., X_N$ be i.i.d. random vectors in a separable Hilbert space $(\mathcal{H}, \|\cdot\|)$ such that $\mathbb{E}\|X_i\|^2 < \infty$. Let μ and $\mathcal{K}$ be their Bochner mean and covariance. Then,

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right) \xrightarrow{w} N(0, \mathcal{K}).$$

(actually pairwise independence is enough)

Assuming $\mathbb{E}[X_i] = 0$ and $\mathbb{E}\|X_i\|^4 < \infty$, and defining $\mathcal{X}_i = X_i \otimes X_i$, we see that $\mathcal{X}_i$ are iid Hilbert-Schmidt operators satisfying $\mathbb{E}\left\|\|\mathcal{X}_i\|\right\|^2 < \infty$, and thus

$$\sqrt{N}\left(\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathcal{X}_i}_{\hat{\mathcal{K}}_n} - \mathcal{K}\right) \xrightarrow{w} \mathcal{Z} \sim N\left(0, \mathbb{E}[\hat{\mathcal{X}}_1 \otimes \hat{\mathcal{X}}_1] - \mathcal{K} \otimes \mathcal{K}\right)$$

Note that the weak limit is a *Gaussian random self-adjoint operator*.

Proof.

Take $\mu = 0$ WLOG. First we check the "projection" part. Let $Z_n = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i - 0\right) = \sqrt{N}\,\bar{X}_N$. For any $f \in \mathcal{H}$, we have:

$$\langle Z_n\,,\, f\rangle \xrightarrow{d} N(\langle \mu, f\rangle, \langle \mathcal{K}f, f\rangle)$$

by the one-dimensional central limit theorem. For the "tightness" part, let $\{e_n\}$ be an ONB of $\mathcal{H}$, and define $\mathcal{S}^K = \text{span}\{e_1, ..., e_K\}$, and $\mathcal{P}_K$ to be the projection operator onto $\mathcal{S}^K$. Then, for any $\epsilon > 0$, and any $K < \infty$

$$\begin{aligned}
\mathbb{P}[Z_n \notin \mathcal{S}_\epsilon^K] &= \mathbb{P}\left[\|(I - \mathcal{P}_K)Z_n\|^2 > \epsilon^2\right] \leq \frac{\mathbb{E}\|(I - \mathcal{P}_K)Z_n\|^2}{\epsilon^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left[\langle \mathcal{P}_K X_i, \mathcal{P}_K X_j\rangle\right]}{N\epsilon^2} = \frac{\mathbb{E}\|(I - \mathcal{P}_K)X_1\|^2}{\epsilon^2}
\end{aligned}$$

by Markov's inequality and then noticing that only the N "diagonal terms" in the double sum survive by (pairwise) independence (and they are all equal). To conclude the proof, note that the RHS can be made smaller than any δ, by choosing K large enough. □

Clearly, we get a Hilbertian law of large numbers (in mean square, and thus also in probability) from the CLT, under second Bochner moment conditions.

But, like in finite dimensions, only first Bochner moments are needed for an almost sure LLN, indeed in the more general context of a separable Banach space:

Theorem (Strong Law of Large Numbers)

Let $\{X_n\}$ be a sequence of i.i.d. random elements in a separable Banach space $(\mathcal{B}, \|\cdot\|)$. Suppose that $\mathbb{E}(\|X_1\|) < \infty$. Then,

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1)\right\| \overset{n\to\infty}{\longrightarrow} 0\right\} = 1.$$

Proof.

Without loss of generality, we take $\mathbb{E}[X_1] = 0$. Bochner integrability implies that for any $\epsilon > 0$ we may approximate X_i by a simple version

$$X_{i,\epsilon} = \sum_{j=1}^{N_\epsilon} x_{j,\epsilon} \mathbf{1}\{X_i \in B_{j,\epsilon}\} \quad \& \quad \mathbb{E}\|X_i - X_{i,\epsilon}\| < \epsilon.$$

Crucially, the coefficients $x_{j,\epsilon}$ are the same for all i, because the random vectors are identically distributed, and the expectation depends only on their common law.

Now we may bound

$$\limsup_{n\to\infty} \left\|\bar{X}_n\right\| \leq \limsup_{n\to\infty} \left\|\bar{X}_n - \bar{X}_{n,\epsilon}\right\| + \limsup_{n\to\infty} \left\|\bar{X}_{n,\epsilon}\right\| \quad \text{almost surely.}$$

For the first term in the RHS, the scalar SLLN yields:

$$\limsup_{n\to\infty} \left\|\bar{X}_n - \bar{X}_{n,\epsilon}\right\| \leq \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \|X_i - X_{i,\epsilon}\| < \epsilon \quad \text{almost surely.}$$

As for the second term in the RHS, the scalar SLLN applied to the Bernoulli random variables $\mathbf{1}\{X_i \in B_{j,\epsilon}\}$ yields, with probability 1,

$$\bar{X}_{n,\epsilon} = \sum_{j=1}^{N_\epsilon} x_j \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in B_{j,\epsilon}\} \overset{n\to\infty}{\longrightarrow} \sum_{j=1}^{N_\epsilon} x_j \mathbb{P}\{X_1 \in B_{j,\epsilon}\} = \mathbb{E}X_{1,\epsilon}$$

Noticing that $\|\mathbb{E}X_{1,\epsilon}\| = \leq \|\underbrace{\mathbb{E}X_1}_{=0} - \mathbb{E}X_{i,\epsilon}\| \leq \mathbb{E}\|X_i - X_{i,\epsilon}\| < \epsilon$ we conclude

$$\limsup_{n\to\infty} \left\|\bar{X}_{n,\epsilon}\right\| < \epsilon, \quad \text{almost surely,}$$

so that, putting things together, $\limsup_{n\to\infty} \left\|\bar{X}_n\right\| < 2\epsilon$. □

In the special case of a separable Hilbert space, we can also obtain a rate of convergence for the SLLN, in the form of the following theorem.

Theorem (Rate of Convergence)

Let $\{X_n\}$ be a sequence of i.i.d. random elements in a separable Hilbert space $(\mathcal{H}, \|\cdot\|)$. Suppose that $\mathbb{E}(\|X_1\|) < \infty$. Then,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}(X_1) \right\| = o\left(\sqrt{\frac{\log n}{n}} \right).$$

(Wish) List of Topics

Our limit theory directly yields consistent and asymptotically Gaussian estimators of the mean function and covariance operator, based on an iid sample $\{X_1, ..., X_n\}$ in a separable Hilbert space:

- The empirical mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.
- The empirical covariance $\mathcal{K}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n) \otimes (X_i - \bar{X}_n)$.

These are unbiased, and indeed are the best linear unbiased estimators (BLUE) of their estimands, in terms of mean Hilbert squared error risk and mean Hilbert-Schmidt squared error risk, respectively.

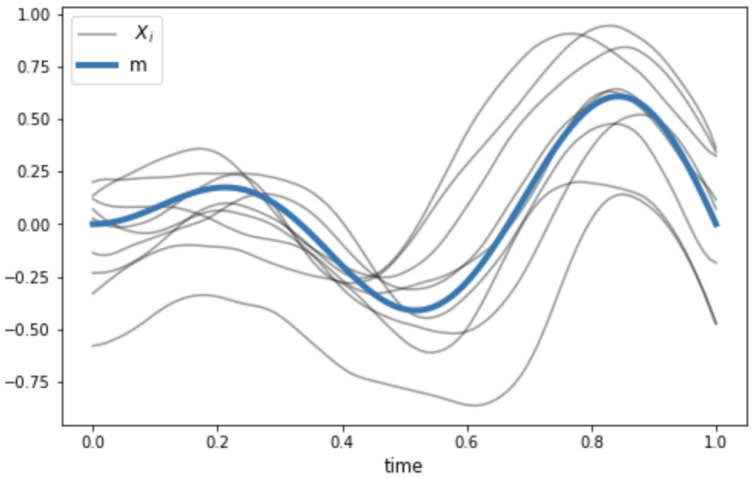
Furthermore, we obtain a.s. rates for the squared loss, of the order $\log n/n$.

- In the case of general separable Hilbert spaces, this is totally legit.
- But in the case of *Hilbert spaces of functions*, one might ask:

do we ever perfectly observe the "Platonic" version of X_i?

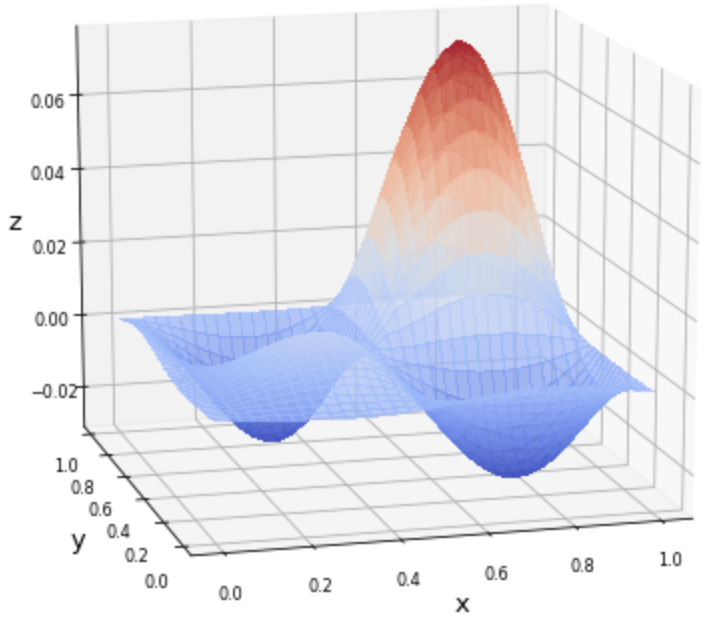
Isn't it, rather, that we may be able to access (noisy) functionals?

Assuming Hölder continuous mean/covariance, evaluations are natural functionals.



The iid replicates X_i and their mean function:

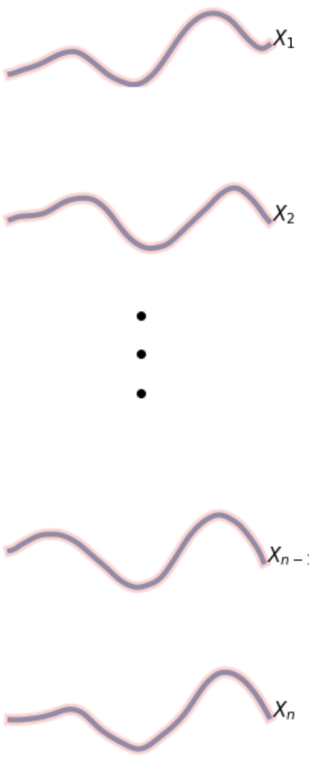
$$\mu(t) = \mathbb{E}[X(t)]$$



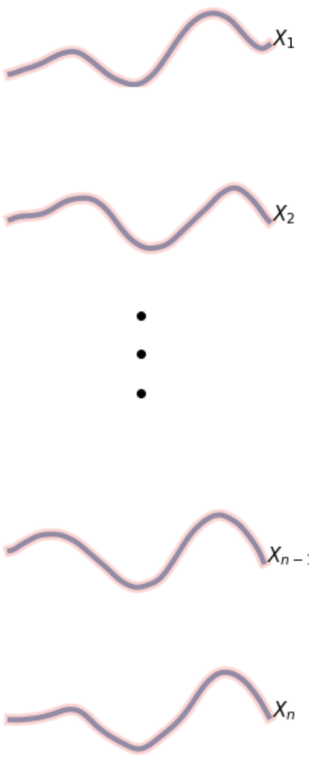
The covariance surface:

$$K(s,t) = \mathbb{E}[X(s)X(t))]$$

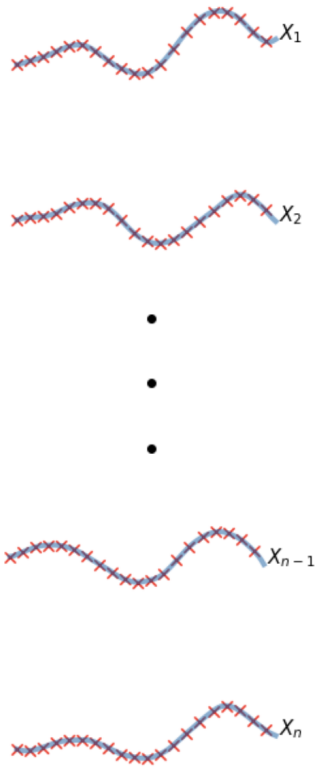
Continuous



Continuous



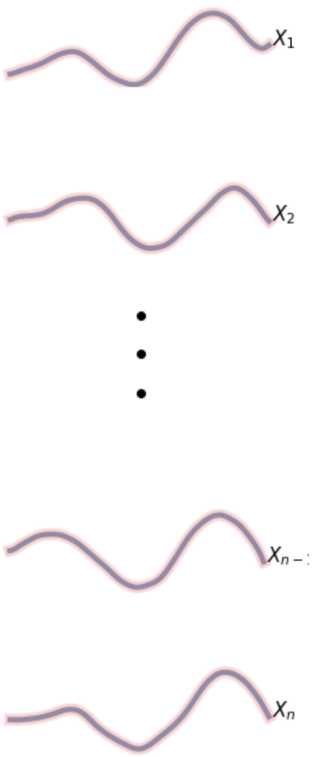
Regular/Dense



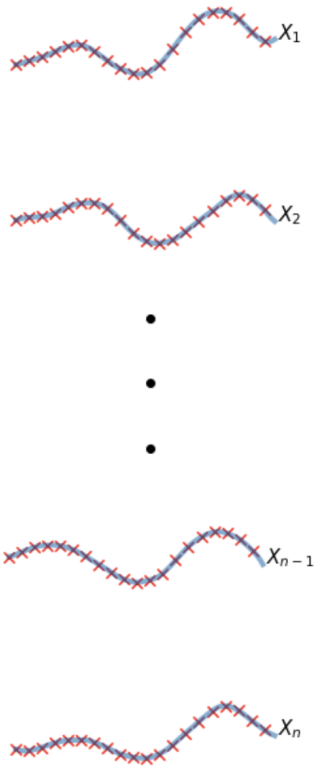
$$Y_{ij} = X_i\left(j/N\right)$$

with $i = 1, \ldots n,\ j = 0, \ldots N$

Continuous



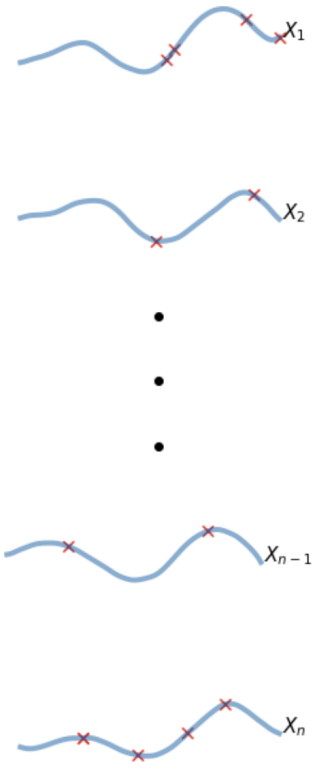
Regular/Dense



$$Y_{ij} = X_i(j/N)$$

with $i = 1, \dots n,\ j = 0, \dots N$

Irregular/Sparse

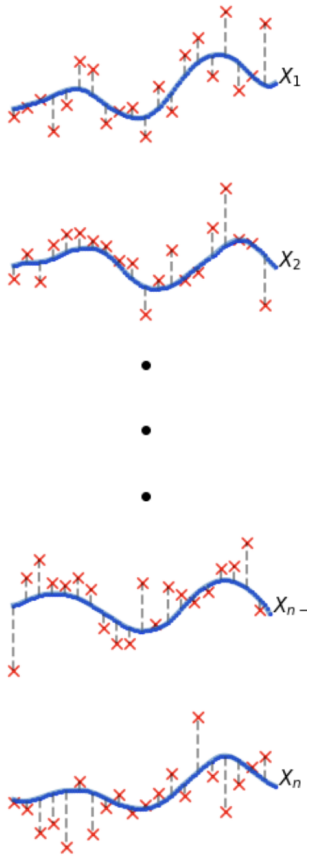


$$Y_{ij} = X_i(T_{ij})$$

with $i = 1, \dots n,\ j = 1, \dots r_i$ and $T_{ij} \in [0, 1]$ and $r_i \in \mathbb{N}$ random.

(some) forms of noisy measurement

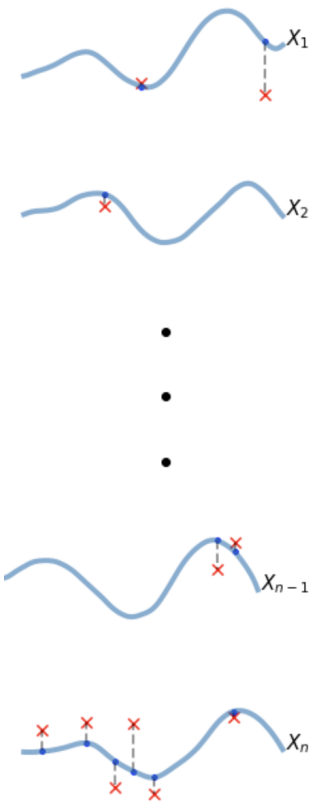
Dense/Regular + *noise*:



$$Y_{ij} = X_i\left(j/N\right) + U_{ij}$$

with $i = 1, \ldots n,\ j = 0, \ldots N$ and U_{ij} random.

Sparse/Irregular + *noise*:



$$Y_{ij} = X_i\left(T_{ij}\right) + U_{ij}$$

with $i = 1, \ldots n,\ j = 1, \ldots r_i$ and $T_i \in [0, 1]$ and $r_i \in \mathbb{N}$ and U_{ij} random.

Two major approaches arose deal with discrete measurements:

1. Smooth discretely observed curves, get proxy mean/covariance

   and,

2. Smooth mean/covariance directly[3], then (if needed) predict curves (kriging)

In either case: typical smoothing assumption was that paths or covariance are C^2

- Led to a *doctrine* that what separates functional from high dimensional data is path smoothness.
- We will see that this is not true, and path smoothness is superfluous.
- It is the trace-class nature of covariances that separates the two data forms.
- And this can be furnished via mean-square continuity in the case of $L^2(E)$.

[3] Especially useful for *sparsely sampled* functions.

Popularised by Ramsay & Silverman (1997), widely used when $E = [0, 1]$ and design is regular/dense.

Assumes sample paths are twice continuously differentiable (C^2).

(in fact, only weak second derivative is necessary)

One defines smoothed curves $\widetilde{X}_i$ as

$$\widetilde{X}_i(t) = \arg \min_{f \in C^2(E)} \left\{ \sum_{j=1}^{r} (f(t_j) - Y_{ij})^2 + \nu \|f''\|^2_{L^2(E)} \right\}, \qquad i = 1, \ldots, n,$$

for $\nu > 0$ a regularising constant.

- Each solution $\widetilde{X}_i$ a cubic spline with knots at t_j
- Can represent in at most r-element basis, useful for computations in R
- Proxy curves $\{\widetilde{X}_i\}$ used in lieu of unobservable $\{X_i\}$
- Take empirical mean/covariance of proxy curves.

Alternatively: Use local polynomical smoothing, or your favourite smoother.

"PACE" (Yao, Müller & Wang (2005), but idea due to Staniswalis & Lee (1997))

Assuming mean is zero, consider the discrete $r \times r$ covariance matrix $\hat{M}$ as

$$\hat{M} = \frac{1}{n} \sum_{i=1}^{n} (Y_{i1}, \ldots, Y_{ir})(Y_{i1}, \ldots, Y_{ir})^\top.$$

Key observation: its population version K satisfies:

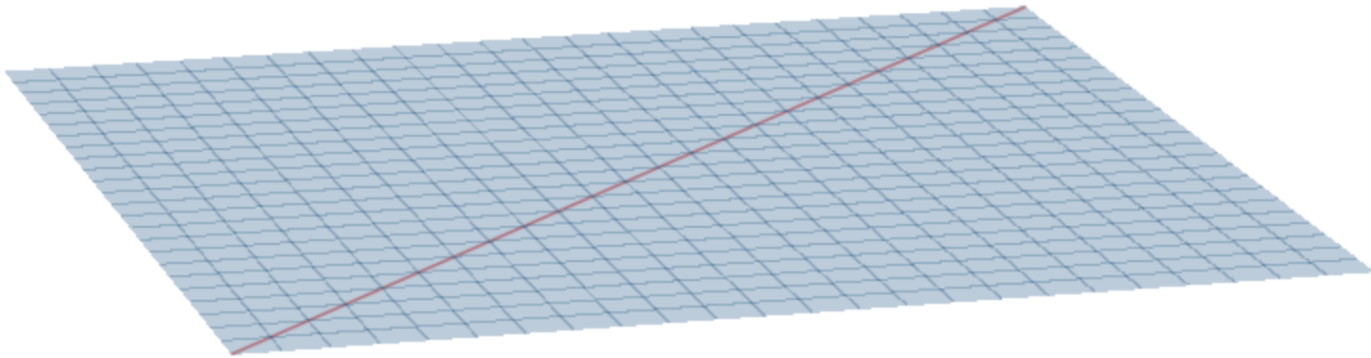
$$M = \mathbb{E}\left[(Y_{i1}, \ldots, Y_{ir})(Y_{i1}, \ldots, Y_{ir})^\top\right] = \{K(t_i, t_j)\}_{i,j=1}^{r} + \sigma^2 I_{r\times r}$$

because U_{ij} are independent of X_i, and iid mean zero, var σ^2!

Motivates estimation strategy via *diagonal removal:*

- Define a discretely sampled kernel $\hat{K}(t_i, t_j) = \hat{M}_{i,j}$, $i \neq j$ ("raw" covariance)
- Assuming $K \in C^2(E^2)$, do 2D smoothing of $\hat{K}$ to get estimator of K.
- If needed, predict X_i by *best linear prediction of KL coefficients* (kriging).

When mean is non-zero, we "pool and smooth" the curves to estimate/subtract.



The "matrix cells" corresponding to the discrete covariance

Notice that in the dense/regular case, smoothness can be entirely circumvented:

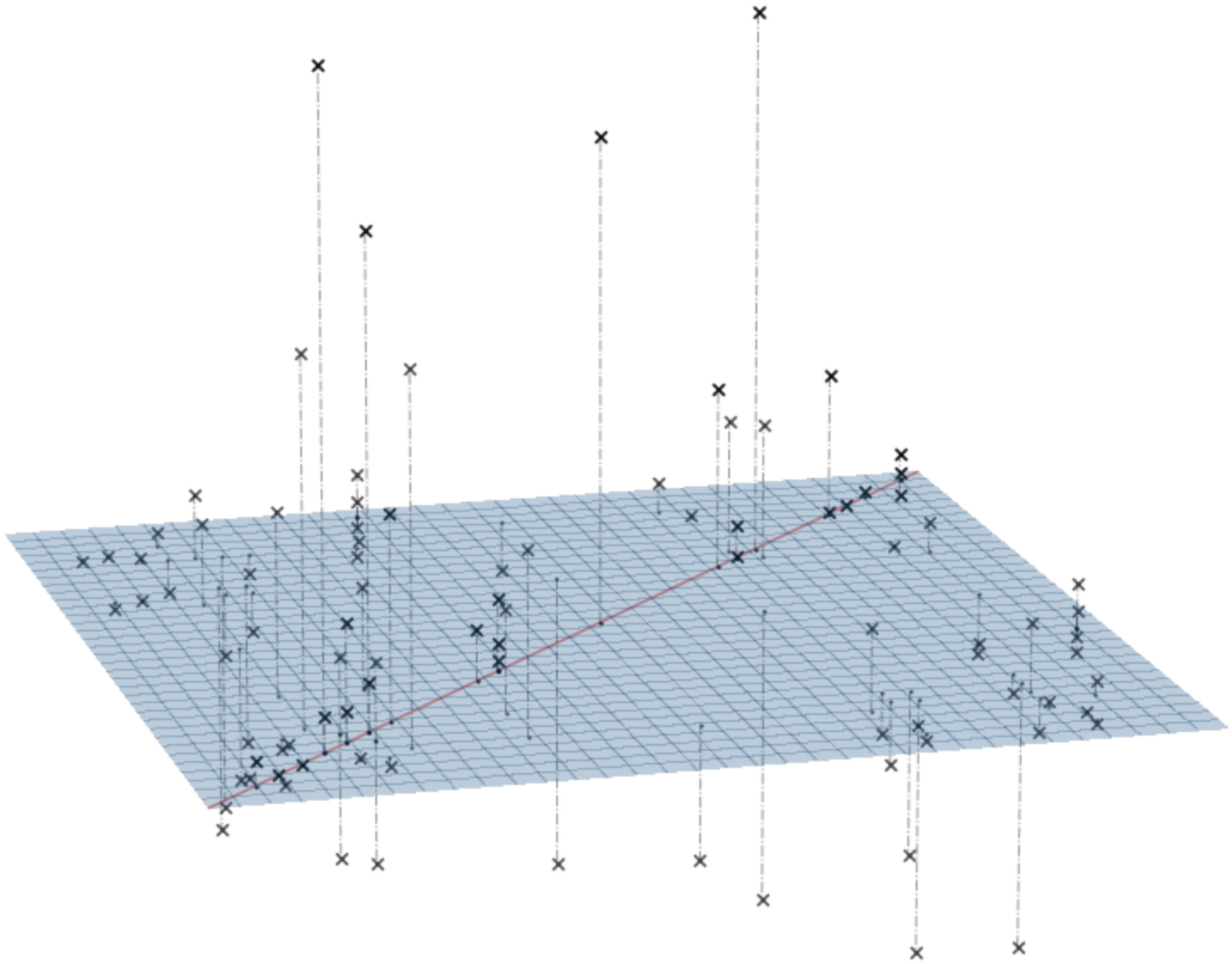
- If we have some Hölder continuity, can approximate by step functions averaging within "pixels" (instead of smoothing)
- This yields rates of convergence that depend on Hölder exponent.

  It can in principle be applied with either order of pooling/averaging but pooling first is more natural.

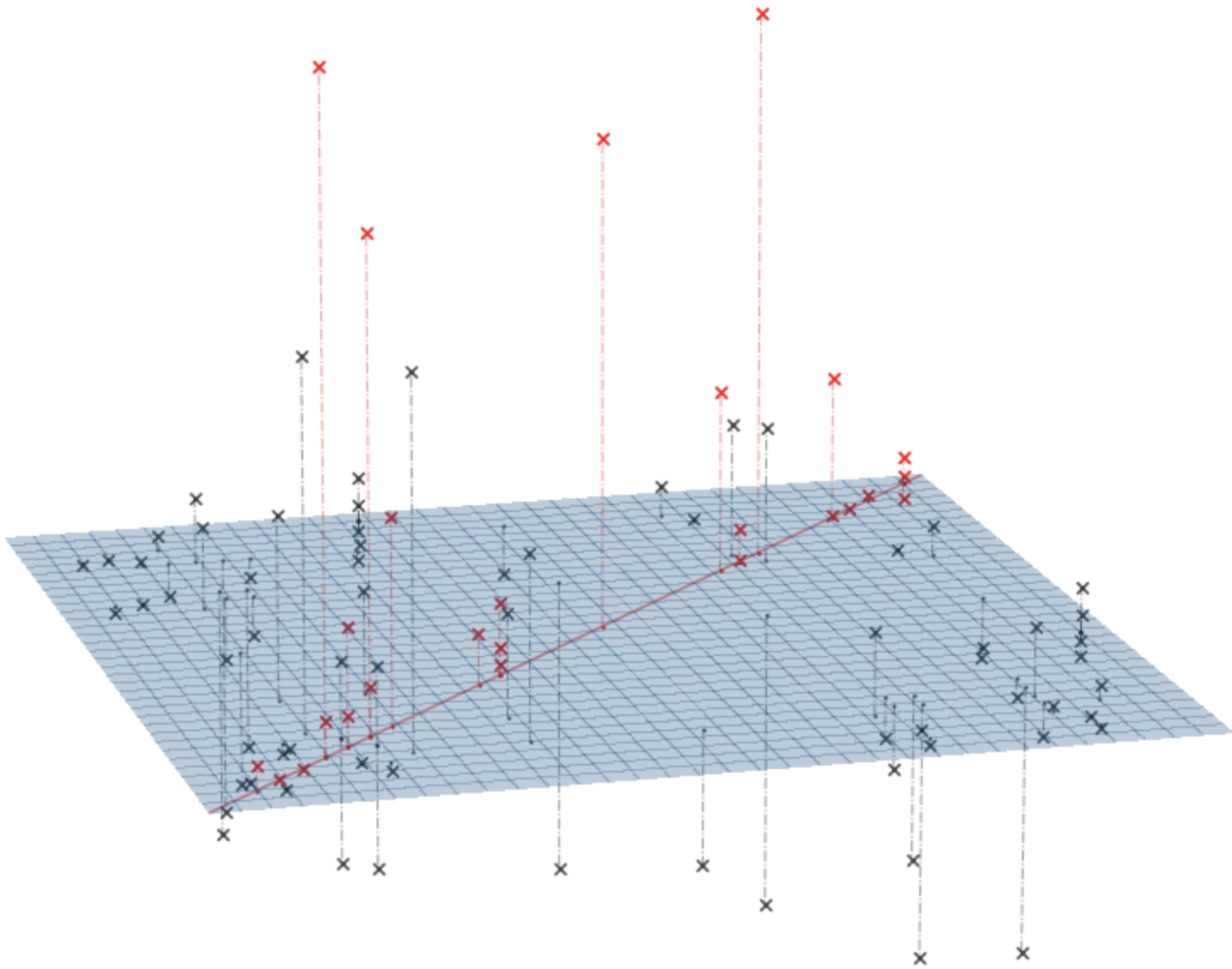
Some remarks:

- The smooth-then-pool approach can apply to irregular dense observations.
- This comes at a computational cost (need $\sim nr$ basis functions).
- Smooth-then-pool cannot apply, however, to irregular and sparse settings.
- By contrast, the pool-then-smooth approach does – it doesn't really require a regular design, all we need is at least $r_i = 2$ measurements per curve.
- In fact, even $r_i = 2$ (fixed) suffices for consisntency.
- Li & Hsing (2010) analysed the pool-then-smooth approach with respect to both r and n, yielding a method indifferent to sparse/dense case (under C^2 assumptions, still)
- Rates of convergence reveal the "phase transition" between sparse and dense sampling regime.
- We will soon relax the C^2 path/covariance assumptions (C^2 mean is)
- In pictures...

Smoothing the covariance with irregular/sparse observations



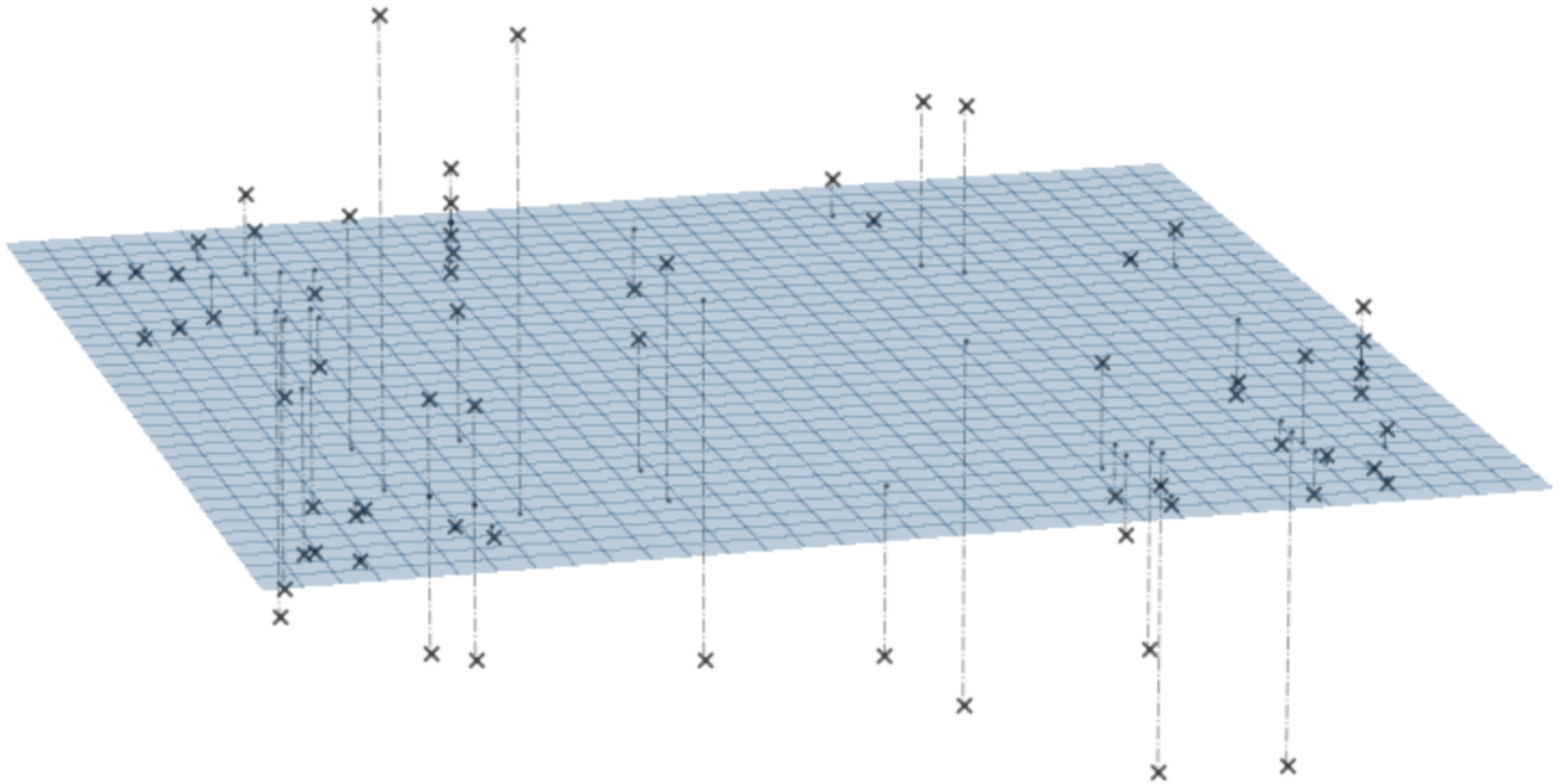
The latent covariance surface and the scatterplot of product observations: $(Y_{ij} \cdot Y_{ik})_{i,j,k}$
where $Y_{ij} = X_i(T_{ij}) + U_{ij}$



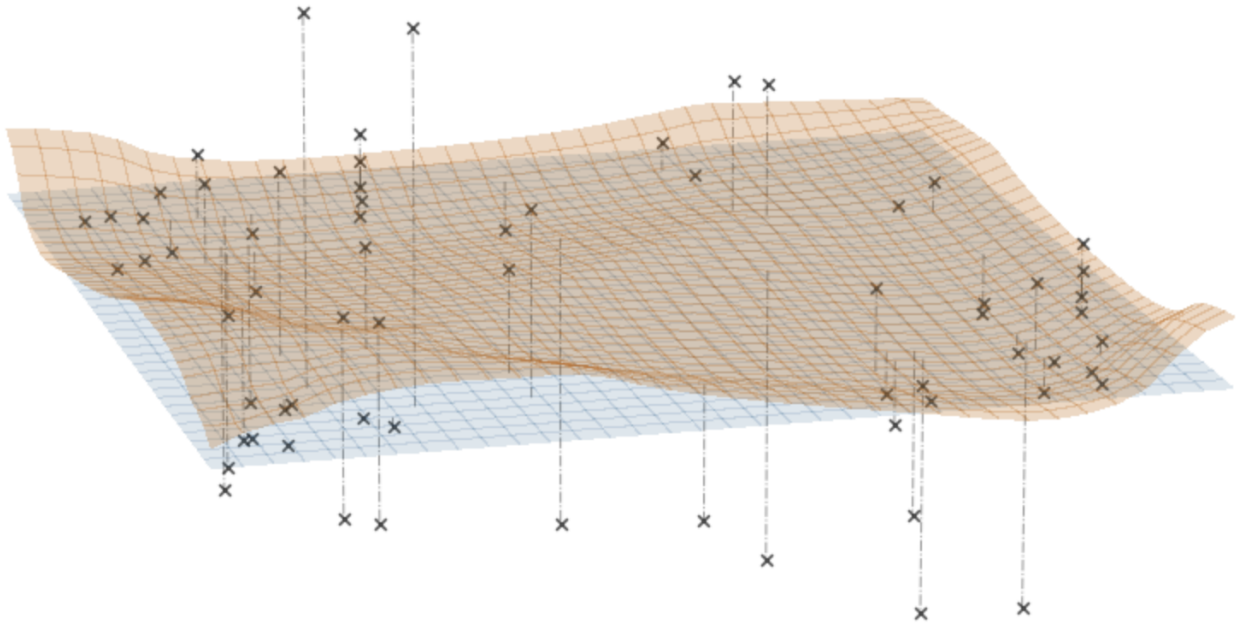
Error contamination affecting diagonal points

$$\mathbb{E}[\,Y_{ij} \cdot Y_{ik}] = K(\,T_{ij}\,,\,T_{ik}) + \delta_{jk}\sigma^2,$$

assuming mean zero.



We exclude diagonal observations.



We can estimate the covariance by 2D smoothing (e.g. locally linear or locally quadratic smoothing) of the product (non diagonal) observations:

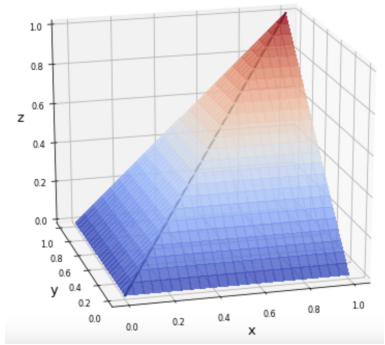
$$(Y_{ij} \cdot Y_{ik})_{i,j \neq k}$$

- Li & Hsing show consistency/rates under C^2 assumption on mean and covariance.
- We will get the same rates while relaxing C^2 covariance assumption (allowing for rough paths) – smoothness will be required for the mean, but this is innocuous, and has no bearing on paths.

Is this really worth fussing about?

- Yes: continuous time Markov processes (diffusions) are the most classic example of random functions, and have nowhere differentiable paths, a.s.
- They should be fair game for us, and not artificially excluded.

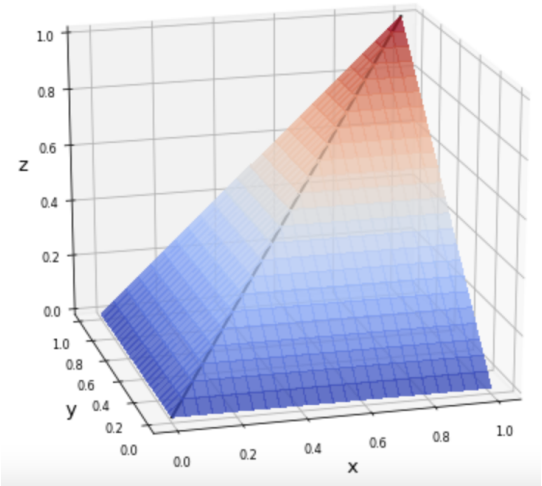
Sample path of standard BM.



The covariance surface of a BM.

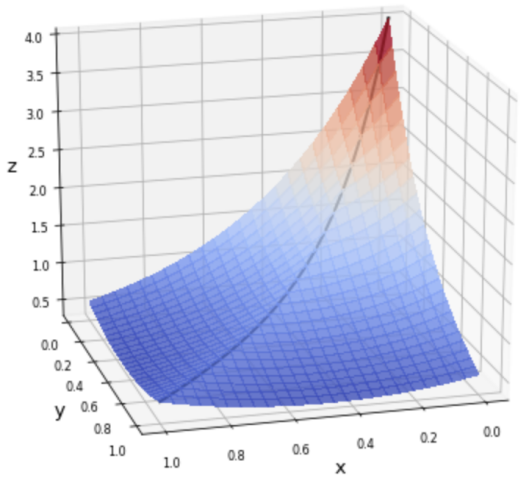
Important observation:

- The BM covariance is non-differentiable near the diagonal. This will happen with all diffusions. Recall that this is what relates to sample path properties.
- When the C^2 assumption fails near a segment of the diagonal, the associated Gaussian process is a.s. non-differentiable almost everywhere on the corresponding interval (Cambanis 1973, Theorem 6).
- However, the BM covariance if flat (infinitely smooth) away from the diagonal. This is where the smoothing is supposed to take place.



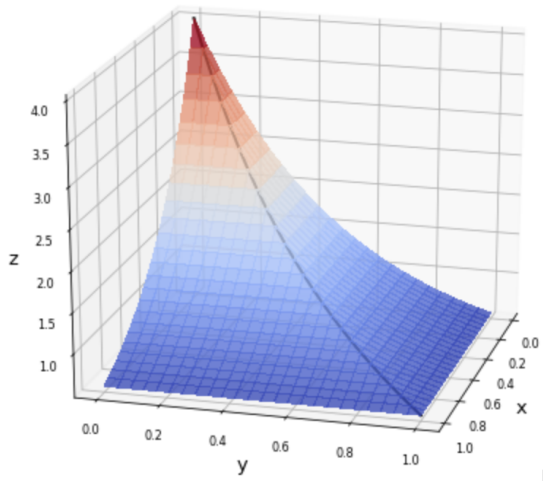
Brownian Motion

$$dX(t) = dW(t)$$



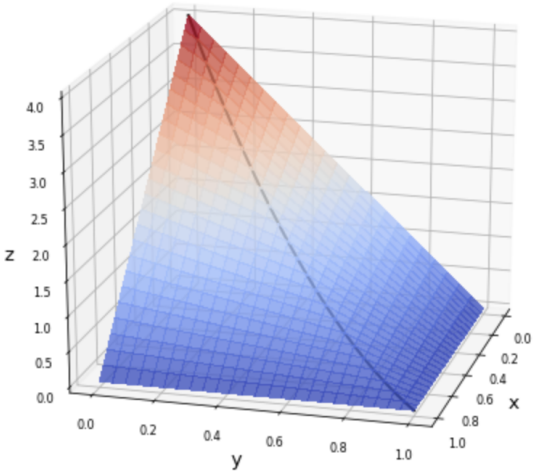
Ornstein Uhlenbeck

$$dX(t) = X(t)\mu\, dt + \sigma\, dW(t)$$



Black Scholes

$$dX(t) = X(t)\mu\, dt + X(t)\sigma\, dW(t)$$



Brownian Bridge

$$dX(t) = \frac{X(t)}{(1-t)}\, dt + dW(t)$$

Define the triangle

$$\triangle = \{(s,t) \in [0,1]^2 \mid 0 \leq t \leq s \leq 1\}$$

and observe that $C|_{\triangle}$ determines K completely.

Now consider some diffusion examples restricted on $\triangle$:

- Brownian motion: $K(s,t) = t$
- Brownian bridge: $K(s,t) = t - st$
- Ornstein-Uhlenbeck process w/ drift β: $K(s,t) = e^{-\beta(t+s)}(e^{2\beta t} - 1)/(2\beta)$
- Geometric Brownian motion: $K(s,t) = \exp\{(t+s)/2\})\exp\{t\} - 1)$

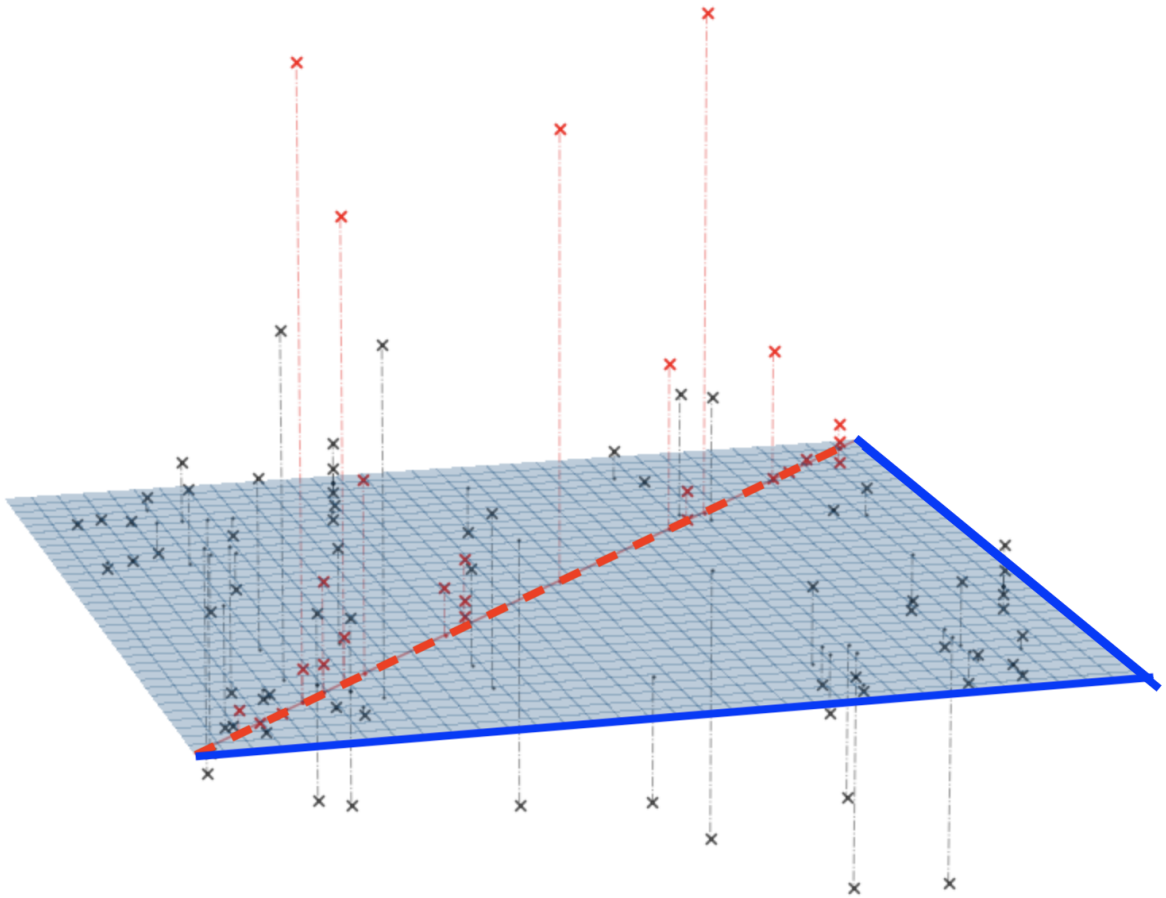
These are instances of a very general setting with a covariance of form

$$K(s,t) = g\{\min(s,t), \max(s,t)\},$$

for some smooth function g.

Any continuous time Gaussian process that is Markov adheres to this form.

Motivates the following modification of the pool-then-smooth procedure:



Smooth and reflect scatterplot restricted to lower triangle:

$$(Y_{ij} \cdot Y_{ik})_{i,j,k}, \qquad 1 \leq k < j \leq r$$

where $Y_{ij} = X_i(T_{ij}) + U_{ij}$

Concretely, let's use local polynomials as our smoothing method of choice.
Define:

$$\widehat{\mathbb{E}}\left(X(s), X(t)\right) = \begin{cases} \widehat{a}_0(s,t) & \text{if } t < s, \\ \widehat{\mathbb{E}}\left(X(t), X(s)\right) & \text{otherwise.} \end{cases}$$

where

$$\begin{aligned}(\widehat{a}_0(s,t), \widehat{a}_1(s,t), \widehat{a}_2(s,t)) \quad &= \\ \underset{a_0, a_1, a_2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{2}{r(r-1)} \sum_{1 \le k < j \le r} &\Big[\left\{ Y_{ik}\, Y_{ij} - a_0 - a_1 \left(T_{ij} - s\right) - a_2 \left(T_{ik} - t\right) \right\}^2 \\ &\times K_{hI_{2\times 2}} \left(\left(T_{ij} - s\right), \left(T_{ik} - t\right) \right) \Big]\end{aligned}$$

Recalling that:

$$Y_i\left(T_{ij}\right) = X_i\left(T_{ij}\right) + U_{ij} \qquad i = 1, \ldots n,\ j = 1, \ldots r(n),$$

This method yields the Li & Hsing rates, without the C^2 covariance assumption:

Theorem (Mohammadi & Panaretos ,2023)

Assume that the evaluation points are sampled uniformly and independently at random and that

- $m \in C^2[0,1]$, $C \in C^2(\Delta)$, $\sup_{t\in[0,1]} \mathbb{E}|X(t)|^{4+\epsilon} < \infty$, and $\mathbb{E}|U_{ij}|^{4+\epsilon} < \infty$.
- kernel function $K(\cdot)$ is suitably chosen (some analytical conditions required).

Then

$$\sup_{0\le s\le t\le 1} \left|\widehat{\mathbb{E}}(X(s)X(t)) - \mathbb{E}(X(s)X(t))\right| = O\left[h^{-4}\frac{\log n}{n}\left(h^4 + \frac{h^3}{r} + \frac{h^2}{r^2}\right)\right]^{1/2} + O\left(h^2\right), \text{ a.s.}$$

Corollary (Dense Sampling Scheme)

If furthermore observations are sufficiently *dense* and bandwidth parameters are tuned at an appropriate rates, then we have, almost surely

$$\sup_{0\le s\le t\le 1} \left|\widehat{\mathbb{E}}(X(s)X(t)) - \mathbb{E}(X(s)X(t))\right| = O\left(\frac{\log n}{n}\right)^{1/2}$$

matching the rate in the SLLN (i.e., the Platonic rate)

(Wish) List of Topics

PCA needs no additional introduction, once we have the KL-expansion

(or the Fourier expansion given by the spectral decomposition, if we are dealing with general vectors)

Assuming X has a Mercer covariance kernel,

$$X(\tau) - m(\tau) = \sum_{n=1}^{\infty} \xi_n \varphi_n(\tau) \qquad \left(\text{i.e. } \mathbb{E} \left\| X_t - m - \sum_{n=1}^{q} \xi_n \varphi_n(\tau) \right\|^2 = \sum_{n>q} \lambda_n \right)$$

where $\xi_n = \langle X - m, \varphi_n \rangle$ are zero mean uncorrelated with variance λ_n.

Captures **complete curve dynamics** – Canonical FDA framework:

- Separation of variables (stochastic vs functional)
- Quantification of smoothness (φ_i contributes as $\lambda_i / \sum_i \lambda_i$)
- Variance components / functional fluctuations around mean
- Optimal finite dimensional representation
  (modeling/methodology+inference/regularization)

Some history:

47/49 Independent introduction by Karhunen and Loève

↪ Linear filtering of stoc. proc. and series rep. of Wiener measure.

1950 Ulf Grenander shows importance in statistics (birth of FDA?)

↪ Uses as coordinate representation for likelihood ratios

1958 C.R. Rao hints potential usefulness for growth curves

↪ Components of variance interpretation

⋮

1973 Kleffe considers empirical version $(\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X}) \otimes (X_t - \bar{X}))$

↪ Large sample convergence

1982 Dauxois, Pousse & Romain develop asymptotics of empirical version

1986 Besse & Ramsey (psychometrics) use as PCA

1991 Rice & Silverman:

"Estimating Mean and Covariance When Data Are Curves"

Subject then takes off, largely thanks to the Ramsey & Silverman book.

Motivates methodology but also shows up in inference:

Functional Regression (Estimation, many different variants)

E.g., estimate $\beta \in L^2$ on the basis of $y_t = \langle X_t, \beta \rangle + \varepsilon_t$, $t = 1, ..., T$.

Functional Analysis of Variance (Testing)

Do several random functions with same covariance share same mean?

Functional Classification (Discrimination)

Given a random function, classify between $\{m_1, \mathcal{R}_1\}, ..., \{m_k, \mathcal{R}_k\}$

As an example, consider the functional linear regression model (with scalar response):

$$y_i = \langle X_i, \beta \rangle + \varepsilon_i, \qquad i = 1, ..., N.$$

where $\mathbb{E}\|X_i\|^2 < \infty$ & ε_i are zero-mean and variance σ^2, uncorrelated with the random vectors X_i in $\mathcal{H}$. Assume we have "complete observations".

Purple: Estimate $\beta \in \mathcal{H}$ given observations $(X_1, y_1), ..., (X_N, y_N)$.

1. Assume that $\mathbb{E}[X_i] = \mathbb{E}[y_i] = 0$ and that covariance $\mathcal{K}$ of X is known and strictly positive definite.
2. Define $\kappa = \mathbb{E}[yX]$ the covariance between y and X, and observe:

$$\kappa = \mathbb{E}[yX] = \mathbb{E}[\langle X, \beta \rangle X] = \mathcal{K}\beta$$

3. Can then easily estimate κ "easily" by $\hat{\kappa} = \frac{1}{N} \sum_{i=1}^{N} y_i X_i$
4. Tempting to then estimate β by

$$\hat{\beta} = \mathcal{K}^{-1} \hat{\kappa}.$$

This naïve approach won't work.

Expand β and κ in the eigenfunction basis given by $\mathcal{K}$:

$$\beta = \sum_{n=1}^{\infty} \underbrace{\langle \beta, \varphi_n \rangle}_{\beta_n} \varphi_n \qquad \kappa = \sum_{n=1}^{\infty} \underbrace{\langle \kappa, \varphi_n \rangle}_{\kappa_n} \varphi_n.$$

Then, employing the estimator $\hat{\beta} = \mathcal{K}^{-1}\hat{f}$ amounts to estimating

$$\hat{\beta}_n = \frac{\hat{\kappa}_n}{\lambda_n}.$$

But

$$\begin{aligned} \hat{\kappa}_n &= \langle \hat{\kappa}, \varphi_n \rangle = \left\langle \frac{1}{N} \sum_{i=1}^{N} y_i \sum_{k=1}^{\infty} \xi_{ik} \varphi_k, \varphi_n \right\rangle \\ &= \frac{1}{N} \sum_{i=1}^{N} y_i \xi_{in} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{k=1}^{\infty} \xi_{ik} \beta_n + \varepsilon_i \right) \xi_{in} \end{aligned}$$

Rearranging and putting things together, we have

$$\hat{\beta}_n = \frac{\hat{\kappa}_n}{\lambda_n} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{\infty}\frac{\xi_{ik}\xi_{in}}{\lambda_n}\beta_n\right) + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\frac{\xi_{in}}{\lambda_n}.$$

Let's consider the variance of this estimator. The two terms are uncorrelated (because $\mathbb{E}[\varepsilon_i] = 0$ and ε_i is uncorrelated with X_i).

Thus variance of second term gives lower bound:

$$\begin{aligned}\text{var}\left[\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\frac{\xi_{in}}{\lambda_n}\right] &= \frac{1}{N\lambda_n^2}\left\{\mathbb{E}[\varepsilon_i^2\xi_{in}^2] - \mathbb{E}^2[\varepsilon_i\xi_{in}]\right\} = \frac{1}{N\lambda_n^2}\mathbb{E}[\varepsilon_i^2]\mathbb{E}[\xi_{in}^2] \\ &= \frac{\sigma^2\lambda_n}{N\lambda_n^2} = \frac{\sigma^2}{N\lambda_n}.\end{aligned}$$

In conclusion:

$$\text{var}[\hat{\beta}_n] \geq \frac{\sigma^2}{N\lambda_n}$$

while $\lambda_n^{-1} \to \infty$, since $\sum_k \lambda_k < \infty$...

- PCA can be used to to interpret which features of β can be estimated well, and to what level of precision.
- It can also be used implicitly to obtain consistent estimators.

The plug-in approach: if we have estimator $\hat{\mathcal{K}}$ of $\mathcal{K}$, then we can

$$\text{estimate the eigenpairs of of } \mathcal{K} \text{ by those of } \hat{\mathcal{K}}$$

- There is a zoo of possible observation scenarios, and a corresponding zoo of estimators.
- If we can get perturbation bounds, then performance of plug-in estimators can be gauged by performance of $\mathcal{K}$ itself.
- Viewing the eigenpairs as functionals of the corresponding covariance, their degree of regularity will yield correspondingly coarse/refined performance guarantees:
  - Contituiny will yield consistency.
  - Lipshitz/Hölder continuity of functionals, will yield rates of convergence.
  - (Fréchet) differentiability, will yield a central limit theorem (in the Platonic case, anyway)
- This leads us to consider perturbation bounds.

Theorem (von Neumann's trace inequality)

Let $\mathcal{A}, \mathcal{B} : \mathcal{H}_1 \to \mathcal{H}_2$ be Hilbert-Schmidt operators between separable Hilbert spaces. Then

$$|\langle \mathcal{A}, \mathcal{B} \rangle| \equiv |\operatorname{tr}\{\mathcal{A}\mathcal{B}^*\}| \leq \sum_{n \geq 1} \sigma_n(\mathcal{A})\sigma_n(\mathcal{B}),$$

where $\sigma_j(\cdot)$ denotes to the jth singular value (always non-negative by convention).

In fact, equality holds if and only if the two operators commute.

Proof.

Since the two operators are compact, they admit SVDs. Take an ONB for $\mathcal{H}_1$ by extending the left singular vectors of $\mathcal{A}$, and an ONB for $\mathcal{H}_2$ by extending the right singular vectors of $\mathcal{B}$. By the isometric isomorphism that identifies each of these bases to the canonical basis $\{e_j\}$ of ℓ_2, we reduce the statement to the setting:

- $\mathcal{H}_1 = \mathcal{H}_2 = \ell_2$, $\mathcal{A} = I\Lambda U$, $\mathcal{B} = V\Omega I$
- Λ, Ω are diagonal square-summable infinite arrays
- $I = \{\delta_{ij}\}$ is the identity array.
- U, V are orthogonal infinite arrays ($U^\top U = V^\top V = I$)

and we need to show that: $|\operatorname{trace}\{\Lambda U \Omega V^\top\}| \leq \operatorname{trace}\{\Lambda\Omega\}$.

We express Λ and Ω as weighted averages of the projectors $P_k = \sum_{i=1}^{k} e_i e_i^\top$, with $\{e_i\}$ the canonical basis of ℓ_2:

$$\Lambda = (\lambda_1 - \lambda_2)P_1 + (\lambda_2 - \lambda_3)P_2 + \ldots + (\lambda_{p-1} - \lambda_p)P_{p-1} + \ldots = \sum_{i=1}^{\infty} \alpha_i P_i$$

$$\Omega = (\omega_1 - \omega_2)P_1 + (\omega_2 - \omega_3)P_2 + \ldots + (\omega_{p-1} - \omega_p)P_{p-1} + \ldots = \sum_{i=1}^{\infty} \beta_i P_i$$

where the telescoping series obviously converges in Hilbert-Schmidt norm. With this representation, our sought inequality becomes

$$\left| \sum_{i,j} \alpha_i \beta_j \mathrm{tr}\{P_i U P_j V^\top\} \right| \leq \sum_{i,j} \alpha_i \beta_j \mathrm{tr}\{P_i P_j\}.$$

This will follow by the triangle inequality if we can bound each term as

$$|\alpha_i \beta_j \mathrm{tr}\{P_i U P_j V^\top\}| \leq \alpha_i \beta_j \mathrm{tr}\{P_i P_j\},$$

To show this, take $i \geq j$ (say), so the RHS is $\alpha_i \beta_j \times j$. Letting u_k be the jth column of U,

$$P_i U P_j = (P_i u_1, P_i u_2, ..., P_i u_j, 0, ...)$$

So it suffices to show $\sum_{k=1}^{j} \langle P_i u_k, v_k \rangle \leq j$. But this follows from the Cauchy-Schwarz inequality since $\|P_i u_k\| \leq \|P_i\|_\infty \|u_k\| \leq \|u_k\| = 1$. □

Corollary (Singular Value Perturbation Bound)

Let $\mathcal{A}, \mathcal{B} : \mathcal{H}_1 \to \mathcal{H}_2$ be Hilbert-Schmidt operators between separable Hilbert spaces. Then,

$$\sup_j |\sigma_j(\mathcal{A}) - \sigma_j(\mathcal{B})| \leq \sqrt{\sum_j |\sigma_j(\mathcal{A}) - \sigma_j(\mathcal{B})|^2} \leq \|\mathcal{A} - \mathcal{B}\|_{HS} \leq \|\mathcal{A} - \mathcal{B}\|_{\mathrm{tr}}.$$

Proof.

The first and last inequality are immediate. For the second inequality, we open the square and use von Neumann's trace inequality:

$$\begin{aligned}
\|\mathcal{A} - \mathcal{B}\|_{HS}^2 &= \|\mathcal{A}\|_{HS}^2 + \|\mathcal{B}\|_{HS}^2 - 2\mathrm{tr}(\mathcal{A}\mathcal{B}^*) \geq \|\mathcal{A}\|_{HS}^2 + \|\mathcal{B}\|_{HS}^2 - 2|\mathrm{tr}(\mathcal{A}\mathcal{B}^*)| \\
&\geq \sum_j \sigma_j^2(\mathcal{A}) + \sum_j \sigma_j^2(\mathcal{B}) - 2\sum_j \underbrace{\sigma_j(\mathcal{A})\sigma_j(\mathcal{B})}_{\geq 0} \\
&= \sum_j (\sigma_j(\mathcal{A}) - \sigma_j(\mathcal{B}))^2
\end{aligned}$$

□

Eigenvectors require a little more work:

Theorem (Eigenvector Perturbation Bound)

Let $\mathcal{K} \succ 0$ and $\hat{\mathcal{K}} \succ 0$ be trace-class on a separable Hilbert space, with eigenpairs (λ_j, u_j) and $(\hat{\lambda}_j, \hat{u}_j)$, respectively, both with distinct eigenvalues. Define $u_j^* = \mathrm{sign}\{\langle u_j, \hat{u}_j \rangle\} u_j$. Then,

$$\|\hat{u}_j - u_j^*\| \leq 2\sqrt{2}\alpha_j \left\|\hat{\mathcal{K}} - \mathcal{K}\right\|_{HS} \leq 2\sqrt{2}\alpha_j \left\|\hat{\mathcal{K}} - \mathcal{K}\right\|_{\mathrm{tr}},$$

where $\alpha_1 = (\lambda_1 - \lambda_2)^{-1}$ and $\alpha_j = \max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\}$, $j \geq 2$.

- Distinct eigenvalues allow for individual eigendirections to be identifiable.
- But eigenvectors are unique only up to a sign change, hence the use of u_j^*

Proof

The second inequality on the RHS is immediate, so we focus on the first.

$$\mathcal{K}\hat{u}_j - \lambda_j \hat{u}_j = (\mathcal{K} - \hat{\mathcal{K}} + \hat{\mathcal{K}})\hat{u}_j - (\lambda_j - \hat{\lambda}_j + \hat{\lambda}_j)\hat{u}_j = (\mathcal{K} - \hat{\mathcal{K}})\hat{u}_j + (\hat{\lambda}_j - \lambda_j)\hat{u}_j$$

Thus, the triangle inequality and the second inequality in the last corollary imply that

$$\|\mathcal{K}\hat{u}_j - \lambda_j \hat{u}_j\| \leq \|(\mathcal{K} - \hat{\mathcal{K}})\hat{u}_j\| + \|(\hat{\lambda}_j - \lambda_j)\hat{u}_j\| \leq \left\|\mathcal{K} - \hat{\mathcal{K}}\right\|_{\infty} + \left\|\mathcal{K} - \hat{\mathcal{K}}\right\|_{HS}$$

and since $\left\|\mathcal{K} - \hat{\mathcal{K}}\right\|_{\infty} \leq \left\|\mathcal{K} - \hat{\mathcal{K}}\right\|_{HS}$ the RHS is majorised by $2\left\|\mathcal{K} - \hat{\mathcal{K}}\right\|_{HS}$.

Now for all $1 \leq j \leq p$ we aim to lower bound $||\mathcal{K}\hat{u}_j - \lambda_j \hat{u}_j||^2$ below by $(2\alpha_j^2)^{-1}||u_j^* - \hat{u}_j||^2$.

$$||\mathcal{K}\hat{u}_j - \lambda_j \hat{u}_j||^2 = \sum_k \langle \mathcal{K}\hat{u}_j - \lambda_j \hat{u}_j, u_k \rangle^2 = \sum_k (\langle \mathcal{K}\hat{u}_j, u_k \rangle - \langle \lambda_j \hat{u}_j, u_k \rangle)^2$$

$$= \sum_k (\lambda_k - \lambda_j)^2 \langle \hat{u}_j, u_k \rangle^2 \geq \min_{k \neq j} (\lambda_k - \lambda_j)^2 \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 \geq \alpha_j^{-2} \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2$$

Recalling that $u_j^* = \text{sign}\{\langle u_j, \hat{u}_j \rangle\} u_j$, observe that $||u_j^* - \hat{u}_j||^2$ can be written as

$$\sum_k \langle u_j^* - \hat{u}_j, u_k \rangle^2 = \{\text{sign}(\langle u_j^*, u_j \rangle) - \langle \hat{u}_j, u_j \rangle\}^2 + \sum_{k \neq j} \langle u_j^* - \hat{u}_j, u_k \rangle^2$$

$$= \{1 - |\langle \hat{u}_j, u_j \rangle|\}^2 + \sum_{k \neq j} (\langle u_j^*, u_k \rangle - \langle \hat{u}_j, u_k \rangle)^2 = \{1 - |\langle \hat{u}_j, u_j \rangle|\}^2 + \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2$$

Since $\sum_{k=1} \langle \hat{u}_j, u_k \rangle^2 = 1$,

$$\{1 - |\langle \hat{u}_j, u_j \rangle|\}^2 = \sum_{k=1} \langle \hat{u}_j, u_k \rangle^2 - 2|\langle \hat{u}_j, u_j \rangle| + |\langle \hat{u}_j, u_j \rangle|^2$$

$$= \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 + 2\underbrace{\{|\langle \hat{u}_j, u_j \rangle|^2 - |\langle \hat{u}_j, u_j \rangle|\}}_{\leq 0} \leq \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2$$

because $\langle \hat{u}_j, u_j \rangle \leq 1$. Thus $2\sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 \geq ||u_j^* - \hat{u}_j||^2$.

Combining the inequalities in blue, and re-arranging the constant factors, we arrive at

$$4|||\hat{\mathcal{K}} - \mathcal{K}|||_{HS}^2 \geq ||\mathcal{K}\hat{u}_j - \lambda_j \hat{u}_j||^2 \geq \alpha_j^{-2} \sum_{k \neq j} \langle \hat{u}_j, u_k \rangle^2 \geq (2\alpha_j^2)^{-1}||u_j^* - \hat{u}_j||^2$$ □

- Note that the derivation of the two bounds do not have any subtleties related to infinite dimensions.
- However the interpretation of the eigenvector perturbation bound does:
  - We get a uniform bound when $\dim(\mathcal{H}) < \infty$.
  - But the factor on the RHS blows up when $\dim(\mathcal{H}) = \infty$.

A crude remedy is given by the following corollary:

Corollary (uniform eigenvector perturbation bound)

In the same context as in the previous theorem, assume that $\lambda_k = \varphi(k)$ for some convex function φ on $[0, +\infty)$. Then,

$$\sup_{j \leq k_n} \|\hat{u}_j - u_j^*\| \to 0 \qquad \text{as} \quad \left\|\hat{\mathcal{K}} - \mathcal{K}\right\|_{HS} \to 0$$

provided

$$\alpha_{k_n}^{-1} \equiv (\lambda_{k_n+1} - \lambda_{k_m})^{-1} = o\left(\left\|\hat{\mathcal{K}} - \mathcal{K}\right\|_{HS}\right).$$

Let's reflect on the result when estimating the covariance by its empirical version:

- With a sample of size n we can hope to estimate up to n eigenvectors.
- But the worst case error for the top n eigenvectors grows too fast with n, certainly much faster than our $\sqrt{n}$ perturbation bound:

$$(\lambda_n - \lambda_{n+1})^{-1} \quad \text{vs} \quad 1/\sqrt{n}$$

- Since $\mathcal{K}$ is trace class, λ_n goes to zero faster than $1/n$, and even then

$$\left(n^{-1} - (n+1)^{-1}\right)^{-1} \sim n^2 \quad \text{vs} \quad 1/\sqrt{n}$$

- For Brownian motion, λ_n goes like $1/n^2$, so

$$\left(n^{-2} - (n+1)^{-2}\right)^{-1} \sim n^3 \quad \text{vs} \quad 1/\sqrt{n}$$

- In other words, if we want to estimate n eigenvectors, we would need our sample size to be orders of magnitude large than n.
- Equivalently, need to "slow down" how many eigenvectors we expect to estimate uniformly well for a sample of size n, and take only k_n.
- The smaller the effective rank the slower we need to go beyond k_n the bound cannot guarantee that we can distinguish between actual modes of variation.
- Of course these are asymptotics. A finite portion of eigenvalues can feature "nice gaps". Luckily empirical eigenvalues are good estimators of true ones (uniformly). Importance of the scree plot...

With "Platonic" case, the rate n^{-1} is obviously optimal for the MSE.

Under partial observation,

- In the irregular and sparse regime, Hall, Müller & Wang (2006) get an optimal rate of

$$n^{-\frac{2r}{2r+1}}$$

when the sample paths are assumed C^r.

- In the regular regime, based on a regular grid of p points, Belhakem et al (2021) get an optimal rate of

$$n^{-1} + p^{-2\alpha}$$

assuming an $\alpha-$Hölder covariance ($0 \leq \alpha \leq 1$), i.e. $\frac{\alpha}{2} - \epsilon$ Hölder paths.

This yields the parametric rate when measurements are sufficiently dense,

$$p \gtrsim n^{\frac{1}{2\alpha}}.$$

For a CLT, we need to show that the eigenpairs are obtained as *smooth transformations* of the underlying covariance, and use the delta method.

Here is a summary of the basic idea when $\lambda_k > \lambda_{k+1}$, for all $k \geq 1$:

1. The triple $\{\lambda_k, e_k, \mathcal{K}\}$ satisfies the equation

$$F(\lambda_k, e_k, \mathcal{K}) = 0$$

where

$$F : \mathbb{R} \times \mathcal{H} \times \mathbb{B}_{HS}(\mathcal{H}) \to \mathcal{H} \times \mathbb{R}, \quad F(\alpha, u, \mathcal{X}) = \begin{pmatrix} (\alpha \mathcal{I} - \mathcal{X})u \\ \langle u, u \rangle - 1 \end{pmatrix}.$$

2. We can verify that this map is continuously Fréchet differentiable with non-vanishing Jacobian when λ_k is simple.
3. Thus we can make use of the Banach implicit function theorem: there is an open set $U \ni \mathcal{K}$, and unique implicit functions $\sigma_k(\cdot) : U \to \mathbb{R}$ and (up to sign) $v_k : U \to \mathcal{H}$ such that
   1. $\sigma_k(\cdot)$ and $v_k(\cdot)$ are continuously differentiable.
   2. $\sigma_k(\mathcal{K}) = \lambda_k$ and $v_k(\mathcal{K}) = e_k$
   3. $\mathcal{X} v_k(\mathcal{X}) = \sigma_k(\mathcal{X}) v_k(\mathcal{X})$ & $\langle v_k(\mathcal{X}), v_k(\mathcal{X}) \rangle = 1$ for all $\mathcal{X} \in U$.

- Via the delta method, these implicit functions now yield $\sqrt{n}$-CLTs for the kth eigenpair.
- Note that the logic in the argument can apply simultaneously to any finite set of eigenpairs, yielding the corresponding joint asymptotic law.
- The exact for of the asymptotic covariance follows from the form of the derivatives of the implicit functions (which can actually be determined).
- Two interesting outcomes of this calculation are:
  1. In the Gaussian case, empirical eigenvalues are asymptotically independent.
  2. Empirical eigenvectors are dependent, even asymptotically (and how could it be otherwise...)
- Apart from that the specific covariance is not of much use, as it depends on the true spectrum.
- The Gaussianity of the limit can be useful, however, to justify bootstrap procedures for inferences on (finitely many) eigenpairs.
- Recall that all this requires a "base case CLT" for the covariance estimator itself – e.g. for the empirical covariance in the Platonic case.

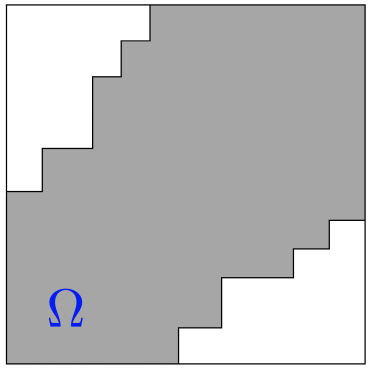
(Wish) List of Topics

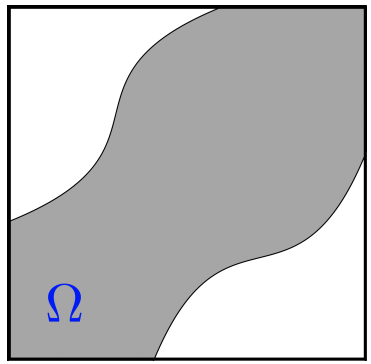
1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

Let $K_\Omega(s,t) : \Omega \to \mathbb{R}$ be a partial covariance kernel[4] on a domain

$$\Omega = \bigcup_{a \in A} (I_a \times I_a)$$

for a (possibly uncountable) cover $\{I_a\}_{a \in A}$ of $[0,1]$.





We consider the following problem:

How can $K_\Omega(s,t)$ be completed to a covariance kernel $K(s,t)$ on $[0,1]^2$?

[4] i.e. $\forall\ I \times I \subset \Omega$, the restriction $K_\Omega|_{I \times I}$ is a covariance kernel

We can refine or vary the question:

- Do there always exist completions? How many?
- Is there canonical choice among them? Is it constructible?
- Is a unique completion necessarily canonical?
- Can we find necessary and sufficient conditions for unique completion?
- Can we constructively characterise all completions?
- How do completions vary when we perturb K_Ω? (estimation)
- How do these questions relate to a process $\{X(t) : t \in [0,1]\}$ such that

$$\mathrm{Cov}\{X(u), X(v)\} = K_\Omega(u,v), \qquad (u,v) \in \Omega.$$

1. Analysis/Probability: continuation of positive definite *functions*

   is a p.d. function ϕ determined by its restriction on $(-\delta, \delta)$?

   Equivalent to our problem in stationary case,

   $$K_\Omega(u, v) = \phi(u - v), \quad \Omega = \{|u - v| < \delta\}$$

   Related to

   - truncated (trigonometric) moment problem
   - continuation of characteristic functions
   - Major results by Carathéodory, Gnedenko, Gneiting, Esseen, Krein, Calderon, Rudin...

2. Matrix Algebra and Statistics: non-negative *matrix completion*

   - Key results by Gohberg, Johnson, Dempster...

Arises naturally in FDA:

Covariance Recovery from Sample Path Fragments

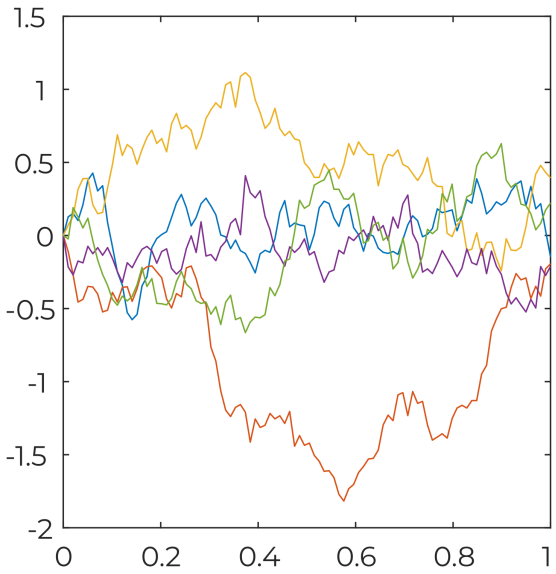
- I_1, I_2, I_3,... a sequence of intervals eventually covering $[0, 1]$
- $\{X_1(u), ..., X_n(u)\}$ are independent sample paths of a Gaussian process X
- Observe

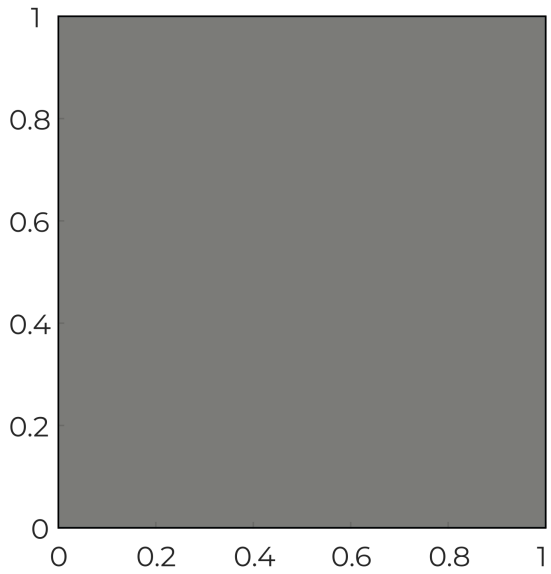
$$X_1|_{I_1} \ , \ldots , \ X_n|_{I_n}$$

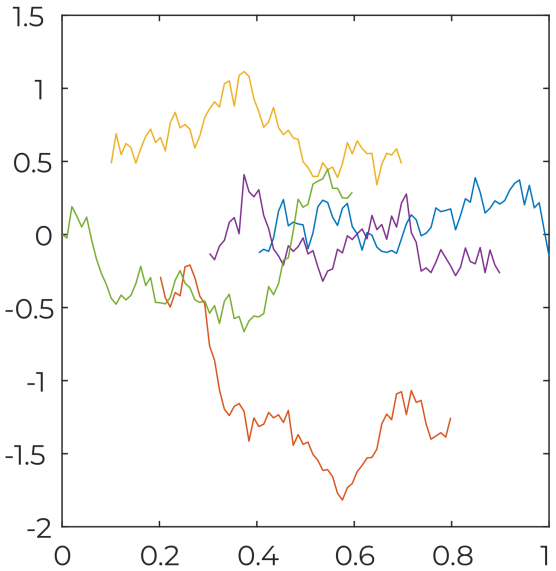
Can we estimate $K = \text{Cov}\{X(u), X(v)\}$ on $[0, 1]^2$ when $|I_i| < 1$, $\forall\ i \geq 1$?

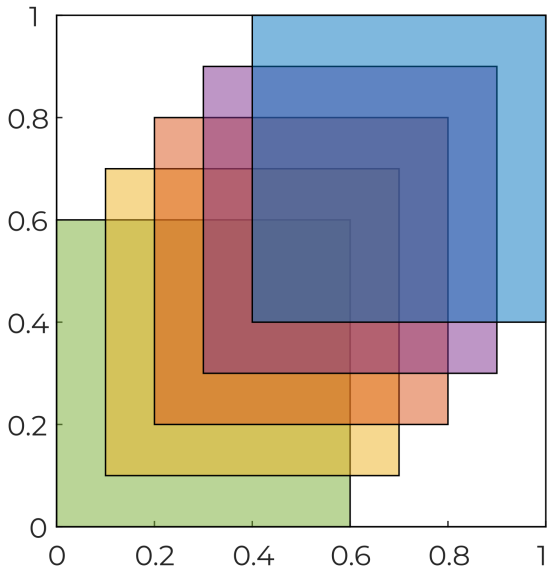
Descary & Panaretos (2019)
Delaigle et al. (2020)
Lin et al (2020)
Kneip & Leibl (2020)
$\vdots$

Recovering Covariance from Sample Path Fragments

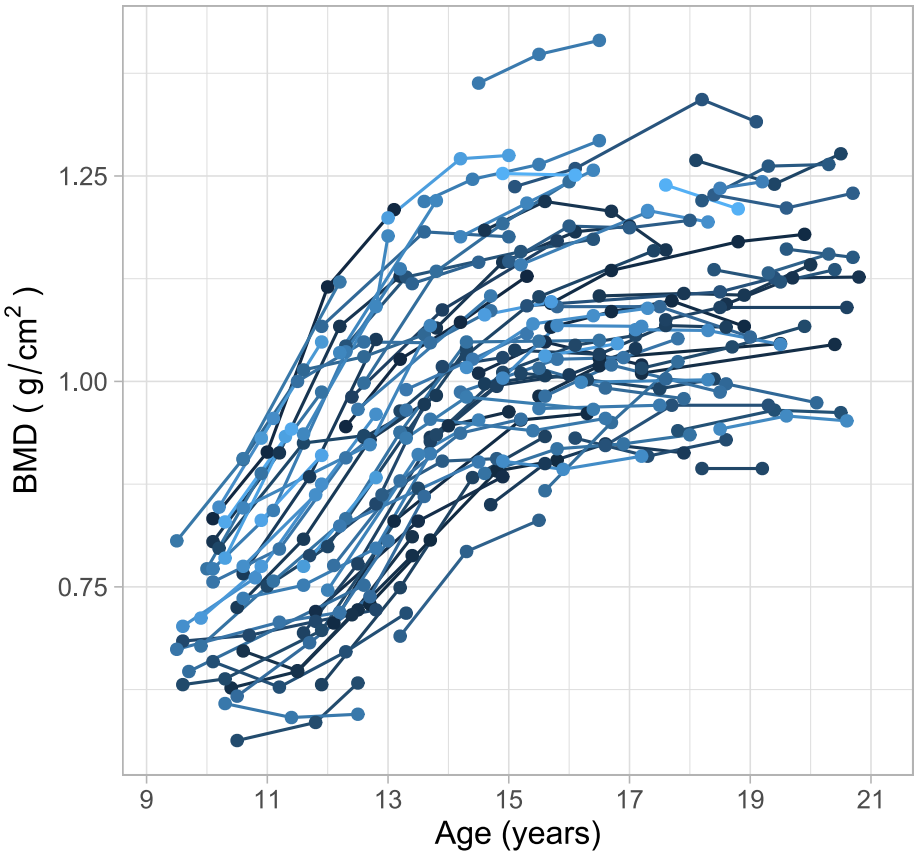


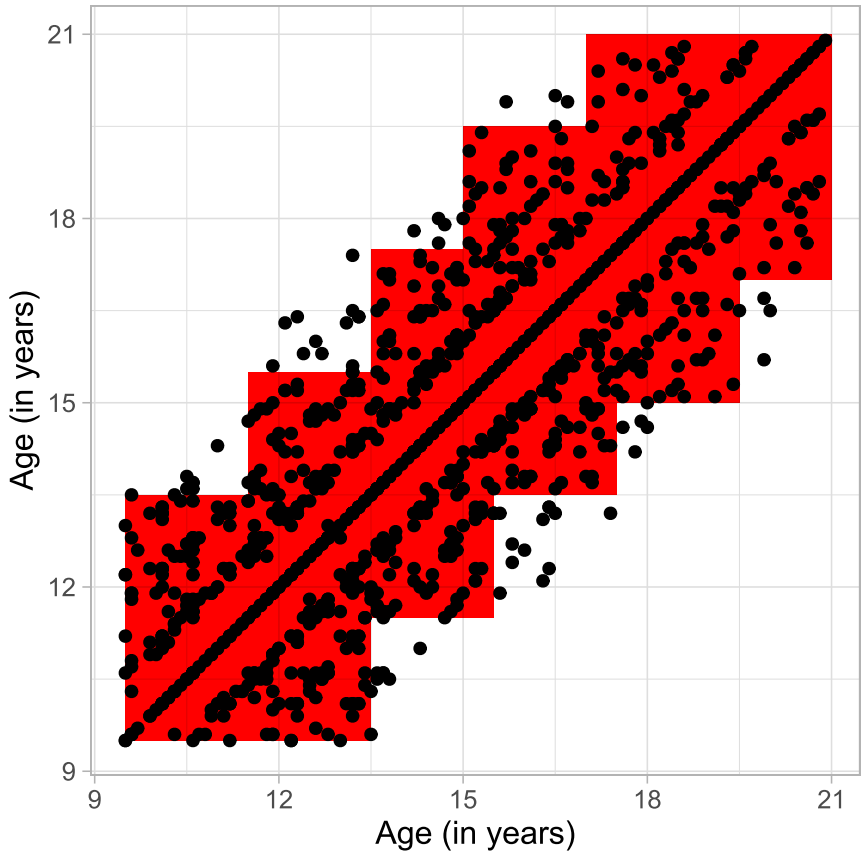






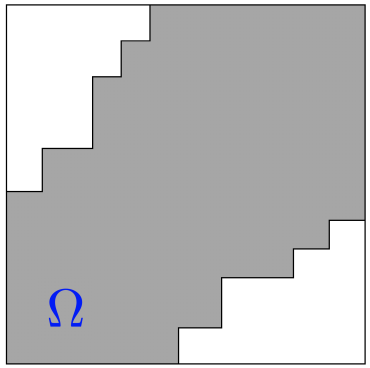
Example: Bone Mineral Density

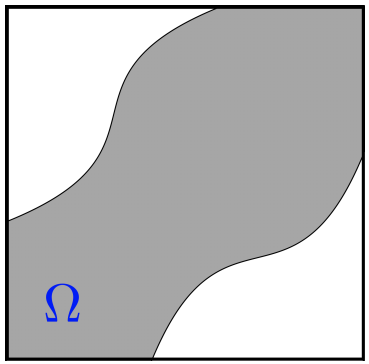




BMD measurements for 117 females taken between the ages of 9.5 and 21 years

Back to our completion problem





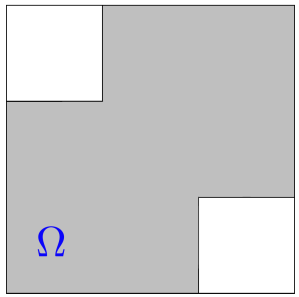
How can $K_\Omega(s,t)$ be completed to a covariance kernel $K(s,t)$ on $[0,1]^2$?

Define the set of completions as

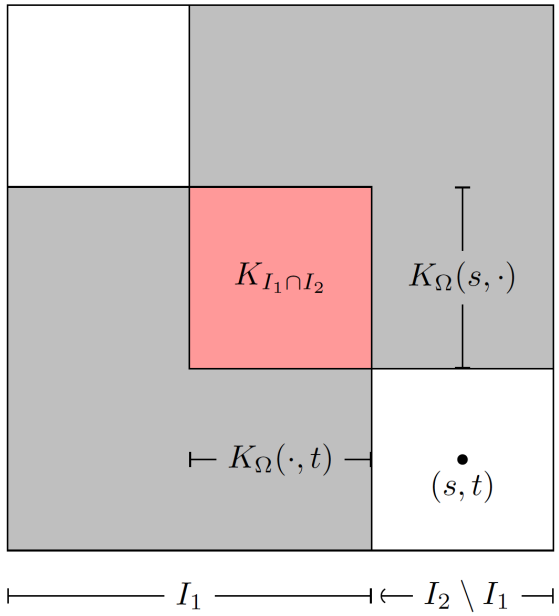
$$\mathcal{C}(K_\Omega) = \{K \succeq 0 \text{ on } [0,1]^2 : K|_\Omega = K_\Omega\}.$$

- Previous work focusses on sufficient conditions for $|\mathcal{C}(K_\Omega)| = 1$.
- We wish to comprehensively understand the set $\mathcal{C}(K_\Omega)$

Let's start with the simples case: the **2-serrated** case.



$$\Omega = (I_1 \times I_1) \cup (I_2 \times I_2) \quad \text{with} \quad I_1 = [0, b],\ I_2 = [a, 1]\ a \leq b.$$



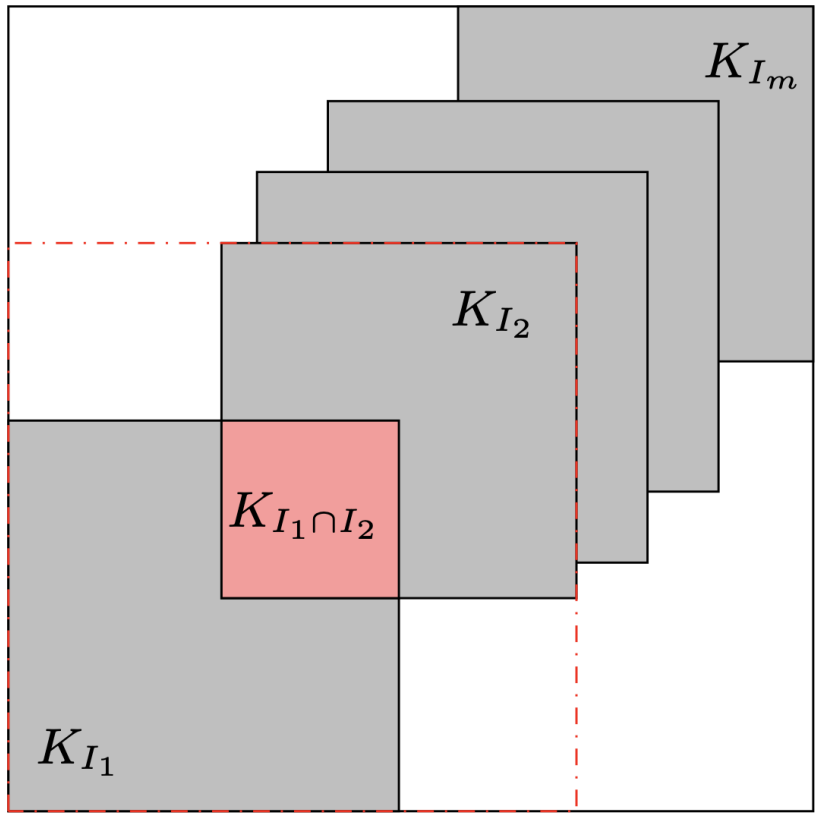
Define $K_\star : [0,1]^2 \to \mathbb{R}$ as

$$K_\star(s,t) = \begin{cases} K_\Omega(s,t), & (s,t) \in \Omega \\ \left\langle K_\Omega(s,\cdot), K_\Omega(\cdot,t) \right\rangle_{\mathcal{H}(K_{I_1 \cap I_2})}, & (s,t) \notin \Omega \end{cases}$$

where $\mathcal{H}(C)$ denotes the RKHS of a covariance C.

Proposition (Waghmare & Panaretos, 2022)

$K_\star$ is a bona fide covariance and $K_\star \in \mathcal{C}(K_\Omega)$.



Theorem (Waghmare & Panaretos, 2022)

Recursive application of the 2-serrated formula yields a valid completion $K^* \in \mathcal{C}(K_\Omega)$, indeed the same completion irrespective of the order it is applied in.

As an example, let $I = [0, 1]$

$$K_\Omega(s,t) = s \wedge t, \quad (s,t) \in \Omega = \underbrace{([0, 2/3] \times [0, 2/3])}_{I_1} \cup \underbrace{([1/3, 1] \times [1/3, 1])}_{I_2}.$$

Clearly, this can be completed to the covariance of standard Brownian motion,

$$K(s,t) = s \wedge t, \qquad (s,t) \in [0,1]^2.$$

- In this case the RKHS is explicitly known to be a Sobolev space
- Can calculaate the two-serrated completion explicitly.
- Turns out to coincide with the standard BM kernel.
- Hence, by iterating, any m-serrated completion is the BM kernel
- So the completion method seems to give the "right" answer in some standard examples (more can be produced)

Theorem (Waghmare & Panaretos, 2022)

The covariance K^* is the only completion of K_Ω such that the associated Gaussian process forms an undirected graphical model w.r.t. $G = ([0,1], \Omega)$

- Equivalently K^* is the unique extension w/ the global Markov property w.r.t. edge set Ω
- Intuitively, relies exclusively on correlations intrinsic to Ω — propagates only "observed" correlations via the Markov property, without introducing arbitrary unseen correlations.
- It is unique in doing so among all possible completions
- Later shown in Waghmare & Panaretos (2024) that perturbations of $K_\star$ decrease the (Fredholm) determinant.

For all these reasons:

We call the completion K^* the *canonical completion*.

Theorem (Waghmare & Panaretos, 2022)

Let K_Ω be a partial covariance kernel on a serrated domain $\Omega = \bigcup_{j=1}^{m}(I_m \times I_m)$

The following three statements are equivalent:

1. K_Ω admits a unique completion K on $[0,1]^2$, i.e. $\mathcal{C}(K_\Omega)$ is a singleton.

2. if $X_j \sim N(0, K_\Omega|_{I_j \times I_j})$, then there exists $r \in \{1, ..., m\}$ such that

$$X_j = \mathcal{A}_j X_r$$

for $m-1$ deterministic linear maps $\{\mathcal{A}_j\}_{j \neq r}$.

3. for some $r \in \{1, \ldots, m\}$, the following Schur complements[a] vanish:

$$K_{I_p}/K_{I_p \cap I_{p+1}} = 0, \text{ for } 1 \leq p < r \quad \text{and} \quad K_{I_{q+1}}/K_{I_q \cap I_{q+1}} = 0, \text{ for } r \leq q < m.$$

[a] For $A \subset B \subset \Omega$, $K_A,(K_B/K_A)(s,t) = K_B(s,t) - \left\langle K_B(s,\cdot), K_B(\cdot,t)\right\rangle_{\mathcal{H}(K_A)}$ i.e. the covariance of the *residuals* $\{X_t - \Pi(X_t|X_A) : t \in B \setminus A\}$.

- Condition (2) implies that $X(t) = \Pi[X(t)|\{X(s) : t \in I_r\}]$ for one of the intervals I_r defining the serrated domain.

- So when unique completion is possible, the process $\{X(t) : t \in [0,1]\}$ is a **deterministic linear transformation** of its restriction $\{X(t) : t \in I_r\}$ to one of the intervals I_r defining the serrated domain.

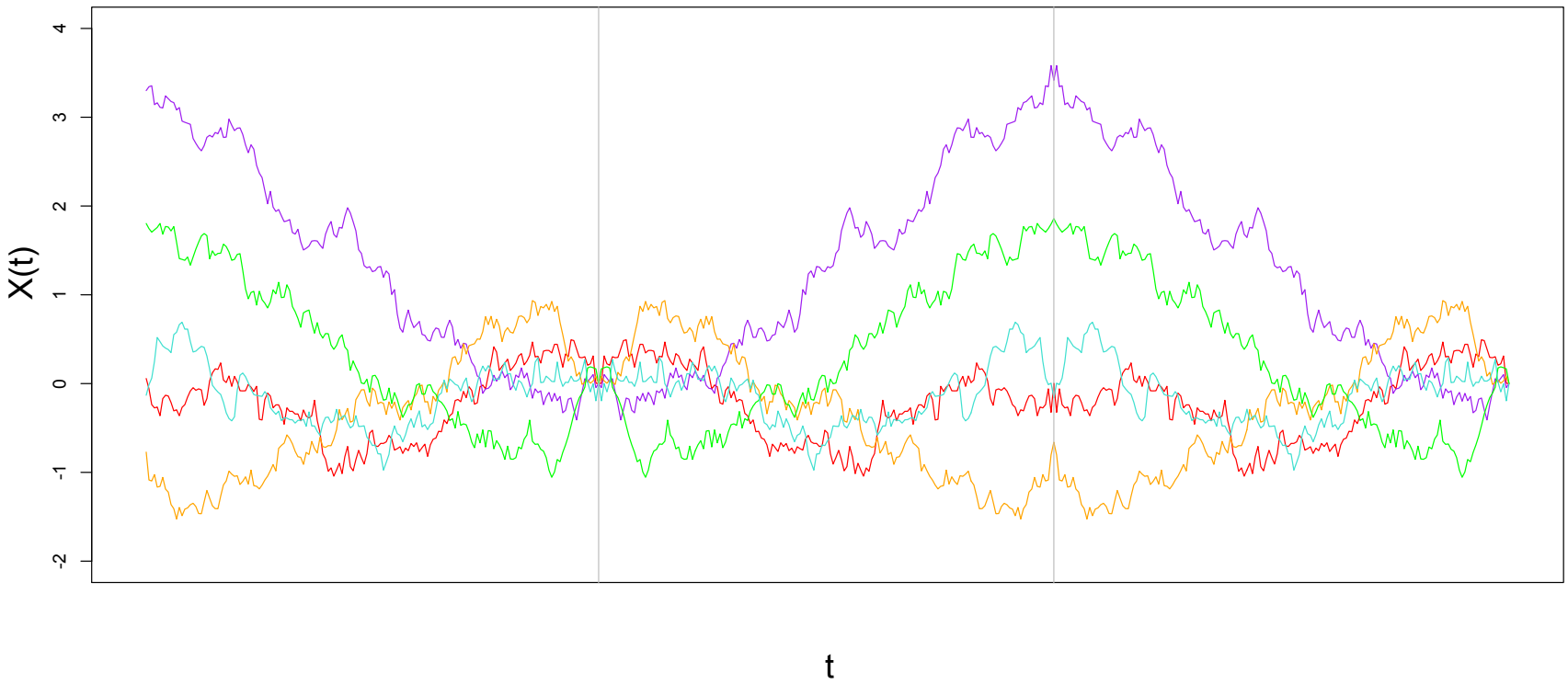
- In any case, when a unique completion exists, it must be the canonical one.

- Condition (3) is checkable at the level of K_Ω, i.e. at the level of observables

- It has nothing to do with smoothness or finite rank assumptions (see next slide).

- Notice that *identifiability* of K from $K|_\Omega$ does not *require* unique completion conditions on $K|_\Omega$ – can assume Ω-Markovianity (a very considerably weaker assumption)

The following "left/right reflection" example shows that dimensionality or smoothness play no role:



Let $I_1 = [0, 2/3]$, $I_2 = [1/3, 1]$ and $\Omega = I_1^2 \cup I_2^2$ be a 2-serrated domain. Define

$$X(t) = B(2/3 - t)\mathbf{1}\{t \in I_1 \setminus I_2\} + B(t)\mathbf{1}\{t \in I_1 \cap I_2\} + B(4/3 - t)\mathbf{1}\{t \in I_2 \setminus I_1\}$$

with $\{B(t) : t \in [1/3, 2/3]\}$ a standard Brownian motion on $I_1 \cap I_2$.

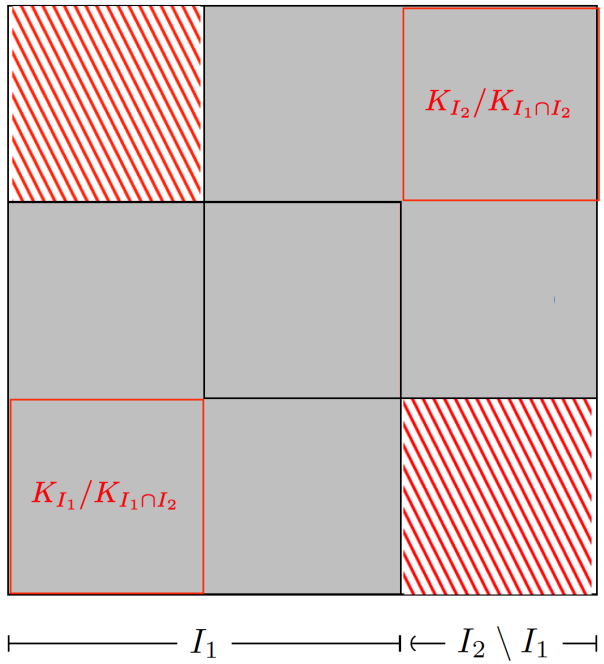
Theorem (Waghmare & Panaretos, 2022)

K is a completion of K_Ω if and only if

$$K = K_\star + C$$

where C is a valid cross-covariance between

$$X_1 \sim N(0, K_{I_1}/K_{I_1 \cap I_2}) \quad \text{and} \quad X_2 \sim N(0, K_{I_2}/K_{I_1 \cap I_2})$$



- Valid C are easily but arbitrarily obtained:
  - *Any* coupling of $X_1 \sim N(0, K_{I_1}/K_{I_1 \cap I_2})$ and $X_2 \sim N(0, K_{I_2}/K_{I_1 \cap I_2})$ will yield valid cross-covariance $C(s,t) = \mathrm{cov}\{X_1(s), X_2(t)\}$
  - Like assigning a correlation to two variances – think of 3×3 matrices

$$\begin{pmatrix} \sigma_1^2 & * & ? \\ * & * & * \\ ? & * & \sigma_2^2 \end{pmatrix}$$

- Can characterise in operator notation – choose $\|\Psi\| \leq 1$ arbitrarily, then

$$\mathbf{K}f = \mathbf{K}_\star f + \underbrace{\left(\begin{array}{c|c|c} 0 & 0 & (\mathbf{L}_1^{1/2}\Psi\mathbf{L}_2^{1/2})^* \\ \hline 0 & 0 & 0 \\ \hline \mathbf{L}_1^{1/2}\Psi\mathbf{L}_2^{1/2} & 0 & 0 \end{array}\right)}_{\mathbf{C}} \begin{pmatrix} f|_{I_1 \setminus I_2} \\ f|_{I_1 \cap I_2} \\ f|_{I_2 \setminus I_1} \end{pmatrix}$$

- Any completion other than canonical one introduces arbitrary correlations
- Valid completions in bijection with $\|\cdot\|_\infty$-unit ball.
- Can build them all once we have $K_\star$

Everything in black depends only on K_Ω (equiv. on its canonical extension $K_\star$):

Theorem (Waghmare & Panaretos, 2022)

Let K_Ω be a continuous partial covariance on a serrated domain Ω of m intervals. Then K is a completion of K_Ω if and only if its operator $f \mapsto \mathcal{K}f$ has the form

$$\mathcal{K}f(t) = \sum_{j:t\in I_j} \mathcal{K}_j f_{I_j}(t) + \sum_{p:t\in S_p} \mathcal{R}_p f_{D_p}(t) + \sum_{p:t\in D_p} \mathcal{R}_p^* f_{S_p}(t) - \sum_{p:t\in I_p\cap I_{p+1}} \mathcal{J}_p f_{J_p}(t) \text{ a.e.}$$

where for $1 \le p < m$,

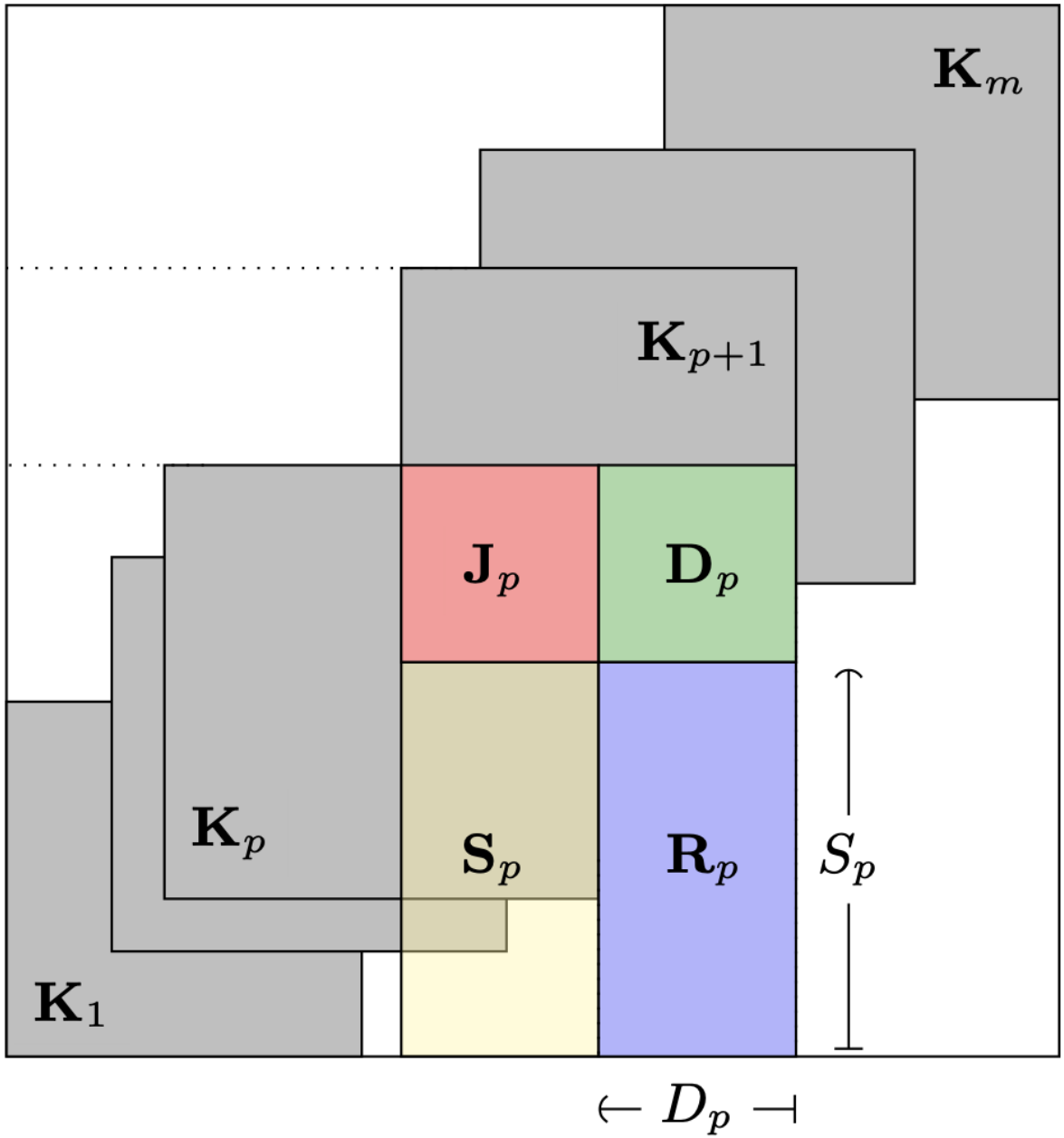
$$\mathcal{R}_p = \underbrace{\left[\mathcal{J}_p^{-1/2}\mathcal{S}_p^*\right]^* \left[\mathcal{J}_p^{-1/2}\mathcal{D}_p\right]}_{\text{w/ kernel } K_\star|_{R_p},\ \text{step } p \text{ of algorithm}} + \mathcal{U}_p^{1/2}\Psi_p\mathcal{V}_p^{1/2}$$

$$\mathcal{U}_p = \mathcal{K}_{S_p} - \left[\mathcal{J}_p^{-1/2}\mathcal{S}_p^*\right]^* \left[\mathcal{J}_p^{-1/2}\mathcal{S}_p^*\right], \qquad \mathcal{V}_p = \mathcal{K}_{D_p} - \left[\mathcal{J}_p^{-1/2}\mathcal{D}_p^*\right]^* \left[\mathcal{J}_p^{-1/2}\mathcal{D}_p^*\right]$$

and $\Psi_p : L^2(D_p) \to L^2(S_p)$ are bounded linear maps with $\|\Psi_p\|_\infty \le 1$.

Furthermore, taking $\Psi_1 = \Psi_2 = \ldots = \Psi_m = 0$ yields the canonical completion.

The Picture that Illustrates the Formula



Makes sense to choose canonical completion as target of estimation:

- When completion is unique, it will be canonical
- When completion non-unique, canonical completion is least presumptuous
- Canonical completion is pivot to construct all completions

$\Longrightarrow$ It is always an identifiable and interpretable target of estimation

Estimating specifically the canonical completion is qualitatively different under non-uniqueness than all previous approaches (which focussed on uniqueness)

1. If we impose uniqueness by way of assumption (a very strong assumption), then one can use, for example, series estimators or matrix completion.
2. However such estimators will yield arbitrary (almost certainly non-canonical) completions if uniqueness does not actually hold.
3. To guarantee canonicity, we need to satisfy the system of operator equations on the previous slide – an inverse problem
4. Can be seen as an adaptive approach: will yield the unique completion when uniqueness holds, and a stable/canonical one otherwise.

Let $\widehat{K}_\Omega$ be an estimator of K_Ω.

Define $\widehat{K}_\star$ to be the estimator of $K_\star$ based on solving a regularised version of the linear operator system defining $K_\star$ (i.e. with all $\Psi_p = 0$), with $\widehat{K}_\Omega$ in lieu of K_Ω.

Regularisation by spectral truncation (at level N_p) of each of $p = 1, ..., m-1$ equations,

$$\mathcal{R}_p = \left[\mathcal{J}_p^{-1/2}\mathcal{S}_p^*\right]^* \left[\mathcal{J}_p^{-1/2}\mathcal{D}_p\right]$$

replacing unknown quantities with their "hat counterparts".

Let $A_{p,k}$ be the squared Hilbert-Schmidt error when approximating $\mathcal{R}_p = \left[\mathcal{J}_p{}^{-1/2}\mathcal{S}_p^*\right]^* \left[\mathcal{J}_p{}^{-1/2}\mathcal{D}_p\right]$ by replacing $\mathcal{J}_p$ with its rank-k truncation.

Theorem (Waghmare & Panaretos, 2022+)

Assume that for every $1 \leq p < m$, *we have*

- $\text{eigen}_k(K_\Omega|_{I_p \times I_p}) = \lambda_{p,k} \sim k^{-\alpha}$
- $A_{p,k} \sim k^{-\beta}$.

then

$$\|\widehat{K}_\star - K_\star\|_{L^2(I\times I)} = O_{\mathbb{P}}\left(\|\hat{K}_\Omega - K_\Omega\|^{\gamma_{m-1}}_{L^2(\Omega)}\right)$$

where

$$\gamma_{m-1} = \frac{\beta}{4\alpha+\beta+3}\left[\frac{\beta}{2\alpha+\beta+1}\right]^{m-2}, \qquad m > 1,$$

provided the regularisation parameters are chosen to satisfy

$$N_p \times \|\hat{K}_\Omega - K_\Omega\|^{2\gamma_p/\beta}_{L^2(\Omega)} = O_{\mathbb{P}}(1)$$

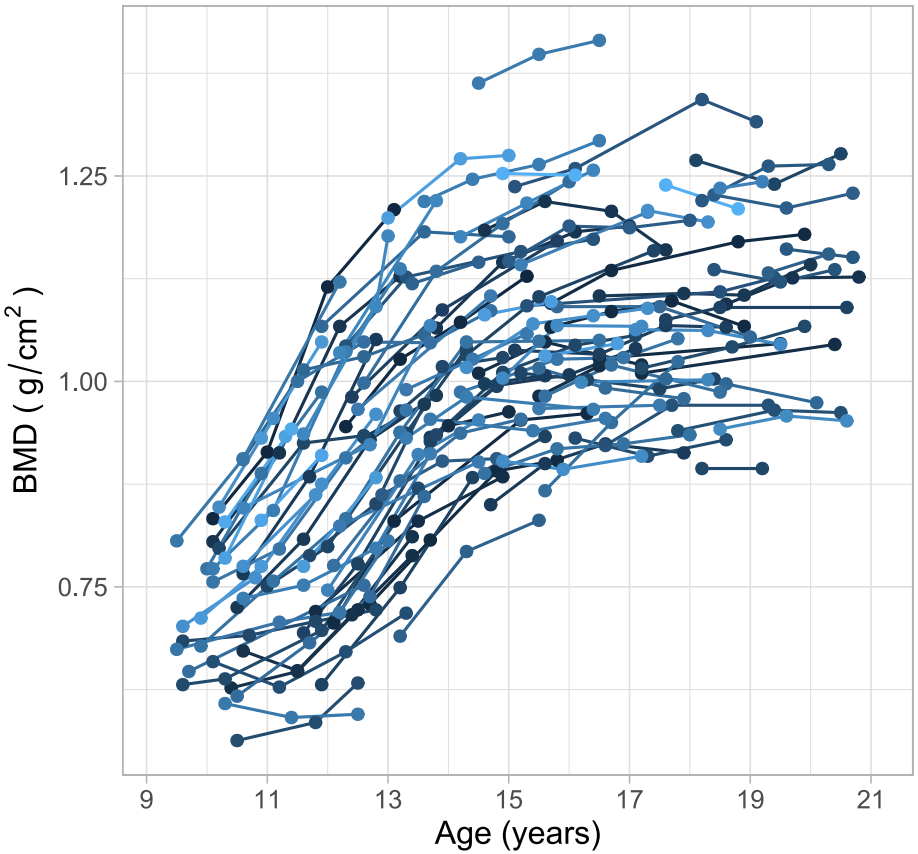
$$\|\widehat{K}_\star - K_\star\|_{L^2(I\times I)} \preceq \|\hat{K}_\Omega - K_\Omega\|^{\gamma_{m-1}}_{L^2(\Omega)}$$

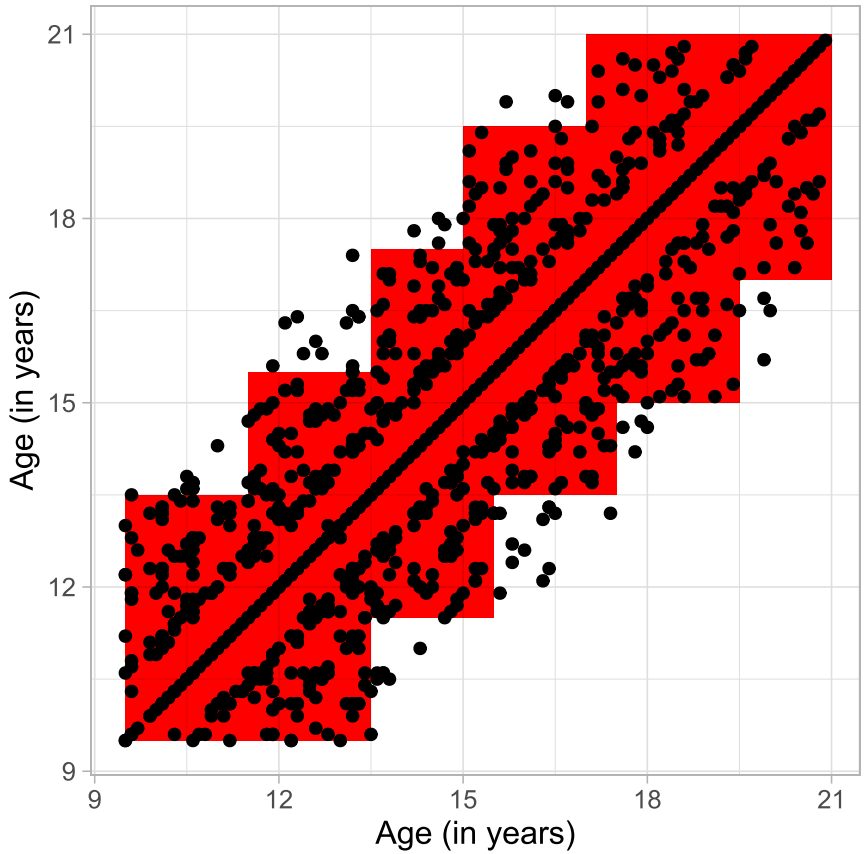
where $\gamma_{m-1} = \frac{\beta}{4\alpha+\beta+3}\left[\frac{\beta}{2\alpha+\beta+1}\right]^{m-2}$ for $m > 1$.

Remarks on the exponent γ_{m-1}:

- It strictly decreases as a function of the number of intervals m
- It can get arbitrarily close to 1 for a large enough rate of decay of approximation errors β.
- An increase in the rate of decay of eigenvalues α is accompanied by a decrease in the rate of convergence.
- If $K_\Omega \in C^r(\Omega)$ then the same applies to the kernels $K_\Omega|_{J_p\times J_p}$ of $\mathcal{J}_p$ implying $\lambda_{p,k}$ is $o(1/k^{r+1})$ for every $1 \leq p < m$ and thus $\alpha = r + 1$.
- All other things being equal, an increase in the smoothness of K_Ω also tends to a decrease in the rate of convergence

Example: Bone Mineral Density





BMD measurements for 117 females taken between the ages of 9.5 and 21 years

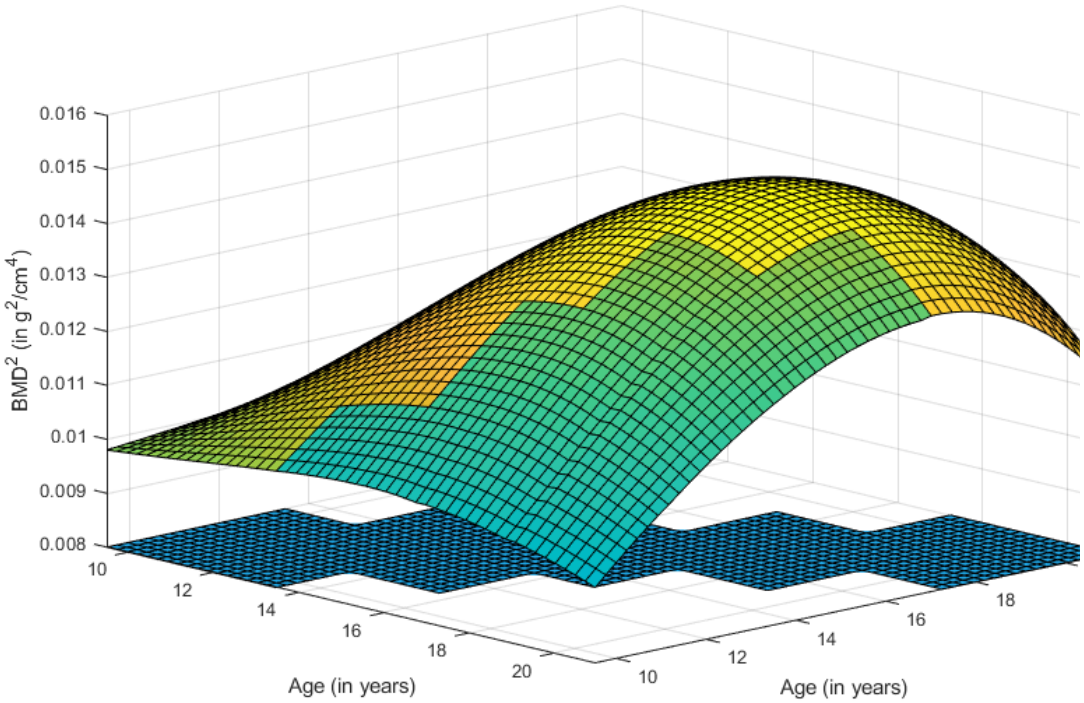
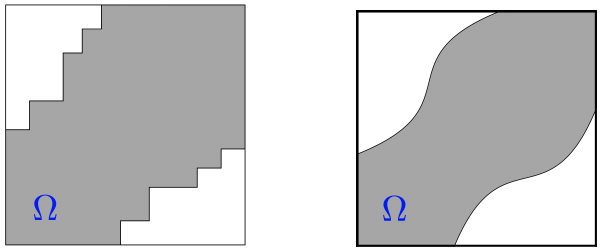


Figure: Completed covariance of the BMD data.



With some more effort, we also have

Theorem (Waghmare & Panaretos, 2024)

Let K_Ω be a continuous partial covariance kernel on a regular domain Ω. Then,

- There exists a canonical completion $K_\star$ of K_Ω.
- There exists an increasing sequence $\Omega_j \subset \Omega$ of serrated domains with $\cup_j \Omega_j = \Omega$, such that the canonical completions K_j of $K_\Omega \big|_{\Omega_j}$ converge pointwise to $K_\star$.

Leads to a novel result even in a classical context:

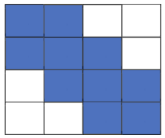
Corollary (Stationarity and ϵ-Markov extensions)

Any positive definite function on $[-\epsilon, \epsilon]$ admits a **canonical** extension to $(-\infty, \infty)$

(Wish) List of Topics

1. Reminder on Normed Vector Spaces
2. Bochner Integration
3. Reproducing kernel Hilbert Spaces
4. Basic operator theory, Mercer's theorem
5. Random vectors and their moments
6. Gaussian measures, the Hajék-Feldman dichotomy, Conditional Independence
7. Mean square contintuity & the Karhunen-Loève theorem
8. Mean square vs pathwise regularity
9. Weak Convergence, tightness, CLT, LLN
10. Moment estimation and the problem of measurement
11. Functional Principal Components
12. The positive definite continuation problem
13. Intrinsic functional graphical models

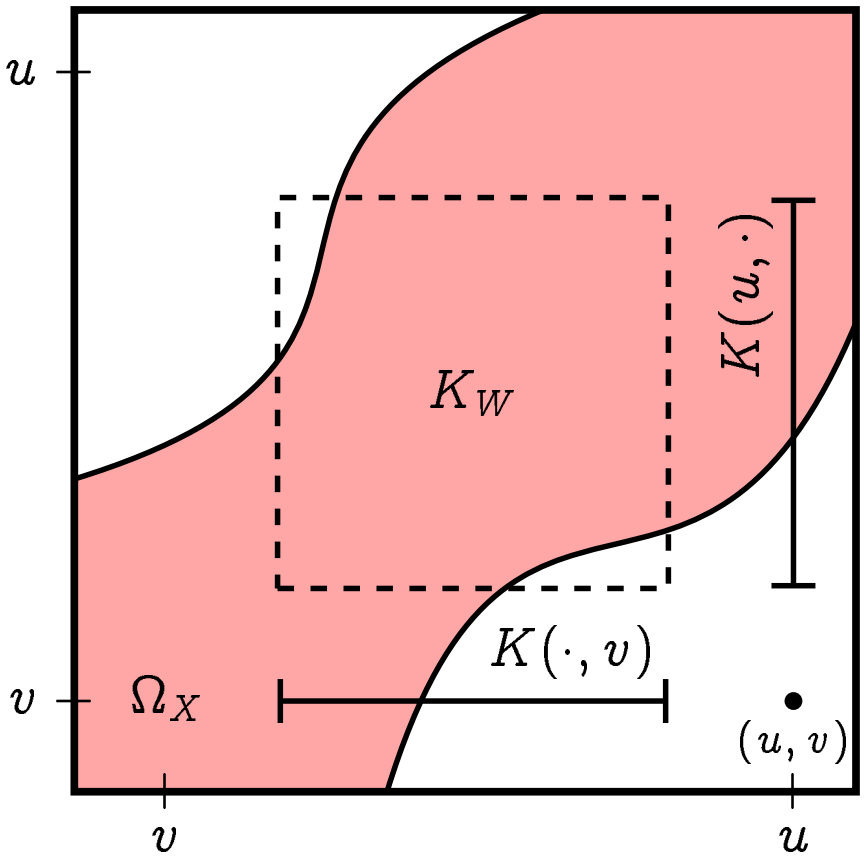
- To gain some intuition, consider the $p \times p$ matrix case



- Ask for completion that maximizes Gaussian differential entropy

$$\frac{1}{2} \log \{(2\pi e)^p \det(K)\}$$

- Solution: "missing entries" of Σ^{-1} should be zero.
- This is precisely what our procedure would produce.
- Except we have no inverse and no usual determinant now...
- Still, clearly there must be a connection to graphical models



Say that $X \sim N(0, K)$ has the graph $\Omega \subset U \times U$ if for every $u, v \in U$ separated in Ω by $W \subset U$, it satisfies the *global Markov property*:

$$X_u \perp\!\!\!\perp X_v \mid X_W$$

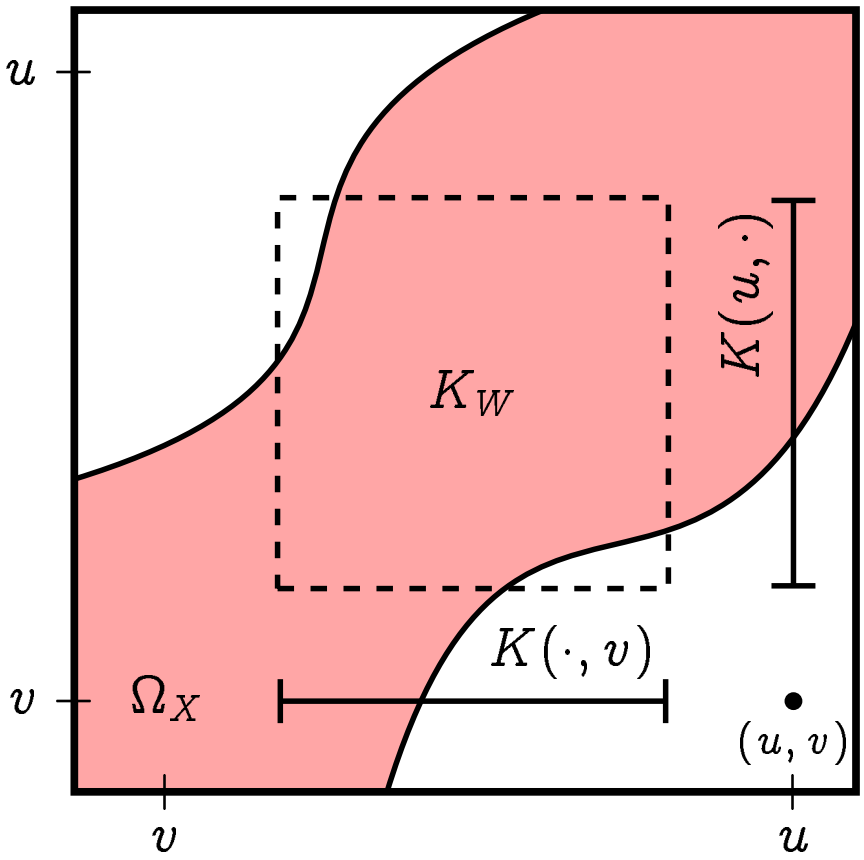
Theorem (Waghmare & Panaretos, 2024+)

Equivalent to satisfying the completion formula

$$K(u,v) = \langle K(u,\cdot), K(\cdot,v) \rangle_{\mathcal{H}(K_W)}$$

for any W separating (u, v) in Ω

Thus the graph Ω of a process can described in terms of its covariance K without reference to the inverse.



Holds whether U is finite, countably infinite, uncountably infinite... in fact for virtualy any U.

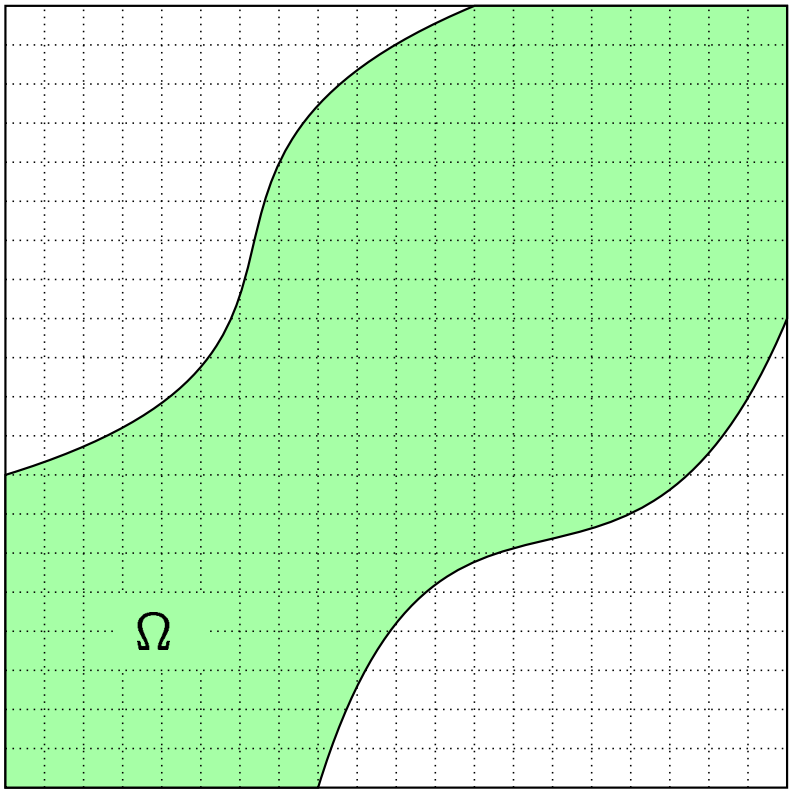
Characterizes graph structure via the covariance, bypassing invertibility

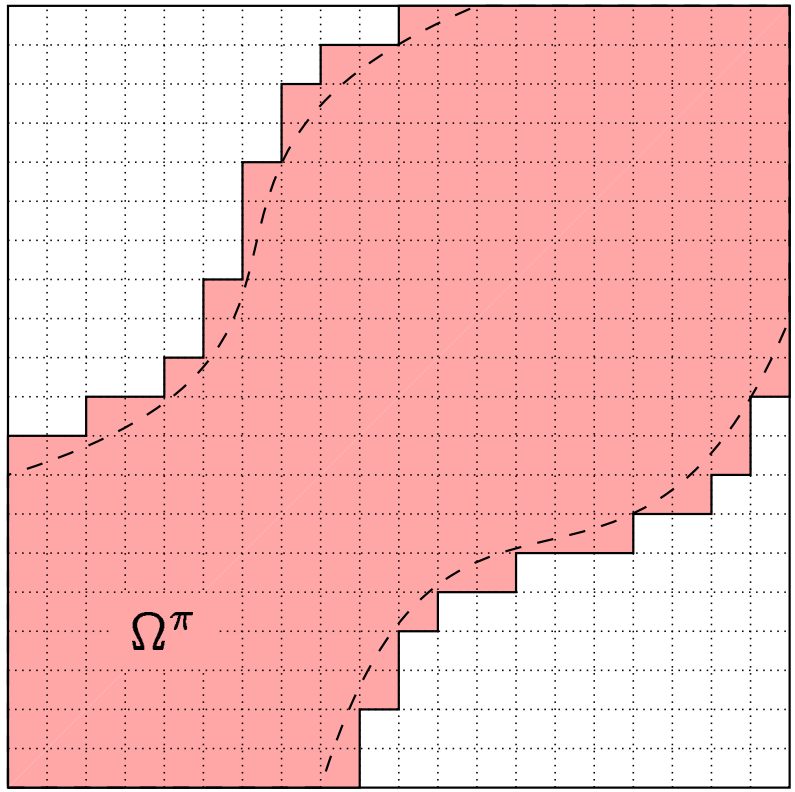
But... can we operationalize it and go the other way round?

Given the covariance K of a process, can we work out the graph Ω?

Here, there is no such thing as an inverse, but the completion equation can help us "approximate" the graph via successive serrated versions.

A pragmatic approach: partition the domain U into $\pi = \{U_1, \ldots, U_p\}$.





Instead of asking whether there is an edge between $u, v \in U$, we ask whether there is an edge between (some point in) U_i and (some point in) U_j.

Turns out Ω^{π} now admits a nice inverse zero characterization: for $1 \leq i, j \leq p$, let

$$\mathcal{K}_{ij} : L^2(U_j, \mu) \to L^2(U_i, \mu)$$

be the integral operator induced by the integral kernel $K_{ij} = K|_{U_i \times U_j}$ given by

$$\mathcal{K}_{ij} f(u) = \int_{U_j} K_{ij}(u, v) f(v) d\mu(v)$$

Define the *covariance operator matrix* $\mathcal{K}_{\pi}$ induced by the partition π as

$$\mathcal{K}_{\pi} = [\mathcal{K}_{ij}]_{i,j=1}^{p}.$$

Define the *correlation operator matrix* $\mathcal{R}_{\pi}$ induced by the partition π as

$$\mathcal{R}_{\pi} = [\mathcal{R}_{ij}]_{i,j=1}^{p} = [\mathcal{K}_{ii}^{-1/2} \mathcal{K}_{ij} \mathcal{K}_{jj}^{-1/2}]_{i,j=1}^{p}$$

$$= \mathrm{dg}(\mathcal{K})^{-1/2} \mathcal{K} \mathrm{dg}(\mathcal{K})^{-1/2} = \mathcal{I} + \mathrm{dg}(\mathcal{K})^{-1/2} (\mathcal{K} - \mathrm{dg}(\mathcal{K})) \mathrm{dg}(\mathcal{K})^{-1/2} = \mathcal{I} + (\mathcal{R}_{\pi} - \mathrm{dg}(\mathcal{R}_{\pi}))$$

because $\mathrm{dg}(\mathcal{R}_{\pi}) = \mathcal{I}$ by definition. So the correlation operator matrix is a bounded and self-adjoint perturbation of the identity.

Thus, contrary to $\mathcal{K}_\pi$, the correlation operator matrix $\mathcal{R}_\pi$ is "typically" invertible:

If $\inf \sigma(\mathcal{R}_\pi - \mathrm{dg}(\mathcal{R}_\pi)) > -1$, then $\mathcal{R}_\pi^{-1}$ is well-defined.

And so now we are legitimised to state:

Theorem (Waghmare & Panaretos, 2024+)

If $\mathcal{R}_\pi$ is invertible, the graph Ω^π is related to the inverse $\mathcal{P}_\pi = \mathcal{R}_\pi^{-1}$ as follows:

$$\Omega^\pi = \cup \{ U_i \times U_j : \|\mathcal{P}_{ij}\| \neq 0 \}.$$

- Choosing a finer partition π yields a higher resolution version Ω^π of Ω.
- Importantly, this characterization behaves coherently under *refinement*.

Corollary (Waghmare & Panaretos, 2024+)

Let $X \sim N(0, K)$ on U with K continuous. If π is a parttiion of U with $\mathcal{R}_\pi$ is invertible, then Ω_X is identifiable up to π-resolution. Furthermore, if there exists a sequence $\{\pi_j\}_{j=1}^\infty$ such that (a) the correlation operators $\mathcal{R}_{\pi_j}$ are invertible and (b) the partitions separate points on U, then Ω_X is identifiable exactly.

As discussed in the context of functional PCA:

- There is a zoo of possible observation scenarios...
- ... and a corresponding zoo of covariance estimators.
- Ideally, we would like to have plug-in estimators.
- First, we estimate $\mathcal{K}$ by the appropriate (for the observation scenario) $\hat{\mathcal{K}}$.
- Then we get $\hat{\mathcal{R}}_\pi$ and $\hat{\mathcal{P}}_\pi$ by applying the steps described to $\hat{\mathcal{K}}$.
- Going from $\mathcal{K}_\pi$ to $\mathcal{R}_\pi$ and $\mathcal{P}_\pi$ involved inverses.
- We are thus going to need regularised plug-in estimation.

Thresholding in the Operator Norm

1. Covariance Operator Matrix Estimator.

$$\hat{\mathcal{K}} = [\hat{\mathcal{K}}_{ij}]_{i,j=1}^{p}$$

2. (Regularised) Correlation Operator Matrix Estimator

$$\hat{\mathcal{R}} = \mathcal{I} + (\kappa\mathcal{I} + \mathrm{dg}\hat{\mathcal{K}})^{-1/2}(\hat{\mathcal{K}} - \mathrm{dg}\hat{\mathcal{K}})(\kappa\mathcal{I} + \mathrm{dg}\hat{\mathcal{K}})^{-1/2}$$

3. Precision Operator Matrix. $\hat{\mathcal{P}} = \hat{\mathcal{R}}^{-1}$.

4. Estimation of Ω by thresholding at level ρ.

$$\hat{\Omega} = \cup\{U_i \times U_j : \|\hat{\mathcal{P}}_{ij}\|_\infty \geq \rho\}$$

What about performance guarantees for our estimators, say at a given resolution

These come in asymptotic and a non-asymptotic forms:

- *Asymptotic Guarantees*. This is well-suited to the plug-in mentality, which is flexible w.r.t. the measurement problem. In this setting we can pursue plug-in rates of convergence, which take as input the rate of convergence of the chosen covariance estimator at the given regime, and yield the rate of convergence of the other estimands.
- *Non-Asymptotic Guarantees*. Beyond rates of convergence, which are asymptotic in nature, we can also consider *finite-sample guarantees* for the various possible observation regimes. Finite sample guarantees are by nature specific to the estimator used, which in turn needs to be tailored to the corresponding sampling regime.

An advantage of non-asymptotic theory is that it makes no reference to limits, hence allows us to address recovery of the continuum version of Ω_X.

I.e. it can tell us how to successively refine the partition π as sample size increases, in order to construct a consistent estimator at infinite resolution.

Since we have an inverse problem at hand we need some regularity conditions.

First that $\inf \sigma(\mathcal{R}_\pi - \mathrm{dg}(\mathcal{R}_\pi)) > -1$, so that $\mathcal{R}_\pi^{-1}$ is well-defined.

Then a "source condition": for some bounded operator matrix Φ_0 with all the diagonal entries zero and $\beta > 0$, we have

$$\mathcal{R}_\pi - \mathrm{dg}\mathcal{R}_\pi = [\mathrm{dg}\mathcal{K}]^\beta \Phi_0 [\mathrm{dg}\mathcal{K}]^\beta.$$

Note that this implies that $\mathcal{R}_\pi - \mathrm{dg}\mathcal{R}_\pi$ is compact.

The assumption simply ensures that $\mathcal{K} - \mathrm{dg}\mathcal{K} = [\mathrm{dg}\mathcal{K}]^{1/2+\beta} \Phi_0 [\mathrm{dg}\mathcal{K}]^{1/2+\beta}$ is linearly well-conditioned for inversion by $[\mathrm{dg}\mathcal{K}]^{1/2}$.

Our first result now relates $\|\hat{\mathcal{R}}_\pi - \mathcal{R}_\pi\|$ to $\|\hat{\mathcal{K}} - \mathcal{K}\|$, $\mathcal{K}$ and $\|\mathcal{R}_\pi\|$:

Theorem (Asymptotic Guarantees – Correlation)

In the presence of our two regularity assumptions, and given any sequences $\kappa_n > 0$ and $\delta_n \geq \|\hat{\mathcal{K}} - \mathcal{K}\|$, we have

$$\|\hat{\mathcal{R}}_\pi - \mathcal{R}_\pi\| \leq 5 \cdot \|\mathcal{R}_\pi\| \cdot \left[(\delta_n/\kappa_n)^2 + (\delta_n/\kappa_n)\right] + 2 \cdot \kappa_n^{\beta\wedge 1} \cdot \|\Phi_0\| \cdot \|\mathcal{K}\|^{2\beta-\beta\wedge 1}.$$

The estimator $\hat{\mathcal{R}}_\pi$ is consistent so long as the regularization parameter κ_n is chosen such that $\kappa_n \to 0$ and $\delta_n/\kappa_n \to 0$ as $n \to \infty$. The optimal rate is given by

$$10 \cdot (\|\mathcal{R}_\pi\| \vee \|\Phi_0\|\|\mathcal{K}\|^{2\beta-\beta\wedge 1}) \cdot \delta_n^{\frac{\beta\wedge 1}{1+\beta\wedge 1}}$$

and it is achieved for the choice $\kappa_n = \delta_n^{\frac{1}{1+\beta\wedge 1}}$.

If $\hat{\mathcal{K}}$ is the empirical covariance,

- δ_n is $O_{\mathbb{P}}(n^{-1/2})$,
- the optimal choice of the regularization parameter is given by $\kappa_n \asymp n^{-1/2(\beta\wedge 1+1)}$
- and we obtain the rate of convergence $\|\hat{\mathcal{R}}_\pi - \mathcal{R}_\pi\| = O_{\mathbb{P}}\left(n^{-\beta\wedge 1/2(\beta\wedge 1+1)}\right)$.

With our assumptions, $\mathcal{R}_\pi$ is strictly positive. The operator $\hat{\mathcal{R}}_\pi$ is also strictly positive for all sufficiently large n, by virtue of being consistent. So for all sufficiently large n, we may write

$$\hat{\mathcal{P}}_\pi - \mathcal{P}_\pi = \hat{\mathcal{R}}_\pi^{-1}\mathcal{R}_\pi\mathcal{R}_\pi^{-1} - \hat{\mathcal{R}}_\pi^{-1}\hat{\mathcal{R}}_\pi\mathcal{R}_\pi^{-1} = \hat{\mathcal{R}}_\pi^{-1}\left[\mathcal{R}_\pi - \hat{\mathcal{R}}_\pi\right]\mathcal{R}_\pi^{-1} = \hat{\mathcal{P}}_\pi\left[\mathcal{R}_\pi - \hat{\mathcal{R}}_\pi\right]\mathcal{P}_\pi.$$

and it can be shown that $\|\hat{\mathcal{P}}_\pi\|$ is bounded in probability under our assumptions. As a result, the convergence rates for $\|\hat{\mathcal{R}}_\pi - \mathcal{R}_\pi\|$ also apply to $\|\hat{\mathcal{P}}_\pi - \mathcal{P}_\pi\|$.

Theorem (Asymptotic Guarantees - Precision and Graph Recovery)

In the same context as the last result, and with the optimal choice of the regularization parameter κ_n, we have

$$\|\hat{\mathcal{P}}_\pi - \mathcal{P}_\pi\| = \|\mathcal{P}_\pi\|(\|\mathcal{R}_\pi\| \vee \|\Phi_0\|)\|\mathcal{K}\|^{2\beta-\beta\wedge 1}) \cdot O_{\mathbb{P}}(\delta_n^{\frac{\beta\wedge 1}{1+\beta\wedge 1}}).$$

If we choose ρ_n such that $\rho_n/\delta_n^{\beta\wedge 1/(1+\beta\wedge 1)} \to 0$, then

$$\mathbb{P}[\hat{\Omega}^\pi(\rho_n) \neq \hat{\Omega}_X^\pi] \overset{n\to\infty}{\longrightarrow} 0.$$