LECTURE ON

BASIC TIME SERIES MODELS

CHUNG-MING KUAN

Institute of Economics
Academia Sinica

This version: July 16, 2005

Address for correspondence: Institute of Economics, Academia Sinica, Taipei 115, Taiwan.
E-mail: ckuan@econ.sinica.edu.tw; URL: www.sinica.edu.tw/~ckuan

Contents

1 Introduction

Time series models are designed to capture various characteristics in time series data. These models have been widely used in many disciplines in the science. In particular, it has been found that time series models are very useful in analyzing economic and financial data. This motivates more and more econometricians and statisticians to devote themselves to the development of new (or refined) time series models and methods. This note serves as an introduction to basic time series models and related issues in model estimation and hypothesis testing. This note does not try to cover all the topics in the time series analysis; readers are referred to other time series textbooks for more detailed discussion of the topics not covered by this note. I acknowledge that many results of this note are taken freely from some textbooks, such as Brockwell and Davis (1987), Hamilton (1994), Fuller (1996), and Tsay (2002).

2 Basic Concepts

2.1 Weak and Strict Stationarity

Let Y denote a random element in some probability space such that for each t, Y_t is a random variable, and for a random outcome ω in this probability space, $Y(\omega)$ is a function of t. Let y_t denote an observation of Y_t. The collection $\{Y_t\}$ is usually referred to as a *stochastic process* or a *time series* with $\{y_t\}$ as its realization (time path). For notational convenience, we will not distinguish between Y_t and y_t; hence $\{y_t\}$ may denote a time series or its realization.

A time series $\{y_t\}$ is said to be *weakly stationary* or *covariance stationary* if $\mathbb{E}(y_t) = \mu$, and *autocovariances*

$$\mathbb{E}[(y_t - \mu)(y_{t-j} - \mu)] = \gamma_j, \qquad j = 0, \pm 1, \pm 2, \ldots,$$

depend on j but not on t. As $\gamma_0 = \mathrm{var}(y_t)$, the *autocorrelations* of y_t are

$$\rho_j = \gamma_j/\gamma_0, \qquad j = 0, \pm 1, \pm 2, \ldots,$$

which are also independent of t. Clearly, a weakly stationary series has $\rho_j = \rho_{-j}$. In particular, a series with zero mean, constant variance, and zero autocorrelations is called a *white noise*.

The finite dimensional distributions of a time series $\{y_t\}$ are the joint distribution functions of $y_{t_1}, y_{t_2}, \ldots, y_{t_n}$ for any finite collection of $t_1, t_2, \ldots, t_n$. A time series $\{y_t\}$ is said to be *strictly stationary* if its finite dimensional distributions are invariant under

time displacements, i.e., for each s,

$$F_{t_1,\ldots,t_n}(c_1,\ldots,c_n) = F_{t_1+s,\ldots,t_n+s}(c_1,\ldots,c_n).$$

A sequence of i.i.d. random variables is strictly stationary, but the converse need not hold. Note that strict stationary imposes no restriction on moments. When a strict stationary series has a finite second moment, it must be weakly stationary. A time series $\{y_t\}$ is Gaussian if its finite dimensional distribution functions are all Gaussian. A white noise with Gaussian marginal distributions is a Gaussian series which is also a sequence of i.i.d. normal random variables. Hence, a Gaussian white noise is strictly and weakly stationary. A sequence of i.i.d. Cauchy random variables is strictly stationary but not weakly stationary.

2.2 Difference Equations

Given the first-order *difference equation*:

$$y_t = \psi_1 y_{t-1} + u_t, \qquad t = 0, 1, 2, \ldots,$$

recursive substitution yields

$$y_t = \psi_1^{t+1} y_{-1} + \psi_1^t u_0 + \psi_1^{t-1} u_1 + \cdots + \psi_1 u_{t-1} + u_t.$$

Similarly,

$$y_{t+j} = \psi_1^{j+1} y_{t-1} + \psi_1^j u_t + \psi_1^{j-1} u_{t+1} + \cdots + \psi_1 u_{t+j-1} + u_{t+j}.$$

Define the *impulse response* (*dynamic multiplier*) of the future observation y_{t+j} to the effect of one unit change of u_t as $\partial y_{t+j}/\partial u_t$. In this case,

$$\frac{\partial y_{t+j}}{\partial u_t} = \psi_1^j,$$

which depends only on j but not on t. A dynamic system is said to be *stable* if its impulse response eventually vanishes as j tends to infinity; a dynamic system is *explosive* if its impulse response diverges as j increases. Clearly, the first-order difference equation is stable (explosive) when $|\psi_1| < 1$ ($|\psi_1| > 1$), and the impulse response converges to zero (diverges) exponentially fast. Only when $\psi_1 = 1$ will a given change of u_t have a constant effect on all future observations.

Consider now the p^{th}-order difference equation:

$$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p} + u_t.$$

To determine whether a p^{th}-order difference equation is stable or not, we write this equation as a first-order vector difference equation:

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}}_{\boldsymbol{\eta}_t} = \underbrace{\begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_{p-1} & \psi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}}_{\boldsymbol{F}} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{bmatrix}}_{\boldsymbol{\eta}_{t-1}} + \underbrace{\begin{bmatrix} u_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\boldsymbol{\nu}_t};$$

that is $\boldsymbol{\eta}_t = \boldsymbol{F}\boldsymbol{\eta}_{t-1} + \boldsymbol{\nu}_t$. Thus,

$$\boldsymbol{\eta}_{t+j} = \boldsymbol{F}^{j+1}\boldsymbol{\eta}_{t-1} + \boldsymbol{F}^{j}\boldsymbol{\nu}_t + \boldsymbol{F}^{j-1}\boldsymbol{\nu}_{t+1} + \cdots + \boldsymbol{F}\boldsymbol{\nu}_{t+j-1} + \boldsymbol{\nu}_{t+j}.$$

The impulse response of $\boldsymbol{\eta}$ is

$$\nabla_{\boldsymbol{\nu}_t}\boldsymbol{\eta}_{t+j} = \boldsymbol{F}^j.$$

Let f_{mn}^t denote the $(m,n)^{\text{th}}$ element of $\boldsymbol{F}^t$. It is straightforward to verify that

$$y_{t+j} = f_{11}^{j+1}y_{t-1} + f_{12}^{j+1}y_{t-2} + \cdots + f_{1p}^{j+1}y_{t-p} + \sum_{i=0}^{j} f_{11}^{i}u_{t+j-i}.$$

The impulse response of y_{t+j} is thus

$$\frac{\partial y_{t+j}}{\partial u_t} = f_{11}^j,$$

the $(1,1)^{\text{th}}$ element of $\boldsymbol{F}^j$.

Recall that the eigenvalues of $\boldsymbol{F}$ are the roots of its *characteristic equation*:

$$\lambda^p - \psi_1\lambda^{p-1} - \cdots - \psi_{p-1}\lambda - \psi_p = 0,$$

and hence are also known as *characteristic roots*. For example, for a second-order difference equation,

$$\boldsymbol{F} = \begin{bmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{bmatrix},$$

and its eigenvalues are the solutions to the characteristic equation:

$$\det(\boldsymbol{F} - \lambda\boldsymbol{I}_2) = -(\psi_1 - \lambda)\lambda - \psi_2 = \lambda^2 - \psi_1\lambda - \psi_2 = 0.$$

Specifically, these two roots are

$$\lambda_1 = \frac{\psi_1 + \sqrt{\psi_1^2 + 4\psi_2}}{2}, \qquad \lambda_2 = \frac{\psi_1 - \sqrt{\psi_1^2 + 4\psi_2}}{2}.$$

An eigenvalue $\lambda^* = a + bi$ is less than one in modulus if $|\lambda^*| = (a^2 + b^2)^{1/2} < 1$; that is, λ^* is inside the inside the unit circle on the complex plane. Suppose that all eigenvalues of $\boldsymbol{F}$ are distinct. Then, $\boldsymbol{F}$ can be diagonalized by a nonsingular matrix $\boldsymbol{C}$ such that $\boldsymbol{C}^{-1}\boldsymbol{F}\boldsymbol{C} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix with all the eigenvalues of $\boldsymbol{F}$ on its main diagonal. It follows that $\boldsymbol{F}^j = \boldsymbol{C}\boldsymbol{\Lambda}^j\boldsymbol{C}^{-1}$. When all the eigenvalues of $\boldsymbol{F}$ are less than one in modulus (inside the unit circle), $\boldsymbol{\Lambda}^j$ (hence $\boldsymbol{F}^j$) converges to a zero matrix as j tends to infinity so that the p th-order difference equation is stable. On the other hand, when there is at least one eigenvalue greater than one in modulus, this eigenvalue eventually dominates so that $\boldsymbol{F}^j$ will explode. In this case, the p th-order difference equation is explosive.

For a stable first-order difference equation, summing the impulse responses yields an *accumulated response* (*interim multiplier*):

$$\sum_{i=0}^{j} \frac{\partial y_{t+j}}{\partial u_{t+i}} = \psi_1^j + \psi_1^{j-1} + \cdots + \psi_1 + 1.$$

Letting j tend to infinity, we have for $|\psi_1| < 1$,

$$\lim_{j\to\infty} \sum_{i=0}^{j} \psi_1^{j-i} = \frac{1}{1-\psi_1},$$

which represents the *long-run effect* (*total multiplier*) of a permanent change in u. This is the total effect resulted from the changes of current and all subsequent innovations. For a first-oder vector difference equation, the long-run effect of a permanent change of $\boldsymbol{\nu}$ is then

$$\lim_{j\to\infty} \sum_{i=0}^{j} \boldsymbol{F}^{j-i} = (\boldsymbol{I}_p - \boldsymbol{F})^{-1}.$$

It can also be shown that the $(1,1)$ th element of $(\boldsymbol{I}_p - \boldsymbol{F})^{-1}$ is

$$\frac{1}{1-\psi_1-\cdots-\psi_p},$$

which is the long-run effect of a permanent change of u in a stable p th-oder difference equation; see Hamilton (1994) for details.

2.3 Back-Shift Operator

The *back-shift operator* $\mathcal{B}$ applied to a time series y_t is defined as $\mathcal{B}y_t = y_{t-1}$. We also write $\mathcal{B}^2 y_t = \mathcal{B}(\mathcal{B}y_t) = y_{t-2}$, $\mathcal{B}^3 y_t = \mathcal{B}(\mathcal{B}^2 y_t) = y_{t-3}$, and so on. Applying this operator

to the constant c, we have $\mathcal{B}(c) = c$. It is easily seen that the back-shift operator has the following linear properties: for constants c and d and two time series y_t and z_t,

$$\mathcal{B}(cy_t + dz_t) = c(\mathcal{B}y_t) + d(\mathcal{B}z_t) = cy_{t-1} + dz_{t-1}.$$

This operator is convenient for representing and manipulating time series.

Using the back-shift operator we can write the first-order difference equation as

$$y_t = \psi_1 \mathcal{B} y_t + u_t, \qquad \text{or} \qquad (1 - \psi_1 \mathcal{B}) y_t = u_t.$$

By pre-multiplying both sides of this equation by $(1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t)$ we have

$$\begin{aligned}(1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t) u_t \\ &= (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t)(1 - \psi_1 \mathcal{B}) y_t \\ &= (1 - \psi_1^{t+1} \mathcal{B}^{t+1}) y_t.\end{aligned}$$

Then provided that y_{-1} is finite and $|\psi_1| < 1$,

$$(1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t)(1 - \psi_1 \mathcal{B}) y_t \approx y_t,$$

when t is large. This suggests that when $|\psi_1| < 1$,

$$\lim_{t \to \infty} (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t)(1 - \psi_1 \mathcal{B}) = \mathcal{I},$$

the identity operator. We may then define the inverse of $(1 - \psi_1 \mathcal{B})$ as

$$(1 - \psi_1 \mathcal{B})^{-1} = \lim_{t \to \infty} (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t).$$

The inverse of other polynomials in $\mathcal{B}$ can be defined in a similar way.

Consider again a second-order difference equation expressed in terms of the back-shift operator:

$$(1 - \psi_1 \mathcal{B} - \psi_2 \mathcal{B}^2) y_t = u_t.$$

We know that this equation is stable if all the roots of $\lambda^2 - \psi_1 \lambda - \psi_2 = 0$ are inside the unit circle. Letting λ_1 and λ_2 denote the characteristic roots, the characteristic polynomial can be factored as

$$\lambda^2 - \psi_1 \lambda - \psi_2 = (\lambda - \lambda_1)(\lambda - \lambda_2).$$

Now, setting $\lambda = z^{-1}$ and multiplying both sides by z^2 yield a polynomial in z:

$$(1 - \psi_1 z - \psi_2 z^2) = (1 - \lambda_1 z)(1 - \lambda_2 z),$$

which has roots: $z_1 = 1/\lambda_1$ and $z_2 = 1/\lambda_2$. This shows that the second-order difference equation is stable if all the roots of $(1 - \psi_1 z - \psi_2 z^2) = 0$ are outside the unit circle. As the polynomial in z corresponds to the polynomial in $\mathcal{B}$ for this difference equation, the condition that all the characteristic roots are inside the unit circle is equivalent to requiring that all the roots of the polynomial in $\mathcal{B}$ are outside the unit circle. More generally, a p^{th}-order difference equation is stable if all the roots of $1-\psi_1 z-\ldots-\psi_p z^p = 0$ are outside the unit circle.

3 Stationary and Invertible ARMA Processes

In this section, we consider processes $\{y_t\}$ that are generated by a white noise $\{\varepsilon_t\}$ which has mean zero and variance σ_ε^2. While y_t are observed random variables, ε_t are unobservable and usually referred to as "innovations" or "random shocks."

3.1 Moving Average Processes

The process $\{y_t\}$ is said to be a *moving average* (MA) process if it can be expressed as

$$y_t = \mu + \Pi(\mathcal{B})\varepsilon_t,$$

where μ is a real number, and $\Pi(\mathcal{B})$ is a polynomial in $\mathcal{B}$. For example, when $\Pi(\mathcal{B}) = \pi_0 - \pi_1\mathcal{B}$, $\{y_t\}$ is an MA process of order one, also known as an MA(1) process. It is typical to set $\pi_0 = 1$ so that an MA(1) process is

$$y_t = \mu + \varepsilon_t - \pi_1\varepsilon_{t-1}.$$

For this MA(1) process, we have $\mathbb{E}(y_t) = \mu$ and the autocovariances:

$$\gamma_0 = \mathbb{E}[(\varepsilon_t - \pi_1\varepsilon_{t-1})^2] = (1 + \pi_1^2)\sigma_\varepsilon^2,$$
$$\gamma_1 = \mathbb{E}[(\varepsilon_t - \pi_1\varepsilon_{t-1})(\varepsilon_{t-1} - \pi_1\varepsilon_{t-2})] = -\pi_1\sigma_\varepsilon^2,$$
$$\gamma_j = 0, \qquad j = 2, 3, \ldots.$$

Hence, the autocorrelations are $\rho_1 = -\pi_1/(1 + \pi_1^2)$ and $\rho_j = 0$ for $j = 2, 3, \ldots$. Note that this series is weakly stationary regardless of the value of π.

In Figure 1 we plot a white noise series and the time paths of three MA(1) processes with $\pi_1 = 0.2, 0.5, 0.8$. It can be seen that these paths are very ragged, and their patterns are similar. When π_1 gets larger, the resulting process has slightly larger first-order autocorrelation, and its time path is less "choppy".

When $\Pi(\mathcal{B}) = 1 - \pi_1\mathcal{B} - \pi_2\mathcal{B}^2 - \cdots - \pi_q\mathcal{B}^q$, we have an MA process of order q, also known as an MA(q) process,

$$y_t = \mu + \varepsilon_t - \pi_1\varepsilon_{t-1} - \pi_2\varepsilon_{t-2} - \cdots - \pi_q\varepsilon_{t-q}.$$

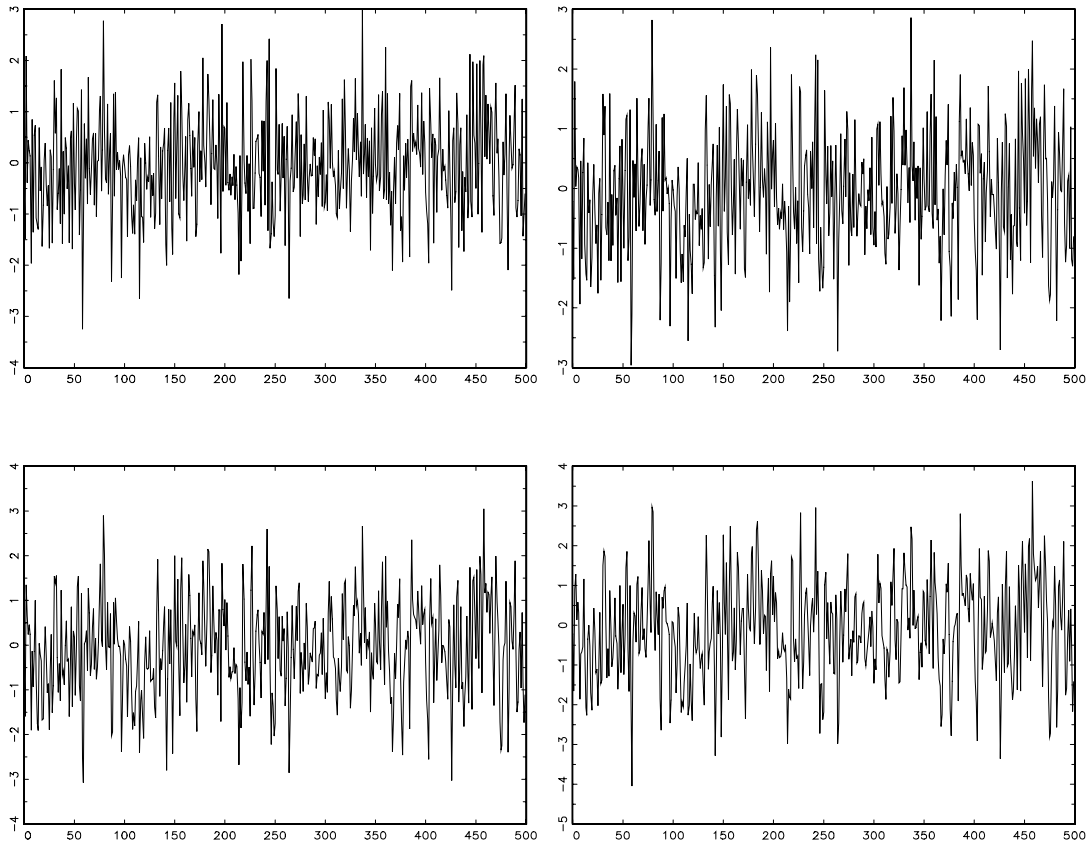

Figure 1: White noise (upper left) and MA processes with $\pi_1 = 0.2$ (upper right), 0.5 (lower left) and 0.8 (lower right).

In this case, $\mathbb{E}(y_t) = \mu$, and

$$\begin{aligned}
\gamma_0 &= (1 + \pi_1^2 + \cdots + \pi_q^2)\sigma_\varepsilon^2, \\
\gamma_1 &= (-\pi_1 + \pi_1\pi_2 + \pi_2\pi_3 + \cdots + \pi_{q-1}\pi_q)\sigma_\varepsilon^2, \\
\gamma_2 &= (-\pi_2 + \pi_1\pi_3 + \pi_2\pi_4 + \cdots + \pi_{q-2}\pi_q)\sigma_\varepsilon^2, \\
&\vdots \\
\gamma_q &= -\pi_q\sigma_\varepsilon^2, \\
\gamma_j &= 0, \qquad j = q+1, q+2, \ldots
\end{aligned}$$

More concisely, for $j = 1, 2, \ldots$,

$$\gamma_j = \left(\sum_{k=1}^{q-j} \pi_k \pi_{k+j} - \pi_j \right) \sigma_\varepsilon^2,$$

with $\pi_j = 0$ if $j > q$. Also, $\rho_j = 0$ for $j = q+1, q+2, \ldots$. Note that an MA(q) process

has only a "fixed" memory for q periods, in the sense that two elements of this series become uncorrelated when they are more than q periods apart. An MA(q) process is also weakly stationary regardless of the values of its MA coefficients.

By letting q tend to infinity, we have an MA(∞) process:

$$y_t = \mu + \varepsilon_t - \sum_{j=1}^{\infty} \pi_j \varepsilon_{t-j},$$

which has $\mathbb{E}(y_t) = \mu$ and

$$\gamma_j = \left(\sum_{k=1}^{\infty} \pi_k \pi_{k+j} - \pi_j \right) \sigma_\varepsilon^2, \qquad j = 0, 1, 2, \ldots$$

Clearly, γ_0 is well defined provided that $\sum_{j=0}^{\infty} \pi_j^2 < \infty$, i.e., π_j are *square summable*. When π_j are square summable, we have from the Cauchy-Schwartz inequality,

$$\sum_{k=0}^{\infty} \pi_k \pi_{k+j} \le \left(\sum_{k=0}^{\infty} \pi_k^2 \right)^{1/2} \left(\sum_{k=0}^{\infty} \pi_{k+j}^2 \right)^{1/2} < \infty,$$

so that all the autocovariances are also well defined. This shows that an MA(∞) process is weakly stationary provided that its MA coefficients π_j are square summable.

3.2 Autoregressive Processes

An *autoregressive* (AR) process is such that

$$\Psi(\mathcal{B}) y_t = c + \varepsilon_t,$$

where c is a real number, $\{\varepsilon_t\}$ is again a white noise with mean zero and variance σ_ε^2, and $\Psi(\mathcal{B})$ is a polynomial in $\mathcal{B}$. When $\Psi(\mathcal{B})$ is of order p, this is an AR process of order p, also known as an AR(p) process.

Consider the AR(1) process with $\Psi(\mathcal{B}) = 1 - \psi_1 \mathcal{B}$:

$$y_t = c + \psi_1 y_{t-1} + \varepsilon_t.$$

From the discussion of the first-order difference equation we know that when $|\psi_1| \ge 1$, the effect of ε_t on y_{t+j} does not die out when j increases, and hence y_t cannot be weakly stationary. To ensure weak stationarity, an AR process is required to have all the roots of $\Psi(z) = 0$ outside the unit circle. For an AR(1) process, this condition is equivalent to $|\psi_1| < 1$. Note that $(1 - \psi_1 \mathcal{B})^{-1} = (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots)$. When $|\psi_1| < 1$, we can then write

$$\begin{aligned} y_t &= (1 - \psi_1 \mathcal{B})^{-1} (c + \varepsilon_t) \\ &= (1 + \psi_1 + \psi_1^2 + \cdots) c + (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots) \varepsilon_t \\ &= c/(1 - \psi_1) + (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots) \varepsilon_t, \end{aligned}$$

where $1/(1-\psi_1)$ is just $\Psi(1)^{-1}$. This shows that a stationary AR(1) process has an MA(∞) representation with square summable MA coefficients.

It is now easy to see that $\mathbb{E}(y_t) = c/(1-\psi_1)$,

$$\begin{aligned}
\gamma_0 &= (1+\psi_1^2+\psi_1^4+\cdots)\sigma_\varepsilon^2 = \sigma_\varepsilon^2/(1-\psi_1^2),\\
\gamma_1 &= (\psi_1+\psi_1^3+\psi_1^5+\cdots)\sigma_\varepsilon^2 = \psi_1[\sigma_\varepsilon^2/(1-\psi_1^2)],\\
\gamma_2 &= (\psi_1^2+\psi_1^4+\psi_1^6+\cdots)\sigma_\varepsilon^2 = \psi_1^2[\sigma_\varepsilon^2/(1-\psi_1^2)],\\
&\vdots
\end{aligned}$$

More concisely,

$$\gamma_j = \psi_1^j \frac{\sigma_\varepsilon^2}{1-\psi_1^2} = \psi_1^j \gamma_0, \qquad j = 0, 1, 2, \ldots,$$

so that $\rho_j = \psi_1^j$. In other words, the autocorrelations (memory) of a weakly stationary AR(1) process dies out exponentially fast.

An alternative approach to deriving the moments of an AR(1) process is as follows. Given that $\{y_t\}$ is weakly stationary, we can write

$$\mu = \mathbb{E}(y_t) = c + \psi_1 \mathbb{E}(y_{t-1}) = c + \psi_1 \mu,$$

so that $\mu = c/(1-\psi_1)$. Hence,

$$(y_t - \mu) = \psi_1 (y_{t-1} - \mu) + \varepsilon_t,$$

i.e., the process in terms of its deviations from the mean has the same AR structure. As

$$\mathbb{E}(\varepsilon_t y_{t-j}) = \begin{cases} \sigma_\varepsilon^2, & j = 0, \\ 0, & j = 1, 2, \ldots, \end{cases}$$

the autocovariances are

$$\gamma_j = \mathbb{E}[(y_t - \mu)(y_{t-j} - \mu)] = \psi_1 \mathbb{E}[(y_{t-1} - \mu)(y_{t-j} - \mu)] = \psi_1 \gamma_{j-1}, \qquad j = 1, 2, \ldots,$$

for $j = 1, 2, \ldots$, with

$$\gamma_0 = \mathbb{E}[(y_t - \mu)^2] = \psi_1^2 \mathbb{E}[(y_{t-1} - \mu)^2] + \sigma_\varepsilon^2 = \psi_1^2 \gamma_0 + \sigma_\varepsilon^2. = \sigma_\varepsilon^2/(1-\psi_1^2).$$

These results are precisely what we obtained earlier. Hence, the autocovariances and autocorrelations of an AR(1) process have the same AR(1) structure.

In Figure 2 we plot a white noise series and the time paths of three AR processes with $\psi_1 = 0.2, 0.5, 0.8$. It can be seen that as ψ_1 increases, the process has stronger autocorrelations through time, and the resulting time path becomes smoother.

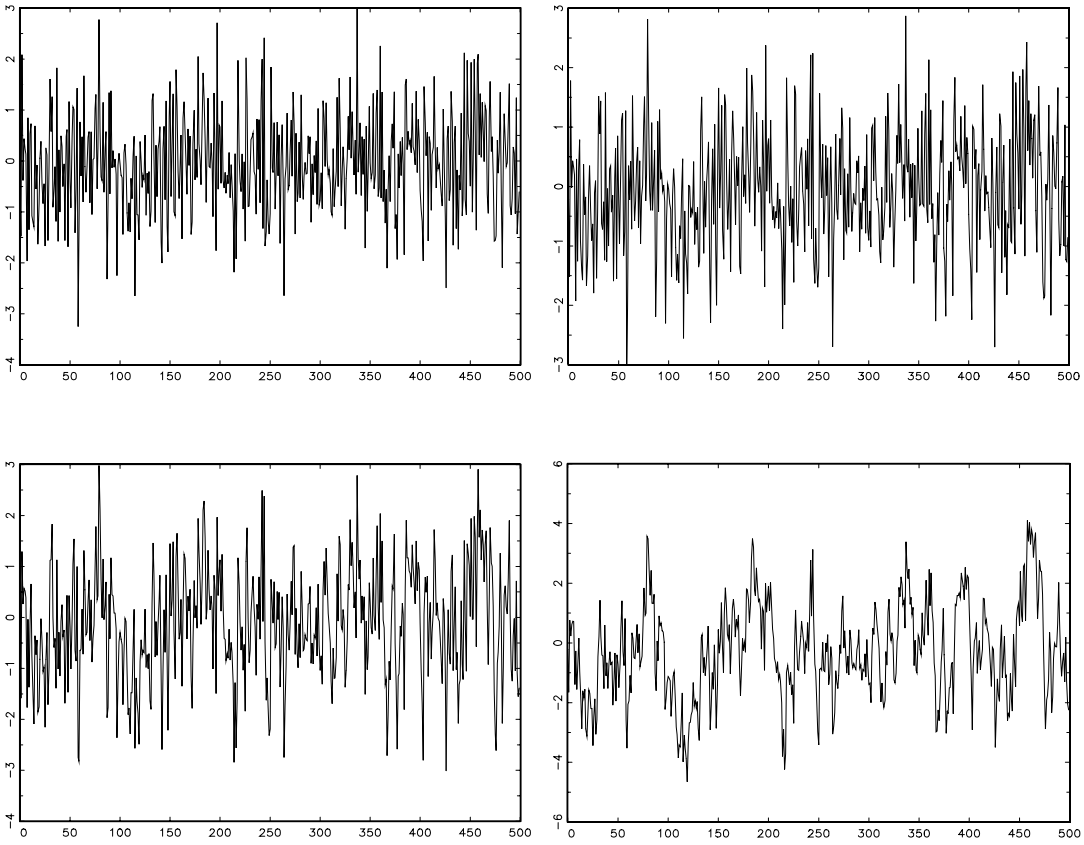

Figure 2: White noise (upper left) and AR processes with $\psi_1 = 0.2$ (upper right), 0.5 (lower left) and 0.8 (lower right).

The previous results extend straightforwardly to the AR(p) process with the polynomial $\Psi(\mathcal{B}) = 1 - \psi_1\mathcal{B} - \psi_2\mathcal{B}^2 - \cdots - \psi_p\mathcal{B}^p$:

$$y_t = c + \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p} + \varepsilon_t.$$

This AR(p) process is weakly stationary if all the roots of $\Psi(z) = 0$ are outside the unit circle. A weakly stationary AR(p) process also has an MA(∞) representation:

$$y_t = \Psi(1)^{-1}c + \Psi(\mathcal{B})^{-1}\varepsilon_t.$$

In this case, $\mu = \Psi(1)^{-1}c = c/(1 - \psi_1 - \psi_2 - \cdots - \psi_p)$, and the autocovariances have the same AR(p) structure:

$$\gamma_j = \psi_1\gamma_{j-1} + \psi_2\gamma_{j-2} + \cdots + \psi_p\gamma_{j-p}, \qquad j = 1, 2, \ldots,$$

and $\gamma_0 = \psi_1\gamma_1 + \psi_2\gamma_2 + \cdots + \psi_p\gamma_p + \sigma_\varepsilon^2$. For autocorrelations, we have

$$\rho_j = \psi_1\rho_{j-1} + \psi_2\rho_{j-2} + \cdots + \psi_p\rho_{j-p}, \qquad j = 1, 2, \ldots$$

These equations are known as the *Yule-Walker equations* which form a p^{th}-order difference equation in ρ_j. This system is also stable because it has the same AR(p) structure as y_t. Given that the initial value of this difference equation is $\rho_0 = 1$, ρ_j must converge to zero exponentially fast as j tends to infinity.

3.3 Autoregressive Moving Average Processes

Combining an AR(p) process and an MA(q) process we obtain a mixed ARMA(p, q) process:

$$\Psi(\mathcal{B})y_t = c + \Pi(\mathcal{B})\varepsilon_t,$$

where $\Psi(\mathcal{B})$ is a p^{th}-order polynomial in $\mathcal{B}$ and $\Pi(\mathcal{B})$ is a q^{th}-order polynomial in $\mathcal{B}$. If these two polynomials have a common factor, these factors would cancel out, resulting in an ARMA process with lower orders. Thus, it is always assumed that the AR and MA polynomials of an ARMA(p, q) process do not have any common factor.

When all the roots of $\Psi(z) = 0$ are outside the unit circle, we can let $\Phi(\mathcal{B}) = \Psi(\mathcal{B})^{-1}\Pi(\mathcal{B})$ and write

$$y_t = \Psi(1)^{-1}c + \Phi(\mathcal{B})\varepsilon_t = \Psi(1)^{-1}c + \sum_{j=0}^{\infty}\phi_j\varepsilon_{t-j},$$

with $\phi_0 = 1$. This process has mean $\mu = c/(1 - \psi_1 - \cdots - \psi_p)$. In terms of the deviations from the mean, we have the ARMA process

$$\Psi(B)(y_t - \mu) = \Pi(B)\varepsilon_t.$$

To compute the autocovariances, we note that

$$\Psi(B)(y_t - \mu)(y_{t-j} - \mu) = \Pi(B)\varepsilon_t(y_{t-j} - \mu).$$

We do not state explicitly the autocovariances γ_j for $j \leq q$ because they are quite complicated; readers are referred to other textbooks for details. Yet we note that for $j = q + 1, q + 2, \ldots$, $\mathbb{E}[\Pi(B)\varepsilon_t(y_{t-j} - \mu)] = 0$, so that

$$\gamma_j = \psi_1\gamma_{j-1} + \cdots + \psi_p\gamma_{j-p}.$$

That is, the autocovariances for $j > q$ obey the AR(p) structure.

3.4 Invertibility of MA Processes

The MA process $y_t = \mu + \Pi(\mathcal{B})\varepsilon_t$ is said to be *invertible* if all the roots of $\Pi(z) = 0$ are outside the unit circle. Similarly, the ARMA process,

$$\Psi(\mathcal{B})y_t = c + \Pi(\mathcal{B})\varepsilon_t,$$

is invertible if all the roots of $\Pi(z) = 0$ are outside the unit circle.

It is easy to see that the MA(1) process $y_t = \mu + (1 - \pi_1\mathcal{B})\varepsilon_t$ with $|\pi_1| < 1$ is invertible. An invertible MA(1) process has the following AR(∞) representation:

$$(1 - \pi_1\mathcal{B})^{-1}(y_t - \mu) = \sum_{j=0}^{\infty} \pi_1^j \mathcal{B}^j (y_t - \mu) = \varepsilon_t.$$

This expression shows that for invertible MA(1) processes, each innovation ε_t can be expressed as a weighted sum of current and all past observations y_t. More generally, each innovation ε_t of MA(q) processes can also be expressed as a weighted sum of current and all past y_t.

On the other hand, an MA(1) process with $|\pi_1| > 1$ is non-invertible. When $|\pi_1| > 1$, $(1 + \pi_1\mathcal{B} + \pi_1^2\mathcal{B}^2 + \cdots)$ can not be defined as $(1 - \pi_1\mathcal{B})^{-1}$. Consider the polynomial of the *forward-shift* operator $\mathcal{B}^{-1}$, where $\mathcal{B}^{-1}$ is such that $\mathcal{B}^{-1}y_t = y_{t+1}$ and $\mathcal{B}^{-1}\mathcal{B} = \mathcal{I}$. Then for the polynomial $(1 - \pi_1^{-1}\mathcal{B}^{-1})$ with root inside the unit circle, its inverse is well defined as

$$(1 - \pi_1^{-1}\mathcal{B}^{-1})^{-1} = (1 + \pi_1^{-1}\mathcal{B}^{-1} + \pi_1^{-2}\mathcal{B}^{-2} + \cdots).$$

Straightforward calculation shows that

$$-\pi_1^{-1}\mathcal{B}^{-1}(1 + \pi_1^{-1}\mathcal{B}^{-1} + \pi_1^{-2}\mathcal{B}^{-2} + \cdots)(1 - \pi_1\mathcal{B}) = \mathcal{I}.$$

This suggests that for $|\pi_1| > 1$, we can define

$$(1 - \pi_1\mathcal{B})^{-1} = -\big(1 - \pi_1^{-1}\mathcal{B}^{-1}\big)^{-1}\big(\pi_1^{-1}\mathcal{B}^{-1}\big).$$

It follows that a non-invertible MA(1) process can be represented as

$$-\pi_1^{-1}\mathcal{B}^{-1}(1 + \pi_1^{-1}\mathcal{B}^{-1} + \pi_1^{-2}\mathcal{B}^{-2} + \cdots)(y_t - \mu) = \varepsilon_t,$$

which is a weighted sum of all future y_t. This result also extends to non-invertible MA(q) processes so that each innovation ε_t depends on all future observations. As far as forecasting is concerned, it make practical sense to consider invertible processes whose innovations depend on what happen in the past. This is a reason why researchers usually confine themselves with stationary and invertible ARMA processes.

3.5 Vector AR processes

In many empirical studies, we usually encounter multiple time series and thus would like to analyze not only the dynamic pattern of a particular time series but also the dynamic

relationships across different time series. In this section we will extend univariate AR processes discussed in Section 3.2 to vector AR (VAR) processes.

Let $\{\boldsymbol{\varepsilon}_t\}$ be a d-dimensional vector time series with mean zero, the variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, and the autocovariances $\mathrm{cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_s) = \boldsymbol{o}$ for all $t \neq s$. Then, $\{\boldsymbol{y}_t\}$ is a VAR process when

$$\boldsymbol{\Psi}(\mathcal{B})\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{c}$ is a vector of constants, and $\boldsymbol{\Psi}(\mathcal{B}) = \boldsymbol{I}_d - \boldsymbol{\Psi}_1\mathcal{B} - \boldsymbol{\Psi}_2\mathcal{B}^2 - \ldots$ is a matrix polynomial in $\mathcal{B}$ with $\boldsymbol{\Psi}_j$ a $d \times d$ matrix. This is a VAR(p) process if the order of $\boldsymbol{\Psi}(\mathcal{B})$ is p.

When $\boldsymbol{\Psi}(\mathcal{B}) = \boldsymbol{I}_d - \boldsymbol{\Psi}_1\mathcal{B}$, we have a VAR(1) process. Similar as before, define $(\boldsymbol{I}_d - \boldsymbol{\Psi}_1\mathcal{B})^{-1}$ as $\boldsymbol{I}_d + \boldsymbol{\Psi}_1\mathcal{B} + \boldsymbol{\Psi}_1^2\mathcal{B}^2 + \cdots$. The MA($\infty$) representation of this process is then

$$\boldsymbol{y}_t = (\boldsymbol{I} - \boldsymbol{\Psi}_1)^{-1}\boldsymbol{c} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_1^j \boldsymbol{\varepsilon}_{t-j}.$$

It is easy to see that the effect of $\boldsymbol{\varepsilon}_{t-j}$ on $\boldsymbol{y}_t$ eventually dies out provided that all the characteristic roots of $\boldsymbol{\Psi}_1$ are less than one in modulus (inside the unit circle). This condition also defines weak stationarity of $\boldsymbol{y}_t$. The VAR(p) process $\boldsymbol{\Psi}(\mathcal{B})\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\varepsilon}_t$ can also be expressed as a pd-dimensional VAR(1) process:

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{bmatrix}}_{\boldsymbol{Y}_t} = \underbrace{\begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{o} \\ \boldsymbol{o} \\ \vdots \\ \boldsymbol{o} \end{bmatrix}}_{\boldsymbol{C}} + \underbrace{\begin{bmatrix} \boldsymbol{\Psi}_1 & \boldsymbol{\Psi}_2 & \cdots & \boldsymbol{\Psi}_{p-1} & \boldsymbol{\Psi}_p \\ \boldsymbol{I}_d & \boldsymbol{o} & \cdots & \boldsymbol{o} & \boldsymbol{o} \\ \boldsymbol{o} & \boldsymbol{I}_d & \cdots & \boldsymbol{o} & \boldsymbol{o} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \boldsymbol{o} & \boldsymbol{o} & \cdots & \boldsymbol{I}_d & \boldsymbol{o} \end{bmatrix}}_{\boldsymbol{F}} \underbrace{\begin{bmatrix} \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \boldsymbol{y}_{t-3} \\ \vdots \\ \boldsymbol{y}_{t-p} \end{bmatrix}}_{\boldsymbol{Y}_{t-1}} + \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \boldsymbol{o} \\ \boldsymbol{o} \\ \vdots \\ \boldsymbol{o} \end{bmatrix}}_{\boldsymbol{E}_t};$$

that is, $\boldsymbol{Y}_t = \boldsymbol{C} + \boldsymbol{F}\boldsymbol{Y}_{t-1} + \boldsymbol{E}_t$. It follows that the VAR(p) process is stationary if all the characteristic roots of $\boldsymbol{F}$ are inside the unit circle.

The properties of VAR processes are similar to those of univariate AR processes. For the VAR(1) process $\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Psi}_1\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t$, we have $\mathbb{E}(\boldsymbol{y}_t) = (\boldsymbol{I}_d - \boldsymbol{\Psi}_1)^{-1}\boldsymbol{c}$, and the autocovariances are

$$\boldsymbol{\Gamma}_j = \mathrm{cov}(\boldsymbol{y}_t, \boldsymbol{y}_{t-j}) = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_1^{i+j} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{\Psi}_1^{i\prime}, \qquad j = 0, 1, 2, \ldots;$$

in particular, $\boldsymbol{\Gamma}_0 = \mathrm{var}(\boldsymbol{y}_t) = \sum_{i=0}^{\infty} \boldsymbol{\Psi}_1^i \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{\Psi}_1^{i\prime}$. Note that $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}_{-j}'$. For $\boldsymbol{\Gamma}_0$, its k^{th} diagonal element is $\gamma_{kk,0}$, the variance of $y_{k,t}$, and its $(h,k)^{\text{th}}$ element is $\gamma_{hk,0}$, the contemporaneous covariance of $y_{h,t}$ and $y_{k,t}$. More concisely, we have the multivariate Yule-Walker

equations:

$$\boldsymbol{\Gamma}_j = \boldsymbol{\Psi}_1\boldsymbol{\Gamma}_{j-1}, \qquad j = 1, 2, \ldots,$$

and $\boldsymbol{\Gamma}_0 = \boldsymbol{\Psi}_1\boldsymbol{\Gamma}_1' + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Let $\boldsymbol{D}$ denote the diagonal matrix with the k^{th} diagonal element $\gamma_{kk,0}$. The autocorrelations of $\boldsymbol{y}_t$ are

$$\boldsymbol{R}_j = \boldsymbol{D}^{-1/2}\boldsymbol{\Gamma}_j\boldsymbol{D}^{-1/2}, \qquad j = 0, 1, 2, \ldots$$

For the VAR(p) process $\boldsymbol{\Psi}(\mathcal{B})\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\varepsilon}_t$, $\mathbb{E}(\boldsymbol{y}_t) = \boldsymbol{\Psi}(1)^{-1}\boldsymbol{c}$, and the multivariate Yule-Walker equations of autocovariances now read

$$\boldsymbol{\Gamma}_j = \boldsymbol{\Psi}_1\boldsymbol{\Gamma}_{j-1} + \boldsymbol{\Psi}_2\boldsymbol{\Gamma}_{j-2} + \cdots + \boldsymbol{\Psi}_p\boldsymbol{\Gamma}_{j-p}, \qquad j = 1, 2, \ldots,$$

and $\boldsymbol{\Gamma}_0 = \boldsymbol{\Psi}_1\boldsymbol{\Gamma}_1' + \boldsymbol{\Psi}_2\boldsymbol{\Gamma}_2' + \cdots + \boldsymbol{\Psi}_p\boldsymbol{\Gamma}_p' + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

3.6 Impulse Responses and Error Variance Decomposition

Characterizing the impulse responses of a VAR process is more involved. Consider the VAR(p) process $\boldsymbol{\Psi}(\mathcal{B})\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\varepsilon}_t$ and write its MA representation as

$$\boldsymbol{y}_t = \boldsymbol{\Phi}(1)\boldsymbol{c} + \boldsymbol{\Phi}(\mathcal{B})\boldsymbol{\varepsilon}_t = \boldsymbol{\Phi}(1)\boldsymbol{c} + \sum_{j=0}^{\infty}\boldsymbol{\Phi}_j\,\boldsymbol{\varepsilon}_{t-j},$$

where the polynomial $\boldsymbol{\Phi}(\mathcal{B}) = \boldsymbol{\Psi}(\mathcal{B})^{-1}$ with $\boldsymbol{\Phi}_0 = \boldsymbol{I}_d$. The impulse response of $\boldsymbol{y}_t$ to one unit shock to the i^{th} equation, $\varepsilon_{i,t-j}$, is the i^{th} column of $\boldsymbol{\Phi}_j$, which can be expressed as $\boldsymbol{\Phi}_j\boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the i^{th} Cartesian unit vector. In particular, the impulse response of the VAR(1) process $\boldsymbol{y}_t$ to $\varepsilon_{i,t-j}$ is the i^{th} column of $\boldsymbol{\Psi}_1^j$. The accumulated response over n periods is

$$\boldsymbol{A}_n\boldsymbol{e}_i = \left(\sum_{j=0}^{n}\boldsymbol{\Phi}_j\right)\boldsymbol{e}_i,$$

and the long-run effect is $\boldsymbol{A}_\infty\boldsymbol{e}_i$, where $\boldsymbol{A}_\infty = \boldsymbol{\Phi}(1) = \boldsymbol{\Psi}(1)^{-1}$.

As $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is not necessarily a diagonal matrix, one unit shock of $\varepsilon_{i,t-j}$ may be accompanied by the shock of other innovations. Merely focusing on the shock of a particular element of $\boldsymbol{\varepsilon}_{t-j}$ and ignoring other shocks may not give a complete description of the dynamic effect of innovations. To circumvent this problem, consider the *Cholesky decomposition* $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is a lower triangular matrix with non-zero diagonal elements, and the transformed innovations $\boldsymbol{v}_t = \boldsymbol{L}^{-1}\boldsymbol{\varepsilon}_t$. As $\text{var}(\boldsymbol{v}_t) = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{L}^{-1\prime} = \boldsymbol{I}_d$,

the elements of $\boldsymbol{v}_t$ are uncorrelated random variables and will be referred to as "orthogonalized innovations." The MA representation of the VAR(p) process $\boldsymbol{y}_t$ in terms of $\boldsymbol{v}_t$ is

$$\boldsymbol{y}_t = \boldsymbol{\Phi}(1)\boldsymbol{c} + \sum_{j=0}^{\infty} \boldsymbol{\Phi}_j \boldsymbol{L}\, \boldsymbol{v}_{t-j} = \boldsymbol{\Phi}(1)\boldsymbol{c} + \sum_{j=0}^{\infty} \boldsymbol{\Theta}_j\, \boldsymbol{v}_{t-j},$$

where $\boldsymbol{\Theta}_j = \boldsymbol{\Phi}_j \boldsymbol{L}$ are transformed coefficient matrices. The *orthogonalized impulse response* is then defined as the impulse response to one unit shock of the orthogonalized innovations. Thus, the orthogonalized impulse response to $\boldsymbol{v}_{t-j}$ is the coefficient matrix $\boldsymbol{\Theta}_j$; In particular, the i^{th} column of $\boldsymbol{\Theta}_j$, $\boldsymbol{\Theta}_j \boldsymbol{e}_i$, is the impulse response of $\boldsymbol{y}_t$ to one unit shock of $v_{i,t-j}$. Note that there is no "scaling" problem here because $\operatorname{var}(\boldsymbol{v}_t)$ is an identity matrix so that one unit shock is equivalent to the shock of one standard deviation. A drawback of the orthogonalized impulse response function is that it depends on the ordering of the elements of $\boldsymbol{y}_t$. To see this, observe that, as $\boldsymbol{\Phi}_0 = \boldsymbol{I}_d$, the immediate response to $\boldsymbol{v}_t$ is $\boldsymbol{\Theta}_0 = \boldsymbol{L}$. Thus, the immediate effect on $y_{1,t}$ is resulted from its own (orthogonalied) innovation $v_{1,t}$, $y_{2,t}$ is immediately affected by the first two (orthogonalized) innovations: $v_{1,t}$ and $v_{2,t}$, and so on. As such, changing the ordering of the elements of $\boldsymbol{y}_t$ results in different impulse responses. That is, the orthogonalized impulse responses are *not* uniquely defined.

Let $\mathcal{F}^t$ denote the information set up to time t. Pesaran and Shin (1998) define the *generalized impulse response* of $\boldsymbol{y}_t$ to the shock $\varepsilon_{i,t-j} = \delta$ as

$$\mathbb{E}\big(\boldsymbol{y}_t \mid \varepsilon_{i,t-j} = \delta, \mathcal{F}^{t-j-1}\big) - \mathbb{E}\big(\boldsymbol{y}_t \mid \mathcal{F}^{t-j-1}\big).$$

From the MA representation it is easily seen that

$$\mathbb{E}\big(\boldsymbol{y}_t \mid \varepsilon_{i,t-j} = \delta, \mathcal{F}^{t-j-1}\big) = \boldsymbol{\Phi}(1)\boldsymbol{c} + \sum_{k=j+1}^{\infty} \boldsymbol{\Phi}_k \boldsymbol{\varepsilon}_{t-k} + \boldsymbol{\Phi}_j\, \mathbb{E}(\boldsymbol{\varepsilon}_{t-j} \mid \varepsilon_{i,t-j} = \delta).$$

This differs from $\mathbb{E}\big(\boldsymbol{y}_t \mid \mathcal{F}^{t-j-1}\big)$ by the last term $\boldsymbol{\Phi}_j\, \mathbb{E}(\boldsymbol{\varepsilon}_{t-j} \mid \varepsilon_{i,t-j} = \delta)$, which is the generalized impulse response to the shock $\varepsilon_{i,t-j} = \delta$. When $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution, it is well known that

$$\mathbb{E}(\varepsilon_{k,t} \mid \varepsilon_{i,t} = \delta) = \frac{\sigma_{ki}}{\sigma_{ii}}\, \delta,$$

where σ_{ki} is the $(k,i)^{\text{th}}$ element of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Using this result we immediately obtain

$$\mathbb{E}(\boldsymbol{\varepsilon}_t \mid \varepsilon_{i,t} = \delta) = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{e}_i \delta / \sigma_{ii},$$

so that the generalized impulse response is

$$\boldsymbol{\Phi}_j\, \mathbb{E}(\boldsymbol{\varepsilon}_{t-j} \mid \varepsilon_{i,t-j} = \delta) = \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{e}_i \delta / \sigma_{ii}.$$

Setting $\delta = \sigma_{ii}^{1/2}$, a shock of one standard deviation to the i th equation, the generalized impulse response of $\boldsymbol{y}_t$ is $\boldsymbol{\Phi}_j \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{e}_i / \sigma_{ii}^{1/2}$. It is clear that this impulse response does not depend on the ordering of the elements of $\boldsymbol{y}_t$ but requires the normality assumption on $\boldsymbol{\varepsilon}_t$.

When $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is diagonal, the Cholesky decomposition is based on $\boldsymbol{L} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{1/2}$ which is also diagonal. In this case, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{e}_i = \sigma_{ii} \boldsymbol{e}_i$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{1/2} \boldsymbol{e}_i = \sigma_{ii}^{1/2} \boldsymbol{e}_i$. The generalized impulse response then becomes

$$\boldsymbol{\Phi}_j \sigma_{ii}^{1/2} \boldsymbol{e}_i = \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{1/2} \boldsymbol{e}_i = \boldsymbol{\Theta}_j \boldsymbol{e}_i,$$

which is also the orthogonalized impulse response. It can also be shown that when $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is not diagonal, the two impulse responses coincide only for the shock entering the first equation but not otherwise.

Recall that a VAR(p) process written in terms of the orthogonalized innovations $\boldsymbol{v}_t$ is $\boldsymbol{y}_t = \boldsymbol{\Phi}(1)\boldsymbol{c} + \sum_{j=0}^{\infty} \boldsymbol{\Theta}_j \, \boldsymbol{v}_{t-j}$. The optimal forecast of $\boldsymbol{y}_{t+h}$ based on the information set $\mathcal{F}^t$ is

$$\hat{\boldsymbol{y}}_t(h) := \mathbb{E}\big(\boldsymbol{y}_{t+h} \mid \mathcal{F}^t\big) = \boldsymbol{\Phi}(1)\boldsymbol{c} + \sum_{j=h}^{\infty} \boldsymbol{\Theta}_j \, \boldsymbol{v}_{t+h-j}.$$

The h-step forecast error is then $\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_t(h) = \sum_{j=0}^{h-1} \boldsymbol{\Theta}_j \, \boldsymbol{v}_{t+h-j}$, with the i th element:

$$y_{i,t+h} - \hat{y}_{i,t}(h) = \sum_{j=0}^{h-1} \boldsymbol{e}_i' \boldsymbol{\Theta}_j \, \boldsymbol{v}_{t+h-j} = \sum_{j=0}^{h-1} \sum_{k=1}^{d} \theta_{ik,j} v_{k,t+h-j}.$$

As $\operatorname{var}(\boldsymbol{v}_t) = \boldsymbol{I}_d$, the elements of $\boldsymbol{v}_t$ are mutually uncorrelated as well as serially uncorrelated, the h-step *forecast error variance* of $y_{i,t+h}$ is defined as

$$\mathbb{E}\big[y_{i,t+h} - \hat{y}_{i,t}(h)\big]^2 = \sum_{j=0}^{h-1} \sum_{k=1}^{d} \theta_{ik,j}^2 = \sum_{k=1}^{d} \sum_{j=0}^{h-1} (\boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{e}_k)^2,$$

in which $\sum_{j=0}^{h-1} (\boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{e}_k)^2$ is due to the contribution of the k th innovation. This error variance can also be expressed in terms of the original coefficient matrices:

$$\sum_{k=1}^{d} (\boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{e}_k)^2 = \boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{\Theta}_j' \boldsymbol{e}_i = \boldsymbol{e}_i' \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{\Phi}_j' \boldsymbol{e}_i.$$

The proportion of the total forecast error variance that can be attributed to the k th innovation is

$$\frac{\sum_{j=0}^{h-1} (\boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{e}_k)^2}{\sum_{j=0}^{h-1} \sum_{k=1}^{d} (\boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{e}_k)^2} = \frac{\sum_{j=0}^{h-1} (\boldsymbol{e}_i' \boldsymbol{\Theta}_j \boldsymbol{e}_k)^2}{\sum_{j=0}^{h-1} \boldsymbol{e}_i' \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \boldsymbol{\Phi}_j' \boldsymbol{e}_i},$$

which is known as the *orthogonalized forecast error variance decomposition*. As these ratios sum to one (over k), this decomposition enables us to determine the relative importance of a particular (orthogonalized) innovation. Note that for different h, the resulting forecast errors and their variance decompositions are also different. Pesaran and Shin (1998) define the *generalized forecast error variance decomposition* as

$$\frac{\sum_{j=0}^{h-1}(\boldsymbol{e}_i'\boldsymbol{\Phi}_j\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{e}_k)^2/\sigma_{ii}}{\sum_{j=0}^{h-1}\boldsymbol{e}_i'\boldsymbol{\Phi}_j\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}_j'\boldsymbol{e}_i},$$

yet these ratios do not sum to one. This decomposition. although does not depend on the ordering of variables, can not be interpreted as a measure of relative importance of innovations.

4 Box-Jenkins Approach

A fundamental result in the time series analysis is *Wold's decomposition* which asserts that any covariance-stationary process can be represented as the sum of two components: an MA(∞) component and a linearly deterministic component; see e.g., Fuller (1996, pp. 96–98). Fitting a model with infinitely many parameters is practically intractable. The Box-Jenkins approach is designed to find a model with a parsimonious ARMA structure (i.e., a small number of parameters) which can well represent the MA(∞) component.

The standard Box-Jenkins approach consists of the following steps:

1. Transform the original time series to a covariance stationary series.
2. Identify a preliminary ARMA(p, q) model for the transformed series.
3. Estimate unknown parameters in this preliminary model.
4. Conduct diagnostic tests to check model adequacy and re-estimate an ARMA model if the preliminary model is found inappropriate.

These steps may be repeated until a suitable model is found.

4.1 Differencing

When a series exhibit trending behavior, the first step of the Box-Jenkins approach amounts to "de-trending" this series. For example, one may transform the original series $\{\eta_t\}$ by taking the first difference:

$$y_t = \eta_t - \eta_{t-1} = (1 - \mathcal{B})\eta_t.$$

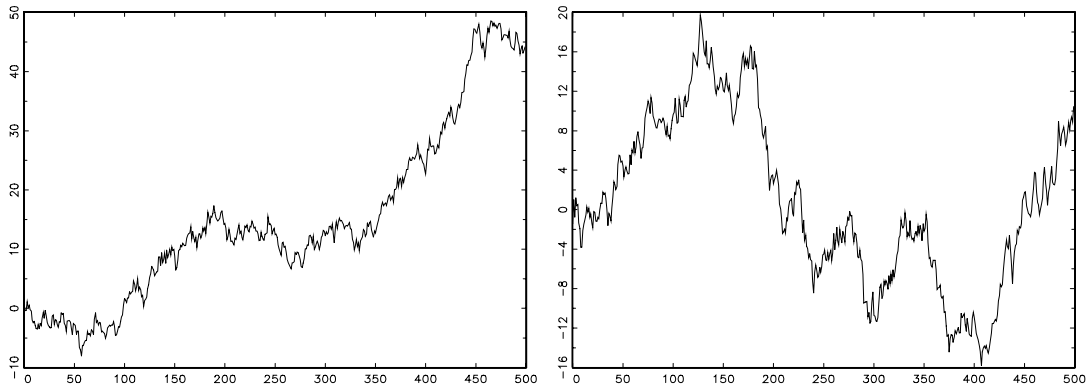

Figure 3: The time paths of a Gaussian random walk.

When y_t are obtained from differencing once, η_t are in fact cumulative sums of y_t. In other words, η_t are integrated y_t. If the differenced series $\{y_t\}$ is an ARMA process, $\{\eta_t\}$ is known as an ARIMA (autoregressive, integrated, moving average) process. When y_t are i.i.d. random variables and hence form an ARMA$(0,0)$ process, then $\{\eta_t\}$ is an ARIMA$(0,1,0)$ process, also known as a *random walk*. We plot two sample paths of a Gaussian random walk in Figure 3, where one exhibits upward trending pattern and the other has large swings. These paths are all relatively smooth and quite different from those of stationary processes.

Suppose that $(1-\mathcal{B})\eta_t = y_t$ and $\{y_t\}$ is a weakly stationary AR(1) process: $y_t = \psi_1 y_{t-1} + \varepsilon_t$. Then, $\{\eta_t\}$ is also an ARMA$(2,0)$ process:

$$(1-\psi_1\mathcal{B})(1-\mathcal{B})\eta_t = [1-(1+\psi_1)+\psi_1\mathcal{B}^2]\eta_t = \varepsilon_t,$$

but the AR polynomial $\Psi(z) = 0$ has a root on the unit circle. Such a root is also known as a *unit root*. Similarly, when $\{y_t\}$ is a stationary ARMA(p,q) process, $\{\eta_t\}$ is an ARIMA$(p,1,q)$ process or an ARMA$(p+1,q)$ process with an AR unit root. In what follows, an ARIMA$(p,1,q)$ process is understood as the process whose first differenced series is a stationary ARMA(p,q) process. An ARIMA$(p,1,q)$ process is also known as an *integrated process*, or simply an $I(1)$ process.

The data may be differenced several times. When y_t are obtained by differencing η_t twice:

$$y_t = (\eta_t - \eta_{t-1}) - (\eta_{t-1} - \eta_{t-2}) = (\eta_t - 2\eta_{t-1} + \eta_{t-2}),$$

$\{\eta_t\}$ is an ARIMA$(p,2,q)$ process or an $I(2)$ process. More generally, an ARIMA(p,d,q) process is an $I(d)$ process, and it must be differenced d times to yield a stationary ARMA representation.

For quarterly data η_t, there may exist quarterly regularity (seasonal pattern). To eliminate this pattern, it is common to conduct the following *seasonal differencing*:

$$y_t = \eta_t - \eta_{t-4} = (1 - \mathcal{B}^4)\eta_t.$$

Note that the polynomial $(1 - z^4) = 0$ contains four unit roots because

$$(1 - z^4) = (1 - z^2)(1 + z^2) = (1 - z)(1 + z)(1 + iz)(1 - iz),$$

where each unit root accounts for the behavior of η_t at different frequencies; we omit the details. For monthly data, seasonal differencing is such that

$$y_t = \eta_t - \eta_{t-12} = (1 - \mathcal{B}^{12})\eta_t.$$

For data at other frequencies, different seasonal differencing operators may be employed.

In practice, differencing is *not* the only way to eliminate trending or seasonal patterns. For example, one may remove the deterministic time trend by regressing η_t on a simple time trend variable t or on trends with different orders: $t, t^2, \ldots, t^p$. It must be emphasized that the trending behavior of an $I(1)$ process, usually known as a *stochastic trend*, is quite different from a deterministic trend. For data exhibiting seasonality, one may also eliminate seasonal patterns by regressing η_t on seasonal dummies or by estimating a seasonal ARMA model. Which methods should be used depend on the properties of the time series being studied.

4.2 Identification

Identifying a proper ARMA model is never an easy task; the original Box-Jenkins approach provides only a quick and easy way to determine a preliminary model. As computing at the present time is much easy, such identification procedures may not be necessary.

Recall that the autocorrelations ρ_j of an AR(p) process decay to zero exponentially fast and that ρ_j of an MA(q) process has an abrupt cut-off at $j = q$ such that $\rho_j = 0$ for $j > q$. An important step in model identification is to examine the sample autocorrelations:

$$\hat{\rho}_j = \hat{\gamma}_j / \hat{\gamma}_0,$$

where $\hat{\gamma}_j = \sum_{t=j+1}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y})/T$ and where $\bar{y} = \sum_{t=1}^{T} y_t/T$. Under regularity conditions,[1]

$$\hat{\rho}_j \xrightarrow{\mathbb{P}} \gamma_j/\gamma_0 = \rho_j,$$

[1]For an MA(∞) process with absolute summable MA coefficients, it is typically required the innovations ε_t to have finite 4^{th} moment when they are i.i.d. random variables or to have uniformly bounded 6^{th} moment when they are independent (but not necessarily identically distributed) random variables. See e.g., Brockwell and Davis (1987, pp. 214–215).

and $\sqrt{T}\hat{\rho}_j$ are asymptotically normally distributed. Thus, the plot of $\hat{\rho}_j$ against j provides a rough diagnostic check of the property of the underlying process. When $\hat{\rho}_j$ of a time series become close to zero from a particular lag, say q, this series may be an MA(q) process; when $\hat{\rho}_j$ exhibit an exponentially decaying (damped) pattern, this series may be an AR process.

A well known result of sample autocorrelations is that, for those j such that $\rho_j \approx 0$, the variances of the corresponding sample autocorrelations are:

$$\operatorname{var}(\hat{\rho}_j) \approx \frac{1}{T}\sum_{i=-\infty}^{\infty} \rho_i^2;$$

see e.g., Fuller (1996, p. 318). In particular, for an MA(q) process,

$$\operatorname{var}(\hat{\rho}_j) \approx \frac{1}{T}\big(1 + 2\rho_1^2 + \cdots + 2\rho_q^2\big), \qquad j = q+1, q+2, \ldots.$$

This result, together with the asymptotic normality of $\sqrt{T}\hat{\rho}_j$, suggest that one may construct the 95% confidence interval of $\hat{\rho}_j$ using

$$\pm\frac{1.96}{\sqrt{T}}\big(1 + 2\hat{\rho}_1^2 + \cdots + 2\hat{\rho}_q^2\big)^{1/2};$$

replacing 1.96 with 1.645 in the bounds above yields the 90% confidence interval. The $\hat{\rho}_j$ falling within this interval is then considered not significantly different from zero. Note also that for a white noise,

$$\operatorname{var}(\hat{\rho}_j) \approx 1/T, \qquad j = 1, 2, \ldots,$$

so that $\pm 1.96/\sqrt{T}$ ($\pm 1.645/\sqrt{T}$) form the 95% (90%) confidence interval.

Remarks:

1. For convenience, many existing programs simply use $\pm 1.96/\sqrt{T}$ (or $\pm 2/\sqrt{T}$) as the 95% confidence interval. From the discussion above one can see that these bounds are in fact appropriate only for checking the autocorrelations of a white noise.

2. The confidence interval obtained above is for checking a *single* sample autocorrelation. When m sample autocorrelations are examined jointly, such an interval is not appropriate because it results in a significance level much larger than 5% (or 10%). To perform a joint test, a confidence region taking into account the variance-covariance structure of these sample autocorrelations is needed.

In addition to checking autocorrelations, the model identification in the Box-Jenkins approach also evaluates the the *partial autocorrelation* function. The partial autocorrelation α_j of a covariance-stationary time series y_t is defined as

$$\begin{aligned}\alpha_1 &= \operatorname{corr}(y_t, y_{t-1}) = \rho_1,\\ \alpha_m &= \operatorname{corr}\big[y_t - \mathrm{M}(y_t \mid \mathcal{Y}_{t-m+1}^{t-1}),\ y_{t-m} - \mathrm{M}(y_{t-m} \mid \mathcal{Y}_{t-m+1}^{t-1})\big], \qquad m = 2, 3, \ldots,\end{aligned}$$

where $\mathrm{M}(y_t \mid \mathcal{Y}_{t-m+1}^{t-1})$ is the *linear projection* of y_t on the space of $1, y_{t-1}, \ldots, y_{t-m+1}$, in the sense that it minimizes the mean squared error:

$$\mathbb{E}\big[y_t - (a_0 + a_1 y_{t-1} + \cdots + a_{m-1} y_{t-m+1})\big]^2.$$

Thus, the m^{th} partial autocorrelation is simply the correlation between y_t and y_{t-m}, after the effects of $y_{t-1}, \ldots, y_{t-m+1}$ are excluded. By the Frisch-Waugh-Lovell Theorem, the partial correlation coefficient α_m is also the last coefficient of the linear projection of y_t on $1, y_{t-1}, \ldots, y_{t-m}$.

Suppose that y_t is an AR(p) process. Then, y_t are correlated with y_{t-m} for $m \leq p$, so that the last coefficient of the linear projection of y_t on $1, y_{t-1}, \ldots, y_{t-m}$ should be different from zero; otherwise, y_t are uncorrelated with y_{t-m}, and the last coefficient of the linear projection above must be zero. That is, the partial autocorrelations of an AR(p) process exhibit an abrupt cut-off at the lag p. On the other hand, all the partial autocorrelations α_m of an invertible MA process are non-zero because this process has an AR(∞) representation. Similar to autocorrelations, α_m also approaches zero exponentially fast when m becomes large. In practice, the first m sample partial autocorrelations $\hat{\alpha}_i$ are the coefficient estimates of $a_{i,i}$ of the following regressions:

$$\begin{aligned}y_t &= a_{1,0} + a_{1,1} y_{t-1} + e_t,\\ y_t &= a_{2,0} + a_{2,1} y_{t-1} + a_{2,2} y_{t-2} + e_t,\\ y_t &= a_{3,0} + a_{3,1} y_{t-1} + a_{3,2} y_{t-2} + a_{3,3} y_{t-3} + e_t,\\ &\vdots\\ y_t &= a_{m,0} + a_{m,1} y_{t-1} + a_{m,2} y_{t-2} + \cdots + a_{m,m} y_{t-m} + e_t.\end{aligned}$$

The plot of $\hat{\alpha}_i$ against i thus provides a rough diagnostic check of the underlying process and may be used to determine the order of an AR process. It can also be shown that

$$\sqrt{T}\operatorname{var}(\hat{\alpha}_i) \xrightarrow{D} \mathcal{N}(0,1). \qquad m = p+1, p+2, \ldots;$$

see e.g., Brockwell and Davis (1987, p. 234). To determine whether $\hat{\alpha}_i$ is sufficiently close to zero, one may construct a confidence interval ($\pm 1.96/\sqrt{T}$ or $\pm 1.645/\sqrt{T}$) for $\hat{\alpha}_i$, analogous to that for sample autocorrelations.

For a stationary and invertible ARMA(p,q) process, both autocorrelations and partial autocorrelations gradually decay to zero and do not have abrupt cut-off points. Identifying ARMA orders using these functions is therefore difficult and somewhat arbitrary.

4.3 Model Estimation

When a preliminary ARMA(p,q) model $\Psi(\mathcal{B})y_t = c+\Pi(\mathcal{B})\varepsilon_t$ is chosen, it remains to estimate the unknown parameters, including the AR parameters $\psi_1,\ldots,\psi_p$, MA parameters $\pi_1,\ldots,\pi_q$, the constant term c, and the variance σ_ε^2. Let $\boldsymbol{\theta}$ denote the vector of these unknown parameters. A typical estimation method is the method of *qausi-maximum likelihood*; the resulting estimator of $\boldsymbol{\theta}$ is known as the quasi-maximum likelihood estimator (QMLE).

We first discuss the estimation of the AR(p) model: $\Psi(\mathcal{B})y_t = c+\varepsilon_t$. To implement the method of quasi-maximum likelihood, it is typical to postulate a condition density function $f(y_t \mid \boldsymbol{Y}^{t-1};\boldsymbol{\theta})$, where $\boldsymbol{Y}^{t-1} = (y_1, y_2, \ldots, y_{t-1})'$. Corresponding to this conditional density, let $f(\boldsymbol{Y}^t;\boldsymbol{\theta})$ denote the joint density function of $\boldsymbol{Y}^t$. In what follows, we may omit $\boldsymbol{\theta}$ in these density functions so as to simplify notations. Given these densities, the joint likelihood function is

$$\begin{aligned} L_T(\boldsymbol{Y}^T;\boldsymbol{\theta}) &= f(y_T \mid \boldsymbol{Y}^{T-1}) f(\boldsymbol{Y}^{t-1}) \\ &= f(y_T \mid \boldsymbol{Y}^{T-1}) f(y_{T-1} \mid \boldsymbol{Y}^{T-2}) f(\boldsymbol{Y}^{t-2}) \\ &\vdots \\ &= \Big(\prod_{j=p+1}^{T} f(y_j \mid \boldsymbol{Y}^{j-1}) \Big) f(\boldsymbol{Y}^p). \end{aligned}$$

To obtain the QMLE of $\boldsymbol{\theta}$, one maximizes the average of the log-likelihood function:

$$\mathcal{L}_T(\boldsymbol{Y}^T;\boldsymbol{\theta}) = \frac{1}{T}\ln L_T(\boldsymbol{Y}^T;\boldsymbol{\theta}) = \frac{1}{T}\Big(\ln f(\boldsymbol{Y}^p) + \sum_{j=p+1}^{T} \ln f(y_j \mid \boldsymbol{Y}^{j-1}) \Big).$$

Note that the postulated density functions need not be the true conditional densities. This is why the resulting estimators are referred to as QMLEs, rather than MLEs.

Assuming that the conditional distributions of y_t given $\boldsymbol{Y}^{t-1}$ are normal, we have

$$f(y_t \mid \boldsymbol{Y}^{t-1};\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(\frac{-(y_t - c - \psi_1 y_{t-1} - \cdots - \psi_p y_{t-p})^2}{2\sigma_\varepsilon^2} \right).$$

Taking the initial $y_1,\ldots,y_p$ as given, we can ignore the joint density $f(\boldsymbol{Y}^p;\boldsymbol{\theta})$ and maximize

$$\mathcal{L}_T^c(\boldsymbol{Y}^T;\boldsymbol{\theta}) = \frac{1}{T}\sum_{j=p+1}^{T} \ln f(y_j \mid \boldsymbol{Y}^{j-1};\theta).$$

It is now easy to see that, conditional on the initial values $y_1, \ldots, y_p$, the resulting QMLEs of $c, \psi_1, \ldots, \psi_p$ are nothing but the OLS estimators, and the QMLE of σ_ε^2 is

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{j=p+1}^{T} \hat{e}_t^2,$$

where $\hat{e}_t$ are the OLS residuals. Such estimators are also known as the *conditional* QMLEs of the AR(p) model.

When $f(\boldsymbol{Y}^p; \boldsymbol{\theta})$ is taken into account, we have under the normality assumption,

$$\begin{aligned} f(\boldsymbol{Y}^p; \boldsymbol{\theta}) =& (2\pi\sigma_e^2)^{-p/2} \det(\boldsymbol{V}_p)^{-1/2} \\ & \exp\left[\frac{-1}{2\sigma_\varepsilon^2}\Big(\boldsymbol{Y}^p - \frac{c}{1-\psi_1-\cdots-\psi_p}\boldsymbol{\ell}\Big)' \boldsymbol{V}_p^{-1}\Big(\boldsymbol{Y}^p - \frac{c}{1-\psi_1-\cdots-\psi_p}\boldsymbol{\ell}\Big)\right], \end{aligned}$$

where $\boldsymbol{\ell}$ is the p-dimensional vector of ones,

$$\sigma_\varepsilon^2 \boldsymbol{V}_p = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{pmatrix},$$

and γ_j are the autocovariances of the AR(p) process. The log-likelihood function is now a complex nonlinear function in unknown parameters and must be solved numerically. The resulting QMLEs are also know as the the *exact* QMLEs.

The estimation of MA and ARMA models is more involved because the innovations ε_t are not observable and can only be computed recursively. We consider the MA(1) model: $y_t = \mu + \varepsilon_t - \pi_1\varepsilon_{t-1}$. Given $\varepsilon_0 = 0$, ε_t can be obtained recursively as

$$\varepsilon_t = y_t - \mu + \pi_1\varepsilon_{t-1},$$

so that they depend only on observed y_t and unknown parameters. That is, $\varepsilon_1 = y_1 - \mu$, $\varepsilon_2 = y_2 - \mu + \pi_1(y_1 - \mu)$, and so on. Under the normality assumption,

$$\begin{aligned} f(y_t \mid \boldsymbol{Y}^{t-1}, \varepsilon_0 = 0; \boldsymbol{\theta}) &= f(y_t | u_{t-1}, \varepsilon_0 = 0; \boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(\frac{-(y_t - \mu + \pi_1\varepsilon_{t-1})^2}{2\sigma_\varepsilon^2}\right), \end{aligned}$$

and the quasi-log-likelihood function conditional on $\varepsilon_0 = 0$ is s

$$\mathcal{L}(\boldsymbol{Y}^T \mid \varepsilon_0 = 0; \boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_\varepsilon^2) - \frac{1}{T}\sum_{t=1}^{T} \frac{(y_t - \mu + \pi_1\varepsilon_{t-1})^2}{2\sigma_\varepsilon^2}.$$

It is clear that plugging the recursive formulae of ε_t results in a highly nonlinear function in parameters. Thus, this likelihood function must be maximized using a numerical algorithm; the resulting solution is the conditional QMLE of the MA(1) model.

Similarly, the conditional QMLE of the MA(q) model is obtained from maximizing the likelihood function conditional on $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0$, and ε_t are computed via the following recursions:

$$\varepsilon_t = y_t - \mu + \pi_1 \varepsilon_{t-1} + \cdots + \pi_q \varepsilon_{t-q}.$$

We omit the details. To compute the conditional QMLE of the ARMA(p,q), we need p initial values of $y_0, y_{-1}, \ldots, y_{-p+1}$ and q initial values of $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1}$, and ε_t are computed via

$$\varepsilon_t = y_t - c - \psi_1 y_{t-1} - \cdots - \psi_p y_{t-p} + \pi_1 \varepsilon_{t-1} + \cdots + \pi_q \varepsilon_{t-q}.$$

It is typical to set the initial ε's to zero, as in the case for MA models, and set the initial y's to the estimates of the expected value $c/(1 - \psi_1 - \cdots - \psi_p)$. For more details of the conditional QMLE and exact QMLE of ARMA(p, q) models, we refer to Hamilton (1994).

4.4 Asymptotic Properties of the QMLE

Recall that the conditional QMLE and the exact QMLE of ARMA models handle initial values in different ways. If the underlying process is indeed weakly stationary, the effect of initial values would eventually die out. This suggest that these two estimators should have the same asymptotic properties, yet their finite-sample properties may be quite different. In this section, we simply use $\tilde{\boldsymbol{\theta}}_T$ to denote the (conditional or exact) QMLE of $\boldsymbol{\theta}$ and sketch its asymptotic properties.

The QMLE $\tilde{\boldsymbol{\theta}}_T$ maximizes the average of the quasi-log-likelihood function and solves the average of the score $\nabla \mathcal{L}_T(\boldsymbol{Y}^T; \tilde{\boldsymbol{\theta}}_T) = \boldsymbol{o}$. Let $\boldsymbol{\theta}^*$ denote the unknown parameter vector that solves

$$\nabla \, \mathbb{E}\big[\mathcal{L}_T(\boldsymbol{Y}^T; \boldsymbol{\theta})\big] = \boldsymbol{o}.$$

Also let

$$\boldsymbol{H}_T(\boldsymbol{\theta}) = \mathbb{E}[\nabla^2 \mathcal{L}_T(\boldsymbol{Y}^T; \boldsymbol{\theta})],$$
$$\boldsymbol{B}_T(\boldsymbol{\theta}) = \operatorname{var}(\sqrt{T} \nabla \mathcal{L}_T(\boldsymbol{Y}^T; \boldsymbol{\theta}))$$

Under suitable regularity conditions that ensure a strong uniform law of large numbers and a central limit theorem, $\tilde{\boldsymbol{\theta}}_T$ is strongly consistent for $\boldsymbol{\theta}^*$, and

$$\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{-1/2} \boldsymbol{H}_T(\boldsymbol{\theta}^*) \sqrt{T} (\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{o}, \boldsymbol{I}).$$

Letting $\boldsymbol{C}_T(\boldsymbol{\theta}^*) = \boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}\boldsymbol{B}_T(\boldsymbol{\theta}^*)\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{-1}$, we have

$$\boldsymbol{C}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{o}, \boldsymbol{I}).$$

When the model is correctly specified, the *information matrix equality* holds: $\boldsymbol{H}_T(\boldsymbol{\theta}^*) + \boldsymbol{B}_T(\boldsymbol{\theta}^*) = \boldsymbol{o}$. In this case, $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$ simplifies to

$$\boldsymbol{C}_T(\theta^*) = -\boldsymbol{H}_T(\boldsymbol{Y}^T; \boldsymbol{\theta}^*)^{-1} = \boldsymbol{B}_T(\boldsymbol{\theta}^*)^{-1},$$

so that

$$\boldsymbol{B}_T(\boldsymbol{\theta}^*)^{1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\boldsymbol{H}_T(\boldsymbol{\theta}^*)^{1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{o}, \boldsymbol{I}).$$

Thus, the QMLE is asymptotically efficient in the sense that its asymptotic covariance matrix $\boldsymbol{C}_T(\theta^*)$ reaches the Cramér-Rao lower bound asymptotically.

In the traditional time series analysis, the information matrix equality is taken for granted. Thus, one only has to consistently estimate either $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ or $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$. It is straightforward to estimate $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ using its sample counterpart: the Hessian matrix of the quasi-log-likelihood function evaluated at $\tilde{\boldsymbol{\theta}}_T$, viz., $\tilde{\boldsymbol{H}}_T = \nabla^2\mathcal{L}_T(\boldsymbol{Y}^T; \tilde{\boldsymbol{\theta}}_T)$. With this estimator, we have

$$-\tilde{\boldsymbol{H}}_T^{1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{o}, \boldsymbol{I});$$

this result is the foundation of likelihood-based tests for model parameters (e.g., tests of parameter significance). Unfortunately, the information matrix equality need not hold in practice. For example, this equality breaks down when an ARMA(p, q) model is estimated but the data series in in fact an ARMA(p', q') process with $p' > p$ and/or $q' > q$. That is, the information matrix equality would not hold when important dynamic structures are ignored in model specification. In this case, $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$ can not be simplified; both $\boldsymbol{H}_T(\boldsymbol{\theta}^*)$ and $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$ must be estimated to form a consistent estimator for $\boldsymbol{C}_T(\boldsymbol{\theta}^*)$. In particular, a Newey-West type estimator is usually needed to estimate $\boldsymbol{B}_T(\boldsymbol{\theta}^*)$.

4.5 Diagnostic Checking

Based on the asymptotic normality result of the QMLE, one can easily test model parameters using the Wald, Lagrange Multiplier, and likelihood ratio tests. Other than these tests, it is also typical to conduct diagnostic checks of the residual series of an estimated ARMA model. An intuition of such diagnostic checks is that, when the estimated model correctly captures the dynamics of y_t, its residuals $\tilde{e}_t$ ought to be "clean" and behave like a white noise.

To test whether the residual series is close to a white noise, it is natural to check the sample autocorrelations of $\tilde{e}_t$:

$$\hat{\rho}_j^r = \sum_{t=1}^{T-j} \frac{(\tilde{e}_t - \bar{e})(\tilde{e}_{t+j} - \bar{e})}{\sum_{t=1}^{T}(\tilde{e}_t - \bar{e})^2},$$

where $\bar{e}$ is the sample average of ARMA residuals. Note that the superscript r of $\hat{\rho}_j^r$ is used to signify that this sample autocorrelation is computed from residuals. A joint test of m sample autocorrelations being zero is the so-called *Box-Pierce test*:

$$\mathcal{Q}_T^r = T\sum_{j=1}^{m}\left(\hat{\rho}_j^r\right)^2,$$

whose asymptotic null distribution is $\chi^2(m-p-q)$. A modified version, known as the *Ljung-Box test*, is

$$\widetilde{\mathcal{Q}}_T^r = T(T+2)\sum_{j=1}^{m}\frac{\left(\hat{\rho}_j^r\right)^2}{T-j}.$$

It is easily seen that the asymptotic behavior of $\widetilde{\mathcal{Q}}_T^r$ would be the same as $\mathcal{Q}_T^r$ because the scaling factors does not matter asymptotically. Thus, the asymptotic null distribution of the Ljung-Box test is also $\chi^2(m-p-q)$. We will discuss these two tests in more details when we discuss the tests of the martingale difference property.

The structure of ARMA models may also be determined using *model selection criteria*. Two leading choices of such criteria are the *Akaike Information Criterion* (AIC) and *Schwartz Information Criterion* (SIC):

$$\text{AIC} = \ln\tilde{\sigma}_T^2 + \frac{2(p+q+1)}{T},$$
$$\text{SIC} = \ln\tilde{\sigma}_T^2 + \frac{(p+q+1)\ln T}{T},$$

where $\tilde{\sigma}_T^2$ is the QMLE of σ_ε^2. These two criteria are in effect the Gaussian log-likelihood values penalized by model complexity (in terms of the number of model parameters). The SIC is also known as the *Bayesian Information Criterion* (BIC). In practice, one may estimate an array of ARMA models and choose the one with the smallest AIC (SIC) as the "best" model. Observe that the SIC penalizes a model more heavily and usually results in a simpler model in finite samples. Note that the SIC has the property of "dimensional consistency," in the sense that it is able to select the correct ARMA orders when the sample is sufficiently large. Although the AIC is not dimensionally consistent, this does not imply that the AIC is not useful. When a more complex model is chosen by the AIC, it still includes the "correct" model as a special case. In this case, the QMLE remains consistent, yet it may be less efficient than that of the correct model.

5 Volatility Models

Volatilty is a crucial determinant of asset pricing. A model that can properly characterize the volatility of asset returns is thus of paramount importance in the finance research. There are some stylized facts about financial variables. First, financial time series usually exhibit *volatility clustering*, in the sense that large (small) changes are followed by large (small) changes, in either sign. Second, the number of outliers of these variables are more than what a normal distribution can describe. This suggests that the marginal distributions of these variables have thicker tails than a normal distribution. Moreover, volatility asymmetry and changing volatility patterns are also quite common in many financial time series. Researchers therefore try to construct volatility models that are able to accommodate these features. In this section, we will present several leading volatility models in the literature.

5.1 ARCH Models

The autoregressive conditional heteroskedasticity (ARCH) model introduced by Engle (1982) was a first attempt in econometrics to capture volatility clustering in time series data. In particular, Engle (1982) used conditional variance to characterize volatility and postulate a dynamic model for conditional variance. We will discuss some generalizations and modifications of the ARCH model in subsequent sections; for a comprehensive review of this class of models we refer to Bollerslev, Chou, and Kroner (1992).

To illustrate the properties of an ARCH process, we first consider a very simple, weakly stationary ARCH(1) process: $y_t = \sqrt{h_t}\, u_t$, where u_t are i.i.d. random variables with mean zero and variance one, and

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2, \qquad \alpha_0 > 0,\ \alpha_1 \geq 0.$$

Note that h_t depend on the past information contained in the information set $\mathcal{F}^{t-1}$. The conditional mean of y_t given $\mathcal{F}^{t-1}$ is

$$\mathbb{E}(y_t \mid \mathcal{F}^{t-1}) = \sqrt{h_t}\, \mathbb{E}(u_t \mid \mathcal{F}^{t-1}) = \sqrt{h_t}\, \mathbb{E}(u_t) = 0.$$

Hence, the conditional variance of y_t is

$$\mathbb{E}(y_t^2 \mid \mathcal{F}^{t-1}) = h_t\, \mathbb{E}(u_t^2 \mid \mathcal{F}^{t-1}) = h_t\, \mathbb{E}(u_t^2) = h_t.$$

As h_t change with y_{t-1}^2, y_t are conditionally heteroskedastic. Writing

$$y_t^2 = h_t + (y_t^2 - h_t) = \alpha_0 + \alpha_1 y_{t-1}^2 + h_t(u_t^2 - 1),$$

we obtain an AR(1) representation of y_t^2 with the innovations $h_t(u_t^2 - 1)$ which have zero mean and are serially uncorrelated. That is, there are correlations among squared y_t when the ARCH effect is present.

By the law of iterated expectations, it is clear that $\mathbb{E}(y_t) = \mathbb{E}[\mathbb{E}(y_t \mid \mathcal{F}^{t-1})] = 0$ and

$$\operatorname{var}(y_t) = \mathbb{E}(h_t) = \alpha_0 + \alpha_1 \operatorname{var}(y_{t-1}).$$

The weak stationarity of y_t then implies that $\operatorname{var}(y_t) = \alpha_0/(1-\alpha_1)$. The autocovariances of y_t are

$$\mathbb{E}(y_t y_{t-j}) = \mathbb{E}\left[\sqrt{h_t h_{t-j}}\, u_{t-j}\, \mathbb{E}(u_t \mid \mathcal{F}^{t-1})\right] = 0, \qquad j = 1, 2, \ldots.$$

This shows that y_t are serially uncorrelated, yet they are not independent because y_t^2 are serially correlated.

Assuming that y_t are conditionally normally distributed, more can be said about their marginal distribution. Under conditional normality, $\mathbb{E}(y_t^4 \mid \mathcal{F}^{t-1}) = 3h_t^2$, so that

$$\mathbb{E}(y_t^4) = 3\big[\alpha_0^2 + 2\alpha_0\alpha_1\, \mathbb{E}(h_t) + \alpha_1^2\, \mathbb{E}(y_{t-1}^4)\big].$$

When $\mathbb{E}(y_t^4)$ is a constant m_4, we have

$$m_4 = 3\alpha_0^2\Big(1 + \frac{2\alpha_1}{1-\alpha_1}\Big) + 2\alpha_0\alpha_1\, \mathbb{E}(h_t) + 3\alpha_1^2 m_4,$$

or equivalently,

$$m_4 = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

This implies that $0 \le \alpha_1^2 < 1/3$. The kurtosis coefficient of y_t is then

$$\frac{m_4}{\operatorname{var}(y_t)^2} = 3\,\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3.$$

This result indicates that the marginal distribution of y_t is *leptokurtic* and has thicker tails than a normal distribution. Consequently, even when y_t are conditionally normally distributed, the resulting ARCH(1) process can not be a Gaussian white noise.

A novel feature of this simple ARCH process is its multiplicative form. Expressing y_t as a product of $\sqrt{h_t}$ and u_t is quite convenient for modeling the behavior of conditional variance. An immediate generalization is the ARCH(p) process: $y_t = \sqrt{h_t}\, u_t$, with the conditional variance:

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2, \qquad \alpha_0 > 0,\ \alpha_1, \ldots, \alpha_p \ge 0.$$

This process also results in an AR(p) representation of y_t^2:

$$y_t^2 = h_t + (y_t^2 - h_t) = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2 + h_t(u_t^2 - 1),$$

where the innovations $h_t(u_t^2 - 1)$ are serially uncorrelated. In this case, $\{y_t\}$ is a white noise with mean zero and $\mathrm{var}(y_t) = \alpha_0/(1 - \alpha_1 - \cdots - \alpha_p)$. It can also be shown that the marginal distribution of y_t is still leptokurtic under conditional normality. An even more general process is the AR(p_1)-ARCH(p_2) process:

$$y_t = c + \psi_1 y_{t-1} + \cdots + \psi_{p_1} y_{t-p_1} + \varepsilon_t,$$

where $\varepsilon_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_{p_2} y_{t-p_2}^2, \qquad \alpha_0 > 0,\ \alpha_1, \ldots, \alpha_{p_2} \geq 0.$$

The process $\{y_t\}$ now has a more complex conditional mean and is no longer a white noise.

5.2 GARCH Models

In many applications it was found that a high order ARCH model is usually needed to describe the dynamics of conditional variances. A natural way to generalize the ARCH model is to consider an ARMA representation of y_t^2, which in turn leads to the generalized ARCH (GARCH) model of Bollerslev (1986).

To illustrate, we consider the GARCH(1,1) process: $y_t = \sqrt{h_t}\, u_t$, with the conditional variance:

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}, \qquad \alpha_0 > 0,\ \alpha_1, \beta_1 \geq 0.$$

This process has the following ARMA(1,1) representation of y_t^2 with the innovations $h_t(u_t^2 - 1)$:

$$y_t^2 = h_t + (y_t^2 - h_t) = \alpha_0 + (\alpha_1 + \beta_1) y_{t-1}^2 + h_t(u_t^2 - 1) - \beta_1 h_{t-1}(u_{t-1}^2 - 1).$$

This again shows that there are correlations among squared y_t. Clearly, y_t have mean zero and

$$\mathrm{var}(y_t) = \mathbb{E}(h_t) = \alpha_0 + \alpha_1\, \mathbb{E}(y_{t-1}^2) + \beta_1\, \mathbb{E}(h_{t-1}).$$

Weak stationarity again implies that

$$\mathrm{var}(y_t) = \frac{\alpha_0}{1 - (\alpha_1 + \beta_1)}.$$

Thus, $\alpha_1+\beta_1$ must be less than one to ensure a finite variance. It is also easy to see that the autocovariances of y_t and y_{t-j}, $j=1,2,\ldots$, are also zero, so that $\{y_t\}$ is still a white noise. Moreover, the kurtosis coefficient is, under conditional normality,

$$\frac{m_4}{\mathrm{var}(y_t)^2} = 3\,\frac{1-(\alpha_1+\beta_1)^2}{1-(\alpha_1+\beta_1)^2-2\alpha_1^2} > 3,$$

provided that $1-(\alpha_1+\beta_1)^2-2\alpha_1^2>0$. This shows that the GARCH(1,1) process has a leptokurtic marginal distribution when y_t are conditionally normally distributed. In fact, this result holds even when y_t are not conditionally normally distributed; for more details see Section 3.14 of Tsay (2002).

The more general GARCH(p,q) process is: $y_t=\sqrt{h_t}\,u_t$, with the conditional variance:

$$h_t = \alpha_0 + \sum_{i=1}^{p}\alpha_i y_{t-i}^2 + \sum_{j=1}^{q}\beta_j h_{t-j}, \qquad \alpha_0>0,\ \alpha_i,\beta_j\geq 0.$$

This leads to the following ARMA representation of y_t^2:

$$y_t^2 = \alpha_0 + \sum_{i=1}^{\max(p,q)}(\alpha_i+\beta_i)y_{t-i}^2 + h_t(u_t^2-1) - \sum_{j=1}^{q}\beta_j h_{t-j}(u_{t-j}^2-1),$$

where we set $\alpha_i=0$ if $i>p$ and $\beta_i=0$ if $i>q$. We also have $\mathbb{E}(y_t)=0$,

$$\mathrm{var}(y_t) = \frac{\alpha_0}{1-\sum_{i=1}^{\max(p,q)}(\alpha_i+\beta_i)},$$

provided that $1-\sum_{i=1}^{\max(p,q)}(\alpha_i+\beta_i)>0$, and zero autocovariances. We may, of course, construct an AR(p_1)-GARCH(p_2,q) process:

$$y_t = c+\psi_1 y_{t-1}+\cdots+\psi_{p_1}y_{t-p_1}+\varepsilon_t,$$

where $\varepsilon_t=\sqrt{h_t}\,u_t$, with

$$h_t = \alpha_0 + \sum_{i=1}^{p_2}\alpha_i y_{t-i}^2 + \sum_{j=1}^{q}\beta_j h_{t-j}, \qquad \alpha_0>0,\ \alpha_i,\beta_j\geq 0.$$

An ARMA(p_1,q_1)-GARCH(p_2,q_2) process is also possible.

Although there are many empirical studies suggesting that a GARCH(1,1) model suffices to describe the conditional variances of a wide variety of financial time series, we note that this need not be true. For example, Chen and Kuan (2002) demonstrated that many well known diagnostic tests have little power against volatility asymmetry and hence lead to an incorrect conclusion on the selected GARCH model. As far as the GARCH(1,1) model is concerned, it is also quite common to observe that the sum

of the estimated α_1 and β_1 is close to one. If $\alpha_1 + \beta_1$ is indeed one, $\mathrm{var}(y_t)$ would be unbounded, and the process of y_t^2 has a unit root. Such a process is referred to as an *integrated* GARCH (IGARCH) process. Taking into account the constraint that $\alpha_1 + \beta_1 = 1$, the IGARCH(1,1) process is $y_t = \sqrt{h_t}\, u_t$, with the conditional variance:

$$h_t = \alpha_0 + (1 - \beta_1) y_{t-1}^2 + \beta_1 h_{t-1}, \qquad \alpha_0 > 0,\ 0 < \beta_1 < 1.$$

The IGARCH phenomenon is, however, difficult to interpret because the unconditional variance of y_t^2 are growing with t. Some researchers recently argue that the IGARCH result may be a consequence of ignoring structural changes (level shifts) in conditional variance. They find that when a model admits two or more regimes of conditional variances, the resulting estimates of $\alpha_1 + \beta_1$ in each regime is typically much less than one. Hence, the observed IGARCH effects in many empirical studies may be dubious.

Another important variant of the GARCH process is the GARCH-in-mean (GARCH-M) process, introduced by Engle, Lilien, and Robins (1987). By noting that asset returns may also depend on their volatility, they proposed the following GARCH(1,1)-M model:

$$y_t = c + \gamma h_t + \varepsilon_t,$$

with $\varepsilon_t = \sqrt{h_t}\, u_t$ and

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}, \qquad \alpha_0 > 0,\ \alpha_1, \beta_1 \geq 0,$$

where γ is called the risk premium parameter. Due to the presence of h_t in the equation of y_t, it is easy to see that y_t in this case are serially correlated. The GARCH(1,1)-M model above can be easily extended by adding an AR structure to the equation of y_t and/or allowing h_t to be a GARCH(p, q) process.

5.3 EGARCH Models

While the GARCH models are capable of capturing volatility clustering, there are some drawbacks of these models. First, the GARCH models are unable to represent volatility asymmetry. Due to the presence of lagged y_t^2 in the variance equation, the positive and negative values of the lagged innovations have the *same* effect on the conditional variance. In the finance literature, however, it has long been recognized that volatility often responds to positive and negative shocks in different ways. For example, Black (1976) observed that the volatility of stock returns tends to increase (decrease) when there is "bad news" ("good news"). Second, to ensure positiveness of h_t in the GARCH model, non-negative constraints are imposed on the coefficients in the variance equation. These constraints are convenient, yet they are not necessary.

Nelson (1991) considered the following weighted innovations:

$$g(u_t) = \theta_1 u_t + \gamma_1(|u_t| - \mathbb{E}\,|u_t|),$$

where $|u_t| - \mathbb{E}\,|u_t|$ are also i.i.d. random variables with mean zero, so that $g(u_t)$ have mean zero. When u_t are normally distributed, for example, we have $\mathbb{E}\,|u_t| = \sqrt{2/\pi}$. Note that $g(u_t)$ can be represented as a threshold function:

$$g(u_t) = \begin{cases} (\theta_1 + \gamma_1)u_t - \gamma_1\,\mathbb{E}\,|u_t|, & u_t \geq 0, \\ (\theta_1 - \gamma_1)u_t - \gamma_1\,\mathbb{E}\,|u_t|, & u_t < 0. \end{cases}$$

Hence, $g(u_t)$ is linear in u_t with slope $\theta_1 + \gamma_1$ when u_t are non-negative, and $g(u_t)$ has slope $\theta_1 - \gamma_1$ when u_t are negative. It should be noted that the asymmetric response of g to u_t is due to θ_1, rather than γ_1.

To avoid the non-negativity constraints on the coefficients in the variance equation, Nelson (1991) proposed the *exponential* GARCH (EGARCH) model in which h_t is an exponential function of lagged h_t and the weighted innovation $g(u_{t-1})$. Specifically, a simple EGARCH(1,1) process is $y_t = \sqrt{h_t}\,u_t$, with the conditional variance:

$$h_t = \exp\left[\alpha_0 + \beta_1 \ln(h_{t-1}) + \left(\theta_1 \frac{y_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \left|\frac{y_{t-1}}{\sqrt{h_{t-1}}}\right|\right)\right];$$

note that $\mathbb{E}\,|u_{t-1}|$ has been absorbed into the constant term. The coefficient θ_1 is usually interpreted as a measure of the "leverage" effect of u_{t-1}, while γ_1 is interpreted as a measure of the "magnitude" effect. In empirical studies, the estimate of θ_1 is typically negative while γ_1 is positive, showing that positive shocks have less impact on volatility.

Although this EGARCH process is still able to capture volatility clustering, it differs from GARCH processes in the following respects. First, the conditional variance of the EGARCH process responds differently to positive and negative innovations.[2] Second, due to the presence of exponential function, an innovation with larger magnitude has much larger impact on h_t. Moreover, there is no constraint on the coefficients in h_t.

Extending the EGARCH(1,1) process above to the EGARCH(p, q) process is straightforward: $y_t = \sqrt{h_t}\,u_t$, with

$$h_t = \exp\left[\alpha_0 + \sum_{i=1}^{q} \beta_i \ln(h_{t-i}) + \sum_{j=1}^{p} \left(\theta_j \frac{y_{t-j}}{\sqrt{h_{t-j}}} + \gamma_j \left|\frac{y_{t-j}}{\sqrt{h_{t-j}}}\right|\right)\right],$$

[2]Engle and Ng (1993) defined the "news impact curve" as the relationship between the conditional variance h_t and u_{t-1}, holding constant the information on and before time $t-2$ (lagged conditional variances are evaluated at the unconditional variance). Based on this idea, it is easy to see that the news impact curve of a GARCH process is symmetric, but that of an EGARCH process is asymmetric.

where θ_j and γ_j characterize the asymmetry and magnitude effects of the shock u_{t-j} on the volatility h_t. One may also add an AR structure to the mean equation of y_t; allowing h_t to enter the mean equation results in an EGARCH-M process.

5.4 GJR-GARCH Models

Focusing on the impacts of positive and negative shocks, Glosten, Jegannathan, and Runkle (1993) proposed a modified GARCH model, now known as the GJR-GARCH model, that also possesses an asymmetric news impact curve. Comparing to the EGARCH model, the GJR-GARCH model utilizes a threshold function to capture volatility asymmetry, but it does not impose an exponential function on the conditional variance equation. In fact, an EGARCH model may generate unreasonably large conditional variances because of the exponential function.

A simple GJR-GARCH(1,1) process is $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \beta_1 h_{t-1} + (\alpha_1 + \theta_1 D_{t-1}) y_{t-1}^2,$$

where $D_{t-1} = 1$ when $y_{t-1} < 0$ and $D_{t-1} = 0$ otherwise. Clearly, h_t of the GJR process also respond differently to positive and negative y_{t-1}; this process reduces to a standard GARCH(1,1) process if $\theta_1 = 0$. This model is thus capable of capturing both volatility clustering and volatility asymmetry. Without the exponential function, this model does not yield ridiculous conditional variances. Non-negativity constraints on the coefficients in h_t are still needed, however. It is also straightforward to extend the simple GJR-GARCH model to an AR(p_1)-GJR-GARCH(p_2, q) model or a GJR-GARCH(p, q)-M model. Finally, we note that the asymmetry effects of positive and negative shocks identified by Glosten, Jegannathan, and Runkle (1993) are different from those based on the EGARCH model.

5.5 Implementing GARCH Models

In estimating GARCH models, one must determine the conditional distribution of y_t (or ε_t). A standard approach is to postulate a conditional normal distribution. Although conditional normality results in a leptokurtic marginal distribution, it can not fully account for the outliers in real data. It is now also typical to assume y_t (or ε_t) in a GARCH model to have a conditional t distribution. Recall that a t distribution with ν degrees of freedom has variance $\nu/(\nu-2)$ when $\nu > 2$, and it does not have a finite variance when $\nu \leq 2$. Normalizing y_t to have conditional variance one yields the density of u_t:

$$f(u) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{(\nu-2)\pi}} \left(1 + \frac{u^2}{\nu-2}\right)^{-(\nu+1)/2},$$

where Γ is the Gamma function such that $\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}\,\mathrm{d}y$.

We now consider the estimation of an ARCH(p) model. Similar to the estimation of an AR(p) model, it is simpler to drop initial p observations and estimate the parameters by maximizing the average of the conditional quasi-log-likelihood function. Under conditional normality, the QMLE is obtained by maximizing

$$\mathcal{L}_T = -\frac{T-p}{2T}\ln(h_t) - \frac{1}{2T}\sum_{i=p+1}^{T}\frac{y_t^2}{h_t},$$

where $h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2$. For an AR($p_1$)-ARCH($p_2$) model, we maximize

$$\mathcal{L}_T = -\frac{T-p^*}{2T}\ln(h_t) - \frac{1}{2T}\sum_{i=p^*+1}^{T}\frac{(y_t - c - \psi_1 y_{t-1} - \cdots - \psi_{p_1} y_{t-p_1})^2}{h_t},$$

where $p^* = \max(p_1, p_2)$ and $h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p_2}^2$. Under conditional t distribution, we simply substitute the t density function for the normal density. We omit the details of estimating GARCH models.

Another commonly used distribution is the generalized error distribution; for example, Nelson (1991) considered this distribution for EGARCH models. Given this assumption, we may normalize y_t to have conditional mean zero and conditional variance one and obtain the following density of u_t:

$$f(u) = \frac{\nu\,\exp[-|u/\lambda|^\nu/2]}{\lambda\, 2^{1+1/\nu}\,\Gamma(1/\nu)}, \qquad \nu > 0,$$

where ν is the parameter characterizing the thickness of tails and

$$\lambda = \left[2^{-2/\nu}\,\Gamma(1/\nu)\,\Gamma(3/\nu)\right]^{1/2}.$$

This is a standard normal distribution when $\nu = 2$; for $\nu < 2$ ($\nu > 2$), the tails of the distribution of u are thicker (thinner) than the normal distribution. In particular, it is a double exponential distribution when $\nu = 1$ and a uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ when $\nu \to \infty$; see Nelson (1991). Using the density of u_t we obtain the quasi-log-likelihood function of y_t, from which the QMLE can be computed numerically.

For a GARCH (EGARCH) model, let $\hat{h}_t$ denote the estimated conditional variances and $\hat{e}_t$ denote the model residuals when there is a mean equation. To examine whether an estimated GARCH model is appropriate, one may evaluate the standardized residuals: $y_t/\hat{h}_t^{1/2}$ (when there is no mean equation) or $\hat{e}_t/\hat{h}_t^{1/2}$ (when there is a mean equation). A Ljung-Box Q test on the standardized residuals is usually used to check if the mean equation is appropriate; a Q test on the squares of the standardized residuals is used to

evaluate the variance equation; see e.g., Tsay (2002, p. 89). It is also typical to employ a test of independence, such as the BDS test of Brock et al. (1987), to determine if the standardized residuals are independent. For example, Bollerslev, Chou and Kroner (1992) concluded that: "most studies tend to find that once ARCH effects are removed the BDS test on standardized residuals exhibits very little evidence of nonlinear dependence" (pp. 22–23). Chen and Kuan (2002) showed, however, that neither the Q-type tests nor the BDS test is powerful enough to detect neglected volatility asymmetry. Thus, one tends to accept the null hypothesis of no correlations (or no dependence) when these tests are applied to the standardized residuals of GARCH and EGARCH models. That is, these tests can not distinguish between the GARCH and EGARCH models. By contrast, the test of time reversibility considered by Chen and Kuan (2002), which is also a test of independence, overcomes this problem. As far as testing volatility asymmetry is concerned, the "sign bias" test and the "positive (negative) size bias" test of Engle and Ng (1993) may also be used.

5.6 Stochastic Volatility Models

The stochastic volatility (SV) model is an important alternative to the GARCH models and has attracted much attention recently. A simple stochastic SV process is $y_t = \sqrt{h_t}\, u_t$, with

$$\ln(h_t) = \alpha_0 + \alpha_1 \ln(h_{t-1}) + v_t, \qquad |\alpha_1| < 1,$$

where v_t are random variables such that $\{v_t\}$ and $\{u_t\}$ are independent of each other. A novel feature of SV processes is that the conditional variances h_t are driven by a different set of innovations v_t, whereas h_t of GARCH processes are not. The inclusion of new innovations v_t admits more flexibility in the model but also renders model estimation difficult. Similar to an EGARCH process, this process also postulates the conditional variance h_t as an exponential function of past information so that no non-negativity constraint on the coefficients is needed.

Assuming that u_t are independent $\mathcal{N}(0,1)$ random variables and v_t are independent $\mathcal{N}(0,\sigma_v^2)$, we have

$$\ln(h_t) \sim \mathcal{N}\left(\frac{\alpha_0}{1-\alpha_1}, \frac{\sigma_v^2}{1-\alpha_1^2}\right).$$

Clearly, $\mathbb{E}(y_t) = 0$. Knowing the mean and variance of the lognormal random variable,

it is easy to derive the unconditional second and fourth moments of y_t as:[3]

$$\mathbb{E}(y_t^2) = \mathbb{E}(h_t)\,\mathbb{E}(u_t^2) = \exp\left(\frac{\alpha_0}{1-\alpha_1} + \frac{\sigma_v^2}{2(1-\alpha_1^2)}\right),$$

$$\mathbb{E}(y_t^4) = \mathbb{E}(h_t^2)\,\mathbb{E}(u_t^4) = 3\,\exp\left(\frac{2\alpha_0}{1-\alpha_1} + \frac{2\sigma_v^2}{1-\alpha_1^2}\right).$$

Thus, y_t are also leptokurtic because

$$\mathbb{E}(y_t^4)/[\mathbb{E}(y_t^2)]^2 = 3\,\exp\left(\frac{\sigma_v^2}{1-\alpha_1^2}\right) > 3.$$

Moreover, y_t are serially uncorrelated, while y_t^2 are serially correlated. Allowing u_t and v_t to be correlated produces valatility asymmetry, similar to that of an EGARCH process.

The estimation of an SV model is typically cumbersome; see e.g., Jacquier, Polson, and Rossi (1994), Harvey, Ruiz, and Shephard (1994), and Harvey and Shephard (1996). Let Y^T denote the collection of all y_t and h^T the collection of all conditional variances h_t. Then, the density of Y^T is

$$P(Y^T) = \int P(Y^T, h^T)\,\mathrm{d}h^T = \int P(Y^T|h^T)\,P(h^T)\,\mathrm{d}h^T,$$

which is a mixture over the density of h^T. Difficulty in estimation arises because a T-dimensional integral must be evaluated. The Markov chain Monte Carlo (MCMC) method suggested by Jacquier, Polson, and Rossi (1994) avoids this difficulty. Other estimation methods are the method of quasi-maximum likelihood and the generalized method of moment; we omit the details.

5.7 Realized Volatility

The parametric volatility models discussed in the preceding sections have their limitations. First, it is hard to evaluate the performance of the estimated conditional variances because the true conditional variances (or the volatility in general) are not observable. Second, different volatility models often yield quite different volatility patterns. As such, the results of parametric volatility models are not robust and hence are vulnerable. Note that neither squared y_t (when there is no mean equation) nor squared $\hat{e}_t$ (when there is a mean equation) is a good approximation to unobserved conditional variances. Therefore, a model-free estimate of conditional variance is highly desirable and may serve as a benchmark of unobserved volatility. The *realized volatility* proposed by Andersen, Bollerslev, Diebold, and Ebens (2001) is such an estimate.

[3]Note that there is a typo in the formula of $\mathbb{E}(y_t^2)$ in Tsay (2002, p. 110).

A standard diffusion model of the logarithm of the asset price p_t is

$$\mathrm{d}p_t = \mu_t \,\mathrm{d}t + \sigma_t \,\mathrm{d}W_t,$$

where μ_t is the drift term, σ_t is the diffusion, and W is a standard Wiener process. Let $r_{t,m} = p_t - p_{t-m}$ denote the m-period returns which is the sum of m one-period returns. The conditional distribution of $r_{t+1,1}$ is

$$\mathcal{N}\left(\int_0^1 \mu_{t+s}\,\mathrm{d}s,\ \int_0^1 \sigma_{t+s}^2\,\mathrm{d}s\right).$$

Let $[c]$ denote the largest integer less than c. Partition the time between t and $t+1$ into $m = [1/\delta]$ non-overlapping sub-periods, each with the length δ (say, 1 minute or 5 minutes). Then, $r_{t+1,1}$ is the one-day return and also the sum of m δ-period returns:

$$r_{t+1,1} = r_{t+\delta,\delta} + r_{t+2\delta,\delta} + \cdots + r_{t+1,\delta} = \sum_{j=1}^{[1/\delta]} r_{t+j\delta,\delta}.$$

The conditional variance of $r_{t+1,1}$ is thus the conditional variance of the sum of m returns. When the partition become finer (i.e., $\delta \to 0$, or equivalently, $m \to \infty$), it is well known that the quadratic variations of r are such that

$$\sum_{j=1}^{[1/\delta]} r_{t+j\delta,\delta}^2 \to \int_0^1 \sigma_{t+s}^2\,\mathrm{d}s,$$

in the almost sure sense.

This convergence result suggests that $\sum_{j=1}^{[1/\delta]} r_{t+j\delta,\delta}^2$ serves as a natural estimate of the conditional variance of $r_{t+1,1}$. Andersen et al. (2001) refer to this estimate as the realized volatility of $r_{t+1,1}$.[4] The realized daily volatility can then be computed as the sum of squared returns of intraday data at higher frequencies (say, squared 5-minute returns). When $\boldsymbol{r}_{t+1,1}$ are vectors of asset returns, one may also define the "realized variance-covariance matrix" as

$$\sum_{j=1}^{[1/\delta]} (\boldsymbol{r}_{t+j\delta,\delta})(\boldsymbol{r}_{t+j\delta,\delta})',$$

where each diagonal term is the realized volatility of the corresponding asset return. This approach was also considered by French, Schwert, and Stambaugh (1987) in computing

[4] Andersen et al. (2001) claimed that the realized volatility is model free. This is not entirely correct because the convergence of the quadratic variations holds under the diffusion model, i.e., r_t are driven by a standard Wiener process which has independent increments. Without this condition, the correlations between the sub-period returns may not be ignored.

the volatility of monthly returns based on daily returns, yet they estimate both the volatility of daily returns and their covariances.

Andersen, Bollerslev, Diebold, and Ebens (2001) found by visual inspection that the unconditional distribution of the stock returns standardized by the realized standard deviations is approximately normal. This is in contrast with the common wisdom that the stock returns standardized by the standard deviations of a GARCH model are leptokurtic in general. They also found that the unconditional distribution of the realized variances is highly skewed to the right, but the logarithm of the unconditional distribution of the logarithm of the realized standard deviations is also approximately normal. This result suggests that the unconditional distribution of stock returns may be well approximated by a continuous lognormal-normal mixture.

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility, *Journal of Financial Economics*, **61**, 43–76.

Black, F. (1976). Studies of stock price volatility changes, Proceedings of the 1976 meeting of the American Statistical Associations, Business and Economic Statistics Section, 177–181.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, **31**, 307–327.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992). ARCH modeling in finance, *Journal of Econometrics*, **52**, 5–59.

Brock, W., W. D. Dechert, and J. Scheinkman (1987). A test for independence based on the correlation dimension, Working paper, Department of Economics, University of Wisconsin, Madison.

Brockwell, P. J. and R. A. Davis (1987). *Time Series: Theory and Methods*, New York, NY: Springer-Verlag.

Chen, Y.-T. and C.-M. Kuan (2002). Time irreversibility and EGARCH effects in U.S. stock index returns, *Journal of Applied Econometrics*, **17**, 565–578.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50**, 987–1006.

Engle, R. F., D. M. Lilien, and R. P. Robins (1987). Estimating time-varying risk premia in the term structure: the ARCH-M model, *Econometrica*, **55**, 391–407.

Engle, R. F. and V. K. Ng (1993). Measuring and testing the impact of news on volatility, *Journal of Finance*, **48**, 1749–1778.

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987). Expected stock returns and volatility, *Journal of Financial Economics*, **19**, 3–29.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, second edition, New York, NY: Wiley.

Glosten, L., R. Jegannathan, and D. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance*, **48**, 1779–1801.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton, NJ: Princeton University Press.

Harvey, A. C., E. Ruiz, and N. Shephard (1994). Multivariate stochastic variance models, *Review of Economic Studies*, **61**, 247–264.

Harvey, A. C. and N. Shephard (1996). Estimation of an asymmetric stochastic volatility model for asset returns, *Journal of Business & Economic Statistics*, **14**, 429–434.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models, *Journal of Business & Economic Statistics*, **12**, 371–389.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach, *Econometrica*, **59**, 347–370.

Pesaran, H. H. and Y. Shin (1998). Generalized impulse response analysis in linear multivariate models, *Economics Letters*, **58**, 17–29.

Tsay, R. S. (2002). *Analysis of Financial Time Series*, New York, NY: John Wiley & Sons.