Time Series Analysis (I)

MEI-YUAN CHEN

Department of Finance
National Chung Hsing University

Feb. 26, 2013
# Contents

1 Introduction
   1.1 Linear Time Series Models . . . . . . . . . . . . . . . . . . . . . . . . . 3
   1.2 Stochastic Processes . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   1.3 Stationary Processes . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
   1.4 The Autocorrelation Function and Partial Autocorrelation Functions . . 7
   1.5 Some Useful Stochastic Processes . . . . . . . . . . . . . . . . . . . . . 7
      1.5.1 Random Walk . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
      1.5.2 Moving Average Processes . . . . . . . . . . . . . . . . . . . . . 8
   1.6 Autocovariance Functions of a MA($q$) . . . . . . . . . . . . . . . . . 8
   1.7 Autoregressive Processes . . . . . . . . . . . . . . . . . . . . . . . . . 11
      1.7.1 First-order Process . . . . . . . . . . . . . . . . . . . . . . . . . . 11
      1.7.2 Autoregressive Moving Average Models . . . . . . . . . . . . . . . 12
   1.8 Autoregressive Integrated Moving Average Models . . . . . . . . . . . . 12

2 Model Identification
   2.1 Estimations for the Time Dependence . . . . . . . . . . . . . . . . . . . 14

3 Model Estimation
   3.1 MLE of an AR(1) Model . . . . . . . . . . . . . . . . . . . . . . . . . . 15
   3.2 MLE for a MA($q$) Model . . . . . . . . . . . . . . . . . . . . . . . . . 16

4 Check White Noise for Filtered Residuals
   4.1 Portmanteau Test . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
   4.2 Testing That a Dependent Process Is Uncorrelated . . . . . . . . . . . . 19
      4.2.1 Testing $H_0 : \mu = 0$ . . . . . . . . . . . . . . . . . . . . . . 19
      4.2.2 Testing $H_0 : \gamma(1) = 0$ . . . . . . . . . . . . . . . . . . . . 20
   4.3 Selection for a Parsimonious Model . . . . . . . . . . . . . . . . . . . . 22
   4.4 Impulse Response Function . . . . . . . . . . . . . . . . . . . . . . . . . 22

5 Forecasting
   5.1 Combined Forecasts . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
   5.2 Forecast Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
   5.3 Test for Martingale Difference . . . . . . . . . . . . . . . . . . . . . . . 27
1 Introduction

A univariate random variable $Y$ is defined as

$$Y : \Omega \to \{y_1, y_2, \ldots\} = D_Y(\mu_Y, \text{var}(Y), \alpha_3(Y), \alpha_4(Y)).$$

The distribution of realizations is demonstrated by measures $E(Y)$ (unconditional mean), $\text{var}(Y)$ (unconditional variance), $\alpha_3(Y)$ (unconditional skewness), $\alpha_4(Y)$ (unconditional kurtosis). Of course, the distribution of all realizations can be captured by the cumulative distribution function $F_Y$ and the probability density function $f_Y$ if $F_Y$ is globally differentiable.

Bivariate random variables $X$ and $Y$ are defined as

$$(X, Y) : \Omega \to \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\} = D_{X,Y}(\mu_X, \sigma^2_X, \alpha_3(X), \alpha_4(X), \mu_Y, \sigma^2_Y, \alpha_3(Y), \alpha_4(Y), \text{cov}(X, Y)).$$

That is, the plot (distribution) of realizations $(x, y)$ can be demonstrated by measures $\mu_X, \sigma^2_X, \alpha_3(X), \alpha_4(X)$ for $X$, $\mu_Y, \sigma^2_Y, \alpha_3(Y), \alpha_4(Y)$ for $Y$, and $\text{cov}(X, Y)$ for the covariance between $X$ and $Y$.

The conditional random variable $Y$ on $X = x$ is defined as

$$(Y|X = x) : \Omega \to \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\} = D_{Y|X=x}(E(Y|X = x), \text{var}(Y|X = x), \alpha_3(Y|X = x), \alpha_4(Y|X = x)).$$

The distribution of realizations is demonstrated by measures $E(Y|X = x)$ (conditional mean), $\text{var}(Y|X = x)$ (conditional variance), $\alpha_3(Y|X = x)$ (conditional skewness), $\alpha_4(Y|X = x)$ (conditional kurtosis). Similarly, the distribution of all realizations can be captured by the cumulative distribution function $F_{Y|X=x}$ and the probability density function $f_{Y|X=x}$ if $F_{Y|X=x}$ is globally differentiable. The purpose of exploring the distribution of conditional random variable $(Y|X = x)$ is to check whether it is different from the distribution of unconditional random variable $Y$, i.e., $D_{Y|X=x} = D_Y$. For example, let $Y$ denote the today’s stock return and $X = 1$ for having positive and $X = 0$ for negative yesterday’s return. Then $D_{Y|X=x} = D_Y$ implies that knowing previous day’s return is positive or negative has no help for understanding the distribution of today’s stock return. In contrast, $D_{Y|X=x} \neq D_Y$ indicates that knowing yesterday’s return is positive or negative is useful to understanding the distribution of today’s return. The information of $X = 1$ or $X = 0$ is valuable to understand the distribution of today’s stock return. In other words, checking $D_{Y|X=x} = D_Y$ or not is equivalent to study whether the information $X$ is valuable or not. Without doubt, the one-dimensional random variable can be extended to be multiple dimensional and denoted
as \(X = [X_1, X_2, \ldots, X_k]'\). When \(X\) is multiple dimensional, the joint distributions \(D_{Y,X_1}, D_{Y,X_2}, \ldots, D_{Y,X_k}\) are explored by studying the conditional distribution \(D_{Y|X}\).

Usually \((Y|X = x)\) is decomposed as

\[
(Y|X = x) = E(Y|X = x) + \sqrt{\text{var}(Y|X = x)} \times e.
\]

It is clear that under the assumption of normality and homogeneity of \(D_{Y|X=x}\), the exploration of \(D_{Y|X=x}\) is reduced to study \(E(Y|X = x)\). Therefore, the null hypothesis of checking whether \(X\) is valuable to know the distribution \(D_y\), \(H_0 : D_{Y|X=x} = D_{Y}\), is reduced to be \(H_0 : E(Y|X = x) = E(Y)\). In statistical analysis, a random sample \(\{(y_i, X_i)', i = 1, \ldots, n\}\) is used to study \(E(Y|X = x)\) under the assumption that \(E(Y|X = x)\) is “fixed” for all realizations (also for sample observations.) In econometric analysis, the content of \(X\) are suggested by the corresponding economic theory and are selected by using model selection criteria in regression analysis.

In a time series analysis, the information contained in \(X\) is the \(\sigma\)-algebra generated by the historical values of \(Y\). That is, the conditional random variable of \(Y_t\) on \(\mathcal{F}_{t-1}\) is the interest, where

\[
\mathcal{F}_{t-1} := \{Y_{t-1}, Y_{t-2}, \ldots\}.
\]

That is the conditional random variable \((Y_t|\mathcal{F}_{t-1})\) is under investigating. In other words, the joint distributions \(D_{Y_t, Y_{t-1}}, D_{Y_t, Y_{t-2}}, \ldots, D_{Y_t, Y_{t-\infty}}\) are going to be studied. To be investigatable, \(D_{Y_t, Y_{t-1}}, D_{Y_t, Y_{t-2}}, \ldots, D_{Y_t, Y_{t-p}}\) is considered for a finite number of \(p\). This is also equivalent to study the joint distributions \(D_{Y_t, Y_1}, D_{Y_t, Y_2}, \ldots, D_{Y_t, Y_{-p}}\). These joint distributions can be demonstrated by the measures \(E(Y_t), \sigma_{Y_t}^2, \alpha_3(Y_t), \alpha_4(Y_t)\), and \(\text{cov}(Y_t, Y_{-j})\) for all \(t\) and \(j = 1, \ldots, k\). Therefore, there are plenty of measures under estimating if they are time-varying, i.e., different for different \(t\). In reality, there is only a time series sample with \(T\) observations, \(\{y_1, y_2, \ldots, T\}\), obtained and used to do the estimations. It is necessary to assume that these measures are all time-invariant. This is the necessity of “strict stationary” for doing time series analysis. Given the condition of stationarity, the joint distribution \(D_{Y_t, Y_{t-1}, \ldots, Y_{t-p}}\) becomes time invariant and so does the distribution of conditional random variable \((Y_t|Y_{t-1}, \ldots, Y_{t-p})\). This also implies that the conditional mean \(E(Y_t|Y_{t-1}, \ldots, Y_{t-p})\) and conditional variance \(\text{var}(Y_t|Y_{t-1}, \ldots, Y_{t-p})\) are also constant over time. Under the assumption of normality, \(E(Y_t|Y_{t-1}, \ldots, Y_{t-p})\) and \(\text{var}(Y_t|Y_{t-1}, \ldots, Y_{t-p})\) are the two measures to be studied for exploring \(D_{Y_t|Y_{t-1}, \ldots, Y_{t-p}}\).
1.1 Linear Time Series Models

In conventional time series analysis, the assumptions of normality and homogeneity (\( \text{var}(Y_t|Y_{t-1}, \ldots, Y_{t-p}) \) is constant for all time \( t \) and all values of \( Y_{t-1}, \ldots, Y_{t-p} \)) are imposed on \( D_{Y_t|Y_{t-1}, \ldots, Y_{t-p}} \). As a result, the conditional mean, \( E(Y_t|Y_{t-1}, \ldots, Y_{t-p}) \), becomes the only object to be studied for exploring \( D_{Y_t|Y_{t-1}, \ldots, Y_{t-p}} \). For simplicity as in linear regression analysis, the linearity is always assumed to model \( E(Y_t|Y_{t-1}, \ldots, Y_{t-p}) \).

That is,

\[
E(Y_t|Y_{t-1}, \ldots, Y_{t-p}) = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_k Y_{t-p},
\]

which is called the **linear autoregressive model**. To be more general, an **autoregressive and moving average** (ARMA model, hereafter) is suggested as

\[
E(Y_t|F_{t-1}) = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_k Y_{t-p} + \pi_1 e_{t-1} + \cdots + \pi_q e_{t-q}.
\]

It is crucial to determine the numbers of \( p \) and \( q \) before doing model estimation.

For financial time series analysis, it is common to recognize that the distribution of \( (Y_t|F_{t-1}) \) is not normal and not homogeneous. Therefore, \( D_{Y_t|Y_{t-1}, \ldots, Y_{t-p}} \) can not be explored completely by studying \( E(Y_t|Y_{t-1}, \ldots, Y_{t-p}) \) only.

1.2 Stochastic Processes

Mathematically, a stochastic process \( \{Y_t, -\infty < t < \infty\} \) may be defined as a collection of random variables \( (Y_t) \) which are ordered in time and defined on a random experiment (information set, \( \Omega_t \)) at a set of time points which may be continuous \( (t \in \mathbb{R}) \) or discrete \( (t = 0, \pm 1, \pm 2, \ldots) \). Definitions of ten types of stochastic process:

1. **Strictly stationary**: The multivariate distribution function for \( k \) consecutive variables does not depend on the time subscript attached to the first variable (any \( k \)).

2. **Stationary**: Means and variances do not depend on time subscripts, covariances depend only on the difference between two subscripts.

3. **Uncorrelated**: The correlation between variables having different time subscripts is always zero.

4. **Autocorrelated**: It is not uncorrelated.

5. **White noise**: The variables are uncorrelated, stationary and have mean equal to 0.
6 Strict white noise: The variables are independent and have identical distributions whose mean is equal to 0.

7 Martingale: The expected value of variable at time $t$, conditional on the information provided by all previous values, equals variables at time $t - 1$.

8 Martingale difference: The expected value of a variable at period $t$, conditional on the information provided by all previous values, always equals 0.

9 Gaussian: All multivariate distributions are multivariate normal.

10 Linear: It is a linear combination of the present and past terms from a strict white noise process.

In time series analysis, a stationary time series data $\{y_t, t = 1, \ldots, T\}$ is used to explore the joint distribution $Y_1, Y_2, \ldots, Y_T$, for examples,

$$E(Y_t), \text{var}(Y_t), \text{cov}(Y_t, Y_s), t, s = 1, \ldots, T, t \neq s.$$ 

However, the linear conditional mean $E(Y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p})$ is focused. It is clear also that

$$E(Y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}) = \sum_{j=1}^{p} \alpha_j y_{t-j} = c$$

implies that $\alpha_j = 0, \forall j$. That is the values of previous times have no effects on the mean behavior of $Y_t$. We say that $y_{t-j}, j = 1, \ldots, p$ have no explanation (prediction) power on the mean of $Y_t$.

Most statistical problem are concerned with estimating the measures (or parameters) of a population from a sample. In time-series analysis there is a rather different situation in that, although it may be possible to vary the “length” of the observed time series– the sample– it is impossible to make more than one observation at any given time. Thus we only have a single outcome of the process and a single observation on the random variable at time $t$. Nevertheless we may regard the observed time series as just one example of the “infinite” set of time series which might have been observed. The infinite set of time series is sometimes called the “ensemble”. Every member of the ensemble is a possible “realization” of the stochastic process.

One way of describing a stochastic process is to specify the joint probability distribution of $Y_{t_1}, \ldots, Y_{t_n}$ for any set of times $t_1, \ldots, t_n$ and any value of $n$. However, a simpler, more useful way of describing a stochastic process is to give the moments of the process, particularly the first and second moments, which are called the mean, variance, and autocovariance functions.
1. Mean: The mean function $\mu_t$ is defined by $\mu_t = E(Y_t)$.

2. Variance: The variance function $\sigma^2_t$ is defined by $\sigma^2_t = \text{var}(Y_t)$.

3. Autocovariance: The covariance function $\gamma(Y_{t_1}, Y_{t_2})$ is defined by

$$\gamma(y_{t_1}, y_{t_2}) = E\{[Y_{t_1} - E(Y_{t_1})][Y_{t_2} - E(Y_{t_2})]\} = \gamma(\tau), \quad \tau = t_2 - t_1.$$ 

4. Autocorrelation: The autocorrelation function $\rho(Y_{t_1}, Y_{t_2})$ is defined by

$$\rho(y_{t_1}, y_{t_2}) = \frac{E\{[Y_{t_1} - E(Y_{t_1})][Y_{t_2} - E(Y_{t_2})]\}}{\sigma_{t_1}\sigma_{t_2}} = \frac{\gamma(\tau)}{\gamma(0)} = \rho(\tau), \quad \tau = t_2 - t_1.$$ 

5. Partial Autocorrelation: The partial autocorrelations measures the excess correlation between $Y_t$ and $Y_{t-p}$ which is not accounted for the correlations among $Y_{t-1}, \ldots, Y_{t-p+1}$.

1.3 Stationary Processes

A time series is said to be “strictly stationary” if the joint distribution of $Y_{t_1}, \ldots, Y_{t_n}$ is the same as the joint distribution of $Y_{t_1+\tau}, \ldots, Y_{t_n+\tau}$ for all $t_1, \ldots, t_n$ and $\tau$. In other words, shifting the time origin by an amount $\tau$ has no effect on the joint distribution. In particular, if $n = 1$, strictly stationary implies that the distribution of $Y_t$ is the same for all $t$, so that provided the first two moments are finite, we have $\mu_t = \mu$ and $\sigma^2_t = \sigma^2$ for all $t$. Furthermore, if $n = 2$, the joint distribution of $Y_{t_1}$ and $Y_{t_2}$ depends only on $(t_2 - t_1)$, which is called the “lag”. $\rho(\tau)$ is called the “autocorrelation function” at lag $\tau$.

In practice it is often useful to define stationarity in a less restricted way than the strictly stationary. A process is called “second-order (weakly) stationary” if its mean is constant and its ACF depends only on the lag, so that $E(Y_t) = \mu$ and $\text{cov}(Y_{t}, Y_{t+\tau}) = \gamma(\tau)$. No assumptions are made about the higher moments than those of second order. By letting $\tau = 0$, the variance is implied to be constant.
1.4 The Autocorrelation Function and Partial Autocorrelation Functions

Suppose a stationary stochastic process $Y_t$ has mean $\mu$, variance $\sigma^2$, autocovariance functions $\gamma(\tau)$ and autocorrelation function $\rho(\tau)$. Then

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \frac{\gamma(\tau)}{\sigma^2}.$$ 

Note that $\rho(0) = 1$. Some properties of autocorrelation function:

1. The autocorrelation function is an even function of the lag $\tau$ in that $\rho(\tau) = \rho(-\tau)$.
2. $|\rho(\tau)| \leq 1$.
3. Lack of uniqueness. Although a given stochastic process has a unique covariance structure, the converse is not in general true.

Consider the regressions of $Y_t$ on $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$,

$$Y_t = \alpha_{01} + \alpha_{11} Y_{t-1} + e_t$$
$$Y_t = \alpha_{02} + \alpha_{12} Y_{t-1} + \alpha_{22} Y_{t-2} + e_t$$
$$Y_t = \alpha_{03} + \alpha_{13} Y_{t-1} + \alpha_{23} Y_{t-2} + \alpha_{33} Y_{t-3} + e_t$$
$$\vdots$$

$$Y_t = \alpha_{0p} + \alpha_{1p} Y_{t-1} + \alpha_{2p} Y_{t-2} + \alpha_{3p} Y_{t-3} + \cdots + \alpha_{pp} Y_{t-p} + e_t$$

The coefficients $\alpha_{11}, \alpha_{22}, \ldots, \alpha_{pp}$ are called the partial autocorrelation coefficients of $\{Y_t\}$. Denote the PACF (partial autocorrelation function) as a function of the partial autocorrelation coefficients on the number of lag $k$, i.e., $PACF(k) = \alpha_{kk}$.

1.5 Some Useful Stochastic Processes

A discrete-time process is called a purely random process or white noise if it consists of a sequence of random variable $\{Z_t\}$ which are mutually independent and identically distributed. That is the process has constant mean and variance and that

$$\gamma(k) = \text{cov}(Z_t, Z_{t+k}) = 0 \text{ for } k = \pm1, \pm2, \ldots, \ldots.$$ 

As the mean and ACF do not depend on time, the process is second-order stationary. In fact it is clear that the process is also strictly stationary. The ACF is given by

$$\rho(k) = \begin{cases} 
1 & k = 0 \\
0 & k = \pm1, \pm2, \ldots 
\end{cases}$$

As $\{Z_t\}$ is mutually independent, the partial autocorrelation coefficients $PACF(k) = \alpha_{kk} = 0, \forall k$. 

7
1.5.1 Random Walk

Suppose that \( \{Z_t\} \) is a discrete, purely random process with mean \( \mu \) and variance \( \sigma_Z^2 \).

A process \( \{Y_t\} \) is said to be a random walk if

\[
Y_t = Y_{t-1} + Z_t.
\]

The process is customarily started at zero when \( t = 0 \), so that \( Y_1 = Z_1 \) and \( Y_t = \sum_{i=1}^{t} Z_i \). Then we find that \( E(Y_t) = t\mu \) and that \( \text{var}(Y_t) = t\sigma_Z^2 \). As the mean and variance change with \( t \), the process is non-stationary. However, it is interesting to note that the first difference of a random walk, given by

\[
\triangle Y_t = Y_t - Y_{t-1} = Z_t
\]

form a purely random process, which is therefore stationary.

1.5.2 Moving Average Processes

A process \( \{Y_t\} \) is said to be a moving average process of order \( q \) (denoted as MA (\( q \))) if

\[
Y_t = Z_t + \pi_1 Z_{t-1} + \cdots + \pi_q Z_{t-q}
\]

where \( \{\pi_i\} \) are constants. It is clear that

\[
E(Y_t) = 0 \\
\text{var}(Y_t) = \sigma_Z^2 \sum_{i=1}^{q} \pi_i^2, \quad \text{with } \pi_0 = 1
\]

since the \( \{Z_t\} \)s are independent.

1.6 Autocovariance Functions of a MA(\( q \))

Besides, we have

\[
\gamma(k) = \text{cov}(Y_t, Y_{t+k}) = \text{cov}(Z_t + \pi_1 Z_{t-1} + \cdots + \pi_q Z_{t-q}, Z_{t+k} + \pi_1 Z_{t+k-1} + \cdots + \pi_q Z_{t+k-q})
\]

\[
= \begin{cases} 
0 & k > q \\
\sigma_Z^2 \sum_{i=1}^{q-k} \pi_i \pi_{i+k} & k = 0, 1, \ldots, q \\
\gamma(-k) & k < 0 
\end{cases}
\]
since
\[
\text{cov}(Z_s, Z_t) = \begin{cases} 
\sigma^2_Z & s = t \\
0 & s \neq t.
\end{cases}
\]

As \(\gamma(k)\) does not depend on \(t\), and the mean is constant, the process is second-order stationary for all values of \(\{\pi_i\}\). The ACF of the MA\((q)\) process is given by
\[
\rho(k) = \begin{cases} 
1 & k = 0 \\
\sum_{i=1}^{q-k} \pi_i \pi_{i+k} / (1 + \sum_{i=1}^{q} \pi_i^2) & k = 1, \ldots, q \\
0 & k > q \\
\rho(-k) & k < 0
\end{cases}
\]

Note that the ACF “cuts off” at lag \(q\), which is a special feature of MA processes. In particular the MA\((1)\) process has an ACF given by
\[
\rho(k) = \begin{cases} 
1 & k = 0 \\
\pi_1 / (1 + \pi_1^2) & k = \pm 1 \\
0 & \text{otherwise}
\end{cases}
\]

No other restriction on the \(\{\pi_i\}\) are required for a (finite-order) MA process to be stationary, but it is generally desirable to impose restrictions on the \(\{\pi_i\}\) to ensure that the process satisfies a condition called “invertibility”.

Consider the following first-order MA processes:

\(A\) : \(Y_t = Z_t + \theta Z_{t-1}\)

\(B\) : \(Y_t = Z_t + \frac{1}{\theta} Z_{t-1}\).

It can easily be shown that these two different processes have exactly the same ACF. Thus we cannot identify an MA process from a given ACF. Now, if we express processes \(A\) and \(B\) by putting \(Z_t\) in terms of \(Y_t, Y_{t-1}, \ldots\), we find by successive substitution that

\(A\) : \(Z_t = Y_t - \theta Y_{t-1} + \theta^2 Y_{t-2} - \cdots\)

\(B\) : \(Z_t = Y_t - \frac{1}{\theta} Y_{t-1} + \frac{1}{\theta^2} Y_{t-2} - \cdots\)

If \(|\theta| < 1\), the process \(A\) converges whereas that the process \(B\) does not. Thus if \(|\theta| < 1\), process \(A\) is said to be invertible whereas process \(B\) is not. The imposition of the invertibility condition ensures that there is a unique MA process for a given ACF.

The invertibility condition for the general-order MA process is best expressed by using the backward shift operator, denoted by \(B\), which is defined by
\[
B^j Y_t = Y_{t-j} \quad \forall j.
\]
Then the general-order MA process may be written as

\[ Y_t = (1 + \pi_1 B + \cdots + \pi_q B^q)Z_t \]

where \( \pi(B) \) is a polynomial of order \( q \) in \( B \). An MA process of order \( q \) is invertible if the roots of the equation

\[ \pi(z) = 1 + \pi_1 z + \pi_2 z^2 + \cdots + \pi_q z^q = 0 \]

all lie outside the unit circle. For example, in the first-order case we have \( \pi(z) = 1 + \pi_1 z = 0 \), which has root \( z = -1/\pi_1 \). Thus the root is outside the unit circle, \(|z| > 1 \), provided that \(|\pi_1| < 1 \).

For a MA(1) process, \( Y_t = Z_t + \pi_1 Z_{t-1} \) may be written as \( Y_t = (1 + \pi_1 B)Z_t \) and then

\[ Z_t = \frac{1}{1 + \pi_1 B} Y_t = (1 - \pi_1 B + \pi_2 B^2 - \pi_3 B^3 + \cdots)Y_t = Y_t - \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots \]

or

\[ Y_t = \pi_1 Y_{t-1} - \pi_2 Y_{t-2} - \pi_3 Y_{t-3} + \cdots + Z_t \]

which is an AR(\( \infty \)) process. The partial autocorrelation function PACF(\( k \)), \( \alpha_{kk} \), will never “cut off” at any lag \( k \) but will decay exponentially to zero as \( k \) gets large.

This result can also be extended to the MA(q) process, since

\[ Y_t = (1 + \pi_1 B + \pi_2 B^2 + \cdots + \pi_q B^q)Z_t \]

can be expressed as

\[ Z_t = \frac{1}{1 + \pi_1 B + \pi_2 B^2 + \cdots + \pi_q B^q} Y_t = (1 - \psi_1 B - \psi_2 B^2 - \psi_3 B^3 - \cdots)Y_t = Y_t - \psi_1 Y_{t-1} - \psi_2 Y_{t-2} - \psi_3 Y_{t-3} - \cdots \]

or

\[ Y_t = \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \psi_3 Y_{t-3} + \cdots + Z_t \]

which is an AR(\( \infty \)) process. The PACF(\( k \)) will also never “cut off” at any lag \( k \) but will decay exponentially to zero as \( k \) gets large.

To summary, an MA(q) process has ACF(\( k \)) cutting off at lag \( k \) and PACF(\( k \)) decaying to zero exponentially.
1.7 Autoregressive Processes

A process \( \{Y_t\} \) is said to be an autoregressive process of order \( p \) if

\[
Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + Z_t,
\]

which is abbreviated to an AR(\( p \)) process.

1.7.1 First-order Process

When \( p = 1 \), \( Y_t = \phi_1 Y_{t-1} + Z_t \). The AR(1) process is sometimes called the Markov process, after the Russian A.A. Markov. By successive substitution, we have

\[
Y_t = \sum_{j=0}^{\infty} \phi_1^j Z_{t-j}
\]

which is an infinite-order MA process provided \(-1 < \phi_1 < 1\). As

\[(1 - \phi_1 B)Y_t = Z_t\]

so that

\[
Y_t = \frac{Z_t}{(1 - \phi_1 B)} = (1 + \phi_1 + \phi_1^2 + \cdots)Z_t = \sum_{j=0}^{\infty} \phi_1^j Z_{t-j}.
\]

It is clear that

\[
E(Y_t) = 0
\]

\[
\text{var}(Y_t) = \sigma_Z^2 \sum_{j=0}^{\infty} \phi_1^{2j} = \sigma_Z^2 / (1 - \phi_1^2)
\]

given \( |\phi_1| < 1 \). The ACF is given by

\[
\gamma(k) = E(Y_t Y_{t+k}) = \sigma_Z^2 \sum_{j=0}^{\infty} \phi_1^j \phi_1^{k+j} \text{ for } k \geq 0
\]

which converges for \( |\phi_1| < 1 \) to

\[
\gamma(k) = \phi_1^k \sigma_Z^2 / (1 - \phi_1^2) = \phi_1^k \sigma_Y^2.
\]
For $k < 0$, we find $\gamma(k) = \gamma(-k)$. Since $\gamma(k)$ does not depend on $t$, an AR(1) process is second-order stationary provided that $|\phi_1| < 1$. The ACF is given by

$$\rho(k) = \phi_1^k, \quad k = 0, 1, 2, \ldots$$

That is the ACF($k$) of an AR(1) process will not “cut off” at any lag $k$ but will decay to zero exponentially since $|\phi_1| < 1$. However, the PACF at $k = 1$, i.e., $\alpha_{11}$, equals to $\phi_1$ and $\alpha_{kk} = 0$ for all $k \geq 2$. More generally, the ACF($k$) of an AR($p$) process will not “cut off” at any lag $k$ but will decay to zero exponentially. However, the PACF at $k = p$, i.e., $\alpha_{pp}$, equals to $\phi_p$ and $\alpha_{kk} = 0$ for all $k \geq p + 1$.

### 1.7.2 Autoregressive Moving Average Models

More generally, a useful class of processes is formed by combining both AR and MA processes. A mixed autoregressive/moving-average process containing $p$ AR terms and $q$ MA terms is said to be an ARMA process of order $(p, q)$. That is

$$\phi(B)Y_t = \pi(B)Z_t,$$

where $\phi(B)$ and $\pi(B)$ are the $p$-th and $q$-th order polynomials in $B$. A process generated from an ARMA($p, q$) model is called an ARMA($p, q$) process; its stationarity and invertibility are determined by the roots of $\phi(z)$ and $\pi(z)$, respectively.

Since an ARMA($p, q$) process may be rewritten as

$$Y_t = [\pi(B)/\phi(B)]Z_t = \psi(B)Z_t,$$

an ARMA process can be viewed as a pure MA($\infty$) process. Similarly, an ARMA process may also be rewritten as

$$Z_t = [\pi(B)/\phi(B)]^{-1}Y_t = \varphi(B)Y_t,$$

an ARMA process can be viewed as a pure AR($\infty$) process. Then the ACF($k$) and PACF($k$) both will not “cut off” at any lag $k$ till $k \to \infty$.

### 1.8 Autoregressive Integrated Moving Average Models

In many applications it has been found that there are many economic time series that are non-stationary but their (first) differences are stationary. This leads to the class of autoregressive integrated moving average (ARIMA) processes. An ARIMA($p, d, q$) process $Y_t$ can be expressed as an ARMA($p, q$) process of $\triangle Y_t = (1 - B)^d Y_t$:

$$\phi(B)(1 - B)^d Y_t = \pi(B)Z_t.$$
The process \{Y_t\} is also called an integrated process of order \(d\), denoted as \(I(d)\), which must be differenced \(d\) times to achieve stationarity. In practice, the integrated time series we usually encounter are \(I(1)\) process.

Consider an ARIMA\((p, 1, q)\) process \(Y_t\) with \((1 - B)Y_t = \Delta Y_t\). By recursive substitution,

\[
Y_t = \sum_{j=0}^{t-1} \Delta Y_{t-j}
\]

has mean \(t\mu_{\Delta Y_t}\) and variance \(t\sigma_{\Delta Y_t}^2\). Hence, \(Y_t\) is non-stationary because its mean \(t\mu_{\Delta Y_t}\) and variance \(t\sigma_{\Delta Y_t}^2\) are growing with \(t\). More importantly, the autocorrelations of \(Y_t\) never die out. Also note that the polynomial \(\phi(z) = \pi(z)(1 - z)\) has a root on the unit circle, and \(\{Y_t\}\) is also called a unit root process with drift \(\mu_{\Delta Y_t}\). In particular, if \(\Delta Y_t\) are i.i.d., \(Y_t\) is a random walk with drift \(\mu_{\Delta Y_t}\).

\[2\text{ Model Identification}\]

In Econometric analyzes, the regression model is set to regress \(Y\) on \(X\) based on “some” economic theoretic results, i.e., \((Y|X = x) = m(x) = x'\beta_0\) for example. To do the model selection, some criteria such as \(R^2\) (or \(\bar{R}^2\), (adjusted) determinant coefficients), AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion, or Schwartz Criterion). However, in time series analysis, there is no theoretic reference result to be used to select a stochastic process and then to say an observed time series data is a realization from the selected stochastic process. A general model ARMA\((p,q)\) is considered first and then the appropriate orders of \(p\) and \(q\) are under identifying based on the time structures implied by observed time series data.

A time series model is set as

\[
y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + u_t + \sum_{j=1}^{q} u_{t-j}
\]

for an observed time series data \(\{y_t, t = 1, \ldots, T\}\). An appropriate time series model, the ARMA\((p,q)\) with appropriate \(p\) and \(q\), is identified by

1. the time structures (or time dependence), i.e., autocovariances (autocorrelations) and partial autocorrelation functions described by the time series model is similar to a stationary stochastic process;
2. the filtered errors from the appropriate model is white-noise.
3. the model is “parsimonious”.

13
2.1 Estimations for the Time Dependence

As the autocovariance (autocorrelation) and partial autocorrelation functions are used to describe the time dependence of a stochastic process. However the autocorrelation and partial autocorrelation functions of an observed time series data are unobservable. Fortunately, they are estimatable from an observed time series data under the condition of stationarity. Given an observed time series data, \( \{y_t, t = 1, \ldots, T\} \), which is assumed to be a realization of a stationary stochastic process. The sample autocorrelations can be estimated by

\[
\hat{\rho}(k) = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y}_T)(y_{t-k} - \bar{y}_{T-k})}{\sum_{t=k+1}^{T} (y_t - \bar{y}_T)^2}, \quad k = 1, 2, \ldots
\]

and that partial autocorrelations \( \hat{\alpha}_{kp} \) can be estimated by running the regression

\[
y_t = \alpha_{p1}y_{t-1} + \alpha_{p2}y_{t-2} + \cdots + \alpha_{kk}y_{t-k} + u_t, \quad k = 1, 2, \ldots
\]

For a stationary AR\((p)\) process, \( \rho(k) \) should decay toward zero as \( k \) increases, and \( \alpha_{kk} \) ought to have an abrupt cutoff with all values “close” to zero for \( k > p \). For an invertible MA\((q)\) process, \( \rho(k) \) have an abrupt cutoff with values “close” to zero for all \( k > q \), and sample partial autocorrelations gradually decay toward zero. Thus, by examining the plots of sample autocorrelations and partial autocorrelations we can determine the order of a pure AR or MA model. Unfortunately, the AR and MA orders of an ARMA process cannot be easily identified from the plots of sample autocorrelations and partial autocorrelations.

It has been shown that if \( y_t \) is generated from an AR\((p)\) process, sample partial autocorrelations \( \hat{\alpha}_{kk} \) are approximately normally distributed with mean zero and variance \( T^{-1} \) for \( k > p \), where \( T \) is the sample size. A simple diagnostic check is then based on the comparison of the estimates with \( \pm 2T^{-1/2} \); those estimates fall within these bounds are considered to be “close” enough to zero. If \( y_t \) is generated from an MA\((q)\) process, sample autocorrelations \( \hat{\rho}(k) \) are approximately normally distributed with mean zero and variance

\[
\text{var}(\hat{\rho}(k)) = \frac{1}{T} \left( 1 + \sum_{i=1}^{q} \rho(i)^2 \right), \quad k = q + 1, q + 2, \ldots
\]

It is thus not easy to implement a simple diagnostic test of autocorrelations, except for the case that the data are generated from a white noise series.
3 Model Estimation

After the diagnostic check for sample autocorrelation and partial autocorrelation functions, the orders of $p$ and $q$ are identified. The next step for establishing an appropriate time series model, the filtered error from subtracting $y_t$ from the value evaluated by the appropriate model should be white-noise. However, the filtered error is not observable since the appropriate model is unobservable as its coefficients are unknown. The proxy of filtered error can be obtained by subtracting $y_t$ from the “fitted” value of the estimated appropriate model. The proxy is also called the filtered residual. Therefore, estimation for the appropriate model is the next step.

An ARMA model is typically estimated using the method of maximum likelihood. Under the assumption of normality, it can be shown that the likelihood function of an ARMA($p,q$) model is non-linear in unknown parameters and more complicated than that of an AR or MA model. Hence, the maximum likelihood estimator (MLE) must be computed using non-linear optimization techniques.

3.1 MLE of an AR(1) Model

For an AR(1) model

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t, \ t = 1, \ldots, T$$

where $\{u_t\}$ is assumed to be a Gaussian white noise series. The joint likelihood function can be written as a product of conditional likelihood functions:

$$f(y_1, \ldots, y_T; \theta) = \prod_{t=2}^{T} f(y_t|y_{t-1}, \ldots, y_1; \theta)$$

where $\theta = [\phi_0 \ \phi_1 \ \sigma_u^2]^T$.

$$f(y_1; \theta) = \frac{1}{\sqrt{2\pi\sigma_u^2/(1-\phi_1^2)}} \exp \left( -\frac{(y_1 - \phi_0/(1-\phi_1)^2}{2\sigma_u^2/(1-\phi_1^2)} \right),$$

$$f(y_t|y_{t-1}; \theta) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp \left( -\frac{(y_t - \phi_0 - \phi_1 y_{t-1})^2}{2\sigma_u^2} \right).$$

The MLE of $\theta$ is obtained by maximizing:

$$L_T(y_1, \ldots, y_T; \theta)$$
\[ -\frac{T}{2} \log(\sigma_u^2) + \frac{1}{2} \log(1 - \phi_1^2) - \frac{(y_1 - \phi_0/(1 - \phi_1))^2}{2\sigma_u^2/(1 - \phi_1^2)} \]
\[ - \sum_{t=2}^{T} \frac{(y_t - \phi_0 - \phi_1 y_{t-1})^2}{2\sigma_u^2}. \]

Obviously, this is a non-linear optimization problem and can be solved using numerical methods.

Alternatively, we may maximize the likelihood function \textit{conditional} on the initial observation \(y_1\):
\[ \mathcal{L}_T(y_2, \ldots, y_T | y_1; \theta) \propto -\frac{T - 1}{2} \log(\sigma_u^2) - \sum_{t=2}^{T} \frac{(y_t - \phi_0 - \phi_1 y_{t-1})^2}{2\sigma_u^2}. \]

It is clear that the resulting MLEs \(\tilde{\phi}_0\) and \(\tilde{\phi}_1\) are nothing but the OLS estimators based on all but the first observations. The MLE of \(\sigma_u^2\) is
\[ \frac{1}{T - 1} \sum_{t=2}^{T} (y_t - \tilde{\phi}_0 - \tilde{\phi}_1 y_{t-1})^2. \]

Note that the difference between the exact MLE and conditional MLE is resulted from the treatment of the first observation; this difference is negligible when sample size \(T\) is sufficiently large.

Similarly, the exact MLEs of an AR(\(p\)) model can be obtained by maximizing the (non-linear) log-likelihood function. The conditional MLEs of AR coefficients are just the OLS estimators computed from the last \(T - p\) observations, and the MLE of \(\sigma_u^2\) is
\[ \frac{1}{T - p} \sum_{t=p+1}^{T} (y_t - \tilde{\phi}_0 - \tilde{\phi}_1 y_{t-1} - \cdots - \tilde{\phi}_p y_{t-p})^2. \]

### 3.2 MLE for a MA(\(q\)) Model

Consider now an MA(1) model: \(y_t = \pi_0 - \pi_1 u_{t-1} + u_t\), where \(\{u_t\}\) is also assumed to be a Gaussian white noise series. Let \(Y = [y_1 \ y_2 \ \cdots \ y_T]'\), \(\theta = [\pi_0 \ \pi_1 \ \sigma_u^2]'\). As \(Y\) has the covariance matrix:
\[ \Sigma = \sigma_u^2 \begin{bmatrix} 1 + \pi_1^2 & -\pi_1 & 0 & \cdots & 0 & 0 \\ -\pi_1 & 1 + \pi_1^2 & -\pi_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + \pi_1^2 \\ \end{bmatrix}, \]
which can be decomposed as \(ADA'\), \(D = \sigma_u^2 \text{diag}[d_1, d_2, \ldots, d_T]\) with
\[ d_t = \frac{\sum_{j=0}^{T} \pi_1^{2j}}{\sum_{j=0}^{T-1} \pi_1^{2j}}, \quad t = 1, \ldots, T, \]
and $A$ is a lower triangular matrix with ones on the main diagonal and $-\pi_1/d_t$, $t = 1, \ldots, T - 1$ on the first sub-diagonal.

The joint likelihood function can then be written as:

$$f(y_1, \ldots, y_T; \theta) \propto \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(Y - \pi_0\ell)'\Sigma^{-1}(Y - \pi_0\ell)\right)$$

$$= \det(D)^{-1/2} \exp\left(-\frac{1}{2}\tilde{Y}'D^{-1}\tilde{Y}\right)$$

$$= (\sigma_u^2)^{-T/2}(\prod_{t=1}^d t)^{-1/2} \exp\left(-\frac{1}{2\sigma_u^2} \sum_{t=1}^T \tilde{y}_t^2/d_t\right),$$

where $\tilde{Y} = A^{-1}(Y - \pi_0\ell)$ with elements $\tilde{y}_t$. The log-likelihood function is then

$$\mathcal{L}(Y; \theta) \propto -\frac{T}{2} \log(\sigma_u^2) - \frac{1}{2} \sum_{t=1}^T \log(d_t) - \frac{1}{2\sigma_u^2} \sum_{t=1}^T \tilde{y}_t^2/d_t.$$ 

This is a complex non-linear function in $\theta$ and must be maximized by non-linear optimization techniques.

Conditional on $u_0 = 0$ (say), we can compute $u_t = y_t - \pi_0 + \pi_1 u_{t-1}$ iteratively. Hence, each $u_t$ depends on unknown parameters $\pi_0$ and $\pi_1$ directly and indirectly through the presence of $u_{t-1}$. This parameter dependence structure makes the conditional likelihood function a very complex non-linear function in $\theta$, in sharp contrast with the conditional likelihood function of an AR($p$) model. Similarly, the exact and conditional MLEs for an MA($q$) model must also be computed from non-linear likelihood functions.

4 Check White Noise for Filtered Residuals

For checking whether an observed time series comes from a white-noise stochastic process, test for the null of zero lag-1 autocorrelation is studied first. If $\{y_t\}$ is an independent and identically distributed (iid) sequence with $E(y_t^2) < \infty$,

$$\hat{\rho}_T(1) = \frac{\sum_{t=2}^T(y_t - \bar{y}_{T-1})(y_{t-1} - \bar{y}_{T-1})}{\sum_{t=2}^T(y_{t-1} - \bar{y}_{T-1})^2} \rightarrow N(0,1), \text{ for a } T \text{ large enough,}$$

see Brockwell and Davis (1991, Theorem 7.2.2).

For testing $H_0: \rho(1) = 0$ against $H_a: \rho(1) \neq 0$, the test statistic under the null

$$\sqrt{T} \hat{\rho}_T(1) \rightarrow^p N(0,1).$$
In general, for testing $H_0 : \rho(l) = 0$ against $H_a : \rho(l) \neq 0$, the test statistic under the null
\[
\sqrt{T} \hat{\rho}_T(l) \to^p N(0, 1),
\]
\[
\hat{\rho}_T(l) = \frac{\sum_{t=l+1}^{T}(y_t - \bar{y}_T)(y_{t-l} - \bar{y}_T-l)}{\sum_{t=l+1}^{T}(y_{t-l} - \bar{y}_T-l)^2} \to N(0,1), \text{ for a } T \text{ large enough,}
\]

4.1 Portmanteau Test

Box and Pierce (1970) propose the portmanteau statistic
\[
Q^*(m) = T \sum_{l=1}^{m} \hat{\rho}_T(l)^2
\]

4.1.1 Portmanteau Test

Box and Pierce (1970) propose the portmanteau statistic
\[
Q^*(m) = T \sum_{l=1}^{m} \hat{\rho}_T(l)^2
\]

4.1.2 Portmanteau Test

Under the null, $Q^*(m) \to \chi^2(m)$ as $T \to \infty$.

Ljung and Box (1978) modify the $Q^*(m)$ statistic to increase the power of the test in finite samples,
\[
Q(m) = T(T+2) \sum_{l=1}^{m} \frac{\hat{\rho}_T(l)^2}{T-l}.
\]

4.1.3 Portmanteau Test

Given the MLE estimation of an ARMA($p,q$) model, the filtered residuals are defined as
\[
\hat{u}_t = y_t - \sum_{i=1}^{p} \hat{\phi}_i y_{t-i} - \sum_{j=1}^{q} \hat{\pi}_j u_{t-j}, t = p + q + 1, \ldots, T.
\]

4.1.4 Portmanteau Test

Replace $\{y_t, t = 1, \ldots, T\}$ with $\{\hat{u}_t, t = p + q + 1, \ldots, T\}$, above portmanteau test is applicable to check the white noise for $\hat{u}_t$. If the null is not rejected, $\{u_t\}$ is inferred to be a white noise and the time series model is to be appropriate. Note that, it is easy to know that if the portmanteau test statistic is significant for the filtered residuals from estimated ARMA($p,q$), the test will also be insignificant for the filtered residuals from a model with higher orders of Ar and/or MA terms than $p$ and $q$. To have the model identified parsimoniously (or have fewer parameters to be estimated, or to have larger degree of freedom), model selection criteria have been applied.
4.2 Testing That a Dependent Process Is Uncorrelated

Labato (2001b) provides a new testing procedure for the null hypothesis that a stochastic process is uncorrelated when the process is possibly dependent. Let $y_t$ be a covariance stationary stochastic process with mean $\mu$ and autocovariance at lag $j$, $\gamma(j) = E[(y_t - \mu)(y_{t-j} - \mu)]$. A sample of $y_t(t = 1, \ldots, T)$ is observed. Denote by $\bar{y}_T = \sum_{t=1}^{T} y_t / T$ as the sample mean. The null hypothesis of interest is $H_0: \mu = 0$ and the alternative is $H_0: \mu \neq 0$.

As $\sqrt{T}(\bar{y}_T - \mu) \Rightarrow N(0, 2\pi f_y(0))$ where $\Rightarrow$ refers to weak convergence, and $f_y(0)$ denotes the spectral density of $y_t$ at zero frequency that satisfies $2\pi f_y(0) = \sum_{j=-\infty}^{\infty} \gamma(j)$, the test statistic $T \bar{y}_T^2 / 2\pi f_y(0)$ converges to $\chi^2(1)$ and is appropriate for testing the null when $f_y(0)$ is known. However, it is usually unknown but can be estimated by kernel smoothing method. The bandwidth parameter for the kernel smoothing estimation for $f_y(0)$ is difficult to determine.

Under weak dependence conditions, and under the null hypothesis, for any $r \in (0, 1)$ an invariance principle or functional central limit theorem (CLT) (FCLT) holds, namely,

$$
\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} y_t \Rightarrow [2\pi f_y(0)]^{1/2} B(r),
$$

where $B(r)$ is a Brownian motion, and $[Tr]$ denotes the integer part of $Tr$. And,

$$
\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} (y_t - \bar{y}_T) \Rightarrow [2\pi f_y(0)]^{1/2} [B(r) - rB(1)],
$$

where $B(r) - rB(1)$ is called a Brownian bridge.

In addition, the statistic $S^M = T^{-2} \sum_{t=1}^{T} [\sum_{s=1}^{t}(y_s - \bar{y}_T)]^2$ converges weakly to $2\pi f_y(0) \int_{0}^{1} [B(r) - rB(1)]^2 dr$. Consider an alternative test statistic for $H_0: \mu = 0$ as

$$
T^M = T \bar{y}_T^2 / S^M \Rightarrow U_1 = \frac{B(1)^2}{\int_{0}^{1} [B(r) - rB(1)]^2 dr},
$$

which “does not depend” on $2\pi f_y(0)$. However the limiting distribution is not standard. Approximate upper critical values of the distribution are tabulated in the first row of Table 1 in Lobato (2001b) by means of simulations. The critical values for 90 %, 95 %, and 99 % are 28.31, 45.4, and 99.76, respectively.

4.2.1 Testing $H_0: \mu = 0$

Let $y_t$ be a covariance stationary stochastic vector process of dimension $K$ with mean $\mu$ and autocovariance matrix at lag $j$, $\Gamma(j) = E[(y_t - \mu)(y_{t-j} - \mu)^\prime]$. A sample of $y_t$
is observed \((t = 1, \ldots, T)\), and the sample mean is denoted by \(\bar{y}_T = \sum_{t=1}^{T} y_t / T\). The null hypothesis is \(H_0 : \mu = 0\).

Under the null, see Davidson 1994, chap. 24; Dhrymes 1998, chap. 9,
\[
\sqrt{T}(\bar{y}_T - \mu) \Rightarrow \mathcal{N}(0, 2\pi f_y(0)),
\]
where \(f_y(0)\) is the spectral density matrix of \(y_t\) at zero frequency that satisfies \(2\pi f_y(0) = \sum_{j=-\infty}^{\infty} \Gamma(j)\). With \(2\pi f_y(0) = \Psi\Psi'\) and under the null,
\[
\sqrt{T}\Psi^{-1}\bar{y}_T \Rightarrow \mathcal{N}(0, I_K).
\]

Denote \(S_M^{[T]} = \sum_{i=1}^{[T]}(y_t - \bar{y}_T)\), under the null and some regular conditions,
\[
\frac{1}{\sqrt{T}} S_M^{[T]} \Rightarrow \Psi [B_K(r) - r B_K(1)],
\]
where \(B_K(\cdot)\) denotes a \(K\)-dimensional Brownian motions. And
\[
C_M = T^{-2} \sum_{t=1}^{T} S_t^M S_t^{M'} \Rightarrow \Psi V K \Psi',
\]
where \(V_K = \int_0^1 [B_K(r) - r B_K(1)][B_K(r) - r B_K(1)]' dr\). The test statistic for \(H_0 : \mu = 0\) suggested by Lobato (2001b) is
\[
T_K^M = T \bar{y}_T' C_M^{-1} \bar{y}_T \Rightarrow U_K = B_K(1)' V^{-1} K B_K(1).
\]

Upper values of the distribution of \(U_K\) are tabulated in Table 1 of Lobato (2001b) by means of simulations

### 4.2.2 Testing \(H_0 : \gamma(1) = 0\)

Denote \(c_1 = T^{-1} \sum_{t=2}^{T}(y_t - \bar{y}_T)(y_{t-1} - \bar{y}_T)\) as the sample autocovariance. As \(c_1\) can be seen as the sample mean of \(z_{1t} = (y_t - \bar{y}_T)(y_{t-1} - \bar{y}_T)\). That is, under weak dependence conditions, the following CLT for \(c_1\) holds:
\[
\sqrt{T}(c_1 - \gamma(1)) \Rightarrow \mathcal{N}(0, 2\pi f_{\tilde{z}_1}(0))
\]
where \(f_{\tilde{z}_1}(0)\) is the spectral density at zero frequency of \(\tilde{z}_1_t = (y_t - \mu)(y_{t-1} - \mu)\). Under the null \(H_0 : \gamma(1) = 0\), the test statistic \(T c_1^2 / 2\pi \hat{f}_{\tilde{z}_1}(0)\) converges weakly to \(\chi^2(1)\), where \(\hat{f}_{\tilde{z}_1}(0)\) is a consistent estimator for \(f_{\tilde{z}_1}(0)\). Under the null \(H_0 : \gamma(1) = 0\) and the FCLT for the partial sum \(\sum_{t=1}^{T}[T] z_{1t}\) holds, i.e.,
\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T}[T] z_{1t} \Rightarrow [2\pi f_{\tilde{z}_1}]^{1/2} B(r),
\]
where \(B(\cdot)\) denotes the standard Brownian motion.
\[
C_1 = T^{-2} \sum_{t=1}^{T} \left( \sum_{s=1}^{t} (z_{1t} - c_1) \right)^2 \Rightarrow [2\pi f_{\tilde{z}}] \int_0^1 [B(r) - rB(1)]^2 dr.
\]

Hence, for one-sided hypothesis testing, Lobato (2001b) considers the statistic \( T^*_1 = T^{1/2}c_1/C_1^{1/2} \), and for two-sided hypothesis testing, Lobato (2001b) considers the statistic \( T_1 = Tc_2^2/C_1 \). Under the null hypothesis, \( T^*_1 \Rightarrow U_1^{1/2} \) and \( T_1 \Rightarrow U_1 \).

In general, the null hypothesis of interest is that the process is uncorrelated up to lag \( k \), \( H_0 : \gamma(1) = \gamma(2) = \cdots = \gamma(K) = 0 \) and the alternative is \( H_a : \gamma(j) \neq 0 \) for some \( j = 1, \ldots, K \). Define the vector of sample autocovariances \( c = (c_1, \ldots, c_K)' \) and the vector of the population autocovariances \( \gamma = (\gamma(1), \ldots, \gamma(K))' \), the vector \( z_t = (z_{1t}, \ldots, z_{Kt})' \) with \( z_{kt} = (y_t - \bar{y})(y_{t-k} - \bar{y}) \), and the vector \( \tilde{z}_t = (\tilde{z}_{1t}, \ldots, \tilde{z}_{Kt})' \) with \( \tilde{z}_{kt} = (y_t - \mu)(y_{t-k} - \mu) \). Under a variety of weak dependence conditions [see Romano and Thombs (1996, p. 592, p. 599) for mixing processes], the following CLT can be derived:

\[
\sqrt{T}(c - \gamma) \Rightarrow \mathcal{N}(0, 2\pi f_{\tilde{z}}(0))
\]

where \( f_{\tilde{z}}(0) \) is the spectral density matrix at zero frequency of the vector \( \tilde{z}_t \). Therefore,

\[
\sqrt{T} \Phi^{-1} c \Rightarrow \mathcal{N}(0, I_K),
\]

the lower triangular \( K \times K \) matrix \( \Phi \) satisfies \( 2\pi f_{\tilde{z}}(0) = \Phi \Phi' \). Given a consistent estimator \( \hat{f}_{\tilde{z}}(0) \) for \( f_{\tilde{z}}(0) \), Lobato (2001a) introduces the test statistic

\[
\tilde{Q}_K = Tc'(2\pi \hat{f}_{\tilde{z}}(0))^{-1} c
\]

converges weakly to \( \chi^2(K) \).

Denote the vector

\[
S_{[Tr]} = \sum_{t=1}^{[Tr]} \begin{pmatrix} z_{1t} - c_1 \\ \\ \vdots \\ z_{Kt} - c_K \end{pmatrix}
\]

Under the null hypothesis,

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} z_t \Rightarrow \Phi B_K(r),
\]

where \( B_K(\cdot) \) denotes a \( K \)-dimensional Brownian motions. Therefore,

\[
\frac{1}{\sqrt{T}} S_{[Tr]} \Rightarrow \Phi [B_K(r) - rB_K(1)]
\]
and

\[ C_K = \frac{1}{T^2} \sum_{t=1}^{T} S_t S_t' \Rightarrow \Phi V_K \Phi', \]

where \( V_K = \int_0^1 [B_K(r) - rB_K(1)][B_K(r) - rB_K(1)]' dr \). The test statistic proposed by Lobato (2001b) is

\[ T_K = T c' C_K^{-1} c \Rightarrow U_K. \]

### 4.3 Selection for a Parsimonious Model

There are different ways to determine whether an estimated ARMA model is “correct”. One approach is to estimate a family of ARMA models and select the “best” (and “most parsimonious”) model among this family according to a model selection criterion such as AIC or SIC. Let \( \hat{u}_t \) denote the residual from a fitted ARMA\((p,q)\) model and \( \hat{\sigma}^2 = \sum_{t=1}^{T} \hat{u}_t^2 / T \) be the variance estimate. We have

\[
\begin{align*}
\text{AIC} & = \log \hat{\sigma}^2 + 2(p + q + 1)/T, \\
\text{SIC} & = \log \hat{\sigma}^2 + [(p + q + 1) \log T]/T.
\end{align*}
\]

Note that SIC is “dimensionally consistent” for ARMA models.

### 4.4 Impulse Response Function

Given a stationary and invertible ARMA process, \( \phi(B)Y_t = \pi(B)Z_t \),

\[ Y_t = [\pi(B)/\phi(B)]Z_t = \psi(B)Z_t, \quad (1) \]

The weights \( \psi_0, \psi_1, \psi_2, \ldots \) in the function

\[ \psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \cdots \]

is called the impulse response function of \( Y_t \). The weight, \( \psi_j \), can be demonstrated as the “response” of \( Y_t \) caused by a unit (one standard deviation) change of \( Z_{t-j} \). Given (1),

\[ Y_{t+j} = \psi_0 Z_{t+j} + \psi_1 Z_{t+j-1} + \psi_2 Z_{t+j-2} + \cdots + \psi_j Z_t + \cdots \]

Define the impulse response (dynamic multiplier) of the future observation \( Y_{t+j} \) to the effect of one unit change of \( Z_t \) as \( \partial y_{t+j}/\partial Z_t = \psi_j \) 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. For a stable dynamic equation, summing the impulse responses yields an accumulated response (interim multiplier):

\[
\sum_{i=0}^{j} \frac{\partial Y_{t+i}}{\partial Z_{t+i}} = \sum_{i=0}^{j} \psi_i
\]

Letting \( j \) tend to infinity, we have

\[
\lim_{j \to \infty} \sum_{i=0}^{j} \frac{\partial Y_{t+i}}{\partial Z_{t+i}} = \lim_{j \to \infty} \sum_{i=0}^{j} \psi_i
\]

which represents the long-run effect (total multiplier) of a permanent change in \( Z_t \). This is the total effect resulted from the changes of current and all subsequent innovations.

5 Forecasting

Once an ARMA model is determined as the “best” model, it is believed that this model is capable of representing the behavior of the variable of interest. Forecasts should then be derived from the fitted model. Suppose that the selected model is ARMA(\( p, q \)), this relationship,

\[
y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \pi_1 \epsilon_{t-1} - \cdots - \pi_q \epsilon_{t-q} + \epsilon_t,
\]

should hold for observations in and out of sample. Let \( \hat{y}_{t+h|t} \) denote the forecast of \( y_{t+h} \) made at time \( t \). Then, \( h \)-step ahead forecasts are:

\[
\hat{y}_{t+h|t} = \hat{\phi}_0 + \hat{\phi}_1 \hat{y}_{t+h-1|t} + \cdots + \hat{\phi}_p \hat{y}_{t+h-p|t} - \hat{\pi}_1 \hat{\epsilon}_{t+h-1|t} - \cdots
\]

\[
- \hat{\pi}_q \hat{\epsilon}_{t+h-q|t} + \hat{\epsilon}_{t+h|t},
\]

where \( \hat{\phi} \) and \( \hat{\pi} \) are MLEs of model parameters. Here, for \( j \leq 0 \), \( \hat{y}_{t+j|t} = y_{t+j} \) are known observations, and \( \hat{\epsilon}_{t+j|t} = \hat{u}_{t+j} \) are residuals from the fitted model; for \( j > 0 \), \( \hat{\epsilon}_{t+j|t} = 0 \) are the best forecast of white noise.

As an example consider the forecasts from a stationary AR(1) model:

\[
\hat{y}_{t+h|t} = \begin{cases} 
\hat{\phi}_0 + \hat{\phi}_1 \hat{y}_{t+h-1|t} + \hat{\epsilon}_{t+h|t} & h = 1 \\
\hat{\phi}_0 + \hat{\phi}_1 \hat{y}_{t+h-1|t} & h = 2, 3, \ldots
\end{cases}
\]

It follows that

\[
\hat{y}_{t+h|t} = (1 + \hat{\phi}_1 + \cdots + \hat{\phi}_1^{h-1}) \hat{\phi}_0 + \hat{\phi}_1 y_t,
\]

23
and consequently, the $h$-step ahead forecasts would be a constant when $h$ is large. Consider also an MA(1) model:

$$
\hat{y}_{t+h|t} = \hat{\pi}_0 - \hat{\pi}_1 e_t, \quad h = 1
$$

$$
= \hat{\pi}_0, \quad h = 2, 3, \ldots
$$

Hence, the forecasts of two or more periods ahead are simply a constant.

### 5.1 Combined Forecasts

Forecasts may be obtained from econometric models, exponential smoothing algorithms, ARIMA models, and other forecasting schemes. As each method utilizes the information of data differently, it may be desirable to combine two or more alternative forecasts to form a composite forecast. Let $f_{i,t+h|t}, i = 1, \ldots, N$, denote the forecast of $y_t$ resulted from the $i$-th method. The forecast combining $N$ individual forecasts can be expressed as

$$
f_{t+h|t} = \frac{1}{N} \sum_{j=1}^{N} w_j f_{j,t+h|t}, \quad \text{with} \quad \sum_{j=1}^{N} w_j = 1, \quad 0 \leq w_j \leq 1.
$$

Let $e_{i,t+h|t} = y_{t+h} - f_{i,t+h|t}$ and $e_{t+h|t} = y_{t+h} - f_{t+h|t}$ be the errors resulted from individual and combined forecasts. It is readily seen that, if each individual forecast is unbiased (i.e., $E(e_{i,t+h|t}) = 0$), then so is the combined forecast (i.e., $E(e_{t+h|t}) = 0$). The simplest way to combine forecasts is to give an equal weight $w_j = 1/N$ to each forecast so that

$$
f_{t+h|t} = \frac{1}{N} \sum_{j=1}^{N} f_{j,t+h|t}.
$$

To see why this combined forecast may be preferred, consider the case $N = 2$ so that

$$
E(e_{t+h|t}^2) = E\left(\frac{e_{1,t+h|t}^2 + e_{2,t+h|t}^2}{2}\right)^2 = \frac{1}{4}(\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2),
$$

where $\sigma_{i,t+h|t}^2 = E(e_{i,t+h|t}^2), i = 1, 2$, are error variances, and $\rho$ is the correlation coefficient of $e_{1,t+h|t}$ and $e_{2,t+h|t}$.

Suppose further that these two forecasts are of the same quality so that they have the same error variance $\sigma_{t+h|t}^2$. Then,

$$
E(e_{t+h|t}^2) = \frac{1}{2}\sigma_{t+h|t}^2(1 + \rho) \leq \sigma_{t+h|t}^2,
$$
and the equality holds when $\rho = 1$. It should be clear that the error variance of the combined forecast is less than that of each individual forecast as long as $\rho < 1$. Hence, it is relatively less likely for the combined forecast to commit huge forecasting errors. On the other hand, when two forecasts are perfectly, positively correlated ($\rho = 1$), combined forecast would not be necessary.

Another approach is to make the combining weight of a given method inversely proportional to the sum of squared forecasting errors achieved by this particular method (over $T'$ periods). That is, for $j = 1, \ldots, N$,

$$w_j = \frac{\left(\sum_{t=1}^{T'} e^2_{j,t+h|t}\right)^{-1}}{\left(\sum_{t=1}^{T'} e^2_{1,t+h|t}\right)^{-1} + \cdots + \left(\sum_{t=1}^{T'} e^2_{N,t+h|t}\right)^{-1}}.$$

Thus, a method that yields a smaller sum of squared forecasting errors receives a larger weight for its forecast, and the combined forecast relies more heavily on the forecasts from “better” methods.

A more natural way is to find the combining weights that minimize the sum of squared forecasting errors. This amounts to fitting a regression model of $y_t$ on the forecasts $f_{i,t+h|t}$, $i = 1, \ldots, N$, subject to the constraint that all weights summing to one, i.e.,

$$y_{t+h} - f_{N,t+h|t} = w_1(f_{1,t+h|t} - f_{N,t+h|t}) + w_2(f_{2,t+h|t} - f_{N,t+h|t}) + \cdots + w_{N-1}(f_{N-1,t+h|t} - f_{N,t+h|t}) + u_t.$$

The combining weights are the estimated regression coefficients, and the combined forecasts are the fitted values of this model:

$$f_{t+h|t} = \hat{w}_1 f_{1,t+h|t} + \hat{w}_2 f_{2,t+h|t} + \cdots + \hat{w}_{N-1} f_{N-1,t+h|t} + (1 - \sum_{j=1}^{N-1} \hat{w}_j) f_{N,t+h|t}.$$

More generally, the combined forecast may be determined by fitting the model:

$$y_{t+h} = \alpha + \beta_1 f_{1,t+h|t} + \beta_2 f_{2,t+h|t} + \cdots + \beta_N f_{N,t+h|t} + u_t,$$

without restricting $\alpha = 0$ and all $\beta$’s summing to one. Note, however, that there is no definite conclusion about the “best” method of combining forecasts. One may have to experiment different methods before settling on a particular one.

5.2 Forecast Evaluation

In practice, it is typical to partition the sample into two non-overlapping parts: the first sub-sample is used to determine an appropriate model for forecasting, and the
second sub-sample is merely used to evaluate the forecasts generated from the previously estimated model. As the information in the second sample is not used for model determination, the resulting forecasts are also referred to as “out-of-sample” forecasts. The forecasting performance of a particular model should be evaluated based on out-of-sample forecasts. There are various descriptive measures of average quality of forecasts. For example:

1. Mean Squared Error (MSE): \[ \text{MSE} = \frac{1}{T'} \sum_{t=1}^{T'} e_{t+h|t}^2 \] where \( T' \) is the number of observations in the second sub-sample. A similar measure is the square root of MSE, also known as RMSE.

2. Mean Absolute Error (MAE): \[ \text{MAE} = \frac{1}{T'} \sum_{t=1}^{T'} |e_{t+h|t}| \]

3. Mean Absolute Percentage Error (MAPE): \[ \text{MAPE} = \frac{100}{T'} \sum_{t=1}^{T'} \frac{|e_{t+h|t}|}{y_{t+h}} \]

To compare forecasts from different methods, we may simply compare the magnitudes of their out-of-sample MSEs (or MAEs or MAPEs). We may also compute the ratio of MSE of some model to the MSE of a benchmark model. Suppose the benchmark forecast is the naive forecast: \( f_{t+h|t} = y_{t-1} \). The RMSE ratio is also known as Theil’s \( U \):
\[
U = \left( \frac{\sum_{t=1}^{T'} (y_{t+h} - f_{t+h|t})^2}{\sum_{t=1}^{T'} (y_{t+h} - y_{t-1})^2} \right)^{1/2}.
\]

A different way to evaluate the effectiveness of forecasts is to perform a regression:
\[
y_{t+h} = a + bf_{1,t+h|t} + u_t,
\]
and test the joint hypothesis \( a = 0 \) and \( b = 1 \). Failing to reject the null hypothesis suggests that the forecasts are capable of describing the systematic variation of \( y_t \). In addition, we may want to evaluate whether combining an alternative forecast \( f_{2,t+h|t} \) with \( f_{1,t+h|t} \) can reduce MSE. For this purpose, we fit the model:
\[
y_{t+h} = a + bf_{1,t+h|t} + cf_{2,t+h|t} + u_t,
\]
and test the joint hypothesis: \( a = 0 \), \( b = 1 \), and \( c = 0 \). This model can of course be generalized to incorporate more alternative forecasts.

Finally, we examine a decomposition of MSE. Let \( S_f = |\sum_{t=1}^{T'} (f_{t+h|t} - \bar{f})^2|^{1/2} / T' \), \( S_y = |\sum_{t=1}^{T'} (y_{t+h} - \bar{y})^2|^{1/2} / T' \), and
\[
r = \frac{1}{T'S_f S_y} \sum_{t=1}^{T'} (y_{t+h} - \bar{y})(f_{t+h|t} - \bar{f}),
\]
26
where $\bar{f}$ and $\bar{y}$ are averages of $f_{t+h|t}$ and $y_{t+h}$. Then,

$$\text{MSE} = \frac{1}{T'} \sum_{t=1}^{T'} [(y_{t+h} - \bar{y}) - (f_{t+h|t} - \bar{f}) + (\bar{y} - \bar{f})]^2$$

$$= (\bar{f} - \bar{y})^2 + (S_f - rS_y)^2 + (1 - r^2)S_y^2.$$ 

Let UM, UR, and UD denote, respectively, the three terms on the right-hand side divided by MSE. Hence, $\text{UM} + \text{UR} + \text{UD} = 1$. Clearly, UM compares the means of $y_{t+h}$ and $f_{t+h|t}$. For the other two terms, we observe that

$$rS_y/S_f = \frac{\sum_{t=1}^{T'} (y_{t+h} - \bar{y})(f_{t+h|t} - \bar{f})}{\sum_{t=1}^{T'} (f_{t+h|t} - \bar{f})^2},$$

is the slope, $\hat{b}$, of the regression of $y_{t+h}$ on $f_{t+h|t}$, and that

$$(1 - r^2)S_y^2 = S_y^2 - \left(\frac{rS_y}{S_f}\right)^2 rS_yS_f$$

$$= \frac{1}{T'} \sum_{t=1}^{T'} [(y_{t+h} - \bar{y})((y_{t+h} - \bar{y}) - \hat{b}(f_{t+h|t} - \bar{f}))$$

$$= \frac{1}{T'} \sum_{t=1}^{T'} [(y_{t+h} - \bar{y}) - \hat{b}(f_{t+h|t} - \bar{f})]^2.$$ 

Therefore, UR measures the difference between $\hat{b}$ and one, and UD measures the magnitude of residuals from the regression of $y_{t+h}$ on $f_{t+h|t}$. This decomposition shows that forecasts with small UM (bias proportion), small UR (regression proportion), and UD (disturbance proportion) close to one will be preferred.

### 5.3 Test for Martingale Difference

Escanciano, J. C. and S. Mayoral (2007) provide the following Data-Driven Smooth Tests for the Martingale Difference Hypothesis. Let $\{Y_t\}$ be a strictly stationary and ergodic time series process defined on the probability space $(\Omega, \mathcal{F}, P)$. The MDH states that the best predictor, in a mean square error sense, of $Y_t$ given $I_{t-1} := (Y_{t-1}, Y_{t-2}, \ldots)'$ is just the unconditional expectation, which is zero for a martingale difference sequence (mds). In other words, the MDH states that $Y_t = X_t - X_{t-1}$, where $X_t$ is a martingale process with respect to the $\sigma$-field generated by $I_{t-1}$, i.e., $\mathcal{F}_{t-1} := \sigma(I_{t-1})$. The null of $Y_t$ is martingale difference series is

$$H_0 : E(Y_t | I_{t-1}) = 0 \text{ almost surely (a.s.)}$$

27
The alternative $H_1$ is the negation of the null, i.e., $Y_t$ is not a mds. Given a sample \( \{y_t\}_{t=1}^T \), consider the marked empirical process (cf. Koul and Stute, 1999)

\[
R_T(\theta) = \frac{1}{\hat{\sigma}_T \sqrt{T}} \sum_{t=1}^T y_t 1(y_{t-1} < \theta), \theta \in \mathcal{R},
\]

where $\hat{\sigma}_T = T^{-1} \sum_{t=1}^T y_t^2$ and $1(A)$ denotes the indicator function, i.e. $1(A) = 1$ if $A$ occurs and 0 otherwise.

Let $B$ be a standard Brownian motion on $[0,1]$, and define

\[
\tau^2(\theta) = \sigma^{-2} E[y_t^2 1(y_{t-1} \leq \theta)]
\]

where $\sigma^2 = E(y_t^2)$. Notice that $\tau^2(\infty) = 1$, $\tau^2(-\infty) = 0$, and is nondecreasing and continuous. Then, under regularity condition and null hypothesis,

\[
R_T(\theta) \Rightarrow B(\tau^2(\theta)).
\]

An immediate consequence of Theorem 1 and Lemma 3.1 in Chang (1990) is that, under some regularity conditions and $H_0$,

\[
CvM_t := \int_{\mathcal{R}} |R_T(\theta)|^2 \tau^2_T(d\theta) \rightarrow^d
\]

\[
CvM_\infty := \int_{\mathcal{R}} |B(\tau^2(\theta))|^2 \tau^2(d\theta) = \int_0^1 |B(u)|^2 du,
\]

where $\tau^2_T(\theta) = \hat{\sigma}_T^2 T^{-1} \sum_{t=1}^T y_t^2 1(y_{t-1} \leq \theta)$. Or,

\[
KS_T := \sup_{\theta \in \mathcal{R}} |R_T(\theta)| \rightarrow^d \sup_{\mathcal{R}} |B(\tau^2(\theta))|^2 = \sup_{u \in [0,1]} |B(u)|.
\]