Time Series Econometrics

Bo Hu

Institute of New Structural Economics
Peking University

May 10, 2021

Contents

1 Single Equation Linear Models

1.1 The Model

In the study of economics, we are frequently interested in the relationships between two sets of variables $\{y_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$. Often, these variables are non-deterministic in nature. Therefore, we usually treat them as (real) random variables or vectors on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.[1] A random variable or a random vector is a *measurable* function from the sample space Ω to $\mathbb{R}$ or $\mathbb{R}^k$.

Suppose the relationship is given by a linear form as

$$y_i = x_i'\beta + \varepsilon_i \tag{1.1}$$

where y_i is a random variable, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ is a k-dimensional random vector, $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$ is a fixed k-dimensional vector, called the parameter (or the regression coefficient), that represents the nature of this relationship, and ε_i, called the "error term", is a random variable that represents the "mismatched" part. We usually call y_i the *dependent variable*, the *response variable*, or the *regressand* and call x_i respectively the *independent variable(s)*, the *response variable(s)*, or the *regressor(s)*. Our goal is to use the data to learn about the relationship β, but what does β stand for?

We may always write $y_i = x_i'\beta^\circ + \varepsilon_i^\circ$ where β° is any vector different from β and ε° defined as $\varepsilon + x'(\beta - \beta^\circ)$ is the new "error term". Note that the new model and the old model are exactly in the same form! Without restrictions on the error term, we won't be able to have a clear meaning or interpretation of the coefficient β.

If we impose the restriction that $\mathbb{E}(\varepsilon_i|x_i) = 0$ for all i, then

$$\beta = \frac{\partial \mathbb{E}(y_i|x_i)}{\partial x_i},$$

or

$$\beta_j = \frac{\partial \mathbb{E}(y_i|x_i)}{\partial x_{ij}}, \qquad j = 1, 2, \ldots, k.$$

That is, β_j is the marginal change in the conditional expectation of y_i when the j-th component of x_i increases by one unit, holding the other components of x_i constant. In econometric terms, we say that β_j is the *partial effect* of x_{ij} on $\mathbb{E}(y_i|x_i)$, or simply the partial effect of x_{ij} on y_i.

The above discussion indicates that when we talk about an econometric model, for example, a linear model as discussed above, we not only mean the linear form of the relationship $y_i = x_i'\beta + \varepsilon_i$, but also the restrictions such as $\mathbb{E}(\varepsilon_i|x_i) = 0$ that we impose on the relationship. This restriction gives us the meaning of the "relationship" β, and is as essential as the form of the model.

[1] In this series of notes, if not mentioned otherwise, all random variables and random vectors are real. For details of the concept of probability spaces, measurable functions and the axiomatic foundations of probability theory, see Billingsley (1995, Chapter 1 and 2) or Shiryaev (1989, Chapter 2).

1.2 Identification

A fundamental question is whether such β is learnable at all, or in econometric terms, identifiable. To give a general definition of identification, suppose a data generating system is governed by a set of parameters θ and generates data $\{z_i\}_{i=1}^n$, whose joint distribution is denoted by $P_n(\theta)$. If the mapping $\theta \mapsto P_n(\theta)$ is invertible for some n, we say that θ is identifiable. In plain words, identification means that we cannot find two different sets of parameter values that generate data with the same joint distribution.

When θ is identifiable, the inverse of the mapping $\theta \mapsto P_n(\theta)$ is necessarily a function of the joint distribution, and therefore, a function of the data $\{z_i\}_{i=1}^n$. If we denote this function by $g_n(z_1, \cdots, z_n)$, then we say that θ is identified by $\theta = g_n(z_1, \cdots, z_n)$. Back to the linear model, when we say that the regression coefficient β is identified, we mean that β can be written as a function of some properties of the data $\{(x_i, y_i)\}_{i=1}^n$.

In this chapter we only focus on the situation where the data (x_i, y_i) are independent and identically distributed (iid). Our results here can be generalized to the case in which $\{(x_i, y_i)\}_{i=1}^n$ is not independent and identically distributed. We introduce corresponding tools in Chapter 2.

In this chapter we make the following assumptions.

Assumption 1.1. $\mathbb{E}y_i^2 < \infty$ and $\mathbb{E}\left\|x_i\right\|^2 < \infty$ where $\|\cdot\|$ denotes the Euclidean norm defined by $\|(a_1, a_2, \ldots, a_k)\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_k^2}$.

Assumption 1.2. $\operatorname{rank}(\mathbb{E}x_i x_i') = k$.

Assumption 1.3. $\mathbb{E}x_i \varepsilon_i = 0$.

Note that by Assumption 1.1, the expectation terms in Assumption 1.2 and 1.3 are well defined. If $\mathbb{E}\varepsilon = 0$, Assumption 1.3 is equivalent to that $\operatorname{Cov}(x_i, \varepsilon_i) = 0$. And it is obvious from (1.1) that as long as x contains the constant regressor, we may always assume that $\mathbb{E}\varepsilon_i = 0$.

Theorem 1.1. *Suppose Assumptions 1.1 - 1.3 hold, then β is identified.*

Proof. Pre-multiply every term in (1.1) by x_i and take expectation. By Assumption 1.3, we have

$$\mathbb{E}x_i y_i = \mathbb{E}x_i x_i' \beta.$$

From Assumption 1.2 it follows that

$$\beta = \left(\mathbb{E}x_i x_i'\right)^{-1} \mathbb{E}x_i y.$$

■

We jump ahead a little bit here by noting that the set of all random variables with finite second moments forms a Hilbert space[2], which may be thought of as a infinite dimensional generalization

[2]The relevant concepts will be introduced in Chapter 3.

of Euclidean spaces. Each random variable should be viewed as a vector in this space. The "dot product" of two random variables w_1 and w_2 is defined to be $\mathbb{E}w_1w_2$, and the two random variables are "orthogonal" if $\mathbb{E}w_1w_2 = 0$. Write

$$x_i'\beta = x_i' \left(\mathbb{E}x_ix_i'\right)^{-1} \mathbb{E}x_iy_i = \left(\mathbb{E}y_ix_i'\right) \left(\mathbb{E}x_ix_i'\right)^{-1} x_i := Py_i, \tag{1.2}$$

then P_x can be interpreted as the orthogonal projection of the regressand y_i onto the subspace spanned by the k regressors in x_i.

1.3 Estimation

The ordinary least square estimation and the maximum likelihood estimations are frequently used to estimate linear models.

1.3.1 The Ordinary Least Squares Estimation

The ordinary least squares (OLS) estimator $\hat{\beta}_{OLS}$ for the coefficient β in the model (1.1) is defined by:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2, \tag{1.3}$$

which yields

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{n} x_ix_i'\right)^{-1} \sum_{i=1}^{n} x_iy_i = \left(\frac{1}{n}\sum_{i=1}^{n} x_ix_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_iy_i\right). \tag{1.4}$$

Note that under Assumption 1.2, the probability that $\frac{1}{n}\sum_{i=1}^{n} x_ix_i'$ is invertible converges to 1. This is because by the weak law of large numbers $\frac{1}{n}\sum_{i=1}^{n} x_ix_i' \to_p \mathbb{E}x_ix_i'$, and the continuous mapping theorem implies that $\det(\frac{1}{n}\sum_{i=1}^{n} x_ix_i') \to_p \det(\mathbb{E}x_ix_i')$. By definition of convergence in probability, $\mathbb{P}\left(\left|\det\left(\frac{1}{n}\sum_{i=1}^{n} x_ix_i'\right) - \det\left(\mathbb{E}xx'\right)\right| < \epsilon\right) \to 1$ for any $\epsilon > 0$. Our result then follows by choosing $\epsilon < |\det(\mathbb{E}xx')|$. The second order sufficient condition for minimization holds under Assumption 1.2: the Hessian of the objective function $\frac{2}{n}\sum_{i=1}^{n} x_ix_i'$ is positive definite with probability approaching to 1. This follows by a similar argument as above, in which the continuous mapping theorem is applied to the eigenvalues of a matrix. Note that $\mathbb{E}x_ix_i'$ is always positive semi-definite as a quadratic form and the full rank condition implies that it is in fact positive definite.

It is obvious that $\hat{\beta}_{OLS}$ defined above is also the *sample analogue* estimator, or the *method of moment* estimator for β. It should be pointed out that by denoting the estimator as $\hat{\beta}_{OLS}$, we have suppressed the dependence of the estimator on the sample size n. In the following asymptotic analysis, the results are obtained for $n \to \infty$.

The next theorem establishes the consistency and asymptotic normality of the OLS estimator $\hat{\beta}_{OLS}$. An estimator is *consistent* if for any true value of β, the estimator converges to the true value in probability if the sample size n goes to infinity. An estimator is asymptotically normal if it *converges in distribution* to a normal random variable or vector after appropriate scaling.

Assumption 1.4. $\mathbb{E}y_i^4 < \infty, \mathbb{E}\left\|x_i\right\|^4 < \infty$.

Theorem 1.2. *Under Assumption 1.1 - 1.3, $\hat{\beta}_{OLS}$ is consistent. If Assumption 1.4 also holds, then*

$$\sqrt{n}\left(\hat{\beta}_{OLS} - \beta\right) \to_d \mathbb{N}\left(0, V\right),$$

where

$$V = \left(\mathbb{E}x_i x_i'\right)^{-1}\left(\mathbb{E}\varepsilon_i^2 x_i x_i'\right)\left(\mathbb{E}x_i x_i'\right)^{-1}.$$

Proof. Since

$$\hat{\beta}_{OLS} - \beta = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i \varepsilon_i\right), \tag{1.5}$$

consistency follows from that

$$\frac{1}{n}\sum_{i=1}^n x_i x_i' \to_p \mathbb{E}x_i x_i'$$

and that

$$\frac{1}{n}\sum_{i=1}^n x_i \varepsilon_i \to_p \mathbb{E}x_i\varepsilon_i = 0.$$

If in addition that Assumption 1.4 holds, then $\mathbb{E}\varepsilon^2 xx' < \infty$. The asymptotic normality follows immediately from the central limit theorem that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i \varepsilon_i \to_d \mathbb{N}(0, \mathbb{E}\varepsilon^2 xx')$$

and the Slutsky's theorem. ■

It is natural to estimate the asymptotic variance V of $\sqrt{n}(\hat{\beta}_{OLS} - \beta)$ by

$$\hat{V} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1},$$

where $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}_{OLS}$ is the residual. The estimator for the asymptotic variance is called the *heteroskedasticity-consistent* estimator by White (1980). Andrews (1991) further considers the case in which ε_i is not independent and studies the *heteroskedasticity and autocorrelation consistent (HAC)* estimation of variance matrix of parameter estimators. We shall encounter such estimators in Chapter 2 once we introduced concepts to handle dependence.

Theorem 1.3. *Under Assumption 1.1 - 1.4,*

$$\hat{V} \to_p V.$$

Proof. Write

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 x_i x_i' = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 x_i x_i' - \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i'(\hat{\beta}_{OLS}-\beta)x_i x_i' + \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{OLS}-\beta)' x_i x_i'(\hat{\beta}_{OLS}-\beta)x_i x_i'.$$

Under Assumption 1.4, we have that $\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|\,\|x_i\|^3 = O_p(1)$ and that $\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^4 = O_p(1)$. Since $\hat{\beta}_{OLS}-\beta = o_p(1)$, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 x_i x_i' = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 x_i x_i' + o_p(1) \to_p \mathbb{E}\varepsilon^2 xx'.$$

Consistency of $\hat{V}$ then follows immediately. ■

Write $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)'$ and let $\hat{V}_{jj}$ be the (j,j)-th entry of $\hat{V}$. We define the asymptotic *standard error* of $\hat{\beta}_j$ by

$$\text{s.e.}(\hat{\beta}_j) = \sqrt{\hat{V}_{jj}/n}.$$

The variance σ^2 of the error term ε_i could be estimated as the sample variance of the OLS residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Using similar techniques as in the proof of Theorem 1.3, we may easily show that under Assumption 1.1 - 1.4, $\hat{\sigma}^2 \to_p \sigma^2$.

In the case of homoskedasticity, i.e., when $\mathbb{E}(\varepsilon_i^2|x_i) = \sigma^2$, the variance matrix V in the asymptotic distribution reduces to $\sigma^2\left(\mathbb{E}x_i x_i'\right)^{-1}$, and we may estimate V by

$$\tilde{V} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\right)^{-1}.$$

If the true data process is heteroskedastic, but we treat it as homoskedastic and use $\tilde{V}$ instead of $\hat{V}$ to estimate the variance matrix V, then the estimation of V is not consistent, and inferences based on this estimator would be incorrect.

1.3.2 Projection Interpretation of the OLS Estimation

If we write

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

then we may write (1.1) as

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon},$$

and (1.4) as

$$\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Then we have that

$$\hat{\boldsymbol{y}} =: \boldsymbol{X}\hat{\beta} = P_{\boldsymbol{X}}\boldsymbol{y}$$

and

$$\hat{\boldsymbol{\varepsilon}} =: \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - P_{\boldsymbol{X}})\boldsymbol{y}$$

where $P_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the orthogonal projection in $\mathbb{R}^n$ onto the subspace $\mathcal{R}(\boldsymbol{X})$ spanned by the column vectors of $\boldsymbol{X}$. Note that $I - P_{\boldsymbol{X}}$ is the orthogonal projection onto the subspace $\mathcal{R}(\boldsymbol{X})^{\perp}$. That is, $I - P_{\boldsymbol{X}}$ is the orthogonal projection onto the subspace that is orthogonal to $\mathcal{R}(\boldsymbol{X})$.

We may partition the model (1.1) as

$$y_i = x_{i1}'\beta_1 + x_{i2}'\beta_2 + \varepsilon_i$$

where $x_i = (x_{i1}', x_{i2}')'$ and $\beta = (\beta_1', \beta_2')'$. If we write $\boldsymbol{X}_1$ as the matrix whose rows are x_{i1}', and $\boldsymbol{X}_2$ as the matrix whose rows are x_{i2}', then we may write the model as

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{\varepsilon}$$

and the fitted model as

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_2\hat{\beta}_2 + \hat{\boldsymbol{\varepsilon}}$$

where $\hat{\beta}_{OLS} = (\hat{\beta}_1', \hat{\beta}_2')'$ is partitioned accordingly. Pre-multiply both sides by $I - P_{\boldsymbol{X}_2}$, and notice that $\hat{\boldsymbol{\varepsilon}} \in \mathcal{R}(\boldsymbol{X})^{\perp} \subset \mathcal{R}(\boldsymbol{X}_2)^{\perp}$, we have that

$$(I - P_{\boldsymbol{X}_2})\boldsymbol{y} = (I - P_{\boldsymbol{X}_2})\boldsymbol{X}_1\hat{\beta}_1 + \hat{\boldsymbol{\varepsilon}}.$$

Also $\hat{\boldsymbol{\varepsilon}} \in \mathcal{R}(\boldsymbol{X})^{\perp} \subset \mathcal{R}(\boldsymbol{X}_1)^{\perp}$, then

$$\boldsymbol{X}_1'(I - P_{\boldsymbol{X}_2})\boldsymbol{y} = \boldsymbol{X}_1'(I - P_{\boldsymbol{X}_2})\boldsymbol{X}_1\hat{\beta}_1.$$

This implies that

$$\hat{\beta}_1 = \left(\boldsymbol{X}_1'(I - P_{\boldsymbol{X}_2})\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'(I - P_{\boldsymbol{X}_2})\boldsymbol{y},$$

or equivalently,

$$\begin{aligned}\hat{\beta}_1 = &\left[\sum_{i=1}^n x_{i1}x_{i1}' - \sum_{i=1}^n x_{i1}x_{i2}'\left(\sum_{i=1}^n x_{i2}x_{i2}'\right)^{-1}\sum_{i=1}^n x_{i2}x_{i1}'\right]^{-1} \\ &\bullet \left[\sum_{i=1}^n x_{i1}y_i - \sum_{i=1}^n x_{i1}x_{i2}'\left(\sum_{i=1}^n x_{i2}x_{i2}'\right)^{-1}\sum_{i=1}^n x_{i2}y_i\right].\end{aligned} \tag{1.6}$$

Note that this is the sample analogue estimator of

$$\left(\mathbb{E}[x_{1i}(I - P_{x_{2i}})x_{1i}']\right)^{-1}\mathbb{E}[x_{1i}(I - P_{x_{2i}})y_i],$$

where $P_{x_{ji}}$ is the orthogonal projection (in the Hilbert space of all random variables with finite second moments) onto the subspace spanned by random variables in $x_{ji}, j = 1, 2$.

1.3.3 Constrained Least Squares

Suppose that we have the constraint $R\beta = r$ on the parameter β where where R is a $q \times k$ matrix with rank $q \leq k$, and r is a q-dimensional vector. We want the estimated β, denoted by $\tilde{\beta}$, to satisfy the restriction $R\tilde{\beta} = r$. This constrained least squares estimator could be obtained by solving the following constrained minimization problem

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\ \text{s.t.} \quad & R\beta = r. \end{aligned}$$

One may solve this constrained optimization problem by any numerical method that is applicable. However, since the constraint is linear, it is straightforward to get the analytical solution, which illuminates the relationship between the constrained least squares estimator $\tilde{\beta}$ and the unconstrained OLS estimator $\hat{\beta}_{OLS}$:[3]

Theorem 1.4. *Suppose that Assumptions 1.1 and 1.2 hold. Then*

$$\tilde{\beta} = \hat{\beta}_{OLS} - \left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\right)^{-1}R'\left(R\left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\right)^{-1}R'\right)^{-1}\left(R\hat{\beta}_{OLS} - r\right).$$

Write $A_n = \left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\right)^{-1}$. Then under the constraint $R\beta = r$, we have

$$\tilde{\beta} - \beta = \left[I - A_n R'\left(RA_n R'\right)^{-1}R\right]\left(\hat{\beta}_{OLS} - \beta\right).$$

It is straightforward to obtain the asymptotic distribution of $\tilde{\beta}$ from Theorem 1.2. Also, we note here that $A_n R'\left(RA_n R'\right)^{-1}R$ is the non-orthogonal projection onto $\mathcal{R}(A_n R')$ along the direction of $\mathcal{R}(R')^{\perp}$, and $I - A_n R'\left(RA_n R'\right)^{-1}R$ is therefore the non-orthogonal projection onto $\mathcal{R}(R')^{\perp}$ along the direction of $\mathcal{R}(A_n R')$.

[3] The Lagrangian can be written as $\mathcal{L} = \frac{1}{2n}\sum(y - x_i'\beta)^2 - \lambda'(R\beta - r)$. The first order condition is then given by $\frac{1}{n}\sum x_i y_i - \frac{1}{n}\sum x_i x_i'\beta - R'\lambda = 0$. We first solve for β as an expression of R, λ and the data through this first order condition, noting that $\frac{1}{n}\sum x_i x_i'$ is invertible with probability converging to one. We then substitute it into the constraint $R\beta = r$ to solve for λ as an expression of R, r and the data. In the end we substitute the expression for λ back into the first order condition to get the optimal β.

1.3.4 Maximum Likelihood Estimation

The method of maximum likelihood has been one of the most popular methods in estimation since the advocacy by Fisher (1922). To state the general idea, let $x = \{x_1, x_2, \ldots, x_n\}$ be a set of random variables or random vectors which we believe to be generated from one of a class of models indexed by the parameter $\theta \in \Theta$. The density of a realization $x^\circ = \{x_1^\circ, x_2^\circ, \ldots, x_n^\circ\}$ of x is dependent on the value of θ and may be written as $f_\theta(x^\circ)$. The *likelihood function* $\mathcal{L}(\theta; x^\circ)$ of θ given the realization x° is the density $f_\theta(x^\circ)$ viewed as a function of θ. The *maximum likelihood estimator (MLE)* of θ is defined by

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ \mathcal{L}(\theta; x^\circ).$$

We may replace the likelihood function $\mathcal{L}(\theta; x^\circ;)$ with the *log-likelihood function* $\ell(\theta; x^\circ) = \ln \mathcal{L}(\theta; x^\circ;)$ in the definition of MLE since the log transformation is strictly increasing.

For the linear model, the likelihood is dependent on the distributional assumptions of the error term.

Assumption 1.5. $\varepsilon_t \sim$ iid $\mathbb{N}(0, \sigma^2)$, and the distribution of $x_1, x_2, \ldots, x_n$ does not depend on the parameters β and σ^2.

Let z and w be random variables or vectors. With a little abuse of notation, we use $f_\theta(z|w)$ to denote the conditional density of observing a realization (z°, w°) of the random element (z, w). We do similar things to density functions and likelihood functions. That is, although in principle z and w are random elements, when they appear in density or likelihood functions we treat them as labels, which represents the realizations of the corresponding random elements.

Under Assumption 1.5 and the model (1.1), the density function of $\{(x_i, y_i)\}_{i=1}^n$ is

$$\begin{aligned} f_{\beta,\sigma^2}(y_1, x_1, y_2, x_2, \ldots, y_n, x_n) &= \prod_{i=1}^n f_{\beta,\sigma^2}(y_i, x_i) \\ &= \prod_{i=1}^n f_{\beta,\sigma^2}(y_i|x_i) f_{\beta,\sigma^2}(x_i) \\ &= \left[\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right)\right] \prod_{i=1}^n f_{\beta,\sigma^2}(x_i). \end{aligned}$$

The likelihood, which is viewed as a function of β, is therefore proportional to the term in the square bracket in the last line above, and the log-likelihood therefore is

$$\ell(\beta, \sigma^2) = C - \frac{n}{2} \ln \sigma^2 - \sum_{i=1}^n \frac{(y_i - x_i'\beta)^2}{2\sigma^2},$$

where C is a constant independent of β and σ^2.

The maximum likelihood estimator is obtained by finding values for β and σ^2 that maximize the log-likelihood function. Note that the value of σ^2 does not affect the choice of the optimal β

in this maximization problem. It is quite straightforward to show that $\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$. Therefore, MLE and OLS are equivalent under Assumption 1.5. The results in Theorem 1.2 also hold for $\hat{\beta}_{MLE}$.

Also, from the maximization problem we have

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{MLE})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2,$$

which is a consistent estimator of σ^2.

Similarly, in view of the relationship between the OLS objective function and the MLE objective function, one concludes immediately the ML estimator of β under the linear constraint $R\beta = r$ is also the same as the OLS estimator under the constraint.

1.3.5 Quasi-Maximum Likelihood Estimation

In our settings above, the OLS estimator of β is consistent, regardless of the distribution of the error term. This implies that even if the distribution of ε_i is not normal, we may still maximize the "false" normal likelihood function, and the resulting "false" ML estimator for β is still consistent, since this "false" ML estimator is the same as the OLS estimator.

Estimation based on a likelihood that is different from the true one is called *quasi-maximum likelihood estimation*, or *pseudo-maximum likelihood estimation*. Whether the likelihood is the true one depends on whether the econometric model is correctly-specified. The specification involves both the form of the model (e.g., whether the relationship between the dependent and independent variables is linear) and the underlying distributional assumptions (e.g., whether the variables are jointly normally distributed.)

Usually, if the "false" or "imprecise" likelihood captures the essence of the model, and is not too far away from the true likelihood, the quasi-maximum likelihood estimator could still be consistent and asymptotically normal. However, it is less efficient than the maximum likelihood estimator, meaning that it has a larger variance than the maximum likelihood estimator. See White (1982) for reference. We shall also return to this issue in a more general setting in Chapter 6.

1.4 Statistical Inferences

Since the OLS and MLE estimator of β is the same in our settings, in this section we denote both of them by $\hat{\beta}$ for the unrestricted estimator.

1.4.1 Student's t-test

Suppose that we want to test the null hypothesis $H_0 : \beta_j = \beta_{j0}$ against the alternative $H_1 : \beta_j \neq \beta_{j0}$, where β_{j0} is a given real number. We define the t-test statistic for H_0 by

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{\text{s.e.}(\hat{\beta}_j)}.$$

The following theorem, which gives the asymptotic distribution of t, follows immediately from Theorem 1.2.

Theorem 1.5. *Suppose that Assumptions 1.1 - 1.4 hold. Under H_0,*

$$t \to_d \mathbb{N}(0, 1).$$

It is known that t-test based on the HAC standard error estimator over rejects when the sample size is small.[4]

1.4.2 The Wald Test

Suppose that we want to test the null hypothesis $H_0 : R\beta = r$ against the alternative $H_1 : R\beta \neq r$ where R is a $q \times k$ matrix with rank $q \leq k$, and r is a q-dimensional vector. The Wald test statistic is defined as

$$W = n(R\hat{\beta} - r)' \left(R\hat{V}R'\right)^{-1} (R\hat{\beta} - r).$$

Theorem 1.6. *Suppose that Assumptions 1.1 - 1.4 hold. Under H_0,*

$$W \to_d \chi^2_q.$$

Proof. The result follows immediately from that under H_0,

$$\sqrt{n}(R\hat{\beta} - r) = \sqrt{n}R(\hat{\beta} - \beta) \Rightarrow \mathbb{N}(0, RVR').$$

■

The Wald test could be easily generalized for nonlinear restrictions. See, e.g., Newey and McFadden (1994, Section 9). Aside from the Wald test, the Lagrange multiplier test and the likelihood ratio test, which we shall not elaborate on at this moment, are popular asymptotic tests that have been carefully studied and are frequently used. See Wooldridge (2010, Chapter 12 and 13).

[4] Hausman and Palmer (2012) states that "it is not uncommon for the actual size of the test to be 0.15 when the nominal size is the usual 0.05." For more discussions on this issue and methods (including bootstrap) to obtain the correct size of the test, see, e.g., Hausman and Palmer (2012) and MacKinnon (2013).

1.5 Endogeneity and Instrumental Variables Estimation

When Assumption 1.3 fails, that is, when $\mathbb{E}x_i\varepsilon_i \neq 0$, we say that x_i is *endogenous*. It is obvious from the proof of Theorem 1.2 that in the presence of endogeneity, the OLS estimator is not consistent any more. Endogeneity usually arises because of the existence of omitted variables, measurable errors, and/or simultaneity. For detailed discussions of the three situations, see Wooldridge (2010, Chapter 4 and 5).

Instrumental variables estimation could be used to treat endogeneity. Suppose that in our model (1.1) we have that Assumptions 1.1 and 1.2 hold but Assumption 1.3 fails. Suppose as before that x_i has dimension k. Suppose that there exists an l dimensional iid random vectors $\{z_i\}$ such that the following assumptions hold:

Assumption 1.6. $\mathbb{E}\left\|z_i\right\|^2 < \infty, \mathbb{E}z_i\varepsilon_i = 0$, $\operatorname{rank}(\mathbb{E}z_iz_i') = l$ and $\operatorname{rank}(\mathbb{E}z_ix_i') = k$.

Explanatory variables in such z_i are called *instrument variables (IV)*, or simply, *instruments*. The condition $\mathbb{E}z_i\varepsilon_i$ requires that the instruments are not correlated with the error term. The condition $\operatorname{rank}(\mathbb{E}z_ix_i') = k$, loosely speaking, requires that the instruments should be correlated with x_i, the endogenous explanatory variables. As one will see soon, this condition is crucial for identification of β. Note that for Assumption 1.6 to hold, it is necessary that $l > k$. If a explanatory variable does not correlate with ε_i, that is, if it is not endogenous, then it is obvious that it could serve as an instrument. Then $l > k$ requires that besides the exogenous explanatory variables, we should have at least as many extra instruments as the number of endogenous explanatory variables.

The following theorem establishes the identification of β.

Theorem 1.7. *Suppose that Assumptions 1.1 and 1.6 hold. Then β is identified.*

Proof. Let P_{z_i} be the orthogonal projection in the Hilbert space of finite second moment random variables onto the space spanned by the random variables in z_i. Premultiply both sides of (1.1) by P_{z_i}. Since $\mathbb{E}z_i\varepsilon_i = 0$, by (1.2) we have that

$$P_{z_i}y_i = P_{z_i}x_i'\beta.$$

Then β could be identified by

$$\beta = \left(\mathbb{E}x_iP_{z_i}x_i'\right)^{-1}\mathbb{E}x_iP_{z_i}y_i$$

since the assumptions that $\mathbb{E}z_iz_i'$ and $\mathbb{E}z_ix_i'$ are full rank implies that $\mathbb{E}x_iP_{z_i}x_i'$ is full rank, and therefore invertible. ■

It is then immediate to propose the IV estimator based on the sample analogue of the above identifying relationship:

$$\hat{\beta}_{IV} = \left[\left(\sum_{i=1}^n x_iz_i'\right)\left(\sum_{i=1}^n z_iz_i'\right)^{-1}\left(\sum_{i=1}^n z_ix_i'\right)\right]^{-1}\left(\sum_{i=1}^n x_iz_i'\right)\left(\sum_{i=1}^n z_iz_i'\right)^{-1}\left(\sum_{i=1}^n z_iy_i\right).$$

Note that the IV estimator could be obtained by the following two-step least square procedure:

1. Run OLS regression of each random variables in x on z. Collect the fitted values (in the original order) in the vector $\hat{x}$.
2. Run OLS regression of y on $\hat{x}$.

Similarly as in the proof of Theorem 1.2 and 1.3, it is straightforward to establish the consistency and asymptotic normality of the IV estimator.

Theorem 1.8. *Suppose that Assumptions 1.1 and 1.6 hold, then $\hat{\beta}_{IV}$ is consistent. If in addition Assumption 1.4 holds, then*

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) \to_d \mathbb{N}\left(0, A^{-1}BA^{-1}\right),$$

where

$$A = (\mathbb{E}xz')(\mathbb{E}zz')^{-1}(\mathbb{E}zx')$$

and

$$B = (\mathbb{E}xz')(\mathbb{E}zz')^{-1}(\mathbb{E}\varepsilon^2 xx')(\mathbb{E}zz')^{-1}(\mathbb{E}zx').$$

A and B could be consistently estimated using sample analogue with ε_i replaced by $\hat{\varepsilon}_i$, the IV estimation residuals.

In practice, it could be difficult to find good instruments. Often, the requirement that the instruments do not correlate with the error term contradicts with the requirement that the instruments correlate with the endogenous explanatory variables. Even if one can find instruments that satisfy these requirements, the instruments could be *weak* in the sense that the correlation between the endogenous explanatory variables and the instruments is small. Such weak instruments may lead to large asymptotic variance of the IV estimator, making inference useless (nothing will be significant). In that case, one must choose between an inconsistent, but small-variance OLS estimator and a consistent, but large-variance IV estimator. For more discussion of issues with instrumental variables estimation, see Wooldridge (2010, p. 107-112).

2 Weakly Stationary Time Series

Starting from this chapter, we look at samples that are not independent, but are temporally correlated.

2.1 Stationary Time Series

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space.

Definition 2.1. A *stochastic process* $X = (X_t)_{t\in T}$ is a collection of random variables or vectors indexed by $t \in T$. In time series econometrics, T usually has the interpretation of a set of time.

If T has a finite or countable number of elements, such as the set of integers, the stochastic process is said to be in discrete time. If T is some interval of the real line, then the stochastic process is said to be in continuous time. A series of data from a discrete-time stochastic process is usually called a time series. In this course, we shall mainly deal with discrete-time stochastic processes.

We may take a stochastic process $X = X(t, \omega), t \in T, \omega \in \Omega$ as a function of two arguments. $X(t, \cdot)$ is simply X_t, the random variable (vector) at time t. $X(\cdot, \omega)$ is a *sample path*, which is the trajectory of this stochastic process for a particular realization.

Definition 2.2. A time series $(X_t)_{t\in\mathbb{Z}}$ is said to be *strictly stationary* if the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is the same as the joint distribution of $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$ for all $k \in \mathbb{Z}_+, h \in \mathbb{Z}$ and $t_1, \ldots, t_k \in \mathbb{Z}$.

The above definition is equivalent to that the joint distribution of $(X_1, X_2, \cdots, X_k)$ is the same as the joint distribution of $(X_{1+h}, X_{2+h}, \cdots, X_{k+h})$ for all $k \in \mathbb{Z}_+$ and $h \in \mathbb{Z}$.

Definition 2.3. For a time series $(X_t)_{t\in\mathbb{Z}}$ such that $\mathrm{Var}(X_t) < \infty$ for all t, define its *autocovariance function* $\gamma(\cdot, \cdot)$ as

$$\gamma(t, s) = \mathrm{Cov}(X_t, X_s) = \mathbb{E}(X_t - \mathbb{E}X_t)(X_s - \mathbb{E}X_s)'.$$

Definition 2.4. A time series $(X_t)_{t\in\mathbb{Z}}$ is said to be *weakly stationary* if for any $t, s, h \in \mathbb{Z}$, $\mathbb{E}X_t = \mathbb{E}X_s$, and $\gamma(t, s) = \gamma(t + h, s + h)$.

Note that the definition of a weakly stationary time series implicitly requires the existence of the first two moments of the random variables (vectors) in the time series. The value of the autocovariance function $\gamma(t, s)$ of such a time series depends only on the time difference $t - s$ but not on the individual values of t and s. Therefore, for a weakly stationary time series, we may redefine its autocovariance function as a function of the time lag by

$$\gamma(k) = \gamma(k, 0) = \mathbb{E}(X_t - \mathbb{E}X_t)(X_{t-k} - \mathbb{E}X_{t-k})'.$$

It is easy to see that $\gamma(-k) = \gamma(k)'$, where A' denotes the transpose of A. When X_t is a (scalar) random variable, $\gamma(k) = \gamma(-k)$.

We may also define the *autocorrelation functions* $\rho(k)$ of a weakly stationary time series with autocovariance function $\gamma(k)$ by

$$\rho_{ij}(k) = \frac{\gamma_{ij}(k)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}},$$

where A_{ij} denotes the (i, j)-th entry if A is a matrix. It is apparent that $|\rho_{ij}(k)| \leq 1$ for all i, j and k, $\rho_{ii}(0) = 1$ for all i, and $\rho(-k) = \rho(k)'$.

Note that neither strict stationarity nor weak stationarity implies each other. A strictly stationary time series can fail to be weakly stationary if it does not have well defined first or second moment. However, for a strictly stationary time series with well defined variance (and therefore mean), it is weakly stationary. On the other hand, it is obvious that weak stationarity does not imply strict stationarity since weak stationarity does not say anything about moments higher than the second. However, in the case where the first two moments completely determines the distribution, weak stationarity implies strict stationarity. For example, strict stationary is equivalent to weak stationary for Gaussian time series.

Definition 2.5. A time series $(X_t)_{t\in\mathbb{Z}}$ is called a *Gaussian time series* if the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is normal for any choice of $t_1, t_2, \ldots, t_k \in \mathbb{Z}$.

Here we provide a note about the joint normal distributions. If $(X_1, X_2, \ldots, X_k)$ is jointly normal, then X_i is normal for each $i \in 1, \ldots, k$. However, if each X_i is normal, $(X_1, \ldots, X_k)$ is not necessarily normal. If the X_i's are mutually independent and each X_i is normal, then $(X_1, \ldots, X_k)$ is normal. As a consequence, if $(X_t)_{t\in\mathbb{Z}}$ is a sequence of independent random variables, and each X_t is normal, then $(X_t)_{t\in\mathbb{Z}}$ is a Gaussian time series.

An important class of weakly stationary time series is the mean-zero serially uncorrelated weakly stationary time series, which are called white noise and serve as the building block of many other classes of time series.

Definition 2.6. A time series $(\varepsilon_t)_{t\in\mathbb{Z}}$ is called a *white noise* process if for any $t, s \in \mathbb{Z}$, $\mathbb{E}\varepsilon_t = 0$, and

$$\mathbb{E}(\varepsilon_t \varepsilon_s') = \begin{cases} \Sigma, & t = s, \\ 0, & t \neq s. \end{cases}$$

A Gaussian white noise process is a white noise process who is also a Gaussian time series. Each ε_t in a Gaussian white noise process is normally distributed, and is independent of any other ε's in the series.

At the end of this section, we give a property of the covariance function.

Theorem 2.7. *Let $\gamma : \mathbb{Z} \to M(n)$ be a mapping from the set of all integers to the set of all n-dimensional square matrices. Then γ is the autocovariance function of a weakly stationary time series if and only if $\gamma(k) = \gamma(-k)'$ for any $k \in \mathbb{Z}$ and $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i'\gamma(t_i - t_j)a_j \geq 0$ for any $t_1, \ldots, t_n \in \mathbb{Z}, a_1, \ldots, a_n \in \mathbb{R}^n$ and $n \in \mathbb{N}$.*

Proof. Necessity. Let γ be the autocovariance function of a weakly stationary time series $(X_t)_{t\in\mathbb{Z}}$. Without loss of generality assume that $\mathbb{E}X_t = 0$. Obviously $\gamma(k) = \gamma(-k)'$. For any $a_1, \ldots, a_n \in \mathbb{R}^n$, $\sum_{i=1}^n \sum_{j=1}^n a_i'\gamma(t_i - t_j)a_j = \mathrm{Var}(\sum_{i=1}^n a_i X_{t_i}) \geq 0$.

Sufficiency. A strictly stationary Gaussian time series with autocovariance function γ that satisfies the conditions could be constructed by the Kolmogorov's existence theorem and the fact that the distribution of Gaussian random vectors are fully determined by their means and covariance matrices. See, e.g., Brockwell and Davis (1991, Theorem 1.5.1). ■

2.2 Some Asymptotic Results

We give some general results that will be useful in asymptotic analysis in time series setting. This exposition here mainly follows White (2001, Chapter 3, 5) and Shiryaev (1989).

2.2.1 Law of Large Numbers

We start from the simplest case.

Theorem 2.8. *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent and identically distributed random vectors with $\mathbb{E}\,|\xi_1| < \infty$. Let $\mu = \mathbb{E}\xi_1$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} \xi_t \to_{a.s.} \mu.$$

In fact, the conclusion of the theorem continues to hold if $\mathbb{E}\xi_1$ exists but is not necessarily finite.

Proof. See, e.g., White (2001, p. 32) and Shiryaev (1989, p. 391). ■

We now relax the identically distributed assumption.

Theorem 2.9. *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with $\mu_t = \mathbb{E}\xi_t$. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be positive, even and continuous such that $\frac{\varphi(x)}{|x|}$ is increasing and $\frac{\varphi(x)}{x^2}$ is decreasing on $\mathbb{R}^+$. If there exists a sequence $0 < a_t \uparrow \infty$ such that*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}\varphi(\xi_t - \mu_t)}{\varphi(a_t)} < \infty,$$

then

$$\sum_{t=1}^{T} \frac{\xi_t - \mu_t}{a_t}$$

converges almost surely. By Kronecker's Lemma, this implies that

$$\frac{1}{a_T}\sum_{t=1}^{T}(\xi_t - \mu_t) \to_{a.s.} 0.$$

Proof. See, e.g., Chung (2001, p. 129-132). ■

Corollary 2.10. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random vectors with $\mu_t = \mathbb{E}\xi_t$. If there exists some $\delta > 0$ such that

$$\sup_t \mathbb{E}\left|\xi_t - \mu_t\right|^{1+\delta} < \infty,$$

then

$$\frac{1}{T}\sum_{t=1}^{T}(\xi_t - \mu_t) \to_{a.s.} 0.$$

Proof. This is a corollary of Theorem 2.9 by taking $\varphi(x) = |x|^{1+\delta}$ and $a_t = t$. ■

The condition in the above theorem could be replaced by $\sup_t \mathbb{E}\left|\xi_t\right|^{1+\delta} < \infty$. In fact, by the c_r inequality we have $\mathbb{E}\left|\xi_t - \mu_t\right|^{1+\delta} \leq 2^{\delta}(\mathbb{E}\left|\xi_t\right|^{1+\delta} + |\mu_t|^{1+\delta})$ and the boundedness of $|\mu_t|$ is due to Jensen's inequality.

Remarks 2.11. The convergence rates in Theorem 2.9 can be sharpened if we choose a_T properly. For example, let $s_T^2 = \sum_{t=1}^{T}\mathrm{Var}(\xi_t) \to \infty$. Then by the Abel-Dini Theorem (Hildebrandt, 1942, Theorem Ia), we have $\sum_{t=1}^{T}\frac{\mathbb{E}(\xi_t-\mu_t)^2}{s_T^2(\ln s_T^2)^{1+2\epsilon}} < \infty$ for any $\epsilon > 0$. Then by the above theorem, we have $\frac{1}{s_T^2(\ln s_T)^{1/2+\epsilon}}\sum_{t=1}^{T}(\xi_t - \mu_t) \to_{a.s.} 0$ for any $\epsilon > 0$. For the case when ξ_t is independent and has the same variance, we have that $\frac{1}{\sqrt{T}(\ln\sqrt{T})^{1/2+\epsilon}}\sum_{t=1}^{T}(\xi_t-\mu_t) \to_{a.s.} 0$. The denominator has a divergence rate a little bit faster than $T^{1/2}$, but slower than $T^{1/2+\delta}$ for any $\delta > 0$.

We may compare the results with the Central Limit Theorem, which will be introduced in the next section. Under certain regularity conditions, the central limit theorem states that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\xi_t-\mu_t) \to_d \mathbb{N}(0,\sigma^2)$ for some $\sigma^2 > 0$. Then we obviously have that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\xi_t - \mu_t) \not\to_{a.s.} 0$. Now the questions is, is there a cutting-edge rate between $\sqrt{T}$ and $\sqrt{T}(\ln\sqrt{T})^{1/2+\epsilon}$ that determines convergence and non-convergence? The law of iterated logarithm gives an affirmative answer. It says that for an iid sequence $\{\xi_t\}$ with finite variance, $\limsup_T \frac{1}{\sqrt{2T\ln\ln T}}\sum_{t=1}^{T}(\xi_t - \mu_t) = \sigma$ a.s..

More general laws of iterated logarithm for the independent but heterogeneous case are given in Wittmann (1985) and Wittmann (1987). We state the result in Wittmann (1985) here.

Theorem 2.12. *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with $\mu_t = \mathbb{E}\xi_t$ and $Var(\xi_t) < \infty$. Let $s_T^2 = \sum_{t=1}^{T} Var(\xi_t) \to \infty$. If $\limsup_T \frac{s_{T+1}}{s_T} < \infty$ and*

$$\sum_{t=1}^{\infty}\frac{\mathbb{E}\left|\xi_t - \mu_t\right|^p}{\left(2s_T^2\ln\ln_+ s_T^2\right)^{p/2}} < \infty$$

for some $2 < p \leq 3$ where $\ln_+ x = \max\{e, \ln x\}$, then

$$\limsup_{T\to\infty}\frac{1}{\sqrt{2s_T^2\ln\ln_+ s_T^2}}\sum_{t=1}^{T}(\xi_t - \mu_t) = 1 \quad a.s..$$

We now turn to the dependent and identically distributed case. Identical distribution, or even strict stationarity, is not sufficient for a law of large numbers to hold. To give an example, let $\{x_t\}$ be

an iid sequence of $\mathbb{N}(0,1)$ random variables, and y be a $\mathbb{N}(0,1)$ random variable independent of $\{x_t\}$. Define $\xi_t = x_t + y$. Then $\{\xi_t\}$ is strictly stationary, and $\mathbb{E}\xi_t = 0$. However, $\frac{1}{T}\sum_{t=1}^{T} \xi_t \to_{a.s.} y$, which is random. The reason why the law of large numbers fail here is that the sequence $\{\xi_t\}$, although strictly stationary, exhibits too much temporal dependence. In particular, the dependence of x_t and x_{t+k} does not die out as k grows big.

One way to restrict the temporal dependence and obtain a law of large numbers is through the concept of ergodicity. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A measurable transformation $S : \Omega \to \Omega$ is called measure-preserving if for every $A \in \mathcal{F}$, $\mathbb{P}(S^{-1}A) = \mathbb{P}(A)$. Pointcaré's recurrence theorem states that if S is a measure-preserving transformation and $A \in \mathcal{F}$, then for almost every $\omega \in \Omega$, $S^n\omega \in A$ for infinitely many $n \geq 1$. To gain some intuition, take Ω as the collection of all molecules in a glass of wine, where each molecule is represented by its position, or a vector ω in $\mathbb{R}^3$, and $\mathbb{P}$ be the measure of volume, where the volume of all liquid in the glass is normalized to be one. Let S represent the action of stirring the wine using a stick. Then $S\omega$ gives the position after one stir of a molecule originally at ω. The action of stirring is a measure-preserving transformation since it changes the location of the molecules in the wine, but it does not affect their volumes. The Pointcaré's recurrence theorem says that if we continue to stir the wine, then any particular molecule in a volume of wine will revisit that volume again and again.

Given a measure-preserving transformation S, a set $A \in \mathcal{F}$ is called invariant if $S^{-1}A = A$. A measure-preserving transformation S is called ergodic if every invariant set has measure zero or one. In our stirring wine example, if the stirring action is ergodic, it means that for any volume of the wine, as long as it is not the whole volume, some of the molecules in that volume will move out of that volume after one stir; No part of the glass of wine can be "autonomous" if we stir. Given that molecules will move out of their old areas and come back again and again, it is expected that molecules will move here and there. Actually, ergodicity implies that every molecule actually will visit "everywhere" in the liquid if keep stirring forever.[1] Each molecule will run through the whole volume of the liquid. Not surprisingly, we have the following theorem.

Theorem 2.13. *Let S be a measure-preserving transformation and $\xi = \xi(\omega)$ be a random variable with $\mathbb{E}|\xi| < \infty$. If S is ergodic, then*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \xi(S^t\omega) = \mathbb{E}\xi \quad a.s..$$

Suppose a cup of wine starts from an initial state in which the liquid is separated in two layers that the bottom layer is pure water (80%), and the top layer is pure alcohol (20%). If we take $\xi(w)$ to be the indicator function that takes value one when the molecule at ω is an alcohol molecule and zero when the molecule at ω is a water molecule in the initial state of the wine, the above theorem implies that as the stirring goes on, the probability of any molecule being in the top layer will approach 20%, regardless of its type. After sufficient stirring, alcohol and water are mixed

[1] For the exact meaning of "everywhere", see Halmos (2006, p. 26, Lemma).

well. That is, the state of the wine after sufficient stirring is irrelevant of its initial state.

A measure-preserving transformation is ergodic if and only if it is weak mixing, i.e., if for any $A, B \in \mathcal{F}$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}(A \cap S^{-t}B) = \mathbb{P}(A)\mathbb{P}(B).$$

We may think $S^{-t}B$ as being the event B shifted t periods back into the past. Ergodicity then requires that for any A and B, A and $S^{-t}B$ should be independent on average in the limit. Obviously, weak mixing is implied by a stronger condition: A measure-preserving transformation S is mixing (in the ergodic-theoretic sense) if for all $A, B \in \mathcal{F}$, $\lim_{t\to\infty} \mathbb{P}(A \cap S^{-t}B) = \mathbb{P}(A)\mathbb{P}(B)$.

For any random variable ξ on $(\Omega, \mathcal{F}, \mathbb{P})$, given a measure-preserving transformation S, we define a sequence $\xi_1, \xi_2, \ldots$ of random variables by $\xi_i(\omega) = \xi(S^i\omega)$. Then it is easy to see that $\xi_1, \xi_2, \ldots$ is a strictly stationary sequence. Conversely, let $\xi_1, \xi_2, \ldots$ be a strictly stationary sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then we can always find $\xi_1', \xi_2', \ldots$ on a probability space $(\Omega', \mathcal{F}', \mathbb{P})$ where $\Omega' = \mathbb{R}^\infty$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^\infty)$ and $\mathbb{P}'$ the distribution of $\{\xi\}$ and a measure-preserving transformation S such that $(\xi_1', \xi_2', \ldots)$ have the same distribution as $(\xi_1, \xi_2, \ldots)$ and that $\xi_{i+1}'(\omega) = \xi_i'(S\omega)$. For details of the proofs, see, e.g., Shiryaev (1989, p. 405). By the above argument, we have associated any strictly stationary time series with a measure-preserving transformation.

Now let $\{\xi_t\}$ be a strictly stationary time series on $(\Omega, \mathcal{F}, \mathbb{P})$, and S the associated measure-preserving transformation as constructed above. Then $\{\xi_t\}$ is called ergodic if its associated measure-preserving transformation is ergodic. Equivalently, if we define a set $A \in \mathcal{F}$ to be invariant with respect to $\{\xi\}$ if there is a set $B \in \mathbb{R}^\infty$ such that for all $n \geq 1$, $A = \{\omega \in \Omega : (\xi_n, \xi_{n+1}, \ldots) \in B\}$, then $\{\xi_t\}$ is ergodic if each of its invariant sets has measure zero or one. For details, see Shiryaev (1989, p. 412-413). Now we have

Theorem 2.14. *Let $\xi_1, \xi_2, \ldots$ be strictly stationary and ergodic with $\mathbb{E}|\xi_1| < \infty$. Then*

$$\frac{1}{T} \sum_{t=1}^{T} \xi_t \to_{a.s.} \mathbb{E}\xi_1.$$

Proof. See, e.g., White (2001, p. 44). ■

We may construct new strictly stationary and ergodic sequences from existing strictly stationary and ergodic sequences.

Theorem 2.15. *Suppose $\xi_1, \xi_2, \ldots$ is a strictly stationary and ergodic sequence. If $\{\eta_t\}$ is a sequence defined by $\eta_t = f(\ldots, \xi_{t-1}, \xi_t, \xi_{t+1}, \ldots)$ where f is a measurable function into $\mathbb{R}^k$, then $\{\eta_t\}$ is strictly stationary and ergodic.*

Proof. See, e.g., White (2001, p. 44). ■

Now we look at general dependent heterogeneous sequences. We first introduce some more detailed concepts on mixing, which implies ergodicity. The first of the strong mixing conditions

is introduced in Rosenblatt (1956). For a sequence of random elements $\{\xi_t\}$, we let $\mathcal{F}_{-\infty}^t = \sigma(\ldots, \xi_{t-1}, \xi_t)$ be the σ-algebra generated by the sequence up to time t, and $\mathcal{F}_t^\infty = \sigma(\xi_t, \xi_{t+1}, \ldots)$ be the σ-algebra generated by the sequence from time t on. Now we define the mixing coefficients

$$\alpha(k) = \sup_t \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+k}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

$$\phi(k) = \sup_t \sup_{\substack{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+k}^\infty, \\ \mathbb{P}(A)>0}} |\mathbb{P}(B|A) - \mathbb{P}(B)|,$$

$$\psi(k) = \sup_t \sup_{\substack{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+k}^\infty, \\ \mathbb{P}(A)>0, \mathbb{P}(B)>0}} \left| \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)\mathbb{P}(B)} - 1 \right|$$

and

$$\rho(k) = \sup_t \sup_{\substack{f \in L^2_{\mathbb{R}}(\mathcal{F}_{-\infty}^t), \\ g \in L^2_{\mathbb{R}}(\mathcal{F}_{t+k}^\infty)}} |\mathrm{Corr}(f, g)|,$$

where $L^2_{\mathbb{R}}(\mathcal{A})$ is the space of (equivalent classes) of square integrable, $\mathcal{A}$-measurable real-valued random variables. We also define

$$\beta(k) = \sup_t \sup_{\substack{\{A_1, \ldots, A_I\} \in \tau(\mathcal{F}_{-\infty}^t), \\ \{B_1, \ldots, B_J\} \in \tau(\mathcal{F}_{t+k}^\infty)}} \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$$

where $\tau(\mathcal{A})$ is the collection of all finite $\mathcal{A}$-measurable partitions of Ω. These coefficients measure the dependence between events separated by at least k periods. If $\alpha(k), \phi(k), \psi(k)$, $\rho(k)$ or $\beta(k)$ converges to zero as $k \to \infty$, the process $\{\xi_t\}$ is called α-mixing (strong mixing), ϕ-mixing (uniform mixing), ψ-mixing, ρ-mixing, or β-mixing (absolutely regular), respectively. If $\alpha(k) = O(k^{-a-\epsilon})$ for some $\epsilon > 0$, then we say that α is of size $-a$. Similarly, we can define sizes of the other mixing coefficients.

It has been well established in the literature of strong mixing conditions (see, e.g., Bradley (2005)) that

$$2\alpha(k) \leq \beta(k) \leq \phi(k) \leq \frac{1}{2}\psi(k),$$

and

$$4\alpha(k) \leq \rho(k) \leq \psi(k).$$

We have the following implications:

$$m\text{-dependence} \Rightarrow \psi\text{-mixing} \Rightarrow \phi\text{-mixing} \begin{array}{l} \Rightarrow \beta\text{-mixing} \Rightarrow \\ \Rightarrow \rho\text{-mixing} \Rightarrow \end{array} \alpha\text{-mixing}.$$

In general, there is no other implications between these mixing conditions except for those that can

be derived through transitivity.

We make a remark here that although the phrase "strong mixing condition" is used to refer to the α-mixing, the phrase "strong mixing conditions" used in the plural form refers to all mixing conditions that are at least as strong as α-mixing.

In the case where $\{\xi_t\}$ is strictly stationary, we clearly have that

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{k}^{\infty}} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|.$$

Similar results hold for the other mixing coefficients.

Under strict stationarity, strong mixing conditions are stronger than ergodicity.

Theorem 2.16. *If $\xi_1, \xi_2, \ldots$ is a strictly stationary and α-mixing, then it is mixing (in the ergodic-theoretic sense) and therefore ergodic.*

Proof. See, e.g., White (2001, p. 48). ■

The following two theorems are due to McLeish (1975). We use the exposition of White (2001, p. 49).

Theorem 2.17. *Let $\{\xi_t\}$ be a sequence of random vectors with $\mu_t = \mathbb{E}\xi_t$. Suppose*

$$\sum_{t=1}^{\infty} \left(\frac{\mathbb{E}\left|\xi_t - \mu_t\right|^{r+\delta}}{t^{r+\delta}} \right)^{\frac{1}{r}} < \infty,$$

for some $r \geq 1$ and $0 < \delta \leq r$. If ϕ is of size $-\frac{r}{2r-1}, r \geq 1$ or α is of size $-\frac{r}{r-1}, r > 1$, then

$$\frac{1}{T} \sum_{t=1}^{T} (\xi_t - \mu_t) \to_{a.s.} 0.$$

The theorem generalizes the results in Theorem 2.9, in which $r = 1$. Also, we have

Theorem 2.18. *Let $\{\xi_t\}$ be a sequence of random vectors. If ϕ is of size $-\frac{r}{2r-1}, r \geq 1$ or α is of size $-\frac{r}{r-1}, r > 1$, and*

$$\sup_t \mathbb{E}\left|\xi_t\right|^{r+\delta} < \infty$$

for some $\delta > 0$, then

$$\frac{1}{T} \sum_{t=1}^{T} (\xi_t - \mu_t) \to_{a.s.} 0.$$

Now we consider functions of mixing processes. We obviously have the following result.

Theorem 2.19. *Suppose $\xi_1, \xi_2, \ldots$ is ϕ-mixing (α-mixing) of size $-a, a > 0$. If $\{\eta_t\}$ is a sequence defined by $\eta_t = f(\ldots, \xi_{t-1}, \xi_t, \xi_{t+1}, \ldots)$ where f is a measurable function into $\mathbb{R}^k$, then $\{\eta_t\}$ is ϕ-mixing (α-mixing) of size $-a$.*

Theorem 2.20. *Suppose $\xi_1, \xi_2, \ldots$ is ϕ-mixing (or α-mixing), and $\eta_t = f(\ldots, \xi_{t-1}, \xi_t, \xi_{t+1}, \ldots)$ where f is a measurable function. Suppose $\mathbb{E}\eta_t = 0$ for all t. If $\sum_{t=1}^{\infty} \frac{\|\eta_t\|_{2r}^2}{t^2} < \infty$ for some $r \geq 1$, $\sup_t \left\| \eta_t - \mathbb{E}(\eta_t | \mathcal{F}_{t-k}^{t+k}) \right\|_2 = O(k^{-1/2-\epsilon})$ for some $\epsilon > 0$, and ϕ is of size $-\frac{r}{2r-1}, r \geq 1$ (or α is of size $-\frac{r}{r-1}, r > 1$), then*

$$\frac{1}{T}\sum_{t=1}^{T} \eta_t \to_{a.s.} 0$$

as $T \to \infty$.

Proof. See McLeish (1975). ■

We comment here that in although many processes we are interested in satisfy some strong mixing conditions, there are some simple processes that are not strong mixing. For example, the autoregressive process $y_t = \frac{1}{2}y_{t-1} + \varepsilon_t$ where ε_t is a sequence of iid Bernoulli distributed random variables.

We now turn to the case of martingale difference sequences.

Definition 2.21. Let $\{\xi_t\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\{\mathcal{F}_t\}$ be a sequence of σ-algebras such that $\cdots \subset \mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$ and $\{\xi_t\}$ be *adapted* to $\{\mathcal{F}_t\}$ (i.e., ξ_t is $\mathcal{F}_t$-measurable for each t). We call $\{(\xi_t, \mathcal{F}_t)\}$ a *predictable sequence* if X_t is $\mathcal{F}_{t-1}$ measurable for all t. We call $\{(\xi_t, \mathcal{F}_t)\}$ a *martingale* if $\mathbb{E}|\xi_t| < \infty$ and $\mathbb{E}(\xi_t|\mathcal{F}_{t-1}) = \xi_{t-1}$ a.s. for all t. We call $\{(\xi_t, \mathcal{F}_t)\}$ a *martingale difference sequence* if $\mathbb{E}|\xi_t| < \infty$ and $\mathbb{E}(\xi_t|\mathcal{F}_{t-1}) = 0$ a.s. for all t.

Note that a martingale difference sequence is serially uncorrelated. Therefore, it is a concept that lies between independence and uncorrelatedness.

Theorem 2.22. *Let $\{(\xi_t, \mathcal{F}_t)\}$ be a martingale difference sequence. If*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}|\xi_t|^{2r}}{t^{1+r}} < \infty$$

for some $r \geq 1$, then

$$\frac{1}{T}\sum_{t=1}^{T} \xi_t \to_{a.s.} 0.$$

Proof. See, e.g., White (2001, p. 60). ■

Theorem 2.23. *Let $\{(\xi_t, \mathcal{F}_t)\}$ be a martingale difference sequence. If*

$$\sup_t \mathbb{E}|\xi_t|^{2+\delta} < \infty$$

for some $\delta > 0$, then

$$\frac{1}{T}\sum_{t=1}^{T} \xi_t \to_{a.s.} 0.$$

Proof. See, e.g., White (2001, p. 60). ■

A concept that is easier to deal with than martingale, first introduced by McLeish (1975), is mixingale, which may be viewed as the asymptotic analogue to martingales.

Definition 2.24. Let $\{\xi_t\}$ be a sequence of random variables, and $\mathcal{F}_t$ a sequence of increasing sub-σ-algebras. Let $p \geq 1$. The sequence $\{\xi_t, \mathcal{F}_t\}$ is called an *L^p-mixingale* if there exists non-negative sequences $\{c_t\}$ and $\{\psi_k\}$ such that $\psi_k \to 0$ as $k \to \infty$ and that for each t and k,

(a) $\|\mathbb{E}(\xi_t|\mathcal{F}_{t-k})\|_p \leq c_t\psi_k$;

(b) $\|\xi_t - \mathbb{E}(\xi_t|\mathcal{F}_{t+k})\|_p \leq c_t\psi_{k+1}$.

If $\psi_k = O(k^{-a-\epsilon})$ for some ϵ, we say that the sequence $\{\psi_k\}$ is of size $-a$. Note that mixingales are necessarily mean-zero. Usually we take $c_t = \|\xi_t\|_p$ and we can always make ψ_k non-increasing. In the case where ξ_t is $\mathcal{F}_t$-adapted, the condition (b) is automatically satisfied. If in addition $\psi_k = 0$ for all k, then $\{\xi_i\}$ is a martingale difference sequence.

The following theorem is due to McLeish (1975).

Theorem 2.25. *Let $\{\xi_t\}$ be an L^2-mixingale with mixingale numbers $\{\psi_k\}$ of size $-\frac{r}{2}, r \geq 1$ and $\sum_{t=1}^{\infty} \frac{c_t^2}{t^2} < \infty$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} \xi_t \to_{a.s.} 0.$$

Andrews (1988) obtains L^1 and weak laws of large numbers for uniformly integrable L^1-mxingales without restrictions on the mixingale numbers.

2.2.2 Central Limit Theorems

We begin this section with the Lindeberg-Lévy Central Limit Theorem.

Theorem 2.26. *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent and identically distributed random variables with mean μ_t and variance σ_ξ^2. Then*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\xi_t - \mu_t) \to_d \mathbb{N}(0, \sigma_\xi^2).$$

Proof. See, e.g., Shiryaev (1989). ■

The following is the Lindeberg-Feller Central Limit Theorem.

Theorem 2.27. *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. Let $s_T^2 = \sum_{t=1}^{T} \sigma_t^2$. If the Lindeberg condition that for any $\epsilon > 0$,*

$$\frac{1}{s_T^2}\sum_{t=1}^{T} \mathbb{E}\left((\xi_t - \mu_t)^2 1_{\{|\xi_t - \mu_t| \geq \epsilon s_T\}}\right) \to 0$$

holds, then

$$\frac{1}{s_T}\sum_{t=1}^{T}(\xi_t - \mu_t) \to_d \mathbb{N}(0,1).$$

Proof. See, e.g., Shiryaev (1989). ■

It is easy to show that the Lindeberg condition can be replaced by the stronger Lyapunov condition

$$\frac{1}{s_T^{2+\delta}}\sum_{t=1}^{T}\mathbb{E}\left|\xi_t - \mu_t\right|^{2+\delta} \to 0$$

for some $\delta > 0$. For details, see, e.g., Shiryaev (1989, Section III.4).

Now we look at various dependent cases.

Theorem 2.28. *Let $\{(\xi_t, \mathcal{F}_t)\}$ be a strictly stationary ergodic adapted L^2-mixingale with ψ_k of size -1. Then*

$$\bar{\sigma}_T^2 = Var\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\right) \to \bar{\sigma}^2 < \infty$$

and

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_t \to_d \mathbb{N}(0,\bar{\sigma}^2).$$

Proof. See, e.g., White (2001, p. 125). ■

Theorem 2.29. *Let $\{\xi_t\}$ be a sequence of random variables with $\mu_t = \mathbb{E}\xi_t = 0$ and $\sigma_t^2 = Var(\xi_t)$. Suppose*

$$\sup_t \mathbb{E}\left|\xi_t\right|^r < \infty$$

for some $r \geq 2$, ϕ is of size $-\frac{r}{2(r-1)}, r \geq 2$, or α is of size $-\frac{r}{r-2}, r > 2$, and

$$\bar{\sigma}_T^2 = Var\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\right) > \delta > 0$$

for T large enough, then

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\xi_t - \mu_t}{\bar{\sigma}_T} \to_d \mathbb{N}(0,1).$$

The following theorem is cited from the online Encyclopedia of Mathematics article by Richard. C. Bradley.

Theorem 2.30. *Let $\{\xi_t\}$ be a α-mixing strictly stationary sequence of random variables such that $\mathbb{E}\xi_t = 0, \mathbb{E}\xi_t^2 < \infty, \sigma_T^2 = Var(\sum_{t=1}^{T}\xi_t) \to \infty$ as $T \to \infty$. Then*

$$\frac{1}{\sigma_T}\sum_{t=1}^{T}\xi_t \to_d \mathbb{N}(0,1)$$

as $T \to \infty$ if and only if $\{\frac{(\sum_{t=1}^T \xi_t)^2}{\sigma_T^2}\}$ is uniformly integrable. If all of the above conditions hold, then $\sigma_T^2 = Th(T)$ for some function $h(T)$ that is slow varying as $T \to \infty$.

Theorem 2.31. *Let $\xi_1, \xi_2, \ldots$ be a martingale difference sequence such that $\mathbb{E}\xi_t^2 = \sigma_t^2$. If $\frac{1}{T}\sum_{t=1}^T \xi_t^2 \to_p \sigma_\xi^2$, and if the conditional Lindeberg condition that for any $\epsilon > 0$,*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\xi_t^2 1_{\{|\xi_t| \geq \epsilon\sqrt{T}\}} \middle| \mathcal{F}_{t-1}\right) \to_p 0$$

holds, then

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_t \to_d \mathbb{N}(0, \sigma_\xi^2).$$

The condition $\frac{1}{T}\sum_{t=1}^T \xi_t^2 \to_p \sigma_\xi^2$ can be replace by $\frac{1}{T}\sum_{t=1}^T \mathbb{E}(\xi_t^2|\mathcal{F}_{t-1}) \to_p \sigma_\xi^2$. See Shiryaev (1989, p. 543). The conditional Lindeberg condition can be replaced by any of the following three conditions:

(a) The unconditional Lindeberg condition. For any $\epsilon > 0$,

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\xi_t^2 1_{\{|\xi_t| \geq \epsilon\sqrt{T}\}}\right) \to 0.$$

(b) The unconditional Lyapunov condition. There exists some $\delta > 0$ such that for all t,

$$\mathbb{E}\,|\xi_t|^{2+\delta} < \infty.$$

(c) The conditional Lyapunov condition. There exists some $\delta > 0$ such that for all t

$$\mathbb{E}\left(|\xi_t|^{2+\delta} \middle| \mathcal{F}_{t-1}\right) < \infty.$$

It is known that the conditional Lyapunov condition implies the unconditional Lyapunov condition, which implies the unconditional Lindeberg condition, which in turn implies the conditional Lindeberg condition. See Alj et al. (2014) for further references.

2.3 Estimation of the Mean of Weakly Stationary Time Series

In this section, we focus on the estimation of the mean of the weakly stationary time series. Suppose we observe $(y_1, y_2, \ldots, y_T)$ from a weakly stationary time series $(y_t)_{t\in\mathbb{Z}}$. The most natural estimator for the mean $\mu = \mathbb{E}y_t$ is the sample average

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^T y_t.$$

This estimator is unbiased, since $\mathbb{E}\hat{\mu} = 1/T \sum_{t=1}^{T} \mathbb{E}y_t = \mu$. This estimator, under some regularity conditions, is also consistent, i.e., $\hat{\mu} \to_p \mu$. In this case, we say that the series $\{y_t\}$ is *ergodic for the mean*. Recall that $X_T \to_p X$ if and only if

$$\lim_{T\to\infty} \mathbb{P}\{\omega \in \Omega : |X_T(\omega) - X(\omega)| > \varepsilon\} = 0,$$

for any choice of $\varepsilon > 0$, where $|\cdot|$ is the Euclidean norm defined as the length of the vector considered. In the case of scalars, $|\cdot|$ is just the absolute value. For a matrix $A = [a_{ij}]$, we use $|A|$ to denote the Frobenius norm $|A| = \sqrt{\sum_i \sum_j a_{ij}^2}$.

Theorem 2.32. *Let $(y_t)_{t\in\mathbb{Z}}$ be a weakly stationary time series with mean μ and autocovariance function $\gamma(\cdot)$. Let $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$. If $\sum_{k=1}^{\infty} |\gamma(k)| < \infty$, then*

(a) $\lim_{T\to\infty} T\mathbb{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)' = \sum_{k=-\infty}^{\infty} \gamma(k)$.

(b) $\hat{\mu} \to_p \mu$ *as* $T \to \infty$.

Proof.

$$\begin{aligned} T\mathbb{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)' &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathbb{E}(y_t - \mu)(y_s - \mu)' \\ &= \frac{1}{T}\big[T\gamma(0) + 2(T-1)\gamma(1) + 2(T-2)\gamma(2) + \ldots + 2\gamma(T-1)\big] \\ &= \sum_{k=-(T-1)}^{T-1} \left(1 - \frac{|k|}{T}\right)\gamma(k). \end{aligned}$$

Given $\sum_{k=1}^{\infty} |\gamma(k)| < \infty$, by Kronecker's Lemma, $\sum_{k=-(T-1)}^{T-1} \frac{|k|}{T}\gamma(k) \to 0$. See Shiryaev (1989, p. 390). Therefore, $T\mathbb{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)' \to \sum_{t=-\infty}^{\infty} \gamma(k)$. As a consequence, $\mathbb{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)' \to 0$. This implies that $\hat{\mu} \to_{L^2} \mu$, which in turn implies that $\hat{\mu} \to_p \mu$. ■

Note that $\hat{\mu}$ is said to *converge in mean square* to μ, or $\hat{\mu} \to_{L^2} \mu$ if and only if $\mathbb{E}\,|\hat{\mu} - \mu|^2 \to 0$. It is a well known result in probability theory that convergence in mean square implies convergence in probability due to Chebyshev's inequality.

The asymptotic variance of the sample mean estimator is given by

$$\mathrm{Var}(\sqrt{T}(\hat{\mu} - \mu)) = T\mathbb{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)' \to \sum_{k=-\infty}^{\infty} \gamma(k).$$

This limit is called the *long-run variance* of the series $\{y_t\}$.

One may be further interested in the asymptotic distribution of $\hat{\mu} - \mu$. However, the asymptotic distribution can be obtained only with more assumptions on the structure of the time series, which we shall not elaborate on at this moment. In the special case where the time series is Gaussian, $\sqrt{T}(\hat{\mu} - \mu) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(y_t - \mu)$ is Gaussian, and it converges in distribution to $\mathbb{N}\left(0, \sum_{k=-\infty}^{\infty} \gamma(k)\right)$. As we shall see in Chapter 4, if $\{y_t\}$ is a linear process, a central limit theorem holds for $\{y_t\}$, and

we also have $\sqrt{T}(\hat{\mu} - \mu) \to_d \mathbb{N}\left(0, \sum_{k=-\infty}^{\infty} \gamma(k)\right)$. Or one may impose regularity assumptions such as ergodicity, mixing, or martingale assumptions to the time series so that a central limit theorem in Section 2.2 holds.

2.4 Estimation of the Autocovariance Function of Weakly Stationary Time Series

With a sample $(y_1, y_2, \ldots, y_T)$ from a weakly stationary time series $(y_t)_{t\in\mathbb{Z}}$, we may estimate the autocovariance function $\gamma(k)$ of the series by

$$\hat{\gamma}(k) = \frac{1}{T-k} \sum_{t=k+1}^{T} (y_t - \hat{\mu})(y_{t-k} - \hat{\mu})'$$

for $0 \leq k \leq T-1$, and $\hat{\gamma}(k) = \hat{\gamma}(-k)'$ for $-(T-1) \leq k < 0$, where $\hat{\mu}$ is the sample average estimator studied in the previous section. Note that because of the existence of $\hat{\mu}$, this $\hat{\gamma}(k)$ is not an unbiased estimator of $\gamma(k)$.

We may also replace the denominator $T-k$ in the above expression with T. The estimator is then given by

$$\hat{\gamma}(k) = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \hat{\mu})(y_{t-k} - \hat{\mu})'.$$

Obviously the two estimators are asymptotically equivalent. In this section we shall use the latter estimator.

In order to obtain the asymptotic properties of the autocovariance estimator, we need to look at the 4-th moment of the process. We therefore define the following stationarity concept. In this section we mainly follow Parzen (1957) and deal with the case in which $(y_t)_{t\in\mathbb{Z}}$ is a univariate process. It is quite straightforward to extend the results to the multivariate case. We call a process *4-th order weakly stationary* if $\mathbb{E}|y_t|^4 < \infty$ and that for any k_1, k_2, k_3, $P(k_1, k_2, k_3) = \mathbb{E}y_t y_{t+k_1} y_{t+k_2} y_{t+k_3}$ is independent of t.

In particular, if y_t is mean zero and Gaussian, then its fourth moment function can be expressed as

$$P_{\mathbb{N}}(k_1, k_2, k_3) = \gamma(k_1)\gamma(k_2 - k_3) + \gamma(k_2)\gamma(k_3 - k_1) + \gamma(k_3)\gamma(k_1 - k_2).$$

by Isserlis' theorem. Then

$$Q(k_1, k_2, k_3) = P(k_1, k_2, k_3) - P_{\mathbb{N}}(k_1, k_2, k_3)$$

of a generic mean zero fourth order weakly stationary process $\{y_t\}$ measures the deviation of the process from Gaussianity in terms of the fourth moment structure. The quantity Q coincide with the so-called *fourth cumulant* of the process.

Theorem 2.33. *Let $\{y_t\}$ be a fourth order weakly stationary time series with absolutely summable autocovariance function $\gamma(\cdot)$ and absolutely summable fourth cumulant $Q(k_1, k_2, k_3)$. Let $\hat{\gamma}(k)$ be*

the estimator of $\gamma(k)$*. Then for any* k, k_1, k_2*,*

$$\mathbb{E}\hat{\gamma}(k) - \gamma(k) = O(T^{-1}),$$

and

$$\begin{aligned}
&\lim_{T\to\infty} T\,Cov(\hat{\gamma}(k_1), \hat{\gamma}(k_2)) \\
&= \sum_{k=-\infty}^{\infty} \left[Q\left(|k_1|, k, k+|k_2|\right) + \gamma(k)\gamma\left(k+|k_2|-|k_1|\right) + \gamma\left(k+|k_2|\right)\gamma(k-|k_1|)\right].
\end{aligned}$$

Proof. Consider the case where $k \geq 0$. Write

$$\begin{aligned}
\hat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T-k}(y_{t+k} - \mu + \mu - \hat{\mu})(y_t - \mu + \mu - \hat{\mu}) \\
&= D(k) + \left(1 - \frac{k}{T}\right)\gamma(k) + R(k)
\end{aligned}$$

where

$$D(k) = \frac{1}{T}\sum_{t=1}^{T-k}[(y_{t+k} - \mu)(y_t - \mu) - \gamma(k)]$$

and

$$R(k) = \frac{1}{T}\sum_{t=1}^{T-k}(y_{t+k} - \mu)(\mu - \hat{\mu}) + \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \mu)(\mu - \hat{\mu}) + \left(1 - \frac{k}{T}\right)(\mu - \hat{\mu})^2.$$

We can show that for any $k_1 \geq 0, k_2 \geq 0$,

$$\begin{aligned}
\mathbb{E}D(k_1)D(k_2) &= \frac{1}{T^2}\sum_{t_1=1}^{T-k_1}\sum_{t_2=1}^{T-k_2}[\tilde{P}(t_1+k_1, t_1, t_2+k_2, t_2) - \gamma(k_1)\gamma(k_2)] \\
&= \frac{1}{T^2}\sum_{t_1=1}^{T-k_1}\sum_{t_2=1}^{T-k_2}[Q(k_1, t_2-t_1, t_2-t_1+k_2) + \gamma(t_2-t_1)\gamma(t_2-t_1+k_2-k_1) \\
&\quad + \gamma(t_2-t_1+k_2)\gamma(t_2-t_1-k_1)] \\
&= \frac{1}{T}\sum_{k=-(T-1-k_1)}^{T-1-k_2} u(k, k_1, k_2)[Q(k_1, k, k+k_2) + \gamma(k)\gamma(k+k_2-k_1) \\
&\quad + \gamma(k+k_2)\gamma(k-k_1)]
\end{aligned}$$

where

$$u(k, k_1, k_2) = \begin{cases} 1 - \frac{\max(k_1, k_2+k)}{T}, & \text{if } k \geq 0, \\ 1 - \frac{\max(k_1+|k|, k_2)}{T}, & \text{if } k < 0. \end{cases}$$

Under the summability of the fourth cumulant, by Kronecker Lemma we have

$$\lim_{T\to\infty} T\mathbb{E}D(k_1)D(k_2) = \sum_{k=-\infty}^{\infty} [Q(k_1, k, k+k_2) + \gamma(k)\gamma(k+k_2-k_1) + \gamma(k+k_2)\gamma(k-k_1)].$$

We can also show that

$$\begin{aligned}
\mathbb{E}R(k)^2 &\leq \left(\frac{9}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}|\gamma(t-s)|\right)\left(\mathbb{E}(\mu-\hat{\mu})^2\right) \\
&= \left(\frac{9}{T}\sum_{k=-T+1}^{T=1}\left(1-\frac{k}{T}\right)|\gamma(k)|\right)\left(\mathbb{E}(\mu-\hat{\mu})^2\right) \\
&\leq CT^{-2}
\end{aligned}$$

for some constant C under the absolute summability of $\gamma(k)$.

The results then follows easily. ■

Note that in the Gaussian case, since $Q(k_1, k_2, k_3) = 0$ for all k_1, k_2, k_3, we have

$$\lim_{T\to\infty} T\text{Cov}(\hat{\gamma}(k_1), \hat{\gamma}(k_2)) = \sum_{k=-\infty}^{\infty} \left[\gamma(k)\gamma\left(k+|k_2|-|k_1|\right) + \gamma\left(k+|k_2|\right)\gamma(k-|k_1|)\right].$$

The above theorem gives the asymptotic bias and variance of the estimator $\hat{\gamma}(k)$. It obviously implies that $\hat{\gamma}(k)$ converges in mean square, and therefore in probability, to $\gamma(k)$. We may also obtain a central limit theorem for the $D(k)$ part, and therefore establish that $\sqrt{T}(\hat{\gamma}(k) - \gamma(k)) \to_d \mathbb{N}(0, V)$ where V is given by the expression in the above theorem if we assume that $\{y_t\}$ satisfies some strong mixing conditions with certain mixing rate. For details, see Section 2.2.2.

Absolute summability of the fourth cumulants holds trivially for Gaussian processes. Also, it is known that absolute summability of the four cumulants holds for fourth order stationary linear processes[2] with absolutely summable coefficients and innovations whose fourth moments exist (Andrews, 1991, p. 823). Absolute summability of the fourth cumulants also holds when the process satisfy some α-mixing conditions. Andrews (1991) shows that if $\{X_t\}$ is mean zero, fourth order weakly stationary with $\sup_t \mathbb{E}|X_t|^{4\nu} < \infty$ for some $\nu > 1$, and α-mixing with mixing coefficients $\alpha(k)$ satisfying $\sum_{k=1}^{\infty} k^2\alpha(k)^{\frac{\nu-1}{\nu}} < \infty$, then its forth cumulants are absolutely summable.

[2]See Chapter 4 for an introduction of linear processes.

3 Spectral Analysis of Weakly Stationary Processes

3.1 Spectral Distributions and Spectral Densities

This chapter introduces the frequency domain analysis of time series. Although time series we usually encounter in real-world applications are real-valued, in spectral analysis, it will be mathematically more convenient to consider them as complex-valued series. We therefore adapt a few concepts we encountered earlier to processes on the complex field $\mathbb{C}$. In this chapter we state results for the univariate case. Their generalization to the vector process should be straightforward.

A complex-valued time series $\{X_t\}$ is said to be weakly stationary if $\mathbb{E}|X_t|^2 < \infty$, and both $\mathbb{E}X_t$ and $\mathbb{E}X_t\bar{X}_{t-h}$ are independent of t, where $\bar{X}_t$ denotes the complex conjugate of X_t. The autocovariance function $\gamma(\cdot)$ of a complex-valued weakly stationary time series $\{X_t\}$ is defined to be

$$\gamma(k) = \mathbb{E}\left(X_t - \mathbb{E}X_t\right)\overline{\mathbb{E}(X_{t-h} - \mathbb{E}X_{t-h})} = \mathbb{E}X_t\bar{X}_{t-h} - \mathbb{E}X_t\mathbb{E}\bar{X}_{t-h}.$$

We have a theorem analogous to Theorem 2.7.

Theorem 3.1. *A mapping $\gamma : \mathbb{Z} \to \mathbb{C}$ is the autocovariance function of a complex-valued weakly stationary time series if and only if $\gamma(k) = \overline{\gamma(-k)}$ for any $k \in \mathbb{Z}$ and $\sum_{r=1}^n \sum_{s=1}^n a_r\gamma(t_r - t_s)\bar{a}_s \geq 0$ for any $t_1, \ldots, t_n \in \mathbb{Z}, a_1, \ldots, a_n \in \mathbb{C}$ and $n \in \mathbb{N}$.*

The following is the spectral characterization of the autocovariance function.

Theorem 3.2. *A mapping $\gamma : \mathbb{Z} \to \mathbb{C}$ is the autocovariance function of a complex-valued weakly stationary time series if and only if*

$$\gamma(k) = \int_{-\pi}^{\pi} e^{i\lambda k}\mathrm{d}F(\lambda), \quad k \in \mathbb{Z}$$

where F is a right-continuous, non-decreasing, bounded function on $[-\pi, \pi]$ with $F(-\pi) = 0$.

In the above theorem, $i = \sqrt{-1}$. Note that by De Moivre's theorem we have $e^{i\alpha} = \cos\alpha + i\sin\alpha$ and $e^{-i\alpha} = \cos\alpha - i\sin\alpha$ for any $\alpha \in \mathbb{R}$.

Proof. Sufficiency: It is easy to verify that $\gamma(k) = \overline{\gamma(-k)}$. Let $a_1, \ldots, a_n \in \mathbb{C}$. Then

$$\begin{aligned}\sum_{r=1}^n \sum_{s=1}^n a_r\gamma(t_r - t_s)\bar{a}_s &= \int_{-\pi}^{\pi} \sum_{r,s=1}^n a_r\bar{a}_s e^{i\lambda(t_r - t_s)}\mathrm{d}F(\lambda) \\ &= \int_{-\pi}^{\pi} \left|\sum_{r=1}^n a_r e^{i\lambda t_r}\right|^2 \mathrm{d}F(\lambda) \geq 0.\end{aligned}$$

Then by the above theorem γ is an autocovariance function.

Necessity: Suppose γ is an autocovariance function. Define

$$\begin{aligned} f_N(\lambda) &= \frac{1}{2\pi N} \sum_{r,s=1}^{N} e^{-i\lambda r} \gamma(r-s) e^{i\lambda s} \\ &= \frac{1}{2\pi} \sum_{|k|<N} \left(1 - \frac{|k|}{N}\right) e^{-i\lambda k} \gamma(k) \end{aligned}$$

on $\lambda \in (-\pi, \pi]$ and 0 everywhere else. By positive definiteness of γ, we have $f_N(\lambda) \geq 0$ on $(-\pi, \pi]$. Define

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(\nu) \mathrm{d}\nu$$

for $\lambda \in (-\pi, \pi]$, $F_N(\lambda) = 0$ for $\lambda \leq -\pi$, and $F_N(\lambda) = F_N(\pi)$ for $\lambda > \pi$. Then it is easy to see that

$$\int_{-\pi}^{\pi} e^{i\lambda k} \mathrm{d}F_N(\lambda) = \begin{cases} \left(1 - \frac{|k|}{N}\right) \gamma(k), & |k| < N, \\ 0 & |k| \geq N. \end{cases}$$

Since $\{F_N\}$ is a sequence of distribution functions supported on $[-\pi, \pi]$, and $F_N(\pi) = \gamma(0)$ for all N, we have that $\{F_N\}$ is tight. It then follows from Shiryaev (1989, Theorem 1, p. 318) that that there exists a subsequence $\{F_{N_m}\}$ such that $F_{N_m} \to_w F$ for some distribution function F where $\to_w$ denotes weak convergence of measures. Since the mapping $\lambda \mapsto e^{i\lambda k}$ is continuous and bounded, this implies that

$$\lim_{m\to\infty} \int_{-\pi}^{\pi} e^{i\lambda k} \mathrm{d}F_{N_m}(\lambda) = \int_{-\pi}^{\pi} e^{i\lambda k} \mathrm{d}F(\lambda).$$

However, the limit on the left-hand-side can only be $\gamma(k)$. Therefore, we have

$$\gamma(k) = \int_{-\pi}^{\pi} e^{i\lambda k} \mathrm{d}F(\lambda).$$

Obviously, F constructed in this way satisfies the required properties. ■

We shall call F the *spectral distribution function* of γ or $\{X_t\}$. If F admits a derivative f such that $F(\lambda) = \int_{-\pi}^{\lambda} f(\nu) \mathrm{d}\nu$, we call f the *spectral density* of γ or $\{X_t\}$.

We note here that the spectral distribution F is uniquely determined by the autocovariance function γ. This is because $\gamma(k) = \int_{-\pi}^{\pi} e^{i\lambda k} \mathrm{d}F(\lambda)$ holds for all k, and $\{e^{i\lambda k}\}_{k\in\mathbb{Z}}$ serves as a set of test functions that can be used to determine the values of F. See Brockwell and Davis (1991, p. 119) for details.

Let $\{X_t\}$ be a weakly stationary time series with absolutely summable autocovariance function $\gamma(\cdot)$. Let $f : (-\pi, \pi] \to \mathbb{R}$ be defined by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\lambda k}.$$

Then we have

$$\begin{aligned}\int_{-\pi}^{\pi} e^{i\lambda k} f(\lambda)\mathrm{d}\lambda &= \int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} \gamma(k) e^{i\lambda(k-m)}\mathrm{d}\lambda \\ &= \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} \gamma(k) \int_{-\pi}^{\pi} e^{i\lambda(k-m)}\mathrm{d}\lambda \\ &= \gamma(k)\end{aligned}$$

where the interchange of the summation and integration is guaranteed by Fubini's theorem since $\int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} \left|\gamma(k)e^{i\lambda(k-m)}\right| \mathrm{d}\lambda < \infty$ under the absolute summability assumption of γ. In view of the above theorem and that $0 \le f_N(\lambda) \to f(\lambda)$, we have that

Theorem 3.3. *An absolutely summable function $\gamma : \mathbb{Z} \to \mathbb{C}$ is the autocovariance function of a weakly stationary complex-valued time series if and only if it has spectral density given by*

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\lambda k} \ge 0, \quad \lambda \in (-\pi, \pi].$$

For any real weakly stationary scalar process with absolutely summable autocovariances, since $\gamma(-k) = \gamma(k)$, we have that $f(\lambda)$ is real, $f(\lambda) = f(-\lambda)$, and

$$f(\lambda) = \frac{1}{2\pi} \left(\gamma(0) + 2 \sum_{k=1}^{\infty} \gamma(k) \cos(\lambda k) \right).$$

Actually we may easily derive the following result.

Theorem 3.4. *A function $f : (-\pi, \pi] \to \mathbb{R}$ is the spectral density of a real-valued weakly stationary process if and only if $f(\lambda) = f(-\lambda)$, $f(\lambda) \ge 0$ and $\int_{-\pi}^{\pi} f(\lambda)\mathrm{d}\lambda < \infty$.*

In the end we relate the spectral density with the bounds of eigenvalues of the covariance matrix of $(X_1, \ldots, X_n)$ when $\{X_t\}$ is a weakly stationary process.

Theorem 3.5. *Let $\{X_t\}$ be a weakly stationary process with spectral density f such that $0 < m \le f(\lambda) \le M < \infty$ for all $\lambda \in (-\pi, \pi]$. Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be the eigenvalues of the covariance matrix Γ_n of $(X_1, X_2, \ldots, X_n)'$. Then*

$$2\pi m \le \lambda_1 \le \lambda_n \le 2\pi M.$$

Proof. Supoose $v = (v_1, \cdots, v_n)'$ is a normalized eigenvector of Γ_n with eigenvalue λ. Let the autocovariance function of $\{X_t\}$ be γ. Then we have

$$\begin{aligned}\lambda &= v'\Gamma_n v \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} v_j v_k \gamma(j-k)\end{aligned}$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} v_j v_k \int_{-\pi}^{\pi} e^{-i(j-k)\lambda} f(\lambda)\mathrm{d}\lambda$$
$$\leq 2\pi M \sum_{j=1}^{n} v_j^2.$$

This implies that $\lambda \leq 2\pi M$. Similarly we may show that $\lambda \geq 2\pi m$. ■

3.2 Spectral Representation

To obtain the spectral representation of a weakly stationary process, we first introduce some concepts and define a stochastic integral. For more details, readers may refer to Shiryaev (1989, Section VI.2).

Throughout this section, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $(E, \mathcal{E})$ be a measurable space.

Definition 3.6. A complex-valued function $Z(\Delta) = Z(\omega; \Delta)$ defined for $\omega \in \Omega, \Delta \in \mathcal{E}$ is called a *stochastic measure* if

(a) $\mathbb{E}\left|Z(\Delta)\right|^2 < \infty$ for all $\Delta \in \mathcal{E}$;
(b) For every disjoint $\Delta_1, \Delta_2 \in \mathcal{E}$, $Z(\Delta_1 \cup \Delta_2) = Z(\Delta_1) + Z(\Delta_2)$;
(c) For all disjoint $\Delta_1, \Delta_2, \ldots$ in $\mathcal{E}$, we have $\lim_{n\to\infty} \mathbb{E}\left|Z(\bigcup_{i=1}^{\infty} \Delta_i) - \sum_{i=1}^{n} Z(\Delta_i)\right|^2 = 0$.

Note that the measurability of $Z(\cdot; \Delta)$ for all $\Delta \in \mathcal{E}$ is implicitly required by the statement of the definition.

Definition 3.7. A stochastic measure Z is called *orthogonal* if for every disjoint $\Delta_1, \Delta_2 \in \mathcal{E}$ we have $\mathbb{E}Z(\Delta_1)\overline{Z(\Delta_2)} = 0$. For an orthogonal stochastic measure Z, we call the function m defined by $m(\Delta) = \mathbb{E}\left|Z(\Delta)\right|^2, \Delta \in \mathcal{E}$ the *structure function* of Z.

It can be shown that the structure function m is a finite measure on $(E, \mathcal{E})$.

Now we define integral with respect to orthogonal stochastic measures. Let $L^2 = L^2(E, \mathcal{E}, m)$ be the Hilbert space of complex-valued square integrable functions on E with inner product given by

$$\langle f, g\rangle_{L^2} = \int_E f\overline{g}\mathrm{d}m$$

and $H^2 = H^2(\Omega, \mathcal{F}, \mathbb{P})$ be the space of complex-valued square integrable random variables on Ω with inner product given by

$$\langle \xi, \eta\rangle_{H^2} = \mathbb{E}\xi\overline{\eta}.$$

Norms are defined in the usual way.

Now for simple function $f = \sum_{i=1}^{n} s_i 1_{\Delta_i}, \Delta_i \in \mathcal{E}$, define the integral with respect to an orthogonal stochastic integral Z by

$$\int_\Omega f\mathrm{d}Z = \sum_{i=1}^{n} s_i Z(\Delta_i).$$

It is easy to verify that for simple f and g, we have

$$\left\langle \int_\Omega f \mathrm{d}Z, \int_\Omega g \mathrm{d}Z \right\rangle_{H^2} = \langle f, g \rangle_{L^2} \tag{3.1}$$

and

$$\left\| \int_\Omega f \mathrm{d}Z \right\|^2 = \|f\|^2 = \int_E |f|^2 \, \mathrm{d}m. \tag{3.2}$$

Now we may approximate any $f \in L^2$ by simple $\{f_n\}$ in L^2. Then for $n, m \to 0$, we have $\|f_n - f_m\| \to 0$, which by the above equivalence implies that $\left\| \int f_n \mathrm{d}Z - \int f_m \mathrm{d}Z \right\| \to 0$. Since $\{\int f_n \mathrm{d}Z\}$ is Cauchy, by completeness of H^2, there exists a unique ($\mathbb{P}$-a.s.) random variable that is the limit of $\{\int f_n \mathrm{d}Z\}$. We define this limit to be the *stochastic integral* of f with respect to Z, and denote it by $\int_\Omega f \mathrm{d}Z$.

Similarly one can easily extend the equations (3.1) and (3.2) to general f and g in L^2. This establishes the isomorphism between L^2 and H^2 defined by $f \mapsto \int_\Omega f \mathrm{d}Z$. Linearity of the integral can also be shown easily.

We now associate an orthogonal stochastic measure with a stochastic process which has orthogonal increments.

Definition 3.8. A set of complex-valued random variables $\{Z_\lambda\}, \lambda \in \mathbb{R}$, is called a *stochastic process with orthogonal increments* if

(a) $\mathbb{E}\,|Z_\lambda|^2 < \infty$ for all $\lambda \in \mathbb{R}$;

(b) For every $\lambda_n \downarrow \lambda \in \mathbb{R}$, $\mathbb{E}\,|Z_\lambda - Z_{\lambda_n}|^2 \to 0$ as $n \to \infty$;

(c) $\mathbb{E}(Z_{\lambda_4} - Z_{\lambda_3})(\overline{Z_{\lambda_2} - Z_{\lambda_1}}) = 0$ for any $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 \in \mathbb{R}$.

Now consider the case where $E = \mathbb{R}$ and $\mathcal{E} = \mathcal{B}(\mathbb{R})$. Let $Z = Z(\Delta), \Delta \in \mathcal{B}(\mathbb{R})$ be an orthogonal stochastic measure with structure function $m = m(\Delta)$, and the distribution function $F(\lambda) = m(-\infty, \lambda]$. Define the stochastic process $\{Z_\lambda\}$ by $Z_\lambda = Z((-\infty, \lambda])$. It is easy to verify that $\{Z_\lambda\}$ is a stochastic process with orthogonal increments. On the other hand, if $\{Z_\lambda\}$ is a stochastic process with orthogonal increments such that $\mathbb{E}\,|Z_\lambda|^2 = F(\lambda), F(-\infty) = 0, F(\infty) < \infty$, then we can verify that Z defined by $Z((a, b]) = Z_b - Z_a$ (and extended to all $\mathcal{B}(\mathbb{R})$-measurable sets) is an orthogonal stochastic measure. We therefore have established a one-to-one correspondence between an orthogonal stochastic measure and a stochastic process with orthogonal increments, and we may now define the stochastic integral $\int_\mathbb{R} f(\lambda) \mathrm{d}Z_\lambda$ with respect to a process $\{Z_\lambda\}$ with orthogonal increments by the value of the stochastic integral $\int_\Omega f \mathrm{d}Z$ where Z is the corresponding orthogonal stochastic measure associated with $\{Z_\lambda\}$.

Now we obtain the spectral representation of weakly stationary processes. Let $\{\xi_t\}$ be a weakly stationary process with spectral distribution function F. One may check that the mapping U given by

$$U\left(\sum_{j=1}^n a_j \xi_{t_j}\right) = \sum_{j=1}^n a_j e^{it_j \cdot}$$

defines an isomorphism between $\bigvee_{t=-\infty}^{\infty} \xi_t$ and $\bigvee_{t=-\infty}^{\infty} e_t$ where $e_t(\lambda) = e^{it\lambda}, \lambda \in (-\pi, \pi]$, and $\bigvee$ denotes the span. By a usual argument we can extend this isomorphism uniquely to an isomorphism between the closure of the two spanned spaces, i.e., between $\overline{\bigvee_{t=-\infty}^{\infty} \xi_t}$ and $L^2((-\pi,\pi], \mathcal{B}((-\pi,\pi]), F)$. We therefore have the following result established:

Theorem 3.9. *Let $\{\xi_t\}$ be a mean-zero weakly stationary process with spectral distribution function F. Then there exists a unique isomorphism U between $\overline{\bigvee_{t=-\infty}^{\infty} \xi_t}$ and $L^2((-\pi,\pi], \mathcal{B}((-\pi,\pi]), F)$ such that*

$$(U\xi_t)(\lambda) = e^{it\lambda}, \quad \lambda \in (-\pi, \pi]$$

for all $t \in \mathbb{Z}$.

We define the set function $Z(\Delta) = U^{-1}1_\Delta$ for $\Delta \in \mathcal{B}((-\pi,\pi])$. Note that $\mathbb{E}\,|Z(\Delta)|^2 = \|1_\Delta\| = F(\Delta)$. We can easily verify that Z thus defined is an orthogonal stochastic measure. Also we may easily show that for simple $f \in L^2((-\pi,\pi], \mathcal{B}((-\pi,\pi]), F)$, $\int f \mathrm{d}Z = U^{-1}f$. A usual argument extends this equality for general $f \in L^2((-\pi,\pi], \mathcal{B}((-\pi,\pi]), F)$. Now take $f(\lambda) = e^{it\lambda}$, in view of Theorem 3.9, we have the following *spectral representation* result.

Theorem 3.10. *Let $\{\xi_t\}$ be a mean-zero weakly stationary process with spectral distribution function F. Then there exists an orthogonal stochastic measure $Z = Z(\Delta), \Delta \in \mathcal{B}((-\pi,\pi])$ with structure function F such that for all $t \in \mathbb{Z}$,*

$$\xi_t = \int_{-\pi}^{\pi} e^{it\lambda} \mathrm{d}Z(\lambda) \qquad \mathbb{P}\text{-}a.s..$$

We make the following remarks regarding this theorem.

(a) The associated stochastic process with orthogonal increments can obviously be defined by $Z_\lambda = U^{-1}1_{(-\pi,\lambda]}$. So obviously the above theorem can also be stated in the form of stochastic integral with respect to the process $\{Z_\lambda\}$.

(b) It can be shown that if $\xi_t = \int e^{it\lambda} \mathrm{d}Z_\lambda, \mathbb{P}$-a.s., and $\xi_t = \int e^{it\lambda} \mathrm{d}Y_\lambda, \mathbb{P}$-a.s. for two stochastic processes Z_λ and Y_λ with orthogonal increments, then $\mathbb{P}(Z_\lambda = Y_\lambda) = 1$ for each $\lambda \in [-\pi, \pi]$.

(c) If F has a discontinuity (jump) at $\lambda = \lambda_0$, then $\xi_t = \int_{[-\pi,\pi]\setminus\{\lambda_0\}} e^{it\lambda} \mathrm{d}Z(\lambda) + e^{it\lambda_0} Z(\{\lambda_0\})$. This implies that there is a deterministic sinusoidal component with frequency λ_0 in the time series. For the general case, we may think that the spectral representation decomposes the time series $\{\xi_t\}$ into (a continuous) "sum" of sine and cosine components with different frequencies λ. The spectral density $f(\lambda)$ of the process $\{\xi_t\}$, if exists, may be viewed as the variance of $|\mathrm{d}Z(\lambda)|$, which gives the "magnitude", or "significance", or "importance" of the components corresponding to different frequencies. In particular, we have that $\gamma(0) = \int_{-\pi}^{\pi} f(\lambda)\mathrm{d}\lambda$, i.e., the variance of $\{\xi_t\}$ is the integral of contributions $f(\lambda)$ from the individual frequencies.

Apparently, if $\{\xi_n\}$ is a real process with the representation $\int_{-\pi}^{\pi} e^{i\lambda t} \mathrm{d}Z$, then we may represent ξ_t as

$$\xi_t = \int_{-\pi}^{\pi} \cos(\lambda t) \mathrm{d}Z_1(\lambda) + \int_{-\pi}^{\pi} \sin(\lambda t) \mathrm{d}Z_2(\lambda) \quad a.s.$$

where $Z = Z_1 + iZ_2$, Z_1 and Z_2 real valued.

Theorem 3.11. *Let $\{\xi_t\}$ be weakly stationary with spectral representation $\xi_t = \int_{-\pi}^{\pi} e^{i\lambda t} \mathrm{d}Z(\lambda)$ and spectral distribution F. Let $H^2 = \overline{\bigvee_{t=-\infty}^{\infty} \xi_t}$. If $\eta \in H^2$, then there exists a function $\varphi \in L^2((-\pi, \pi], \mathcal{B}((-\pi, \pi]), F)$ such that*

$$\eta = \int_{-\pi}^{\pi} \varphi(\lambda) \mathrm{d}Z(\lambda) \quad a.s..$$

If there exists $\{h_j\}_{j=-\infty}^{\infty}$ such that the sequence $\eta_t = \sum_{j=-\infty}^{\infty} h_j \xi_{t-j}$ is well defined in mean square, then

$$\eta_t = \int_{-\pi}^{\pi} e^{i\lambda t} h(e^{-i\lambda}) \mathrm{d}Z(\lambda) \quad a.s.$$

where $h(z) = \sum_{j=-\infty}^{\infty} h_j z^j$.

Proof. See Shiryaev (1989, p. 433). ■

The Fourier transform $h(e^{-i\lambda})$ is called the *transfer function* associated with the *linear filter* $h(L)$.

Theorem 3.12. *Let $\{\eta_t\}$ be a weakly stationary time series with spectral density $f_\eta(\lambda)$. Then (possibly at the expense of enlarging the original probability space) we can find a white noise $\{\varepsilon_t\}$ and a linear filter $h(z) = \sum_{j=-\infty}^{\infty} h_j z^j$ such that $\eta_t = h(L)\varepsilon_t = \sum_{j=-\infty}^{\infty} h_j \varepsilon_{t-j}$.*

In particular, if $f_\eta(\lambda) > 0$ almost everywhere with respect to the Lebesgue measure and $f_\eta(\lambda) = \frac{1}{2\pi} \left| h(e^{-i\lambda}) \right|$ for some $h(z) = \sum_{j=0}^{\infty} h_j z^j, \sum_{j=0}^{\infty} |h_j|^2 < \infty$, then $\eta_t = \sum_{j=0}^{\infty} h_j \varepsilon_{t-j}$ for some white noise $\{\varepsilon_t\}$ (on the same probability space).

Proof. See Shiryaev (1989, p. 435). ■

In the end, we use spectral representation to prove some ergodic theorems for weakly stationary time series.

Theorem 3.13. *Let $\{\xi_t\}$ be a weakly stationary time series with zero mean, autocovariance function $\gamma(\cdot)$, spectral distribution F and spectral representation $\xi_t = \int_{-\pi}^{\pi} e^{it\lambda} \mathrm{d}Z(\lambda)$. Then*

$$\frac{1}{T} \sum_{t=1}^{T} \xi_t \to_{L^2} Z(\{0\})$$

and

$$\frac{1}{T} \sum_{k=1}^{T} \gamma(k) \to F(\{0\}).$$

Proof. Since $\frac{1}{T}\sum_{t=1}^{T} e^{it\lambda} \to 1_{\{0\}}(\lambda)$ in $L^2((-\pi,\pi],\mathcal{B}((-\pi,\pi]),F)$, and $\left|\frac{1}{T}\sum_{t=1}^{T} e^{it\lambda}\right| \leq 1$ for all T, therefore we have

$$\frac{1}{T}\sum_{t=1}^{T}\xi_t = \frac{1}{T}\int_{-\pi}^{\pi}\sum_{t=0}^{T-1} e^{it\lambda}\mathrm{d}Z(\lambda) \to_{L^2(\mathbb{P})} \int_{-\pi}^{\pi} 1_{\{0\}}(\lambda)\mathrm{d}Z(\lambda) = Z(\{0\}).$$

We can prove the result for autocovariance function similarly. ■

Note that if the spectral distribution F is continuous at 0, then $F(\{0\}) = Z(\{0\}) = 0$, and therefore the two limits are zero. A corollary of the above theorem is that $\frac{1}{T}\sum_{k=1}^{T}\gamma(k) \to 0$ is a sufficient and necessary condition for $\frac{1}{T}\sum_{t=1}^{T}\xi_t \to_{L^2} 0$. If $\mathbb{E}\xi_t = \mu$, then $\frac{1}{T}\sum_{k=1}^{T}\gamma(k) \to 0$ or the spectral distribution F continuous at 0 is a sufficient and necessary condition for $\frac{1}{T}\sum_{t=1}^{T}\xi_t \to_{L^2} \mu$. This gives a condition by which the sample mean estimator is consistent (in mean square, and therefore in probability) for the true mean.

Also note that if $\mathbb{E}\xi_t = 0$, and $Z(\{0\}) \neq 0$, then there is a random variable $\zeta = Z(\{0\})$ such that $\xi_t = \zeta + \eta_t$, $\eta_t = \int_{-\pi}^{\pi} e^{i\lambda t}\tilde{Z}(\mathrm{d}\lambda)$, $\tilde{Z}(\{0\}) = 0$, and $\frac{1}{T}\sum_{t=1}^{T}\xi_t \to_{L^2} \zeta$ as $T \to \infty$.

3.3 Estimating the Spectral Densities

In this section we first introduce the periodogram and its properties. Our spectral density estimator will be based on smoothing the periodogram.

Let $X = (X_1, X_2, \ldots, X_T)$ be the data. We may view any realization of X as elements in the space $\mathbb{C}^T$ over the field of complex numbers. Consider the vector of the form $v_j = \frac{1}{\sqrt{T}}(e^{i\omega_j}, e^{2i\omega_j}, \ldots, e^{iT\omega_j})', j \in \mathbb{Z}$. Let $\omega_j = \frac{2\pi j}{T}$ and $J_T = \{j \in \mathbb{Z} | -\pi < \omega_j \leq \pi\}$. Then $\{v_j\}_{j\in J_T}$ is an orthonormal basis of $\mathbb{C}$, and we therefore may express any $x = (x_1, \ldots, x_T)$ in $\mathbb{C}^T$ as

$$x = \sum_{j\in J_T} a_j v_j$$

where

$$a_j = \langle x, v_j\rangle = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_j e^{-it\omega_j}.$$

The mapping from x to $\{a_j\}_{j\in J_T}$ is called the *discrete Fourier transformation* of x.

Definition 3.14. Let $x \in \mathbb{C}^T$ and $\{a_j\}_{j\in J_T}$ be the discrete Fourier transformation of x. The *periodogram* $I(\omega_j)$ of x at frequency $\omega_j = \frac{2\pi j}{T}, j \in J_T$ is defined to be

$$I(\omega_j) = |a_j|^2 = \frac{1}{T}\left|\sum_{t=1}^{T} x_t e^{-it\omega_j}\right|^2.$$

Since $\|x\|^2 = \sum_{j\in J_T} I(\omega_j)$, the decomposition can be viewed as a form of "variance decomposition analysis".

Theorem 3.15. *Let $x \in \mathbb{C}^T$ and $I(\omega_j)$ its periodogram. Then for any $\omega_j \neq 0, j \in J_T$,*

$$I(\omega_j) = \sum_{k=-T+1}^{T-1} \gamma_x(k) e^{-ik\omega_j},$$

where $\gamma_x(k) = \frac{1}{T}\sum_{t=1}^{T-k}(x_{t+k} - m)\overline{(x_t - m)}$ for $k \geq 0$, $m = \frac{1}{T}\sum_{t=1}^{T} x_t$, and $\gamma_x(k) = \overline{\gamma_x(-k)}$ for $k < 0$.

Proof. Since $\sum_{t=1}^{T} e^{it\omega_j} = \sum_{t=1}^{T} e^{-it\omega_j} = 0$ for any $\omega_j \neq 0$, we have

$$\begin{aligned} I(\omega_j) &= \frac{1}{T}\left(\sum_{s=1}^{T}(x_s - m)\, e^{-is\omega_j}\right)\left(\sum_{t=1}^{T}\overline{(x_t - m)} e^{it\omega_j}\right) \\ &= \frac{1}{T}\sum_{1\leq s,t\leq t}(x_s - m)\overline{(x_t - m)} e^{-i(s-t)\omega_j} \\ &= \sum_{k=-T+1}^{T-1}\gamma_x(k) e^{-ik\omega_j}. \end{aligned}$$

■

The similarity between the above equation and the representation of the spectral density in Theorem 3.3 hints that we may construct an estimator of the spectral density of a weakly stationary process with absolutely summable autocovariance function based on the periodogram, i.e., the discrete Fourier transformation of the data. Since the discrete Fourier transformation defines the periodogram for discrete frequencies, in order to estimate spectral density, we extend the periodogram for all frequencies in $(-\pi, \pi]$.

Let $X_1, X_2, \ldots, X_T$ be a real time series. The periodogram $I_T(\omega), \omega \in (-\pi, \pi]$ is defined as follows: Let

$$I_T(\omega) = \begin{cases} T\left|\frac{1}{T}\sum_{t=1}^{T} X_t\right|^2, & \text{if } \omega = 0, \\ \frac{1}{T}\sum_{k=-T+1}^{T-1}\sum_{t=1}^{T-|k|}\left(X_{t+|k|} - \frac{1}{T}\sum_{t=1}^{T} X_t\right)\left(X_t - \frac{1}{T}\sum_{t=1}^{T} X_t\right)e^{-ik\omega}, & \text{if } \omega = \frac{2\pi j}{T}, j \in J_T. \end{cases}$$

Then define

$$I_T(\omega) = \begin{cases} I_T(\frac{2\pi j}{T}), & \text{if } \frac{\pi(2j-1)}{T} < \omega \leq \frac{\pi(2j+1)}{T}, \omega \in [0, \pi] \\ I_T(-\omega), & \text{if } \omega \in (-\pi, 0). \end{cases}$$

Note that we extend $I_T(\omega)$ in a piece-wise constant way. We have the following results.

Theorem 3.16. *Let $\{X_t\}_{t=1}^T$ be a (real) weakly stationary time series with mean μ, absolutely summable autocovariance function $\gamma(\cdot)$, spectral density $f(\cdot)$, and periodogram $I_T(\omega)$, $\omega \in (-\pi, \pi]$.*

Then

$$\mathbb{E}I_T(0) - T\mu^2 \to 2\pi f(0)$$

and

$$\mathbb{E}I_T(\omega) \to 2\pi f(\omega) \quad \text{if } \omega \neq 0$$

as $T \to \infty$.

In addition, if $\mu = 0$, *then* $\mathbb{E}I_T(\omega)$ *converges uniformly to* $2\pi f(\omega)$ *on* $(-\pi, \pi]$.

Proof. Write $g(T, \omega)$ to be the multiple of $\frac{2\pi}{T}$ closest to ω. Note that $g(T, \omega) \to \omega$ as $T \to \infty$. Also note that for $\omega = \frac{2\pi j}{T}, j \in J_T$, $I_T(\omega)$ can be equivalently written as

$$\frac{1}{T} \sum_{k=-T+1}^{T-1} \sum_{t=1}^{T-|k|} \left(X_{t+|k|} - \mu\right)\left(X_t - \mu\right).$$

The results then follows easily. Uniform convergence follows from the uniform convergence of $g(T, \omega)$ and the uniform continuity of $f(\cdot)$. ■

The above theorem suggests that $\frac{I_T(\omega)}{2\pi}$ (with small modification at $w = 0$) may serve as an asymptotically unbiased estimator of the spectral density f. However, as the next theorem shows, it is not a consistent estimator.

Theorem 3.17. *Let* $X_t \sim$ *i.i.d.*$(0, \sigma^2)$ *be a real sequence of random variables, and let* $f(\lambda) = \frac{\sigma^2}{2\pi}$ *be its spectral density and* $I_T(\cdot)$ *be its periodogram as defined above. Then for any* $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_n < \pi$, $\frac{1}{2}(\frac{I_T(\lambda_1)}{2\pi f(\lambda_1)}, \frac{I_T(\lambda_2))}{2\pi f(\lambda_2)}, \ldots, \frac{I_T(\lambda_n)}{2\pi f(\lambda_n)})$ *converges in distribution to a vector of independent* χ^2_2*-distributed random variables.*

Proof. For $\lambda > 0$, we have

$$I_T(\lambda) = \left|\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t e^{-itg(T,\lambda)}\right|^2 = \left|\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \cos(tg(T,\lambda))\right|^2 + \left|\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \sin(tg(T,\lambda))\right|^2.$$

We therefore consider the joint distribution of

$$Z_T = \left(\left\{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \cos(tg(T,\lambda_j))\right\}_{j=1}^{n}, \left\{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \sin(tg(T,\lambda_j))\right\}_{j=1}^{n}\right).$$

We first look at $Y_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \cos(tg(T,\lambda_j))$. Note that $\mathrm{Var}(Y_T) = \frac{\sigma^2}{2}$ since X_t is i.i.d. and $\sum_{t=1}^{T} \cos^2(t(g(T,\lambda_j))) = \sum_{t=1}^{T} \left(\frac{e^{itg(T,\lambda_j)} + e^{-itg(T,\lambda_j)}}{2}\right)^2 = \sum_{t=1}^{T} \frac{e^{2itg(T,\lambda_j)} + e^{-2itg(T,\lambda_j)} + 2}{4} = \frac{T}{2}$. Also, the Lindeberg condition holds by

$$\lim_{T\to\infty} \sum_{t=1}^{T} \mathbb{E}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \cos(tg(T,\lambda_j))\right]^2 1_{\left\{\left|\frac{1}{\sqrt{T}} X_t \cos(tg(T,\lambda_j))\right| > \epsilon\right\}}$$

$$\leq \lim_{T\to\infty} \mathbb{E}X_t^2 1_{\{|X_t|\}>\epsilon\sqrt{T}} = 0.$$

Therefore a central limit theorem holds for Y_T and $Y_T \to_d \mathbb{N}(0, \frac{\sigma^2}{2})$. We can follow similar procedures and use the Cramer-Wold device to show that $Z_T \to_d \mathbb{N}(0, \frac{\sigma^2}{2} I_{2n})$. The results then follows immediately. ■

The above theorem shows that even for simple processes, $\frac{I_T(\lambda)}{2\pi} = \frac{1}{2\pi}\sum_{k=-T+1}^{T-1} \hat{\gamma}(k) e^{-i\lambda k}$ converges to some non-degenerate random variables instead of the true spectral density $f(\lambda)$. Therefore, $\frac{I_T(\lambda)}{2\pi}$ is not a consistent estimator of $f(\lambda)$. The result can be extended to cases where X_t is not necessarily i.i.d., but a (possibly dependent and heterogeneous) sequence such that a central limit theorem holds for $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t e^{-itg(T,\lambda_j)}$. This holds when the process $\{X_t\}$ satisfies some mixing conditions, mixingale conditions, or martingale conditions. See Section 2.2.2 for details.

To improve the estimation performance, we need to smooth the periodogram. In the following we state a theorem from Parzen (1957). For properly chosen b_T, which is a function of the sample size T, and function $w(\cdot)$, we define the estimator of spectral density by

$$\hat{f}(\lambda) = \frac{1}{2\pi} \sum_{|k|<T} e^{-ik\lambda} w(b_T k) \hat{\gamma}(k). \tag{3.3}$$

Theorem 3.18. *Let X_t be a fourth-order stationary time series, $\gamma(\cdot)$ be its autocovariance function, and $f(\cdot)$ be its spectral density. Let f be estimated by $\hat{f}$ as in (3.3). Suppose that for some $q > 0$, $\sum_{k=-\infty}^{\infty} |k|^q |\gamma(k)| < \infty$, $\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\sum_{t=-\infty}^{\infty} |Q(r,s,t)|$ where $Q(r,s,t)$ is the joint cumulant of (X_0, X_r, X_s, X_t), $w : \mathbb{R} \to \mathbb{R}_+$ is an even, bounded, square integrable function such that $w(0) = 1$ and that for every b and T we have $b\sum_{|k|<T} w(bk) \leq C(bT)^{1/2-\epsilon}$ for some $\epsilon > 0$. Suppose that there is a largest positive number r (could be infinite, meaning that the required condition holds for all positive number r) such that $w^{(r)} = \lim_{z\to 0} \frac{1-w(z)}{|z|^r}$ is finite and non-zero and let $q \leq r$. Let b_T be a sequence such that $b_T \to 0$, $b_T T \to \infty$ and $0 < \lim_{T\to\infty} b_T^{1+2q} T < \infty$. Let $f^{(q)}(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} |k|^q \gamma(k) e^{-i\lambda k}$. Then we have that*

$$\lim_{T\to\infty} b_T^{-q} \left|\mathbb{E}\hat{f}(\lambda) - f(\lambda)\right| = \begin{cases} C\left|w^{(r)} f^{(r)}(\lambda)\right|, & \text{if } q = r, \\ 0, & \text{if } q < r, \end{cases}$$

and that

$$\lim_{T\to\infty} b_T T\, Cov(\hat{f}(\lambda_1), \hat{f}(\lambda_2)) = \begin{cases} 0, & \text{if } \lambda_1 \neq \pm\lambda_2, \\ f(\lambda_1)^2 \int_{-\infty}^{\infty} w^2(x)\mathrm{d}x, & \text{if } \lambda_1 = \pm\lambda_2 \neq 0, \\ 2f(\lambda_1)^2 \int_{-\infty}^{\infty} w^2(x)\mathrm{d}x, & \text{if } \lambda_1 = \lambda_2 = 0. \end{cases}$$

Proof. The proof here mainly follows Parzen (1957) and Hannan (1970, p. 280, Theorem 9). We

cite some of the results from Theorem 2.33 in the following. For $k \geq 0$, we can write

$$\begin{aligned}\hat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T-k}(y_{t+k} - \mu + \mu - \hat{\mu})(y_t - \mu + \mu - \hat{\mu}) \\ &= D(k) + \left(1 - \frac{k}{T}\right)\gamma(k) + R(k)\end{aligned}$$

where

$$D(k) = \frac{1}{T}\sum_{t=1}^{T-k}[(y_{t+k} - \mu)(y_t - \mu) - \gamma(k)]$$

and

$$R(k) = \frac{1}{T}\sum_{t=1}^{T-k}(y_{t+k} - \mu)(\mu - \hat{\mu}) + \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \mu)(\mu - \hat{\mu}) + \left(1 - \frac{k}{T}\right)(\mu - \hat{\mu})^2.$$

We have that $\mathbb{E}R(k)^2 \leq CT^{-2}$ for all k. Also, we have that for $k_1, k_2 \geq 0$,

$$\begin{aligned}\mathbb{E}D(k_1)D(k_2) = \frac{1}{T}\sum_{k=-(T-1-k_1)}^{T-1-k_2} & u(k, k_1, k_2)[Q(k_1, k, k + k_2) + \gamma(k)\gamma(k + k_2 - k_1) \\ & + \gamma(k + k_2)\gamma(k - k_1)]\end{aligned}$$

where

$$u(k, k_1, k_2) = \begin{cases} 1 - \frac{\max(k_1, k_2 + k)}{T}, & \text{if } k \geq 0, \\ 1 - \frac{\max(k_1 + |k|, k_2)}{T}, & \text{if } k < 0. \end{cases}$$

For convenience, we write

$$\begin{aligned}\mathbb{E}D(k_1)D(k_2) = \frac{1}{T}\sum_{k=-\infty}^{\infty} & u(k, k_1, k_2)[Q(k_1, k, k + k_2) + \gamma(k)\gamma(k + k_2 - k_1) \\ & + \gamma(k + k_2)\gamma(k - k_1)]\end{aligned}$$

where

$$u(k, k_1, k_2) = \begin{cases} 0, & \text{if } k > T - 1 - k_2, \\ 1 - \frac{\max(k_1, k_2 + k)}{T}, & \text{if } 0 \leq k \leq T - 1 - k_2, \\ 1 - \frac{\max(k_1 + |k|, k_2)}{T}, & \text{if } -(T - 1 - k_1) < k < 0, \\ 0, & \text{if } k \leq -(T - 1 - k_1). \end{cases}$$

Now we proceed to prove the theorem. First, by Minkowski inequality, we have that

$$\lim_{T\to\infty} b_T T \mathbb{E}\left|\sum_{|k|<T} e^{-i\lambda k} w(b_T k) R(k)\right|^2$$

$$\leq \lim_{T\to\infty} b_T T \left(\sum_{|k|<T} \left(\mathbb{E} \left| e^{-i\lambda k} w(b_T k) R(k) \right|^2 \right)^{1/2} \right)^2$$
$$\leq \lim_{T\to\infty} b_T T \left(\sum_{|k|<T} |w(b_T k)| \left(\mathbb{E} R(k)^2 \right)^{1/2} \right)^2$$
$$\leq \lim_{T\to\infty} C(b_T T)^{-2\epsilon} = 0.$$

Note that the convergence is uniformly in λ.

Next, write

$$b_T^{-q}(\mathbb{E}\hat{f}(\lambda) - f(\lambda)) = R_1 + R_2 + R_3 + B$$

where

$$R_1 = \frac{b_T^{-q}}{2\pi} \sum_{|k|<T} e^{-i\lambda k} w(b_T k) R(k),$$

$$R_2 = -\frac{b_T^{-q}}{2\pi T} \sum_{|k|<T} e^{-i\lambda k} |k| \, w(b_T k) \gamma(k),$$

$$R_3 = -\frac{b_T^{-q}}{2\pi} \sum_{|k|\geq T} e^{-i\lambda k} \gamma(k),$$

and

$$B = -\frac{b_T^{-q}}{2\pi} \sum_{|k|<T} e^{-i\lambda k} (1 - w(b_T k)) \gamma(k).$$

By what we have just shown, we see that $R_1 \to_{L^2} 0$ uniformly in λ. Since w is bounded, if $q \geq 1$, then

$$|R_2| \leq \frac{C}{2\pi T b_T^q} \sum_{|k|<T} |k| \, |\gamma(k)| \leq \frac{C}{2\pi T b_T^q} \sum_{|k|<T} |k|^q \, |\gamma(k)| \to 0.$$

If $q < 1$, then

$$|R_2| = \frac{C}{2\pi (b_T T)^q} \sum_{|k|<T} \left(\frac{|k|}{T} \right)^{1-q} |k|^q \, |\gamma(k)| \leq \frac{C}{2\pi (b_T T)^q} \sum_{|k|<T} |k|^q \, |\gamma(k)| \to 0.$$

Therefore $R_2 \to 0$ uniformly in λ. Write

$$|R_3| = \frac{1}{2\pi (b_T T)^q} \sum_{|k|\geq T} \left(\frac{T}{|k|} \right)^q |k|^q \, |\gamma(k)| \, .$$

Since $\frac{T}{|k|} \to 1$, $R_3 \to 0$ uniformly in λ.

For the essential bias term B, note that

$$B = -\frac{1}{2\pi} \sum_{|k|<T} (b_T |k|)^{r-q} \left(\frac{1 - w(b_T k)}{|b_T k|^r} \right) |k|^q \gamma(k) e^{-i\lambda k},$$

it is then easy to see that

$$\lim_{T\to\infty} |B| = \begin{cases} C \left| w^{(r)} f^{(r)}(\lambda) \right|, & \text{if } q = r, \\ 0, & \text{if } q < r \end{cases}$$

uniformly in λ. We have thus established that

$$\lim_{T\to\infty} b_T^{-q} \left| \mathbb{E}\hat{f}(\lambda) - f(\lambda) \right| = \begin{cases} C \left| w^{(r)} f^{(r)}(\lambda) \right|, & \text{if } q = r, \\ 0, & \text{if } q < r. \end{cases}$$

Next, for non-negative λ_1 and λ_2,

$$\begin{aligned}
&\lim_{T\to\infty} b_T T \mathrm{Cov}(\hat{f}(\lambda_1), \hat{f}(\lambda_2)) \\
&= \lim_{T\to\infty} \frac{b_T T}{4\pi^2} \mathbb{E} \left(\sum_{|k|<T} e^{-i\lambda_1 k} w(b_T k) D(k) \right) \left(\sum_{|k|<T} e^{-i\lambda_2 k} w(b_T k) D(k) \right) \\
&= \lim_{T\to\infty} \frac{b_T T}{4\pi^2} \sum_{k_1=-(T-1)}^{T-1} \sum_{k_2=-(T-1)}^{T-1} e^{-i\lambda_1 k_1} e^{-i\lambda_2 k_2} w(b_T k_1) w(b_T k_2) \mathbb{E} D(k_1) D(k_2) \\
&= \lim_{T\to\infty} \frac{b_T}{4\pi^2} \sum_{k_1=-(T-1)}^{T-1} \sum_{k_2=-(T-1)}^{T-1} \sum_{k=-\infty}^{\infty} e^{-i\lambda_1 k_1} e^{-i\lambda_2 k_2} w(b_T k_1) w(b_T k_2) \\
&\qquad \cdot u(k, |k_1|, |k_2|)[Q(|k_1|, k, k + |k_2|) + \gamma(k)\gamma(k + |k_2| - |k_1|) + \gamma(k + |k_2|)\gamma(k - |k_1|)].
\end{aligned}$$

Since the sine and cosine functions, $w(\cdot)$ and $u(\cdot,\cdot,\cdot)$ are bounded, and that $Q(r,s,t)$ is summable, we may ignore the term $Q(k_1, k, k + k_2)$ in the above expression. Also we change z for $|k_1| - |k_2|$ and rewrite the second part as

$$\begin{aligned}
&\lim_{T\to\infty} b_T T \mathrm{Cov}(\hat{f}(\lambda_1), \hat{f}(\lambda_2)) \\
&= \lim_{T\to\infty} \frac{b_T}{4\pi^2} \sum_{z=-2(T-1)}^{T-1} \sum_{k_2=-(T-1)}^{T-1} \sum_{k=-\infty}^{\infty} e^{-i\lambda_1 (z+|k_2|)} e^{-i\lambda_2 k_2} w(b_T(z + |k_2|)) w(b_T k_2) \\
&\qquad \cdot \left[u(k, |z + |k_2||, |k_2|) + u(k - |k_2|, |z + |k_2||, |k_2|) \right] \left[\gamma(k)\gamma(k - z) \right] \\
&= \lim_{T\to\infty} \frac{1}{4\pi^2} \sum_{z=-2(T-1)}^{T-1} \sum_{k=-\infty}^{\infty} \gamma(k)\gamma(k - z) e^{-i\lambda_1 z} \sum_{k_2=-(T-1)}^{T-1} e^{-i\lambda_1 |k_2|} e^{-i\lambda_2 k_2} \\
&\qquad \cdot b_T w(b_T(z + |k_2|)) w(b_T k_2) u^*(k, k_2, z).
\end{aligned}$$

where

$$u^*(k,k_2,z)=u(k,|z+|k_2||,|k_2|)+u(k-|k_2|,|z+|k_2||,|k_2|).$$

Note that $\lim_{T\to\infty}u^*(k,k_2,z)=2$.

Let

$$L(z,k,k_2)=e^{-i\lambda_1|k_2|}e^{-i\lambda_2k_2}b_Tw(b_T(z+|k_2|))w(b_Tk_2)u^2(k,k_2,z).$$

Note that

$$\begin{aligned}&\left(\sum_{k_2=-(T-1)}^{T-1}b_Tw(b_T(z+|k_2|))w(b_Tk_2)\right)^2\\ \le&\left(\sum_{k_2=-(T-1)}^{T-1}w^2(b_T(z+|k_2|))b_T\right)\left(\sum_{k_2=-(T-1)}^{T-1}w^2(b_Tk_2)b_T\right)\\ \le&\left(\sum_{k_2=-\infty}^{\infty}w^2(b_Tk_2)b_T\right)^2=\left(\int_{-\infty}^{\infty}w^2(x)\mathrm{d}x\right)^2.\end{aligned}$$

Then by dominate convergence theorem, $\sum L(z,k,k_2)$ converges uniformly in z,k. To evaluate the value of this sum, we write $\lim_{T\to\infty}\sum_{k_2=-(T-1)}^{T-1}L(z,k,k_2)$ as

$$\begin{aligned}&\lim_{T\to\infty}\sum_{k_2=1}^{T-1}(e^{-i(\lambda_2-\lambda_1)k_2}+e^{-i(\lambda_1+\lambda_2)k_2})b_Tw(b_T(z+k_2))w(b_Tk_2)u^2(k,k_2,z)\\ =&\lim_{T\to\infty}2\int_0^{b_TT}(e^{-i\frac{\lambda_2-\lambda_1}{b_T}x}+e^{-i\frac{\lambda_2+\lambda_1}{b_T}x})w^2(x)\mathrm{d}x\end{aligned}$$

Recall that we consider the case when $\lambda_1\ge0$ and $\lambda_2\ge0$. By Riemann-Lebesgue Lemma, if $\lambda_1\neq\lambda_2$, the above limit is zero. If $\lambda_1=\lambda_2\neq0$, the above limit is $\int_{-\infty}^{\infty}w^2(x)dfx$. If $\lambda_1=\lambda_2=0$, the above limit is $2\int_{-\infty}^{\infty}w^2(x)\mathrm{d}x$.

Noting that $\frac{1}{2\pi}\sum_{k=-\infty}^{\infty}e^{-i\lambda_1k}\gamma(k)=f(\lambda_1)$, we have that

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{4\pi^2}\sum_{z=-2(T-1)}^{T-1}\sum_{k=-\infty}^{\infty}\gamma(k)\gamma(k-z)e^{-i\lambda_1z}\sum_{k_2=-(T-1)}^{T-1}e^{-i\lambda_1|k_2|}e^{-i\lambda_2k_2}\\ &\quad\cdot b_Tw(b_T(z+|k_2|))w(b_Tk_2)u^*(k,k_2,z)\\ =&\lim_{T\to\infty}\sum_{z=-2(T-1)}^{T-1}\sum_{k=-\infty}^{\infty}e^{-i\lambda_1k}\gamma(k)e^{-i\lambda_1(z-k)}\gamma(k-z)\sum_{k_2=-(T-1)}^{T-1}e^{-i\lambda_1|k_2|}e^{-i\lambda_2k_2}\\ &\quad\cdot b_Tw(b_T(z+|k_2|))w(b_Tk_2)u^*(k,k_2,z)\\ =&Cf(\lambda_1)^2\int_{-\infty}^{\infty}w^2(x)\mathrm{d}x\end{aligned}$$

where

$$C = \begin{cases} 0, \text{ if } \lambda_1 \neq \lambda_2, \\ 1, \text{ if } \lambda_1 = \lambda_2 \neq 0, \\ 2, \text{ if } \lambda_1 = \lambda_2 = 0. \end{cases}$$

Note that we derive the above result in the case when both λ_1 and λ_2 are non-negative. Given that $f(\lambda)$ is symmetric, it is easy to adapt the results to general λ_1, λ_2. ■

We make the following remarks regarding the above theorem.

(a) If we redefine the periodogram $I_T(\omega)$

$$I_T(\omega) = \sum_{k=-T+1}^{T-1} e^{-ik\omega} \hat{\gamma}(k)$$

so that it is not piece-wise constant anymore, then we have

$$\hat{\gamma}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\omega} I_T(\omega) \mathrm{d}\omega.$$

Let

$$W_T(\omega) = \frac{1}{2\pi} \sum_{|k|<T} w(b_T k) e^{-ik\omega}.$$

Then it is easy to verify that

$$\hat{f}(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} W_T(\lambda - \omega) I_T(\omega) \mathrm{d}w.$$

(b) We call $w(\cdot)$ the *lag window* of the spectral density estimator and call $W_N(\cdot)$ the spectral window. The following are a list of popular window functions.

(1) The rectangular or truncated window. The lag window function is given by

$$w(x) = \begin{cases} 1, & \text{if } |x| \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

For this window, we have

$$W_T(\omega) = \frac{1}{2\pi} \frac{\sin\left(\frac{1}{b_T} + \frac{1}{2}\right)\omega}{\sin \frac{1}{2}\omega},$$

which is the Dirichlet kernel.

(2) The Bartlett window or the triangular window. The lag window function is given by

$$w(x) = \begin{cases} 1 - |x|, & \text{if } |x| \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

For this window, we have

$$W_T(\omega) = \frac{b_T}{2\pi} \frac{\sin^2 \frac{\omega}{2b_T}}{\sin^2 \frac{\omega}{2}} = \frac{b_T}{2\pi} \frac{1 - \cos \frac{\omega}{b_T}}{1 - \cos \omega},$$

which is the Fejér kernel.

(3) The Daniell window. The lag window function is given by

$$w(x) = \begin{cases} \frac{\sin \pi x}{\pi x}, & \text{if } |x| \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

For this window, we have

$$W_T(x) = \begin{cases} \frac{1}{2\pi b_T}, & \text{if } |\omega| \leq b_T \pi, \\ 0, & \text{otherwise.} \end{cases}$$

(4) The Blackman-Tukey window. The lag window function is given by

$$w(x) = \begin{cases} 1 - 2a + 2a \cos x, & \text{if } |x| \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

for some a. For this window we have

$$W_T(\omega) = aD(\omega - b_T \pi) + (1 - 2a)D(\omega) + aD(\omega + b_T \pi)$$

where D is the Dirichlet kernel. When $a = 0.25$, this window function is called the Tukey-Hanning window, and when $a = 0.23$, this window function is called the Tukey-Hamming window.

(5) The Parzen window. The lag window function is given by

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, & \text{if } |x| < 1/2, \\ 2(1 - |x|)^3, & \text{if } 1/2 \leq |x| \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

For this window we have

$$W_T(\omega) = \frac{6b_T^3}{\pi} \frac{\sin^4 \frac{\omega}{4b_T}}{\sin^4 \frac{\omega}{2}}.$$

For more discussions, see Brockwell and Davis (1991, p. 359-362), Hannan (1970, p. 275-280) and Priestley (1981).

(c) The theorem shows that by appropriately choosing b_T and w, the spectral density estimator $\hat{f}(\lambda)$, under the assumptions in the theorem, is consistent. Its asymptotic mean square error is of order $b_T^{2q} + (b_T T)^{-1}$. This implies that that the optimal bandwidth b_T is given by $b_T = CT^{-\frac{1}{1+2q}}$. Then we have that

$$\lim_{T\to\infty} T^{\frac{2q}{1+2q}} \mathbb{E}(\hat{f}(\lambda) - f(\lambda))^2 = C_1 \left| w^{(r)} f^{(r)}(\lambda) \right| 1_{q=r} + C_2 f^2(\lambda) \int_{-\infty}^{\infty} w^2(x) \mathrm{d}x.$$

For $q > 0$, $\frac{2q}{1+2q} \in (0, 1)$. This implies that by choosing appropriate window function, convergence rates between 0 and $\sqrt{T}$ could be attained, with the highest attainable rate determined by the largest q such that $\sum |k|^q \gamma(k) < \infty$ holds.

(d) Absolute summability of fourth order cumulants holds when the series satisfies some α-mixing conditions. See notes after Theorem 2.33.

(e) It is also possible to obtain central limit theorems for spectral density estimator. See Rosenblatt (1984) for example.

3.4 Estimating Long-Run Variances

The spectral density estimator could be used to estimate the long-run variance of a weakly stationary time series. If a central limit theorem holds for a weakly stationary time series $\{y_t\}$, the asymptotic variance must be its long run variance $J = \sum_{k=-\infty}^{\infty} \gamma(k)$. It is therefore necessary to develop an estimator for the long run variance. If we observe $y_1, y_2, \ldots, y_T$, a naive estimator for the long run variance is $\sum_{k=-(T-1)}^{T-1} \hat{\gamma}(k)$ where $\hat{\gamma}(k) = \frac{1}{T} \sum_{t=k+1}^{T} (y_t - \hat{\mu})(y_{t-k} - \hat{\mu})'$ is the sample average estimator of the autocovariance function. However, just like the periodogram, this naive estimator is not consistent. To see this point, we assume without loss of generality that y_t is mean zero. Then

$$\begin{aligned}
\sum_{k=-(T-1)}^{T-1} \hat{\gamma}(k) &= \frac{1}{T} \sum_{k=-(T-1)}^{T-1} \sum_{t=k+1}^{T} y_t y_{t-k} \\
&= \frac{1}{T} \left(\sum_{t=1}^{T} y_t \right)^2 + o_p(1) \\
&= \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_t \right)^2 + o_p(1).
\end{aligned}$$

If a central limit theorem holds for $\{y_t\}$, then the naive estimator converges to a squared normal distribution instead of the deterministic value $\sum_{k=-\infty}^{\infty} \gamma(k)$. That is, the naive estimator is not consistent.

To solve this issue, we utilize the relationship that if $\{X_t\}$ is weakly stationary with absolutely summable autocovariance function $\gamma(\cdot)$ and spectral density $f(\cdot)$, then $\sum_{k=-\infty}^{\infty} \gamma(k) = 2\pi f(0)$.

Therefore we may estimate the long run variance of X_t by $2\pi\hat{f}(0)$. As long as $\hat{f}$ is consistent, the long run variance estimator is consistent.

4 Linear Processes

Starting from this chapter, we focus on time series of (real) random variables. In this chapter, we develop a decomposition of a weakly stationary time series, which motivates the study of linear processes. Before we proceed to the decomposition, we first introduce some theory of Hilbert spaces.

4.1 Hilbert Spaces

We present in this section the theory of Hilbert spaces over the field of complex numbers. This presentation can be easily accommodated to the case of Hilbert spaces over the field of real numbers. For proofs of theorems in this section, see, e.g., Rudin (1987, Chapter 4).

Definition 4.1. A complex vector space H is called an *inner product space* if to each ordered pair of vectors x and $y \in H$ there is associated a complex number $\langle x, y\rangle$, called the *inner product* of x and y, such that

1. $\langle y, x\rangle = \overline{\langle x, y\rangle}$.
2. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ for any $z \in H$.
3. $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$ if α is a scalar.
4. $\langle x, x\rangle \geq 0$ and equality holds only if $x = 0$.

Notice that the above definition implies that

1. $\langle x, \alpha y\rangle = \bar{\alpha}\langle x, y\rangle$.
2. $\langle z, x + y\rangle = \langle z, x\rangle + \langle z, y\rangle$.

We may define $\|x\|$, the *norm* of the vector $x \in H$, to be

$$\|x\| = \sqrt{\langle x, x\rangle}. \tag{4.1}$$

Theorem 4.2. *(The Schwarz Inequality) If H is an inner product space and $x, y \in H$, then*

$$|\langle x, y\rangle| \leq \|x\| \, \|y\| .$$

Theorem 4.3. *If H is an inner product space and $x, y \in H$, then*

$$\|x + y\| \leq \|x\| + \|y\| .$$

It follows from the triangle inequality that

$$\|x - z\| \leq \|x - y\| + \|y - z\| \qquad (x, y, z \in H.) \tag{4.2}$$

Definition 4.4. Let H be an inner product space with norm $\|\cdot\|$ defined by (4.1). Inequality (4.2) suggests that we may define the *distance* between x and y in H to be $\|x - y\|$. It is easy to verify

that with this distance, H is a *metric space*. If this metric space is complete, that is, if every Cauchy sequence converges in H, then H is called a *Hilbert space*.

Theorem 4.5. *Let H be a Hilbert space. For any fixed $y \in H$, the mappings*

$$x \to \langle x, y\rangle, \quad x \to \langle y, x\rangle, \quad x \to \|x\|$$

are continuous functions on H.

Definition 4.6. Let H be a Hilbert space and M be a subset of H. The *orthogonal complement* of M, denoted by $M^\perp$, is the set of all elements $x \in H$ such that $\langle x, y\rangle = 0$ for all $y \in M$.

If $\langle x, y\rangle = 0$ for some $x, y \in H$, then we say that x is orthogonal to y, and write $x \perp y$.

Theorem 4.7. *Let H be a Hilbert space and $M \subset H$. Then $M^\perp$ is a closed subspace of H.*

Theorem 4.8. *Let H be a Hilbert space and M be a closed subspace of H. For any $x \in H$, we have the followings.*

(a) We may uniquely decompose x as

$$x = P_M x + P_{M^\perp} x$$

where $P_M x \in M$ and $P_{M^\perp} \in M^\perp$.

(b) $P_M x$ and $P_{M^\perp} x$ are the nearest points to x in M and $M^\perp$, respectively. That is, $\|x - P_M x\| = \min_{y \in P_M} \|x - y\|$, $\|x - P_{M^\perp} x\| = \min_{y \in M^\perp} \|x - y\|$.

(c) The mappings $P_M : H \to M$ and $P_{M^\perp} : H \to M^\perp$ are linear.

(d) $\|x\|^2 = \|P_M x\|^2 + \|P_{M^\perp} x\|^2$.

The mappings P_M are called the *orthogonal projections* onto the subspace M and onto the subspace $M^\perp$, respectively. Note that the linearity proposition implies that $P_{M^\perp} = I - P_M$.

4.2 Projections on Spaces Spanned by a Sequence

Theorem 4.9. *Let $\{x_n\}$ be a sequence in an inner product space H, and let $M_k = \bigvee_{i=1}^k x_i$. Then M_k is a Hilbert space.*

Proof. It is easy to see that M_k is an inner product space (inherits from H). Since it is finite-dimensional, it has an orthogonal basis $\{e_1, \cdots, e_j\}$ such that $\|e_i\| = 1$ for all $i = 1, \cdots, j$. Let $\{y_t\}, t = 1, 2, \cdots$ be a Cauchy sequence in M_k and let $(\alpha_{t1}, \cdots, \alpha_{tj})$ be the coordinates of y_t with respect to the orthogonal basis above.

For any y_m, y_n in M_k,

$$\|y_m - y_n\| = \sum_{s=1}^{j} (\alpha_{ms} - \alpha_{ns})^2.$$

Then $\{y_t\}$ is Cauchy implies that $\{\alpha_{ts}\}$ is Cauchy for each $s = 1, \cdots, j$. Since the complex plain $\mathbb{C}$ is complete, the sequence $\{\alpha_t s\}$ converges to a complex number, denoted by α_s for $s = 1, \cdots, j$.

Let

$$y = \sum_{s=1}^{j} \alpha_s e_s.$$

It is obvious that $\{y_t\}$ converges to y. Since $y \in M_k$, M_k is complete. Therefore, M_k is a Hilbert space. ■

Theorem 4.10. *Let $\{x_n\}$ be a sequence in a Hilbert space H, let $M_\infty = \overline{\bigvee_{i=1}^{k} x_i}$, the closure of the space of all finite linear combinations of elements in $\{x_n\}$. Then M_∞ is a Hilbert space.*

Proof. Let $\{z_n\}$ be a Cauchy sequence in M_∞. Since H is complete, $\{z_n\}$ converges to a point z in H. Since M_∞ is closed, if $z \notin M_\infty$, there is an ϵ-ball of z that does not intersect with M_∞. This implies that all but finitely many points of $\{z_n\}$ should lie in this ϵ-ball, which contradicts with the assumption that $\{z_n\}$ is a sequence in M_∞. ■

Theorem 4.11. *Let H be a Hilbert space and $\{x_j\}, j = 1, 2, \cdots$ be a sequence in H. Let $M_k, k = 1, 2, \cdots$ and M_∞ be defined as in Theorem 4.9 and 4.10. For any $z \in H$, let $\hat{z}$ be the orthogonal projection of z on M_∞ and $\hat{z}_n$ be the orthogonal projection of z on M_n. Then*

$$\lim_{n\to\infty} \|\hat{z} - \hat{z}_n\| = 0.$$

That is, $\{\hat{z}_n\}$ converges to $\hat{z}$.

Proof. By uniqueness of orthogonal projection, $\hat{z}_n$ is also the orthogonal projection of $\hat{z}$ on M_n. Since M_∞ is the closure of $\bigcup M_k$ and $M_k \subset M_{k+1}$, then for any $\hat{z} \in M_\infty$, there is a sequence $\{z_n\}$ with $z_n \in M_n$ such that $\|\hat{z} - z_n\| \to 0$ as $n \to \infty$. Since $\|\hat{z} - z_n\| \geq \|\hat{z} - \hat{z}_n\|$, it follows that $\|\hat{z} - \hat{z}_n\| \to 0$ as $n \to \infty$. ■

Definition 4.12. A set of vectors e_α in a Hilbert space H, where α runs through some index set A, is called *orthonormal* if for all $\alpha, \beta \in A$ we have

$$\langle e_\alpha, e_\beta \rangle = \begin{cases} 1, & \text{if } \alpha = \beta, \\ 0, & \text{if } \alpha \neq \beta. \end{cases}$$

Theorem 4.13. *If a sequence $\{e_i\}_{i=1}^{\infty}$ in a Hilbert space H is orthonormal, and $\hat{z}$ is the orthogonal projection of $z \in H$ on $\overline{\bigvee_{i=1}^{\infty} e_i}$, then $\hat{z}$ has the representation*

$$\hat{z} = \sum_{i=1}^{\infty} \theta_i e_i$$

where $\theta_i = \langle z, e_i \rangle$ and $\sum_{i=1}^{\infty} \theta_i^2 < \infty$.

Proof. Let $M_n = span\left(\{e_i\}_{i=1}^n\right)$ and

$$z_n = \sum_{i=1}^{n} \langle z, e_i \rangle e_i.$$

We shall prove that z_n is the orthogonal projection of $\hat{z}$ on M_n. First, notice that $z_n \in M_n$. For any $e_j, j = 1, \cdots, n$, we have

$$\begin{aligned}
\langle \hat{z} - z_n, e_j \rangle &= \langle \hat{z}, e_j \rangle - \sum_{i=1}^{n} \left\langle \langle z, e_i \rangle e_i, e_j \right\rangle \\
&= \langle \hat{z}, e_j \rangle - \sum_{i=1}^{n} \langle z, e_i \rangle \langle e_i, e_j \rangle \\
&= \langle \hat{z}, e_j \rangle - \langle z, e_j \rangle \langle e_j, e_j \rangle \\
&= \langle \hat{z}, e_j \rangle - \langle \hat{z}, e_j \rangle \langle e_j, e_j \rangle \\
&= 0.
\end{aligned}$$

The third equality follows from that $e_i \perp e_j$ if $i \neq j$, the fourth equality follows from that $\hat{z}$ is the orthogonal projection of z on $\overline{\bigvee_{i=1}^{\infty} e_i}$ and the last equality follows from that $\|e_j\| = 1$.

Now by Theorem 4.11, $z_n \to \hat{z}$, that is,

$$\lim_{n \to \infty} \sum_{i=1}^{n} \theta_i e_i = \hat{z}.$$

Then $\hat{z} = \sum_{i=1}^{\infty} \theta_i e_i$.

Since $z_n \to \hat{z}$, and $\|\cdot\|$ is continuous, $\|z_n\| \to \|\hat{z}\|$. Since $\|z_n\| = \sum_{i=1}^{n} \theta_i^2$, we have

$$\sum_{i=1}^{\infty} \theta_i^2 = \lim_{n \to \infty} \sum_{i=1}^{n} \theta_i^2 = \|\hat{z}\|.$$

The square summability of $\{\theta_i\}$ then follows from the fact that $\|\hat{z}\| < \infty$. ■

Theorem 4.14. *Suppose X and Y are orthogonal sub-Hilbert spaces of H. Then $X+Y$ is a Hilbert space.*

Proof. Let $\{z_i\}$ be a Cauchy sequence in $X + Y$. Since $z_i = x_i + y_i$ for some $x_i \in X, y_i \in Y$ and X and Y are orthogonal,

$$\|z_m - z_n\|^2 = \|x_m - x_n\|^2 + \|y_m - y_n\|^2.$$

This implies that $\{z_i\}$ is Cauchy only if $\{x_i\}$ and $\{y_i\}$ are both Cauchy. Since X and Y are complete, $\{x_i\}$ converges to some $x \in X$ and $\{y_i\}$ converges to some $y \in Y$. Then $\{z_i\} = \{x_i + y_i\}$ converges to $x + y$. Since $x + y \in X + Y$, our conclusion then follows. ■

Theorem 4.15. *Let X be a subspace of a Hilbert space H and let $X^\perp$ be the orthogonal complement of X in H. Then $X^\perp$ is a Hilbert space.*

Proof. Let y_i be a Cauchy sequence in $X^\perp$. Then for any $x \in X$, $\langle y_i, x\rangle = 0$. Since H is a Hilbert space, $\{y_i\}$ converges to some $y \in H$. Since $\langle \cdot, x\rangle$ is continuous for all fixed $x \in H$, $\langle y, x\rangle = 0$. Therefore $y_i \to y \in X^\perp$ and our conclusion follows. ■

Note that we did not require that X to be a Hilbert space in Theorem 4.15.

4.3 The Wold Decomposition Theorem

Let the underlying probability space be $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ of square integrable real random variables on Ω. The space is a Hilbert space in the a.s. sense. The inner product of two random variables ξ, η in this space is given by $\mathbb{E}\xi\eta$. For more information, see Billingsley (1995, Section 19). In this section from now on, we talk about random variables in the $\mathbb{P}$-a.s. sense, or, in the sense of the usual equivalent classes in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Now consider a time series $\{X_t\}$ such that $0 < \mathbb{E}X_t^2 < \infty$. $\{X_t\}$ could be viewed as sequence of points in the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Let $H_n(X) = \overline{\bigvee_{t=-\infty}^{n} X_t}$, $H(X) = \overline{\bigvee_{t=-\infty}^{\infty} X_t}$, and $H_{-\infty}(X) = \bigcap_{n=-\infty}^{\infty} H_n(X)$. Note that this space is closed. For any $\xi \in H(X)$, let $\hat{\xi}$ be the orthogonal projection of ξ on $H_{-\infty}(X)$. Then we may write $\xi = \hat{\xi} + (\xi - \hat{\xi})$. This implies that we may write

$$H(X) = H_{-\infty}(X) \oplus R(X)$$

where $R(X) = \{\xi - \hat{\xi} | \xi \in H(X)\}$.

Definition 4.16. A weakly stationary time series $X = \{X_t\}$ is called *deterministic* if $H(X) = H_{-\infty}(X)$, and is called *purely non-deterministic* if $H(X) = R(X)$.

If X is a deterministic series, then the whole series is completely predictable with certainty from an arbitrary distant past. In this sense we are using the word "deterministic".

Theorem 4.17. *Let $X = \{X_t\}$ be a weakly stationary time series. Then it has a decomposition $X_t = X_t^d + X_t^p$ such that $X^d = \{X_t^d\}$ is deterministic and $X^p = \{X_t^p\}$ is purely non-deterministic. Also, $\mathbb{E}X_t^d X_s^p = 0$ for any s, t.*

Proof. Let X_t^d be the orthogonal projection of X_t on $H_{-\infty}(X)$, and let $X_t^p = X_t - X_t^p$ for all t.

Since $X_t^p \perp H_{-\infty}(X)$ for all t, we have $H_{-\infty}(X^p) \perp H_{-\infty}(X)$. On the other hand, $X_n^p \in H_n(X)$, then $H_n(X_n^p) \in H_n(X)$, then $H_{-\infty}(X^p) \subset H_{-\infty}(X)$. We then have $H_{-\infty}(X^p) = \{0\}$. This implies that X^p is purely non-deterministic.

Since $H_n(X) \subset H_n(X^d) \oplus H_n(X^p)$, and $H_n(X^d) \subset H_n(X)$, $H_n(X^p) \subset H_n(X)$, we have that $H_n(X) = H_n(X^d) \oplus H_n(X^p)$. We therefore have $H_{-\infty}(X) \subset H_n(X^d) \oplus H_n(X^p)$ for all n. Since Since $X_t^p \perp H_{-\infty}(X)$ for all t, we have $H_{-\infty}(X) \subset H_n(X^d)$ for all n. This implies that $H_{-\infty}(X) \subset H_{-\infty}(X^d) \subset H(X^d)$. Since $X_t^d \in H_{-\infty}(X)$ for all t, we have $H_{-\infty}(X^d) \subset H_{-\infty}(X)$. This implies that $H_{-\infty}(X) = H_{-\infty}(X^d) = H(X^d)$. This shows that X^d is deterministic.

Since for each t, $X_t^p \perp H_{-\infty}(X) = H(X^d)$, we have that $\mathbb{E}X_t^d X_s^p = 0$ for all t, s. ■

Definition 4.18. Let $X = \{X_t\}$ be a weakly stationary time series with positive variance. A random sequence $\varepsilon = \{\varepsilon_t\}$ is an *innovation sequence* for X if ε is a unit variance white noise process and $H_t(X) = H_t(\varepsilon)$ for all t.

We may understand the innovation ε_{t+1} as the new information to $H_t(X)$ that is need to form $H_{t+1}(X)$.

Theorem 4.19. *A weakly stationary time series $X = \{X_t\}_{t\in\mathbb{Z}}$ with positive variance is purely non-deterministic if and only if there is an innovation sequence $\varepsilon = \{\varepsilon_t\}_{t\in\mathbb{Z}}$ and a sequence of real numbers $\{a_k\}_{k\in\mathbb{N}}$ with $\sum_{k=0}^{\infty} a_k^2 < \infty$ such that*

$$X_t = \sum_{k=0}^{\infty} a_k \varepsilon_{t-k} \quad a.s..$$

Proof. Necessity. Write $H_t(X) = H_{t-1}(X) \oplus B_n$. Obviously, B_t has dimension either zero or one. However, if B_t has dimension zero, then by stationarity B_s has dimension zero for all $s \in \mathbb{Z}$. This then implies that $H_t(X) = H_s(X)$ for all t, s, and consequently $H(X) = H_{-\infty}(X)$, contradicting with the assumption that X is purely non-deterministic. Therefore, the dimension of B_t is one, and we therefore let ε_t be an element in B_t such that $\mathbb{E}\varepsilon_t^2 = 1$.

For any t, we have

$$H_t = H_{t-k}(X) \oplus B_{t-k+1} \oplus \cdots \oplus B_t.$$

Note that $\varepsilon_{t-k+1}, \ldots, \varepsilon_t$ is an orthogonal basis in $B_{t-k+1} \oplus \cdots \oplus B_t$ and we therefore may represent

$$X_t = \sum_{i=0}^{k-1} a_i \varepsilon_{t-i} + \pi_{t-k}(X_t)$$

where π_{t-k} is the orthogonal projection onto $H_{t-k}(X)$, and $a_i = \mathbb{E}X_t X_{t-i}$. Note that a_i is independent of t because X is weakly stationary. Since $\{\varepsilon_{t-i}\}_{i=0}^{\infty}$ form an orthonormal sequence, by Bessel's inequality we have that $\sum_{i=0}^{\infty} a_i^2 < \infty$. Therefore $\sum_{i=0}^{\infty} a_i \varepsilon_{t-i}$ converges in mean square, and we only need to show $\pi_{t-k}(X_t) \to_{L^2} 0$ as $k \to \infty$.

Without loss of generality we may just consider the case $t = 0$. Write

$$\pi_{-k} = \pi_0 + \sum_{i=0}^{k} (\pi_{-i} - \pi_{-i+1}).$$

The $k+1$ terms in the right hand side of the equation are orthogonal, we then have

$$\sum_{i=0}^{k} \|(\pi_{-i} - \pi_{-i+1})(X_0)\|^2 = \left\| \left(\sum_{i=0}^{k} (\pi_{-i} - \pi_{-i+1}) \right)(X_0) \right\|^2 = \|\pi_{-k}(X_0) - \pi_0(X_0)\|^2 \leq 4\mathbb{E}X_0^2 < \infty,$$

then $\lim_{k\to\infty} \pi_{-k}(X_0)$ exists in mean square. Since $\pi_{-k}(X_0) \in H_{-k}(X)$ for each k, $H_{-k}(X)$ is

decreasing in k, we therefore have $\pi_{-s}(X_0) \in H_{-k}(X)$ for all $s > k$. Since $H_{-k}(X)$ is closed, we have $\lim_{s\to\infty} \pi_{-s}(X_0) \in H_{-k}(X_0)$. Since this holds for all k, we have $\lim_{k\to\infty} \pi_{-k}(X_0) \in \bigcap_{k\geq 0} H_{-k}(X) = H_{-\infty}(X)$. Since X is purely non-deterministic, $H_{-\infty}(X) = \{0\}$. Therefore we have $\pi_{t-k}(X_t) \to_{L^2} 0$.

Sufficiency. Let $X = \{X_t\}$ satisfy the representation. Then $H_t(X) \subset H_t(\varepsilon)$ and therefore $H_{-\infty}(X) \subset H_t(\varepsilon)$ for all t. Since $\varepsilon_{t+1} \perp H_t(\varepsilon)$, $\varepsilon_{t+1} \perp H_{-\infty}(X)$. This then implies that $H(\varepsilon) \perp H_{-\infty}(X)$. By the representation, ε is a orthonormal basis in $H(X)$. It then follows that $H_{-\infty}(X) = \{0\}$, which implies that X is purely non-deterministic. It is easy to show that this series is weakly stationary, and because it is non-deterministic, it must have variance greater than zero. ■

It follows from the proof of the theorem that X is a purely non-deterministic weakly stationary time series if and only if it admits a representation

$$X_t = \sum_{k=0}^{\infty} \tilde{a}_k \tilde{\varepsilon}_{t-k}$$

where $\tilde{\varepsilon}$ is a white noise process not necessarily satisfies $H_t(X) = H_t(\tilde{\varepsilon})$. Therefore, the above theorem gives a stronger result in terms of the necessary condition.

Now we obtain a full version of the Wold decomposition.

Theorem 4.20 (Wold Decomposition)**.** *Let $X = \{X_t\}_{t\in\mathbb{Z}}$ be a weakly stationary time series with positive variance. Then we may represent*

$$X_t = \sum_{k=0}^{\infty} a_k \varepsilon_{t-k} + W_t$$

where $W = \{W_t\}_{t\in\mathbb{Z}}$ is deterministic, $\varepsilon = \{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an innovation sequence, and $\sum_{k=0}^{\infty} a_k^2 < \infty$.

4.4 Linear Processes

The Wold decomposition theorem justifies the study of an important class of processes called *linear processes*. A process $\{X_t\}_{t\in\mathbb{Z}}$ is linear if it takes the form of

$$X_t = \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i} \tag{4.3}$$

where $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$. We usually need to impose some restrictions on the coefficients ϕ_i. One frequently used restriction is square summability:

$$\sum_{i=0}^{\infty} \phi_i^2 < \infty.$$

Sometimes we also work with the absolute summability condition

$$\sum_{i=0}^{\infty} |\phi_i| < \infty,$$

which is slightly stronger than the square summability condition.

Before we proceed, we show that such X_t is well defined in some probabilistic sense under the square summability condition. We shall show the right hand side of (4.3) converges in mean square to some random variable under the condition. Note that in the Hilbert space of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ of square integrable random variables, convergence is given by the Cauchy criteria. That is, we need to show that

$$\mathbb{E}\left(\sum_{i=0}^{m} \phi_i \varepsilon_{t-i} - \sum_{i=0}^{n} \phi_i \varepsilon_{t-i}\right)^2 \to 0$$

as $m, n \to \infty$. This is obviously the case given square summability of ϕ_i. Therefore, $\sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i}$ converges in mean square to some random variable, which we denoted by X_t.

It is easy to show that for $k \geq 0$,

$$\gamma_X(t, t-k) = \sigma^2 \sum_{i=0}^{\infty} \phi_{i+k}\phi_i,$$

which is well defined under square summability of ϕ_i since

$$\left|\sum_{i=0}^{\infty} \phi_{i+k}\phi_i\right| \leq \left(\sum_{i=k}^{\infty} \phi_i^2\right)^{1/2} \left(\sum_{i=0}^{\infty} \phi_i^2\right)^{1/2},$$

and is independent of the time t. Therefore, under the square summability condition, $\{X_t\}_{t\in\mathbb{Z}}$ is well defined in the mean square sense, and is weakly stationary.

Now

$$\sum_{k=0}^{\infty} |\gamma(k)| \leq \sum_{k=0}^{\infty}\sum_{i=0}^{\infty} |\phi_{i+k}\phi_i| \leq \sum_{i=0}^{\infty} |\phi_i| \sum_{k=0}^{\infty} |\phi_{i+k}| \leq \left(\sum_{i=0}^{\infty} |\phi_i|\right)^2.$$

Therefore, the autocovariances are absolutely summable under the absolute summability of ϕ_i. By our results in Section 2.3, the absolute summability of ϕ_i is sufficient for that

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to \mathbb{E}X_t = 0.$$

4.5 The Lag Operator

A time series operator transforms one or multiple time series into a new time series. Suppose we have two time series $\{x_t\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$. The scalar multiplication operator $x \mapsto \alpha x$ for some

$\alpha \in \mathbb{R}$ transforms a time series into a new one:

$$\alpha\{x_t\}_{t\in\mathbb{Z}} = \{\alpha x_t\}_{t\in\mathbb{Z}}.$$

The addition operator + transforms the two time series into a new time series:

$$\{x_t\}_{t\in\mathbb{Z}} + \{y_t\}_{t\in\mathbb{Z}} = \{x_t + y_t\}_{t\in\mathbb{Z}}.$$

There is a very frequently used operator in time series econometrics which transforms a time series

$$\cdots, x_{-1}, x_0, x_1, x_2, x_3, \cdots$$

to

$$\cdots, x_{-2}, x_{-1}, x_0, x_1, x_2, \cdots.$$

Such an operator is called a *lag operator*, and is usually denoted as L. Written formally, we have that

$$L\{x_t\}_{t\in\mathbb{Z}} = \{x_{t-1}\}_{t\in\mathbb{Z}}.$$

We usually use Lx_t to denote the time-t element of the transformed series $L\{x_t\}_{t\in\mathbb{Z}}$. Then Lx_t should be understood as $(Lx)_t$ where $x = \{x_t\}_{t\in\mathbb{Z}}$. With this definition, we have $Lx_t = x_{t-1}$.

Let $x = \{x_t\}_{t\in\mathbb{Z}}$ and $y = \{y_t\}_{t\in\mathbb{Z}}$ be two time series whose elements are random variables. Let $\alpha \in \mathbb{R}$. Then

$$L(x+y) = L\{x_t + y_t\} = \{x_{t-1} + y_{t-1}\} = \{x_{t-1}\} + \{y_{t-1}\} = Lx + Ly,$$

and

$$L(\alpha x) = L\{\alpha x_t\} = \{\alpha x_{t-1}\} = \alpha\{x_{t-1}\} = \alpha Lx.$$

Therefore, we have

$$L(\alpha x + y) = \alpha Lx + Ly,$$

or in element-wise format,

$$L(\alpha x_t + y_t) = \alpha x_{t-1} + y_{t-1}.$$

This shows that the lag operator is a linear operator.

For any $d \geq 0$, we write $L^d x = L(L^{d-1}x)$. Using this notation, we have, e.g., $L^2 x_t = L(Lx_t) = x_{t-2}$. We define L^{-1} to be the inverse of L, that is, if $L^{-1}x = y$, then $Ly = x$. It is easy to see that $L^{-1}x_t = x_{t+1}$, and we define $L^d x = L^{-1}(L^{d+1}x)$ for $d < 0$.

It is easy to show that for any $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ and $p, q \in \mathbb{Z}$,

$$(\alpha + \beta L^p)(\gamma + \delta L^q) = \alpha\gamma + \beta\gamma L^p + \alpha\delta L^q + \beta\delta L^{(p+q)}.$$

This calculation could be generalized and it is easy to see that the basic rules of algebra apply to

polynomials of the lag operator.

Using the lag operator, we may write a linear process as

$$X_t = \sum_{i=0}^{\infty} \phi_i L^i \varepsilon_t,$$

or

$$X_t = \phi(L)\varepsilon_t$$

where

$$\phi(L) = \sum_{i=0}^{\infty} \phi_i L^i.$$

Note that up to this point, $\phi(L)$ as an infinite sum is a "formal expression". It makes sense only when it is applied to a time series. The next section shows that $\phi(L)$ itself could be viewed as a power series of the lag operator under certain circumstances.

4.6 Linear Filters

Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is a time series, and $\{Y_t\}$ is generated by

$$Y_t = \sum_{i=-\infty}^{\infty} \phi_i X_{t-i},$$

then we say that $\{Y_t\}$ is obtained by applying the *linear filter* $\phi(L) = \sum_{i=-\infty}^{\infty} \phi_i L^i$ to $\{X_t\}$.

Theorem 4.21. *Let $\{X_t\}_{t\in\mathbb{Z}}$ be a sequence of random variables and $\phi(L) = \sum_{i=-\infty}^{\infty} \phi_i L^i$. If $\sum_{i=-\infty}^{\infty} |\phi_i| < \infty$ and $\sup_t \mathbb{E}X_t^2 < \infty$, then $\phi(L)X_t$ converges in mean square. If in addition $\{X_t\}$ is weakly stationary, so is $\{\phi(L)X_t\}$.*

Proof. The mean square convergence follows from the Lebesgue's dominated convergence theorem and

$$\begin{aligned}
\mathbb{E}\left(\phi(L)X_t\right)^2 &= \lim_{T\to\infty} \mathbb{E}\left(\sum_{i=-T}^{T}\sum_{j=-T}^{T} \phi_i\phi_j X_i X_j\right) \\
&\le \lim_{T\to\infty} \left(\sum_{i=-T}^{T} |\phi_i|\right)^2 \sup_t \mathbb{E}X_t^2 \\
&< \infty.
\end{aligned}$$

It is easy to check that when $\{X_t\}$ is weakly stationary, the covariance function of $Y_t = \phi(L)X_t$ is given by

$$\gamma_Y(k) = \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} \phi_i\phi_j\gamma_X(i-j+k),$$

where γ_X is the autocovariance function of $\{X_t\}$. Note that the right hand side of the above equation converges absolutely since $|\gamma_X(k)| \leq \gamma_X(0)$ for all k, and ϕ_i is absolutely summable by assumption. The covariance function is independent of t, which implies that $\{Y_t\}$ is weakly stationary. ■

The above results shows that for two linear filters $\phi(L) = \sum_{i=-\infty}^{\infty} \phi_i L^i$ and $\psi(L) = \sum_{i=-\infty}^{\infty} \psi_i L^i$ with $\sum_{i=-\infty}^{\infty} |\phi_i| < \infty$ and $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$, if $\{X_t\}$ is weakly stationary, then $\{\phi(L)\psi(L)X_t\}$ and $\{\psi(L)\phi(L)X_t\}$ are well defined and are weakly stationary. Also, it can be shown that

$$\phi(L)\psi(L)X_t = \psi(L)\phi(L)X_t = \eta(L)X_t,$$

where

$$\eta = \sum_{i=-\infty}^{\infty} \theta_i L^i,$$

$$\eta_i = \sum_{k=-\infty}^{\infty} \phi_k \psi_{i-k} = \sum_{k=-\infty}^{\infty} \psi_k \phi_{i-k}.$$

Linear operators with absolutely summable coefficients inherit algebraic properties of power series.

Now consider $\phi(L) = \sum_{i=0}^{\infty} \phi_i L^i$ such that $\phi(z) \neq 0$ for all $|z| \leq 1$ on the complex plane. Then there exists $\epsilon > 0$ such that $1/\phi(z)$ has a power series expansion

$$\frac{1}{\phi(z)} = \sum_{i=0}^{\infty} \varphi_i z^i = \varphi(z)$$

for $|z| < 1+\epsilon$. This implies that $\varphi_i(1+\epsilon/2)^i \to 0$ as $i \to \infty$, which in turn implies that there exists C such that $|\varphi_i| < C(1+\epsilon/2)^{-i}$ for all i. As a consequence, $\sum_{i=0}^{\infty} i\,|\varphi_i| < \infty$.

By construction, $\varphi(z)\phi(z) = 1$ for $|z| \leq 1$. Since linear operators with absolutely summable coefficients inherit algebraic properties of power series, we have

$$\varphi(L)\phi(L)X_t = X_t.$$

Therefore, we may view $\varphi(L)$ as the *inverse* of $\phi(L)$. That is, $\phi^{-1}(L) = \varphi(L)$. The inverse exists if $\phi(z) \neq 0$ for all $|z| \leq 1$.

4.7 The Beveridge-Nelson Decomposition

The Beveridge-Nelson decomposition is an important tool in studying linear processes. To develop the decomposition, write

$$\begin{aligned}\phi(L) &= \sum_{i=0}^{\infty} \phi_i L^i \\ &= \sum_{i=0}^{\infty} \phi_i - \sum_{i=1}^{\infty} (\phi_i - \phi_i L^i)\end{aligned}$$

$$= \sum_{i=0}^{\infty} \phi_i - \sum_{i=1}^{\infty} \phi_i (1-L)(1+L+L^2+\cdots+L^{i-1})$$
$$= \phi(1) - (1-L) \sum_{i=0}^{\infty} \left(\sum_{j=i+1}^{\infty} \phi_j \right) L^i.$$

For the latter part of the last line to give a well defined linear process, we need

$$\sum_{i=0}^{\infty} \left(\sum_{j=i+1}^{\infty} \phi_j \right)^2 < \infty.$$

This is the case if $\sum_{i=0}^{\infty} j^2 \phi_j^2 < \infty$. See Phillips and Solo (1992) for details. Furthermore, $\sum_{i=0}^{\infty} \left| \sum_{j=i+1}^{\infty} \phi_j \right| < \infty$ if $\sum_{i=0}^{\infty} j \left| \phi_j \right| < \infty$.

Given $\sum_{i=0}^{\infty} j^2 \phi_j^2 < \infty$, we may write

$$X_t = \phi(1)\varepsilon_t - (\tilde{X}_t - \tilde{X}_{t-1})$$

where $\tilde{X}_t = \sum_{i=0}^{\infty} \tilde{\phi}_i \varepsilon_{t-i}$, $\tilde{\phi}_i = \sum_{j=i+1}^{\infty} \phi_j$. Note that $\{\tilde{X}_t\}$ is a stationary linear process. The above decomposition, first introduced by Beveridge and Nelson (1981), is called the *Beveridge-Nelson decomposition* or the *permanent-transitory decomposition* of the linear process $\{X_t\}$.

4.8 Asymptotics for Linear Processes

The Beveridge-Nelson decomposition could be used to obtain asymptotics for linear processes. This approach was developed by Phillips and Solo (1992). The law of large numbers for linear processes has been given in Section 4.4. Given $\sum_{i=0}^{\infty} j^2 \phi_i^2 < \infty$, we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(\phi(1)\varepsilon_t - (\tilde{X}_t - \tilde{X}_{t-1}) \right)$$
$$= \phi(1) \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t - \frac{1}{\sqrt{T}} (\tilde{X}_T - \tilde{X}_0).$$

Note that

$$\frac{1}{\sqrt{T}} (\tilde{X}_T - \tilde{X}_0) = O_p(1/\sqrt{T}),$$

we have a central limit theorem for $\{X_t\}$ as long as we have a central limit theorem hold for $\{\varepsilon_t\}$. Readers may wish to consult Section 2.2.2 for a variety of central limit theorems frequently used in econometrics. For example, under the iid assumption or mds assumption with appropriate regularity conditions, we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \to_d \mathbb{N}(0, \phi^2(1)\sigma^2) =_d \mathbb{N}\left(0, \left(\sum_{i=0}^{\infty} \phi_i \right)^2 \sigma^2 \right).$$

5 Stationary ARMA Processes

5.1 Moving Average Processes

The Wold decomposition theorem implies that any pure non-deterministic weakly stationary process can be represented as an infinite order moving average process. However, since it is impossible to fit an arbitrary infinite order moving average process with finite number of data points, we further approximate infinite order moving average processes by finite order moving average processes.

A q-th order *moving average process* $\{X_t\}$, denoted by MA(q), is given by

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$. Taking expectations on both sides, we have $\mathbb{E}X_t = \mu$. Then the demeaned process $\{X_t - \mu\}$ is just a special case of the linear process introduced in the previous chapter. For any choice of the parameters $\theta_1, \ldots, \theta_q$, they are absolutely summable. So an MA(q) process inherits directly all the properties of linear processes. As a consequence, $\{X_t\}$ is weakly stationary, and its autocovariance function is given by (taking $\theta_0 = 1$)

$$\gamma(k) = \begin{cases} \sigma^2 \sum_{i=0}^{q-|k|} \theta_i \theta_{i+|k|}, & |k| \le q, \\ 0, & |k| > q. \end{cases}$$

In particular, we have

$$\mathrm{Var}(X_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2.$$

Since the condition $\sum_{i=0}^{q} i\,|\theta_i| < \infty$ (which implies $\sum_{i=0}^{q} |\theta_i| < \infty$) is always satisfied, we have that

$$\frac{1}{T} \sum_{t=1}^{T} X_t \to_p \mu,$$

and

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (X_t - \mu) \to_d \mathbb{N}\left(0, (1 + \theta_1 + \cdots + \theta_q)^2 \sigma^2\right)$$

under appropriate assumptions on $\{\varepsilon_t\}$. (For example, if we assume that $\{\varepsilon_t\}$ is an iid sequence.)

5.2 Autoregressive Processes

Another class of linear process is the *autoregressive process*, which connects the present value of the time series variable with its past values in a linear way. A p-th order autoregressive process $\{X_t\}$, denoted by AR(p), is given by

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \tag{5.1}$$

where $\varepsilon_t \sim \text{WN}(0, \sigma^2)$.

If a weakly stationary economic time series really behaves as in equation (5.1), i.e., the economic variable X_t interacts with its past values in the linear way specified in (5.1), then it means that the equation (5.1), viewed as a stochastic difference equation in X_t, should have a sensible solution. To see when this is the case, we first consider the simplest AR model given by

$$X_t = \alpha X_{t-1} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise process.

Suppose $|\alpha| < 1$. We first show that the AR(1) model has a weakly stationary solution. Consider the proposed solution

$$X_t^* = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}.$$

Since $|\alpha| < 1$, by our discussion in the previous chapter, this proposed solution is well defined in the mean square sense, and is weakly stationary. It is also easy to check that it actually solves the difference equation in the sense that

$$X_t^* - \alpha X_{t-1}^* = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i} - \alpha \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-1-i} = \varepsilon_t.$$

That is, we have found a weakly stationary solution. To show that this is the only weakly stationary solution, we iterate backward the difference equation. Any solution $\{X_t^\circ\}$ to the difference equation should satisfy

$$X_t^\circ = \varepsilon_t + \alpha\varepsilon_{t-1} + \cdots + \alpha^k \varepsilon_{t-k} + \alpha^{k+1} X_{t-k-1}^\circ.$$

Then by triangular inequality,

$$\begin{aligned}
\|X_t^\circ - X_t^*\|_{L^2} &= \left\| X_t^\circ - \sum_{i=0}^{k} \alpha^i \varepsilon_{t-i} + \sum_{i=0}^{k} \alpha^i \varepsilon_{t-i} - X_t^* \right\|_{L^2} \\
&\le \left\| X_t^\circ - \sum_{i=0}^{k} \alpha^i \varepsilon_{t-i} \right\|_{L^2} + \left\| \sum_{i=0}^{k} \alpha^i \varepsilon_{t-i} - X_t^* \right\|_{L^2} \\
&= \alpha^{k+1} \left\| X_{t-k-1}^\circ \right\|_{L^2} + \left\| \sum_{i=0}^{k} \alpha^i \varepsilon_{t-i} - X_t^* \right\|_{L^2}.
\end{aligned}$$

for any $k > 0$. By definition of X_t^*, $\left\| \sum_{i=0}^{k} \alpha^i \varepsilon_{t-i} - X_t^* \right\|_{L^2} \to 0$ as $k \to \infty$. If $\{X_t^\circ\}$ is weakly stationary, then $\left\| X_{t-k-1}^\circ \right\|_{L^2} = \mathbb{E} X_{t-k-1}^{\circ 2} < \infty$. This then implies that the right hand side of the above equation converges to 0. This in turn implies that $\|X_t^\circ - X_t^*\|_{L^2} = 0$, that is, $X_t^\circ = X_t^*$ in mean square sense. This shows that $X_t^* = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}$ is the only weakly stationary solution (in the mean square sense) to the AR(1) difference equation. Note that the above statement does not rule out the possibility that there are non-weakly stationary solutions to the difference equation.

For example, it is easy to verify that $\tilde{X}_t = \alpha^t + \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}$ is a solution to the difference equation. However, $\{\tilde{X}_t\}$ is not weakly stationary.

Now consider the case that $|\alpha| > 1$. Apparently,

$$X_t = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}$$

is not a well defined solution (in mean square sense) in this case, and it cannot be weakly stationary. However, by similar arguments, we can show that

$$X_t = -\sum_{i=1}^{\infty} \frac{1}{\alpha^i} \varepsilon_{t+i}$$

is the only weakly stationary solution to the AR(1) difference equation.

It can be shown that if $|\alpha| = 1$, the AR(1) difference equation does not have any weakly stationary solution.

If we give t the interpretation as an index for time, then in the case $|\alpha| > 1$, the weakly stationary is determined by the future innovations $\{\varepsilon_s\}_{s>t}$. This is unnatural in terms of causality. Therefore, we focus on models in which the solutions are *causal*, meaning that the solutions can be expressed as functions of past and current variables but not future variables.

According to the above discussion, if we want an AR(1) model to have a causal weakly stationary solution (the solution turns out to be unique), we need to restrict α to be smaller than one in absolute value.

Now we come back to the general AR model (5.1) and ask when the model has a causal weakly stationary solution. The following theorem, which is a corollary of Theorem 5.2, answers this question.

Theorem 5.1. *Let $\{X_t\}$ be an AR(p) process given by $\alpha(L)X_t = c + \varepsilon_t$ where $\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots \alpha_p L^p$ and $\varepsilon_t \sim WN(0, \sigma^2)$. If $\alpha(z) \neq 0$ for all $|z| \leq 1$, then $\{X_t\}$ has a unique causal weakly stationary solution given by*

$$X_t = \alpha(L)^{-1}(c + \varepsilon_t) = \frac{c}{\alpha(1)} + \sum_{i=0}^{\infty} \varphi_i \varepsilon_{t-i}$$

where $\sum_{i=0}^{\infty} \varphi_i z^i$ is the power series expansion of $\alpha^{-1}(z)$, and $\sum_{i=0}^{\infty} i\,|\varphi_i| < \infty$.

Lütkepohl (2005) calls a process *stable* if it satisfies the condition that $\alpha(z) \neq 0$ for all $|z| \leq 1$. It is worth noting that the AR(p) model, even with the stability condition, does allow for other solutions that are not causal and weakly stationary. However, it only allows for a unique solution that is causal and weakly stationary.

From the above theorem, we have that for a weakly stationary AR(p) process $\{X_t\}$, its expec-

tation is

$$\mu = \frac{c}{\alpha(1)}.$$

The AR process has a demeaned representation:

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \alpha_2(X_{t-2} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + \varepsilon_t.$$

Since the solution is causal, X_t is uncorrelated with ε_s for any $s > t$. We therefore can multiple both sides of the above equation by $X_{t-k} - \mu$ for some $k > 1$ and take expectations. We then obtain the *Yule-Walker equations* of the AR(p) process:

$$\gamma(k) = \alpha_1\gamma(k-1) + \alpha_2\gamma(k-2) + \cdots + \alpha_p\gamma(k-p), \quad k = 1, 2, \cdots,$$

where $\gamma(\cdot)$ is the autocovariance function of $\{X_t\}$. Dividing both sides by $\gamma(0)$, we get the Yule-Walker equations for the autocorrelations:

$$\rho(k) = \alpha_1\rho(k-1) + \alpha_2\rho(k-2) + \cdots + \alpha_p\rho(k-p), \quad k = 1, 2, \cdots.$$

The Yule-Walker equations could be used to calculate the autocorrelations of AR(p) processes. For example, for an AR(2) process given by $X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t$, we have

$$\mathbb{E}X_t = \frac{c}{1 - \alpha_1 - \alpha_2}.$$

Also, $\rho(1) = \alpha_1\rho(0) + \alpha_2\rho(-1)$. Since $\rho(0) = 1$, and $\rho(-1) = \rho(1)$, we have $\rho(1) = \frac{\alpha_1}{1-\alpha_2}$. Given $\rho(0)$ and $\rho(1)$, we may calculate $\rho(k)$ for any k using the Yule-Walker equations.

Note that from the theorem above, the demeaned AR(p) process is a pure non-deterministic linear process with coefficients satisfying the summability condition for the Beveridge-Nelson decomposition. Therefore, $\{X_t\}$ has absolutely summable autocovariances, and as a result,

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to_p \mu,$$

and under appropriate conditions of ε_t (for example, $\varepsilon_t \sim$ iid),

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(X_t - \mu) \to_d \mathbb{N}\left(0, \frac{\sigma^2}{\alpha^2(1)}\right).$$

Note that, by the arguments in Section 2.3, $\sum_{k=-\infty}^{\infty}\gamma(k) = \frac{\sigma^2}{\alpha^2(1)}$.

5.3 ARMA Processes

The *ARMA model* contains both an autoregressive part and a moving average part. If $\{X_t\}$ follows an ARMA(p, q) model, then

$$X_t = c + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\varepsilon_t \sim \text{WN}(0, \sigma^2)$. We could also write it as

$$\alpha(L) X_t = c + \theta(L) \varepsilon_t \tag{5.2}$$

where $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$.

Theorem 5.2. *Let $\{X_t\}$ be an ARMA(p, q) process given by (5.2), and assume that the two polynomials $\alpha(z)$ and $\theta(z)$ have no common zeros. If $\alpha(z) \neq 0$ for all $|z| \leq 1$, then $\{X_t\}$ has a unique causal weakly stationary solution given by*

$$X_t = \alpha^{-1}(L)(c + \theta(L)\varepsilon_t) = \frac{c}{\alpha(1)} + \sum_{i=0}^{\infty} \varphi_i \varepsilon_{t-i}$$

where $\sum_{i=0}^{\infty} \varphi_i z^i$ is the power series expansion of $\frac{\theta(z)}{\alpha(z)}$, and $\sum_{i=0}^{\infty} i\,|\varphi_i| < \infty$.

Proof. Suppose that $\alpha(z) \neq 0$ for all $|z| \leq 1$. According to Section 4.6, we may apply $\alpha^{-1}(L)$ to both sides of (5.1) and all the results follows. Note that uniqueness comes from the invertibility of $\alpha(L)$ with respect to the class of weakly stationary series. ■

When $\alpha(z)$ and $\theta(z)$ have common zeros, if the zeros are outside the unit disk, and the reduced ARMA process (by canceling the common factors in $\alpha(z)$ and $\theta(z)$) satisfies the conditions in Theorem 5.2, then all the arguments in the proof of Theorem 5.2 go through, and the results still holds, with $\alpha(L)$ and $\theta(L)$ replaced with their reduced version. In fact, if none of the common zeros are on the unit circle, then the ARMA process has a unique causal weakly stationary solution. However, if there is some common zeros that are on the unit disk, the ARMA process may have multiple causal weakly stationary solutions, even if the reduced ARMA satisfies the conditions in Theorem 5.2. For identification, we shall work with models in which $\alpha(z)$ and $\theta(z)$ have no common zeros. See Theorem 5.3 and the remarks following it.

The mean of the ARMA process (5.2) is given by

$$\mu = \frac{c}{\alpha(1)},$$

and the model could be written in the demeaned version:

$$X_t - \mu = \alpha_1(X_{t-1} - \mu) + \cdots + \alpha_p(X_{t-p} - \mu) + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

The Yule-Walker equations are given by

$$\gamma(k) = \alpha_1\gamma(k-1) + \alpha_2\gamma(k-2) + \cdots + \alpha_p\gamma(k-p), \quad k = q+1, q+2, \cdots .$$

We may solve the Yule-Walker equations (which are some homogeneous p-th order linear difference equations) and obtain the general solution

$$\gamma(k) = \sum_{i=1}^{n}\sum_{j=0}^{m_i-1} r_{ij}k^j z_i^{-k}, \qquad k \geq \max(p, q+1) - p,$$

where $z_i, i = 1, 2, \ldots, n$ are the distinct zeros of $\alpha(z)$, m_i is the multiplicity of z_i so that $\sum_{i=1}^{k} m_i = p$, and r_{ij} are some constants that could be determined by the boundary conditions. See Brockwell and Davis (1991, Section 3.3 and 3.6) for details. It can be seen that the covariance function of a causal ARMA process is a sum of some geometrically decaying terms, which implies that causal ARMA processes (and hence AR and MA processes) have geometrically decaying autocovariance functions as the lag k goes to infinity.

Similarly as the AR model, for a causal weakly stationary ARMA(p, q) process satisfying conditions in Theorem 5.2,

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to_p \mu,$$

and under appropriate assumptions on $\{\varepsilon_t\}$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(X_t - \mu) \to_d \mathbb{N}\left(0, \frac{\theta^2(1)}{\alpha^2(1)}\sigma^2\right).$$

Note that $\sum_{k=-\infty}^{\infty}\gamma(k) = \frac{\theta^2(1)}{\alpha^2(1)}\sigma^2$.

The ARMA process (5.2) is said to be invertible if $\{\varepsilon_t\}$ can be represented by

$$\varepsilon_t = \sum_{i=0}^{\infty}\vartheta_i(X_{t-i} - \mu)$$

with $\sum_{i=0}^{\infty}|\vartheta_i| < \infty$. Naturally, if $\alpha(z)$ and $\theta(z)$ have no common zeros, and $\theta(z) \neq 0$ for all $|z| \leq 1$, then we may apply $\theta^{-1}(L)$ to both sides of (5.2) and obtain

$$\varepsilon_t = \theta^{-1}(L)\alpha(L)(X_t - \mu).$$

Similarly, $\theta^{-1}(z)\alpha(z)$ has a power series representation with absolutely summable coefficients. That is, the ARMA is invertible.

5.4 The Autocovariance Generating Function

Let $\{X_t\}$ be a weakly stationary process with covariance function $\gamma(\cdot)$. Its *autocovariance generating function* is defined as

$$G(z) = \sum_{k=-\infty}^{\infty} \gamma(k) z^k$$

provided that the power series converges for $\frac{1}{r} < |z| < r$ with some $r > 1$. If we know the autocovariances, we can calculate the autocovariance generating function. On the other hand, if we know the autocovariance generating function, we can back out the autocovariances by looking at the coefficients of the power series.

It is obvious that a process is white noise if and only if its autocovariance generating function is constant. In this case, the constant is equal to the variance of the process. For $X_t = \sum_{i=-\infty}^{\infty} \phi_i \varepsilon_{t-i}$ where $\varepsilon \sim \text{WN}(0, \sigma^2)$ and $\sum_{i=-\infty}^{\infty} |\phi_i| < \infty$, we have that $\gamma(k) = \sigma^2 \sum_{i=-\infty}^{\infty} \phi_i \phi_{i+|k|}$. Then

$$\begin{aligned}
G(z) &= \sigma^2 \sum_{k=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \phi_i \phi_{i+|k|} z^k \\
&= \sigma^2 \left(\sum_{i=-\infty}^{\infty} \phi_i^2 + \sum_{k=1}^{\infty} \sum_{i=-\infty}^{\infty} \phi_i \phi_{i+k} (z^k + z^{-k}) \right) \\
&= \sigma^2 \left(\sum_{i=-\infty}^{\infty} \phi_i z^i \right) \left(\sum_{j=-\infty}^{\infty} \phi_j z^{-j} \right) \\
&= \sigma^2 \phi(z) \phi(z^{-1}).
\end{aligned}$$

For a causal ARMA(p, q) process $\{X_t\}$ given by $\alpha(L) X_t = c + \theta(L) \varepsilon_t$, it can be written as $X_t = \mu + \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i}$ where $\sum_{i=0}^{\infty} \phi_i z_{t-i} = \frac{\theta(z)}{\alpha(z)}$. Then its autocovariance generating function is given by

$$G(z) = \frac{\theta(z)\theta(z^{-1})}{\alpha(z)\alpha(z^{-1})} \sigma^2.$$

5.5 Non-Causal and Non-Invertible Stationary ARMA Processes

In this section we give a brief discussion of non-causal and non-invertible stationary ARMA processes. We use a well-know result in complex analysis that for a finite order polynomial $\phi(z)$ such that $\phi(z) \neq 0$ for all $|z| = 1$, its reciprocal admits a Laurent series expansion given by

$$\phi(z)^{-1} = \sum_{i=-\infty}^{\infty} \psi_i z^i = \psi(z)$$

where the Laurent series converges absolutely on $r^{-1} < |z| < r$ for some $r > 1$. In particular, $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$. The difference between the case discussed here and the case in which the polynomial has roots only outside the unit circle is that the expansion of the reciprocal is now a

two-sided infinite series.

Theorem 5.3. *Let $\{X_t\}$ be an ARMA(p, q) process given by (5.2), and assume that $\alpha(z)\theta(z) \neq 0$ for all $|z| = 1$. Then $\{X_t\}$ has a unique weakly stationary solution given by*

$$X_t = \alpha(L)^{-1}(c + \theta(L)\varepsilon_t) = \frac{c}{\alpha(1)} + \sum_{i=-\infty}^{\infty} \varphi_i \varepsilon_{t-i}$$

where $\sum_{i=-\infty}^{\infty} \varphi_i z^i$ is the Laurent series expansion of $\frac{\theta(z)}{\alpha(z)}$, and $\sum_{i=-\infty}^{\infty} |\varphi_i| < \infty$.

Proof. The solution given in the statement of the theorem is obviously weakly stationary. Also, by applying $\alpha(L)$ on both sides of solution, we may easily verify that it is indeed a solution to the ARMA equation. To show uniqueness, suppose that $\{Y_t\}$ is also a stationary solution to the ARMA equation. Then we have $\alpha(L)(Y_t - X_t) = 0$. By the stationarity of $\{Y_t - X_t\}$ and therefore the stationarity of $\{\alpha(L)(Y_t - X_t)\}$, we may apply the operator $\alpha(L)^{-1}$ on both sides of $\alpha(L)(Y_t - X_t) = 0$, which gives $Y_t - X_t = 0$. ■

We allow for common zeros in the above theorem, as long as the polynomials do not have roots on the unit circle. The uniqueness of the stationary solution is not affected by common zeors. This implies that in the case of common zeros, as long as there is no roots on the unit circle, the ARMA representation that contains common zero gives the same weakly stationary solution as the ARMA representation obtained by cancelling the common factors. However, if there is at least one common zeros that lie on the unit circle, then the ARMA equation may have more than one weakly stationary solution. To give an example, let $\{X_t\}$ be the weakly stationary solution to $\alpha(L)X_t = \theta(L)\varepsilon_t$ where $\alpha(z)$ and $\theta(z)$ have no roots on the unit circle. Let $|z_0| = 0$. For any mean-zero random variable A uncorrelated with $\{X_t\}$, it is easy to verify that $\{X_t + Az_0^t\}$ is a weakly stationary solution to the ARMA equation $(1 - z_0 L)\alpha(L)X_t = (1 - z_0 L)\theta(L)\varepsilon_t$.

We next give a theorem that transforms an arbitrary stationary ARMA process into a causal and invertible one.

Theorem 5.4. *Let $\{X_t\}$ be an ARMA(p, q) process given by (5.2), and assume that $\alpha(z)\theta(z) \neq 0$ for all $|z| = 1$. Then there exists polynomials $\tilde{\alpha}(z)$ and $\tilde{\theta}(z)$ of order p and q, respectively, a constant $\tilde{c}$, and a white noise process $\{\varepsilon_t^*\}$ such that $\tilde{\alpha}(z)\tilde{\theta}(z) \neq 0$ for all $|z| \leq 1$ and that*

$$\tilde{\alpha}(L)X_t = \tilde{c} + \tilde{\theta}(L)\varepsilon_t^*.$$

Proof. Without loss of generalizty we may assume that X_t is mean zero (i.e., $c = 0$). The generalization to the non-mean-zero case is obvious.

Assume $\alpha(L)X_t = \theta(L)\varepsilon_t$. Let the zeros of $\alpha(z)$ be $a_1, a_2, \ldots, a_p$ such that $a_{r+1}, a_{r+2}, \ldots, a_p$ are the zeros that lie in the unit circle. Let the zeros of $\theta(z)$ be $b_1, b_2, \ldots, b_q$ such that $b_{s+1}, b_{s+2}, \ldots, b_q$

are the zeros that lie in the unit circle. Define

$$\tilde{\alpha}(z) = \alpha(z) \prod_{r<i\leq p} \frac{1 - a_i z}{1 - a_i^{-1} z} \quad \text{and} \quad \tilde{\theta}(z) = \theta(z) \prod_{s<i\leq p} \frac{1 - b_i z}{1 - b_i^{-1} z}.$$

Note that all roots of $\tilde{\alpha}(z)$ and $\tilde{\theta}(z)$ lie outside the unit circle. Now define

$$\varepsilon_t^* = \frac{\tilde{\alpha}(L)}{\tilde{\theta}(L)} X_t = \left(\prod_{r<i\leq p} \frac{1 - a_i L}{1 - a_i^{-1} L} \right) \left(\prod_{s<i\leq p} \frac{1 - b_i^{-1} L}{1 - b_i L} \right) \varepsilon_t.$$

The autocovariance generating function of ε_t^* is given by

$$\begin{aligned} G_{\varepsilon_t^*}(z) &= \left(\prod_{r<i\leq p} \frac{1 - a_i z}{1 - a_i^{-1} z} \right) \left(\prod_{r<i\leq p} \frac{1 - a_i z^{-1}}{1 - a_i^{-1} z^{-1}} \right) \left(\prod_{s<i\leq p} \frac{1 - b_i^{-1} z}{1 - b_i z} \right) \left(\prod_{s<i\leq p} \frac{1 - b_i^{-1} z^{-1}}{1 - b_i z^{-1}} \right) \sigma^2 \\ &= \left(\prod_{r<i\leq p} a_i^2 \right) \left(\prod_{s<i\leq p} b_i^2 \right) \sigma^2, \end{aligned}$$

which is constant. Therefore, $\{\varepsilon_t^*\}$ is a white noise. And by construction, $\tilde{\alpha}(L) X_t = \tilde{\theta}(L) \varepsilon_t^*$. ■

We shall point out here that in general ε_t and ε_t^* have different variances. Also, $\{\varepsilon_t^*\}$, which is constructed by applying a linear filter on $\{\varepsilon_t\}$, could be dependent even if $\{\varepsilon_t\}$ is an iid sequence.

The above theorem shows that that for any weakly stationary ARMA process without roots on the unit circle, we can always find a white noise process $\{\varepsilon_t^*\}$ such that the ARMA process has a causal and invertible ARMA(p, q) representation with respect to this new white noise process. Also, we may conduct similar procedures to a causal and invertible ARMA(p, q) process to obtain its non-causal and/or non-invertible representations. Therefore, an ARMA process could have multiple representations, each represents the process equally well (in the sense of characterizing the first two moments of the series.) However, for practical reasons, most of the time we shall only focus on ARMA processes that are both causal and invertible. The white noise process that corresponds to the causal and invertible representation of a series $\{X_t\}$ is called the *fundamental innovation* process of $\{X_t\}$.

5.6 Spectral Densities of ARMA Processes

Theorem 5.5. *Let $\{X_t\}$ be a mean-zero, complex-valued weakly stationary process with autocovariance function γ_X and spectral distribution function F_X. Let*

$$Y_t = \phi(L) X_t = \sum_{j=-\infty}^{\infty} \phi_j X_{t-j}$$

where $\phi(z) = \sum_{j=-\infty}^{\infty} \phi_j z^j$ *and* $\sum_{j=-\infty}^{\infty} |\phi_j| < \infty$*, then* $\{Y_t\}$ *is weakly stationary with autocovariance function*

$$\gamma_Y(k) = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \phi_r \bar{\phi}_s \gamma_X(k+s-r),$$

and the spectral distribution function

$$F_Y(\lambda) = \int_{-\pi}^{\lambda} \left|\phi(e^{-i\nu})\right|^2 \mathrm{d}F_X(\nu).$$

Proof. Obviously, $\mathbb{E}Y_t = 0$ for all t. We may follow the proof of Theorem 4.21 to show that

$$\mathbb{E}Y_t \bar{Y}_{t-k} = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \phi_r \bar{\phi}_s \gamma_X(k+s-r),$$

which is independent of t. This show that $\{Y_t\}$ is weakly stationary.

Also, we have

$$\begin{aligned}
\gamma_Y(k) &= \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \phi_r \bar{\phi}_s \gamma_X(k+s-r) \\
&= \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \phi_r \bar{\phi}_s \int_{-\pi}^{\pi} e^{i\lambda(k+s-r)} \mathrm{d}F_X(\lambda) \\
&= \int_{-\pi}^{\pi} e^{i\lambda k} \left(\sum_{r=-\infty}^{\infty} \phi_r e^{-i\lambda r} \right) \left(\sum_{s=-\infty}^{\infty} \bar{\phi}_s e^{i\lambda s} \right) \mathrm{d}F_X(\lambda) \\
&= \int_{-\pi}^{\pi} e^{i\lambda k} \left|\phi(e^{-i\lambda})\right|^2 \mathrm{d}F_X(\lambda).
\end{aligned}$$

It is then easy to verify that

$$F_Y(\lambda) = \int_{-\pi}^{\lambda} \left|\phi(e^{-i\nu})\right|^2 \mathrm{d}F_X(\nu).$$

■

In the setting of the above theorem, if f_X is the spectral density of $\{X_t\}$, then the spectral density of $\{Y_t\}$ is

$$f_Y(\lambda) = \left|\phi(e^{-i\lambda})\right|^2 f_X(\lambda) = \phi(e^{-i\lambda})\phi(e^{i\lambda}) f_X(\lambda).$$

We call the function $\lambda \mapsto \phi(e^{-i\lambda})$ the *transfer function* of the filter $\phi(L)$, and the function $\lambda \mapsto \left|\phi(e^{-i\lambda})\right|^2$ the *power transfer function* of the filter.

It is easy to see that a white noise with variance σ^2 has a constant spectral density $\frac{\sigma^2}{2\pi}$. From the above theorem we can easily show that that if $\{X_t\}$ is a causal ARMA process given by

$\alpha(L)X_t = c + \theta(L)\varepsilon_t$ where $\varepsilon_t \sim \text{WN}(0, \sigma^2)$, it has a spectral density

$$\frac{\sigma^2}{2\pi}\frac{\theta(e^{-i\lambda})\theta(e^{i\lambda})}{\alpha(e^{-i\lambda})\alpha(e^{i\lambda})} = \frac{\sigma^2}{2\pi}\frac{\left|\theta(e^{-i\lambda})\right|^2}{\left|\alpha(e^{-i\lambda})\right|^2}.$$

For weakly stationary processes $\{X_t\}$ with absolutely summable autocovariances, we have that $f(0) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)$. This implies that the long-run variance of $\{X_t\}$ is given by $2\pi f(0)$.

Let $\{X_t\}$ be a weakly stationary time series with autocovariance generating function $G_X(z)$ and spectral density $f_X(\lambda)$. Let $Y_t = \phi(L)X_t$ where $\phi(L) = \sum_{i=-\infty}^{\infty}\phi_i L^i$ with $\sum_{i=-\infty}^{\infty}|\phi_i| < \infty$. With a completely similar argument as in Section 5.4 we may show that

$$G_Y(z) = \phi(z)\phi(z^{-1})G_X(z)$$

where $G_Y(z)$ is the autocovariance generating function of $\{Y_t\}$.

Now we state a approximation result. Using the fact that trigonometric polynomials are uniformly dense in the space of continuous even function on $[-\pi, \pi]$, we have the following result.

Theorem 5.6. *If f is a symmetric continuous spectral density function on $[-\pi, \pi]$, then for any $\epsilon > 0$ there exists a p-th order polynomial $a(z) = 1 + a_1 z + \cdots + a_p z^p$ whose coefficients are real and whose zeros are strictly outside the unit circle such that*

$$\sup_{\lambda\in[-\pi,\pi]}\left|A\left|a(e^{-i\lambda})\right|^2 - f(\lambda)\right| < \epsilon$$

where $A = \frac{\int_{-\pi}^{\pi} f(\nu)\mathrm{d}\nu}{2\pi(1+a_1^2+\cdots+a_p^2)}$.

For the proof of the theorem, see Brockwell and Davis (1991, p. 130). With the above theorem we may conclude immediately the following approximation result.

Theorem 5.7. *If f is a symmetric continuous spectral density function on $[-\pi, \pi]$, then for any $\epsilon > 0$, there exists an invertible MA(q) process*

$$X_t = \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_q\varepsilon_q, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

with spectral density f_X such that

$$\sup_{\lambda\in[-\pi,\pi]}|f_X(\lambda) - f(\lambda)| < \epsilon$$

where $\sigma^2 = \frac{\int_{-\pi}^{\pi} f(\nu)\mathrm{d}\nu}{1+a_1^2+\cdots+a_p^2}$.

It is straightforward to show that if $\{X_t\}$ and $\{Y_t\}$ are two independent weakly stationary time series with autocovariance generating functions $G_X(z)$ and $G_Y(z)$ and spectral densities $f_X(\lambda)$ and $f_Y(\lambda)$, then the series $\{Z_t\} = \{X_t + Y_t\}$ has autocovariance generating function $G_Z(z) = G_X(z) + G_Y(z)$ and spectral density $f_Z(\lambda) = f_X(\lambda) + f_Y(\lambda)$.

We also have an approximation result using AR processes.

Theorem 5.8. *If f is a symmetric continuous spectral density function on $[-\pi, \pi]$, then for any $\epsilon > 0$, there exists a causal AR(p) process*

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

with spectral density f_X such that

$$\sup_{\lambda \in [-\pi, \pi]} |f_X(\lambda) - f(\lambda)| < \epsilon.$$

Proof. Let $f^\epsilon(\lambda) = \max\{f(\lambda), \frac{\epsilon}{2}\}$. Then we have $f^\epsilon(\lambda) \geq \frac{\epsilon}{2}$ and $|f(\lambda) - f^\epsilon(\lambda)| \leq \frac{\epsilon}{2}$. Fix $0 < \delta < \frac{\epsilon}{4}$. Then there exists a polynomial $a(z) = 1 + a_1 z + \cdots + a_p z^p$ whose coefficients are real and whose roots are outside the unit disk such that

$$\sup_{\lambda \in [-\pi, \pi]} \left| A \left| a(e^{-i\lambda}) \right|^2 - \frac{1}{f^\epsilon(\lambda)} \right| < \delta$$

where $A = \frac{\int_{-\pi}^{\pi} f(\nu) \mathrm{d}\nu}{2\pi(1 + a_1^2 + \cdots + a_p^2)}$. Then we have $A \left| a(e^{-i\lambda}) \right|^2 > \frac{\epsilon}{4}$. Now we have

$$\sup_{\lambda \in [-\pi, \pi]} \left| \frac{1}{A \left| a(e^{-i\lambda}) \right|^2} - f^\epsilon(\lambda) \right| = \sup_{\lambda \in [-\pi, \pi]} \frac{\left| A \left| a(e^{-i\lambda}) \right|^2 - \frac{1}{f^\epsilon(\lambda)} \right|}{A \left| a(e^{-i\lambda}) \right|^2 \frac{1}{f^\epsilon(\lambda)}} < 8\delta\epsilon^2.$$

Therefore,

$$\sup_{\lambda \in [-\pi, \pi]} \left| \frac{1}{A \left| a(e^{-i\lambda}) \right|^2} - f(\lambda) \right| < 8\delta\epsilon^2 + \frac{\delta}{2}.$$

We may choose δ small enough so that the right hand side of the above inequality is smaller than ϵ. Note that $\frac{1}{A \left| a(e^{-i\lambda}) \right|^2}$ is the spectral density of the process $a(L)X_t = \varepsilon_t, \varepsilon_t \sim \mathrm{WN}(0, \frac{2\pi}{A})$. We therefore have found the desired AR process. ∎

Now we show that an ARMA process with a unit root cannot be stationary.

Theorem 5.9. *If $\phi(z)$ and $\theta(z)$ are polynomials with no common zeros and if $\phi(z) = 0$ for some $|z| = 1$, then the ARMA equation $\phi(L)X_t = \theta(L)\varepsilon_t, \varepsilon_t \sim WN(0, \sigma^2)$ has no weakly stationary solution.*

Proof. Let the unit root of $\phi(z)$ be z_0. We may pick $\lambda = \lambda_0$ so that $e^{-i\lambda_0} = z_0$. Since $\phi(z)$ and $\theta(z)$ have no common zeros, $\left| \theta(e^{-i\lambda_0}) \right| \neq 0$. We denote this non-zero value by θ_0. Since $\cdot \mapsto \left| \theta(e^{-i\cdot}) \right|$ is continuous, we may choose $\epsilon_0 > 0$ such that $\left| \theta(e^{-i\lambda}) \right|^2 > \frac{1}{2}\theta_0$ for all $\lambda \in [\lambda_0, \lambda_0 + \epsilon_0]$. Since $\phi(e^{-i\lambda_0}) = \phi(e^{i\lambda_0}) = 0$ and both $\left| \phi(e^{-i\lambda}) \right|$ and $\left| \phi(e^{i\lambda}) \right|$ are continuously differentiable, there exists C such that $\left| \phi(e^{-i(\lambda_0 + \epsilon)}) \right| \leq C\epsilon$ and $\left| \phi(e^{i(\lambda_0 + \epsilon)}) \right| \leq C\epsilon$ for all $\epsilon \in [0, \epsilon_0]$.

Suppose the ARMA equations have a weakly stationary solution $\{X_t\}$ with spectral distribution function F. Then we have

$$\int_{\lambda_0}^{\lambda_0+\epsilon} \left|\phi(e^{-i\nu})\right|^2 \mathrm{d}F(\nu) = \int_{\lambda_0}^{\lambda_0+\epsilon} \left|\theta(e^{-i\nu})\right|^2 \frac{\sigma^2}{2\pi}\mathrm{d}\nu$$

for any $\epsilon \in [0, \epsilon_0]$. This implies that

$$\int_{\lambda_0}^{\lambda_0+\epsilon} C^2\epsilon^2 \mathrm{d}F(\nu) > \frac{\theta_0\sigma^2\epsilon}{4\pi}.$$

This implies that

$$F(\lambda_0+\epsilon) - F(\lambda_0) > \frac{\theta_0\sigma^2}{4\pi C^2\epsilon}.$$

Since it holds for all $\epsilon \in [0, \epsilon_0]$, we may take $\epsilon \to 0$ and conclude that $F(\lambda_0+\epsilon) = \infty$, contradicting with the fact that the spectral distribution function of a weakly stationary process is bounded above by the variance of the weakly stationary process. Therefore, the ARMA equation cannot have any weakly stationary solution. ■

To estimate the spectral density of an ARMA(p, q) process $\alpha(L)X_t = \theta(L)\varepsilon_t$, we may took the general non-parametric estimator (3.3), or we could first estimate the ARMA parameters of the process, and obtain the estimated polynomials $\hat{\alpha}(z)$ and $\hat{\theta}(z)$. Then we may estimate the spectral density as

$$\hat{f}(\lambda) = \frac{1}{2\pi}\hat{G}(e^{-i\lambda}) = \frac{\hat{\sigma}^2}{2\pi}\frac{\hat{\theta}(e^{-i\lambda})\hat{\theta}(e^{i\lambda})}{\hat{\alpha}(e^{-i\lambda})\hat{\alpha}(e^{i\lambda})}.$$

If the parameter estimators are consistent, one would expect that the spectral density estimator is consistent under some regularity conditions.

5.7 Forecasting

5.7.1 Principles of Forecasting

Suppose we are interested in forecasting Y given a set of variables $X_1, X_2, \ldots$. Let $\hat{Y}$ be a forecast of Y. To evaluate the performance of the forecast, we need to specify a measurement of forecast error. A frequently used measure is the *mean square error* defined as

$$\mathrm{MSE}(\hat{Y}) = \mathbb{E}(Y - \hat{Y})^2.$$

The optimal forecast $\hat{Y}$ is a function g of $X_1, X_2, \ldots$ that minimizes

$$\mathbb{E}\left(Y - g(X_1, X_2, \ldots)\right)^2.$$

Note that

$$\begin{aligned}\mathbb{E}\left[Y - g(X_1, X_2, \ldots)\right]^2 &= \mathbb{E}\left[Y - \mathbb{E}(Y|X_1, X_2, \ldots) + \mathbb{E}(Y|X_1, X_2, \ldots) - g(X_1, X_2, \ldots)\right]^2 \\ &= \mathbb{E}[Y - \mathbb{E}(Y|X_1, X_2, \ldots)]^2 + \mathbb{E}[\mathbb{E}(Y|X_1, X_2, \ldots) - g(X_1, X_2, \ldots)]^2 \\ &\quad + 2\mathbb{E}[(Y - \mathbb{E}(Y|X_1, X_2, \ldots))(\mathbb{E}(Y|X_1, X_2, \ldots) - g(X_1, X_2, \ldots))].\end{aligned}$$

Since

$$\mathbb{E}\left[\left[Y - \mathbb{E}(Y|X_1, X_2, \ldots)\right]\left[\mathbb{E}(Y|X_1, X_2, \ldots) - g(X_1, X_2, \ldots)\right]\middle| X_1, X_2, \ldots\right] = 0,$$

the unconditional expectation

$$\mathbb{E}[(Y - \mathbb{E}(Y|X_1, X_2, \ldots))(\mathbb{E}(Y|X_1, X_2, \ldots) - g(X_1, X_2, \ldots))]$$

is zero, and $\mathbb{E}\left[Y - g(X_1, X_2, \ldots)\right]^2$ is maximized at $g(X_1, X_2, \ldots) = \mathbb{E}(Y|X_1, X_2, \ldots)$. Therefore, the best forecast of Y given $X_1, X_2, \ldots$ is the conditional expectation of Y given $X_1, X_2, \ldots$.

If $Y, X_1, X_2, \ldots \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathbb{E}(Y|X_1, X_2, \ldots)$ could be viewed as the orthogonal projection of Y on the subspace of all (measurable) functions of $X_1, X_2, \ldots$.

In practice, it is not always clear what the conditional expectations should be. Therefore, we usually restrict our attention to the class of forecasts that could be expressed as a linear function of $X_1, X_2, \ldots$. The linear forecast, denoted by $\mathbb{L}(Y|X_1, X_2, \cdots)$, is obtained by the orthogonal projection of Y on the subspace of all linear functions of $X_1, X_2, \ldots$. Obviously, $\text{MSE}(\mathbb{E}(Y|X_1, X_2, \ldots)) \leq \text{MSE}(\mathbb{L}(Y|X_1, X_2, \ldots))$.

5.7.2 Linear Forecasting Based on an Infinite Number of Observations

Now let $X_t = \mu + \phi(L)\varepsilon_t = \mu + \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i}$ where $\varepsilon_t \sim \text{WN}(0, \sigma^2)$ and $\sum_{i=0}^{\infty} |\phi_i| < \infty$. Suppose we know the parameters μ and ϕ_i's and would like to forecast X_{t+h} based on $\varepsilon_t, \varepsilon_{t-1}, \ldots$. The optimal linear forecast $\mathbb{L}(X_{t+h}|1, \varepsilon_t, \varepsilon_{t-1}, \ldots) = \nu + \sum_{i=0}^{\infty} \beta_i \varepsilon_{t-i}$ should satisfy the orthogonal conditions

$$\mathbb{E}\left(X_{t+h} - \left(\nu + \sum_{i=0}^{\infty} \beta_i \varepsilon_{t-i}\right)\right) = 0$$

and

$$\mathbb{E}\left(X_{t+h} - \left(\nu + \sum_{i=0}^{\infty} \beta_i \varepsilon_{t-i}\right)\right) \varepsilon_j = 0$$

for all $j \leq t$. This implies that $\nu = \mu$ and $\beta_i = \phi_{i+h}$. That is,

$$\mathbb{L}(X_{t+h}|1, \varepsilon_t, \varepsilon_{t-1}, \ldots) = \mu + \phi_h \varepsilon_t + \phi_{h+1}\varepsilon_{t-1} + \cdots.$$

We may write it as

$$\mathbb{L}(X_{t+h}|1, \varepsilon_t, \varepsilon_{t-1}, \ldots) = \mu + \left[\frac{\phi(L)}{L^h}\right]_+ \varepsilon_t$$

where for any power series $\alpha(z)$, $[\alpha(z)]_+$ replaces the terms with negative powers in $\alpha(z)$ with zeros.

Now for $\{X_t\}$ that follows a causal ARMA(p,q) process $\alpha(L)X_t = c + \theta(L)\varepsilon_t$, given all the parameters and $\varepsilon_t, \varepsilon_{t-1}, \ldots$,

$$\mathbb{L}(X_{t+h}|1, \varepsilon_t, \varepsilon_{t-1}, \ldots) = \frac{c}{\alpha(1)} + \left[\frac{\theta(L)}{\alpha(L)L^h}\right]_+ \varepsilon_t.$$

If the process is also invertible and we know $X_t, X_{t-1}, \ldots$ instead of $\varepsilon_t, \varepsilon_{t-1}, \ldots$,

$$\hat{X}_{t+h|t} = L(X_{t+h}|1, X_t, X_{t-1}, \ldots) = \frac{c}{\alpha(1)} + \left[\frac{\theta(L)}{\alpha(L)L^h}\right]_+ \frac{\alpha(L)}{\theta(L)}\left(X_t - \frac{c}{\alpha(1)}\right).$$

This is known as the Wiener-Kolmogorov prediction formula.

The multiple-step-ahead forecasts could be computed in a recursive way. We first show that when $h = 1$, that is, when we make one-step-ahead forecast,

$$\left[\frac{\theta(L)}{\alpha(L)L}\right]_+ \frac{\alpha(L)}{\theta(L)} = \frac{1-\alpha(L)}{L} + \frac{\theta(L)-1}{L}\left(1 - \left[\frac{\theta(L)}{\alpha(L)L}\right]_+ \frac{\alpha(L)}{\theta(L)}L\right).$$

This can be shown by plugging

$$\begin{aligned}\left[\frac{\theta(L)}{\alpha(L)L}\right]_+ &= \frac{\theta(L)-1}{L}\alpha^{-1}(L) + \left[\frac{\alpha^{-1}(L)-1+1}{L}\right]_+ \\ &= \frac{\theta(L)-1}{L}\alpha^{-1}(L) + \frac{\alpha^{-1}(L)-1}{L}\end{aligned}$$

into the both sides of the equation we want to establish. This new representation implies that

$$\hat{X}_{t+1|t} - \mu = \alpha_1(X_t - \mu) + \alpha_2(X_{t-1} - \mu) + \cdots + \alpha_p(X_{t-p+1} - \mu) + \theta_1\hat{\varepsilon}_t + \cdots + \theta_q\hat{\varepsilon}_{t-q+1},$$

where

$$\hat{\varepsilon}_t = \left(1 - \left[\frac{\theta(L)}{\alpha(L)L}\right]_+ \frac{\alpha(L)}{\theta(L)}L\right)(X_t - \mu) = X_t - \hat{X}_{t|t-1},$$

which serves as an approximation or estimate of the unobserved ε_t. Now, by law of iterated projections,

$$\hat{X}_{t+2|t} = \mathbb{L}\Big(\mathbb{L}(X_{t+2}|1, X_{t+1}, X_t, \ldots)\Big|1, X_t, X_{t-1}, \ldots\Big),$$

then

$$\hat{X}_{t+2|t} - \mu = \alpha_1(\hat{X}_{t+1|t} - \mu) + \alpha_2(X_t - \mu) + \cdots + \alpha_p(X_{t-p+2} - \mu) + \theta_2\hat{\varepsilon}_t + \cdots + \theta_q\hat{\varepsilon}_{t-q+2}.$$

Note that $\mathbb{L}(\hat{\varepsilon}_{t+1}|1, X_t, X_{t-1}, \ldots) = 0$. In general, for $h = 1, 2, \ldots, q$,

$$\hat{X}_{t+h|t} - \mu = \alpha_1(\hat{X}_{t+h-1|t} - \mu) + \alpha_2(\hat{X}_{t+h-2|t} - \mu) + \cdots + \alpha_p(\hat{X}_{t+h-p|t} - \mu) + \theta_h\hat{\varepsilon}_t + \cdots + \theta_q\hat{\varepsilon}_{t+h-q},$$

and for $h = q+1, q+2, \ldots$,

$$\hat{X}_{t+h|t} - \mu = \alpha_1(\hat{X}_{t+h-1|t} - \mu) + \alpha_2(\hat{X}_{t+h-2|t} - \mu) + \cdots + \alpha_p(\hat{X}_{t+h-p|t} - \mu).$$

Note that we have used the fact that $X_{s|t} = X_s$ for $s \le t$.

5.7.3 Linear Forecasting Based on a Finite Number of Observations

When we have only a finite number of observations $X_t, X_{t-1}, \cdots, X_{t-m+1}$ (but continue to assume that we know the parameters), we may still use the recursive method to make forecasts by setting $\hat{\varepsilon}_{t-m} = \hat{\varepsilon}_{t-m-1} = \cdots = 0$. The performance of this approximation depends on the model, and we shall not investigate in detail here.

Another way to make forecast based on a finite number of observations is to directly project X_{t+h} onto the space spanned by $1, X_t, X_{t-1}, \cdots, X_{t-m+1}$. Suppose that the forecast is given by

$$X_{t+h} - \mu = \beta_t^h(X_t - \mu) + \beta_{t-1}^h(X_{t-1} - \mu) + \cdots + \beta_{t-m+1}^h(X_{t-m+1} - \mu) + u_t^h = X_t^{h\prime}\beta_t^h + u_t^h.$$

The orthogonal condition

$$\mathbb{E}X_t^h(X_{t+h} - \mu - X_t^{h\prime}\beta_t^h) = 0$$

implies that

$$\begin{aligned}\beta_t^h &= \left(\mathbb{E}X_{t+h}X_{t+h}'\right)^{-1}\mathbb{E}X_{t+h}(X_{t+h} - \mu) \\ &= \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(m-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(m-2) \\ \vdots & \vdots & & \vdots \\ \gamma(m-1) & \gamma(m-2) & \cdots & \gamma(0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+m-1) \end{bmatrix}.\end{aligned}$$

5.7.4 Optimal Forecasting for Gaussian Processes

Theorem 5.10. *Let $Y =_d \mathbb{N}(\mu, \Sigma)$ and partition Y as $Y = (Y_1', Y_2')'$. Suppose μ and Σ are accordingly partitioned as*

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2, \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then the conditional distribution of Y_1 given Y_2 is $\mathbb{N}(\mu_{1\cdot2}, \Sigma_{11\cdot2})$ where $\mu_{1\cdot2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2)$ and $\Sigma_{11\cdot2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proof. It is easy to show that $Y_1 - \Sigma_{12}\Sigma_{22}^{-1}Y_2$ is uncorrelated with Y_2, and therefore, independent of Y_2. Write

$$Y_1 = (Y_1 - \Sigma_{12}\Sigma_{22}^{-1}Y_2) + \Sigma_{12}\Sigma_{22}^{-1}Y_2.$$

The conditional distribution of the first term on the right hand side above given Y_2, due to independence, is equal to its unconditional distribution, which is $\mathbb{N}(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

The second term is constant given Y_2. The results then follows immediately. ■

The above theorem shows that $\mathbb{E}(Y_1|Y_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2)$, which is a linear function of Y_2. In the forecast setting, when the time series $\{X_t\}$ is Gaussian, the optimal linear forecast (orthogonal projection, including the constant term) coincides with the optimal forecast (conditional expectation).

5.8 Estimation

In this section, we shall consider the Maximum Likelihood Estimation, to which the OLS estimation will be connected.

5.8.1 Estimating AR Models

Suppose we have data $X_1, X_2, \cdots, X_T$ from an AR(p) model

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t$$

where $\varepsilon_t \sim$ iid $\mathbb{N}(0, \sigma^2)$. We shall estimate $\Theta = (c, \alpha_1, \cdots, \alpha_p, \sigma^2)$ using MLE. Since X_t can be expressed as an infinity sum of $\{\varepsilon_t\}$, we have that $\{X_t\}$ is Gaussian, and the density function of $X^\circ = (X_1, X_2, \cdots, X_p)'$ given the parameters is

$$f(X_p, X_{p-1}, \ldots, X_1; \Theta) = (2\pi)^{-p/2} \det(\Sigma_p)^{-1/2} \exp\left(-\frac{(X^\circ - \mu^\circ)'\Sigma_p^{-1}(X^\circ - \mu^\circ)}{2}\right)$$

where $\mu = 1/(1 - \alpha_1 - \cdots - \alpha_p)$, $\mu^\circ = (\mu, \mu, \cdots, \mu)'$ is the mean of X°, and

$$\Sigma_p = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(p-2) \\ \vdots & \vdots & & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{bmatrix}$$

is the covariance matrix of X°. The autocovariances could be obtained by the Yule-Walker equations or using the autocovariance generating function.

For $t > p$, the conditional density function of X_t given $X_{t-1}, \cdots, X_1$ and the parameters is

$$f(X_t|X_{t-1,X_t-2}, \cdots, X_1; \Theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_t - c - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p})^2}{2\sigma^2}\right).$$

Now the likelihood of the full sample is

$$f(X_T, X_{T-1}, \cdots, X_1; \Theta) = \left[\prod_{t=p+1}^{T} f(X_t|X_{t-1,X_t-2}, \cdots, X_1; \Theta)\right] f(X_p, X_{p-1}, \ldots, X_1; \Theta),$$

or in the form of log-likelihood (normalized by $1/T$):

$$\ell(\Theta) = \frac{1}{T} \ln f(X_p, X_{p-1}, \ldots, X_1; \Theta) + \frac{1}{T} \sum_{t=p+1}^{T} \ln f(X_t | X_{t-1, X_t-2}, \cdots, X_1; \Theta).$$

The MLE estimator $\hat{\Theta} = (\hat{c}, \hat{\alpha}_1, \cdots, \hat{\alpha}_p, \hat{\sigma}^2)$ maximizes the log-likelihood $\ell(\Theta)$. Note that the mean of the process is estimated as $\hat{\mu} = 1/(1 - \hat{\alpha}_1 - \cdots - \hat{\alpha}_p)$.

Unfortunately, due to the existence of the unconditional density $f(X_p, X_{p-1}, \ldots, X_1; \Theta)$, the maximization problem does not have an analytical solution, and we have to rely on numerical optimization methods. For an introduction of popular numerical optimization algorithms, see Section 6.3 or Hamilton (1994, Section 5.7) for example.

The conditional maximum likelihood estimates (CMLE) are also frequently used. In CMLE, we look at

$$f(X_T, X_{T-1}, \cdots, X_{p+1} | X_p, X_{p-1}, \cdots, X_1; \Theta),$$

that is, we look at the likelihood conditional on the first p observations. This (normalized) log-likelihood is given by

$$\begin{aligned} &\frac{1}{T} \sum_{t=p+1}^{T} \ln f(X_t | X_{t-1}, X_{t-2}, \cdots, X_1; \Theta) \\ =&\frac{C}{T} - \frac{T-p}{2T} \ln \sigma^2 - \frac{1}{2\sigma^2 T} \sum_{t=p+1}^{\infty} (X_t - c - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p})^2, \end{aligned}$$

where C is a constant. Note that the CMLE estimators $\hat{c}, \hat{\alpha}_1, \ldots, \hat{\alpha}_p$ are the same as the OLS estimators. The CMLE estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{\infty} (X_t - \hat{c} - \hat{\alpha}_1 X_{t-1} - \cdots - \hat{\alpha}_p X_{t-p})^2$$

We make a remark here that the MLE estimator and the CMLE estimator have the same asymptotic distribution, and are both consistent.[1] Since CMLE estimator is the same as the OLS estimator, and the OLS estimator does not rely on the distributional assumptions on ε_t, then even if the distribution of ε_t is not normal, we may still obtain the quasi-maximum likelihood estimator under the normality assumption, and the QMLE is consistent.

[1] The two (normalized) log-likelihood functions differ by $\frac{1}{T} \ln f(X_p, X_{p-1}, \ldots, X_1; \Theta)$, which is $o_p(1)$. This implies that their maximizers should be very close to each other in probability when T is large, under some uniform convergence condition on the log-likelihood functions. For details and a general theory of extremum estimator, and in particular, the maximum likelihood estimator, see Newey and McFadden (1994).

5.8.2 Estimating MA Models

Suppose we have data $X_1, X_2, \ldots, X_T$ from an MA(q) model

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where $\varepsilon_t \sim$ iid $\mathbb{N}(0, \sigma^2)$. We shall estimate $\Theta = (\mu, \theta_1, \cdots, \theta_q, \sigma^2)$ using MLE.

To obtain the exact likelihood function, we write $Y^\circ = \mu^\circ + A\varepsilon^\circ$ where $Y^\circ = (Y_1, Y_2, \ldots, Y_T)'$, $\mu^\circ = (\mu, \mu, \ldots, \mu)'$, $\varepsilon^\circ = (\varepsilon_{1-q}, \varepsilon_{2-q}, \ldots, \varepsilon_T)'$, and A is an appropriate matrix whose entries contain the θ's. Since ε° is normal with mean zero and variance $\sigma^2 I$, where I is the $(T+q)$ by $(T+q)$ dimensional identity matrix, then $Y^\circ =_d \mathbb{N}(\mu^\circ, \sigma^2 AA')$, and the log likelihood function is given by

$$(2\pi)^{-T/2} \det(\sigma^2 AA')^{-1/2} \exp\left(-\frac{1}{2\sigma^2} (Y^\circ - \mu)'(AA')^{-1}(Y^\circ - \mu) \right).$$

The conditional likelihood is obtained as the density of $X_T, X_{T-1}, \cdots, X_1$ conditional on the initial innovations $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{1-q}$. Given the initial innovations, we may obtain $\{\varepsilon_t\}_{t=1}^T$ recursively through

$$\varepsilon_t = X_t - \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}.$$

The conditional log likelihood is given by

$$\begin{aligned}
\ell(\Theta) &= \frac{1}{T} \ln f(X_T, X_{t-1}, \ldots, X_1 | \varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{1-q}; \Theta) \\
&= \frac{1}{T} \sum_{t=1}^{T} \ln f(X_t | X_{t-1}, X_{t-2}, \ldots, X_1, \varepsilon_0, \ldots, \varepsilon_{1-q}; \Theta) \\
&= \frac{C}{T} - \frac{1}{2} \ln \sigma^2 - \frac{1}{T} \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2},
\end{aligned}$$

where C is a constant. Since the initial innovations are not observed, we may set $\varepsilon_0 = \varepsilon_{-1} = \ldots = \varepsilon_{1-q} = 0$ as an approximation. This approximation is sensible only if the MA process is invertible. Otherwise, the effect of $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{1-q}$ will accumulate over time.

5.8.3 Estimating ARMA Processes

Suppose we have data $X_1, X_2, \ldots, X_T$ from an ARMA(p, q) model

$$X_t = c + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\varepsilon_t \sim$ iid $\mathbb{N}(0, \sigma^2)$. We shall estimate $\Theta = (c, \alpha_1, \alpha_1, \cdots, \alpha_p, \theta_1, \cdots, \theta_q, \sigma^2)$ using MLE.

We may obtain the conditional likelihood of the $X_T, X_{T-1}, \ldots, X_{p+1}$ given $X_p, X_{p-1}, \ldots, X_1$ and $\varepsilon_p, \varepsilon_{p-1}, \ldots, \varepsilon_{p-q+1}$. With these initial values, we may recursively obtain $\{\varepsilon_t\}_{t=p+1}^T$ by

$$\varepsilon_t = X_t - c - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}.$$

The conditional log likelihood is then

$$\ell(\Theta) = \frac{C}{T} - \frac{T-p}{2T}\ln\sigma^2 - \frac{1}{T}\sum_{t=p+1}^{T}\frac{\varepsilon_t^2}{2\sigma^2},$$

where C is a constant. Once again, we may set $\varepsilon_p = \varepsilon_{p-1} = \cdots = \varepsilon_{p-q+1} = 0$ if the ARMA process is invertible.

To obtain the exact likelihood for an ARMA process, we will use the Kalman filter, which will be introduced in a later chapter.

5.8.4 Asymptotic Properties of the Estimators

Under some regularity conditions, the MLE or CMLE estimators turn out to be consistent and asymptotically normal. If fact,

$$\sqrt{T}(\hat{\Theta}_n - \Theta_0) \to_d \mathbb{N}\left(0, -H^{-1}\right)$$

where H is the probability limit of $\frac{\partial^2\ell(\Theta)}{\partial\Theta\partial\Theta'}$ evaluated at the true value Θ_0 of Θ, which happens to be non-random, i.e.,

$$\frac{\partial^2\ell(\Theta_0)}{\partial\Theta\partial\Theta'} \to_p H.$$

We may estimate H by $\frac{\partial^2\ell(\hat{\Theta}_n)}{\partial\Theta\partial\Theta'}$, the Hessian of the log likelihood function evaluated at the estimated Θ.

The theory of maximum likelihood estimation justifies another estimator for the asymptotic variance: $I = -H$ where I is the probability limit of $T\frac{\partial\ell(\Theta)}{\partial\Theta}\frac{\partial\ell(\Theta)}{\partial\Theta'}$ evaluated at the true value Θ_0 of Θ, which also happens to be non-random, i.e.,

$$T\frac{\partial\ell(\Theta_0)}{\partial\Theta}\frac{\partial\ell(\Theta_0)}{\partial\Theta'} \to_p I.$$

We may estimate I by $T\frac{\partial\ell(\hat{\Theta}_n)}{\partial\Theta}\frac{\partial\ell(\hat{\Theta}_n)}{\partial\Theta'}$.

For a careful treatment of the theory of maximum likelihood estimation, see Chapter 6.

5.9 Model Selection

To determine the order of an MA process $\{X_t\}$, we may check the autocorrelation function (ACF) of $\{X_t\}$. If $\{X_t\}$ follows an MA(q_0), then we have $\rho(k) = 0$ for all $k > q_0$. We may estimate the sample autocorrelation function by $\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$, and for a MA($q_0$) process, under appropriate conditions (see Theorem 2.33 and the comments below it), we can show that

$$\sqrt{T}\hat{\rho}(k) \to_d \mathbb{N}\left(0, \sum_{i=-q}^{q}\rho^2(i)\right)$$

for $k > q_0$.

To determine the order of an causal AR process $\{X_t\}$, we may check the *partial autocorrelation function* of $\{X_t\}$. The partial autocorrelation function PACF(k) at lag k for a weakly stationary time series $\{X_t\}$ is defined to be the coefficient in front of X_{t-k} in $\mathbb{L}(X_t|1, X_{t-1}, X_{t-2}, \ldots, X_{t-k})$. It is easy to see that if the true model is AR(p_0), then PACF$(k) = 0$ for all $k > p_0$. The lag-k partial autocorrelation could be obtained as the OLS estimate of α_{kk} in the regression model

$$X_t = \alpha_{k0} + \alpha_{k1} X_{t-1} + \cdots + \alpha_{kk} X_{t-k} + \varepsilon_{kt},$$

and for an AR(p_0) model, using the two-step regression method introduced in Section 1.3.2, we can show that

$$\sqrt{T}\hat{\alpha}_{kk} \to_d \mathbb{N}(0, 1)$$

for $k > p_0$.

To determine the order (p, q) of a causal and invertible ARMA(p, q) model, we may apply the Akaike Information Criteria (AIC) or the Bayesian Information Criteria (BIC). For any p and q, we estimate the model using MLE, getting the fitted residuals $\hat{\varepsilon}_t$, and calculate

$$\hat{\sigma}^2_{p,q} = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}^2_t.$$

The AIC and BIC are defined respectively as

$$\text{AIC}(p, q) = \ln \hat{\sigma}^2_{p,q} + \frac{2(p + q)}{T}$$

and

$$\text{BIC}(p, q) = \ln \hat{\sigma}^2_{p,q} + \frac{(p + q) \ln T}{T}.$$

The estimated orders $\hat{p}, \hat{q}$ are the pair of (p, q) that minimizes AIC or BIC.

Suppose that the true model is an ARMA(p_0, q_0) model. Under the assumption that ε_t is iid normal, we may show that $\hat{p} \to_p p_0$ and $\hat{q} \to_p q_0$ if we use the BIC. Actually we may replace the penalty term in BIC with $\frac{(p+q)C(T)}{T}$ for any $C(T)$ that diverges to infinity as $T \to \infty$ and still get the consistency result hold. Unfortunately the order estimators using AIC are not consistent. The model selected by AIC tends to overfit. That is, the AIC tends suggest orders that are greater than the true orders. See Hannan (1980) for details. Of course, the information criteria also work for the AR and the MA models.

6 Extremum Estimation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space and $x_1, x_2, \cdots, x_n$ be random variables or random vectors taking values in $\mathcal{X}$. Let Θ be the parameter space. Let $Q_n(X, \theta)$ be a function from $\mathcal{X}^n \times \Theta$ to $\mathbb{R}$. Then $\hat{\theta}_n$ is called an extremum estimator if

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(X, \theta).$$

In particular, if $Q_n(X, \theta) = \frac{1}{n}\sum_{i=1}^n q(x_i, \theta)$ for some $q(x, \theta)$ from $\mathcal{X} \times \Theta$ to $\mathbb{R}$, then $\hat{\theta}_n$ is called an M-estimator. Obviously, M-estimators are extremum estimators. In this chapter we present a general asymptotic theory for extremum estimation and M-estimation.

6.1 Asymptotic Consistency

The following is a general result for consistency. This theorem holds for general minimization estimators, including the M-estimators.

Theorem 6.1. *Let Θ be a metric space and $\{Q_n\}$ be a sequence of real random functions defined on Θ. Let $\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta)$. If there exists a function $Q : \Theta \to \mathbb{R}$ and $\theta_0 \in \Theta$ such that*

(a) for each open ball N of θ_0, $Q(\theta_0) < \inf_{\theta \in \Theta \backslash N} Q(\theta)$, and

(b) $Q_n(\theta)$ converges uniformly in probability to $Q(\theta)$ as $n \to \infty$, i.e.,

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \to_p 0,$$

then $\hat{\theta}_n \to_p \theta_0$ as $n \to \infty$.

It should be noted here that Q_n is a function from $\Omega \times \Theta$ to $\mathbb{R}$. When we write $Q_n(\theta)$, we understand it as $Q_n(\cdot, \theta)$ and treat it as a random function. The condition (a) in the theorem is a separation assumption. It requires not only that θ_0 is the unique minimizer of Q, but the minimum is well separated. The convergence in condition (b) should be understood as in outer probability $\mathbb{P}^*$. Note that even if $Q_n(\theta)$ is measurable for each $\theta \in \Theta$ and for each n, $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)|$ may still be not measurable. For more details on outer probability (outer measure), one may refer to Billingsley (1995, Section 11).

Proof of Theorem 6.1. Let ρ be the metric in Θ. By the separation assumption (a), for any $\epsilon > 0$, there exists $\delta > 0$ such that $Q(\theta) - Q(\theta_0) > \delta$ for all $\theta \in \Theta \backslash N$, where N is the ϵ-ball in Θ entered at θ_0. Thus,

$$\mathbb{P}\Big(\rho(\hat{\theta}_n, \theta_0) \geq \varepsilon\Big) \leq \mathbb{P}^*\Big(\Big|Q(\hat{\theta}_n) - Q(\theta_0)\Big| > \delta\Big).$$

Then consistency of $\hat{\theta}_n$ follows immediately from that

$$\begin{aligned}\left|Q(\hat{\theta}_n) - Q(\theta_0)\right| &= Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\hat{\theta}_n) - Q(\theta_0) \\ &\leq \sup_{\theta\in\Theta} |Q(\theta) - Q_n(\theta)| + Q_n(\theta_0) - Q(\theta_0) \\ &\leq 2\sup_{\theta\in\Theta} |Q(\theta) - Q_n(\theta)| \to_p 0.\end{aligned}$$

■

It can be easily seen from the proof that $\hat{\theta}_n$ do not have to be the exact minimizer of the objective function $Q_n(\theta)$. This condition can be replaced by that $Q_n(\hat{\theta}_n) \leq \inf_{\theta\in\Theta} Q_n(\theta) + o_p(1)$ without affecting the result. If we replace the uniform convergence in probability of Q_n to uniform almost sure convergence, then we get strong consistency of $\hat{\theta}_n$, i.e., $\hat{\theta}_n \to_{a.s.} \theta_0$.

Condition (a) holds if Θ is a compact metric space, $Q : \Theta \to \mathbb{R}$ is lower semicontinuous,[1] and θ_0 is the unique minimizer of Q. In fact, fix $\epsilon > 0$ and let N be the ϵ-ball centered at θ_0. Then $\Theta\backslash N$, being a closed subset of a compact set, is compact. A lower semicontinuous real function on a compact set attains its minimum. The results then follows easily. This result leads to the following corollary of Theorem 6.1 immediately, which appears in Newey and McFadden (1994).

Corollary 6.2. Let Θ be a compact metric space and $\{Q_n\}$ be a sequence of real random functions defined on Θ. Let $\hat{\theta}_n = \arg\min_{\theta\in\Theta} Q_n(\theta)$. If there exists a continuous function $Q : \Theta \to \mathbb{R}$ such that $Q(\theta)$ is uniquely minimized at $\theta = \theta_0$, and that $Q_n(\theta)$ converges uniformly in probability to $Q(\theta)$ as $n \to \infty$, then $\hat{\theta}_n \to_p \theta_0$ as $n \to \infty$.

The assumption of compactness of Θ may be replaced by a local convexity assumption. For details, see Newey and McFadden (1994, Section 2.6).

The uniform convergence in probability assumption in the above theorem is an abstract one. In the context of M-estimators, the uniform law of large numbers (ULLN) provides simple conditions that guarantee uniform convergence in probability.

Definition 6.3. Let $\{Q_n\}$ be a sequence of real random functions on Θ, where Θ is a metric space with metric ρ. The sequence $\{Q_n\}$ is said to be *stochastically equicontinuous* if for any $\epsilon > 0$ and $\eta > 0$ there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\theta\in\Theta} \sup_{\{\theta'\in\Theta:\rho(\theta,\theta')<\delta\}} \left|Q_n(\theta) - Q_n(\theta')\right| > \epsilon\right) < \eta.$$

Definition 6.4. A metric space (Θ, ρ) is said to be *totally bounded* if for any $\epsilon > 0$ there exists a finite collection of open balls in Θ of radius ϵ whose union is Θ.

Obviously, any compact metric space is totally bounded.

[1] A function $f : X \to \mathbb{R}$ is called lower semicontinuous if $f^{-1}((a, \infty))$ is open for each $a \in \mathbb{R}$. A continuous function is lower semicontinuous.

Theorem 6.5. *Let (Θ, ρ) be a metric space and $\{Q_n\}$ be a sequence of real random functions defined on Θ.*

(a) If

$$\sup_{\theta\in\Theta} |Q_n(\theta)| \to_p 0,$$

then $\{Q_n\}$ is stochastically equicontinuous.

(b) If $\{Q_n\}$ is stochastically equicontinuous, $Q_n(\theta) \to_p 0$ for all $\theta \in \Theta$, and Θ is totally bounded, then

$$\sup_{\theta\in\Theta} |Q_n(\theta)| \to_p 0.$$

Proof. The proof for part (a) is trivial. For part (b), fix $\epsilon > 0$ and $\eta > 0$. By stochastic equicontinuity of $\{Q_n\}$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\theta\in\Theta} \sup_{\{\theta'\in\Theta:\rho(\theta,\theta')<\delta\}} \left|Q_n(\theta) - Q_n(\theta')\right| > \frac{\epsilon}{2} \right) < \eta.$$

Since Θ is totally bounded, there exists δ-balls centered at $t_1, t_2, \cdots t_r$ that covers Θ. Define the function $h : \Theta \to \{t_1, \cdots, t_r\}$ such that $\rho(h(\theta), \theta) < \delta$. Then for all n, we have that

$$\begin{aligned}
\sup_{\theta\in\Theta} |Q_n(\theta)| &\le \sup_{\theta\in\Theta} |Q_n(\theta) - Q_n(h(\theta))| + \sup_{\theta\in\Theta} |Q_n(h(\theta))| \\
&\le \sup_{\theta\in\Theta} \sup_{\{\theta'\in\Theta:\rho(\theta,\theta')<\delta\}} \left|Q_n(\theta) - Q_n(\theta')\right| + \max_{1\le i\le r} |Q_n(t_i))| .
\end{aligned}$$

Then

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\theta\in\Theta} |Q_n(\theta)| > \epsilon \right) \\
&\le \limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\theta\in\Theta} \sup_{\{\theta'\in\Theta:\rho(\theta,\theta')<\delta\}} \left|Q_n(\theta) - Q_n(\theta')\right| > \frac{\epsilon}{2} \right) + \limsup_{n\to\infty} \mathbb{P}^*\left(\max_{1\le i\le r} |Q_n(t_i))| > \frac{\epsilon}{2} \right) \\
&< \eta.
\end{aligned}$$

Note that in the last inequality we have used the fact that $Q_n(\theta) \to_p 0$ for all θ implies that $\mathbb{P}(\max_{1\le i\le r} |Q_n(t_i))| > \frac{\epsilon}{2}) \to 0$. Since η is arbitrary, our conclusion follows immediately. ■

Theorem 6.6. *Suppose that $\{x_i\}$ is a sequence of random variables or vectors taking values in $\mathcal{X}$, and $q_i : \mathcal{X} \times \Theta \to \mathbb{R}$ is a function. Suppose $q_i(x_i, \theta)$ is measurable for each i and each $\theta \in \Theta$ so that we may define*

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n q_i(x_i, \theta) \quad \text{and} \quad \bar{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}q_i(x_i, \theta).$$

If

(a) Θ is a compact metric space,

(b) for each $\theta \in \Theta, Q_n(\theta) - \bar{Q}_n(\theta) \to_p 0$, *and*

(c) there exists measurable functions $b_t : \mathcal{X} \to \mathbb{R}_+$ *and* $h : \mathbb{R}_+ \to \mathbb{R}_+$ *with* $h(0) = 0$ *and* h *continuous at* 0 *such that* $\frac{1}{n}\sum_{i=1}^n \mathbb{E}b(x_i) = O(1)$, *and* $|q_i(x_i, \theta) - q_i(x_i, \theta')| \leq b_i(x_i)h(\rho(\theta, \theta'))$ *for* $\theta, \theta' \in \Theta$, *where* ρ *is the metric on* Θ.

Then $\bar{Q}_n(\theta)$ *is stochastically equicontinuous, and*

$$\sup_{\theta \in \Theta} \left| Q_n(\theta) - \bar{Q}_n(\theta) \right| \to_p 0.$$

Proof. Let $B_n = \frac{1}{n}\sum_{i=1}^n b(x_i)$. By assumption, we have $B_n = O_p(1)$. Note

$$\left| Q_n(\theta) - Q_n(\theta') \right| \leq B_n h(\rho(\theta, \theta')).$$

This implies that $Q_n(\theta)$ is stochastically equicontinuous. Also,

$$\begin{aligned} \left| \bar{Q}_n(\theta) - \bar{Q}_n(\theta') \right| &\leq \mathbb{E} \left| Q_n(\theta) - Q_n(\theta') \right| \\ &\leq O(1)h(\rho(\theta, \theta')), \end{aligned}$$

This implies that $\bar{Q}_n(\theta)$ is equicontinuous. Then $Q_n(\theta) - \bar{Q}_n(\theta)$ is stochastically equicontinuous. The results follows immediately from Part (b) of Theorem 6.5. ■

Note that this theorem gives conditions under which pointwise convergence in probability (part (b) in the assumption) implies uniform convergence in probability. For further details, see Newey (1991). Also, if we consider Borel σ-algebras, then continuous functions are measurable. Here we make this assumption implicit.

Corollary 6.7. Suppose $\{x_i\}$ is a sequence of iid or strictly stationary and ergodic sequence of random variables or vectors taking values in $\mathcal{X}$. Suppose $q : \mathcal{X} \times \Theta$ is continuous at each $\theta \in \Theta$ with probability one, Θ is a compact metric space, and $|q(x_i, \theta)| \leq d(x_i)$ for some $d : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}d(x_i) < \infty$. Then

$$\sup_{\theta \in \Theta} \left| \frac{1}{n}\sum_{i=1}^n q(x_i, \theta) - \mathbb{E}q(x_i, \theta) \right| \to_p 0,$$

and $\mathbb{E}(x_i, \theta)$ is continuous.

Note that strictly stationarity and ergodicity of the data process and almost sure continuous of the q function guarantee the pointwise convergence in probability of $\frac{1}{n}\sum_{i=1}^n q(x_i, \theta)$.

Theorem 6.8. *Let* $\{x_i\}$ *be a sequence of random variables or vectors taking values in* $\mathcal{X}$, $q : \mathcal{X} \times \Theta \to \mathbb{R}, Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n q(x_i, \theta), \hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta)$. *If*

(a) Θ *is a compact metric space,*

(b) $\{x_i\}$ *is iid or strictly stationary and ergodic,*

(c) $\mathbb{E}q(x_i, \theta) > \mathbb{E}q(x_i, \theta_0)$ *for all* $\theta \neq \theta_0$, *and*

(d) $q : \mathcal{X} \times \Theta$ *is continuous at each* $\theta \in \Theta$ *with probability one, and* $|q(x_i, \theta)| \leq d(x_i)$ *for some* $d : \mathcal{X} \to \mathbb{R}$ *such that* $\mathbb{E}d(x_i) < \infty$.

Then $\hat{\theta}_n \to_p \theta_0$.

Proof. Note that if $\{x_t\}$ is strictly stationary and ergodic, then the process $\{y_t\}$ defined by $y_t = \phi(x_t, x_{t-1}, \cdots)$ is strictly stationary and ergodic for any measurable function $\phi : \mathbb{R}^\infty \to \mathbb{R}^k$. A law of large number then holds. Then the results follows immediately from the previous results. ■

6.2 Asymptotic Normality

Theorem 6.9. *Let* $\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta)$ *where* Q_n *is a function of the data points* $x_1, \ldots, x_n$ *and the parameters* θ, *and* $\hat{\theta}_n \to_p \theta_0$ *for some* θ_0 *in the interior of the convex set* $\Theta \subset \mathbb{R}^m$. *Suppose that* $Q_n(\theta)$ *is twice continuously differentiable in (the interior of)* Θ, $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \to_d \mathbb{N}(0, \Sigma)$, *and that* $H_n(\tilde{\theta}_n) \to_p H$ *as long as* $\tilde{\theta}_n \to_p \theta_0$, *where* $H_n(\cdot) = \frac{\partial^2 Q_n(\cdot)}{\partial \theta \partial \theta'}$ *is the Hessian of* Q_n, *and* H *is a nonstochastic positive definite matrix. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathbb{N}(0, H^{-1}\Sigma H^{-1}).$$

Proof. By definition of the maximizer and the differentiability of $Q_n(\theta)$, we have that

$$\sqrt{n}\frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta^i} = 0$$

for each $i = 1, 2, \ldots, m$, where θ^i is the i-th component of θ. The mean value theorem gives that

$$\sqrt{n}\frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta^i} = \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta^i} + \frac{\partial^2 Q_n(\tilde{\theta}_n^i)}{\partial \theta^i \partial \theta'}\sqrt{n}(\hat{\theta}_n - \theta_0),$$

where $\tilde{\theta}_n^i$ is some point in Θ such that $\left\|\tilde{\theta}_n^i - \theta_0\right\| \leq \left\|\hat{\theta}_n - \theta_0\right\|$. If we stack the equation for $i = 1, \ldots, m$ together, we have

$$\sqrt{n}\frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta} = \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + H_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_m)\sqrt{n}(\hat{\theta}_n - \theta_0)$$

where $H_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$ is a matrix whose i-th row is $\frac{\partial^2 Q_n(\tilde{\theta}_n^i)}{\partial \theta^i \partial \theta'}$. By assumption, we have $H_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$ converges in probability to H.

Now we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -H_n(\tilde{\theta}_1, \ldots, \tilde{\theta}_d)^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \to_d \mathbb{N}(0, H^{-1}\Sigma H^{-1}).$$

■

A sufficient condition for "$H_n(\tilde{\theta}_n) \to_p H$ as long as $\tilde{\theta}_n \to_p \theta_0$" is $\sup_{\theta \in \Theta} |H_n(\theta) - H(\theta)| \to_p 0$

for some $H(\cdot)$ continuous at θ_0. To see this, note

$$\left|H_n(\hat{\theta}_n) - H(\theta_0)\right| \leq \left|H_n(\hat{\theta}_n) - H(\hat{\theta}_n)\right| + \left|H_n(\hat{\theta}_n) - H(\theta_0)\right|.$$

We also note here that if the minimum appears on the boundary of Θ, the asymptotic normality can fail. See Newey and McFadden (1994, p. 2144) for examples.

6.3 Numerical Optimization Methods

In this section we introduce some numerical optimization methods in computing extremum estimators. Our goal is to find

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta)$$

where θ is a p-dimensional parameter in $\mathbb{R}^p$, given data $(x_1, \ldots, x_n)$.

One way to obtain the minimizer is to find the analytical solution to the optimization problem, if such a solution exists. If such a task is difficult, then one probably has to rely on some numerical methods. One may conduct a brute-force grid search for minimum. However, this only works in the case where p is small. When p is large, the curse-of-dimensionality kicks in, and in general, it is not possible to conduct the search within computing powers that are accessible to most of us. In the following we introduce some useful methods. For simplicity, we treat data as fixed values instead of random elements in the following, until we start to examine the probabilistic properties of these methods.

6.3.1 Newton-Raphson Method

Suppose we want to use the first order condition

$$f_n(\hat{\theta}_n) = \frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta} = 0$$

to solve for $\hat{\theta}_n$. Suppose we have $\theta_{n,k}$ that is close to $\hat{\theta}_n$. If f_n admits a Taylor expansion around $\theta_{n,k}$, we have

$$0 = f_n(\hat{\theta}_n) = f_n(\theta_{n,k}) + F_n(\theta_{n,k})(\hat{\theta}_n - \theta_{n,k}) + o\left(\left\|\hat{\theta}_n - \theta_{n,k}\right\|\right),$$

where

$$F_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'}.$$

This implies that

$$\hat{\theta}_n \approx \theta_{n,k} - F_n^{-1}(\theta_{n,k}) f_n(\theta_{n,k}),$$

which suggests an iterative algorithm to calculate $\hat{\theta}_n$ by

$$\theta_{n,k+1} = \theta_{n,k} - F_n^{-1}(\theta_{n,k}) f_n(\theta_{n,k}), \quad k = 1, 2, \ldots.$$

We iterate until it converges. However, in general, Newton-Raphson method may or may not converge. In cases where it does not converge, we may try some modified Newton-Raphson method instead. The algorithm of default Newton-Raphson is given by

$$\theta^d_{n,k+1} = (1 - J_k)\theta^*_{n,k+1} + J_k\theta^+_{n,k+1}$$

where

$$\theta^*_{n,k+1} = \theta^d_{n,k} - F_n^{-1}(\theta^d_{n,k})f_n(\theta^d_{n,k}),$$
$$\theta^+_{n,k+1} = \theta^d_{n,k} - F_{n,k}^{-1}f_n(\theta^d_{n,k}),$$
$$J_k = 1\{Q_n(\theta^*_{n,k+1}) \geq Q_n(\theta^d_{n,k})\},$$

and $F_{n,k}$ is chosen so that $Q_n(\theta^+_{n,k+1}) < Q_n(\theta^d_{n,k})$.

The algorithm of line search Newton-Raphson is given by

$$\theta^M_{n,k+1} = \theta^{\alpha_k}_{n,k+1},$$

where

$$\alpha_k = \arg\min_{\alpha\in\mathcal{A}} Q_n(\theta_{n,k+1})^\alpha$$

for some $\mathcal{A} \subset [0, 1]$ and

$$\theta^\alpha_{n,k+1} = \theta^M_{n,k} - \alpha F_n^{-1}(\theta^M_{n,k})f_n(\theta^M_{n,k}).$$

6.3.2 Gauss-Newton Method

Suppose $Q_n(\theta)$ takes the form of

$$Q_n(\theta) = \frac{1}{2}r_n(\theta)'r_n(\theta)$$

where $r_n(\theta) = (r_{n1}(\theta), \ldots, r_{np}(\theta))'$. The first order derivative is

$$f_n(\theta) = \frac{\partial r_n'(\theta)}{\partial\theta}r_n(\theta),$$

and the second order derivative is

$$F_n(\theta) = \frac{\partial r_n'(\theta)}{\partial\theta}\frac{\partial r_n(\theta)}{\partial\theta'} + \sum_{i=1}^{k}\frac{\partial^2 r_{ni}(\theta)}{\partial\theta\partial\theta'}r_{ni}(\theta).$$

At around $\hat{\theta}_n$ we may ignore the second term in $F_n(\theta)$ and obtain the algorithm

$$\theta^G_{n,k+1} = \theta^G_{n,k} - G_n^{-1}(\theta^G_{n,k})f_n(\theta^G_{n,k})$$

where

$$G_n(\theta) = \frac{\partial r_n'(\theta)}{\partial\theta}\frac{\partial r_n(\theta)}{\partial\theta'}.$$

Theorem 6.10. *Suppose that conditions of Theorem 6.9 holds. Let $\bar{\theta}_n$ be an initial estimator for θ and $\bar{H}_n$ be an estimator of $\operatorname{plim}_{n\to\infty} \frac{\partial^2 \hat{Q}_n(\theta_0)}{\partial\theta\partial\theta'}$ such that $\bar{H}_n \to_p H$. Let $\tilde{\theta}_n = \bar{\theta}_n - \bar{H}_n^{-1}\frac{\partial \hat{Q}_n(\bar{\theta}_n)}{\partial\theta}$. If $\sqrt{n}(\bar{\theta}_n - \theta_0)$ is bounded in probability, then $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d \mathbb{N}(0, H^{-1}\Sigma H^{-1})$.*

Proof. We have that

$$\begin{aligned}
\sqrt{n}(\tilde{\theta}_n - \theta_0) &= \sqrt{n}(\bar{\theta}_n - \theta_0) - \sqrt{n}\bar{H}_n^{-1}\frac{\partial \hat{Q}_n(\bar{\theta}_n)}{\partial\theta} \\
&= \sqrt{n}(\bar{\theta}_n - \theta_0) - \sqrt{n}\bar{H}_n^{-1}\frac{\partial \hat{Q}_n(\theta_0)}{\partial\theta} - \sqrt{n}\bar{H}_n^{-1}\frac{\partial^2 \hat{Q}_n(\theta_n^*)}{\partial\theta\partial\theta'}(\bar{\theta}_n - \theta_0) \\
&= \left(I - \bar{H}_n^{-1}\frac{\partial^2 \hat{Q}_n(\theta_n^*)}{\partial\theta\partial\theta'}\right)\sqrt{n}(\bar{\theta}_n - \theta_0) - \sqrt{n}\bar{H}_n^{-1}\frac{\partial \hat{Q}_n(\theta_0)}{\partial\theta}
\end{aligned}$$

where θ_n^* is the mean value. Since $\bar{H}_n^{-1} \to_p H^{-1}$ and $\frac{\partial^2 \hat{Q}_n(\theta_n^*)}{\partial\theta\partial\theta'} \to_p H$, and $\sqrt{n}(\bar{\theta}_n - \theta_0)$ is bounded in probability, we have that the first term in the last expression converge in probability to zero. The second term converge in distribution to $\mathbb{N}(0, H^{-1}\Sigma H^{-1})$. This completes the proof of the theorem. ■

Note that $\tilde{\theta}_n$ is the one step iteration of $\hat{\theta}_n$ in the Newton-Raphson method or its variant. The theorem then shows that if we can start from a reasonably good estimator $\bar{\theta}_n$ (there is no restriction on the variance of the estimator), after one iteration, we would end up with an estimator that has the same asymptotic variance as the extreme estimator.

6.4 Asymptotics for Maximum Likelihood Estimation

We first consider the consistency and asymptotic normality of the conditional maximum likelihood estimators. Note that the conditional likelihood functions (normalized by $1/T$) in Chapter 5 are of the form

$$\frac{1}{T}\sum_{t=1}^{T} q(Y_t; \theta)$$

where θ is the vector of parameters and Y_t is a vector of data that appear in the conditional density in period t. To be consistent with notations used in the previous sections in this Chapter, I used θ instead of Θ to denote the parameter vector. Note that the normalizing constant $1/T$ may be replaced by $1/(T-p)$ in the AR cases. The q function in general takes the form of

$$q(Y_t, \theta) = r(\theta) + \frac{(X_t - c - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q})^2}{2\sigma^2}$$

where $Y_t = (X_t, \ldots, X_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q})'$, and $r(\theta)$ is some function of the parameter θ given by $\theta = (c, \alpha_1, \ldots, \alpha_p, \theta_1, \ldots, \theta_q, \sigma^2)'$. If we impose conditions such that terms like

$$\frac{1}{T}\sum_{t=1}^{T} X_i X_j \qquad \text{and} \qquad \frac{1}{T}\sum_{t=1}^{T} X_i \varepsilon_j$$

converge in probability respectively to $\mathbb{E}X_iX_j$ and $\mathbb{E}X_i\varepsilon_j$, then

$$\frac{1}{T}\sum_{t=1}^{T} q(Y_t;\theta) \to_p \mathbb{E}q(Y_t,\theta),$$

and we may use the results in the above two sections to establish the consistency and asymptotic normality of the MLE estimators, if we can show that $\mathbb{E}q(Y_t,\theta) > \mathbb{E}q(Y_t,\theta_0)$ for all $\theta \neq \theta_0$, where θ_0 is the true value of θ.

For the above condition to hold, we only need that θ_0 is identified. If we use $f(\cdot,\theta)$ to denote the conditional density function (such that $q(Y_t,\theta) = -\ln f(Y_t,\theta)$), identification of θ_0 means that if $\theta \neq \theta_0$ and $\theta \in \Theta$, then $f(Y_t,\theta) \neq f(Y_t,\theta_0)$. In the MLE setting, identification implies uniqueness of maximizer (minimizer) due to the information inequality:

$$\begin{aligned}
\mathbb{E}q(Y_t,\theta) - \mathbb{E}q(Y_t,\theta_0) &= -\mathbb{E}\left(\ln\frac{f(Y_t,\theta)}{f(Y_t,\theta_0)}\right) \\
&> -\ln\left(\mathbb{E}\frac{f(Y_t,\theta)}{f(Y_t,\theta_0)}\right) \\
&= -\ln\int\frac{f(y,\theta)}{f(y,\theta_0)}f(y,\theta_0)\mathrm{d}y \\
&= 0.
\end{aligned}$$

Note that the information inequality is a consequence of the Jensen's inequality, and the strictness comes from the identification assumption.

For MLE estimators, we may simplify its "sandwich" asymptotic variance using the relationship between the Fisher information and the Hessian. For notation convenience, we denote the observed data by Y, and denote $\mathbb{E}_\theta$ as the expectation with respect to the density $f(y;\theta)$. Furthermore, we introduce the following definitions related to the (normalized) log-likelihood function $\ell(\theta) = \ln f(y;\theta)$.

(a) (Score function) $s(\theta) = \frac{\partial\ell(\theta)}{\partial\theta}$.
(b) (Hessian) $h(\theta) = \frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'}$.
(c) ((Fisher) information) $I(\theta) = \mathbb{E}_\theta[s(\theta)s'(\theta)]$.
(d) (Expected Hessian) $H(\theta) = \mathbb{E}_\theta h(\theta)$.

Theorem 6.11. *Suppose* $\frac{\partial f(y;\theta)}{\partial\theta}$ *exists and* $\frac{\partial\int f(y;\theta)\mathrm{d}y}{\partial\theta} = \int\frac{\partial f(y;\theta)}{\partial\theta}\mathrm{d}y$. *Then*

$$\mathbb{E}_\theta s(\theta) = 0.$$

Furthermore, if $\frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta'}$ *exists and* $\frac{\partial^2\int f(y;\theta)\mathrm{d}y}{\partial\theta\partial\theta'} = \int\frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta'}\mathrm{d}y$, *then*

$$I(\theta) = -H(\theta).$$

Proof. We have

$$\begin{aligned}\mathbb{E}_\theta \frac{\partial \ell(\theta)}{\partial \theta} &= \int \frac{\partial \ln f(y;\theta)}{\partial \theta} f(y;\theta) \mathrm{d}y \\ &= \int \frac{\partial f(y;\theta)}{\partial \theta} \mathrm{d}y \\ &= \frac{\partial \int f(y;\theta) \mathrm{d}y}{\partial \theta} \\ &= \frac{\partial 1}{\partial \theta} = 0.\end{aligned}$$

The other result could be obtained by noting that

$$\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} = \frac{1}{f(y;\theta)} \frac{\partial^2 f(y;\theta)}{\partial \theta \partial \theta'} - \frac{\partial \ln f(y;\theta)}{\partial \theta} \frac{\partial \ln f(y;\theta)}{\partial \theta'}.$$

■

In our conditional maximum likelihood framework,

$$H_n(\tilde{\theta}_n) = \frac{\partial^2 Q_n(\tilde{\theta}_n)}{\partial \theta \partial \theta'} = \frac{\partial^2 \left(-\frac{1}{T} \sum_{t=1}^T \ell(\tilde{\theta}_n) \right)}{\partial \theta \partial \theta'} \to_p -H(\theta_0)$$

as $\tilde{\theta}_n \to_p \theta_0$, and

$$\sqrt{T} \frac{\partial Q_n(\theta_0)}{\partial \theta} = -\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{\partial \ell(\theta)}{\partial \theta} \to_d \mathbb{N}\left(0, I\left(\theta_0\right)\right).$$

Then the above theorem shows that

$$\sqrt{T}(\hat{\theta}_n - \theta_0) \to_d \mathbb{N}(0, -H^{-1}(\theta_0))$$

and

$$\sqrt{T}(\hat{\theta}_n - \theta_0) \to_d \mathbb{N}(0, I^{-1}(\theta_0)).$$

Note that if the objective function of conditional maximum likelihood estimation is given by $Q_n(\theta)$, then the objective function of maximum likelihood estimation is given by $Q_n(\theta) + o_p(1)$. Then by the remark after the proof of Theorem 6.1, the consistency and asymptotic normality of maximum likelihood estimators follow directly from the consistency and asymptotic normality of the corresponding conditional maximum likelihood estimator.

6.5 Other Topics

For conciseness, we have omitted many important topics in the general theory of extremum estimation, such as efficiency, two-step estimators and so on. For more discussion about these topics, see Newey and McFadden (1994, Section 5, 6, 8).

There are also estimators whose objective function is not as nice as required by the assumptions of the theorems in this chapter. Properties of estimators under such objective functions has been studied by various authors. For example, for results for asymptotic normality under non-smooth objective functions, see Newey and McFadden (1994, Section 7). Johansen and Nielsen (2019) give conditions for uniform boundedness of M-estimator in sample size in time series linear regressions where the objective function is possibly non-convex and non-continuous.

7 Vector Autoregressions

A large part of the theory for weakly stationary vector process is completely analogous to its counterpart of weakly stationary scalar process. Whenever this is the case, we shall point it out and not elaborate on the very details.

7.1 Vector Linear Processes

An n-dimensional *vector linear process* $\{X_t\}$ is a stochastic process that takes the form of

$$X_t = \sum_{i=0}^{\infty} \Phi_i \varepsilon_{t-i}$$

where $\varepsilon_t \sim \text{WN}(0, \Sigma)$ is a white noise process with covariance matrix Σ, and Φ_i are matrices whose (j,k)-th entry will be denoted by Φ_i^{jk}. The above process may be written using the lag operator as

$$X_t = \sum_{i=0}^{\infty} \Phi_i L^i \varepsilon_t = \Phi(L)\varepsilon_t,$$

where L is the lag operator that shift the series $\{\varepsilon_t\}$ backwards.

Following similar arguments as in Chapter 4, we may easily show that if $\sum_{i=0}^{\infty}(\Phi_i^{jk})^2 < \infty$ for all j and k, then $\sum_{i=0}^{\infty} \Phi_i \varepsilon_{t-i}$ is a well defined random vector in the space of all n-dimensional random vectors with finite covariance matrices in the mean square sense, and $\{X_t\}$ is weakly stationary. If $\sum_{i=0}^{\infty} \left|\Phi_i^{jk}\right| < \infty$ for all j and k, we have that $\sum_{k=-\infty}^{\infty} |\gamma(k)| < \infty$ where $\gamma(\cdot)$ is the autocovariance function of $\{X_t\}$. The autocovariance function of the above linear process is given by

$$\gamma(k) = \sum_{i=0}^{\infty} \Phi_{i+k} \Sigma \Phi_i'.$$

If $\sum_{i=0}^{\infty} i \left|\Phi_i^{jk}\right| < \infty$ for all j and k, we have the Beveridge-Nelson decomposition of X_t given by

$$X_t = \Phi(1)\varepsilon_t - (\tilde{X}_t - \tilde{X}_{t-1})$$

where $\tilde{X}_t = \sum_{i=0}^{\infty} \tilde{\Phi}_i \varepsilon_{t-i}, \tilde{\Phi}_i = \sum_{j=i+1}^{\infty} \Phi_j$.

The asymptotics for vector linear processes are also similar to their scalar counterpart. Given $\sum_{i=0}^{\infty} \left|\Phi_i^{jk}\right| < \infty$ for all j and k, we have that

$$\frac{1}{T} \sum_{t=1}^{T} X_t \to_p 0.$$

Given $\sum_{i=0}^{\infty} i \left|\Phi_i^{jk}\right| < \infty$ for all j and k, if $\{\varepsilon_t\}$ is an iid sequence, or a martingale difference sequence

satisfying some Lindeberg or Lyapunov conditions (see Section 4.8), we have that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t \to_d \mathbb{N}(0, \Phi(1)\Sigma\Phi(1)').$$

If

$$Y_t = \sum_{i=-\infty}^{\infty} \Psi_i X_{t-i},$$

we say that $\{Y_t\}$ is obtained by applying the linear filter $\Psi(L) = \sum_{i=-\infty}^{\infty} \Psi_i L^i$ to $\{X_t\}$. If $\Phi(L)$ and $\Psi(L)$ are two linear filters with absolutely summable coefficients, then $\Phi(L)\Psi(L)X_t$ is well defined for all weakly stationary $\{X_t\}$, as long as the dimensions of Φ_i, Ψ_i and X_t are appropriate for matrix multiplication. Also, the filters, viewed as lag operator polynomials with matrix-valued coefficients, when applied to weakly stationary time series, satisfies the usual algebraic rules of regular polynomials with matrix-valued coefficients. However, we note here that $\Phi(L)\Psi(L)$ does not necessarily equals $\Psi(L)\Phi(L)$ since the matrix multiplication operation is not commutative.

Now consider a linear filter $\Phi(L) = \sum_{i=0}^{\infty} \Phi_i L^i$ such that $\det(\Phi(z)) \neq 0$ for all $|z| \leq 1$ on the complex plane. Then there exists $\epsilon > 0$ such that $\Phi(z)$ is invertible for all $|z| < 1 + \epsilon$ and $\Phi^{-1}(z)$ has a power series expansion

$$\Phi^{-1}(z) = \sum_{i=0}^{\infty} \Upsilon_i z^i.$$

Then by a completely analogous argument as in the scalar case, we have that $\sum_{i=0}^{\infty} i \left|\Upsilon_i^{jk}\right| < \infty$ for all j and k, and we may define $\Upsilon(L) = \sum_{i=0}^{\infty} \Upsilon_i L^i$ as the inverse of $\Phi(L)$, denoted by $\Phi^{-1}(L)$.

The autocovariance generating function for a weakly stationary vector processes $\{X_t\}$ is defined as

$$G(z) = \sum_{k=-\infty}^{\infty} \gamma(k) z^k$$

where $\gamma(k)$ is the autocovariance function of $\{X_t\}$.

It can be shown that if $\{X_t\}$ is weakly stationary with autocovariance generating function $\Gamma_X(z)$, and $Y = \Phi(L)X_t$ where $\Phi(L)$ has absolutely summable coefficients, then the autocovariance generating function of $\{Y_t\}$ is given by

$$G_Y(z) = \Phi(z)G_X(z)\Phi(z^{-1})',$$

where $\Phi(z^{-1})'$ denotes the transpose of $\Phi(z^{-1})$. As a consequence, it is easy to see that the linear process $Y_t = \Phi(L)\varepsilon_t$ where $\varepsilon_t \sim \text{WN}(0, \Sigma)$ has autocovariance generating function $\Phi(z)\Sigma\Phi(z^{-1})'$.

Also, if $\{X_t\}$ and $\{Y_t\}$ are independent weakly stationary processes with autocovariance generating functions $\Gamma_X(z)$ and $\Gamma_Y(z)$ respectively, then the process $\{X_t + Y_t\}$ has autocovariance generating function $G_X(z) + G_Y(z)$.

The spectral density of a weakly stationary vector process with autocovariance function $\gamma(\cdot)$

and autocovariance generating function $G(z)$ is defined to be

$$f(\lambda) = \frac{1}{2\pi} G(e^{-i\lambda}).$$

And similarly as in the scalar case, we have that

$$\gamma(k) = \int_{-\pi}^{\pi} e^{i\lambda k} f(\lambda) \mathrm{d}\lambda.$$

7.2 Vector Moving Average Processes

A *vector moving-average process* of order q, abbreviated as VMA(q), is a process $\{X_t\}$ following

$$X_t = \mu + \Theta(L)\varepsilon_t = \mu + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \cdots + \Theta_q \varepsilon_{t-q},$$

where μ is a constant vector, and $\varepsilon_t \sim \mathrm{WN}(0, \Sigma)$. This process is always weakly stationary with mean μ and autocovariance function

$$\gamma(k) = \begin{cases} \sum_{i=0}^{q-k} \Theta_{i+k} \Sigma \Theta_i', & 0 \le k \le q, \\ \sum_{i=0}^{q+k} \Theta_i \Sigma \Theta_{i-k}', & q \le k < 0, \\ 0, & |k| > q \end{cases}$$

where Θ_0 is understood to be the identity matrix. In particular,

$$\mathrm{Cov}(X_t) = \gamma(0) = \sum_{i=0}^{q} \Theta_i \Sigma \Theta_i'.$$

Also, we have that

$$\frac{1}{T} \sum_{t=1}^{T} X_t \to_p \mu,$$

and under appropriate conditions for ε_t, we have that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (X_t - \mu) \to_d \mathbb{N}\Big(0, \Theta(1)\Sigma\Theta(1)'\Big).$$

A VMA model could be estimated by MLE or CMLE. Suppose $\varepsilon_t \sim$ iid $\mathbb{N}(0, \Sigma)$, and we observe $X_1, \ldots, X_T$. Then we may write our model as $X^\circ = \mu^\circ + A\varepsilon^\circ$ where $X^\circ = (X_1', X_2', \ldots, X_T')'$, $\mu^\circ = (\mu', \mu', \ldots, \mu')'$, $\varepsilon^\circ = (\varepsilon_{1-q}', \varepsilon_{2-q}', \ldots, \varepsilon_T')'$, and A is an appropriate matrix whose entries contain the parameters of the VMA process. Then we have $X^\circ =_d \mathbb{N}(\mu^\circ, A(I \otimes \Sigma)A')$, where I is the $(T+q)$-dimensional identity matrix. The log likelihood function is

$$C - \frac{1}{2} \ln \det(A(I \otimes \Sigma)A') - \frac{1}{2}(X^\circ - \mu^\circ)'(A(I \otimes \Sigma)A')^{-1}(X^\circ - \mu^\circ)$$

where C is some constant independent of the parameters.

Similarly, the (normalized) conditional log-likelihood function given the initial innovations $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{1-q}$ is

$$\frac{C}{T} - \frac{1}{2}\ln\det(\Sigma) - \frac{1}{2T}\sum_{t=1}^{T}\varepsilon_t'\Sigma\varepsilon_t.$$

When the initial values are not observed, we may set them to zero as an approximation. This approximation is sensible only if the VMA process is invertible, which is the case if $\det(\Theta(z)) \neq 0$ for all $|z| \leq 1$ where $\Theta(z) = I + \Theta_1 z + \ldots + \Theta_q z^q$.

7.3 Vector Autoregressive Processes

A *vector autoregressive process* of order p, abbreviated as VAR(p), is a process $\{X_t\}$ following

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \varepsilon_t, \tag{7.1}$$

where c is a constant vector, and $\varepsilon_t \sim \mathrm{WN}(0, \Sigma)$.

The following theorem is a direct generalization of its univariate counterpart.

Theorem 7.1. *Let $\{X_t\}$ be a VAR(p) process given by $\Phi(L)X_t = c + \varepsilon_t$ where $\Phi(L) = 1 - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p$ and $\varepsilon_t \sim$ WN$(0, \sigma^2)$. If $\det\Phi(z) \neq 0$ for all $|z| \leq 1$, then $\{X_t\}$ has a unique causal weakly stationary solution given by*

$$X_t = \Phi(L)^{-1}(c + \varepsilon_t) = \Phi^{-1}(1)c + \sum_{i=0}^{\infty}\Upsilon_i\varepsilon_{t-i}$$

where $\sum_{i=0}^{\infty}\Upsilon_i z^i$ is the power series expansion of $\Phi^{-1}(z)$, and $\sum_{i=0}^{\infty} i\,|\Upsilon_i| < \infty$.

Then mean μ of the VAR process is $\Phi^{-1}(1)c$, and the VAR process could be expressed in its demeaned form:

$$X_t - \mu = \Phi_1(X_{t-1} - \mu) + \Phi_2(X_{t-2} - \mu) + \cdots + \Phi_p(X_{t-p} - \mu) + \varepsilon_t.$$

The autocovariance function $\gamma(\cdot)$ could be obtained from the Yule-Walker equations

$$\gamma(k) = \Phi_1\gamma(k-1) + \Phi_2\gamma(k-2) + \cdots + \Phi_p\gamma(k-p), \quad k = 1, 2, \ldots.$$

The first few autocovariances could be obtained using the VAR(1) representation of general VAR(p) given by

$$\begin{bmatrix} X_t \\ X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-p+1} \end{bmatrix} = \begin{bmatrix} c \\ c \\ c \\ \vdots \\ c \end{bmatrix} + \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \cdots & \Phi_p \\ I & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \\ X_{t-3} \\ \vdots \\ X_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

If we write the above representation as $Y_t = a + AY_{t-1} + U_t$, and let $\Gamma = \mathrm{Var}(Y_t)$, $\Sigma_U = \mathrm{Var}(U_t) = \mathrm{diag}(\Sigma, O)$, then we have $\Gamma = A\Gamma A' + \Sigma_U$, and therefore $\mathrm{vec}(\Gamma) = (I - A \otimes A)^{-1}\mathrm{vec}(\Sigma_U)$. Note that Γ contains $\gamma(0), \ldots, \gamma(p-1)$. Once we obtain the initial values, we can obtain other autocovariances by recursion using the Yule-Walker equation.

Also, we have that

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to_p \mu,$$

and under appropriate conditions for ε_t, we have that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(X_t - \mu) \to_d \mathbb{N}\Big(0, \Phi^{-1}(1)\Sigma\Phi^{-1}(1)'\Big).$$

The VAR models may be estimated by conditional maximum likelihood. Suppose that $\varepsilon_t \sim$ iid $\mathbb{N}(0, \Sigma)$. We rewrite our VAR(p) model as

$$X_t = \Pi Y_t + \varepsilon_t$$

where $\Pi = [c\ \Phi_1\ \Phi_2\ \cdots \Phi_p]$ and $Y_t = (1, X'_{t-1}, X'_{t-2}, \ldots, X'_{t-p})'$. The conditional density $f(X_t|X_{t-1}, X_{t-2}, \ldots)$ is given by

$$(2\pi)^{-n/2}\big(\det(\Sigma)\big)^{-1/2}\exp\left(-\frac{1}{2}(X_t - \Pi Y_t)'\Sigma^{-1}(X_t - \Pi Y_t)\right).$$

The (normalized) log-likelihood function is

$$\frac{C}{T} - \frac{1}{2}\ln\det(\Sigma) - \frac{1}{2T}\sum_{t=1}^{T}(X_t - \Pi Y_t)'\Sigma^{-1}(X_t - \Pi Y_t).$$

Maximizing the log-likelihood function yields the CMLE estimators

$$\hat{\Pi} = \left(\sum_{t=1}^{T} X_t Y_t'\right)\left(\sum_{t=1}^{T} Y_t Y_t'\right)^{-1}$$

and

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'$$

where $\hat{\varepsilon}_t = X_t - \hat{\Pi}Y_t$ is the residual. It is easy to see that the CMLE estimator is the same as the OLS estimator. We may write our n-dimensional VAR system as n single equations, and do OLS regression for each of the individual equations.

Under the iid normality assumption we have that

$$Q_T = \frac{1}{T}\sum_{t=1}^{T} Y_t Y_t' \to_p \begin{bmatrix} 1 & \mu & \mu & \mu & \cdots & \mu \\ \mu & \gamma(0)+\mu^2 & \gamma(1)+\mu^2 & \gamma(2)+\mu^2 & \cdots & \gamma(p-1)+\mu^2 \\ \mu & \gamma(-1)+\mu^2 & \gamma(0)+\mu^2 & \gamma(1)+\mu^2 & \cdots & \gamma(p-2)+\mu^2 \\ \mu & \gamma(-2)+\mu^2 & \gamma(-1)+\mu^2 & \gamma(0)+\mu^2 & \cdots & \gamma(p-3)+\mu^2 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \mu & \gamma(1-p)+\mu^2 & \gamma(2-p)+\mu^2 & \gamma(3-p)+\mu^2 & \cdots & \gamma(0)+\mu^2 \end{bmatrix}.$$

We denote the limit by Q. Now we write

$$\sqrt{T}(\text{vech}(\hat{\Pi}) - \text{vech}(\Pi)) = \begin{bmatrix} Q_T^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t\varepsilon_{t1} \\ Q_T^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t\varepsilon_{t2} \\ \vdots \\ Q_T^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t\varepsilon_{tn} \end{bmatrix} = (I_n \otimes Q_T^{-1})\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t$$

where $\text{vech}(A) = (A_{11}, A_{12}, \ldots, A_{1\ell}, A_{21}, A_{22}, \ldots, A_{2\ell}, \ldots, A_{m1}, A_{m2}, \ldots, A_{m\ell})'$ for any $m \times \ell$ matrix, ε_{ti} is the i-th component of ε_t, and $Z_t = (Y_t'\varepsilon_{t1}, Y_t'\varepsilon_{t2}, \ldots, Y_t'\varepsilon_{tn})'$. It can be shown that $\{Z_t\}$ is a martingale difference sequence, has finite fourth moment (and therefore satisfies the Lyapunov condition), and a typical element of $\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'$ takes the form of

$$\frac{1}{T}\sum_{t=1}^{T} X_{i_1,t-k_1}\varepsilon_{j_1,t}X_{i_2,t-k_2}\varepsilon_{j_2,t} = \frac{1}{T}\sum_{t=1}^{T} W_t + \mathbb{E}(\varepsilon_{j_1,t}\varepsilon_{j_2,t})\frac{1}{T}\sum_{t=1}^{T} X_{i_1,t-k_1}X_{i_2,t-k_2},$$

where

$$W_t = \big(\varepsilon_{j_1,t}\varepsilon_{j_2,t} - \mathbb{E}(\varepsilon_{j_1,t}\varepsilon_{j_2,t})\big)X_{i_1,t-k_1}X_{i_2,t-k_2}$$

is a martingale difference sequence with finite second moments. Therefore, by a law of large numbers for martingale difference sequence, $\frac{1}{T}\sum_{t=1}^{T} W_t \to_p 0$, and

$$\frac{1}{T}\sum_{t=1}^{T} X_{i_1,t-k_1}\varepsilon_{j_1,t}X_{i_2,t-k_2}\varepsilon_{j_2,t} \to_p \mathbb{E}(\varepsilon_{j_1,t}\varepsilon_{j_2,t})\mathbb{E}(X_{i_1,t-k_1}X_{i_2,t-k_2}).$$

As a consequence,

$$\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t' \to_p \mathbb{E}Z_t Z_t' = \Sigma \otimes Q.$$

Then by a central limit theorem for martingale difference sequence, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \to_d \mathbb{N}(0, \Sigma \otimes Q),$$

and

$$\sqrt{T}(\text{vech}(\hat{\Pi}) - \text{vech}(\Pi)) = \mathbb{N}(0, \Sigma \otimes Q^{-1}).$$

The asymptotic results could be derived under weaker conditions. For example, for independent u_t with bounded fourth moments (see Lütkepohl (2005, Lemma 3.1)), or martingale difference sequence u_t with bounded conditional $2+\delta$ moments (see Fuller (1996, Theorem 8.2.3)). The essence is that we may find weaker conditions to guarantee that $\frac{1}{T}\sum Y_t Y_t' \to_p Q$ and $\frac{1}{\sqrt{T}}\sum \varepsilon_t Y_t' \to_d \mathbb{MN}(0, \Sigma, Q)$ where $\mathbb{MN}$ is the matrix normal distribution. Note that if $\xi \sim \mathbb{MN}(M, U, V)$, then $\text{vec}(\xi) \sim \mathbb{N}(\text{vec}(M), V \otimes U)$, and for non-random matrix C, D with appropriate dimensions, $D\xi C \sim \mathbb{MN}(DMC, DUD', CVC')$.

If we would like to test for the hypothesis that $R \cdot \text{vech}(\Pi) = r$, we may apply the Wald test statistic

$$W = T\left(R \cdot \text{vech}(\hat{\Pi}) - r\right)\left(R(\Sigma \otimes Q^{-1})R'\right)^{-1}\left(R \cdot \text{vech}(\hat{\Pi}) - r\right)'.$$

Under the null,

$$W \to_d \chi^2_m$$

where m is the number of restrictions tested. Note that under the null,

$$\sqrt{T}(R \cdot \text{vech}(\hat{\Pi}) - r) \to_d \mathbb{N}(0, R(\Sigma \otimes Q^{-1})R').$$

In particular, if we want to test the null hypothesis that $\Pi = \Pi_0$, we may form the Wald test statistic

$$\begin{aligned} W &= T\left(\text{vech}(\hat{\Pi} - \Pi_0)\right)'\left(\Sigma \otimes Q^{-1}\right)^{-1}\left(\text{vech}(\hat{\Pi} - \Pi_0)\right) \\ &= T\,\text{tr}\left[\left(\hat{\Pi} - \Pi_0\right)'\Sigma^{-1}\left(\hat{\Pi} - \Pi_0\right)Q\right]. \end{aligned}$$

Note that in the above expressions, Σ and Q can be replace respectively by $\frac{1}{T}\sum_{t=1}^T \hat{\varepsilon}_t^2$ and $\frac{1}{T}\sum_{t=1}^T Y_t Y_t'$.

Also, since

$$\sqrt{T}(\hat{\Sigma} - \Sigma) = \sqrt{T}\left(\hat{\Sigma} - \frac{1}{T}\sum_{t=1}^T \varepsilon_t\varepsilon_t' + \frac{1}{T}\sum_{t=1}^T \varepsilon_t\varepsilon_t' - \Sigma\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^T(\varepsilon_t\varepsilon_t' - \Sigma) + o_p(1),$$

and $\{\varepsilon_t\varepsilon_t'\}$ is iid, we have that

$$\sqrt{T}(\text{vech}(\hat{\Sigma}) - \text{vech}(\Sigma)) \to_d \mathbb{N}(0, \Xi)$$

where $\Xi = \mathbb{E}[\text{vech}(\varepsilon_t\varepsilon_t' - \Sigma)][\text{vech}(\varepsilon_t\varepsilon_t' - \Sigma)]'$.

In fact, since

$$\begin{bmatrix} \text{vech}(\hat{\Pi}) - \text{vech}(\Pi) \\ \text{vech}(\hat{\Sigma}) - \text{vech}(\Sigma) \end{bmatrix}$$

is a martingale difference sequence satisfying the Lyapunov condition and a law of large number for its second moment, we have that

$$\sqrt{T}\begin{bmatrix}\text{vech}(\hat{\Pi}) - \text{vech}(\Pi) \\ \text{vech}(\hat{\Sigma}) - \text{vech}(\Sigma)\end{bmatrix} \to_d \mathbb{N}\left(0, \begin{bmatrix}\Sigma \otimes Q & 0 \\ 0 & \Xi\end{bmatrix}\right).$$

Note that $\text{vech}(\hat{\Pi}) - \text{vech}(\Pi)$ and $\text{vech}(\hat{\Sigma}) - \text{vech}(\Sigma)$ are asymptotically orthogonal since

$$\mathbb{E}Y_t\varepsilon_{t\ell}(\varepsilon_{ti}\varepsilon_{tj} - \Sigma_{ij}) = 0$$

for all i, j, ℓ. In particular, we can show that $\mathbb{E}\varepsilon_{t\ell}\varepsilon_{ti}\varepsilon_{tj} = 0$ for a mean-zero jointly normal ε_t by sequential conditioning.

In the end of this section, we make a remark that since any VAR(p) model has a VAR(1) representation, when we study properties of VAR processes, we may focus on the VAR(1) model, and many results for VAR(p) models can be easily derived from the study of their VAR(1) representation.

7.4 Forecasting

The forecasting theory for multivariate linear processes is completely analogous to that of the univariate case. We therefore shall not elaborate on it here.

7.5 Granger Causality

A concept that is directly related to forecasting in multivariate setting is the *Granger-causality*. Granger-causality is a concept that tries to describe whether one economic variable helps to forecast another. Let $\{x_t\}$ and $\{y_t\}$ be two (possibly multivariate) time series. Let $\mathcal{F}_t$ be the information available for prediction up to time t. Let $\ell(x_{t+h}|\mathcal{G})$ be some measure of forecast imprecision for the optimal h-step forecast of x based on the information set $\mathcal{G}$. For example, we could take ℓ to be the mean square error, as we shall do in the following. If

$$\ell(x_{t+h}|\mathcal{F}_t) < \ell(x_{t+h}|\mathcal{F}_t \backslash \{y_s\}_{s\leq t})$$

for at least one $h = 1, 2, \ldots$, we say that y causes x in Granger's sense, or y Granger-causes x. If both x Granger-causes y and y Granger-causes x, we say that $(x', y')'$ is a feedback system.

Though different measures of forecast imprecision lead to different definitions of Granger causality, we usually take ℓ to be the mean square error and replace the optimal forecast with the best linear forecast. Also, we usually restrict our information set to be the information generated by current and past values of x and y. In this setting, we say that y Granger-cause x if for at least one $h > 0$,

$$\text{MSE}[\mathbb{L}(x_{t+h}|1, x_{t-1}, x_{t-2}, \ldots)] > \text{MSE}[\mathbb{L}(x_{t+h}|1, x_{t-1}, x_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots)].$$

Apparently if $(x_t', y_t')'$ is a system that admits an MA representation

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \mu + \sum_{i=1}^{\infty} \begin{bmatrix} \Theta_{i,11} & \Theta_{i,12} \\ \Theta_{i,21} & \Theta_{i,22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-i} \\ \varepsilon_{2,t-i} \end{bmatrix},$$

where $(\varepsilon_{1t}', \varepsilon_{2t}')'$ is a white noise process with non-singular variance matrix, then y does not Granger-cause x if and only if $\Theta_{i,12} = 0$ for all $i = 1, 2, \ldots$. If $(x_t, y_t)'$ admits a VAR(p) representation

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = c + \sum_{i=1}^{p} \begin{bmatrix} \Phi_{i,11} & \Phi_{i,12} \\ \Phi_{i,21} & \Phi_{i,22} \end{bmatrix} \begin{bmatrix} x_{t-i} \\ y_{t-i} \end{bmatrix} + \varepsilon_t, \tag{7.2}$$

then y does not Granger-cause x if and only if $\Phi_{i,12} = 0$ for all $i = 1, 2, \ldots, p$. To test for Granger-causality in this case, we may run a VAR of (7.2) and test for $\Phi_{1,12} = \Phi_{2,12} = \cdots = \Phi_{p,12} = 0$. Wald test could be used.

For more concepts and tests related to Granger-causality, see Lütkepohl (2005, Sec. 2.3.1 and 3.6). It should be pointed out that Granger-causality, first proposed by Granger (1969), is not an indicator of the causality in the usual sense but an indicator of predictability. Also, the result of Granger-causality test could be very sensitive to the choice of the lag p. See Hamilton (1994, p. 305) for details. See Dufour and Renault (1998) and Dufour et al. (2006) for discussions on the issue of Granger causality when there are multiple sets of variables of interest.

7.6 Structural VAR

A *structural* VAR model for an n-dimensional vector process $\{X_t\}$ is a model of the form

$$BX_t = b + B_1 X_{t-1} + B_2 X_{t-2} + \cdots + B_p X_{t-p} + e_t$$

where $\mathrm{Var}(e_t) = \Lambda$ is a *diagonal* matrix. The corresponding VAR model in *reduced form*, given by

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \varepsilon_t$$

where $\mathrm{Var}(\varepsilon_t) = \Sigma$, relates to its *structural form* through

$$b = Bc,$$

$$B_i = B\Phi_i,$$

and

$$e_t = B\varepsilon_t.$$

Due to the existence of B, structural VAR allows for contemporaneous relationships in the components of X_t, while reduced form VAR does not. Also, the diagonality of Λ restricts components in e_t to be uncorrelated, while in the reduced form VAR, components in ε_t can be correlated. In

particular, when e_t is Gaussian, components in e_t are independent. We shall call e_t the *structural innovations*, and ε_t the *reduced form errors*.

If we allow B to be any $n \times n$ matrix, then B is not *identified*, meaning that we have more than one structural VAR that gives the same reduced form VAR. For example, if we pre-multiply both sides of the structural VAR by any $n \times n$ matrix C such that its column vectors are mutually orthogonal, we still get a structural VAR. However, as long as $C \neq I$, the two structural VARs are not the same, although they have the same reduced form VAR.

Suppose we have identified a reduced form VAR. Then if we can uniquely identify B, we can identify b and B_i's by the relationships we observed above. To identify a structural VAR model, we need to put enough restrictions on the contemporaneous relationship matrix B. Since

$$B\Sigma B' = \Lambda,$$

the diagonality of Λ generates $r(r-1)/2$ restrictions on B. So we need to impose additional $r(r+1)/2$ restrictions on B for identification. One of the most simple and popular ways to restrict B is to specify B to be a lower triangular matrix such that its diagonal contains all ones. These zero-one restrictions then serve as the additional $r(r+1)/2$ restrictions that makes the structural VAR identified. That B is lower triangular and $\varepsilon_t = Be_t$ implies that the first component in ε_t, denoted by ε_{t1}, is more "causal" than any other ε_{ti}'s. And ε_{ti} provides information about ε_{tj} for any $j > i$. This then implies a *recursive causal chain* of the innovations we consider in our model. Or put it in another way, when we formulate X_t, we should order the variables under consideration according to the causal chain we have in mind. Structural VARs identified through the above scheme are called *recursive* structural VARs.

For a covariance matrix Σ, we can decompose it as

$$\Sigma = LL'$$

where L is a unique lower triangular matrix. Such a decomposition is called the Cholesky decomposition. Also, we have $L^{-1}\Sigma L^{-1'} = I$. This implies that

$$B = \Lambda^{1/2} L^{-1}.$$

Therefore, to estimate a recursive structural VAR model, we may first estimate its reduced form by OLS, obtain its MLE estimator for Σ and conduct the Cholesky decomposition on the estimated Σ. In the end, we solve the above equality for the undetermined parameters in B and Λ. Since the model is just identified, estimators obtained in this way is exactly the full-information maximum likelihood estimator (FIML).

There are many other identification strategies. Interested readers may referred to, e.g., Hamilton (1994, Section 11.6) or Stock and Watson (2016, Chapter 4), among many others.

As an example, Sims (1980) consider a system of six variables for an empirical macroeconomic

model: money, real GNP, unemployment, wages, price level and import prices. To see the responses of the system to random shocks, the author utilize the recursive identification strategy where the variables are ordered as above. That is, money shocks are assumed to affect all other variables of the system instantly, while the import price shocks only affect the import prices.

7.7 Impulse Responses and Variance Decomposition

We may write

$$X_t = \mu + \sum_{i=0}^{\infty} \Upsilon_i \varepsilon_{t-i} = \mu + \sum_{i=0}^{\infty} A_i e_t^*$$

where e_t^* is e_t normalized to have identity variance,

$$A_i = \Upsilon_i B^{-1} \Lambda^{1/2}.$$

Then the (p,q)-th entry of A_i, denoted by $A_{i,pq}$, can be interpreted as the response of the p-th variable i periods later to an impulse in the q-th structural innovations.

The k-step (linear) forecast error may be written as

$$X_{t+k} - \hat{X}_{t+k|t} = \sum_{i=0}^{k-1} \Upsilon_i \varepsilon_{t+k-i} = \sum_{i=0}^{k-1} A_i e_{t+k-i}^*.$$

The variance of the forecast error for the p-th variable is

$$\sum_{i=0}^{k-1} \left(\sum_{j=1}^{r} A_{i,pj}^2 \right).$$

The ratio

$$\frac{\sum_{i=1}^{k-1} A_{i,pq}^2}{\sum_{i=0}^{k-1} \left(\sum_{j=1}^{r} A_{i,pj}^2 \right)}$$

then gives the contribution of the q-th structural innovation to the forecast error variance of the p-th variable.

Once we establish the asymptotic normality of the coefficient estimator of the VAR process, we may use the Delta method to establish the asymptotic normality and obtain the (pointwise) confidence interval (bands) of the impulse responses and the variance decomposition ratios. See Lütkepohl (1990) for details. Runkle (1987) suggests using simulation or bootstrap methods to obtain the confidence intervals. See also Kilian (1998, 1999), Sims and Zha (1999), Gonçalves and Kilian (2004) and Inoue and Kilian (2020).

7.8 Order Selection for VAR Models

We may use the information criteria to select the order of VAR models. We consider three criteria. The Akaike Information Criterion (AIC) for an VAR(p) model is defined as

$$\text{AIC}(p) = \ln\det(\hat{\Sigma}) + 2\frac{pn^2}{T}$$

where $\hat{\Sigma}$ is the ML estimate of Σ, n is the dimension of the random vector X_t, and T is the sample size. Note that pn^2 is the number of free parameters to estimate in the model (if we ignore the constant term). The Bayesian Information Criteria, or the Schwarz Criterion, for an VAR(p) model is defined as

$$\text{BIC}(p) = \ln\det(\hat{\Sigma}) + \frac{pn^2 \ln T}{T}.$$

The Hannan-Quinn Criterion for an VAR(p) model is defined as

$$\text{HQ}(p) = \ln\det(\hat{\Sigma}) + 2\frac{pn^2 \ln\ln T}{T}.$$

We choose the order p so that either AIC or BIC or HQ is minimized. It is know that AIC tends to choose an order that minimized the forecast mean square error, while BIC or HQ tends to choose an order that is consistent. For details, see, e.g., Lütkepohl (2005, Chapter 4).

7.9 Bayesian VAR

A very popular way to estimate VAR models is through Bayesian approaches. We shall not go into details here but refer readers to references, e.g., Hamilton (1994, Chapter 12) or Lütkepohl (2005, Sec. 5.4).

7.10 Vector Autoregressive Moving-Average Model

Just as in the univariate case, we have a vector version of autoregressive moving-average model, abbreviated as VARMA. The theory of VARMA models is more involved, and we shall refer the readers to, e.g., Lütkepohl (2005, Part IV).

7.11 Some Results of Matrix Algebra

Suppose that A, B, C, D are matrices. Suppose that in each of the following entries the dimensions of the matrices are appropriate such that the operations of matrices are well defined. We have

1. $(A \otimes B)' = A' \otimes B'$.
2. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
3. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.
4. $\text{vech}(A) = \text{vec}(A')$.
5. $(A \otimes B)\text{vech}(C) = \text{vech}(ACB')$.

6. $\text{tr}(ABC) = \text{tr}(CAB)$.
7. $\text{vech}(A)'\text{vech}(B) = \text{tr}(AB')$.
8. $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$.

8 State Space Models and the Kalman Filter

8.1 State Space Models

A typical *state space model* consists of a *measurement equation*

$$y_t = Aw_t + Bx_t + u_t$$

where y_t is a random vector of interest, x_t is a vector of exogenous or predetermined variables, w_t is a vector of possibly hidden states, and a *transition equation*

$$w_t = Tw_{t-1} + v_t$$

which specifies the law of motion of the state. The two error terms u_t and v_t are assumed to be serially and crossly uncorrelated with mean zero and variances R and Q, respectively. We assume that the initial value w_0 is uncorrelated with u_t and v_t, and that $\{u_t, v_t\}$ is jointly normal.

The state space models are very useful since many systems in economics involve unobserved variables, and many econometric models have state space model representation. For example, an ARMA(p, q) model specified by $\alpha(L)y_t = \theta(L)\varepsilon_t$ with $\alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p$ and $\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$ may be written in the state space representation

$$y_t = \theta(L)z_t,$$

$$\alpha(L)z_t = \varepsilon_t$$

where $z_t = \alpha^{-1}(L)\varepsilon_t$. z_t, as an $AR(p)$ process, has an VAR(1) representation (as in the transition equation above). As we shall see later in this chapter, we may obtain the exact likelihood of an ARMA process by its state space model representation.

8.2 The Kalman Filter

We define the *filtration* $\{\mathcal{F}_t\}$ (an increasing sequence of σ-algebras or more intuitively, "information sets") by

$$\mathcal{F}_t = \sigma\left(y_1, y_2, \ldots, y_t\right).$$

By exogeneity or predeterminacy of x_t, we mean that x_t is $\mathcal{F}_{t-1}$-measurable, or x_t is "known" given the information $\mathcal{F}_{t-1}$. We denote

$$w_{s|t} = \mathbb{E}(w_s|\mathcal{F}_t),$$

$$y_{s|t} = \mathbb{E}(y_s|\mathcal{F}_t),$$

$$\Omega_{s|t} = \mathrm{Var}(w_s|\mathcal{F}_t),$$

and

$$\Sigma_{s|t} = \mathrm{Var}(y_s|\mathcal{F}_t)$$

where $\mathbb{E}(\cdot|\mathcal{F}_t)$ is the conditional expectation given $\mathcal{F}_t$ and $\mathrm{Var}(\cdot|\mathcal{F}_t)$ is the conditional variance given $\mathcal{F}_t$. For a random variable X, its conditional variance given a σ-algebra $\mathcal{F}$ is defined by

$$\mathrm{Var}(X|\mathcal{F}) = \mathbb{E}\left[\big(X - \mathbb{E}(X|\mathcal{F})\big)^2 \middle| \mathcal{F}\right]$$

Since $y_t = A(T^t w_0 + T^{t-1}v_1 + \cdots + Tv_{t-1} + v_t) + Bx_t + u_t$, we have that $\mathbb{E}(y_s v_t') = \mathbb{E}(y_s u_t') = 0$ for all $s < t$. Under normality, this implies that $\{y_s\}_{s=1}^{t-1}$ is independent of v_t and u_t. Then we have that

$$\mathbb{E}(v_t|\mathcal{F}_{t-1}) = \mathbb{E}(u_t|\mathcal{F}_{t-1}) = 0.$$

Taking conditional expectations with respect to $\mathcal{F}_{t-1}$ on both sides of the measurement equation and the transition equation, we obtain

$$w_{t|t-1} = Tw_{t-1|t-1},$$

$$y_{t|t-1} = Aw_{t|t-1} + Bx_t.$$

It is also easy to calculate the conditional variances by definition and obtain

$$\Omega_{t|t-1} = T\Omega_{t-1|t-1}T' + Q$$

and

$$\Sigma_{t|t-1} = A\Omega_{t|t-1}A' + R.$$

The above step of getting the conditional distributions of w_t and y_t given information up to time $t-1$ is labeled *prediction*.

Write

$$y_t - y_{t|t-1} = A(w_t - w_{t|t-1}) + u_t.$$

The prediction step implies that

$$\begin{bmatrix} w_t \\ y_t \end{bmatrix} \middle| \mathcal{F}_{t-1} =^d \mathbb{N}\left(\begin{bmatrix} w_{t|t-1} \\ y_{t|t-1} \end{bmatrix}, \begin{bmatrix} \Omega_{t|t-1} & \Omega_{t|t-1}A' \\ A\Omega_{t|t-1} & \Sigma_{t|t-1} \end{bmatrix}.\right)$$

Since $w_{t|t} = \mathbb{E}(w_t|y_t, \mathcal{F}_{t-1})$ and $\Omega_{t|t} = \mathrm{Var}(w_t|y_t, \mathcal{F}_{t-1})$, by Theorem 5.10, we have

$$w_{t|t} = w_{t|t-1} + \Omega_{t|t-1}A'\Sigma_{t|t-1}^{-1}(y_t - y_{t|t-1})$$

and

$$\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}A'\Sigma_{t|t-1}^{-1}A\Omega_{t|t-1}.$$

The above procedure is labeled *updating* since we updated the conditional distributions of w_t given that we have a new observation y_t. We may see that the updating rule for $\{w_t\}$, given by $w_{w|t} - w_{t|t-1}$, is proportion to the forecast error of $\{y_t\}$:

$$w_{w|t} - w_{t|t-1} = \Omega_{t|t-1} A' \Sigma_{t|t-1}^{-1} (y_t - y_{t|t-1}).$$

We sometimes call the proportion $K_t = \Omega_{t|t-1} A' \Sigma_{t|t-1}^{-1}$ the *Kalman gain*. The Kalman gain can thus be interpreted as the weight assigned to the information that is newly available at time t.

8.3 Kalman Filter and Maximum Likelihood Estimation

We may obtain the maximum likelihood function of $(y_1, \ldots, y_T)$ as follows:

(a) Start from an initialization value $w_{0|0}, \Omega_{0|0}$. For given parameters (denoted by θ), set the step-zero log likelihood $\mathcal{L}_0(\theta) = 0$.

(b) Given $w_{t-1|t-1}, \Omega_{t-1|t-1}$, update to get $w_{t|t-1}, \Omega_{t|t-1}, y_{t|t-1}, \Sigma_{t|t-1}$. Then $y_t|\mathcal{F}_{t-1} =_d \mathbb{N}(y_{t|t-1}, \Sigma_{t|t-1})$. Then we may write down the log likelihood function $\ell_t(\theta)$ of $y_t|\mathcal{F}_{t-1}$ as

$$\ell_t(\theta) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln \det \Sigma_{t|t-1} - \frac{1}{2} (y_t - y_{t|t-1})' \Sigma_{t|t-1}^{-1} (y_t - y_{t|t-1})$$

and set $\mathcal{L}_t(\theta) = \mathcal{L}_{t-1}(\theta) + \ell_t(\theta)$.

(c) Update to obtain $w_{t|t}, \Omega_{t|t}$.

(d) Repeat (b) and (c) until we get $\mathcal{L}_T(\theta)$.

8.4 Smoothing

Often we are interested in estimating w_t given all the observations $(y_1, \ldots, y_T)$. To obtain $w_{t|T} = \mathbb{E}(w_t|\mathcal{F}_T)$, we note that since

$$y_{t+k} = A(T^{k-1} w_{t+1} + T^{k-2} v_{t+2} + \cdots + v_{t+k}) + B x_{t+k} + u_{t+k},$$

given w_{t+1} and $\mathcal{F}_t$, y_{t+k} is uncorrelated with, and therefore independent of, w_t, we have that

$$\mathbb{E}(w_t|w_{t+1}, \mathcal{F}_T) = \mathbb{E}(w_t|w_{t+1}, \mathcal{F}_t).$$

Since $w_{t+1} - w_{t+1|t} = T(w_t - w_{t|t}) + v_{t+1}$, we have $\mathrm{Cov}(w_t, w_{t+1}|\mathcal{F}_t) = \Omega_{t|t} T'$, and then

$$\mathbb{E}(w_t|w_{t+1}, \mathcal{F}_t) = w_{t|t} + J_t(w_{t+1} - w_{t+1|t})$$

where $J_t = \Omega_{t|t} T' \Omega_{t+1|t}^{-1}$. Note J_t is obtained by projection. Then

$$w_{t|T} = \mathbb{E}\left(\mathbb{E}\left(w_t|w_{t+1}, \mathcal{F}_T\right) \middle| \mathcal{F}_T \right) = w_{t|t} + J_t(w_{t+1|T} - w_{t+1|t}).$$

To obtain $\Omega_{t|T} = \mathrm{Var}(w_t|\mathcal{F}_T)$, write the above equation as $w_{t|T} - w_{t|t} = J_t(w_{t+1|T} - w_{t+1|t})$. Since $\mathbb{E}(w_{s|T}|\mathcal{F}_t) = w_{s|t}$, we have that

$$\mathbb{E}(w_{s|T} - w_{s|t})(w_{s|T} - w_{s|t})' = \mathbb{E}w_{s|T}w'_{s|T} - \mathbb{E}w_{s|t}w'_{s|t}.$$

Similarly, we may derive

$$\mathbb{E}w_{s|t}w'_{s|t} = \mathbb{E}w_s w'_s - \mathbb{E}(w_s - w_{s|t})(w_s - w_{s|t})' = \mathbb{E}w_s w'_s - \Omega_{s|t}.$$

Now it is easy to deduce that

$$\Omega_{t|T} = \Omega_{t|t} + J_t(\Omega_{t+1|T} - \Omega_{t+1|t})J'_t.$$

It should be noted that without normality, the Kalman filter does not provide the conditional mean and variance of the state variables. However, if we take $\mathcal{F}_t$ as the *linear* span of $(y_1, \cdots, y_t)$, and view $\mathbb{E}(\cdot|\mathcal{F}_t)$ and $\mathrm{Var}(\cdot|\mathcal{F}_t)$ respectively as the projection on $\mathcal{F}_t$ and the variance of the $\mathcal{F}_t$ projection error, then all previous derivation continue to hold, and the Kalman filter yields the minimum MSE linear estimate of the state variables.

8.5 Markov Chains

Let the underlying probability space be $(\Omega, \mathcal{F}, \mathbb{P})$. For our purposes it suffices to consider time-homogeneous Markov chains with finite state spaces in discrete-time.

Let $\{X_n\}$ be a sequence of random variables taking values in a finite set $\{a_1, a_2, \ldots, a_N\}$. Let $\{\mathcal{G}_n\}$ be an increasing sequence of σ-algebras (information sets) such that $\cdots \subset \mathcal{G}_t \subset \mathcal{G}_{t+1} \subset \cdots \subset \mathcal{F}$. We say that $\{(X_t, \mathcal{G}_t)\}$ is a *Markov chain* if

$$\mathbb{P}(X_t = a_j|\mathcal{G}_{t-1}) = \mathbb{P}(X_t = a_j|X_{t-1})$$

for all j and t. If in addition, $\mathbb{P}(X_t = a_i|X_{t-1})$ is independent of time t, we say that the Markov Chain is *time-homogeneous*. We usually take $\mathcal{G}_t = \sigma(X_1, X_2, \ldots, X_t)$, just as in the previous chapter. In that case, we say that $\{X_t\}$ *itself* is a Markov chain, or simply that $\{X_t\}$ is a Markov chain.

From now on, we assume that $\{X_t\}$ is time-homogeneous and denote

$$\mathbb{P}(X_t = a_j|X_{t-1} = a_i) = p_{ij}.$$

The *transition matrix* of the Markov chain is a matrix defined as

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN}. \end{bmatrix}$$

It is obvious that each row of P sums up to one.

8.6 Hamilton's Markov Switching Model

The Hamilton's *Markov switching model* is given by

$$y_t = \mu_{s_t} + w_t,$$

and

$$\alpha(L) w_t = \varepsilon_t$$

where y_t is the observed variable, $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$, $\{s_t\}$ denotes the state, which is assumed to be a Markov chain and independent of $\mathcal{F}_{t-1} = \sigma(y_1, y_2, \ldots, y_{t-1})$, and $\varepsilon_t \sim$ iid $\mathbb{N}(0, \sigma^2)$. Note that in this model, the mean of y_t is dependent on the state s_t.

The Hamilton's filter has two steps, i.e., prediction and updating. We first consider the case in which w_t is AR(1). For notation convenience, we denote by $p(X|\mathcal{G})$ the conditional density function of X given information set $\mathcal{G}$. To predict, we note

$$p(s_t, s_{t-1}|\mathcal{F}_{t-1}) = p(s_t|s_{t-1}) p(s_{t-1}|\mathcal{F}_{t-1})$$

and

$$p(y_t|\mathcal{F}_{t-1}) = \sum_{s_t, s_{t-1}} p(y_t|s_t, s_{t-1}, \mathcal{F}_{t-1}).$$

To update, note that

$$p(s_t, s_{t-1}|\mathcal{F}_t) = \frac{p(y_t|s_t, s_{t-1}, \mathcal{F}_{t-1}) p(s_t, s_{t-1}|\mathcal{F}_{t-1})}{p(y_t|\mathcal{F}_{t-1})}.$$

To start the algorithm, we set $p(s_0|\mathcal{F}_0)$ to be the unconditional probability of the Markov chain. Also note that $p(s_t|\mathcal{F}_t)$ can be obtained by marginalizing $p(s_t, s_{t-1}|\mathcal{F}_t)$.

For general AR(p) process of w_t, we may look at $p(s_t, \ldots, s_{t-p}|\mathcal{F}_{t-1})$ and $p(s_t, \ldots, s_{t-p}|\mathcal{F}_t)$ in place of $p(s_t, s_{t-1}|\mathcal{F}_{t-1})$ and $p(s_t, s_{t-1}|\mathcal{F}_t)$, respectively.

In the smoothing step, we obtain $p(s_t|\mathcal{F}_T)$. We look at the AR(1) case first. We start from $p(s_T, s_{T-1}|\mathcal{F}_T)$. We have that

$$p(s_{t+1}, s_t, s_{t-1}|\mathcal{F}_T) = p(s_{t-1}|s_{t+1}, s_t, \mathcal{F}_T) p(s_{t+1}, s_t|\mathcal{F}_T)$$

where

$$p(s_{t-1}|s_{t+1}, s_t, \mathcal{F}_T) = p(s_{t-1}|s_{t+1}, s_t, \mathcal{F}_t) = p(s_{t-1}|s_t, \mathcal{F}_t).$$

Note that the first equality is due to the AR(1) property that

$$p(y_{t+k}|s_{t+1}, s_t, s_{t-1}, \mathcal{F}_t) = p(y_{t+k}|s_{t+1}, s_t, \mathcal{F}_t)$$

for all $k > 0$. The second equality is due to that

$$p(s_{t+1}|s_t, s_{t-1}, \mathcal{F}_t) = p(s_{t+1}|s_t, \mathcal{F}_t)$$

and the Bayes formula. The term $p(s_{t-1}|s_t, \mathcal{F}_t)$ can be obtained by

$$p(s_{t-1}|s_t, \mathcal{F}_t) = \frac{p(s_t, s_{t-1}|\mathcal{F}_t)}{p(s_t|\mathcal{F}_t)}$$

using results in the prediction and updating step. In the end, $p(s_t, s_{t-1}|\mathcal{F}_T)$ and $p(s_{t-1}|\mathcal{F}_T)$ can be obtained by marginalization.

For general AR(p) process of w_t, we may use

$$p(s_{t+1}, \ldots, s_{t-p}|\mathcal{F}_T) = p(s_{t-p}|s_{t+1}, \ldots, s_{t-p+1}, \mathcal{F}_T)p(s_{t+1}, \ldots, s_{t-p+1}|\mathcal{F}_T).$$

Note that

$$\begin{aligned} p(s_{t-p}|s_{t+1}, \ldots, s_{t-p+1}, \mathcal{F}_T) &= p(s_{t-p}|s_{t+1}, \ldots, s_{t-p+1}, \mathcal{F}_t) \\ &= p(s_{t-p}|s_t, \ldots, s_{t-p+1}, \mathcal{F}_t) \\ &= \frac{p(s_t, \ldots, s_{t-p}, \mathcal{F}_t)}{p(s_t, \ldots, s_{t-p+1}, \mathcal{F}_t)}. \end{aligned}$$

9 Conditional Heteroskedasticity

Sometimes we are interested in not only the level of a time series, but also the second moment information of it. For example, volatility is the standard deviation of (log) returns of a financial asset. Volatility of financial assets is usually time varying, meaning that it is subject to large changes over time. Since volatility is a measure of risk, it is an important factor in asset pricing. For example, the well-known Black-Scholes formula for the price of a European call option is given by

$$c_t = P_t\Phi(x) - Kr^{-\ell}\Phi(x - \sigma_t\sqrt{\ell})$$

and

$$x = \frac{\ln(P_t/Kr^{-\ell})}{\sigma_t\sqrt{\ell}} + \frac{1}{2}\sigma_t\sqrt{\ell},$$

where P_t is the current price of the underlying stock corresponds to the option, r is the risk-free interest rate, ℓ is the time to expiration, K is the strike price, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and σ_t is the conditional standard deviation of the log return of the underlying stock. This chapter is devoted to models of conditional standard deviation.

9.1 The ARCH Model

Let $\{y_t\}$ be the time series of interest and $\{\mathcal{F}_t\}$ be a filtration representing the information flow. Let $\mu_t = \mathbb{E}(y_t|\mathcal{F}_{t-1})$ and $\sigma_t^2 = \mathrm{Var}(y_t|\mathcal{F}_{t-1})$ be the conditional mean and the conditional variance of y_t given $\mathcal{F}_{t-1}$, respectively.

Suppose $\{y_t\}$ follows an AR(p) process

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t,$$

where $u_t \sim \mathrm{WN}(0, \sigma^2)$. If all roots of the polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ lie outside the unit circle, the series is weakly stationary, and has constant unconditional mean and unconditional variance. If we let $\mathcal{F}_{t-1}$ be the σ-algebra generated by all innovations u_t (and therefore all y_t) prior to t, we have

$$\mu_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p},$$

i.e., the conditional mean is time varying. Similarly, the conditional variance of a weakly stationary time series can also be time varying. A popular model for time varying conditional variance is the *Autoregressive Conditional Heteroskedastic (ARCH) model* proposed by Engle (1982).

The ARCH model assumes that u_t is serially uncorrelated but dependent. To be specific, an ARCH(m) model assumes that

$$u_t = \sigma_t\varepsilon_t,$$

where $\{\varepsilon_t\}$ is a sequence of iid random variables with mean zero and variance one, and

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_m u_{t-m}^2$$

for some $\alpha_0 > 0$ and $\alpha_i \geq 0$. The coefficients α_i must satisfy some conditions so that the unconditional variance of u_t is finite. We usually take ε_t to be a normal or a Student's t random variable. It is easy to verify that $\sigma_t^2 = \text{Var}(u_t|\mathcal{F}_{t-1})$. Note that we may also represent the above ARCH process as

$$u_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_m u_{t-m}^2 + w_t \tag{9.1}$$

where w_t is a white noise process. Obviously, we have $\sigma^2 = \alpha_0/(1 - \alpha_1 - \cdots - \alpha_p)$. For the unconditional variance to exist, it requires that $\alpha_1 + \cdots + \alpha_p < 1$.

ARCH model is a simple and easy to use model. One good feature of the ARCH model is that even if ε_t is normal, the unconditional distribution of u_t generated from ARCH has a fatter tail than a normal distribution. This fat tail phenomenon appears frequently in finance. However, it does have some drawbacks. First, ARCH models assume that positive and negative shocks have the same effects on the variance of the time series since the shocks enter into volatility in the form of squares. Second, the ARCH models are restrictive in term of the possible ranges of the ARCH parameters. Third, ARCH models are likely to overpredict the volatility in finance (see Tsay (2010, p. 119)).

The test for the existence of ARCH effect is based on equation (9.1). Once we obtain u_t or estimated $\hat{u}_t$, we test for the null hypothesis that $\beta_i = 0, i = 1, 2, \ldots, m$ jointly in the regression

$$u_t^2 = \beta_0 + \beta_1 u_{t-1}^2 + \cdots + \beta_m u_{t-m}^2 + e_t.$$

A Wald test or an F-test may be applied.

9.2 Estimating ARCH Models

The ARCH models can be estimated by conditional maximum likelihood. Suppose that we have data up to time T. The conditional density function of $(u_{m+1}, u_{m+2}, \ldots, u_T)$ given $(u_1, u_2, \cdots, u_m)$ is

$$f(u_{m+1}, \ldots, u_T|u_1, \ldots, u_m) = f(u_T|\mathcal{F}_{t-1}) f(u_{T-1}|\mathcal{F}_{t-2}) \cdots f(u_{m+1}|\mathcal{F}_m).$$

Under normality of ε_t, we have

$$f(u_{m+1}, \ldots, u_T|u_1, \ldots, u_m) = \prod_{t=m+1}^{T} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{u_t^2}{2\sigma_t^2}\right),$$

and the log likelihood function (ignoring the constant part) is given by

$$\ell(\varepsilon_{m+1}, \varepsilon_{m+2}, \ldots, \varepsilon_T | \alpha_0, \alpha_1, \ldots, \alpha_m) = -\frac{1}{2} \sum_{t=m+1}^{T} \ln \sigma_t^2 - \frac{1}{2} \sum_{t=m+1}^{T} \frac{u_t^2}{\sigma_t^2},$$

where σ_t^2 is a function of $\alpha_0, \alpha_1, \ldots, \alpha_m$.

When u_t is not directly observable but comes from a regression

$$y_t = x_t' \beta + u_t,$$

we may replace u_t with $y_t - x_t'\beta$ in the above expression of likelihood and estimate β and the α's jointly. We may also assume that the conditional distributions are Student's t distributions or the generalized error distribution (GED). The likelihood functions could be obtained similarly by sequential conditioning. Under some conditions, the quasi-maximum likelihood estimation also generates consistent estimators. See Hamilton (1994, p. 663).

The ARCH order m could be determined using the PACF of u_t^2.

9.3 The GARCH Models

There are many variants to the ARCH models. One of the most popular class of variants to ARCH is the generalized ARCH model by Bollerslev (1986). If $u_t = y_t - \mu_t$ follows a GARCH(m, s) model, then

$$u_t = \sigma_t \varepsilon_t,$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_s \sigma_{t-s}^2,$$

where ε_t is a sequence of iid random variables with mean zero and variance one, $\alpha > 0, \alpha_i \geq 0, \beta_j \geq 0$, and $\sum_{i=1}^{m \vee s} (\alpha_i + \beta_i) < 1$ (α_i or β_i are set to zero if they do not exist). The unconditional variance is $\mathbb{E} u_t^2 = \alpha_0 / (\sum_{i=1}^{m \vee s} (\alpha_i + \beta_i))$. Note that GARCH equation could be written as

$$u_t^2 = \alpha_0 + (\alpha_1 + \beta_1) u_{t-1}^2 + \cdots + (\alpha_p + \beta_p) u_{t-p}^2 + w_t - \beta_1 w_{t-1} - \cdots - \beta_s w_{t-s}$$

where $p = m \vee s$ and $w_t = u_t^2 - \sigma_t^2$ is a white noise process. This implies that $\{u_t^2\}$ is an ARMA$(m \vee s, s)$ process.

We may continue to use the maximum likelihood estimation to estimate GARCH models, providing that the starting values of the volatility are assumed to be known. For issues regarding the choice of initialization values, one may refer to, e.g., Bollerslev (1986).

In both ARCH and GARCH models, when the conditional mean function is not known, we may first estimate the conditional mean function, treating the ARCH/GARCH effect as non-existent. Then we use the fitted residuals as an observed series and use their ARMA representations to estimate the ARCH/GARCH parameters. Although the properties of such estimators are complex and not clearly known, in practice it turns out to provide good approximations. See Tsay (2010,

p. 140).

9.4 The GARCH-M Model

Since volatility is a measure of risk, it is usually priced in finance. Financial returns are therefore dependent on the volatility of their underlying assets. One model for financial returns that incorporates volatility as a price factor is the GARCH-M model, or the GARCH-in-mean model. A version of the model was first proposed by Engle et al. (1987). A GARCH-M model takes the form of

$$y_t = x_t'\gamma + \delta\sigma_t^2 + u_t,$$

$$u_t = \sigma_t\varepsilon_t,$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2 + \beta_1\sigma_{t-1}^2 + \cdots + \beta_s\sigma_{t-s}^2.$$

Note that $y_t|x_t, \mathcal{F}_{t-1} =_d \mathbb{N}(x_t'\gamma + \delta\sigma_t^2, \sigma_t^2)$ under the normality assumption of ε_t. The GARCH-M model could therefore be estimated by conditional maximum likelihood.

We could also have other specification. For example, we could have $y_t = x_t'\gamma + \delta\sigma_t + u_t$ or $y_t = x_t'\gamma + \delta \ln\sigma_t^2 + \varepsilon_t$.

9.5 The EGARCH Model

Nelson (1991) propose the exponential GARCH (EGARCH) model. If u_t follows an EGARCH model, then

$$u_t = \sigma_t\varepsilon_t,$$

$$\ln\sigma_t^2 = \pi_0 + \sum_{i=1}^{\infty}\pi_i\left(|\varepsilon_{t-i}| - \mathbb{E}\,|\varepsilon_{t-i}| + \nu\varepsilon_{t-i}\right),$$

where ε_t is iid with mean zero and unit variance.

The parameter ν in the EGARCH model can generate asymmetric dynamics in volatility. When $\nu = 0$, a positive shock to ε (and therefore the level y) has the same effect on the volatility as a negative shock of the same magnitude. When $-1 < \nu < 0$, a positive shock has a weaker effect on the volatility than a negative shock of the same magnitude. When $\nu < -1$, positive shocks and negative shocks generate effects of different directions. If the πs are positive and $\nu < 1$, the model generates the so called leverage effect: the negative correlation between asset returns and their volatilities.

To estimate the EGARCH models, one has to give a parametric specification for the infinite sum. Usually, as in the ARMA models, we assume that $\pi(L) = \sum_{i=1}\pi_i L^i$ can be expressed as the ratio of two finite order lag polynomials. As a consequence, the model may be parameterized in an ARMA form as

$$\begin{aligned}\ln\sigma_t^2 = \beta_0 + \beta_1\ln\sigma_{t-1}^2 + \cdots + \beta_s\ln\sigma_{t-m}^2 + \alpha_1\left(|\varepsilon_{t-1}| - \mathbb{E}\,|\varepsilon_{t-1}| + \nu\varepsilon_{t-1}\right)\\ + \cdots + \alpha_m\left(|\varepsilon_{t-m}| - \mathbb{E}\,|\varepsilon_{t-m}| + \nu\varepsilon_{t-m}\right).\end{aligned}$$

Given a distributional specification of ε_t, we may estimate the model with conditional maximum likelihood estimation.

9.6 Other Models of Conditional Heteroskedasticity

There are many other variants to the ARCH models that appear in the study of financial econometrics. The threshold GARCH model proposed by Glosten et al. (1993) captures the asymmetric effect by specifying the dynamics of conditional variance on top of the basic GARCH setting as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m}(\alpha_i + \gamma_i I_{t-i})u_{t-i}^2 + \sum_{i=1}^{s}\sigma_{t-i}^2,$$

where I_t is an indicator function defined by

$$I_t = \begin{cases} 1, & \text{if } u_t \leq c, \\ 0, & \text{otherwise} \end{cases}$$

where c is a constant that represents some threshold. This model can be estimated by conditional maximum likelihood.

A time series y_t is said to follow an RCA(p) model if

$$y_t = \phi_0 + \sum_{i=1}^{p}(\phi_i + \delta_{it})y_{t-i} + \varepsilon_t$$

where $\delta_t = (\delta_{1t}, \delta_{2t}, \ldots, \delta_{pt})'$ is a sequence of independent random vectors with mean zero and variance Ω_δ, and $\{\delta_t\}$ is independent of $\{\varepsilon_t\}$. The conditional mean of y_t is

$$\mu_t = \phi_0 + \sum_{i=1}^{p}\phi_i y_{t-i},$$

and the conditional variance of the model is

$$\sigma_t^2 = \sigma_\varepsilon^2 + (y_{t-1}, \ldots, y_{t-p})\Omega_\delta(y_{t-1}, \ldots, y_{t-p})'.$$

This model could be estimated by conditional maximum likelihood estimation.

Another way to generate time-varying volatility into the model is to introduce an innovation to the conditional variance equation of u_t. Such models are called stochastic volatility models. A stochastic volatility model takes the form of

$$u_t = \sigma_t \varepsilon_t,$$

$$\ln \sigma_t^2 = \alpha_0 + \alpha_1 \ln \sigma_{t-1}^2 + \cdots + \alpha_m \ln \sigma_{t-m}^2 + v_t,$$

where v_t are iid normal random variables with mean zero and variance σ_v^2, and $\{\varepsilon_t\}$ and $\{v_t\}$ are independent. To estimate these models, we need to use quasi-maximum likelihood with the Kalman filter, or use Markov chain Monte Carlo (MCMC) method.

9.7 Multivariate GARCH

We may easily generalize the GARCH model to the multivariate setting. A multivariate GARCH model takes the form of

$$y_t = \Pi x_t + u_t,$$

$$H_t = \mathbb{E}(u_t u_t' | y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots),$$

$$H_t = A_0 + A_1 u_{t-1} u_{t-1}' A_1' + \cdots + A_m u_{t-1} u_{t-1}' A_m' + B_1 H_{t-1} B_1' + \cdots + B_s H_{t-s} B_s',$$

where y and x are vectors, u is a vector white noise, and H, A and B are matrices. To estimate the model, we continue to use conditional maximum likelihood, but usually we need to restrict the parameters so that the numerical maximization becomes feasible. For example, we usually restrict Hs to be diagonal. Sometimes we also restrict A and Bs to be diagonal

10 Nonstationary Time Series

Studies have shown that many of the financial and aggregate macroeconomic time series exhibit features of random walk or nonstationarity. Influential examples include Hall (1978), Nelson and Plosser (1982) and Marsh and Merton (1986).

10.1 The Invariance Principle

Let the probability space be $(\Omega, \mathcal{F}, \mathbb{P})$. A continuous-time stochastic process $W = (W_t)_{t\in\mathbb{R}_+}$ is called a standard Brownian motion (or a Wiener process) if

(a) $W_0(\omega) = 0$ for all $\omega \in \Omega$,

(b) The mapping $t \mapsto W_t(\omega)$ is a continuous function for all $\omega \in \Omega$ (continuous sample paths).

(c) For every $t, h \geq 0$, $W_{t+h} - W_t$ is independent of $(W_s)_{s\leq t}$, and has a normal distribution with mean zero and variance h (independent Gaussian increments).

A vector Brownian motion $B = (B_t)_{t\in\mathbb{R}_+}$ with covariance matrix Ξ is the stochastic process $(\Xi^{1/2}W_t)$ where W_t is a vector Brownian motion whose components are independent standard Brownian motions.

Let $\{w_t\}_{t=1,2,...}$ be a sequence of m-dimensional random variables. Define

$$B_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_t$$

for $0 \leq r \leq 1$. If $B_n \to_d B$ where B is a vector Brownian motion with covariance matrix

$$\Xi = \lim_{n\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=1}^{T} w_t\right)\left(\sum_{t=1}^{T} w_t\right)'$$

provided that Ξ exists, then we say that $\{w_t\}$ satisfies the invariance principle (IP) or functional central limit theorem (FCLT). It is easy to see that Ξ is the long-run variance of $\{w_t\}$. It is well know that if $\{w_t\}$ is an iid sequence with $\text{Cov}(w_t) = \Sigma$, then it satisfies the IP with covariance Σ. This is called the Donsker's theorem. See, e.g., Billingsley (1999, Chapter 14).

The invariance principle becomes very powerful when applied with the continuous mapping theorem: if $X_n \to_d X$ where X has distribution $\mathbf{P}$ and f is continuous -a.s., then $f(X_n) \to_d f(X)$. For example, let $B_n \to_d B$. Since the function f from the space of all continuous functions on $[0, 1]$ to $\mathbb{R}$ defined by $f(g) = \int_0^1 g(r)\mathrm{d}r$ is continuous, we have that $\int_0^1 B_T(r)\mathrm{d}r \to_d \int_0^1 B(r)\mathrm{d}r$. Similarly, we have results like $\int_0^1 rB_T(r)\mathrm{d}r \to_d \int_0^1 rB(r)\mathrm{d}r$. Also, we have $B_T(1) \to_d B(1)$, which implies that

$$B_T(1) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \to_d \mathbb{N}(0, \Xi).$$

Therefore, the functional central limit theorem implies the central limit theorem. Actually, the

functional central limit theorem implies much more than the central limit theorem.

The invariance principle holds not only for iid sequences, but also for heteroskedastic and serially dependent series, given that the heteroskedasticity and serial dependence are appropriately controlled. For example, IP holds for sequences that satisfy certain mixing conditions and also for martingale difference sequences that satisfies the conditions in Theorem 2.31. See, e.g., McLeish (1974, 1977), Kuelbs and Philipp (1980), Herrndorf (1983, 1984a,b, 1985), Peligrad (1985), Eberlein (1986), Phillips and Durlauf (1986) and Phillips and Solo (1992). Also, IP holds for general linear processes.

Theorem 10.1. *Let*

$$w_t = \Phi(L)\varepsilon_t = \sum_{i=0}^{\infty} \Phi_i \varepsilon_{t-i}$$

where $\sum_{i=1}^{\infty} i\,|\Phi_i| < \infty$ *and* $\varepsilon_t \sim$ *iid* $(0, \Sigma)$. *Then*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_t \to_d B$$

where B *is a Brownian motion with covariance matrix* $\Phi(1)\Sigma\Phi(1)'$.

Proof. The series $\{w_t\}$ admits the Beveridge-Nelson decomposition representation

$$w_t = \Phi(1)\varepsilon_t - (e_t - e_{t-1})$$

where $e_t = \sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} \phi_j \varepsilon_{t-i}$. Note that by Chebyshev's inequality $e_t = O_p(1)$ uniformly in t. Then we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_t = \Phi(1) \frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} \varepsilon_t + \frac{1}{\sqrt{T}}(e_{Tr} - e_0) \to_d B$$

where B is a Brownian motion with covariance matrix $\Phi(1)\Sigma\Phi(1)'$. ■

Phillips (1987a) shows that under suitable weak dependence and heteroskedasticity conditions for w_t, we may establish an asymptotic expansion for $B_T(r)$ given by

$$B_T(r) =_d B(r) + R$$

where R could be $O_p(T^{-1/2})$ or $O_p(T^{-1})$, depending on the third-order cumulant of the process $B_T(r)$.

10.2 Introduction to Stochastic Calculus

Before we introduce our important results, we first introduce the idea of stochastic integral. The formal theory of stochastic integral is very involved. We will just give some of the main ideas here. Let $M = \{M(t)\}$ be a continuous-time martingale process with respect to some filtration $(\mathcal{F}_t)_{t\geq 0}$

and $X = \{X(t)\}$ be a continuous-time stochastic process that is adapted to $(\mathcal{F}_t)$. Recall that if g is a continuous real function and h is of bounded variation on $[a, b]$, then for the Riemann sum

$$\sum_{i=0}^{n-1} g(s_i)(h(t_{i+1}) - h(t_i))$$

where $a = t_0 < t_1 < \cdots < t_n = b$ is a partition of $[a, b]$ and $s_i \in [t_i, t_{i+1}]$, its limit exists as the partition becomes finer and finer, and the limit, which we denote by the Riemann-Stieltjes integral $\int_a^b g(t)\mathrm{d}h(t)$, is independent of the choice of s_i. However, if we apply this idea to X and M and try to figure out the limit (in some probability sense) of the partial sum

$$\sum_{i=0}^{n-1} X(s_i)(M(t_{i+1}) - M(t_i)),$$

we find that the limit depends on the choice of s_i. This is because M is not of bounded variation. We therefore define the Ito integral as

$$\int_a^b X(t)\mathrm{d}M(t) = \operatorname{plim} \sum_{i=0}^{n-1} X(s_i)(M(t_{i+1}) - M(t_i)),$$

where the limit is taking as the partition becomes finer and finer. It is known that the Ito integral is well defined if M is square integrable, that is, $\mathbb{E}M^2(t) < \infty$ for all t.

To state the Ito formula, which plays the role of the fundamental theorem of calculus in the stochastic integral setting, we define the quadratic variation $[X]_t$ of a stochastic process (X_t) to be the limit in probability of

$$\sum_{k=1}^{n} (X_{t_k} - X_{t_{k-1}})^2$$

where the limit is taken over all partitions of the interval $[0, t]$ such that the mesh of the partition goes to zero. It is well known that $[W]_t = t$ for a standard Brownian motion W, and it is easy to derive that if B_t is a Brownian motion with variance σ^2, then $[B]_t = \sigma^2 t$. The covariation $[X, Y]_t$ of two processes X and Y is defined to be the limit in probability of

$$\sum_{k=1}^{n} (X_{t_k} - X_{t_{k-1}})(Y_{t_k} - Y_{t_{k-1}})$$

where the limit is taken over all partitions of the interval $[0, t]$ such that the mesh of the partition goes to zero.

Now if f is a twice continuously differentiable function and M is a square integrable martingale process,

$$\mathrm{d}f(M_t) = f'(M_t)\mathrm{d}M_t + \frac{1}{2} f''(M_t)\mathrm{d}[M]_t.$$

If (M) and (N) are two square integrable martingale process with respect to some filtration

(F_t), then the integration by parts formula for Ito integral is given by

$$\int_0^s N(t)\mathrm{d}M(t) = N(t)M(t) - N(0)M(0) - \int_0^s M(t)\mathrm{d}N(t) - [N,M]_t.$$

This formula can be generalized to the case in which the two processes are semi-martingale. In particular, if $X = R + M$ and $Y = S + N$ where M and N are defined as above and R and S are $\mathcal{F}_t$-adapted continuous processes of bounded variation, then

$$\int_0^s Y(t)\mathrm{d}X(t) = Y(t)X(t) - Y(0)X(0) - \int_0^s X(t)\mathrm{d}Y(t) - [N,M]_t.$$

For a full treatment of stochastic integration, see, e.g., Karatzas and Shreve (2000, Chapter 3). We now introduce a result from Chan and Wei (1988), whose proof we shall follow next.

Theorem 10.2. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables such that $\{(X_t, Y_t)\}$ is a martingale difference sequence with respect to a filtration $\{\mathcal{F}_t\}$ with $\mathbb{E}(X_t^2|\mathcal{F}_{t-1}) < c$ and $\mathbb{E}(Y_t^2|\mathcal{F}_{t-1}) < c$ for some $c > 0$. Suppose*

$$\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} X_t, \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} Y_t\right) \to_d (H, W)$$

where H and W are two Brownian motions with respect to a filtration $(\mathcal{G}_t)$. Let $Z_t = \sum_{k=1}^t X_t$. Then

$$\frac{1}{T}\sum_{t=2}^{T} Z_{t-1}Y_t \to_d \int_0^1 H(r)\mathrm{d}W(r).$$

Proof. For notation convenience, write $\tilde{H}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} X_t$ and $\tilde{W}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} Y_t$. Since $(\tilde{H}_T, \tilde{W}_T) \to_d (H, W)$, by the Skorokhod representation theorem, there exist a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and random elements H_T, W_T with values in $D[0,1]$ such that $(H_T, W_T) =_d (\tilde{H}_T, \tilde{W}_T)$ and (H_T, W_T) converges in the Skorokhod topology to (H, W) almost surely. Since H and W have sample paths in $C[0,1]$, convergence in the Skorokhod topology implies uniform convergence, i.e.,

$$\sup_r \|(H_T(r), W_T(r)) - (H(r), W(r))\| \to 0 \quad \text{a.s..}$$

See Billingsley (1999, Section 12). Let

$$G_T = \sum_{t=1}^{T} H_T\left(\frac{t-1}{T}\right)\left(W_T\left(\frac{t}{T}\right) - W_T\left(\frac{t-1}{T}\right)\right).$$

Then $G_T =_d \frac{1}{T}\sum_{t=2}^{T} Z_{t-1}Y_t$ (note that $\tilde{H}_T(0)$, therefore $H_T(0)$, are zero for all T). We need to show that $G_T \to_d \int_0^1 H\mathrm{d}W$.

Fix $\epsilon > 0$. By Egorov's theorem, there exists $A \in \mathcal{F}'$ such that $\mathbb{P}'(A^c) < \varepsilon$ such that

$$\sup_{\omega \in A} \sup_r \|(H_T(r,\omega), W_T(r,\omega)) - (H(r,\omega), W(r,\omega))\| = \delta_T \to 0$$

where $\{\delta_T\}$ is a sequence of non-random numbers. For each T, we may choose an integer k_T such that $k_T \to \infty, k_T\delta_T^2 \to 0$ and $k_T/T \to 0$ as $T \to \infty$. Then for each T we may choose a sequence of integers $\{n_{T1}, n_{T2}, \ldots, n_{Tk_T}\}$ in $\{1, \ldots, T\}$, which in turn defines a partition $\{t_{T0}, t_{T1}, \ldots, t_{Tk_T}\}$ of $[0,1]$, such that $t_{Ti} = n_{Ti}/T$, $0 = t_{T0} < t_{T1} < \cdots < t_{Tk_T} = 1$ and that $\max_i |t_{T,i+1} - t_{Ti}| \to 0$ as $T \to \infty$.

Let

$$\begin{aligned} J_T &= G_T - \sum_{k=1}^{k_T} H_T(t_{k-1})(W_T(t_k) - W_T(t_{k-1})) \\ &= \sum_{k=1}^{k_T} \sum_{i=n_{T,k-1}}^{n_{Tk}-1} \left(H_T\left(\frac{i}{T}\right) - H_T(t_{k-1})\right)\left(W_T\left(\frac{i+1}{T}\right) - W_T\left(\frac{i}{T}\right)\right) \end{aligned}$$

Properties of martingale difference sequences implies that

$$\begin{aligned} \mathbb{E}J_T^2 &= \sum_{k=1}^{k_T} \sum_{i=n_{T,k-1}}^{n_{Tk}-1} \mathbb{E}\left(H_T\left(\frac{i}{T}\right) - H_T(t_{k-1})\right)^2 \left(W_T\left(\frac{i+1}{T}\right) - W_T\left(\frac{i}{T}\right)\right)^2 \\ &\le \sum_{k=1}^{k_T} \sum_{i=n_{T,k-1}}^{n_{Tk}-1} c^2 \left(\frac{i}{T} - \frac{n_{T,k-1}}{T}\right)\frac{1}{T} \\ &\le c^2 \sum_{k=1}^{k_T} (t_k - t_{k-1})^2 = o(1). \end{aligned}$$

Therefore,

$$G_T = \sum_{k=1}^{k_T} H_T(t_{k-1})(W_T(t_k) - W_T(t_{k-1})) + o_p(1).$$

By Cauchy-Schwartz inequality,

$$\begin{aligned} &\mathbb{E}\left(I_A \sum_{k=1}^{k_T} (H_T(t_{k-1}) - H(t_{k=1}))(W_T(t_k) - W_T(t_{k-1}))\right)^2 \\ \le &\mathbb{E}\left(\sum_{k=1}^{k_T} (H_T(t_{k-1}) - H(t_{k-1}))^2 I_A\right)\left(\sum_{k=1}^{k_T} (W_T(t_k) - W_T(t_{k-1}))^2\right) \\ \le &k_T\delta_T^2 \sum_{k=1}^{k_T} \mathbb{E}(W_T(t_k) - W_T(t_{k-1}))^2 \\ \le &k_T\delta_T^2 c \to 0. \end{aligned}$$

Therefore,

$$I_A G_T = I_A \sum_{k=1}^{k_T} H(t_{k-1})(W_T(t_k) - W_T(t_{k-1})) + o_p(1).$$

With summation/integration by parts and a similar argument as above, we have

$$\begin{aligned}
& I_A \sum_{k=1}^{k_T} H(t_{k-1})(W_T(t_k) - W_T(t_{k-1})) \\
=& - I_A \sum_{k=1}^{k_T} W_T(t_k)(H(t_k) - H(t_{k-1})) + H(1)W_T(1) \\
=& - I_A \sum_{k=1}^{k_T} W(t_k)(H(t_k) - H(t_{k-1})) + H(1)W(1) + o_p(1) \\
=& I_A \sum_{k=1}^{k_T} H(t_{k-1})(W(t_k) - W(t_{k-1})) + o_p(1) \\
=& I_A \int_0^1 H(r)\mathrm{d}W(r) + o_p(1).
\end{aligned}$$

The last equality follows from the definition of Ito integral. Since ϵ is arbitrary, we have that $G_T \to_d \int_0^1 H\mathrm{d}W$. This completes the proof. ■

We make a few remarks here.

(a) We make replace the martingale difference sequence assumption on X_t by the assumption that $\{X_t\}$ is an $\mathcal{F}_t$-adapted mean-zero weakly stationary time series with absolute convergent autocovariances without essentially affecting the proof. (The only difference is that in some places of the proof c may need to be replaced by $c\sum_{k=-\infty}^{\infty}|\gamma(k)|$ where $\gamma(\cdot)$ is the autocovariance function of X_t.) Phillips (1988b) provides similar results of convergence to stochastic integral for innovations that satisfying some strong mixing conditions.

(b) This proof can be directly generalized to the multidimensional case. In particular, if H and W are two vector Brownian motions whose components are denoted respectively by H^i and W^i, then the $\int_0^1 H\mathrm{d}W'$ is defined to be the random matrix whose (i,j)-th entry is $\int_0^1 H^i\mathrm{d}W^j$.

10.3 Some Important Asymptotic Results

Now we assume that w_t is a linear process defined as in Section 10.1. Let $z_t = \sum_{i=1}^t w_i$. The process $\{z_t\}$ is said to be an *integrated* process. Such a process contains a *stochastic trend*, which should be obvious if one plots the process.

Besides stochastic trends, we also need to deal with deterministic trends. Let $\{c_t\}$ be a sequence of deterministic vectors whose i-th component is denoted by $\{c_{it}\}$. Suppose there exists $\delta_i \geq 0$ and $f_i \in L^2[0,1]$ of bounded variation such that $f_{Ti} \in D[0,1]$ defined by

$$f_{Ti}(r) = \frac{c_{i[Tr]}}{T^{\delta_i}}$$

satisfies

$$f_{Ti} \to_{L^2} f_i.$$

We assume that the f_i's are linearly independent, and write $f = (f_1, \ldots, f_\ell)$.

Now we introduce some important asymptotics.

Theorem 10.3. *Let w_t, z_t, c_t, B be defined as above. Then*

$$\frac{1}{T^{\delta_i+3/2}} \sum_{t=1}^{T} c_{it} z_t \to_d \int_0^1 f_i(r) B(r) \mathrm{d}r,$$

$$\frac{1}{T^2} \sum_{t=1}^{T} z_t z_t' \to_d \int_0^1 B(r) B(r)' \mathrm{d}r,$$

$$\frac{1}{T^{\delta_i+1/2}} \sum_{t=1}^{T} c_{it} w_t' \to_d \int_0^1 f_i(r) \mathrm{d}B(r)',$$

$$\frac{1}{T} \sum_{t=1}^{T} z_{t-1} w_t' \to_d \int_0^1 B(r) \mathrm{d}B(r)' + \Lambda'$$

and

$$\frac{1}{T} \sum_{t=1}^{T} z_t w_t' \to_d \int_0^1 B(r) \mathrm{d}B(r)' + \Lambda^{\circ\prime}$$

where $\Lambda = \sum_{k=1}^{\infty} \Gamma(k)$ and $\Lambda^\circ = \sum_{k=0}^{\infty} \Gamma(k)$.

Proof. Write $B_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{Tr} w_t$. Note that $B_T \to_d B$. We have

$$\begin{aligned}
\frac{1}{T^{\delta_i+3/2}} \sum_{t=1}^{T} c_{it} z_t &= \frac{1}{T} \sum_{t=1}^{T} \frac{c_{it}}{T^{\delta_i}} \frac{z_t}{\sqrt{T}} \\
&= \sum_{t=1}^{T} \int_{(t-1)/T}^{t/T} f_{Ti}(r) B_T(r) \mathrm{d}r \\
&= \int_0^1 f_{Ti}(r) B_T(r) \mathrm{d}r \\
&\to_d \int_0^1 f_i(r) B(r) \mathrm{d}r.
\end{aligned}$$

Since the mapping $(g_1, g_2) \mapsto \int_0^1 g_1(r) g_2(r) \mathrm{d}r$ is continuous, the last step follows from the continuous mapping theorem. The integration is in the Lebesgue sense.

Similarly, we have

$$\frac{1}{T^2} \sum_{t=1}^{T} z_t z_t' = \frac{1}{T} \sum_{t=1}^{T} \frac{z_t}{\sqrt{T}} \frac{z_t'}{\sqrt{T}} \to_d \int_0^1 B(r) B(r)' \mathrm{d}r.$$

Use summation and integration by parts, we have

$$\begin{aligned}
\frac{1}{T^{\delta_i+1/2}}\sum_{t=1}^{T} c_{it}w_t' &= \frac{1}{T^{\delta_i+1/2}}\sum_{t=1}^{T} c_{it}(z_t - z_{t-1}) \\
&= \frac{c_{iT}w_T}{T^{\delta_i+1/2}} - \sum_{t=1}^{T}\frac{c_{it}-c_{i,t-1}}{T^{\delta_i}}\frac{z_t}{\sqrt{T}} \\
&= f_{Ti}(1)B_T(1)' - \int_0^1 \mathrm{d}f_{Ti}(r)B_T(r)' + o_p(1) \\
&\to_d f_i(1)B(1)' - \int_0^1 \mathrm{d}f_i(r)B(r) \\
&= \int_0^1 f_i(r)\mathrm{d}B(r)'.
\end{aligned}$$

Note that the third equality is due to the definition of Stieltjes integrals and that f_{Ti} and f_i are of bounded variation. The convergence in distribution is due to the continuous mapping theorem, and the last equality is due to integration by parts of Ito integral and the fact that the cross variation of a continuous martingale with a process (function) of bounded variation is zero.

Write $w_t = \Phi(1)\varepsilon_t - (e_t - e_{t-1}) = \tilde{w}_t - (e_t - e_{t-1})$. Then $\tilde{w}_t$ is a martingale difference sequence with $\tilde{B}_T(r) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]}\tilde{w}_t \to_d B(r)$. Using summation by parts, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} z_{t-1}w_t' &= \frac{1}{T}\sum_{t=1}^{T} z_{t-1}\tilde{w}_t' - \frac{1}{T}\sum_{t=1}^{T} z_{t-1}(e_t - e_{t-1})' \\
&= \frac{1}{T}\sum_{t=1}^{T} z_{t-1}\tilde{w}_t' - \frac{1}{T}z_T e_T' + \frac{1}{T}\sum_{t=1}^{T} w_t e_t'.
\end{aligned}$$

The first term converges in distribution to $\int_0^1 B(r)\mathrm{d}B(r)'$ by Theorem 10.2. The second term converge in probability to zero since $z_T = O_p(\sqrt{T})$ and $e_T = O_p(1)$. The third term converges in probability to $\mathbb{E}w_t e_t'$ while

$$\begin{aligned}
\mathbb{E}w_t e_t' &= \mathbb{E}\left(\sum_{i=0}^{\infty}\Phi_i\varepsilon_{t-i}\right)\left(\sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\Phi_j\varepsilon_{t-i}\right)' \\
&= \sum_{i=0}^{\infty}\sum_{j=i+1}^{\infty}\Phi_i\Sigma\Phi_j' \\
&= \sum_{k=1}^{\infty}\sum_{i=0}^{\infty}\Phi_i\Sigma\Phi_{i+k}' \\
&= \sum_{k=1}^{\infty}\Gamma(k)'.
\end{aligned}$$

This completes the proof that $\frac{1}{T}\sum_{t=1}^{T} z_{t-1}w_t' \to_d \int_0^1 B(r)\mathrm{d}B(r)' + \Lambda'$. And this implies immediately

that $\frac{1}{T}\sum_{t=1}^{T} z_t w_t' \to_d \int_0^1 B(r)\mathrm{d}B(r)' + \Lambda^{\circ\prime}$ by noting that $\frac{1}{T}\sum_{t=1}^{T} w_t w_t' \to_p \Gamma(0)$. ■

The martingale approximation idea in the above proof is due to Phillips (1988a). Also, note that $\mathbb{E}\frac{1}{T}\sum_{t=1}^{T} z_{t-1} w_t' \to_p \Lambda'$ and $\mathbb{E}\frac{1}{T}\sum_{t=1}^{T} z_t w_t' \to_p \Lambda^{\circ\prime}$. So Λ' and $\Lambda^{\circ\prime}$ may be viewed as the bias terms (the stochastic integrals in this theorem are mean zero). For reference for the later part of this chapter, let $\{w_{1t}\}, \{w_{2t}\}$ be two linear processes such that $(w_{1t}', w_{2t}')'$ satisfies an invariance principle, whose limit is the Brownian motion (B_1', B_2') with long run variance $\Xi = \sum_{k=-\infty}^{\infty} \Gamma(k)$. Let $z_{1t} = \sum_{s=1}^{t} w_{1t}$ and $z_{2t} = \sum_{s=1}^{t} w_{2t}$. Let $\Lambda = \sum_{k=1}^{\infty} \Gamma(k)$ and $\Lambda^{\circ} = \sum_{k=0}^{\infty} \Gamma(k)$ and partition them respectively as

$$\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \Lambda_{11}^{\circ} & \Lambda_{12}^{\circ} \\ \Lambda_{21}^{\circ} & \Lambda_{22}^{\circ} \end{bmatrix}$$

to match the dimension of w_{1t} and w_{2t}. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} z_{2,t-1} w_{1t}' \to_d \int_0^1 B_2(r)\mathrm{d}B_1(r) + \Lambda_{21}$$

and

$$\frac{1}{T}\sum_{t=1}^{T} z_{2t} w_{1t}' \to_d \int_0^1 B_2(r)\mathrm{d}B_1(r) + \Lambda_{21}^{\circ}.$$

10.4 Unit Roots

We call a time series $\{X_t\}$ integrated of order k, and write $\{X_t\} \sim I(k)$, if the k-th differencing is needed to make $\{X_t\}$ stationary. That is, $\{\Delta^k X_t\}$ is stationary, or, $\{\Delta^k X_t\} \sim I(0)$.

The $I(1)$-ness of a univariate time series $\{y_t\}$ may be tested through the regression

$$y_t = \alpha y_{t-1} + u_t$$

where u_t is stationary. Under the null, $\alpha = 1$, and the above equation becomes an autoregression with a unit root. Therefore, an $I(1)$ process is often called a *unit root* process. The tests for $I(1)$-ness would therefore be referred to as the *unit root tests*.

Now we consider the OLS estimate of α given by

$$\hat{\alpha} = \frac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Note that this is also the maximum likelihood estimator under the assumption of normality. Suppose that $\{u_t\}$ satisfies the invariance principle with $\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} w_t \to_d B(r)$. Then it is easy to see that

$$T(\hat{\alpha} - 1) = \frac{\frac{1}{T}\sum_{t=1}^{T} y_{t-1} u_t}{\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2} \to_d \frac{\int_0^1 B(r)\mathrm{d}B(r) + \lambda}{\int_0^1 B(r)^2 \mathrm{d}r},$$

where $\lambda = \sum_{k=1}^{\infty} \gamma(k)$, $\gamma(\cdot)$ is the autocovariance function of $\{u_t\}$. Also, let $\hat{u}_t$ be the residual of

the OLS regression, and estimate the error variance by $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^T \hat{u}_t^2$, we have

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^T u_t^2 - \frac{1}{T}\frac{\left(\frac{1}{T}\sum_{t=1}^T y_{t-1}u_t\right)^2}{\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2} = \frac{1}{T}\sum_{t=1}^T u_t^2 + O_p(1/T) \to_p \sigma^2$$

where $\sigma^2 = \mathbb{E}u_t^2$. Then we may define the t statistic (the likelihood ratio test statistics) of α and have

$$t(\alpha) = \frac{\hat{\alpha} - 1}{\hat{\sigma}\left(\sum_{t=1}^T y_{t-1}^2\right)^{-1/2}} = \frac{\frac{1}{T}\sum_1^T y_{t-1}u_t}{\hat{\sigma}\left(\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2\right)^{1/2}} \to_d \frac{\int_0^1 B(r)\mathrm{d}B(r) + \lambda}{\sigma\left(\int_0^1 B(r)^2\mathrm{d}r\right)^{1/2}}.$$

10.5 The Dickey-Fuller Test

If $u_t \sim \mathrm{WN}(0, \sigma^2)$, then $\lambda = 0$, and $B = \sigma W$ where W is the standard Brownian motion. Then we have

$$T(\hat{\alpha} - 1) \to_d \frac{\int_0^1 W(r)\mathrm{d}W(r)}{\int_0^1 W(r)^2\mathrm{d}r}$$

and

$$t(\alpha) \to_d \frac{\int_0^1 W(r)\mathrm{d}W(r)}{\left(\int_0^1 W(r)^2\mathrm{d}r\right)^{1/2}}.$$

Note that the limit distribution the above two statistics are free of nuisance parameters and may be obtained by simulation. They are usually referred to as the Dickey-Fuller distributions (Dickey and Fuller, 1979) and the test for unit root based on the $t(\alpha)$ statistic of the AR(1) model is called the Dickey-Fuller test. For more details including exact distribution as well as power of tests based directly on the OLS estimator of α under the assumption of normality, see Evans and Savin (1981) and Evans and Savin (1984).

10.6 The Augmented Dickey-Fuller Test

We may allow u_t to be an AR(p) process. That is, $\alpha(L)u_t = \varepsilon_t$ where $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$ and $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_P z^p$. In this case, we may test the null hypothesis $\alpha = 1$ based on the regression

$$y_t = \alpha y_{t-1} + \sum_{k=1}^p \alpha_k \Delta y_{t-k} + \varepsilon_t.$$

The t statistic for α is given by

$$t(\alpha) = \frac{\hat{\alpha} - 1}{\hat{\sigma}\left(\sum_{t=1}^T y_{t-1}^2\right)^{-1/2}} = \frac{T(\hat{\alpha} - 1)}{\hat{\sigma}\left(\frac{1}{T^2}\sum_{t=1}^T y_{t-1}^2\right)^{-1/2}}.$$

To obtain $\hat{\alpha}$, write $v_t = (\Delta y_{t-1}, \ldots, \Delta y_{t-p})'$ and $\gamma = (\alpha_1, \ldots, \alpha_p)'$. Note that v_t is $I(0)$. Then our regression model may be written as $y_t = \alpha y_{t-1} + v_t'\gamma + \varepsilon_t$. By two-step regression

$$\begin{aligned}
T(\hat{\alpha}-1) =& T\left[\sum_{t=1}^{T} y_{t-1}^2 - \sum_{t=1}^{T} y_{t-1}v_t' \left(\sum_{t=1}^{T} v_t v_t'\right)^{-1} \sum_{t=1}^{T} v_t y_{t-1}\right]^{-1} \\
& \bullet \left[\sum_{t=1}^{T} y_{t-1}\varepsilon_t - \sum_{t=1}^{T} y_{t-1}v_t' \left(\sum_{t=1}^{T} v_t v_t'\right)^{-1} \sum_{t=1}^{T} v_t \varepsilon_t\right] \\
=& \left[\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2 - \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}v_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T} v_t v_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} v_t y_{t-1}\right)\right]^{-1} \\
& \bullet \left[\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\varepsilon_t - \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}v_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T} v_t v_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} v_t \varepsilon_t\right)\right] \\
=& \left[\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2 + O_p\left(\frac{1}{T}\right)\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\varepsilon_t + o_p(1)\right] \\
=& \left(\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}\varepsilon_t\right) + o_p(1).
\end{aligned}$$

Note that if $\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]}\varepsilon_t \to_d B(r)$, then $\frac{1}{\sqrt{T}}\sum_{t=1}^{[Tr]} u_t \to_d \frac{B(r)}{\alpha(1)}$. Then

$$T(\hat{\alpha}-1) \to_d \frac{\int_0^1 \frac{B(r)}{\alpha(1)}\mathrm{d}B(r)}{\int_0^1 \left(\frac{B(r)}{\alpha(1)}\right)^2 \mathrm{d}r} = \frac{\alpha(1)\int_0^1 W(r)\mathrm{d}W(r)}{\int_0^1 W(r)^2\mathrm{d}r}$$

and

$$t(\alpha) \to_d \frac{\int_0^1 \frac{B(r)}{\alpha(1)}\mathrm{d}B(r)}{\sigma\left(\int_0^1 \left(\frac{B(r)}{\alpha(1)}\right)^2 \mathrm{d}r\right)^{1/2}} = \frac{\int_0^1 W(r)\mathrm{d}W(r)}{\left(\int_0^1 W(r)^2\mathrm{d}r\right)^{1/2}}.$$

We may conduct our test based on $\frac{T(\hat{\alpha}-1)}{\hat{\alpha}(1)}$ or $t(\alpha)$ where $\hat{\alpha}(1)$ is the OLS estimate for $\alpha(1)$. This t test procedure above is called the augmented Dickey-Fuller test after Dickey and Fuller (1981). Said and Dickey (1984) study the case when u_t is an MA process with unknown order, and the augmented Dickey-Fuller test procedures can still be used. See also Solo (1984).

When u_t is in general serially correlated (not necessarily follows an AR process), the Phillips-Perron unit root test directly corrects the bias term by nonparametric methods. Interested readers are invited to refer to Phillips (1987b), Phillips and Perron (1988) and Hamilton (1994, Section 17.6).

One may also conduct unit root tests based on the residuals. See, for example, Sargan and Bhargava (1983) and Bhargava (1986).

10.7 Testing for a Unit Root with Maintained Time Trends

Now we suppose that the time series $\{y_t\}$ is generated as

$$y_t = \pi' c_t + y_t^\circ$$

where $\{c_t\}$ is a vector of deterministic sequence and y_t° is the stochastic component of the process $\{y_t\}$. To test for the presence of a unit root in the stochastic component of $\{y_t\}$, we utilize the regression

$$y_t = \pi' c_t + \alpha y_{t-1} + u_t \tag{10.1}$$

When $c_t = (1, t)'$ and $|\alpha| < 1$, the process is said to be *trend stationary*. When $c_t = 1$ and $\alpha = 1$, the process is said to be *difference stationary*.

We may also conduct the test based on

$$y_t^\circ = \alpha y_{t-1}^\circ + u_t. \tag{10.2}$$

Of course, the unobserved y_t° needs to be replaced with the OLS residual $\hat{y}_t^\circ$ from regressing y_t on c_t. That is,

$$\begin{aligned}
\hat{y}_t^\circ &= y_t - \left(\sum_{t=1}^T y_t c_t'\right)\left(\sum_{t=1}^T c_t c_t'\right)^{-1} c_t \\
&= y_t^\circ - \left(\sum_{t=1}^T y_t^\circ c_t'\right)\left(\sum_{t=1}^T c_t c_t'\right)^{-1} c_t
\end{aligned}$$

Then it is obvious that the OLS estimator of α in (10.1) is equivalent to the OLS estimator of the regression (10.2) using the residuals.

On the other hand, let δ_i be the smallest number such that $c_{i[Tr]}/T^\delta$ converges to f_i in $L^2[0,1]$. Let $D_T = \mathrm{diag}(T^{\delta_1}, \ldots, T^{\delta_m})$ and $c_t^* = D_T^{-1} c_t$. Then $c_{[Tr]} \to_{L^2} f(r) = (f_1(r), \ldots, f_m(r))$.

Then

$$\begin{aligned}
T(\hat{\alpha} - 1) &= \frac{T\sum_{t=1}^T y_{t-1}^\circ u_t}{\sum_{t=1}^T (y_{t-1}^\circ)^2} \\
&= T\frac{\sum_{t=1}^T y_{t-1}^\circ u_t - \left(\sum_{t=1}^T y_{t-1}^\circ c_t'\right)\left(\sum_{t=1}^T c_t c_t'\right)^{-1}\sum_{t=1}^T c_t u_t}{\sum_{t=1}^T \left(y_{t-1}^\circ - \left(\sum_{t=1}^T y_{t-1}^\circ c_t'\right)\left(\sum_{t=1}^T c_t c_t'\right)^{-1} c_t\right)^2} \\
&= \frac{\frac{1}{T}\sum_{t=1}^T y_{t-1}^\circ u_t - \left(\sum_{t=1}^T \frac{y_{t-1}}{\sqrt{T}}^\circ c_t^{*\prime}\right)\left(\sum_{t=1}^T c_t^* c_t^{*\prime}\right)^{-1}\sum_{t=1}^T c_t^* \frac{u_t}{\sqrt{T}}}{\frac{1}{T}\sum_{t=1}^T \left(\frac{y_{t-1}^\circ}{\sqrt{T}} - \left(\sum_{t=1}^T \frac{y_{t-1}^\circ}{\sqrt{T}} c_t^{*\prime}\right)\left(\sum_{t=1}^T c_t^* c_t^{*\prime}\right)^{-1} c_t^*\right)^2} \\
&\to_d \frac{\int_0^1 \tilde{B}(r)\mathrm{d}\tilde{B}(r) + \lambda}{\int_0^1 \tilde{B}(r)^2 \mathrm{d}r}
\end{aligned}$$

where

$$\tilde{B}(r) = B(r) - \int_0^1 B(r)f(r)'\mathrm{d}r \left(\int_0^1 f(r)f(r)'\mathrm{d}r \right) f(r)$$

is the residual of the Hilbert space projection of B on f, and $\lambda = \sum_{k=1}^{\infty} \mathbb{E}(u_t u_{t-k})$.

The unit root test may be based on the above asymptotic result. There are nuissance parameters in the above result, which we need to estimate.

10.8 Unit Root Test in the Multivariate Case

This part is mainly based on Phillips and Durlauf (1986). Suppose that the multivariate time series $\{y_t\}$ where y_t is given by

$$y_t = y_{t-1} + u_t,$$

an invariance principle holds for $\{u_t\}$:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} u_t \to_d B(r),$$

where B is a Brownian motion with variance Xi, which is the long-run variance of the process $\{u_t\}$.

Now consider the regression

$$y_t = Ay_{t-1} + u_t$$

and estimate A by the usual OLS estimator for VAR(1) given as

$$\hat{A} = \left(\frac{1}{T} \sum_{t=1}^{T} y_t y_{t-1}' \right) \left(\frac{1}{T} \sum_{t=1}^{T} y_{t-1} y_{t-1}' \right)^{-1}.$$

It is straightforward to show that

$$T(\hat{A} - I) \to_d \left(\int_0^1 \mathrm{d}B(r)B(r)' + \Lambda \right) \left(\int_0^1 B(r)B(r)'\mathrm{d}r \right)^{-1}$$

where $\Lambda = \sum_{k=1}^{\infty} \Gamma(k)$ with $\Gamma(\cdot)$ being the autocovariance function of $\{u_t\}$.

We may also consider a symmetrized version of the estimator given by

$$\tilde{A} = \left(\frac{1}{2T} \sum_{t=1}^{T} (y_t y_{t-1}' + y_{t-1} y_t') \right) \left(\frac{1}{T} \sum_{t=1}^{T} y_{t-1} y_{t-1}' \right)^{-1}.$$

Using summation/integration by parts techniques as in the proof of Theorem 10.3, we may show that

$$T(\tilde{A} - I) \to_d \frac{1}{2} \left(B(1)B(1)' - \Sigma \right) \left(\int_0^1 B(r)B(r)'\mathrm{d}r \right)^{-1}$$

where $\Sigma = \mathbb{E}u_t u_t'$.

The usual Wald test statistics for the null hypothesis is given by

$$W = T \operatorname{tr}\left[(\tilde{A} - I)'\hat{\Sigma}^{-1}(\tilde{A} - I)\left(\frac{1}{T}\sum_{t=1}^{T} y_t y_t'\right)\right]$$
$$\to_d \frac{1}{4}\operatorname{tr}\left[(B(1)B(1)' - \Sigma)\Sigma^{-1}(B(1)B(1)' - \Sigma)\left(\int_0^1 B(r)B(r)'\mathrm{d}r\right)^{-1}\right],$$

where $\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \Delta y_t \Delta y_t'$ is a consistent estimator of Σ under the null. Note that the limit distribution of the Wald test statistic has nuissance parameters. Only in the case when $\{u_t\}$ is a white noise, the limit distribution can be reduced to one without nuissance parameter:

$$\frac{1}{4}\operatorname{tr}\left[(W(1)W(1)' - I)^2\left(\int_0^1 W(r)W(r)'\mathrm{d}r\right)^{-1}\right]$$

where W is the standard vector Brownian motion.

To accommodate the case of general $\{u_t\}$, we consider an consistent positive semi-definite estimator $\hat{\Xi}$ of Ξ. For a construction of such an estimator, see Phillips and Durlauf (1986). Then it is easy to see that the modified Wald test statistic

$$W_s = W - \frac{1}{4}\operatorname{tr}\left[\left(\frac{1}{T^2}y_T y_T'(\hat{\Sigma}^{-1} - \hat{\Xi}^{-1})y_T y_T' + (\hat{\Sigma} - \hat{\Xi})\right)\left(\frac{1}{T^2}\sum_{t=1}^{T} y_t y_t'\right)^{-1}\right]$$
$$\to_d \frac{1}{4}\operatorname{tr}\left[(W(1)W(1)' - I)^2\left(\int_0^1 W(r)W(r)'\mathrm{d}r\right)^{-1}\right].$$

Alternatively, we may use

$$G = \operatorname{tr}\left[\left(T(\tilde{A} - I) - \frac{1}{2}\left(\frac{1}{T}y_T y_T' - \hat{\Sigma}\right)\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}y_{t-1}'\right)^{-1}\right)'\right.$$
$$\left.\left(T(\tilde{A} - I) - \frac{1}{2}\left(\frac{1}{T}y_T y_T' - \hat{\Sigma}\right)\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{t-1}y_{t-1}'\right)^{-1}\right)\right] + \frac{1}{T}y_T'\hat{\Xi}^{-1}y_T$$
$$\to_d \chi^2_m$$

where m is the dimension of y_t. Note that under the null, the first part of G converges in probability to zero while the second part of G converges in distribution of $\operatorname{tr}(\mathbb{N}(0, I)^2)$.

10.9 General Unstable Autoregressive Process

Chan and Wei (1988) study the limit distribution of the OLS estimator of a general AR process

with roots on the unit circle. To be specific, they consider the process

$$\phi(L)y_t = \varepsilon_t$$

where $\{\varepsilon_t\}$ is a martingale difference sequence satisfying some moment conditions and $\phi(z)$ can be factorized as

$$\phi(z) = (1-z)^a(1+z)^b \prod_{k=1}^{l}(1 - 2\cos\theta_k z + z^2)^{d_k}\psi(z)$$

where $\psi(z)$ is a polynomial with roots outside the unit circle. It has been shown that the different components of the process that correspond to the 1 root, the -1 root, the $e^{\pm i\theta}$ roots, and the stationary roots, respectively, if properly normalized, converge in distribution to functionals of Brownian motions. Also, these different components are asymptotically independent from each other.

10.10 Fractionally Integrated Series

An integrated series $\{y_t\}$ may be written in the form of

$$(1-L)^d y_t = \Phi(L)\varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise process and $\Phi(L)$ is a lag operator polynomial with absolutely summable coefficients, and d is an integer which represents the order of integration. Granger and Joyeux (1980) and Hosking (1981) suggest to generalize the above model to non-integral value of $-1 < d < 1, d \neq 0$ by defining y_t through

$$y_t = (1-L)^{-d}\Phi(L)\varepsilon_t \tag{10.3}$$

where $(1-L)^{-d}$ is defined through the Taylor expansion of the polynomial $(1-z)^{-d}$ at around $z = 0$ given by

$$\begin{aligned}(1-z)^{-d} &= 1 + dz + \frac{d(d+1)}{2}z^2 + \frac{d(d+1)(d+2)}{6}z^3 + \cdots \\ &= 1 + \sum_{j=1}^{\infty}\frac{\Gamma(d+j)}{\Gamma(d)\Gamma(j+1)}z^j,\end{aligned}$$

and $\Gamma(\cdot)$ is the gamma function. Note that we have used the relationships that $j! = \Gamma(j+1)$ and that $d(d+1)\cdots(d+j-1) = \frac{\Gamma(d+j)}{\Gamma(d)}$, given that d and $d+j$ are not negative integers. The latter term is called a rising factorial.

Write $(1-z)^{-d} = \sum_{j=0}^{\infty} h_j z^j$. Using Stirling's approximation formula for the gamma function, which states that $\Gamma(z) = \sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z\left(1 + O(\frac{1}{z})\right)$, we can show that $h_j \sim (j+1)^{d-1}$ for j large, or $\lim_{j\to\infty}\frac{h_j}{(j+1)^{d-1}} = C$ for some constant C.

This above asymptotic approximation shows that if $d < \frac{1}{2}$, $\sum_{j=0}^{\infty} h_j^2 < \infty$. Therefore, (10.3) defines a weakly stationary series. Note that such a series is also causal if $\Phi(\cdot)$ is causal. In the

case of $d \geq \frac{1}{2}$, $\sum_{j=0}^{\infty} h_j^2$ diverges and the right hand side of (10.3) has infinite variance. Therefore, it cannot represent a weakly stationary time series.

When $-\frac{1}{2} < d < 0$, $(1-L)^d = \sum_{j=0}^{\infty} h_j' L^j$ has square summable coefficients. Therefore, if (10.3) represents a weakly stationary time series $\{y_t\}$ with $-\frac{1}{2} < d < 0$, then $\{y_t\}$ is invertible if $\Phi(\cdot)$ is invertible.

Now let $y_t = (1-L)^{-d} x_t$ for $-\frac{1}{2} < d < \frac{1}{2}$ and $\{x_t\}$ be a weakly stationary time series with spectral density f_x. Then we have

$$f_y(\lambda) = \left|1 - e^{i\lambda}\right|^{-2d} f_x(\lambda).$$

Taylor expansion shows that $\left|1 - e^{i\lambda}\right|^{-2d} \sim C\lambda^{-2d}$ as $\lambda \to 0$ for some constant C. If $d \in (0, \frac{1}{2})$ and u_t is a white noise, then the spectral density of y_t at $\lambda = 0$ is infinity. This corresponds to a stationary process that has a large low frequency component, or has a "long memory". If $d \in (-\frac{1}{2}, 0)$ and u_t is a white noise, then the spectral density of y_t at $\lambda = 0$ is zero and has derivative ∞. This corresponds to a stationary process that has almost no low frequency component, or has a "short memory", or "anti-persistent" in the terminology of Mandelbrot (1977).In contrast, a stationary ARMA process has spectral density converging to a constant C as the frequency λ converges to 0, where C is related to its long-run variance.

The fractionally integrated model can therefore be used to describe processes that has long memory or short memory. Note that the impulse response coefficients decay asymptotically at the hyperbolic speed $(j+1)^{d-1}$, as opposed to the ARMA case in which the decaying rate of the impulse response is geometric asymptotically. In general, if $(1-L)^d y_t$ is an ARMA(p, q) process, then we call y_t an ARIMA(p, d, q) process in the terminology of Hosking (1981), or more explicitly, an ARFIMA(p, d, q) process.

In practice, the covariance structure of a fractionally integrated process can be well approximated by a large order ARMA process. However, fractionally integrated model can serve as a parsimonious model for series that has slow decaying multipliers.

In the case of $d > \frac{1}{2}$ or $d < -\frac{1}{2}$, we may difference or sum the original series so that the resulting series fits a fractionally integrated model of order between $-\frac{1}{2}$ and $\frac{1}{2}$.

Granger (1980) shows that long memory processes can arise from aggregating many individual AR(1) series whose AR coefficients follow a Beta distribution.

10.11 Explosive Roots

White (1958) consider the asymptotic distribution of the estimator $\hat{\alpha}$ under the explosive root assumption that $|\alpha| > 1$. Using moment generating function technique, the author shows that the limit distribution of

$$\frac{|\alpha|^T}{\alpha^2 - 1}(\hat{\alpha} - \alpha)$$

is the standard Cauchy distribution. White (1959) shows that the corresponding distribution of

the t-statistic, as opposed to the unit root case, is asymptotically normal. Anderson (1959) and Rao (1961) consider the case of higher order autoregressive processes with explosive roots.

11 Cointegration

11.1 Cointegration

Consider the regression

$$y_t = x_t'\beta + u_t$$

in which x_t and y_t are $I(1)$, but u_t is $I(0)$. In this case, we say that y_t and x_t are cointegrated, and the cointegration relationship is given by $(-1, \beta')'$. The concept of cointegration was introduced in Granger (1981) and Engle and Granger (1987).

Suppose that $w_t = (u_t, \Delta x_t')'$ satisfies an invariance principle to a vector Brownian motion $(B_1, B_2')'$ where B_1 is one dimensional, and the dimension of B_2 is the same as that of x_t. Let $\hat{\beta}$ be the OLS estimator of β. We may easily derive that

$$T(\hat{\beta} - \beta) \to_d \left(\int_0^1 B_2(r) B_2(r) \mathrm{d}r \right)^{-1} \left(\int_0^1 B_2(r) \mathrm{d}B_1(r) + \Lambda_{21}^{\circ} \right) \tag{11.1}$$

where $\sum_{k=0}^{\infty} \Delta x_t u_{t-k}$.

We note here that the above result shows that $\hat{\beta} \to_p \beta$ even in the case where $\Lambda_{21}^{\circ} \neq 0$, i.e., in the case where x_t and u_t are correlated. This is very different to the regression in the stationary setting in which endogeneity will lead to inconsistent OLS estimate for the regression coefficient. See Phillips and Hansen (1990) for more discussion.

Phillips and Park (1988) show that in the case when $\{u_t\}$ and $\{\Delta x_t\}$ are independent and u_t follows an AR process, the OLS estimator $\hat{\beta}$ is efficient in the sense that it is asymptotically equivalent to the GLS estimator.

Now we consider cointegration regression with additional regressors:

$$y_t = z_t'\gamma + x_t'\beta + u_t.$$

In the case when z_t is $I(0)$, using two step regression and similar arguments as in Section 10.6, it is easy to show that the OLS estimator of β converges in distribution to the same distribution as if there were not the additional $I(0)$ term. If $z_t = c_t$, a deterministic sequence as in Section 10.7, then a similar argument show that the OLS estimator $\tilde{\beta}$ for β has asymptotics given by

$$T(\hat{\beta} - \beta) \to_d \left(\int_0^1 \tilde{B}_2(r) \tilde{B}_2(r) \mathrm{d}r \right)^{-1} \left(\int_0^1 \tilde{B}_2(r) \mathrm{d}B_1(r) + \Lambda_{21}^{\circ} \right)$$

where

$$\tilde{B}(r) = B(r) - \int_0^1 B(r) f(r)' \mathrm{d}r \left(\int_0^1 f(r) f(r)' \mathrm{d}r \right) f(r).$$

For a detailed discussion of the general cases, see, e.g., Phillips and Durlauf (1986) and Park and Phillips (1988). Park and Phillips (1989) consider the case in which there are regressors that are

integrated of higher orders.

11.2 Spurious Regression

Let y_t and x_t be $I(1)$, and let $w_t = (\Delta y_t, \Delta x_t')'$ satisfy an invariance principle. Suppose that y_t is not integrated with x_t, i.e., for any choice of β, $e_t = y_t - x_t'\beta$ is $I(1)$. If we run an OLS regression on

$$y_t = x_t'\beta + e_t$$

and obtain the OLS estimator $\hat{\beta}$ of β, we have

$$\hat{\beta} = \left(\frac{1}{T^2}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T^2}\sum_{t=1}^{T} x_t y_t\right) \to_d \left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1}\left(\int_0^1 B_2(r)B_1(r)\mathrm{d}r\right).$$

Since $\hat{\beta}$ converges to something random, $\hat{\beta}$ cannot be consistent.

Note that in the spurious regression setting,

$$\begin{aligned}
\frac{1}{T}\hat{\sigma}^2 &= \frac{1}{T^2}\sum_{t=1}^{T}(y_t - x_t'\hat{\beta})^2 \\
&= \frac{1}{T^2}\sum_{t=1}^{T} y_t^2 - \frac{1}{T^2}\sum_{t=1}^{T}\hat{\beta}' x_t x_t'\hat{\beta} \\
&\to_d \int_0^1 B_1(r)^2\mathrm{d}r - \left(\int_0^1 B_1(r)B_2(r)'\mathrm{d}r\right)\left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1}\left(\int_0^1 B_2(r)B_1(r)\mathrm{d}r\right) \\
&= \int_0^1 \tilde{B}_1(r)^2\mathrm{d}r
\end{aligned}$$

where

$$\tilde{B}_1(r) = B_1(r) - B_2(r)'\left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1}\left(\int_0^1 B_2(r)B_1(r)\mathrm{d}r\right).$$

Then for the t-statistic under the null hypothesis $\beta_i = 0$, we have

$$\begin{aligned}
\frac{1}{\sqrt{T}}t(\beta_i) &= \frac{\hat{\beta}}{\sqrt{T}\hat{\sigma}\left[\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1/2}\right]_{ii}} \\
&\to_d \frac{\left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1}\left(\int_0^1 B_2(r)B_1(r)\mathrm{d}r\right)}{\left(\int_0^1 \tilde{B}_1(r)^2\mathrm{d}r\right)^{1/2}\left[\left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1/2}\right]_{ii}}.
\end{aligned}$$

This shows that the t-statistic $t(\beta_i)$ diverges to infinity as T goes to infinity. That is, we will always reject the null that $\beta_i = 0$ when the sample size is large enough.

Similarly, we may want to use the F-test to test for the joint significance of β. However,

$$\begin{aligned}\frac{1}{T}F(\beta) &= \frac{\sum_{t=1}^T y_t x_t' \left(\sum_{t=1}^T x_t x_t'\right)^{-1} \sum_{t=1}^T x_t y_t/m}{\sum_{t=1}^T y_t^2 - \sum_{t=1}^T y_t x_t' \left(\sum_{t=1}^T x_t x_t'\right)^{-1} \sum_{t=1}^T x_t y_t} \\ &\to_d \frac{\int_0^1 B_1(r)B_2(r)'\mathrm{d}r \left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1} \int_0^1 B_2(r)B_1(r)'\mathrm{d}r/m}{\int_0^1 B_1(r)^2\mathrm{d}r - \int_0^1 B_1(r)B_2(r)'\mathrm{d}r \left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1} \int_0^1 B_2(r)B_1(r)'\mathrm{d}r}.\end{aligned}$$

This shows that $F(\beta)$ diverges to infinity as T goes to infinity. That is, we will always reject the null that $\beta = 0$ when the sample size is large enough, even though y_t and x_t are not connected in any meaningful way.

Also,

$$\begin{aligned}R^2 &= \frac{\sum_{t=1}^T y_t x_t' \left(\sum_{t=1}^T x_t x_t'\right)^{-1} \sum_{t=1}^T x_t y_t}{\sum_{t=1}^T y_t^2} \\ &= \frac{\int_0^1 B_1(r)B_2(r)'\mathrm{d}r \left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1} \int_0^1 B_2(r)B_1(r)'\mathrm{d}r}{\int_0^1 B_1(r)^2\mathrm{d}r},\end{aligned}$$

which remains random as $T \to \infty$. Usually its value is very close to one.

Lastly, consider the Durbin-Watson statistic

$$DW = \frac{\sum_{t=1}^T (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^T \hat{e}_t^2}$$

where

$$\hat{e}_t = y_t - \left(\sum_{t=1}^T y_t x_t'\right)\left(\sum_{t=1}^T x_t x_t'\right)^{-1} x_t$$

and

$$\hat{e}_t - \hat{e}_{t-1} = \Delta y_t - \left(\sum_{t=1}^T y_t x_t'\right)\left(\sum_{t=1}^T x_t x_t'\right)^{-1} \Delta x_t.$$

It can be easily seen that $\frac{1}{T^2}\sum_{t=1}^T \hat{e}_t^2 = O_p(1)$, while $\frac{1}{T}\sum_{t=1}^T (\Delta\hat{e}_t)^2 = O_p(1)$. Note that

$$\frac{1}{T}\sum_{t=1}^T \left([1, -\hat{\beta}']w_t\right)^2 \to_d \eta'\Sigma_w\eta$$

where

$$\eta = \begin{bmatrix} 1 \\ -\left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1}\left(\int_0^1 B_2(r)B_1(r)\mathrm{d}r\right) \end{bmatrix}$$

and $\Sigma_w = \mathbb{E} w_t w_t'$. Then

$$T \cdot DW \to_d \frac{\eta' \Sigma_w \eta}{\int_0^1 \tilde{B}_1(r)^2 \mathrm{d}r}.$$

This shows that as sample size gets large, the Durbin-Watson statistic gets close to zero, indicating that there is strong positive first-order serial correlation in the residuals, which is something well expected in the spurious regression setting.

The following is quoted from Granger and Newbold (1974), the paper that introduced the concept of spurious regressions.

> *It is very common to see reported in applied econometric literature time series regression equations with an apparently high degree of fit, as measured by the coefficient of multiple correlation R^2 or the corrected coefficient $\bar{R}^2$, but with an extremely low value for the Durbin-Watson statistic. We find it very curious that whereas virtually every textbook on econometric methodology contains explicit warnings of the dangers of autocorrelated errors, this phenomenon crops up so frequently in well-respected applied work.*
>
> *...*
>
> *There are, in fact, as is well-known, three major consequences of autocorrelated errors in regression analysis:*
>
> *(i) Estimates of the regression coefficients are inefficient.*
>
> *(ii) Forecasts based on the regression equations are sub-optimal.*
>
> *(iii) The usual significance tests on the coefficients are invalid.*

In such situations, Granger and Newbold (1974) propose to difference the time series before running regressions. See also Plosser and Schwert (1978) for a discussion of possible effects of underdifferencing and over-differencing. To be specific, if the variables are under-differenced, we have the spurious regression problem. If the variables in the regression is over-differenced, the coefficient estimator is still consistent, but inefficient, and the inference based on the usual t-statistic is problematic. The reason is that when the model is over-differenced, the new error term, which is the original error term first differenced, is autocorrelated. In this situation, we need to use HAC standard errors if we would like to conduct inference based on OLS estimator, or use GLS estimator to achieve efficiency. For the situation of spurious regression when an intercept term is included in the regression, see Phillips (1986). Also, note that in the case when y_t and x_t are cointegrated, the long-run variance matrix of w_t is singular, and the asymptotics in this section does not hold anymore. The asymptotics are given in the previous section.

11.3 Testing for Cointegration

We may test for cointegration between y_x and x_t by applying the unit root test to the OLS residuals $\{\hat{e}_t\}$ of the regression

$$y_t = x_t' \beta + e_t.$$

If y_t and x_t are cointegrated, there exists β such that e_t is $I(0)$. Otherwise, e_t contains a unit root for any choice of β. We run the augmented Dickey-Fuller test on $\hat{e}_t$ through the regression

$$\hat{e}_t = \alpha \hat{e}_{t-1} + \sum_{k=1}^{p} \alpha_k \Delta \hat{e}_{t-k} + \varepsilon_t.$$

Under the null hypothesis that y_t and x_t are *not* cointegrated, $\alpha = 1$

Suppose that $(\Delta y_t, \Delta x_t')'$ satisfies an invariance principle such that

$$\begin{bmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} \Delta y_t \\ \frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} \Delta x_t \end{bmatrix} \to_d \begin{bmatrix} B_1(r) \\ B_2(r) \end{bmatrix}$$

with the long run covariance matrix

$$\begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \Omega_{22} \end{bmatrix}$$

where the long run covariance matrix is appropriately partitioned so that ω_{11} is one-by-one dimensional. It is straightforward to show that

$$\frac{1}{\sqrt{T}} \hat{e}_{[Tr]} \to_d B_1(r) - \left(\int_0^1 B_1(r) B_2(r)' \mathrm{d}r \right) \left(\int_0^1 B_2(r) B_2(r)' \mathrm{d}r \right)^{-1} B_2(r) := \tilde{B}.$$

Then it is easy to follow Section 10.6 to establish that

$$t(\alpha) = \frac{T(\hat{\alpha} - 1)}{\hat{\sigma} \left(\frac{1}{T^2} \sum_{t=1}^{T} \hat{e}_{t-1}^2 \right)^{-1/2}} \to_d \frac{\int_0^1 W(r) \mathrm{d}W(r)}{\left(\int_0^1 W(r)^2 \mathrm{d}r \right)^{1/2}}$$

where $\hat{\sigma}^2$ is a consistent estimator of the variance of ε_t.

However, it is not straightforward to get/simulate the distribution of $\tilde{B}$ since B_1 and B_2 are correlated. Therefore, we write

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \omega_{11}^{1/2} \sqrt{1 - (\omega_{12} \Omega_{22}^{-1} \omega_{21} / \omega_{11})^2} & \omega_{12} \Omega_{22}^{-1/2} \\ 0 & \Omega_{22}^{1/2} \end{bmatrix} \begin{bmatrix} W_{11} \\ W_{22} \end{bmatrix}$$

where W_1 and W_2 are two independent Brownian motions. Then we may show that

$$\tilde{B}(r) = \omega_{11}^{1/2} \sqrt{1 - (\omega_{12} \Omega_{22}^{-1} \omega_{21} / \omega_{11})^2} \tilde{W}(r)$$

where

$$\tilde{W}(r) = W_1(r) - \left(\int_0^1 W_1(r) W_2(r)' \mathrm{d}r \right) \left(\int_0^1 W_2(r) W_2(r)' \mathrm{d}r \right)^{-1} W_2(r).$$

Then

$$t(\alpha) \to_d \frac{\int_0^1 \tilde{W}(r)\mathrm{d}\tilde{W}(r)}{\left(\int_0^1 \tilde{W}(r)^2\mathrm{d}r\right)^{1/2}}.$$

Phillips and Ouliaris (1988) propose another test for cointegration based on the eigenvalues of the long-run variance estimator of $(\Delta x_t, \Delta y_t)$. The idea is that if x_t and y_t are cointegrated, then the long-run variance matrix of its innovations should be singular.

11.4 Inference in Cointegrated Models

The problem with statistical inference in cointegrated models is that the limit distribution in (11.1) is non-standard, and the distributions of the conventional test statistics contains nuisance parameters. We introduce in this section two popular approaches that can be used to conduct inference in cointegrated models.

11.4.1 Phillips and Hansen's Fully Modified OLS

Phillips and Hansen (1990) modifies the OLS estimator in the cointegration regression. The fully modified OLS estimator (FM-OLS) is

$$\hat{\beta}_{\text{FM-OLS}} = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t y_t^+ - \Lambda^+\right)$$

where

$$y_t^+ = y_t - \omega_{12}\Omega_{22}^{-1}\Delta x_t,$$

and

$$\Lambda^+ = \Lambda_{21}^\circ - \Lambda_{22}^\circ\Omega_{22}^{-1}\omega_{21},$$

following the notations introduced in the earlier sections of this chapter.

Theorem 11.1. *We have that*

$$T(\hat{\beta}_{FM\text{-}OLS} - \beta) \to_d \left(\int_0^1 B_2(r)B_2(r)'\mathrm{d}r\right)^{-1}\int_0^1 B_2(r)\mathrm{d}B_{1\cdot 2}(r)$$

where $B_{1\cdot 2} = B_1 - \omega_{12}\Omega_{22}^{-1}B_2$ *is independent of* B_2.

Proof. Note

$$\begin{aligned} T(\hat{\beta}_{\text{FM-OLS}} - \beta) &= \left(\frac{1}{T^2}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t(u_t - \omega_{12}\Omega_{22}^{-1}\Delta x_t) - \Lambda^+\right) \\ &= \left(\frac{1}{T^2}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t w_t'\begin{bmatrix}1\\ -\Omega_{22}^{-1}\end{bmatrix} - \Lambda^+\right) \end{aligned}$$

. The result follows immediately by noting

$$\frac{1}{T^2}\sum_{t=1}^{T} x_t x_t' \to_d \int_0^1 B_2(r)B_2(r)'\mathrm{d}r$$

and

$$\frac{1}{T}\sum_{t=1}^{T} x_t w_t' \to_d \begin{bmatrix} \int_0^1 B_2(r)\mathrm{d}B_1(r) + \Lambda_{21}^{\circ} \\ \int_0^1 B_2(r)\mathrm{d}B_2(r) + \Lambda_{22}^{\circ} \end{bmatrix}.$$

■

11.4.2 Park's Canonical Cointegrating Regression

Instead of modifying the OLS estimator, Park (1992) transforms the data. Define

$$y_t^* = y_t - \beta'\Lambda_2^{\circ\prime}\Sigma^{-1}w_t - \omega_{12}\Omega_{22}^{-1}\Delta x_t$$

and

$$x_t^* = x_t - \Lambda_2^{\circ\prime}\Sigma^{-1}w_t$$

where $\Sigma = \mathbb{E}u_t u_t'$, Λ_2° is Λ° partitioned as $\Lambda^{\circ} = [\Lambda_1^{\circ}\ \Lambda_2^{\circ}]$, and the rest notations are as earlier. The canonical cointegrating regression (CCR) estimator $\hat{\beta}_{\text{CCR}}$, is defined to be the OLS estimator of the regressing y_t^* on x_t^*. The CCR estimator utilizes the fact that the cointegration relationship between two $I(1)$ variables is unchanged by adding $I(0)$ (i.e., stationary) components to the two variables. The CCR error term $u_t^* = y_t^* - x_t^*\beta = u_t - \omega_{12}\Omega_{22}^{-1}\Delta x_t$, is asymptotically independent of $\{\Delta x_t\}$ or $\{\Delta x_t^*\}$.

It can be easily proved (following the proof of asymptotic distribution of the FM-OLS estimator) that $\hat{\beta}_{\text{CCR}}$ has exactly the same asymptotic distribution as that of $\hat{\beta}_{\text{FM-OLS}}$.

In *feasible* FM-OLS and CCR estimator, the omega's and lambda's are replaced with their consistent (and possibly nonparametric) estimators.

11.4.3 The Wald Test

To test for null hypothesis $H_0 : R\beta = r$ against the alternative $R\beta \neq r$, we apply the Wald statistic

$$W = \frac{(R\hat{\beta} - r)'\left(R(\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1})R'\right)^{-1}(R\hat{\beta} - r)}{\omega_{11} - \omega_{12}\Omega_{22}^{-1}\omega_{21}}$$

where $\hat{\beta}$ is either $\hat{\beta}_{\text{FM-OLS}}$ or $\hat{\beta}_{\text{CCR}}$.

Theorem 11.2. *Under the null, $W \to_d \chi_q^2$ where q is the number of restrictions.*

Proof. The result follows from that under the null

$$T(R\hat{\beta} - \beta) = TR(\hat{\beta} - \beta)$$

and that

$$\left(R\left(\int_0^1 B_2(r)B_2(r)\mathrm{d}r\right)^{-1}R'\right)^{-1/2}R\int_0^1 B_2(r)\mathrm{d}B_{1\cdot 2} =_d \mathbb{N}(0,\omega_{11}-\omega_{12}\Omega_{22}^{-1}\omega_{21}).$$

Note that if $A(r)$ is a deterministic process, then $\int_0^1 A\mathrm{d}B =_d \mathbb{N}\left(0,\int_0^1 A(r)\Xi A'(r)\mathrm{d}r\right)$ if B is a Brownian motion with covariance Ξ. Since B_2 is independent of $B_{1\cdot 2}$, the distribution of the above term conditioning on B_2 is $\mathbb{N}(0,\omega_{11}-\omega_{12}\Omega_{22}^{-1}\omega_{21})$. Therefore, this is also the unconditional distribution. ■

11.5 Cointegrated VAR and the Error Correction Models

We consider a vector autoregressive system in which variables contain unit roots and are possibly cointegrated. To be specific, let $\{y_t\}$ be an r-dimensional VAR(p) process given by

$$\Phi(L)y_t = \varepsilon_t$$

where $\varepsilon_t \sim \mathrm{WN}(0,\Sigma)$ and

$$\Phi(z) = I - \Phi_1 z - \cdots - \Phi_p z^p.$$

Instead of assuming that all roots of $\det\Phi(z)=0$ lie outside the unit circle, we introduce unit roots in this system, and assume that there are m roots that are one, $0<m\le r$, and all the other roots are outside the unit circle. Since $z=1$ is a root of $\det\Phi(z)=0$, $\Phi(1)$ is a reduced rank matrix and we assume that $\operatorname{rank}\Phi(1)=\ell$.

We may write

$$\Phi(z) = -z\Phi(1) + (1-z)\Gamma(z)$$

where

$$\Gamma(z) = I - \Phi_2 z - \cdots - \Phi_p z^{p-1}.$$

This is the VAR written in the error correction form, which was first studied in Granger and Weiss (1983) and Engle and Granger (1987).

References

Alj, A., Azrak, R., and Mélard, G. (2014). On conditions in central limit theorems for martingale difference arrays. *Economics Letters*, 123(3):305–307.

Anderson, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Annals of Mathematical Statistics*, 30(3):676–687.

Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, 4(3):458–467.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Beveridge, S. and Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary Economics*, 7:151–174.

Bhargava, A. (1986). On the theory of testing for unit roots in observed time series. *Review of Economic Studies*, 53(3):369–384.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York, 3rd edition.

Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley & Sons, New York, 2nd edition.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

Bradley, R. C. (2005). Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, 2:107–144.

Brockwell, P. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, 2nd edition.

Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annal of Statistics*, 16(1):367–401.

Chung, K. L. (2001). *A Course in Probability Theory*. Academic Press, 3rd edition.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49(4):1057–1072.

Dufour, J.-M., Pelletier, D., and Renault, É. (2006). Short run and long run causality in time series: inference. *Journal of Econometrics*, 132(2):337–362.

Dufour, J.-M. and Renault, E. (1998). Short run and long run causality in time series: Theory. *Econometrica*, 66(5):1099–1125.

Eberlein, E. (1986). On strong invariance principles under dependence assumptions. *Annals of Probability*, 14(1):260–270.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50:987–1007.

Engle, R. F. and Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Engle, R. F., Lilien, D. M., and Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, 55:391–407.

Evans, G. B. A. and Savin, N. E. (1981). Testing for unit roots: 1. *Econometrica*, 49(3):753–779.

Evans, G. B. A. and Savin, N. E. (1984). Testing for unit roots: 2. *Econometrica*, 52(5):1241–1269.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Phsical and Engineering Science*, 222:309–368.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*. Wiley, 2nd edition.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

Gonçalves, S. and Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 123(1):89–120.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14:227–238.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16:121–130.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-momory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29.

Granger, C. W. J. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120.

Granger, C. W. J. and Weiss, A. A. (1983). Time series analysis of error-correction models. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, pages 255–278. Academic Press.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy*, 86(6):971–987.

Halmos, P. R. (2006). *Lectures on Ergodic Theory*. AMS Chelsea Publishing, chelsea edition.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hannan, E. J. (1970). *Multiple Time Series*. John Wiley & Sons.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *The Annals of Statistics*, 8(5):1071–1081.

Hausman, J. and Palmer, C. (2012). Heteroskedasticity-robust inference in finite samples. *Economics Letters*, 116:232–235.

Herrndorf, N. (1983). The invariance principle for ϕ-mixing sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 63:97–108.

Herrndorf, N. (1984a). A functional central limit theorem for ρ-mixing sequences. *Journal of Multivariate Analysis*, 15:141–146.

Herrndorf, N. (1984b). A functional central limit theorem for weakly dependent sequences of random variables. *Annals of Probability*, 12(1):141–153.

Herrndorf, N. (1985). A functional central limit theorem for strongly mixing sequences of random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 69:541–550.

Hildebrandt, T. H. (1942). Remarks on the abel-dini theorem. *American Mathematical Monthly*, 49(7):441–445.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, 68(1):165–176.

Inoue, A. and Kilian, L. (2020). The uniform validity of impulse response inference in autoregressions. *Journal of Econometrics*, 215(2):450–472.

Johansen, S. and Nielsen, B. (2019). Boundedness of M-estimators linear regression in time series. *Econometric Theory*, 35(3):653–683.

Karatzas, I. and Shreve, S. E. (2000). *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition.

Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics*, 80(2):218–230.

Kilian, L. (1999). Finite-sample properties of percentile and percentile-t bootstrap confidence intervals for impulse responses. *Review of Economics and Statistics*, 81(4):652–660.

Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *Annals of Probability*, 8(6):1003–1036.

Lütkepohl, H. (1990). Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *The Review of Economics and Statistics*, 72(1):116–125.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In Chen, X. and Swanson, N. R., editors, *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pages 437–461. Springer, New York.

Mandelbrot, B. B. (1977). *Fractals: Form, Chance and Dimension*. Freeman.

Marsh, T. A. and Merton, R. C. (1986). Dividend variability and variance bounds tests for the rationality of stock market prices. *American Economic Review*, 76(3):483–498.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability*, 2(4):620–628.

McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *Annals of Probability*, 3(5):829–839.

McLeish, D. L. (1977). On the invariance principle for nonstationary mixingales. *Annals of Probability*, 5(4):616–621.

Nelson, C. R. and Plosser, C. I. (1982). Trends and random walks in macroeconomic time series. *Journal of Monetary Economics*, 10(2):139–162.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2111–2245. Elsevier Science.

Park, J. Y. (1992). Canonical cointegrating regressions. *Econometrica*, 60(1):119–143.

Park, J. Y. and Phillips, P. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4:468–497.

Park, J. Y. and Phillips, P. (1989). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 5:95–131.

Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *Annals of Mathematical Statistics*, 28(2):329–348.

Peligrad, M. (1985). An invariance principle for ϕ-mixing sequences. *Annals of Probability*, 13(4):1304–1313.

Phillips, P. and Durlauf, S. (1986). Multiple time series regression with integrated processes. *Review of Economic Studies*, 53(4):473–495.

Phillips, P. C. B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics*, 33:311–340.

Phillips, P. C. B. (1987a). Asymptotic expansions in nonstationary vector autoregressions. *Econometric Theory*, 3(1):45–68.

Phillips, P. C. B. (1987b). Time series regression with a unit root. *Econometrica*, 55(2):277–301.

Phillips, P. C. B. (1988a). Weak convergence of sample covariance matrices to stochastic integrals via martingale approximations. *Econometric Theory*, 4:528–533.

Phillips, P. C. B. (1988b). Weak convergence to the matrix stochatic integral $\int_0^1 BdB'$. *Journal of Multivariate Analysis*, 24:252–264.

Phillips, P. C. B. and Hansen, B. E. (1990). Statistical inference in instrumental variables regression with i(1) processes. *Review of Economic Studies*, 57(1):99–125.

Phillips, P. C. B. and Ouliaris, S. (1988). Testing for cointegration using principal components methods. *Journal of Economic Dynamics and Control*, 12(2-3):205–230.

Phillips, P. C. B. and Park, J. Y. (1988). Asymptotic equivalence of ordinary least squares and generalized least squares in regressions with integrated regressors. *Journal of the American Statistical Association*, 83(401):111–115.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Phillips, P. C. B. and Solo, V. (1992). Asymptotics for linear processes. *The Annals of Statistics*, 20(2):971–1001.

Plosser, C. I. and Schwert, G. W. (1978). Money, income, and sunspots: Measuring economic relationships and the effects of differencing. *Journal of Monetary Economics*, 4:637–660.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*, volume I. Academic Press.

Rao, M. M. (1961). Consistency and limit distributions of estimators of parameters in explosive stochastic difference equations. *Annals of Mathematical Statistics*, 32(1):195–218.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1):43–47.

Rosenblatt, M. (1984). Asymptotic normality, strong mixing and spectral density estimates. *Annals of Probability*, 12(4):1167–1180.

Rudin, W. (1987). *Real and Complex Analysis*. McGraw-Hill Companies, Inc., 3rd edition.

Runkle, D. E. (1987). Vector autoregressions and reality. *Journal of Business & Economic Statistics*, 5(4):437–442.

Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.

Sargan, J. D. and Bhargava, A. (1983). Testing residuals from least squares regression for being generated by the gaussian random walk. *Econometrica*, 51(1):153–174.

Shiryaev, A. N. (1989). *Probability*. Springer, 2nd edition.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sims, C. A. and Zha, T. (1999). Error bands for impulse responses. *Econometrica*, 67(5):1113–1155.

Solo, V. (1984). The order of differencing in arima models. *Journal of the American Statistical Association*, 79(388):916–921.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, volume 2A, chapter 8, pages 415–525. Elsevier.

Tsay, R. S. (2010). *Analysis of Financial Time Series*. Wiley, 3rd edition.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

White, H. (2001). *Asymptotic Theory for Econometricians*. Academic Press, revised edition.

White, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Annals of Mathematical Statistics*, 29(4):1188–1197.

White, J. S. (1959). The limiting distribution of the serial correlation coefficient in the explosive case ii. *Annals of Mathematical Statistics*, 30(3):831–834.

Wittmann, R. (1985). A general law of iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 68:521–543.

Wittmann, R. (1987). Sufficient moment and truncated moment conditions for the law of the iterated logarithm. *Probability Theory and Related Fields*, 75(4):509–530.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge.