Hilbert Spaces

Overview

1. Ideas
2. Preliminaries: vector spaces, metric spaces, normed linear spaces
3. Hilbert spaces
4. Projection
5. Orthonormal bases
6. Separability and the fundamental isomorphism
7. Applications to random variables

Ideas

Rationale for studying Hilbert spaces:

1. Formalize intuition from regression on prediction, orthogonality
2. Define infinite sums of random variables, such as $\sum_j \psi_j w_{t-j}$
3. Frequency domain analysis
4. Representation theorems

The related Fourier analysis establishes an *isometry* between the collection of stationary stochastic processes $\{X_t\}$ and squared integrable functions on $[-\pi, \pi]$. This isometry lets us replace

$$X_t \quad \Rightarrow \quad e^{it\lambda}$$

in a way that preserves covariances. In the notation of inner-products, $\langle x, y \rangle$

$$\begin{aligned}
\mathrm{Cov}(X_{t+h}, X_t) &= \langle X_{t+h}, X_t \rangle \\
&= \langle e^{i(t+h)\lambda}, e^{it\lambda} \rangle_f \\
&= \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda
\end{aligned}$$

$$\text{more generally} \quad = \quad \int_{-\pi}^{\pi} e^{ih\lambda} dF(\lambda)$$

for a suitably defined function f, the spectral density function, or F, the spectral measure. (The spectral density might not exist.)

Fourier transform The relationship $\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda$ indicates that the s.d.f. is the Fourier transform of the covariances. The Fourier transform is an isometry between Hilbert spaces.

Related ideas For more reading see Appendix C in Shumway and Stoffer as well as these classics (and one newer edition)

- Halmos, P. R. (1958). *Finite Dimensional Vector Spaces*, Springer.
- Lax, P. (2002). *Functional Analysis*, Wiley.
- Reed and Simon (1972). *Functional Analysis*, Academic Press.
- Rudin, W. (1973). *Functional Analysis*, McGraw-Hill.

and for time series, see Brockwell and Davis, *Time Series: Theory and Methods*.

Vector Spaces

Key ideas associated with a vector space (a.k.a, *linear space*) are subspaces, basis, and dimension.

Define. A complex vector space is a set $\mathcal{V}$ of elements called vectors that satisfy the following axioms on *addition of vectors*

1. For every $x, y \in \mathcal{V}$, there exists a sum $x + y \in \mathcal{V}$
2. Addition is commutative, $x + y = y + x$.
3. Addition is associative, $x + (y + z) = (x + y) + z$.
4. There exists an *origin* denoted 0 such that $x + 0 = x$.
5. For every $x \in \mathcal{V}$, there exists $-x$ such that $x + (-x) = 0$, the origin.

and the following axioms on *multiplication by a (complex) scalar*

1. For every $x \in \mathcal{V}$ and $\alpha \in \mathcal{C}$, the product $\alpha x \in \mathcal{V}$.

2. Multiplication is commutative.
3. $1\,x = x$.
4. Multiplication is distributive: $\alpha(x+y) = \alpha x + \alpha y$ and $(\alpha+\beta)x = \alpha x + \beta x$.

A vector space must have at least one element, the origin.

Examples. Some common examples of vector spaces are:

1. Set of all complex numbers $\mathbb{C}$ or real numbers $\mathbb{R}$.
2. Set of all polynomials with complex coefficients.
3. Euclidean n-space ("vectors") whose elements are complex numbers, often labelled $\mathbb{C}^n$.

Subspaces. A set $\mathcal{M} \subset \mathcal{V}$ is a subspace or *linear manifold* if it is *algebraically closed*,

$$x, y \in \mathcal{M} \quad \Rightarrow \quad \alpha x + \beta y \in \mathcal{M}\ .$$

Consequently, each subspace must include the origin, since $x - x = 0$. The typical way to *generate* a subspace is to begin with a collection of vectors and consider the set of all possible linear combinations of this set; the resulting collection is a subspace. Intersections of subspaces are also subspaces. (Note: "closure" here is not in the sense of open and closed sets. A vector space need not have a topology.)

Linear dependence. A countable set of vectors $\{x_i\}$ is linearly dependent if there exists a set of scalars, not all zero, s.t.

$$\sum_i \alpha_i x_i = 0.$$

The sum $\sum_i \alpha_i x_i$ is known as a *linear combination* of vectors. Alternatively, the collection of vectors $\{x_i\}$ is linearly dependent iff some member x_k is a linear combination the preceding vectors.

Bases and dimension. A basis for $\mathcal{V}$ is a set of linear independent vectors $\{x_i\}$ such that every vector $v \in \mathcal{V}$ can be written as a linear combination of the basis,

$$v = \sum_i \alpha_i x_i.$$

The *dimension* of a vector space is the number of elements in a basis for that space.

Isometry. Two vector spaces are *isomorphic* if there exists a linear bijection (one-to-one, onto) $T : X \to Y$,

$$T(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 T(x_1) + \alpha_2 T(x_2).$$

Metric spaces

Distance metric defines topology (open, closed sets) and convergence.

Define. A metric space X combines a set of elements (that need not be a vector space) with the notion of a distance, called the *metric*, d. The metric $d : (X \times X) \to R$ must satisfy:

- Non-negative: $d(x, y) \geq 0$, with equality iff $x = y$.
- Symmetric: $d(x, y) = d(y, x)$.
- Triangle: $d(x, y) \leq d(x, z) + d(z, y)$.

Examples. Three important metrics defined on space of continuous functions $C[a, b]$ are

- $d_\infty(f, g) = \max |f(x) - g(x)|$. (Uniform topology)
- $d_1(f, g) = \int_a^b |f(x) - g(x)| dx$.
- $d_2(f, g)^2 = \int_a^b (f(x) - g(x))^2 dx$.

Convergence is defined in the metric, $x_n \to x$ if $d(x_n, x) \to 0$. Different metrics induce different notions of convergence. The "triangle functions" h_n defined on $\frac{1}{2n+1}, \frac{1}{2n}, \frac{1}{2n-1}$ converge to zero in d_1, but not in d_∞.

Cauchy sequences and completeness. The sequence x_n is Cauchy if for all $\epsilon > 0$, there exists an N such that $n, m \geq N$ implies $d(x_n, x_m) < \epsilon$. All convergent sequences must be Cauchy, though the converse need not be true. If all Cauchy sequences converge to a member of the metric space, the space is said to be *complete*.

Since the triangle functions have disjoint support, $d_\infty(h_n, h_m) = 1$, $\{h_n\}$ is not Cauchy, and thus does not converge. With d_∞, $C[a,b]$ is complete since sequences like $\{h_n\}$ do not converge in this metric; $C[a,b]$ is not complete with d_1 (or d_2) as the metric.

Continuous functions. With the notion of convergence in hand, we can define a function f to be continuous if it preserves convergence of arguments. The function f is continuous iff

$$x_n \to x \quad \Rightarrow \quad f(x_n) \to f(x) \ .$$

Isometry. Two metric spaces (X, d_x) and (Y, d_y) are isometric if if there exists a bijection (1-1, onto) $f : X \to Y$ which preserves distance,

$$d_x(a,b) = d_y(f(a), f(b)) \ .$$

Isomorphism between two vector spaces requires the preservation of linearity (linear combinations), whereas in metric spaces, the metric must preserve distance. These two notions — the algebra of vector spaces and distances of metric spaces — combine in normed linear spaces.

Normed linear spaces.

Combine the algebra of vector spaces and distance of metric spaces.

Define. A normed vector space $\mathcal{V}$ is a vector space together with a real-valued function $\|x\|$, the "norm" which is

1. Non-negative: $\|x\| \ge 0$, with equality iff $x = 0$.
2. Scalar mult: $\|\alpha x\| = |\alpha| \ \|x\|$.
3. Triangle: $\|x + y\| \le \|x\| + \|y\|$.

Continuity of norm. If $x_n \to x$, then $||x_n|| \to ||x||$. This follows from the triangle inequality (noting $x = x - x_n + x_n$)

$$| \, \|x_n\| - \|x\| \, | \le \ \|x_n - x\| \ .$$

Complete space A normed vector space $\mathcal{V}$ is complete if all Cauchy sequences $\{X_i\} \in \mathcal{V}$ have limits within the space:

$$\lim \|X_i - X_j\| \to 0 \implies \lim X_i \in \mathcal{V}$$

Examples. Several common normed spaces are ℓ_1 and the collection L^1 of Legesgue integrable functions with the norm

$$\|f\| = \int_{-\infty}^{\infty} |f(x)| dx < \infty \,.$$

While it is true that $\ell_1 \subset \ell_2$, the same does not hold for functions due to problems at the origin. There is not a nesting of L^2 and L^1. For example,

$$1/(1 + |x|) \in L^2 \text{ but not in } L^1.$$

Conversely,

$$|x|^{-1/2} e^{-|x|} \in L^1 \text{ but not in } L^2.$$

Remarks. Using the Lebesgue integral, L^1 is complete and is thus a *Banach space.* Also, the space of continuous functions $C[a, b]$ is *dense* is $L^1[a, b]$.

Hilbert Spaces

Geometry Hilbert spaces conform to our sense of geometry and regression. For example, a key notion is the orthogonal decomposition (data = fit + residual, or $Y = \hat{Y} + (Y - \hat{Y})$).

Inner product space A vector space $\mathcal{H}$ is an *inner-product space* if for each $x, y \in \mathcal{H}$ there exists a real-valued, bilinear function $\langle x,\, y\rangle$ which is

1. Linear: $\langle x + y,\, z\rangle = \langle x,\, z\rangle + \langle y,\, z\rangle$
2. Scalar multiplication: $\langle \alpha x,\, y\rangle = \alpha\langle x,\, y\rangle$
3. Non-negative: $\langle x,\, x\rangle \geq 0$, with $\langle x,\, x\rangle = 0$ iff $x = 0$
4. Conjugate symmetric: $\langle x,\, y\rangle = \overline{\langle y,\, x\rangle}$ (symmetry in real-valued case)

Hilbert space A Hilbert space is a *complete* inner-product space. An inner-product space can always be "completed" to a Hilbert space by adding the limits of its Cauchy sequences to the space.

Examples The most common examples of Hilbert spaces are

1. Euclidean $\mathbb{R}^n$ and $\mathbb{C}^n$ with inner products defined by the dot-product $\langle x,\, y\rangle = \sum_i x_i \bar{y}_i$.
2. ℓ_2 sequences (square summable sequences). This is the canonical Hilbert space.
3. $L_2[a,b]$; $f \in L_2$ iff $\int_a^b f^2 < \infty$. L_2 is complete, and is thus a Hilbert space. Note that the inner-product $\langle f,\, g\rangle = \int f\overline{g}$ is valid (integrable) since
$$(f-g)^2 \geq 0 \implies |f(x)\overline{g(x)}| \leq (|f(x)|^2 + |g(x)|^2)/2$$
so that the product $f\,\overline{g}$ is integrable (lies in L_1).
4. Random variables with finite variance, an idea that we will explore further. The inner product is $\langle X,\, Y\rangle = \text{Cov}(X, y)$.

Norm Every Hilbert space has an associated *norm* defined using its inner product,
$$\|x\|^2 = \langle x,\, x\rangle,$$
which reduces to the (squared) length of a vector in $\mathbb{R}^n$. Observe that $\|\alpha x\| = |\alpha|\,\|x\|$, as in the definition of a normed space. Norms in i.p. spaces are special; in particular, they also satisfy the Parallelogram Law
$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

Orthonormal If $\langle x,\, y\rangle = 0$, then x and y are orthogonal, often written as $x \perp y$. A collection of orthogonal vectors having norm 1 is an *orthonormal* set. For example, in $\mathbb{R}^n$, the columns of an orthogonal matrix form an orthonormal set.

Pythagorean theorem Let $\mathcal{X} = \{x_j\}_{j=1}^n$ denote an orthonormal set in $\mathcal{H}$. Then for any $x \in \mathcal{H}$,
$$\|x\|^2 = \sum_{j=1}^n |\langle x,\, x_j\rangle|^2 + \|x - \textstyle\sum_j \langle x,\, x_j\rangle x_j\|^2\;. \tag{1}$$

We know this in statistics as the ANOVA decomposition in statistics, Total SS = Fit SS + Resid SS. Furthermore, the vector $r = x - \sum_i \langle x,\, x_i\rangle x_i$ is orthogonal to the subspace spanned by $\mathcal{X}$. Compare to

$$x = (I - H + H)x = Hx + (I - H)x$$

where H is a projection (idempotent) matrix.

Proof. Begin with the identity (once again, add and subtract)

$$x = \sum_j \langle x,\, x_j\rangle x_j + \left(x - \sum_j \langle x,\, x_j\rangle x_j \right) \;,$$

The two on the r.h.s are orthogonal and thus (1) holds. The coefficients $\langle x,\, x_j\rangle$ seen in (1) are known as Fourier coefficients.

Bessel's inequality If $\mathcal{X} = \{x_1, \ldots, x_n\}$ is an orthonormal set in $\mathcal{H}$, $x \in \mathcal{H}$ is any vector, and the Fourier coefficients are $\alpha_j = \langle x,\, x_j\rangle$, then

$$\|x\|^2 \geq \sum_j |\alpha_j|^2 \;.$$

Proof. Immediate from Pythagorean theorem.

Cauchy-Schwarz inequality For any $x, y \in \mathcal{H}$,

$$|\langle x,\, y\rangle| \;\leq\; \|x\|\,\|y\|.$$

Equality occurs when $\{x, y\}$ are linearly dependent. Hence we can think of the norm as an upper bound on the size of inner products:

$$\|x\| = \max_{\|y\|=1} |\langle x,\, y\rangle|$$

Proof. The proof suggests that the C-S inequality is closely related to the ideas of projection. The result is immediate if $y = 0$. Assume $y \neq 0$ and consider the orthonormal set $\{y/||y||\}$. Bessel's inequality implies

$$|\alpha|^2 = \langle x,\, y/\|y\|\rangle^2 \leq \|x\|^2.$$

Equality occurs when x is a multiple of y, for then the term omitted from the Pythagorean theorem that leads to Bessel's inequality is zero.

Some results that are simple to prove with the Cauchy-Schwarz theorem are:

1. The inner product is continuous, $\langle x_n, y_n\rangle \to \langle x, y\rangle$. The proof essentially uses $x_n = x_n - x + x$ and the Cauchy-Schwarz theorem. Thus, we can always replace $\langle x, y\rangle = \lim_n \langle x_n, y\rangle$.
2. Every inner product space is a normed linear space, as can be seen by using the C-S inequality to verify the triangle inequality for the implied norm.

Isometry between Hilbert spaces combines linearity from vector spaces with distances from metric spaces. Two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are isomorphic if there exists a linear function U which preserves inner products,

$$\forall x, y \in \mathcal{H}_1, \quad \langle x, y\rangle_1 \;=\; \langle Ux, Uy\rangle_2 \;.$$

Such an operator U is called *unitary*. The canonical example of an isometry is the linear transformation implied by an orthogonal matrix.

Summary of theorems For these, $\{x_j\}_{j=1}^n$ denotes a finite orthonormal set in an inner product space $\mathcal{H}$:

1. Pythgorean theorem: $\|x\|^2 = \sum_{j=1}^n |\langle x, x_j\rangle|^2 + \|x - \sum_j \langle x, x_j, x\rangle_j\|^2$.
2. Bessel's inequality: $\|x\|^2 \ge \sum_j |\langle x, x_j\rangle|^2$.
3. Cauchy-Schwarz inequality: $|\langle x, y\rangle| \;\le\; \|x\|\,\|y\|$.
4. Inner product spaces are normed spaces with $\|x\|^2 = \langle x, x\rangle$. This norm satisfies the parallelogram law.
5. The i.p. is continuous, $\lim_n \langle x_n, y\rangle = \langle x, y\rangle$.

Projection

Orthogonal complement Let $\mathcal{M}$ denote any subset of $\mathcal{H}$. Then the set of all vectors orthogonal to $\mathcal{M}$ is denoted $\mathcal{M}^\perp$, meaning

$$x \in \mathcal{M}, y \in \mathcal{M}^\perp \quad \Rightarrow \quad \langle x, y\rangle = 0 \;.$$

Notice that

1. $\mathcal{M}^\perp$ is a subspace since the i.p. is linear.
2. $\mathcal{M}^\perp$ is closed (contains limit points) since the i.p. is continuous:

$$y \in \mathcal{M}, x_n \in \mathcal{M}^\perp \to x \implies \langle x, y \rangle = \lim_n \langle x_n, y \rangle = 0$$

Projection lemma Let $\mathcal{M}$ denote a closed subspace of $\mathcal{H}$. Then for any $x \in \mathcal{H}$, there exists a *unique* element $\hat{x} = P_{\mathcal{M}} x \in \mathcal{M}$ closest to x,

$$d = \inf_{y \in \mathcal{M}} \|x - y\|^2 = \|x - \hat{x}\|^2, \quad \hat{x} \text{ is unique.}$$

The vector $\hat{x}$ is known as the projection of x onto $\mathcal{M}$. A picture suggests the "shape" of closed subspaces in a Hilbert space is very regular (not "curved"). This lemma only says that such a closest element exists; it does not attempt to describe it.

Proof. Relies on the parallelogram law and closure properties of the subspace. The first part of the proof shows that there is a Cauchy sequence y_n in $\mathcal{M}$ for which $\lim \|x - y_n\| = \inf_{\mathcal{M}} \|x - y\|$. To see unique, suppose there were two, then use the parallelogram law to show that they are the same:

$$\begin{aligned} 0 \le \|\hat{x} - \hat{z}\|^2 &= \|(\hat{x} - x) - (\hat{z} - x)\|^2 \\ &= -\|(\hat{x} - x) + (\hat{z} - x)\|^2 + 2(\|\hat{x} - x\|^2 + \|\hat{z} - x\|^2) \\ &= -4\|(\hat{x} + \hat{z})/2 - x)\|^2 + 2(\|\hat{x} - x\|^2 + \|\hat{z} - x\|^2) \\ &\le -4d + 4d = 0 \end{aligned}$$

Projection theorem. Let $\mathcal{M}$ denote a closed subspace of $\mathcal{H}$. Then every $x \in \mathcal{M}$ can be uniquely written as

$$x = P_{\mathcal{M}} x + z \text{ where } z \in \mathcal{M}^\perp$$

Proof. Let $P_{\mathcal{M}} x$ be the vector identified in the lemma so that uniqueness is established. Define $z = x - P_{\mathcal{M}} x$. The challenge is to show that $\langle z, y \rangle = 0$ for all $y \in \mathcal{M}$ so that z indeed lies in $\mathcal{M}^\perp$. Again, the proof is via a contradiction. Suppose $\exists y \in \mathcal{M}$ such that $\langle x - \hat{x}, y \rangle \ne 0$. This contradicts $\hat{x}$ being the closest to x. Let $b = \langle x - \hat{x}, y \rangle / \|y\|^2$, the "regression coefficient of the residual $x - \hat{x}$ on y. Using real numbers,

$$\|x - \hat{x} - by\|^2 = \|x - \hat{x}\|^2 - \frac{\langle x - \hat{x}, y \rangle^2}{\|y\|^2} < \|x - \hat{x}\|^2 ,$$

a contradiction.

Properties of projection mapping Important properties of the projection mapping $P_{\mathcal{M}}$ are

1. Linear: $P_{\mathcal{M}}(\alpha x + \beta y) = \alpha P_{\mathcal{M}} x + \beta P_{\mathcal{M}} y$.
2. Anova decomposition: $\|x\|^2 = \|P_{\mathcal{M}} x\|^2 + \|(I - P_{\mathcal{M}})x\|^2$.
3. Representation $x = P_{\mathcal{M}} x + (I - P_{\mathcal{M}})x$ is unique from the projection theorem.
4. Continuous: $P_{\mathcal{M}} x_n \to P_{\mathcal{M}} x$ if $x_n \to x$. (use linearity and the anova decomposition)
5. Idempotent: $P_{\mathcal{M}} x = x \Leftrightarrow x \in \mathcal{M}$ and $P_{\mathcal{M}} x = 0 \Leftrightarrow x \in \mathcal{M}^{\perp}$
6. Subspaces: $P_{\mathcal{M}_1} P_{\mathcal{M}_2} x = P_{\mathcal{M}_1} x \Leftrightarrow \mathcal{M}_1 \subseteq \mathcal{M}_2$.

Regression Least squares regression fits nicely into the Hilbert space setting. Let $\mathcal{H}$ denote real Euclidean n-space R^n with the usual dot-product as inner product, and let $\mathcal{M}$ denote the subspace formed by linear combinations of the vectors $x_1, x_2, \ldots, x_k$.

Consider a vector $y \in \mathcal{H}$. The projection theorem tells us that we can form an orthogonal decomposition of y as

$$y = P_{\mathcal{X}} y + z \text{ where } P_{\mathcal{X}} y = \sum \alpha_j x_j\,,$$

and $z = y - P_{\mathcal{X}} y$. Since $\langle z,\, x_j \rangle = 0$, we obtain a system of equations (*the normal equations* — it's also clear now why these are called the normal equations!)

$$\langle z,\, x_j \rangle = \langle y - \textstyle\sum \alpha_i x_i,\, x_j \rangle = 0, \quad j = 1, \ldots, k$$

Solving this system gives the usual OLS regression coefficients. Notice that we can also express the projection theorem explicitly as

$$y = Hy + (I - H)y\,,$$

where the idempotent projection matrix $P_{\mathcal{X}} = H$ is $H = X(X'X)^{-1}X'$, the "hat matrix".

Orthonormal bases.

Regression is most easy to interpret and compute if the columns $x_1, x_2, \ldots, x_k$ are orthonormal. In that case, the normal equations are diagonal and regression coefficients are simply $\alpha_j = \langle y, x_j \rangle$. This idea of an *orthonormal basis* extends to *all* Hilbert spaces, not just those that are finite dimensional. If the o.n. basis $\{x_j\}_{j=1}^n$ is finite, though, the projection is $P_{\mathcal{M}} y = \sum \langle y, x_j \rangle x_j$ as in regression with orthogonal X.

Theorem. Every Hilbert space has an orthonormal basis.

The *proof* amounts to Zorn's lemma or the axiom of choice. Consider the collection of all orthonormal sets, ...

Fourier representation Let $\mathcal{H}$ denote a Hilbert space and let $\{x_\alpha\}$ denote an orthonormal basis (Note: α is a member of some set A, not just integers.) Then for any $y \in \mathcal{H}$, we have

$$y = \sum_A \langle y, x_\alpha \rangle x_\alpha \quad \text{and} \quad \|y\|^2 = \sum_A |\langle y, x_\alpha \rangle|^2$$

The latter equality is called *Parseval's identity.*

Proof. Bessel's inequality works for half of the equality for any finite subsets $A' \subset A$,

$$\sum_{A'} |\langle y, x_\alpha \rangle|^2 \le \|y\|^2.$$

This implies that $\langle y, x_\alpha \rangle >\neq 0$ for at most countable α's so that (with some ordering of the elements of A, $j = \alpha_j$) $\sum_{j=1}^n |\langle y, x_j \rangle|^2$ is a monotone series with an upper bound and is thus convergent as $n \to \infty$. The proof continues by showing that the resulting approximation $\hat{y}_n = \sum_{j=1}^n \langle y, x_j \rangle x_j$ converges to y.

Now show it's Cauchy, and use completeness of $\mathcal{H}$ to conclude that the limit y' must be y,

$$\begin{aligned}
\langle y - y', x_k \rangle &= \lim_n \langle y - \textstyle\sum_{j=1}^n \langle y, x_j \rangle x_j, x_k \rangle \\
&= \langle y, x_k \rangle - \langle y, x_k \rangle \\
&= 0.
\end{aligned}$$

For any other $\alpha \neq \alpha_j$, the same argument shows $\langle y - y', x_\alpha \rangle = 0$. Since $y - y'$ is orthogonal to all of the x_α's, it must be zero (or we could extend the orthonormal basis).

To prove the norm relationship, use the continity of the norm and orthogonality,

$$0 = \lim_n \|y - \textstyle\sum_{j=1}^n \langle y, x_j \rangle x_j\|^2 = \|y\|^2 - \displaystyle\sum_A |\langle y, x_\alpha \rangle|^2$$

Construction The *Gram-Schmidt* construction converts a set of vectors into an orthonormal basis. The method proceeds recursively,

$$\begin{array}{lcl} x_1 & \Rightarrow & o_1 = x_1/\|x_1\| \\ x_2 & \Rightarrow & u_2 = x_2 - \langle x_2, o_1 \rangle o_1, o_2 = u_2/\|u_2\| \\ & \cdots & \\ x_n & \Rightarrow & u_n = x_n - \displaystyle\sum_{j=1}^{n-1} \langle x_n, o_j \rangle o_j, o_n = u_n/\|u_n\| \end{array}$$

QR decomposition In regression analysis, a modified version of the Gram-Schmidt process leads to the so-called QR decomposition of the matrix X. The QR decomposition expresses the covariate matrix X as

$$X = QR \text{ where } Q'Q = I,$$

and R is upper-triangular. With X in this form, one solves the modfied system

$$Y = X\beta + \epsilon \quad \Rightarrow \quad Y = Q(\alpha = R\beta) + \epsilon$$

using $\hat{\alpha_j} = \langle Y, q_j \rangle$. The β's come via back-substitution if needed.

Separability and the Fundamental Isomorphism.

Separable A Hilbert space is *separable* if it has a countable dense subset. Examples: (1) real number system (rationals), (2) Continuous functions $C[a, b]$ (polynomials with rational coefs). A Hilbert space is separable iff it has a countable orthonormal basis.

Proof. If its separable, use G-S to convert the countable dense subset to an orthonormal set (removing those that are dependent). If it has a countable basis, use the Fourier representation to see that it is dense.

Isomorphisms If a separable Hilbert space is finite dimensional, it is isomorphic to $\mathbb{C}^n$. If it not finite dimensional, it is isomorphic to ℓ_2.

Proof. Define the isomorphism that maps $y \in \mathcal{H}$ to ℓ_2 by

$$U y = \{\langle y,\, x_j\rangle\}_{j=1}^{\infty}$$

where $\{x_j\}$ is an orthonormal basis. The sequence in ℓ_2 is the sequence of Fourier coefficients in the chosen basis. Note that the inner product is preserved since

$$\langle y,\, w\rangle = \langle \textstyle\sum_j \langle y,\, x_j\rangle x_j,\, \sum_k \langle w,\, x_k\rangle x_k\rangle = \displaystyle\sum_j \langle y,\, x_j\rangle \overline{\langle w,\, x_j\rangle}$$

which is the i.p. on ℓ_2.

L_2 Space of Random variables

Define the inner product space of random variables with finite variance $L_2 = L_2(\Omega, F, P)$ as the collection of measureable complex-valued functions f for which

$$\int f^2(\omega) P(d\omega) = \int f^2 dP < \infty \ .$$

With the inner product $\langle f,\, g\rangle = \int f\overline{g} dP$, L_2 is a Hilbert space.

Translated to the language of random variables, we form an i.p. space from random variables X for which $E\,X^2 < \infty$ with the inner product

$$\langle X,\, Y\rangle = E\,X\,Y$$

If the random variables have mean zero, then $\langle X,\, Y\rangle = \text{Cov}(X, Y)$.

Equivalence classes Observe that $\langle X,\, X\rangle = E\,X^2 = 0$ does not imply that X is identically zero. It only implies that $X = 0$ a.e. In L_2, the symbol X really stands for an *equivalence class* of functions which are equal almost everywhere. The inner product retains the important property that $\langle X,\, X\rangle = 0$ iff $X = 0$, but the claim only holds for X a.e.

Mean square convergence Convergence in L_2 is convergence in mean square (m.s.),

$$X_n \to X \quad \Leftrightarrow \quad \|X_n - X\| \to 0.$$

That is, $\mathbb{E}\,(X_n - X)^2$ must go to zero.

Properties of mean square convergence derive from those of the associated inner product. We can interchange limits with means, variances and covariances. If $\|X_n - X\| \to 0$, then

1. Mean: $\lim_n EX_n = \lim_n \langle X_n,\, 1\rangle = \langle \lim_n X_n,\, 1\rangle = EX$.
2. Variance: $\lim_n EX_n^2 = \lim_n \langle X_n,\, X_n\rangle = \langle X,\, X\rangle = EX^2$.
3. Covariance: $\lim_n E\, X_n Y_n = \lim_n \langle X_n,\, Y_n\rangle = \langle X,\, Y\rangle = EXY$

The first two are consequences of the third, with $Y_n = 1$ or $Y_n = X_n$.

Note: Probabilistic modes of convergence are:

- Convergence in probability: $\lim_n P\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\} = 1$.
- Convergence almost surely:

$$P\{\omega : \lim_n X_n(\omega) = X\} = 1 \quad \text{or} \quad \lim_n P\{\omega : \sup_{m>n} |X_m(\omega) - X(\omega)| < \epsilon\} = 1.$$

Chebyshev's inequality implies that convergence in mean square implies convergence in probability; also, by definition, a.s. convergence implies convergence in probability. The reverse holds for subsequences. For example, the Borel-Cantelli lemma implies that if a sequence converges in probability, then a subsequence converges almost everywhere. Counter-examples to converses include the "rotating functions" $X_n = I_{[(j-1)/k,j/k]}$ and "thin peaks" $X_n = nI_{[0,1/n]}$. I will emphasize mean square convergence, a Hilbert space idea. However, m.s. convergence also implies a.s. convergence along a subsequence.

Projection and conditional expectation

Conditional mean is the *minimum mean squared predictor* of any random variable Y given a collection $\{X_1, \ldots, X_n\}$ is the conditional expectation of Y given the X's. Need to assume that $\mathrm{Var}(Y) < \infty$.

Proof. We need to show that for *any* function g (not just linear)

$$\min_g E\,(Y - g(X_1,\ldots,X_n))^2 = E\,(Y - E[Y|X_1,\ldots,X_n])^2.$$

As usual, one cleverly adds and substracts, writing (with X for $\{X_1,\ldots,X_n\}$)

$$\begin{aligned}
\mathbb{E}\,(Y-g(X))^2 &= \mathbb{E}\,(Y \pm \mathbb{E}\,[Y|X] - g(X))^2 \\
&= \mathbb{E}\,(Y - \mathbb{E}\,[Y|X])^2 + \mathbb{E}\,(\mathbb{E}\,[Y|X]-g(X))^2 \\
&\quad +2\mathbb{E}\,[(\mathbb{E}\,[Y|X]-g(X))\mathbb{E}\,(Y-\mathbb{E}\,[Y|X])] \\
&= \mathbb{E}\,(Y - \mathbb{E}\,[Y|X])^2 + \mathbb{E}\,(\mathbb{E}\,[Y|X]-g(X))^2 \\
&> \mathbb{E}\,(Y - \mathbb{E}\,[Y|X])^2
\end{aligned}$$

Projection The last step in this proof suggests that we can think of the conditional mean as a projection into a subspace. Let $\mathcal{M}$ denote the closed subspace associated with the $X's$, where by closed we mean random variables Z that can be expressed as functions of the X's. Define a "projection" into $\mathcal{M}$ as

$$P_{\mathcal{M}}Y = \mathbb{E}\,[Y|\{X_1,\ldots X_j,\ldots\}]\ .$$

This operation has the properties seen for projection in a Hilbert space,

1. Linear ($\mathbb{E}\,[aY+bX|Z] = a\mathbb{E}\,[Y|Z] + b\mathbb{E}\,[X|Z]$)
2. Continuous ($Y_n \to Y$ implies $P_{\mathcal{M}}Y_n \to P_{\mathcal{M}}Y$).
3. Nests ($\mathbb{E}\,[Y|X] = \mathbb{E}\,[\ \mathbb{E}\,[Y|X,Z]\ |X]$).

Indeed, we also obtain a form of orthogonality in that we can write

$$Y = Y \pm \mathbb{E}\,[Y|X] = \mathbb{E}\,[Y|X] + (Y - \mathbb{E}\,[Y|X])$$

with

$$\langle \mathbb{E}\,[Y|X],\, Y - \mathbb{E}\,[Y|X]\rangle = 0\ .$$

Since $\mathbb{E}\,[Y|\mathit{nothing}] = \mathbb{E}\,Y$, the subspace $\mathcal{M}$ should contain the constant vector 1 for this sense of projection to be consistent with our earlier definitions.

Tie to regression The fitted values in regression (with a constant) preserve the covariances with the predictors,

$$\text{Cov}(Y,X_j) = \text{Cov}(Y \pm \hat{Y}, X_j) = \text{Cov}(\hat{Y},X_j)\ .$$

Similiarly, for any $Z = g(X_1,\ldots) \in \mathcal{M}$,

$$\mathbb{E}\,[Y\ Z] = \mathbb{E}\,[(Y \pm \mathbb{E}\,[Y|X])\ Z] = \mathbb{E}\,[\ \mathbb{E}\,[Y|X]\ Z\ ]\ . \tag{2}$$

Best linear prediction

Linear projection. We need to make it easier to satisfy the orthogonality conditions. Simplest way to do this is to project onto a space formed by linear operations rather than *any* measureable function. Consider the projection defined as

$$P_{\overline{sp}(1,X_1,\ldots,X_n)} Y = \sum_{j=0}^{n} \alpha_j X_j, \quad X_0 = 1,$$

where the coefficients are chosen as in regression to make the "residual" orthogonal to the X's; that is, the coefficients satisfy the normal equations

$$\langle Y, X_k \rangle = \langle \textstyle\sum_j \alpha_j X_j, X_k \rangle \implies \langle Y - \textstyle\sum_j \alpha_j X_j, X_k \rangle = 0, \quad k = 0, 1, \ldots, n.$$

Note that

- The m.s.e. of the linear projection will be at least as large as that of the conditional mean, and sometimes much more (see below).
- The two are the same if the random variables are Gaussian.

Example Define $Y = X^2 + Z$ where $X, Z \sim N(0,1)$, and are independent. In this case, $E[Y|X] = X^2$ which has m.s.e 1. In contrast, the best linear predictor into $\overline{sp}(1,X)$ is the combination $b_0 + b_1 X$ with, from the normal equations,

$$\begin{aligned} \langle Y, 1 \rangle &= 1 = \langle b_0 + b_1 X, 1 \rangle \\ \langle Y, X \rangle &= 0 = \langle b_0 + b_1 X, X \rangle , \end{aligned}$$

$b_0 = 1$ and $b_1 = 0$. The m.s.e of this predictor is $E(Y-1)^2 = EY^2 - 1 = 3$.

Predictors for ARMA processes

Infinite past In these examples, the Hilbert space is defined by a stationary process $\{X_t\}$. We wish to project members of this space into the closed subspace defined by the process up to time n, $\mathcal{X}_n = \overline{sp}\{X_n, X_{n-1}, \ldots\}$

AR(p) Let $\{X_t\}$ denote the covariance stationary *AR(p)* process

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + w_t$$

where $w_t \sim WN(0, \sigma^2)$. What is the best linear predictor of X_{n+1} in $\mathcal{X}_n$? The prediction/orthogonality equations that the predictor $\hat{X}_{n+1}$ must satisfy are

$$\langle \hat{X}_{n+1},\, X_k \rangle = \langle X_{n+1},\, X_k \rangle, \quad k = n, n-1, \ldots .$$

Since $w_{n+1} \perp \mathcal{X}_n$ we have

$$\begin{aligned} \langle X_{n+1},\, X_j \rangle &= \langle w_t + \textstyle\sum_{j=1}^p \phi_j X_{t-j},\, X_k \rangle \\ &= \langle \textstyle\sum_{j=1}^p \phi_j X_{t-j},\, X_k \rangle, \end{aligned}$$

so that $\hat{X}_{n+1} = \sum_{j=1}^p \phi_j X_{t-j}$. These lead back to the *Yule-Walker equations*. Note that this argument does not require that the order of the autoregression p be finite.

MA(1) Let $\{X_t\}$ denote the *invertible MA(1)* process

$$X_t = w_t - \theta w_{t-1}$$

with $|\theta| < 1$ and $w_t \sim WN(0, \sigma^2)$. Since the process is invertible, express it as the autoregression $(1 - \theta\, B)^{-1} X_t = w_t$, or

$$X_t = w_t - \theta X_{t-1} - \theta^2 X_{t-2} - \ldots .$$

From the AR example, it follows that $\hat{X}_{n+1} = -\sum_{j=1}^\infty \theta^j X_{n-j+1}$.

Role for Kalman filter? What about conditioning on the finite past? That's what the Kalman filter is all about.