– Lecture Notes 1

Review of fundamental concepts in time series analysis

Please let me know if you have any comments or if you find typos.

These notes contains a rough review of linear time series analysis. The techniques reviewed here will be useful for various topics discussing during the length of the course, inclusing the analysis of structural vector auoregressions (SVAR) and the solution and estimation of linearized dynamic stochastic general equilibrium models (DSGE). Here we cover the following topics:

a) Definitions and basic building blocks of linear time series.

b) Linear least squares and recursive projections.

c) Wold representation theorem.

d) Brief review of limit theorems.

Basic time series concepts

A time series is a set of repeated observations of a variable (say, GDP) over a number of periods $t = 1, 2, .., T$. We denote the time series by $\{x_1, x_2, ..., x_T\}$ (or $\{x_t\}_{t=1}^{T}$) and think of it as the realized value of (a chunk of) a stochastic process.

A stochastic process is a collection of random variables indexed by a number t in some set $\mathcal{T}$. We take $\mathcal{T}$ to be the set of integer numbers and think of each $t \in \mathcal{T}$ as a "time period". A time period could be one year, one quarter, one month, and so on. A stochastic process X_t is a collection of random variables

$$\boldsymbol{X} = \{X_t\}_{t=-\infty}^{\infty} = \{...X_{-2}, X_{-1}, X_0, X_1, X_2, ...\}. \tag{1.1}$$

This process extends infinitely into the past and the future. With each drawing of the stochastic process, we draw an entire *sequence* $\{x_t\}_{t=-\infty}^{\infty} \in \boldsymbol{X}$. Our interest lies in studying probability distributions over such sequences. By conceptualizing the stochastic process in this manner, we can use Hilbert space theory's tools to develop formal arguments and proofs, although we will mostly employ an informal approach.

To use Hilbert space theory, it is necessary to impose certain restrictions on the set of random variables under consideration. We assume that each X_t satisfies the condition

$$E\left[X_t^2\right] < \infty. \tag{1.2}$$

Random variables satisfying this condition are referred to as belonging to L^2, which represents the set of all random variables with finite second moments.

Hilbert spaces are the natural generalization of Euclidean spaces into infinite dimensions. In formal terms, a Hilbert space is a complete normed linear space in which the norm is defined through an *inner product*. For a given pair of elements X, Y belonging in L^2, we define the inner product as

$$\langle X, Y\rangle = E\left[XY\right].$$

The norm associated with this inner product is given by:

$$\|X\| = \langle X, X\rangle^{1/2} = \left(E\left[X^2\right]\right)^{1/2}.$$

We now prove a classic lemma that is widely used.

Lemma (Cauchy-Schwarz inequality)**.** *Let $X, X \in L^2$. Then,*

$$\left|E\left[xy\right]\right| \leq E\left[x^2\right]^{1/2} E\left[y^2\right]^{1/2}. \tag{1.3}$$

Proof. Note that, for any realization of X and Y,

$$\begin{aligned} 0 \leq & \left(\frac{|X|}{E\left[X^2\right]^{1/2}} - \frac{|Y|}{E\left[Y^2\right]^{1/2}}\right)^2 \\ = & \frac{|X|^2}{E\left[X^2\right]} + \frac{|Y|^2}{E\left[Y^2\right]} - 2\frac{|X|\,|Y|}{E\left[x^2\right]^{1/2} E\left[Y^2\right]^{1/2}}. \end{aligned}$$

Rearranging gives

$$\frac{|X|\,|Y|}{E\left[X^2\right]^{1/2} E\left[Y^2\right]^{1/2}} \leq \frac{1}{2}\left[\frac{|X|^2}{E\left[X^2\right]} + \frac{|Y|^2}{E\left[Y^2\right]}\right].$$

Taking expectations on both sides of the inequality and multiplying by $E\left[X^2\right]^{1/2} E\left[Y^2\right]^{1/2}$ gives $E\left|X\right|\left|Y\right| \leq E\left[X^2\right]^{1/2} E\left[Y^2\right]^{1/2}$. But $\left|E\left[XY\right]\right| \leq E\left[\left|X\right|\left|Y\right|\right]$, which leads to $\left|E\left[XY\right]\right| \leq E\left[X^2\right]^{1/2} E\left[Y^2\right]^{1/2}$. ■

In linear time series analysis, we usually characterize the probability distribution of a stochastic process using means and covariances. With normal distributions, this is all we need to characterize the probability distribution. We denote the (unconditional) mean and covariances of the process $\{X_t\}$ by

$$\begin{aligned} \mu_t = & E\left[X_t\right] \\ \sigma_{t,s} = & E\left[\left(X_t - \mu_t\right)\left(X_s - \mu_s\right)\right]. \end{aligned}$$

We say that a stochastic process in L^2 is *covariance stationary* if μ_t is the same for all t (that is, $E[x_t] = \mu$), and the covariance between X_t and X_s depends only on $t-s$, that is, the distance between time periods and not the particular calendar dates t and s. Other names for the same property—which we use interchangeably—are *weakly stationary*, *wide sense stationary*, *second order stationary*, or simply *stationary*.

The *covariogram* or *autocovariance function* is the sequence of covariances

$$\gamma(\tau) \equiv \sigma_{t,t-\tau} = E[(X_t - \mu)(X_{t-\tau} - \mu)], \tag{1.4}$$

where we already assumed stationarity (otherwise we should use $\gamma_t(\tau)$). The covariogram is symmetric, $\gamma(\tau) = \gamma(-\tau)$.[1]

Moreover, the Cauchy-Schwarz inequality implies

$$|E[(X_t - \mu)(X_{t-\tau} - \mu)]| \leq E\left[(X_t - \mu)^2\right]^{1/2} E\left[(X_{t-\tau} - \mu)^2\right]^{1/2}$$

or $|\gamma(\tau)| \leq \gamma(0)$ for all τ.

Because we can always extract the mean from any stationary time series, it saves a lot on notation to work with mean zero stochastic processes. If we want to recover the mean, we simply add it back to the process. So, from now on, we will mostly consider stochastic processes with zero mean.

A usual way to attach a probability distribution to X_t is through linear combinations of a serially uncorrelated stochastic process ε_t satisfying

$$\begin{aligned} E[\varepsilon_t] &= 0 \text{ for all } t, \\ E\left[\varepsilon_t^2\right] &= \sigma^2 \text{ for all } t, \\ E[\varepsilon_t \varepsilon_{t-\tau}] &= 0 \text{ for all } t \text{ and } \tau \neq 0. \end{aligned}$$

The process ε_t is covariance stationary and is referred to as a **white noise**.

Consider the random variable X_t defined as

$$x_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{1.5}$$

where $\{\varepsilon_t\}$ are white noise shocks, and the sequence $\{a_j\}$ satisfies $\sum_{j=0}^{\infty} a_j^2 < \infty$. We need this assumption to make sure that the variance of x_t is finite and, therefore, that x_t belongs to L^2 (proved later).

We call the stochastic process (1.5) an infinite order moving average and denote it by $MA(\infty)$. The Wold representation theorem (discussed below) proves that stochastic processes of the form (1.5) are sufficiently general to capture, in some sense to be discussed later, all linear properties of any covariance stationary stochastic process.

A special, but very important case of the process (1.5), is given by the family of ARMA models:

$$\begin{aligned} AR(p): &\quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t \\ MA(q): &\quad x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \\ ARMA(p,q): &\quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}. \end{aligned}$$

[1] $\gamma(\tau) = E[(X_t - \mu)(X_{t-\tau} - \mu)] = E[(X_{t+\tau} - \mu)(X_t - \mu)] = E\left[(X_t - \mu)\left(X_{t-(-\tau)} - \mu\right)\right] = \gamma(-\tau)$, where the second equality uses stationarity.

Lag operators

Once we start taking linear combinations of current and lagged variables, algebra becomes messy pretty fast. Therefore, it is useful to introduce the concept of the **lag operator**. The lag operator takes a sequence as input and delivers another sequence as output which is equal to the original sequence with the index lagged one period. That is $L\{x_t\} = \{y_t\}$ where $y_t = x_{t-1}$. But to avoid clutter, we simply write

$$Lx_t = x_{t-1}.$$

Clearly, $L(Lx_t) = Lx_{t-1} = x_{t-2}$. Let $L^2 x_t$ denote this double application of the lag operator. More generally, $L^p x_t = x_{t-p}$ for any $p \geq 1$. We also have $L^{-p} x_t = x_{t+p}$ and this defines the forward operator.[2]

We can also *define* a polynomial in the lag operator $a(L)$ as

$$a(L) = a_0 + a_1 L + a_2 L^2 + ... = \sum_{j=1}^{\infty} a_j L^j,$$

where $L^0 \equiv 1$. With this notation, the process (1.5) can be written as

$$y_t = a(L)\varepsilon_t = \left(\sum_{j=0}^{\infty} a_j L^j \right) \varepsilon_t.$$

ARMA models can be written in terms of the lag operator as

$$\begin{array}{rc} AR(p): & (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) x_t = \varepsilon_t \\ MA(q): & x_t = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q) \varepsilon_t \\ ARMA(p,q): & (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) x_t = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q) \varepsilon_t. \end{array}$$

We manipulate lag polynomials as if they were regular polynomials. For example, take an AR(1) model. By repeated substitution we have

$$\begin{aligned} x_t =& \phi x_{t-1} + \varepsilon_t \\ =& \phi^2 x_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\ =& \phi^3 x_{t-3} + \phi^2 \varepsilon_{t-2} + \phi \varepsilon_{t-1} + \varepsilon_t \\ & \vdots \\ =& \phi^{s+1} x_{t-s-1} + \phi^s \varepsilon_{t-s} + \phi^{s-1} \varepsilon_{t-s+1} + ... + \phi \varepsilon_{t-1} + \varepsilon_t, \end{aligned}$$

so that, if $|\phi| < 1$, the term $\phi^{s+1} x_{t-s-1}$ tends to zero in the mean-squared sense:

$$\lim_{s \to \infty} E\left(\phi^{s+1} x_{t-s-1}\right)^2 = \lim_{s \to \infty} \phi^{2(s+1)} E\left(x^2\right) = E\left(x^2\right) \lim_{s \to \infty} \phi^{2(s+1)} = 0.$$

[2] Formally, if we define the forward operator as $L^{-1} x_t = x_{t+1}$, it is easy to see that L^{-1} is the inverse of the lag operator L—and hence the notation. Indeed, recall that a function $h(\cdot)$ is called the inverse of a function $f(\cdot)$ if $h(f(x)) = x$. Therefore, we immediately see that $L^{-1}(Lx_t) = L^{-1}(x_{t-1}) = x_t$. This proves that the forward operator is the inverse of the lag-operator.

Taking the limit as $s \to \infty$ we obtain the $MA(\infty)$ representation of the AR(1) model

$$x_t = \sum_{s=0}^{\infty} \phi_s \varepsilon_{t-s} = \left[1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + ...\right] \varepsilon_t,$$

where $\phi_0 = 1$. Note that we can obtain the same expression using the lag operator as follows: write the AR(1) model as

$$(1 - \phi L)\, x_t = \varepsilon_t.$$

We need to *invert* the lag polynomial $(1 - \phi L)$. That is, we want to find an operator, which we denote by $(1 - \phi L)^{-1}$ or $1/(1 - \phi L)$, such that $(1 - \phi L)^{-1}(1 - \phi L) = 1$.

Recall that, for a real number c with $|c| < 1$, we have the geometric series expansion

$$\frac{1}{1-c} = 1 + c + c^2 + c^3 + \ldots.$$

This expansion *suggests* treating ϕL like a real number, with the hope that $|\phi| < 1$ implies $|\phi L| < 1$ in some sense. If this interpretation is correct[3], we obtain

$$(1 - \phi L)^{-1} = 1 + \phi L + \phi^2 L^2 + \ldots.$$

Therefore,

$$x_t = \frac{\varepsilon_t}{1 - \phi L} = \left[1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + ...\right] \varepsilon_t.$$

Multiplication of lag polynomials works in the obvious way. Suppose $a(L) = a_0 + a_1 L + a_2 L^2$ and $b(L) = b_0 + b_1 L + b_2 L^2$, then

$$\begin{aligned} a(L)\, b(L) &= \left(a_0 + a_1 L + a_2 L^2\right)\left(b_0 + b_1 L + b_2 L^2\right) \\ &= a_0 b_0 + a_0 b_1 L + a_0 b_2 L^2 + b_0 a_1 L + b_1 a_1 L^2 + b_2 a_1 L^3 + b_0 a_2 L^2 + b_1 a_2 L^3 + b_2 a_2 L^4 \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0) L + (a_0 b_2 + b_1 a_1 + b_0 a_2) L^2 + (b_2 a_1 + b_1 a_2) L^3 + b_2 a_2 L^4 \end{aligned}$$

Here is another trick for lag operators. Let's write the following AR(2) model,

$$\left(1 - \phi_1 L - \phi_2 L^2\right) x_t = \varepsilon_t,$$

in terms of a $MA(\infty)$ representation. Rather than inverting the AR(2) by brute force, we write

$$\begin{aligned} 1 - \phi_1 L - \phi_2 L^2 &= (1 - \lambda_1 L)(1 - \lambda_2 L) \\ &= 1 - (\lambda_1 + \lambda_2) L + \lambda_1 \lambda_2 L^2. \end{aligned}$$

Thus, λ_1 and λ_2 solve

$$\lambda_1 + \lambda_2 = \phi_1;\ \lambda_1 \lambda_2 = -\phi_2.$$

Therefore,

$$(1 - \lambda_1 L)(1 - \lambda_2 L)\, x_t = \varepsilon_t.$$

[3] And it is indeed correct. But to make this statement formal we need to define and analyze properties of the $z-$transform. See, for example, Gabel and Roberts, chapter 4.

These polynomials are invertible as long as $|\lambda_1| < 1$ and $|\lambda_2| < 1$. If this is the case, we can write

$$\begin{aligned} x_t &= (1-\lambda_1 L)^{-1}(1-\lambda_2 L)^{-1}\varepsilon_t \\ &= \left(\sum_{j=0}^{\infty}\lambda_1^j L^j\right)\left(\sum_{i=0}^{\infty}\lambda_2^i L^i\right)\varepsilon_t. \end{aligned}$$

This is still ugly. If $\lambda_1 \neq \lambda_2$, we can use another trick: partial fractions expansions

$$\begin{aligned} \frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} &= \frac{a}{1-\lambda_1 L}+\frac{b}{1-\lambda_2 L} \\ &= \frac{a(1-\lambda_2 L)+b(1-\lambda_1 L)}{(1-\lambda_1 L)(1-\lambda_2 L)} \\ &= \frac{a+b-(a\lambda_2+b\lambda_1)L}{(1-\lambda_1 L)(1-\lambda_2 L)}, \end{aligned}$$

which is true as long as $a+b=1$ and $a\lambda_2+b\lambda_1=0$ or

$$a=\frac{\lambda_1}{\lambda_1-\lambda_2},\ b=\frac{\lambda_2}{\lambda_2-\lambda_1}.$$

Therefore,

$$\begin{aligned} x_t &= (1-\lambda_1 L)^{-1}(1-\lambda_2 L)^{-1}\varepsilon_t \\ &= \left[\left(\frac{\lambda_1}{\lambda_1-\lambda_2}\right)\frac{1}{1-\lambda_1 L}+\left(\frac{\lambda_2}{\lambda_2-\lambda_1}\right)\frac{1}{1-\lambda_2 L}\right]\varepsilon_t \\ &= \sum_{j=0}^{\infty}\left[\frac{\lambda_1}{\lambda_1-\lambda_2}\lambda_1^j+\frac{\lambda_2}{\lambda_2-\lambda_1}\lambda_2^j\right]\varepsilon_{t-j}. \end{aligned}$$

When $\lambda_1=\lambda_2$, the algebra is different and we have to use binomial expansions with negative exponents (see Sargent (1987), pages 194-195).

In general, for an $AR(p)$ process we need to find the p roots of the polynomial $1-\phi_1 z-\phi_2 z^2-...-\phi_p z^p=0$. The λ's of the previous decomposition are the reciprocal of these roots. The $AR(p)$ is invertible as long as all roots of the above polynomial are greater than 1 in absolute value (so that the reciprocal of the roots, or the λ's, are *less* than one in absolute value). In this case we can write the $AR(p)$ model as

$$\begin{aligned} y_t &= \phi(L)^{-1}\varepsilon_t = \left[1-\phi_1 L-\phi_2 L^2-...-\phi_p L^p\right]^{-1}\varepsilon_t \\ &= \left[(1-\lambda_1 L)(1-\lambda_2 L)...(1-\lambda_p L)\right]^{-1}\varepsilon_t. \end{aligned}$$

If all the λ's are different, the partial fractions trick implies

$$\frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)...(1-\lambda_p L)}=\sum_{i=1}^{p}\frac{a_i}{1-\lambda_i L},$$

where

$$a_i=\frac{\lambda_i}{\prod_{j\neq i}(\lambda_i-\lambda_j)}\text{ for all } i,$$

so that

$$x_t = \sum_{j=0}^{\infty} \left(\sum_{i=1}^{p} a_i \lambda_i^j \right) \varepsilon_{t-j}.$$

More tricks using the lag operator can be found in Sargent (1987) and Cochrane (2005).

We now consider conditions under which a MA process is weakly stationary. We need to show that unconditional means and covariances are finite and do not depend on time. Write the $MA(\infty)$ as

$$x_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$$

where ε_t satisfies $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] = \sigma^2$ and $E[\varepsilon_t \varepsilon_{t-j}] = 0$ for $j \neq 0$. First, note that $E[x_t] = \sum_{j=0}^{\infty} \theta_j E[\varepsilon_t] = 0$, so the unconditional mean does not depend on time. The variance of x_t is given by

$$E\left[x_t^2\right] = E[\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}]^2 = \sum_{j=0}^{\infty} \theta_j^2 E\left[\varepsilon_{t-j}^2\right] = \sigma^2 \sum_{j=0}^{\infty} \theta_j^2$$

which is finite if and only if $\sum_{j=0}^{\infty} \theta_j^2 < \infty$.

The autocovariance $\gamma(\tau)$ is

$$\begin{aligned} E[x_t x_{t-\tau}] &= E\left[\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} \sum_{h=0}^{\infty} \theta_h \varepsilon_{t-\tau-h}\right] = \sum_{j=0}^{\infty} \sum_{h=0}^{\infty} \theta_j \theta_h E[\varepsilon_{t-j} \varepsilon_{t-\tau-h}] \\ &= \sum_{j=0}^{\infty} \theta_j \theta_{j-\tau} E[\varepsilon_{t-j} \varepsilon_{t-j}] = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j-\tau}. \end{aligned}$$

From the second to the third equality we use that $E[\varepsilon_{t-j} \varepsilon_{t-\tau-h}] = 0$ for all $j \neq \tau - h$ and $E[\varepsilon_{t-j} \varepsilon_{t-\tau-h}] = \sigma^2$ for $j = h + \tau \Rightarrow h = j - \tau$. This proves that $\gamma(\tau)$ depends only on τ and not on t. Moreover, by the Cauchy-Schwarz inequality we know that $|\gamma(\tau)| \leq \gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \theta_j^2$ which implies that all autocovariances are finite. Thus, the $MA(\infty)$ is covariance stationary as long as $\sum_{j=0}^{\infty} \theta_j^2 < \infty$.

Remark: One has to be careful when passing the expectation operator through the summation operator—and also, when we claimed above that in the AR(1) case, $\lim_{s \to \infty} \phi^s x_{t-s} = 0$. This requires using some theorems from the Lebesgue theory of integration which are beyond the scope of these notes. But if are a nerd, see pages 26-39 of Fuller (1995).

Linear projections

Here we consider a problem that will appear in various forms throughout the course. Let y, x_1, x_2, ... , x_n denote a set of random variables in L^2. The Cauchy-Schwarz inequality (1.3) implies that the second moments $E[yx_j]$ and $E[x_i x_j]$ exist and are finite as well. Moreover, it can be shown that, if second moments exist, first moments exist as well, so $E[y]$, $[x_1]$, $E[x_2]$, ..., $E[x_n]$ are also finite.

Consider estimating the random variable y on the basis of knowing the values of the random variables x_1, x_2,..., x_n. In particular, we want to compute the **best linear projection** defined as the linear (affine) function

$$\hat{y} = a_0 + a_1 x_1 + ... + a_n x_n$$

that best approximates y.

By best approximation we mean the following: we choose numbers a_i that makes the random variable $\hat{y}$ as close as possible to y in the least squares sense $E\left(y - \hat{y}\right)^2$:

$$\min_{\{a_i\}} E\left[\left(y - a_0 x_0 - a_1 x_1 - ... - a_n x_n\right)^2\right], \tag{1.6}$$

where we created a trivial random variable $x_0 \equiv 1$. This is equivalent to ordinary least squares but using population rather than sample moments.

Theorem 1.1 — Orthogonality principle

The numbers $a_0, a_1, a_2..., a_n$ minimize (1.6) if and only if

$$E\left[\left(y - a_0 x_0 - a_1 x_1 - ... - a_n x_n\right) x_i\right] = 0 \text{ for } i = 0, 1, 2, ..., n. \tag{1.7}$$

Proof. Let $a = (a_0, a_1, ..., a_n)'$ and consider the minimization problem

$$\min_a J(a) = \min_a \frac{1}{2} E[(y - \sum_{j=0}^{n} a_j x_j)^2].$$

Differentiating (1.6) with respect to a_i gives

$$\frac{\partial J(a)}{\partial a_i} = -E(y - \sum_{j=0}^{n} a_j x_j) x_i = 0 \text{ for } i = 0, 1, 2, ..., n.$$

This shows that (1.7) is a necessary condition. If we show that the problem is strictly convex, the minimizer is unique and sufficiency also holds.

Let $x = (x_0, x_1, x_2, ..., x_n)'$ be an $(n+1) \times 1$ column vector. The first order conditions can be written in matricial form as

$$\nabla_a J(a) = -\left[E(xy) - E(xx')\, a\right] = \mathbf{0}_{n+1 \times 1}.$$

Differentiating this expression with respect to a' gives

$$\nabla_{aa'} J(a) = E(xx'),$$

which is positive definite because $E[xx']$ is a covariance matrix. Therefore, the minimization problem is convex and the first order conditions are sufficient. This completes the proof. ■

Assuming that $E[xx']$ is invertible[4], we can obtain the optimal weights a by solving

$$a = E(xx')^{-1} E(xy). \tag{1.8}$$

The random variable $\sum_{j=0}^{n} a_j x_j$ is called the **projection** of y onto $\{1, x_1, x_2, ..., x_n\}$.

The orthogonality principle implies that the projection error, $y - \sum_{j=0}^{n} a_j x_j$, is **orthogonal** to each of the x_i and, therefore, to any linear combination of them. (Two random variables x, y are orthogonal if $E[xy] = 0$.) Defining the projection error as ε, it follows that

$$y = \sum_{j=0}^{n} a_j x_j + \varepsilon \tag{1.9}$$

where $E[\varepsilon \sum_{i=0}^{n} \phi_i x_i] = 0$ for any $\{\phi_i\}$ (*why?*). Thus, equation (1.9) decomposes y into two orthogonal components: $\sum_{j=0}^{n} a_j x_j$ and ε.

It then follows that

$$E\left(y^2\right) = E\left(\sum_{j=0}^{n} a_j x_j\right)^2 + E\left(\varepsilon^2\right).$$

In addition, note that $E(\varepsilon) = 0$. This follows from the orthogonality condition for $i = 0$ above. The key for this result is to include a constant in the projection; without the constant, the forecast error ε need not have zero mean.

Sometimes we use the following notation for the projection

$$P[y|1, x_1, x_2, ..., x_n] \equiv x'a = \sum_{j=0}^{n} a_j x_j.$$

Lemma. *The projection is a linear operator:*

$$P[\alpha y + \beta z|1, x_1, x_2, ..., x_n] = \alpha P[y|1, x_1, x_2, ..., x_n] + \beta P[z|1, x_1, x_2, ..., x_n].$$

Proof. Let $P[y|1, x_1, x_2, ..., x_n] = \sum_{j=0}^{n} a_j x_j$ and $P[z|1, x_1, x_2, ..., x_n] = \sum_{j=0}^{n} b_j x_j$. The orthogonality principle implies

$$\begin{aligned} E\left(y - \sum_{j=0}^{n} a_j x_j\right) x_i &= 0 \text{ for all } i \\ E\left(z - \sum_{j=0}^{n} b_j x_j\right) x_i &= 0 \text{ for all } i \end{aligned}$$

[4] $E[xx']$ not invertible means that (at least) one of the random variables x_i is a linear combination of the others. If a random variables is a linear combinations of the others it does not add anything to the linear projection. We can simply delete these variables until we obtain an invertible covariance matrix $E[xx']$.

Multiplying the first condition by α and the second by β gives

$$\begin{aligned} E\left(\alpha y - \alpha \sum_{j=0}^{n} a_j x_j\right) x_i &= 0 \text{ for all } i \\ E\left(\beta z - \beta \sum_{j=0}^{n} b_j x_j\right) x_i &= 0 \text{ for all } i \end{aligned}$$

Adding these equations gives

$$E\left[\alpha y + \beta z - \sum_{j=0}^{n} (\alpha a_j + \beta b_j) x_j\right] x_i = 0 \text{ for all } i.$$

This means that the numbers $(\alpha a_j + \beta b_j)$ for $j = 0, 1, 2, ..., n$ satisfy the orthogonality principle of a projection of $\alpha y + \beta z$ onto $\{1, x_1, x_2, ..., x_n\}$. Therefore,

$$\begin{aligned} P[\alpha y + \beta z | 1, x_1, x_2, ..., x_n] &= \sum_{j=0}^{n} (\alpha a_j + \beta b_j) x_j \\ &= \alpha \sum_{j=0}^{n} a_j x_j + \beta \sum_{j=0}^{n} b_j x_j \\ &= \alpha P[y | 1, x_1, x_2, ..., x_n] + \beta P[z | 1, x_1, x_2, ..., x_n], \end{aligned}$$

which completes the proof. ■

Recursive projections

Here we show how to update a projection when new information arrives. This result will be useful to derive the updating formula for the Kalman filter.

We observe a set of random variables $\Omega = \{1, x_1, x_2, ..., x_n\}$ (we include the constant 1 in Ω) and compute the projection $P[y|\Omega]$. Suppose that we are given a new set of random variables $\mathbf{z} = (z_1, z_2, ..., z_m)'$ and want to compute (update) the linear projection $P[y|\Omega, \mathbf{z}]$ based on our knowledge of $P[y|\Omega]$.

Consider the decomposition (1.9) for the updated projection:

$$\begin{aligned} y =& P[y|\Omega, \mathbf{z}] + \varepsilon \\ =& \sum_{j=0}^{n} a_j x_j + \sum_{s=1}^{m} \delta_s z_s + \varepsilon. \end{aligned} \tag{1.10}$$

where $E(\varepsilon) = 0$, $E(\varepsilon x_j) = 0$ for $j = 1, 2, ..., n$, and $E(\varepsilon z_s) = 0$ for $s = 1, 2, ...m$. The orthogonality principle guarantees that the a_j's and δ_s's are the least square parameter values. Now project both sides of (1.10) on the smaller set Ω,

$$\begin{aligned} P[y|\Omega] =& P\left[\sum_{j=0}^{n} a_j x_j + \sum_{s=1}^{m} \delta_s z_s + \varepsilon | \Omega\right] \\ =& \sum_{j=0}^{n} a_j P[x_j|\Omega] + \sum_{s=1}^{m} \delta_s P[z_s|\Omega] + P[\varepsilon|\Omega]. \end{aligned}$$

where we used that the projection is a linear operator. Moreover, $P[x_j|\Omega] = x_j$ for $j = 1, 2, ..., n$ and $P[\varepsilon|\Omega] = 0$. To see the former, consider the objective function of the least square problem for

$$\min_{a_0,a_1,..,a_n} E\left[(x_j - a_0x_0 - a_1x_1 - a_2x_2 - ... - a_nx_n)^2\right].$$

This is clearly minimized when $a_j = 1$ and $a_i = 0$ for all $i \neq j$. To see the latter, note that the orthogonality conditions of (1.10) imply that all the coefficients of the linear projection of ε on x_j for $j = 1, 2, ..., n$ are zero. Thus, we obtain

$$P[y|\Omega] = \sum_{j=0}^{n} a_j x_j + \sum_{s=1}^{m} \delta_s P[z_s|\Omega]. \tag{1.11}$$

Subtracting (1.11) from (1.10) then gives

$$y - P[y|\Omega] = \sum_{s=1}^{m} \delta_s (z_s - P[z_s|\Omega]) + \varepsilon. \tag{1.12}$$

This equation looks like a projection of $y - P[y|\Omega]$ on $z_s - P[z_s|\Omega]$. To confirm this conjecture, we need to show that ε is orthogonal to $z_s - P[z_s|\Omega]$ for all s. But this is obvious because ε is orthogonal to z_s and x_j, and $P[z_s|\Omega]$ is a linear function of x_j, hence $E[(z_s - P[z_s, \Omega])\varepsilon] = 0$ for all s. Therefore, the orthogonality principle implies that δ_s for $s = 1, 2, ..., m$ are the coefficients of the projection of $(y - P[y|\Omega])$ on $(z_s - P[z_s, \Omega])$,

$$P[(y - P[y|\Omega]) \,|\, (\mathbf{z} - P[\mathbf{z}|\Omega])] = \sum_{s=1}^{m} \delta_s (z_s - P[z_s|\Omega]).$$

Rearranging (1.12) and using the previous result gives

$$y = P[y|\Omega] + P[(y - P[y|\Omega]) \,|\, (\mathbf{z} - P[\mathbf{z}|\Omega])] + \varepsilon.$$

Because ε is orthogonal to $\{\Omega, \mathbf{z}\}$, it then follows that

$$P[y|\Omega, z] = \underbrace{P[y|\Omega]}_{\text{Original projection}} + \underbrace{P[(y - P[y|\Omega]) \,|\, (\mathbf{z} - P[\mathbf{z}|\Omega])]}_{\text{Projection of prediction errors on prediction errors}}. \tag{1.13}$$

In words, to update a linear projection when new information $\mathbf{z}$ arrives, one adds to the original projection, $P[y|\Omega]$, the projection of the prediction errors of the original projection, $y - P[y|\Omega]$, on the prediction errors of the projection of the new variables $\mathbf{z}$ on the original set of variables, $\mathbf{z} - P(\mathbf{z}|\Omega)$.

For future use, let's write (1.13) in vector notation. Let $x = (1, x_1, x_2, ..., x_n)'$ and $z = (z_1, z_2, ..., z_m)'$. Furthermore, stack the projection coefficients as $a = (a_0, a_1, ..., a_n)'$ and $b^s = (b_0^s, b_1^s, b_2^s, ..., b_n^s)'$ for all s. The normal equations are

$$\begin{aligned} a =& E(xx')^{-1} E(xy) \\ b^s =& E(xx')^{-1} E(xz_s) \end{aligned}$$

which implies

$$\begin{aligned} P[y|x] =& x'a = x'E(xx')^{-1}E(xy) \\ P[z_s|x] =& x'b^s = x'E(xx')^{-1}E(xz_s). \end{aligned}$$

The projection errors are thus

$$y - P[y|x] = y - x'a = y - x'E(xx')^{-1}E(xy)$$

$$u_s \equiv z_s - P[z_s|x] = z_s - x'b^s = z_s - x'E(xx')^{-1}E(xz_s).$$

Let $u = (u_1, u_2, ..., u_m)'$ denote the vector of the projections errors of z_s on x for $s = 1, 2, ..., m$, and $B = [b^1\ b^2\ ...\ b^m]$ be the $(n+1) \times m$ matrix whose column s contains the vector of projection coefficients of z_s on x. Then, we can write

$$u = z - B'x.$$

It then follows that the $m \times 1$ vector δ of coefficients of the projection of the forecast errors $y - P[y|x]$ on the forecast errors u is given by

$$\begin{aligned} \delta =& E(uu')^{-1}E[u(y - P[y|x])] \\ =& E(uu')^{-1}E\left[u\left(y - x'E[xx']^{-1}E[xy]\right)\right] \end{aligned}$$

so that

$$P\left[\left(y - x'E[xx']^{-1}E[xy]\right)|u\right] = u'\delta = u'E(uu')^{-1}E\left[u\left(y - x'E(xx')^{-1}E(xy)\right)\right].$$

Therefore

$$\begin{aligned} P[y|x,z] =& P[y|x] + P[(y - P[y|x])\,|\,(z - P[z|x])] \\ =& x'E(xx')^{-1}E(xy) + u'E(uu')^{-1}E\left[u\left(y - x'E(xx')^{-1}E(xy)\right)\right], \end{aligned}$$

where $u = z - B'x$.

Wold Representation Theorem

Above we constructed a covariance stationary process by combining white noise shocks

$$x_t = a(L)\varepsilon_t,$$

where $a(L) = 1 + a_1L + a_2L^2 + ...$ is a polynomial on the lag operator and $\sum_{j=0}^{\infty} a_j^2 < \infty$. The Wold representation theorem reverses the procedure: any weakly stationary process can be written as an infinite order moving average plus a perfectly predictable term.

Theorem 1.2 — Wold Representation Theorem

Any mean zero, covariance stationary process $\{x_t\}$ can be represented in the form

$$x_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} + \eta_t \tag{1.14}$$

where

a) $\varepsilon_t = x_t - P[x_t|x_{t-1}, x_{t-2}, ...]$ is the prediction error of the projection of x_t on all its lags,

b) $P[\varepsilon_t|x_{t-1}, x_{t-2}, ...] = 0$, $E(\varepsilon_t x_{t-j}) = 0$ for all $j \geq 1$; $E(\varepsilon_t^2) = \sigma^2$ for all t; $E(\varepsilon_t) = 0$ for all t; $E(\varepsilon_t \varepsilon_s) = 0$ for all $s \neq t$,

c) $\theta_0 = 1$; $\sum_{j=0}^{\infty} \theta_j^2 < \infty$,

d) $\{\theta_j\}$ and $\{\varepsilon_t\}$ are unique,

e) η_t is linearly deterministic: $\eta_t = P[\eta_t|x_{t-1}, x_{t-2}, x_{t-3}, ...]$.

Before going to the proof of the theorem, we mention what the theorem says and what the theorem does not say (mostly from Cochrane, 2005):

a) The ε_t's are a white noise but need not be i.i.d. or normally distributed.

b) Although $E(\varepsilon_t x_{t-j}) = 0$ (ε_t and x_{t-j} are orthogonal) $E(\varepsilon_t|x_{t-j})$ need not be zero. This is the difference between orthogonality and independence: two random variables x and y can be orthogonal but not independent. For example, let x be normal with mean zero and variance σ^2 and let $y = x^2$. Then $E(xy) = E(x^3) = 0$ but $E(y|x) = x^2$.

c) The innovations ε_t are prediction errors. They do not have a structural interpretation as the shocks of a model. The Wold decomposition is a probabilistic decomposition.

d) The Wold decomposition is *one linear representation* of the process $\{x_t\}$. There could be other non-linear representations that may be better in some sense. Moreover, the Wold decomposition is not even the *unique linear* $MA(\infty)$ representation of the process (see below).

e) We usually ignore η_t.

We provide a sketch of the proof following Sargent (1987)–the formal proof requires being more careful in some steps.

Proof of the Wold Representation Theorem. The proof is constructive. We divide the proof in a number of steps:

Step 1: construct the white noise process ε_t.

Using the orthogonality principle, write

$$x_t = P\left[x_t | x_{t-1}, x_{t-2}, ...\right] + \varepsilon_t,$$

where $P\left[x_t | x_{t-1}, x_{t-2}, ...\right]$ is the projection of x_t on the entire history of past x's, and ε_t is a prediction error orthogonal to x_{t-j} for $j = 1, 2, ...$ This defines the sequence of unique innovations $\{\varepsilon_t\}$ because the projection is unique. Furthermore, since each x_t has mean zero and ε_t is a linear combination of x's, then $E(\varepsilon_t) = 0$.

Let σ^2 be the mean squared error of the projection,

$$\sigma^2 = E\left(\varepsilon_t^2\right) = E\left(x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right]\right)^2.$$

Note that σ^2 does not depend on t because ε_t is a linear combination of current and past x's, and x is covariance stationary.

Using that

$$\varepsilon_{t-s} = x_{t-s} - P\left[x_{t-s} | x_{t-s-1}, x_{t-s-2}, ...\right]$$

is a linear combination of $x_{t-s}, x_{t-s-1}, ...$ and that the orthogonality principle implies $E\left[\varepsilon_t x_{t-s}\right] = 0$ for all $s \geq 1$, it follows that $E\left[\varepsilon_t \varepsilon_{t-s}\right] = 0$ for all $s \geq 1$. This proves that $\{\varepsilon_t\}$ is a serially uncorrelated process.

Step 2: Construct the coefficients θ_j of the projection of x_t on past ε

Consider projecting x_t on a sequence of (finite) sets $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-m}\}$ for successively larger m's. Denote the projection of x_t on such set as

$$\hat{x}_t^m = P\left[x_t | \varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-m}\right] = \sum_{j=0}^{m} \theta_j \varepsilon_{t-j}.$$

The orthogonality principle implies that the prediction error is orthogonal to each ε in the set,

$$E[(x_t - \sum_{j=0}^{m} \theta_j \varepsilon_{t-j})\varepsilon_{t-k}] = 0 \text{ for } k = 0, 1, 2, ..., m.$$

Since $E\left[\varepsilon_{t-j}\varepsilon_{t-k}\right] = 0$ for all $j \neq k$ we have

$$E\left[x_t \varepsilon_{t-k}\right] - \theta_k E\left[\varepsilon_{t-k}^2\right] = 0 \text{ for } k = 0, 1, 2, ..., m$$

so that

$$\theta_k = \frac{E\left(x_t \varepsilon_{t-k}\right)}{\sigma^2} \text{ for } k = 0, 1, 2, ..., m.$$

Let $k = 0$. Since $\varepsilon_t = x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right]$, it follows that $E\left(\varepsilon_t x_{t-j}\right) = 0$ for all $j \geq 1$ (orthogonality principle). Therefore,

$$\begin{aligned} E\left[\varepsilon_t x_t\right] &= E\left[\varepsilon_t \left(\varepsilon_t + P\left[x_t | x_{t-1}, x_{t-2}, ...\right]\right)\right] \\ &= E\left(\varepsilon_t^2\right) + \underbrace{E\left(\varepsilon_t P\left[x_t | x_{t-1}, x_{t-2}, ...\right]\right)}_{=0}. \end{aligned}$$

Thus, $\theta_0 = E\left(\varepsilon_t x_t\right) / E\left(\varepsilon_t^2\right) = E\left(\varepsilon_t^2\right) / E\left(\varepsilon_t^2\right) = 1$. *A key property of these projections is that* θ_k *does not depend on* m*, the length of the projection set* $\{\varepsilon_t, \varepsilon_{t-1}, ..., \varepsilon_{t-m}\}$. This property reflects the lack of serial correlation of the ε's.

We now compute the variance of the prediction error

$$\begin{aligned}
0 \leq E\left(x_t - \sum_{j=0}^{m} \theta_j \varepsilon_{t-j}\right)^2 &= E[x_t^2 - 2x_t \sum_{j=0}^{m} \theta_j \varepsilon_{t-j} + (\sum_{j=0}^{m} \theta_j \varepsilon_{t-j})^2] \\
&= E\left(x_t^2\right) - 2\sum_{j=0}^{m} \theta_j E\left(x_t \varepsilon_{t-j}\right) + \sum_{j=0}^{m} \theta_j^2 E\left(\varepsilon_{t-j}\right)^2 \\
&= E\left(x_t^2\right) - 2\sum_{j=0}^{m} \theta_j \sigma^2 \frac{E\left(x_t \varepsilon_{t-j}\right)}{\sigma^2} + \sum_{j=0}^{m} \theta_j^2 \sigma^2 \\
&= E\left(x_t^2\right) - 2\sigma^2 \sum_{j=0}^{m} \theta_j^2 + \sigma^2 \sum_{j=0}^{m} \theta_j^2 \\
&= E\left(x_t^2\right) - \sigma^2 \sum_{j=0}^{m} \theta_j^2,
\end{aligned}$$

where the second equality follows because the ε_{t-j} are uncorrelated, and the fourth equality uses the definition of θ_j. Since $E\left(x_t^2\right) < \infty$, the previous inequality implies

$$\sum_{j=0}^{m} \theta_j^2 \leq \frac{E\left(x_t^2\right)}{\sigma^2} < \infty \text{ for all } m.$$

Taking the limit as $m \to \infty$ proves that $\sum_{j=0}^{\infty} \theta_j^2 < \infty$—the sequence $\{\theta_j\}$ is square summable.

The square summability of $\{\theta_j\}$ implies that the projection $\hat{x}_t^m$ is a Cauchy sequence. To see this, take $n > m$ and compute

$$\begin{aligned}
\|\hat{x}_t^n - \hat{x}_t^m\|^2 &= E\left(\hat{x}_t^n - \hat{x}_t^m\right)^2 \\
&= E\left(\sum_{j=0}^{n} \theta_j \varepsilon_{t-j} - \sum_{j=0}^{m} \theta_j \varepsilon_{t-j}\right)^2 \\
&= E\left(\sum_{j=m+1}^{n} \theta_j \varepsilon_{t-j}\right)^2 \\
&= \sum_{j=m+1}^{n} \theta_j^2 \sigma^2 \leq \sigma^2 \sum_{j=m+1}^{\infty} \theta_j^2
\end{aligned}$$

Since $\sum_{j=0}^{m} \theta_j^2 < \infty$ for all m, it follows that we can take m big enough to make $\sum_{j=m+1}^{\infty} \theta_j^2$ arbitrarily close to zero. This means that $\hat{x}_t^m$ is Cauchy. Therefore, the completeness of the Hilbert space L^2 implies that there exist an element $\hat{x}_t \in L_2$ such that

$$\hat{x}_t^m \to \hat{x}_t \equiv \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}. \tag{1.15}$$

Step 3: Construct the component η_t

Let η_t be the difference between x_t and the projection of x_t onto the current and past ε_t's

$$\eta_t \equiv x_t - \hat{x}_t = x_t - \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} \tag{1.16}$$

We first establish that $E(\eta_t \varepsilon_s) = 0$ for all s and t. It should be clear that $E(\eta_t \varepsilon_s) = 0$ for $s > t$ because η_t is a linear function of x_t and the sequence of shocks $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, ...$ However, ε_s is orthogonal to all other ε_t's and to x_t when $s > t$. Consider now the case $s \leq t$,let $s = t - k$ for the appropriate $k \geq 0$ and compute

$$\begin{aligned} E(\eta_t \varepsilon_{t-k}) &= E(x_t \varepsilon_{t-k}) - E(\hat{x}_t \varepsilon_{t-k}) \\ &= E(x_t \varepsilon_{t-k}) - E[(\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}) \varepsilon_{t-k}] \\ &= \sigma^2 \theta_k - \sigma^2 \theta_k = 0. \end{aligned}$$

where we are using the definition of θ_k. Thus, the entire $\{\varepsilon_t\}$ process is orthogonal to the process $\{\eta_t\}$.

We next establish that η_t is perfectly predictable from past observations of x_t's. In particular, project η_t on $\{x_{t-1}, x_{t-2}, ...\}$

$$P[\eta_t | x_{t-1}, x_{t-2}, ...] = P[x_t | x_{t-1}, x_{t-2}, ...] - \sum_{j=0}^{\infty} \theta_j P[\varepsilon_{t-j} | x_{t-1}, x_{t-2}, ...], \tag{1.17}$$

where we used linearity of the projection. Consider the second term on the right hand side. In the case $j = 0$ we need to compute $P[\varepsilon_t | x_{t-1}, x_{t-2}, ...]$. However, we already established that ε_t is orthogonal to all past x_t's, so that $E(\varepsilon_t x_{t-j}) = 0$ for all $j \geq 1$ which implies

$$P[\varepsilon_t | x_{t-1}, x_{t-2}, ...] = 0.$$

Consider now computing the projection $P[\varepsilon_t | x_t, x_{t-1}, x_{t-2}....]$. By step 1 above, ε_t is a linear combination of the current and past x_t's. Therefore, the orthogonality principle implies $P[\varepsilon_t | x_t, x_{t-1}....] = \varepsilon_t$ because, ε_t being a linear combination of current and past x_t's, means that we can set the objective function of the projection problem to exactly zero. A similar argument can be made to argue that

$$P[\varepsilon_t | x_{t+j-1}, x_{t+j-2}, ...] = \varepsilon_t \text{ for } j \geq 0.$$

In effect, ε_t is a linear combination of a subset of the projecting variables $x_{t+j-1}, x_{t+j-2}, ...$ It will then be possible to make the objective function of the projection exactly equal to zero as well.

This implies that

$$\begin{aligned} P[\varepsilon_t | x_{t-1}, x_{t-2}, ...] &= \varepsilon_t \\ P[\varepsilon_{t-1} | x_{t-1}, x_{t-2}, ...] &= \varepsilon_{t-1} \\ &\vdots \\ P[\varepsilon_{t-j} | x_{t-1}, x_{t-2}, ...] &= \varepsilon_{t-j}. \end{aligned}$$

Therefore, (1.17) becomes

$$P[\eta_t | x_{t-1}, x_{t-2}, ...] = P[x_t | x_{t-1}, x_{t-2}, ...] - \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j}.$$

Subtracting (1.16) from this expression gives

$$\begin{aligned}
\eta_t - P\left[\eta_t | x_{t-1}, x_{t-2}, ...\right] &= \left(x_t - \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}\right) - \left(P\left[x_t | x_{t-1}, x_{t-2}, ...\right] - \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j}\right) \\
&= \underbrace{\left(x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right]\right)}_{\varepsilon_t} - \left(\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} - \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j}\right) \\
&= \underbrace{\left(x_t - P\left[x_t | x_{t-1}, x_{t-2}, ...\right]\right)}_{\varepsilon_t} - \theta_0 \varepsilon_t \\
&= \varepsilon_t - \theta_0 \varepsilon_t = 0.
\end{aligned}$$

All this algebra proves that

$$\eta_t = P\left[\eta_t | x_{t-1}, x_{t-2}, ...\right],$$

which means that the term η_t is "linearly deterministic" in the sense that it can be predicted without error using past information of x's. The remaining term, $\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$ is called the "linearly indeterministic" component of the process. This completes the proof of the theorem. ■

Wold representation for vector time series

The same decomposition holds for vector processes. Let $X_t = [x_{1t}, x_{2t}, ..., x_{nt}]'$ where each x_{it} is an individual stochastic process. We say that X_t is covariance stationary if $E[X_t] = \mu$ is independent of time and the matrix of autocovariances $E\left[(X_t - \mu)(X_{t-\tau} - \mu)'\right] = \Gamma_\tau$ only depends on τ and not on t.[5] As above, assume $\mu = 0$. Then, any covariance stationary vector process X_t can be represented as

$$X_t = \sum_{j=0}^{\infty} \Theta_j \varepsilon_{t-j} + \eta_t \tag{1.18}$$

where

a) $\varepsilon_t = X_t - P\left[X_t | X_{t-1}, X_{t-2}, X_{t-3}, ...\right]$ is the forecast error of the projection of the vector X_t on its lagged values,

b) $P\left[\varepsilon_t | X_{t-1}, X_{t-2}, X_{t-3}, ...\right] = 0$, $E(\varepsilon_t X_{t-j}) = 0$ for all $j \geq 1$; $E(\varepsilon_t^2) = \Sigma$ for all t is a constant covariance matrix; $E(\varepsilon_t) = 0$ for all t; $E(\varepsilon_t \varepsilon_s') = 0$ for all $s \neq t$,

c) Θ_j are $n \times n$ matrices that satisfy $\Theta_0 = I$; $\sum_{j=0}^{\infty} \Theta_j \Theta_j' < \infty$,

d) $\{\Theta_j\}$ and $\{\varepsilon_t\}$ are unique,

e) η_t is linearly deterministic; that is, $\eta_t = P\left[\eta_t | X_{t-1}, X_{t-2}, X_{t-3}, ...\right]$.

[5] Please note that the symmetry property now reads $\Gamma_\tau = \Gamma_{-\tau}'$.

An important remark: The Wold representation theorem shows that there is a unique representation of a covariance stationary process as a $MA(\infty)$ satisfying 1-5 above. This *does not mean* that (1.18) is the *unique* moving average representation of the process $\{X_t\}$. To see this, note that we can always write (1.18) as

$$X_t = \sum_{j=0}^{\infty} \Theta_j \varepsilon_{t-j} + \eta_t = \sum_{j=0}^{\infty} \Theta_j \Lambda \Lambda^{-1} \varepsilon_{t-j} + \eta_t = \sum_{j=0}^{\infty} \Phi_j \nu_{t-j} + \eta_t,$$

where Λ is an arbitrary $n \times n$ invertible matrix, $\Phi_j = \Theta_j \Lambda$ and $\nu_{t-j} = \Lambda^{-1}\varepsilon_{t-j}$. The innovation ν_t now satisfies $E(v_t) = \Lambda^{-1}E(\varepsilon_t) = 0$, $E\left(v_t v'_{t-s}\right) = 0$ for $s \neq 0$ and $E(v_t v'_t) = E\left(\Lambda^{-1}\varepsilon_t \varepsilon'_t (\Lambda^{-1})'\right) = \Lambda^{-1}\Sigma(\Lambda^{-1})'$. Therefore $X_t = \sum_{j=0}^{\infty} \Phi_j \nu_{t-j} + \eta_t$ is *another* infinite moving average representation of the vector process X_t. How do we relate this to the uniqueness claim in the Wold theorem? What happens here is that the residual ν_t *is not* the forecast error of projecting X_t on its infinite history.

This non-uniqueness result of the moving average representation of X_t will be used when discussing structural vector autoregressions later in the course.

Limit theorems

We use different versions of two limit theorems: Laws of Large Numbers (LLN) and Central Limit Theorems (CLT). Both are concerned with the behavior of sample means under different assumptions. The LLN is about convergence—in probability, almost surely, in L^2—of the sample mean to the population mean. The CLT is about convergence in distribution (the asymptotic distribution) of the sample mean. By appropriately weighting the sample mean by a function of the sample size (typically $\sqrt{T}$), the central limit theorem provides a non-degenerate distribution theory that can be used to test hypotheses, compute asymptotic confidence bands, etc.

Properties of the sample mean of a vector process (Hamilton, p. 279): Suppose that we have a sample of size T, $\{X_1, X_2, ..., X_T\}$ of an n dimensional vector process $\{X_t\}$, where X_t is covariance stationary with

$$\begin{aligned} E[X_t] &= \mu, \\ E\left[(X_t - \mu)(X_{t-v} - \mu)'\right] &= \Gamma_v. \end{aligned}$$

Assume also that the autocovariances are absolutely summable, that is $\sum_{v=-\infty}^{\infty} |\Gamma_v| < \infty$. If we let $\gamma_{ij}^{(v)}$ denote the element (i,j) of Γ_v, the requirement is that $\sum_{v=\infty}^{\infty} |\gamma_{ij}^{(v)}| = c_{ij} < \infty$. Recall also that for a vector process $\Gamma_v = \Gamma'_{-v}$.

Consider the sample mean

$$\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t.$$

Clearly, $E\left[\bar{X}_T\right] = \frac{1}{T}\sum_{t=1}^{T} E[X_t] = \mu$.

The covariance matrix of the sample mean is

$$E\left[\left(\bar{X}_T-\mu\right)\left(\bar{X}_T-\mu\right)'\right]=E\left[\left(\frac{1}{T}\sum_{t=1}^{T}X_t-\mu\right)\left(\frac{1}{T}\sum_{t=1}^{T}X_t-\mu\right)'\right]$$

$$=\frac{1}{T^2}E\begin{bmatrix}(X_1-\mu)\left[(X_1-\mu)'+(X_2-\mu)'+(X_3-\mu)'+...+(X_T-\mu)'\right]+\\(X_2-\mu)\left[(X_1-\mu)'+(X_2-\mu)'+(X_3-\mu)'+...+(X_T-\mu)'\right]+\\(X_3-\mu)\left[(X_1-\mu)'+(X_2-\mu)'+(X_3-\mu)'+...+(X_T-\mu)'\right]+\\...+\\(X_T-\mu)\left[(X_1-\mu)'+(X_2-\mu)'+(X_3-\mu)'+...+(X_T-\mu)'\right]\end{bmatrix},$$

or

$$\begin{aligned}T^2E\left[\left(\bar{X}_T-\mu\right)\left(\bar{X}_T-\mu\right)'\right]=&\Gamma_0+\Gamma_{-1}+\Gamma_{-2}+...+\Gamma_{-(T-1)}+\\&\Gamma_1+\Gamma_0+\Gamma_{-1}+...+\Gamma_{-(T-2)}+\\&\Gamma_2+\Gamma_1+\Gamma_0+\Gamma_{-1}+...+\Gamma_{-(T-3)}+\\&+...+\\&\Gamma_{T-1}+\Gamma_{T-2}+...+\Gamma_0\\=&T\Gamma_0+(T-1)\Gamma_1+(T-1)\Gamma_{-1}+(T-2)\Gamma_2+(T-2)\Gamma_{-2}+\\&(T-3)\Gamma_3+(T-3)\Gamma_{-3}+...+(T-(T-1))\Gamma_{T-1}+(T-(T-1))\Gamma_{-(T-1)}\\=&\sum_{v=-(T-1)}^{T-1}(T-|v|)\Gamma_v.\end{aligned}$$

Thus

$$TE\left[\left(\bar{X}_T-\mu\right)\left(\bar{X}_T-\mu\right)'\right]=\sum_{v=-(T-1)}^{T-1}\left(1-\frac{|v|}{T}\right)\Gamma_v=\sum_{v=-(T-1)}^{T-1}\Gamma_v-\sum_{v=-(T-1)}^{T-1}\frac{|v|}{T}\Gamma_v.$$

The following is an important result that is used to prove the asymptotic distribution of several estimator,

Proposition 1:

$$\lim_{T\to\infty}TE\left[\left(\bar{X}_T-\mu\right)\left(\bar{X}_T-\mu\right)'\right]=\sum_{v=-\infty}^{\infty}\Gamma_v$$

Proof: Consider

$$\begin{aligned}\sum_{v=-\infty}^{\infty}\Gamma_v-TE\left[\left(\bar{X}_T-\mu\right)\left(\bar{X}_T-\mu\right)'\right]&=\sum_{v=-\infty}^{\infty}\Gamma_v-\sum_{v=-(T-1)}^{T-1}\Gamma_v+\sum_{v=-(T-1)}^{T-1}\frac{|v|}{T}\Gamma_v\\&=\sum_{|v|\geq T}\Gamma_v+\sum_{v=-(T-1)}^{T-1}\frac{|v|}{T}\Gamma_v.\end{aligned}$$

The (i,j) element of the above expression can be written as

$$\sum_{|v|\geq T}\gamma_{ij}^{(v)}+\sum_{v=-(T-1)}^{T-1}\frac{|v|}{T}\gamma_{ij}^{(v)}.$$

We need to prove that the absolute value of this term converges to zero for each i, j. Consider

$$\left| \sum_{|v|\geq T} \gamma_{ij}^{(v)} + \sum_{v=-(T-1)}^{T-1} \frac{|v|}{T} \gamma_{ij}^{(v)} \right| \leq \sum_{|v|\geq T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=-(T-1)}^{T-1} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| \tag{1.19}$$

Absolute summability of $\{\Gamma_v\}$ means that for any $\varepsilon > 0$ there exist an index q such that

$$\sum_{|v|>q}^{\infty} \left| \gamma_{ij}^{(v)} \right| < \frac{\varepsilon}{2},$$

for otherwise the sum will not converge. Now choose $T - 1 > q$ and write

$$\begin{aligned} \sum_{v=-(T-1)}^{T-1} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| &= \sum_{v=-q}^{q} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=q+1}^{T-1} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=-(T-1)}^{-(q+1)} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| \\ &\leq \sum_{v=-q}^{q} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=q+1}^{T-1} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=-(T-1)}^{-(q+1)} \left| \gamma_{ij}^{(v)} \right| \end{aligned}$$

The inequality above uses that $|v|/T < 1$ for all $|v| = \{q+1, q+2, ..., T-1\}$.

Therefore, the right hand side of (1.19) can be written as

$$\begin{aligned} \sum_{|v|\geq T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=-(T-1)}^{T-1} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| &\leq \sum_{|v|\geq T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=-q}^{q} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=q+1}^{T-1} \left| \gamma_{ij}^{(v)} \right| + \sum_{v=-(T-1)}^{-(q+1)} \left| \gamma_{ij}^{(v)} \right| \\ &= \sum_{v=-q}^{q} \frac{|v|}{T} \left| \gamma_{ij}^{(v)} \right| + \sum_{|v|>q} \left| \gamma_{ij}^{(v)} \right| \\ &< \frac{1}{T} \sum_{v=-q}^{q} |v| \left| \gamma_{ij}^{(v)} \right| + \frac{\varepsilon}{2}. \end{aligned}$$

where the last inequality uses that q satisfies the inequality $\sum_{|v|>q} \left| \gamma_{ij}^{(v)} \right| < \varepsilon/2$. Putting together these inequalities gives

$$\left| \sum_{|v|\geq T} \gamma_{ij}^{(v)} + \sum_{v=-(T-1)}^{T-1} \frac{|v|}{T} \gamma_{ij}^{(v)} \right| < \frac{1}{T} \sum_{v=-q}^{q} |v| \left| \gamma_{ij}^{(v)} \right| + \frac{\varepsilon}{2}.$$

But the term $\sum_{v=-q}^{q} |v| \left| \gamma_{ij}^{(v)} \right|$ is a number that does not depend on T, so that $\frac{1}{T} \sum_{v=-q}^{q} |v| \left| \gamma_{ij}^{(v)} \right|$ can be made smaller than $\varepsilon/2$ for sufficiently large T. We thus conclude that, for any $\varepsilon > 0$, there is a T such that

$$\left| \sum_{|v|\geq T} \gamma_{ij}^{(v)} + \sum_{v=-(T-1)}^{T-1} \frac{|v|}{T} \gamma_{ij}^{(v)} \right| < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this proves that

$$\lim_{T\to\infty} TE\left[\left(\bar{X}_T - \mu \right) \left(\bar{X}_T - \mu \right)' \right] = \sum_{j=-\infty}^{\infty} \Gamma_j.$$

The above algebra is ~~tedious and boring~~ important as it shows the formula for the asymptotic covariance of the sample mean of a vector process. This is used, for example, for computing asymptotics of GMM with time dependent data or for computing robust standard errors of many estimators (e.g. OLS estimators where residuals could have autocorrelation and/or heteroskedasticity).

We now recall the two fundamental limit theorems: suppose that $X_1, X_2, ...$ are i.i.d. random variables with $E(X_t) = \mu$ and $E(X_t - \mu)^2 = \sigma^2 < \infty$. Then,

Law of large numbers (LLN): $\frac{1}{T}\sum_{t=1}^{T} X_t \to \mu$ (converges in probability, a.s., in L^2)

Central limit theorem (CLT): $\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} X_t - \mu\right) \Rightarrow N(0, \sigma^2)$ (converges in distribution)

The above results use independence and gives an idea of how quickly and in what sense the sample average converges to the population mean. In time series we typically don't have independence. There are, however, versions of the above theorems for dependent data. Let's consider how we prove the (weak) law of large number. For this we need the following results:

Markov inequality: Let $\phi(x) \geq 0$ be a non-decreasing function on R_+. Then, for any random variable $X \geq 0$ and constant $a > 0$,

$$\Pr(X \geq a) \leq \frac{E[\phi(X)]}{\phi(a)}.$$

Proof:

$$\begin{aligned} E[\phi(X)] &= \Pr(\phi(X) \geq \phi(a)) E[\phi(X) | \phi(X) \geq \phi(a)] + \Pr(\phi(X) < \phi(a)) E[\phi(X) | \phi(X) < \phi(a)] \\ &\geq \Pr(\phi(X) \geq \phi(a)) E[\phi(X) | \phi(X) \geq \phi(a)] \\ &\geq \Pr(\phi(X) \geq \phi(a)) \phi(a) \end{aligned}$$

where the first inequality uses that $E[\phi(X) | \phi(X) < \phi(a)] \geq 0$ and the second inequality uses $E[\phi(X) | \phi(X) \geq \phi(a)] \geq \phi(a)$. Therefore,

$$\Pr(\phi(X) \geq \phi(a)) \leq \frac{E[\phi(X)]}{\phi(a)}.$$

To finish the proof, note the following inclusion of events

$$\{\omega \in \Omega : X(\omega) \geq a\} \subseteq \{\omega \in \Omega : \phi(X(\omega)) \geq \phi(a)\}.$$

If $\phi(x)$ is strictly increasing, the two events are equal, but if $\phi(x)$ is constant over some range, the inclusion can be strict. Therefore, the set inclusion implies

$$\Pr(X \geq a) \leq \Pr(\phi(X) \geq \phi(a)) \leq \frac{E[\phi(X)]}{\phi(a)}. \square$$

Chebyshev inequality: For any random variable X and constant $a > 0$,

$$\Pr\left(|X - E[X]| \geq a\right) \leq \frac{Var(X)}{a^2}.$$

Proof: Let $Z = |X - E[X]|$ and $\phi(z) = z^2$. Now apply Markov's inequality.□

We now provide a proof of the weak law of large numbers with iid data.

Theorem (WLLN): Let $X_1, X_2, ...$ be *i.i.d.* random variables with $E[X_t] = \mu$ and uniformly bounded variance $E\left[(X_t - \mu)^2\right] \leq B < \infty$. Let $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$. Then, for any $\varepsilon > 0$,

$$\lim_{T\to\infty} \Pr\left[\left|\bar{X}_T - \mu\right| > \varepsilon\right] = 0.$$

Proof: Note that $E\left[\bar{X}_T\right] = \mu$ and

$$\begin{aligned} Var\left(\bar{X}_T\right) &= Var\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right) = \frac{1}{T^2} Var\left(\sum_{t=1}^{T} X_t\right) \\ &= \frac{1}{T^2}\sum_{t=1}^{T} Var(X_t) \leq \frac{TB}{T^2} = \frac{B}{T} \end{aligned}$$

where the third equality uses that X_t is iid and the inequality uses that the variance is bounded. Therefore, for any $\varepsilon > 0$, Chebyshev's inequality implies

$$\lim_{T\to\infty} \Pr\left[\left|\bar{X}_T - \mu\right| > \varepsilon\right] \leq \lim_{T\to\infty} \frac{Var\left(\bar{X}_T\right)}{\varepsilon^2} \leq \lim_{T\to\infty} \frac{B}{\varepsilon^2 T} = 0.\square$$

Note that we used that X_t is iid in two parts. First, for using that $E[X_t] = \mu$ for all t and, more importantly, for writing $Var\left(\sum_{t=1}^{T} X_t\right) = \sum_{t=1}^{T} Var(X_t)$.

What happens if the random variables are not i.i.d.? The variance of $\bar{X}_T$ is given by

$$Var\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} Cov(X_t, X_s).$$

A WLLN can be proved for covariance stationary processes by imposing restrictions on the autocovariances. In particular, suppose that $\{X_t\}$ is a covariance stationary process, so that $E(X_t) = \mu$ for all t and $Cov(X_s, X_t) = \gamma_{|s-t|}$ for all s, t with absolutely summable autocovariances, so that

$$\sum_{j=-\infty}^{\infty} |\gamma_j| = c < \infty.$$

Then,

$$\begin{aligned} T^2 Var\left(\bar{X}_T\right) &= \sum_{t=1}^{T}\sum_{s=1}^{T} Cov\left(X_t, X_s\right) = \sum_{t=1}^{T}\sum_{s=1}^{T} \gamma_{|t-s|} \\ &\leq \sum_{t=1}^{T}\sum_{s=1}^{T} \left|\gamma_{|t-s|}\right| \\ &\leq \sum_{t=1}^{T}\sum_{s=-\infty}^{\infty} \left|\gamma_{|t-s|}\right| \\ &\leq Tc \end{aligned}$$

where the last inequality uses absolute summability of autocovariances. It then follows that

$$Var\left(\bar{X}_T\right) \leq \frac{c}{T},$$

which is then used to prove the WLLN for dependent random variables under the assumption of stationarity and absolute summability of autocovariances.

A similar reasoning can be used to argue that the variance of the central limit theorem should change. Indeed, if the autocovariance function is absolutely summable, we showed above for the vector case that

$$\lim_{T\to\infty} T\ Var\left(\bar{X}_T\right) = \lim_{T\to\infty} Var\left(\sqrt{T}\ \bar{X}_T\right) = \sum_{j=-\infty}^{\infty} \gamma_j$$

where, $\gamma_j = Cov\left(X_t, X_{t-j}\right)$. This implies that, under an appropriate CLT,

$$\sqrt{T}\ \bar{X}_T = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t \to N\left(0, \sum_{j=-\infty}^{\infty} \gamma_j\right)$$

After reading the notes about time series in the frequency domain, please note that the asymptotic covariance of the sample mean is the spectral density of the process evaluated at frequency zero.

References:

[1] Cochrane, John (2005). Time Series for Macroeconomics and Finance. Manuscript.

[2] Gabel, R. A. and R. A. Roberts (1986). *Signals and Linear Systems*, 3rd Edition.

[3] Hamilton, James (1994). *Time Series Analysis*. Chapters 2 and 3

[4] Hansen, L. and T. Sargent (1991). Lecture notes on least squares predictions theory. In Rational Expectations Econometrics (Underground Classics in Economics). These notes provides detailed proofs with all the technicalities of most of the stuff we use. It can be found online here: http://www.tomsargent.com/books/TOMchpt.2.pdf

[5] Fuller, Wayne. (1995). *Introduction to Statistical Time Series*. Second Edition. Wiley Series in Probability and Statistics.

[6] Sargent, Thomas (1987). *Macroeconomic Theory*. Chapters 10 and 11.