Chapter 3

FORECASTING

3.1 Projections

In macroeconomics, forecasting is important in many ways. For structural macroeconomic models, we usually need to specify the forecasting rules that economic agents are using and the information set used by them to forecast future economic variables. Taking the conditional expectation is one way to model forecasting. This method generally requires nonlinear forecasting rules which are difficult to estimate. For the purpose of testing the models and parameter estimation, it is sometimes possible for an econometrician to use a simpler forecasting rule and a smaller information set.

In this section, we study projections as a forecasting method. Projections are used to explain the Wold representation, which forms a basis for studying linear and nonlinear stochastic processes.

3.1.1 Definitions and Properties of Projections

In this chapter, we consider random variables with finite second moments unless otherwise noted. We consider the problem of forecasting y, using a set H of random variables. Typically, y is a future random variable such as the growth rate of the Gross Domestic Product (GDP) or the growth rate of a stock price, and H contains current and past economic variables that are observed by economic agents and/or econometricians. Let us denote a forecast of y based on H by y^f, so that the forecasting error is $y - y^f$. In most economic applications, we choose the forecast, y^f, so that y^f minimizes

$$E[(y - y^f)^2]. \tag{3.1}$$

In other words, y^f is in H, and for all h in H,

$$E[(y - y^f)^2] \leq E[(y - h)^2]. \tag{3.2}$$

The expression (3.1) is called the *mean squared error* associated with the forecast, y^f.

When two random variables h_1 and h_2 satisfy

$$E(h_1 h_2) = 0, \tag{3.3}$$

they are said to be *orthogonal* to each other. When either h_1 or h_2 has mean zero, orthogonality means that they are uncorrelated. The concept of orthogonality is closely related to the problem of minimizing the mean squared error. Under certain conditions on H, the Classical Projection Theorem (see, e.g., Luenberger, 1969) states that there exists a unique random variable y^f in H that minimizes the mean squared error, and that y^f is the minimizer if and only if the forecasting error is orthogonal to all members of H:

$$E((y - y^f)h) = 0 \tag{3.4}$$

for all h in H; this is called the *orthogonality condition.* When such a forecast exists, we call the forecast, y^f, a *projection* of y onto H, and denote it by $\hat{E}(y|\mathrm{H})$. When $\mathbf{Y}$ is a random vector with finite second moments, we apply the projection to each element of $\mathbf{Y}$ and write $\hat{E}(\mathbf{Y}|\mathrm{H})$.

Some properties of projections are very important:

Proposition 3.1 (Properties of Projections)

(a) Projections are linear: $\hat{E}(aX + bY|\mathrm{H}) = a\hat{E}(X|\mathrm{H}) + b\hat{E}(Y|\mathrm{H})$ for any random variables, X and Y, with finite variance and constants, a and b.

(b) If a random variable Z is in the information set H, then

$$\hat{E}(ZY|\mathrm{H}) = Z\hat{E}(Y|\mathrm{H}).$$

(c) *The Law of Iterated Projections:* If the information set H is smaller than the information set G ($\mathrm{H} \subset G$), then

$$\hat{E}(Y|\mathrm{H}) = \hat{E}[\hat{E}(Y|G)|\mathrm{H}].$$

∎

3.1.2 Linear Projections and Conditional Expectations

The meaning of projection depends on how the information set H used for the projection is constructed. Let $\mathbf{X}$ be a $p \times 1$ vector of random variables with finite second moments. Let $\mathrm{H} = \{h$ is a random variable such that $h = \mathbf{X}'\mathbf{b}$ for some p-dimensional vector of real numbers $\mathbf{b}\}$. Since $\hat{E}(y|\mathrm{H})$ is also a member of H, there exists $\mathbf{b}_0$ such that

$$\hat{E}(y|\mathrm{H}) = \mathbf{X}'\mathbf{b}_0. \tag{3.5}$$

In this sense, $\hat{E}(y|\mathrm{H})$ uses a linear forecasting rule. When we use an information set such as H, which only allows for linear forecasting rules, the projection based on such an information set is called a *linear projection*. We write $\hat{E}(y|\mathrm{H}) = \hat{E}(y|\mathbf{X})$.

Let $\mathrm{H}^N = \{h$ is a random variable with a finite variance such that $h = f(\mathbf{X})$ for a function $f\}$.[1] In this case, there exists a function $f_0(\cdot)$ such that

$$\hat{E}(y|\mathrm{H}^N) = f_0(\mathbf{X}). \tag{3.6}$$

In this sense, $\hat{E}(y|\mathrm{H}^N)$ allows for a nonlinear forecasting rule. It can be shown that

$$\hat{E}(y|\mathrm{H}^N) = E(y|\mathbf{X}). \tag{3.7}$$

Hence the projection and conditional expectation coincide when we allow for nonlinear forecasting rules. For this reason, the projections we use in this book are linear projections unless otherwise noted.

An important special case is when y and $\mathbf{X}$ are jointly normally distributed. In this case, the expectation of y conditional on $\mathbf{X}$ is a linear function of $\mathbf{X}$. Hence the linear projection of y onto the information set generated by $\mathbf{X}$ is equal to the expectation of y conditional on $\mathbf{X}$.

When it is necessary to distinguish the information set I generated by $\mathbf{X}$ for conditional expectations introduced in Chapter 2 and the information set H generated by $\mathbf{X}$ for linear projections, H will be called the *linear information set* generated by $\mathbf{X}$. (????? Unclear! from Billy)

Masao needs to check this!

Linear projections are important because it is easy to estimate them in many applications. Note that the orthogonality condition states that

$$E[(y - \mathbf{X}'\mathbf{b}_0)h] = 0 \tag{3.8}$$

for any h in H. Since each element of $\mathbf{X}$ is in H, using the i-th element $\mathbf{X}_i$ for h, we obtain

$$E[(y - \mathbf{X}'\mathbf{b}_0)\mathbf{X}_i] = 0 \tag{3.9}$$

[1] As in Proposition 2.2, we require that the function f is measurable.

for $i = 1, 2, \cdots, p$, or

$$E[\mathbf{X}(y - \mathbf{X}'\mathbf{b}_0)] = 0. \tag{3.10}$$

Therefore

$$E(\mathbf{X}y) = E(\mathbf{X}\mathbf{X}')\mathbf{b}_0. \tag{3.11}$$

Assuming that $E(\mathbf{X}\mathbf{X}')$ is nonsingular, we obtain

$$\mathbf{b}_0 = E(\mathbf{X}\mathbf{X}')^{-1}E(\mathbf{X}y) \tag{3.12}$$

and

$$\hat{E}(y|\mathrm{H}) = \mathbf{X}'\mathbf{b}_0, \tag{3.13}$$

where H is the linear information set generated by $\mathbf{X}$. As we will discuss, if $\mathbf{X}$ and y are strictly stationary, Ordinary Least Squares (OLS) can be used to estimate $\mathbf{b}_0$.

Following examples show differences between conditional expectation and linear projection.

Youngsoo needs to check this!

Example 3.1 Let X and Y be random variables with non-zero mean. The linear projection of Y on X is

$$\hat{E}(Y|1, X) = a + bX. \tag{3.14}$$

Then, from (3.12) and $E(Y) = a + bE(X)$ we have

$$b = \frac{E(XY)}{E(X^2)} = \frac{Cov(X,Y)}{Var(X)} \tag{3.15}$$

$$a = E(Y) - bE(X). \tag{3.16}$$

∎

Note that the linear projection is a population regression; that is, a and b are defined by population moments. corresponding sample moments can be used to estimate $\hat{a}$ and $\hat{b}$.

Example 3.2 Let X be a standard Normal random variable, and $Y = X^2$. Note that Y is $\chi(1)$ random variable and $E(Y) = 1$. The linear projection of Y on X is

$$\hat{E}(Y|1, X) = a + bX = 1. \tag{3.17}$$

This is becasue from Example 3.1, we have

$$b = \frac{E(XY)}{E(X^2)} = \frac{E(X^3)}{Var(X)} = 0 \tag{3.18}$$

and

$$a = E(Y) = E(X^2) = 1. \tag{3.19}$$

Note that $E(X^3) = 0$ because the distribution of X is symmetric. Whereas the conditional expectation of Y on X is[2]

$$E(Y|X) = X^2. \tag{3.20}$$

∎

Example 3.3 Let X_0 be a standard Normal random variable, and ε_1 be a Normal random variable with mean 0 and variance σ^2. Assume that X_0 and ε are independent each other. Define $X_1 = a + bX_0 + cX_0^2 + \varepsilon_1$. Then, the unconditional expectation of X_1, the linear projection and the conditional expectation of X_1 on X_0 are, respectively,

$$E(X_1) = E(a + bX_0 + cX_0^2 + \varepsilon_1) = a + c, \tag{3.21}$$

[2] Note that since $X^0 = 1$, 1 is always in the information set for conditional expectation. However, 1 may not be in the linear information set.

$$\hat{E}(X_1|1, X_0) = \hat{E}(a + bX_0 + cX_0^2 + \varepsilon_1|1, X_0) \tag{3.22}$$
$$= a + bX_0 + c.$$

Note that $\hat{E}(X_0^2|1, X_0) = 1$ by (3.17).

$$E(X_1|X_0) = a + bX_0 + cX_0^2. \tag{3.23}$$

∎

3.2 Some Applications of Conditional Expectations and Projections

This section presents some applications of conditional expectations and projections in order to illustrate their use in macroeconomics. More explanations of some of these applications and presentations of other applications will be given in later chapters. In this chapter, all random variables are assumed to have finite second moments.

3.2.1 Volatility Tests

Many rational expectations models imply

$$X_t = E(Y_t|\mathrm{I}_t) \tag{3.24}$$

for economic variables X_t and Y_t. Here X_t is in the information set I_t which is available to the economic agents at date t while Y_t is not. A testable implication of (3.24) can be obtained by comparing the volatility of X_t with that of Y_t. Relation (3.24) implies

$$Y_t = X_t + \epsilon_t \tag{3.25}$$

where $\epsilon_t = Y_t - E(Y_t|\mathrm{I}_t)$ is the forecast error. Since $E(\epsilon_t|\mathrm{I}_t) = 0$,

$$E(\epsilon_t h_t) = 0 \tag{3.26}$$

for any random variable h_t that is in I_t. We can interpret (3.26) as an orthogonality condition. The forecast error must be uncorrelated with any variable in the information set. Since X_t is in I_t, (3.26) implies $E(\epsilon_t X_t) = 0$. Therefore, from (3.25) we obtain

$$E(Y_t^2) = E(X_t^2) + E(\epsilon_t^2). \tag{3.27}$$

Since (3.24) implies that $E(X_t) = E(Y_t)$, (3.27) implies

$$Var(Y_t) = Var(X_t) + E(\epsilon_t^2). \tag{3.28}$$

Since $E(\epsilon_t^2) \geq 0$, we conclude

$$Var(Y_t) \geq Var(X_t). \tag{3.29}$$

Thus, if X_t forecasts Y_t, X_t must be less volatile than Y_t. Various volatility tests have been developed to test this implication of (3.24).

LeRoy and Porter (1981) and Shiller (1981) started to apply volatility tests to the present value model of stock prices. Let p_t be the real stock price (after the dividend is paid) in period t and d_t be the real dividend paid to the owner of the stock at the beginning of period t. Then the no-arbitrage condition is

$$p_t = E[b(p_{t+1} + d_{t+1})|\mathrm{I}_t], \tag{3.30}$$

where b is the constant real discount rate, and I_t is the information set available to economic agents in period t. Solving (3.30) forward and imposing the no bubble condition,[3] we obtain the present value formula:

$$p_t = E(\sum_{i=1}^{\infty} b^i d_{t+i}|\mathrm{I}_t). \tag{3.31}$$

[3] It rules out the exploding solution of the difference equation

Applying the volatility test, we conclude that the variance of $\sum_{i=1}^{\infty} b^i d_{t+i}$ is greater than or equal to the variance of p_t. One way to test this is to directly estimate these variances and compare them. However, $\sum_{i=1}^{\infty} b^i d_{t+i}$ involves infinitely many data points for the dividend. When we have data for the stock price and dividend for $t = 1, \cdots, T$, we use (3.31) to obtain

$$p_t = E(\sum_{i=1}^{T-t} b^i d_{t+i} + b^{T-t} p_T | \mathrm{I}_t). \tag{3.32}$$

Let $Y_t = \sum_{i=1}^{T-t} b^i d_{t+i} + b^{T-t} p_T$. Then we have data on Y_t from $t = 1$ to $t = T$ when we choose a reasonable number for the discount rate b. We can estimate the variance of p_t and the variance of Y_t, and compare them to form a test statistic.[4]

3.2.2 Parameterizing Expectations

As discussed in Section 3.1, conditional expectations allow for nonlinear forecasting rules. For example, consider $E(Y|\mathrm{I})$ for a random variable Y and an information set I generated from a random variable X. Then $E(Y|\mathrm{I})$ can be written as a function of $X : E(Y|\mathrm{I}) = f(X)$. The function $f(\cdot)$ can be nonlinear here. In most applications involving nonlinear forecasting rules, the functional form of $f(\cdot)$ is not known. In order to simulate rational expectations models, it is often necessary to have a method to estimate $f(\cdot)$.

Marcet's (1989) parameterizing expectations method (also see den Haan and Marcet, 1990) is based on the fact that the conditional expectation is a projection, and thus minimizes the mean square error. We take a class of functions that approximate any function. For example, take a class of polynomial functions and let $f_N(X) = a_0 + a_1 X + a_2 X^2 + \cdots + a_N X^N$. We choose $a_0, \cdots, a_N$ to minimize the mean square

[4] There are some problems with this procedure. One problem is nonstationarity of p_t and Y_t. For more detailed explanation of volatility tests, see Campbell, Lo, and MacKinlay (1997).

error, $E[(Y - f_N(X))^2]$. Intuitively, $f_N(\cdot)$ should approximate $f(X)$ for a large enough N. This method is used to simulate economic models with rational expectations.

3.2.3 Noise Ratio

In econometrics, we often test an economic model with test statistics whose probability distributions are known under the null hypothesis that the model is true. Hansen's J test, which will be discussed in Chapter 9, is an example. Given that all economic models are meant to be approximations, however, it seems desirable to measure how good a model is in approximating reality. Durlauf and Hall (1990) and Durlauf and Maccini (1995) propose such a measure called the noise ratio.[5]

Consider an economic model which states

$$E(g(\mathbf{Y})|\mathrm{I}) = 0 \tag{3.33}$$

for an information set I and a function $g(\cdot)$ of a random vector $\mathbf{Y}$. For example, let S be the spot exchange rate of a currency in the next period, F be the forward exchange rate observed today for the currency to be delivered in the next period, $g(S, F) = S - F$, and I be the information set available to the economic agents today. Then under the assumption of risk neutral investors, we obtain (3.33).

Let $\nu = g(\mathbf{Y}) - E(g(\mathbf{Y})|\mathrm{I})$. If the model is true, then $g(\mathbf{Y}) = \nu$. Since this model is an approximation, however, $g(\mathbf{Y})$ deviates from ν. Let N be the deviation: $N = g(\mathbf{Y}) - \nu$, which is called the model noise. A natural measure of how well the model approximates reality is $Var(N)$. Durlauf and Hall (1990) propose a method to estimate a lower bound of $Var(N)$ using $\eta = Var(\hat{E}(g(\mathbf{Y})|\mathrm{H}))$, where H is an information set generated from some variables in I.[6]

[5] See Konuki (1999) for an application of the noise ratio to foreign exchange rate models.

[6] For example, in the forward exchange rate model mentioned above, some lagged values of $S - F$

Using the law of iterated projections[7], we have $\hat{E}(\nu|\mathrm{H}) = 0$. Thus, $\hat{E}(g(\mathbf{Y})|\mathrm{H}) = \hat{E}(N|\mathrm{H})$, and therefore $\eta = Var(\hat{E}(N|\mathrm{H}))$. Because $N = \hat{E}(N|\mathrm{H}) + (N - \hat{E}(N|\mathrm{H}))$, and the forecast error, $N-\hat{E}(N|\mathrm{H})$, is orthogonal to $\hat{E}(N|\mathrm{H})$, $E(N^2) = E[(\hat{E}(N|\mathrm{H}))^2] + E[(N-\hat{E}(N|\mathrm{H}))^2]$. Since $E[(N-\hat{E}(N|\mathrm{H}))^2] \geq 0$, $E(N^2) \geq E[(\hat{E}(N|\mathrm{H}))^2]$. Therefore, $Var(N) = E(N^2) - (E(N))^2 \geq E[(\hat{E}(N|\mathrm{H}))^2] - \{E[\hat{E}(N|\mathrm{H})]\}^2 = \eta$.[8] Thus η is a lower bound of $Var(N)$.

In a sense, η is a sharp lower bound. Since we do not know much about the model noise, N, it may or may not be in H. If N happens to be in H, then $\hat{E}(N|\mathrm{H}) = N$. Therefore, in this case $Var(N) = \eta$.

The noise ratio, NR, is defined by $NR = \frac{\eta}{Var(g(\mathbf{Y}))}$. Since $\hat{E}(g(\mathbf{Y})|\mathrm{H})$ is orthogonal to $g(\mathbf{Y}) - \hat{E}(g(\mathbf{Y})|\mathrm{H})$,

$$Var(g(\mathbf{Y})) = \eta + Var(g(\mathbf{Y}) - \hat{E}(g(\mathbf{Y})|\mathrm{H})). \tag{3.34}$$

Therefore, the $0 \leq NR \leq 1$.

Appendix

3.A Introduction to Hilbert Space

This Appendix explains Hilbert space techniques used in this book.[9] Projections explained in this chapter are defined in a Hilbert space. In Appendix B, we will consider another Hilbert space, which provides the foundation for the lag operator methods and the frequency domain analysis which are useful in macroeconomics and time series econometrics.

---

and a constant can be used to generate a linear information set H.

[7] We assume that the second moment exists and is finite. Therefore, the conditional expectation is a projection.

[8] Here, we assumed that the constants are included in H, so that $E(S) = E[\hat{E}(S|\mathrm{H})]$.

[9] All proofs of the results can be found in Luenberger (1969) or Hansen and Sargent (1991).

A pre-Hilbert space is a vector space on which an inner product is defined. The inner product is used to define a distance. If all Cauchy sequences of a pre-Hilbert space converge, then it is said to be complete. A Hilbert space is a complete pre-Hilbert space. One reason why a Hilbert space is useful is that the notion of orthogonality can be defined with the inner product. Since a Hilbert space is complete, we can prove that the limit of a sequence exists once we prove that the sequence is Cauchy. For example, this technique can be used to prove that a projection can be defined.

Section 3.A.1 reviews definitions regarding vector spaces. Section 3.A.2 gives an introduction to Hilbert space.

3.A.1 Vector Spaces

Given a set of scalars K (either the real line, R, or the complex plane, C)[10], a *vector space* (or a *linear space*) X on K is a set of elements, called vectors, together with two operations (addition and scalar multiplication) which satisfy the following conditions: For any $\mathbf{x}, \mathbf{y}, \mathbf{z}$ in X and for any α, β in K, we require

(3.A.1) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ (commutative law)

(3.A.2) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ (associative law)

(3.A.3) There is a null vector $\mathbf{0}$ in X such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x}$ in X.

(3.A.4) $\left.\begin{array}{l} \alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y} \\ (\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x} \end{array}\right\}$ (distributive laws)

(3.A.5) $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x})$ (associative law)

(3.A.6) $0\mathbf{x} = \mathbf{0}, 1\mathbf{x} = \mathbf{x}$.

[10] In general, an additive group X for which scalar multiplication satisfies (3.A.4)-(3.A.6) for any field K is a vector space on K. In this book K is either the real line or the complex plane.

Using $\alpha = -1$, we define $\mathbf{x} - \mathbf{y} = \mathbf{x} + (-1)\mathbf{y}$. In this Appendix, we give examples of vector spaces on R, but state results that are applicable when $K = C$. Examples of vector spaces on C are given in Appendix B.

A nonempty subset H of a vector space X is called a *(linear) subspace* of X if every vector of the form $\alpha\mathbf{x} + \beta\mathbf{y}$ is in H whenever $\mathbf{x}$ and $\mathbf{y}$ are both in H and α and β are in K. A subspace always contains the null vector $\mathbf{0}$, and satisfies conditions (3.A.1)-(3.A.6). Hence a subspace is itself a vector space.

If a subset H of X is not a subspace, it is often convenient to construct the smallest subspace containing H. For this purpose, we use linear combinations of vectors in H. A *linear combination* of the vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ is a sum of the form $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_n\mathbf{x}_n$ where α_i is a scalar $(i = 1, \cdots, n)$. The set consisting of all vectors in X which are linear combinations of vectors in H is called the *(linear) subspace generated* by H.

A *normed vector space* is a vector space X on which a norm is defined. The norm is a real-valued function that maps each element of $\mathbf{x}$ in X into a real number $\|\mathbf{x}\|$, which satisfies

(3.A.7) $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x}$ in X and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

(3.A.8) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (The triangle inequality)

(3.A.9) $\|\alpha\mathbf{x}\| = |\alpha| \; \|\mathbf{x}\|$ for all α in K and $\mathbf{x}$ in X.

A norm can be used to define a metric d on X by $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

A sequence $\{\mathbf{x}_n\}_{n=1}^{\infty}$ in a normed vector space *converges* to $\mathbf{x}_0$ if the sequence $\{\|\mathbf{x}_n - \mathbf{x}_0\|\}_{n=1}^{\infty}$ of real numbers converges to zero, which is denoted by $\mathbf{x}_n \to \mathbf{x}_0$ or $\lim \mathbf{x}_n = \mathbf{x}_0$. A sequence $\{\mathbf{x}_n\}_{n=1}^{\infty}$ in a normed vector space is a *Cauchy sequence* if for any $\epsilon > 0$, there exists an integer N such that $\|\mathbf{x}_n - \mathbf{x}_m\| < \epsilon$ for all $n, m > N$. In a normed vector space, every convergent sequence is a Cauchy sequence. A space in which every Cauchy sequence has a limit is said to be *complete*. A complete normed vector space is called a *Banach space*.

Example 3.A.1 The real line, R, is a vector space on $K = \mathrm{R}$ with addition and scalar multiplication defined in the usual way. When the norm of a real number is defined as its absolute value, R is a Banach space. ∎

Example 3.A.2 Vectors in the space consist of sequences of n real numbers, R^n, which is a vector space on R when $\mathbf{x}+\mathbf{y}$ for $\mathbf{x} = (x_1, x_2, \cdots, x_n)'$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ is defined by $(x_1+y_1, x_2+y_2, \cdots, x_n+y_n)'$ and $\alpha\mathbf{x}$ for α in R is defined by $(\alpha x_1, \alpha x_2, \cdots, \alpha x_n)'$. When we define a norm of $\mathbf{x}$ as $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$, R^n is a Banach space. ∎

3.A.2 Hilbert Space

A *pre-Hilbert space* is a vector space X on K for which an inner product is defined. The inner product is a scalar-valued function that maps each element of $(\mathbf{x}, \mathbf{y})$ in $X \times X$ into an element $(\mathbf{x}|\mathbf{y})$ in K, which satisfies

$$(\mathbf{x}|\mathbf{y}) = \overline{(\mathbf{y}|\mathbf{x})} \tag{3.A.10}$$

$$(\mathbf{x}+\mathbf{z}\ |\mathbf{y}) = (\mathbf{x}|\mathbf{y}) + (\mathbf{z}|\mathbf{y}) \tag{3.A.11}$$

$$(\alpha\mathbf{x}|\mathbf{y}) = \alpha(\mathbf{x}|\mathbf{y}) \tag{3.A.12}$$

$$(\mathbf{x}|\mathbf{x}) \geq 0 \text{ and } (\mathbf{x}|\mathbf{x}) = 0 \text{ if and only if } \mathbf{x} = \mathbf{0}. \tag{3.A.13}$$

for any $\mathbf{x}, \mathbf{y}, \mathbf{z}$ in X and α in K. The bar on the right side on (3.A.10) denotes complex conjugation, which can be ignored if K is R. By (3.A.10), $(\mathbf{x}|\mathbf{x})$ is real for each $\mathbf{x}$ even when K is C.

A norm can be defined from an inner product by $\|\mathbf{x}\| = \sqrt{(\mathbf{x}|\mathbf{x})}$. Thus a pre-Hilbert space is a normed vector space. A complete pre-Hilbert space is called a *Hilbert space.*

Example 3.A.3 When we define $(\mathbf{x}|\mathbf{y}) = \sum_{i=1}^{n} x_i y_i$, R^n is a Hilbert space on R. ▮

The following Hilbert space of random variables with finite second moments is the one we used in Chapter 3.

Example 3.A.4 Let $(\mathrm{S}, \mathcal{F}, Prob)$ be a probability space. Let $L^2(Prob) = \{h : h$ is a (real-valued) random variable and $E(|h|^2) < \infty\}$. Then with an inner product defined by $(h_1|h_2) = E(h_1 h_2)$, $L^2(Prob)$ is a Hilbert space on R. If two different random variables h_1 and h_2 satisfy $E[(h_1 - h_2)^2] = 0$, then h_1 and h_2 are the same element in this space. If $E[(h_1 - h_2)^2] = 0$, then $h_1 = h_2$ with probability one. Hence this definition does not cause problems for most purposes. In this space, the distance is defined by the mean square, so the convergence in this space is the convergence in mean square. ▮

One reason why an inner product is useful is that we can define the notion of orthogonality. In a Hilbert space, two vectors $\mathbf{x}$ and $\mathbf{y}$ are said to be *orthogonal* if $(\mathbf{x}|\mathbf{y}) = 0$. A vector $\mathbf{x}$ is said to be orthogonal to a set H if $\mathbf{x}$ is orthogonal to each element h in H. Some useful results concerning the inner product are:[11]

Proposition 3.A.1 *(The Cauchy-Schwarz Inequality)* For all $\mathbf{x}$, $\mathbf{y}$ in a *Hilbert space*, $|(\mathbf{x}|\mathbf{y})| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|$. Equality holds if and only if $\mathbf{x} = \lambda\mathbf{y}$ for some λ in K, or $\mathbf{y} = \mathbf{0}$. ▮

[11] These three propositions hold for a pre-Hilbert space. See Luenberger (1969, p.47 and p.49).

Proposition 3.A.2 *(Continuity of the Inner Product)* Suppose that $\mathbf{x}_n \to \mathbf{x}$ and $\mathbf{y}_n \to \mathbf{y}$ in a Hilbert space. Then $(\mathbf{x}_n|\mathbf{y}_n) \to (\mathbf{x}|\mathbf{y})$. ∎

Proposition 3.A.3 If $\mathbf{x}$ is orthogonal to $\mathbf{y}$ in a Hilbert space, then $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$. ∎

Example 3.A.5 In $L^2(Prob)$, the Cauchy-Schwarz Inequality becomes $|E(xy)| \leq \sqrt{E(x^2)}\sqrt{E(y^2)}$ for any random variables with finite second moments. Proposition 3.A.3 states that if x and y satisfy $E(xy) = 0$, then $E[(x+y)^2] = E(x^2) + E(y^2)$. ∎

Projections can be defined on a Hilbert space due to the following result:

Proposition 3.A.4 *(The Classical Projection Theorem)* Let X be a Hilbert space and H be a closed linear subspace of X. Corresponding to any vector $\mathbf{x}$ in X, there is a unique vector $\mathbf{h}_0$ in H such that $\|\mathbf{x}-\mathbf{h}_0\| \leq \|\mathbf{x}-\mathbf{h}\|$. Furthermore, a necessary and sufficient condition that $\mathbf{h}_0$ in H be the unique minimizing vector is that $\mathbf{x}-\mathbf{h}_0$ be orthogonal to H. ∎

Given a closed linear space H, we define a function $\hat{E}(\cdot|\mathrm{H})$ on X by $\hat{E}(\mathbf{x}|\mathrm{H}) = \mathbf{h}_0$ where $\mathbf{h}_0$ is an element in H such that $\mathbf{x}-\mathbf{h}_0$ is orthogonal to H. $\hat{E}(\mathbf{x}|\mathrm{H})$ is the projection of $\mathbf{x}$ onto H. The projection defined in Section 3.1 in $L^2(Prob)$ is one example.

If a sequence $\{\mathbf{e}_t\}_{t=1}^{\infty}$ in a Hilbert space satisfies $\|\mathbf{e}_t\| = 1$ for all t and $(\mathbf{e}_t|\mathbf{e}_s) = 0$ for all $t \neq s$, then it is said to be an *orthonormal sequence*. We are concerned with an infinite series of the form $\sum_{t=1}^{\infty} \alpha_t \mathbf{e}_t$. An infinite series of the form $\sum_{t=1}^{\infty} \mathbf{x}_t$ is said to *converge* to the element $\mathbf{x}$ in a Hilbert space if the sequence of partial sums $s_T = \sum_{t=1}^{T} \mathbf{x}_t$ converges to $\mathbf{x}$. In that case we write $\mathbf{x} = \sum_{t=1}^{\infty} \mathbf{x}_t$. A necessary and sufficient condition for an infinite series of orthonormal sequence to converge in Hilbert space is known (see Luenberger, 1969, p.59):

Proposition 3.A.5 Let $\{\mathbf{e}_j\}_{j=1}^{\infty}$ be an orthonormal sequence in a Hilbert space X. A series of the form $\sum_{j=1}^{\infty} \alpha_j \mathbf{e}_j$ converges to an element $\mathbf{x}$ in X if and only if $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$, and in that case we have $\alpha_j = (\mathbf{x}|\mathbf{e}_j)$. ∎

Example 3.A.6 Applying the above proposition in $L^2(Prob)$, we obtain a necessary and sufficient condition for an MA(∞) representation $\sum_{j=0}^{\infty} b_j v_{t-j}$ to converge for a white noise process $\{v_{t-j}\}_{j=0}^{\infty}$ with $E(v_t^2) = \sigma_v^2 > 0$. Define $e_t = \frac{v_t}{\sigma_v}$, and $\alpha_j = b_j \sigma_v$, so that $\{e_{t-j}\}_{j=0}^{\infty}$ is orthonormal because $E(e_t^2) = 1$ and $E(e_t e_s) = 0$ for $t \neq s$. From the above proposition, $\sum_{j=1}^{\infty} b_j v_j = \sum_{j=1}^{\infty} \alpha_j e_j$ converges in $L^2(Prob)$, if and only if $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$. Since $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$ if and only if $\sum_{j=1}^{\infty} |b_j|^2 < \infty$, $\sum_{j=1}^{\infty} b_j v_j$ converges in mean square if and only if $\{b_j\}_{j=1}^{\infty}$ is square summable. ∎

Given an orthonormal sequence $\{\mathbf{e}_j\}_{j=1}^{\infty}$, we started from a square summable sequence $\{\alpha_j\}$ and constructed $\mathbf{x} = \sum_{j=1}^{\infty} \alpha_j \mathbf{e}_j$ in X in the above proposition. We now start with a given $\mathbf{x}$ in X and consider a series

$$\sum_{j=1}^{\infty} (\mathbf{x}|\mathbf{e}_j)\mathbf{e}_j. \tag{3.A.14}$$

The series is called the *Fourier series* of $\mathbf{x}$ relative to $\{\mathbf{e}_j\}_{j=1}^{\infty}$, and $(\mathbf{x}|\mathbf{e}_j)$ is called the *Fourier coefficient* of $\mathbf{x}$ with respect to $\mathbf{e}_j$.

In general, $\mathbf{x}$ is not equal to its Fourier series. Given a subset H of a Hilbert space, the *closed subspace generated by* H is the closure of the linear subspace generated by H. Let M be the closed subspace generated by $\{\mathbf{e}_j\}_{j=1}^{\infty}$. If $\mathbf{x}$ is in M, then $\mathbf{x}$ is equal to its Fourier series as implied by the next proposition:

Proposition 3.A.6 Let $\mathbf{x}$ be an element in a Hilbert space X and $\{\mathbf{e}_j\}_{j=1}^{\infty}$ be an orthonormal sequence in H. Then the Fourier series $\sum_{j=1}^{\infty}(\mathbf{x}|\mathbf{e}_j)\mathbf{e}_j$ converges to an element $\hat{\mathbf{x}}$ in the closed subspace M generated by $\{\mathbf{e}_j\}_{j=1}^{\infty}$. The difference vector $\mathbf{x}-\hat{\mathbf{x}}$ is orthogonal to M. ∎

This proposition shows that the Fourier series of $\mathbf{x}$ is the projection of $\mathbf{x}$ onto M: $\hat{E}(\mathbf{x}|M) = \sum_{j=1}^{\infty}(\mathbf{x}|\mathbf{e}_j)\mathbf{e}_j$.[12]

Exercises

3.1 Let S_t be a spot exchange rate at time t and F_t be a forward exchange rate observed at time t for delivery of one unit of a currency at $t + 1$. Assume that $F_t = E(S_{t+1}|\mathrm{I}_t)$ where I_t is the information set available for the economic agents at t. Prove that $Var(F_t) \leq Var(S_{t+1})$.

3.2 Let $i_{n,t}$ be the n year interest rate observed at time t. The expectations hypothesis of the term structure of interest rates states that $i_{n,t} = E(A_t|\mathrm{I}_t)$ where

$$A_t = \frac{1}{n}\sum_{\tau=0}^{n-1} i_{1,t+\tau}, \tag{3.E.1}$$

where I_t is the information available at time t. Imagine that data on interest rates clearly indicate that $Var(i_{n,t}) \leq Var(A_t)$. Does the data support the expectations theory? Explain your answer.

3.3 Let p_t be the real stock price, d_t be the real dividend, and b be the constant ex ante discount rate. Assume that p_t and d_t are stationary with zero mean and finite

[12] See Luenberger (1969, p.60).

second moments. Let

$$p_t^e = \sum_{\tau=1}^{\infty} b^\tau E(d_{t+\tau}|\mathrm{I}_t), \tag{3.E.2}$$

where I_t is the information set available in period t that includes the present and past values of p_t and d_t. Let $\hat{E}(\cdot|\mathrm{H}_t)$ be the linear projection onto an information set H_t. Define the model noise N_t by

$$N_t = p_t - p_t^e. \tag{3.E.3}$$

Let $\eta = Var(\hat{E}(N_t|\mathrm{H}_t))$.

(**a**) Assume that H_t is generated by $\{d_t\}$. Show that $\eta \leq Var(N_t)$ for any noise N_t.

(**b**) Assume that H_t is generated by $\{d_t, d_{t-1}, d_{t-2}\}$. Show that $\eta \leq Var(N_t)$ for any noise N_t.

3.4 Derive (3.34) in the text.

References

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997): *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey.

den Haan, W. J., and A. Marcet (1990): "Solving the Stochastic Growth Model by Parameterizing Expectations," *Journal of Business and Economic Statistics*, 8, 31–34.

Durlauf, S. N., and R. E. Hall (1990): "Bounds on the Variances of Specification Errors in Models with Expectations," Manuscript.

Durlauf, S. N., and L. J. Maccini (1995): "Measuring Noise in Inventory Models," *Journal of Monetary Economics*, 36, 65–89.

Hansen, L. P., and T. J. Sargent (1991): *Rational Expectations Econometrics*. Westview, London.

Konuki, T. (1999): "Measuring Noise in Exchange Rate Models," *Journal of International Economics*, 48(2), 255–270.

LeRoy, S. F., and R. D. Porter (1981): "The Present-Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, 49(3), 555–574.

Luenberger, D. G. (1969): *Optimization by Vector Space Methods*. Wiley, New York.

Marcet, A. (1989): "Solving Non-Linear Stochastic Models by Parameterizing Expectations," Manuscript.

Shiller, R. J. (1981): "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, 71, 421–436.