CHAPTER 2
Hilbert Spaces

Although it is possible to study time series analysis without explicit use of Hilbert space terminology and techniques, there are great advantages to be gained from a Hilbert space formulation. These are largely derived from our familiarity with two- and three-dimensional Euclidean geometry and in particular with the concepts of orthogonality and orthogonal projections in these spaces. These concepts, appropriately extended to infinite-dimensional Hilbert spaces, play a central role in the study of random variables with finite second moments and especially in the theory of prediction of stationary processes.

Intuition gained from Euclidean geometry can often be used to make apparently complicated algebraic results in time series analysis geometrically obvious. It frequently serves also as a valuable guide in the development and construction of algorithms.

This chapter is therefore devoted to a study of those aspects of Hilbert space theory which are needed for a geometric understanding of the later chapters in this book. The results developed here will also provide an adequate background for a geometric approach to many other areas of statistics, for example the general linear model (see Section 2.6). For the reader who wishes to go deeper into the theory of Hilbert space we recommend the book by Simmons (1963).

§2.1 Inner-Product Spaces and Their Properties

Definition 2.1.1 (Inner-Product Space). A complex vector space $\mathscr{H}$ is said to be an inner-product space if for each pair of elements x and y in $\mathscr{H}$, there is a complex number $\langle x, y\rangle$, called the inner product of x and y, such that

(a) $\langle x, y \rangle = \overline{\langle y, x \rangle}$, the bar denoting complex conjugation,
(b) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ for all $x, y, z \in \mathscr{H}$,
(c) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in \mathscr{H}$ and $\alpha \in \mathbb{C}$,
(d) $\langle x, x \rangle \geq 0$ for all $x \in \mathscr{H}$,
(e) $\langle x, x \rangle = 0$ if and only if $x = 0$.

Remark 1. A real vector space $\mathscr{H}$ is an inner-product space if for each $x, y \in \mathscr{H}$ there exists a *real* number $\langle x, y \rangle$ satisfying conditions (a)–(e). Of course condition (a) reduces in this case to $\langle x, y \rangle = \langle y, x \rangle$.

Remark 2. The inner product is a natural generalization of the inner or scalar product of two vectors in n-dimensional Euclidean space. Since many of the properties of Euclidean space carry over to inner-product spaces, it will be helpful to keep Euclidean space in mind in all that follows.

EXAMPLE 2.1.1 (Euclidean Space). The set of all column vectors

$$\mathbf{x} = (x_1, \ldots, x_k)' \in \mathbb{R}^k,$$

is a real inner-product space if we define

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{k} x_i y_i. \tag{2.1.1}$$

Equation (2.1.1) defines the usual scalar product of elements of $\mathbb{R}^k$. It is a simple matter to check that the conditions (a)–(e) are all satisfied.

In the same way it is easy to see that the set of all complex k-dimensional column vectors

$$\mathbf{z} = (z_1, \ldots, z_k)' \in \mathbb{C}^k$$

is a complex inner-product space if we define

$$\langle \mathbf{w}, \mathbf{z} \rangle = \sum_{i=1}^{k} w_i \bar{z}_i. \tag{2.1.2}$$

Definition 2.1.2 (Norm). The norm of an element x of an inner-product space is defined to be

$$\|x\| = \sqrt{\langle x, x \rangle}. \tag{2.1.3}$$

In the Euclidean space $\mathbb{R}^k$ the norm of the vector is simply its length, $\|\mathbf{x}\| = (\sum_{i=1}^k x_i^2)^{1/2}$.

The Cauchy–Schwarz Inequality. If $\mathscr{H}$ is an inner-product space, then

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| \quad \text{for all } x, y \in \mathscr{H}, \tag{2.1.4}$$

and

$$|\langle x, y \rangle| = \|x\| \, \|y\| \quad \text{if and only if } x = y\langle x, y \rangle / \langle y, y \rangle. \tag{2.1.5}$$

PROOF. The following proof for complex $\mathscr{H}$ remains valid (although it could be slightly simplified) in the case when $\mathscr{H}$ is real.

Let $a = \|y\|^2$, $b = |\langle x, y\rangle|$ and $c = \|x\|^2$. The polar representation of $\langle x, y\rangle$ is then

$$\langle x, y\rangle = be^{i\theta} \quad \text{for some } \theta \in (-\pi, \pi].$$

Now for all $r \in \mathbb{R}$,

$$\begin{aligned}\langle x - re^{i\theta}y, x - re^{i\theta}y\rangle &= \langle x, x\rangle - re^{i\theta}\langle y, x\rangle - re^{-i\theta}\langle x, y\rangle + r^2\langle y, y\rangle \\ &= c - 2rb + r^2 a,\end{aligned} \tag{2.1.6}$$

and using elementary calculus, we deduce from this that

$$0 \le \min_{r \in \mathbb{R}}(c - 2rb + r^2 a) = c - b^2/a,$$

thus establishing (2.1.4).

The minimum value, $c - b^2/a$, of $c - 2rb + r^2 a$ is achieved when $r = b/a$. If equality is achieved in (2.1.4) then $c - b^2/a = 0$. Setting $r = b/a$ in (2.1.6) we then obtain

$$\langle x - ye^{i\theta}b/a, x - ye^{i\theta}b/a\rangle = 0,$$

which, by property (e) of inner products, implies that

$$x = ye^{i\theta}b/a = y\langle x, y\rangle/\langle y, y\rangle.$$

Conversely if $x = y\langle x, y\rangle/\langle y, y\rangle$ (or equivalently if x is *any* scalar multiple of y), it is obvious that there is equality in (2.1.4). □

EXAMPLE 2.1.2 (The Angle between Elements of a Real Inner-Product Space). In the inner-product space $\mathbb{R}^3$ of Example 2.1.1, the angle between two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ is the angle in $[0, \pi]$ whose cosine is $\sum_{i=1}^{3} x_i y_i / (\|\mathbf{x}\| \|\mathbf{y}\|)$. Analogously we define the angle between non-zero elements x and y of any real inner-product space to be

$$\theta = \cos^{-1}[\langle x, y\rangle/(\|x\| \|y\|)]. \tag{2.1.7}$$

In particular x and y are said to be *orthogonal* if and only if $\langle x, y\rangle = 0$. For non-zero vectors x and y this is equivalent to the statement that $\theta = \pi/2$.

The Triangle Inequality. If $\mathscr{H}$ is an inner-product space, then

$$\|x + y\| \le \|x\| + \|y\| \quad \text{for all } x, y \in \mathscr{H}. \tag{2.1.8}$$

PROOF.

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y\rangle \\ &= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\ &\le \|x\|^2 + 2\|x\| \|y\| + \|y\|^2\end{aligned}$$

by the Cauchy–Schwarz inequality. □

Proposition 2.1.1 (Properties of the Norm). *If $\mathscr{H}$ is a complex (respectively real) inner-product space and $\|x\|$ is defined as in* (2.1.3), *then*

(a) $\|x + y\| \le \|x\| + \|y\|$ *for all* $x, y \in \mathscr{H}$,
(b) $\|\alpha x\| = |\alpha| \, \|x\|$ *for all* $x \in \mathscr{H}$ *and all* $\alpha \in \mathbb{C}$ ($\alpha \in \mathbb{R}$),
(c) $\|x\| \ge 0$ *for all* $x \in \mathscr{H}$,
(d) $\|x\| = 0$ *if and only if* $x = 0$.

(*These properties justify the use of the terminology "norm" for* $\|x\|$.)

Proof. The first property is a restatement of the triangle inequality and the others follow at once from Definition (2.1.3). □

The Parallelogram Law. If $\mathscr{H}$ is an inner-product space, then

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \text{for all } x, y \in \mathscr{H}. \tag{2.1.9}$$

Proof. Problem 2.1. Note that (2.1.9) is not a consequence of the properties (a), (b), (c) and (d) of the norm. It depends on the particular form (2.1.3) of the norm as defined for elements of an inner-product space. □

Definition 2.1.3 (Convergence in Norm). A sequence $\{x_n, n = 1, 2, \ldots\}$ of elements of an inner-product space $\mathscr{H}$ is said to converge in norm to $x \in \mathscr{H}$ if $\|x_n - x\| \to 0$ as $n \to \infty$.

Proposition 2.1.2 (Continuity of the Inner Product). *If $\{x_n\}$ and $\{y_n\}$ are sequences of elements of the inner-product space $\mathscr{H}$ such that $\|x_n - x\| \to 0$ and $\|y_n - y\| \to 0$ where $x, y \in \mathscr{H}$, then*

$$\text{(a)} \quad \|x_n\| \to \|x\|$$

and

$$\text{(b)} \quad \langle x_n, y_n \rangle \to \langle x, y \rangle.$$

Proof. From the triangle inequality it follows that $\|x\| \le \|x - y\| + \|y\|$ and $\|y\| \le \|y - x\| + \|x\|$. These statements imply that

$$\|x - y\| \ge \big| \|x\| - \|y\| \big|, \tag{2.1.10}$$

from which (a) follows immediately. Now

$$\begin{aligned} |\langle x_n, y_n \rangle - \langle x, y \rangle| &= |\langle x_n, y_n - y \rangle + \langle x_n - x, y \rangle| \\ &\le |\langle x_n, y_n - y \rangle| + |\langle x_n - x, y \rangle| \\ &\le \|x_n\| \, \|y_n - y\| + \|x_n - x\| \, \|y\|, \end{aligned}$$

where the last line follows from the Cauchy–Schwarz inequality. Observing from (a) that $\|x_n\| \to \|x\|$, we conclude that

$$|\langle x_n, y_n \rangle - \langle x, y \rangle| \to 0 \quad \text{as } n \to \infty.$$ □

§2.2 Hilbert Spaces

An inner-product space with the additional property of completeness is called a Hilbert space. To define completeness we first need the concept of a Cauchy sequence.

Definition 2.2.1 (Cauchy Sequence). A sequence $\{x_n, n = 1, 2, \ldots\}$ of elements of an inner-product space is said to be a Cauchy sequence if

$$\|x_n - x_m\| \to 0 \quad \text{as } m, n \to \infty,$$

i.e. if for every $\varepsilon > 0$ there exists a positive integer $N(\varepsilon)$ such that

$$\|x_n - x_m\| < \varepsilon \quad \text{for all } m, n > N(\varepsilon).$$

Definition 2.2.2 (Hilbert Space). A Hilbert space $\mathscr{H}$ is an inner-product space which is complete, i.e. an inner-product space in which every Cauchy sequence $\{x_n\}$ converges in norm to some element $x \in \mathscr{H}$.

Example 2.2.1 (Euclidean Space). The completeness of the inner-product space $\mathbb{R}^k$ defined in Example 2.1.1 can be verified as follows. If $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nk})' \in \mathbb{R}^k$ satisfies

$$\|\mathbf{x}_n - \mathbf{x}_m\|^2 = \sum_{i=1}^{k} |x_{ni} - x_{mi}|^2 \to 0 \quad \text{as } m, n \to \infty,$$

then each of the components must satisfy

$$|x_{ni} - x_{mi}| \to 0 \quad \text{as } m, n \to \infty.$$

By the completeness of $\mathbb{R}$, there exists $x_i \in \mathbb{R}$ such that

$$|x_{ni} - x_i| \to 0 \quad \text{as } n \to \infty,$$

and hence if $\mathbf{x} = (x_1, \ldots, x_k)$, then

$$\|\mathbf{x}_n - \mathbf{x}\| \to 0 \quad \text{as } n \to \infty.$$

Completeness of the complex inner-product space $\mathbb{C}^k$ can be checked in the same way. Thus $\mathbb{R}^k$ and $\mathbb{C}^k$ are both Hilbert spaces.

Example 2.2.2 (The Space $L^2(\Omega, \mathscr{F}, P)$). Consider a probability space $(\Omega, \mathscr{F}, P)$ and the collection C of all random variables X defined on Ω and satisfying the condition,

$$EX^2 = \int_{\Omega} X(\omega)^2 P(d\omega) < \infty.$$

With the usual notion of multiplication by a real scalar and addition of random variables, it is clear that C is a vector space since

$$E(aX)^2 = a^2 EX^2 < \infty \quad \text{for all } a \in \mathbb{R} \text{ and } X \in C,$$

and, from the inequality $(X + Y)^2 \le 2X^2 + 2Y^2$,

$$E(X + Y)^2 \le 2EX^2 + 2EY^2 < \infty \quad \text{for all } X, Y \in C.$$

The other properties required of a vector space are easily checked. In particular C has a zero element, the random variable which is identically zero on Ω.

For any two elements X, $Y \in C$ we now define

$$\langle X, Y \rangle = E(XY). \tag{2.2.1}$$

It is easy to check that $\langle X, Y \rangle$ satisfies all the properties of an inner product except for the last. If $\langle X, X \rangle = 0$ then it does not follow that $X(\omega) = 0$ for all ω, but only that $P(X = 0) = 1$. This difficulty is circumvented by saying that the random variables X and Y are equivalent if $P(X = Y) = 1$. This equivalence relation partitions C into classes of random variables such that any two random variables in the same class are equal with probability one. The space L^2 (or more specifically $L^2(\Omega, \mathscr{F}, P)$) is the collection of these equivalence classes with inner product defined by (2.2.1). Since each class is uniquely determined by specifying any one of the random variables in it, we shall continue to use the notation X, Y for elements of L^2 and to call them random variables (or functions) although it is sometimes important to remember that X stands for the collection of all random variables which are equivalent to X.

Norm convergence of a sequence $\{X_n\}$ of elements of L^2 to the limit X means

$$\|X_n - X\|^2 = E|X_n - X|^2 \to 0 \quad \text{as } n \to \infty.$$

Norm convergence of X_n to X in an L^2 space is called mean-square convergence and is written as $X_n \xrightarrow{\text{m.s.}} X$.

To complete the proof that L^2 is a Hilbert space we need to establish completeness, i.e. that if $\|X_m - X_n\|^2 \to 0$ as $m, n \to \infty$, then there exists $X \in L^2$ such that $X_n \xrightarrow{\text{m.s.}} X$. This is indeed true but not so easy to prove as the completeness of $\mathbb{R}^k$. We therefore defer the proof to Section 2.10.

Example 2.2.3 (Complex L^2 Spaces). The space of complex-valued random variables X on $(\Omega, \mathscr{F}, P)$ satisfying $E|X|^2 < \infty$ is a complex Hilbert space if we define an inner product by

$$\langle X, Y \rangle = E(X\bar{Y}). \tag{2.2.2}$$

In fact if μ is any finite non-zero measure on the measurable space $(\Omega, \mathscr{F})$, and if D is the class of complex-valued functions on Ω such that

$$\int_\Omega |f|^2 \, d\mu < \infty \tag{2.2.3}$$

(with identification of functions f and g such that $\int_\Omega |f - g|^2 \, d\mu = 0$), then D becomes a Hilbert space if we define the inner product to be

$$\langle f, g\rangle = \int_\Omega f\bar{g}\,d\mu. \tag{2.2.4}$$

This space will be referred to as the complex Hilbert space $L^2(\Omega, \mathscr{F}, \mu)$. (The real Hilbert space $L^2(\Omega, \mathscr{F}, \mu)$ is obtained if D is replaced by the real-valued functions satisfying (2.2.3). The definition of $\langle f, g\rangle$ then reduces to $\int_\Omega fg\,d\mu$.)

Remark 1. The terms $L^2(\Omega, \mathscr{F}, P)$ and $L^2(\Omega, \mathscr{F}, \mu)$ will be reserved for the respective real Hilbert spaces unless we state specifically that reference is being made to the corresponding complex spaces.

Proposition 2.2.1 (Norm Convergence and the Cauchy Criterion). *If $\{x_n\}$ is a sequence of elements belonging to a Hilbert space $\mathscr{H}$, then $\{x_n\}$ converges in norm if and only if $\|x_n - x_m\| \to 0$ as $m, n \to \infty$.*

PROOF. The sufficiency of the Cauchy criterion is simply a restatement of the completeness of $\mathscr{H}$. The necessity is an elementary consequence of the triangle inequality. Thus if $\|x_n - x\| \to 0$,

$$\|x_n - x_m\| \le \|x_n - x\| + \|x - x_m\| \to 0 \quad \text{as } m, n \to 0. \qquad \square$$

EXAMPLE 2.2.4. The Cauchy criterion is used primarily in checking for the norm convergence of a sequence whose limit is not specified. Consider for example the sequence

$$S_n = \sum_{i=1}^{n} a_i X_i \tag{2.2.5}$$

where $\{X_i\}$ is a sequence of independent $N(0, 1)$ random variables. It is easy to see that with the usual definition of the L^2-norm,

$$\|S_m - S_n\|^2 = \sum_{i=n+1}^{m} a_i^2, \qquad m > n,$$

and so by the Cauchy criterion $\{S_n\}$ has a mean-square limit if and only if for every $\varepsilon > 0$, there exists $N(\varepsilon) > 0$ such that $\sum_{i=n+1}^{m} a_i^2 < \varepsilon$ for $m > n > N(\varepsilon)$. Thus $\{S_n\}$ converges in mean square if and only if $\sum_{i=1}^{\infty} a_i^2 < \infty$.

§2.3 The Projection Theorem

We begin this section with two examples which illustrate the use of the projection theorem in particular Hilbert spaces. The general result is then established as Theorem 2.3.1.

EXAMPLE 2.3.1 (Linear Approximation in $\mathbb{R}^3$). Suppose we are given three vectors in $\mathbb{R}^3$,

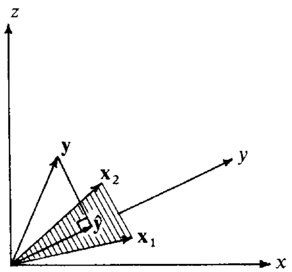

Figure 2.1. The best linear approximation $\hat{\mathbf{y}} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$, to $\mathbf{y}$.

$$\mathbf{y} = (\tfrac{1}{4}, \tfrac{1}{4}, 1)',$$
$$\mathbf{x}_1 = (1, 0, \tfrac{1}{4})',$$
$$\mathbf{x}_2 = (0, 1, \tfrac{1}{4})',$$

Our problem is to find the linear combination $\hat{\mathbf{y}} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$ which is closest to $\mathbf{y}$ in the sense that $S = \|\mathbf{y} - \alpha_1\mathbf{x}_1 - \alpha_2\mathbf{x}_2\|^2$ is minimized.

One approach to this problem is to write S in the form $S = (\frac{1}{4} - \alpha_1)^2 + (\frac{1}{4} - \alpha_2)^2 + (1 - \frac{1}{4}\alpha_1 - \frac{1}{4}\alpha_2)^2$ and then to use calculus to minimize with respect to α_1 and α_2. In the alternative geometric approach to the problem we observe that the required vector $\hat{\mathbf{y}} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$ is the vector in the plane determined by $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $\mathbf{y} - \alpha_1\mathbf{x}_1 - \alpha_2\mathbf{x}_2$ is orthogonal to the plane of $\mathbf{x}_1$ and $\mathbf{x}_2$ (see Figure 2.1). The orthogonality condition may be stated as

$$\langle \mathbf{y} - \alpha_1\mathbf{x}_1 - \alpha_2\mathbf{x}_2, \mathbf{x}_i \rangle = 0, \qquad i = 1, 2, \tag{2.3.1}$$

or equivalently

$$\alpha_1\langle \mathbf{x}_1, \mathbf{x}_1 \rangle + \alpha_2\langle \mathbf{x}_2, \mathbf{x}_1 \rangle = \langle \mathbf{y}, \mathbf{x}_1 \rangle,$$
$$\alpha_1\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + \alpha_2\langle \mathbf{x}_2, \mathbf{x}_2 \rangle = \langle \mathbf{y}, \mathbf{x}_2 \rangle.$$

For the particular vectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{y}$ specified, these equations become

$$\tfrac{17}{16}\alpha_1 + \tfrac{1}{16}\alpha_2 = \tfrac{1}{2},$$
$$\tfrac{1}{16}\alpha_1 + \tfrac{17}{16}\alpha_2 = \tfrac{1}{2},$$

from which we deduce that $\alpha_1 = \alpha_2 = \frac{4}{9}$, and $\hat{\mathbf{y}} = (\frac{4}{9}, \frac{4}{9}, \frac{2}{9})'$.

Example 2.3.2 (Linear Approximation in $L^2(\Omega, \mathscr{F}, P)$). Now suppose that X_1, X_2 and Y are random variables in $L^2(\Omega, \mathscr{F}, P)$. If only X_1 and X_2 are observed we may wish to estimate the value of Y by using the linear combination $\hat{Y} = \alpha_1 X_1 + \alpha_2 X_2$ which minimizes the mean squared error,

$$S = E|Y - \alpha_1 X_1 - \alpha_2 X_2|^2 = \|Y - \alpha_1 X_1 - \alpha_2 X_2\|^2.$$

As in Example 2.3.1 there are at least two possible approaches to this problem. The first is to write

$$S = EY^2 + \alpha_1^2 EX_1^2 + \alpha_2^2 EX_2^2 - 2\alpha_1 E(YX_1) - 2\alpha_2 E(YX_2) + \alpha_1\alpha_2 E(X_1 X_2),$$

and then to minimize with respect to α_1 and α_2 by setting the appropriate derivatives equal to zero. However it is also possible to use the same geometric approach as in Example 2.3.1. Our aim is to find an element $\hat{Y}$ in the set

$$\mathscr{M} = \{X \in L^2(\Omega, \mathscr{F}, P) : X = a_1 X_1 + a_2 X_2 \text{ for some } a_1, a_2 \in \mathbb{R}\},$$

whose "squared distance" from Y, $\|Y - \hat{Y}\|^2$, is as small as possible. By analogy with Example 2.3.1 we might expect $\hat{Y}$ to have the property that $Y - \hat{Y}$ is orthogonal to all elements of $\mathscr{M}$. The validity of this analogy, and the extent to which it may be applied in more general situations, is established in Theorem 2.3.1 (the projection theorem). Applying it to our present problem, we can write

$$\langle Y - \alpha_1 X_1 - \alpha_2 X_2, X \rangle = 0 \quad \text{for all } X \in \mathscr{M}, \tag{2.3.2}$$

or equivalently, by the linearity of the inner product,

$$\langle Y - \alpha_1 X_1 - \alpha_2 X_2, X_i \rangle = 0, \qquad i = 1, 2. \tag{2.3.3}$$

These are the same equations for α_1 and α_2 as (2.3.1), although the inner product is of course defined differently in (2.3.3). In terms of expectations we can rewrite (2.3.3) in the form

$$\alpha_1 E(X_1^2) + \alpha_2 E(X_2 X_1) = E(YX_1),$$
$$\alpha_1 E(X_1 X_2) + \alpha_2 E(X_2^2) = E(YX_2),$$

from which α_1 and α_2 are easily found.

Before establishing the projection theorem for a general Hilbert space we need to introduce a certain amount of new terminology.

Definition 2.3.1 (Closed Subspace). A linear subspace $\mathscr{M}$ of a Hilbert space $\mathscr{H}$ is said to be a closed subspace of $\mathscr{H}$ if $\mathscr{M}$ contains all of its limit points (i.e. if $x_n \in \mathscr{M}$ and $\|x_n - x\| \to 0$ imply that $x \in \mathscr{M}$).

Definition 2.3.2 (Orthogonal Complement). The orthogonal complement of a subset $\mathscr{M}$ of $\mathscr{H}$ is defined to be the set $\mathscr{M}^\perp$ of all elements of $\mathscr{H}$ which are orthogonal to every element of $\mathscr{M}$. Thus

$$x \in \mathscr{M}^\perp \text{ if and only if } \langle x, y \rangle = 0 \text{ (written } x \perp y) \text{ for all } y \in \mathscr{M}. \tag{2.3.4}$$

Proposition 2.3.1. *If $\mathscr{M}$ is any subset of a Hilbert space $\mathscr{H}$ then $\mathscr{M}^\perp$ is a closed subspace of $\mathscr{H}$.*

PROOF. It is easy to check from (2.3.4) that $0 \in \mathscr{M}^\perp$ and that if $x_1, x_2 \in \mathscr{M}^\perp$ then all linear combinations of x_1 and x_2 belong to $\mathscr{M}^\perp$. Hence $\mathscr{M}^\perp$ is a subspace of $\mathscr{H}$. If $x_n \in \mathscr{M}^\perp$ and $\|x_n - x\| \to 0$, then by continuity of the inner product (Proposition 2.1.2), $\langle x, y \rangle = 0$ for all $y \in \mathscr{M}$, so $x \in \mathscr{M}^\perp$ and hence $\mathscr{M}^\perp$ is closed. □

Theorem 2.3.1 (The Projection Theorem). *If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$, then*

(i) *there is a unique element $\hat{x} \in \mathcal{M}$ such that*

$$\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|, \tag{2.3.5}$$

and

(ii) *$\hat{x} \in \mathcal{M}$ and $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \in \mathcal{M}^{\perp}$.*

[*The element $\hat{x}$ is called the (orthogonal) projection of x onto $\mathcal{M}$.*]

Proof. (i) If $d = \inf_{y \in \mathcal{M}} \|x - y\|^2$ then there is a sequence $\{y_n\}$ of elements of $\mathcal{M}$ such that $\|y_n - x\|^2 \to d$. Apply the parallelogram law (2.1.9), and using the fact that $(y_m + y_n)/2 \in \mathcal{M}$, we can write

$$\begin{aligned} 0 \le \|y_m - y_n\|^2 &= -4\|(y_m + y_n)/2 - x\|^2 + 2(\|y_n - x\|^2 + \|y_m - x\|^2) \\ &\le -4d + 2(\|y_n - x\|^2 + \|y_m - x\|^2) \\ &\to 0 \text{ as } m, n \to \infty. \end{aligned}$$

Consequently, by the Cauchy criterion, there exists $\hat{x} \in \mathcal{H}$ such that $\|y_n - \hat{x}\| \to 0$. Since $\mathcal{M}$ is closed we know that $\hat{x} \in \mathcal{M}$, and by continuity of the inner product

$$\|x - \hat{x}\|^2 = \lim_{n \to \infty} \|x - y_n\|^2 = d.$$

To establish uniqueness, suppose that $\hat{y} \in \mathcal{M}$ and that $\|x - \hat{y}\|^2 = \|x - \hat{x}\|^2 = d$. Then, applying the parallelogram law again,

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

Hence $\hat{y} = \hat{x}$.

(ii) If $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \in \mathcal{M}^{\perp}$ then $\hat{x}$ is the unique element of $\mathcal{M}$ defined in (i) since for any $y \in \mathcal{M}$,

$$\begin{aligned} \|x - y\|^2 &= \langle x - \hat{x} + \hat{x} - y, x - \hat{x} + \hat{x} - y \rangle \\ &= \|x - \hat{x}\|^2 + \|\hat{x} - y\|^2 \\ &\ge \|x - \hat{x}\|^2, \end{aligned}$$

with equality if and only if $y = \hat{x}$.

Conversely if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \notin \mathcal{M}^{\perp}$ then $\hat{x}$ is *not* the element of $\mathcal{M}$ closest to x since

$$\tilde{x} = \hat{x} + ay/\|y\|^2$$

is closer, where y is any element of $\mathcal{M}$ such that $\langle x - \hat{x}, y \rangle \ne 0$ and

$a = \langle x - \hat{x}, y \rangle$. To see this we write

$$\begin{aligned}\|x - \tilde{x}\|^2 &= \langle x - \hat{x} + \hat{x} - \tilde{x}, x - \hat{x} + \hat{x} - \tilde{x} \rangle \\ &= \|x - \hat{x}\|^2 + |a|^2/\|y\|^2 + 2\,\mathrm{Re}\langle x - \hat{x}, \hat{x} - \tilde{x} \rangle \\ &= \|x - \hat{x}\|^2 - |a|^2/\|y\|^2 \\ &< \|x - \hat{x}\|^2.\end{aligned}$$

□

Corollary 2.3.1 (The Projection Mapping of $\mathscr{H}$ onto $\mathscr{M}$). *If $\mathscr{M}$ is a closed subspace of the Hilbert space $\mathscr{H}$ and I is the identity mapping on $\mathscr{H}$, then there is a unique mapping $P_{\mathscr{M}}$ of $\mathscr{H}$ onto $\mathscr{M}$ such that $I - P_{\mathscr{M}}$ maps $\mathscr{H}$ onto $\mathscr{M}^{\perp}$. $P_{\mathscr{M}}$ is called the projection mapping of $\mathscr{H}$ onto $\mathscr{M}$.*

PROOF. By Theorem 2.3.1, for each $x \in \mathscr{H}$ there is a unique $\hat{x} \in \mathscr{M}$ such that $x - \hat{x} \in \mathscr{M}^{\perp}$. The required mapping is therefore

$$P_{\mathscr{M}} x = \hat{x}, \qquad x \in \mathscr{H}. \tag{2.3.6}$$

□

Proposition 2.3.2 (Properties of Projection Mappings). *Let $\mathscr{H}$ be a Hilbert space and let $P_{\mathscr{M}}$ denote the projection mapping onto a closed subspace $\mathscr{M}$. Then*

(i) $P_{\mathscr{M}}(\alpha x + \beta y) = \alpha P_{\mathscr{M}} x + \beta P_{\mathscr{M}} y, \qquad x, y \in \mathscr{H}, \quad \alpha, \beta \in \mathbb{C}$,
(ii) $\|x\|^2 = \|P_{\mathscr{M}} x\|^2 + \|(I - P_{\mathscr{M}})x\|^2$,
(iii) *each $x \in \mathscr{H}$ has a unique representation as a sum of an element of $\mathscr{M}$ and an element of $\mathscr{M}^{\perp}$*, i.e.

$$x = P_{\mathscr{M}} x + (I - P_{\mathscr{M}})x, \tag{2.3.7}$$

(iv) $P_{\mathscr{M}} x_n \to P_{\mathscr{M}} x$ *if* $\|x_n - x\| \to 0$,
(v) $x \in \mathscr{M}$ *if and only if* $P_{\mathscr{M}} x = x$,
(vi) $x \in \mathscr{M}^{\perp}$ *if and only if* $P_{\mathscr{M}} x = 0$,
and
(vii) $\mathscr{M}_1 \subseteq \mathscr{M}_2$ *if and only if* $P_{\mathscr{M}_1} P_{\mathscr{M}_2} x = P_{\mathscr{M}_1} x$ *for all* $x \in \mathscr{H}$.

PROOF. (i) $\alpha P_{\mathscr{M}} x + \beta P_{\mathscr{M}} y \in \mathscr{M}$ since $\mathscr{M}$ is a linear subspace of $\mathscr{H}$. Also

$$\begin{aligned}\alpha x + \beta y - (\alpha P_{\mathscr{M}} x + \beta P_{\mathscr{M}} y) &= \alpha(x - P_{\mathscr{M}} x) + \beta(y - P_{\mathscr{M}} y) \\ &\in \mathscr{M}^{\perp},\end{aligned}$$

since $\mathscr{M}^{\perp}$ is a linear subspace of $\mathscr{H}$ by Proposition 2.3.1. These two properties identify $\alpha P_{\mathscr{M}} x + \beta P_{\mathscr{M}} y$ as the projection $P_{\mathscr{M}}(\alpha x + \beta y)$.

(ii) This is an immediate consequence of the orthogonality of $P_{\mathscr{M}} x$ and $(I - P_{\mathscr{M}})x$.

(iii) One such representation is clearly $x = P_{\mathscr{M}} x + (I - P_{\mathscr{M}})x$. If $x = y + z$, $y \in \mathscr{M}$, $z \in \mathscr{M}^{\perp}$ is another, then

$$y - P_{\mathscr{M}} x + z - (I - P_{\mathscr{M}})x = 0.$$

Taking inner products of each side with $y - P_{\mathcal{M}}x$ gives $\|y - P_{\mathcal{M}}x\|^2 = 0$, since $z - (I - P_{\mathcal{M}})x \in \mathcal{M}^{\perp}$. Hence $y = P_{\mathcal{M}}x$ and $z = (I - P_{\mathcal{M}})x$.

(iv) By (ii), $\|P_{\mathcal{M}}(x_n - x)\|^2 \le \|x_n - x\|^2 \to 0$ if $\|x_n - x\| \to 0$.

(v) $x \in \mathcal{M}$ if and only if the unique representation $x = y + z$, $y \in \mathcal{M}$, $z \in \mathcal{M}^{\perp}$, is such that $y = x$ and $z = 0$, i.e. if and only if $P_{\mathcal{M}}x = x$.

(vi) Repeat the argument in (v) with $y = 0$ and $z = x$.

(vii) $x = P_{\mathcal{M}_2}x + (I - P_{\mathcal{M}_2})x$. Projecting each side onto $\mathcal{M}_1$ we obtain

$$P_{\mathcal{M}_1}x = P_{\mathcal{M}_1}P_{\mathcal{M}_2}x + P_{\mathcal{M}_1}(I - P_{\mathcal{M}_2})x.$$

Hence $P_{\mathcal{M}_1}x = P_{\mathcal{M}_1}P_{\mathcal{M}_2}x$ for all $x \in \mathcal{H}$ if and only if $P_{\mathcal{M}_1}y = 0$ for all $y \in \mathcal{M}_2^{\perp}$, i.e. if and only if $\mathcal{M}_2^{\perp} \subseteq \mathcal{M}_1^{\perp}$, i.e. if and only if $\mathcal{M}_1 \subseteq \mathcal{M}_2$. □

The Prediction Equations. Given a Hilbert space $\mathcal{H}$, a closed subspace $\mathcal{M}$, and an element $x \in \mathcal{H}$, Theorem 2.3.1 shows that the element of $\mathcal{M}$ closest to x is the unique element $\hat{x} \in \mathcal{M}$ such that

$$\langle x - \hat{x}, y \rangle = 0 \quad \text{for all } y \in \mathcal{M}. \tag{2.3.8}$$

The equations (2.3.1) and (2.3.2) which arose in Examples 2.3.1 and 2.3.2 are special cases of (2.3.8). In later chapters we shall constantly be making use of the equations (2.3.8), interpreting $\hat{x} = P_{\mathcal{M}}x$ as the best predictor of x in the subspace $\mathcal{M}$.

Remark 1. It is helpful to visualize the projection theorem in terms of Figure 2.1, which depicts the special case in which $\mathcal{H} = \mathbb{R}^3$, $\mathcal{M}$ is the plane containing $\mathbf{x}_1$ and $\mathbf{x}_2$, and $\hat{\mathbf{y}} = P_{\mathcal{M}}\mathbf{y}$. The prediction equation (2.3.8) is simply the statement (obvious in this particular example) that $\mathbf{y} - \hat{\mathbf{y}}$ must be orthogonal to $\mathcal{M}$. The projection theorem tells us that $\hat{y} = P_{\mathcal{M}}y$ is uniquely determined by this condition for *any* Hilbert space $\mathcal{H}$ and closed subspace $\mathcal{M}$. This justifies in particular our use of equations (2.3.2) in Example 2.3.2. As we shall see later (especially in Chapter 5), the projection theorem plays a fundamental role in all problems involving the approximation or prediction of random variables with finite variance.

EXAMPLE 2.3.3 (Minimum Mean Squared Error Linear Prediction of a Stationary Process). Let $\{X_t, t = 0, \pm 1, \dots\}$ be a stationary process on $(\Omega, \mathcal{F}, P)$ with mean zero and autocovariance function $\gamma(\cdot)$, and consider the problem of finding the linear combination $\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$ which best approximates X_{n+1} in the sense that $E|X_{n+1} - \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}|^2$ is minimum. This problem is easily solved with the aid of the projection theorem by taking $\mathcal{H} = L^2(\Omega, \mathcal{F}, P)$ and $\mathcal{M} = \{\sum_{j=1}^{n} \alpha_j X_{n+1-j} : \alpha_1, \dots, \alpha_n \in \mathbb{R}\}$. Since minimization of $E|X_{n+1} - \hat{X}_{n+1}|^2$ is identical to minimization of the squared norm $\|X_{n+1} - \hat{X}_{n+1}\|^2$, we see at once that $\hat{X}_{n+1} = P_{\mathcal{M}}X_{n+1}$. The prediction equations (2.3.8) are

$$\left\langle X_{n+1} - \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}, Y \right\rangle = 0 \quad \text{for all } Y \in \mathcal{M},$$

which, by the linearity of the inner product, are equivalent to the n equations

$$\left\langle X_{n+1} - \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}, X_k \right\rangle = 0, \qquad k = n, n-1, \ldots, 1.$$

Recalling the definition $\langle X, Y \rangle = E(XY)$ of the inner product in $L^2(\Omega, \mathscr{F}, P)$, we see that the prediction equations can be written in the form

$$\Gamma_n \phi_n = \gamma_n \tag{2.3.9}$$

where $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})'$, $\gamma_n = (\gamma(1), \ldots, \gamma(n))'$ and $\Gamma_n = [\gamma(i-j)]_{i,j=1}^n$. The projection theorem guarantees that there is *at least* one solution ϕ_n of (2.3.9). If Γ_n is singular then (2.3.9) will have infinitely many solutions, but the projection theorem guarantees that every solution will give the same (uniquely defined) predictor $\hat{X}_{n+1}$.

Example 2.3.4. To illustrate the last assertion of Example 2.3.3, consider the stationary process

$$X_t = A\cos(\omega t) + B\sin(\omega t),$$

where $\omega \in (0, \pi)$ is constant and A, B are uncorrelated random variables with mean 0 and variance σ^2. We showed in Example 1.3.1 that for this process, $\gamma(h) = \sigma^2 \cos(\omega h)$. It is easy to check from (2.3.9) (see Problem 2.6) that

$$\phi_1 = \cos\omega \quad \text{and} \quad \phi_2 = (2\cos\omega, -1)'.$$

Thus

$$\hat{X}_3 = (2\cos\omega)X_2 - X_1.$$

The mean squared error of $\hat{X}_3$ is

$$E(X_3 - (2\cos\omega)X_2 + X_1)^2 = 0,$$

showing that for this process we have the identity,

$$X_3 = (2\cos\omega)X_2 - X_1. \tag{2.3.10}$$

The same argument and the stationarity of $\{X_t\}$ show that

$$\hat{X}_4 = (2\cos\omega)X_3 - X_2, \tag{2.3.11}$$

again with mean squared error zero. Because of the relation (2.3.10) there are infinitely many ways to reexpress $\hat{X}_4$ in terms of X_1, X_2 and X_3. This is reflected by the fact that Γ_3 is singular for this process and (2.3.9) has infinitely many solutions for ϕ_3.

§2.4 Orthonormal Sets

Definition 2.4.1 (Closed Span). The closed span $\overline{\text{sp}}\{x_t, t \in T\}$ of any subset $\{x_t, t \in T\}$ of a Hilbert space $\mathscr{H}$ is defined to be the smallest closed subspace of $\mathscr{H}$ which contains each element x_t, $t \in T$.

Remark 1. The closed span of a finite set $\{x_1, \ldots, x_n\}$ is the set of all linear combinations, $y = \alpha_1 x_1 + \cdots + \alpha_n x_n$, $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ (or $\mathbb{R}$ if $\mathscr{H}$ is real). See Problem 2.7. If for example $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^3$ and $\mathbf{x}_1$ is not a scalar multiple of $\mathbf{x}_2$ then $\overline{\text{sp}}\{\mathbf{x}_1, \mathbf{x}_2\}$ is the plane containing $\mathbf{x}_1$ and $\mathbf{x}_2$.

Remark 2. If $\mathscr{M} = \overline{\text{sp}}\{x_1, \ldots, x_n\}$, then for any given $x \in \mathscr{H}$, $P_{\mathscr{M}} x$ is the unique element of the form

$$P_{\mathscr{M}} x = \alpha_1 x_1 + \cdots + \alpha_n x_n,$$

such that

$$\langle x - P_{\mathscr{M}} x, y \rangle = 0, \qquad y \in \mathscr{M},$$

or equivalently such that

$$\langle P_{\mathscr{M}} x, x_j \rangle = \langle x, x_j \rangle, \qquad j = 1, \ldots, n. \tag{2.4.1}$$

The equations (2.4.1) can be rewritten as a set of linear equations for $\alpha_1, \ldots, \alpha_n$, viz.

$$\sum_{i=1}^{n} \alpha_i \langle x_i, x_j \rangle = \langle x, x_j \rangle, \qquad j = 1, \ldots, n. \tag{2.4.2}$$

By the projection theorem the system (2.4.2) has at least one solution for $\alpha_1, \ldots, \alpha_n$. The uniqueness of $P_{\mathscr{M}} x$ implies that all solutions of (2.4.2) must yield the same element $\alpha_1 x_1 + \cdots + \alpha_n x_n$.

Definition 2.4.2 (Orthonormal Set). A set $\{e_t, t \in T\}$ of elements of an inner-product space is said to be orthonormal if for every $s, t \in T$,

$$\langle e_s, e_t \rangle = \begin{cases} 1 & \text{if } s = t, \\ 0 & \text{if } s \neq t. \end{cases} \tag{2.4.3}$$

EXAMPLE 2.4.1. The set of vectors $\{(1,0,0)', (0,1,0)', (0,0,1)'\}$ is an orthonormal set in $\mathbb{R}^3$.

EXAMPLE 2.4.2. Any sequence $\{Z_t, t \in \mathbb{Z}\}$ of independent standard normal random variables is an orthonormal set in $L^2(\Omega, \mathscr{F}, P)$.

Theorem 2.4.1. *If $\{e_1, \ldots, e_k\}$ is an orthonormal subset of the Hilbert space $\mathscr{H}$ and $\mathscr{M} = \overline{\text{sp}}\{e_1, \ldots, e_k\}$, then*

$$P_{\mathscr{M}} x = \sum_{i=1}^{k} \langle x, e_i \rangle e_i \quad \textit{for all } x \in \mathscr{H}, \tag{2.4.4}$$

$$\|P_{\mathscr{M}} x\|^2 = \sum_{i=1}^{k} |\langle x, e_i \rangle|^2 \quad \textit{for all } x \in \mathscr{H}, \tag{2.4.5}$$

$$\left\| x - \sum_{i=1}^{k} \langle x, e_i \rangle e_i \right\| \leq \left\| x - \sum_{i=1}^{k} c_i e_i \right\| \quad \textit{for all } x \in \mathscr{H}, \tag{2.4.6}$$

and for all $c_1, \ldots, c_k \in \mathbb{C}$ (or $\mathbb{R}$ if $\mathscr{H}$ is real). Equality holds in (2.4.6) if and only if $c_i = \langle x, e_i \rangle$, $i = 1, \ldots, k$.

The numbers $\langle x, e_i \rangle$ are sometimes called the Fourier coefficients of x relative to the set $\{e_1, \ldots, e_k\}$.

PROOF. To establish (2.4.4) it suffices by Remark 2 to check that $P_{\mathscr{M}}x$ as defined by (2.4.4) satisfies the prediction equations (2.4.1), i.e. that

$$\left\langle \sum_{i=1}^{k} \langle x, e_i \rangle e_i, e_j \right\rangle = \langle x, e_j \rangle, \qquad j = 1, \ldots, k.$$

But this is an immediate consequence of the orthonormality condition (2.4.3).

The proof of (2.4.5) is a routine computation using properties of the inner product and the assumed orthonormality of $\{e_1, \ldots, e_k\}$.

By Theorem 2.3.1 (ii), $\|x - P_{\mathscr{M}}x\| \le \|x - y\|$ for all $y \in \mathscr{M}$, and this is precisely the inequality (2.4.6). By Theorem 2.3.1 (ii) again, there is equality in (2.4.6) if and only if

$$\sum_{i=1}^{k} c_i e_i = P_{\mathscr{M}}x = \sum_{i=1}^{k} \langle x, e_i \rangle e_i. \tag{2.4.7}$$

Taking inner products of each side with e_j and recalling the orthonormality assumption, we immediately find that (2.4.7) is equivalent to the condition $c_j = \langle x, e_j \rangle, j = 1, \ldots, k$. □

Corollary 2.4.1 (Bessel's Inequality). *If x is any element of a Hilbert space $\mathscr{H}$ and $\{e_1, \ldots, e_k\}$ is an orthonormal subset of $\mathscr{H}$ then*

$$\sum_{i=1}^{k} |\langle x, e_i \rangle|^2 \le \|x\|^2. \tag{2.4.8}$$

PROOF. This follows at once from (2.4.5) and Proposition 2.3.2 (ii). □

Definition 2.4.3. (Complete Orthonormal Set). If $\{e_t, t \in T\}$ is an orthonormal subset of the Hilbert space $\mathscr{H}$ and if $\mathscr{H} = \overline{\mathrm{sp}}\{e_t, t \in T\}$ then we say that $\{e_t, t \in T\}$ is a complete orthonormal set or an orthonormal basis for $\mathscr{H}$.

Definition 2.4.4 (Separability). The Hilbert space $\mathscr{H}$ is separable if $\mathscr{H} = \overline{\mathrm{sp}}\{e_t, t \in T\}$ with $\{e_t, t \in T\}$ a finite or countably infinite orthonormal set.

Theorem 2.4.2. *If $\mathscr{H}$ is the separable Hilbert space $\mathscr{H} = \overline{\mathrm{sp}}\{e_1, e_2, \ldots\}$ where $\{e_i, i = 1, 2, \ldots\}$ is an orthonormal set, then*

(i) *the set of all finite linear combinations of $\{e_1, e_2, \ldots\}$ is dense in $\mathscr{H}$, i.e. for each $x \in \mathscr{H}$ and $\varepsilon > 0$, there exists a positive integer k and constants $c_1, \ldots, c_k$ such that*

$$\left\| x - \sum_{i=1}^{k} c_i e_i \right\| < \varepsilon, \tag{2.4.9}$$

(ii) $x = \sum_{i=1}^{\infty} \langle x, e_i \rangle e_i$ *for each* $x \in \mathscr{H}$, *i.e.* $\|x - \sum_{i=1}^{n} \langle x, e_i \rangle e_i\| \to 0$ *as* $n \to \infty$,
(iii) $\|x\|^2 = \sum_{i=1}^{\infty} |\langle x, e_i \rangle|^2$ *for each* $x \in \mathscr{H}$,
(iv) $\langle x, y \rangle = \sum_{i=1}^{\infty} \langle x, e_i \rangle \langle e_i, y \rangle$ *for each* $x, y \in \mathscr{H}$, *and*
(v) $x = 0$ *if and only if* $\langle x, e_i \rangle = 0$ *for all* $i = 1, 2, \ldots$.

The result (iv) *is known as* ***Parseval's identity***.

PROOF. (i) If $S = \bigcup_{j=1}^{\infty} \overline{\text{sp}}\{e_1, \ldots, e_j\}$, the set of all finite linear combinations of $\{e_1, e_2, \ldots\}$, then the closure $\bar{S}$ of S is a closed subspace of $\mathscr{H}$ (Problem 2.17) containing $\{e_i, i = 1, 2, \ldots\}$. Since $\mathscr{H}$ is by assumption the smallest such closed subspace, we conclude that $\bar{S} = \mathscr{H}$.

(ii) By Bessel's inequality (2.4.8), $\sum_{i=1}^{k} |\langle x, e_i \rangle|^2 \leq \|x\|^2$ for all positive integers k. Hence $\sum_{i=1}^{\infty} |\langle x, e_i \rangle|^2 \leq \|x\|^2$. From (2.4.6) and (2.4.9) we conclude that for each $\varepsilon > 0$ there exists a positive integer k such that

$$\left\| x - \sum_{i=1}^{k} \langle x, e_i \rangle e_i \right\| < \varepsilon.$$

Now by Theorem 2.4.1, $\sum_{i=1}^{n} \langle x, e_i \rangle e_i = P_{\mathscr{M}} x$ where $\mathscr{M} = \overline{\text{sp}}\{e_1, \ldots, e_n\}$, and since for $k \leq n$, $\sum_{i=1}^{k} \langle x, e_i \rangle e_i \in \mathscr{M}$, we also have

$$\left\| x - \sum_{i=1}^{n} \langle x, e_i \rangle e_i \right\| < \varepsilon \quad \text{for all } n \geq k. \tag{2.4.10}$$

(iii) From (2.4.10) we can write, for $n \geq k$,

$$\begin{aligned} \|x\|^2 &= \left\| x - \sum_{i=1}^{n} \langle x, e_i \rangle e_i \right\|^2 + \left\| \sum_{i=1}^{n} \langle x, e_i \rangle e_i \right\|^2 \\ &< \varepsilon^2 + \sum_{i=1}^{\infty} |\langle x, e_i \rangle|^2. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, we deduce that

$$\|x\|^2 \leq \sum_{i=1}^{\infty} |\langle x, e_i \rangle|^2,$$

which together with the reversed inequality proved in (ii), establishes (iii).

(iv) The result (2.4.10) established in (iii) states that $\|\sum_{i=1}^{n} \langle x, e_i \rangle e_i - x\| \to 0$ as $n \to \infty$ for each $x \in \mathscr{H}$. By continuity of the inner product we therefore have, for each $x, y \in \mathscr{H}$,

$$\begin{aligned} \langle x, y \rangle &= \lim_{n \to \infty} \left\langle \sum_{i=1}^{n} \langle x, e_i \rangle e_i, \sum_{j=1}^{n} \langle y, e_j \rangle e_j \right\rangle \\ &= \lim_{n \to \infty} \sum_{i=1}^{n} \langle x, e_i \rangle \langle e_i, y \rangle \\ &= \sum_{i=1}^{\infty} \langle x, e_i \rangle \langle e_i, y \rangle. \end{aligned}$$

(v) This result is an immediate consequence of (ii). □

Remark 3. Separable Hilbert spaces are frequently encountered as the closed spans of countable subsets of possibly non-separable Hilbert spaces.

§2.5 Projection in $\mathbb{R}^n$

In Examples 2.1.1, 2.1.2 and 2.2.1 we showed that $\mathbb{R}^n$ is a Hilbert space with the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i, \tag{2.5.1}$$

the corresponding squared norm

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2, \tag{2.5.2}$$

and the angle between $\mathbf{x}$ and $\mathbf{y}$,

$$\theta = \cos^{-1}\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \, \|\mathbf{y}\|}\right). \tag{2.5.3}$$

Every closed subspace $\mathscr{M}$ of the Hilbert space $\mathbb{R}^n$ can be expressed by means of Gram–Schmidt orthogonalization (see for example Simmons (1963)) as $\mathscr{M} = \overline{\text{sp}}\{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ where $\{\mathbf{e}_1, \dots, \mathbf{e}_m\}$ is an orthonormal subset of $\mathscr{M}$ and m $(\leq n)$ is called the dimension of $\mathscr{M}$ (see also Problem 2.14). If $m < n$ then there is an orthonormal subset $\{\mathbf{e}_{m+1}, \dots, \mathbf{e}_n\}$ of $\mathscr{M}^\perp$ such that $\mathscr{M}^\perp = \overline{\text{sp}}\{\mathbf{e}_{m+1}, \dots, \mathbf{e}_n\}$. By Proposition 2.3.2 (iii) every $\mathbf{x} \in \mathbb{R}^n$ can be expressed uniquely as a sum of two elements of $\mathscr{M}$ and $\mathscr{M}^\perp$ respectively, namely

$$\mathbf{x} = P_{\mathscr{M}}\mathbf{x} + (I - P_{\mathscr{M}})\mathbf{x}, \tag{2.5.4}$$

where, by Theorem 2.4.1,

$$P_{\mathscr{M}}\mathbf{x} = \sum_{i=1}^{m} \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i \tag{2.5.5}$$

and

$$(I - P_{\mathscr{M}})\mathbf{x} = \sum_{i=m+1}^{n} \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i. \tag{2.5.6}$$

The following theorem enables us to compute $P_{\mathscr{M}}\mathbf{x}$ directly from any specified set of vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ spanning $\mathscr{M}$.

Theorem 2.5.1. *If* $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, \dots, m$, *and* $\mathscr{M} = \overline{\text{sp}}\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ *then*

$$P_{\mathscr{M}}\mathbf{x} = X\boldsymbol{\beta}, \tag{2.5.7}$$

where X *is the* $n \times m$ *matrix whose* j^{th} *column is* $\mathbf{x}_j$ *and*

$$X'X\boldsymbol{\beta} = X'\mathbf{x}. \tag{2.5.8}$$

Equation (2.5.8) *has at least one solution for* $\boldsymbol{\beta}$ *but* $X\boldsymbol{\beta}$ *is the same for all solutions. There is exactly one solution of* (2.5.8) *if and only if* $X'X$ *is non-singular and in this case*

$$P_{\mathscr{M}}\mathbf{x} = X(X'X)^{-1}X'\mathbf{x}. \tag{2.5.9}$$

PROOF. Since $P_{\mathscr{M}}\mathbf{x} \in \mathscr{M}$, we can write

$$P_{\mathscr{M}}\mathbf{x} = \sum_{i=1}^{m} \beta_i \mathbf{x}_i = X\boldsymbol{\beta}, \qquad \text{for some } \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)' \in \mathbb{R}^m. \tag{2.5.10}$$

The prediction equations (2.3.8) are equivalent in this case to

$$\langle X\boldsymbol{\beta}, \mathbf{x}_j \rangle = \langle \mathbf{x}, \mathbf{x}_j \rangle, \qquad j = 1, \ldots, m, \tag{2.5.11}$$

and in matrix form these equations can be written

$$X'X\boldsymbol{\beta} = X'\mathbf{x}. \tag{2.5.12}$$

The existence of at least one solution for $\boldsymbol{\beta}$ is guaranteed by the existence of the projection $P_{\mathscr{M}}\mathbf{x}$. The fact that $X\boldsymbol{\beta}$ is the same for all solutions is guaranteed by the uniqueness of $P_{\mathscr{M}}\mathbf{x}$. The last statement of the theorem follows at once from (2.5.7) and (2.5.8). □

Remark 1. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is an orthonormal set then $X'X$ is the identity matrix and so we find that

$$P_{\mathscr{M}}\mathbf{x} = XX'\mathbf{x} = \sum_{i=1}^{m} \langle \mathbf{x}, \mathbf{x}_i \rangle \mathbf{x}_i,$$

in accordance with (2.5.5)

Remark 2. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is a linearly independent set then there must be a unique vector $\boldsymbol{\beta}$ such that $P_{\mathscr{M}}\mathbf{x} = X\boldsymbol{\beta}$. This means that (2.5.8) must have a unique solution, which in turn implies that $X'X$ is non-singular and

$$P_{\mathscr{M}}\mathbf{x} = X(X'X)^{-1}X'\mathbf{x} \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

The matrix $X(X'X)^{-1}X'$ must be the same for all linearly independent sets $\{x_1, \ldots, x_m\}$ spanning $\mathscr{M}$ since $P_{\mathscr{M}}$ is a uniquely defined mapping on $\mathbb{R}^n$.

Remark 3. Given a real $n \times n$ matrix M, how can we tell whether or not there is a subspace $\mathscr{M}$ of $\mathbb{R}^n$ such that $M\mathbf{x} = P_{\mathscr{M}}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$? If there is such a subspace we say that M is a projection matrix. Such matrices are characterized in the next theorem.

Theorem 2.5.2. *The* $n \times n$ *matrix* M *is a projection matrix if and only if*

$$\text{(a)}\ M' = M$$

and

$$\text{(b)}\ M^2 = M.$$

Proof. If M is the projection matrix corresponding to some subspace $\mathcal{M}$ then by Remark 2 it can be written in the form $X(X'X)^{-1}X'$ where X is any matrix having linearly independent columns which span $\mathcal{M}$. It is easily verified that (a) and (b) are then satisfied.

Suppose now that (a) and (b) are satisfied. We shall show that $M\mathbf{x} = P_{\mathcal{M}}\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^n$ where $\mathcal{M}$ is the range of M defined by

$$R(M) = \{M\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\}.$$

First observe that $M\mathbf{x} \in R(M)$ by definition. Secondly we know that for any $\mathbf{y} \in R(M)$ there exists $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{y} = M\mathbf{w}$. Hence

$$\langle \mathbf{x} - M\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x} - M\mathbf{x}, M\mathbf{w} \rangle = \mathbf{x}'(I - M)'M\mathbf{w} = 0 \quad \text{for all } \mathbf{y} \in R(M),$$

showing that $M\mathbf{x}$ is indeed the required projection. □

§2.6 Linear Regression and the General Linear Model

Consider the problem of finding the "best" straight line

$$y = \theta_1 x + \theta_2, \tag{2.6.1}$$

or equivalently the best values $\hat{\theta}_1$, $\hat{\theta}_2$ of θ_1, $\theta_2 \in \mathbb{R}$, to fit a given set of data points (x_i, y_i), $i = 1, \ldots, n$. In least squares regression the best estimates $\hat{\theta}_1$, $\hat{\theta}_2$ are defined to be values of θ_1, θ_2 which minimize the sum,

$$S(\theta_1, \theta_2) = \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_2)^2,$$

of squared deviations of the observations y_i from the fitted values $\theta_1 x_i + \theta_2$.

This problem reduces to that of computing a projection in $\mathbb{R}^n$ as is easily seen by writing $S(\theta_1, \theta_2)$ in the equivalent form

$$S(\theta_1, \theta_2) = \|\mathbf{y} - \theta_1 \mathbf{x} - \theta_2 \mathbf{1}\|^2, \tag{2.6.2}$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$, $\mathbf{1} = (1, \ldots, 1)'$ and $\mathbf{y} = (y_1, \ldots, y_n)'$. By the projection theorem there is a unique vector of the form $(\hat{\theta}_1 \mathbf{x} + \hat{\theta}_2 \mathbf{1})$ which minimizes $S(\theta_1, \theta_2)$, namely $P_{\mathcal{M}}\mathbf{y}$ where $\mathcal{M} = \overline{\text{sp}}\{\mathbf{x}, \mathbf{1}\}$.

Defining X to be the $n \times 2$ matrix $X = [\mathbf{x}, \mathbf{1}]$ and $\hat{\boldsymbol{\theta}}$ to be the column vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)'$, we deduce from Theorem 2.5.1 that

$$P_{\mathcal{M}}\mathbf{y} = X\hat{\boldsymbol{\theta}}$$

where

$$X'X\hat{\boldsymbol{\theta}} = X'\mathbf{y}. \tag{2.6.3}$$

There is a unique solution $\hat{\boldsymbol{\theta}}$ if and only if $X'X$ is non-singular. In this case

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1}X'\mathbf{y}. \tag{2.6.4}$$

If $X'X$ is singular there are infinitely many solutions of (2.6.4), however by the uniqueness of $P_{\mathcal{M}}\mathbf{y}$, $X\hat{\boldsymbol{\theta}}$ is the same for all of them.

The argument just given applies equally well to least squares estimation for the general linear model. The general problem is as follows. Given a set of data points

$$(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}, y_i), \qquad i = 1, \ldots, n; \ m \le n,$$

we are required to find a value $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$ of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ which minimizes

$$\begin{aligned} S(\boldsymbol{\theta}) &= \sum_{i=1}^{n} (y_i - \theta_1 x_i^{(1)} - \cdots - \theta_m x_i^{(m)})^2 \\ &= \|\mathbf{y} - \theta_1 \mathbf{x}^{(1)} - \cdots - \theta_m \mathbf{x}^{(m)}\|^2, \end{aligned}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{x}^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})'$, $j = 1, \ldots, m$. By the projection theorem there is a unique vector of the form $(\hat{\theta}_1 \mathbf{x}^{(1)} + \cdots + \hat{\theta}_m \mathbf{x}^{(m)})$ which minimizes $S(\boldsymbol{\theta})$, namely $P_{\mathscr{M}}\mathbf{y}$ where $\mathscr{M} = \overline{\mathrm{sp}}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$.

Defining X to be the $n \times m$ matrix $X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}]$ and $\hat{\boldsymbol{\theta}}$ to be the column vector $\hat{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_m)'$, we deduce from Theorem 2.5.1 that

$$P_{\mathscr{M}}\mathbf{y} = X\hat{\boldsymbol{\theta}}$$

where

$$X'X\hat{\boldsymbol{\theta}} = X'\mathbf{y} \tag{2.6.5}$$

As in the special case of fitting a straight line, $\hat{\boldsymbol{\theta}}$ is uniquely defined if and only if $X'X$ is non-singular, in which case

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1}X'\mathbf{y}. \tag{2.6.6}$$

If $X'X$ is singular then there are infinitely many solutions of (2.6.5) but $X\hat{\boldsymbol{\theta}}$ is the same for all of them.

In spite of the assumed linearity in the parameters $\theta_1, \ldots, \theta_m$, the applications of the general linear model are very extensive. As a simple illustration, let us fit a quadratic function,

$$y = \theta_1 x^2 + \theta_2 x + \theta_3,$$

to the data

x	0	1	2	3	4
y	1	0	3	5	8

The matrix X for this problem is

$$X = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \end{bmatrix} \quad \text{giving } (X'X)^{-1} = \frac{1}{140}\begin{bmatrix} 10 & -40 & 20 \\ -40 & 174 & -108 \\ 20 & -108 & 124 \end{bmatrix}.$$

The least squares estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)'$ is therefore unique and is found from (2.6.6) to be

$$\hat{\boldsymbol{\theta}} = (0.5, -0.1, 0.6)'.$$

The vector of fitted values $X\hat{\boldsymbol{\theta}} = P_{\mathscr{M}}\mathbf{y}$ is given by

$$X\hat{\boldsymbol{\theta}} = (0.6, 1, 2.4, 4.8, 8.2)',$$

as compared with the vector of observations,

$$\mathbf{y} = (1, 0, 3, 5, 8)'.$$

§2.7 Mean Square Convergence, Conditional Expectation and Best Linear Prediction in $L^2(\Omega, \mathscr{F}, P)$

All results in this section will be stated for the real Hilbert space $L^2 = L^2(\Omega, \mathscr{F}, P)$ with inner product $\langle X, Y \rangle = E(XY)$. The reader should have no difficulty however in writing down analogous results for the complex space $L^2(\Omega, \mathscr{F}, P)$ with inner product $\langle X, Y \rangle = E(X\bar{Y})$. As indicated in Example 2.2.2, mean square convergence is just another name for norm convergence in L^2, i.e. if $X_n, X \in L^2$, then

$$X_n \xrightarrow{\text{m.s.}} X \quad \text{if and only if } \|X_n - X\|^2 = E|X_n - X|^2 \to 0 \text{ as } n \to \infty. \tag{2.7.1}$$

By simply restating properties already established for norm convergence we obtain the following proposition.

Proposition 2.7.1 (Properties of Mean Square Convergence).
(a) *X_n converges in mean square if and only if* $E|X_m - X_n|^2 \to 0$ as $m, n \to \infty$.
(b) *If $X_n \xrightarrow{\text{m.s.}} X$ and $Y_n \xrightarrow{\text{m.s.}} Y$ then as $n \to \infty$,*

$$\text{(i)}\ EX_n = \langle X_n, 1 \rangle \to \langle X, 1 \rangle = EX,$$

$$\text{(ii)}\ E|X_n|^2 = \langle X_n, X_n \rangle \to \langle X, X \rangle = E|X|^2,$$

and

$$\text{(iii)}\ E(X_n Y_n) = \langle X_n, Y_n \rangle \to \langle X, Y \rangle = E(XY).$$

Definition 2.7.1 (Best Mean Square Predictor of Y). If $\mathscr{M}$ is a closed subspace of L^2 and $Y \in L^2$, then the best mean square predictor of Y in $\mathscr{M}$ is the element $\hat{Y} \in \mathscr{M}$ such that

$$\|Y - \hat{Y}\|^2 = \inf_{Z \in \mathscr{M}} \|Y - Z\|^2 = \inf_{Z \in \mathscr{M}} E|Y - Z|^2. \tag{2.7.2}$$

The projection theorem immediately identifies the unique best predictor of Y in $\mathscr{M}$ as $P_{\mathscr{M}} Y$. By imposing a little more structure on the closed subspace

$\mathcal{M}$, we are led from Definition 2.7.1 to the notions of conditional expectation and best linear predictor.

Definition 2.7.2 (The Conditional Expectation, $E_{\mathcal{M}}X$). If $\mathcal{M}$ is a closed subspace of L^2 containing the constant functions, and if $X \in L^2$, then we define the conditional expectation of X given $\mathcal{M}$ to be the projection,

$$E_{\mathcal{M}}X = P_{\mathcal{M}}X. \tag{2.7.3}$$

Using the definition of the inner product in L^2 and the prediction equations (2.3.8) we can state equivalently that $E_{\mathcal{M}}X$ is the unique element of $\mathcal{M}$ such that

$$E(WE_{\mathcal{M}}X) = E(WX) \quad \text{for all } W \in \mathcal{M}. \tag{2.7.4}$$

Obviously the operator $E_{\mathcal{M}}$ on L^2 has all the properties of a projection operator, in particular (see Proposition 2.3.2)

$$E_{\mathcal{M}}(aX + bY) = aE_{\mathcal{M}}X + bE_{\mathcal{M}}Y, \qquad a, b \in \mathbb{R}, \tag{2.7.5}$$

$$E_{\mathcal{M}}X_n \xrightarrow{\text{m.s.}} E_{\mathcal{M}}X \quad \text{if } X_n \xrightarrow{\text{m.s.}} X \tag{2.7.6}$$

and

$$E_{\mathcal{M}_1}(E_{\mathcal{M}_2}X) = E_{\mathcal{M}_1}X \quad \text{if } \mathcal{M}_1 \subseteq \mathcal{M}_2. \tag{2.7.7}$$

Notice also that

$$E_{\mathcal{M}}1 = 1 \tag{2.7.8}$$

and if $\mathcal{M}_0$ is the closed subspace of L^2 consisting of all the constant functions, then an application of the prediction equations (2.3.8) gives

$$E_{\mathcal{M}_0}X = EX. \tag{2.7.9}$$

Definition 2.7.3 (The Conditional Expectation $E(X|Z)$). If Z is a random variable on $(\Omega, \mathcal{F}, P)$ and $X \in L^2(\Omega, \mathcal{F}, P)$ then the conditional expectation of X given Z is defined to be

$$E(X|Z) = E_{\mathcal{M}(Z)}X, \tag{2.7.10}$$

where $\mathcal{M}(Z)$ is the closed subspace of L^2 consisting of all random variables in L^2 which can be written in the form $\phi(Z)$ for some Borel function $\phi: \mathbb{R} \to \mathbb{R}$. (For the proof that $\mathcal{M}(Z)$ is a closed subspace see Problem 2.25.)

The operator $E_{\mathcal{M}(Z)}$ has all the properties (2.7.5)–(2.7.8), and in addition

$$E_{\mathcal{M}(Z)}X \geq 0 \quad \text{if } X \geq 0. \tag{2.7.11}$$

Definition 2.7.3 can be extended in a fairly obvious way as follows: if $Z_1, \ldots, Z_n$ are random variables on $(\Omega, \mathcal{F}, P)$ and $X \in L^2$, then we define

$$E(X|Z_1, \ldots, Z_n) = E_{\mathcal{M}(Z_1, \ldots, Z_n)}(X), \tag{2.7.12}$$

where $\mathcal{M}(Z_1, \ldots, Z_n)$ is the closed subspace of L^2 consisting of all random variables in L^2 of the form $\phi(Z_1, \ldots, Z_n)$ for some Borel function $\phi: \mathbb{R}^n \to \mathbb{R}$. The properties of $E_{\mathcal{M}(Z)}$ listed above all carry over to $E_{\mathcal{M}(Z_1, \ldots, Z_n)}$.

Conditional Expectation and Best Linear Prediction. By the projection theorem, the conditional expectation $E_{\mathcal{M}(Z_1, \ldots, Z_n)}(X)$ is the best mean square predictor of X in $\mathcal{M}(Z_1, \ldots, Z_n)$, i.e. it is the best function of $Z_1, \ldots, Z_n$ (in the m.s. sense) for predicting X. However the determination of projections on $\mathcal{M}(Z_1, \ldots, Z_n)$ is usually very difficult because of the complex nature of the equations (2.7.4). On the other hand if $Z_1, \ldots, Z_n \in L^2$, it is relatively easy to compute instead the projection of X on $\overline{\text{sp}}\{1, Z_1, \ldots, Z_n\} \subseteq \mathcal{M}(Z_1, \ldots, Z_n)$ since we can write

$$P_{\overline{\text{sp}}\{1, Z_1, \ldots, Z_n\}}(X) = \sum_{i=0}^{n} \alpha_i Z_i, \quad Z_0 = 1, \tag{2.7.13}$$

where $\alpha_0, \ldots, \alpha_n$ satisfy

$$\left\langle \sum_{i=0}^{n} \alpha_i Z_i, Z_j \right\rangle = \langle X, Z_j \rangle, \quad j = 0, 1, \ldots, n, \tag{2.7.14}$$

or equivalently,

$$\sum_{i=0}^{n} \alpha_i E(Z_i Z_j) = E(X Z_j), \quad j = 0, 1, \ldots, n. \tag{2.7.15}$$

The projection theorem guarantees that a solution $(\alpha_0, \ldots, \alpha_n)$ exists. Any solution, when substituted into (2.7.13) gives the required projection, known as the best *linear* predictor of X in terms of $1, Z_1, \ldots, Z_n$. As a projection of X onto a subspace of $\mathcal{M}(Z_1, \ldots, Z_n)$ it can never have smaller mean squared error than $E_{\mathcal{M}(Z_1, \ldots, Z_n)}X$. Nevertheless it is of great importance for the following reasons:

(a) it is easier to calculate than $E_{\mathcal{M}(Z_1, \ldots, Z_n)}(X)$,
(b) it depends only on the first and second order moments, EX, EZ_i, $E(Z_i Z_j)$ and $E(XZ_j)$ of the joint distribution of $(X, Z_1, \ldots, Z_n)$,
(c) if $(X, Z_1, \ldots, Z_n)'$ has a multivariate normal distribution then (see Problem 2.20),

$$P_{\overline{\text{sp}}\{1, Z_1, \ldots, Z_n\}}(X) = E_{\mathcal{M}(Z_1, \ldots, Z_n)}(X).$$

Best linear predictors are defined more generally as follows:

Definition 2.7.4 (Best Linear Predictor of X in Terms of $\{Z_\lambda, \lambda \in \Lambda\}$). If $X \in L^2$ and $Z_\lambda \in L^2$ for all $\lambda \in \Lambda$, then the best linear predictor of X in terms of $\{Z_\lambda, \lambda \in \Lambda\}$ is defined to be the element of $\overline{\text{sp}}\{Z_\lambda, \lambda \in \Lambda\}$ with smallest mean square distance from X. By the projection theorem this is just $P_{\overline{\text{sp}}\{Z_\lambda, \lambda \in \Lambda\}}X$.

EXAMPLE 2.7.1. Suppose $Y = X^2 + Z$ where X and Z are independent standard normal random variables. The best predictor of Y in terms of X

is $E(Y|X) = X^2$. (The reader should check that the defining properties of $E(Y|X) = E_{\mathscr{M}(X)}Y$ are satisfied by X^2, i.e. that $X^2 \in \mathscr{M}(X)$ and that (2.7.4) is satisfied with $\mathscr{M} = \mathscr{M}(X)$.) On the other hand the best *linear* predictor of Y in terms of $\{1, X\}$ is

$$P_{\overline{\text{sp}}\{1,X\}} Y = aX + b,$$

where, by the prediction equations (2.7.15),

$$\langle aX + b, X \rangle = \langle Y, X \rangle = E(YX) = 0$$

and

$$\langle aX + b, 1 \rangle = \langle Y, 1 \rangle = E(Y) = 1.$$

Hence $a = 0$ and $b = 1$ so that

$$P_{\overline{\text{sp}}\{1,X\}} Y = 1.$$

The mean squared errors of the two predictors are

$$\|E(Y|X) - Y\|^2 = E(Z^2) = 1,$$

and

$$\|Y - P_{\overline{\text{sp}}\{1,X\}} Y\|^2 = \|Y\|^2 - 1 = E(X^4) + E(Z^2) - 1 = 3,$$

showing the substantial superiority of the best predictor over the best linear predictor in this case.

Remark 1. The conditional expectation operators $E_{\mathscr{M}(Z)}$ and $E_{\mathscr{M}(Z_1,\ldots,Z_n)}$ are usually defined on the space $L^1(\Omega, \mathscr{F}, P)$ of random variables X such that $E|X| < \infty$ (see e.g. Breiman (1968), Chapter 4). The restrictions of these operators to $L^2(\Omega, \mathscr{F}, P)$ coincide with $E_{\mathscr{M}(Z)}$ and $E_{\mathscr{M}(Z_1,\ldots,Z_n)}$ as we have defined them.

§2.8 Fourier Series

Consider the complex Hilbert space $L^2[-\pi, \pi] = L^2([-\pi, \pi], \mathscr{B}, U)$ where $\mathscr{B}$ consists of the Borel subsets of $[-\pi, \pi]$, U is the uniform probability measure $U(dx) = (2\pi)^{-1}\,dx$, and the inner product of f, $g \in L^2[-\pi, \pi]$ is defined as usual by

$$\langle f, g \rangle = Ef\bar{g} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\bar{g}(x)\,dx. \tag{2.8.1}$$

The functions $\{e_n,\ n \in \mathbb{Z}\}$ defined by

$$e_n(x) = e^{inx}, \tag{2.8.2}$$

are orthonormal in $L^2[-\pi,\pi]$ since

$$\begin{aligned}\langle e_m, e_n\rangle &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(m-n)x}\,dx\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}[\cos(m-n)x + i\sin(m-n)x]\,dx\\ &= \begin{cases}1 & \text{if } m=n\\ 0 & \text{if } m\neq n.\end{cases}\end{aligned}$$

Definition 2.8.1 (Fourier Approximations and Coefficients). The n^{th} order Fourier approximation to any function $f\in L^2[-\pi,\pi]$ is defined to be the projection of f onto $\overline{\text{sp}}\{e_j, |j|\le n\}$, which by Theorem 2.4.1 is

$$S_n f = \sum_{j=-n}^{n}\langle f, e_j\rangle e_j. \tag{2.8.3}$$

The coefficients

$$\langle f, e_j\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ijx}\,dx \tag{2.8.4}$$

are called the Fourier coefficients of the function f.

We can write (2.8.3) a little more explicitly in the form

$$S_n f(x) = \sum_{j=-n}^{n}\langle f, e_j\rangle e^{ijx}, \qquad x\in[-\pi,\pi], \tag{2.8.5}$$

and one is naturally led to investigate the senses (if any) in which the sequence of functions $\{S_n f\}$ converges to f as $n\to\infty$. In this section we shall restrict attention to mean square convergence, deferring questions of pointwise and uniform convergence to Section 2.11.

Theorem 2.8.1. (a) *The sequence $\{S_n f\}$ has a mean square limit as $n\to\infty$ which we shall denote by $\sum_{j=-\infty}^{\infty}\langle f, e_j\rangle e_j$ or Sf.*

(b) $Sf = f$.

PROOF. (a) From Bessel's inequality (2.4.8) we have $\sum_{|j|\le n}|\langle f, e_j\rangle|^2 \le \|f\|^2$ for all n which implies that $\sum_{j=-\infty}^{\infty}|\langle f, e_j\rangle|^2 < \infty$. Hence for $n > m \ge 1$,

$$\|S_n f - S_m f\|^2 \le \sum_{|j|>m}|\langle f, e_j\rangle|^2 \to 0 \quad \text{as } m\to\infty,$$

showing that $\{S_n f\}$ is a Cauchy sequence and therefore has a mean square limit.

(b) For $|j|\le n$, $\langle S_n f, e_j\rangle = \langle f, e_j\rangle$, so by continuity of the inner product

$$\langle Sf, e_j\rangle = \lim_{n\to\infty}\langle S_n f, e_j\rangle = \langle f, e_j\rangle \quad \text{for all } j\in\mathbb{Z}.$$

In Theorem 2.11.2 we shall show that $\langle g, e_j \rangle = 0$ for all $j \in \mathbb{Z}$ implies that $g = 0$. Hence $Sf - f = 0$. □

Corollary 2.8.1. $L^2[-\pi, \pi] = \overline{\text{sp}}\{e_j, j \in \mathbb{Z}\}$.

PROOF. Any $f \in L^2[-\pi, \pi]$ can be expressed as the mean square limit of $S_n f$ where $S_n f \in \overline{\text{sp}}\{e_j, j \in \mathbb{Z}\}$. Since $\overline{\text{sp}}\{e_j, j \in \mathbb{Z}\}$ is by definition closed it must contain f. Hence $\overline{\text{sp}}\{e_j, j \in \mathbb{Z}\} \supseteq L^2[-\pi, \pi]$. □

Corollary 2.8.2. (a) $\|f\|^2 = \sum_{j=-\infty}^{\infty} |\langle f, e_j \rangle|^2$.
(b) $\langle f, g \rangle = \sum_{j=-\infty}^{\infty} \langle f, e_j \rangle \overline{\langle g, e_j \rangle}$.

PROOF. Corollary 2.8.1 implies that the conditions of Theorem 2.4.2 are satisfied. □

§2.9 Hilbert Space Isomorphisms

Definition 2.9.1 (Isomorphism). An isomorphism of the Hilbert space $\mathscr{H}_1$ onto the Hilbert space $\mathscr{H}_2$ is a one to one mapping T of $\mathscr{H}_1$ onto $\mathscr{H}_2$ such that for all $f_1, f_2 \in \mathscr{H}_1$,

$$\text{(a)} \quad T(af_1 + bf_2) = aTf_1 + bTf_2 \text{ for all scalars } a \text{ and } b$$

and

$$\text{(b)} \quad \langle Tf_1, Tf_2 \rangle = \langle f_1, f_2 \rangle.$$

We say that $\mathscr{H}_1$ and $\mathscr{H}_2$ are isomorphic if there is an isomorphism T of $\mathscr{H}_1$ onto $\mathscr{H}_2$. The inverse mapping T^{-1} is then an isomorphism of $\mathscr{H}_2$ onto $\mathscr{H}_1$.

Remark 1. In this book we shall always use the term isomorphism to indicate that both (a) and (b) are satisfied. Elsewhere the term is frequently used to denote a mapping satisfying (a) only.

EXAMPLE 2.9.1 (The Space l^2). Let l^2 denote the complex Hilbert space of sequences $\{z_n, n = 1, 2, \ldots\}$, $z_n \in \mathbb{C}$, $\sum_{n=1}^{\infty} |z_n^2| < \infty$, with inner product

$$\langle \{y_n\}, \{z_n\} \rangle = \sum_{i=1}^{\infty} y_i \bar{z}_i.$$

(For the proof that l^2 is a separable Hilbert space see Problem 2.23.) If now $\mathscr{H}$ is any Hilbert space with an orthonormal basis $\{e_n, n = 1, 2, \ldots\}$ then the mapping $T: \mathscr{H} \to l^2$ defined by

$$Th = \{\langle h, e_n \rangle\} \tag{2.9.1}$$

is an isomorphism of $\mathscr{H}$ onto l^2 (see Problem 2.24). Thus every separable Hilbert space is isomorphic to l^2.

Properties of Isomorphisms. Suppose T is an isomorphism of $\mathscr{H}_1$ onto $\mathscr{H}_2$. We then have the following properties, all of which follow at once from the definitions:

(i) If $\{e_n\}$ is a complete orthonormal set in $\mathscr{H}_1$ then $\{Te_n\}$ is a complete orthonormal set in $\mathscr{H}_2$.
(ii) $\|Tx\| = \|x\|$ for all $x \in \mathscr{H}_1$.
(iii) $\|x_n - x\| \to 0$ if and only if $\|Tx_n - Tx\| \to 0$.
(iv) $\{x_n\}$ is a Cauchy sequence if and only if $\{Tx_n\}$ is a Cauchy sequence.
(v) $TP_{\overline{\mathrm{sp}}\{x_\lambda, \lambda \in \Lambda\}}(x) = P_{\overline{\mathrm{sp}}\{Tx_\lambda, \lambda \in \Lambda\}}(Tx)$.

The last property is the basis for the spectral theory of prediction of a stationary process $\{X_t, t \in \mathbb{Z}\}$ (Section 5.6), in which we use the fact that the mapping

$$TX_t = e^{it\cdot}$$

defines an isomorphism of a certain Hilbert space of random variables onto a Hilbert space $L^2([-\pi, \pi], \mathscr{B}, \mu)$ with μ a finite measure. The problem of computing projections in the former space can then be tranformed by means of (v) into the problem of computing projections in the latter.

§2.10* The Completeness of $L^2(\Omega, \mathscr{F}, P)$

We need to show that if $X_n \in L^2$, $n = 1, 2, \ldots$, and $\|X_n - X_m\| \to 0$ as $m, n \to \infty$, then there exists $X \in L^2$ such that $X_n \xrightarrow{\text{m.s.}} X$. This will be shown by identifying X as the limit of a sufficiently rapidly converging subsequence of $\{X_n\}$. We first need a proposition.

Proposition 2.10.1. *If $X_n \in L^2$ and $\|X_{n+1} - X_n\| \le 2^{-n}$, $n = 1, 2, \ldots$, then there is a random variable X on $(\Omega, \mathscr{F}, P)$ such that $X_n \to X$ with probability one.*

Proof. Let $X_0 = 0$. Then $X_n = \sum_{j=1}^{n} (X_j - X_{j-1})$. Now $\sum_{j=1}^{\infty} |X_j - X_{j-1}|$ is finite with probability one since, by the monotone convergence theorem and the Cauchy–Schwarz inequality,

$$E \sum_{j=1}^{\infty} |X_j - X_{j-1}| = \sum_{j=1}^{\infty} E|X_j - X_{j-1}| \le \sum_{j=1}^{\infty} \|X_j - X_{j-1}\| \le \|X_1\| + \sum_{j=1}^{\infty} 2^{-j} < \infty.$$

It follows that $\lim_{n\to\infty} \sum_{j=1}^{n} |X_j - X_{j-1}|$ (and hence $\lim_{n\to\infty} \sum_{j=1}^{n} (X_j - X_{j-1}) = \lim_{n\to\infty} X_n$) exists and is finite with probability one. □

Theorem 2.10.1. *$L^2(\Omega, \mathscr{F}, P)$ is complete.*

Proof. If $\{X_n\}$ is a Cauchy sequence in L^2 then we can find integers $n_1, n_2, \ldots$, such that $n_1 < n_2 < \cdots$ and

$$\|X_n - X_m\| \le 2^{-k} \quad \text{for } n, m > n_k. \tag{2.10.1}$$

(First choose n_1 to satisfy (2.10.1) with $k = 1$, then successively choose n_2, n_3, ..., to satisfy the appropriate conditions.)

By Proposition 2.10.1 there is a random variable X such that $X_{n_k} \to X$ with probability one as $k \to \infty$. Now

$$\|X_n - X\|^2 = \int |X_n - X|^2 \, dP = \int \liminf_{k\to\infty} |X_n - X_{n_k}|^2 \, dP,$$

and so by Fatou's lemma,

$$\|X_n - X\|^2 \le \liminf_{k\to\infty} \|X_n - X_{n_k}\|^2. \tag{2.10.2}$$

The right-hand side of (2.10.2) can be made arbitrarily small by choosing n large enough since $\{X_n\}$ is a Cauchy sequence. Consequently $\|X_n - X\|^2 \to 0$. The fact that $E|X|^2 < \infty$ follows from the triangle inequality

$$\|X\| \le \|X_n - X\| + \|X_n\|,$$

the right-hand side of which is certainly finite for large enough n. □

§2.11* Complementary Results for Fourier Series

The terminology and notation of Section 2.8 will be retained throughout this section. We begin with the classical result that trigonometric polynomials are uniformly dense in the space of continuous functions f which are defined on $[-\pi, \pi]$ and which satisfy the condition $f(\pi) = f(-\pi)$.

Theorem 2.11.1. *Let f be a continuous function on $[-\pi, \pi]$ such that $f(\pi) = f(-\pi)$. Then*

$$n^{-1}(S_0 f + S_1 f + \cdots + S_{n-1} f) \to f \tag{2.11.1}$$

uniformly on $[-\pi, \pi]$ as $n \to \infty$.

Proof. By definition of the n^{th} order Fourier approximation,

$$\begin{aligned} S_n f(x) &= \sum_{|j| \le n} \langle f, e_j \rangle e_j \\ &= (2\pi)^{-1} \int_{-\pi}^{\pi} f(y) \sum_{|j| \le n} e^{ij(x-y)} \, dy, \end{aligned}$$

which by defining $f(x) = f(x + 2\pi)$, $x \in \mathbb{R}$, can be rewritten as

$$S_n f(x) = (2\pi)^{-1} \int_{-\pi}^{\pi} f(x - y) D_n(y) \, dy, \tag{2.11.2}$$

where $D_n(y)$ is the Dirichlet kernel,

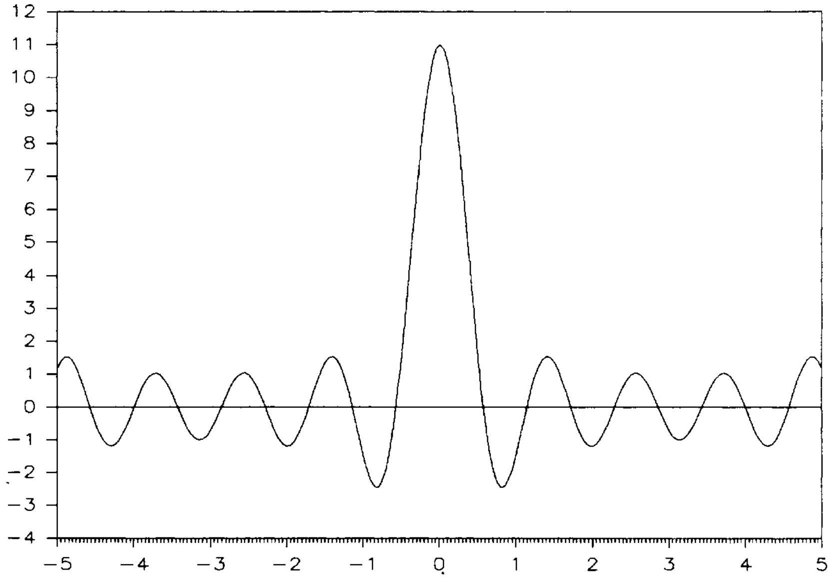

Figure 2.2. The Dirichlet kernel $D_5(x)$, $-5 \le x \le 5$ ($D_n(\cdot)$ has period 2π).

$$D_n(y) = \sum_{|j|\le n} e^{ijy} = \frac{e^{i(n+1/2)y} - e^{-i(n+1/2)y}}{e^{iy/2} - e^{-iy/2}} = \begin{cases} \dfrac{\sin[(n+\frac{1}{2})y]}{\sin(\frac{1}{2}y)} & \text{if } y \ne 0, \\ 2n+1 & \text{if } y = 0. \end{cases} \tag{2.11.3}$$

A graph of the function D_n is shown in Figure 2.2. For the function $f(x) \equiv 1$, $\langle f, e_0\rangle = 1$ and $\langle f, e_j\rangle = 0, j \ne 0$. Hence $S_n 1(x) \equiv 1$, and substituting this in (2.11.2) we find that

$$(2\pi)^{-1} \int_{-\pi}^{\pi} D_n(y)\,dy = 1. \tag{2.11.4}$$

Making use of (2.11.2) we can now write

$$n^{-1}(S_0 f(x) + \cdots + S_{n-1} f(x)) = \int_{-\pi}^{\pi} f(x-y) K_n(y)\,dy, \tag{2.11.5}$$

where $K_n(y)$ is the Fejer kernel,

$$K_n(y) = \frac{1}{2\pi n} \sum_{j=0}^{n-1} D_j(y) = \frac{\sum_{j=0}^{n-1} \sin[(j+\frac{1}{2})y]}{2\pi n \sin(\frac{1}{2}y)}.$$

Evaluating the sum with the aid of the identity,

$$2\sin(\tfrac{1}{2}y)\sin[(j+\tfrac{1}{2})y] = \cos(jy) - \cos[(j+1)y],$$

we find that

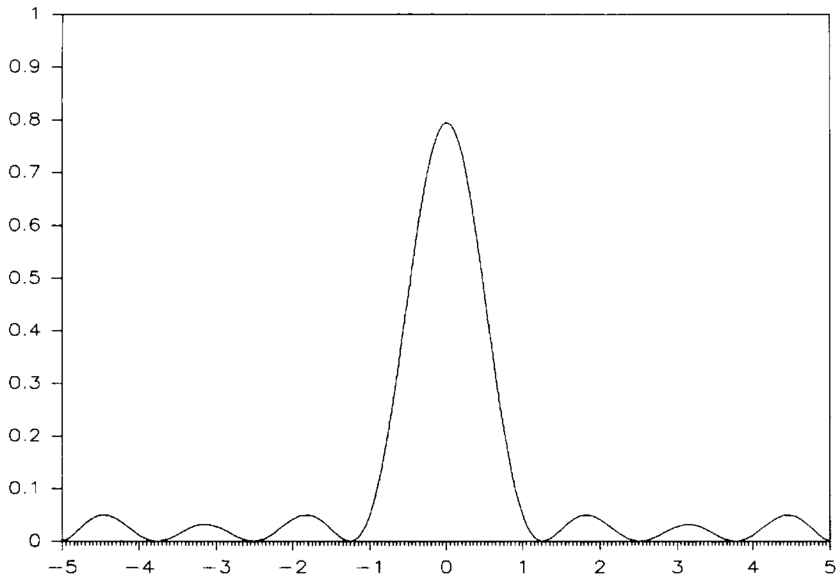

Figure 2.3. The Fejer kernel $K_5(x)$, $-5 \le x \le 5$ ($K_n(\cdot)$ has period 2π).

$$K_n(y) = \begin{cases} \dfrac{1}{2\pi n} \dfrac{\sin^2(ny/2)}{\sin^2(y/2)} & \text{if } y \neq 0, \\ \dfrac{n}{2\pi} & \text{if } y = 0. \end{cases} \tag{2.11.6}$$

The Fejer kernel is shown in Figure 2.3. It has the properties,

(a) $K_n(y) \ge 0$ (unlike $D_n(y)$),
(b) $K_n(\cdot)$ has period 2π,
(c) $K_n(\cdot)$ is an even function
(d) $\int_{-\pi}^{\pi} K_n(y)\, dy = 1$,
(e) for each $\delta > 0$, $\int_{-\delta}^{\delta} K_n(y)\, dy \to 1$ as $n \to \infty$.

The first three properties are evident from (2.11.6). Property (d) is obtained by setting $f(x) \equiv 1$ in (2.11.5). To establish (e), observe that

$$K_n(y) \le \frac{1}{2\pi n \sin^2(\delta/2)} \quad \text{for } 0 < \delta < |y| \le \pi.$$

For each $\delta > 0$ this inequality implies that

$$\int_{-\pi}^{-\delta} K_n(y)\, dy + \int_{\delta}^{\pi} K_n(y)\, dy \to 0 \quad \text{as } n \to \infty,$$

which, together with property (d), proves (e).

Now for any continuous function f with period 2π, we have from (2.11.5) and property (d) of $K_n(.)$,

$$\begin{aligned}\Delta_n(x) &\equiv |n^{-1}(S_0 f(x) + \cdots + S_{n-1}f(x)) - f(x)| \\ &= \left|\int_{-\pi}^{\pi} f(x-y)K_n(y)\,dy - f(x)\right| \\ &= \left|\int_{-\pi}^{\pi} [f(x-y) - f(x)]K_n(y)\,dy\right|.\end{aligned}$$

Hence for each $\delta > 0$,

$$\begin{aligned}\Delta_n(x) \le &\left|\int_{-\delta}^{\delta} [f(x-y) - f(x)]K_n(y)\,dy\right| \\ &+ \left|\int_{[-\pi,\pi]\setminus(-\delta,\delta)} [f(x-y) - f(x)]K_n(y)\,dy\right|.\end{aligned} \tag{2.11.7}$$

Since a continuous function with period 2π is uniformly continuous, we can choose for any $\varepsilon > 0$, a value of δ such that $\sup_{-\pi\le x\le\pi}|f(x-y) - f(x)| < \varepsilon$ whenever $|y| < \delta$. The first term on the right of (2.11.7) is then bounded by $\varepsilon\int_{-\pi}^{\pi} K_n(y)\,dy$ and the second by $2M(1 - \int_{-\delta}^{\delta} K_n(y)\,dy)$ where $M = \sup_{-\pi\le x\le\pi}|f(x)|$. Hence

$$\begin{aligned}\sup_{-\pi\le x\le\pi} \Delta_n(x) &\le \varepsilon\int_{-\delta}^{\delta} K_n(dy)\,dy + 2M\left(1 - \int_{-\delta}^{\delta} K_n(y)\,dy\right) \\ &\to \varepsilon \quad \text{as } n \to \infty.\end{aligned}$$

But since ε was arbitrary and $\Delta_n(x) \ge 0$, we conclude that $\Delta_n(x) \to 0$ uniformly on $[-\pi, \pi]$ as required. □

Remark 1. Under additional smoothness conditions on f, $S_n f$ may converge to f in a much stronger sense. For example if the derivative f' exists and $f' \in L^2[-\pi, \pi]$, then $S_n f$ converges absolutely and uniformly to f (see Churchill (1969) and Problem 2.22).

Theorem 2.11.2. *If $f \in L^2[-\pi, \pi]$ and $\langle f, e_j\rangle = 0$ for all $j \in \mathbb{Z}$, then $f = 0$ almost everywhere.*

Proof. It sufficies to show that $\int_A f(x)\,dx = 0$ for all Borel subsets A of $[-\pi, \pi]$ or, equivalently, by a monotone class argument (see Billingsley (1986)),

$$(2\pi)^{-1}\int_a^b f(x)\,dx = \langle f, I_{[a,b]}\rangle = 0 \tag{2.11.8}$$

for all subintervals $[a, b]$ of $[-\pi, \pi]$. Here $I_{[a,b]}$ denotes the indicator function of $[a, b]$.

To establish (2.11.8) we first show that $\langle f, g\rangle = 0$ for any continuous function g on $[-\pi, \pi]$ with $g(-\pi) = g(\pi)$. By Theorem 2.11.1 we know that

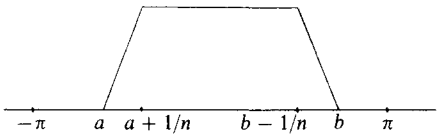

Figure 2.4. The continuous function h_n approximating $I_{[a,b]}$.

for g continuous, $g_n = n^{-1}(S_0 g + \cdots + S_{n-1} g) \to g$ uniformly on $[-\pi, \pi]$, implying in particular that

$$g_n \xrightarrow{\text{m.s.}} g.$$

By assumption $\langle f, g_n \rangle = 0$, so by continuity of the inner product,

$$\langle f, g \rangle = \lim_{n \to \infty} \langle f, g_n \rangle = 0.$$

The next step is to find a sequence $\{h_n\}$ of continuous functions such that $h_n \xrightarrow{\text{m.s.}} I_{[a,b]}$. One such sequence is defined by

$$h_n(x) = \begin{cases} 0 & \text{if } -\pi \le x \le a, \\ n(x-a) & \text{if } a \le x \le a + 1/n, \\ 1 & \text{if } a + 1/n \le x \le b - 1/n, \\ -n(x-b) & \text{if } b - 1/n \le x \le b, \\ 0 & \text{if } b \le x \le \pi, \end{cases}$$

since $\|I_{[a,b]} - h_n\|^2 \le (1/2\pi)(2/n) \to 0$ as $n \to \infty$. (See Figure 2.4.) Using the continuity of the inner product again,

$$\langle f, I_{[a,b]} \rangle = \lim_{n \to \infty} \langle f, h_n \rangle = 0. \qquad \square$$

Problems

2.1. Prove the parallelogram law (2.1.9).

2.2. If $\{X_t, t = 0, \pm 1, \ldots\}$ is a stationary process with mean zero and autocovariance function $\gamma(\cdot)$, show that $Y_n = \sum_{k=1}^{n} a_k X_k$ converges in mean square if $\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i a_j \gamma(i-j)$ is finite.

2.3. Show that if $\{X_t, t = 0, \pm 1, \ldots\}$ is stationary and $|\theta| < 1$ then for each n, $\sum_{j=1}^{m} \theta^j X_{n+1-j}$ converges in mean square as $m \to \infty$.

2.4. If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$, show that $(\mathcal{M}^\perp)^\perp = \mathcal{M}$.

2.5. If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$, prove that

$$\min_{y \in \mathcal{M}} \|x - y\| = \max\{|\langle x, z \rangle| : z \in \mathcal{M}^\perp, \|z\| = 1\}.$$

2.6. Verify the calculations of ϕ_1 and ϕ_2 in Example 2.3.4. Also check that $X_3 = (2\cos\omega)X_2 - X_1$.

2.7. If $\mathscr{H}$ is a complex Hilbert space and $x_i \in \mathscr{H}$, $i = 1, \ldots, n$, show that $\overline{\text{sp}}\{x_1, \ldots, x_n\} = \{\sum_{j=1}^n \alpha_j x_j : \alpha_j \in \mathbb{C}, j = 1, \ldots, n\}$.

2.8. Suppose that $\{X_t, t = 1, 2, \ldots\}$ is a stationary process with mean zero. Show that $P_{\overline{\text{sp}}\{1, X_1, \ldots, X_n\}} X_{n+1} = P_{\overline{\text{sp}}\{X_1, \ldots, X_n\}} X_{n+1}$.

2.9. (a) Let $\mathscr{H} = L^2([-1,1], \mathscr{B}[-1,1], \mu)$ where $d\mu = dx$ is Lebesgue measure on $[-1,1]$. Use the prediction equations to find constants α_0, α_1 and α_2 which minimize

$$\int_{-1}^{1} |e^x - \alpha_0 - \alpha_1 x - \alpha_2 x^2|^2 \, dx.$$

(b) Find $\max_{\{g \in \mathscr{M}^\perp, \|g\| = 1\}} \int_{-1}^{1} e^x g(x)\, dx$ where $\mathscr{M} = \overline{\text{sp}}\{1, x, x^2\}$.

2.10. If $X_t = Z_t - \theta Z_{t-1}$, where $|\theta| < 1$ and $\{Z_t, t = 0, \pm 1, \ldots\}$ is a sequence of uncorrelated random variables, each with mean 0 and variance σ^2, show by checking the prediction equations that the best mean square predictor of X_{n+1} in $\overline{\text{sp}}\{X_j, -\infty < j \le n\}$ is

$$\hat{X}_{n+1} = -\sum_{j=1}^{\infty} \theta^j X_{n+1-j}.$$

What is the mean squared error of $\hat{X}_{n+1}$?

2.11. If X_t is defined as in Problem 2.10 with $\theta = 1$, find the best mean square predictor of X_{n+1} in $\overline{\text{sp}}\{X_j, 1 \le j \le n\}$ and the corresponding mean squared error.

2.12. If $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$, $t = 0, \pm 1, \ldots$ where $\{Z_t\}$ is a sequence of uncorrelated random variables, each with mean zero and variance σ^2 and such that Z_t is uncorrelated with $\{X_j, j < t\}$ for each t, use the prediction equations to show that the best mean square predictor of X_{n+1} in $\overline{\text{sp}}\{X_j, -\infty < j \le n\}$ is

$$\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1} + \cdots + \phi_p X_{n+1-p}.$$

2.13. (Gram–Schmidt orthogonalization). Let $x_1, x_2, \ldots, x_n$ be linearly independent elements of a Hilbert space $\mathscr{H}$ (i.e. elements for which $\|\alpha_1 x_1 + \cdots + \alpha_n x_n\| = 0$ implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$). Define

$$w_1 = x_1$$

and

$$w_k = x_k - P_{\overline{\text{sp}}\{w_1, \ldots, w_{k-1}\}} x_k, \quad k > 1.$$

Show that $\{e_k = w_k/\|w_k\|, k = 1, \ldots, n\}$ is an orthonormal set and that $\overline{\text{sp}}\{e_1, \ldots, e_k\} = \overline{\text{sp}}\{x_1, \ldots, x_k\}$ for $1 \le k \le n$.

2.14. Show that every closed subspace $\mathscr{M}$ of $\mathbb{R}^n$ which contains a non-zero vector can be written as $\mathscr{M} = \overline{\text{sp}}\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ where $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ is an orthonormal subset of $\mathscr{M}$ and $m\ (\le n)$ is the same for all such representations.

2.15. Let X_1, X_2 and X_3 be three random variables with mean zero and covariance matrix,

$$V = \begin{bmatrix} 14 & -1 & 3 \\ -1 & 5 & -1 \\ 3 & -1 & 1 \end{bmatrix}.$$

Use the Gram–Schmidt orthogonalization process of Problem 2.13 to find three uncorrelated random variables Z_1, Z_2 and Z_3 such that $\overline{\text{sp}}\{X_1\} = \overline{\text{sp}}\{Z_1\}$, $\overline{\text{sp}}\{X_1, X_2\} = \overline{\text{sp}}\{Z_1, Z_2\}$ and $\overline{\text{sp}}\{X_1, X_2, X_3\} = \overline{\text{sp}}\{Z_1, Z_2, Z_3\}$.

2.16. (Hermite polynomials). Let $\mathscr{H} = L^2(\mathbb{R}, \mathscr{B}, \mu)$ where $d\mu = (2\pi)^{-1/2}e^{-x^2/2}\,dx$. Set $f_0(x) \equiv 1$, $f_1(x) = x$, $f_2(x) = x^2$, $f_3(x) = x^3$. Using the Gram–Schmidt orthogonalization process, find polynomials $H_k(x)$ of degree k, $k = 0, 1, 2, 3$ which are orthogonal in $\mathscr{H}$. (Do not however normalize $H_k(x)$ to have unit length.) Verify that $H_k(x) = (-1)^k e^{x^2/2} \dfrac{d^k}{dx^k} e^{-x^2/2}$, $k = 0, 1, 2, 3$.

2.17. Prove the first statement in the proof of Theorem 2.4.2.

2.18. (a) Let x be an element of the Hilbert space $\mathscr{H} = \overline{\text{sp}}\{x_1, x_2, \ldots\}$. Show that $\mathscr{H}$ is separable and that

$$P_{\overline{\text{sp}}\{x_1, \ldots, x_n\}} x \to x \quad \text{as } n \to \infty.$$

(b) If $\{X_t, t = 0, \pm 1, \ldots\}$ is a stationary process show that

$$P_{\overline{\text{sp}}\{X_j, -\infty < j \le n\}} X_{n+1} = \lim_{r \to \infty} P_{\overline{\text{sp}}\{X_j, n-r < j \le n\}} X_{n+1}.$$

2.19. (General linear model). Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the vector of observations, X is a known $n \times m$ matrix of rank $m < n$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ is an m-vector of parameter values, and $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ is the vector of noise variables. The least squares estimator of $\boldsymbol{\theta}$ is given by equation (2.6.4), i.e.

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1}X'\mathbf{Y}.$$

Assume that $\mathbf{Z} \sim N(\mathbf{0}, \sigma^2 I_n)$ where I_n is the n-dimensional identity matrix.

(a) Show that $\mathbf{Y} \sim N(X\boldsymbol{\theta}, \sigma^2 I_n)$.

(b) Show that $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2(X'X)^{-1})$.

(c) Show that the projection matrix $P_{\mathscr{M}} = X(X'X)^{-1}X'$ is non-negative definite and has m non-zero eigenvalues all of which are equal to one. Similarly, $I_n - P_{\mathscr{M}}$ is also non-negative definite with $(n - m)$ non-zero eigenvalues all of which are equal to one.

(d) Show that the two vectors of random variables, $P_{\mathscr{M}}(\mathbf{Y} - X\boldsymbol{\theta})$ and $(I_n - P_{\mathscr{M}})\mathbf{Y}$ are independent and that $\sigma^{-2}\|P_{\mathscr{M}}(\mathbf{Y} - X\boldsymbol{\theta})\|^2$ and $\sigma^{-2}\|(I_n - P_{\mathscr{M}})\mathbf{Y}\|^2$ are independent chi-squared random variables with m and $(n - m)$ degrees of freedom respectively. ($\|\mathbf{Y}\|$ here denotes the Euclidean norm of $\mathbf{Y}$, i.e. $(\sum_{i=1}^n Y_i^2)^{1/2}$.)

(e) Conclude that

$$\frac{(n-m)\|P_{\mathcal{M}}(Y - X\boldsymbol{\theta})\|^2}{m\|\mathbf{Y} - P_{\mathcal{M}}\mathbf{Y}\|^2}$$

has the F distribution with m and $(n-m)$ degrees of freedom.

2.20. Suppose $(X, Z_1, \ldots, Z_n)'$ has a multivariate normal distribution. Show that

$$P_{\overline{\text{sp}}\{1, Z_1, \ldots, Z_n\}}(X) = E_{\mathcal{M}(Z_1, \ldots, Z_n)}(X),$$

where the conditional expectation operator $E_{\mathcal{M}(Z_1, \ldots, Z_n)}$ is defined as in Section 2.7.

2.21. Suppose $\{X_t, t = 0, \pm 1, \ldots\}$ is a stationary process with mean zero and autocovariance function $\gamma(\cdot)$ which is absolutely summable (i.e. $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$). Define f to be the function,

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-ih\lambda}, \qquad -\pi \le \lambda \le \pi,$$

and show that $\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda$.

2.22. (a) If $f \in L^2([-\pi, \pi])$, prove the Riemann–Lebesgue lemma: $\langle f, e_h \rangle \to 0$ as $h \to \infty$, where e_n was defined by (2.8.2).

(b) If $f \in L^2([-\pi, \pi])$ has a continuous derivative $f'(x)$ and $f(\pi) = f(-\pi)$, show that $\langle f, e_h \rangle = (ih)^{-1} \langle f', e_h \rangle$ and hence that $h\langle f, e_h \rangle \to 0$ as $h \to \infty$. Show also that $\sum_{h=-\infty}^{\infty} |\langle f, e_h \rangle| < \infty$ and conclude that $S_n f$ (see Section 2.8) converges uniformly to f.

2.23. Show that the space l^2 (Example 2.9.1) is a separable Hilbert space.

2.24. If $\mathcal{H}$ is any Hilbert space with orthonormal basis $\{e_n, n = 1, 2, \ldots\}$, show that the mapping defined by $Th = \{\langle h, e_n \rangle\}$, $h \in \mathcal{H}$, is an isomorphism of $\mathcal{H}$ onto l^2.

2.25.* Prove that $\mathcal{M}(Z)$ (see Definition 2.7.3) is closed.

http://www.springer.com/978-1-4419-0319-8

Time Series: Theory and Methods
Brockwell, P.J.; Davis, R.A.
1991, XVI, 580 p., Softcover
ISBN: 978-1-4419-0319-8