

Hilbert Spaces

Overview

1. Ideas
2. Hilbert spaces
3. Projection
4. Orthonormal bases
5. Separability and the fundamental isomorphism
6. Applications to random variables

Ideas

Rationale for studying Hilbert spaces:

1. Formalize intuition from regression
2. Define infinite sums 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, $\langle x, y \rangle$

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

for a suitably defined function f known as the spectral density function.

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.

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)

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

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

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. Any ℓ_2 sequence (square summable sequence). 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 $\langle f, g \rangle = \int f \bar{g}$ 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 $f \bar{g}$ 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 the the (squared) length of a vector in $\mathbb{R} 2^n$. 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 $x \in \mathcal{H}$,

$$\|x\|^2 = \sum_{j=1}^n |\langle x, x_j \rangle|^2 + \|x - \sum_j \langle x, x_j \rangle x_j\|^2. \tag{1}$$

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 (??) holds. The coefficients $\langle x, x_j \rangle$ seen in (??) are known as Fourier coefficients.

Bessel's inequality If $\mathcal{X} = \{x_1, \dots, x_n\}$ 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, $\langle x_n, y_n \rangle \rightarrow \langle x, y \rangle$ (P2.1.2). 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. Pythagorean theorem: $\|x\|^2 = \sum_{j=1}^n |\langle x, x_j \rangle|^2 + \|x - \sum_j \langle x, x_j \rangle x_j\|^2$.
2. Bessel's inequality: $\|x\|^2 \geq \sum_j |\langle x, x_j \rangle|^2$.
3. Cauchy-Schwarz inequality: $|\langle x, y \rangle| \leq \|x\| \|y\|$.
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:

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

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:

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

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 $\langle x - \hat{x}, y \rangle \neq 0$. 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: $P_{\mathcal{M}}x_n \rightarrow P_{\mathcal{M}}x$ if $x_n \rightarrow x$. (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, \dots, 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!)

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

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, \dots, 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 $A' \subset A$,

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

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 $n \rightarrow \infty$. 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 ,

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

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 continuity of the norm and orthogonality,

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

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

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

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 modified system

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

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

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

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

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

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 measurable complex-valued functions f for which

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

With the inner product $\langle f, g \rangle = \int f\bar{g}dP$, 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.),

$$X_n \rightarrow X \Leftrightarrow \|X_n - X\| \rightarrow 0.$$

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 $\|X_n - X\| \rightarrow 0$, 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: $\lim_n EX_n Y_n = \lim_n \langle X_n, Y_n \rangle = \langle X, Y \rangle = EXY$

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 $\{X_1, \dots, X_n\}$ is the conditional expectation of Y given the X 's. Need to assume that $\text{Var}(Y) < \infty$.

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

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

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

$$\mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y \pm \mathbb{E}[Y|X] - g(X))^2$$

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

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

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

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 ($Y_n \rightarrow Y$ implies $P_{\mathcal{M}}Y_n \rightarrow P_{\mathcal{M}}Y$).
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

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

with

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

Since $\mathbb{E}[Y|\text{nothing}] = \mathbb{E}Y$, 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,

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

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

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

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* measurable function. Consider the projection defined as

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

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

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

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 $\overline{\text{sp}}(1, X)$ 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 , $\mathcal{X}_n = \overline{\text{sp}}\{X_n, X_{n-1}, \dots\}$

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

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

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

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

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

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

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.