NATIONAL KAPODISTRIAN UNIVERSITY OF ATHENS

MASTER THESIS

Bayesian unit root testing

Author:
Vasiliki GEORGOPOULOU

Supervisor:
Dr. Loukia MELIGKOTSIDOU

Msc in Statistics and Operational Research
Department of Mathematics

October 12, 2021

NATIONAL KAPODISTRIAN UNIVERSITY OF ATHENS

Abstract

Department of Mathematics

Msc in Statisics and Operational reasearch

Bayesian unit root testing

by Vasiliki GEORGOPOULOU

The current dissertation analyzes the Bayesian unit root tests in autoregressive processes as an alternate to the classical autoregressive unit root tests. The Bayesian approach to unit root testing was mainly motivated by the power and size distortions of the classical tests under the Dickey-Fuller distribution. Initially, basic principles and theory concerning time series and Bayesian analysis are introduced which are then followed by the structure of the classical autoregressive tests and the Dickey-Fuller distribution. The subsequent chapter is devoted to the methods applied in the Bayesian unit root testing. Simulation results and conclusions based on two Bayesian methods are finally reported.

Acknowledgements

I would like to thank my supervisor Dr. Loukia Meligkotsidou for her advices and support as well as Dr. Burnetas and Dr. Siannis for accepting being members of my thesis committee.

I am also grateful to the people who helped and supported me during all this process.

Contents

Abstract **iii**

Acknowledgements **v**

1 Time Series - Basic Concepts **1**
1.1 Time Series as a stochastic process . . . 1
1.2 Stationarity . . . 1
1.3 Autocovariance and autocorrelation . . . 2
1.4 Stationary Models . . . 3
1.4.1 White noise . . . 3
1.4.2 Moving average process . . . 4
1.4.2.1 Invertibility of MA models . . . 5
1.4.3 Autoregressive process . . . 6
1.4.3.1 AR(1) process with constant . . . 6
1.4.3.2 AR(1) process without constant . . . 8
1.4.3.3 AR(p) process . . . 9
1.4.4 ARMA process . . . 11
1.5 Non-Stationary Models . . . 11
1.5.1 Introduction . . . 11
1.5.2 Stochastic trends . . . 11
1.5.2.1 Random walk model without drift . . . 12
1.5.2.2 Random walk with drift model . . . 14
1.5.3 Deterministic trends . . . 15
1.6 Unit root tests . . . 16

2 Bayesian Theory **19**
2.1 Basics of Bayesian theory . . . 19
2.2 Specifying the prior . . . 20
2.2.1 Conjugate priors . . . 20
2.2.2 Non-informative priors . . . 20
2.3 Bayesian Inference . . . 22
2.3.1 Decision Theory . . . 22
2.3.2 Point Estimation . . . 23
2.3.3 Interval estimation . . . 24
2.3.4 Hypothesis testing . . . 25
2.4 Bayesian Model Comparison . . . 26
2.4.1 Basic information and motives . . . 26
2.4.2 Motivation for Bayesian model comparison . . . 27
2.4.3 Standard framework and modelling . . . 28

3 Autoregressive unit root tests **33**
3.1 Introduction . . . 33
3.2 Basic concepts for unit root tests . . . 33
3.3 Wiener processes . . . 35
3.3.1 From random walk to Wiener process . . . 35
3.3.2 Definition of the Wiener process . . . 36
3.3.3 The Functional Central Limit Theorem (FCLT) . . . 37
3.3.4 The Continuous Mapping Theorem (CMT) . . . 38
3.3.5 Basic results of the Wiener process . . . 39
3.4 Unit root tests based on the Dickey-Fuller distribution . . . 41
3.4.1 Random walk model without drift and trend . . . 41
3.4.2 Models with drift . . . 43
3.4.3 Models with linear trend . . . 45
3.5 Other unit root tests . . . 46
3.5.1 The Augmented Dickey Fuller test . . . 46
3.5.2 The Phillips-Perron test . . . 49
3.6 Issues associated with the classical unit root tests . . . 50

4 Bayesian unit root testing **53**
4.1 A preface to the Bayesian unit root testing . . . 53
4.1.1 Introduction . . . 53
4.1.2 From Classical to Bayesian point of view . . . 54
4.2 Posterior distribution of an AR(1) parameter using a diffuse prior . . . 54
4.3 The "proper" model for unit root testing . . . 56
4.3.1 Schotman's structural model parameterization . . . 56
4.3.2 Phillip's reduced form parameterization . . . 57
4.4 Selecting the appropriate prior . . . 58
4.4.1 The flat prior . . . 58
4.4.2 Phillip's ignorance prior . . . 58
4.4.3 Reference prior . . . 60
4.4.4 The Schotman and Van Dijk prior . . . 60
4.4.5 Normal-Wishart and Lubrano priors . . . 61
4.5 Bayesian methods for unit root testing . . . 61
4.5.1 Providing the evidence . . . 62
4.5.2 Model selection criteria . . . 64
4.6 Criticism on Bayesian unit root testing and alternatives . . . 66

5 Simulation study **67**

6 Conclusions **75**

Bibliography **77**

List of Figures

1.1 White noise 3
1.2 (a) ACF plot (b) Partial ACF plot 3
1.3 (a) ACF plot (b) Partial ACF plot 7
1.4 AR(1) model with and without constant 9
1.5 AR(1) and Random walk processes 13
1.6 (a) Random walk with drift $\delta = 0.5$ (b) Random walk without drift . . 14
1.7 A trend-stationary process 16

3.1 Simulation for (a) T=10 (b) T=50 and (c) T=250 38

5.1 (a) Prior of ϕ (b) Prior of σ 70
5.2 (a) Marginal posterior of ϕ (b) Marginal posterior of σ 71
5.3 Normal prior distribution for ϕ 73
5.4 Prior and marginal posterior distribution of ϕ 74
5.5 Prior and marginal posterior distribution for σ^2 74

List of Tables

5.1 Posterior odds probabilities in favour of the random walk 68
5.2 Probabilities in favour of the random walk using ADF 69
5.3 Jeffreys' rule for $T = 25$ in favour of the random walk 69
5.4 Jeffreys' rule for $T = 100$ in favour of the random walk 69
5.5 Posterior probabilities of $Pr(\phi \geq 1|x)$ 69
5.6 BIC probabilities in favour of the random walk 70
5.7 Posterior odds probabilities in favour of the random walk 72
5.8 BIC probabilities in favour of the random walk 72
5.9 Probabilities in favour of the random walk using ADF 72
5.10 Jeffreys' rule for T=25 73
5.11 Jeffreys' rule for T=100 73

Chapter 1

Time Series - Basic Concepts

1.1 Time Series as a stochastic process

A time series is a collection of random variables $\{X_t\}$ recorded in chronological order. Such a collection is called a *stochastic process*. If the collection of the random variables is continuous, the time series is said to be continuous. If the collection is discrete, the time series is said to be discrete.

The distributional properties of a time series are completely described by the joint distribution function, $F(X_{t_1}, X_{t_2}, ..., X_{t_n})$, for any positive integer n and for any subset $(t_1, t_2, ..., t_n)$ of T, where T is an infinite set of time.

1.2 Stationarity

The stationarity of a time series arises from the similar statistical properties of $\{X_t, t = 0, \pm1, \pm2, ...\}$ and $\{X_{t+h}, t = 0, \pm1, \pm2, ...\}$.

Definition 1.2.1. (Strict stationarity) A $\{X_t\}$ process is said to be **strictly stationary** if

$$F_{X_{t_1}, X_{t_2}, ..., X_{t_n}}(x_{t_1}, x_{t_2}, ..., x_{t_n}) = F_{X_{t_1+h}, X_{t_2+h}, ..., X_{t_n+h}}(x_{t_1}, x_{t_2}, ..., x_{t_n})$$

for any $(x_{t_1}, x_{t_2}, ..., x_{t_n})$ in the range of X_t and for all possible subsets of indices $t_1, t_2, ..., t_n$ and $t_1 + h, t_2 + h, ..., t_n + h$, where h integer.

Due to the fact that for a normal stationary process second-order stationarity is equivalent to strict stationarity, the concept of stationarity in time series is predominantly based on the first- and second- order moments of X_t. In addition, there is intrisic difficulty in defining exactly the joint distribution of a strictly stationary process. The following definitions make the above claims precise.

Definition 1.2.2. Let X_t be a time series with $E[X^2] < \infty$. The **mean function** of X_t is

$$\mu_X(t) = E(X_t)$$

The **covariance function** of $\{X_t\}$ is

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))],$$

for all integers r and s.

Definition 1.2.3. (Weak stationary process) A $\{X_t\}$ process is said to be **weakly stationary** if

(i) $\mu_X(t)$ is independent of t

and

(ii) $\gamma_X(t+h,t)$ is independent of t for each h.

Weak stationarity implies finite moments and hence the time series is *covariance stationary*, which is not necessarily guaranteed in the case of a strictly stationary time series. In terms of time series analysis, *weak stationarity* is called simply *stationarity*.

1.3 Autocovariance and autocorrelation

The stationarity assumption implies that the joint distribution of the time series is the same for every $t_1, t_2, .., t_n$ no matter how distant they are. The autocovariance is the covariance of the time series with itself at constant intervals of time, denoted by *lags*.

Under the stationarity assumption, the covariance between X_t and X_{t+h} separated by h intervals must be the same for any t. Due to the dependence by the units of measurement, the autocovariance function may not analyze accurately the basic properties of a time series. To this end, the autocorrelation function is commonly preferred, which is precisely defined below.

Definition 1.3.1. Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (ACVF) at lag h is

$$\gamma_X(h) = Cov(X_{t+h}, X_t)$$

Definition 1.3.2. The **autocorrelation function** (ACF) of $\{X_t\}$ at lag h is

$$\rho_X(h) = Corr(X_{t+h}, X_t) = \frac{E[(X_{t+h} - \mu_X(t+h))(X_t - \mu_X(t))]}{E[(X_t - \mu_X(t))^2]} = \frac{\gamma_X(h)}{\gamma_X(0)},$$

where, $\gamma_X(0) = Cov(X_t, X_t) = E[(X_t - \mu_X(t))^2] = \sigma^2$ for all values of t.

Some basic properties of the $\gamma(\cdot)$ are:

(i) $\gamma_X(0) \geq 0$,

(ii) $|\gamma_X(h)| \leq \gamma_X(0)$ for all h,

(iii) $\gamma_X(-h) = \gamma_X(h)$ for all h.

Since the analysis of the time series properties is based on *observed* data, the following definitions portray the mean, the $acvf$ and acf of a sample.

Definition 1.3.3. Let $x_1, x_2, ..., x_n$ be observations of a time series. The **sample mean** of $x_1, x_2, ..., x_n$ is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The **sample autocovariance function** is

$$\hat{\gamma(h)} = \frac{1}{n}\sum_{i=1}^{n-|h|} (x_{t+|h|-\bar{x}})(x_t - \bar{x}), \qquad -n < h < n$$

1.4 Stationary Models

1.4.1 White noise

The basic building block of time series analysis is a sequence of uncorrelated and identically distributed random variables, $\{\epsilon_t\}_{t=-\infty}^{\infty}$, denoted *white noise process*. The mean and the variance of ϵ_t are $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma_\epsilon^2$, respectively. A realisation of a stationary white noise process is shown in the following figure

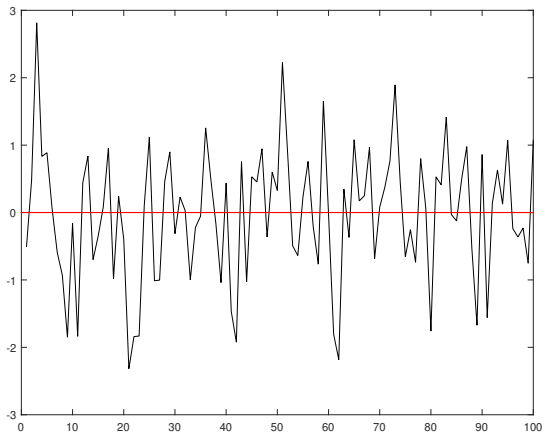


FIGURE 1.1: White noise

The acvf of this second-order stationary process is

$$\gamma_X(h) = cov(X_t, X_{t+h}) = E(X_t X_{t+h}) = 0,$$

for $h \neq 0$ and the acf is

$$\rho_X(h) = \begin{cases} 1, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

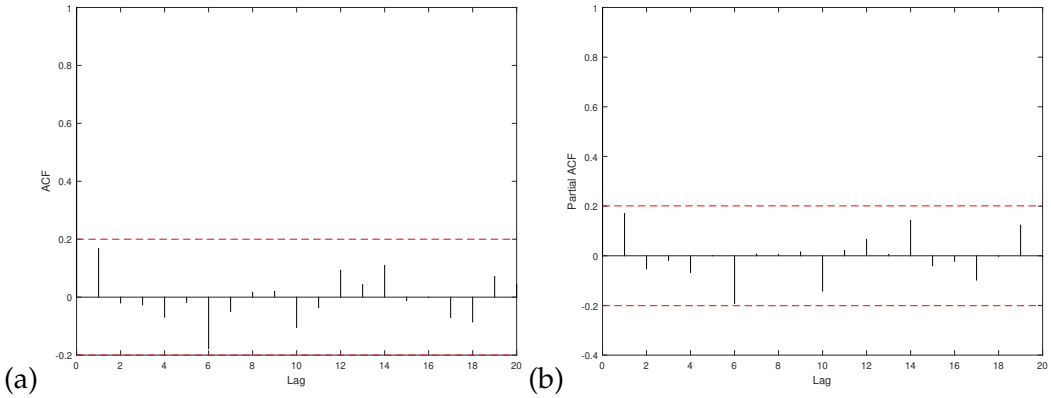


FIGURE 1.2: (a) ACF plot (b) Partial ACF plot

In the special case of independent ϵ_ts the above process is an *iid noise* which is identical to *white noise* only when the random variables are normally distributed.

1.4.2 Moving average process

MA(1) process

Let ϵ_t be a white noise as defined above. The process

$$X_t = \mu + \epsilon_t + \theta\epsilon_{t-1},$$

where μ and θ could be any constants, is called a *first-order moving average process*, denoted MA(1). The mean and the variance of the process are

$$E(X_t) = E(\mu + \epsilon_t + \theta\epsilon_{t-1}) = \mu + E(\epsilon_t) + \theta E(\epsilon_{t-1}) = \mu,$$

$$\begin{aligned}\gamma(0) = Var(X_t) &= E(X_t - \mu)^2 = E(\epsilon_t + \theta\epsilon_{t-1})^2 \\ &= E(\epsilon_t^2) + 2\theta E(\epsilon_t\epsilon_{t-1}) + \theta^2 E(\epsilon_{t-1}^2) \\ &= (1+\theta^2)\sigma_\epsilon^2,\end{aligned}$$

since $E(\epsilon_t) = 0$ for all t and $E(\epsilon_t\epsilon_s) = 0$ for $t \neq s$. The first order autocovariance of the process is

$$\begin{aligned}\gamma(1) = Cov(X_t, X_{t-1}) &= E(X_t - \mu)(X_{t-1} - \mu) \\ &= E(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2}) \\ &= E(\epsilon_t\epsilon_{t-1}) + \theta E(\epsilon_t\epsilon_{t-2}) + \theta E(\epsilon_{t-1}^2) + \theta^2 E(\epsilon_{t-1}\epsilon_{t-2}) \\ &= \theta\sigma_\epsilon^2\end{aligned}$$

Higher than first-order autocovariances are all equal to zero:

$$\begin{aligned}\gamma_X(h) = Cov(X_t, X_{t-h}) &= E(X_t - \mu)(X_{t-h} - \mu) \\ &= E(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-h} + \theta\epsilon_{t-h-1}) \\ &= 0\end{aligned}$$

for all $h > 1$. Thus, the mean and the autocovariances of the process are independent of time, which implies that MA(1) is second-order stationary regardless of the values of θ. The first autocorrelation is

$$\rho_X(1) = \frac{\gamma_X(1)}{\gamma_X(0)} = \frac{\theta\sigma_\epsilon^2}{(1+\theta^2)\sigma_\epsilon^2}.$$

It can be inferred that positive values of θ imply positive autocorrelation between X_t, X_{t-1}.

MA(q) process

Inductively, the *qth order moving average model* is

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q},$$

where $\{\epsilon_t\}$ is a white noise and $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_q)$ is a vector of real numbers. The mean and the variance of the process are $E(X_t) = \mu$ and $\gamma_X(0) = Var(X_t) = (1 + \theta_1 + \theta_2 + \theta_q)\sigma^2$, respectively. The *acvf* is:

$$\gamma_X(h) = Cov(X_t, X_{t-h}) = \begin{cases} \sigma^2 \sum_{i=0}^{g-h} \theta_i \theta_{i+h}, & h = 0, 1, 2, ..., q, \\ 0, & h > q, \end{cases}$$

and the *acf* can be obtained by dividing $\gamma_X(h)$ by $\gamma_X(0)$ for $h < q$.

MA(∞) process

It is worth mentioning, that the model which occurs as $q \to \infty$,

$$X_t = \mu + \sum_{i=0}^{\infty} \theta_i \epsilon_{t-i} = \mu + \theta_0 \epsilon_t + \theta_1 \epsilon_{t-1} + ... \tag{1.1}$$

is described as a MA(∞) process. The equation 1.1 represents a well defined, weak stationary process only if

$$\sum_{i=0}^{\infty} \theta_i^2 < \infty. \tag{1.2}$$

However, a slightly stronger condition than 1.2 is preferred:

$$\sum_{i=0}^{\infty} |\theta_i| < \infty. \tag{1.3}$$

The mean, variance and autocovariance of this model are a generalization of those of the MA(q) model. More precisely,

$$\begin{aligned} E(X_t) &= \lim_{T \to \infty} E(\mu + \theta_0 \epsilon_0 + \theta_1 \epsilon_1 + ... + \theta_T \epsilon_{t-T}) = \mu \\ \gamma_X(0) &= Var(X_t) = E(X_t - \mu)^2 = \lim_{T \to \infty} E(\theta_0 \epsilon_0 + \theta_1 \epsilon_1 + ... + \theta_T \epsilon_{t-T}) \\ &= \lim_{T \to \infty} (\theta_0 \epsilon_0 + \theta_1 \epsilon_1 + ... + \theta_T \epsilon_{t-T}) \sigma_\epsilon^2 \\ \gamma_X(h) &= E(X_t - \mu)(X_{t-h} - \mu) \\ &= \sigma_\epsilon^2 (\theta_h \theta_0 + \theta_{h+1} \theta_t + \theta_{h+2} \theta_2 + \theta_{h+3} \theta_3 + ..) \end{aligned}$$

1.4.2.1 Invertibility of MA models

A time series is said to be *invertible* if the white noise part can be represented as a linear function of current and past observations of the process. The conditions for the invertibility of the MA(1) and MA(q) process are demonstrated below.

Invertibility of MA(1) process

Let $X_t = \mu + \epsilon + \theta \epsilon_{t-1}$ be a first order MA process, which can be written as

$$X_t - \mu = (1 + \theta L) \epsilon_t \tag{1.4}$$

in terms of the lag operator L. By multiplying 1.4 by $(1 + \theta L)^{-1}$, the above equation is expressed as

$$\begin{aligned} (1 + \theta L)^{-1} (X_t - \mu) = \epsilon_t \Rightarrow \\ \pi(L)(X_t - \mu) = \epsilon_t, \end{aligned} \tag{1.5}$$

where, $\pi(L) = \sum_{i=0}^{t} (-\theta)^i L^i = 1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + ...)$. The convergence of the sum depends on the values of θ. If $|\theta| < 1$, then the process is invertible, otherwise

1.5 would not be well defined. Equivalently, the root of the $1 + \theta L = 0$ has to lie outside the unit circle.

Hence, the MA(1) process can be expressed as a linear function of present and past observations simply by inverting the moving average operator $(1 + \theta L)$ (Hamilton, 1994).

Invertibility of MA(q) process

In the general case of a MA(q) process,

$$X_t = \mu + \epsilon_t + \theta\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q} \Rightarrow \tag{1.6}$$

$$X_t = \mu + (1 + \theta L + \theta^2 L^2 + ... + \theta_q L^q)\epsilon_t \Rightarrow \tag{1.7}$$

$$X_t - \mu = \Theta(L)\epsilon_t \tag{1.8}$$

the *invertibility condition* for 1.8 is that, all the roots of the equation $\Theta(L) = 0 \Rightarrow 1 + \theta z + \theta^2 z^2 + ... + \theta_q z^q = 0$ have to lie outside the unit circle. By multiplying 1.6 by $\Theta^{-1}(L)$, arises the following equation:

$$\epsilon_t = \Theta^{-1}(L)(X_t - \mu) \tag{1.9}$$

Hence, 1.6 is an *invertible* process.

1.4.3 Autoregressive process

The current subsection demonstrates some basic *autoregressive processes*. The characteristic of these models is that the present value is expressed as a linear aggregate of past observations and a white noise process,$\{\epsilon_t\}$. Since the model is *regressed* on its own past values, it is called *autoregressive*.

1.4.3.1 AR(1) process with constant

A process which satisfies the following difference equation,

$$X_t = \delta + \phi X_{t-1} + \epsilon_t \tag{1.10}$$

is called *first-order autoregressive model*, denoted by AR(1).

An AR(1) process can be represented as a MA(∞) process as follows:

$$\begin{aligned} X_t &= \delta + \phi X_{t-1} + \epsilon_t \\ &= \delta + \phi(\delta + \phi X_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= ... = \delta \sum_{i=0}^{k} \phi^i + \sum_{i=0}^{k} \phi^i \epsilon_{t-i} + \phi^{k+1} X_{t-k-1}, \end{aligned}$$

The asymptotic behaviour of the time series as $k \to \infty$ depends on the values of ϕ. Specifically, if $|\phi| > 1$ the series is explosive whereas the MA(∞) representation is achieved only for $|\phi| < 1$. Then, $\phi^{k+1} \to 0$ and $\sum_{i=0}^{k} \phi^i = \frac{1}{1-\phi}$ as $k \to \infty$. Hence, the AR(1) model can be described by the equation

$$X_t = \frac{\delta}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}. \tag{1.11}$$

Let X_0 be zero. By taking expectations of the 1.11, the mean of AR(1) process is $\mu = E(X_t) = \frac{\delta}{1-\phi}$, under the assumption of $|\phi| < 1$. The variance of the model 1.10 is

$$\begin{aligned} \gamma_X(0) = Var(X_t) &= \sigma^2(1 + \phi^2 + \phi^4 + \phi^6 + ...) \\ &= \frac{\sigma_\epsilon^2}{1-\phi^2}, \end{aligned}$$

while the hth autocovariance is

$$\begin{aligned} \gamma_X(h) &= Cov(X_t, X_{t-h}) = E(X_t - \mu)(X_{t-h} - \mu) \\ &= (\epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + ...)(\epsilon_{t-h} + \phi\epsilon_{t-h-1} + \phi^2\epsilon_{t-h-2} + ...) \\ &= \sigma_\epsilon^2 \phi^h (1 + \phi^2 + \phi^4 + ...) \\ &= \frac{\phi^h}{1-\phi^2}\sigma_\epsilon^2. \end{aligned}$$

The hth autocorrelation function derives from

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^h. \tag{1.12}$$

The *acf* decays geometrically as h increases. Hence, the acf of the AR(1) model is different from zero and the sign of it cycles infinitely between positive and negative, depending on the value of h. The following figures demonstrate the *acf* and *pacf* plots of an AR(1) process with $d = 5$ and $\phi = 0.6$ for 100 realisations of the process.

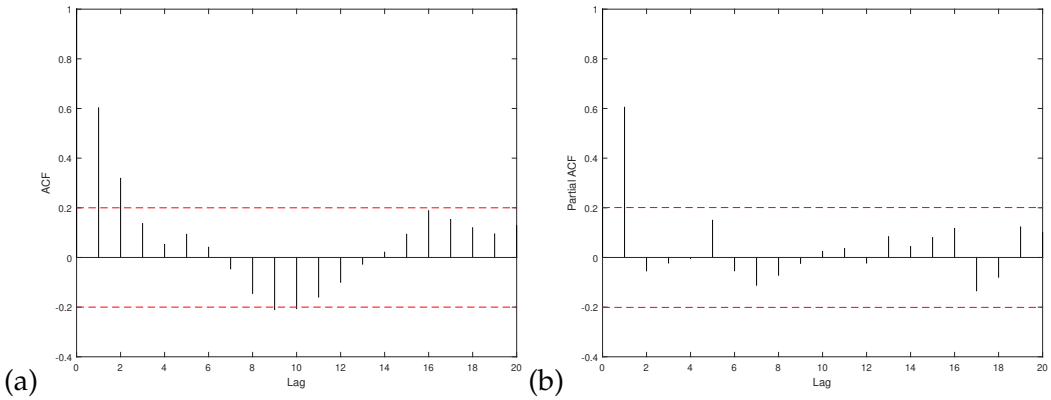


FIGURE 1.3: (a) ACF plot (b) Partial ACF plot

Consequently, AR(1) is a covariance-stationary process under the assumption of $|\phi| < 1$, since the mean and the autocovariances are independent of time. The stationarity of the AR(1) model can also be proved in terms of the *lag operator L*,

which is defined as $L^i X_t = X_{t-i}$ for all i. Thus, 1.10 becomes

$$X_t = \delta + \phi X_{t-1} + \epsilon_t \tag{1.13}$$
$$\Rightarrow (1 - \phi L)(X_t - \delta) = \epsilon_t \tag{1.14}$$
$$\Rightarrow \Phi(L)(X_t - \delta) = \epsilon_t, \tag{1.15}$$

where $\Phi(L) = 1 - \phi_1 L$ is the *characteristic polynomial* of the AR(1) process. If $\phi(L)$ converges, 1.15 will be

$$X_t - \delta = [\Phi(L)]^{-1}\epsilon_t. \tag{1.16}$$

Let $\Psi(L) = [\Phi(L)]^{-1}$. Hence,

$$\Psi(L) = (1 - \phi L)^{-1} = \sum_{i=0}^{\infty} \phi_1^i L^i. \tag{1.17}$$

$\Psi(L)$ converges if and only if $|\phi| < 1$ and $|L| < 1$. Therefore, the root, $z = \frac{1}{\phi}$ of the equation $\Phi(z) = 0 \Rightarrow 1 - \phi z = 0$ must lie outside the unit circle. That is,

$$|z| > 1 \Leftrightarrow |\frac{1}{\phi}| > 1 \Leftrightarrow |\phi| < 1.$$

The constant δ is considered for facilitation zero.

1.4.3.2 AR(1) process without constant

The AR(1) model without constant is described by the equation

$$X_t = \phi X_{t-1} + \epsilon_t \tag{1.18}$$

where, ϵ_t is a white noise process. The above process comprises a subcase of the general AR(1) process with constant, simply by substituting $\delta = 0$. Therefore, the representation of it as a MA(∞) process is

$$X_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}, \tag{1.19}$$

and implies that the process is stationary only if $|\phi| < 1$. The following figure illustrates an AR(1) model with $\delta = 5$ opposed to a process without constant regarding a zero initial value.

The first and second-order moments are:

$$E(X_t) = 0$$
$$\gamma_X(0) = Var(X_t) = \frac{\sigma_\epsilon^2}{1 - \phi^2}$$
$$\gamma_X(h) = Cov(X_t, X_{t-h}) = \phi^h \frac{\sigma_\epsilon^2}{1 - \phi^2}$$

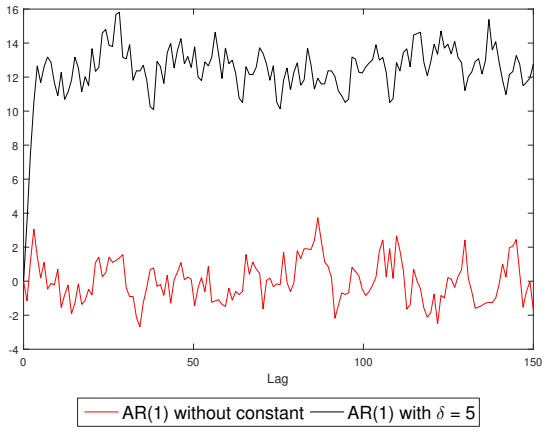


FIGURE 1.4: AR(1) model with and without constant

1.4.3.3 AR(p) process

The pth-order autoregressive model, denoted AR(p) satisfies

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t \tag{1.20}$$

where, $\vec{\phi} = (\phi_1, \phi_2, ..., \phi_p)$ any real number and ϵ_t a white noise. The use of the *lag operator* L facilitates the notation of the 1.20. Hence, the AR(p) can be described by

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t \tag{1.21}$$
$$\Rightarrow (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) X_t = \epsilon_t + \delta \tag{1.22}$$
$$\Rightarrow \Phi(L) X_t = \epsilon_t + \delta. \tag{1.23}$$

The $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p$ is called *characteristic polynomial* of AR(p) model. If $\Phi(L)$ converges, 1.23 takes the form of a stable difference equation

$$X_t = \phi(L)^{-1} \epsilon_t + \phi(L)^{-1} \delta$$

The convergence of $\Phi(L)$ can be achieved only when the roots of the equation

$$\Phi(z) = 0 \Rightarrow 1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p = 0$$

lie outside the unit circle. Essentially, the stationarity of an AR(p) model depends on the value of $\vec{\phi} = (\phi_1, \phi_2, ..., \phi_p)$, while MA models do not need any restrictions on $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_q)$ to achieve stationarity. Under the assumption of stationarity the mean of the process is a constant $\mu = E(X_t) = E(X_{t-1}) = E(X_{t-2}) = ...$, where,

$$\mu = \delta + \phi_1 \mu + \phi_2 \mu + ... + \phi_p \mu \Rightarrow$$
$$\mu = \frac{\delta}{1 - \phi_1 - \phi_2 - ... - \phi_p}.$$

The equation 1.20 via 1.23 takes the form:

$$X_t - \mu = \phi_1 (X_{t-1} - \mu) + \phi_2 (X_{t-2} - \mu) + ... + \phi_p (X_{t-p} - \mu) + \epsilon_t \tag{1.24}$$

Multiplying 1.24 with $(X_{t-j} - \mu)$ and taking expectations, the autocovariances of AR(p) process are

$$\gamma_j = \phi_1\gamma_1 + \phi_2\gamma_2 + ... + \phi_p\gamma_p + \sigma_\epsilon^2, \qquad for \quad j = 0 \tag{1.25}$$

and

$$\gamma_X(j) = \phi_1\gamma_X(j-1) + \phi_2\gamma_X(j-2) + ... + \phi_p\gamma_X(j-p), \qquad for \quad j = 1, 2, .. \tag{1.26}$$

The autocorrelations of the process are produced by dividing 1.26 by $\gamma_X(0)$

$$\rho_X(j) = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} + ... + \phi_p\rho_{j-p} \qquad for \quad j = 1, 2, ... \tag{1.27}$$

The above equations are known as *Yule-Walker equations*. The solution of these equations with respect to $\phi_1, \phi_2, ..$, produce the *partial autocorrelations* for every value of order p.

Definition 1.4.1. The **partial autocorrelation function** (PACF) of a stationary process is the function $\alpha(\cdot)$ defined by the equations

$$\alpha(0) = 1, \quad for \quad h = 0$$

and

$$\alpha(h) = \phi_{hh}, \quad for \quad h \geq 1$$

where, ϕ_{hh} is the last component of $\phi_h = \Gamma_h^{-1}$, Γ_h is the autocovariance matrix and γ_h is the variance vector of the process.

More analytically, the Yule-Walker equation for the AR(1) model i.e. $p = 1$, is $\rho_1 = \phi_1$. Thus, the partial autocorrelation will be $\alpha(1) = \phi_1 = \rho_1$, different from zero, while $\alpha(h) = 0$ for higher order lags. Similarly, the Yule-Walker equations for an AR(2) process are

$$\rho_1 = \phi_1 + \phi_2\rho_1$$
$$\rho_2 = \phi_1\rho_1 + \phi_2$$

the solution of which, gives the non-zero partial autocorrelation at lag 2

$$\alpha(2) = \phi_2 = \frac{\hat{\rho_2} - \hat{\rho_1^2}}{1 - \hat{\rho_1^2}}$$

and $\alpha(h) = 0$ for $h > 2$.

Intuitively, the partial autocorrelation at lag h denotes the autocorrelation between X_t and X_{t+h}, with the effect of correlation at shorter lag terms removed.

A notable inference about acf and $pacf$ is that in an AR(p) model all autocorrelations are different from zero, while the partials autocorrelations are all zero for lags higher than p.

1.4.4 ARMA process

The model which contains p autoregressive and q moving-average terms is called *autoregressive moving-average process*, denoted by ARMA(p,q), and satisfies the equation

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}, \tag{1.28}$$

where $\{\epsilon_t\}$ is a white noise. In terms of the lag operator L, the 1.28 can be expressed as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} \tag{1.29}$$
$$\Rightarrow X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = \epsilon_t + \theta_1 X_{t-1} + ... + \theta_p X_{t-p} \tag{1.30}$$
$$\Rightarrow (1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) X_t = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q) \epsilon_t \tag{1.31}$$
$$\Rightarrow \Phi(L) X_t = \Theta(L) \epsilon_t, \tag{1.32}$$

where $\Phi(L)$ and $\Theta(L)$ are the characteristic polynomials of order p and q respectively. The equation 1.32 is stationary only if the roots of $\Phi(L)$ lie outside the unit circle, as in the AR(p) model. For the stationarity of the MA component no restrictions are needed. The invertibility of the MA component requires, however, that the roots of $\Theta(L)$ lie outside the unit circle.

The calculations of $acvf$ and acf for an ARMA(p,q) model are more complicated than those of an AR(R) or MA(q) model, however they are based on the same theory.

1.5 Non-Stationary Models

1.5.1 Introduction

Up to this point, the models presented satisfied the stationarity assumption since their mean, variance and autocovariance were all independent of time. The scope of this section is to outline some basic cases where weak stationarity is violated.

The interest is predominantly focused on the case where the mean of the time series does not vary about a constant value but presents a *trend*. A *trend* in a process exhibits the sustained and systematic variations over time and can be hardly or precisely predictable. In the first case, the trend is called *deterministic* and in the second *stochastic*.

The current section presents the models with *trend* and demonstrates how they could be transformed in order to achieve stationarity. The transformed models are divided into *trend stationary* and *difference stationary* models, depending on the kind of trend, *deterministic* or *stochastic*.

1.5.2 Stochastic trends

Let

$$\Phi(L) X_t = \Theta(L) \epsilon_t \tag{1.33}$$

be an ARMA(p,q) process where $\Phi(L)$ and $\Theta(L)$ are the characteristic polynomials of order p and q respectively. The stationarity condition for an ARMA(p,q) model is that all the roots of the AR polynomial lie outside the unit circle. If at least one root

of the polynomial does not lie outside the unit circle then the series will be explosive and the polynomial has at least one *unit root*.

Let the polynomial $\Phi(L)$ have one *unit root*, that is, $\Phi(L)$ can be factorized as $\Phi(L) = \Phi^*(L)(1-L)$. Hence, the equation 1.33 becomes

$$\Phi^*(L)(1-L)X_t = \Theta(L)\epsilon_t \Rightarrow$$

$$\Phi^*(L)\Delta X_t = \Theta(L)\epsilon_t \tag{1.34}$$

where $\Phi^*(L)$ is a polynomial of order $p-1$ and $\Delta = 1-L$ is the *first order difference operator*. If $\Phi^*(L)$ does not have any more unit roots, the 1.34 expresses a stationary process. In the case of more than one, say d, unit roots the AR polynomial will be factorized as

$$\Phi(L) = \Phi^*(L)(1-L)^d, \tag{1.35}$$

where $\Phi^*(L)$ is a $p-d$ order polynomial. Hence, the initial equation 1.33 will be expressed as

$$\Phi^*(L)\Delta^d X_t = \Theta(L)\epsilon_t, \tag{1.36}$$

where $\Delta^d(L)$ is the dth order *lag operator*.

In short, 1.33 exhibits a non-stationary process with a stochastic trend, which implies permanent shocks on the time series, while the dth difference of it, $d \geq 1$ follows a stationary and invertible ARMA(p,d,q) model. When a non-stationary process needs to be d times differenced in order to achieve stationarity, the series is said to be *integrated of order d*, denoted I(d). An I(0) process is a stationary process.

An ARMA(p,q) model which has been differenced d times is called *autoregressive integrated moving average model*, denoted ARIMA(p,d,q).

Two fundamental ARIMA processes are: the random walk without drift and the random walk with drift.

1.5.2.1 Random walk model without drift

Let $X_t = \phi X_{t-1} + \epsilon_t$ be an AR(1) model without drift, where ϵ_t is a white noise process. It has been proved that the AR(1) model is stationary only if $|\phi| < 1$.

In the case of $|\phi| > 1$ the series is explosive, which means that the *shocks* to the system are not only persistent over time, but they have an extremely large impact on the behaviour of the series. If $\phi = 1$, the process

$$X_t = X_{t-1} + \epsilon_t, \tag{1.37}$$

is called *random walk model* or *AR(1) model with a unit root*. A simulated random walk process without drift and *iid* normally generated innovations is illustrated in figure 1.5.

The equation 1.37 is a non-stationary process since the mean and variance of the process is *time dependent*. By repeated substitution 1.37 yields

$$X_t = \sum_{i=0}^{t-1} \epsilon_{t-j}, \tag{1.38}$$

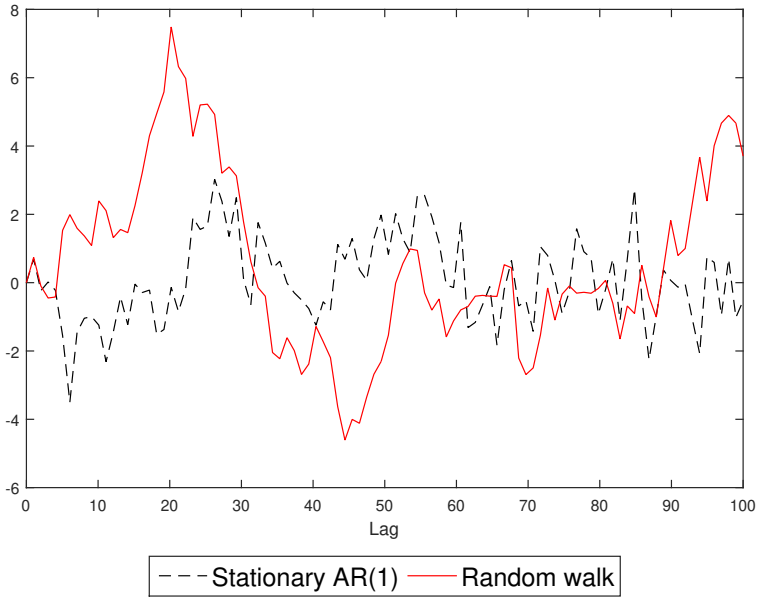


FIGURE 1.5: AR(1) and Random walk processes

where the initial observation, x_0, is fixed at zero. The first and second moments of 1.38 are thus

$$E(X_t) = E(\sum_{i=0}^{t-1} \epsilon_{t-j}) = 0, \tag{1.39}$$

$$\gamma_X(0) = Var(X_t) = Var(\sum_{i=0}^{t-1} \epsilon_{t-j}) = t\sigma_\epsilon^2 \tag{1.40}$$

and

$$\gamma_X(h) = Cov(X_t, X_{t-h}) = E(\epsilon_1 + ... + \epsilon_t)(\epsilon_1 + ... + \epsilon_{t-h}) = (t-h)\sigma_\epsilon^2. \tag{1.41}$$

From the autocorrelation function,

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \frac{(t-h)\sigma_\epsilon^2}{t\sigma_\epsilon^2}, \tag{1.42}$$

it can be easily inferred that as the sample size tends to infinity the acf of the process approaches one, which is a sign for a non-stationary process.

Generally, although the process has a constant mean of zero, the autocovariances and autocorrelation depend on time t, which proves the existence of a unit root.

Hence, the statistical properties of the random walk cannot be analyzed by using the conventional asymptotic theory, since all the asymptotic theorems require weak stationarity.

The process of *differencing* transforms the model in 1.37 into a stationary process, as follows:

$$\begin{aligned} X_t &= X_{t-1} + \epsilon_t, \\ (1-L)X_t &= \epsilon_t, \\ \Delta X_t &= \epsilon_t, \end{aligned} \tag{1.43}$$

and therefore,

$$E(X_t) = E(\epsilon_t) = 0$$

$$\gamma_X(0) = Var(X_t) = Var(\epsilon_t) = \sigma_\epsilon^2$$
$$\gamma_X(h) = Cov(X_t, X_{t-h}) = E(\epsilon_t \epsilon_{t-h}) = 0$$

Hence, the model in 1.37 is a *difference stationary* I(1) process.

1.5.2.2 Random walk with drift model

The random walk model with drift is essentially an AR(1) model with drift and $\phi = 1$,

$$X_t = \delta + X_{t-1} + \epsilon_t. \tag{1.44}$$

The drift imposes a direction to the random walk, as shown in figure 1.6(a), which illustrates 10 random walks with drift, δ, equal to 0.5. For the calculation of the

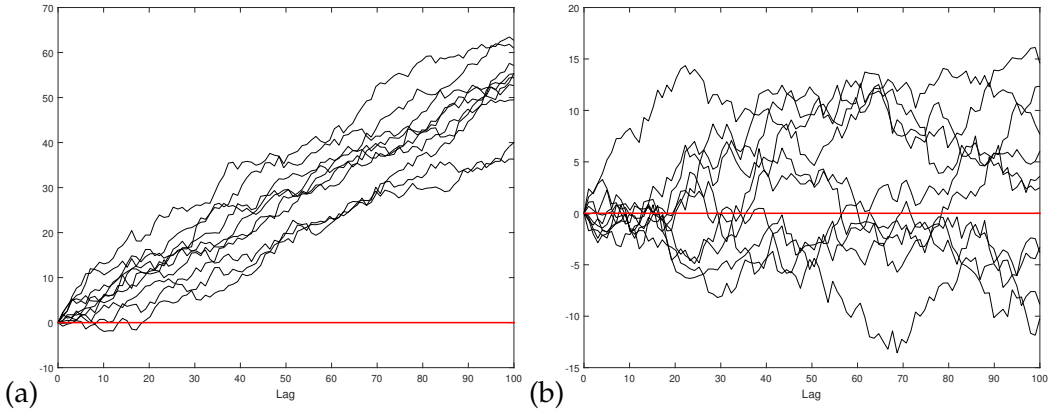


FIGURE 1.6: (a) Random walk with drift $\delta = 0.5$ (b) Random walk without drift

mean, variance and autocovariance of the process it is convenient to express 1.44 as a linear function of the ϵ_ts, as follows:

$$X_t = X_0 + \delta t + \sum_{i=1}^{t} \epsilon_i. \tag{1.45}$$

For simplicity, the initial observation X_0 is taken to be zero. Hence,

$$E(X_t) = E(\delta t + \sum_{i=1}^{t} \epsilon_i), = \delta t$$

$$Var(X_t) = Var(\delta t + \sum_{i=1}^{t} \epsilon_i) = t\sigma^2$$

and

$$\begin{aligned}\gamma_X(h) = Cov(X_t, X_{t-h}) &= E(X_t - \delta t)(X_{t-h} - (t-h)\delta) \\ &= E(\sum_{i=1}^{t} \epsilon_i)(\sum_{i=1}^{t-h} \epsilon_i) \\ &= (t-h)\sigma^2.\end{aligned}$$

Since the first- and second-order moments are functions of time, the process is non-stationary. By using the lag operator L, the above process is expressed as

$$(1 - L)X_t = \delta + \epsilon_t \Rightarrow,$$

$$\Delta X_t = \delta + \epsilon_t \tag{1.46}$$

It can be easily proved that the mean and autocovariances of the differenced process resemble those of a stationary time series. Thus, the 1.46 is a *difference stationary* I(1) process.

1.5.3 Deterministic trends

A simple model with a deterministic trend has the form

$$X_t = \mu_t + \epsilon_t, \tag{1.47}$$

where ϵ_t follows a white noise process and μ_t is a deterministic function of time. This function may have a linear form, for instance

$$\mu_t = \alpha + \beta t,$$

where, α denotes the *intercept* and β the slope of the deterministic trend, or a nonlinear form, such as a polynomial time trend,

$$\mu_t = \sum_{i=0}^{k} \beta_i t^i.$$

A short analysis on the statistic behaviour of

$$X_t = \alpha + \beta t + \epsilon_t \tag{1.48}$$

portrays that the mean of the process is a function of time, whereas the variance remains stable over time.

$$E(X_t) = E(\alpha + \beta t + \epsilon_t) = \alpha + \beta t,$$

$$Var(X_t) = Var(\alpha + \beta t + \epsilon_t) = \sigma^2.$$

A process with these properties is called a *trend-stationary process*. Forecasts can be precisely calculated, provided that the values of α, β and t are known and the subtraction of $\alpha + \beta t$ from X_t produces a *stationary process*. Hence, a kind of trend-stationarity is implied.

The choice of ϵ_t is not restrictive. In the place of it there may be any other stationary process, such as an ARMA(p,q) process. The following figure illustrates 50 paths of length 200 of an ARMA(1,1) trend stationary process with $\alpha = 0$, $\beta = 0.5$, $\phi = 0.9$ and $\theta = 0.4$. The innovations ϵ_t are considered to be normally distributed with zero mean and variance $\sigma^2 = 9$.

It can be inferred that the sample paths fluctuate around the theoretical trend line with constant variance and the simulation mean is quite close to the true trend line.

It is worth mentioning that the process of differencing the mean $\alpha + \beta t$ could not be applied in a unit root process, as in 1.45, since the linear time-trend may be

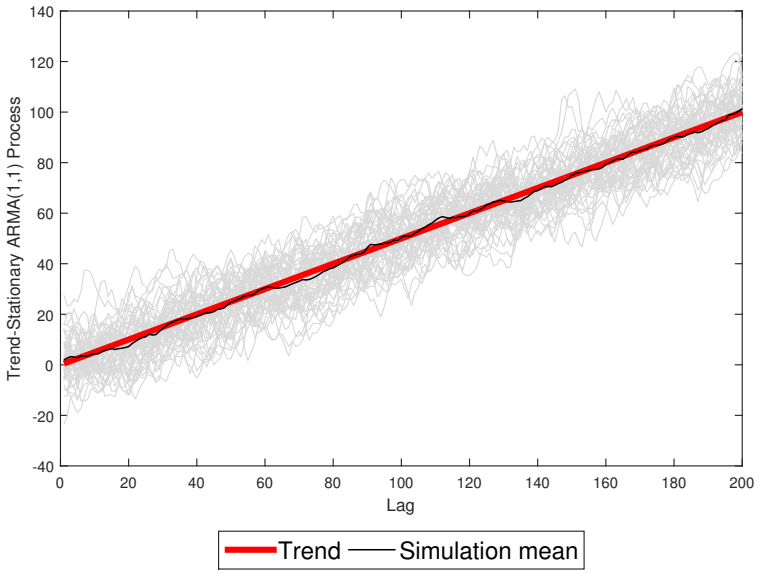


FIGURE 1.7: A trend-stationary process

subtracted, whereas the variance grows over time. Thus, the process would not be stationary.

1.6 Unit root tests

A non-stationary time series can be transformed into a stationary time series either by de-trending, that is, by subtracting the time trend, or by differencing. In the first case, the process is *trend-stationary*, while in the second the process is *difference-stationary*. The question lies on the kind of non-stationarity of the process.

If the time series is *trend-stationary* and a dth-difference is applied, then the series would be *overdifferenced*. On the other hand, if the time series is *difference-stationary* and a de-trending method is used, then the series will be *underdifferenced*.

The majority of the series in economics or finance are mainly described by *stochastic* non-stationary models, i.e. models with stochastic trends, rather than models with deterministic non-stationarity. Hence, the detection of a *unit root* in an autoregressive process determines whether the trending data should be first differenced or regressed on a function of time in order to achieve stationarity.

A *unit root test* tests the hypothesis of $\phi = 1$ against the one-sided alternative $|\phi| < 1$ in the extended time series model:

$$\begin{aligned} X_t &= \alpha + \beta t + z_t, \\ z_t &= \phi z_{t-1} + \epsilon_t, \end{aligned}$$

where ϵ_t follows a white noise process. If $|\phi| < 1$, then X_t presents a trend-stationary process. Conversely, under the hypothesis of $|\phi| = 1$, X_t contains a stochastic trend and therefore, has to be differenced, as follows:

$$\begin{aligned} X_t - X_{t-1} &= \alpha + \beta t + z_t - \alpha - \beta(t-1) - z_{t-1} \\ &= \beta + \epsilon_t, \end{aligned}$$

since $z_t = z_{t-1} + \epsilon_t$. Hence, under the null hypothesis X_t is an I(1) with drift process.

With regards to the nature of the test, none of the conventional tests, such as the t-test, can be applied under the hypothesis that $|\phi| = 1$, since X_t does not follow an asymptotic normal distribution. Thus, the t-statistic for the unit root test arises form a relevant asymptotic distribution based on the *Wiener process* (Maddala Kim, 1998).

A detailed description of the *unit root tests* and specifically, of the *autoregressive unit root tests* will be presented in next chapter.

Chapter 2

Bayesian Theory

2.1 Basics of Bayesian theory

The process of observing and analyzing past experience in order to predict unobserved phenomena comprises the cornerstone of the inductive theory. Based on the above claim, the Bayesian approach quantifies the uncertainty included in an experiment, using probability theory, which, in combination with observed data, contributes to making both inferences about the parameter of interest and reliable predictions.

The purpose of the current chapter is to outline some basic principles and characteristics of the Bayesian theory and subsequently focus on the model comparison via different methods.

Bayes' theorem is based on conditional probabilities and takes the following form for continuous random variables.

Theorem 1. (**Bayes' Theorem**) *Let $f(x,\theta)$ denote the joint probability density function (pdf) for a random observation vector x and a random parameter vector θ. Then,*

$$\begin{aligned} f(x,\theta) &= f(x|\theta)f(\theta) \\ &= f(\theta|x)f(x) \end{aligned}$$

and thus,

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}, \tag{2.1}$$

where $f(x) = \int f(x|\theta)f(\theta)\mathrm{d}\theta$ is the normalizing constant for the pdf in 2.1.

The conditional probability $f(\theta|x)$ is the *posterior pdf* for the parameter vector θ, while $f(\theta)$ represents the *a priori* knowledge about θ, denoted *prior pdf*. Furthermore, if the conditional probability $f(x|\theta)$ is viewed as a function of θ, it will yield the *likelihood function*, $L(\theta|x)$, of the observation vector x. All the information about θ in the data can be derived from the likelihood function.

The posterior distribution $f(\theta|x)$ is a genuine distribution, since $f(x)$ plays the role of a normalizing constant, which depends only on x. Hence, the relationship in 2.1 could be expressed as:

$$f(\theta|x) = cf(x|\theta)f(\theta) \tag{2.2}$$

and thus,

$$f(\theta|x) \propto f(x|\theta)f(\theta) \tag{2.3}$$

which implies that the posterior distribution is proportional to the likelihood times the prior.

2.2 Specifying the prior

Prior beliefs about parameter θ are reflected on the posterior inference. More specifically, the rational choice of the prior is not an objective issue, since it depends on the subjective beliefs of the analyst, which thus renders the Bayesian approach a subjective analysis. The influence of the prior, though, would be quite eliminated, as more data is available.

Even in the case of ignorance regarding *a priori* knowledge about θ, the choice of the prior distribution should reflect this ignorance. Hence, the process of *prior elicitation* should be based on how informative the prior is.

2.2.1 Conjugate priors

In the case of an *informative* prior, all the information included in the likelihood-prior blend is reflected on the posterior distribution. Nevertheless, a precise calculation of a numerical value of $f(\theta|x)$ may be quite difficult, mostly due to the complexity of integrals or the dimension of the parameter space.

These inherent difficulties do not comprise any huge drawback in the Bayesian inference, since a specific selection of priors are *conjugate* to the likelihood, $f(x|\theta)$, and permit the posterior distribution to rise without complicated integral computations. In essence, *conjugate priors* belong to the same probability distributional family as the posterior distributions.

The only case where the prior and the posterior are *conjugate* distributions, is when the data originates from models within the *exponential family*.

2.2.2 Non-informative priors

If a priori knowledge concerning θ is not available and the choice of one value of θ would not be favoured over another, then the use of a *non-informative prior* would be advisable. According to Kass and Wasserman (1996), non-informative priors could formally *represent the ignorance* about the parameter. However, the choice of them is neither *objective* nor *unique* and *are chosen by public agreement much like units of length and weight*.

Hence, the process of eliciting an appropriate non-informative prior distribution, is focused on the comparison of different priors, since the choice is not *unique*. A non-informative prior may be more efficient than another one, but it cannot be said that some priors are less informative than others (Robert,2007).

Two major drawbacks concerning non-informative priors are the *invariance under re-parameterisations* and the emergence of *improper priors*.

Improper priors

While bounded parameter spaces could generate *genuine* non-informative priors, unbounded parameter spaces mostly lead to *improper priors* or *flat priors*. More analytically, the form of the prior in the latter case is

$$f(\theta) = c, \tag{2.4}$$

for any $c > 0$. Hence,

$$\int f(\theta)\mathrm{d}\theta = \infty \tag{2.5}$$

that is, the distribution does not integrate to one and consequently does not comprise a *proper distribution*.

Priors not invariant to transformations

Some non-informative priors are not invariant under re-parameterisation. If that is the case, the insufficient knowledge about θ, implies no information about any transformation of θ, $g(\theta)$. The corresponding prior density for a one-to-one function $\phi = g(\theta)$ is given by the following transformation formula:

$$f_{\Phi}(\phi) = f(\theta) \times \left| \frac{d\theta}{d\phi} \right|. \tag{2.6}$$

As an application of the above statement let the prior be a uniform distribution over $(0, 1)$ that is $f(\theta) = 1$. If the prior is non-informative about θ, it emerges that there would not be any information about $\phi = 1/\theta$. By using the relationship 2.6, the distribution for ϕ is

$$f_{\Phi}(\phi) = \frac{1}{\phi^2}. \tag{2.7}$$

It can be easily inferred that a uniform prior for θ does not corresponds to a uniform distribution for $1/\theta$.

Jeffreys' prior

A prior which satisfies the requirement of invariant monotone transformations of the parameter is *Jeffreys' prior*, which is based on the Fisher information $I(\theta)$ given by

$$I_{\Theta}(\theta) = E\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 = -E\left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} \right), \tag{2.8}$$

$I(\theta)$ indicates the amount of information which resides in the model and thus the values which are more likely for the prior distribution are these, which enlarge Fisher's information. Hence, *the influence of the prior is minimized and the prior becomes as non-informative as possible*. (Robert, p.130, 2007)

Jeffreys' prior is defined as

$$\pi(\theta) \propto |I_{\Theta}(\theta)|^{1/2} \tag{2.9}$$

and is invariant under one-to-one re-parameterisation of θ. By using the transformation $\phi = g(\theta)$ and the chain rule of difference, it emerges that

$$\pi_{\Phi}(\phi) = \pi_{\Theta}(g^{-1}(\phi)) \times \left| \frac{\mathrm{d}g^{-1}(\phi)}{\mathrm{d}\phi} \right| = \pi_{\Theta}(\theta) \times \left| \frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right|. \tag{2.10}$$

In the vast majority of cases, Jeffreys' prior is *improper*, which is not a drawback for Bayesian inference since it could be handled by using proper mathematical devices.

2.3 Bayesian Inference

2.3.1 Decision Theory

The vast majority of statistical studies aim to provide analysts with *decisions*, according to which a specific action would be favoured over another one. Ideally, the decision could be made according to a utility function which would optimize a specific choice.

Due to the difficulty in specifying such a function, the interest focuses on an evaluation criterion which depends on the parameters of the model and the consequences this decision may cause. The process of comparing and eliminating unreasonable decisions, using an evaluation criterion, comprises the main scope of the *Bayesian theory*.

This criterion is called *loss function*, denoted by $L(\theta, \alpha)$, where α is the action adopted from an *action space* $\mathcal{A}$ and θ is the true state of nature of the parameter, which originates from a *parameter space* Θ. It represents the loss incurred when the action α is adopted.

Having observed data x, the posterior expected value of the loss function, called *posterior expected loss* or *posterior risk* is defined as:

$$\rho(\alpha, x) = E_{\theta|X}\left[L(\theta, \alpha)|x\right] = \int_{\Theta} L(\theta, \alpha) f(\theta|x) \mathrm{d}\theta. \tag{2.11}$$

In contrast to the frequentist approach, which imposes integration over the space X, as x is known, the Bayesian method integrates over the parameter space Θ , as θ is unknown, based on the posterior distribution.

The *Expected loss principle* advocates the selection of the action, $\alpha^*(x)$, with the minimum posterior expected loss, after data x has been observed. The action with this property is called *Bayes' estimator* or *Bayes' action* and the decision rule $\delta(x) = \alpha^*(x)$ is called *Bayesian decision rule*.

Some of the most frequently used loss function forms are:

0-1 Loss function

The *zero-one loss function* assigns uniform unit loss to an incorrect decision and zero loss to a correct one, as follows:

$$Ł(\theta, \alpha) = \begin{cases} 1, & \theta = \alpha, \\ 0, & \theta \neq \alpha. \end{cases}$$

This nonquantitative type of loss is widely used in the common frequentist hypothesis testing. The penalty is the same for every estimator that does correspond to the the true parameter and none otherwise, despite the distance between them.

The above loss function is minimized when α is identical to the *maximum a posteriori* probability.

The quadratic loss

The *quadratic loss* or *squared error loss* is given by

$$L(\theta, \alpha) = (\theta - \alpha)^2 \tag{2.12}$$

and is one of the most commonly used loss functions. Unlike the zero-one loss function, the quadratic loss indicates the difference between the estimation α and the real parameter θ. This type of function is more appropriate in the case of a continuous parameter space, where the distance of the values of θ and α is well-defined.

A drawback of this evaluation criterion is that, due to the quadratically increase, errors are not generally penalized.

The minimization of 2.12 is achieved for the *expectation of the posterior distribution*, that is $\alpha = E[f(\theta|x)]$.

Absolute error loss

An alternative, equally popular type of loss function is the *absolute error loss function*,

$$L(\theta, \alpha) = |\theta - \alpha| . \tag{2.13}$$

The optimization of the above function implies that the *Bayes' estimator* of θ is the *median* of the posterior distribution.

In frequentist usage, the expectation of the loss function over data x, with the parameter θ regarded as a constant, is defined as *frequentist risk* and is a function of θ. The frequentist risk function of a decision rule $\delta(x)$ is

$$R(\theta, \delta) = E_{X|\theta}[L(\theta, \delta(x)] = \int_X L(\theta, \delta(x)) f(x|\theta) \mathrm{d}x. \tag{2.14}$$

On the contrary, the *Bayesian risk* is the expected loss integrated over *both* θ and x and produces a single number. That is,

$$r(\theta, \delta) = E_\theta[R(\theta, \delta(x)] = \int_\Theta R(\theta, \delta(x) f(\theta) \mathrm{d}\theta. \tag{2.15}$$

Bayesian risk can also be expressed as,

$$r(\theta, \delta) = E_X[\rho(\theta, \delta)] = \int_X \rho(\theta, \delta(x)) f(x) \mathrm{d}x. \tag{2.16}$$

Consequently, the estimator of the parameter θ, $\hat{\theta}$, which minimizes the *Bayes risk* among all estimators is a *Bayes estimator*.

The updated inference about the parameter θ is summarized in the posterior distribution. The interpretation and handling of the posterior density (or cumulative distribution) may, however, provide more information than usually required.

Hence, the information derived from the posterior distribution could be summarized via point estimation, credible sets or hypothesis testing, the analogues of the common frequentist estimation techniques.

2.3.2 Point Estimation

The process of *point estimation* incorporates sample data and statistical methods in order to produce an approximation of the unknown parameter θ of the population. The *Bayesian point estimation* includes summary features of the posterior distribution, such as the posterior mean, the posterior mode, the posterior median and the posterior variance.

The *posterior mode* or *maximum a posteriori estimator* (MAP) is derived quite easily by calculating the numerator of 2.1, that is the non-standardized posterior distribution. Asymptotically, the MAP estimators coincide with the classical maximum likelihood estimators, since the information contained in large samples dominate the fixed knowledge provided by the prior. In the case of a flat prior, the posterior mode identifies with the maximum likelihood estimator (MLE) of θ.

As aforementioned, the posterior risk for the absolute error loss is minimized by the *posterior median*. Additionally, the *posterior mean* identifies with the Bayes' estimator which minimizes the squared loss function and thus the Bayesian risk.

It can be proved that the posterior mean minimizes the posterior variance over all possible estimators $\hat{\theta}$. Let $\mu = \mu(x)$ be the posterior mean $E_{\theta|x}(\theta)$. Therefore,

$$\begin{aligned} Var_{\theta|x}(x) = E_{\theta|x}(\theta - \hat{\theta})^2 &= E_{\theta|x}(\theta - \mu + \mu - \hat{\theta})^2 \\ &= Var_{\theta|x}(\theta) + (\mu - \hat{\theta})^2. \end{aligned}$$

Hence, $Var_{\theta|x}$ is minimized only when $\hat{\theta} = \mu$.

The approximation of the point estimators is achieved through several computation techniques which are relevant to Bayesian approach, such as *Markov Chain Monte Carlo* (MCMC) algorithms.

2.3.3 Interval estimation

Another inferential procedure in classical statistics is the *confidence interval* (CI). Under the Bayesian approach to inference, the analogue to the classical confidence interval is the *credible set* and it is defined as follows:

Definition 2.3.1. A $100 \times (1 - \alpha)\%$ credible set for θ is a subset C of the parameter space Θ such that

$$1 - \alpha \leq P(C|x) = \int_C P(\theta|x)\mathrm{d}\theta \tag{2.17}$$

where integration is replaced by summation for discrete components of θ.

The interpretation of the *credible set* differs starkly from that of the frequentist *confidence interval*. In the latter case, the meaning of the CI would be:

If the CI could be recalculated many times, then in $(1 - \alpha) \times 100\%$ cases the true parameter θ would be contained in the interval.

On the contrary, the credible set could be interpreted as follows:

The probability that the parameter θ lies within the interval given the data x is at least $1 - \alpha$.

The Bayesian analogue of the CI can be undoubtedly viewed as a probabilistic statement, since it is based only on the prior knowledge about θ and the observed data.

In opposition to the probabilistic standpoint, the CI requires a great amount of repetitions of the experiment in order to be precise, which is practically impossible. Since only one data set is available, the fact that θ lies in or out of the interval constitutes just an indication for this type of estimation and the accuracy of the procedure.

Nevertheless, not every set C with probability equal or greater than $1 - \alpha$ is legitimate, as the definition 2.3.1 implies. Due to the fact that the credible set is not uniquely defined, that is, any set C with probability $1 - \alpha$ would satisfy the definition, there exists the possibility that regions with 'plausible' values of θ are excluded from the interval.

Therefore, a preciser estimation and at the same time a tighter credible set could be derived by using as a region the *highest posterior density* (HPD), defined as:

$$C = \{\theta \in \Theta : f(\theta|x) \geq l(\alpha)\} \tag{2.18}$$

where, l is the largest constant such that $P(C|x) \geq 1 - \alpha$.

2.3.4 Hypothesis testing

A principal method of statistical inference is *hypothesis testing*. Classical hypothesis testing examines whether the *null* hypothesis H_0 can be *rejected* or *not* against an *alternative* hypothesis, H_1,

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

where, Θ_0 and Θ_1 are subspaces of the parameter space Θ. In the case wherein Θ_0 and Θ_1 consist of a single point, the hypothesis test becomes,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

The probabilistic nature of Bayesian hypothesis testing implies that the inference is based on the posterior probability of θ under hypotheses H_0 and H_1.

A substantial advantage of the Bayesian approach is that more than one hypotheses can be tested at the same time and they are not bounded to be *nested* one within the other, unlike with the frequentist approach. Thus, the alternative hypotheses are better described by the term 'models' M_i, $i = 1, .., m$, which, as detailed below, demonstrate a fundamental role in the Bayesian hypothesis testing.

To facilitate the notation, let the subspaces Θ_0 and Θ_1 contain a single point. A common frequentist way to compare the statistical models portrayed by the null and the alternative hypothesis is to evaluate the *likelihood ratio*,

$$L = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

Large values of the above ratio favor the choice of θ_1 against θ_0. The Bayesian analogue of this quantity is called *posterior odds* and is defined as follows,

$$L_B = \frac{f(\theta_0|x)}{f(\theta_1|x)}.$$

The posterior distributions for both hypotheses are,

$$f(\theta_1|x) = \frac{f(x|\theta_1)}{f(x)} = \frac{f(\theta_1)f(x|\theta_1)}{f(x\theta_0)\pi(x|\theta_0) + \pi(\theta_1)f(|\theta_1)}$$
$$f(\theta_0|x) = 1 - f(\theta_1|x)$$

The plausibility of the null and the alternative hypothesis is assessed by a quantity which involves the prior knowledge and the observed data termed as *Bayes' factor*. Essentially, it is the ratio of the posterior odds over the prior odds in favor of H_1.

Definition 2.3.2. (Bayes' factor) The Bayes' factor is the ratio of the posterior probability of the null and the alternative hypothesis over the ratio of the prior probabilities of the null and the alternative hypothesis.

$$BF = \frac{f(x|\theta_0)}{f(x|\theta_1)} = \frac{f(\theta_0|x)}{f(\theta_1|x)} \Big/ \frac{\pi(\theta_0)}{\pi(\theta_1)}. \tag{2.19}$$

In words,

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

In the general case, when the hypotheses to be tested are not simple ones, the Bayes' factor can be expressed as,

$$BF = \frac{f(x|\theta \in \Theta_0)}{f(x|\theta \in \Theta_1)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}. \tag{2.20}$$

The quantity $f(x|\theta \in \Theta_i)$ is the *marginal probability of the data* under each hypothesis H_i, $i = 0, 1$ which arises by *integrating* over the corresponding parameter space Θ_i, $i = 0, 1$. Thus,

$$f(x|H_i) = \int_{\Theta_i} f(x|\theta_i, H_i)\pi(\theta_i|H_i).\, \mathrm{d}\theta \tag{2.21}$$

If the a priori probabilities for H_0 and H_1 are equal, that is, $\pi(\theta \in \Theta_0) = \pi(\theta \in \Theta_1) = .5$, then the Bayes factor is identical to the posterior odds of H_1.

Jeffreys proposed a scale according to which, the null hypothesis is accepted or not. For computational convenience, the logarithmic version of the Bayes' factor is commonly used. Hence, the evidence *in favor of* H_0 is:

- *poor*, if log_{10}BF varies between 0 and 0.5.
- *substantial*, if it varies between 0.5 and 1.
- *strong*, if it is between 1 and 2.
- *decisive*, if the value of the logarithm is greater than 2.

Bayesian hypothesis testing differs from frequentist testing in terms of the comparative nature of the first and the fact that at least two models can be involved. Additionally, the evidence is *weighed* not only *in favor the null hypothesis* but also *against it*.

2.4 Bayesian Model Comparison

2.4.1 Basic information and motives

In the general case, model comparison can be applied to numerous statistical problems, such as variable selection in a regression model, the choice of the parametric family or the determination of the number of components in a mixture model.

Bayesian model comparison will not determinate whether a model is *true* or not, however it can assign *preference* to a particular model, including all the information in the data and the prior.

Model choice is the process of identifying the index $i \in \{1, 2, .., k\}$ which corresponds to the most appropriate model for the data among $M_1, .., M_k$, where,

$$M_i : x \sim f_i(x|\theta_i), \quad \theta_i \in \Theta_i, \quad i \in I. \tag{2.22}$$

In the case that the index set I is finite and a small number of models are compared, the process is quite facilitated. Hence, model comparison simulates better a point estimation process rather than hypothesis testing.

The complexity of the structures involved in the model comparison makes the computation of the unknown quantities quite difficult. Therefore, advanced theoretical and computational tools such as the Bayes' factor, the posterior distributions of the models and the MCMC methods are used for the model comparison.

2.4.2 Motivation for Bayesian model comparison

There is significant distinction between the classical and the Bayesian approach to model comparison. A few indicative reasons which favor the choice of Bayesian model comparison are demonstrated below.

- **Consistency of the Bayesian model selection.** The selection of a model under the Bayesian framework is a consistent process. Specifically, if the model that 'fits' the data is one of the entertained models, then the Bayesian method, under conditions and if enough data is provided, will conclude to this model. Classical inference could not guarantee the selection of the true model, even if statistical tools such as *p-value* and AIC support this claim.

- **Direct and intuitive mathematical structures**. The main statistical tools in the Bayesian perspective are the Bayes' factor and the posterior model probabilities. The inference is based on these intuitive and readily understandable quantities, unlike with frequentist statistic quantities, such as *p-value*, which are not always interpreted directly.

- **Parsimony**. The Bayesian model comparison leads to *parsimonious* models, that is, models which are simple and not over-parameterized. The principle of *parsimony* is directly linked with the principle of *Occam's razor*, which favors the simple models against complex ones. In this context, the difficulty that lurks within the construction of a model, is that over-parameterized and implausibly detailed models have usually better 'fit' than simpler ones, comprising thus a significant goodness of fit measure. A balance between *parsimony* and *goodness of fit* should be essential.

  In classical statistics, the problem of overfitting is handled by adding a *penalty term*, for example in some model selection criterion, such as AIC, whose value increases if the complexity of the model increases. Bayesian model comparison automatically and quantitatively embodies, since it acts like an Occam's razor and over-parameterized models are automatically penalized by the Bayes' rule.

- **Bayesian model selection is unaffected by the number of models.** Contrarily, the classical approach advocates that, when only two models are involved, the model choice is identified by a hypothesis testing. In the challenging case of more than two models, the framework differs from hypothesis testing and other statistical tools are required to handle model comparison.

- **The entertained models do not have to be nested.** A substantial motive for choosing the Bayesian model comparison is the ability to compare models which, unlike with the classical framework, do not satisfy any initial assumptions, such as the *nested models* or *standardized distributions*.

- **The uncertainty of a model is taken into account.** In the classical approach, the estimations of the model parameters may be biased due to the fact that the model is based on the data and the same data (or part of it) is then used for the estimation of the parameters. It thus arises, that, under the assumption of the correctness of the model, the estimates and their interpretations are implausibly optimistic.

  Under the Bayesian approach, a mechanism named *Bayesian model averaging* is accounting for the model uncertainty, since the parameter estimates (or the predictions) are obtained by using the *weighted average* of the parameters (or the predictions) for each model. The weights for each model are determined from the posterior probabilities of each model. Hence, not only the uncertainty concerning a parameter given a particular model is taken into consideration, but also the uncertainty across all models is combined.

2.4.3 Standard framework and modelling

Let k denote the number of different models for the data x involved, where

$$M_i \sim f_i(x|\theta_i), \quad i = 1, .., k. \tag{2.23}$$

Under the ith model the domain of the parameter θ is the subspace Θ_i of Θ. Then, the prior associated with each model is $f(\theta|M_i)$ and reflects the knowledge about the model parameter under the hypothesis of the ith model. Inference under model uncertainty is divided into two levels.

1. **Model fitting**. Initially, a model M_i is selected to 'fit' the data x. The model's parameters are given by the posterior probability of θ,

$$f(\theta|x, M_i) = \frac{f(x|\theta, M_i) f(\theta|M_i)}{f(x|M_i)}, \tag{2.24}$$

   where $f(x|M_i)$ is the *marginal* or *predictive* density of x, namely,

$$f(x|M_i) = \int_{\Theta_i} f(x|\theta_i, M_i) f(\theta_i|M_i), \, \mathrm{d}\theta_i \tag{2.25}$$

   As shown later, the normalizing constant $f(x|M_i)$ is called *evidence* and indicates the preference for a model, since it represents the probability that the data originate from model M.

   The most probable values of θ can be obtained by finding the maximum a posteriori value (MAP).

2. **Model comparison**. The next level of inference after the construction of the models is the detection of the *true* model, which actually 'fits' the data. Therefore, each model needs an *a priori* knowledge before the involvement of the

data, which is denoted be $f(M_i)$. The *posterior probability* of the model M_i is

$$f(M_i|x) \propto f(x|M_i)f(M_i). \tag{2.26}$$

The denominator of the right member of the equation is the *evidence of all incorporating models* and is given by,

$$f(x) = \sum_{i=1}^{k} f(M_i) \int_{\Theta_i} f(x|\theta_i, M_i) f(\theta_i|M_i)\, \mathrm{d}\theta_i. \tag{2.27}$$

The omission of the denominator in the above quantity would not induce any significant difficulty, since the process of looking for new models is persistent and thus, the hypothesis space does not need to be completely defined.

In the case where the prior probabilities $f(M_i)$ are close, the posterior probabilities of the models are mainly based on the *evidence of* M_i.

Prior choice

In Bayesian analysis, the question whether a prior for a model is correct or not is not of great importance. Different priors correspond to different posteriors. However, if the priors do not differ significantly, then the relationship in 2.26 implies that the inference for the posterior model probability is mainly based on the evidence.

In the specific but common case of equal prior model probabilities for $i = 1, ..k$, that is $f(M_i) = \frac{1}{k}$, the posterior model probability is given by,

$$f(M_i|x) = \frac{f(x|M_i)f(M_i)}{f(x)} = \frac{f(x|M_i)}{\sum_{i=1}^{k} \int_{\Theta_i} f(x|\theta_i, M_i) f(\theta_i|M_i)\, \mathrm{d}\theta_i},$$

and are the same as the renormalized probabilities (Berger Pericchi, 2001).

Evaluating the evidence

The *marginal likelihood* or *evidence* of model M_i denotes the preference shown to this model and as detailed below, is inextricably bounded up with the concept of *Occam's razor*. It is calculated by marginalizing over the parameters, as follows

$$f(x|M_i) = \int_{\Theta_i} f(x|\theta_i, M_i) f(\theta_i|M_i).\, \mathrm{d}\theta_i \tag{2.28}$$

The posterior distribution of the parameter θ, $f(\theta|x, M_i)$, is maximized, as aforementioned, at the *posterior mode*, denoted by θ_{PM}. Then the evidence could be approximated by the maximum value of $f(x|\theta, M_i)f(\theta|M_i)$ times the width $\Delta\theta$. Hence,

$$f(x|M_i) \propto \underbrace{f(x|\theta_{PM}, M_i)}\, \underbrace{f(\theta_{PM}|H_i)\Delta\theta_{PM}}. \tag{2.29}$$

In words,

$$\text{Evidence} \propto \text{Likelihood's maximum} \times \text{Occam's factor}$$

The *Occam's factor* is a less than one quantity which automatically penalizes the model M_i for having the parameter θ. It emerges that, the Occam's factor penalizes model complexity and thus, the evidence becomes a measure of *plausibility* of the model.

Bayes' factor

After the construction of the models and the selection of the priors, the inference concerning the testing of the models is provided by the *Bayes' factor*, which quantifies the preference for one model over another.

As aforementioned, the Bayes' factor is the ratio of the posterior odds over the prior odds for the null and the alternative hypotheses. In terms of models, the Bayes' factor could be interpreted as the *odds provided by the data for model M_i versus model M_j* and is defined as,

$$BF_{ij} = \frac{f(x|M_i)}{f(x|M_j)} = \frac{f(M_i|x)}{f(M_j|x)} \Big/ \frac{f(M_i)}{f(M_j)}. \tag{2.30}$$

Hence,

$$BF_{ij} = \frac{\int_{\Theta_i} f(x|\theta_i, M_i) f(\theta_i|M_i)\, \mathrm{d}\grave{}_{\mathrm{i}}}{\int_{\Theta_j} f(x|\theta_i, M_i) f(\theta_i|M_i)\, \mathrm{d}\theta_i}. \tag{2.31}$$

After the evaluation of the evidences for each model, the preference for model M_i is indicated by the value classification in subsection 2.3.4.

Schwartz's criterion

Another criterion that can be used for model selection is the *Bayesian Information Criterion* (BIC), proposed by Schwartz (1978) and Akaike (1977, 1978). In the case of a large number of parameters, the criterion adds a penalty term on the number of parameters and favours parsimonious models over complex, overfitted models.

The mathematical formula for BIC under the hypothesis of model M_i is,

$$BIC = -2\log L_i(\hat{\theta}_i) + p_i \log n \tag{2.32}$$

where, $L_i(\hat{\theta}_i)$ is the maximum likelihood of the data, p_i denotes the number of the unknown parameters and n the size of the data.

The maths behind the formula do not require any Bayesian method, although the name of the criterion is Bayesian. The procedure is briefly described below.

As already mentioned, the posterior probability of the model M_i is,

$$f(M_i|x) = \frac{f(x|M_i) f(M_i)}{\sum_{i=1}^{k} f(M_i) f(x|M_i)}. \tag{2.33}$$

The maximazation of the above relationship is equivalent to the maximization of

$$f(M_i) f(x|M_i) = f(M_i) \int_{\Theta_i} f(x|\theta_i, M_i) \cdot f(\theta_i|M_i)\, \mathrm{d}\theta_i. \tag{2.34}$$

The evidence $f(x|M_i)$ can be expressed as,

$$f(x|M_i) = \int_{\Theta_i} \exp\{n \cdot m_i(\theta_i|x, M_i)\}\, \mathrm{d}\theta_i, \tag{2.35}$$

where

$$m_i(\theta_i|x, M_i) = n^{-1}\{\log f(x|\theta_i, M_i) + \log f(\theta_i|M_i)\}.$$

Under regularity conditions the *Laplace approximation* can be applied to 2.35. By substituting the Laplace approximation to 2.34 and taking the logarithm of the resulted formula, the Schwartz's criterion is derived.

If errors are identically, independently and normally distributed and the derivative of the log-likelihood with respect to the true variance is zero, the criterion under the hypothesis of model M_i becomes

$$BIC = \log\hat{\sigma_i^2} - p_i\frac{\log n}{n},$$

where $\hat{\sigma^2}$ is the error variance, p_i is the number of parameters and n the sample size.

The Bayesian Information Criterion penalizes more heavily the over-parameterized models compared to Akaike's Information Criterion (AIC), since it contains the parameter p_i which resembles the number of unknown parameters and thus increases the goodness of fit of the model.

The selection criterion when comparing models is denoted by the value of BIC. The preference is assigned to the model with the *lowest* value, since it indicates lower penalty terms.

Bayesian deviance criterion

The *Deviance Information Criterion* (DIC) was proposed by Spiegelhalter (2002) and could be interpreted as a generalization of the AIC and BIC criteria. It explicitly balances model complexity and goodness-of-fit, based on the *deviance*,

$$D(\theta) = -2\log f(x|\theta). \tag{2.36}$$

Since the deviance cannot merely discriminate models due to its bias in favour of over-parameterized models, a penalty term is introduced and the criterion becomes

$$DIC = E\left[D(\theta|x)\right] + q_D, \tag{2.37}$$

where $q_D = \{E\left[D(\theta|x)\right] - D(E\left[\theta|x\right])$.

The above criterion is the Bayesian alternative of AIC and BIC and is motivated mainly by the fact that is accounting for the priors (even when they are improper) and the simple and direct calculation from a sample generated by a MCMC simulation. In contrast to other criteria where the evaluation of the maximum likelihood value of the evidence is essential, the deviance criterion is automatically derived from the data.

With regards to the selection criterion, models with low value should be preferred to those with larger value.

Chapter 3

Autoregressive unit root tests

3.1 Introduction

The trending behavior or the non-stationary mean appear to be dominant characteristics of the time series and are usually detected in financial autoregressive time series. An irrational handling of the non-stationarity of an asset price or a GDP time series would render the statistical analysis misleading and unreliable.

The type of non-stationarity in a time series can be detected via the autoregressive unit root tests, which are thoroughly described below. That is, I(1) time series require first differencing and trend-stationary processes need a time-trend regression in order to remove the polynomial time of the long-run component and achieve stationarity.

The subject of the current chapter is to outline the theory and formulas behind the *standard* autoregressive unit root tests, which were first introduced by Dickey-Fuller and Phillips-Perron and perform these tests to the non-stationary autoregressive models mentioned in the first chapter, that is, models with drift, linear trend etc.

As a tailpiece to this chapter, some *issues* concerning the standard unit root testing are discussed, such as size distortion and low power of the tests.

3.2 Basic concepts for unit root tests

A *trend-stationary* process can be written as

$$X_t = DT_t + \epsilon_t, \tag{3.1}$$

where $DT_t = \alpha + \beta t$ is a time trend and ϵ_t is a stationary ARMA process. A first-difference process, on the other hand, can be expressed as,

$$X_t = \alpha + X_{t-1} + \epsilon_t, \tag{3.2}$$

where ϵ_t is a stationary ARMA process. The above equations can be condensed into the following expression,

$$\begin{aligned} X_t &= \alpha + \beta t + z_t, \\ z_t &= \phi z_{t-1} + \epsilon_t, \end{aligned}$$

where ϵ_t denotes a stationary process. With $|\phi| < 1$, z_t is stationary which implies that X_t is a trend stationary process.

A time series is said to have a *unit root* if $\phi = 1$. The non-stationarity can be removed by taking first differences on the above set of equations. Then, the process is said to be $I(1)$. Tests using $I(1)$ as the null hypotheses are the autoregressive unit root tests and comprise the main theme of the current chapter.

Conversely, if y_t is a trend stationary process,

$$y_t = \alpha + \beta t + z_t \tag{3.3}$$

and first difference of y_t is applied, the process becomes overdifferenced, as follows,

$$\Delta y_t = \alpha + \beta t - (\alpha + \beta(t-1)) + z_t - z_{t-1} \tag{3.4}$$
$$= \beta + z_t - z_{t-1}. \tag{3.5}$$

It emerges that the differenced process has a *moving average unit root*. Tests using $I(0)$ as the null hypothesis are hereunto not an object of interest.

The hypothesis test for the unit root testing is

$$H_0 : \phi = 1,$$
$$H_1 : |\phi| < 1.$$

The LSE (Least Squares Estimator) estimator of ϕ is

$$\hat{\phi} = \frac{\sum_{t=1}^{T} X_{t-1} X_t}{\sum_{t=1}^{T} X_{t-1}^2} \tag{3.6}$$

and can be equally expressed as,

$$\hat{\phi} = \phi + \frac{\sum_{t=1}^{T} X_{t-1} \epsilon_t}{\sum_{t=1}^{T} X_{t-1}^2}. \tag{3.7}$$

The variance of the LSE $\hat{\phi}$ is

$$Var(\hat{\phi}) = \frac{1}{T}\left(1 - \phi^2\right). \tag{3.8}$$

The proof follows from the fact that $E(X_t) = 0$, $Var(X_t) = \frac{\sum_{t=1}^{T} X^2 t}{n}$ and X_{t-1} is independent of ϵ_t. Hence,

$$V(\hat{\phi}) = V\left(\phi + \frac{\sum_{t=1}^{T} X_{t-1}\epsilon_t}{\sum_{t=1}^{T} X_{t-1}^2}\right) = V\left(\frac{\sum_{t=1}^{T} X_{t-1}\epsilon_t / T}{\sum_{t=1}^{T} X_{t-1}^2 / T}\right)$$
$$= \left(\frac{1-\phi^2}{\sigma_\epsilon}\right)^2 \frac{1}{T} V(X_{t-1}\epsilon_t) = \left(\frac{1-\phi^2}{\sigma_\epsilon}\right)^2 \frac{1}{T} E(X_{t-1}^2 \epsilon_t^2) =$$
$$= \left(\frac{1-\phi^2}{\sigma_\epsilon^2}\right)^2 \frac{1}{T} \frac{(\sigma_\epsilon^2)^2}{1-\phi^2} = \frac{1}{T}(1-\phi^2).$$

If $|\phi| < 1$, it can be shown (Hamilton, 1994, p. 216), that the distribution of the estimator $\hat{\phi}$ is,

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{\mathrm{d}} N(0, 1-\phi^2), \tag{3.9}$$

which can be expressed as,

$$\hat{\phi} \sim N(\phi, \frac{1}{T}(1-\phi^2)). \tag{3.10}$$

Under the hypothesis of $\phi = 1$, the distribution of $\hat{\phi}$ becomes

$$\hat{\phi} \sim N(1, 0), \tag{3.11}$$

which does not really make any sense. The test statistic under the non-stationarity hypothesis, $\phi = 1$, is based on the *Wiener process*, as elaborated below.

3.3 Wiener processes

3.3.1 From random walk to Wiener process

The notion of the *Wiener process* arises from the *limiting properties and behaviour* of the random walk process. It is used in physics, denoted as *Brownian motion*, to describe the motion of a particle that is subject to a large number of molecular shocks (Patterson, 2011).

In concrete terms, the random walk attains limiting properties, as the length of it is divided into arbitrarily small time steps, the width of which tends to zero.

Let X_t be a random walk process of length T,

$$X_t = X_{t-1} + \epsilon_t, \quad t = 0, 1, .., T, \tag{3.12}$$

where ϵ_t is white noise process. In the case when T is fixed, the equation 3.12 resembles a *discrete* time process with independent increments. By dividing the fixed time space T into smaller and smaller parts, the random walk is converted into a *continuous* time process, denoted *Wiener process*.

Let X_t be a random walk of fixed length T, as in 3.12. The *limiting process* is obtained by dividing the length T into M small time steps and letting M increase. The width of each step is,

$$\Delta t = \frac{T}{M}, \quad t = 0, 1, .., T. \tag{3.13}$$

Without loss of generality, T is set equal to one. Hence,

$$\Delta t = \frac{1}{M}, \quad t = 0, 1, .., T. \tag{3.14}$$

Thus, the width of each time step, Δt, is 'shrunk' to zero, as M increases. The random walk is then defined at I=[0,1], as follows,

$$I = [t_0 = 0, t_1, .., t_{M-1}, t_M = 1], \quad where, \quad t_i = \frac{i}{M}.$$

The time steps are, thus, described by the equation

$$t_i = t_{i-1} + \Delta t. \tag{3.15}$$

With regards to the *size* of each step, it is taken to be

$$\Delta X_t = (\sqrt{\Delta t})\epsilon_t, \tag{3.16}$$

with $Var(\Delta X_t) = \Delta t \sigma_\epsilon^2$. Under the hypotheses that $\Delta t = 1$ and that ϵ_t's are identically and normally distributed with $E(\epsilon_t) = 0$ and $\sigma_\epsilon^2 = Var(\epsilon_t) = 1$, i.e. $\epsilon_t \sim iid(0, \sigma_\epsilon^2)$, the difference between every step, ΔX_t, becomes ϵ_t, which denotes the size step of the *standard random walk*.

Taking the above scaling from zero to unity into account, the random walk takes the form,

$$X_{t_i} = X_{t_{i-1}} + (\sqrt{\Delta t})\epsilon_t \tag{3.17}$$

and with backward substitution,

$$X_{t_i} = X_{t_0} + \sum_{j=1}^{i} (\sqrt{\Delta t})\epsilon_t. \tag{3.18}$$

The unconditional variance of X_{t_j} is,

$$\begin{aligned} V(X_{t_i}) &= V(X_{t_0} + \sum_{j=1}^{i} (\sqrt{\Delta t})\epsilon_t) = \sum_{j=1}^{i} V\left[(\sqrt{\Delta t})\epsilon_t\right] \\ &= i\Delta t\sigma_\epsilon^2 = \frac{j}{N}\sigma_\epsilon^2 = t_i\sigma_\epsilon^2. \end{aligned}$$

The limiting result of the process is obtained by taking a *scaled version* of X_{t_i},

$$Y_{t_i} = \frac{X_{t_i}}{\sigma_\epsilon \sqrt{M}}. \tag{3.19}$$

By the Central Limit Theorem (CLT), it can be proved that, as $M \to \infty$ and T is fixed, Y_{t_i} converges asymptotically to a normal distribution, as follows,

$$X_{t_i} \sim N(0, t_i)$$

and hence,

$$Y_{t_i} \sim \sqrt{t_i} N(0, 1). \tag{3.20}$$

Thus, by dividing 3.20 by $\sqrt{t_i}$ it emerges that,

$$Z_{t_i} = \frac{X_{t_i}}{\sigma_\epsilon \sqrt{M} \sqrt{t_i}}$$

has a *limiting standard normal distribution*,

$$Z_{t_i} \sim N(0, 1).$$

3.3.2 Definition of the Wiener process

Definition 3.3.1. (*Wiener Process*) The *Wiener process*, denoted *W(t)*, is a continuous-time stochastic process defined as

$$\Delta W(t) = \epsilon\sqrt{\Delta t},$$

where $\Delta W(t)$ is the change of $W(t)$ over the interval Δt and $\epsilon \sim N(0, \sigma_\epsilon^2)$. The process satisfies the following conditions:

1. $W(0) = 0$
2. $W(t)$ has stationary and independent increments over time,
3. $W(t) \sim N(0, t\sigma_\epsilon^2)$

If $\epsilon_t \sim N(0,1)$, the process is named a *standard or Gaussian Wiener process*, which implies that $W(t) \sim N(0,t)$, $0 \leq t \leq 1$. Consequently, the Wiener process reflects a continuous random walk defined on $[0,1]$.

The second condition in definition 3.3.1 implies, that for any two different time intervals the value of ΔW is independent and normally distributed. In fact, for given times t and k, with $0 < k \leq t$,

$$W(t) - W(k) = \epsilon\sqrt{(t-k)} \tag{3.21}$$

that is,

$$W(t) - W(k) \sim N(0, t-k) \equiv \sqrt{(t-k)}N(0,1). \tag{3.22}$$

Hence, for a small interval Δt, 3.22 becomes,

$$W(t+\Delta t) - W(t) \sim \sqrt{\Delta t}N(0,1). \tag{3.23}$$

The connection between relationships 3.18 and 3.23 could now be more obvious. By comparing both equations, the first one is defined in discrete time and the assumption of the ϵ_t's is not necessary for the asymptotic normality of X_{t_i}, whereas the second is specified in continuous time and the Wiener process, $W(t)$, *is* normally distributed.

Notwithstanding, these differences are asymptotically eliminated, as $\Delta t \to 0$, which indicates that as the time steps approach infinity the random walk simulates a Wiener process. This intuitive result can be proved by using the *functional central limit theorem* (FCLT).

3.3.3 The Functional Central Limit Theorem (FCLT)

The *functional central limit theorem*, also known as *Donsker's theorem*, is the analogue of the *central limit theorem* and is applied to stochastic processes rather than random variables.

Let X_t be a standard random walk process with $X_0 = 0$ and $\epsilon_t \sim N(0,1)$. The X_t process can be written as $X_t = \sum_{t=1}^{T} \epsilon_t$ for $t = 0, 1, .., T$.

The FCLT is mapping the interval from 0 to T to a fixed interval [0,1] (Maddala,1998). The division of [0,1] into $T+1$ parts is exactly as described in the previous subsection. Hence, if $0 \leq k \leq 1$,

$$k = \frac{i}{T}, \quad i = 0, 1, .., T. \tag{3.24}$$

The X_t process is now converted to a function of k, which is denoted as *step function*

$$Y_T(k) \equiv \frac{X_T(k)}{\sqrt{T}}, \tag{3.25}$$

where

$$X_T(k) = \sum_{t=1}^{[kT]} \epsilon_t. \tag{3.26}$$

The notation $[kT]$ refers to the integer part of kT. For instance, if $T = 100$ and $k = 0.543$, then $[kT] = 54$.

In the limit, *as* $T \to \infty$, *the step function* $Y_T(k)$ *becomes increasingly dense on* $[0, 1]$ *and converges weakly to the standard Wiener process* (Maddala,1998). Hence, the FCLT implies that

$$Y_T(k) \Rightarrow W(k). \tag{3.27}$$

Figure 3.1 illustrates an approximation of the random walk process by the step function $Y_T(k)$ for $T = 10$, $T = 50$ and $T = 250$.

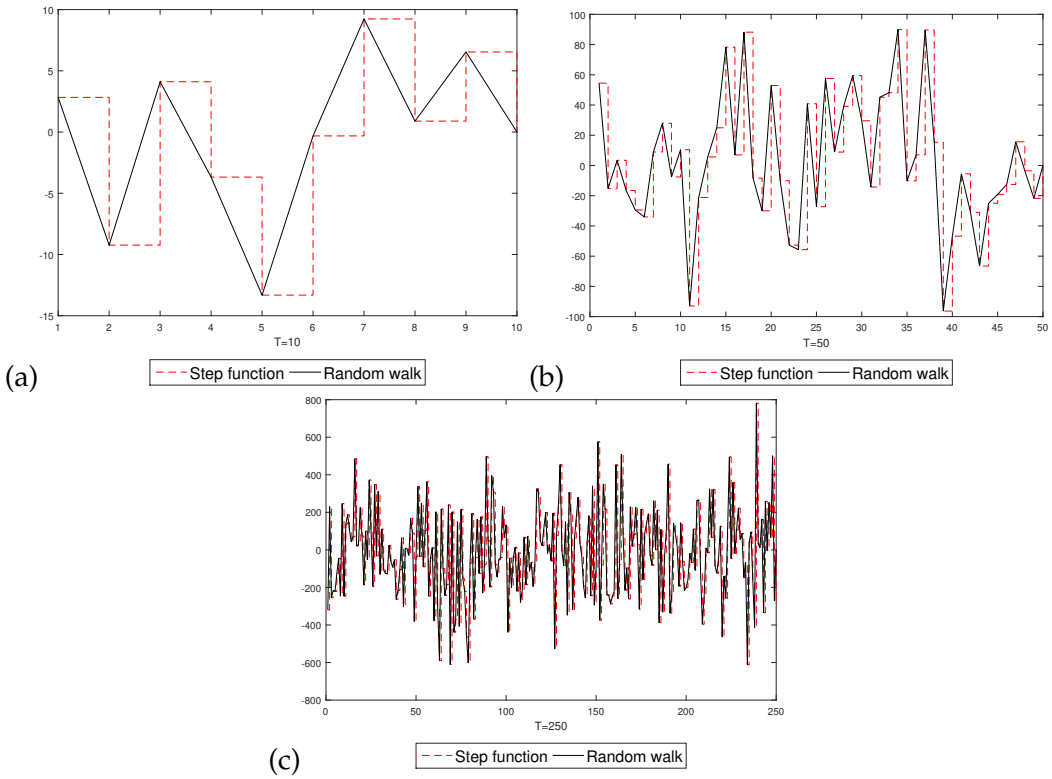


FIGURE 3.1: Simulation for (a) T=10 (b) T=50 and (c) T=250

3.3.4 The Continuous Mapping Theorem (CMT)

An essential theorem for test statistics in unit roots is the *Continuous mapping theorem*, which is usually used in combination with the FCLT. The CMT states that

$$\text{if } X_t \Rightarrow X \text{ and } P(X \in A) = 0, \text{ then } f(X_t) \Rightarrow f(X),$$

where $f(\cdot)$ is a continuous function and A denotes the set of discontinuity points of f. The CMT could be used as a consequence of FCLT, for a continuous function $f(\cdot)$ defined on $[0, 1]$.Then, the relationship 3.27 becomes

$$f(Y_T(k)) \Rightarrow f(W(k)). \tag{3.28}$$

3.3.5 Basic results of the Wiener process

Let X_t be a random walk process, $X_t = X_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0,1)$ and $X_t = \sum_{t-1}^{T} \epsilon_t$. Thus, $X_t \sim N(0,T)$ and

$$Y_T(1) \equiv \frac{X_T}{\sqrt{T}} \sim N(0,1). \tag{3.29}$$

From the FCLT it arises that

$$Y_T(1) \Rightarrow W(1). \tag{3.30}$$

Some very useful results for the unit root test statistic distribution derive from the following two lemmas. Both lemmas are expressed the assumption of independent ϵ_t's, $\epsilon_t \sim N(0,1)$ and $X_0 = 0$.

The first one, which is cited without proof, portrays the transition from the step function to the Wiener process, under the FCLT and the CMT.

Lemma 2. (i) $T^{-3/2} \sum_{t=0}^{T} X_t \Rightarrow \int_0^1 W(k)dk$

(ii) $T^{-2} \sum_{t=1}^{T} X_t^2 \Rightarrow \int_0^1 [W(k)]^2 dk$

(iii) $T^{-\frac{5}{2}} \sum_{t=1}^{T} tX_t \Rightarrow \int_0^1 rW(k)dk$

(iv) $T^{-\frac{3}{2}} \sum_{t=1}^{T} t\epsilon_t \Rightarrow \int_0^1 rdW(k)$

(v) $T^{-1} \sum_{t=2}^{T} X_{t-1}\epsilon_t \Rightarrow \int_0^1 W(k)dW(k)$

The following lemma demonstrates the connection between the Wiener process and the normal distribution.

Lemma 3. (i) $\frac{X_T}{\sqrt{T}} \equiv W(1) \sim N(0,1)$

(ii) $\int_0^1 W(k)dk \sim N(0, \frac{1}{3})$

(iii) $\int_0^1 kdW(k) \sim N(0, \frac{1}{3})$

(iv) $\int_0^1 W(k)dW(k) = \frac{1}{2}\left[W(1)^2 - 1\right] \sim \frac{1}{2}\left[\mathcal{X}^2(1) - 1\right]$

Proof. (i) The FCLT states that the step function converges weakly to the Wiener process; that is,

$$\frac{X_T(k)}{\sqrt{T}} \Rightarrow W(k). \tag{3.31}$$

Evaluated at $k = 1$, $\frac{X_T(1)}{\sqrt{T}}$ is the *sample mean*, $\sum_{t=1}^{T} \epsilon_t$, which divided by $\sqrt{T}$ follows a normal distribution,

$$\frac{X_T(1)}{\sqrt{T}} = \frac{X_T}{\sqrt{T}} \sim N(0,1).$$

(ii) The sum of X_t, $\sum_{t=1}^{T} X_t$ can be expressed as,

$$\sum_{t=1}^{T} X_t = T\epsilon_1 + (T-1)\epsilon_2 + .. + \epsilon_T, \tag{3.32}$$

since $X_t = \epsilon_1 + \epsilon_2 + ... + \epsilon_T$. Since ϵ_t are identically and normally distributed, $\epsilon_t \sim N(0,1)$, it emerges that

$$Var\left(\sum_{t=0}^{T} X_t\right) = T^2 + (T-1)^2 + ... + 1^2 = \sum_{t=1}^{T} t^2.$$

Hence,

$$\sum_{t=1}^{T} X_t \sim N\left(0, \sum_{t=1}^{T} t^2\right) \Rightarrow$$
$$T^{-\frac{3}{2}} \sum_{t=1}^{T} X_t \sim N\left(0, T^{-3} \sum_{t=1}^{T} t^2\right).$$

By using the sum $\sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6}$, it can be easily proved that $T^{-3} \sum_{t=1}^{T} t^2 \approx \frac{1}{3}$. Thus,

$$T^{-\frac{3}{2}} \sum_{t=1}^{T} X_t \sim N\left(0, \frac{1}{3}\right).$$

Combining the above relationship with lemma 2, it emerges that

$$\int_0^1 W(k)dk \sim N(0, \frac{1}{3}).$$

(iii) The proof is based on (i). The variance of the sum $\sum_{t=1}^{T} t\epsilon_t$ is

$$Var\left(\sum_{t=1}^{T} t\epsilon_t\right) = \sum_{t=1}^{T} t^2 Var(\epsilon_t) = \sum_{t=1}^{T} t^2.$$

Hence,

$$T^{-\frac{3}{2}} \sum_{t=1}^{T} t\epsilon_t \sim N\left(0, T^{-3} \sum_{t=1}^{T} t^2\right).$$

The approximation of the variance is, as described above, equal to $\frac{1}{3}$. Provided that $T^{-\frac{3}{2}} \sum_{t=1}^{T} t\epsilon_t \Rightarrow \int_0^1 rdW(k)$, it emerges that

$$\int_0^1 kdW(k) \sim N\left(0, \frac{1}{3}\right).$$

(iv) The proof is based on the relationship (v) of lemma 2. The product $X_{t-1}\epsilon_t$ is obtained by

$$X_t^2 = (X_{t-1} + \epsilon_t)^2 = X_{t-1}^2 + 2X_{t-1}\epsilon_t + \epsilon_t^2 \Rightarrow$$
$$X_{t-1}\epsilon_t = \frac{1}{2}\left(X_t^2 - X_{t-1}^2 - \epsilon_t^2\right).$$

Hence,

$$T^{-1} \sum_{t=0}^{T} X_{t-1}\epsilon_t = \frac{T^{-1}}{2} \sum_{t=1}^{T} \left(X_t^2 - X_{t-1}^2\right) - \frac{T^{-1}}{2} \sum_{t=1}^{T} \epsilon_t^2. \tag{3.33}$$

The term $\sum_{t=1}^{T}\left(X_t^2 - X_{t-1}^2\right)$ is equal to $\frac{X_T^2}{T}$. It is already known that $\frac{X_T}{\sqrt{T}} \Rightarrow W(1) \equiv N(0,1)$. Consequently, by use of the CMT, it arises that,

$$\frac{X_T^2}{T} \Rightarrow [N(0,1)]^2 \equiv \mathcal{X}_1^2.$$

Moreover,

$$\frac{\sum_{t=1}\epsilon_t^2}{T} \Rightarrow Var(\epsilon_t) = 1.$$

Thus,

$$T^{-1}\sum_{t=0}^{T} X_{t-1}\epsilon_t \Rightarrow \frac{1}{2}\left[\mathcal{X}_1^2 - 1\right].$$

Taking into consideration the above result and the relationship (v) of lemma 2 it is proved, that,

$$\int_0^1 W(k)dW(k) = \frac{1}{2}\left[W(1)^2 - 1\right] \sim \frac{1}{2}\left[\mathcal{X}^2(1) - 1\right]. \tag{3.34}$$

□

3.4 Unit root tests based on the Dickey-Fuller distribution

The most common autoregressive unit root tests are considered to be the *Dickey-Fuller unit root test*, denoted as DF and the *Augmented Dickey-Fuller unit root test*, denoted as ADF. The latter is a semi-parametric test which is mostly used for higher order AR or ARMA models.

For the simple DF test, Dickey (1976) and Dickey and Fuller (1979) derived the the test statistic distribution of the unit root test in three cases:

- AR(1) model without drift or deterministic trend
- AR(1) model with drift
- AR(1) model with a linear trend

The distribution tables and critical values are not the same for each case and have been evaluated by Dickey and Fuller.

3.4.1 Random walk model without drift and trend

In the case of estimating an AR(1) process without trend or drift

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2), \tag{3.35}$$

the null hypothesis is

$$X_t = X_{t-1} + \epsilon_t \tag{3.36}$$

Defined on an interval of length T, the LSE of the parameter ϕ of an AR(1) model is,

$$\hat{\phi} = \frac{\sum_{t=1}^{T} X_{t-1} X_t}{\sum_{t=1}^{T} X_{t-1}^2}. \tag{3.37}$$

Under the assumption of $|\phi| < 1$ and independent and identically distributed errors, the standardised LSE follows a standard normal distribution, as initially proved by Mann and Wild (1943),

$$\frac{\sqrt{T}\left(\hat{\phi} - \phi\right)}{\sqrt{1-\phi^2}} \Rightarrow N(0,1). \tag{3.38}$$

Under the assumption of $|\phi| = 1$, on the contrary, the limiting distribution of $\hat{\phi}$ could not be easily evaluated. Phillips (1987) derived the standardised distribution of the quantity,

$$T\left(\hat{\phi} - \phi\right), \tag{3.39}$$

which, led Fuller (1976) and Dickey and Fuller (1979) to derive the limiting distribution of 3.39, under the assumptions of $X_0 = 0$ and independent and identical distributed errors.

The Dickey-Fuller distribution

Under the assumption of the null hypothesis, $H_0 : \phi = 1$, the test statistic is,

$$T\left(\hat{\phi} - 1\right) = T\left(\frac{\sum_{t=1}^{T} X_{t-1} X_t}{\sum_{t=1}^{T} X_{t-1}^2} - 1\right) = \frac{T\sum_{t=1}^{T} X_{t-1}\epsilon_t}{\sum_{t=1}^{T} X_{t-1}^2}. \tag{3.40}$$

The above relationship could be expressed as,

$$T\left(\hat{\phi} - 1\right) = \frac{T^{-1}\sum_{t=1}^{T} X_{t-1}\epsilon_t}{T^{-2}\sum_{t=1}^{T} X_{t-1}^2}. \tag{3.41}$$

By using CMT and relationships (ii) and (v) of lemma 2, it emerges that the quantity $T\left(\hat{\phi} - 1\right)$ has a distribution related to the Wiener process, which is the Dickey-Fuller distribution.

$$T\left(\hat{\phi} - 1\right) \Rightarrow \frac{\int_0^1 W(k)dW(k)}{\int_0^1 \left[W(k)\right]^2 dk} \tag{3.42}$$

The above relationship combined with (iv) of lemma 3 produces that,

$$T\left(\hat{\phi} - 1\right) \Rightarrow \frac{\left[W^2(1) - 1\right]}{2\int_0^1 \left[W(k)\right]^2 dk}. \tag{3.43}$$

It can be noticed that the convergence of $\left(\hat{\phi} - 1\right)$ is of order T, whereas, under the assumption of $|\phi| < 1$ it is of order $T^{1/2}$.

The tables for the DF distribution are evaluated by Dickey and Fuller using Monte Carlo simulation. Obviously, the critical values of the distribution differ from those of a simple $t-$test or a $F-$test.

There are two different versions to the DF unit root test. The test statistic for the first one is,

$$T\left(\hat{\phi} - 1\right) \tag{3.44}$$

and the critical values origin from the tables of the distribution in 3.43. The other

approach involves the standard deviation of $\hat{\phi}$ and is denoted as DF $t-$test, since the quantity $T\left(\hat{\phi}-1\right)$ is divided by $\hat{\sigma}_{\hat{\phi}}$ and the test becomes a $t-$type test; that is,

$$t_{\hat{\phi}} = \frac{\hat{\phi}-1}{\hat{\sigma}_{\hat{\phi}}}, \tag{3.45}$$

where $\hat{\sigma}_{\hat{\phi}}$ denotes the standard error of $\hat{\phi}$. The distribution $t_{\hat{\phi}}$ has tables and critical values which are totally different from those of *Student* distribution.

As shown above, $\hat{\sigma}^2_{\hat{\phi}} = \frac{1}{T}(1-\phi^2)$. Hence, the test statistic t_T in 3.45 becomes

$$t_{\hat{\phi}} = \frac{\sqrt{T}(\hat{\phi}-1)}{\sqrt{1-\hat{\phi}^2}} = \frac{T(\hat{\phi}-1)}{\sqrt{T(1-\hat{\phi}^2)}}. \tag{3.46}$$

Given the above two lemmas and relationship 3.43, the distribution of t_T is a DF distribution as follows,

$$\frac{T(\hat{\phi}-1)}{\sqrt{T(1-\hat{\phi}^2)}} \Rightarrow \frac{\int_0^1 W(k)dW(k) \,/\, \int_0^1 [W(k)]^2\,dk}{1 \,/\, \left(\int_0^1 [W(k)]^2\,dk\right)^{1/2}} = \frac{\int_0^1 W(k)dW(k)}{\left(\int_0^1 [W(k)]^2\,dk\right)^{1/2}}. \tag{3.47}$$

Hence, $t_{\hat{\phi}}$ converges weakly to the DF $t-$distribution,

$$\frac{T(\hat{\phi}-1)}{\sqrt{T(1-\hat{\phi}^2)}} \Rightarrow \frac{W^2(1)-1}{2\left(\int_0^1 [W(k)]^2\,dk\right)^{1/2}}. \tag{3.48}$$

DF distribution and rejection of H_0

In both approaches the critical values, under the hypothesis of $\phi = 1$, are not those of the Student-t distribution. The DF distribution is skewed to the left, since the probability that $\mathcal{X}^2(1)$ is less than one in the numerator, is 0.68, which implies that the quantity $T(\hat{\phi}-1)$ is negative with probability 0.68. Hence, the DF distribution is *asymmetric* and the use of the critical values of the symmetric *Student-t* distribution could lead to overrejection of the null hypothesis.

The null hypothesis H_0 in a DF unit root test is *rejected*, if the value of the test statistic is less than the critical value in the corresponding table, since the test is one-sided. Typically, the null hypothesis is rejected at 5% significance level if the value of $t_{\hat{\phi}}$ is less than -1.992.

Dickey and Fuller have provided different tables with critical values for *each* random walk cases presented below (with constant, with trend).

3.4.2 Models with drift

In the second case the estimated AR(1) process includes drift

$$X_t = \beta + \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma^2_\epsilon). \tag{3.49}$$

There are two cases for the null-hypothesis: the random walk without drift and the random walk with drift.

Case 1: Random walk without drift

The null hypothesis is

$$X_t = X_{t-1} + \epsilon_t. \tag{3.50}$$

The estimated model under which, the limiting distribution for the null and the alternative hypotheses are derived, includes a drift. Hence, the null hypothesis for this case is

$$H_0\text{: } \beta = 0 \textit{ and } \phi = 1.$$

If the mean of the process, $\mu = E(X_t) = \frac{\beta}{1-\phi}$, is subtracted from both sides of 3.49 the process becomes,

$$X_t - \mu = \phi(X_{t-1} - \mu) + \epsilon_t. \tag{3.51}$$

The 'demeaned' process is denoted as $X_t^\star$ and the previous equation can be expressed as

$$X_t^\star = \phi X_{t-1}^\star + \epsilon_t. \tag{3.52}$$

Based on the theory presented in 3.4.1 it emerges that

$$T\left(\hat{\phi} - 1\right) = \frac{T^{-1}\sum_{t=1}^{T} X_{t-1}^\star \epsilon_t}{T^{-2}\sum_{t=1}^{T}\left(X_{t-1}^\star\right)^2}. \tag{3.53}$$

The limiting distribution of $T(\hat{\phi} - 1)$ is then

$$T\left(\hat{\phi} - 1\right) \Rightarrow \frac{\int_0^1 W^\star(k)dW(k)}{\int_0^1 \left(W_{t-1}^\star(k)\right)^2 dk}. \tag{3.54}$$

where $W^\star(k) = W(k) - \int_0^1 W(k)dk$ is the 'demeaned' Wiener process. Correspondingly, the limiting distribution of the $t-$test becomes

$$t_{\hat{\phi}} \Rightarrow \frac{\int_0^1 W^\star(k)dW(k)}{\left[\int_0^1 \left(W_{t-1}^\star(k)\right)^2 dk\right]^{1/2}}. \tag{3.55}$$

The test statistic for the drift, $t_{\hat{\beta}}$, is proved to be also a function of the Wiener process. The critical values for this distribution differ from those of the standard DF distribution and are derived by Dickey and Fuller.

Case 2: Random walk with drift

If the null hypothesis is assumed to be a random walk with drift

$$X_t = \beta + X_{t-1} + \epsilon_t, \tag{3.56}$$

then,

$$H_0\text{: } \beta \neq 0 \text{ and } \phi = 1.$$

By sequential backward substitution, the process $X_t = \beta + X_{t-1} + \epsilon_t$ can be expressed as

$$X_t = \beta t + \sum_{i=0}^{t} \epsilon_i$$

with the initial value X_0 set to zero. In the limit, the deterministic trend βt tends to dominate the process and renders the distribution of the LSE to be asymptotically normal. The exact proof for this intuitive result is based on the asymptotic distributions of LSE of β and ϕ.

By using the asymptotic results of lemmas 2 and 3, it emerges that

$$\begin{bmatrix} T^{1/2}\left(\hat{\beta}-\beta\right) \\ T^{3/2}\left(\hat{\phi}-1\right) \end{bmatrix} \Rightarrow N\left(0, \sigma_\epsilon^2 Q^{-1}\right),$$

where

$$Q = \begin{bmatrix} 1 & \beta/2 \\ \beta/2 & \beta^2/3 \end{bmatrix}.$$

Hence, under the null hypothesis $H_0 : \phi = 1$, the DF test statistic has the following distribution:

$$T^{3/2}\left(\hat{\phi}-1\right) \Rightarrow N\left(0, \frac{12\sigma_\epsilon^2}{\beta^2}\right) \tag{3.57}$$

which is rational result only if $\beta \neq 0$.

3.4.3 Models with linear trend

In the general case of estimating an AR(1) model with linear trend,

$$X_t = \beta + \gamma t + \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma_\epsilon^2) \tag{3.58}$$

there are two alternative models under the null-hypothesis of non-stationarity: random walk only with drift and random walk with linear trend.

Case 1: Random walk with drift only

In the first case the null-hypothesis is described by the process

$$X_t = \beta + X_{t-1} + \epsilon_t \tag{3.59}$$

The asymptotic distribution of $\hat{\phi}$ is invariant to the value of β which renders the null hypothesis of the unit root test as follows:

$$H_0: \ \beta \neq 0, \gamma = 0 \ and \ \phi = 1.$$

The asymptotic distribution of $\hat{\phi}$ and the corresponding $t-$test are proved to be (Maddala, 1998):

$$T\left(\hat{\phi}-1\right) \Rightarrow \frac{\int_0^1 W(k)dW(k) + C}{A} \tag{3.60}$$

and

$$t_{\hat{\phi}} \Rightarrow \frac{\int_0^1 W(k)dW(k) + C}{A^{1/2}}, \tag{3.61}$$

where

$$A = \int_0^1 (W(k))^2 dk - 12 \left(\int_0^1 kW(k)dk \right)^2 + 12 \int_0^1 W(k)dk \int_0^1 kW(k)dk - 4 \left(\int_0^1 W(k)dk \right)^2$$

and

$$C = 12 \left[\int_0^1 kW(k)dk - \frac{1}{2} \int_0^1 W(k)dk \right] \times \left[\int_0^1 W(k)dk - \frac{1}{2} W(1) \right] - W(1) \int_0^1 W(k)dk.$$

The asymptotic distributions of the LSE $\hat{\beta}$ and $\hat{\gamma}$ are also functional forms of the Wiener process.

Case 2: Random walk with linear trend

For the case of the random walk with trend,

$$X_t = \beta + \gamma t + X_{t-1} + \epsilon_t, \tag{3.62}$$

the null hypothesis of the unit root testing is

$$H_0: \quad \beta \neq 0, \gamma \neq 0 \text{ and } \phi = 1$$

It can be shown (Maddala, 1998) that the LSE $\hat{\phi}$ follows an asymptotic normal distribution.

3.5 Other unit root tests

The results of the aforementioned unit root tests are based on the independent and identically distributed errors with mean zero and constant variance, $\epsilon_t \sim iid(0, \sigma^2)$. The strong assumption of no auto-correlation in errors, though, hardly responds to real data.

In addition, there exists inherent difficulty in performing unit root tests to higher order AR models or more complex ARMA models. The fact that the standard Dickey - Fuller test can only be applied to a simple random walk model with iid errors gave birth to a number of different unit root tests. Some common tests are, the *augmented Dickey - Fuller test*, the *Phillips - Perron test*, the *Sargan-Bhargava and Bhargava test* and others. Only the first two tests are discussed below.

3.5.1 The Augmented Dickey Fuller test

In many time series the errors are not identically and independently distributed and thus, are not comprising a white noise process. Since the standard Dickey - Fuller tests are based on *iid* errors with $\sigma = \sigma_\epsilon^2$, they could not capture the autocorrelation and other complexities.

A solution to such difficulties is proposed by Dickey and Fuller (1979) and Said and Dickey (1984), who augmented the basic unit root tests in order to respond to a wide class of series, such as AR(p) or ARMA(p,q) with unknown orders, and take the issue of auto-correlation into consideration. The tests derived are named *Augmented Dickey Fuller tests*, or abbreviated, ADF.

The AR(p) process

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t, \tag{3.63}$$

under the assumption of $\phi_0 = 0$ and $\epsilon_t \sim iid(0, \sigma_\epsilon^2)$, can be expressed as

$$A(L)X_t = \epsilon_t, \tag{3.64}$$

where $A(L) = 1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p$ is the lag-polynomial or order p. If X_t has a unit root, then $A(1) = 0$ and

$$1 - \sum_{i=0}^{p} \phi_i L^i = 0 \Rightarrow \sum_{i=0}^{p} \phi_i = 1$$

Hence, the equation 3.64 can be written as

$$(1 - L)A^\star(L)X_t = \epsilon_t$$

where $A^\star(L) = \sum_{i=1}^{p-1} \alpha_i L^i$. The model can be, thus, rewrited as

$$\Delta X_t = B(L)^{-1}\epsilon_t, \tag{3.65}$$

where $B(L) = (1 - L)A^\star(L)$. By expressing the regression in this way, the autocorrelation among the errors could be now evident.

The ADF test is based on finding the LSE of the model

$$\Delta X_t = \beta + \phi X_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta X_{t-i} + \epsilon_t. \tag{3.66}$$

sketch of the proof could be given by the following illustrative example.

Let $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \epsilon_t$ be an AR(3) model with $\phi_0 = 0$. By taking first differences on the model it emerges that

$$\Delta X_t = (\phi_1 - 1)X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \epsilon_t \tag{3.67}$$

Then, by adding and subtracting firstly the term $\phi_3 X_{t-2}$ and consequently the term $(\phi_2 + \phi_3)_{t-1}$ the equation becomes

$$\begin{aligned}\Delta X_t &= (\phi_1 - 1)X_{t-1} + (\phi_2 + \phi_3)X_{t-2} - \phi_3(X_{t-2} - X_{t-3}) + \epsilon_t \\ &= (\phi_1 + \phi_2 + \phi_3 - 1)X_{t-1} - (\phi_2 + \phi_3)(X_{t-1} - X_{t-2}) \\ &\quad - \phi_3(X_{t-2} - X_{t-3}) + \epsilon_t.\end{aligned}$$

Hence, the equation takes the form

$$\Delta X_t = \phi X_{t-1} + \sum_{i=1}^{2} \alpha_i \Delta X_{t-1} + \epsilon_t \tag{3.68}$$

where $\phi = \phi_1 + \phi_2 + \phi_3$, $\alpha_1 = -(\phi_2 + \phi_3)$ and $\alpha_2 = -\phi_3$, Inductively, the regression 3.68 is generalised to the *augmented Dickey-Fuller regression*

$$\Delta X_t = \phi X_{t-1} + \sum_{i=1}^{p-1} \alpha_i \Delta X_{t-i} + \epsilon_t. \tag{3.69}$$

Under the null hypothesis $H_0 : \phi = 1$, the distribution of the LSE $\hat{\alpha_i}$, for $i = 1, .., p-1$, is the normal distribution and the distribution of $\hat{\phi}$ is the DF distribution as in 3.45.

If constant and trend are added to the model, the ADF regression becomes respectively,

$$\Delta X_t = \beta + \phi X_{t-1} + \sum_{i=1}^{p-1} \alpha_i \Delta X_{t-i} + \epsilon_t,$$

$$\Delta X_t = \beta + \gamma t + \phi X_{t-1} + \sum_{i=1}^{p-1} \alpha_i \Delta X_{t-i} + \epsilon_t.$$

The tables of the critical values for these cases are the same as in the cases with constant and trend mentioned above.

Lag selection

The size and the power of the ADF tests are affected by the number of lags, p, contained in the estimating regression. Specifically, if the number of lags is small, then the remaining auto-correlation in the errors will bias the test and the size of it will grow. On the other hand, a very large number of lags would diminish the power of the test. In general, the number of lags included in the regression usually increases with the size of the data.

Ng and Perron (1995) proposed a rule for selecting the number of lags, which would *preserve the size* stable and *maximize the power of the test*. That is, the number of lags is set to have an upper bound, denoted as p_{max}. The ADF regression is estimated with $p = p_{max}$ and if the absolute value of the test statistic for the lagged difference is greater than 1.6, then the ADF test is normally performed. Otherwise, the number of lags is reduced by one and the process is repeated.

Schwert (1989) proposed a common rule for the calculation of the upper bound. This is,

$$p_{max} = \left[12 \cdot \left(\frac{T}{100} \right)^{1/4} \right], \tag{3.70}$$

where $[\cdot]$ denotes the integer part.

The lag length could be also determined by the information criteria AIC and BIC. The corresponding functions for the model in 3.69 are

$$AIC = (T - p - 1)\log\hat{\sigma^2} + 2k,$$
$$BIC = (T - p - 1)\log\hat{\sigma^2} + k \cdot \log(T - p - 1),$$

where k denotes the number of parameters in the model. The value of p which gives the lowest value of AIC or BIC signifies the *proper* number of lags in the regression.

3.5.2 The Phillips-Perron test

Another test which accounts for serially correlated and heteroscedastic errors is the Phillips-Perron test (1987), denoted as PP test. The test statistic of the PP test is based on the DF distribution and *corrects* for possible heteroskedasticity or serial autocorrelation in the errors. In the case when there is no autocorrelation, the PP test is identical to the DF test.

An advantage of the PP test is its non-parametric character; that is, the level of serial correlation, i.e. the lags, does not need to be defined, as in the ADF test. Hence, the autocorrelation in the errors is *corrected* without the specific choice of lags, p, and the biased opinion of the researcher. However, this test is more reliable only for a large amount of data, since it is based on asymptotic results of the DF distribution. Another drawback of this test is the sensitivity to structural breaks.

The idea behind the PP test is that the non-augmented DF test is initially used and if any autocorrelation in the errors is detected, it is *adjusted* by proper test statistics.

The test statistic for the PP test includes the *variance of sum of errors*, denoted by σ^2 and the *variance of errors*, denoted by σ_ϵ^2. Since the errors are not independent,

$$Var(\epsilon_t) = \sigma_\epsilon^2, \tag{3.71}$$

and more specifically,

$$\sigma_\epsilon^2 = \lim_{T\to\infty} T^{-1} \sum_{i=1}^{T} E(\epsilon_j^2). \tag{3.72}$$

The variance of the AR(1) model without constant in the general case is

$$\lim_{T\to\infty} E\left(\frac{X_T^2}{T}\right) = \sigma^2, \tag{3.73}$$

which, by backward substitution, becomes

$$\sigma^2 = \lim_{T\to\infty} T^{-1} E\left[\left(\sum_{i=1}^{T} \epsilon_i\right)^2\right]. \tag{3.74}$$

Hence,

$$\frac{X_T}{\sqrt{T}} \Rightarrow N(0, \sigma^2). \tag{3.75}$$

As a result,

$$\frac{X_T^2}{T} \Rightarrow \succ^2(1). \tag{3.76}$$

The distribution of the numerator of the test statistic $T(\hat{\phi} - 1)$ is then

$$T^{-1} \sum_{t=1}^{T} X_{t-1}\epsilon_t \Rightarrow \frac{\sigma^2}{2}\left(\succ^2(1) - 1\right) + \left(\frac{\sigma^2 - \sigma_\epsilon^2}{2}\right). \tag{3.77}$$

The distribution of the denominator is

$$T^{-2} \sum_{t=1}^{T} X_{t-1}^2 \Rightarrow \sigma^2 \int_0^1 \left(W(k)\right)^2 dk.$$

Consequently, if the errors are not independently distributed, the DF test statistic $T(\hat{\phi}-1)$ takes the form

$$T(\hat{\phi}-1) \Rightarrow \frac{\frac{1}{2}\left[\succ_1^2 - 1\right] + \mu\sigma^{-2}}{\int_0^1 \left[W(k)\right]^2 dk},$$

where

$$\mu = \frac{\sigma^2 - \sigma_\epsilon^2}{2}.$$

In the case of uncorrelated errors, the above quantities σ^2 and σ_ϵ^2 are equal.

The modified test statistic for the PP test is

$$B_{\hat{\phi}} = T(\hat{\phi}-1) - \mu \left[T^{-2} \sum_{t=1}^{T} X_{t-1}^2 \right]^{-1},$$

which then implies that $B_{\hat{\phi}}$ follows the Dickey-Fuller distribution.

Phillips and Perron proposed the use of the consistent estimators s_ϵ^2 and σ^2 of the parameters σ_ϵ^2 and σ^2 respectively. Hence, they provided non-parametric test statistics for each case of the unit root testing: AR(1) model without drift or trend, AR(1) model with drift and AR(1) model with drift and trend.

The limiting distributions of the test statistics for each case identify with those of $T(\hat{\phi}-1)$ and $t_{\hat{\phi}}$ introduced by Dickey and Fuller, when $\sigma^2 = \sigma_\epsilon^2$.

3.6 Issues associated with the classical unit root tests

Size distortion and low power

In general, the ADF and PP tests suffer from *size distortions*, which means that the correct null hypothesis is, in most cases, rejected and hence, the stationary I(0) series is regarded as non-stationary I(1). Hence, the conventional asymptotic critical values for distorted tests cannot be employed. In which cases, however, the standard DF and PP tests suffer from size distortions?

It has been argued (Schwert, 1989) that the DF distribution with a moving average (MA) component differs from the standard DF distribution mentioned previously. Especially, when the errors are negatively correlated, the size of the test is highly affected.

If the *first-differenced* time series has an ARMA representation and includes a large and negative MA or AR component, the PP test leads to overrejecting the null hypothesis and thus, to size distortion.

The *power of the unit root test* is the probability of correct rejection of an invalid null hypothesis. The ADF and PP tests have generally low power, especially when the alternative is a highly persistent stationary process with ϕ close to unity or when the series contains a time trend. Typically, the power of the PP and ADF tests against trend-stationary alternatives fluctuates around 0.10 and 0.30 respectively. Hence, the ADF test is more likely to be used in practice.

Discontinuity near the unit root $\phi = 1$

The test statistics of the classical unit root tests are *continuous* for $|\phi| \leq 1$, in the case of finite samples. At the limit, however, the asymptotic distributions of the test statistics change *discontinuously* near the unit root.

According to Cohrane (1991), any test in the form of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_0 - \epsilon$ has relatively low power, especially in small samples. This drawback, combined with the discontinuity between the null and the alternative hypothesis, leads to arbitrarily poor power of the classical unit root tests.

Power problems and modifications to the standard unit root test

Inference based on the basic autoregressive unit root tests, ADF and PP, may be quite fragile, mainly due to the distorted size and the low power of these tests. As aforementioned, the presence of large negative MA components can cause size and power problems.

A way to handle this adversity has been developed by Yap and Reinsel (1995). They suggested the *Likelihood Ratio* (LR) tests which are performed on ARMA models and encompass MA components. The test statistics of LM unit root tests follow asymptotically the Dickey-Fuller distribution, which implies that the addition of MA terms to an AR process does not change the distribution that could be obtained by a simple AR model.

Perron and Ng (1996) proposed some modifications to the test statistic of the basic PP tests. The *modified PP tests* are proved to maintain their power and size, when there exist negative MA components in the model. In addition, the order of the MA component does not need to be a priori defined.

The *ADF-GLS* (or DF-GLS) test suggested by Elliott, Rothenberg and Stock (1996) is a modification of the standard ADF test which is more efficient due to its large power and diminished size distortions. In essence, the test is based on the standard *augmented* DF test; however, the time series is transformed by using a *generalized least squares* (GLS) regression before performing the test.

Stationarity as null hypothesis

The common unit root tests have as the null hypothesis the assumption of non-stationary, I(1) time series, against the stationary alternative. The hypothesis test is performed by rejecting or not the null hypothesis, providing evidence against it. By experience, the vast majority of economic series are not very informative, and thus, the standard unit root tests cannot reject the null hypothesis, even in the case of a random walk series. Hence, the basic unit root tests have *low power* against the relevant alternatives.

The above claim led many researchers to conduct hypothesis tests using *stationarity* as the null hypothesis and *unit root* as the alternative. Some widely known *stationarity tests* are: Tanaka (1990), KPSS (Kwiatkowski, Phillips, Schmidt and Shin, 1992), Park (1900) and others.

Chapter 4

Bayesian unit root testing

4.1 A preface to the Bayesian unit root testing

4.1.1 Introduction

As mentioned above, the presence of a unit root in a time series leads to persistent non-stationary models, which are in practice encountered in many economic time series. The detection of a unit root comprises a crucial issue in the statistics literature.

The previous chapter exhibited some of the most widely used autoregressive unit root tests which are based on the distribution of the LSE of the autoregressive parameter ϕ. In the stationarity cases, the distributions are symmetric and standardized, unlike with the unit root case, where the distribution is non-standard and asymmetric. This *anomaly* in the asymptotic distribution under the null hypothesis renders the inference fragile and many times unreliable.

Bayesian methods for autoregressive processes have been a major object of interest for the statisticians over the past few years. Zellner and Tiao (1964) were the first who studied AR(1) models, giving the stimulus to Zellner (1971), Box et al. (1976), Marriot and Smith (1992) and Monahan (1984) to analyze AR(1) models with Bayesian methods.

Phillips (1991) provided alternative choices of priors for the evaluation of the posterior distribution, when non-stationarity is assumed. Since then, many analysts, including Ghosh and Heo (2000), Schotman and Van Dijk (1991) and Sims and Uhlig (1991) try noninformative priors without considering stationarity to derive the posterior distribution of the parameters.

The Bayesian framework regards the autoregressive parameter as a random variable and draws inference from the *posterior distribution* of it, which is evaluated by taking the data and any prior knowledge about the parameter as fixed and known.

The posterior distribution of ϕ is proved to be *symmetric* and *standard*, under specific choice of priors and may lead to trustworthy conclusions and p-values, since the posterior distribution of ϕ *is* the inference under the Bayesian approach. In multiparameter problems, all the inference is summarized in the *joint posterior distribution*. Inference about model indicator i is made through the *marginal likelihood*, which is obtained by integrating out the other parameters.

A significant advantage of the Bayesian approach to model selection is the *consistency* of the selected model. If there is enough data available, then, under mild conditions, the posterior probability focuses on the *real* model. That is, the the posterior probability of the models containing the truth approaches 1.

The selection of the proper prior has employed a great amount of literature and comprises a controversial issue among the statisticians. Sims and Phillips have made significant contributions to the substantial issue of *prior selection*.

Sims (1989) argued that the discontinuity of the LSE distribution under the hypothesis of a unit root is a serious drawback of the classical approach and could lead to unreasonable results. He suggested the use of the *simple flat prior*, as a non-informative and rational choice.

On the other hand, Phillips (1991) claims that the use of flat prior is the reason for the divergence between the results of the classical and Bayesian methods. He proposed the Jeffreys' prior as a *satisfactory representation of ignorance*.

With regards to testing the hypothesis of a unit root, it comprises a model comparison with a point null hypothesis of $\phi = 1$. An inherent difficulty in tests is that the *point null hypothesis* can not be effectively checked with *continuous priors* due to the continuity of the posterior, which assigns zero weight to the unit root hypothesis.

There exist tests that either are based on discontinuous priors and thus, do not assign zero weight to the unit root hypothesis (SVD 1991, Dejong and Whiteman, 1991), or are closely related to the point null hypotheses without being actually point null hypotheses, using continuous priors (Koop, 1994).

Hence, the inference is based either on odd ratios and Bayes factors or on model selection criteria, such as BIC, PIC, FIC and others.

4.1.2 From Classical to Bayesian point of view

The classical approach calculates the distribution of $\hat{\phi}$ conditional on a particular value of ϕ, $\phi = 1$. On the contrary, the Bayesian unit root theory focuses on the distribution of $\hat{\phi}|\phi$ for every possible value of ϕ and prior distribution of ϕ.

Skewness and Dispersion

If the distribution of $\hat{\phi}|\phi$ preserves the skewness and dispersion of $\hat{\phi}|\phi = 1$, then the *true* value of ϕ would be on a regular basis larger than the estimator $\hat{\phi}$ (Hamilton, 1994).

The values of ϕ affect the distribution of $\hat{\phi}|\phi$. The skewness of $\hat{\phi}|\phi$ becomes smaller for small values of ϕ. However, the dispersion increases, since the variance of $\sqrt{T}(\hat{\phi} - \phi)$ approximates $1 - \phi^2$.

For instance, under the assumption of no skewness, a given observation $\hat{\phi} = 0.90$ is more likely to origin from a distribution centered at $\phi = 0.85$ and large dispersion, than from a distribution centered at $\phi = 1$ and small dispersion.

This characteristic is fading when the *flat prior* is used, since all values of ϕ will be equally weighted and an observed value of $\hat{\phi} = 0.90$ will give the same probability to ϕ to be less or more than 0.90.

4.2 Posterior distribution of an AR(1) parameter using a diffuse prior

Let

$$X_t = \phi X_{t-1} + \epsilon_t \tag{4.1}$$

be an AR(1) model with drift zero and normally distributed errors, $\epsilon_t \sim N(0, \sigma^2)$. For the posterior distribution of ϕ, the distribution of the vector parameter $\theta = (\phi, \sigma)$ needs first to be evaluated.

The *joint likelihood* of the AR(1) model can be derived either by calculating the *exact likelihood* of the model, regarding the initial value as *an unknown parameter*, or by evaluating the *conditional likelihood*; that is, the initial value X_0 is taken to be *known*.

Based on the latter approach, if the initial observation X_0 is known, then the likelihood can be evaluated by conditioning on this initial value. For instance, the joint function of (X_1, X_2) given X_0 and θ is factorized as the product of the conditional density of X_3 and the marginal likelihood of X_2, as follows

$$f(X_1, X_2|X_0, \theta) = f(X_2|X_1, X_0, \theta) f(X_1|X_0, \theta).$$

Inductively,

$$f(X_T, X_{T-1}, ..., X_1|X_0, \theta) = \prod_{t=1}^{T} f(X_t|I_{t-1}, \theta),$$

where I_{t-1} denotes the information available at time $t-1$. Since the errors are normally distributed, the likelihood becomes

$$L(X_t|\phi, \sigma, X_0) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(X_t - \phi X_{t-1})^2}{2\sigma^2}\right)$$

and thus,

$$L(X_t|X_0, \phi, \sigma) = (2\pi)^{-T/2} \sigma^{-T}\, exp\left(-\frac{\sum_{t=1}^{T}(X_t - \phi X_{t-1})^2}{2\sigma^2}\right).$$

A diffuse prior for the parameter vector $\theta = (\phi, \sigma)$ is selected to represent the *a priori* ignorance about the parameters. Jeffreys (1961) proposed that for location parameters the prior should be taken proportional to a constant, and for scale parameters the prior should be taken proportional to their inverses.

Hence, the parameters are uniformly and independently distributed as follows,

$$f(\phi, \sigma) \propto \frac{1}{\sigma}, \quad -1 < \phi < 1 \quad and \quad \sigma > 0.$$

In the general case when the autoregressive parameter ϕ can be any real number, $-\infty < \phi < \infty$, *the analysis applies to both explosive and nonexplosive cases* (Zellner, 1996, p.187).

The posterior distribution of θ can be evaluated up to a normalising constant by using Bayes' theorem,

$$f(\phi, \sigma|X_t, X_0) \propto f(X_t|\phi, \sigma, X_0) f(\phi, \sigma).$$

Hence,

$$f(\phi, \sigma|X_t, X_0) \propto \sigma^{-T-1} exp\left(-\frac{\sum_{t=1}^{T}(X_t - \phi X_{t-1})^2}{2\sigma^2}\right).$$

The marginal posterior distributions of ϕ and σ are obtained by integrating with respect to σ and ϕ, respectively. Hence,

$$f(\phi|X_t, X_0) = \int_0^\infty f(\phi, \sigma|X_t, X_0)d\sigma \propto \left(\sum_{t=1}^{T}(X_t - \phi X_{t-1})^2\right)^{-T/2}.$$

and

$$f(\sigma|X_t, X_0) = \int_{-1}^{1} f(\phi, \sigma|X_t, X_0)d\phi \propto \sigma^{-T} exp\left(-\frac{\sum_{t=1}^{T}(X_t - \phi X_{t-1})^2}{2\sigma^2}\right)$$

The marginal posterior of ϕ is a *univariate, symmetric about the LSE t-distribution*, while the posterior of σ is an *inverse gamma distribution with 2 degrees of freedom* (Zellner, 1971).

The classical approach to unit root testing relies on the inference made from *the behaviour of the estimators in repeated samples viewing the true parameter of interest as unknown and fixed* (Maddala, Kim). On the contrary, the Bayesian framework draws inference from the posterior distribution of the autoregressive parameter and regards the data and the parameter estimator as known.

Under the presence of a unit root, the Bayesian methods facilitate the process of inference since the posterior distribution is *symmetric* and *standard*, whereas the LSE distribution in the classical framework is *asymmetric* and *non-standard*, as stated in the previous section.

4.3 The "proper" model for unit root testing

A key characteristic of the processes with stochastic trend is the long run fluctuations. The stochastic character of the process implies that the mean does not exist. Consequently, *there is no deterministic trend that serves as a long run anchor around which the process would have a tendency to revert* (Lubrano, 1995).

Initiated by Schotman in 1992, there are two common parameterizations of the standard AR process: *the structural parameterization* and *the reduced form parameterization*. It can be shown that the former, compared with the latter, reflects effectively the properties of the unit root stochastic process.

4.3.1 Schotman's structural model parameterization

Let AR(p) be an autoregressive process of order p. The key question if the model is difference stationary (DS) or trend stationary (TS) can be answered by unit root tests. A more general expression of the process is

$$X_t = \mu + \delta t + v_t, \tag{4.2}$$

$$\Phi(L)v_t = \epsilon_t, \tag{4.3}$$

where ϵ_t is a White noise and $\Phi(L)$ is a lag polynomial of order p. This parameterization has been mostly discussed by Schotman and Van Dijk (1991) and Schotman (1992) for testing a unit root hypothesis. In the simple case of $p = 1$ the model becomes

$$X_t = \mu + \delta t + v_t,$$
$$v_t - \phi v_{t-1} = \epsilon_t,$$

which produces the following equation,

$$X_t = \phi X_{t-1} + [(1-\phi)\mu + \delta\phi] + (1-\phi)\delta t + \epsilon_t. \tag{4.4}$$

There are four cases with regards to the value of ϕ:

- $\phi = 1$: The process is a DS process since there is a unit root, $X_t = \delta + X_{t-1} + \epsilon_t$
- $\phi = 0$: The process is a *pure* TS process, $X_t = \mu + \delta t + \epsilon_t$
- $\phi < 1$: The process is *stationary* with a deterministic trend,
- $|\phi| > 1$: The process is *explosive*.

The model with the unit root has neither intercept μ nor trend. Only the drift δ remains. As outlined in 1.5.2.2, the mean of the process is $E(X_t) = \delta t$ which denotes that the process does not revert to a constant, but grows over time. This is a characteristic feature of the stochastic process with a unit root.

4.3.2 Phillip's reduced form parameterization

The following parameterization was initated by Phillips in 1991 and has been used both in classical and Bayesian framework. It is expressed as a reduced form of 4.2 and 4.3,

$$X_t = \alpha_0 + \alpha_1 t + \phi X_{t-1} + \epsilon_t, \tag{4.5}$$

where, $\alpha_0 = (1-\phi)\mu + \phi\delta$ and $\alpha_1 = (1-\phi)\delta$.

Under the null hypothesis of a unit root the trend coefficient α_1 does not vanish. Hence, X_t will not be of the same *order of magnitude* under each hypothesis (Davidson and Mackinnon, 1993). Specifically, X_t in 4.5 is $O(T)$ under the alternative, $\phi < 1$, and $O(T^2)$ under the null hypothesis, since the presence of a unit root in the process increases the order of X_t (Lubrano, 1995).

The difference between the two parameterizations is not obvious in classical inference, since the estimation of ϕ and its standard deviation would be the same. However, the distributions of Dickey and Fuller are based on asymptotic theory, where this distinction is crucial. Only with the second parameterization the distributional results are obtained.

From the Bayesian perspective, the inference is the posterior density of the parameters. The model in 4.5 with a flat prior can be analyzed and explained with the standard classical framework, providing results similar to those in a stationary case.

On the other hand, the structural model in 4.2 and 4.3 encompass in the joint posterior density one of the major characteristic features of the stochastic process with a unit root: *the absence of the mean*. Hence, the results concluded via this parameterization are similar to those obtained in the classical inference.

4.4 Selecting the appropriate prior

extensive literature and research has been devoted to the Bayesian unit root analysis and specifically to the composite issue of the *prior selection*. Sims (1989), Sims and Uhlig (1991) and Zellner (1971) are some of the Bayesian-statisticians who support the flat prior for the unit root tests.

4.4.1 The flat prior

According to Sims and Uhlig (1991), one-tailed unit root tests cannot usually reject the null hypothesis of a unit root when classical theory is used, compared to the corresponding test based on the t distribution. If the information about the pdf parameters is *diffuse*, then a flat prior is used, that is,

$$f(\phi, \sigma) \propto \frac{1}{\sigma}. \tag{4.6}$$

The posterior distribution of ϕ is a t distribution and hence, the usual t tests are used. Commonly, the flat prior does not easily reject the null hypothesis and thus, favors large values of ϕ.

Sims and Uhlig (1991) compared the classical and the Bayesian approach by evaluating the two following distributions. The first is the distribution of $\hat{\phi}|\phi = 1$, which is the distribution of the estimated parameter under the hypothesis of a unit root and the latter is that of $\phi|\hat{\phi} = 1$, which is the distribution of the true parameter with the estimated parameter as known. They argued that the Bayesian approach provides more *logical* conclusions than the classical approach in the case of a flat prior, especially due to the asymmetry and non-standardness of the classical distribution.

4.4.2 Phillip's ignorance prior

On the contrary, Phillips (1991) claimed that the flat prior does not represent properly the uninformativeness in the data. He argued that the flat prior actually *is* informative, since it favours stationarity, which may explain the discrepancy between the inferences obtained by the Bayesian and classical methods.

One of the basic arguments of Phillip stems from the interpretation of the autoregressive parameter in a AR(1) time series model $X_t = \phi X_{t-1} + \epsilon_t$, and a linear regression model $Y_t = \beta X_t + \epsilon_t$. In the case that $|\phi|$ is large, the data is quite informative about the parameter. Flat priors ignore the impact of the coefficients on the information contained in the data. Hence, a flat prior which gives equal weight to every value of ϕ may *downgrade* higher values of ϕ and hence, the possibility of a unit root.

Phillips proposed that the Jeffreys' invariant to transformations prior should be used, as a satisfactory representation of ignorance. The Jeffreys' prior for the standard AR(1) model in 4.1 is

$$f_J(\phi, \sigma) \propto \frac{1}{\sigma} I_{\phi\phi}^{1/2}, \tag{4.7}$$

where

$$I_{\phi\phi} = \frac{T}{1-\phi^2} - \frac{1}{1-\phi^2}\frac{1-\phi^{2T}}{1-\phi^2} + \left(\frac{X_0}{\sigma}\right)^2 \frac{1-\phi^{2T}}{1-\phi^2}, \quad \text{if } \phi \neq 1$$

and

$$I_{\phi\phi} = \frac{T(T-1)}{2} + T\left(\frac{X_0}{\sigma}\right)^2, \quad \text{if } \phi = 1.$$

The initial value X_0 is often assumed to be zero, otherwise a prior distribution for X_0 would seem a reasonable choice. In the latter case the Jeffreys' prior depends on the prior distribution of the initial value.

The confidence sets that emerge from the Jeffrey's prior are *tighter* when the values of $|\phi|$ become larger. This prior is also invariant to transformations of the parameter. Another noticeable feature is the dependence of Jeffreys' prior of the lag length T, in comparison to the flat prior. If $X_0 = 0$ and $|\phi|$ is large, then

$$f_J(\phi, \sigma) \approx \frac{\sqrt{2}}{\sigma}\phi^{(T-2)},$$

which grows exponentially when $T > 2$.

Under the assumption of *normally distributed errors*, the marginal posterior distribution of ϕ based on the observed data X_t and conditioning on the initial value X_0 is,

$$f(\phi|X_t, X_0) \propto \alpha^{1/2}\left[\sum_{t=1}^{T}(X_t - \phi X_{t-1})^2\right]^{-T/2},$$

where

$$\alpha = \begin{cases} \frac{T}{1-\phi^2} - \frac{1}{1-\phi^2}\frac{1-\phi^{2T}}{1-\phi^2} & \text{if } \phi \neq 1, \\ \frac{T(T-1)}{2} & \text{if } \phi = 1. \end{cases}$$

The Jeffreys' posterior for the AR(1) Gaussian model is less susceptible to downweighting the unit root hypothesis than the posterior based on the flat prior.

Inference is quite sensitive to the model used in the analysis. For instance, the conclusions made above apply only to the AR(1) model without trend or/and intercept. The Jeffreys' ignorance prior proposed by Phillips does not provide rational results for the richer structure

$$X_t = \beta + \gamma t + \phi X_{t-1} + \epsilon_t.$$

It has been argued by Phillips and Schotman and van Dijk (SVD) (1991) that, compared to the flat prior, the prior proposed by Phillips *downweights* the hypothesis of a unit root when the model includes a trend and an intercept.

SVD have provided the Jeffreys' prior for every AR(1) model:

- model with no constant and no trend
- model with constant only
- model with constant and trend

In the first case the prior is biased toward $\phi > 1$. In the other two cases the prior has zero value at $\phi = 1$, which could explain the bias towards stationarity.

4.4.3 Reference prior

Reference priors have been introduced by Berger and Bernando as a motivation to overcome the sensitivity of Jeffreys' prior in larger models. In multiparameter problems, Jeffreys' prior may need to be modified in order to handle the arising difficulties (of downweighting the unit root etc).

Reference priors, on the contrary, are able to *break up* the multiparameter problems into smaller conditional models with a single parameter. This approach applies effectively under iid error assumptions. For dependent data the prior exists only for the stationary case $|\phi| < 1$ and does not exist for the explosive case $|\phi| > 1$.

The non-asymptotic reference prior for a large time length T and for the AR(1) model is

$$\pi(\phi) \propto exp\left\{\frac{1}{2}E_\phi\left[log\sum_{t=1}^{T}X_{t-1}^2\right]\right\},$$

where E_ϕ denotes the expectation over the data $(X_1, .., X_T)$ given ϕ. In the asymptotic case $T \to \infty$ there exists no reliable reference prior for the explosive case $|\phi| > 1$.

4.4.4 The Schotman and Van Dijk prior

Schotman and Van Dijk (1991b) (SVD) developed an approach similar to that of Sims and Uhlig with regards to the prior choice. They were based similarly on the AR(1) model, $X_t = \phi X_{t-1} + \epsilon_t$, and provided a *non informative* prior based on the following assumptions,

(i) The initial value x_0 is a constant and the likelihood is calculated conditionally on x_0.

(ii) ϵ_t is a sequence of independent and identically distributed random variables with zero mean and unknown variance σ^2.

(iii) $\phi \in \{P, 1\}$, where, $P = \{\phi | -1 < \alpha \leq \phi < 1\}$ and α is a significance level

The priors as defined by SVD for ϕ are:

$$Pr(\phi = 1) = \pi_0 \text{ and } f(\phi|\phi \in P) = \tfrac{1}{1-\alpha}$$

and for σ

$$f(\sigma) \propto \tfrac{1}{\sigma}$$

Namely, ϕ is uniformly distributed over P and gives a probability mass on $\phi = 1$. The quantity π_0 is equal to $\frac{K_0}{1+K_0}$, where K_0 denotes the prior odds ratio in favor of the hypothesis $\phi = 1$. A flat prior is also used for σ.

SVD perform the Bayesian unit root testing by initially considering a *mass point* at $\phi = 1$ and consequently comparing the posterior odds between the null and the alternative hypothesis.

4.4.5 Normal-Wishart and Lubrano priors

A widely known non-flat prior choice is the *Normal-Wishart* prior which belongs to the conjugate family. That is, the posterior distribution of the autoregressive parameters follows a Normal-Wishart distribution, too. This prior is quite informative about properties of the model and gives low probability to an explosive root when it centered around $\phi = 1$ (Uhlig, 1994).

There are many discussions about the importance of the initial value. Lubrano (1995) argues that, by regarding the following *structural model* and the initial observation as a random variable, the Bayesian unit root test provides results which almost coordinate with the results of the classical approach. The structural model is,

$$X_t = \alpha + \beta t + v_t,$$
$$v_t = \phi v_{t-1} + \epsilon_t,$$

where $\epsilon_t \sim iid(0, \sigma^2)$. According to Lubrano, the distribution of the initial value, X_0, is

$$X_0 \sim N\left(\alpha, \frac{\sigma^2}{1-\phi^2}\right).$$

The distribution is valid only for $|\phi| < 1$. In the case of $|\phi| > 1$ the distribution is modified as follows

$$X_0 \sim N\left(\alpha, \frac{\sigma^2}{h(1-\phi^2)}\right),$$

where

$$h(1-\phi^2) = \begin{cases} 0, & \text{if } |\phi| > \sqrt{1+u}. \\ (1-\phi^2+u)^2/4u, & \text{if } \sqrt{1-u} \leq |\phi| \leq \sqrt{1+u}, \\ 1-\phi^2, & \text{if } |\phi| < \sqrt{1-u}, \end{cases}$$

where u is suggested to be 0.5 so that $\phi \in [-1.225, 1.225]$.

With regards to the prior, Lubrano uses a diffuse prior which is evaluated as the limit of an informative prior; specifically of a Beta distribution,

$$\pi(\phi|u) \propto (\sqrt{1+u}+\phi)^{p-1}(\sqrt{1+u}-\phi)^{q-1},$$

where $p > 0$ and $q > 0$. By setting $p = 1$ and $q = 0$, Lubrano obtains the non-informative prior

$$\pi(\phi|u) \propto \frac{1}{\sqrt{1+u}-\phi}.$$

This prior is simnilar to Jeffreys' prior in shape, but it does not depend on the length T. The results obtained, when real data is involved, by comparing other priors with the ADF test are closer to the classical framework.

4.5 Bayesian methods for unit root testing

In Bayesian hypothesis testing the null and the alternative hypotheses are on *equal footing*, that is, the purpose of the test in not to *reject or not* the null hypothesis, as it happens with the classical tests. The Bayesian test *compares* the two hypotheses with different methods.

Widespread approaches for the Bayesian unit root testing are the *posterior odds* using posterior or predictive distribution and the model selection via *Information Criteria*, such as PIC, BIC and others.

4.5.1 Providing the evidence

Let $H_0 : \phi = \phi_0$ and $H_1 : \phi = \phi_1$ be the null and the alternative hypotheses of a typical hypothesis test. There exist quantities that measure the evidence and analogously can reject or not a hypothesis. A widely used one by both Bayesians and frequentists is the *p-value*.

Exclusively from the Bayesian point of view the evidence is provided by the *posterior odds ratio* which is based either on the *posterior distribution* or on the *predictive distribution* and depicts which hypothesis dominates the other.

The p-value

Let $t(x)$ be the value of the test statistic when data $X = x$ is observed. If $T(x)$ is more extreme than expected under the assumption of a true H_0, then the null hypothesis is rejected over the alternative. Specifically, H_0 is rejected if the probability of $T(X)$ being greater than the observed $t(x)$ is lower than a significance level α, when H_0 is true. That is,

$$Pr\left[|T(X)| \geq t(x)\right] \leq \alpha.$$

Fisher proposed a measure of evidence according to the significance level. Namely, $\alpha = 0.99$ denotes a very strong evidence against H_0, 0.95 denotes strong evidence and 0.90 implies a neutral evidence.

Posterior odds using the posterior distribution

The posterior probability of H_1 (and analogously for H_0) can be expressed as,

$$f(H_1|x) = \frac{f(x|H_1)f(H_1)}{f(x)},$$

where $f(x) = f(x|H_0)f(H_0) + f(x|H_1)f(H_1)$.

As outlined in Section 2, the comparison between the null and the alternative hypothesis can be performed by evaluating the *posterior odds ratio* based on observed data x, which is the ratio of the posterior probabilities under H_0 and H_1,

$$L_{01} = \frac{f(H_0|x)}{f(H_1|x)} = \frac{f(x|H_0)f(H_0)}{f(x|H_1)f(H_1)}.$$

The posterior odds provide the evidence of the null over the alternative hypothesis when data x is observed. If $L_{01} > 1$ or $f(H_0|x) \geq 0.50$ then H_0 is accepted, otherwise H_1 is accepted. That is, the null and the alternative hypotheses are treated in a symmetric way.

Similarly, the *prior odds* is the ratio $f(H_0)/f(H_1)$ and denotes the prior plausibility of H_0 over H_1 before the data has been observed.

Bayes' factor for unit root testing

Bayes' factor provides evidence with regards to the degree the prior odds update the observed data. It is defined as the ratio

$$BF_{01} = \frac{f(x|H_0)}{f(x|H_1)},$$

or equivalently

$$BF_{01} = \frac{f(H_0|x)}{f(H_1|x)} \Big/ \frac{f(H_0)}{f(H_1)}.$$

In the case of setting equal prior probabilities to the hypotheses, $P(H_0) = P(H_1) = 1/2$, the Bayes' factor coincides with the posterior odds.

Bayes' factor can suitably replace p-value in terms of the *decision rule* in a hypothesis test, since a large value of the first provides stronger evidence in favor of H_0 over H_1 which leads to a decision. In the latter case the decision over a hypothesis is based on the significance level α.

Bayes factor and posterior odds apply to the unit root test

$$H_0 : \phi = 1,$$
$$H_1 : \phi \in A,$$

for any generic AR model and alternative hypothesis. A weight π_0 is assigned to the null hypothesis and the complement $1 - \pi_0$ to the alternative. In the general case, let $K = \{\phi, \rho\}$, $\rho \in \mathbb{R}^{k-1}$ depict the k-dimensional parameter vector, where {K,A}$\in$ $\mathbb{R}^k$. The evaluation of the Bayes' factor includes a multidimensional marginalization over the parameters of Θ. The posterior odds conditioning on the initial value x_0 are thus,

$$L_{01} = \frac{f(\phi = 1|\rho)}{f(\phi \in A|\rho)} = \frac{\pi_0}{1 - \pi_0} \frac{\int_K f(x|\phi = 1, \rho, x_0) f(\rho|\phi = 1) d\rho}{\int_A \int_K f(x|\phi, \rho, x_0) f(\rho|\phi) f(\phi) d\rho d\phi}.$$

The factorization of the joint prior $f(\phi, \rho) = f(\rho|\phi)f(\phi)$ can be applied due to the assumed conditional indepedence of the parameters.

In the simple case of an AR(1) model, $K = \{\phi, \sigma\}$ and $k = 2$. Hence, the posterior odds ratio is,

$$L_{01} = \frac{f(\phi = 1|\sigma)}{f(\phi \in A|\sigma)} = \frac{\pi_0}{1 - \pi_0} \frac{\int_0^\infty f(x|\phi = 1, \sigma, x_0) f(\sigma) d\sigma}{\int_A \int_0^\infty f(x|\phi, \rho, x_0) f(\sigma) f(\phi) d\rho d\phi}.$$

Commonly, the conditional independence among the parameters is extended to *unconditional variance*, which provides that σ follows a prior distribution $f(\sigma|\phi) = f(\sigma)$.

Posterior odds using the predictive distribution

The predictive distribution is used in the Bayesian framework as an alternative for the posterior distribution in terms of the posterior odds ratio. The predictive density of a future observation $X^\star$ is $f(x^\star|x)$ and is evaluated as

$$f(x^\star|X) = \int f(x^\star|\phi, X) f(\phi|X) d\phi,$$

where $f(X^\star|\phi, X)$ denotes the *updated* likelihood function and $f(\phi|X)$ is the posterior distribution of ϕ.

The predictive odds ratio of the models in the two hypotheses is

$$L_{01} = \frac{f(H_0|x^\star)}{f(H_1|x^\star)} = \frac{f(H_0)f(x^\star|x,H_0)}{f(H_1)f(x^\star|x,H_1)}. \tag{4.8}$$

Specifically, the predictive density under the null and the alternative hypotheses is calculated as,

$$f(x^\star|x,H_i) = \int f(x^\star|x,\phi_i,H_i)f(\phi_i|x,H_i)d\phi_i, \quad i = 0,1$$

Thus, the predictive odds ratio in 4.8 becomes

$$L_{01} = \frac{f(H_0)}{f(H_1)} \frac{\int f(x^\star|x,\phi_0,H_0)f(\phi_0|x,H_0)d\phi_0}{\int f(x^\star|x,\phi_1,H_1)f(\phi_1|x,H_1)d\phi_1}$$

In the posterior odds approach the selection of the noninformative priors is arbitrary. For the predictive odds approach Maddala and Kim (1998) have stated the following:

"Initial noninformative priors of the form $P(\alpha,\sigma^2) \propto \sigma^2$ are applied to each model over periods 1 to m (m <n) to produce proper posterior densities for (α,σ^2) . These are then used as priors to evaluate proper predictive densities for periods (m+1) to n."

That is, a subsample is used to produce *proper* posterior distributions which are then used as a prior for the rest of the sample.

In essence, the approach of predictive odds ratios is a *complete* Bayesian approach and can provide trustworthy conclusions even in the case of a noninformative prior, which in the posterior odds approach may cause difficulties.

4.5.2 Model selection criteria

In the field of model selection, the hypotheses should not be viewed and tested as if they were *true* entities, but as potential models that probably or hopefully could fit the data.

As discussed in Section 2 there exist several model selection methods and criteria. Below there are demonstrated some further information about the model selection which mostly pertain to the unit root hypothesis testing.

The comparison of the involved models M_i, $i = 1,2,..,k$ requires the determination of prior model probabilities of the models $f(M_i)$ and prior probabilities for the parameters $f(\phi)$. The posterior density obtained is then used in criteria for the selection of the model which is most probable to fit the data.

The PSR quasi-likelihood and PSR quasi-Bayes criteria

The vast majority of time series analysis aim eventually at the *prediction* of upcoming occurrences. Therefore, apart from the fundamental question of *which model fits the observed data best*, the question that, in most circumstances is of real interest is *which of the models yields the best predictions for future observations from the same process which generated the given set of data* (Geisser and Eddy, 1979).

Geisser and Eddy (1979) have suggested two criteria based on the *predictive sample reuse methods* (PSR), which were first introduced by Geisser (1975). PSR is a combination of two widely known approaches in data analysis, *cross-validatory assessment* and *function fitting*.

Based on the above theory, the first criterion is the *PSR quasi-likelihood* criterion which selects the model that maximizes the quantity

$$\hat{L}_k = \prod_{i=1}^{N} f(x_i|\hat{\phi}_{(k)}, M_k),$$

where $\hat{\phi}_{(k)}$ is the maximum likelihood estimator of ϕ_k if x_i is not included.

The second criterion is termed as *PSR* quasi-Bayes and selects the model that maximizes

$$L_k = \prod_{i=1}^{N} f_p(x_i|x_{(i)}, M_k),$$

where $x_{(i)}$ is the data when x_i is omitted and f_p denotes the predictive distribution,

$$f_p(x_i|x_{(i)}, M_k) = \int f(x_i|\phi_k, M_k) df(\phi_k|x_{(i)}, M_k),$$

where $f(\phi_k|x_{(i)}, M_k)$ is the posterior distribution of ϕ_k based on data $x_{(i)}$ and a diffuse prior on the parameter ϕ_k.

The BIC and FIC criteria

According to the *Bayesian Information Criterion*, BIC or *Schwatz's criterion*, the model M_i that minimizes

$$BIC = \log\hat{\sigma^2_{p_i}} - p_i\frac{\log n}{n},$$

where $\hat{\sigma^2_{p_i}}$ is the estimate of σ^2 in a model with p_i parameters, is the preferred one.

A criterion similar to the BIC was proposed by Fisher. The Fisher information criterion (FIC) suggests that the model M_i that minimizes

$$FIC = n\hat{\sigma_l^2} + \hat{\sigma_L^2}\ln|A_l|,$$

where $A_{p_l} = X'X$ and X is the design matrix for the data and $\hat{\sigma^2_{p_l}}$ and σ^2_L are the estimates of the variances for the models with l and L parameters respectively. The model with L parameters is the *richest* model.

FIC seems quite reliable in selecting the true model and is close to the behavior of BIC.

The PIC criterion

Phillips and Ploberger suggested another model selection criterion, the *posterior information criterion* (PIC). The selected model is the model that minimizes

$$P = c_L|A_l/\hat{\sigma}^2_L|^{1/2}\, exp\left[-(1/2\hat{\sigma}_L)\hat{\beta_l}'A_l\hat{\beta}_l\right]$$

where, c_L denotes a constant which depends on L, the maximum number of explanatory variables, $\hat{\sigma_L^2}$ is the estimate variance of the richest model, $A_l = X_l'X_l$ is the data matrix of the model with l regressors.

The PIC is based directly on the data matrix A_l , unlike with the BIC which depends on the number of parameters and the variance estimator. However, the way the PIC is formulated is not completely Bayesian due to the estimation process of the residual variance.

Other criteria have been discussed by Phillips and Koop (1994). Phillips proposed a variation of PIC, the PICF which is close to PIC and depends on the predictive distributions, as they have been discussed earlier. It can be used for a *sequential model choice*, when a single model cannot be applied for the whole period, but changes over time.

The criterion suggested by Koop is the *Bayesian likelihood ratio* criterion (BLR) which tests the null hypothesis of a unit root against the stationary alternative.

4.6 Criticism on Bayesian unit root testing and alternatives

The Bayesian perspective in unit root testing has apparently many advantages over the classical unit root tests, especially in terms of model comparison and selection. However, a relatively small number of studies regarding Bayesian unit root testing have appeared. The reasons may be the difficulty in deriving numerically the likelihood function and the controversial issue of the prior selection. (Ahking, 2009)

The arbitrary selection of the priors and computational adversity renders the Bayesian perspective in unit root testing more controversial and less 'objective'. Phillips emphasized the subjectivity of the prior choice as a major drawback of the Bayesian unit root testing and argued for a test which would be more objective. Koop's unit root test is based on the work of Zellner and Siow and represents an 'objective' alternative in Bayesian unit root testing. This test is computationally simple and does not require the use of improper non-informative priors, since it involves *informative priors*. Additionally, to all hypotheses is assigned equal probability, which diminishes the subjective character of the test.

Chapter 5

Simulation study

The purpose of the following study is to implement Bayesian unit root tests in Monte Carlo simulated data from an AR(1) process without constant and for different values of the length T and the parameter ϕ. In particular, there are generated $N = 1000$ AR(1) time series of length $T = \{25, 50, 100, 500, 1000, 5000\}$ and autoregressive parameter values $\phi = \{0.7, 0.8, 0.9, 0.95, 0.99, 1\}$. The simulation study aims at discriminating between a process with a unit root, i.e. a random walk and a stationary process. The hypotheses examined are: $H_0 : \phi = 1$ and $H_1 : |\phi| < 1$.

Initially, the decision rule underlying the test is the *posterior odds ratio* criterion; namely, the model with the highest posterior probability is chosen. The prior odds are regarded balanced with $P(H_0) = P(H_1) = 0.5$ which means that the chosen model is the one which is most favoured by the data,

$$\frac{Pr(H_0|data)}{Pr(H_1|data)} = \frac{f(data|H_0)}{f(data|H_1)},$$

since $\frac{Pr(H_0)}{Pr(H_1)} = 1$. Hence, the posterior odds is identical to Bayes' factor.

Subsequently, the hypotheses tests are performed according to the BIC criterion.

Diffuse prior

Initially, the diffuse, non-informative prior of Schotman and Van Dijk (SVD) (1991b) is employed to test the null hypothesis: $H_0 : X_t = X_{t-1} + \epsilon_t$ against $H_1 : X_t = \phi X_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$ and σ^2 unknown. The constant X_0 is considered known and $\phi \in P \cup \{1\}$, where $P = \{\phi | -1 < \alpha \leq \phi < 1\}$. The marginal priors for ϕ and σ^2 are defined as

(i) Pr(ϕ=1)=π_0

(ii) Pr($\phi|\phi \in P$)=$\frac{1}{1-\alpha}$

(iii) f(σ) $\propto \frac{1}{\sigma}$

The probability of the random walk hypothesis is denoted by the positive mass π_0. Hence, the prior odds in favour of H_0 are:

$$Prior\,odds = \frac{Pr(H_0)}{Pr(H_1)} = \frac{\pi_0}{1 - \pi_0}.$$

The prior on σ is diffuse and the prior on ϕ is uniform in $[\alpha, 1)$ and has a probability π_0 at $\phi = 1$. The posterior odds ratio therefore becomes,

$$Posterior\,odds = \frac{Pr(\phi = 1|data)}{Pr(\phi \in P|data)} = \frac{\int_0^\infty f(x|H_0, \phi = 1, \sigma) f(\phi = 1) f(\sigma) \mathrm{d}\sigma}{\int_0^\infty \int_\alpha^1 f(x|\phi, \sigma, H1) f(\phi) f(\sigma) \mathrm{d}\text{Œœ}}.$$

Let P_0 and P_1 denote the prior and posterior odds respectively. Table 5.1 demonstrates *the percentages of random walk cases* for every value of ϕ and T. The entries of the table designate the probability of $P_1 > 1$ if $N = 1000$ time series for every value of ϕ and T are generated. The value of π_0 is taken to be 0.5 and the lower bound -1 is chosen for the parameter α.

TABLE 5.1: Posterior odds probabilities in favour of the random walk

ϕ \ T	25	50	100	500	1000	5000
0.7	.380	.075	.000	.000	.000	.000
0.8	.623	.332	.028	.000	.000	.000
0.9	.833	.786	.520	.000	.000	.000
0.95	.886	.885	.892	.058	.000	.000
0.99	.922	.937	.981	.980	.914	.000
1	.948	.967	.978	.999	1.000	1.000

Table 5.1 indicates that, when the length T is equal to 25 or 50 the posterior odds are strongly in favour of the random walk hypothesis for almost every value of ϕ. Unlike with the equally balanced prior model probabilities, the probability of the unit root hypothesis for $\phi = 0.8$ exceeds 0.6. The results are generally biased towards the random walk hypothesis.

It can be easily inferred that, as length T and ϕ increase, the probability of correctly identifying the random walk model approaches unity. Indicatively, the probabilities of being in favor of H_0 are almost zero when ϕ is 0.7, 0.8, 0.9 and 0.95 and T exceeds 500.

Comparing the relative results derived from the ADF test with those in Table 5.1, it can be concluded that the acceptance probabilities of H_0 follow almost the same pattern. That is, there is still great possibility of incorrectly accepting the random walk model (0.457 for $\phi = 0.7$ and $T = 25$). However, the increase in length and ϕ values leads to high acceptance probabilities slightly lower than the relative ones in Table 5.1.

According to Jeffreys' rule outlined in subsection 2.3.4 the decision in favour of H_0 can be discriminated in *poor*, *substantial*, *strong* and *decisive*. The results for $T = 25$ and $T = 1000$ are demonstrated below.

A worth mentioned outcome from tables 5.3 and 5.4 is the difference in the decisiveness of being in favor of the null hypothesis using the posterior odds. In the case of $T = 25$ only 3.6% accepts strongly the nonstationary model when $\phi = 1$. On the other hand, the corresponding percentage for $T = 100$ is 85.5%.

TABLE 5.2: Probabilities in favour of the random walk using ADF

ϕ \ T	25	50	100	500	1000	5000
0.7	.457	.037	.000	.000	.000	.000
0.8	.675	.234	.001	.000	.000	.000
0.9	.862	.685	.203	.000	.000	.000
0.95	.903	.831	.703	.000	.000	.000
0.99	.932	.921	.939	.702	.262	.000
1	.960	.944	.949	.959	.947	.953

TABLE 5.3: Jeffreys' rule for $T = 25$ in favour of the random walk

ϕ	poor	substantial	strong	decisive
0.7	.291	.086	.003	.000
0.8	.359	.235	.029	.000
0.9	.247	.378	.208	.000
0.95	.176	.335	.375	.000
0.99	.137	.257	.508	.020
1	.102	.269	.541	.036

TABLE 5.4: Jeffreys' rule for $T = 100$ in favour of the random walk

ϕ	poor	substantial	strong	decisive
0.7	.000	.000	.000	.000
0.8	.000	.000	.000	.000
0.9	.000	.000	.000	.000
0.95	.000	.000	.000	.000
0.99	.126	.213	.547	.028
1	.006	.012	.127	.855

Table 5.4 depicts the posterior probabilities of $\phi \geq 1|x$ for every value of $T \leq 500$ and ϕ. The results when ϕ approaches unity are moderate. That is, the highest posterior probability is 32,5% in the unit root case ($\phi = 1$).

TABLE 5.5: Posterior probabilities of $Pr(\phi \geq 1|x)$

ϕ \ T	25	50	100	500
0.7	.001	.000	.000	.000
0.8	.003	.000	.000	.000
0.9	.034	.002	.000	.000
0.95	.128	.033	.001	.000
0.99	.278	.207	.143	.002
1	.355	.333	.318	.325

Table 5.6 demonstrates the probabilities in favour of the null hypothesis according to the BIC. The formula for the BIC calculation is

$$BIC = (T - z - 1) \cdot log(\sigma^2_{\hat{\phi}}) + p \cdot log(T - z - 1),$$

where T is the length of the series, z is the lag length, $\sigma^2_{\hat{\phi}}$ denotes the variance of the model and p is the number of the estimated parameters.

The entries in Table 5.6 denote the probabilities of accepting the null hypothesis according to BIC for all different values of ϕ and T. The BIC criterion has difficulty in discriminating a random walk from a stationary process for relatively small samples (25, 50). For instance, the exceeding value for $\phi = 0.8$ and $T = 25$ is 0.621.

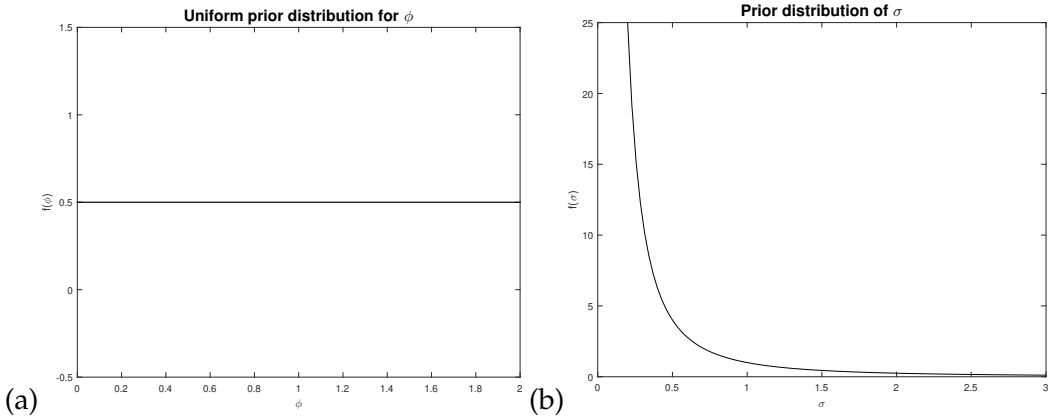
For moderate and large samples the BIC tends to have the same bahaviour as the posterior odds criterion.

TABLE 5.6: BIC probabilities in favour of the random walk

ϕ \ T	25	50	100	500	1000	5000
0.7	.378	.039	.000	.000	.000	.000
0.8	.621	.259	.011	.000	.000	.000
0.9	.844	.720	.383	.000	.000	.000
0.95	.891	.862	.791	.006	.000	.000
0.99	.918	.930	.958	.907	.731	.000
1	.916	.945	.973	.983	.988	.993

The results for a specific data set are presented below. The data generating process is $X_t = \phi X_{t-1} + \epsilon_t$ with $\phi = 0.9$ and ϵ_t gaussian white noise. The SVD prior is used. After applying OLS the estimated model becomes

$$X_t = 0.9733 X_{t-1} + \epsilon_t$$



(a) (b)

FIGURE 5.1: (a) Prior of ϕ (b) Prior of σ

The marginal posterior distributions of ϕ and σ are

$$\phi|\sigma, x \sim t(0.973, 0.005, 99)$$
$$\sigma|\phi, x \sim IG(49.5, 43.8)$$

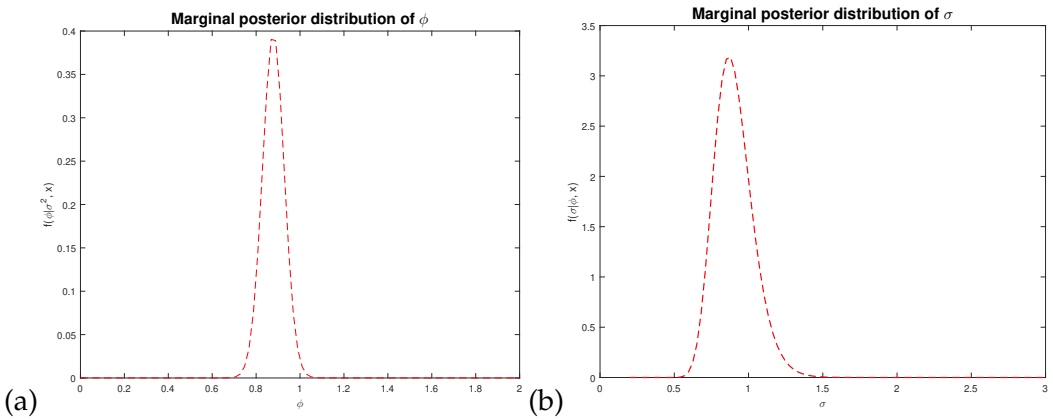


FIGURE 5.2: (a) Marginal posterior of ϕ (b) Marginal posterior of σ

It can be concluded from figure 5.2(a) that although the SVD prior of ϕ was uniform (diffuse), the posterior t distribution is centered about $\hat{\phi}$ and has a very small variance.

Informative prior

The hypothesis of the unit root in the simple AR(1) model $X_t = X_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$ with σ^2 unknown, is now tested using a subjective and informative prior. The joint prior distribution is

$$f(\phi, \sigma^2) = f(\phi|\sigma^2) f(\sigma^2),$$

where

$$\begin{aligned} \phi|\sigma^2 &\sim N(\mu, \sigma^2) \\ \sigma^2 &\sim InverseGamma(\alpha, \beta) \end{aligned}$$

and ϕ, σ^2 independent parameters.

After many trials the parameters for the current simulation study are $\mu = 0$, $\alpha = 4$ and $\beta = 0.5$.

Table 5.9 includes the probabilities of selecting the random walk hypothesis according to the *Bayes Factor*, which is equal to the posterior odds ratio, since the models are equally balanced with $P(H_0) = P(H_1) = 1/2$. Specifically,

$$Posterior\,odds = \frac{\int_0^{\infty} f(x|\phi = 1, \sigma^2, H_0) \mathrm{d}\sigma^2}{\int_0^{\infty} \int_{-\infty}^{+\infty} f(x|\phi, \sigma, H_1) f(\phi|\sigma) f(\sigma) \mathrm{d}\phi \mathrm{d}\sigma^2}.$$

The entries in Table 5.7 indicate that the probabilities of accepting the unit root model are progressively higher as ϕ approaches unity for every value of T. The probabilities at small lengths (T=25 and T=50) imply that it is quite probable to accept a random walk hypothesis for values above 0.9. However, there is still a 25% probability of rejecting the random walk for $T = 25$ when $\phi = 1$.

In essence, the subjectivity in the prior renders the posterior odds criterion *more reliable* compared to the relative results coming from the diffuse prior.

The BIC seems, on the other hand, more *flexible* with regards to the acceptance of a random walk. For instance, the length $T = 100$ and $\phi = 0.8$ produce an acceptance

TABLE 5.7: Posterior odds probabilities in favour of the random walk

ϕ \ T	25	50	100	500	1000	5000
0.7	.030	.000	.000	.000	.000	.000
0.8	.136	.053	.000	.000	.000	.000
0.9	.452	.395	.170	.000	.000	.000
0.95	.595	.672	.630	.014	.000	.000
0.99	.695	.850	.913	.919	.770	.000
1	.738	.880	.942	.986	.993	.999

probability of almost 50%. This tendency may be quite rational, since this criterion is biased towards the *simpler* models.

TABLE 5.8: BIC probabilities in favour of the random walk

ϕ \ T	25	50	100	500	1000	5000
0.7	.294	.116	.004	.000	.000	.000
0.8	.557	.469	.126	.000	.000	.000
0.9	.781	.809	.729	.000	.000	.000
0.95	.861	.930	.933	.397	.000	.000
0.99	.912	.960	.996	.994	.986	.058
1	.933	.968	.987	1.000	1.000	.999

Comparing the Bayesian methods of the Bayes factor and the BIC to the classical ADF unit root test for the same data set, it can be concluded that for small series length the frequentist approach is pretty unreliable due to the high probabilities in favour of the random walk. Indicatively, for $\phi = 0.7$ and $T = 25$ the relative probability is 0.427.

TABLE 5.9: Probabilities in favour of the random walk using ADF

ϕ \ T	25	50	100	500	1000	5000
0.7	.427	.031	.000	.000	.000	.000
0.8	.684	.258	.004	.000	.000	.000
0.9	.857	.677	.230	.000	.000	.000
0.95	.918	.848	.676	.397	.000	.000
0.99	.960	.932	.927	.727	.221	.000
1	.957	.951	.954	.951	.942	.955

The decisiveness of the Bayes factor for $T = 25$ is generally poor, as Table 5.10 indicates. Indeed, only in 2,2% of the cases there is decisive choice in favour of the random walk when $\phi = 1$. However, the probability of *strong* evidence against the stationary alternative reaches 0.372.

As T increases, the probabilities of strong and decisive acceptance of the random walk hypothesis increase. Indicatively, the highest *decisive* acceptance probability

derived from table 5.11 for $T = 100$ and $\phi = 1$ is 65.6%. Nevertheless, the corresponding percentage under the diffuse prior hypothesis reaches 85.5%.

TABLE 5.10: Jeffreys' rule for T=25

ϕ \ T	poor	substantial	strong	decisive
0.7	.024	.006	.000	.000
0.8	.077	.051	.003	.000
0.9	.150	.162	.094	.001
0.95	.149	.207	.227	.011
0.99	.134	.174	.372	.013
1	.126	.191	.361	.022

TABLE 5.11: Jeffreys' rule for T=100

ϕ \ T	poor	substantial	strong	decisive
0.7	.000	.000	.000	.000
0.8	.000	.001	.000	.000
0.9	.109	.056	.021	.001
0.95	.128	.156	.280	.045
0.99	.057	.077	.252	.524
1	.034	.051	.196	.656

The graphs for a specific data set are demonstrated below. The prior for ϕ is the standard Normal distribution and for σ^2 is the InverseGamma distribution with shape parameter $\alpha = 4$ and scale parameter $\beta = 0.5$.

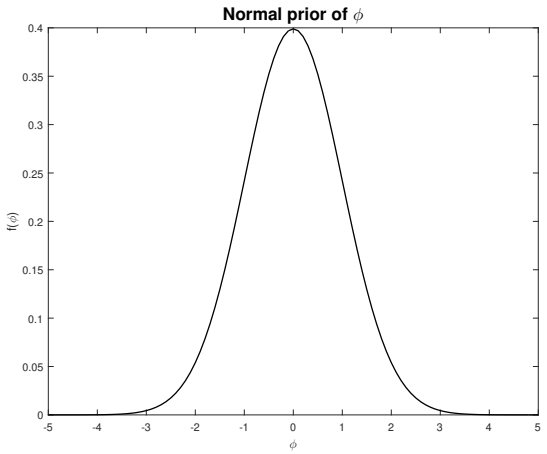


FIGURE 5.3: Normal prior distribution for ϕ

The *real* parameter of the data generating process is $\phi = 0.9$. The implementation of OLS method on ϕ produces the estimated AR(1) model:

$$X_t = 0.8512\, X_{t-1} + \epsilon_t$$

The posterior distributions of ϕ and σ are

$$\phi|\sigma, x \sim t(0.8512, 1.06, 108)$$
$$\sigma|\phi, x \sim IG(54, 45.7)$$

Figure 5.4 depicts the prior and posterior distribution for ϕ.

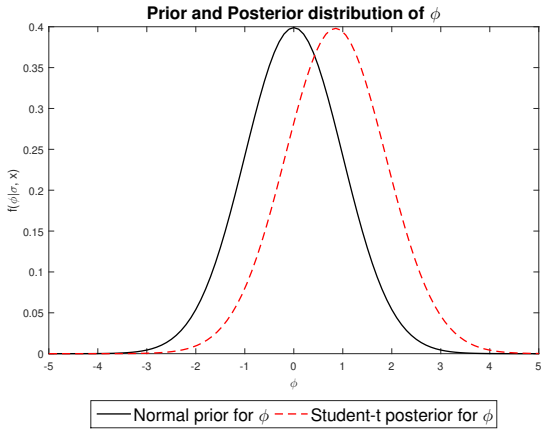


FIGURE 5.4: Prior and marginal posterior distribution of ϕ

The prior and posterior distribution of σ^2 are demonstrated below.

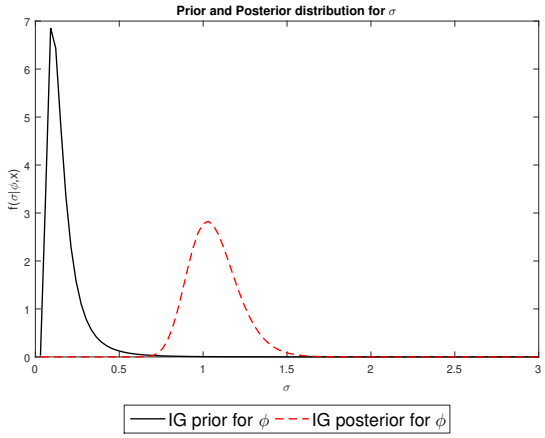


FIGURE 5.5: Prior and marginal posterior distribution for σ^2

Chapter 6

Conclusions

The Bayesian approach in the unit root testing theory has been of great interest and research, especially in economic fields. The controversial issue of discriminating a random walk process from a stationary process is examined by the simulation study in the previous chapter. The AR(1) process without drift was tested for a unit root under the assumption of a *diffuse* and an *informative prior*.

The Bayesian inference *tools* employed for the study were the *posterior odds*, which coincides with the Bayes factor due to the equally balanced hypotheses, and the BIC. The probabilities in favor of the random walk process were, subsequently, compared to those from the classical ADF test.

For relatively *small values of* ϕ the two priors for the AR(1) model yielded the following results.

The *posterior odds criterion* for the SVD diffuse prior yielded that the incorrect identification of the random walk has *high* probability. On the contrary, the informative prior can detect more accurately a stationary process when ϕ and T are small. Hence, the probabilities on favour of the random walk process were *lower*.

By assigning *medium to large values to* ϕ, the behaviour of both priors does not differ significantly. The informative prior tends to detect slightly better the stationary process when ϕ is less than one. The BIC probabilities in favour of the random walk are significantly higher comparing to those coming from the posterior odds, which yields the inference according to the Information Criterion more unreliable.

As T is getting higher the probabilities in favour of the random walk are getting lower when ϕ is not equal to one. Hence, the random walk in the most cases is correctly detected.

Bibliography

[1] Francis W Ahking. "The power of the'objective'Bayesian unit-root test". In: *The Open Economics Journal* 2 (2009), pp. 71–79.

[2] Francis W Ahking. "The power of the" objective" Bayesian unit-root test". In: (2004).

[3] Clair Alston et al. "Bayesian model comparison: Review and discussion". In: *International Statistical Insitute, 55th session* (2005).

[4] James O Berger et al. "Objective Bayesian methods for model selection: Introduction and comparison". In: *Lecture Notes-Monograph Series* (2001), pp. 135–207.

[5] José M Bernardo and Adrian FM Smith. *Bayesian theory*. Vol. 405. John Wiley & Sons, 2009.

[6] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[7] Peter J Brockwell et al. *Introduction to time series and forecasting*. Springer, 2016.

[8] Lyle D Broemeling. *Bayesian analysis of time series*. CRC Press, 2019.

[9] Jiahua Chen and Zehua Chen. "Extended Bayesian information criteria for model selection with large model spaces". In: *Biometrika* 95.3 (2008), pp. 759–771.

[10] Kosei Fukuda. "Practical Unit-Root Analysis Using Information Criteria: Simulation Evidence". In: *Journal of Modern Applied Statistical Methods* 6 (2007), p. 24.

[11] Wayne A Fuller. *Introduction to statistical time series*. Vol. 428. John Wiley & Sons, 2009.

[12] Seymour Geisser and William F Eddy. "A predictive approach to model selection". In: *Journal of the American Statistical Association* 74.365 (1979), pp. 153–160.

[13] Andrew Gelman et al. *Bayesian data analysis*. CRC press, 2013.

[14] Margherita Gerolimetto and Isabella Procidano. "Bayesian Unit Root Tests: a Monte Carlo Study". In: *46TH SCIENTIFIC MEETING OF THE ITALIAN STATISTICAL SOCIETY* (2012).

[15] James Douglas Hamilton. *Time series analysis*. Princeton university press, 1994.

[16] Hossein Masoumi Karakani, Janet van Niekerk, and Paul van Staden. "Bayesian Analysis of AR (1) model". In: *arXiv preprint arXiv:1611.08747* (2016).

[17] Robert E Kass and Adrian E Raftery. "Bayes factors". In: *Journal of the american statistical association* 90.430 (1995), pp. 773–795.

[18] Denis Kwiatkowski et al. "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" In: *Journal of econometrics* 54.1-3 (1992), pp. 159–178.

[19] Michel Lubrano. "Testing for unit roots in a Bayesian framework". In: *Journal of Econometrics* 69.1 (1995), pp. 81–109.

[20] Gangadharrao S Maddala and In-Moo Kim. "Unit roots, cointegration, and structural change". In: (1998), pp. 8–154, 263–295.

[21] Pierre-Alexandre Mattei. "A parsimonious tour of bayesian model uncertainty". In: *arXiv preprint arXiv:1902.05539* (2019).

[22] Charles R Nelson and Charles R Plosser. "Trends and random walks in macroeconmic time series: some evidence and implications". In: *Journal of monetary economics* 10.2 (1982), pp. 139–162.

[23] Luis Raúl Pericchi. "Model selection and hypothesis testing based on objective probabilities and Bayes factors". In: *Handbook of statistics* 25 (2005), pp. 115–149.

[24] Peter CB Phillips and Pierre Perron. "Testing for a unit root in time series regression". In: *Biometrika* 75.2 (1988), pp. 335–346.

[25] Peter CB Phillips and Zhijie Xiao. "A primer on unit root testing". In: *Journal of Economic Surveys* 12.5 (1998), pp. 423–470.

[26] Don van Ravenzwaaij and Alexander Etz. "Simulation studies as a tool to understand Bayes Factors". In: *Simulation* 1 (2020), pp. 0–385.

[27] P Schotman and HK van Dijk. *A Bayesian analysis of the unit root hypothesis*. Tech. rep. 1989.

[28] Peter C Schotman. "Priors for the AR (1) model: parameterization issues and time series considerations". In: *Econometric Theory* 10.3-4 (1994), pp. 579–595.

[29] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Cengage learning, 2015.