STOCKHOLM DOCTORAL PROGRAM IN ECONOMICS

Handelshögskolan i Stockholm   Stockholms universitet

Paul Klein[1]
Email: pklei@iies.su.se
URL: http://sites.google.com/site/matfu2site

Hilbert spaces and the projection theorem

We now introduce the notion of a projection. Projections are used a lot by economists; they define linear regressions as well as the conditional expectation.

1 Vector spaces

Definition 1. A *vector space* is a non-empty set $\mathcal{S}$ associated with an addition operation $+ : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$ and a scalar multiplication operation $\cdot : \mathbb{R} \times \mathcal{S} \to \mathcal{S}$ such that for all $x, y \in \mathcal{S}$ and all scalars α, $\alpha(x+y) \in \mathcal{S}$. For $+$ and $\cdot$ to qualify as addition and scalar multiplication operators, we require the following axioms.

1. $x + y = y + x$ for all $x, y \in \mathcal{S}$
2. $x + (y + z) = (x + y) + z$ for all $x, y, z \in \mathcal{S}$
3. There exists an element $\theta \in \mathcal{S}$ such that $x + \theta = x$ for all $x \in \mathcal{S}$
4. $\alpha(x + y) = \alpha x + \alpha y$ for all $\alpha \in \mathbb{R}$ and all $x, y \in \mathcal{S}$
5. $(\alpha + \beta) x = \alpha x + \beta x$ for all $\alpha, \beta \in \mathbb{R}$ and all $x \in \mathcal{S}$

[1] Thanks to Mark Voorneveld for helpful comments.

6. $\alpha\,(\beta x) = (\alpha\beta)\,x$ for all $\alpha, \beta \in \mathbb{R}$ and all $x \in \mathcal{S}$

7. $0x = \theta$ for all $x \in \mathcal{S}$

8. $1x = x$ for all $x \in \mathcal{S}$

Definition 2. A *norm* on a vector space $\mathcal{S}$ is a function $\|\cdot\| : \mathcal{S} \to \mathbb{R}$ such that

1. For each $x \in \mathcal{S}$, $\|x\| \geq 0$ with equality for and only for the zero element $\theta \in \mathcal{S}$.
2. For each $x \in \mathcal{S}$ and every scalar α, $\|\alpha x\| = |\alpha|\|x\|$.
3. (The triangle inequality.) For all $x, y \in \mathcal{S}$, $\|x + y\| \leq \|x\| + \|y\|$.

Proposition 1. Let $(\mathcal{S}, \|\cdot\|)$ be a vector space with an associated norm. Define the function $\mu : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ via

$$\mu\,(x, y) = \|x - y\| \tag{1}$$

Then $(\mathcal{S}, \mu)$ is a metric space, and μ is called the metric generated by $\|\cdot\|$.

Proof. Exercise. ■

Definition 3. A *Banach space* is normed vector space $(\mathcal{S}, \|\cdot\|)$ that is complete in the metric generated by the norm $\|\cdot\|$.

Obvious examples of Banach spaces include the Euclidean spaces. A less obvious example is the following. Let X be the set of bounded and continuous real-valued functions $f : [0, 1] \to \mathbb{R}$. We define addition and scalar multiplication via

$$(f + g)(x) = f(x) + g(x)$$

and

$$(\alpha f)(x) = \alpha \cdot f(x),$$

and the norm via

$$\|f\| = \sup_{x \in [0,1]} |f(x)|.$$

2 Hilbert spaces

An *inner product* on a vector space $\mathcal{H}$ is a function from $\mathcal{H} \times \mathcal{H}$ into $\mathbb{R}$ such that, for all $x, y \in \mathcal{H}$,

1. $(x, y) = (y, x)$.
2. For all scalars α, β, $(\alpha x + \beta y, z) = \alpha (x, z) + \beta (y, z)$.
3. $(x, x) \geq 0$ with equality iff $x = \theta$.

Proposition 2. The function $\| \cdot \|$ defined via $\|x\| = \sqrt{(x, x)}$ is a norm. This norm is called the norm *generated* by $(\cdot, \cdot)$.

Proof. To show the triangle inequality, square both sides and use the Cauchy-Schwarz inequality (see below). ■

Definition 4. A Hilbert space is a pair $(\mathcal{H}, (\cdot, \cdot))$ such that

1. $\mathcal{H}$ is a vector space.
2. $(\cdot, \cdot)$ is an inner product.
3. The normed space $(\mathcal{H}, \| \cdot \|)$ is complete, where $\| \cdot \|$ is the norm generated by $(\cdot, \cdot)$.

Henceforth whenever the symbol $\mathcal{H}$ appears, it will denote a Hilbert space (with some associated inner product).

Intuitively, a Hilbert space is a generalization of $\mathbb{R}^n$ with the usual inner product $(x, y) = \sum_{i=1}^{n} x_i y_i$ (and hence the Euclidean norm), preserving those properties which have to do with *geometry* so that we can exploit our ability to visualize (our intuitive picture of) physical space in order to deal with problems that have nothing whatever to do with physical space. In particular, the ideas of *distance, length,* and *orthogonality* are preserved. As expected, we have the following.

Definition 5. The distance between two elements $x, y \in \mathcal{H}$ is defined as $\|x - y\|$, where $\| \cdot \|$ is the norm generated by the inner product associated with $\mathcal{H}$.

Definition 6. Two elements $x, y \in \mathcal{H}$ are said to be orthogonal if $(x, y) = 0$.

Some further definitions and facts are needed before we can proceed.

Definition 7. A Hilbert subspace $\mathcal{G} \subset \mathcal{H}$ is a subset $\mathcal{G}$ of $\mathcal{H}$ such that $\mathcal{G}$, too, is a Hilbert space.

The following proposition is very useful, since it guarantees the well-definedness of the inner product between two elements of finite norm.

Proposition 3. (The Cauchy-Schwarz inequality). Let $\mathcal{H}$ be a Hilbert space. Then, for any $x, y \in \mathcal{H}$, we have

$$|(x, y)| \leq \|x\| \|y\| \tag{2}$$

Proof. If $x = \theta$ or $y = \theta$ the inequality is trivial. So suppose $\|x\|, \|y\| > 0$ and let $\lambda > 0$ be a real number. We get

$$\begin{aligned} 0 \leq \| x - \lambda y\|^2 &= (x - \lambda y, x - \lambda y) = \\ &= \|x\|^2 + \lambda^2 \|y\|^2 - 2\lambda \, (x, y) \end{aligned} \tag{3}$$

Dividing by λ, it follows that

$$2 \, (x, y) \leq \frac{1}{\lambda} \|x\|^2 + \lambda \|y\|^2 \tag{4}$$

Clearly this is true for all $\lambda > 0$. In particular it is true for $\lambda = \dfrac{\|x\|}{\|y\|}$ which is strictly positive by assumption. It follows that $(x, y) \leq \|x\| \|y\|$. To show that $-(x, y) \leq \|x\| \|y\|$, note that $-(x, y) = (-x, y)$, and since $(-x) \in \mathcal{H}$ we have just shown that $(-x, y) \leq \|x\| \|y\|$. ■

Proposition 4. [The parallelogram identity.] Let $\mathcal{H}$ be a Hilbert space. Then, for any $x, y \in \mathcal{H}$, we have

$$\|x+y\|^2 + \|x-y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right) \tag{5}$$

Proof. Exercise. ■

The following proposition states that the inner product is (uniformly) continuous.

Proposition 5. Let $\mathcal{H}$ be a Hilbert space, and let $y \in \mathcal{H}$ be fixed. Then for each $\varepsilon > 0$ there exists a $\delta > 0$ such that $|(x_1, y) - (x_2, y)| \leq \varepsilon$ for all $x_1, x_2 \in \mathcal{H}$ such that $\|x_1 - x_2\| \leq \delta$.

Proof. If $y = \theta$ the result is trivial, so let $y \neq \theta$.

$$\begin{aligned} |(x_1, y) - (x_2, y)| &= |(x_1 - x_2, y)| \leq \{\text{Cauchy-Schwarz}\} \\ &\leq \|x_1 - x_2\|\|y\| \end{aligned} \tag{6}$$

Set $\delta = \dfrac{\varepsilon}{\|y\|}$, and we are done. ■

Definition 8. Let $\mathbb{G} \subset \mathcal{H}$ be an arbitrary subset of $\mathcal{H}$. Then the orthogonal complement $\mathbb{G}^{\perp}$ of $\mathbb{G}$ is defined via $\mathbb{G}^{\perp} = \{y \in \mathcal{H} : (x, y) = 0 \text{ for all } x \in \mathbb{G}\}$.

Proposition 6. The orthogonal complement $\mathbb{G}^{\perp}$ of a subset $\mathbb{G}$ of a Hilbert space $\mathcal{H}$ is a Hilbert subspace.

Proof. It is easy to see that $\mathbb{G}^{\perp}$ is a vector space. By the fact that any closed subset of a Hilbert space is complete, all that remains to be shown is that $\mathbb{G}^{\perp}$ is closed. To show that, we use the continuity of the inner product. Consider first the set

$$x^{\perp} = \{y \in \mathcal{H} : (x, y) = 0\} \tag{7}$$

But this is easily recognized as the inverse image of the closed set $\{0\}$ under a continuous mapping. Hence $x^{\perp}$ is closed. Now notice that

$$\mathbb{G}^{\perp} = \bigcap_{x \in \mathbb{G}} x^{\perp} \tag{8}$$

Hence $\mathbb{G}^{\perp}$ is the intersection of a family of closed sets. It follows that $\mathbb{G}^{\perp}$ is itself closed. ■

We now state the most important theorem in Hilbert space theory.

2.1 The projection theorem

Theorem 1. [The projection theorem.] Let $\mathcal{G} \subset \mathcal{H}$ be a Hilbert subspace and let $x \in \mathcal{H}$. Then

1. There exists a unique element $\hat{x} \in \mathcal{G}$ (called the *projection* of x onto $\mathcal{G}$) such that
$$\|x - \hat{x}\| = \inf_{y \in \mathcal{G}} \|x - y\| \tag{9}$$
where $\| \cdot \|$ is the norm generated by the inner product associated with $\mathcal{H}$.
2. $\hat{x}$ is (uniquely) characterized by
$$(x - \hat{x}) \in \mathcal{G}^{\perp} \tag{10}$$

Proof. In order to prove part 1 we begin by noting that $\mathcal{G}$, since it is a Hilbert subspace, is both complete and convex. Now fix $x \in \mathcal{H}$ and define

$$d = \inf_{y \in \mathcal{G}} \|x - y\|^2 \tag{11}$$

Clearly d exists since the set of squared norms $\|x - y\|^2$ is a set of real numbers bounded below by 0. Now since d is the greatest lower bound of $\|x - y\|^2$ there exists a sequence $(y_k)_{k=1}^{\infty}$ from $\mathcal{G}$ such that, for each $\varepsilon > 0$, there exists an N_ε such that

$$\|x - y_k\|^2 \leq d + \varepsilon \tag{12}$$

for all $k \geq N_\varepsilon$. We now want to show that any such sequence (y_k) is a Cauchy sequence. For that purpose, define

$$\begin{aligned} u &= x - y_m & (13) \\ v &= x - y_n & (14) \end{aligned}$$

Now applying the parallelogram identity to u and v, we get

$$\|2x - y_m - y_n\|^2 + \|y_n - y_m\|^2 = 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) \tag{15}$$

which may be manipulated to become

$$4\|x - \frac{1}{2}\left(y_m + y_n\right)\|^2 + \|y_m - y_n\|^2 = 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) \tag{16}$$

Now since $\mathcal{G}$ is convex, $\dfrac{1}{2}\left(y_m + y_n\right) \in \mathcal{G}$ and consequently $\|x - \frac{1}{2}\left(y_m + y_n\right)\|^2 \geq d$. It follows that

$$\|y_m - y_n\|^2 \leq 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) - 4d \tag{17}$$

Now consider any $\varepsilon > 0$, choose a corresponding N_ε such that $\|x - y_k\|^2 \leq d + \varepsilon/4$ for all $k \geq N_\varepsilon$ (such an N_ε exists as we have seen). Then, for all $n, m \geq N_\varepsilon$, we have

$$\|y_m - y_n\|^2 \leq 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) - 4d \leq \varepsilon \tag{18}$$

Hence (y_k) is a Cauchy sequence. By the completeness of $\mathcal{G}$, it converges to some element $\widehat{x} \in \mathcal{G}$. By the continuity of the inner product, $\|x - \widehat{x}\|^2 = d$. Hence $\widehat{x}$ is the projection we seek. To show that $\widehat{x}$ is unique, consider another projection $y \in \mathcal{G}$ and the sequence $(\widehat{x}, y, \widehat{x}, y, \widehat{x}, y...)$. By the argument above, this is a Cauchy sequence. But then $\widehat{x} = y$. Hence (1) is proved. The proof of part (2) comes in two parts. First we show that any $\widehat{x}$ that satisfies (9) also satisfies (10). Suppose, then, that $\widehat{x}$ satisfies (9). Define $w = x - \widehat{x}$ and consider an element $y = \widehat{x} + \alpha z$ where $z \in \mathcal{G}$ and $\alpha \in \mathbb{R}$. Since $\mathcal{G}$ is a vector space, it follows that $y \in \mathcal{G}$. Now since $\widehat{x}$ satisfies (9), y is no closer to x than $\widehat{x}$ is. Hence

$$\begin{aligned} \|w\|^2 &\leq \|w - \alpha z\|^2 = (w - \alpha z, w - \alpha z) = \\ &= \|w\|^2 + \alpha^2\|z\|^2 - 2\alpha\left(\varepsilon, z\right) \end{aligned} \tag{19}$$

Simplifying, we get

$$0 \leq \alpha^2\|z\|^2 - 2\alpha\left(w, z\right) \tag{20}$$

This is true for all scalars α. In particular, set $\alpha = (\varepsilon, z)$. We get

$$0 \leq (w, z)^2\left(\|z\|^2 - 2\right) \tag{21}$$

For this to be true for all $z \in \mathcal{G}$ we must have $(w,z) = 0$ for all $z \in \mathcal{G}$ such that $\|z\|^2 < 2$. But then (why?) we must have $(w,z) = 0$ for all $z \in \mathcal{G}$. Hence $w \in \mathcal{G}^\perp$. Now we want to prove the converse, i.e. that if $\widehat{x}$ satisfies (10), then it also satisfies (9). Thus consider an element $\widehat{x} \in \mathcal{G}$ which satisfies (10) and let $y \in \mathcal{G}$. Mechanical calculations reveal that

$$\begin{aligned} \|x-y\|^2 &= (x-\widehat{x}+\widehat{x}-y, x-\widehat{x}+\widehat{x}-y) = \\ &= \|x-\widehat{x}\|^2 + \|\widehat{x}-y\|^2 + 2\,(x-\widehat{x}, \widehat{x}-y) \end{aligned} \tag{22}$$

Now since $(x-\widehat{x}) \in \mathcal{G}^\perp$ and $(\widehat{x}-y) \in \mathcal{G}$ (recall that $\mathcal{G}$ is a vector space), the last term disappears, and our minimization problem becomes (disregarding the constant term $\|x-\widehat{x}\|^2$)

$$\min_{y \in \mathcal{G}} \|\widehat{x}-y\| \tag{23}$$

Clearly $\widehat{x}$ solves this problem. (Note that it doesn't matter for the solution whether we minimize a norm or its square.) Indeed, since $\|\widehat{x}-y\| = 0$ implies $\widehat{x} = y$ we may conclude that if some $\widehat{x}$ satisfies (10), then it is the *unique* solution. ■

A good way to understand intuitively why the projection theorem is true is to visualize the projection of a point in $\mathbb{R}^3$ onto a 2-dimensional plane through the origin.

Corollary 1. [The repeated projection theorem.] Let $\mathcal{G} \subset \mathcal{F} \subset \mathcal{H}$ be Hilbert spaces. Let $\hat{x}_G$ be the projection of $x \in \mathcal{H}$ onto $\mathcal{G}$ and let $\hat{x}_F$ be the projection of x onto $\mathcal{F}$. Then the projection of $\hat{x}_F$ onto $\mathcal{G}$ is simply $\hat{x}_G$.

Proof. By the characterization of the projection, it suffices to prove that $(\hat{x}_F - \hat{x}_G) \in \mathcal{G}^\perp$. To do this, we rewrite $(\hat{x}_F - \hat{x}_G) = (x - \hat{x}_G) - (x - \hat{x}_F)$. Since the orthogonal complement of a Hilbert subspace is a vector space and hence closed under addition and scalar multiplication, it suffices to show that $(x - \hat{x}_G) \in \mathcal{G}^\perp$ and that $(x - \hat{x}_F) \in \mathcal{G}^\perp$. The truth of the first statement follows immediately from the characterization of the projection. The truth of the second one can be seen by noting that $\mathcal{F}^\perp \subset \mathcal{G}^\perp$. ■

3 Applications of Hilbert space theory

3.1 Projecting onto a line through the origin

A *line through the origin* in $\mathbb{R}^n$ is a set that, for some fixed $x_0 \in \mathbb{R}^n$, can be written as

$$L = \{x \in \mathbb{R}^n : x = \alpha x_0 \text{ for some } \alpha \in \mathbb{R}\}$$

Now let's find the projection $\widehat{y}$ of an arbitrary $y \in \mathbb{R}^n$ onto L. Evidently $\widehat{y} = \alpha x_0$ for some scalar α. But what scalar? Orthogonality of the error vector says that

$$(y - \alpha x_0) \cdot x_0 = 0.$$

Solving for α, we get

$$\alpha = \frac{x_0 \cdot y}{\|x_0\|^2}$$

and hence

$$\widehat{y} = \frac{x_0 \cdot y}{\|x_0\|^2} x_0.$$

What if $y \in L$? For example, what if we project $\widehat{y}$ onto L? Then we had better get the same thing back! Fortunately, that is true. To see that, suppose $y = tx_0$. We get

$$\alpha = \frac{tx_0 \cdot x_0}{\|x_0\|^2} = t$$

proving the assertion.

How would you proceed if the line doesn't go through the origin? You will find some hints in the next section.

3.2 Projecting onto a hyperplane

Definition 9. A hyperplane in $\mathbb{R}^n$ is a set of the form

$$H = \{x \in \mathbb{R}^n : p \cdot x = \alpha\}.$$

A hyperplane is not a subspace unless $\alpha = 0$. But it is convex and complete, so the problem

$$\min_{z \in H} \|x - z\|$$

has a unique solution for any $x \in \mathcal{H}$. The only problem is how to characterize it. Intuitively, we should have orthogonality of the error with every vector *along* (rather than *in*) the plane. Meanwhile, every vector that is orthogonal to every vector along the plane should point in the direction of the normal vector of the plane. So the error vector should point in the direction of the normal vector of the plain:

$$x - \widehat{x}_H = tp$$

for some $t \in \mathbb{R}$.

To show that rigorously, we proceed as follows. Let $y \in H$. By definition, every $z \in H$ can be written as $z = y + w$ where $(w, p) = 0$ i.e. $w \in \{p\}^\perp$. But then our projection problem can be written as

$$\min_{w \in \{p\}^\perp} \|x - y - w\|.$$

Evidently the solution to this problem is the projection of $x - y$ onto $\{p\}^\perp$, which is a Hilbert space by Proposition 6. The projection theorem says that the error $x - y - w$ lies in the orthogonal complement of $\{p\}^\perp$, which is just the span of $\{p\}$, i.e. $x - y - w = tp$ for some $t \in \mathbb{R}$. This establishes the intuitive claim.

Moroever, since $\widehat{x}_H \in H$ we have

$$(\widehat{x}_H, p) = \alpha.$$

It follows that

$$(x - \widehat{x}_H, p) = (p, x) - \alpha$$

Since we already know that

$$(x - \widehat{x}_H, p) = (tp, p) = t\|p\|^2$$

we may conclude that

$$(p, x) - \alpha = t\|p\|^2$$

which can easily be solved for t and we get

$$x - \widehat{x}_H = \frac{(p, x) - \alpha}{\|p\|^2} p$$

and, finally,

$$\widehat{x}_H = x - \frac{(p, x) - \alpha}{\|p\|^2} p. \tag{24}$$

Can you show that if $x \in H$ then x is unchanged by projection onto H?

3.3 Ordinary least squares estimation as projection

Let $y \in \mathbb{R}^m$, let X be an $m \times n$ matrix. Denote the ith row of X by x_i. To make the problem interesting, let $n < m$. Suppose we want to minimize the sum of squared errors with respect to the parameter vector $\beta \in \mathbb{R}^n$.

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{m} (y_i - x_i \beta)^2 = (y - X\beta)'(y - X\beta) \tag{25}$$

We can solve this by doing some matrix differential calculus to compute the first order condition. What we get is

$$X'X\beta = X'y$$

and it is worth knowing how to do this for its own sake. However, an alternative is to think about (25) as a projection problem. What we want to do is find the point $\widehat{y}$ in the column space of X that minimizes the distance between y and $\widehat{y}$. In other words, we want to project y onto the subspace

$$G = \{z \in \mathbb{R}^m : z = X\beta \text{ for some } \beta \in \mathbb{R}^n\}.$$

According to the projection theorem, the error vector $\varepsilon = y - X\beta$ is orthogonal to every member of G. In particular, it is orthogonal to every column of X. Formally,

$$X'(y - X\beta) = 0.$$

Suppose $X'X$ has an inverse. Then the projection $\widehat{y} = X\beta$ can be written as

$$\widehat{y} = X(X'X)^{-1}X'y$$

and we call the matrix

$$P = X(X'X)^{-1}X'$$

a projection matrix.

It should be obvious that repeated multiplication by P shouldn't do anything beyond the first application. That's easy to prove.

$$P^2 = PP = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = P.$$

This property of P is called **idempotence**, Greek for self-power.

3.4 Projecting onto an arbitrary subspace of $\mathbb{R}^n$

Let $V \subset \mathbb{R}^n$ be a subspace with basis $\{v^1, v^2, \ldots, v^m\}$. Let $x \in \mathbb{R}^n$ be arbitrary. We now seek an expression for $\widehat{x}_V$, the projection of x onto V. Evidently (why?) it suffices that the error vector be orthogonal to each of the basis vectors v^k.

$$(v^k) \cdot (x - \widehat{x}_V) = 0$$

for each $k = 1, 2, \ldots, m$. Meanwhile, since $\widehat{x}_V \in V$, there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_m$ such that

$$\widehat{x}_V = \alpha_1 v^1 + \alpha_2 v^2 + \ldots + \alpha_m v^m.$$

Inserting this expression into the orthogonality conditions and using matrix notation, we get

$$A^T(x - A\alpha) = 0$$

where

$$A = \left[\begin{array}{cccc} v^1 & v^2 & \ldots & v^m \end{array} \right]$$

which means that A is an $n \times m$ matrix of rank $m \leq n$. It follows that $A^T A$ is an $m \times m$ invertible matrix.

Solving for α, we get

$$\alpha = (A^T A)^{-1} A^T x$$

and it follows that

$$\widehat{x}_V = A(A^T A)^{-1} A^T x.$$

In this context I want you to notice two things:

1. The projection mapping, i.e. the function $f : \mathbb{R}^n \to V$ that takes you from x to $\widehat{x}_V$ is linear and is represented by the matrix $P = A(A^T A)^{-1} A^T$.

2. The matrix P is symmetric ($P^T = P$) and idempotent ($P^2 = P$). Can you show that any $n \times n$ matrix with these properties represents a projection mapping onto some subspace? What subspace is that?

The lessons of this section can be used to find the projection onto a hyperplane through the origin in a slightly more straightforward fashion than we did above. To see how, we first need to notice the following. Not only is the projection mapping linear, but the mapping to the error is linear also. In fact the following is true for any $x \in \mathbb{R}^n$ and any subspace $V \subset \mathbb{R}^n$.

$$x = Px + Qx$$

where Px is the projection of x onto V and Qx is the projection of x onto $V^\perp$. We can use this to find a shortcut to the formula for the projection onto a hyperplane H through the origin with normal vector p. Instead of finding P directly, we will look for Q and then note that $P = I - Q$. Evidently $\{p\}$ is a basis for $H^\perp$. Hence

$$Q = p(p^T p)^{-1} p^T = \frac{1}{\|p\|^2} p p^T$$

and hence

$$\widehat{x}_H = (I - \frac{1}{\|p\|^2} p p^T) x = x - \frac{p^T x}{\|p\|^2} p$$

which is of course the same formula as Equation (24) when the hyperplane goes through the origin, i.e. is a subspace of $\mathbb{R}^n$.

3.5 Minimal mean squared error prediction

The results in this subsection are given and derived in Judge et al. (1988), pages 48-50. Some of what I say relies on abstract probability theory; if you can't follow now, then stay tuned for a later lecture on that topic.

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and let $Z : \Omega \to \mathbb{R}^n$ be a non–degenerate normally distributed vector. What this means is that there is a vector $\mu \in \mathbb{R}^n$ and a positive definite square matrix $\Sigma \in \mathbb{R}^{n \times n}$ such that, for each Borel set $B \subset \mathbb{R}^n$,

$$\mathsf{P}(Z \in B) = \int_B f(x)\, dm(x)$$

where m is the Lebesgue measure on $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}_+$ is given by

$$f(x) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} (x-\mu)' \, \Sigma^{-1} (x-\mu) \right\}.$$

Now partition

$$Z = \left[\begin{array}{c} X \\ Y \end{array} \right]$$

$$\mu = \left[\begin{array}{c} \mu_X \\ \mu_Y \end{array} \right]$$

and

$$\Sigma = \left[\begin{array}{cc} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \Sigma_{yy}. \end{array} \right]$$

Proposition 7. Let X, Y, Z be as above. Then the conditional expectation (which by definition coincides with the minimal mean square error predictor) is given by

$$\mathsf{E}[Y|X] = \mu_Y + \Sigma_{xy}' \Sigma_{xx}^{-1} (X - \mu_x). \tag{26}$$

Proposition 8. Let X, Y, Z be as above except that they might not be normal. Then (26) delivers the best (in the sense of minimal mean square error) affine predictor, and the error is uncorrelated with (but not necessarily independent of) X.

To derive this formula, we use the orthogonality property of the projection. Suppose, for simplicity, that μ_X and μ_Y are both zero and let P be the unknown matrix such that

$$\mathsf{E}[Y|X] = PX.$$

Apparently the error $Y - PX$ is orthogonal to every random variable that is a (Borel) function of X. In particular, it is orthogonal to every element of X itself. Thus

$$\mathsf{E}[X(Y - PX)'] = 0.$$

Notice that this is an $n_x \times n_y$ matrix of equations. It follows that

$$\mathsf{E}[XY'] - \mathsf{E}[XX']P' = 0.$$

Solving for P, we get

$$P = \mathsf{E}[YX']\mathsf{E}[XX']^{-1}.$$

Incidentally, the Hilbert space we are talking about here is the set of square integrable random variable. This is indeed a Hilbert space provided that we identify any two random variables that are equal almost surely.

3.6 Fourier analysis

In time series analysis, an interesting problem is how to decompose the variance of a series by contribution from each frequency. How important are seasonal fluctuations compared to business cycle fluctuations? Are cycles of length 3-5 years more important than longer term or shorter term fluctuations? A proper answer to this question is given by the *spectral analysis of time series*. A good book on that is Brockwell and Davis (1998). The basis for spectral analysis is an excellent 18th century idea: Fourier analysis.

3.6.1 General Fourier analysis

You will recall from linear algebra in $\mathbb{R}^n$ that a useful way of representing a vector $x \in \mathbb{R}^n$ is as a linear combination of some orthogonal basis $\mathbb{B} = \{x_k;\ k = 1, 2, ..., n\}$. You will remember that if $\mathbb{B}$ spans $\mathbb{R}^n$, then for each $x \in \mathbb{R}^n$ there exist scalars $\{\varphi_k;\ k = 1, 2, ..., n\}$ such that

$$x = \sum_{k=1}^{n} \varphi_k x_k \tag{27}$$

where the x_k are the elements of $\mathbb{B}$. A popular choice of orthogonal basis vectors is of course the unit vectors, but there are many other orthogonal bases of $\mathbb{R}^n$.

These ideas can easily be generalized to any Hilbert space, since the only essential part of the whole program is orthogonality.

Definition 10. Let $(\mathcal{H}, (\cdot, \cdot))$ be a Hilbert space and let $\mathbb{B} = \{x_k : k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Then $\mathbb{B}$ is called *orthogonal* if, for all $j \neq k$ we have

$$(x_j, x_k) = 0. \tag{28}$$

If, in addition, $(x_j, x_j) = 1$ for all $j = 1, 2, ...$, then $\mathbb{B}$ is called *orthonormal.*

Definition 11. Let $(\mathcal{H}, (\cdot, \cdot))$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Then the *closed span* $\overline{\mathsf{sp}}(\mathbb{B})$ of $\mathbb{B}$ is defined as follows. Let $\mathsf{sp}(\mathbb{B})$ be the set of elements $x \in \mathcal{H}$ such that there exist scalars $\{\varphi_k;\ k = 1, 2, ...\}$ with the property that

$$x = \sum_{k=1}^{n} \varphi_k x_k \tag{29}$$

where the x_k are elements of $\mathbb{B}$. We then define $\overline{\mathsf{sp}}(\mathbb{B})$ as the closure of $\mathsf{sp}(\mathbb{B})$. Intuitively, taking the closure means that we include all the infinite sums as well as the finite ones.

Definition 12. Let $(\mathcal{H}, (\cdot, \cdot))$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Then $\mathbb{B}$ is said to *span* $\mathcal{H}$ if $\overline{\mathsf{sp}}(\mathbb{B}) = \mathcal{H}$.

Proposition 9. Let $(\mathcal{H}, (\cdot,\cdot))$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Suppose $\mathbb{B}$ spans $\mathcal{H}$. Then, for each $x \in \mathcal{H}$, there are scalars $\{\varphi_k;\ k = 1, 2, ...\}$ such that

$$\lim_{n\to\infty} \left\| x - \sum_{k=1}^{n} \varphi_k x_k \right\| = 0 \tag{30}$$

and we sometimes write

$$x = \sum_{k=1}^{\infty} \varphi_k x_k. \tag{31}$$

Proof. Obvious. ■

Definition 13. Let $(\mathcal{H}, (\cdot,\cdot))$ be a Hilbert space and suppose $\mathbb{B}$ is orthogonal (orthonormal) and spans $\mathcal{H}$. Then $\mathbb{B}$ is called an orthogonal (orthonormal) basis for $\mathcal{H}$.

Definition 14. A Hilbert space $(\mathcal{H}, (\cdot,\cdot))$ is called separable if it has a countable dense subset.

Proposition 10. A Hilbert space has a countable orthogonal basis iff it is separable.
Proof. Omitted. ■

Spanning is hard to check directly, but often we have a standard result to fall back on. An example is the Stone-Weierstrass theorem which asserts that the set of polynomials of arbitrary degree spans the set of continuous functions on an arbitrary interval $[a, b]$. We also have the remarkable result that the set of trigonometric polynomials (linear combinations of $\sin nx$ and $\cos nx$ where n is an arbitrary integer) spans the set of square integrable functions defined on a compact interval; see below.

In any case, suppose $(\mathcal{H}, (\cdot,\cdot))$ is a Hilbert space with the (countable!) orthonormal subset $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$. Now consider an arbitrary element $x \in \mathcal{H}$. Our project now is to find the projection $\widehat{x}$ onto $\overline{\mathsf{sp}}\,(\mathbb{B})$. By the definition of the closed span, there are scalars $\{\varphi_k;\ k=1,2,...\}$ such that

$$\widehat{x}=\sum_{k=1}^{\infty}\varphi_k x_k. \tag{32}$$

The scalars $\{\varphi_k;\ k=1,2,...\}$ are called the Fourier coefficients of x (with respect to $\mathbb{B}$). But what values do they have? To find out, recall the characterization of the projection. The idea is to choose the φ_k so that the projection error is orthogonal to every vector $y\in\overline{\mathsf{sp}}\left(\mathbb{B}\right)$. Actually, it suffices to set the projection error orthogonal to every vector $x_k\in\mathbb{B}$. Then, for every $j=1,2,...$ we have

$$\left(x-\sum_{k=1}^{\infty}\varphi_k x_k, x_j\right)=0. \tag{33}$$

But

$$\begin{aligned}\left(x-\sum_{k=1}^{\infty}\varphi_k x_k, x_j\right) &= (x,x_j)-\sum_{k=1}^{\infty}\varphi_k\left(x_j,x_k\right)= \\ =\{\text{orthogonality!}\} &= (x,x_j)-\varphi_j\left(x_j,x_j\right)=\{\text{orthonormality!}\}= \\ =(x,x_j)-\varphi_j\end{aligned} \tag{34}$$

Hence $\varphi_j=(x,x_j)$ for each $j=1,2,...$, and we have the following proposition.

Proposition 11. Let $(\mathcal{H},(\cdot,\cdot))$ be a Hilbert space and let $\mathbb{B}=\{x_k;\ k=1,2,...\}$ be a countable orthonormal subset of $\mathcal{H}$. Let $x\in\mathcal{H}$. Then the projection $\widehat{x}$ of x onto $\overline{\mathsf{sp}}\left(\mathbb{B}\right)$ is

$$\widehat{x}=\sum_{k=1}^{\infty}\left(x,x_k\right)x_k. \tag{35}$$

Corollary 2. [Bessel's inequality.] Since (why?) $\|x\|=\|\widehat{x}\|+\|x-\widehat{x}\|$, we have $\|\widehat{x}\|\leq\|x\|$ and consequently

$$\sum_{k=1}^{\infty}\left(x,x_k\right)^2\leq\|x\|^2.$$

Corollary 3. [Parseval's identity.] If $\mathbb{B}$ spans $\mathcal{H}$ then $x = \widehat{x}$ and hence

$$\sum_{k=1}^{\infty} (x, x_k)^2 = \|x\|^2 .$$

Remark. This is a generalization of Pythagoras' theorem.

3.6.2 Classical Fourier analysis

Suppose we have a more or less arbitrary[2] periodic function $f : \mathbb{R} \to \mathbb{R}$ with period 2π. What that means is that

$$f(\omega + 2k\pi) = f(\omega)$$

for all $k \in \mathbb{N}$ and all $\omega \in \mathbb{R}$.

We would like to decompose the fluctuations of this function by pure frequencies. (Apparently our brains can do this when we listen to music!) More precisely, we would like to find coefficients a_k and b_k such that

$$\widehat{f}(\omega) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \sin(k\omega) + b_k \cos(k\omega)]$$

is a good approximation of $f(\omega)$.

To begin with, we define the inner product between two (square integrable) functions on $[-\pi, \pi]$ to be

$$(f, g) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\omega) g(\omega) d\omega.$$

In that case, the functions $\cos(k\omega)$ and $\sin(k\omega)$ are orthonormal. (Suppose again that we identify functions that are equal almost everywhere.)

Hence

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\omega) \cos(n\omega) d\omega, \quad n \geq 0$$

[2] It should be square integrable on $[-\pi, \pi]$.

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\omega) \sin(n\omega) d\omega, \quad n \geq 1$$

3.7 Orthogonal polynomials

We are now interested in approximating arbitrary continuous real-valued functions on various domains that are subsets of $\mathbb{R}$, again identifying functions that are equal almost everywhere.

The motivation for this is that, when solving macroeconomic models, some of the unknowns are functions rather than numbers, and we would like to approximate them by functions that have finite-dimensional representations. When there is a finite set of points ("nodes") where the exact function and the approximating function agree, we call this approach "interpolation". We write

$$f(x) \approx \widehat{f}(x) = \sum_{k=0}^{n} \theta_k \psi_k(x).$$

To avoid problems of collinearity, it makes sense to choose as basis functions ψ_k to be *orthogonal polynomials*.

There are various sets of orthogonal polynomials; they differ with respect to their domains and the weighting function that defines the inner product. Throughout, the space concerned is the set of real-valued functions that are square integrable real-valued functions with respect to a weighting function $w(x)$. More precisely, we define a Hilbert space of continuous functions on an interval $[a, b]$ (not necessarily a bounded one) as follows.

- $(\alpha f)(x) = \alpha f(x)$,
- $(f + g)(x) = f(x) + g(x)$, and
- $(f, g) = \int_a^b w(x) f(x) g(x) dx$.

Provided that this Hilbert space is well-defined, which it is for well-behaved weighting functions, there is a unique (up to a scaling factor) set of orthogonal polynomials, one for each degree $d = 0, 1, 2, \ldots$. What does it mean to be well-behaved? The weighting function should be (strictly) positive and continuous, and all "moments" should exist, i.e.

$$\left| \int_a^b w(x) x^n dx \right| < \infty$$

for all $n = 0, 1, \ldots$.

To understand how a set of orthogonal polynomials is constructed, we begin by recalling the formula for a projection of an arbitrary vector $y \in \mathbb{R}^n$ onto the space spanned by another arbitrary vector $x \in \mathbb{R}^n$. We found that

$$\widehat{y}_x = \frac{x \cdot y}{\|x\|^2} x.$$

This can be generalized to a Hilbert space, where the projection becomes

$$\widehat{y}_x = \frac{(x, y)}{\|x\|^2} x.$$

This is the key idea behind so-called Gram-Schmidt orthogonalization, a method that we will apply to the set of monomials $\{1, x, x^2, \ldots\}$ in order to construct a set of orthogonal polynomials.

The core idea of Gram-Schmidt orthogonalization is to take two vectors and then remove the component of the second that points in the direction of the first, resulting in two orthogonal vectors. Suppose the two vectors are x and y. Now project y onto x, i.e.

$$\widehat{y}_x = \frac{(x, y)}{(x, x)} x.$$

Next, remove this projection from y to define

$$z := y - \frac{(x, y)}{(x, x)} x.$$

The new pair, x and z, should now be orthogonal. Indeed,

$$(x, z) = \left(x, y - \frac{(x, y)}{(x, x)} x \right) = (x, y) - (x, y) = 0.$$

The only thing that could go "wrong" here is if z turns out to be the zero vector. That happens if x and y are collinear, i.e. each is a scalar multiple of the other. Even so, x and z will span the same space as x and y. This process can be repeated for multiple vectors, and that is what we will do in order to introduce a whole sequence of orthogonal polynomials.

We introduce the following notation. Let $q_k(x) \equiv x^k$ be the set of monomials and denote by $\{p_0, p_1, p_2, \ldots\}$ our set of orthogonal polynomials. Notice that the set of monomials is linearly independent, so we don't need to worry about anything going wrong.

We can then define

$$p_k = q_k - \sum_{j=0}^{k-1} \frac{(q_k, p_j)}{(p_j, p_j)} p_j.$$

Can you prove that this defines a set of orthogonal polynomials? Does it matter that $q_k \equiv x^k$?

A more useful, recursive, formula is the following.

$$p_{-1}(x) \equiv 0,$$

$$p_0(x) \equiv 1,$$

and

$$p_{i+1}(x) \equiv (x - \alpha_i)p_i(x) - \beta_i p_{i-1}(x)$$

for $i = 0, 1, 2, \ldots$, where

$$\alpha_i = \frac{(xp_i, p_i)}{(p_i, p_i)}$$

for $i = 0, 1, 2, \ldots$,

$$\beta_0 = 0$$

and

$$\beta_i = \frac{(p_i, p_i)}{(p_{i-1}, p_{i-1})}$$

for $i = 1, 2, 3, \ldots$, Can you prove that this defines a set of orthogonal polynomials?

Denoting the set of polynomials of degree no greater than k by Π_k, note that $(p, p_n) = 0$ for any $p \in \Pi_{n-1}$.

For instance, suppose $[a, b] = [-1, 1]$ and $w(x) \equiv 1$. Then $\alpha_0 = 0$ and we get

$$p_1(x) = (x - 0) \cdot 1 - 0 \cdot 0 = x.$$

Moreover, $\alpha_1 = 0$ and $\beta_1 = \frac{1}{3}$ and so

$$p_2(x) = (x - 0) \cdot x - \frac{1}{3} \cdot 1 = x^2 - \frac{1}{3}$$

where it is perhaps worth noting that any scalar multiple of these polynomials will do just as well. The version defined here implies that the polynomials are all *monic*, i.e. the leading coefficient (the nonzero coefficient of highest degree) is equal to 1. An alternative is to make the polynomials orthonormal, i.e. of norm 1.

Proposition. Suppose $p_0, p_1, \ldots$ is a sequence of orthogonal polynomials (given the weighting function w and the interval $[a, b]$) of degree $0, 1, 2, \ldots$. Then, for each n, p_n has n distinct (real) roots in the interval $[a, b]$.

Proof. Take that subset $\{x_1, x_2, \ldots, x_m\}$ of the roots of p_n that (i) are in the interval (a, b) and (ii) where p_n changes sign. Now construct a new polynomial as follows.

$$q(x) := \prod_{j=1}^{m} (x - x_j)$$

and apparently this polynomial is of degree m. By construction, q is such that $p_n(x)q(x)$ does not change sign in $[a, b]$. Also, it is not identically zero. Hence

$$\int_a^b w(x)p_n(x)q(x)dx \neq 0$$

So it must be that the degree of q, m, is equal to n because otherwise $(p_n, q) = 0$. ■

3.7.1 Laguerre

Laguerre polynomials are defined on the interval $[0, \infty)$ and the inner product is defined via

$$(f, g) = \int_0^\infty e^{-x} f(x)g(x)dx.$$

They can be defined via

$$L_0(x) = 1$$

$$L_1(x) = 1 - x$$

and the recurrence relation, valid for any $k \geq 1$,

$$L_{k+1}(x) = \frac{1}{k+1} \left((2k + 1 - x)L_k(x) - kL_{k-1}(x)\right).$$

The Laguerre polynomials are orthonormal, meaning that their inner product is

$$(L_m, L_n) = \delta_{m,n}$$

where $\delta_{m,n}$ is the Kronecker delta function which takes the value one when $m = n$ and zero otherwise.

3.7.2 Legendre

Legendre polynomials are defined on the interval $[-1, 1]$ and the inner product is defined via

$$(f, g) = \int_{-1}^1 f(x)g(x)dx.$$

They can be defined recursively via

$$P_0(x) = 1$$

$$P_1(x) = x$$

and

$$(n + 1)P_{n+1}(x) = (2n + 1)xP_n(x) - nP_{n-1}(x)$$

These polynomials are orthogonal but not orthonormal. Their inner product is

$$(P_m, P_n) = \frac{2}{2n + 1} \cdot \delta_{m,n}.$$

3.7.3 Chebyshev

Chebyshev polynomials are defined on the interval $[-1, 1]$ and are orthogonal under the following inner product.

$$(f, g) = \int_{-1}^{1} \frac{f(x) \cdot g(x)}{\sqrt{1 - x^2}} dx.$$

One beautiful but fairly useless definition of the Chebyshev polynomials is the following.

$$T_n(x) = \cos(n \arccos x).$$

You may want to amuse yourself by proving that this really does define a sequence of polynomials. Hint: use de Moivre's formula. Incidentally, Chebyshev polynomials are denoted by $T_n(x)$ in honour of Pafnutij Tjebysjoff (Swedish transliteration).

A more helpful definition (because it is more amenable to computation) is the following.

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Can you verify this three-term recursion by using the following well-known trigonometric identity?

$$\cos(\alpha + \beta) + \cos(\alpha - \beta) \equiv 2 \cos \alpha \cos \beta.$$

Hint: Set $\alpha = n\theta$ and $\beta = \theta$ where of course $\theta = \arccos x$.

These polynomials are orthogonal but not orthonormal.

$$(T_n, T_m) = \begin{cases} 0 & \text{if} \quad n \neq m \\ \pi & \text{if} \quad n = m = 0 \\ \pi/2 & \text{if} \quad n = m \neq 0 \end{cases}$$

The n zeros (roots) of T_n can be written as

$$x_k = \cos\left(\frac{2k - 1}{n} \cdot \frac{\pi}{2}\right),$$

where $k = 1, 2, \ldots, n$.

Proof.

$$T_n(x_k) = \cos\left(k\pi - \frac{\pi}{2}\right) = \sin(k\pi) = 0.$$

Moreover, the z_k are distinct and so there can be no other roots. ■

Besides the orthogonality property that the Chebyshev polynomials have by construction, they also have a remarkable *discrete* orthogonality property:

$$\sum_{k=0}^{n-1} T_i(x_k)T_j(x_k) = \begin{cases} 0 & \text{if} \quad i \neq j \\ n & \text{if} \quad i = j = 0 \\ n/2 & \text{if} \quad i = j \neq 0 \end{cases}$$

where the x_k are defined as above.

The discrete orthogonality property means that Chebyshev polynomial interpolation is particularly simple when the nodes are chosen to be the roots of the $(n+1)$th Chebyshev polynomial. If we define $x_0, x_1, \ldots, x_n$ as the $n+1$ zeros of T_{n+1} and

$$\theta_0 = \frac{1}{n+1}\sum_{k=0}^{n} f(x_k)$$

and, for $k = 1, 2, \ldots, n$,

$$\theta_k = \frac{2}{n+1}\sum_{i=0}^{n} T_k(x_i)f(x_i),$$

then

$$f(x_k) = \widehat{f}(x_k) = \sum_{k=0}^{n} \theta_k T_k(x_k)$$

for all $k = 0, 1, \ldots, n$.

3.7.4 Hermite

Hermite polynomials are defined on $(-\infty, \infty)$ and are orthogonal with respect to the inner product

$$(f, g) = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} f(x)g(x)dx.$$

They can be defined via

$$H_0(x) = 1,$$

$$H_1(x) = x$$

and

$$H_{n+1}(x) = xH_n(x) - nH_{n-1}(x).$$

The Hermite polynomials are orthogonal but not orthonormal;

$$(H_n, H_n) = n!\sqrt{2\pi}.$$

Hermite polynomial approximation is ideally suited for computing approximate (conditionally) expected values when the source of uncertainty is a normally distributed random variable. In fact, if X is a normal random variable with mean μ and variance 1 then

$$\mathsf{E}[H_n(X)] = \mu^n.$$

References

Brockwell, P. and R. Davis (1998). *Time Series: Theory and Methods*. New York: Springer-Verlag.

Judge, G. G., R. Hill, W. E. Griffiths, H. Lütkepohl, and T.-C. Lee (1988). *Introduction to the Theory and Practice of Econometrics*. John Wiley and Sons.