Lecture notes on Time Series Econometrics

Riccardo (Jack) Lucchetti

30th October 2026

Preface

This book has a long history. It originated in the 1990s as a set of lecture notes, in Italian, for my Econometrics course, and gradually grew into a text that I could use to give students something that could be "read" as well as "studied." That version had mainly two merits: (a) it was freely downloadable, and (b) it was one of the few tutorials on time series econometrics available in Italian. As a result, it was used by quite a few people in the Italian econometric and statistical community.

The need for an English translation grew stronger and stronger over the years, but I really didn't feel up to the mammoth task it entailed. Fortunately, at some point, my benefactor Giulio Palomba stepped in and gave me a draft translation in which 90% of the work had already been done. Thank you, bro. I owe you one.

So I seized the opportunity and revised the draft, not only in terms of language but also in terms of content. Those familiar with the Italian version will notice several changes here and there (hopefully, for the better): most examples were updated, some material was added, some was ditched. The book's structure has changed a little, too. That said, please note: **this is not finished work**. Any feedback is appreciated, and I'll try to do my best to incorporate any suggestions you may have. That goes, of course, for the English too: I'm not a native speaker, so some sentences may sound a little awkward. Again, advice and corrections are very welcome.

My main goal has always been to describe various concepts in an intuitive way and to try to motivate definitions and main results as explicitly as possible. At the same time, I tried to avoid being too technical: most chapters contain a final section, called "Assorted results", for the material that was not indispensable, but nice to have around just in case the reader wanted to dig a little deeeper. Also, I tried to limit references to the literature as much as possible and confine bibliographical references to a final chapter, which would also serve to guide the reader through the *mare magnum* of time series econometrics.

Over the years, I have received a great deal of feedback from many people, whom I would like to thank for helping to improve its contents. Among my peers, I would like to mention in particular (without holding them responsible for any remaining errors) Gianni Amisano, Marco Avarucci, Emanuele Bacchiocchi, Nunzio Cappuccio, Francesca Di Iorio, Luca Fanelli, Massimo Franchi, Carlo Favero, Roberto Golinelli, Diego Lubian, Giulio Palomba, Matteo Pelagatti, Eduardo Rossi, Maurizio Serva, Stefano Siviero, and Gennaro

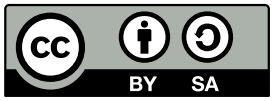
Zezza. Carlo Giannini deserves a special mention because, without him, I would probably have done something completely different with my life, and these lecture notes would never have existed. I would certainly have been a worse person. And, of course, I must also thank the hundreds of students down whose throats these lecture notes were forced.

All the examples in the book are worked out using gretl, which wouldn't exist without my friend Allin Cottrell and the rest of the gretl development gang. Thank you, guys.

As for the prerequisites, I assume that readers already have a certain degree of familiarity with the main concepts of probability (random variables, expected values, conditioning, and various notions of convergence), with OLS, and with some basic concepts of estimation theory, such as identification and the properties of estimators. Therefore, those who have no clue what I'm talking about might as well close this text right here and go hit the books. The rest of you, make yourselves comfortable, and let's get started.

Some passages are written in a smaller font and arranged in two columns, like this one. They are not essential and can be skipped without affecting your understanding of the rest. But then, if I wrote them, they must serve some purpose. It's up to you.

This work is licensed under a Creative Commons "Attribution-ShareAlike 3.0 Unported" licence.



Contents

Preface iii

1 Introduction 1
- 1.1 What is a stochastic process? What is it good for? . . . 1
- 1.2 Characteristics of Stochastic Processes . . . 2
- 1.3 Moments . . . 5
- 1.4 Some examples . . . 6

2 ARMA processes 13
- 2.1 The lag operator . . . 13
- 2.2 White noise processes . . . 16
- 2.3 MA processes . . . 17
- 2.4 AR processes . . . 22
- 2.5 ARMA processes . . . 27
  - 2.5.1 Seasonal ARMA . . . 28
- 2.A Assorted results . . . 30
  - 2.A.1 Inverting polynomials . . . 30
  - 2.A.2 An intro to complex numbers . . . 32

3 ARMA models 37
- 3.1 Numerical techniques . . . 38
- 3.2 Lag Selection . . . 39
- 3.3 Computing the log-likelihood . . . 42
- 3.4 Using ARMA models . . . 45
  - 3.4.1 Forecasting . . . 45
  - 3.4.2 Persistence analysis . . . 49
- 3.5 In practice . . . 52

4 Unit roots 61
- 4.1 Typical features of macro time series . . . 61
- 4.2 Unit root processes . . . 64
- 4.3 Beveridge-Nelson decomposition . . . 68
- 4.4 Unit root tests . . . 71
  - 4.4.1 Distribution of the test statistic . . . 72
  - 4.4.2 Short-run persistence . . . 73
  - 4.4.3 Deterministic component . . . 74
  - 4.4.4 ADF variants . . . 75

4.4.5 The KPSS test . . . 76
4.4.6 An example . . . 77
4.4.7 Use your brain . . . 79
4.5 Spurious regression . . . 80
4.A Assorted results . . . 83
4.A.1 The long-run variance . . . 83
4.A.2 Gretl script for replicating Table 4.2 . . . 83

5 VAR processes **85**
5.1 Multivariate processes . . . 85
5.2 VAR processes . . . 87
5.3 VAR estimation . . . 93
5.3.1 An example . . . 94
5.4 Unit root VARs . . . 96
5.5 Usage of VARs . . . 99
5.5.1 Forecasting . . . 99
5.5.2 Causality analysis . . . 101
5.6 Dynamic Analysis and Structural VARs . . . 104
5.6.1 Triangular SVARs and the Cholesky Decomposition . . . 106
5.6.2 Long-run constraints . . . 110
5.6.3 Other approaches to identification . . . 112
5.A Assorted results . . . 114
5.A.1 Simultaneous equation systems . . . 114
5.A.2 Forecast error variance and confidence intervals . . . 115
5.A.3 Orthogonal matrices . . . 117

6 Cointegration **119**
6.1 Definitions . . . 119
6.2 Properties of cointegration vectors . . . 120
6.3 Error-correction models . . . 122
6.4 The Granger Representation Theorem . . . 126
6.4.1 Some matrix algebra . . . 127
6.4.2 The theorem itself . . . 128
6.4.3 Deterministic component . . . 130
6.5 Estimation techniques . . . 133
6.5.1 The Johansen method . . . 134
6.6 Alternative approaches . . . 138
6.7 An example . . . 140
6.A Assorted results . . . 143
6.A.1 Outline of the proof of the Granger Representation Theorem . . . 143
6.A.2 The Frisch-Waugh theorem . . . 145

7 Persistent volatility **147**
7.1 Stylised facts . . . 147
7.2 ARCH and GARCH processes . . . 150
7.2.1 ARCH processes . . . 151
7.2.2 Testing for volatility persistence . . . 153

7.2.3 GARCH processes 154
7.2.4 Estimation 155
7.2.5 An example 156
7.3 Extensions 159
7.3.1 Non-normal distributions 159
7.3.2 The GARCH zoo 160
7.4 Stochastic volatility 164
7.5 Realised volatility 165
7.6 Multivariate models 165

8 Further reading (not really updated very much) 167
8.1 In general 167
8.2 Univariate Stochastic Processes 168
8.3 VAR processes 168
8.4 I(1) processes and cointegration 168
8.5 Conditional heteroskedastic processes 169

Bibliography 171

List of Figures

1.1 US industrial production: percentage changes . . . 7
1.2 US industrial production: percentage changes — correlogram . 7
1.3 US inflation . . . 9
1.4 US Inflation – correlogram . . . 9
1.5 Nasdaq stock market index – daily returns . . . 10
1.6 Nasdaq stock market index – Correlogram . . . 10
1.7 Nasdaq stock market index – absolute values of daily returns . 11
1.8 Nasdaq stock market index – correlogram of absolute values . 11

2.1 MA(1) with different values of θ . . . 19
2.2 MA(1): Autocorrelation of order 1 as a function of θ . . . 20
2.3 AR(1) with different values of φ . . . 25
2.4 AR(2): $\varphi_1 = 1.8$; $\varphi_2 = -0.9$ . . . 26
2.5 Box and Jenkins' airline data . . . 30
2.6 Graphical representation of a complex number . . . 33

3.1 Impulse response function for $y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$ 51
3.2 USA industrial production (from 1919) . . . 52
3.3 Logarithm of US monthly industrial production . . . 53
3.4 Industrial production growth rate . . . 54
3.5 Industrial production growth rate — correlogram . . . 54
3.6 Impulse response functions . . . 58
3.7 Forecasts . . . 59

4.1 log(GDP) . . . 61
4.2 log(GDP) and deterministic trend . . . 62
4.3 Residuals . . . 62
4.4 Δlog(GDP) . . . 63
4.5 Random walk . . . 66
4.6 Random walk — trending properties . . . 67
4.7 Density function of the DF test . . . 72
4.8 Density function of the DF test with an intercept . . . 75
4.9 White noise with a break . . . 76

5.1 Eigenvalues of the companion matrix . . . 92
5.2 GDP and consumption in the Euro Area . . . 95

5.3 Non-structural impulse response functions 108
5.4 Structural impulse response functions 109

6.1 Stationary VAR(1): simulated time series 124
6.2 Bivariate random walk: simulated time series 125
6.3 Cointegrated process: simulated time series 125
6.4 Oil product spot prices 141
6.5 Oil product spot prices — scatterplots 141

7.1 Nasdaq composite index – logarithms 148
7.2 Nasdaq composite index – daily returns 148
7.3 Nasdaq composite index – absolute returns 149
7.4 Nasdaq returns – marginal distribution 150
7.5 Nasdaq index returns – residuals and estimated standard deviation 158
7.6 Nasdaq returns standardised residuals 159
7.7 Alternative Distributions 161

Chapter 1

Introduction

1.1 What is a stochastic process? What is it good for?

The data we use in econometrics generally fall into two categories: *cross-sectional*, when the available observations are relative to different individuals, or *time series*, when what we have are observations, on one or more variables, collected over time[1].

In the first case, thinking of a set of N observed data as one of the possible realisations of N independent and identical random variables is not too unsustainable a hypothesis: when recording the weight and height of N individuals, there is no reason to think that

1. the physical characteristics of the i-th individual are in some way connected to those of the other individuals (*independence*);

2. the relationship between weight and height that holds for the i-th individual is different from that which holds for all the others (*identity*).

In these cases, we use the concept of the realisation of a random variable as a metaphor for the i-th observation, and the appropriate inferential apparatus is no different from the elementary one, in which independence and identity allow us to write

$$f(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f(x_i),$$

that is, that the density function of our sample is simply the product of the density functions of the individual observations (which are all the same). When we use linear regression, this framework substantially leads us to the so-called "classical assumptions", extensively analysed at the beginning of any Econometrics course. Note that this type of reasoning is perfectly appropriate in the cases where the observed data come from a controlled experiment, of the type of those used by physicians or biologists.

[1] In fact, an intermediate case is given by the so-called *panel* data, but we do not deal with it here.

When data are time series, however, there is a basic conceptual difference that requires a revision of the probabilistic concepts to be used as a metaphor for the data: time has a direction, and history exists. Indeed, the natural tendency of many phenomena to evolve smoothly over time suggests that the data observed at a given time t is more similar to that recorded at time $t-1$ rather than in distant periods; in practice, it can be said that the time series we observe possesses some degree of "memory of itself". This characteristic is generally known as **persistence**, and it sets time series data apart from cross-sectional ones, because in the former the order of the data is of fundamental importance, while in the latter it is completely irrelevant.

The tool we use as a probabilistic metaphor for the observed time series is the **stochastic process**. A non-rigorous definition of a stochastic process, which is intuitive and substantially correct for our needs, can be the following: *a stochastic process is an infinitely long sequence of random variables* or, equivalently, a random vector of infinite dimension. A sample of T consecutive observations over time is therefore not conceived as a realisation of T distinct random variables, but rather as part of *a single* realisation of a stochastic process, whose memory is given by the degree of connection between the random variables that it contains.

1.2 Characteristics of Stochastic Processes

The definition above (where I cunningly avoided a series of technical complications) implies several properties of stochastic processes that are rather important for our work: given a stochastic process whose t-th element[2] is denoted by x_t,

- conceptually, it is possible to define a density function for the process $f(\ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$;
- it is possible to marginalise such a density function for any subset of its components; from this it follows that the marginal density functions are defined for each of the x_t, but also for any pair of elements (x_t, x_{t+1}), any triple and so on; furthermore, the fact that the observations may not be independent of one another implies that the sample density can no longer be written as a simple product of the marginals;
- if moments exist, it is possible to say, for example, that $E(x_t) = \mu_t$, $V(x_t) = \sigma_t^2$, $Cov(x_t, x_{t-k}) = \gamma_{k,t}$ and so forth;
- in the same way, it is possible to define conditional distributions, with objects like $f(x_t|x_{t-1})$, and possibly the relative moments.

The properties I just described refer to stochastic processes as probabilistic structures. However, when we want to use these structures in inferential procedures, two main problems arise:

[2] To be pedantic, we should use two different notations for the whole *stochastic process*, and each generic element in it. If the latter is denoted by x_t, the process to which it belongs should be written $\{x_t\}_{-\infty}^{+\infty}$. I consider this refinement superfluous, and I will use the same notation both for a process and for its t-th element; no confusion should arise.

1. if what we observe is just *one segment* of a *single* realisation of the many possible ones, the logical possibility of making inference on the process cannot be taken for granted: there is no way to tell which features of the observed time series are specific to *that* realisation, and which instead would recur even when observing other, equally possible, ones;

2. even if it were possible, somehow, to use a single realisation to perform statistical inference on the characteristics of the process, it is necessary for it to be stable over time. In other words, the probabilistic features of such a process must remain unchanged, at least relatively stable.

These two issues lead to the definition of two properties that stochastic processes may have or may not.

Stationarity We speak of a stationary stochastic process in two senses: **strong** (or **strict**) and **weak** stationarity.

To define strong stationarity, let us examine any possible subset of the random variables in the process; these do not necessarily have to be consecutive, but to help intuition, let us pretend that they are. Let us consider, therefore, a 'window' of length k on the process, that is a subset defined as $W_t^k = (x_t, \ldots, x_{t+k-1})$. Clearly, this is a k-dimensional random variable, with its own density function which, in general, may depend on t. If, however, it does not, then the distribution of W_t^k is equal to that of W_{t+1}^k, W_{t+2}^k and so on. We this invariance holds for any k, we speak of **strong stationarity**. In other words, when a process is strongly stationary, the distributional characteristics of all the marginals remain constant over time.

Weak stationarity, on the other hand, concerns only windows of length 2, $W_t^2 = (x_t, x_{t+1})$, and implies the further requisite that the first two moments exist. In this case, we don't require all the pairs $W_t^2 = (x_t, x_{t+1})$ to have the same distribution: we just want their means and covariances to be the same.[3]

Despite their names, one definition does not imply the other; for example, a process can be strongly stationary but not possess moments;[4] conversely, the constancy of the moments over time does not imply that the various marginals have the same distribution. In one case, however, the two definitions coincide: this case — which is of particularly importance in practice — is the one in which the process is **Gaussian**, that is, when the joint distribution of any subset of elements of the process is a multivariate normal. If a process is Gaussian, establishing that it is weak stationary is equivalent to establishing strong stationarity. Since Gaussianity is very often assumed in empirical applications, from an operational point of view the definition of weak stationarity is generally

[3] For this reason, weak stationarity is also defined as **covariance stationarity**.

[4] For example: think of a sequence of random variables $y_t = 1/x_t$, where the x_t are independent standard normal variables. The sequence of y_t is a sequence of identical and independent random variables without moments. It's easy to see that the requirements for strong stationarity are met, but the ones for weak stationarity are not.

adopted. When speaking of stationarity without adjectives, we mean "weak".

Limited memory Generally speaking, we will want to work with processes where elements "far apart" can be thought of as "essentially" independent random variables. That is, the information that x_{t-k} carries on the distribution of x_t is practically nil, if k is large.

The most general concept is called ergodicity: in a non-ergodic process, persistence is so strong that one realisation of the process, no matter how long, is insufficient to describe the process distribution. In an ergodic process, on the contrary, the memory of the process is weak over long horizons, and as the sample size increases, the information we get also increases significantly.

The formal conditions under which a stationary stochastic process is ergodic are too complex for a book like this one; I'll just say that, by observing the process for a "sufficiently" long period of time, it is possible to observe "almost all" the subsequences that the process is capable of generating. In other words, it can be said that, in an ergodic system, if something *can* happen, then sooner or later it *must* happen. The fact that events far apart in time can be considered independent for all practical purposes is then often summarised in the following property of ergodic processes (which is sometimes used as a *definition* of an ergodic process):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathrm{Cov}(x_t, x_{t-k}) = 0.$$

Consequently, if a process is ergodic, it is possible (at least in principle) to use the information contained in its observations over time to infer its characteristics. There is a theorem (appropriately called the 'ergodic theorem') which states that, if a process is ergodic, for inferential purposes the observation of a "long enough" realisation is equivalent to the observation of a large number of shorter, separate realisations. If, for example, an ergodic process x_t has an expected value of μ, then its average *over time* is a consistent estimator of μ (in formulas, $T^{-1}\sum_{t=1}^{T} x_t \xrightarrow{\mathrm{p}} \mu$), and therefore μ can be estimated consistently as if many realisations of the process are available instead of only one.

From a practical point of view, however, is much more common to rely on an alternative (and a bit more restrictive) concept, known as "mixing". The nice thing about it is that mixing processes can be shown to be ergodic, but mixing is much easier to understand and handle in practice.

There are several ways to define mixing processes. All of them rely on choosing two events A and B as follows: A is a statement pertaining on something that happened at time $t-m$ or before; B is a statement relative to something that happens at time t or after. Therefore, $P(A)$ is something you can compute using $f(\ldots, x_{t-m-1}, x_{t-m})$, while for $P(B)$ you need $f(x_t, x_{t+1}, \ldots)$. In other words, A is something that may have

happened in the past (a very distant past, if m is large). B is an event that may happen from now on.

Of course, the two events are independent if $P(A \cap B) = P(A)P(B)$. Therefore, consider the quantity

$$\alpha_m = |P(A \cap B) - P(A)P(B)|$$

If $\lim_{m\to\infty} \alpha_k = 0$ for all possible definitions of A and B, then whatever happens up to some point in time becomes irrelevant (or nearly so) after k periods (this particular definition is known as **strong mixing** — there are several more).

In general, it can be said that inference is possible only if the stochastic process we study is stationary and ergodic. It must be said, moreover, that while there are methods to test the hypothesis of non-stationarity (at least in certain contexts, which we will examine later on), the hypothesis of ergodicity is not testable if all we observe is one realisation of the process is available, no matter how long.

1.3 Moments

By definition, for weakly stationary processes every element of the process x_t has a finite and constant expectation $E(x_t) = \mu$ and a finite and constant variance $V(x_t) = \sigma^2$. Furthermore, all covariances between different elements

$$\gamma_k = E\left[(x_t - \mu)(x_{t-k} - \mu)\right] \tag{1.1}$$

exist. These are known as **autocovariances**. Stationarity guarantees that these quantities are not functions of t; however, they are functions of k, so we can define the "autocovariance function" as a function of k such that $\gamma(k) = \gamma_k$. Clearly, the autocovariance of order 0 is nothing but the variance. Furthermore, the definition is such that $\gamma_k = \gamma_{-k}$, since

$$E\left[(x_t - \mu)(x_{t-k} - \mu)\right] = E\left[(x_t - \mu)(x_{t+k} - \mu)\right].$$

Autocorrelations are trivially defined as

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma^2}, \tag{1.2}$$

and obviously, $\rho_0 = 1$. These quantities describe the memory of the process: if $\gamma_1 \neq 0$, then we have that

$$f(x_t|x_{t-1}) \neq f(x_t)$$

and consequently

$$E(x_t|x_{t-1}) \neq E(x_t), \tag{1.3}$$

which can be translated as: if x_{t-1} is known, the expected value of x_t is not the same as what we would expect if x_{t-1} were unknown. We could also extend the set of random variables on which we condition to x_{t-2}, x_{t-3}, and so on. This set of random variables is sometimes called the **information set** at time $t-1$, and is denoted by $\mathcal{F}_{t-1}$.

Strictly speaking, the exact definition of an information set is a bit more complex: we would need to deal with σ-algebras and be more rigorous about what is meant by conditioning. The topic is truly fascinating, but is way beyond the scope of this book. It is not difficult, however, to think of the information set as a collection of random variables with respect to which conditioning is possible. In a time series setting, it is reasonable to assume that the past is known; consequently, it makes sense to condition a random variable at time t on its own past values, because if x_t is known, then so are x_{t-1}, x_{t-2}, and so on. Naturally, the information set at time t may also include variables other than the one we are conditioning: for example, in the theory of rational expectations, the idea of an information set at time t is used as a synonym for *all* available information about the world at time t.

After observing a realisation of size T of a stochastic process x_t, we can define the sample counterparts of the theoretical moments:

$$\begin{aligned} \text{sample mean} \quad & \hat{\mu} = T^{-1} \textstyle\sum_{t=1}^{T} x_t \\ \text{sample variance} \quad & \hat{\sigma}^2 = T^{-1} \textstyle\sum_{t=1}^{T} (x_t - \hat{\mu})^2 \\ \text{sample autocovariance} \quad & \hat{\gamma}_k = T^{-1} \textstyle\sum_{t=k}^{T} (x_t - \hat{\mu})(x_{t-k} - \hat{\mu}) \end{aligned}$$

If the process is stationary and ergodic, it can be shown that these quantities are consistent estimators of the moments of the process[5] However, these statistics are also useful as descriptive statistics. Apart from the sample mean, whose interpretation I take for granted, analysing the autocorrelation function often allows us to make interesting qualitative assessments about the persistence characteristics of the series under consideration. In the next subsection, I will give a few examples.

1.4 Some examples

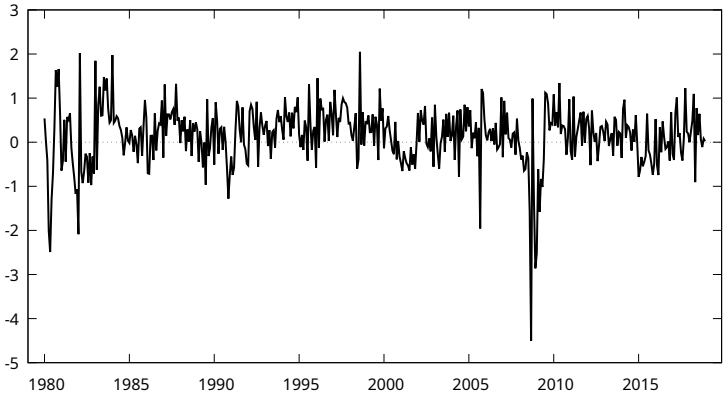
In this section, I'll use a few real-world time series to illustrate the concepts described above. We observe a time series and assume that it can be considered a realisation of a stationary stochastic process; our aim is to find the process that represents this series "best" (whatever that may mean). More precisely, we will look for a stochastic process whose realisation are "qualitatively most similar" to our observed series.

Let us consider, for example, the time series shown in Figure 1.1, which reports the monthly percentage change in the industrial production index from the previous month for the United States, from 1980 to 2018. As can be seen, the series fluctuates fairly regularly around a central value, approximately between -5% and 3% (to be precise, its average is about 0.15%). If we were willing to assume that the process that generated these data is stationary and ergodic, we could say that this average is an estimate of the unconditional expected value of the process.

But is this process (provided that it exists) stationary? And if so, is it also ergodic? More generally: what are its persistence characteristics? It's difficult

[5]The careful reader will have noticed the absence of the 'degrees of freedom correction': in the denominator of the sample variance, for example, there is T instead of $T-1$. Asymptotically, the two formulations are obviously equivalent. What happens in finite samples is usually considered irrelevant or too complicated to be studied.

Figure 1.1: US industrial production: percentage changes



to give an answer just by looking at the graph, at least for the untrained eye. The analysis of the sample autocorrelations, reported in Figure 1.2, comes to the rescue.

Figure 1.2: US industrial production: percentage changes — correlogram

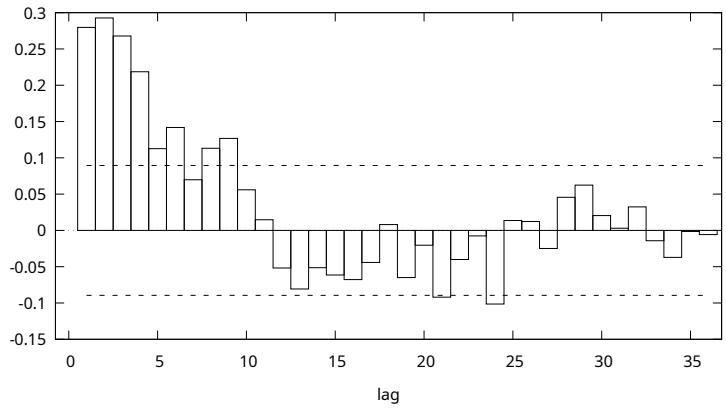


Figure 1.2 shows the **correlogram**, a histogram in which each bar shows the value of the autocorrelation ρ_k against the order k. For example, the correlogram for the industrial production series y_t, tells us that ρ_1, the correlation between y_t and y_{t-1}, is about 28.3%, ρ_2, the one between y_t and y_{t-2} is 29.5%, and so on.

In an inferential setting, we would ask ourselves whether these statistics (the autocorrelation of the process ρ_k) that are significantly different from 0, and proceed with some kind of hypothesis test. However, we can simply take them as descriptive statistics, with a clear meaning: nearby observations are positively correlated, and therefore they are unlikely to be independent, so there is clear evidence of some persistence.

Moreover, persistence seems to fade over time: one could say that, as the distance between observations increases, the absolute value of their correlation

(which we can, at this stage, take as a persistence indicator) tends to decrease: at a distance of 18 months, their correlation is considerably smaller (1.6%). All in all, one could say that from a qualitative point of view, this is exactly what we expect to see in a realisation of a stationary and ergodic process: a persistence that substantially affects the series in the short run, but remains an essentially 'local' phenomenon.

Hence, one might ask whether the time series we are observing can be statistically modelled by studying its conditional mean, just as is done in a linear regression model. Indeed, if in a linear model the equation $y_t = x_t'\beta + \epsilon_t$ splits the dependent variable into a conditional mean plus an error term, nothing prevents us from making the conditional mean a function of the information set $\mathcal{F}_{t-1}$, so we would use OLS on a model like

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \epsilon_t. \tag{1.4}$$

If we did so, using, for example, the values up to four months prior as the conditioning set, we would obtain the results shown in Table 1.1.

Table 1.1: OLS estimation of equation (1.4)

Coefficient	coefficient	std. error	t-statistic	p-value
β_0	0.0631	0.0299	2.109	0.0355
β_1	0.1685	0.0463	3.640	0.0003
β_2	0.1817	0.0465	3.907	0.0001
β_3	0.1369	0.0464	2.948	0.0034
β_4	0.0881	0.0462	1.906	0.0572

Mean dependent var	0.149175	S.D. dependent var	0.666875
Sum squared resid	175.0909	S.E. of regression	0.614952
R^2	0.156941	Adjusted R^2	0.149658

The numbers in Table 1.1 can be taken as simple descriptive statistic, but also as estimators of *something*. In this case, then, their properties must necessarily be studied within an appropriate inferential framework. It is precisely for this reason that we need to study stochastic processes: to give a probabilistic meaning, if possible, to statistics like the ones we have just seen. In the following chapters, I will show how and why estimates like these actually makes sense, and how they should be interpreted.

Now let's turn to another example, where things are less straightforward: Figure 1.3 reports the time series of inflation (defined as the annual percentage change in the consumer price index), again for the US.

Are we sure that a time series like this can be generated by a stationary process? As can be seen, periods (sometimes rather long ones) of high and low inflation alternate. Is it legitimate to think that the hypothetical process generating this series has a constant mean, as required for stationarity? Moreover: the correlogram, displayed in Figure 1.4, shows that treating persistence as a short-run phenomenon is decidedly more questionable. The autocorrelation corresponding to the 24th month is 48.1%, and the plot shows no sign of significant decay.

Highly persistent time series, like this one, are extremely common in economics and finance; to analyse them, one must find some way to represent

Figure 1.3: US inflation

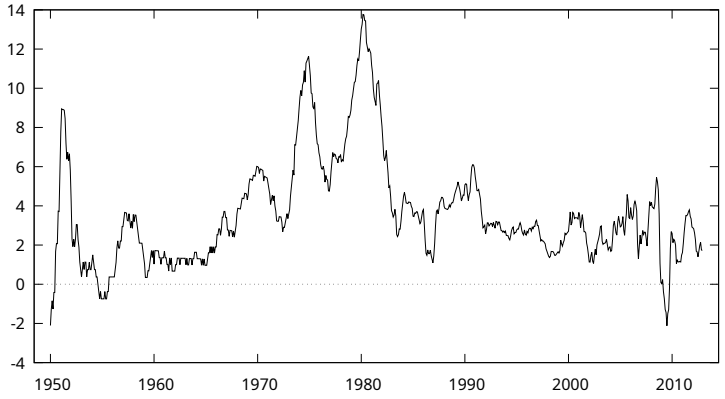
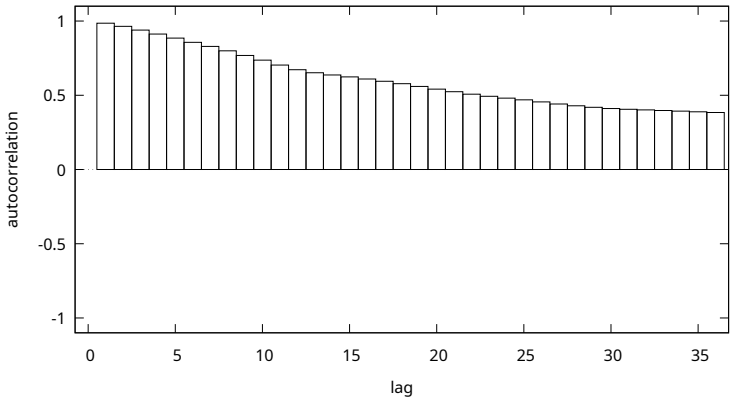
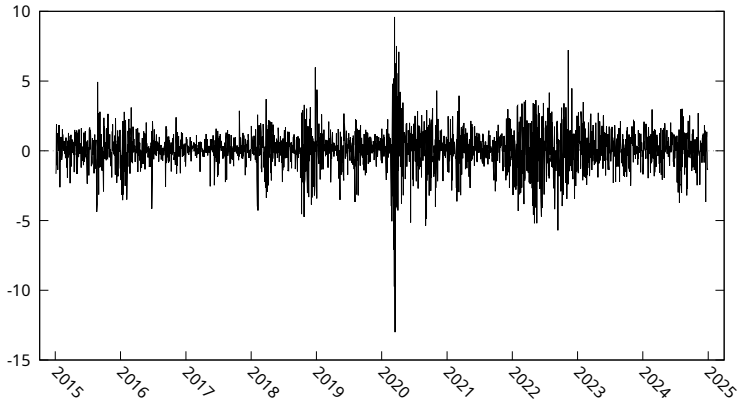


Figure 1.4: US Inflation – correlogram



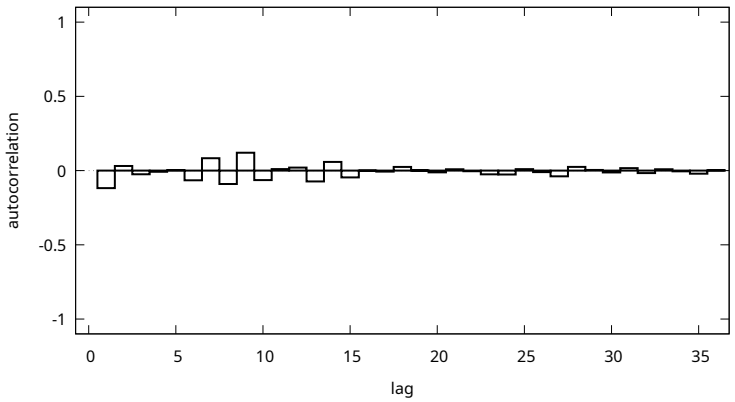
them as realisations of stationary processes, possibly after some suitable modification. In many cases, this can be done using techniques that dominated time series econometrics in the last two decades of the 20th century. These techniques, however, are somewhat complex and I'll defer their treatment to Chapter 4. Please, be patient.

Figure 1.5: Nasdaq stock market index – daily returns



I conclude this overview of examples with a quite different case: the (daily) percentage change in the Nasdaq stock index between 2015 and 2024, shown in Figure 1.5. The behaviour of this time series is — obviously — very different from the ones shown earlier: the data fluctuate around a value slightly above zero (the arithmetic mean is 0.061% — in other words, the Nasdaq index grew on average 0.061% per day over that period), and without the long fluctuations that characterised the industrial production or inflation series. This impression is confirmed by visual inspection of the correlogram in Figure 1.6.

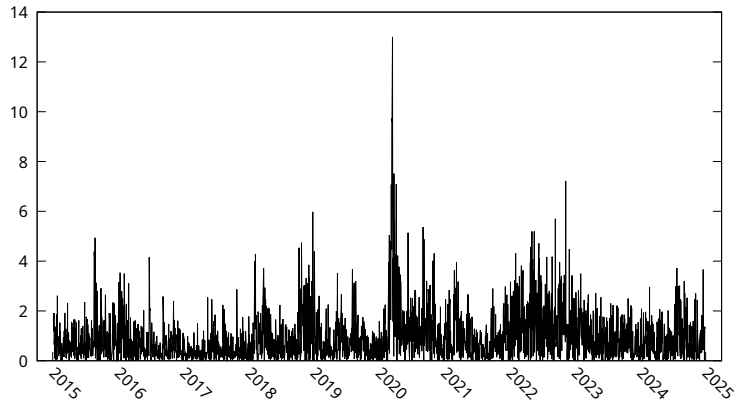
Figure 1.6: Nasdaq stock market index – Correlogram



We see very little persistence here. But then, there's a very good reason for this: no matter what technical analysis fans may say here, if there were a

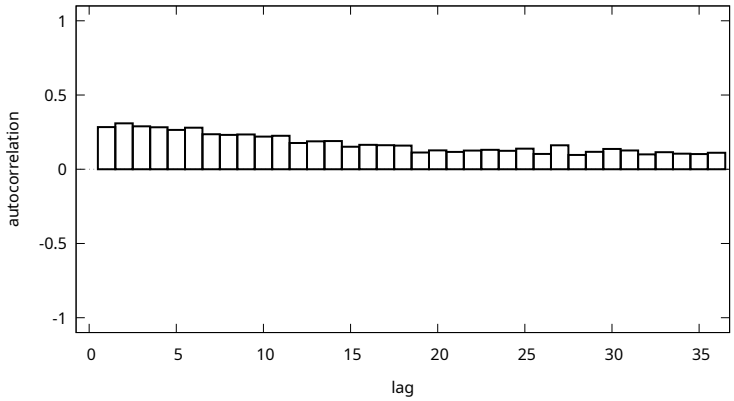
'simple' rule linking returns to their past values, any idiot could follow that rule and make unlimited profits.[6]

Figure 1.7: Nasdaq stock market index – absolute values of daily returns



And yet, we have something interesting here too: in fact, the plot in Figure 1.5 is typical of almost all high-frequency financial time series. The most interesting feature is that periods when absolute returns are larger alternate with periods when they are smaller. Technically, these periods are referred to as high-volatility or low-volatility regimes, respectively. Unsurprisingly, the high-volatility episodes coincide with events that had a major impact on financial markets: for example, the COVID pandemic and the start of the Ukraine war. This feature is clearly visible by considering the time series of the absolute values of the returns (Figure 1.7).

Figure 1.8: Nasdaq stock market index – correlogram of absolute values



Evidently, we have quite a lot of persistence here. In this case, the aim is not so much to model the persistence of the series itself, but rather that of its

[6]Experts will notice that I am oversimplifying the so-called 'efficient market hypothesis' (EMH) with a recklessness you'd only expect in a random YouTube video. My apologies.

volatility.

The statistical concept into which the word 'volatility' translates is, clearly, the variance (provided that the second moments exist). As will be seen later, to analyse series of this type, we use stochastic processes of a specific nature, in which any persistence exhibited by the series translates into the time varying dependence of the variance, rather than the mean. In other words, the persistence of these processes is summarised by the fact that

$$V(x_t|x_{t-1}) \neq V(x_t). \tag{1.5}$$

Compare this with equation (1.3): in these processes, which are called **conditionally heteroskedastic processes**, what makes the difference between the marginal and conditional distributions with respect to the information set $\mathcal{F}_{t-1}$ is precisely the structure of the second moments, rather than the first moments. Processes of this type have been widely used in advanced empirical finance for nearly four decades.

By now, our rationale should be clear enough. In the following chapter, we will introduce a class of stochastic processes which is the cornerstone of all time series econometrics: ARMA processes.

Chapter 2

ARMA processes

ARMA processes are the cornerstone of time series econometrics, for theoretical as well as practical reasons, as we'll see. To describe the main characteristics of these processes, a few basic definitions are necessary.

2.1 The lag operator

Time series are nothing but sequences of numbers, with a natural ordering given by time. In the same way, stochastic processes are sequences of random variables. Therefore, it is natural to look for a general way to manipulate sequences by means of appropriate operators. The **lag operator** is generally denoted by the letter L by econometricians (statisticians use B instead — savages!); it's an operator that turns a sequence x_t into another sequence, that contains the same objects as x_t, but shifted back by one period.[1] If you apply L to a constant, the result is the same constant. In formulae,

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

Repeated application of the L operator n times is indicated by L^n, and therefore $L^n x_t = x_{t-n}$. By convention, $L^0 = 1$. The L operator is *linear*, which means that, if a and b are constant, then $L(ax_t + b) = aLx_t + b = ax_{t-1} + b$. These simple properties have the nice consequence that, in many cases, we can manipulate the L operator algebraically as if it was a number. Such a trick proves especially useful when dealing with *polynomials* in L, that is operators of the form

$$a_0 + a_1 L + a_2 L^2 + \ldots a_p L^p,$$

where the maximum exponent p is known as the "order" of the polynomial (also known as its "degree"). Polynomials in the lag operators are also known as **filters**, so we speak of "filtered" sequences.

Allow me to exemplify:

[1] In certain cases, you might want to use the "lead operator", usually notated as F, which is defined as the inverse to the lag operator ($Fx_t = x_{t+1}$, or $F = L^{-1}$). I'm not using it in this book, but its usage is very common in economic models with rational expectations.

Example 2.1.1 *Call b_t the money you have at time t, and s_t the difference between the money you earn and the money you spend between $t-1$ and t (in other words, your savings). Of course,*

$$b_t = b_{t-1} + s_t.$$

Now the same thing with the lag operator:

$$b_t = Lb_t + s_t \to b_t - Lb_t = (1-L)b_t = \Delta b_t = s_t$$

The Δ operator,[2] which I suppose not unknown to the reader, is defined *as $(1-L)$, that is a polynomial in L of order 1. The above expression simply says that the variation in the money you have is your net saving.*

Example 2.1.2 *Call q_t the GDP for the Kingdom of Verduria in quarter t. Obviously, yearly GDP is given by*

$$y_t = q_t + q_{t-1} + q_{t-2} + q_{t-3} = (1 + L + L^2 + L^3)q_t.$$

Since $(1 + x + x^2 + x^3)(1 - x) = (1 - x^4)$, if you "multiply" the equation above[3] by $(1 - L)$ you get

$$\Delta y_t = (1 - L^4)q_t = q_t - q_{t-4};$$

The variation in yearly GDP between quarters is just the difference between the quarterly figures a year apart from each other.

A polynomial $P(x)$ may be evaluated at any value, but two cases are of special interest. Obviously, if you evaluate $P(x)$ at $x = 0$ you get the "constant" coefficient of the polynomial, since $P(0) = p_0 + p_1 \cdot 0 + p_2 \cdot 0 + \cdots = p_0$; instead, if you evaluate $P(1)$ you get the sum of the polynomial coefficients:

$$P(1) = \sum_{j=0}^{n} p_j 1^j = \sum_{j=0}^{n} p_j.$$

This turns out to be quite handy when you apply a lag polynomial to a constant, since

$$P(L)\mu = \sum_{j=0}^{n} p_j \mu = \mu \sum_{j=0}^{n} p_j = P(1)\mu.$$

In many cases, we will want to invert polynomials. More specifically, given a polynomial $P(x)$ we will look for a polynomial $Q(x)$ (possibly, of infinite order), such that

$$P(x)Q(x) = 1.$$

If such a polynomial exists, we will say that $Q(x)$ is the inverse of $P(x)$ (and vice versa) and write

$$Q(x) = q_0 + q_1 x + q_2 x^2 + \cdots = \frac{1}{P(x)}.$$

[2] Statisticians who use B instead of L also use ∇ instead of Δ.

[3] To be precise, we should say: 'if you apply the $(1 - L)$ operator to the expression above'.

Two results on polynomial inversion are often quite useful: the first one has to do with inverting polynomials of order 1. It can be proven that, if $|\alpha| < 1$,

$$(1 - \alpha L)^{-1} = (1 + \alpha L + \alpha^2 L^2 + \cdots) = \sum_{i=0}^{\infty} \alpha^i L^i; \tag{2.1}$$

the other one is that a polynomial $P(x)$ is invertible if and only if all its roots are greater than one in absolute value:

$$\frac{1}{P(x)} \quad \text{exists iff} \quad P(x) = 0 \Rightarrow |x| > 1. \tag{2.2}$$

The proofs are in subsection 2.A.1.

Example 2.1.3 (The Keynesian multiplier) *Let me illustrate a possible use of polynomial manipulation by a very old-school macro example: the simplest possible version of the Keynesian multiplier idea. Suppose that*

$$Y_t = C_t + I_t; \tag{2.3}$$
$$C_t = \alpha Y_{t-1}; \tag{2.4}$$

where Y_t is GDP, C_t is aggregate consumption and I_t is investment; $0 < \alpha < 1$ is the marginal propensity to consume; note that, in this version of the Keynesian model, equation (2.4) defines consumption as a fraction of past *income, as opposed to* current *income.*

By combining the two equations,

$$Y_t = \alpha Y_{t-1} + I_t \rightarrow (1 - \alpha L) Y_t = I_t.$$

therefore, by applying the first degree polynomial $A(L) = (1 - \alpha L)$ to the Y_t sequence (national income), you get the time series for investments, simply because $I_t = Y_t - C_t = Y_t - \alpha Y_{t-1}$.

Now invert the $A(L) = (1 - \alpha L)$ operator and use equation (2.1)

$$Y_t = \frac{1}{1 - \alpha L} I_i = (1 + \alpha L + \alpha^2 L^2 + \cdots) I_t = \sum_{i=0}^{\infty} \alpha^i I_{t-i} :$$

aggregate demand at time t can be seen as a weighted sum of past and present investment. Suppose that investment goes from 0 to 1 at time 0. This brings about a unit increase in GDP via equation (2.3); but then, at time 1 consumption goes up by α, by force of equation (2.4), so at time 2 it increases by α^2 and so on. Since $0 < \alpha < 1$, $a^n \rightarrow 0$ and this effect eventually dies out.

If investments were constant through time, then $I_t = \bar{I}$; therefore, $A(L) Y_t = \bar{I}$ becomes

$$Y_t = \frac{1}{A(L)} \bar{I} = \frac{1}{A(1)} \bar{I} = \frac{\bar{I}}{1 - \alpha}$$

where the second equality comes from the fact that $\bar{I}$ is constant. The rightmost expression is nothing but the familiar "Keynesian multiplier" formula.

A word of caution: in many cases, it's OK to manipulate L algebraically as if it was a number, but sometimes it's not: the reader should always keep in mind that the expression Lx_t does not mean 'L times x_t', but 'L applied to x_t'. The following example should, hopefully, convince you.

Example 2.1.4 *Given two sequence x_t and y_t, define the sequence z_t as $z_t = x_t \cdot y_t$. Obviously, $z_{t-1} = x_{t-1}y_{t-1}$; however, one may be tempted to argue that*

$$z_{t-1} = x_{t-1}y_{t-1} = Lx_tLy_t = L^2x_ty_t = L^2z_t = z_{t-2}$$

which is obviously nonsensical (it may be a good exercise for the reader to find which of the equalities above is wrong).

2.2 White noise processes

The **white noise** (WN) process is the simplest stochastic process there is:[4] it possesses moments of (at least) up to the second order, and these are constant over time (hence the process is stationary); however, they give the process no memory of its own past.

More formally: a white noise process, whose t-th element will be denoted by ϵ_t, exhibits the following characteristics:

$$E(\epsilon_t) = 0 \tag{2.5}$$

$$E(\epsilon_t^2) = V(\epsilon_t) = \sigma^2 \tag{2.6}$$

$$\gamma_k = 0 \quad \text{for } |k| > 0 \tag{2.7}$$

A white noise process is therefore, in essence, a sequence containing infinitely many random variables, all with zero mean and constant variance; furthermore, these random variables are all uncorrelated with one another. Strictly speaking, this does not mean that they are independent. If, however, the process is a *Gaussian* white noise, where the joint distribution of all pairs $(\epsilon_t, \epsilon_{t+k})$ is a bivariate normal, then they are. Two aspects deserve special attention:

- in the case of normality, a realisation of size N of a white noise can also be quite legitimately considered a realisation of N independent and identically distributed random variables. In this sense, a cross-sectional sample can be viewed as a special case.
- There is no substantial difference between the conditions that define a white noise and the so-called 'classical assumptions' on the disturbance in the OLS model, except for the lack of correlation between regressors and disturbances; the OLS classical assumptions could be stated as 'the disturbance term is a white noise uncorrelated with the regressors'.

A white noise process, therefore, is a stochastic process with no persistence. As such, one might think it inadequate for achieving the goal we set ourselves

[4] The reason why this process bears the evocative name of 'white noise' would be very interesting, but the analytical tools we use in this book are too limited to develop this point. Oh well.

in the introduction, namely finding a probabilistic structure that can be used as a metaphor for persistent time series. The decisive step forward, described in the next section, lies in considering what happens when we filter a white noise process via a lag polynomial.

If I had aimed at impeccable rigour, I should have made a distinction between different types of 'memoryless' stochastic processes. Strictly speaking, in fact, the only type of process with no trace of persistence is whose elements are independent random variables. Often, however, it is useful to consider processes that are not so restrictive: for example, the so-called *martingale difference*, which is a concept very commonly employed both in statistics (especially in asymptotic theory), in economics (theory of rational expectations) and finance (asset pricing).

In a martingale difference sequence (MDS), the distribution is left unspecified; what characterises this type of sequence is the property $E(x_t|\mathcal{F}_{t-1}) = 0$. In this context, our primary interest is the conditional mean of the process, which must not depend in any way on the past.

A white noise, on the other hand, is a different concept altogether: the lack of correlation between different elements ensures only that the conditional expectation is not a *linear* function of the past. Zipped proof:

$$\begin{aligned} E(x_t|\mathcal{F}_{t-1}) = bx_{t-1} \quad &\Longrightarrow \\ E(x_t x_{t-1}|\mathcal{F}_{t-1}) = bx_{t-1}^2 \quad &\Longrightarrow \\ E(x_t x_{t-1}) = E\left[E(x_t x_{t-1}|\mathcal{F}_{t-1})\right] \quad &= \\ = bE(x_{t-1}^2) \neq 0 & \end{aligned}$$

(we thank the law of iterated expectations for its kind cooperation).

However, the conditional could well be a non-zero non-linear function. In fact, one can construct examples of white noise processes that are not martingale difference sequences. Furthermore, not all martingale difference sequences are white noise processes: indeed, the definition of white noise entails very specific conditions on the second moments, which may not even exist in a MDS.

That being said, it must be noted that in practice these concepts can be overlapped fairly painlessly: a Gaussian WN, for example, is a sequence of independent random variables with a mean of 0, so it is also a MDS.

2.3 MA processes

A **moving average** process, or **MA process** for short, is a sequence of random variables that can be written in the form

$$y_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i} = C(L)\epsilon_t,$$

where $C(L)$ is a q-order lag polynomial and ϵ_t is a white noise process. In most cases, we set $C(0) = \theta_0 = 1$. If $C(L)$ is a polynomial of degree q, y_t is also said to be an MA(q) process, which is read as 'MA process of order q'.

Let us examine its moments: for the first moment, we have

$$E(y_t) = E\left[\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right] = \sum_{i=0}^{q} \theta_i E(\epsilon_{t-i}) = 0.$$

Therefore, an MA process has mean 0. At first glance, one might think that this characteristic is a serious shortcoming for applying MA processes to real-world situations, since in general observed time series do not necessarily fluctuate around the value 0. However, this issue can be easily dealt with: for any process x_t for which $E(x_t) = \mu_t$, one can always define a new zero-mean

process $y_t = x_t - \mu_t$.[5] If y_t is covariance-stationary, then all you have to do is consider y_t and then add back the mean to obtain x_t.

As for the variance, the fact that the $E(y_t) = 0$ makes it possible to express it as the second moment, namely

$$V(y_t) = E(y_t^2) = E\left[\left(\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right)^2\right]$$

By expanding the square,[6] the sum can be split into two distinct parts:

$$\left(\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right)^2 = \sum_{i=0}^{q} \theta_i^2 \epsilon_{t-i}^2 + \sum_{i=0}^{q} \sum_{j \neq i} \theta_i \theta_j \epsilon_{t-i} \epsilon_{t-j}$$

It should be obvious, from the "no-memory" property of white noise processes, that the expected value of the second summation is zero, so that

$$V(y_t) = E(y_t^2) = E\left[\sum_{i=0}^{q} \theta_i^2 \epsilon_{t-i}^2\right] = \sum_{i=0}^{q} \theta_i^2 E(\epsilon_{t-i}^2) = \sum_{i=0}^{q} \theta_i^2 \sigma^2 = \sigma^2 \sum_{i=0}^{q} \theta_i^2; \quad (2.8)$$

if q is a finite number, then $\sum_{i=0}^{q} \theta_i^2 < \infty$ and $V(y_t)$ is finite.

Finally, the autocovariances can be worked out by a completely analogous argument: the k-order autocovariance is

$$\begin{aligned} E(y_t y_{t+k}) &= E\left[\left(\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right)\left(\sum_{j=0}^{q} \theta_j \epsilon_{t-j+k}\right)\right] = \\ &= \sum_{i=0}^{q} \theta_i \left[\sum_{j=0}^{q} \theta_j E(\epsilon_{t-i} \epsilon_{t-j+k})\right]. \end{aligned} \quad (2.9)$$

By using, again, the properties of white noise, we get $E(\epsilon_{t-i}\epsilon_{t-j+k}) = \sigma^2$ for $j = i + k$ and 0 in all other cases, so that the previous expression reduces to:

$$\gamma_k = E(y_t y_{t+k}) = \sigma^2 \sum_{i=0}^{q} \theta_i \theta_{i+k},$$

where $\theta_i = 0$ for $i > q$.

Note that:

- the equation for the variance is a special case of the previous formula, setting $k = 0$;
- for $k > q$, all autocovariances are zero.

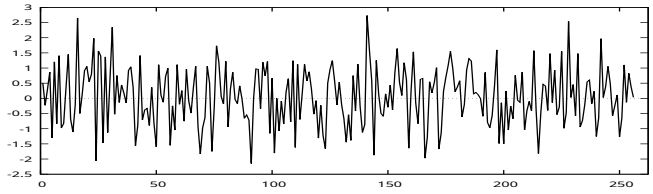
An MA(q) process, therefore, is a process obtained by combining different elements of the same white noise sequence. The higher its order, the longer is its persistence. Note that the order q may be infinite; in this case, however, the existence of the second moments (and therefore stationarity) is guaranteed only if $\sum_{i=0}^{\infty} \theta_i^2 < \infty$, a condition known as **square-summability**.

[5] Note in passing that in this simple example the process x_t is not stationary, according to the definition we have already provided, unless μ_t is constant, but the process y_t is.

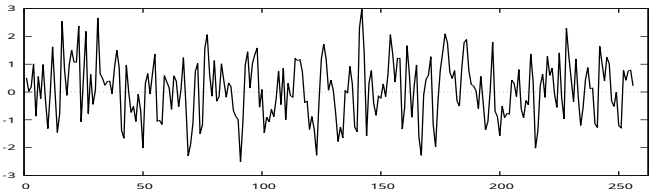
[6] Note: by the argument I made a few pages back, $[C(L)\epsilon_t]^2$ is *different* from $C(L)^2\epsilon_t^2$. Think of the simple case $C(L) = L$ and you will be immediately convinced.

Figure 2.1: MA(1) with different values of θ

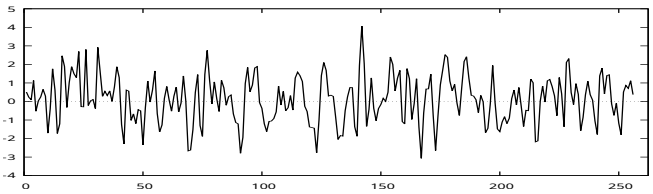
(a) $\theta = 0$ (white noise)



(b) $\theta = 0.5$



(c) $\theta = 0.9$



Example 2.3.1 *Consider the MA(1) process $x_t = \epsilon_t + \theta\epsilon_{t-1}$ and calculate its autocovariances: its variance is given by*

$$E(x_t^2) = E\left(\epsilon_t + \theta\epsilon_{t-1}\right)^2 = E(\epsilon_t^2) + \theta^2 E(\epsilon_{t-1}^2) + 2\theta E(\epsilon_t\epsilon_{t-1}) = (1+\theta^2)\sigma^2.$$

By its definition, the autocovariance of order 1 is

$$E(x_t x_{t-1}) = E\left[\left(\epsilon_t + \theta\epsilon_{t-1}\right)\left(\epsilon_{t-1} + \theta\epsilon_{t-2}\right)\right].$$

By expanding the product, we obtain:

$$E(\epsilon_t\epsilon_{t-1}) + \theta E(\epsilon_t\epsilon_{t-2}) + \theta E(\epsilon_{t-1}^2) + \theta^2 E(\epsilon_{t-1}\epsilon_{t-2}) = \theta\sigma^2.$$

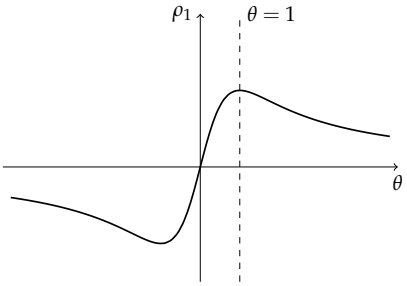
Consequently, the autocorrelation of order 1 is

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2}.$$

Similarly, it can be shown that the higher-order autocovariances are all zero.

To illustrate this more concretely, let us take an example of an MA(1) process and plot it: if the process is $y_t = \epsilon_t + \theta\epsilon_{t-1}$, the behaviour of y_t for different values of θ is shown in Figure 2.1. Of course, when $\theta = 0$ (as in Figure 2.1(a)) the process is white noise. As can be seen, as θ increases, the persistence characteristics become more visible (the path becomes 'smoother')

Figure 2.2: MA(1): Autocorrelation of order 1 as a function of θ



and its variance (roughly measured by the order of magnitude of the ordinates) increases. If we had simulated a higher-order MA process, this would have been even more evident.

Some interesting considerations can be made by digging a little deeper. As the example shows, the autocorrelation of order 1 of an MA(1) process is given by the formula $\rho_1 = \frac{\theta}{1+\theta^2}$, plotted in Figure 2.2. It can be noted that the maximum value reached by ρ_1 is 0.5, corresponding to $\theta = 1$; an analogous argument, with inverted signs, holds for the minimum. Furthermore, we know from previous arguments (see equation (2.9)), that all autocorrelations of order greater than 1 are zero. This means that the correlogram of an MA(1) process has only one interesting spike (the first one), and even that is constrained to lie between $-1/2$ and $1/2$.

Let's now turn to an inferential problem: suppose we were certain that an observed time series is the realisation of an MA(1) process. Which statistic could we use to obtain estimates of the process parameters (in this case, the parameter θ)? Obviously, this procedure would be justifiable only if the empirical correlogram had moderate values for the first-order autocorrelation and negligible ones for the others. If this were the case, we could adopt the following strategy:

- if the process that generated the data is in fact an MA(1), then,
- it is stationary and ergodic, then,
- the sample autocorrelation converges in probability to the theoretical one, then,
- in an infinitely-sized sample, $\hat{\rho}_1$ should be exactly equal to ρ_1.

More formally,

$$\hat{\rho}_1 \xrightarrow{\text{P}} \rho_1 = \frac{\theta}{1+\theta^2};$$

since this is a continuous function of θ, it can be inverted to give a consistent estimator of θ via the method of moments. This amounts to finding the value $\hat{\theta}$ which satisfies the equation

$$\hat{\rho}_1 = \frac{\hat{\theta}}{1+\hat{\theta}^2}; \tag{2.10}$$

It can be easily seen that the solution to (2.10) is[7]

$$\hat{\theta} = \frac{1}{2\hat{\rho}_1}\left(1 - \sqrt{1 - 4\hat{\rho}_1^2}\right).$$

Note that, for the existence of the estimator, it is necessary that $|\hat{\rho}_1| \leq 0.5$, but in this case there is no problem, since we are assuming that we are dealing with a series in which the first-order autocorrelation is not too strong.

For example: if the sample autocorrelation was 0.4, and I was convinced that the process that generated the data is an MA(1), then I would as my estimate of θ the number such that the theoretical autocorrelation is also 0.4, namely $\hat{\theta} = 0.5$. Clearly, this strategy is perfectly justified because the time series actually has the required covariance characteristics, *i.e.* a not-too-large first-order autocorrelation and negligible subsequent autocorrelations.

Now, we know that things are not always this easy: just take a look at Figures 1.2 on page 7 and 1.4 on page 9. It is true, however, that a higher-order MA process has more complex autocovariances, and therefore one may conjecture that the same strategy could be feasible, at least in theory, provided that a sufficiently high order of the polynomial $C(L)$ is specified.

In fact, one might wonder whether the conjecture holds for any autocovariance structure. The answer lies in the never-sufficiently-celebrated **Wold representation theorem**, of which I provide only the statement.

Theorem 1 (Wold representation theorem) *For any covariance-stationary, 0-mean stochastic process y_t, it is always possible to find a sequence (not necessarily finite) of coefficients θ_i such that*

$$y_t = \sum_{i=0}^{\infty} \theta_i \epsilon_{t-i},$$

where ϵ_t is white noise. Moreover, the sum of the squared coefficients is finite: $\sum_{i=0}^{\infty} \theta_i^2 < \infty$.

In other words: any stationary stochastic process can be represented as a moving average process with the original autocovariance structure. This result is of enormous importance: it tells us, essentially, that whatever the 'true' form of a stationary stochastic process, we can always represent it as an MA process. The process may be of infinite order, but square-summability ensures that the effect of the very distant past should not be very important for the present. It is for this reason that, by studying MA processes, we are in fact studying *all* possible stationary processes, at least as far as their mean and covariance characteristics are concerned.

The picture I have just given of Wold's theorem is not really accurate: if you look at 'proper' textbooks, you will notice that in fact the theorem does not say exactly what I wrote above. To be precise, one should say that every second-order stationary process can be decomposed into a 'deterministic' part (i.e. perfectly predictable given the past) plus a moving average part.

[7] In fact, there are two values, because the actual solution would be $\hat{\theta} = \frac{1 \pm \sqrt{1 - 4\hat{\rho}_1^2}}{2\hat{\rho}_1}$ (note the $\pm$ symbol), but to keep things simple let's agree on choosing the solution inside the interval $[-1, 1]$, *i.e.* the one reported in the text.

I have cunningly simplified the definition by adding the words 'with a mean of 0', thus excluding the existence of the deterministic part but the message remains essentially the same.

2.4 AR processes

AR (AutoRegressive) processes form another important class. These processes provide representation of a persistent time series that some may find more intuitive than the one implied by MA processes, because in AR processes the main idea is that the value of the series at time t is a linear function of its own past values, plus white noise. The name derives precisely from the fact that an AR process can be represented in the same way as a regression model, where the explanatory variables are the past values of the dependent variable:

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t. \tag{2.11}$$

Note that, in this context, the white noise process ϵ_t is like the disturbance term in a regression model, that is, as the difference between y_t and its conditional mean; in this case, the random variables that constitute the information set are simply the past of y_t.

AR processes are, to some extent, a natural counterpart to MA processes because an MA process is what you get by filtering a white noise process, whereas an AR process is a process that yields white noise if suitably filtered. In symbols

$$A(L)y_t = \epsilon_t,$$

where $A(L)$ is, as usual, a polynomial in L (with degree p) with $A(0) = 1$ and $a_i = -\varphi_i$.

To describe these processes, let us begin by considering the simplest case, where $p = 1$:

$$y_t = \varphi y_{t-1} + \epsilon_t \longrightarrow (1 - \varphi L)y_t = \epsilon_t.$$

Let's start from mean and variance. These can be obtained in several ways: one rather intuitive way is to assume the stationarity of the process (so we know they exist), and then develop the consequences of this hypothesis. Let us assume, therefore, that the process has a constant mean μ. This hypothesis implies

$$\mu = E(y_t) = \varphi E(y_{t-1}) + E(\epsilon_t) = \varphi\mu.$$

Th expression above can be true in two cases: either $\mu = 0$, in which case it is true for any value of φ, or in the case $\varphi = 1$, and then the expression is true for any value of μ, and the mean of the process is undefined. In the latter case, the process is characterised by a **unit root**, because the value of z for which $A(z) = 0$ is precisely 1; the analysis of this situation, in which bizarre things happen, kept the best econometricians busy during the final two decades of the 20th century. For a long time, it was considered by applied economists as rough terrain, upon which it is inadvisable to venture without a native guide, but nowadays unit roots are a standard tool in applied economics (especially macro). We shall discuss this case in depth later (Chapters 4 and 6). For the

moment, let's rule out the $A(1) = 0$ case. It follows that — in the cases we analyse here — the process has a zero mean.

Another method to obtain $E(y_t)$ is to re-express y_t as a MA process (under stationarity, Wold's theorem guarantees this is legitimate). To do this, we use the results reported above on the manipulation of polynomials. If we focus on the cases where $|\varphi| < 1$ (so unit roots are excluded), we have

$$A(L)^{-1} = (1 - \varphi L)^{-1} = 1 + \varphi L + \varphi^2 L^2 + \cdots = C(L),$$

therefore the MA representation of y_t is

$$y_t = \left(1 + \varphi L + \varphi^2 L^2 + \cdots\right) \epsilon_t = C(L)\epsilon_t,$$

that is, an MA process with $\theta_i = \varphi^i$, which has zero mean;[8] therefore, $E(y_t) = 0$.

As for the second moments, we proceed as above; let us call $V(\epsilon_t) = \sigma^2$. If we call ω the variance of y_t, and assume that it exists and is constant over time, it follows that

$$\omega = V(y_t) = E(y_t^2) = E\left[(\varphi y_{t-1} + \epsilon_t)^2\right] = \varphi^2\omega + \sigma^2 + 2\varphi E(y_{t-1}\epsilon_t).$$

The last element of the sum is 0, since $y_{t-1} = C(L)\epsilon_{t-1}$, and therefore $E(y_{t-1}\epsilon_t)$ is a linear combination of autocovariances of a white noise (all of which are zero by definition). It follows that

$$\omega = \varphi^2\omega + \sigma^2 \Longrightarrow \omega = \frac{\sigma^2}{1 - \varphi^2}.$$

The same result could also have been obtained from the MA representation, noting that

$$\omega = \sigma^2 \sum_{i=0}^{\infty} \theta_i^2 = \sigma^2 \sum_{i=0}^{\infty} \varphi^{2i} = \sigma^2 \sum_{i=0}^{\infty} (\varphi^2)^i = \frac{\sigma^2}{1 - \varphi^2}.$$

The expression $\omega = \frac{\sigma^2}{1-\varphi^2}$ provides several insights. First, it tells us that only if $|\varphi| < 1$ the variance is stable over time (for $|\varphi| \geq 1$ the last equality no longer holds); this condition rules out from the category of stationary AR(1) processes not only those with a unit root, but also those with a so-called explosive root ($|\varphi| > 1$).

Second, by comparing ω, which is the *unconditional* variance of y_t, with σ^2, which is the variance of $y_t|\mathcal{F}_{t-1}$, it can be seen that ω is always greater than σ^2. Furthermore, the difference is greater the closer φ is to 1: the more persistent the process is, the smaller its variance *conditional* on its own past will be compared to its unconditional variance. That is to say, knowing the value of y_{t-1} reduces the uncertainty about the value of y_t the more persistent the series is.

[8] The moving average representation of an AR(1) process can also be obtained by the so-called method of 'recursive substitutions', which is cruder and less elegant, but perhaps more intuitive. Let us consider that, if $y_t = \varphi y_{t-1} + \epsilon_t$, then we will also have $y_{t-1} = \varphi y_{t-2} + \epsilon_{t-1}$; we substitute the second expression into the first and proceed iteratively.

All it remains to do is to compute the autocovariances: γ_0 is ω, which we already know; the autocovariance of order 1 is given by

$$\gamma_1 = E(y_t y_{t-1}) = E\left[(\varphi y_{t-1} + \epsilon_t) y_{t-1}\right] = \varphi E(y_{t-1}^2) + E(y_{t-1}\epsilon_t) = \varphi\omega$$

and, more generally,

$$\gamma_k = E(y_t y_{t-k}) = E\left[(\varphi y_{t-1} + \epsilon_t) y_{t-k}\right] = \varphi\gamma_{k-1},$$

thus obtaining

$$\gamma_k = \varphi^k \frac{\sigma^2}{1-\varphi^2}.$$

Alternatively, starting from the MA representation, we get

$$E(y_t y_{t+k}) = \sigma^2 \sum_{i=0}^{q} \theta_i \theta_{i+k} = \sigma^2 \sum_{i=0}^{q} \varphi^i \varphi^{i+k} = \sigma^2 \sum_{i=0}^{q} \varphi^{2i+k}$$

which is equal to

$$\gamma_k = \varphi^k \sigma^2 \sum_{i=0}^{q} \varphi^{2i} = \varphi^k \frac{\sigma^2}{1-\varphi^2}.$$

As a consequence, the autocorrelations take a very simple form:

$$\rho_k = \varphi^k;$$

autocorrelation decline exponentially towards 0 as k grows larger; this feature is often referred to as "exponential decay".

Again, there is an intuitive interpretation of the result above: autocorrelations, which we take as an index of the memory of the process, are larger (in absolute value) the larger (in absolute value) φ is, so we may think of φ as a persistence parameter. Furthermore, autocorrelations are nonzero for all k, although $\lim_{k\to\infty} \gamma_k = 0$. In practice, the memory of the process is infinite, but the very remote past plays a negligible role.

Let's look at one more example: Figure 2.3(a) shows the white noise process I already presented in Figure 2.1(a) as an example of MA(1) processes with $\theta = 0$. Let us apply the operator $(1-\varphi L)^{-1}$ to this process, with $\varphi = 0.5$ and $\varphi = 0.9$. You can see persistence increasing as the parameter (φ in this case) grows, and in fact the effect is much easier to spot here.

Like MA processes, AR processes can be easily adapted to the non-zero mean case: let us add an 'intercept' to the AR(1) model:

$$y_t = \mu + \varphi y_{t-1} + \epsilon_t \to (1-\varphi L) y_t = \mu + \epsilon_t.$$

By inverting the polynomial $A(L)$, we get

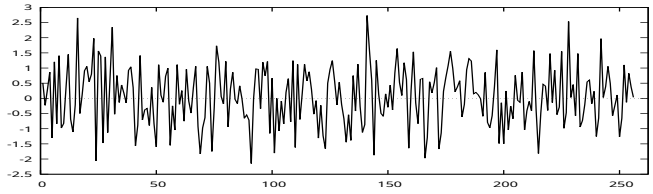
$$y_t = \left(1 + \varphi L + \varphi^2 L^2 + \cdots\right)(\mu + \epsilon_t) = C(L)(\mu + \epsilon_t);$$

since applying L to a constant leaves it unchanged, it follows that

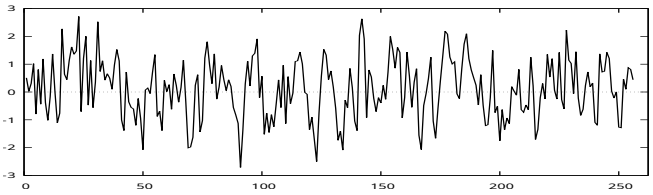
$$y_t = (1 + \varphi + \varphi^2 + \cdots)\mu + C(L)\epsilon_t = \frac{\mu}{1-\varphi} + C(L)\epsilon_t,$$

Figure 2.3: AR(1) with different values of φ

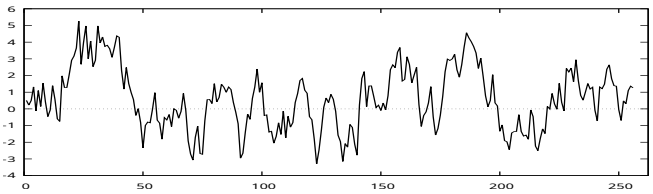
(a) $\varphi = 0$ (white noise)



(b) $\varphi = 0.5$



(c) $\varphi = 0.9$



therefore $E(y_t) = \frac{\mu}{1-\varphi}$.

Generalising to the AR(p) case is very boring, since the necessary algebraic manipulations, stemming from the necessity to invert a p-th degree polynomial to obtain the moving average representation, are a bit of a pain. We have

$$C(L) = A(L)^{-1} = \prod_{j=1}^{p}(1 - \nu_j L)^{-1}$$

where the ν_j are the reciprocals of the roots of $A(L)$, that is

$$\nu_i = 1/\lambda_i, \qquad A(\lambda_i) = 0.$$

Moreover, this generalisation does not bring great advantages to the intuitive understanding of the salient characteristics of these processes. The fundamental point is that if $|\nu_j| < 1$ for all j (and therefore, all the roots of $A(L)$ are greter than 1), then the expression above makes sense because the $A(L)$ polynomial does in fact have an inverse that we can write as $C(L)$ (in this case, $A(L)$ is said to be "stable"); in this case, the AR(p) process is stationary. A rigorous proof is omitted here, but the curious reader should note that all is needed is equation (2.16) in the Appendix of this chapter.[9]

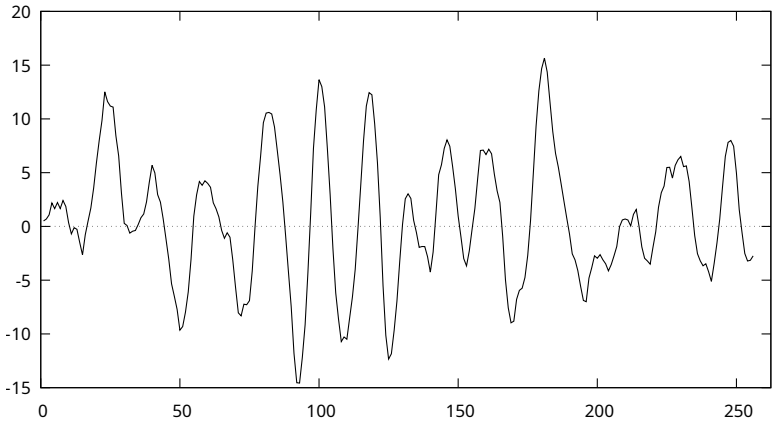
[9]Note: ν_j can be a complex number: in this case, allow me to remind the reader that that its absolute value is given by the formula $|a + bi| = \sqrt{a^2 + b^2}$. For those unfamiliar with the topic, a brief overview is provided in the Appendix to this chapter, on page 32.

Other interesting properties (I won't prove these either) are that an AR(p) process

- has infinite memory, but the autocorrelations decrease with k geometrically like in AR(1) processes;
- in the case of an 'intercept' different from 0, it has an expected value of $\frac{\mu}{A(1)}$, where $A(1)$ is the polynomial $A(z)$ evaluated at $z = 1$ instead of $z = L$ as usual; in practice, $A(1) = \sum_{i=0}^{p} a_i$.

The only aspect worth emphasising when the autoregressive order p is greater than 1 is that AR(p) processes can have cyclical trends: this occurs if and only the roots of the polynomial $A(z)$ contain one or more pairs of complex conjugate numbers. In this case, the process assumes a cyclical trend in which the amplitude of the oscillations varies around a mean value. Clearly, processes of this type are the natural candidates to model economic phenomena characterised by cyclical phases.

Figure 2.4: AR(2): $\varphi_1 = 1.8$; $\varphi_2 = -0.9$



The reason why complex numbers and cyclical trends are linked would be wonderful to explain, but unfortunately the average student in an Economics faculty considers complex numbers and trigonometric functions to be arcane witchcraft, so I'll pass. For the eccentric few who like these things, there's an Appendix at the end of this chapter so that the majority are left unfazed.

Let us take a look at an example: consider the white noise process in Figure 2.3(a) and use it to simulate an AR(2) process in which the polynomial $A(z)$ has two complex roots:

$$y_t = 1.8y_{t-1} - 0.9y_{t-2} + \epsilon_t,$$

where the roots are

$$\lambda = \frac{1.8 \pm \sqrt{3.24 - 3.6}}{1.8} = 1 \pm \frac{i}{3},$$

both greater than 1 in absolute value (they are "one plus something"). As can be seen in Figure 2.4, there is a more or less regular alternation of 'peaks' and 'troughs'.

2.5 ARMA processes

The class of ARMA processes includes both AR processes and MA processes as special cases. An ARMA(p,q) process is defined as

$$A(L)y_t = C(L)\epsilon_t, \tag{2.12}$$

where p is the order of the polynomial $A(L)$ and q is the order of the polynomial $C(L)$. Both are finite numbers. AR or MA processes are therefore special cases ($q = 0$ and $p = 0$, respectively).

If all the roots of $A(L)$ are greater than 1 in modulus, then y_t can also be represented in MA form

$$y_t = A(L)^{-1}C(L)\epsilon_t = C^*(L)\epsilon_t,$$

where $C^*(L)$ is a polynomial of infinite order if $p > 0$. This condition on $A(L)$ is necessary and sufficient for the process to be stationary. Similarly, if the polynomial $C(L)$ is invertible, then y_t admits an autoregressive representation (of infinite order if $q > 0$)

$$C(L)^{-1}A(L)y_t = A^*(L)y_t = \epsilon_t.$$

In this case, the *process* is said to be invertible.

The moments of an ARMA(p,q) process can be derived in a conceptually simple (but algebraically cumbersome) manner from its moving average representation, and I shall not waste your time with this. The only one worth mentioning is that, if we add an intercept, it can be easily shown[10] that only the AR part counts, and the mean of the process is $\frac{\mu}{A(1)}$. Optionally, you can add a touch of further sophistication by allowing for the possibility of a non-zero, time-varying mean, namely something like

$$A(L)y_t = \mu(x_t,\beta) + C(L)\epsilon_t,$$

which looks very much like a regression model. An alternative way of writing the above is to consider a model of the form

$$y_t = \mu_t + u_t,$$

where $\mu_t = \frac{1}{A(L)}\mu(x_t,\beta)$ and $u_t = \frac{C(L)}{A(L)}\epsilon_t$. This is a sort of a regression model where the errors are given by an ARMA(p,q) process. As can be seen, it is easy to switch from one representation to the other.

[10] Blitz proof: $A(L)y_t = \mu + C(L)\epsilon_t \Longrightarrow E[A(L)y_t] = \mu + E[C(L)\epsilon_t]$. By the linearity of the operators E e L, we get $A(L)E[y_t] = \mu + C(L)E[\epsilon_t] = \mu$. However, if y_t is stationary, then $E[y_t]$ exists, is finite and constant, so that $A(L)E[y_t] = A(1)E[y_t]$, and $E[y_t] = \frac{\mu}{A(1)}$.

Why are ARMA processes useful? In theory, they're not: Wold's representation theorem ensures that any stationary process has an MA representation. In practice, however, Wold's representation is, in general, of infinite order. This is a serious problem for inference: if we assume that y_t can be represented in MA form as

$$y_t = B(L)\epsilon_t,$$

nothing prevents the polynomial $B(L)$ from being of infinite order. If an observed time series is considered as a realisation of some stationary process, using an MA process to summarise its covariance characteristics involves the inferential problem of estimating a potentially unbounded number of parameters.

However, one could consider a finite approximation of $B(L)$: in applied mathematics, it is well known that sometimes functions can be approximated very well by "rational functions", that is, the ratio of two polynomials. So, it may be possible to find two polynomials of finite (and, hopefully, low) order, $A(z)$ and $C(z)$, such that

$$B(z) \simeq \frac{C(z)}{A(z)}.$$

If the equality were exact, one could then write

$$y_t = B(L)\epsilon_t \Longrightarrow A(L)y_t = C(L)\epsilon_t.$$

If the equality holds only approximately, then

$$A(L)y_t = C(L)\epsilon_t^*,$$

where

$$\epsilon_t^* = \frac{A(L)}{C(L)}B(L)\epsilon_t.$$

Strictly speaking, the process ϵ_t^* is not white noise; however, if its autocovariances are "small enough", it can be treated as such for all practical purposes. Alternatively, one could argue that treating ϵ_t^* as a white noise process is a only slightly more misleading metaphor than the one based on ϵ_t (namely, Wold's representation), whilst offering the distinct advantage of relying on a finite number of parameters.

In practice, an ARMA model is built by making an *a priori* assumption about the degrees of the two polynomials, $A(L)$ and $C(L)$, and then — once the polynomial coefficients have been estimated — by examining the sample autocorrelations of the time series corresponding to ϵ_t^*. If these are not too large, then ϵ_t^* can be treated as white noise without adverse consequences[11].

2.5.1 Seasonal ARMA

The need to keep the number of parameters under control leads, in certain cases, to the usage of models known as **multiplicative ARMA** models, used primarily for time series characterised by seasonal persistence, so that they are also known as **seasonal ARMA**, or **SARMA**.

[11] For estimation techniques, see Section 3.

For example, let us consider the monthly time series of hotel arrivals for some skiing holiday destination (say, Gstaad or Cortina). We would expect strong seasonality, in the sense that the data for December likely resembles those of December from the previous year much more than that of May from the same year, which is closer in time but qualitatively different.

Say, for simplicity, that we wish to use a pure autoregressive model, that is, without an MA component. A naive application of the ideas presented so far would evidently lead to the use of a polynomial of (at least) order 12, and thus to a structure with a fairly large number of parameters. Many of these, however, are likely redundant: while the conditional mean for the month of December is likely to depend on previous year's December, there is no obvious reason why the data for August or October should be equally relevant. This consideration would simply lead us to use a constrained model, with a "gappy" $A(L)$ polynomial — that is, with some coefficients equal to 0.

A more elegant and efficient approach is to write the polynomial by separating the seasonal effects from the ordinary ones. Take the following polynomial:

$$S(L) = (1 - \varphi L)(1 - \Phi L^s) = 1 - \varphi L - \Phi L^s + (\varphi \cdot \Phi) L^{s+1},$$

where s is the number of sub-periods (*i.e.* 12 for the months in a year, and so on). Obviously, in this case, the order of $A(L)$ is $s + 1$, and the only non-zero coefficients are those of order 1, s, and $s + 1$. The number of unknown parameters is, however, just 2, because the coefficient of order $s + 1$ is the product of the other two (with reverse sign); this feature makes it possible to model a possibly long seasonal pattern while keeping the necessary number of parameters under control.

Evidently, a similar trick can also be applied to the polynomial $C(L)$, so you get a remarkable degree of flexibility without having the number parameters explode. By generalising the above expression in an obvious way, we obtain a model that can be written as

$$A(L)B(L^s)y_t = C(L)D(L^s)\epsilon_t$$

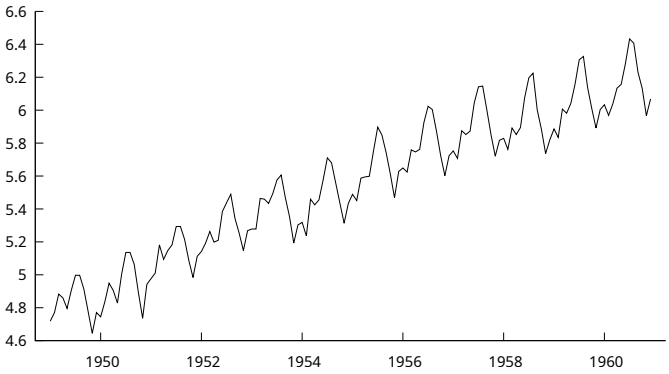
which contains the seasonal autoregressive part $B(L^s)$ and the seasonal moving average part $D(L^s)$. If the order of the polynomials $B(\cdot)$ and $D(\cdot)$ is zero, we go back to the simple ARMA case.

Example 2.5.1 (The airline model) *The most celebrated example of SARMA model is probably the so-called "airline" model, famously used by Box and Jenkins on the series depicted in Figure 2.5: the logarithm of monthly airline passengers from January 1949 to December 1960. Using the techniques I'll describe later in this chapter, they estimated the following model:*

$$\Delta\Delta^{12} y_t = (1 - 0.402\ L)(1 - 0.557\ L^{12})\epsilon_t$$

Where $A(L) = 1 - L$ and $B(L) = 1 - L^{12}$, so the AR polynomial is $1 - L^{12} + L^{13}$; similarly, $C(L) = 1 - 0.402L$ and $D(L) = 1 - 0.557L^{12}$, so the MA polynomial is $1 - 0.402L - 0.557L^{12} + 0.224L^{13}$.

Figure 2.5: Box and Jenkins' airline data



2.A Assorted results

2.A.1 Inverting polynomials

Let us begin by noting that, for any $a \neq 1$,

$$\sum_{i=0}^{n} a^i = \frac{1 - a^{n+1}}{1 - a}, \tag{2.13}$$

which is easy to prove: call

$$S = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n; \tag{2.14}$$

of course

$$a \cdot S = a + a^2 + \cdots + a^{n+1}. \tag{2.15}$$

By subtracting (2.15) from (2.14), $S(1-a) = 1 - a^{n+1}$, you get equation (2.13).

If a is a small number ($|a| < 1$), then $a^n \to 0$, and therefore $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$. By setting $a = \alpha L$, you may say that, for $|\alpha| < 1$, the inverse of $(1 - \alpha L)$ is $(1 + \alpha L + \alpha^2 L^2 + \cdots)$, that is

$$(1 - \alpha L)(1 + \alpha L + \alpha^2 L^2 + \cdots) = 1,$$

or, alternatively,

$$\frac{1}{1 - \alpha L} = \sum_{i=0}^{\infty} \alpha^i L^i$$

provided that $|\alpha| < 1$. Now consider a n-th degree polynomial $P(x)$:

$$P(x) = \sum_{j=0}^{n} p_j x^j$$

If $P(0) = p_0 = 1$, then $P(x)$ can be written as the product of n first-degree polynomials as follows:[12]

$$P(x) = \prod_{j=1}^{n} \left(1 - \frac{1}{\lambda_j} x\right), \tag{2.16}$$

where the numbers λ_j are the roots of $P(x)$: if $x = \lambda_j$, then $1 - \frac{1}{\lambda_j} x = 0$ and consequently $P(x) = 0$. Therefore, if $P(x)^{-1}$ exists, it must satisfy

$$\frac{1}{P(x)} = \prod_{j=1}^{n} \left(1 - \frac{1}{\lambda_j} x\right)^{-1};$$

but if at least one of the roots λ_j is smaller than 1 in modulus,[13] then $1/|\lambda_j|$ is larger than 1 and, as a consequence, $\left(1 - \frac{1}{\lambda_j} x\right)^{-1}$ does not exist, and neither does $P(x)^{-1}$.

For example, let's consider the polynomial $A(x) = 1 - 1.2x + 0.32x^2$; is it invertible? Let's check its roots:

$$A(x) = 0 \Longleftrightarrow x = \frac{1.2 \pm \sqrt{1.44 - 1.28}}{0.64} = (1.2 \pm 0.4)/0.64$$

so $\lambda_1 = 2.5$ and $\lambda_2 = 1.25$. Both are larger than 1 in modulus, so the polynomial is invertible. Specifically,

$$A(x) = (1 - \lambda_1^{-1} x)(1 - \lambda_2^{-1} x) = (1 - 0.4x)(1 - 0.8x)$$

and

$$\begin{aligned} \frac{1}{A(x)} &= (1 - 0.4x)^{-1}(1 - 0.8x)^{-1} = \\ &= (1 + 0.4x + 0.16x^2 + \cdots)(1 + 0.8x + 0.64x^2 + \cdots) = \\ &= 1 + 1.2x + 1.12x^2 + 0.96x^3 + 0.7936x^4 + \cdots \end{aligned}$$

In practice: if the sequence a_t is defined as the result of the application of the operator $P(L)$ to the sequence u_t, that is $a_t = P(L)u_t$, then reconstructing the sequence u_t from a_t is only possible if $P(L)$ is invertible. In this case,

$$u_t = P(L)^{-1} a_t = \frac{1}{P(L)} a_t.$$

[12] If you don't believe me, ask your favourite search engine for "Fundamental theorem of algebra".

[13] Warning: the roots may be complex, so if you're needing some guidance on complex numbers, section 2.A.2 just below is for you.

To be honest, I gave the reader only a partial account on what happens when you try to invert a lag polynomial, since I only focussed on what economists call the "backwards" expansion, where you express a series as a function of the history of another one, like in example 2.3.
In some cases, however, one may want to adopt a different strategy and go for the "forwards" expansion, where you express a series as a function of the *future* of another one. This lets you handle cases where your polynomial has roots smaller than one. Take for example the case $P(L) = 1 - \alpha L$, where $\alpha > 1$. This can be rewritten as

$$P(L) = -\alpha L\left(1 - \frac{1}{\alpha}L^{-1}\right) = -\alpha L(1 - \beta F),$$

where $\beta = 1/\alpha < 1$ and $F = L^{-1}$. By a straightforward argument you get

$$(1 - \beta F)^{-1} = 1 + \beta F + \beta^2 F^2 + \dots,$$

so

$$\begin{aligned} P(L)^{-1} &= -\beta F\left(1 + \beta F + \beta^2 F^2 + \dots\right) = \\ &= -\sum_{i=1}^{\infty} \beta^i F^i = -\sum_{i=1}^{\infty} \alpha^{-i} F^i \end{aligned}$$

(note that the sum starts from $i = 1$). For example, a relationship like

$$(1 - \alpha L)p_t = -d_t,$$

where $\alpha > 1$, cannot be solved "backwards", but does have a "forwards" solution:

$$p_t = -(1 - \alpha L)^{-1} d_t = \sum_{i=1}^{\infty} \alpha^{-i} d_{t+i}.$$

I'm fairly confident that readers who are into rational expectation models will have recognised what this example is about.

2.A.2 An intro to complex numbers

By popular request, you will find in this appendix a rough and concise introduction to how complex numbers work and, incidentally, the reason why complex roots in an autoregressive polynomial give rise to cyclical phenomena.

Let's start with the number i, which is defined by the property $i^2 = -1$. The number i is called the *imaginary unit*, largely for historical reasons (Renaissance mathematicians found i too weird to exist "for real"). Any real multiple of it is called an *imaginary number*. For example, $3i$ is a number whose square is -9.

A complex number is the sum of a real number and an imaginary one: something like $z = a + bi$, where a and b are real numbers, called the *real part* (often written as $\Re(z)$) and the *imaginary part* (often written as $\Im(z)$), respectively.

Every complex number has its *conjugate*, which is simply the same number with reverse sign in the imaginary part. The conjugation operation is written with a bar. Therefore, if $z = a + bi$, then $\bar{z} = a - bi$. Note that $z \cdot \bar{z}$ is by construction a positive real number:

$$z \cdot \bar{z} = (a + bi)(a - bi) = a^2 - (b^2 i^2) = a^2 + b^2.$$

The absolute value, or modulus, of a complex number is defined as

$$|z| = \sqrt{z \cdot \bar{z}} = \sqrt{a^2 + b^2} = \rho;$$

the definition above includes the one for real numbers as a special case, when $b = 0$. Nice.

Since, a complex number is an object identified by two real numbers, one can think of it as a point on a plane (sometimes known as the *Argand plane*). By convention, this plane has with the real part on the x-axis and the imaginary part on the y-axis. This provides an alternative representation for $z = a + bi$, that can be written in terms of ρ, its distance from the origin and ω, the angle it forms with the x-axis, that is, as $z = \rho(\cos\omega + i\sin\omega)$. Evidently, this translates into the two relationships

$$\cos\omega = \frac{\Re(z)}{\rho} \qquad \sin\omega = \frac{\Im(z)}{\rho}.$$

Figure 2.6: Graphical representation of a complex number

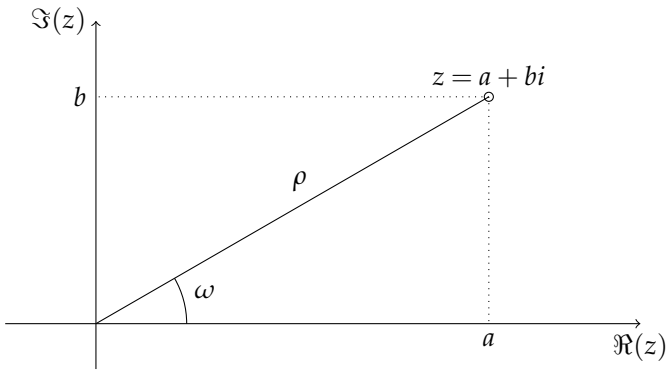


Figure 2.6 should make the dual representation quite easy to visualise. I should add that, if $\rho = 1$, the corresponding point lies on a circle of radius 1 centred at the origin, hence the expression "unit circle". Obviously, if $\rho > 1$, the point lies outside this circle, and if $\rho < 1$, it lies inside.

For reasons that I shall not go into, the following equality also holds:

$$z = a + bi = \rho\exp(i\omega),$$

where the conjugation operation translates into a change of sign in the argument of the exponential function:

$$\bar{z} = a - bi = \rho\exp(-i\omega).$$

Having two representations is convenient because one works well for addition:

$$z_1 + z_2 = (a_1 + b_1 i) + (a_2 + b_2 i) = (a_1 + a_2) + (b_1 + b_2)i$$

and the other one is handy for multiplication and exponentiation:

$$z_1 \cdot z_2 = [\rho_1\exp(i\omega_1)][\rho_2\exp(i\omega_2)] = \rho_1\rho_2 e^{i(\omega_1+\omega_2)}$$

$$z^k = \rho^k e^{i\omega k}.$$

Finally, since $\Re(z) = \frac{z+\bar{z}}{2} = \rho\frac{e^{i\omega}+e^{-i\omega}}{2}$, we also have that

$$\cos\omega = \frac{e^{i\omega}+e^{-i\omega}}{2}$$

and there is also a similar formula for the sine function:

$$\sin\omega = \frac{e^{i\omega}-e^{-i\omega}}{2i}.$$

More generally, it can be shown that

$$\frac{e^{i\omega k}+e^{-i\omega k}}{2} = \cos(\omega k) \implies \frac{z^k+\bar{z}^k}{2} = \rho^k\cos(\omega k) \tag{2.17}$$

$$\frac{e^{i\omega k}-e^{-i\omega k}}{2i} = \sin(\omega k) \implies \frac{z^k-\bar{z}^k}{2i} = \rho^k\sin(\omega k). \tag{2.18}$$

The link between complex numbers and trigonometric functions was discovered in the 18th century by the Swiss genius, Leonhard Euler. The identity named after him

$$e^{i\pi}+1=0$$

is considered by many the most beautiful formula in mathematics. This equation is not only true (it's a special case of the ones above), but it contains all the most basic ingredients of mathematics: the five most important constants (0, 1, e, π and i), the most important operations (sum, product and exponentiation) and the equal sign.
Isn't this wonderful? At some point, I even thought of having this equation tattooed on my arm, although in the end I decided it would have been too nerdy.

At this point, we have all the tools we need to shed light on the link between autoregressive polynomials with complex roots and cyclical phenomena: suppose we have a second-order polynomial in L with complex conjugate roots:

$$A(L) = 1-\varphi_1 L-\varphi_2 L^2 = (1-zL)(1-\bar{z}L).$$

Provided that $|z|$ is less than 1 to ensure invertibility, let us calculate the form of the inverse of $A(L)$, which leads us to the Wold representation. We have

$$A(L)^{-1} = (1-zL)^{-1}(1-\bar{z}L)^{-1} = (1+zL+z^2L^2+\cdots)(1+\bar{z}L+\bar{z}^2L^2+\cdots)$$

As we will now see, all the coefficients of this infinite polynomial are real, and they fluctuate around zero:

$$\begin{aligned} A(L)^{-1} &= 1+ \\ &\quad (z+\bar{z})L+ \\ &\quad (z^2+z\bar{z}+\bar{z}^2)L^2+ \\ &\quad (z^3+z^2\bar{z}+z\bar{z}^2+\bar{z}^3)L^3+\cdots \end{aligned}$$

and, more generally, if we write

$$A(L)^{-1} = 1+\theta_1 L+\theta_2 L^2+\cdots,$$

we also get

$$\theta_k = \sum_{j=0}^{k} z^j\bar{z}^{k-j}.$$

The result we are interested in can be found by noting that

$$\theta_{k+1} = z^{k+1} + \bar{z}\theta_k = \bar{z}^{k+1} + z\theta_k,$$

from which

$$z^{k+1} - \bar{z}^{k+1} = (z - \bar{z})\theta_k;$$

using equation (2.18) on can obtain

$$\theta_k = \rho^k \frac{\sin[\omega(k+1)]}{\sin\omega}.$$

In practice, the impulse response function is the product of two functions: $\frac{\sin[\omega(k+1)]}{\sin\omega}$, which is a periodic function in k with a period of $2\pi/\omega$ (and therefore, the smaller ω is, the longer the cycle). The other, ρ^k, attenuates the fluctuations as k increases since, by assumption, $|\rho| < 1$, and thus obviously $\rho^k \to 0$.

Chapter 3

ARMA models

So far, we have been pretending that the parameters of the stochastic process of our interest were known. But in practice, all we have are time series, so we want to use ARMA processes as *statistical models*. The theoretical apparatus we developed so far will be instrumental for choosing the ARMA process that we are going to employ as a Data Generating Process for what we observe.

Of course, the parameters of the process we assume as our Data Generating Process are unknown, so we have to estimate them. The standard technique for estimating ARMA parameters is maximum likelihood, where the distribution of the data is usually assumed normal to make the problem tractable.[1] Therefore, we assume that the white noise sequence $\ldots,\epsilon_{t-1},\epsilon_t,\epsilon_{t+1},\ldots$ has a joint Gaussian distribution, so, by virtue of the Wold representation, y_t is Gaussian too.

As thew reader should know, the likelihood is the sample density function, calculated for the observed sample. It will depend on a vector ψ of unknown parameters, which determine its shape. For this reason, we write it as $L(\psi)$. Maximising this function with respect to ψ yields the maximum likelihood estimate.

When we observe a realisation of a stochastic process (or, more pedantically, a time series that we can think of as such) $x_1,\ldots,x_T$, the likelihood function is nothing but the joint density function of the observed part of the process, namely the marginal density function of the random vector $x_1,\ldots,x_T$, calculated for the observed values; in the case of an ARMA process of the type

$$A(L)x_t = \mu + C(L)\epsilon_t$$

it will depend on the parameter vector $\psi = \{\mu;\varphi_1\ldots\varphi_p;\theta_1\ldots\theta_q;\sigma^2\}$.

If the process is Gaussian, the likelihood function is simply the density function of a multivariate normal:

$$L(\psi) = f(x;\psi) = (2\pi)^{-\frac{T}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-k)'\Sigma^{-1}(x-k)\right\},$$

where x is the vector $(x_1,\ldots,x_T)$ of the T observations; k and Σ are its first and second moments, which depend on ψ. For example, the ij-th element of

[1] Well, except for a few special cases; we'll see some in Chapter 7.

the matrix Σ is simply $\gamma_{|i-j|}$, the autocovariance of order $|i-j|$ which, as we know, is a function of the parameters of the ARMA process. The ML estimator is therefore

$$\hat{\psi} = \underset{\psi}{\text{Argmax}}\ L(\psi). \tag{3.1}$$

It can be shown that the maximum likelihood estimators of Gaussian ARMA processes are consistent, asymptotically normal, and asymptotically efficient. Furthermore, under fairly mild conditions, the properties of consistency and asymptotic normality are preserved even when the true distribution of the process is not normal (in this case, we refer to quasi-maximum likelihood estimates). This is the reason why, in many cases, people don't bother checking for normality.

From the viewpoint of theory, that's is all. In practice, the problems are only just beginning. First of all, it should be noted that, as usual, we do not work with the function $L(\psi)$, but with its logarithm

$$\ell(\psi) = \log L(\psi) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\left[\log|\Sigma| + (x-k)'\Sigma^{-1}(x-k)\right],$$

but this is of no consequence, since the optimal vector $\hat{\psi}$ is the same for both functions. There are three main relevant problems:

1. the system of equations resulting from setting the score vector $s(\psi) = \frac{\partial l(\psi)}{\partial \psi}$ to 0 is non-linear, and in general cannot be solved analytically, so that no closed-form expressions are available to calculate the elements of $\hat{\psi}$ as simple functions of the data; therefore, numerical optimisation techniques are needed.

2. the order of the $A(L)$ and $C(L)$ polynomials, that we need for a suitable representation of the process of which we believe x_t to be a realisation, are unknown;

3. to calculate $\ell(\psi)$, it is necessary to know the matrix Σ; now, the matrix Σ is a $T \times T$ matrix, and when T is large (even a few dozens, but time series samples can reach sizes in the order of *tens of thousands* of observations), even the simple calculation of its elements as a function of ψ may be a problem. The same applies to its determinant and its inverse.[2]

3.1 Numerical techniques

The first of the problems listed above would deserve a comprehensive treatment, but this is really not the place for it. Suffice it to say that this problem becomes negligible if adequate computing resources are available. Numerical algorithms exist that allow the maximum of likelihood functions to be found without great difficulty, and all econometric packages contain efficient implementations.

[2] It is true that Σ has a very special structure: it's a "symmetric Toeplitz" matrix. This simplifies the problem to some extent, but doesn't get rid of it.

To simplify as much as possible, one can say that these algorithms are used to find the maximum of a function, provided that we can compute the function and its first derivative for any point in the parameter space. If the function is smooth, its derivative (the *gradient*) is 0 at the maximum.

Roughly speaking, the procedure is as follows:

1. Start from a "random" point, ψ_0;
2. calculate the score $s(\psi_0) = \frac{\partial \ell(\psi)}{\partial \psi}$;[3]
3. if $s(\psi_0)$ is "small enough", stop. Otherwise, use it to figure out a direction $d(s(\psi_0))$ in which to move;
4. calculate $\psi_1 = \psi_0 + d(s(\psi_0))$;
5. check that $\ell(\psi_1) > \ell(\psi_0)$;
6. replace ψ_0 with ψ_1 and start again from step 2.

When you stop at step 3, it is usually said that the algorithm "has converged".

There are many algorithms of this type: essentially, each one calculates the direction vector $d(s(\theta_0))$ in its own way, aiming to ensure that $\ell(\psi_1) > \ell(\psi_0)$ at each iteration, so that the maximum is eventually reached.

Some readers may have noticed that the algorithm above depends crucially on what happens at step 3. As long as $s(\psi_0)$ is not 'small enough', the algorithm keeps going. Hence, we must decide what it means for a vector to be 'small', to avoid the risk of the algorithm running forever and our CPU turning into a fireball.

This is one of those deceptively simple questions: there are various ways to answer it, and as usual, none of them is 'right'. One of the most common criteria is to decide that $s(\psi_0)$ is 'zero' when none of its elements exceeds a pre-specified threshold in absolute value, for example, something like 1.0E-07. However, it is not the only one, nor is it necessarily the best.

Usually, the likelihood functions encountered in estimating ARMA models are smooth, and the possibility of having multiple maxima is negligible. Consequently, it is sufficient to be able to calculate the first derivatives of the likelihood for any vector ψ to reach the maximum, sooner or later.

3.2 Lag Selection

As for the second problem at the end of section 3, choosing the orders of the polynomials $A(L)$ and $C(L)$ (call them p and q, respectively) used to be a matter of highly skilled craftsmanship, which requires theoretical knowledge, experience, and a keen eye. Luckily, things are much simple nowadays.

One may be tempted to think that p and q can be estimated along with the rest of the parameters, but things are not this simple, because of the so-called so-called **common factors** problem: start from an ARMA(p,q) process

$$A(L)x_t = C(L)\epsilon_t,$$

[3] The derivatives can be hard-coded in software or computed via a numerical approximation; in the former case, they are said to be "analytical", otherwise "numerical". Analytical derivatives are, when available, almost always preferable.

and apply the filter $(1-\lambda L)$ to both sides of the equality: call

$$A_\lambda(L) = (1-\lambda L)A(L)$$

and

$$C_\lambda(L) = (1-\lambda L)C(L),$$

so the resulting representation is

$$A_\lambda(L)x_t = C_\lambda(L)\epsilon_t, \tag{3.2}$$

that is, an ARMA$(p+1,q+1)$ process. The Wold representation for x_t you get with both representations is exactly the same, since

$$B(L) = \frac{C_\lambda(L)}{A_\lambda(L)} = \frac{(1-\lambda L)C(L)}{(1-\lambda L)A(L)} = \frac{C(L)}{A(L)}$$

and the $(1-\lambda L)$ factors cancel out. Note that λ can be literally any number and the math works just the same.

The process x_t, therefore, has a completely equivalent ARMA$(p+1,q+1)$ representation *for all possible values of* λ. Therefore, for every ARMA(p,q) representation, you have infinitely many ARMA$(p+1,q+1)$ representations that are absolutely equivalent, because the shape of the autocorrelation function is the same for every λ. As we say in econometrics, the model is under-identified.[4]. Therefore, the value of the likelihood function is the same for any λ, and thus there is no unique maximum: there are infinitely many maxima, one for each value of λ.

Example 3.2.1 (Common factors) *A simple example: what kind of process is*

$$y_t = 0.5y_{t-1} + \epsilon_t - 0.5\epsilon_{t-1}?$$

Easy, you will say: it is an ARMA(1,1). Quite right. However, it is also white noise; in fact,

$$y_t = \frac{1-0.5L}{1-0.5L}\epsilon_t = \epsilon_t.$$

In practice, I wrote a white noise process as an ARMA(1,1). This latter representation is redundant, but not wrong. *The important aspect to note is that my choice $\lambda = 0.5$ is completely irrelevant: I could have used 0.7, 0.1, or whatever. There are infinitely many "not wrong" representations.*

From a practical point of view, modelling an ARMA(p,q) as an ARMA$(p+1,q+1)$ leads to all sorts of problems. First, numerical algorithms often

[4]I will briefly recap what is meant by the identification of a model in econometrics: a model is said to be under-identified if there is more than one data representation consistent with what is observed. In practice, it is not possible to decide on the basis of the data whether representation A or representation B is more correct; in these cases, the expression "observational equivalence" is used. If the model is parametric (as in the majority of cases), it is identified if the likelihood function has a unique global maximum; consequently, a necessary condition for identification is the non-singularity of the Hessian matrix at the maximum point. Identification is, clearly, itself a necessary condition for the existence of a consistent estimator.

struggle to converge (which is unsurprising, given that there is no unique maximum). Second, even assuming that convergence is eventually achieved somehow, the maximum point we find is only one of the infinitely many possible representations of the model[5].

Therefore, before we embark on likelihood maximisation, we must make a preliminary choice on p and q. Years ago, when CPUs were toys, RAM was precious and software was scarce, textbooks used to recommend a rather involved procedure,[6] based on the fact that there are highly specific relationships between the order of the polynomials and the autocorrelation structure. By examining of the sample autocorrelations, an initial hypothesis can be made about the orders of the polynomials. If, for example, you note that the sample autocorrelations cut off abruptly beyond a certain order q, the you may consider using an MA(q) model, whose theoretical autocorrelations exhibit the same feature. If, on the other hand, the autocorrelations decay smoothly, an AR process might be better.

At this stage, statistics known as partial autocorrelations (which, in practice, are never used for any other purpose) were also employed. Defining partial autocorrelations rigorously is somewhat boring. It is quicker to say how they are calculated: the partial autocorrelation of order p is calculated by running a regression of y_t on a constant and $y_{t-1} \dots y_{t-p}$. The resulting coefficient associated with y_{t-p} is the partial autocorrelation of order p. These quantities cut off abruptly in the case of pure AR models, and decay gradually in the case of pure MA models.

Even today, some teachers inflict on their students these old-fashioned techniques for choosing p and q by making elaborate considerations about the shape of the autocorrelation function, while conveniently glossing over that, in the majority of practical cases, you either have a highly trained eye or you are left completely in the dark. The fact is that these techniques were invented in an era when ML estimation of ARMA models was difficult and expensive; as a result, it was essential to have as precise an idea as possible of the potential values of p and q before attempting estimation. Today, ML estimation of an ARMA model is ridiculously easy, and the art of interpreting correlograms is for the fans of vintage items, like telephone boxes and vinyl records.

The modern approach to the choice of p and q is based on special statistics called **information criteria**. These are based on concepts taken from information theory, and I am happy to refer to the literature for further details. Information criteria start from the value of the log-likelihood (multiplied by -2) at the maximum and add a *penalty function*, which is increasing in the number of parameters.

[5] It is true that all these representations have the same Wold representation, so the forecasts they lead to and the impulse responses they generate are identical, but we have the problem that any kind of testing in this case is precluded. Indeed, the maximum point we find is only one of the infinitely many possible ones, and therefore the Hessian of the likelihood function is singular. Since all test statistics are based in some way on the curvature of the likelihood function, it is clear that tests cannot be performed.

[6] In the statistical literature this stage is known as the identification phase. This term sometimes causes a bit of confusion, because in econometrics the word 'identification' usually means something else.

Many different criteria exist, depending on the form of the penalty function. The most common ones, in the context of ARMA modelling are the Akaike IC (AIC), the Schwartz IC (BIC, where B is for "Bayesian") and the one by Hannan and Quinn (HQC):

$$\begin{aligned} \text{AIC} &= -2\ell(\hat{\psi}) + 2k && (3.3) \\ \text{BIC} &= -2\ell(\hat{\psi}) + k\log T && (3.4) \\ \text{HQC} &= -2\ell(\hat{\psi}) + 2k\log\log T && (3.5) \end{aligned}$$

where k is the dimension of the vector ψ (the number of estimated parameters) and T is the sample size. The rule is to pick the model that minimises information criteria, so you strike a balance between model fit (given by the value of the log-likelihood) and the model size. Therefore, between two models, you should choose the one with a lower IC.

In the light of what we said earlier, the logic should be clear. If p and/or q are too small, the fit will be poor, and the IC will be large; if, conversely, p or q are too large, the model will be over-identified and will give nearly the same fit as a correctly identified one, while using more parameters. By choosing the model with minimum IC, you try to pick a model that fits the data nicely without being too complex.

Therefore, the modern approach is simply a brute-force one: you select "maximal" orders p and q on the basis of economic/statistical common sense and try all the possible combinations. The model that minimises your IC of choice is the winner. Sometimes, the three criteria may not agree, but even if they don't your problem is reduced to picking one out of two or three at most.

Once estimation has been carried out, one usually checks whether the 'residuals' (the one-step-ahead forecast errors) look like a white noise process. The approach usually adopted is based on testing procedures which allow us to determine whether a process is a white noise or possesses persistence. The most widely used is the so-called **Ljung-Box** test, which is based on the property that, with large T, sample autocorrelations for a white noise process tend to 0: the actual test statistic is

$$LB(p) = T(T+2)\sum_{i=1}^{p}\frac{\hat{\rho}_i^2}{T-i};$$

it will be noted that it is substantially a weighted sum of the squared sample autocorrelations up to order p. The smaller these are, the smaller the test statistic will be; under the null hypothesis the true value of all autocorrelations up to order p is 0, and the critical values are those of the χ_p^2 distribution.

3.3 Computing the log-likelihood

The third problem I hinted at near the end of section 3 is more intriguing: essentially, we need to write the likelihood function in such a way that the calculation of prohibitively large matrices is not required. The literature studied the specific problem of evaluating log-likelihood functions for Gaussian

ARMA models in great detail between the 1970s and the 1980s, so that the built-in algorithms found in econometric packages today are very stable and efficient. Here, however, I want to illustrate a rather interesting technique (the so-called sequential factorisation) because of its great ingenuity and its generality (it is used in much more general cases than ARMA models).

To illustrate sequential factorisation, start from the definition of conditional probability, which is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

from which

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

If we apply this rule to the density function of a bivariate random variable, we obtain

$$f(x,y) = f(y|x)f(x) = f(x|y)f(y). \tag{3.6}$$

This little trick can also be repeated with a trivariate random variable, thus obtaining

$$f(x,y,z) = f(x|y,z)f(y,z) = f(y|x,z)f(x,z) = f(z|x,y)f(x,y). \tag{3.7}$$

By combining (3.6) and (3.7), one can clearly write, for example,

$$f(x,y,z) = f(z|x,y)f(x,y) = f(z|x,y)f(y|x)f(x)$$

and therefore a joint density function of n random variables can be written as

$$f(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i|x_1,\ldots,x_{i-1})$$

thereby transforming a function of many variables into a product of functions of one variable. In the simple case when the variables x_i are independent $f(x_i|x_1,\ldots,x_{i-1}) = f(x_i)$ and (as is well known), the joint density function is the product of the marginals.

Since the likelihood function is nothing but the density function of the sample, this same decomposition can be applied to the likelihood function, and in the case of ARMA models it takes an especially nice form. Since in ARMA models

$$y_t = E(y_t|\mathcal{F}_{t-1}) + \epsilon_t$$

(see equation (3.8)) we have that, conditionally on $\mathcal{F}_{t-1}$, the distribution of y_t is the same as that of ϵ_t, and therefore the likelihood can be written equivalently in terms of the one-step-ahead forecast errors rather than the y_t values.

In the case of pure AR processes, the result is straightforward: in this case, if the process is

$$y_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t$$

and ϵ_t is normal, the density function $f(y_t|\mathcal{F}_{t-1})$ is simply a normal distribution:

$$y_t|\mathcal{F}_{t-1} \sim N\left(\mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}, \sigma^2\right).$$

In the case of an AR(1) process, we will have that

$$y_t|\mathcal{F}_{t-1} \sim N\left(\mu + \varphi y_{t-1}, \sigma^2\right)$$

and therefore $f(y_t|\mathcal{F}_{t-1}) = f(y_t|y_{t-1})$: the only element of $\mathcal{F}_{t-1}$ that matters is the most recent one. Consequently, the likelihood can be written as

$$L(\mu, \varphi, \sigma^2) = f(y_1)f(y_2|y_1)f(y_3|y_2)\cdots = f(y_1) \times \prod_{t=2}^{T} f(y_t|y_{t-1}).$$

Taking logarithms, we obtain

$$\begin{aligned} l(\mu, \varphi, \sigma^2) &= \log f(y_1) + \tilde{l}(\mu, \varphi, \sigma^2) = \\ &= \log f(y_1) - \frac{1}{2}\sum_{i=2}^{T}\left(\log 2\pi + \log\sigma^2 + \frac{(y_t - \mu - \varphi y_{t-1})^2}{\sigma^2}\right) \\ &= \log f(y_1) - \frac{1}{2}\sum_{i=2}^{T}\left(\log 2\pi + \log\sigma^2 + \frac{e_t^2}{\sigma^2}\right), \end{aligned}$$

where I adopted the notation $e_t = y_t - E(y_t|\mathcal{F}_{t-1})$.

If the first term were zero, the remainder would be equal to a completely standard log-likelihood function for a linear model in which the dependent variable is y_t, the explanatory variable is y_{t-1} (plus the intercept), and the disturbance is normal with mean 0 and variance σ^2.

As is well known, in this case OLS is the maximum likelihood estimator, so that, if it were not for the $\log f(y_1)$ term, we could simply run a regression to get the job done. But then, for very large samples, this term becomes negligible in determining the total log-likelihood $l(\mu, \varphi, \sigma^2)$, since (unlike the second term), it does not grow as T increases. The OLS statistic, which maximise $\tilde{l}(\mu, \varphi, \sigma^2)$, therefore converge towards the ML estimator, and asymptotically share their properties[7].

This line of reasoning also holds for AR(p) models: in this case, the first term of the likelihood becomes $\log f(y_1, \ldots, y_p)$, but the argument remains unchanged. Moreover, it is true that, under fairly general conditions, the OLS estimator of the parameters of a stationary autoregressive process is a consistent and asymptotically normal estimator even if the process ϵ_t is not Gaussian; in these cases, the OLS estimator as asymptotically equivalent to the quasi-maximum likelihood estimator, and it's OK to use it (with the appropriate corrections to standard errors). This is the case of *dynamic regression*, where the asymptotic properties of OLS estimators can be proven via limit theorems that are usually dealt with elsewhere in an Econometrics course, so that I don't have to do it here.

In the case of models with a moving average component, the discussion becomes only slightly more intricate. As I have already pointed out in section

[7] In cases such as this, one speaks of **conditional** likelihood, to indicate that the likelihood function we are using treats the first p observations as fixed, and therefore is based on the distribution of $y_{p+1} \ldots y_T$ *conditional on* $y_1 \ldots y_p$. There are also techniques that maximise the so-called **exact** likelihood, that is, the one which also takes into account the distribution of the first p observations, but asymptotically it makes no difference.

3.4.1, in an ARMA model, the white noise process driving the y_t process can be interpreted as the difference between y_t and its expected value conditional on $\mathcal{F}_{t-1}$ (see eq. (3.8)). Consequently, if we assume that the distribution of y_t conditional on $\mathcal{F}_{t-1}$ is normal, it follows that the one-step-ahead forecast errors are a sequence of uncorrelated (and therefore independent) normal variables with mean 0 and constant variance, so that the likelihood can be calculated very simply using forecast errors.

3.4 Using ARMA models

If the parameters of an ARMA process are known, the model can be used for two purposes: forecasting its future values and/or analysing its dynamic characteristics.

3.4.1 Forecasting

We'll build on a simple and intuitive argument: if there is persistence, and the past contains information on the present, then the present contains information on the future, and we may be able to say something sensible on the distribution of future values $y_{T+1}, y_{T+2}, \ldots$ on the basis of the available data. Clearly, we'd like to do so by following some "optimal" rule.

The problem can be framed as follows: let us define a **forecast function** for y_t as some function of the variables contained in the information set $\mathcal{F}_{T-1}$. That is, a forecast function is some rule yielding the forecast we make for y_t given its previous values, which we assume known. We call this value $\hat{y}_t = f(y_{t-1}, y_{t-2}, \ldots)$. Our problem is choosing a forecast function that works "well".

To make this idea less fuzzy, let's define the *forecast error*, as

$$e_t = y_t - \hat{y}_t.$$

Clearly, we'd like e_t to be zero or very small as often as possible, so we must consider its distribution. If y_t is an ARMA process (or representable as such), once we have the model in the form $A(L)y_t = C(L)\epsilon_t$, an assumption about the distribution of ϵ_t makes it possible, at least in principle, to determine the distribution of y_t conditional on $\mathcal{F}_{T-1}$. In turn, this determines the conditional distribution $e_t|\mathcal{F}_{T-1}$, which is crucial for choosing a forecast function.

Strictly speaking, an optimal choice should be made according to the following criterion:

1. first, choose a function $c(e_t)$ (the so-called *loss* function), which associates a cost with the forecast error. In general, we have that $c(0) = 0$ (the cost of a perfect forecast is 0) and $c(e_t) \geq 0$ for $e_t \neq 0$;

2. then, define the expected loss as

$$c^* = E\left[c(e_t)|\mathcal{F}_{T-1}\right] = E\left[c(y_t - \hat{y}_t)|\mathcal{F}_{T-1}\right].$$

   The quantity c^* is the cost that, on average, we incur because of incorrect forecasts. Clearly, we want it to be as small as possible;

3. since c^* is a function of $\hat{y}_t$, we choose $\hat{y}_t$ in such a way as to minimise c^*; that is, we *define* $\hat{y}_t$ as that function which minimises the expected cost of the forecast error.

It should be clear that the best forecast function depends on what the loss function looks like: for example, if this function is asymmetric, choosing the optimal forecast function may be non-trivial. The example I often use is booking a restaurant: since the loss function in this case is asymmetric (it is better to have empty chairs than people left standing), it is always advisable to book for a slightly higher number of people than actually expected.

Fortunately, the matter becomes far less intricate if the loss function is quadratic, that is, if $C(e_t) = \kappa e_t^2$ for some positive number κ. In this case (which can often be taken as a satisfactory approximation to the most appropriate cost function), it can be shown that $\hat{y}_t$ coincides with the conditional expectation:

$$C(e_t) = \kappa e_t^2 \Longrightarrow \hat{y}_{T+1} = E(y_{T+1}|\mathcal{F}_T).$$

This property is so convenient that, in the vast majority of cases, the conditional mean is adopted as "the" forecast function with no justification necessary.

Therefore, from now we'll use the expected value of y_{T+1} as our forecast function, assuming we can use a set of observations ranging from 1 to T:

$$\hat{y}_{T+1} = E(y_{T+1}|\mathcal{F}_T). \tag{3.8}$$

In the case of a pure AR model, the solution is trivial, since all values of y_t up to time T are known, and therefore $E(y_{t-k}|\mathcal{F}_T) = y_{t-k}$ for any $k \geq 0$:

$$E(y_{T+1}|\mathcal{F}_T) = \varphi_1 y_T + \cdots + \varphi_p y_{T-p+1} + E(\epsilon_{T+1}|\mathcal{F}_T).$$

The value of $E(\epsilon_{T+1}|\mathcal{F}_T)$ is evidently 0, since the memoryless property of white noise ensures that there is no information available at present regarding the future of ϵ; consequently, $E(\epsilon_{T+1}|\mathcal{F}_T) = E(\epsilon_{T+1}) = 0$. Therefore, the forecast of y_{T+1} is

$$\hat{y}_{T+1} = \varphi_1 y_T + \cdots + \varphi_p y_{T-p+1}. \tag{3.9}$$

For the two-step-ahead forecast, start from the expression

$$\hat{y}_{T+2} = E(y_{T+2}|\mathcal{F}_T) = \varphi_1 E(y_{T+1}|\mathcal{F}_T) + \cdots + \varphi_p y_{T-p+2} + E(\epsilon_{T+2}|\mathcal{F}_T),$$

which, using the same logic, can be easily shown to be equal to

$$\hat{y}_{T+2} = \varphi_1 \hat{y}_{T+1} + \cdots + \varphi_p y_{T-p+2}$$

and more generally,

$$\hat{y}_{T+k} = \varphi_1 \hat{y}_{T+k-1} + \cdots + \varphi_p \hat{y}_{T+k-p},$$

where, $\hat{y}_{T+k} = y_{T+k}$ for $k \leq 0$. Note the intriguing parallelism between $A(L)y_t = \epsilon_t$ and $A(L)\hat{y}_t = 0$, which is easily obtained by considering the expected value (conditional on $\mathcal{F}_{t-1}$) of the first of the two expressions.

Example 3.4.1 *Given the following AR(2) process*

$$y_t = 0.9y_{t-1} - 0.5y_{t-2} + \epsilon_t,$$

suppose we observe a realisation of this process, and that the last two observations are equal to: $y_{T-1} = 2$ *and* $y_T = 1$*. The best forecast for* y_{T+1} *is*

$$\hat{y}_{T+1} = 0.9 \times 1 - 0.5 \times 2 = -0.1,$$

while the forecast for y_{T+2} *is*

$$\hat{y}_{T+2} = 0.9 \times (-0.1) - 0.5 \times 1 = -0.59,$$

and going on is easy. For the record, the following five values are -0.481*,* -0.1379*, 0.11639, 0.173701, 0.098136.*

Of course, evaluating the conditional mean provides a point forecast, but says nothing about the reliability of the forecast — that is, about the variance of the forecast error. This is not a topic on which I wish to dwell for too long (although this point is explored in more detail in Section 5.A.2 in the more general case of multivariate processes). For the interested reader, I will simply suggest, in addition to the usual bibliographic references at the end, that a useful exercise would be to prove that, in the case of an AR(1) process,

$$V(\hat{y}_{T+k}) = \sigma^2 \frac{1 - \varphi^{2k}}{1 - \varphi^2}.$$

It may be interesting to note that the forecast error variance is always smaller than the unconditional variance of y_t: leveraging on the persistence characteristics of the time series makes its future behaviour less uncertain. Moreover, as $k \to \infty$, the two variances tend to coincide; this occurs because, in stationary AR(1) processes, persistence is always of a short-term nature. Knowledge of the current state of the system is no more informative about its remote future than its unconditional distribution is: for a sufficiently large k, y_t and y_{t+k} are virtually uncorrelated (and therefore, if Gaussian, virtually independent).

Note that here we are treating the true parameters of the process as known, whereas in fact all we usually have are estimates; therefore, the variance of the forecast error in principle depends not only on the intrinsic variability of the process σ^2, but also on the uncertainty on the parameters of the process itself.
For example, take an AR(1) process for which we have a parameter estimate $\hat{\varphi}$; the argument above would have us analyse

$$\hat{y}_{T+k} = E(y_{T+k}|\mathcal{F}_T) = E(\hat{\varphi} \cdot y_{T+k-1}|\mathcal{F}_T),$$

where $\hat{\varphi}$ cannot be "taken out" of the expectation operator, because it is an estimator and not a constant.
That said, this distinction is important in theory, but in practice nobody bothers, and people normally uses the estimates as if they were the true parameters. This is justified because, with consistent estimators, for very large T the distinction becomes immaterial.

In the more general case of processes with an MA part, forecasting can still be carried out by applying the same concept. In particular, it should be noted that if $\mathcal{F}_{t-1}$ has no lower bound in time, it includes not only all past values

of y_t but also those of ϵ_t. This is quite obvious by considering that ϵ_t can be *defined* as

$$\epsilon_t = y_t - E(y_t|\mathcal{F}_{t-1}).$$

If we assume the process y_t to be invertible, then it can be written as

$$C(L)^{-1}A(L)y_t = G(L)y_t = \epsilon_t$$

from which we obtain

$$y_t = \underbrace{g_1 y_{t-k-1} + g_2 y_{t-k-2} + \cdots}_{E(y_t|\mathcal{F}_{t-1})} + \epsilon_t$$

and therefore, given $\mathcal{F}_t$, all the elements of the sequence ϵ_t up to time t can be considered known (as in "available from $\mathcal{F}_t$")

With this proviso, the conditional expectation operator can be applied to every component of an ARMA model. The fact that the available information set is not infinite represents only a minor issue. Indeed, if we only have observations over the time span $\{0 \cdots T\}$, a very convenient solution is to extend our information set backwards by using the unconditional mean values of y_{-1}, y_{-2}, $\cdots$ *etc*. If the process is stationary and ergodic, as the sample size increases, the difference is negligible.[8]

I shall illustrate this point using an ARMA(1,1) process: once the concept is clear, its generalisation is trivial. Let us take, therefore, a process like

$$y_t = \varphi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}.$$

Let us place ourselves at time 0, when no observations are available. What is the best forecast we can make for y_1? Since we have no data, the conditional mean coincides with the marginal mean, and therefore $\hat{y}_1 = E(y_1) = 0$. One period passes, and we observe the actual value y_1. At this point, we can calculate the forecast error for period 1, namely $e_1 = y_1 - \hat{y}_1$; since $\hat{y}_1$ is 0, for the reasons just mentioned, it follows that $e_1 = y_1$. Our forecast for $\hat{y}_2$ is given by the following formula:

$$\hat{y}_2 = E(y_2|\mathcal{F}_1) = E(\varphi y_1 + \epsilon_2 + \theta\epsilon_1|\mathcal{F}_1) = \varphi E(y_1|\mathcal{F}_1) + E(\epsilon_2|\mathcal{F}_1) + \theta E(\epsilon_1|\mathcal{F}_1).$$

Let's consider these terms one by one, bearing in mind that $\mathcal{F}_1 = \{y_1\}$: clearly, the first two terms create no problems, since $E(y_1|\mathcal{F}_1) = y_1$ (this is obvious) and $E(\epsilon_2|\mathcal{F}_1) = 0$ (by assumption). But what about $E(\epsilon_1|\mathcal{F}_1)$? Since ϵ_1 can also be interpreted as the forecast error that would be made at time 0 if the information set were infinite, then *the best possible forecast we can make regarding the forecast error at time 1 is exactly the forecast error that we actually made*. Based on this concept, we can formulate our forecast for y_2 as

$$\hat{y}_2 = \varphi y_1 + \theta e_1.$$

[8]The *exact* calculation can be carried out, if desired. There are many ways to do this, but the most common — also because it is easily automated — is to use a tool called the **Kalman filter**. For those who wish to know more, further details can be found in the literature.

Let another period pass, and we observe y_2; from this, we calculate e_2, and the process continues, in the sense that at this stage we have everything we need to calculate $\hat{y}_3 = \varphi y_2 + \theta e_2$, and so forth. In practice, one-step-ahead forecasts for processes of the type

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

are computed as follows:

$$\hat{y}_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}, \tag{3.10}$$

that is to say, by using the actually observed values of y_t and the values of past forecast errors in place of ϵ_{t-i}.

A brief digression. One might legitimately wonder what the practical value is of forecasts made in this way, given that ARMA processes are merely a stylised and approximate representation of the time series we observe. In other words, the *historical* persistence characteristics of the time series are summarised by pretending that our data series is a realisation of suitably-chosen ARMA process.

There is no *logical* reason, however, why an approximation that worked well for the past should continue to work well for the future. Yes, we assume that the data-generating process is stationary, but this may well be just wishful thinking. The stationarity assumption means, more or less implicitly, that the set of circumstances that have hitherto ensured that a given process is a good approximation of the behaviour of that given time series will continue to hold over the forecast horizon of interest.

This condition is often plausible when the series is a description of a reasonably stable physical phenomenon (for example, the temperature recorded daily at a certain geographical location at a certain hour). However, in the case of economic phenomena, this may be a rather bold assumption, as the causal chain of events that combine to determine the value of the series at any given moment is likely to be more unstable. I would consider it unprofessional to make a forecast for the price of crude oil based exclusively on an ARMA process that does not, for instance, take into account the political situation in Venezuela, or the state of affairs between Iran and the US.

More precisely, an ARMA model forecast should be accepted as a conditional forecast for a specific scenario: *if and only if* the political situation in Venezuela (and in Iran, and in the United States, *etc.*) remains more or less as it is today, *then* it can be said that, and so on and so forth.

3.4.2 Persistence analysis

There are many ways a persistent process can behave:[9] you can have cycles, seasonality, and all sorts of different features. To assess qualitatively these features, you can use the so-called **impulse response function**. This object can be described by considering the equation

$$y_t = E\left[y_t | \mathcal{F}_{t-1}\right] + \epsilon_t = \hat{y}_t + \epsilon_t,$$

that comes from equation (3.8).

The value of y_t can therefore be interpreted as the sum of two components: one ($\hat{y}_t$) which, at least in principle, is perfectly predictable given the past; the other (ϵ_t) which is absolutely unpredictable. In other words, the value of

[9] All white noise processes are alike; each persistent process is persistent in its own way. With apologies to L. Tolstoy.

y_t can be thought of as depending on a persistence component to which a random disturbance or, as it is commonly known, a random **shock** is added, summarising everything unpredictable that occurred between time $t-1$ and time t. The effect of this component, however, also propagates into the future of the series y_t through the persistence effect. This is why the white noise process ϵ_t is often called, in a more neutral form, the *one-step-ahead forecast error* or *innovation*.

Therefore, by writing the process in MA form (or, if you like, by considering its Wold representation)

$$y_t = A(L)^{-1}C(L)\epsilon_t = B(L)\epsilon_t$$

the i-th coefficient of the polynomial $B(L)$ can be thought of as the effect that a shock occurring i periods ago has on the current value of y_t, or, equivalently, as the impact that today's events will have on the series under study in i periods' time:

$$b_k = \frac{\partial y_t}{\partial \epsilon_{t-k}} = \frac{\partial y_{t+k}}{\partial \epsilon_t}. \tag{3.11}$$

The impulse response function is simply given by the coefficients of the MA representation of the process as a function of the lag k, and it is generally represented by a plot with the values of k on the horizontal axis and the values of b_i on the vertical axis.

To compute the Wold representation of an ARMA process, in principle we need the inverse polynomial of $A(L)$. This can be rather tedious, especially if the order of the autoregressive component is high. Fortunately, the same computation can be performed recursively, by a algorithm so simple that it can even be implemented on a spreadsheet:

1. define a series e_t that contains all zeros except for one period, in which it equals 1. In other words, define a series e_t for which $e_0 = 1$ and $e_t = 0$ for $t \neq 0$;
2. define a series b_t, be equal to 0 for $t < 0$; for $t \geq 0$, instead, let $A(L)b_t = C(L)e_t$ hold.

The values obtained for the series b_t are exactly the Wold representation coefficients, that is the values of the impulse response function.

Example 3.4.2 *Let us take, for example, the following ARMA(2,1) process:*

$$y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$$

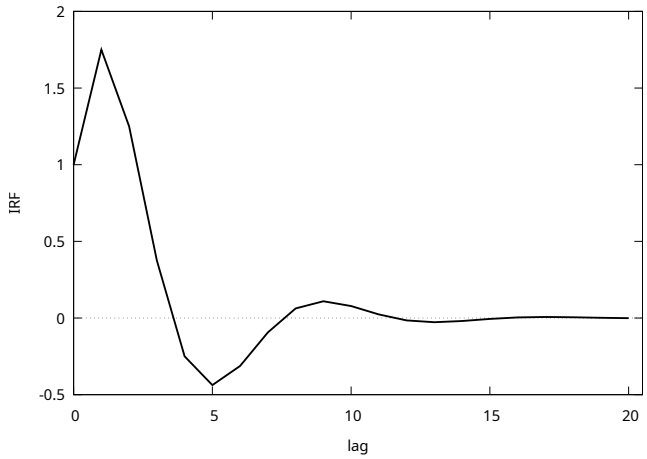
and say that, at time t, an "unpredictable event" equal to 1 has occurred (that is to say, $\epsilon_t = 1$). What effect does this have on the values of y from time t onwards? Let's go one step at a time. At time t, clearly, the effect is 1, since ϵ_t acts directly on y_t and does not affect its other components. At time $t+1$, we will have that

$$y_{t+1} = y_t - 0.5y_{t-1} + \epsilon_{t+1} + 0.75\epsilon_t,$$

and the effect of ϵ_t on y_{t+1} will be twofold:

$$\frac{\partial y_{t+1}}{\partial \epsilon_t} = \frac{\partial y_t}{\partial \epsilon_t} + 0.75.$$

Figure 3.1: Impulse response function for $y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$



On the one hand, there is a direct effect, associated with a coefficient of 0.75; on the other hand, one must take into account the fact that the effect of ϵ_t is also contained within y_t, which is associated with a coefficient equal to 1. Therefore, the total effect will be 1.75.

Moving forward by another period, the direct effect disappears, and only the effect generated by the lagged values of y remains, which turns out to be 1.25 for lag 2, .375 for lag 3, etc. The following small table may be helpful:

t	e_t	i_t
-2	0	0
-1	0	0
0	1	$b_{-1} - 0.5b_{-2} + e_0 + 0.75e_{-1} = 1$
1	0	$b_0 - 0.5b_{-1} + e_1 + 0.75e_0 = 1.75$
2	0	$b_1 - 0.5b_0 + e_2 + 0.75e_1 = 1.25$
3	0	$b_2 - 0.5b_1 + e_3 + 0.75e_2 = 0.375$
⋮	⋮	⋮

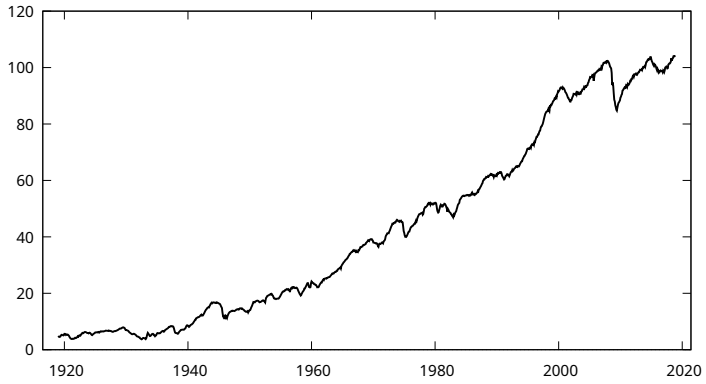
If you have the patience to carry on up to 20 periods, you will be able to draw a plot like the one in Figure 3.1: it can be seen quite clearly that, after 8 periods, the function reproduces (with some obvious decay) more or less the same dynamics. Consequently, it is reasonable to expect that a realisation of this process will exhibit cyclical patterns with a wavelength of approximately 8 periods.

A final remark: one might be led to believe that the Wold representation and the impulse response function are the same thing. Well, not necessarily: the impulse response function is defined as in equation (3.11), and can always be calculated, even if the process is not covariance-stationary, and, therefore, does not possess a Wold representation. However, if the process is in fact stationary, then the impulse response and the Wold representation coefficients coincide.

3.5 In practice

To illustrate how everything works, let us take the case where we want to set up a model for a time series that we have already spent some time on in Section 1.4, namely US industrial production[10]. The available data are plotted in Figure 3.2.

Figure 3.2: USA industrial production: seasonally adjusted index (1919–2018)



To begin with, you will notice that the time series is very long: it runs from January 1919 to December 2018. 'Wow, over a thousand observations!', some might say, 'when it comes to sample size, the more the better!'. Well, not really. Let me remind the reader that what we are trying to do is find a stochastic process such that these data can be thought of as a single realisation.

Now, the "true" stochastic process that generated this time series (assuming it exists) is something that has surely changed a lot over our observation window. Within the timespan of our data, you get Charlie Chaplin, World War II, the invention of the ballpoint pen, Marilyn Monroe, personal computers, *Smells Like Teen Spirit*, and YouTube.

I find it somewhat bold to postulate the existence of a data representation that survives all this intact. In the terms of Chapter 1, we could say with a certain degree of confidence that the "true" stochastic process that generated the data is non stationary. If we really want to force this series into the realisation of a stationary process, it is advisable to shorten the sample to make our fictional assumption a little less unrealistic. In this case, economists like to say that we are trying to avoid *strucural breaks*[11]; it should be noted that this line of reasoning is possible *without even looking at the data*.

By a completely arbitrary decision (this is just an example, after all), I declare that the world we live in today began in January 1980. I also decided to use the log-transformed series. This is a very common procedure, and it

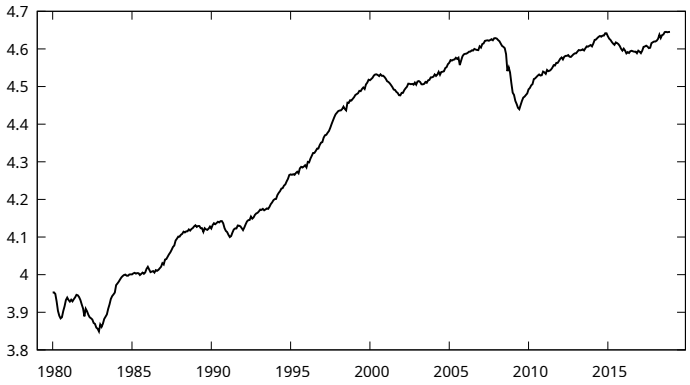
[10] For those who are familiar with these matters: I have taken a shortcut and used the seasonally adjusted series. These people will easily guess why.

[11] Some methods exist for working with time series containing structural breaks, but these are a bit too esoteric for these lecture notes.

allows for a more natural interpretation, since first differences are (approximately) percentage changes.[12]

The result is the time series in Figure 3.3, which is long enough to allow us to say something interesting (480 observations), but at the same time tells us a reasonably homogeneous story.

Figure 3.3: Logarithm of US monthly industrial production



Can we now claim to be observing a realisation of a stationary process? Again, the answer is "probably not". Here, however, the problem does not arise from the heterogeneity of the sample, but from the upward trend you see in Figure 3.3, which evidently precludes us from thinking that the process has a constant mean. One might consider including a deterministic trend in our descrition of the data, that is, something of the type $Y_t = (a + bt) + u_t$, where u_t is some ARMA process.

This wouldn't be unreasonable, especially because the parameter b could be interpreted as the long-run exogenous rate of technical progress. However, this is not such a good idea, for reasons that I will explain in Chapter 4. For now, will simply hint at the fundamental problem: even after removing a deterministic trend, this time series is too persistent to conclude that its generating process is stationary.

Another possibility is to transform the time series in such a way to get rid of the problem, but retaining economic interpretability. We brilliantly solve the problem by taking first differences and considering $y_t = 100\,\Delta Y_t$; you can admire the result in Figure 3.4 and, like I said earlier, is roughly the monthly growth rate in percentage points.

Figure 3.5 shows the correlogram for y_t. The two horizontal dashed lines surrounding the actual correlogram show the constants $\pm 1.96/\sqrt{T}$, where T is the sample size; since $T = 468$ observations, this value is around 0.091. These two lines are often included in correlograms to provide a quick-and-dirty hypothesis test: sample autocorrelations $\hat{\rho}_k$ are consistent estimators of the true autocorrelations ρ_k; if $\rho_k = 0$, then it can be shown that $\sqrt{T}\hat{\rho}_k \xrightarrow{\text{d}} N(0,1)$. Consequently, the horizontal band $\pm 1.96/\sqrt{T}$ is the 95% non-rejection

[12] Let me recall that $\log(y_t) - \log(y_{t-1}) = \log(1 + \frac{\Delta y_t}{y_{t-1}}) \simeq \frac{\Delta y_t}{y_{t-1}}$

Figure 3.4: Industrial production growth rate

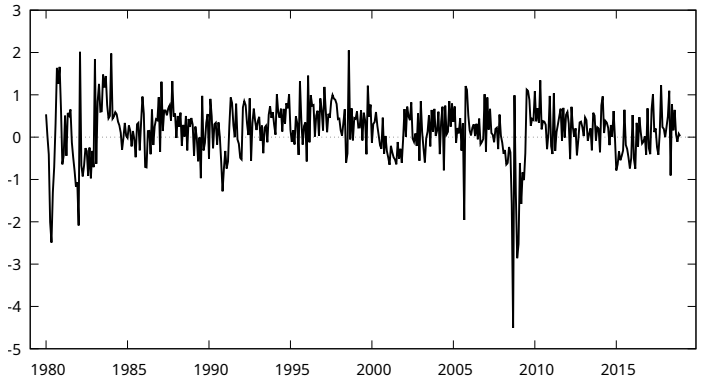
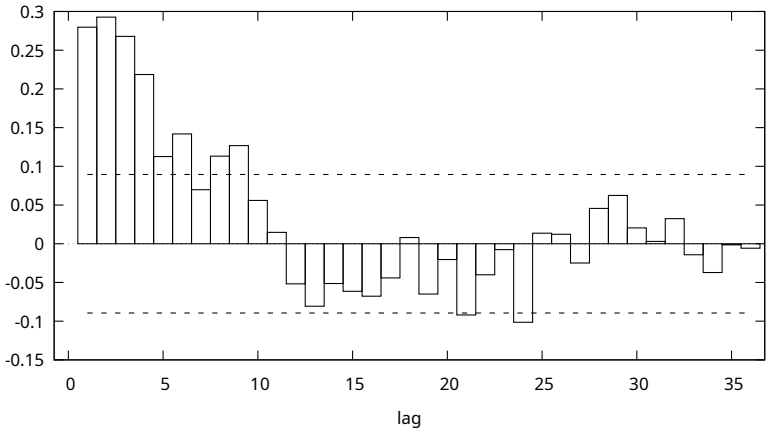


Figure 3.5: Industrial production growth rate — correlogram



region for the hypothesis $\rho_k = 0$; in practice, autocorrelations outside the band are "statistically significant".

As can be seen in Figure 3.5, there are at least six significant bars (possibly more) so we can reasonably rule out the hypothesis that y_t possesses no persistence (something which a trained eye could have spotted from Figure 3.4). Furthermore, the Ljung-Box test up to 12 lags equals 169.117; under the null, this would be a realisation of a χ^2_{12} so the p-value is 0 and rejecting the null hypothesis is the only possible choice.

Persistence, therefore, is indeed present and an ARMA model is worth trying. Let's choose p and q by the "brute-force" approach I described in Section 3.2. Set (quite arbitrarily) the maximum order for p and q to 4 and try all possible combinations using use the Akaike (AIC), Schwarz (BIC), and the Hannan–Quinn (HQC) information criteria.

Table 3.1: ARMA models for the USA industrial production growth rate: information criteria

Akaike (AIC)

AR \ MA	0	1	2	3	4
0	951.902	927.161	910.325	897.223	881.182
1	914.737	881.204	880.187	881.677	882.289
2	890.737	880.565	881.900	883.666	883.323
3	882.024	881.726	883.308	**873.083**	874.720
4	880.354	880.748	879.687	882.956	876.011

Schwartz (BIC)

AR \ MA	0	1	2	3	4
0	960.199	939.607	926.919	917.965	906.073
1	927.183	**897.798**	900.929	906.568	911.328
2	907.331	901.308	906.790	912.706	916.510
3	902.766	906.617	912.347	906.270	912.056
4	905.245	909.788	912.875	920.292	917.496

Hannan-Quinn (HQC)

AR \ MA	0	1	2	3	4
0	955.167	932.059	916.855	905.385	890.976
1	919.634	887.734	888.349	891.471	893.715
2	897.267	888.728	891.694	895.093	896.382
3	890.186	891.520	894.735	**886.142**	889.412
4	890.148	892.175	892.747	897.647	892.335

As I argued back in Section 3.2, the choice is made by selecting the model for which the criterion is lowest. Now, take a look at Table 3.1, where rows indicate the AR order and the columns the MA order; for each table, I highlighted the minimum by framing it in a small box. As can be seen, the choice is not unique: while AIC and HQC suggest an ARMA(3,3), BIC opts for a more conservative ARMA(1,1). This is no coincidence: as a rule, AIC usually tends to be rather permissive in terms of the total number of parameters, whereas

BIC has the opposite tendency. HQC generally opts for a middle ground.[13]

The information criteria, therefore, don't give us a clear winner, but at least the choice is restricted to two models. Let's have a closer look at them. In both cases, the correlogram of the one-step-ahead forecast errors, what we would call the residuals in a regression context, is fairly flat and the relative Ljung-Box test accepts the null without any problems. I'm not reporting these statistics here to avoid interrupting the flow of the argument; the core point is that is that both these models adequately capture the series' persistence. That said: is one preferable to the other, for whatever reason?

Table 3.2: ARMA(1,1) model

	Coefficient	Std. Error	z	p-value
const	0.1454	0.0698	2.0818	0.0374
ϕ_1	0.8538	0.0442	19.3217	0.0000
θ_1	−0.6379	0.0623	−10.2466	0.0000

Mean dependent var	0.149175	S.D. dependent var	0.666875
Mean of innovations	0.001287	S.D. of innovations	0.614897
R^2	0.147993	Adjusted R^2	0.146165
Log-likelihood	−436.6021	Akaike criterion	881.2042
Schwarz criterion	897.7980	Hannan–Quinn	887.7338

		Real	Imaginary	Modulus	Frequency
AR					
	Root 1	1.1712	0.0000	1.1712	0.0000
MA					
	Root 1	1.5677	0.0000	1.5677	0.0000

The ARMA(1,1) model, shown in Table 3.2, has the virtue of being the more parsimonious one. Its goodness of fit is admittedly limited: the standard deviation of one-step-ahead forecast errors ($\hat{\sigma}$) is approximately 0.615, which is only slightly smaller than the observed time series' standard deviation, which is 0.667. The R^2 index, which has the same interpretation as in an ordinary regression model, is just 14.8%. This means that the unconditional dispersion of y_t is barely lower than the dispersion of the distribution conditional on $\mathcal{F}_{t-1}$. Some persistence is certainly there, and it does in fact help us somehow to predict the series, but only to some extent (more on this later).

The ARMA(3,3) model (Table 3.3) shows a slightly different scenario where the standard deviation of forecast errors is slightly smaller (0.604) and the R^2 index goes up to 17.8%. Therefore, you get a barely noticeable improvement. One may be tempted to point out that all the parameters in the table appear to be very significant and therefore consider the model in Table 3.3 as preferable on these grounds, despite being larger in size; this is, after all, what you would do when you interpret the output from a regression.

[13]Moreover, it is true that the Akaike information criterion, which has its own historical importance because it was the first one to be invented, has the problem, compared to the others, of not being *consistent*: it can be shown that the probability of it selecting the "correct" model does not necessarily converge to 1 asymptotically, as happens with the other two; consequently, it is slightly out of favour today.

Table 3.3: ARMA(3,3) model

	Coefficient	Std. Error	z	p-value
const	0.1474	0.0682	2.1594	0.0308
ϕ_1	1.9957	0.0929	21.4839	0.0000
ϕ_2	−1.8314	0.1466	−12.4898	0.0000
ϕ_3	0.7309	0.0826	8.8488	0.0000
θ_1	−1.8274	0.0935	−19.5476	0.0000
θ_2	1.7225	0.1181	14.5839	0.0000
θ_3	−0.6370	0.0790	−8.0613	0.0000

Mean dependent var	0.149175	S.D. dependent var	0.666875
Mean of innovations	0.000702	S.D. of innovations	0.603982
R^2	0.178003	Adjusted R^2	0.169107
Log-likelihood	−428.5413	Akaike criterion	873.0827
Schwarz criterion	906.2704	Hannan–Quinn	886.1419

			Real	Imaginary	Modulus	Frequency
AR						
	Root	1	1.1719	0.0000	1.1719	0.0000
	Root	2	0.6668	−0.8501	1.0805	−0.1441
	Root	3	0.6668	0.8501	1.0805	0.1441
MA						
	Root	1	1.4836	0.0000	1.4836	0.0000
	Root	2	0.6102	−0.8281	1.0287	−0.1489
	Root	3	0.6102	0.8281	1.0287	0.1489

However, consider the bottom part of the table, where the roots of the polynomials are shown. For example, the estimated AR polynomial

$$A(z) = 1 - 1.9957z + 1.8314z^2 - 0.7309z^3$$

has three roots, that is

$$A(z) = 0 \Longleftrightarrow z = \begin{cases} \lambda_1 = 1.1719 \\ \lambda_2 = 0.6668 - 0.8501i \\ \lambda_3 = \bar{\lambda}_2 = 0.6668 + 0.8501i \end{cases}$$

so it can be written as

$$A(z) = \left(1 - \frac{z}{\lambda_1}\right)\left(1 - \frac{z}{\lambda_2}\right)\left(1 - \frac{z}{\lambda_3}\right),$$

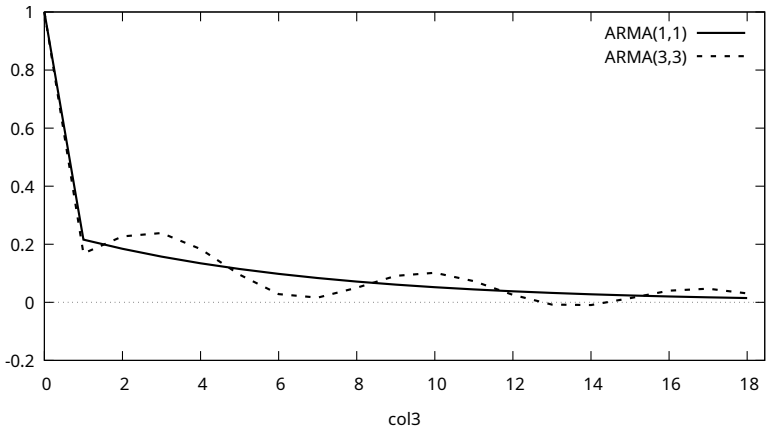
and a similar expression holds for $C(L)$.

Two of the roots of $A(L)$ are the pair of complex numbers $0.6668 \pm 0.8501i$ (which, incidentally, are suspiciously close to 1 in modulus), which is accompanied by the pair of complex numbers $0.6102 \pm 0.8221i$ among the roots of $C(L)$. One may suspect that we are are very close to the "common factors" situation here (go back to page 40 if needed). If those complex roots were the same, they could be "simplified" from the Wold representation and we'd be back to an ARMA(1,1) model.

In other words, the estimate in Table 3.3 may be an inflated estimate with parameters that serve no purpose at all. In fact, the other roots of the two

polynomials are very close to those of the ARMA(1,1) model; in other words, what you see in table 3.3 could well be the exact same model from table 3.2 in disguise.

Figure 3.6: Impulse response functions



Is this the case? Well, here is where the different information criteria disagree: BIC says "they're the same model", AIC and HQC say "no, they're not" so you have to make a decision yourself. To make a choice, it may be helpful to compute a likelihood-ratio test between the two models. I'll leave that to you as an exercise, but for the lazier part of my readers, I'll say that the associated p-value is 0.0028, so the null would be rejected. Moreover, consider Figure 3.6, which contains, for the two models examined here, the impulse response function up to 18 steps, or, in other words, a graphical representation of the first 18 coefficients of the polynomial $\frac{C(L)}{A(L)}$. Notice how the impulse responses for the ARMA(1,1) and ARMA(3,3) models are very close to one another, especially for the first five terms. The only noticeable difference is a slight undulating pattern for the ARMA(3,3) model, which is due to the the complex roots in the AR and MA polynomials not cancelling out exactly and seems to suggest a faint 7-month cycle. In practice, these two models are two alternative possibilities for summarising a feature of persistence in the series that is roughly equivalent.

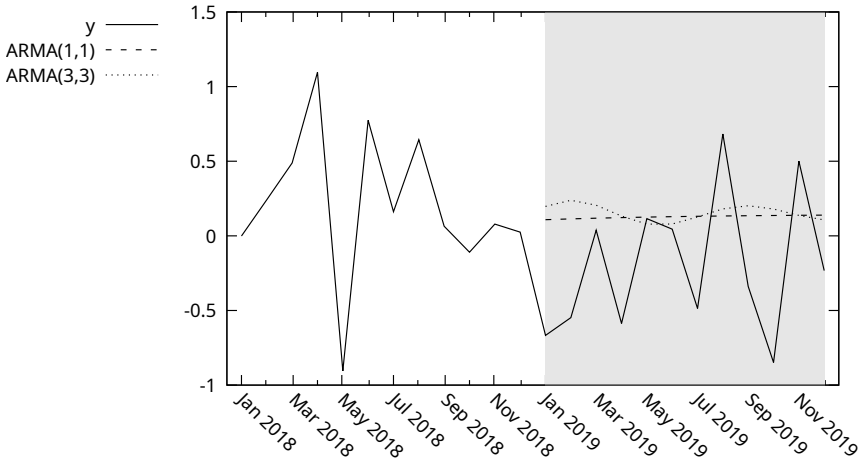
Another way to see the substantial equivalence of the two models is to consider what happens when using them as forecasting models: let's see what happens if we try to use the two models for out-of-sample forecasting into the year 2019: Table 3.4 displays the actual values and the forecasts for our two models. The same date are used to plot Figure 3.7, where the shaded area is the period outside $\mathcal{F}_T$.

The two forecasts are very similar to each other. This should come as no surprise, since we already stablished that the two Wold representations are very much alike. More worryingly, the two forecasts bear little resemblance to the actual series. For each model, we can compute a quantity called **RMSE**

Table 3.4: Out-of-sample forecasts

	y_t	ARMA(1,1)	ARMA(3,3)
2019:1	-0.667	0.108	0.197
2019:2	-0.548	0.114	0.238
2019:3	0.038	0.118	0.205
2019:4	-0.587	0.122	0.133
2019:5	0.115	0.126	0.078
2019:6	0.045	0.129	0.079
2019:7	-0.486	0.131	0.127
2019:8	0.682	0.133	0.181
2019:9	-0.339	0.135	0.202
2019:10	-0.848	0.136	0.180
2019:11	0.501	0.138	0.137
2019:12	-0.233	0.139	0.107

Figure 3.7: Forecasts



(Root Mean Squared Error), defined as

$$RMSE = \sqrt{\frac{1}{H}\sum_{i=1}^{H} e_{T+i}^2},$$

where H is the forecast horizon (12 in this example) and e_t are the forecast errors. Here, it is the ARMA(1,1) model that has a slight edge: its RMSE is 0.555, slightly smaller than the one for the ARMA(3,3) model (0.587). But again, it's a very small difference and the two models are essentially equivalent.

On these grounds, the choice between the two models becomes essentially a matter of taste. Personally, I would probably go for the ARMA(3,3) model because the weak evidence for a 10-month cycle may have some some interesting economic interpretation (provided it's not an artifact), but if one chose the ARMA(1,1) model on the grounds of parsimony, I'd have no objections. For all practical purposes, the two models work equally well (that is, not a lot).

The reader may be struck by the poor forecasting performance of the two models I considered. The reason is that the only source of information we're using here is the past of the series itself: the persistence of the series, whilst non-negligible, provides little information on the future (let me remind you that the R^2 indices for the two model were rather small). There is very little, in the information set used by our two ARMA models, that could have predicted the all the countless macroeconomic factors affecting industrial production. Almost nothing of these events could have been predicted in the industrial production data alone; in other words, our information set $\mathcal{F}_T$ is simply too limited. In fact, this is a much more general point: the quality of the predictions that *any* model can make, no matter how fancy and sophisticated, depends primarily on how much information the model incorporates. Forecasting is no black magic, and, generally speaking, univariate time-series models only predict decently a few steps ahead.

The lack in predicting power of these two models would have become even more spectacular if I had considered the year 2020, with the COVID outbreak and the consequent major collapse in industrial production that took place before of that. The moral of the story is: in economics, an ARMA model can serve, at most, for short-term forecasting during tranquil periods. If you really want a job as a soothsayer, use ARMAs, but follow the news too.

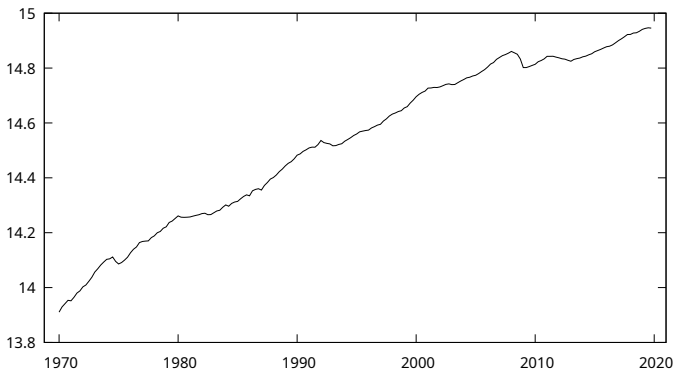
Chapter 4

Unit roots

4.1 Typical features of macro time series

All the apparatus described in the previous chapters hinges on the idea that economic time series can be modelled as realisations of stationary stochastic processes. Unfortunately, this is not the standard situation with macroeconomic time series. Just for example, let's take a look at quarterly real GDP for the Euro area (log, seasonally adjusted), which is shown in Figure 4.1.

Figure 4.1: log(GDP)



As can be seen, the series exhibits a clear upward trend over time, which precludes the possibility of modelling it via a stationary process. In this case, it would be more appropriate to use a process whose mean changes over time. However, one might think of modelling log GDP as follows: suppose that this series follows a stable deterministic growth trend over time (arguably driven by technical progress and capital accumulation), that we can, at least as a first approximation, represent as a linear function of time. The difference between the actual series and the long-term trend would be the "business cycle", which can be thought of as a stationary process, because economic cycles are a zero-mean, short-term phenomenon by definition.

Therefore, we would have a representation of the type

$$y_t = \alpha + \beta t + u_t,$$

where u_t is some stationary, zero-mean stochastic process.

Strictly speaking, the process y_t just described is not a stationary process, since $E(y_t) = \alpha + \beta t$, and therefore the mean of y_t is not constant (for $\beta \neq 0$). However, the non-stationarity of the process is restricted to this aspect. If we consider deviations from the trend, what remains is a stationary process, which can be analysed using the techniques I described earlier. It is for this reason that processes of this type are called trend-stationary processes, or **TS processes**.

Figure 4.2: log(GDP) and deterministic trend

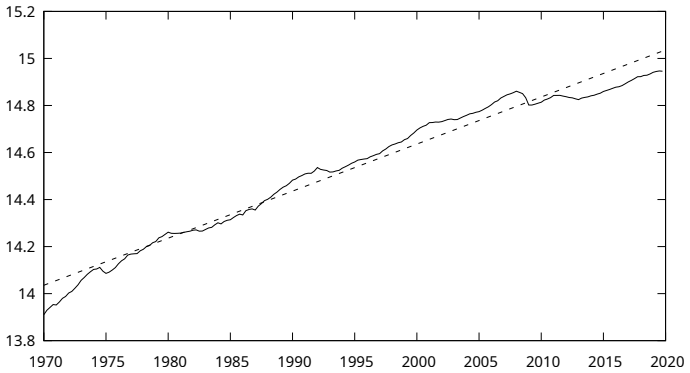
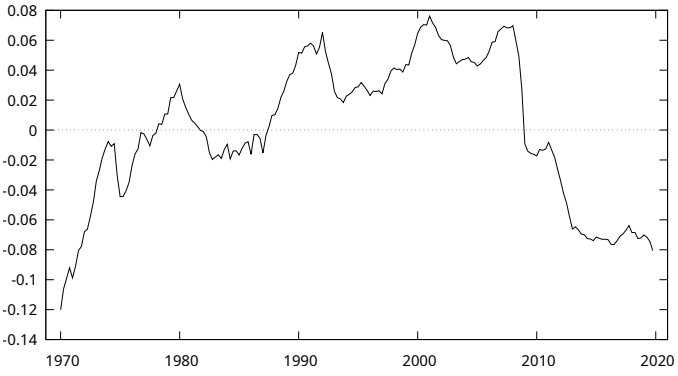


Figure 4.3: Residuals



If the process y_t is TS, then the parameters α and β can be consistently estimated using OLS, and standard inference applies. Once $\hat{\alpha}$ and $\hat{\beta}$ are obtained, it's easy to obtain a trend-cycle decomposition of the series: the series $\hat{y}_t$ yields the trend as a linear function of time, while the cycle is be given by

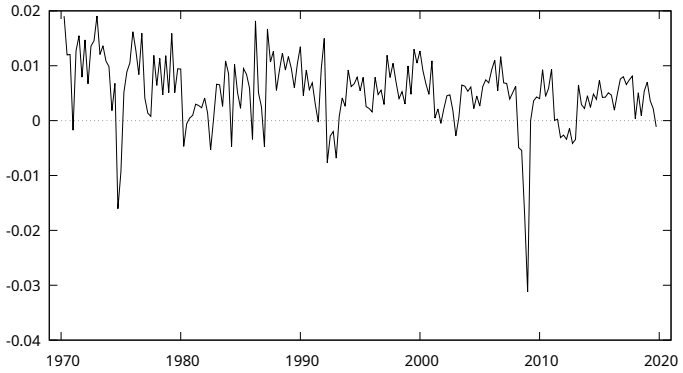
the residuals

$$\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta}t.$$

In our case, the two time series are shown in Figures 4.2 and 4.3. Looking at the latter, even a beginner's eye will conclude that the residuals are not white noise. However, it is conceivable that one could find an ARMA process of some order that accounts for the persistence characteristics of $\hat{u}_t$.

An alternative option for representing the series would be to consider the series Δy_t. Given that we are working with logarithms, this series (the trend of which is shown in Figure 4.4) can be interpreted as the time series of quarterly growth rates.

Figure 4.4: Δlog(GDP)



Since the growth rate should (arguably) fluctuate within a band, one could perhaps represent the series Δy_t as a stationary process, possibly with a non-zero mean. Note that, in this context, y_t is a unit root process: if Δy_t were indeed stationary, it would have a Wold representation of the type

$$\Delta y_t = \mu + C(L)\epsilon_t,$$

where μ is the average growth rate. This expression describes y_t as an ARMA process, where $A(L) = 1 - L$; consequently, the value of z for which $A(z) = 0$ is 1. As we know (see Section 2.4), a unit root process is not stationary, but its first difference is. Processes like these, that can be made stationary by differencing are said to be **difference-stationary**, or **DS**. Another frequently used expression is that y_t is an $I(1)$ process (read as **integrated of order one**), meaning that y_t must be differenced once for the result to be stationary. This is often abbreviated as $y_t \sim I(1)$.

What happens when you difference an $I(0)$ process? Obviously, the resulting process is stationary, but we know that, by construction, it has a unit root in its MA representation, so it is not invertible.
In this case, some argue that it makes sense to extend the notation just illustrated to negative integers. For example, if $u_t \sim I(0)$, following this convention one could say that Δu_t is $I(-1)$. Like all conventions, it is good insofar as it is useful. I personally have not yet made up my mind about it, but I feel I should I tell you.

To properly appreciate the different consequences that arise from the choice of modelling a series as a TS process or as a DS process, it is necessary to analyse in detail the characteristics of unit root processes, which is the subject of the following two sections.

4.2 Unit root processes

As I have just said, an $I(1)$ process is a process that is not stationary, but its first difference is. Therefore, you have a particularly benign form of non-stationarity, in that it can be easily treated by differentiation. More generally, an $I(d)$ process is defined as a process whose d-th difference is stationary. As far as we are concerned, we will focus only on the cases where d is 0 or 1, although there is a vast body of literature dedicated to more exotic cases.[1]

The most basic $I(1)$ process we consider is the so-called **random walk**. Its definition is simple: y_t is a random walk process if Δy_t is white noise. Therefore, $y_t = y_{t-1} + \epsilon_t$, and as a consequence, by repeatedly substituting the past values of y_{t-1}, we obtain

$$y_t = y_{t-n} + \sum_{i=0}^{n-1} \epsilon_{t-i}.$$

What does a process like this look like? To make things easier, assume that the process started at some remote time, which we call time 0, and that at that date the value of y_t was $y_0 = 0$. In this case, the previous expression reduces to

$$y_t = \sum_{i=1}^{t} \epsilon_i. \tag{4.1}$$

Note that this expression can be viewed as some sort of moving average representation of y_t, where all the coefficients are equal to 1. As such, this is not a Wold representation, because the coefficients are not square-summable.

How about its moments? If we assume, like we just did, that $y_0 = 0$, its conditional moments are not hard to find:

$$\begin{aligned} E(y_t|y_0) &= E\left[\sum_{i=1}^{t} \epsilon_i\right] = \sum_{i=1}^{t} E(\epsilon_i) = 0 \\ V(y_t|y_0) &= V\left[\sum_{i=1}^{t} \epsilon_i\right] = \sum_{i=1}^{t} V(\epsilon_i) = t\sigma^2 \end{aligned}$$

where the various equalities above are simple consequences of ϵ_t being white noise. It is clear that, at any time t, the conditional mean of the process is 0: OK so far. The variance, however, is not constant, as y_t is the sum of t independent random variables with a variance of (say) σ^2; it follows that the

[1] An obvious one is $I(2)$, which needs no explanation, but it is also possible, using concepts slightly more complex than those I am discussing here, to define processes where the order of integration d is not an integer. Finally, it is relatively easy to construct examples of processes that *cannot* be made stationary by differentiation, no matter how many times you repeat it.

variance of y_t is $t\sigma^2$, and therefore increases over time. From this, and from $Cov(y_t, y_s) = \sigma^2 \min(t,s)$ (proving this is a useful exercise), it follows that y_t is not stationary. Even worse, its *unconditional* moments don't exist at all (by the argument developed in section 2.4).

You may also see a random walk process as a limiting case of an AR(1) process where $\varphi = 1$ and the persistence characteristics are so extreme as to alter the qualitative features of the process. Furthermore, one can consider its associated impulse response function: although the Wold representation does not exist, the impulse response function is perfectly defined as $IRF_k = \frac{\partial y_{t+k}}{\partial \epsilon_t}$. In this case, it is equal to 1 for all values of k, so it is flat and does not decay exponentially as in the stationary case; this means that the effect of a shock at time t persists indefinitely into the future.

This "infinite memory" feature also implies that random walk processes do not share with stationary processes the property of being **mean-reverting**. A process is mean-reverting tends to move preferentially towards its expected value; for a process with a mean of 0, this means that the graph of the process crosses the horizontal axis 'often'. More formally, the phrase is usually employed to describe a process whose impulse response function tends asymptotically to 0.

Table 4.1: Stationary AR(1) *versus* random walk

$y_t = \varphi y_{t-1} + \epsilon_t$

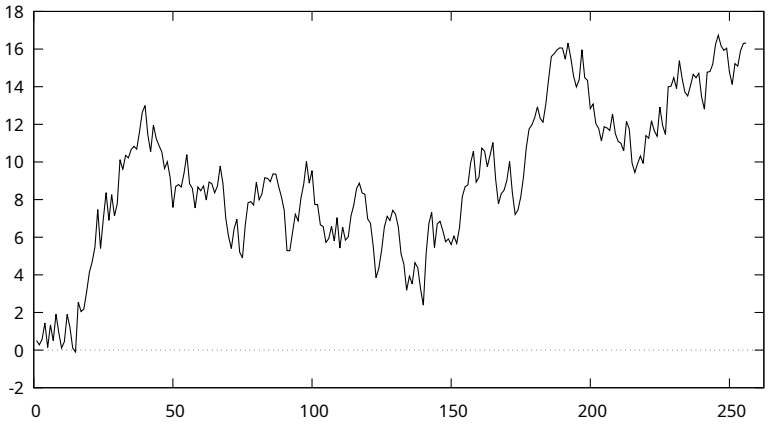
	$\lvert\varphi\rvert < 1$	$\varphi = 1$
Variance	Finite	Unlimited
Autocorrelations	$\rho_i = \varphi^i$	$\rho_i = \sqrt{1 - \frac{i}{t}}$
mean-reverting	Yes	No
Memory	Temporary	Permanent

Table 4.1 (adapted from Banerjee et al. (1993), page 85) highlights the differences between a stationary AR(1) process and a random walk.

Evidently, the choice between a TS process and a DS process for modelling a variable such as, say, log GDP entails significant consequences for analysing the long-run trend of that variable. If log GDP were representable as a realisation of a TS process, in the long run the trajectory of the exogenous trend (technology or whatever else) is the only thing that matters; a cyclical crisis may have a depressing effect, but this is only **temporary**: the system has an intrinsic tendency to return to its long-run trend. Conversely, if the most appropriate metaphor for the GDP time series were a DS process, we would have to conclude that shocks are **permanent** and will never be reabsorbed: the sins (or merits) of the fathers will fall upon the children of the children of the children, and even beyond.

This may well be a disturbing vision, but in certain cases, it can be entirely appropriate: who said that technology marches at an exogenous and fixed rate? The Early Middle Ages was a very long time, but technological innovations were somewhat lacking. And besides, once something has been invented, it cannot be uninvented (bar nuclear wars or other kinds of catastrophe): once technological progress has occurred, it remains for all those who follow. The debate on this topic is rich and flourishing, but I am contented to have given you the general idea.

Figure 4.5: Random walk



What do random walk realisations look like? Compare Figure 2.3 on page 25 and compare it with Figure 4.5. As my astute readers will have already guessed, Figure 4.5 depicts a random walk process whose increments (the ϵ_t process) are the white noise process used to generate the series shown in Figure 2.3. In practice, the only difference between Figure 2.3 and Figure 4.5 is the autoregressive coefficient, which is smaller than 1 in Figure 2.3 and equal to 1 here. Note the absence of mean reversion induced by the unit root.

A characteristic aspect of random walk processes is that the absence of mean reversion causes periods — possibly, very long — in which the resulting series exhibits strong upward or downward trends. An observer unaware that the series plotted in Figure 4.5 is just the cumulation of pseudo-random numbers might easily be tempted to comment on the different "phases" the series has gone through. It is due to this characteristic that one often speaks of **stochastic** trends, as opposed to **deterministic** trends, which are simple functions of time.

Let me make this point more explicit by considering, again, the series plotted in Figure 4.5. As we know, this series was generated via the formula (4.1), where $\epsilon_t \sim N(0,1)$, so $(y_t|y_0 = 0) \sim N(0,t)$. The conditional distribution of y_t is therefore such that

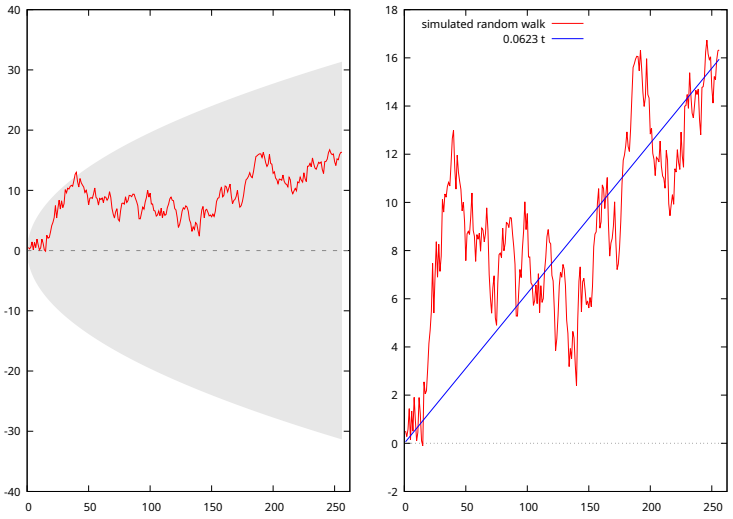
$$P\left[-1.96\sqrt{t} \leq y_t \leq 1.96\sqrt{t}\right] = 0.95.$$

Consider now Figure 4.6: the equation above implies that each point on the observed path of the series should be within the gray band of the left-hand plot with 95% probability.[2] However, the band grows larger and larger with time, so the probability that y_t should be near 0 becomes smaller and smaller as time goes on. As the reader realises, this is just a convoluted way to say that y_t is not mean-reverting, and tends to get farther and farther away from

[2]Mind you: these are *marginal unconditional* probabilities. Because of persistence, however, the *joint* probability of the entire observed path being within the gray band is different.

0 in time, so it *must* display some kind of trend. In this example, the trend happens to be upwards (look at the right-hand pane of Figure 4.6, where the series is fitted against a linear time trend via OLS), but it could have been negative equally likely.

Figure 4.6: Random walk — trending properties



It it also perfectly possible to have deterministic trends superimposed on purely stochastic ones. This occurs, for example, in so-called random walk processes with a *drift*. These processes are simply processes for which one has

$$\Delta y_t = \mu + \epsilon_t$$

and y_t is a random walk combined with a linear function of time. To see why, note that the equality above implies

$$y_t = \sum_{i=0}^{t} \epsilon_i + \mu \cdot t, \tag{4.2}$$

which is a modified version of (4.1), where the second term of the sum is nothing but a linear trend. If the drift term (the constant μ), is positive, the process tends to go up, but with fluctuations around this trend that become progressively wider over time.

In more general cases, one may consider objects like

$$\Delta y_t = d_t + \epsilon_t,$$

where d_t is some deterministic function of time: polynomial trends of various kinds (that is, things of the type $d_t = \delta_0 + \delta_1 t + \cdots$), seasonal dummies, and so on. Here, it is interesting to note that a result which encompasses the simple drift case as a special one: if d_t is a polynomial in t of degree p, then a polynomial in t of degree $p + 1$ will be present within y_t. For example, in

the simple drift case, $d_t = \mu$, which is a polynomial of order 0; similarly, it can be shown that a linear trend included in Δy_t produces a quadratic trend in y_t, and so forth. This occurs simply because, if μ_t is a polynomial in t of order p, then $\Delta\mu_t$ is a polynomial of order $p-1$ (exercise: verify this by yourself).

The random walk generalises rather painlessly to the case where the increments of the process are not white noise, but more generally any stationary stochastic process that can be represented as an ARMA process like

$$A(L)\Delta y_t = C(L)\epsilon_t,$$

where I'm omitting any deterministic component to keep things simple.

In these cases, we speak generically of $I(1)$ processes; as we will see shortly, many of the salient characteristics of random walks extend to the general case: for example, they don't have constant second moments, they're not mean-reverting, and their memory is infinite. However the polynomials $A(L)$ and $C(L)$ give the process short-run memory features, in addition to the long-run memory ones.

A word of caution: although the distinction between integrated and stationary processes is perfectly defined and absolutely unequivocal, when finite realisations of $I(1)$ processes are observed, situations can arise where the differences become more subtle and telling $I(0)$ and $I(1)$ processes apart based on the observed data may be problematic. This is because short-run memory may overlaps with long-run memory and create peculiar effects: let us consider, for example, two processes defined by

$$\begin{aligned} x_t &= 0.99999x_{t-1} + \epsilon_t \\ y_t &= y_{t-1} + \epsilon_t - 0.99999\epsilon_{t-1} \end{aligned}$$

Strictly speaking, x_t is $I(0)$, while y_t is $I(1)$. However, in the case of x_t, the root of the polynomial $A(L)$ is so close to 1 that, for all practical purposes, in finite samples a realisation of x_t is qualitatively identical to that of a random walk. Conversely, y_t can be written in the form

$$(1-L)y_t = (1-0.99999L)\epsilon_t = A(L)y_t = C(L)\epsilon_t,$$

thus making it evident that the polynomials $A(z)$ and $C(z)$ are oh so close to being 'simplifiable'; in practice, any realisation of y_t exhibits no appreciable differences from that of a white noise process.

4.3 Beveridge-Nelson decomposition

Based on what we saw in the previous section, it should come as no surprise that the sum of a random walk and an $I(0)$ process should be $I(1)$. For example, suppose that x_t is a random walk and u_t is some stationary process; their sum $y_t = x_t + u_t$ is $I(1)$, because it is not stationary, and $\Delta y_t = \Delta x_t + \Delta u_t$ is $I(0)$, although not white noise.

The features of an $I(1)$ process of this kind would be quite easy to analyse: the random walk x_t would be responsible for the non-stationarity characteristics (such as trending, infinite memory, etc.), while the short-run persistence

could be attributable to u_t. Therefore, you could split y_t into two components: a permanent, or long-run, component given by x_t, and a transitory, or short-run, component given by u_t. Since u_t is by definition a process with a mean of 0, y_t can be thought of as a process that fluctuates more or less regularly around x_t.

Amazingly enough, such a decomposition is *always* possible, so that *any* $I(1)$ process can be thought of as the sum of a random walk and an $I(0)$ process. This decomposition is known as the **Beveridge-Nelson decomposition**, or **BN decomposition** for short, and has nothing to do with the stochastic nature of processes: it is simply an algebraic artefact.

The BN decomposition can be illustrated by starting from a neat property of polynomials: given a polynomial $C(z)$ of order q, it is always possible to find a polynomial $C^*(z)$ of order $q-1$ such that

$$C(z) = C(1) + C^*(z)(1-z). \tag{4.3}$$

The proof is not difficult: obviously, $D(z) = C(z) - C(1)$ is still a polynomial of order q, since $C(1)$ is a constant (the sum of the coefficients of $C(z)$).

Now, by definition $D(1) = 0$, and therefore 1 is one of the roots of $D(z)$. Therefore, there exists is a polynomial $C^*(z)$ of degree $q-1$ such that $D(z) = C^*(z)(1-z)$.

It follows that

$$C^*(z) = \frac{C(z) - C(1)}{1-z},$$

hence equation (4.3). Proving that

$$c_i^* = -\sum_{j=i+1}^{q} c_j$$

can be a nice exercise.

Let us now take an arbitrary $I(1)$ process and call it y_t. By definition, Δy_t is $I(0)$, and therefore it must have a Wold representation:

$$\Delta y_t = C(L)\epsilon_t.$$

Applying the polynomial decomposition above to $C(L)$ yields

$$\Delta y_t = [C(1) + C^*(L)(1-L)]\,\epsilon_t = C(1)\epsilon_t + C^*(L)\Delta\epsilon_t. \tag{4.4}$$

If we define a process μ_t such that $\Delta\mu_t = \epsilon_t$ holds (that is, a random walk whose increments are given by ϵ_t), we get

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t = P_t + T_t, \tag{4.5}$$

where $P_t = C(1)\mu_t$ is a random walk (possibly, with a drift) known as "permanent component", and $T_t = C^*(L)\epsilon_t$ is an $I(0)$ process know as "transitory component".

Example 4.3.1 (Simple) *Let us consider an integrated process of order 1, y_t, for which the following holds:*

$$\Delta y_t = \epsilon_t + 0.5\epsilon_{t-1} = (1 + 0.5L)\epsilon_t = C(L)\epsilon_t,$$

where ϵ_t is white noise. Since

$$C(1) = 1.5 \qquad C^*(L) = -0.5,$$

one gets

$$y_t = 1.5\mu_t - 0.5\epsilon_t.$$

Example 4.3.2 (More difficult) *Suppose that Δy_t has the ARMA(1,1) representation*

$$(1 - \varphi L)\Delta y_t = (1 + \theta L)\epsilon_t,$$

and therefore $C(L) = \frac{1+\theta L}{1-\varphi L}$.

$C(1)$ is easy to calculate, and is equal to $\frac{1+\theta}{1-\varphi}$. The calculation of $C^*(L)$ is a bit longer but no more difficult; one can show that

$$C^*(L) = -\frac{\varphi + \theta}{1 - \varphi}(1 - \varphi L)^{-1}.$$

The final result is

$$\begin{aligned} y_t &= P_t + T_t \\ P_t &= \frac{1+\theta}{1-\varphi}\mu_t \\ T_t &= -\frac{\varphi+\theta}{1-\varphi}(1-\varphi L)^{-1}\epsilon_t. \end{aligned}$$

Note that T_t is an autoregressive process of order 1, which is the more persistent the larger $|\varphi|$ is. Consequently, y_t can be represented as a random walk plus a stationary AR(1) process fluctuating around it.

An interesting interpretation of the quantity $C(1)$ is to use it as a measure of the persistence of a given process, since it measures the fraction of the shock that remains in the process after an 'infinite' time. It is possible to verify that, by applying the decomposition described here to a stationary process, $C(1) = 0$, whereas $C(1) \neq 0$ in the case of $I(1)$ processes. Intuitively, this interpretation can also be motivated by observing that $C(1)$ is a coefficient that determines the importance of the random walk component in the process. In the case of y_t, the $I(1)$ example process I used at the end of the previous section, which was $I(1)$ but looked very much like white noise, the coefficient $C(1)$ turns out to be just 0.00001.

The utility of the BN decomposition is twofold: as a computational tool, it is frequently used in macroeconometrics to split a time series into a trend and a cycle component: you fit an ARMA model on the first differences, and then the BN decomposition is applied based on the estimated parameters. The BN decomposition is not the only tool to achieve this purpose, and it is not immune to criticism[3], but on this, as usual, I refer the reader to the literature.

The other use of the BN decomposition is theoretical. Under a different name (*martingale decomposition*), it plays a fundamental role in the probabilistic literature on stochastic processes when analysing certain asymptotic properties. While this is not of interest to us here, we will make systematic use of the BN decomposition in the analysis of cointegrated systems, which we will discuss later on.

[3]One is: who says that the long-run component must necessarily be a random walk, rather than some other type of I(1) process? Another one is: it can be easily proven that the correlation between the innovations of P_t and T_t is -1. Is this a reasonable/desirable feature?

4.4 Unit root tests

Integrated processes are very interesting: not only they provide an example of non-stationary processes, but their realisations are remarkably similar to the time series we usually encounter in macroeconomics.

What are the consequences of non-stationarity in $I(1)$ processes for the possibility of making inference on their realisations? So far, we have always assumed the stationarity of the processes that we adopt as our data generating process. In the case of $I(1)$ processes, things become more complex.

One may say: easy. If y_t is $I(1)$, then Δy_t is $I(0)$ by definition, and therefore all we know on parameter estimation for stationary processes can be recycled without any problems by estimating a model of the type

$$A(L)\Delta y_t = C(L)\epsilon_t.$$

Therefore, instead of log GDP we will model a growth rate, like we did with industrial production in Section 3.5. Models of this kind are known in the statistical literature as **ARIMA** models, that is, integrated ARMA. More specifically, you use the phrase "ARIMA(p,d,q)" to say: an ARMA(p,q) on the series, differenced d times.

A strategy of this type, however, assumes that one knows the proper value of d to use, that is, exactly whether a time series come from an integrated or stationary process[4]. In the absence of supernatural revelations, d is usually unknown; more exactly, it is almost never possible to establish *a priori* whether a given time series can be better represented by an $I(0)$ or an $I(1)$ process.

This decision, however, can be made on the basis of the data themselves. A first idea could simply be to observe the plot of the trend of the series over time. If a process is stationary, it cannot exhibit a regular upward or downward trend, and therefore one might consider a process that oscillates around a constant value to be stationary, and non-stationary otherwise.

This rule, (not utterly useless, with a bit of experience) is however too simplistic, and this for at least three reasons: first, because such a judgement is rather subjective and difficult to formalise; second, because it may well be that a process is stationary around a deterministic trend (as seen a few pages ago); finally, it is also possible that an $I(1)$ process may give rise to realisations that do not exhibit a particularly evident upward or downward tendency. For all these reasons, a less arbitrary and more reliable decision data-based rule is required. Decision rules of this type are known as **unit root tests**.

Unit root tests were a hot topic in the 1980s and 1990s, so people put a lot of work into inventing their own version, and you get dozens of them. However, the most widely used ones stem from a common approach: let us start with an AR(1) process y_t:

$$y_t = \varphi y_{t-1} + u_t. \tag{4.6}$$

[4]Remember: integration is not a characteristic of the time series itself, but rather of the stochastic process we adopt to provide a statistical representation of it. To be rigorous, we should say "... whether the observed time series is better represented by a stationary stochastic process or one that is integrated of order 1", and the issue, at this point, could shift to the meaning of "better". Let's do what vast majority of contemporary macroeconomics does, and ignore these subtleties.

By definition, the following relation must hold:

$$\Delta y_t = \rho y_{t-1} + u_t, \tag{4.7}$$

where u_t is white noise and $\rho = \varphi - 1$, so that the process is stationary only if $\rho < 0$. Conversely, if $\rho = 0$, the process is $I(1)$.

Note that the parameter ρ in equation (4.7) has a very nice and intuitive interpretation in terms of mean reversion. If $\rho < 0$, when y_{t-1} is positive, on average Δy_t will be negative (and vice versa): if at some point y_t is above zero, on average it will go down in the next period (and vice versa). If $\rho = 0$, this won't happen. Therefore, one may think of ρ as a measure of how strong mean reversion is.

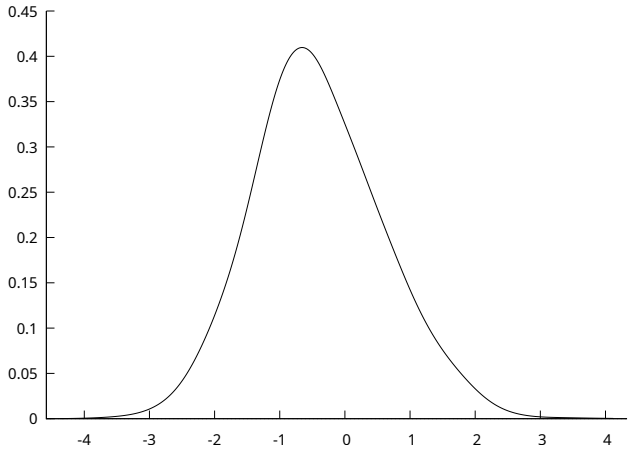
Of course, the equation above looks suspiciously like a regression model, and one might conjecture that a unit root test is nothing other than a t-test for ρ, that is, the following statistic

$$t_\rho = \frac{\hat{\rho}}{\sqrt{\widehat{\text{Var}(\hat{\rho})}}}, \tag{4.8}$$

where the 'hats' indicate, as usual, the OLS estimates. In this case, the test would have the non-stationarity of the process as the null hypothesis (if $\rho = 0$, then $\varphi = 1$), and stationarity as the alternative hypothesis ($\rho < 0$, and therefore $\varphi < 1$). The conjecture is indeed correct, although there are at least three observations to be made.

4.4.1 Distribution of the test statistic

Figure 4.7: Density function of the DF test



The first observation concerns the fact that, under the null hypothesis, the distribution of the t-test for the parameter ρ being equal to zero is neither a Student's t distribution in finite samples, as occurs in the classical linear model, nor asymptotically Gaussian, as is the case for stationary processes.

Its asymptotic distribution is instead a special one, for which there is no compact expression for either the probability density function or the cumulative distribution function[5].

A numerical estimate of the density function of this statistic is shown in Figure 4.7. Note that this density is a little unusual: it is not centred on 0, and it's a bit skewed to the left. The quantiles of this distribution must be calculated using numerical simulations, and the first to do so were Dickey and Fuller in 1976, which is why this distribution is sometimes called the DF, or Dickey-Fuller, distribution. For the same reason, the test is also known as the **DF test**. Subsequent research also found rather effective methods for calculating p-values, and modern econometric software packages provide all you need to decide if to reject the null or not.

4.4.2 Short-run persistence

A second observation concerns that, in general, y_t is not necessarily a random walk (that is, u_t is not necessarily a white noise). In other words, it is possible that Δy_t could have some short-run persistence. In these cases, which are obviously the most commonly ones in practice, it is necessary to ensure that the distribution of the test is not affected by the short-run memory contained in u_t. Therefore, let's generalise equation (4.7) by relaxing the requirement of u_t being white noise, and let's say that u_t is some unspecified $I(0)$ process. Of course, the interpretation of ρ as a mean-reversion indicator remains valid,

The most common approach is to assume that u_t has a finite AR representation of the form

$$H(L)u_t = \epsilon_t,$$

so

$$\Delta y_t = \rho y_{t-1} + \frac{1}{H(L)}\epsilon_t,$$

and consequently $A(L)y_t = \epsilon_t$, where $A(L) = H(L)(1 - \varphi L)$, or, equivalently,

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \epsilon_t. \tag{4.9}$$

Of course, under the null $A(L)$ will have a unit root.

Using the property $y_{t-k} = y_{t-1} - \sum_{i=1}^{k-1} \Delta y_{t-i}$ (it is easy to convince oneself of this by checking the case $k = 2$), equation (4.9) can be reparameterised as

[5]This distribution is defined in terms of integrals of Brownian motions. A Brownian motion, or Wiener process, is a continuous-time stochastic process of which I will not provide the definition, but which can essentially be thought of as a random walk where the interval between observations is infinitesimal.

follows:[6]

$$\Delta y_t = \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p+1} + \epsilon_t, \tag{4.10}$$

where $\rho = (\varphi_1 + \cdots + \varphi_p) - 1$ and $\gamma_i = -(\varphi_{i+1} + \cdots + \varphi_p)$. In this case, the test is named the **ADF test** (Augmented Dickey–Fuller), and it is the t-test of the significance of the parameter ρ in regression (4.10). The name comes from "augmenting" equation (4.7) with an appropriate number of lags of Δ_t.

If the chosen value of p is large enough, and therefore the correction is effective, the asymptotic distribution of the ADF test is the same as that of the DF test. What do I mean by "large enough"? It simply means that u_t must be, at least for all practical purposes, a white noise process. In practice, the same selection criteria for p are often used as for the analogous problem in the stationary framework, which I discussed in Section 3; that is, p is chosen so as to minimise criteria of the Akaike or Schwarz type.

4.4.3 Deterministic component

An extra difficulty comes from the fact that the distribution of the test is not invariant to the deterministic component included in the regression we use for computing the test statistic. So far, we have examined the case of a random walk without drift. But if a drift is actually present in the process that generated the series under examination, it must also be included in the regression used to compute the test. In practice, we don't know if the drift is there or not, so common sense dictates that we keep on the safe side and therefore a regression of the type

$$\Delta y_t = \mu + \rho y_{t-1} + \varphi_1 \Delta y_{t-1} + \cdots + \varphi_p \Delta y_{t-p} + u_t \tag{4.11}$$

where $\mu = 0$ in the "driftless" case.

The trouble is that the test statistic you get by applying OLS to equation (4.11) and computing the t-statistic for $\hat{\rho}$ is not asymptotically distributed like in the previous case.

In fact, there are actually two cases to consider: one — which is also non-standard, and shown in Figure 4.8 — when the true value of μ is 0; if, instead, $\mu \neq 0$, it turns out that the asymptotic distribution of the test is (perhaps surprisingly) Gaussian, although in practice the sample size must be quite large for the approximation to be satisfactory.

One may say: can't we use equation (4.11) for testing whether $\mu = 0$ beforehand, and then proceed? Heh, no, sorry. The problem is that under then null ($\rho = 0$) the distribution of the t-statistic for $\hat{\mu}$ is non-standard too. In short: it's a bit of a mess.

[6] Particularly motivated readers can verify that here we are doing nothing other than applying the BN decomposition: indeed, by writing

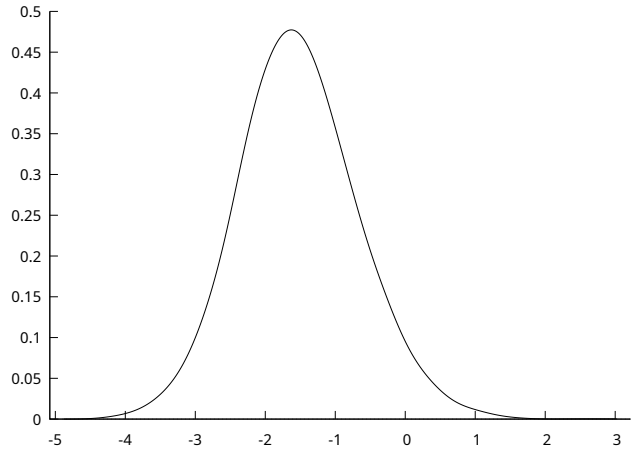
$$y_t = B(L) y_{t-1} + \epsilon_t$$

one can proceed by noting that

$$\Delta y_t = [B(L) - 1] y_{t-1} + \epsilon_t.$$

The result in the text follows by applying the BN decomposition to the polynomial $H(L) = B(L) - 1$.

Figure 4.8: Density function of the DF test with an intercept



Things become even more complicated when considering the case where a linear deterministic trend is also added to (4.11), because yet another distribution (not shown) is obtained. This multiplicity of situations is perhaps one of the most puzzling aspects when first approaching unit root tests. Of course, it is possible to navigate through it, but you need to take great car; however, on this point I refer the reader to the specialised literature, as I consider the task of a popular introduction that I set myself here to be complete. A very similar problem will return in Chapter 6, where I'll describe the possible strategies to get the job done.

4.4.4 ADF variants

As I said earlier, the number of available unit-root tests is enormous, and one cannot do justice to them all, especially in an introductory textbook as this. However, it's important to describe at least a few of them because they've been regularly used in applied work.

The first one tackles the problem of short-run dependency I described in Section 4.4.2 in a different way than ADF, and was proposed by Phillips and Perron and is commonly known as the **PP test**). Instead of augmenting equation (4.7) with lags, you run it as-is and then apply a correction to the test statistic you obtain, based used some ingenious non-parametric techniques. The essential ingredient for this correction is a quantity known as "long-run variance" (more details, if you're interested, in Section 4.A.1). The main problem is that for estimating this quantity you need to choose an integer m called the "window size", and it's difficult to state a precise rule for choosing it. My advice is to try various values of m until the statistic stabilises.

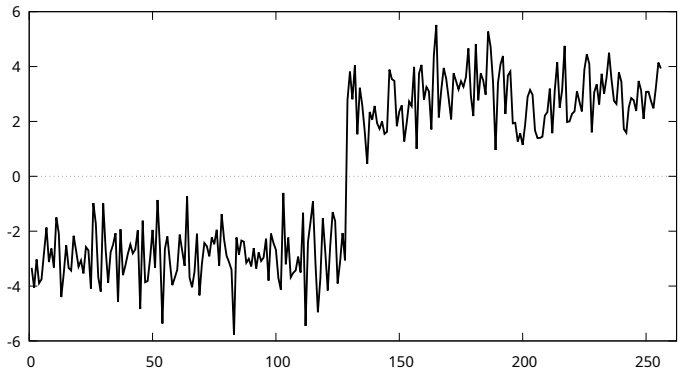
An important evolution of the ADF test is the so-called **ADF-GLS** test, by Elliott, Rothenberg and Stock, in which you use a GLS-like technique to estimate the deterministic part of the model; then, you use the residuals from this preliminary regression in an ordinary ADF regression with no con-

stant. This test was shown to have better power than the ordinary ADF test in many empirically-relevant cases, so it's become the weapon of choice for many macro practitioners. The important thing to know is that the GLS variant used in the first stage depends on a "magic number" ρ, that everybody sets to

$$\rho = \begin{cases} 1 - \frac{7}{T} & \text{with constant only} \\ 1 - \frac{13.5}{T} & \text{with constant and trend} \end{cases}$$

(don't ask me to explain why here, it's way too complicated. But believe me, there's a good reason — it's in the original article). The critical values for the ADF-GLS test are, again, nonstandard, but this is, again, not a huge problem nowadays.

Figure 4.9: White noise with a break



Finally, I'd like to mention the **Zivot-Andrews** test, because it addresses a very frequent problem in economics, namely structural breaks. Suppose you observe a series like the one in Figure 4.9: the series you see is artificially generated by the following rule:

$$y_t = \begin{cases} -3 + \epsilon_t & \text{for } t \leq 128 \\ 3 + \epsilon_t & \text{for } t > 128 \end{cases}$$

where ϵ_t is white noise. I'm willing to bet that you could throw any unit-root test at it, and you'll always conclude the series is $I(1)$. And yet, any human observer will see immediately that this series is just a stationary process around a mean, which is constant apart from one period. The Zivot-Andrews test uses some clever tricks to estimate the breakpoint and run a corrected version of the ADF test afterwards.

4.4.5 The KPSS test

The ADF test and its variants use $y_t \sim I(1)$ as their null hypothesis, but there are also tests whose null hypothesis is stationarity. The most important one of the latter kind is the so-called **KPSS test**, that hinges on the following idea. If y_t were stationary around a deterministic trend, then a regression of the type

$$y_t = \beta_0 + \beta_1 \cdot t + u_t$$

should produce $I(0)$ residuals. After having run regression, you take the OLS residuals and cumulate them, producing a new series $S_t = \frac{1}{T}\sum_{s=1}^{t} \hat{u}_s$; under the null hypothesis, this time series has a very peculiar feature. Since the model contains an intercept, the OLS residuals sum to 0. Therefore, if we set $S_0 = 0$, we know with certainty that $S_T = 0$.[7]

For very large samples, such a time series can be thought of as a realisation of a certain stochastic process (called a *Brownian bridge*), whose theoretical properties are known. It can be shown that the sum of squares of S_t (appropriately normalised) converges in distribution to a random variable, whose distribution is the same for all cases when $u_t \sim I(0)$. If, on the other hand, y_t is non-stationary, the statistic diverges. Consequently, the rejection region goes from a certain critical value (that has been tabulated) to infinity. In other words, when the test statistic is "large", you reject the null that y_t is $I(0)$. Like in the case of the PP test, here everything hinges on the estimation of the long-run variance of u_t, so the remarks I made in the previous subsection also apply here.

The test can also be carried out without a trend, so that the $\hat{u}_s$ are simply the deviations of y_t from its mean. Evidently, in this case, the null hypothesis is that the process is stationary, without a trend. The critical values are different, but these are available too.

In my view, it is always good practice to attempt to test a series' integration order in both ways. Usually, the indications coincide, in the sense that if the KPSS test does not reject, the ADF test rejects, and vice versa. However, it is not uncommon for these tests to yield inconsistent indications; that is, it frequently happens that both the ADF test and the KPSS test reject (or accept) their respective null hypotheses. However, this is not surprising, given the way hypothesis testing works in general: when you don't reject the null, you haven't proven it's true, but merely failed to prove it's false. Therefore, if neither test rejects the null, you're probably in a situation when your data contain too little information to make a decision.

4.4.6 An example

In order to provide an example of unit root tests, I'll use the same series I used in Section 4.1, namely the log GDP at constant prices for the Euro area (see Figure 4.1).

The first thing we have to decide on is the type of deterministic component we want: since the series displays an obvious upward trend, we should consider two possibilities:

1. if the series is stationary,[8] then the upward trend can only come from a deterministic component of the type $d_t = \beta_0 + \beta_1 t$.

[7] Note that these equalities are a consequence of the algebraic properties of OLS, so they are true whether H_0 holds or not.

[8] Allow me to repeat a somewhat pedantic remark: a time series is neither stationary or integrated. These are properties of the stochastic process we use to represent it. But I hope you'll allow me to use the phrase in the text for brevity, as long as we agree on its meaning.

2. Alternatively, if the series comes from a unit-root process, then the trend may be a stochastic one, but even more likely it could contain a drift, so $d_t = \beta_0$.

The best thing is try both possibilities and see what happens, hoping that the results can be put together in a coherent picture.

Let's begin with the ADF test on the series in levels. We estimate all models with the two deterministic options described above and all ADF orders from 0 to a maximum of 8 (2 years, which seems adequate), that is

$$\Delta y_t = d_t + \rho y_{t-1} + \sum_{i=0}^{p} \Delta y_{t-i}.$$

The following table contains the Akaike criterion for each model:[9]

p	$d_t = \beta_0$	$d_t = \beta_0 + \beta_1 t$
0	-1409.26	-1407.78
1	-1444.39	-1443.90
2	-1445.47	-1445.53
3	-1443.49	-1443.53
4	-1441.70	-1441.62
5	-1444.33	-1443.60
6	-1442.47	-1441.84
7	-1441.35	-1440.53
8	-1440.88	-1439.89

As can be seen, the AIC criterion is minimised, in both cases, by the choice $p = 2$. The two regressions used for computing the test, therefore, take the form

$$\Delta y_t = d_t + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \epsilon_t$$

and are displayed in Table

	Constant only			Constant and trend		
Parameter	estimate	s.e.	t-stat	estimate	s.e.	t-stat
β_0	0.0481	0.0217	2.222	0.2146	0.1262	1.914
β_1				0.00006	0.00004	1.556
ρ	-0.0031	0.0015	-2.118	-0.0169	0.0090	-1.884
γ_l	0.3609	0.0721	5.006	0.3634	0.0718	5.059
γ_2	0.1330	0.0716	1.716	0.1432	0.0717	1.5560

The test statistics are the two t-statistic relating to the parameter ρ, namely -2.118 and -1.884, respectively. The corresponding p-values are *not* those for the normal or the t distribution, but must be calculated according to the proper density (see Section 4.4.1). In this case the two p-values turn out to be 23.76% and 66.3%, so we conclude that the most adequate process to represent the series is not stationary.

[9] One could say: why the Akaike criterion, instead of something else? Well, The Akaike criterion is usually very conservative, so when it comes to testing I'd rather trade a little power in exchange for a better chance of getting the size of the test right.

If you repeat the process with the ADF-GLS test, you get much higher p-values in both cases (99.1% and 88.28%). On the contrary, KPSS rejects its null hypothesis (that is $y_t \sim I(0)$) very strongly, so we conclude that the log GDP series does in fact contain (at least) one unit root. The same procedure, applied to Δy_t instead of y_t, and therefore to the growth rates of GDP (see Figure 4.4), produces instead a clear rejection of the null hypothesis for ADF (p-value = 5.15×10^{-08}) and ADF-GLS (p-value = 0.001), and also a failure to reject for KPSS.

Conclusion? Here, things seem rather clear: y_t has a unit root, Δy_t does not, so $y_t \sim I(1)$.

4.4.7 Use your brain

A short comment on unit root tests: it happens very often that, when applying a unit root test to a time series which, reasonably, should fluctuate within a band, it is not possible to reject the unit root hypothesis. This occurs, for example, almost always with unemployment rates, inflation rates, or interest rates (real or nominal). It is common, at this point, for someone to raise their hand and say: "How is it possible for an interest rate to be $I(1)$? It was already 12% in Babylonian times. If it had been $I(1)$, it'd have drifted away by now!"

A very general answer to this point could be based on the observation that most unit root test have the $I(1)$ hypothesis as the null, but failing to reject the null is *not* equivalent to having proven that it's true.[10] A smarter answer is to point out that *within the available sample* the yield on (say) Treasury bills is so persistent that it may be preferable, when using a stochastic process as a metaphor for the data, to think of it as a realisation of an $I(1)$ process rather than an $I(0)$ one.

Treating a series as $I(1)$ doesn't really mean that we actually believe it's $I(1)$, with all the implied consequences. Provided it makes sense at all to consider an interest rate time series as a realisation of some stochastic process, heaven only knows what process it is; we are merely choosing, from within a limited class of processes (ARIMAs), the most appropriate parameterisation to describe the data. If our data spanned thousands of years, I suspect that the most adequate process to represent the trajectory of interest rates over time from 3000 BCE onwards would be an $I(0)$ one, but I do not believe we will ever be in a position to establish this.[11]

It is a *data representation* problem: with a unit root test, we are not truly deciding whether the process is $I(1)$ or $I(0)$. We are merely deciding whether it is more convenient to represent our data as a stationary or as an integrated process. A metaphor that I find quite useful in conveying my point is that of the curvature of the Earth. Some people believe the Earth is flat simply

[10] This may sound like an obvious remark, but unfortunately this misconception is not uncommon among applied economists: every time someone says "this parameter is not significant, therefore it's zero", a kitten dies.

[11] I originally wrote this sentence many years ago. In the meantime, a few papers have appeared where an extremely competent econometrician analysed a time series for interest rates going as far back as the 14th century. Unsurprisingly, she concludes that interest rates are indeed $I(0)$. There you go.

because there is no compelling reason *for them* to think it's round *given their everyday experience*:[12] the Earth's curvature becomes relevant only when dealing with large distances, so much so that, for instance, it must be taken into account when calculating the routes of transoceanic ships or planes. However, if one needs to build a house, or even a stadium, a car park, or a shopping centre, the scale of the problem is such that the curvature of the globe becomes negligible (and indeed, engineers neglect it without their building collapsing).

In the same way, a stationary process can have a degree of persistence such that its realisations become "evidently" stationary only after a very long time. A unit root test in a finite sample may just fail to reject the null hypothesis because there's not enough evidence against non-stationarity. In this case, the test is simply telling us that it might be better, within our sample, to be on the safe side and difference the series.

Finally, I should mention that, when you have many data series, the idea of testing for unit roots one series at a time is considered outdated by many. Sure, you can proceed with a battery of ADF tests or similar on each of them. However, it may be preferable to start from a multivariate representation (which I will discuss in Chapter 5), which leads to the so-called Johansen test (which I will discuss in Chapter 6).

4.5 Spurious regression

As should be clear from the brief treatment of unit root tests I gave in the previous section, inference with integrated processes may imply results far away from the comfortable certainties that we enjoy in the world of stationarity. This state of things is even more striking when considering a phenomenon known as **spurious regression**.

Let us take two processes, y_t and x_t, defined as:

$$\begin{cases} y_t = y_{t-1} + \eta_t \\ x_t = x_{t-1} + \epsilon_t, \end{cases} \tag{4.12}$$

where η_t and ϵ_t are two mutually independent white noiseprocesses. It is evident that y_t and x_t are two random walks that have very little to do with one another: any element of the y_t sequence is independent of any element of the x_t sequence. Intuitively, one wouldn't expect to find any trace of a statistically significant relationship between y_t and x_t. And in fact, there isn't any, but matters are a bit more intricate that that.

The customary way we have to investigate the possible presence of a relationship between x_t and y_t is by setting up a linear regression model like

$$y_t = \alpha + \beta x_t + u_t, \tag{4.13}$$

and at a first glance, one might think that independence between y_t and x_t implies

1. that the R^2 index should be "low";

[12] Having said this: yes, they ought to see a shrink.

2. that the OLS estimator of β should converge in probability to 0;

3. that a t-test of the significance of β, at least in large samples, should be within the usual acceptance region of the null hypothesis $H_0 : \beta = 0$; put simply, that the t-statistic relating to the coefficient β is between -2 and 2 in most cases (95% of the times, to be exact).

Well, none of these three things occur; on the contrary:

1. the R^2 index converges in distribution to a non-degenerate random variable;

2. the OLS estimator of β converges in distribution to a random variable;

3. a standard t-test for the significance[13] of β leads to rejecting H_0, with a probability that is *increasing* with the sample size (!!!).

With results like this, failing to take into account the possibility of the variable being integrated may lead one to "discover" relationships between variables that are just a by-product of a probabilistic artefact: hence the expression 'spurious regression'[14].

Table 4.2: Spurious regression: Monte Carlo experiment

Sample size	Rejection rate
20	47.11%
50	66.48%
100	76.25%
200	83.44%
1000	92.47%

100000 simulations for each sample size

In order to get a better understanding of what is going on here, take a look at Table 4.2, which contains the results of a simple Monte Carlo experiment: I simulated a system like the one in (4.12), with $E(\epsilon_t^2) = E(\eta_t^2) = 1$, for various sample sizes, ranging from very small ($T = 20$) to rather large ($T = 1000$). In each simulation, I ran a regression of y_t on a constant and x_t, as in (4.13)); then, I computed the value of the t-test for $H_0 : \beta = 0$ and compared it with the appropriate critical value of Student's t distribution. Table 4.2 contains the frequency of rejections: not only this number is much higher than 5% but, amazingly, increases as the sample size grows.

Upon closer inspection, however, these results should not be all that surprising. For $\beta = 0$, equation (4.13) reduces to $y_t = \alpha + u_t$; if y_t is $I(1)$, then u_t must also $I(1)$, but then it doesn't have finite moments and then all the

[13] That is: using the critical values of the standard normal distribution.

[14] The phenomenon had already been observed in the 1920s. It was only during the 1970s and 1980s, however, that it was brought to general attention (thanks to Granger and Newbold) and analysed in depth (thanks to P. C. B. Phillips).

limit theorems used in the elementary cases become totally useless. In other words, there is no value of β that makes equation (4.13) a correct description of the data; the value $\beta = 0$ is no more right or wrong than $\beta = 1$, $\beta = -1$ or $\beta = 1000$; there's no such thing as a "true" β. The problem of equation (4.13) is that there is no mechanism to account for the persistence of y_t, whatever the value of β.

In other words, OLS estimation of equation (4.13) dumps all the persistence of y_t onto the residuals $\hat{u}_t$, which turn out to be highly autocorrelated. In fact, it can be shown that, in the presence of this phenomenon, the Durbin-Watson statistic[15] converges in probability to 0. Actually, a rough but effective alert rule for spurious regression is to compare the R^2 index with the DW statistic. If $R^2 > DW$, there is reason to worry (this should not be taken as a proper test; it is simply a heuristic suggestion contained in the original paper by Granger and Newbold).

Now: OLS can yield misleading results if used with realisations of $I(1)$ processes, and macroeconomic time series typically look like $I(1)$ processes. Does this mean that you cannot use OLS on macroeconomic time series? Not necessarily.

First of all, note that a large part of the seemingly paradoxical aspects described above come from the fact there is no value of β compatible with a correct description of the data, as I said before. If we had estimated something of the type

$$y_t = \alpha + \varphi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

we would have found that

$$\begin{pmatrix} \hat{\varphi} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \xrightarrow{\text{p}} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

that is, the OLS estimates converge to the 'true' values of the parameters.

A correct dynamic specification (that is, a specification that allows the disturbances to resemble a white noise process, or at least a stationary one) is a major step forward. We will discuss this result, and others, in greater detail in Section 5.4. The most important one, however, is that OLS regression between integrated variables *may* make sense, and under certain circumstances represents a quick but effective way of measuring meaningful statistical relationships. This happens when the variables on the left- and right-hand side of the equals sign happen to be **cointegrated**. We'll tackle cointegration in Chapter 6.

[15] The Durbin-Watson statistic is a venerable statistic devised in the prehistory of econometrics to check for first-order autocorrelation in OLS residuals. Back in the day, the DW statistic was also used as a test statistic. Today, we have more sophisticated tools to do the same thing better, but software packages still report it to honour tradition.

4.A Assorted results

4.A.1 The long-run variance

The "long-run variance" is defined as the sum of all its autocovariances (from minus to plus infinity). It is often useful because it provides the asymptotic variance of the average of the realisation of a stochastic process.

This quantity is typically estimated non-parametrically via

$$\hat{\omega}^2(m) = T^{-1} \sum_{t=m}^{T-m} \left[\sum_{i=-m}^{m} w_i \hat{u}_t \hat{u}_{t-i} \right],$$

where m is known as the *window size* and the w_i terms are the so-called *Bartlett weights*, defined by $w_i = 1 - \frac{|i|}{m+1}$. It can be shown that, for a sufficiently large m, $\hat{\omega}^2(m)$ provides a consistent estimate of the long-run variance.

What "sufficiently large" means in practice is not very clear: asymptotic theory only says that, for the estimator to be consistent, m must be proportional to $T^{1/3}$, which in practice translates into "do as you please". What I personally do when in doubt is try different values and make a qualitative assessment.

4.A.2 Gretl script for replicating Table 4.2

```
set verbose off                 # turn off unwanted output
set seed 10082026               # make example replicable

SS = {20,50,100,200,1000}       # sample sizes
Tmax = max(SS)

H = 100000                      # number of replications
D = nelem(SS)

RR = zeros(1, D)                # vector for storing rejections
R2 = zeros(H, D)                # matrix for storing R2 indices

nulldata Tmax --preserve        # create a fake dataset

loop h = 1 .. H
    # generate y and x as independent random walks
    y = cum(normal())
    x = cum(normal())

    loop i = 1 .. D
        # for the various sample sizes stored in SS
        T = SS[i]
        # use the first T observations only
        smpl 1 T
        # do OLS quietly and store results
        ols y 0 x --quiet
        R2[h,i] = $rsq          # R^2
        restrict                # significance test for beta
            b2 = 0
        end restrict --silent
        RR[i] += ($pvalue < 0.05)
    endloop
    smpl full                   # restore full sample
endloop

RR = 100 * RR./H                # compute percentage frequency of rejections
```

```
loop i = 1 .. D
    printf "%d\t%5.2f%%\n", SS[i], RR[i]
endloop
```

Chapter 5

VAR processes

5.1 Multivariate processes

Complex phenomena, such as those we are usually interested in as economists, can seldom be described by a single variable. Far more commonly, the facts of interest must be described by more that one quantity. In a static context, this naturally leads to using the concept of multivariate random variables, aka random vectors. The techniques for studying realisations of random vectors are countless, among which regression analysis has a prominent place.

With time-series data, we need to define to extend the setup described in Chapter 1 to **multivariate stochastic processes**. As always, I shall not even try to be rigorous here: suffice it to say that a multivariate stochastic process is a stochastic process whose elements are not single random variables, but vectors; alternatively, one can think of a multivariate stochastic process as a vector whose elements are univariate stochastic processes. For instance, if we think of the daily recording of the euro/dollar exchange rate as the realisation of a univariate process, we can think of the daily recording of the exchange rates for euro/dollar, euro/yen, euro/Swiss franc, *etc.*, as the realisation of a multivariate process. Therefore, if y_t is an n-element vector, we call it a stochastic process of dimension, or size, n.

This definition makes it painless to extend to the multivariate case several concepts I put forward in Chapter 1: for instance, the definitions of stationarity, ergodicity and mixing remain unchanged. Moments are more interesting: for an n-dimensional weakly stationary process y_t the first and second moments are defined as follows:

$$\begin{cases} E(y_t) = \mu \\ E\left[(y_t - \mu)(y_{t-k} - \mu)'\right] = \Gamma_k, \end{cases}$$

where μ is an $n \times 1$ vector and Γ_k (the k-order **autocovariance matrix**) is an $n \times n$ matrix. Of course, Γ_0, is just the covariance matrix of y_t; for $k \neq 0$, the ij-th element of Γ_k is the covariance between the i-th element of y_t and the j-th element of y_{t-k}. Note that, by this definition, $\Gamma_k = \Gamma'_{-k}$ and therefore, in general, $\Gamma_k \neq \Gamma_{-k}$.

Likewise, the definition of multivariate white noise is rather straightforward: we define a process ϵ_t to be white noise if

$$\begin{aligned} E(\epsilon_t) &= 0 \\ \Gamma_k = E(\epsilon_t \epsilon_{t-k}) &= \begin{cases} \Sigma & \text{if } k = 0 \\ 0 & \text{if } k \neq 0. \end{cases} \end{aligned}$$

It should be noted that Σ is a generic covariance matrix, and is therefore symmetric and positive semi-definite, but not necessarily diagonal. Consequently, in a multivariate white noise process, the correlation between any element of the process and the past history of the *entire* process is by definition 0, but the correlation between *contemporaneous* elements may well be non-zero.

Similarly, the behaviour of the lag operator L needs no further qualification: $Lx_t = x_{t-1}$ even when x_t is a vector. Matters become interesting if we consider expressions of the type

$$x_t + Ax_{t-1} = (I + AL)x_t,$$

where A is a square matrix. In this case, the expression $(I + AL)$ is a matrix operator, a function of the lag operator L. Such an object can be viewed in two equivalent ways:

Polynomial with matrix coefficients: the operator $(I + AL)$ is the sum of two matrices, each of which "multiplies" the operator L by a different power. Therefore, one can think of $(I + AL)$ as a first-order polynomial in the lag operator L, where the coefficients are matrices: the identity matrix for L^0 and the matrix A for L^1;

Matrix of polynomials: the operator $(I + AL)$ is a matrix whose elements are first-order polynomials: the ij-th element of $(I + AL)$ is $\delta_{ij} + a_{ij}L$, where δ_{ij} is the so-called 'Kronecker delta', which equals 1 for $i = j$ and 0 otherwise.

Generalisation to the p-th order case should be immediate, so that the meaning of an expression like

$$C(L)x_t = C_0 x_t + C_1 x_{t-1} + \cdots + C_p x_{t-p}$$

should be obvious.

As for inversion, this is easy to define by considering $C(L)$ as a matrix of polynomials: the inverse operator $C(L)^{-1}$, if it exists, is a matrix polynomial such that

$$C(L)C(L)^{-1} = C(L)^{-1}C(L) = I$$

and, at least in principle, can be computed using standard algebraic techniques.

Example 5.1.1 (Matrix polynomial inversion) *Consider*

$$C(L) = \begin{pmatrix} 1 & aL \\ b & 1 - cL \end{pmatrix}.$$

By applying standard rules for inverting 2×2 *matrices, you get*

$$C(L)^{-1} = \frac{1}{1-\varphi L} \begin{pmatrix} 1-cL & -aL \\ -b & 1 \end{pmatrix},$$

where $\varphi \equiv ab + c$*, so that, for example,*

$$\left[C(L)^{-1}\right]_{1,1} = \frac{1-cL}{1-\varphi L}.$$

Note that, for the inverse to exist, the determinant $1 - \varphi L$ *must be itself invertible, which happens if and only if* $|\varphi| = |ab + c| < 1$*. In this case,*

$$\frac{1}{1-\varphi L} = 1 + \varphi L + \varphi^2 L^2 + \dots$$

and therefore

$$C(L)^{-1} = \left[1 + \varphi L + \varphi^2 L^2 + \dots\right] \begin{pmatrix} 1-cL & -aL \\ -b & 1 \end{pmatrix}.$$

Neat, huh?

Fun fact: unlike in the univariate case, the inverse of a matrix polynomial of finite order is not necessarily of infinite order. Consider this example:

$$\begin{aligned} y_t &= \theta x_{t-1} + \epsilon_{1,t} \\ x_t &= \epsilon_{2,t}. \end{aligned}$$

As one can easily verify, the inverse of

$$C(L) = \begin{pmatrix} 1 & -\theta L \\ 0 & 1 \end{pmatrix}$$

is

$$C(L)^{-1} = \begin{pmatrix} 1 & \theta L \\ 0 & 1 \end{pmatrix}$$

and in fact, it can be verified by direct substitution that the first equation can be written as

$$y_t = \epsilon_{1,t} + \theta \epsilon_{2,t-1}.$$

It can be proven that this is a consequence of the fact that $|C(z)| = 1$.

Finally, it should be noted that a multivariate version of Wold's decomposition theorem exists, which guarantees that any zero-mean, weakly stationary multivariate process y_t has a representation like

$$C(L)x_t = C_0\epsilon_t + C_1\epsilon_{t-1} + C_2\epsilon_{t-2} + \cdots$$

where the (matrix) coefficients are square-summable.

5.2 VAR processes

In principle, the most obvious multivariate generalisation of the ARMA processes we considered in Chapter 2 would be a class of processes known as **Vector ARMA**, or **VARMA** for short. These take the following form:

$$\Phi(L)y_t = \Theta(L)\epsilon_t,$$

where y_t is an n-dimensional stochastic process, ϵ_t is a white noise process of the same size, and the two objects $\Phi(L)$ and $\Theta(L)$ are appropriately-sized matrix polynomials, of order p and q respectively. VARMA processes, however,

are not very popular (at least in econometrics) because when $q > 0$ the practical issues that arise in estimating them are decidedly non-trivial. The good news is that all these issues magically disappear when $q = 0$, so economists routinely work with models like

$$A(L)y_t = \epsilon_t \rightarrow y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t,$$

that are known as **VARs** (Vector AutoRegressive), so the one above would be a p-order VAR.

As we will see, working with VARs is very simple in many respects: first (as we will see later) they can be estimated simply and consistently using OLS.[1] In fact, many features of AR processes extend in a rather trivial manner to VARs.

One possibly tricky question is: it is possible, like in the univariate case, to express a VAR process as a multivariate moving average (VMA) process? The answer is clearly linked to the invertibility of the operator $A(L)$, which leads us to "what conditions for the operator $A(L)$ being invertible?". If it were, we could legitimately consider the VMA representation as a Wold representation, thereby ensuring stationarity of the process.

In the case of univariate processes, we saw in Section 2.4 that we needed to check if the roots of $A(L)$ were all greater (in absolute value) than 1. Let's see if we can come up with a similar criterion. For simplicity, let us then consider a first-order VAR process:

$$y_t = Ay_{t-1} + \epsilon_t. \tag{5.1}$$

In this case, $A(L) = I - AL$ is a first-degree matrix polynomial in L. Observe that $y_{t-1} = Ay_{t-2} + \epsilon_{t-1}$, and substitute into (5.1):

$$y_t = A^2 y_{t-2} + \epsilon_t + A\epsilon_{t-1},$$

where $A^2 = A \cdot A$; repeating this procedure n times yields

$$y_t = A^{n+1} y_{t-n-1} + \epsilon_t + A\epsilon_{t-1} + \cdots + A^n \epsilon_{t-n}.$$

As n increases, the first term vanishes if $\lim_{n\to\infty} A^n = 0$; this condition is equivalent to all the eigenvalues[2] of A being less than 1 in absolute value (note that this is just a generalisation of the corresponding criterion for AR(1) processes, where the MA representation is well-defined if $|\alpha| < 1$). It can be shown that this condition on the eigenvalues of A is necessary and sufficient for the process to be covariance-stationary. Since the determinant of a matrix is equal to the product of its eigenvalues, this condition can also be equivalently expressed by saying that the process is stationary if $|A(z)| = 0$ has no solutions for $|z| \leq 1$. Therefore, if the process is stationary, we can define the VMA representation of y_t as

$$y_t = \epsilon_t + A\epsilon_{t-1} + \cdots = \sum_{i=0}^{\infty} A^i \epsilon_{t-i}.$$

[1] Some readers will remember that in Section 3.3 I showed that OLS estimation of univariate AR(p) models can be legitimately be considered a perfectly valid variant of ML estimation.

[2] Allow me to remind you: the eigenvalues of a square matrix A are the values of λ that satisfy the expression $|A - \lambda I| = 0$).

The general VAR(p) case is more intricate algebraically. Fortunately, there is a way to express any VAR process of order p as a VAR(1), which is known as the **companion form** representation. Let us consider, for example, a VAR(3) process

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + A_3 y_{t-3} + \epsilon_t$$

where $V(\epsilon_t) = \Sigma$, and consider the two identities

$$\begin{aligned} y_{t-1} &= y_{t-1} \\ y_{t-2} &= y_{t-2}. \end{aligned}$$

These three relationships can be combined into a 3-equation system, that can be written in matrix form as follows:

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & A_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \\ 0 \end{bmatrix} \tag{5.2}$$

or, in a more compact form,

$$x_t = \overline{A} x_{t-1} + \eta_t,$$

where

$$\begin{aligned} x_t &= \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \end{bmatrix} \\ \overline{A} &= \begin{bmatrix} A_1 & A_2 & A_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \\ \eta_t &= \begin{bmatrix} \epsilon_t \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

The matrix $\overline{A}$ is known as the **companion matrix**.

Equation 5.2 defines a VAR(1) for the process x_t, which is the result of the vertical stacking of y_t, y_{t-1}, and y_{t-2}. If A is a square matrix of order n, then the companion matrix $\overline{A}$ is a square matrix of order $3n$, and η_t is a multivariate white noise process whose covariance matrix is

$$V(\eta_t) = \begin{bmatrix} \Sigma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

which is singular, but symmetric and positive semi-definite. Building on this representation, the stationarity condition is a condition imposed on the $3n$ eigenvalues of $\overline{A}$.

The generalisation to the case of a VAR process of order p should be trivial: in this case, the companion matrix has the following structure:

$$\overline{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_p \\ I & 0 & \cdots & 0 \\ 0 & I & \cdots & 0 \\ \vdots & \vdots & & \ddots \end{bmatrix},$$

which has $n \cdot p$ eigenvalues: all must be less than 1 in modulus for the VAR process to be stationary.

Example 5.2.1 (An AR(2) process rewritten as a VAR(1) process) *For this example, we'll use the trivial fact that a univariate AR process can be thought of as a special case of a VAR process of dimension 1. Let us take an AR(2) process, rewrite it in companion form and check the equivalence of the stationarity conditions:*

$$y_t = 1.3y_{t-1} - 0.4y_{t-2} + \epsilon_t.$$

Following the arguments in Section 2.4, the polynomial whose roots we must find is $A(z) = 1 - 1.3z + 0.4z^2$; since the polynomial is of degree two, $A(z) = 0$ is easily solved using the familiar formula

$$z = \frac{1.3 \pm \sqrt{1.69 - 1.6}}{0.8} \Longrightarrow \begin{cases} z_1 = 2 \\ z_2 = 1.25. \end{cases}$$

The roots of the polynomial are both greater than 1 in absolute value. The process is stationary.

Now use the companion form representation: we have

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1.3 & -0.4 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}$$

that is, a VAR(1) process in x_t, where

$$x_t = \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix}.$$

The process is stationary if the eigenvalues of A are less than one in absolute value. The eigenvalues of A are the numbers λ satisfying the determinantal equation

$$|A - \lambda I| = \begin{vmatrix} 1.3 - \lambda & -0.4 \\ 1 & -\lambda \end{vmatrix} = 0,$$

which implies

$$\lambda^2 - 1.3\lambda + 0.4 = 0,$$

whose solutions are

$$\lambda = \frac{1.3 \pm \sqrt{1.69 - 1.6}}{2} \Longrightarrow \begin{cases} \lambda_1 = 0.8 \\ \lambda_2 = 0.5. \end{cases}$$

Note that $\lambda_1 = z_2^{-1}$ and $\lambda_2 = z_1^{-1}$ (this is no coincidence). Since both eigenvalues are less than 1 in absolute value, we conclude that the process is stationary.

Following on the considerations made in Section 5.1, it is also instructive to inspect the inverse of the polynomial matrix implicit in the companion form representation, since it gives us back y_t as a univariate AR(2) process: from

$$\left(I - \begin{bmatrix} 1.3 & -0.4 \\ 1 & 0 \end{bmatrix} L \right) x_t = \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix},$$

one can write

$$x_t = \begin{bmatrix} 1-1.3L & 0.4L \\ -L & 1 \end{bmatrix}^{-1} \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}.$$

Applying the rules for matrix inversion gives

$$\begin{bmatrix} 1-1.3L & 0.4L \\ -L & 1 \end{bmatrix}^{-1} = \frac{1}{1-1.3L+0.4L^2} \begin{bmatrix} 1 & -0.4L \\ L & 1-1.3L \end{bmatrix},$$

from which

$$(1-1.3L+0.4L^2) \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & -0.4L \\ L & 1-1.3L \end{bmatrix} \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix},$$

therefore, $y_t = 1.3y_{t-1} - 0.4y_{t-2} + \epsilon_t$.

The VMA representation is also easy to calculate using the companion matrix. I will illustrate the procedure with a VAR(2) process, since once the principle is understood, its generalisation is trivial. Start from:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \epsilon_t,$$

where the matrices A_1 and A_2 are known. We wish to calculate the sequence of matrices C_i of its VMA representation, namely

$$y_t = \epsilon_t + C_1\epsilon_{t-1} + C_2\epsilon_{t-2} + \cdots$$

As we have seen above, for a VAR(1) process the C_i matrices obey $C_i = A^i$. Consequently, rewriting the VAR process in companion form as in (5.2) yields

$$x_t = \overline{A}x_{t-1} + \eta_t = \eta_t + \overline{A}\eta_{t-1} + \overline{A}^2\eta_{t-2} + \cdots$$

or

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix} + \overline{A}\begin{bmatrix} \epsilon_{t-1} \\ 0 \end{bmatrix} + \overline{A}^2\begin{bmatrix} \epsilon_{t-2} \\ 0 \end{bmatrix} + \cdots;$$

it can be worked out, from the expression above, that C_i is simply the $n \times n$ block occupying the top-left corner of $\overline{A}^i$; in practice, to calculate the first k elements of the sequence of matrices C_i, you just build the companion matrix, multiply it by itself k times, and extract the top-left corner each time.

Example 5.2.2 *Suppose that*

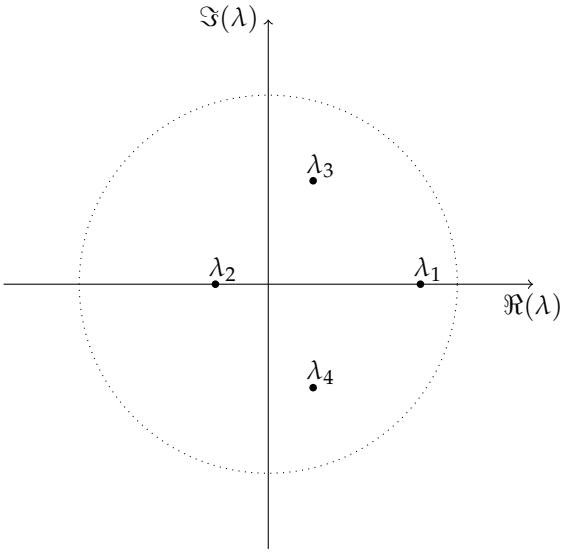
$$A_1 = \begin{bmatrix} 0.8 & -0.6 \\ 0.2 & 0.2 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.1 & 0.3 \\ -0.2 & -0.2 \end{bmatrix}.$$

Therefore, the companion matrix is

$$\overline{A} = \begin{bmatrix} 0.8 & -0.6 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & -0.2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Let's check for stationarity first, to make sure the Wold representation exists: using

Figure 5.1: Eigenvalues of the companion matrix



appropriate software, it takes relatively little effort to establish that the eigenvalues of $\overline{A}$ *are*

$$\lambda = [0.80481, \quad -0.27953, \quad 0.23736 \pm 0.54705i].$$

Two of them are complex, so it's nice to represent them graphically as in Fig. 5.1.[3] *They all lie within the unit circle (identified by a dotted line), hence the VAR process is stationary, and we can calculate the moving average representation.*

By using, again, appropriate software, the first powers of the companion matrix

[3] If you're in trouble with this, check out Section 2.A.2.

can be computed as

$$
\begin{aligned}
\overline{A}^0 &= I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
\overline{A}^1 &= \overline{A} = \begin{bmatrix} 0.8 & -0.6 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & -0.2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \\
\overline{A}^2 &= \overline{A} \cdot \overline{A} = \begin{bmatrix} 0.62 & -0.3 & -0.04 & 0.36 \\ 0.4 & -0.28 & 0.06 & 0.02 \\ 0.8 & -0.6 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & -0.2 \end{bmatrix} \\
\overline{A}^3 &= \overline{A} \cdot \overline{A} \cdot \overline{A} = \begin{bmatrix} 0.396 & -0.072 & 0.002 & 0.246 \\ 0.324 & -0.276 & -0.016 & 0.176 \\ 0.62 & -0.3 & -0.04 & 0.36 \\ 0.4 & -0.28 & 0.06 & 0.02 \end{bmatrix}.
\end{aligned}
$$

See? Since the companion matrix has an identity matrix in its South-West corner, the top 2 rows shift downwards to the bottom at each iteration.

By extracting the respective 2×2 *North-West blocks, we obtain the matrices of the MA representation:*

$$
C_0 = I \quad C_1 = \begin{bmatrix} 0.8 & -0.6 \\ 0.2 & 0.2 \end{bmatrix} \quad C_2 = \begin{bmatrix} 0.62 & -0.3 \\ 0.4 & -0.28 \end{bmatrix} \quad C_3 = \begin{bmatrix} 0.396 & -0.072 \\ 0.324 & -0.276 \end{bmatrix}
$$

Evidently, this procedure is a classic case where calculations are messy and tedious to do by hand, but carried out swiftly by software.

5.3 VAR estimation

To begin from the end: the parameters of a VAR process can be consistently estimated via a battery of OLS regressions. Let us see why: if you unravel the matrix notation I used from equation (5.1) onwards, an n-variate VAR process of order p can be considered as a system of n equations of the form

$$
y_{it} = \sum_{j=1}^{p} \left(a_{i1j} y_{1t-j} + \cdots + a_{inj} y_{nt-j} \right) + \epsilon_{it}. \tag{5.3}
$$

For $n = 2$ and $p = 1$, we would have, for example,

$$
\begin{aligned}
y_{1t} &= a_{11} y_{1t-1} + a_{12} y_{2t-1} + \epsilon_{1t} \\
y_{2t} &= a_{21} y_{1t-1} + a_{22} y_{2t-1} + \epsilon_{2t}.
\end{aligned}
$$

Since p is normally unknown, people usually employ the same methods I discussed in section 3: information criteria, optionally supplemented by auto-correlation tests. From now on, let us assume that the order of the VAR is known.

Once a decision is reached on the VAR order p, each of the n equations in system (5.3) could be viewed as a dynamic regression model (see the discussion at the end of subsection 3.3); in this case, it can be shown that applying OLS yields consistent and asymptotically normal estimates of all the a_{ij} parameters.

From an econometric perspective, much insight can be gained by interpreting a VAR as the reduced form of a simultaneous equation system[4] with *no* identification restrictions, where all the elements of the vector y_t are the endogenous variables and their lags are exogenous. The aim of a VAR model is very often to find a statistically accurate description of the persistence characteristics of a set of time series, without the need for a structural interpretation of the estimated parameters. It is for this reason that, when they first appeared on the scene, VAR models were labelled as "atheoretical" models.

Additional regressors are not usually are a problem: therefore, VAR models usually employed in practice contain deterministic components that account for certain characteristics of the data — such as a constant if the y_{it} series have a non-zero mean, a trend, or dummy variables to indicate exceptional events or seasonal effects.

Estimation of the matrix Σ is fairly straightforward too. You just get the OLS residuals from each equation $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ and compute their sample covariance matrix. It can be proven that

$$\frac{1}{T}\hat{\epsilon}_i'\hat{\epsilon}_j \xrightarrow{\text{p}} \Sigma_{ij}$$

is a consistent estimator of the ij-th element of the covariance matrix of ϵ_t.

5.3.1 An example

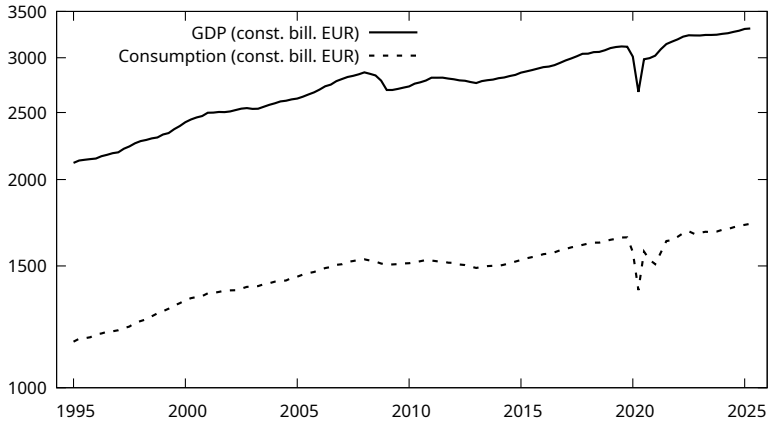
Let us take quarterly data on income and consumption for the Euro Area from 1995:1 to 2025:3 (the data come from `https://mahmuts.bilkent.edu.tr/`, and are at constant prices and seasonally adjusted). The time series are shown in Figure 5.2. The three most evident stylised fact evident when inspecting the time plot are:

1. Both series are trending upwards, and seem to be roughly parallel.
2. GDP takes a small hit in 2008, at the time of the global financial crisis; consumption, maybe not so much. Same pattern (but smaller) for the 2012 fiscal crisis.
3. The COVID pandemic affected both series quite dramatically, but it would appear to have been a transitory episode: after a while, both series go "back to normal".

We take (as usual) the logarithms of the data and decide to insert eight deterministic terms: a constant, a time trend (to account for the trending

[4] In the early days of econometrics, simultaneous equation models were the hot thing. Nowadays, they are considered somehow *passé*, so I cannot assume my readers know what I'm talking about. If you need some help, have a look at Section 5.A.1 in the appendix to this chapter.

Figure 5.2: GDP and consumption in the Euro Area



behaviour), plus 6 dummies for each quarter from 2020:1 to 2021:3, so that the events during the pandemic are treated as a special period, structurally different from the rest. In practice, the deterministic part for each equation is

$$d_{i,t} = \beta_{i,0} + \beta_{i,1} t + \sum_{j=1}^{6} \gamma_{i,j} C_{i,j},$$

where $C_{i,1}$ is a dummy variable for 2020:1, $C_{i,2}$ is a dummy variable for 2020:2, and so on.

As for the choice of the VAR order, the information criteria give contrasting answers: BIC selects 5 lags, HQC chooses 6 and AIC goes up to 9 quarters. In the absence of a clear indication, since this is just an example, I decided to go for simplicity and choose $p = 2$, so bear in mind that the subsequent result should be treated as illustrative and are not meant to be taken as "the truth".[5] In practice, we will assume that our data are a realisation of the following stochastic process:

$$\begin{bmatrix} c_t \\ y_t \end{bmatrix} = \begin{bmatrix} d_{1,t} \\ d_{2,t} \end{bmatrix} + A_1 \begin{bmatrix} c_{t-1} \\ y_{t-1} \end{bmatrix} + A_2 \begin{bmatrix} c_{t-2} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix},$$

where, obviously, A_1 and A_2 are a 2×2 matrix of parameters. These, plus the parameters in the deterministic terms, can be estimated using ordinary least squares. We arrive at the results shown in Table 5.1, where the deterministic parts are omitted for simplicity. The estimated A_1 and A_2 matrices are, therefore,

$$\widehat{A}_1 = \begin{bmatrix} 1.296 & 0.152 \\ 0.326 & 1.347 \end{bmatrix} \quad \widehat{A}_1 = \begin{bmatrix} -0.366 & -0.106 \\ -0.202 & -0.504 \end{bmatrix}$$

[5] Those "in the know" may now complain I've taken no special precautions for unit roots, despite all the fuss I made about them in Chapter 4. I'll take care of those in Chapter 6. What I'm trying to focus on here are the mechanics of a VAR model, so be lenient.

and the companion matrix is

$$\overline{A} = \begin{bmatrix} 1.296 & 0.152 & -0.366 & -0.106 \\ 0.326 & 1.347 & -0.202 & -0.504 \\ 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \end{bmatrix}$$

Table 5.1: OLS results

Equation for c_t				
	Coeff.	S.E.	t-stat.	p-value
c_{t-1}	1.29606	0.0999128	12.97	6.58e-24
c_{t-2}	-0.366140	0.0859935	-4.258	4.36e-05
y_{t-1}	0.152004	0.0908519	1.673	0.0972
y_{t-2}	-0.105747	0.0882436	-1.198	0.2334
	$R^2 = 0.9986$		$\hat{\sigma} = 0.003977$	

Equation for y_t				
	Coeff.	S.E.	t-stat.	p-value
c_{t-1}	0.326256	0.105193	3.101	0.0024
c_{t-2}	-0.201944	0.0905382	-2.230	0.0277
y_{t-1}	1.34705	0.0956534	14.08	2.27e-26
y_{t-2}	-0.503626	0.0929072	-5.421	3.55e-07
	$R^2 = 0.9961$		$\hat{\sigma} = 0.00521754$	

Finally, by calculating the sample covariance matrix of the residuals from the two equations above, we obtain the estimate of the matrix Σ:

$$\hat{\Sigma} = \begin{bmatrix} 1.426 \cdot 10^{-5} & 0.814 \cdot 10^{-5} \\ 0.814 \cdot 10^{-5} & 1.581 \cdot 10^{-5} \end{bmatrix},$$

5.4 Unit root VARs

In Chapter 4, we explored in some detail the characteristics of integrated processes and the reasons why they are so important in macroeconomics. We now turn to the implications of unit roots in VAR processes.

As argued in Section 5.2, we know that an n-variable VAR process $A(L)y_t = \epsilon_t$ is stationary if the solutions to the equation are all greater than one in modulus; an equivalent condition applies to the eigenvalues of the companion matrix, which must all be less than one. What happens if $|A(z)| = 0$ has some solutions for $|z| = 1$?

For a univariate process, either $A(1)$ is zero, or it is not. There's no other possibility. In a multivariate framework, instead, it is important to establish *how many* solutions to $|A(z)| = 0$ exist or, equivalently, how many unit eigenvalues the companion matrix has. As we already know, if all the eigenvalues are less than 1, then the process is stationary. OK. If, on the other hand, there are one or more unit eigenvalues, several cases of interest arise.

The first case occurs when $A(1)$ is a matrix of zeros. In this scenario, it is convenient to employ the BN decomposition I introduced in Section 4.3, which is also valid for multivariate processes:

$$A(L) = A(1) + A^*(L)\Delta,$$

where, obviously, $A(L)$ is a matrix polynomial of order p, and, as a consequence, $A(1)$ and $A^*(L)$ are also matrices. If we apply this decomposition to our VAR process, we obtain

$$A^*(L)\Delta y_t = \epsilon_t,$$

so the process Δy_t has a VAR($p-1$) representation (because the order of $A^*(L)$ is $p-1$).

If Δy_t is stationary, we conclude that y_t becomes stationary after one differentiation, and is therefore $I(1)$. Under these circumstances, the obvious course of action is performing our empirical analysis on the time series in differences rather than in levels. In this case, it can be shown[6] that the number of unit eigenvalues of the companion matrix is exactly n.

It is worth noting that the BN decomposition can also be used on Δy_t in Wald form: starting from

$$\Delta y_t = C(L)\epsilon_t$$

and using equation (4.3),

$$C(z) = C(1) + C^*(z)(1-z),$$

we can write the process y_t as

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t = P_t + T_t,$$

where μ_t is a multivariate random walk.

In these cases, naturally, a Wold representation does not exist, but nothing prevents us from calculating a non-Wold moving average representation for y_t via the impulse response function as a sequence of matrices M defined as

$$(M_n)_{ij} = \frac{\partial y_{it}}{\partial \epsilon_{jt-n}}.$$

It can be shown that the sequence M_n does not converge to 0 like in the stationary case, but rather to a non-zero limit matrix; evidently, this prevents square-summability. Of course, this is a generalisation of what happens in the univariate $I(1)$ case, where the impulse response function remains indefinitely at a level different from 0.

Example 5.4.1 *Consider the following VAR(2) process:*

$$\begin{aligned} y_t &= 1.1y_{t-1} - 0.1x_{t-1} - 0.1y_{t-2} + 0.1x_{t-2} + \epsilon_{1t} \\ x_t &= 1.3x_{t-1} - 0.3x_{t-2} + \epsilon_{2t}, \end{aligned}$$

[6] Warning: the converse is not true. One can construct examples of companion matrices with n unit eigenvalues without the resulting process being $I(1)$.

and start from its companion matrix:

$$\overline{A} = \begin{bmatrix} 1.1 & -0.1 & -0.1 & 0.1 \\ 0 & 1.3 & 0 & -0.3 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Its eigenvalues are 1, 1, 0.3, and 0.1.[7] *We have two unit eigenvalues, with the remaining ones being smaller than 1, so we definitely have unit roots.*

The $A(L)$ polynomial can be written as

$$A(L) = \begin{bmatrix} 1 - 1.1L + 0.1L^2 & 0.1L - 0.1L^2 \\ 0 & 1 - 1.3L + 0.3L^2 \end{bmatrix}.$$

Evaluating this matrix polynomial at 1, you get

$$A(1) = \begin{bmatrix} 1 - 1.1 + 0.1 & 0.1 - 0.1 \\ 0 & 1 - 1.3 + 0.3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Therefore, by applying the BN decomposition,

$$A(L) = A(1) + A^*(L)\Delta = 0 + \begin{bmatrix} 1 - 0.1L & 0.1L \\ 0 & 1 - 0.3L \end{bmatrix}\Delta,$$

so that

$$\begin{aligned} \Delta y_t &= 0.1\Delta y_{t-1} - 0.1\Delta x_{t-1} + \epsilon_{1t} \\ \Delta x_t &= 0.3\Delta x_{t-1} + \epsilon_{2t}. \end{aligned}$$

which is a nice, stationary VAR(1) process.

Finally, if we wish to represent the process in VMA form, by applying the little trick of calculating the powers of $\overline{A}$ as in Example 5.2.2, it turns out that

$$C_0 = I, \quad C_1 = \begin{bmatrix} 1.1 & -0.1 \\ 0 & 1.3 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 1.11 & -0.14 \\ 0 & 1.39 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 1.111 & -0.153 \\ 0 & 1.417 \end{bmatrix}$$

and so forth; anyone wishing to continue will find that C_n does not converge to 0 when $n \to \infty$.

Matters would become more interesting if the number of unit eigenvalues were between 0 and n. In this case, $A(1)$ is non-zero but not invertible either, so its rank is a number smaller than n but greater than 0. This leads to something called **cointegration**, where the process is $I(1)$ but possesses peculiar characteristics that we will examine in Chapter 6. A wholly separate chapter is necessary because the reduced rank of $A(1)$ has major implications for inference as well as for the interpretation of the process behaviour, which becomes very interesting. Understandably, cointegration was *the* hot topic in time series econometrics at the end of the XX century.

Fortunately, the possible presence of integration causes fewer disasters for inference than one might fear. To be more specific:

[7] Computing them by hand is possible, but long and excruciating. If you want to check by yourself, use some appropriate software.

- the use of information criteria for selecting the VAR order still works OK, regardless of stationarity of the VAR process;
- OLS yields consistent parameter estimates in all cases, even though the limiting distribution of the OLS statistics may become non-standard in the presence of unit roots. If we are only interested in a point estimate (for example, when we use the estimated VAR merely as a forecasting machine), this issue is moot. If, however, we need to run hypothesis testing (such as the so-called Granger-causality test, which I will discuss in Section 5.5.2), then one must proceed with caution.

These results (among many others) are all contained in a paper (Sims et al., 1990), which in my opinion is one of the finest time series econometrics papers ever written. To cut a long story short, the paper shows that the consistency of OLS estimators for VAR parameters is not jeopardised by unit roots, provided that the lag structure of the estimated VAR is rich enough to allow the error terms to be — more or less — white noise. The same paper also offers a highly interesting discussion on the implications that the presence of unit roots has for the distribution of these estimators, and the distributional properties of the DF and ADF tests emerge, most elegantly, as a special case.

5.5 Usage of VARs

VARs — like their univariate counterparts — are widely used for forecasting and for analysing the dynamic characteristics of time series. As I'll show you, many of the concepts we will use are natural extensions of those I already described in the context of univariate processes, but sometimes the multivariate nature of VARs gives rise to interesting situations that we need to examine in detail.

The three most common applications in macroeconometrics are:

1. forecasting,
2. causality analysis,
3. dynamic analysis (aka "Structural VARs").

I'll give he first two a rather cursory treatment. Structural VARs, instead, deserve more space and therefore are postponed to section 5.6.

5.5.1 Forecasting

Here, we don't need much more than what I said in Section 3.4.1. The one remark we need to make is that the information set $\mathcal{F}_{t-1}$ used for producing the forecast includes the past of all the time series that make up the VAR, rather than just one.

Assuming (as usual) that our weapon of choice is the conditional expectation, once the matrices $\hat{A}_1, \ldots, \hat{A}_p$ have been estimated using the technique

described above, the forecast for y_{T+k} is given by a simple generalisation of the univariate case

$$\hat{y}_{T+k} = \hat{A}_1 \hat{y}_{T+k-1} + \cdots + \hat{A}_p \hat{y}_{T+k-p}, \tag{5.4}$$

where $\hat{y}_{T+k} = y_{T+k}$ for $k \leq 0$. It is possible — by methods I won't describe in the main text, but I illustrate in Section 5.A.2 — to calculate confidence intervals for the forecasts, so that the whole job can be carried out under a proper statistical setup.

As you can see, the matter is rather simple. It should come as no surprise that VAR analysis has been the standard tool for forecasting in small-scale macroeconomic models at least for forty years, since the parameters contained in the matrices A_i can be simply estimated via OLS. Obviously, this is not the whole story. There are more sophisticated ways of working with VARs when it comes to forecasting, but they lie beyond the scope of this book.

Another minor digression (a close relative of the one made in subsection 3.4.1). The use of VARs as a tool for forecasting is the subject of a certain degree of irony within the profession. Indeed, a "baseline" VAR model like those presented here constitutes a very drastic simplification of empirical facts, which is ultimately inadequate for capturing complex phenomena; suffice it to think, for example, of that in a VAR there is no simple way to incorporate *a priori* information that we might have about the future, such as policy regime changes and so on.

A famous in-joke goes like this:

> Forecasting is like trying to drive a car blindfolded and following directions given by a person who is looking out of the back window.

In fact, this is a bit unfair: no one would dream of labelling as mechanical or ridiculous a statement like "a trade balance surplus today will contribute to reducing unemployment in a year's time, to the extent that past surpluses (deficits) have influenced unemployment one year later". Certainly, such a statement is partial and mechanical, and does not take many things into account, but it can make for an excellent basis for finer reasoning.
A VAR model is a tool for making this kind of reasoning "semi-automatic". A forecast obtained with a model always needs to be assessed in light of the qualitative characteristics of the phenomenon. In this sense, one would be almost tempted to argue the seemingly paradoxical point that wrong forecasts are more useful than correct ones, because it is through the former that we receive signals as to whether the world is changing (or whether, more trivially, our forecasting model has a problem).

Example 5.5.1 (GDP and consumption in the Euro Area) *Let's see how the VAR model we estimated in Section 5.3.1 fares in terms of forecasting power. We used data up to 2025:2, so we will try to forecast the two series up until the end of 2025. Since in this example a VAR(2) was used, all we need for computing our forecasts are the last two observed values, that is:*

Period	y_t	c_t
2025:01	*7.4500*	*8.1023*
2025:02	*7.4536*	*8.1037*

With these in hand, the forecasts up to 2025:4 can be computed easily via equation 5.4; of course, you should also take into account the deterministic terms (constant and trend) that I did not show in the example, for the sake of clarity.

The table below shows, for both variables, the actual value, the forecast and a 95% confidence interval for the forecast.

	c_t	$\hat{c}_t$	95% conf. Int.	y_t	$\hat{y}_t$	95% conf. Int.
2025:3	7.4565	7.4562	[7.4490-7.4639]	8.1057	8.1068	[8.0978-8.1136]
2025:4	7.4591	7.4607	[7.4463-7.4719]	8.1080	8.1088	[8.0936-8.1225]

All in all, results look rather good. All point forecasts are very near the actual value, and the 95% confidence interval always include the actual realisations quite comfortably.

5.5.2 Causality analysis

At the beginning of the 21st century, econometrics was more or less revolutionised by the advent of what is commonly called "causal inference". VARs can also be used for a special flavour of causality analysis although, as we will see presently, several qualifications are in order.

In general, cause-and-effect relationships are highly complex to establish in an empirical analysis of economic data. If we observe a high correlation between two variables X and Y, we can at most say that those two variables exhibit a marked tendency to move together, but in the absence of other information, we cannot say anything about the causal links connecting them. It could be that X causes Y, that Y causes X, or even that there is a third (unobserved or unconsidered) variable Z which is the common cause of both. All three situations would give rise to the same observable phenomenon, namely a high degree of correlation between X and Y.

Sometimes economic theory comes to our aid: if, for example, we observe that the price of a good systematically increases as the quantity traded grows, we could suggest an interpretation based on a rightward shift of the demand curve, corresponding to a stable supply curve. In this case, we may argue that it was the increase in demand that caused the quantity to rise and, *consequently*, the price of the good.

In many circumstances, however, theory does not offer unambiguous indications: in such cases, there is a definition of causality that gives the possibility of determining the direction of the cause-and-effect relationship on purely statistical grounds, and it is based on the following principle: **the cause always precedes the effect**. In other words, it is assumed that if X causes Y, the causal link requires a minimum amount of time to occur, during which we observe the movement of X, and only afterwards its effect, namely the movement of Y. Conversely, if X did not cause Y, changes in X should not produce systematic changes in the *future* values of Y.

To be more precise, **Granger-causality**, can be defined as follows:[8]

$$X \rightsquigarrow Y \iff E(y_t|y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots) \neq E(y_t|y_{t-1}, y_{t-2}, \ldots),$$

where the $\rightsquigarrow$ symbol is taken to mean "causes in the Granger sense", or

[8]This definition was introduced in the 1970s by C.W.J. Granger (2003 Nobel prize winner), and is therefore named after him. Granger, however, pointed out repeatedly that he had simply introduced a definition by the mathematician Norbert Wiener into econometrics. Granger said several times that he considered this concept of causation valid only as a first approximation, and that he didn't like its indiscriminate use.

"Granger-causes".[9] The slashed version $\not\rightsquigarrow$ means of course "does not cause". If we take conditional expectation as synonym for "forecasting function", the above definition implies that, if the past of X helps in predicting Y, then X is one of that causes of Y.[10]

It is immediate to adapt the definition above to the case of a bivariate VAR: if the vector $z_t = (y_t, x_t)'$ can be represented as a VAR process

$$\begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix},$$

we have that $x \rightsquigarrow y \iff A_{12}(L) \neq 0$.

Example 5.5.2 *Suppose we consider the following bivariate VAR:*

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} 0.8 & -0.4 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix},$$

with $\Sigma = E\left(\begin{bmatrix} u_t \\ v_t \end{bmatrix} \begin{bmatrix} u_t & v_t \end{bmatrix} \right) = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$.

In this case, we have that $X \rightsquigarrow Y$*, but not vice versa, because the past values of* Y *do not help to predict* X*. Note, moreover, that*

$$E(x_t | y_t, \mathcal{F}_{t-1}) = 0.96 x_{t-1} + 0.4 y_t - 0.32 y_{t-1}$$

and therefore the past values of Y *are of use in predicting* X*, but only when the contemporaneous value of* Y *also enters the conditioning set.*

Note that, if the VAR is stationary, a test of the hypothesis $x \not\rightsquigarrow y$ is equivalent to a simple significance test: if we write out the first equation of the VAR in full,

$$y_t = \alpha_1 y_{t-1} + \beta_1 x_{t-1} + \alpha_2 y_{t-2} + \beta_2 x_{t-2} + \cdots + \alpha_p y_{t-p} + \beta_p x_{t-p} + \epsilon_{1t}.$$

The hypothesis of no Granger-causality from x to y is equivalent to the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

which is a linear restriction on the parameters of the equation, testable via the usual methods. Since (under stationarity) asymptotic inference can be performed using the standard OLS tools, the Granger-causality test is simply a special case of the hypothesis testing apparatus.

Due to its ease of use, the Granger-causality test has been employed in countless applied papers. At some point, between the 1970s and 1990s, it was widely used as a tool to prove or disprove economic theory. Its usage, however, has declined after the profession realised that the concept of Granger-causality is too mechanical to be considered synonymous with the idea of

[9] In fact, the concept was originally framed in more general terms, using the conditional distribution of y_t rather than simply its conditional mean. However, the formulation I adopt here is in line with common applied practice.

[10] Note that from the definition of Granger-causality it follows that if $X \rightsquigarrow Y$, it does not necessarily mean that $Y \not\rightsquigarrow X$. Likewise, if $X \not\rightsquigarrow Y$, it does not necessarily mean that $Y \rightsquigarrow X$.

causality that we, as economists, have in mind. More specifically, two main criticisms arise.

The first one is a, so to speak, statistics-based criticism, and is centred on the possibility that a variable X can be found to be Granger-causal for another variable Y or not depending on which other variables are present in the system. To clarify this concept, let us give an example (which has nothing to do with Granger-causality by itself): suppose we have a trivariate normal distribution (Y, X, Z) with mean 0 and covariance matrix $\begin{bmatrix} 1 & 0 & \alpha \\ 0 & 1 & \beta \\ \alpha & \beta & 1 \end{bmatrix}$. It is immediately clear that

$$E(Y|X) = 0 \qquad E(Y|X,Z) = \frac{-\alpha\beta}{1-\beta^2}X + \frac{\alpha}{1-\beta^2}Z.$$

If α and β are non-zero, the statement X 'does not explain' Y is correct only if Z is not considered. In terms of Granger-causality, a test that rejects (or not) the null hypothesis of no Granger-causality in a bivariate VAR could reject (or not) it in a trivariate VAR; therefore, GC tests must always be considered valid *within the chosen conditioning set*. By modifying the content of the vector of variables in the VAR, the results of the Granger-causality test may change.[11]

The second criticism is more, let's say, structural; it refers to the *logical* concept of cause-and-effect, which is independent of what happens in physical time. In fact, it is possible for the cause to manifest itself only after the effect, when the latter is influenced by expectations. Edward Leamer made the point that there would have been fewer problems if Granger had used the word "precedence" rather than the word causality. The example usually given is somewhat overused, but it conveys the idea: the fact that Christmas shopping is done before Christmas does not allow us to say that the celebration of Christmas on 25 December is caused by the increase in sales of toys, socks and ties.[12]

A more interesting example for us economists is provided by Italian economic history: it is quite natural to think that the cause of the decline in interest rates on Italian sovereign debt between 1995 and 1997 was due to the fact that Italy's entry into the European Monetary Union was seen by market operators as less and less improbable. However, if we were to rigidly apply the *post hoc, ergo propter hoc*[13] criterion, we would have to conclude that Italy's entry into the EMU was *caused* by the decline in rates. To some extent this is true (the reduction in rates brought about a lightening of the debt service burden and, consequently, an improvement in the public budget), but this is certainly not the whole story: there is an important part of the explanation that is not captured by the influence of falling rates on debt servicing, which has to do with expectations, with the credibility of the government, and with the international economic situation; but you will find this story — better told — in many other places, and therefore you will not find it here.

[11] If one wishes, the entire argument can be read as a special case of the so-called omitted variables problem in the linear model.

[12] Another nice one is: do weather forecasts cause rain?

[13] After this, therefore because of this.

In conclusion, I like to quote one of my favourite econometricians, Adrian Pagan, who in one of his splendid survey articles (Pagan, 1989) wrote:

> *There was a lot of high-powered analysis of this topic, but I came away from a reading of it with the feeling that it was one of the most unfortunate turnings for econometrics in the last two decades, and it has probably generated more nonsense results than anything else during that time.*

5.6 Dynamic Analysis and Structural VARs

As in the univariate case, the primary tool for the dynamic analysis of a VAR process is the impulse response function, which was previously defined and analysed in section 3.4.2 for univariate processes. The reader will recall that, in that section, we motivated the use of the impulse response function by interpreting the white noise appearing in the ARMA representation of the process as the additional information that enters the memory of the time series at each period. Here, however, matters are more complex because we have n variables and n shocks, meaning that the impulse response is an $n \times n$ matrix for any given period. The good thing is that, in return for this additional complication, we get the possibility of interpreting results in a very interesting way.

These ideas give rise to the so-called **structural VARs**, or **SVARs** for short, that provide one of the most important tools in modern macroeconometrics; the literature is enormous, and it has achieved remarkable levels of sophistication. Here, I shall confine myself to a highly introductory discussion. Consider a stationary VAR[14]

$$A(L)y_t = \epsilon_t \tag{5.5}$$

and start from its moving average representation:

$$y_t = \sum_{i=0}^{\infty} \Theta_i \epsilon_{t-i}.$$

is it possible to attach some meaning to the Θ_n matrices? One can attempt a parallel with the univariate case: $(\Theta_n)_{i,j}$ is the response of the i-th variable to the j-th shock after n periods. Since the vector ϵ_t is the difference between y_t and $E(y_t|\mathcal{F}_{t-1})$, it can be thought of the one-step-ahead forecast error (see the discussion in Subsection 3.4.2).

Differently from the univariate case, however, there is more that can be said here. With a single time series, the source of uncertainty that make our forecasts wrong remains obscure: quite simply, the causes must be sought outside the conditioning information set. This is why we confine ourselves to interpreting the shock as the effect of unpredictable causes.

In the multivariate setting, we can go a little further: for each period, we have n forecast errors, one for each time series. Consequently, we can ask what the underlying reasons for the forecast errors are. More formally, we could

[14] Deterministic terms would just be a nuisance here, so let's say the process has zero mean.

imagine that our vector of forecast errors is a linear function of unobservable shocks, to which we can attach some "structural" interpretation, so they are known as we call **structural shocks**:

$$\epsilon_t = Cu_t, \tag{5.6}$$

where we assume that C is square and invertible. Note that the vector ϵ_t is observable (or at least estimable), whereas the vector u_t and the matrix C are not.

Example 5.6.1 *Suppose we observe the price of a good p_t and the quantity traded q_t over time, and that the following representation holds:*

$$A(L)\begin{bmatrix} p_t \\ q_t \end{bmatrix} = \begin{bmatrix} \mu_{p,t} \\ \mu_{q,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{q,t} \end{bmatrix}.$$

where

$$\begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{q,t} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{2,1} & c_{2,2} \end{bmatrix} \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix}.$$

More precisely: let us imagine that demand and supply are unobservable quantities, subject to mutually uncorrelated random shocks, that we call u_t^d and u_t^s, respectively. With a fixed supply curve, an unpredictable shift in the demand curve (a "demand shock") would produce a forecast error of the same sign *for both price and quantity. Clearly, by reversing the argument, a "supply shock"would imply forecast errors* of opposite sign. *Therefore, we would expect c_{11} and c_{21} to have the same sign, while c_{12} and c_{22} should be opposite in sign. In practice, both structural shocks to the demand and supply curves propagate to the forecast errors of both equations.*

If C were known, very interesting possibilities would open up. Clearly, we could reconstruct the history of structural shocks (via $u_t = C^{-1}\epsilon_t$), but most importantly, we could calculate structural impulse responses: combining equations (5.5) and (5.6) yields

$$A(L)y_t = Cu_t$$

and therefore

$$y_t = [A(L)]^{-1} Cu_t = Cu_t + \Theta_1 \cdot Cu_{t-1} + \Theta_2 \cdot Cu_{t-2} + \cdots \tag{5.7}$$

from which

$$IRF(i,j,n) = \frac{\partial y_{it}}{\partial u_{jt-n}} = (\Theta_n \cdot C)_{ij} = [M_n]_{i,j}$$

While the impulse response with respect to the forecast error is hardly interpretable, the impulse response to the structural shock would would be useful in evaluating how the observable quantities (in the previous example, price and quantity) respond over time to a shock impacting a behavioural relation (in the previous example, the demand function or the supply function).[15]

[15] There are more by-products of this representation, namely the Forecast Error Variance Decomposition, commonly abbreviated as FEVD, and the Historical Decomposition. For these, read the literature.

Since the matrix C is unobservable, it must be estimated. However, this is by no means a straightforward task, as the only observable statistic that can serve as a basis for estimation is the covariance matrix of ϵ_t, namely Σ.[16] If, without loss of generality, we normalise the structural shocks to have unit variance, it follows from equation (5.6) that

$$\Sigma = CC'; \tag{5.8}$$

given that Σ can be estimated very easily via the sample covariance matrix of the VAR residuals, a matrix $\hat{C}$ for which $\hat{\Sigma} = \hat{C}\hat{C}'$ holds is a perfectly good estimate of C.

The problem with this idea is that, for every symmetric and positive-definite matrix Σ, there exist infinitely many matrices C that satisfy equation (5.8): suppose that C is a solution to equation (5.8). Then define

$$C^* = CQ,$$

where Q is an arbitrary orthogonal matrix (if your reaction is "what?", then Section 5.A.3 is for you). Since $QQ' = I$ by definition, then C^* is also a solution, as can be verified by direct multiplication.

In the econometric jargon, tihs is called an under-identification problem: there is an uncountably infinite number of matrices $\hat{C}$ that are *observationally equivalent*, whcih means that any of them yields an equally good trepresentation of the data. If we wish to obtain *a single* estimate of C, we must impose constraints such that equation (5.8) has one and only one solution. This is known as the **identification problem** in the SVAR literature. In practice, since the matrix C is also the structural IRF of order 0 (see equation (5.7)), in most cases we use some *a priori* knowledge to impose the necessary constraints on C.

5.6.1 Triangular SVARs and the Cholesky Decomposition

The first solution historically proposed for this problem, which is still widely used today,[17] is based on **triangularising** the VAR. It can be shown that equation (5.8) has a unique solution if assume the matrix C to be lower triangular, meaning that all its elements c_{ij} are zero for $j > i$. In this case, the tool we use to write Σ as CC' is known as the **Cholesky decomposition**: it establishes that any symmetric and positive-definite matrix V can always be written as the product of a lower triangular matrix L and its transpose, *i.e.* $V = LL'$, and that L is unique.[18]

In our case, given a consistent estimate of Σ, it is possible to obtain $\hat{C}$ from $\hat{\Sigma}$ because C is a continuous function of Σ, and therefore $\hat{C}$ is consistent too.

[16] Strictly speaking, this is only true under normality of ϵ_t. A recent strand of literature has explored possible departures from the normality assumption and found some clever tricks to identify C using those, but these techniques are too advanced to be discussed here. I'll just make a passing reference later, in Section 5.6.3.

[17] This is due to both historical reasons and its computational simplicity, which means that many econometric software packages implement it as the default option.

[18] The Cholesky decomposition can be calculated by hand without much effort only when the matrix V is small. However, this is a task best left to computers.

Once these estimates have been obtained, we can proceed to the analysis of the impulse responses, which can be computed from the matrices $\hat{\Theta}_n\hat{C}$, like in equation (5.7).

The meaning of the triangular decomposition can be illustrated by writing (5.6) as follows: since C is lower triangular,

$$\begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{nt} \end{bmatrix} = \begin{bmatrix} c_{11} & 0 & \cdots & 0 \\ c_{21} & c_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{nt} \end{bmatrix} = \begin{bmatrix} c_{11}u_{1t} \\ c_{21}u_{1t} + c_{22}u_{2t} \\ \vdots \\ c_{n1}u_{1t} + c_{n2}u_{2t} + \cdots + c_{nn}u_{nt} \end{bmatrix}.$$

As can be clearly seen, C being lower-triangular implies that the forecast error for the first variable is just the first structural shock (rescaled), the forecast error for the second variable is a combination of the first two, and in general, $\epsilon_{it} = \sum_{j=1}^{i} c_{ij}u_{jt}$ (note that the upper limit of the summation is i). In other words, the ordering of the elements within the vector y_t is not arbitrary but follows a precise logic: the one-step-ahead forecast error on the first variable of the VAR is a function of the first structural shock only, so that *by construction* the first structural shock coincides (up to a scale factor) with the forecast error of the first variable. The second structural shock, on the other hand, is identified by subtraction: given that only the first two structural shocks impact the second variable, the second structural shock u_{2t} is defined as the second forecast error ϵ_{2t} "net of" u_{1t} (or ϵ_{1t}, which is the same).

Hopefully, this argument should make it obvious that the choice of variable ordering is absolutely crucial in interpreting the results of a triangular VAR. A word of caution, however. The above assumption is less restrictive than it appears at first sight. Indeed, if C is lower triangular, the constraint only affects the impulse responses at time 0; after one period, each structural shock can have a non-zero impact on each of the observables.

Identification, therefore, is achieved by specifying a hierarchy among the variables of the VAR based on their (instant) responsiveness to structural shocks, and the variables are ordered by increasing responsiveness: in the jargon of applied economists, the precept is "the most exogenous variables go first".

Example 5.6.2 *Let us continue with Example 5.6.1. One could imagine, for instance, that a shock to the demand for a good does not instantly impact its price, perhaps because producers need some time to perceive the shock, adjust the prices, and allow the new price to propagate along the distribution chain. In this case, a demand shock influences — instantaneously — only the quantity traded.*

The forecast error on the price, therefore, will not contain the demand shock but only the supply shock, whereas the forecast error on the quantity will be a function of both. In formulas:

$$A(L)\begin{bmatrix} p_t \\ q_t \end{bmatrix} = \begin{bmatrix} \mu_{p,t} \\ \mu_{q,t} \end{bmatrix} + \begin{bmatrix} c_{11} & 0 \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

Note that this assumption does not imply that the demand shock never impacts the

price. It only implies the instantaneous *neutrality of the price to demand, meaning that a demand shock affects price only after one period.*

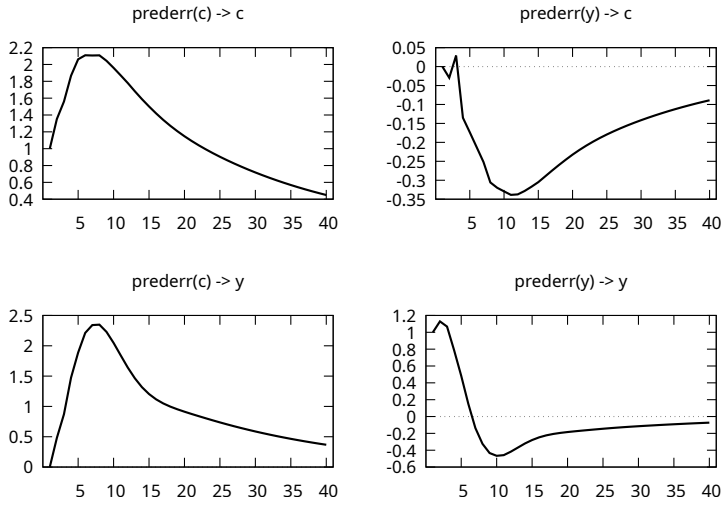
Naturally, there is a certain degree of arbitrariness in the ordering one chooses. Often, institutional factors or information asymmetries are referenced to justify a particular ordering, and it is clear that, in general, the idea that a variable takes at least one period to move in response to a shock is all the more defensible the shorter the period: in the price/quantity example I used above, the identification restriction (price does not respond instantly to a demand shock) could be perfectly justified with weekly data, somewhat less so with monthly data, and would become far less defensible with quarterly or annual data.

Example 5.6.3 *Let us return to Example 5.5.1, where we have consumptiona nd income in the EU. As you may recall, we used a VAR(2) for simplicity, but here we're going to make the example juicier (and more realistic) by adopting the VAR order suggested by the HQ criterion, ie 6. The estimated $\hat{A}_i$ matrices are*

$$A_1 = \begin{bmatrix} 1.353 & -0.029 \\ 0.481 & 1.129 \end{bmatrix} \quad A_2 = \begin{bmatrix} -0.255 & 0.102 \\ -0.321 & -0.193 \end{bmatrix} \quad A_3 = \begin{bmatrix} 0.079 & -0.266 \\ 0.268 & -0.225 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} -0.090 & 0.233 \\ -0.260 & 0.138 \end{bmatrix} \quad A_5 = \begin{bmatrix} -0.180 & -0.063 \\ 0.039 & -0.135 \end{bmatrix} \quad A_6 = \begin{bmatrix} 0.107 & -0.014 \\ -0.015 & 0.059 \end{bmatrix}$$

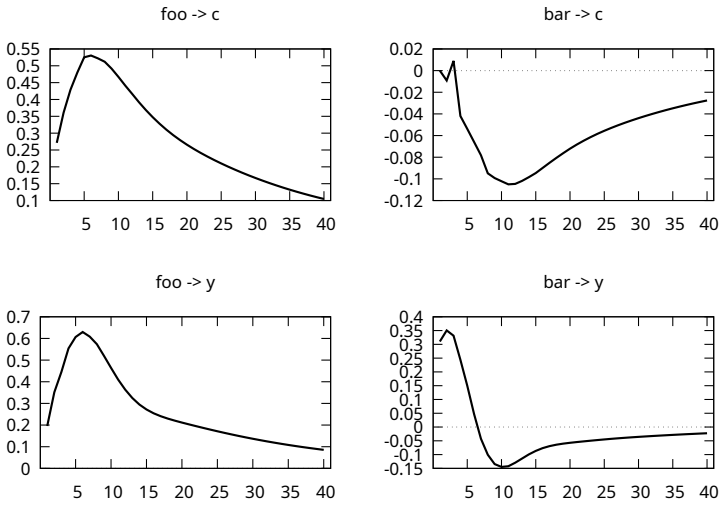
Figure 5.3: Non-structural impulse response functions



By computing powers of the companion matrix, we obtain the Wold representation, that is the non-structural impulse response functions; its coefficient matrices are

$$\Theta_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Theta_1 = \begin{bmatrix} 1.353 & -0.029 \\ 0.481 & 1.129 \end{bmatrix} \quad \Theta_2 = \begin{bmatrix} 1.560 & 0.029 \\ 0.872 & 1.068 \end{bmatrix} \quad \cdots$$

Figure 5.4: Structural impulse response functions



and so forth. The impulse responses are plotted in Figure 5.3. The top-left graph shows the response of c_t with respect to ϵ_{1t} (called "`prederr(c)`" in the plot), and so on.

What could be the economic interpretation of Figure 5.3? Very little, I would say. Those plotted are just the responses of the observed variables to the one-step-ahead forecast errors. Consider, instead, the triangularised version. The covariance matrix of the residuals $\hat{\Sigma}$ and its Cholesky factor (that we take as an estimate of C) are

$$\hat{\Sigma} = \begin{bmatrix} 0.074 & 0.053 \\ 0.053 & 0.135 \end{bmatrix} \qquad \hat{C} = \begin{bmatrix} 0.271 & 0.000 \\ 0.196 & 0.310 \end{bmatrix}$$

Therefore, the structural VMA coefficient matrices are $M_i = \Theta_i \hat{C}$:

$$M_0 = \begin{bmatrix} 0.271 & 0.000 \\ 0.196 & 0.310 \end{bmatrix} \quad M_1 = \begin{bmatrix} 0.361 & -0.009 \\ 0.351 & 0.350 \end{bmatrix} \quad M_2 = \begin{bmatrix} 0.429 & 0.009 \\ 0.445 & 0.331 \end{bmatrix} \quad \cdots$$

I don't want to assign an economic interpretation to the structural shocks just yet. For the moment, I'm going to call them "foo" and "bar". The structural IRFs are plotted in Figure 5.4.

The first structural shock, "foo" is identified as the only shock impacting c_t contemporaneously, and therefore coincides with the one-step-ahead prediction error for c_t. The reader can verify that the top two plots in Figure 5.4 look exactly the same as those in Figure 5.3, apart from a slight scale adaptation. However, under the triangular identification scheme, the prediction error for y_t is a mix of "foo" and "bar", so that the impulse responses for y_t look different (especially the one for "foo", that doesn't start from 0).

How do we interpret Figure 5.4? We need to find some theoretical unobservable time series that we can associate with "bar". That is, some latent variable, to which we can attach a meaning, whose movement does affect y_t instantaneously, but not c_t.

Can you think of one? I don't have a fireproof answer. Possibly, using a rather old-fashioned Keynesian interpretation, one could suggest interpreting the second structural shock "bar" as a shock to the components of autonomous demand — for instance, an unexpected improvement in the trade balance. The forecast errors for c_t are to be ascribed exclusively to the first structural shock; therefore, we may call "foo" the "consumption shock" and "bar" the "autonomous demand shock".

In the world we are imagining, households do not immediately notice the increase in income and only begin spending their extra income after one period. Given a shock to autonomous demand, households do not notice it immediately, and indeed the curve starts from zero, because the Cholesky decomposition sets C to be lower triangular. After one quarter, however, a sort of Keynesian multiplier gets into action, leading households to increase consumption for 2-3 periods, after which the effect begins to decline. Do you find this interpretation convincing? Well, I'll leave it to your judgement.

Two final remarks: the triangular identification scheme can be generalised to more general constraints onto the instantaneous IRF, that is, the M_0 matrix. In the so-called "AB-model", equation (5.6) is replaced by its more flexible counterpart

$$A\epsilon_t = Bu_t, \tag{5.9}$$

where A and B are generic square matrices. A triangular SVAR becomes a special case where $A = I$ and B is lower-triangular. Models of this kind are quite popular in some sub-literatures (eg. fiscal models), but the conditions you need for identification are more complex and demanding, and their estimation is much less straightforward than just taking the Cholesky factor of a matrix.

Finally: it is customary, when plotting the IRF like in Figure 5.4, to surround them by confidence bands so as to make their evaluation a little more appropriate from a statistical viewpoint. This is, in most cases, done by means of a computational technique called *bootstrapping*, whose description is far outside the scope of the present book. Go read the literature.

5.6.2 Long-run constraints

An alternative method of achieving identification is to impose constraints not on what happens instantaneously, but rather on what happens in the long run: this idea was introduced by Blanchard and Quah in a celebrated paper published in 1989, and since then, it has been exploited many times and developed in numerous directions.

Rather than being motivated by institutional considerations, these long-run restrictions stem from an approach more inclined to take seriously economic theory, which is often unable to make predictions about what happens in the short run, but only in the long run.

To illustrate how it all works, I shall return to the original idea of Blanchard and Quah: to describe the state of the real economy, we use two variables (assumed to be stationary), namely the GDP growth rate (ΔY_t) and the

unemployment rate (U_t). The movements of these two variables over time originate from aggregate demand shocks, such as fiscal or monetary policy, and from aggregate supply shocks, such as technological innovations.

The identification strategy, coming from economic theory, is based on the idea that, in the long run, the *level* of GDP is entirely determined by technology, namely by aggregate supply. Fluctuations in aggregate demand, such as those induced by economic policy, impact GDP in only a transitory manner.

To make this idea operational, let us start from the moving average representation:

$$y_t = \Theta(L)\epsilon_t = \epsilon_t + \Theta_1\epsilon_{t-1} + \cdots$$

which in this case is

$$\begin{bmatrix} \Delta Y_t \\ U_t \end{bmatrix} = B \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix} + \Theta_1 B \begin{bmatrix} u_{t-1}^s \\ u_{t-1}^d \end{bmatrix} + \dots$$

If the VAR is stationary, obviously $\Theta_n \to 0$. However, we are not interested in the impact of the shocks on ΔY_t, but rather on the level Y_t. Evidently,

$$Y_{t+k} = Y_t + \Delta Y_{t+1} + \Delta Y_{t+2} + \cdots + \Delta Y_{t+k}$$

and therefore the long-run impact matrix is given by

$$\lim_{k\to\infty} \frac{\partial Y_{t+k}}{\partial u_t^d} = \frac{\partial \sum_{i=0}^{\infty} \Delta Y_{t+i}}{\partial u_t^d} = \sum_{i=0}^{\infty} \frac{\partial \Delta Y_{t+i}}{\partial u_t^d} = \sum_{i=0}^{\infty} [\Theta_i B]_{1,2} = [\Theta(1)B]_{1,2}.$$

Consequently, to analyse the impact of the shocks on the level of GDP, one must consider the *cumulative* impact of the shocks on ΔY_t. The theoretical assumption of long-run neutrality of GDP to demand translates into this quantity being 0:

$$\lim_{k\to\infty} \frac{\partial Y_{t+k}}{\partial u_t^d} = 0. \tag{5.10}$$

The constraint we need, therefore, is obtained by imposing that the matrix $\Theta(1)B$ be lower triangular. The matrix $\Theta(1)$ is easy to estimate once the parameters of the VAR have been estimated: since

$$\Theta(L) = A(L)^{-1},$$

an estimate of $\Theta(1)$ is simply

$$\widehat{\Theta(1)} = \hat{A}(1)^{-1}.$$

Obviously, for this to work, it is necessary that $A(1)$ be invertible, which rules out unit roots processes.[19]

To illustrate the meaning of the constraint, it is useful to mention that the matrix Σ is, by definition, the covariance matrix of ϵ_t, but it can also be viewed as $V(y_t|\mathcal{F}_{t-1})$, *i.e.* the covariance matrix of the conditional distribution

[19] This is not strictly true: there exists a highly interesting generalisation to the cointegrated case which, however, I shall not address here. Anyone interested should google for "King Plosser Stock Watson". I shall say no more.

of y_t. If we focused on the marginal distribution of y_t, namely $V(y_t)$, instead, what would the covariance matrix look like? Let us consider, once again, the moving average representation:

$$y_t = \Theta(L)\epsilon_t = \epsilon_t + \Theta_1\epsilon_{t-1} + \cdots$$

Since ϵ_t is a vector white noise process, its elements are uncorrelated with each other over time; it follows that

$$V(y_t) = V(\epsilon_t) + \Theta_1 V(\epsilon_{t-1})\Theta_1' + \cdots = \Sigma + \Theta_1\Sigma\Theta_1' + \cdots = \Theta(1)\,\Sigma\,\Theta(1)'.$$

Given that $\Sigma = CC'$, we can write

$$\Theta(1)\,\Sigma\,\Theta(1)' = [\Theta(1)C][\Theta(1)C]'.$$

The constraint corresponding to the long-run restriction is that the matrix $[\Theta(1)B]$ be lower triangular. In practice, a long-run restriction can be thought of as a constraint on the *marginal* covariance matrix instead of a constraints on the *conditional* covariance matrix.

At this point, at least in principle, the elements of C can be recovered. In the Blanchard-Quah case, it follows that

$$\Theta(1)_{11}C_{12} + \Theta(1)_{12}C_{22} = 0 \Longrightarrow C_{12} = -\frac{\Theta(1)_{12}}{\Theta(1)_{11}}C_{22},$$

so that, if we define $\theta = -\frac{\Theta(1)_{12}}{\Theta(1)_{11}}$, it follows that

$$CC' = \begin{bmatrix} C_{11}^2 + \theta^2 C_{22}^2 & C_{11}C_{21} - \theta C_{22} \\ C_{11}C_{21} - \theta C_{22} & C_{21}^2 + C_{22}^2 \end{bmatrix} = \Sigma = \begin{bmatrix} \sigma_{\Delta_Y}^2 & \sigma_{\Delta_Y,U} \\ \sigma_{\Delta_Y,U} & \sigma_U^2 \end{bmatrix}$$

and to find the matrix C, the following system of equations

$$\begin{aligned} \sigma_{\Delta_Y}^2 &= C_{11}^2 + \theta^2 C_{22}^2 \\ \sigma_{\Delta_Y,U} &= C_{11}C_{21} - \theta C_{22} \\ \sigma_U^2 &= C_{21}^2 + C_{22}^2 \end{aligned}$$

must be solved.

5.6.3 Other approaches to identification

In recent years, the difficulties inherent in finding justifiable identification restrictions have prompted scholars to pursue other approaches. This text is too introductory to give a reasonable account of the other ideas that have been put forward, so I'll just show you a list of three main alternative approaches, some of which have become extremely popular:

Proxy SVARs The main intuition is that, in some cases, it could be possible to find observable variables that are, arguably, well correlated with structural shocks. For example, people have done very clever things by applying text analysis to the statements issued by central banks to recover

an observable variable that should be act as a reasonably good "proxy" to a genuine monetary policy shock. It is interesting to note that this strand of literature has a fair number of contact points with the one on instrumental variables.

Non-white noise innovations If ϵ_t is normal white noise, the the only source of information we have in the matrix C in equation (5.6) is the sample covariance matrix of the VAR residuals. However, departures form that assumption can be quite beneficial when it comes to identifying C. For example, one could use higher-order moments.

Another idea, that has become quite popular, is **identification by heteroskedasticity**. The simplest example is: suppose that in some periods $V(\epsilon_t) = \Sigma_1$ and in some other periods $V(\epsilon_t) = \Sigma_2 \neq \Sigma_1$, and that you know the time boundaries between the two regimes (for instance: high-uncertainty vs low-uncertainty periods). In this case, it can be proven that you can estimate C directly from the data, and no identification restrictions are necessary.

Sign restrictions In some cases (see eg example 5.6.1) we don't really know very much, apart from the *sign* that a certain IRF should display, according to economic theory. An adverse technology shock should pull GDP down, almost by definition; a restriction in monetary policy has the precise aim of curbing inflation; and so on. In these cases, one could explore the set of matrices C that are consistent with these views. One of the simplest ways to do so is to start with K, the Cholesky factor of Σ, and then use an appropriate computer algorithm to generate random orthogonal matrices $Q_1, Q_2, \ldots, Q_H$. Of course, all the matrices

$$C_i = KQ_i$$

have the property $\Sigma = C_i C_i'$, but only a subset of them will be consistent with the sign restriction we have in mind. Typically, you only keep the "good" ones and see what happens, via specialised techniques I won't describe here.

From the considerations presented in these final pages, it is evident that Structural VARs mat produce results that depend crucially on the identification conditions that have been imposed. It is by no means impossible (indeed, it is far from rare) for two researchers to arrive at completely different interpretations of the dynamic properties of the very same system, simply because they have chosen different identification hypotheses (for example, a different variable ordering in a triangular system). Personally, I do not think this is a weak point of the SVAR methodology. In the analysis of a complex problem, I believe there is no single "right" or "wrong" way to proceed.[20] Consequently, different identification strategies may be reasonable and, obviously, results may change across the various cases because the definitions of the shocks whose responses are being measured also change. Whether an identification assumption more useful, reasonable, informative, or illuminating than another depends largely on the problem under consideration and on personal preference.

---

[20] "...there is always a well-known solution to every human problem — neat, plausible, and wrong". H. L. Mencken.

5.A Assorted results

5.A.1 Simultaneous equation systems

That estimation of systems of equations is the first autonomous contribution of econometrics to the general arsenal of statistical tools.

A simultaneous equation system can be represented in two ways: in the **structural form**, the system can be represented as

$$\Gamma y_t = B x_t + u_t, \tag{5.11}$$

where the vector y_t contains n endogenous variables, the vector x_t contains k exogenous variables, and u_t is a vector of innovations. The matrices Γ and B contain behavioural parameters, to which we like to give an economic interpretation, and these are the ones we would wish to estimate.

For example, consider a two-equation demand-supply system, written as

$$q_t = \alpha_0 - \alpha_1 p_t + \alpha_2 y_t + u_t \tag{5.12}$$

$$p_t = \beta_0 + \beta_1 q_t + \beta_2 m_t + v_t; \tag{5.13}$$

equation (5.12) is the demand equation (quantity at time t as a function of price at time t), (5.13) is supply (price as a function of quantity). Demand depends on price and y_t (the per-capita income at time t), while in the supply equation price depends on production and m_t (the price of raw materials at time t).

The two disturbance terms u_t and v_t are random shocks to the two curves. For example, u_t could incorporate random fluctuations in demand due to shifting customer preferences, fluctuations in disposable income and so forth; v_t, instead could be non-zero because of productivity shifts due to random events (think for example weather for agricultural produce). Assume that $E(u_t) = E(v_t) = 0$.

The system above can be written using matrix notation as

$$\begin{bmatrix} 1 & \alpha_1 \\ -\beta_1 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \alpha_0 & \alpha_2 & 0 \\ \beta_0 & 0 & \beta_2 \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ m_t \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix},$$

where for example $\Gamma = \begin{bmatrix} 1 & \alpha_1 \\ -\beta_1 & 1 \end{bmatrix}$ and so on.

The parameters of interest are the elements of Γ and B. These, however, cannot be consistently estimated using OLS, and this is why many ingenious alternative estimation methods were devised, including Instrumental Variable (IV) estimation.

By premultiplying eq. (5.11) by Γ^{-1} you can also defined the **reduced form** of the system:

$$y_t = \Pi x_t + w_t. \tag{5.14}$$

In the supply-demand example above the reduced form would be

$$q_t = \pi_{1,1} + \pi_{1,2}\, y_t + \pi_{1,3}\, m_t + w_{1,t} \tag{5.15}$$

$$p_t = \pi_{2,1} + \pi_{2,2}\, y_t + \pi_{2,3}\, m_t + w_{2,t}. \tag{5.16}$$

Unlike the structural form, the reduced form can be estimated using OLS for each equation in the system; unfortunately, the coefficients contained in the matrix Π lack an economic interpretation. Nevertheless, once a consistent estimate of Π is obtained (let us call it $\widehat{\Pi}$), we could implicitly define consistent estimators of Γ and B (let us call them $\widehat{\Gamma}$ and $\widehat{B}$) as those statistics that satisfy the relationship $\widehat{\Gamma}\widehat{\Pi} = \widehat{B}$.

However, the number of elements in the matrix Π is smaller than the number of parameters contained in Γ and B; hence, it is impossible to define these statistics uniquely, unless constraints are imposed on the matrices Γ and B. The so-called "identification conditions" are nothing more than the imposition of a certain number of restrictions on the elements of Γ and B. But perhaps you had better pick up a good econometrics textbook.

5.A.2 Forecast error variance and confidence intervals

Let's begin by the simplest case: we have data from 1 to T and we want to forecast y_{T+1} using a VAR(1). Based on Section 5.5.1, our forecast is

$$\hat{y}_{T+1} = \hat{A} y_T;$$

consequently, the forecast error we're going to make is

$$e_{T+1} = y_{T+1} - \hat{y}_{T+1} = A y_T + \epsilon_{T+1} - \hat{A} y_T = \left(A - \hat{A}\right) y_T + \epsilon_{T+1}.$$

Our objective is to compute the covariance matrix of the expression above. Since that the two elements it contains should be independent (or nearly so),[21] the variance of the sum is the sum of the variances, so

$$V(e_{T+1}) = V\left[\left(A - \hat{A}\right) y_T\right] + \Sigma.$$

The first component comes from the fact that we don't observe the true A matrix, but we need to estimate it. The second one is just the covariance matrix of ϵ_t, the white noise driving the process.

In principle, both items should be taken into account. However, it should be noted that in ordinary circumstances consistency of $\hat{A}$ implies that

$$\lim_{T\to\infty} V\left[\left(A - \hat{A}\right) y_T\right] = [0],$$

and what commonly happens is that, with decently-sized samples, the uncertainty stemming from estimation is negligible compared to the one coming from ϵ_t. In practice, nobody bothers and everybody uses the approximation

$$V(e_{T+1}) \simeq V(\epsilon_{T+1}),$$

which is equivalent to assuming that we're using the DGP parameters and not an estimate.

OK, so let's pretend we have the true parameters. There are two more issues we have to tackle:

[21] Why? As lazy writers like me are fond of saying, "this is left as an exercise to the reader".

1. How does this generalise to VAR(p) models?
2. Can we compute the variance for k-step-ahead predictions?

The first issue is, in fact, a non-issue: all we have to do is rewrite the model in companion form and we're done (of course, we have to extract the appropriate sub-matrices from the ones we compute, like we did in Section 5.2, but that's trivial).

As for multi-step forecasts, here's how we do it: let's continue with the example above and calculate our forecast error for $T+2$. The first thing to note is that[22]

$$y_{T+2} = Ay_{T+1} + \epsilon_{T+2} = A^2 y_T + A\epsilon_{T+1} + \epsilon_{T+2}$$

and therefore our forecast can be written as

$$\hat{y}_{T+2} = E(y_{T+2}|\mathcal{F}_T) = A^2 y_T$$

so the forecast error becomes

$$e_{T+2} = A\epsilon_{T+1} + \epsilon_{T+2};$$

again, the white noise property ensures that the covariance between ϵ_{T+1} and ϵ_{T+2} is zero (a matrix of zeros, to be pedantic). Therefore,

$$V(e_{T+2}) = A\Sigma A' + \Sigma. \tag{5.17}$$

The extension to the general case is immediate:

$$e_{T+k} = \epsilon_{T+k-1} + A\epsilon_{T+k-2} + A^2\epsilon_{T+k-3} + \ldots + A^{k-1}\epsilon_{T+1},$$

and the expression corresponding to 5.17 can be easily found.

In fact, the whole argument could be rephrased as follows: once we can express y_t as a vector moving average, the forecast error variance can be computed easily from the choice of the conditional expectation as our forecast function:

$$\hat{y}_{T+k} = E(y_{T+k}|\mathcal{F}_T),$$

which implies $e_{T+k} = y_{T+k} - E(y_{T+k}|\mathcal{F}_T)$. Consider now the VMA representation of y_{T+k} and split it as follows:

$$\begin{aligned} y_{T+k} &= \underbrace{\epsilon_{T+k-1} + C_1\epsilon_{T+k-2} + \cdots + C_{k-1}\epsilon_{T+1}}_{e_{T+k}} + \\ &+ \underbrace{C_k\epsilon_T + C_{k+1}\epsilon_{T-1} + \cdots + C_{k+2}\epsilon_{T-2}}_{E(y_{T+k}|\mathcal{F}_T)}; \end{aligned}$$

(bear in mind that all elements of ϵ_t up to time T are contained in $\mathcal{F}_T$). Using the representation above, computing $V(e_{T+k})$ is boring but trivial (software is very good at this):

$$V(e_{T+k}) = \sum_{i=0}^{k-1} C_i \Sigma C_i',$$

[22] Let me remind that we're pretending we observe the "true" A.

where of course $C_0 = I$.

Confidence intervals are routinely computed using the expression above under the normality assumption. So, for example, if you want to compute a 95% confidence interval for k-period ahead forecast for variable 1, you take the $(1,1)$ element of $V(e_{T+k})$, compute its square root, multiply the result by 1.96 and use that number to build a band around the forecast. Of course, this is justified if the white noise ϵ_t has a Gaussian distribution. Unfortunately, there are no good reasons why this should be true in most cases, but alternative approaches are quite complicated, or lead to inconveniently large confidence intervals, so raising this point is considered as somehow discourteous.

5.A.3 Orthogonal matrices

Linear algebra is full of amazing objects, and orthogonal matrices are one of them. A square $n \times n$ matrix Q is said to be *orthogonal* if

$$QQ' = Q'Q = I$$

or, equivalently, if $Q' = Q^{-1}$, its transpose is also its inverse. They are commonly referred to as "rotation matrices", because premultiplying a vector by Q leaves its length unchanged: if $y = Qx$, then $y'y = x'Q'Qx = x'x$ (more on this later).

If $n = 1$, Q is a real scalar, and can only be 1 or -1.[23] If $n = 2$, instead, there are infinitely many orthogonal matrices, which belong to one of two categories:

$$Q_1 = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \quad Q_2 = \begin{bmatrix} c & s \\ s & -c \end{bmatrix}$$

where $c = \cos(\omega)$, $s = \sin(\omega)$ and ω is an arbitrary real. Therefore, there are infinitely many 2×2 orthogonal matrices. The "rotation" idea can be verified by checking that, for example, if $\omega = \pi/4$, the position on the plane of the vector

$$y = Q_1 x$$

is equal to the position of x, rotated counterclockwise by 45°.

This concept can be generalised to larger matrices, so for example a 3×3 orthogonal matrix can be thought of as performing a rotation on a sphere, and so on.

[23] Yes, I could consider complex matrices too, but not here, ok?

Chapter 6

Cointegration

6.1 Definitions

In this chapter, we will frequently consider *linear combinations* of univariate processes. More specifically, we will discuss the integration properties of processes that can be written as

$$z_t = \beta' y_t,$$

where y_t is a multivariate stochastic process and β is a matrix of constants.

It should be intuitively clear that the linear combination of two or more stationary processes is also stationary. This can be easily proven, but for our purposes, intuition will suffice. Furthermore, a linear combination, such as a sum, of an $I(1)$ process and an $I(0)$ process is also an $I(1)$ process (this follows quite easily from the BN representation of the $I(1)$ process). Following this line of reasoning, one might be tempted to conclude that a linear combination of two $I(1)$ processes should be $I(1)$ too. In fact, this may not always be the case. Let us consider the following example:

$$\begin{cases} x_{1t} = x_{1t-1} + \epsilon_t \\ x_{2t} = x_{1t} + u_t, \end{cases}$$

where ϵ_t and u_t are two generic $I(0)$ processes.

Evidently, x_{1t} is $I(1)$; it is also evident that x_{2t} is also $I(1)$, as it is the sum of an $I(1)$ process and an $I(0)$ process. Let us, however, consider their difference between x_{1t} and x_{2t}, namely: $z_t = x_{2t} - x_{1t}$. From the definitions given above, it is obvious that $z_t = u_t$, which is $I(0)$ by assumption. Therefore, you have a linear combination of $I(1)$ processes that is stationary.

We can now state a more general principle regarding linear combinations of processes: if x_t is $I(d)$ and y_t is $I(b)$, then $z_t = x_t + \lambda y_t$ is $I(c)$, where

$$\begin{cases} c = \max(d,b) & \text{if } d \neq b \\ c \leq \max(d,b) & \text{if } d = b. \end{cases}$$

When $c < \max(d,b)$ (that is, when the inequality holds strictly), **cointegration** occurs. As most literature does, I will focus on the case where a linear combination of $I(1)$ processes yields an $I(0)$ process. Let us therefore assume we

have an n-dimensional multivariate $I(1)$ process, which we denote as y_t; this is equivalent to saying that each of the n processes contained in the vector y_t is, individually, $I(1)$.

Cointegration is when there exists at least one vector β such that the combination $z_t = \beta' y_t$ is $I(0)$; if so, β is termed a **cointegrating vector**. As we shall see in the next section, there may be more than one such vector: if these vectors are collected into a matrix, the latter is called the **cointegrating matrix**; the number of linearly independent cointegrating vectors (that is, the rank of the cointegrating matrix) is termed the **cointegration rank**. In the previous example, $y_t' = [x_{1,t}, x_{2,t}]$ and $\beta' = [-1, 1]$; it can be proven that there are no other linearly independent cointegration vectors, so the cointegration rank is 1, and therefore β also coincides with the cointegrating matrix. A multivariate $I(1)$ process for which there is at least one cointegrating vector is also known as a **cointegrated system**.

6.2 Properties of cointegration vectors

Even at this preliminary stage, a couple of interesting insights cointegrating vectors are possible: first, for an $n \times 1$ dimensional $I(1)$ vector y_t, the cointegration rank can be at most equal to $n - 1$: if n linearly independent cointegrating vectors existed, they would form a non-singular cointegrating matrix B: in this case, we would have $By_t = z_t$. However, if B is invertible, then $y_t = B^{-1} z_t$ must also hold, which is evidently absurd, given that on the left-hand side of the equals sign we have a quantity that is $I(1)$ (by definition), and on the right-hand side a quantity that is $I(0)$, as it is a linear combination of $I(0)$ processes.

The second remark is that, given one or more cointegrating vectors, any linear combination of them is likewise a cointegrating vector: if β is a cointegrating matrix with r rows and n columns, then

$$\beta' y_t = z_t,$$

where $z_t \sim I(0)$; now take a linear combination of the rows of β', defined as $\gamma' = K\beta'$, where K is any matrix with r columns, and define a process w_t as

$$w_t = \gamma' y_t = K z_t.$$

Since w_t is a linear combination of $I(0)$ processes, it is also an $I(0)$ process, so that γ' is itself a cointegrating vector (or matrix). To put it more formally, the cointegrating matrix is defined only up to an invertible linear transformation. As we shall see, this has significant implications for estimation and testing.

The algebraic properties of cointegration vectors are nice and cool, but their usefulness in terms of economic interpretation is even more interesting. Consider what happens to two cointegrated $I(1)$ variables: being DS processes, they will possess all the characteristics discussed in section 4.2, such as the absence of *mean-reversion* and so forth. However, there exists a unique (and, in this specific example, the only) linear combination of them that is stationary. That is to say, this very specific combination fluctuates around a

well-defined mean, possesses only short-term rather than long-term memory, and shows no trending tendency.

This behaviour can be rephrased in terms of a concept that we economists like to use a lot: equilibrium. The original variables may drift freely, yet a relationship exists between them that (perhaps approximately) always holds and can be thought of as an **equilibrium relationship**. If we choose to think of $I(1)$ variables as perpetual wanderers, the more romantically inclined among us might view two cointegrated variables as a **pair** of perpetual wanderers. Cointegration binds their fate forever: they will wander through eternity, always close to one another[1].

Example 6.2.1 (The quantity theory of money) *Let's begin from an old-fashioned example, close to the heart of vintage right-wing economists. The quantity theory of money, in its most basic form, states that a stable proportion exists between money supply and GDP: this ratio is termed the "velocity of money". In formulas:*

$$MV = PY.$$

Alternatively, it can be stated that the velocity of money is given by the ratio of GDP to the real money supply:

$$V = \frac{Y}{M/P}.$$

In logarithms, we obtain:

$$v = y - m,$$

where y is the logarithm of GDP and m is the logarithm of the real money supply. A quantity theory extremist would argue that v is a fixed quantity. A less extreme proponent might suggest that v is a quantity which, when observed over time, exhibits more or less persistent fluctuations around a central value: the time series of v_t therefore should resemble the realisation of an $I(0)$ process. If y_t and m_t can be represented as $I(1)$ processes, supporting the quantity theory of money is roughly equivalent to saying that the two processes y_t and m_t cointegrate, and that the cointegrating vector is $(1, -1)$.

Example 6.2.2 (Purchasing Power Parity) *Consider two countries, A and B and denote the price levels in the two countries at time t as P_t^A and P_t^B. If P_t^A rises while P_t^B remains unchanged, the goods produced in country B clearly become more competitive than those produced in country A. To cut a long story short (if you want the full explanation, you can study it in International Economics), this triggers a chain of consequences by which the currency of country A should depreciate progressively against that of country B, and hence the exchange rate E_t should also tend to rise over time.*

If Purchasing Power Parity (also known as PPP) holds between these two economies, the exchange rate between the currencies of countries A and B will move over time in such a way as to offset the differences in inflation rates between the two nations. We could therefore conjecture that the quantity

$$\frac{P_t^B E_t}{P_t^A}$$

[1] A commendable article published in *The American Statistician* suggests the analogy of a drunk and her dog, which Quentin Tarantino fans may find not as corny.

should remain more or less constant over time. As in the previous example, supporting this hypothesis is equivalent to asserting that, if the series are $I(1)$, the relationship

$$p_t^B + e_t - p_t^A$$

(where, once again, lowercase letters indicate logarithms) is a cointegrating relationship.

The examples above should are meant to illustrate that cointegration is a statistical concept with a very natural link to the economic theory, because it makes it describes long-run equilibrium relationships among two or more variables: if a cointegrating relationship exists, the equilibrium relationship may never hold exactly, but any deviations will nonetheless be limited and temporary.

More explicitly: suppose you have two cointegrated processes y_t and x_t, with cointegration vector equal to $[1, -\beta]$. Cointegration implies that

$$z_t = y_t - \beta x_t$$

is stationary (suppose for simplicity that $E(z_t) = 0$). You can rewrite the equation above as

$$y_t = \beta x_t + z_t$$

and read it as: "in equilibrium, the proportion between y_t and x_t should be β; this may not be true at a particular time t, so we call z_t the deviation from equilibrium at time t, which is zero on average".

The parameters defining the cointegration relationship can sometimes be dictated *a priori* by economic theory, as in the examples above; however, there may be cases when economic theory does not dictate precise numerical values for the parameters appearing in the equilibrium relationship, and indicates their sign at most. Hence, at the end of the 20th century an enormous effort was made to find a way to estimate cointegrating relationships, rather than merely postulating them. After all we've been saying on unit root tests and spurious regression, one should expect that estimating cointegrating vectors is far from trivial, and that is indeed the case. To understand how to proceed, it will be useful to turn our attention to the representation of cointegrated systems before moving on to statistical inference.

6.3 Error-correction models

Let us start with the simplest case, where the system admits a VAR(1) representation. We therefore consider an n-dimensional process y_t of the form

$$y_t = Ay_{t-1} + \epsilon_t. \tag{6.1}$$

Subtracting y_{t-1} from both sides yields

$$\Delta y_t = \Pi y_{t-1} + \epsilon_t, \tag{6.2}$$

where $\Pi = A - I$. Note that the equation above is just a multivariate version of equation (4.7), that I used to illustrate the DF test. If y_t is a scalar process, Π

can only be either zero, or non-zero. If y_t is a vector, you get an intermediate case, which is precisely the one of interest, depending on the rank of the matrix Π. If we denote this rank as r, we have three possibilities:

$r = 0$:	in this case, $\Pi = 0$, so y_t is a multivariate random walk, and there's no cointegration;
$r = n$:	y_t is $I(0)$, since Π is invertible (see section 6.2);
$0 < r < n$:	y_t is a cointegrated system.

The analysis of this final case is the object of this chapter: as we shall see, r is the cointegration rank; furthermore, Π can be written as $\alpha\beta'$, where α and β are two $n \times r$ matrices; in particular, β is the cointegrating matrix. By using $\Pi = \alpha\beta'$, equation (6.2) can be rewritten as

$$\Delta y_t = \alpha z_{t-1} + \epsilon_t, \tag{6.3}$$

where $z_t = \beta' y_t$ is an $r \times 1$ vector. Note that Δy_t is $I(0)$ and ϵ_t is white noise, so z_t must be $I(0)$ too, the matrix β being the cointegrating matrix.

Can we attach a meaning to the variables z_t? As we saw earlier, they can be seen as the deviations from the cointegrating relationships. Put this way, equation (6.3) states something almost trivial: the movement of a cointegrated system is determined by two factors. One (ϵ_t) is random and, in the specific case of our example, is a white noise process; the other (αz_{t-1}) is determined by the magnitude — lagged by one period — of the deviation from the cointegrating relationship, which can be interpreted as the level of the disequilibrium at time $t-1$. The matrix α is called the **loading matrix**, as its ij-th element indicates the effect of the j-th element of z_{t-1} on the i-th variable.

Example 6.3.1 *Let us go back to example 6.2.1. Suppose that, at a given time t, there is more money in the economy than predicted by the equilibrium relationship. In other words, the velocity of money is below its mean value. Consequently, the change in GDP and real money supply (in logarithms) between t and $t+1$ will be given by the sum of two components, which can be summarised as*

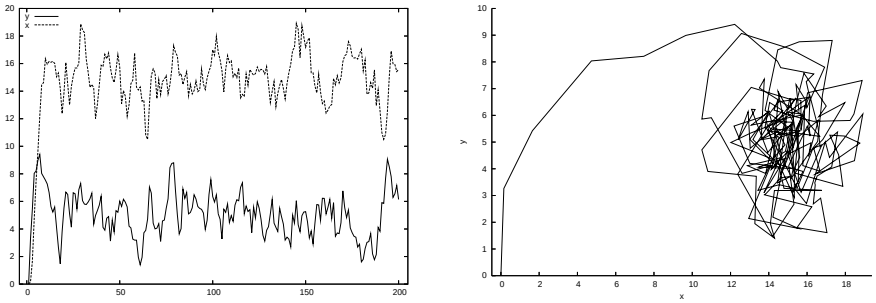
$$\begin{cases} \Delta y_{t+1} = \alpha_1 (y_t - m_t) + \epsilon_{1t} \\ \Delta m_{t+1} = \alpha_2 (y_t - m_t) + \epsilon_{2t}. \end{cases}$$

Recall that $(m_t - y_t)$ can be interpreted as the logarithm of the money velocity; a lower-than-normal value triggers an adjustment, in the next period, of both real income and the real money supply. If, for instance, α_2 were positive, it would imply that in the presence of excess money, Δm_t must be negative, meaning that m_t must fall. This could occur, for example, through a rise in prices. According to hard-core monetarists from the 1970s, this is all you need to explain inflation.

As you may already know, the mechanism I just described has a distinct name, namely the **Error Correction Mechanism**, or more briefly **ECM**; a VAR rewritten in ECM form is often also called a **VECM** (vector ECM). ECM models rightfully occupy a prominent position in modern econometrics, precisely because they provide a contact point between time series analysis and economic theory, between short- and long-run. If a the rank of a cointegrated

system is r, then there exist r long-run equilibrium relationships, and therefore r stationary processes describing the evolution over time of the disequilibrium from these relationships. Whenever the disequilibrium (which we have denoted as z_t) is non-zero, a movement in the vector y_{t+1} will occur such that the disequilibrium tends to be absorbed.

Figure 6.1: Stationary VAR(1): simulated time series
(a) Over time (b) Scatter path



To understand this idea, the concept of an **attractor** is quite useful. In this context, the idea can be conveyed by defining an attractor as a set of points around which a dynamic system tends to hover. For example, let us consider the following stationary VAR(1) process:

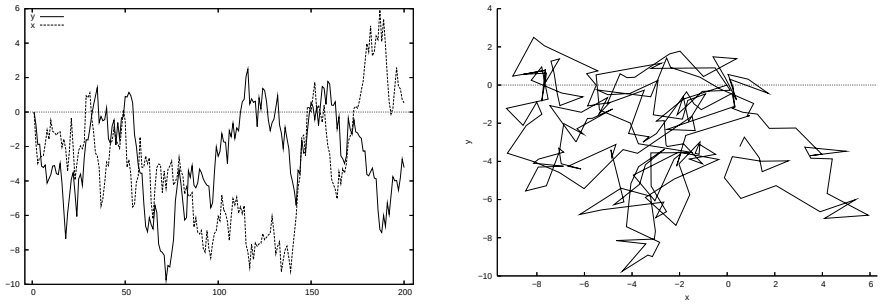
$$\begin{aligned} y_t &= 4 + 0.8y_{t-1} - 0.2x_{t-1} + \epsilon_{1t} \\ x_t &= 2 + 0.2y_{t-1} + 0.8x_{t-1} + \epsilon_{2t}. \end{aligned}$$

With a minimum of algebra, it can be seen that the unconditional mean of the vector $[y_t, x_t]$ is $[5,15]$.

Figure 6.1 displays a 200-observation simulation of this process, starting from the point $[0,0]$. Panel (a) on the left is the customary time-series plot; as can be seen, the two series first jump away form zero, and then oscillate around the mean, which acts as an attractor. Indeed, starting from a "distant" point, their immediate response is to go back towards the attractor, which is precisely the pair of values $[5,15]$. This can be seen even more clearly in panel (b), in which the values of the series are plotted on a plane, that is, as a point for each observation with segments joining consecutive points in time. Here, the path taken by the system over time is clearly visible: it starts from the point $[0,0]$, rapidly moves towards the attractor point, and then lingers around it.

If, on the contrary, we perform the same exercise with a bivariate random walk, things look quite different: in this case, there is no point towards which the series tend, as the attractor is the entire plane. Indeed, if two drunks start from the bar $[0,0]$, they wander through town without aiming anywhere (see Figure 6.2). I would point out that panel (b) of Figure 6.2 also provides an intuitive idea of why the regression between y_t and x_t is, in this case, spurious: the line that passes as close as possible through the middle of the points will, in all probability, have a non-zero slope, despite there being no systematic link between the two variables.

Figure 6.2: Bivariate random walk: simulated time series
(a) Over time (b) Scatter path



Let us now examine a cointegrated process. In this case, the data-generating process is

$$\begin{aligned} y_t &= 0.5y_{t-1} + 0.5x_{t-1} + \epsilon_{1t} \\ x_t &= -0.2y_{t-1} + 1.2x_{t-1} + \epsilon_{2t}, \end{aligned}$$

which can be written in VECM form as

$$\begin{aligned} \Delta y_t &= -0.5z_{t-1} + \epsilon_{1t} \\ \Delta x_t &= -0.2z_{t-1} + \epsilon_{2t}, \end{aligned}$$

where $\beta' = [1, -1]$, and therefore $z_t = y_t - x_t$.

Figure 6.3: Cointegrated process: simulated time series
(a) Over time (b) Scatter path

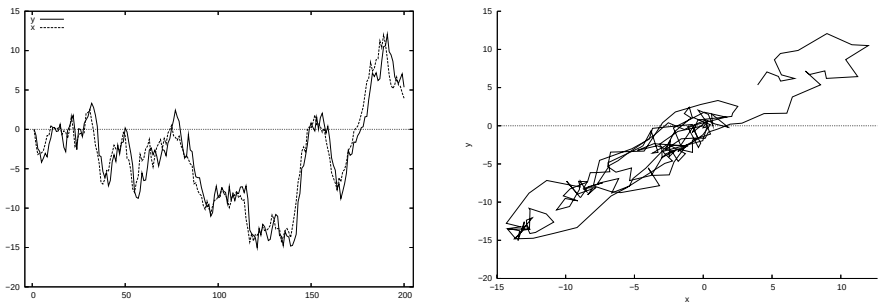


Figure 6.3 shows the typical behaviour of two cointegrated series with a cointegrating vector of $[1, -1]$ (or, if you prefer, the 'master and dog' analogy). Even more interesting, however, is what can be observed in the XY scatter plot: in this case, the equilibrium relationship is $y_t = x_t + z_t$; if z_t was 0, that would be a straight line with a 45° slope. Here, the attractor is precisely this line. Note that the two series, taken separately, display the classic behaviour of an $I(1)$ process; however, the system never drifts away from the equilibrium line. In other words, should a disequilibrium occur ($z_t \neq 0$), the system would go back to the attractor on its own accord. In contrast to the stationary case,

however, the attractor is not a single point but the infinite set of points along the equilibrium line. Consequently, in the long run, it is impossible to know exactly where the system will be, but we do know that it will be close to some point on the attractor.[2]

A straightforward generalisation that we can make at this stage is that an equation of the form (6.3), where the white noise process ϵ_t is replaced by a generic $I(0)$ process, can also be considered an ECM. The interpretation remains more or less the same. However, the ECMs most frequently encountered in the applied literature are those where the short-run persistence in the disturbance term is eliminated via an autoregressive process in the following way:

$$\Delta y_t = \alpha z_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t. \tag{6.4}$$

If z_t were an observable series, parameter estimation could be straightforwardly carried out using OLS, since all the series in equation (6.4) are stationary. Unfortunately, z_t is often unobservable because β in unknown. We'll take care of estimating β in Section 6.5.

It is interesting to note that an equation such as (6.4) can be derived from a VAR representation of y_t using the BN decomposition (see section 4.3). Assume that $A(L)y_t = \epsilon_t$; then,

$$y_t = B(L)y_{t-1} + \epsilon_t,$$

where $B(z) = \frac{I - A(z)}{z}$. Applying the BN decomposition to $B(L)$, we can write

$$y_t = [B(1) + \Gamma(L)\Delta]\, y_{t-1} + \epsilon_t.$$

Subtracting y_{t-1} from both sides yields

$$\Delta y_t = [B(1) - I]\, y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \epsilon_t.$$

From the definition, it is straightforward to verify that $B(1) = I - A(1)$, so that

$$\Delta y_t = -A(1) y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \epsilon_t \tag{6.5}$$

and note, once again, the similarity of this equation to (4.10). The presence of cointegration is linked to the rank of the matrix $A(1)$, which plays the same role that Π played in equation (6.2).

6.4 The Granger Representation Theorem

As we have seen in the previous section, the presence of cointegration in a multivariate stochastic process allows for the interpretation of some parameters (the cointegrating matrix) in an extremely insightful way. This interpretation becomes immediate when the process is written in VECM form. In the

[2] More generally, it can be shown that in a cointegrated system with n variables and a cointegration rank r, the attractor is an $(n - r)$-dimensional hyperplane.

previous section, we assumed that this was possible. In this section, we will further clarify the nature of the link between cointegrated processes and ECM models: this is what the **Granger representation theorem**[3] is about. As we shall see, the analysis of this theorem yields, as a by-product, a substantial body of insights regarding cointegrated systems.

To explore the properties of a cointegrated system y_t, at least two possibilities come to mind: one stems from the possibility that the system could be written as a VAR (possibly of infinite order),

$$A(L)y_t = \epsilon_t \tag{6.6}$$

with its associated ECM representation; and another linked to the property that, if y_t is $I(1)$, then Δy_t is $I(0)$, and must therefore possess a Wold representation of the form

$$\Delta y_t = C(L)\epsilon_t. \tag{6.7}$$

Are these two representations related? And, if so, how? The answer is precisely contained within the Granger representation theorem. To tell the truth, this theorem is a sort of nine-headed hydra, as it lacks a single, well-defined statement; rather, it refers generally to everything that can be said regarding the representation of a cointegrated system. This makes it somewhat difficult to digest, yet it is immensely rewarding, as it demonstrates how various aspects of a cointegrated system can be viewed from multiple perspectives.

It is not my ambition to improve upon what has already been excellently achieved by minds far better than mine, so I will not offer a formal proof of this theorem; instead, I shall merely confine myself to exploring its salient points. Before doing so, however, a brief preamble is necessary.

6.4.1 Some matrix algebra

In the following, I will use a matrix operator that may not be familiar to everyone, denoted by the symbol "$\perp$" and read as "orthogonal". Thus, $X_\perp$ is read as "X orthogonal".

If X is any $l \times m$ matrix, with $l > m$ and rank m, $X_\perp$ is an $l \times (l-m)$ matrix whose columns are linearly independent and orthogonal to those of X.[4] In other words, $X'_\perp X = 0$ by definition[5], and there are no vectors orthogonal to X that are not linear combinations of the columns of $X_\perp$.

The "$\perp$" operator will serve us primarily for one purpose: if a matrix A possesses the property

$$x'A = 0,$$

then it follows that every column of A must be orthogonal to x. However, since any vector orthogonal to x is a linear combination of $x_\perp$, then A must be expressible as $A = x_\perp B$, where B is "something else" that will vary depending on the context.

[3] This is the same Granger of causality and spurious regression fame. In the motivation for the Nobel Prize he was awarded in 2003, cointegration is explicitly mentioned.

[4] For algebra enthusiasts: the columns of $X_\perp$ span a vector space known as the **null space** of X. Consequently, $X_\perp$ is not unique. However, this is of no importance.

[5] Note: $X'_\perp X = 0$ **does not** imply that $XX'_\perp = 0$.

Example 6.4.1 *To gain an intuitive understanding, the reader is invited to verify the properties of the $\perp$ operator on the matrices*

$$w = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and

$$w'_\perp = \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}.$$

Now consider the matrix

$$A = \begin{bmatrix} 2 & -1 \\ -2 & 1 \\ -2 & 1 \end{bmatrix};$$

it can be easily verified that $w'A = 0$ (just multiply). This is enough to state that a matrix B must exist such that $A = w_\perp B$. For the record, this matrix is $B = [2 \quad -1]$ in the present case.

6.4.2 The theorem itself

Theorem 2 (The Granger Representation Theorem) *Every cointegrated system admits an ECM representation; conversely, if an ECM representation exists and the series are integrated, then they are cointegrated.*

The proof is rather complex, but I provide an outline in Section 6.A.1 to help understand the key points. Also, in this subsection I'll concentrate on the special case of a process y_t with no deterministics ($E[\Delta y_t] = 0$), so as to focus on the algebraic part of the theorem. General cases are in Section 6.4.3.

The core of the theorem is as follows: a cointegrated system can be expressed in two equivalent ways, corresponding to the autoregressive and moving average representations. The representation derived from the autoregressive form is the VECM (see eq. (6.4)), reproduced here for your convenience:

$$\Delta y_t = \alpha\beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \epsilon_t.$$

The representation derived from the moving average form is, in a way, an adaptation to the cointegrated case of the Beveridge-Nelson decomposition, which can be written as

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t,$$

where

$$C(1) = \beta_\perp H \alpha'_\perp \tag{6.8}$$

and μ_t is defined by the property $\Delta\mu_t = \epsilon_t$. The key equation here is (6.8), where H is an invertible matrix whose details are, for now, unimportant.[6]

[6] More on this, if you want, in Section 6.A.1.

Naturally, since ϵ_t is a vector white noise process, μ_t is a vector random walk by definition. Note that $C(1)$ matrix has the property of vanishing if pre-multiplied by β', on account of to the properties of the $\perp$ operator. In other words, $C(1)$ is singular, with rank $(n-r)$.

Putting the two expression above together

$$y_t = \left[\beta_\perp H \alpha'_\perp\right] \mu_t + C^*(L)\epsilon_t \tag{6.9}$$

,

you get what is known as Stock and Watson representation, or the **common trends** representation, and now we shall see why. The series μ_t is, as we previously argued, an n-dimensional random walk, and therefore a non-cointegrated $I(1)$ process. Let us now construct a series η_t as

$$\eta_t = \alpha'_\perp \mu_t.$$

Evidently, η_t is an $(n-r)$-dimensional random walk, because the matrix $\alpha_\perp$ is an $n \times (n-r)$ matrix. It is therefore possible to rewrite (6.9) as follows

$$y_t = F\eta_t + u_t,$$

where $F = \beta_\perp H$ is an $n \times (n-r)$ matrix and $u_t = C^*(L)\epsilon_t$ is, by assumption, a stationary process.

Writing y_t in this manner clearly shows that a cointegrated system can be conceptualised as a system driven by a certain number $(n-r)$ of unobservable stochastic trends, which drive the observable series through the matrix F. The time series we observe contain an $I(1)$ component (given by linear combinations of these stochastic trends) and an $I(0)$ component (given by u_t). Cointegration exists precisely because $n - r < n$, so that the linear combination $\beta' y_t$ is a stationary process simply by virtue of the relationship $\beta' F = 0$. In practice, by pre-multiplying y_t by β', we simply kill the common trends.

Example 6.4.2 *Take the pair of cointegrated series plotted in Figure 6.3, where $\beta = (1,-1)'$; consequently, we will find that $z_t = x_t - y_t$ is $I(0)$. In this case, one possible choice for $\beta_\perp$ is $\beta_\perp = (1,1)'$; H is a scalar, so $F = (H,H)'$.*

The two time series can therefore be written as the sum of an $I(1)$ process, which is identical for both, and a stationary component. In practice, the two time series fluctuate around the same stochastic trend and never drift far apart precisely because both tend to remain in the proximity of this common trend. It is clear that cointegration arises from the fact that by taking the difference between the two series, the stochastic trend cancels out, leaving only the fluctuations around it.

It is worth noting that this highly simplified case is ideally suited to being contextualised within the example of the quantity theory of money introduced a few pages earlier. In this scenario, both the money supply and income travel along a common trend, which — with a little imagination — could be thought of as the long-run growth path of the entire economy.

6.4.3 Deterministic component

The situation becomes more complex when the process Δy_t has a non-zero mean. Typically, we consider processes of the form

$$\Delta y_t = d_t + \alpha\beta' y_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t$$

that is, a vector ECM process to which a deterministic component is added, which usually contains an intercept and a linear trend, namely $d_t = \delta_0 + \delta_1 t$.

In analogy with what was discussed in Section 4.2, one might think that if d_t is a polynomial of order p in time (that is, something of the form $\delta_0 + \delta_1 t + \cdots + \delta_p t^p$), then a polynomial of order $p+1$ will, in general, be present in the series in levels y_t. In the case of cointegration, however, this is not necessarily true. Let us see why by considering again, for simplicity, the case where the series in levels can be represented as a VAR(1). If

$$y_t = d_t + A y_{t-1} + \epsilon_t$$

we can rewrite the model in ECM form by merging the deterministic component with the disturbance term as follows:

$$\Delta y_t = \alpha\beta' y_{t-1} + (\epsilon_t + d_t).$$

If we define $u_t = \epsilon_t + d_t$, we can replicate the entire previous reasoning using u_t instead of ϵ_t. In this case, we would have that

$$y_t = \left[\beta_\perp H \alpha'_\perp\right] \tilde{\mu}_t + C^*(L) u_t,$$

where $\Delta\tilde{\mu}_t = u_t = \epsilon_t + d_t$. The process $\tilde{\mu}_t$ thus results from the sum of a multivariate random walk (nothing new so far) plus a deterministic component whose first difference is a polynomial in time of order p, that is, a polynomial of order $p+1$.

Let us consider the special case where $d_t = \delta_0$; the polynomial in t present in $\tilde{\mu}_t$ will be something of the form $\delta_0 t + k$; in practice, a multivariate random walk with a drift. Since in a cointegrated system the matrix $\alpha'_\perp$ has rank $(n-r)$, it may well be that $\alpha'_\perp \delta_0 = 0$.

If this is the case, premultiplying $\tilde{\mu}_t$ kills the drift but not the random walk, so $\alpha'_\perp \tilde{\mu}_t$ will be a *driftless* random walk with dimension $n-r$.

In other words, if $\alpha'_\perp \delta_0 = 0$, then δ_0 can be expressed as a linear combination of the columns of the matrix α, as in

$$\alpha'_\perp \delta_0 \Longrightarrow \delta_0 = \alpha d.$$

If the above holds, the component of order 1 of the polynomial contained in $\tilde{\mu}_t$ vanishes, and the linear trend present in $\tilde{\mu}_t$ does not appear in y_t. Moreover, if $\delta_0 = \alpha d$, then the VECM representation can be written as follows:

$$\Delta y_t = \alpha(d + \beta' y_{t-1}) + \epsilon_t$$

and therefore you can think of the long-run equilibrium relationship as containing an intercept, given by the vector d.

More generally, given a polynomial function of time like

$$d_t = \delta_0 + \delta_1 t + \delta_2 t^2 + \ldots + \delta_p t^p,$$

where the various δ_i are n-element vectors, it is not necessarily the case that the polynomial $\alpha'_\perp d_t$ itself has degree p: it depends on whether $\alpha'_\perp \delta_p$ is zero or not. As we now know, if $\alpha'_\perp \delta_p = 0$, it means that δ_p is some linear combination of the columns of α.

In practice, you get the famous "five cases", usually ordered from the most to the least restrictive:

Case	Deterministics	Description
1	$d_t = 0$	There is no deterministic component at all. The common trends are driftless random walk processes and the z_t have zero mean. The data should show no evidence of deterministic trends.
2	$d_t = \delta_0; \quad \alpha'_\perp \delta_0 = 0$	Here, things are more complex: the VECM representation contains an intercept, which, however, does not give rise to a linear trend in the common trends representation, because the latter do not have a drift. The data do not exhibit deterministic trends, but fluctuate around a non-zero value. The disequilibria z_t have a non-zero mean, so that one can think of an intercept in the cointegration relationship.
3	$d_t = \delta_0; \quad \alpha'_\perp \delta_0 \neq 0$	In this situation, the VECM intercept is unrestricted, and therefore, in principle, we have an intercept both inside the cointegration relationship (which does not generate a trend in the observed time series) and outside it (giving rise to a drift in the common trends representation, and thus a trend in the observed series). In practice, however, these two intercepts are usually combined, and the cointegration relationship is presented without an intercept.
4	$d_t = \delta_0 + \delta_1 t; \quad \alpha'_\perp \delta_1 = 0$	This case runs parallel to Case 2: the cointegration relationship has a linear trend, which, however, does not translate into a quadratic trend in levels.
5	$d_t = \delta_0 + \delta_1 t; \quad \alpha'_\perp \delta_1 \neq 0$	Here, there are no restrictions of any kind, and the time series exhibit a quadratic trend.

Case 2 is the one that usually remains most obscure: I will try to explain how it works by means of an example.

Example 6.4.3 *Suppose we have a process x_t, which is a random walk with drift*

$$x_t = m + x_{t-1} + \varepsilon_t; \tag{6.10}$$

As is well known, x_t is an $I(1)$ process that fluctuates around a deterministic trend; the latter has a slope of m. If $m = 0$, the deterministic trend disappears and we have a pure random walk.

Now consider a second process:

$$y_t = k + x_t + u_t, \tag{6.11}$$

where u_t is a white noise. Since u_t is stationary by definition, x_t and y_t are cointegrated because their difference $z_t = y_t - x_t$ is stationary, namely a white noise with a mean of k.

By plugging equation (6.10) into (6.11), we represent the system as a VAR(1)

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} k+m \\ m \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix}, \tag{6.12}$$

which can be transformed into VECM form by subtracting the vector $[y_{t-1}, x_{t-1}]'$ from both sides:

$$\begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} = \begin{bmatrix} k+m \\ m \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix}, \tag{6.13}$$

where the matrix $\begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}$ is the one we referred to as Π in equation (6.2). The matrix Π is singular (its rank is 1), which makes it possible to write it as $\alpha\beta'$:

$$\begin{aligned} \begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} &= \begin{bmatrix} k+m \\ m \end{bmatrix} + \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix} = \\ &= \mu_0 + \alpha\beta' \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \eta_t = \mu_0 + \alpha z_{t-1} + \eta_t. \end{aligned}$$

Now consider the three possibilities:

1. *$m \neq 0$: As previously stated, x_t has a trend. It follows (from (6.11)) that y_t also has one, because on average it remains at a distance of k from x_t. The vector μ_0 is not subject to any restrictions, and we are in Case 3.*

2. *$m = 0$ and $k \neq 0$: Here, x_t has no deterministic trend, and consequently, neither does y_t. However, the average distance between y_t and x_t is non-zero. The vector μ_0 is also non-zero:*

   $$\mu_0 = \begin{bmatrix} k \\ 0 \end{bmatrix},$$

   so the VECM in (6.13) has an intercept. However, its second element is zero, which makes μ_0 is a multiple of the vector α. We can therefore rewrite the VECM as follows:

   $$\begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & -k \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \\ 1 \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix},$$

   namely, by incorporating the intercept into the cointegration vector. This is Case 2 (the one with a "restricted constant").

3. $m = 0$ *and* $k = 0$*: This is Case 1, the most restrictive one.* x_t *and* y_t *have no trend, and the distance between them is, on average, 0. The VECM has no intercept.*

When building an empirical model, the choice among these different possibilities is based both on the observation of the time series and on some kind of a priori *reasoning. If the two series share a clear trending behaviour, then imposing restrictions on the intercept is inappropriate. Otherwise, one might wonder whether it makes sense to include an intercept in the cointegration relationship. The textbook case arises when we examine two interest rates: normally, these do not have a trend,*[7] *but the VAR may well have an intercept, because their difference (the "spread") can have a non-zero mean (for instance, due to a risk or liquidity premium).*

This may seem like an algebraic quirk without any practical relevance; in reality, these complications become important for estimation, as we shall see in the next section.

6.5 Estimation techniques

For estimating the parameters of a cointegrated system, there are three main cases to consider:

1. The cointegration matrix is known, and consequently so is its rank: this may be, for instance, the case where one starts from a working hypothesis that assumes *a priori* the validity of the purchasing power parity theory.

2. The cointegration rank is known, but the cointegration matrix is not. This case may arise, for example, if we believe that oil price and fuel price form a cointegrated system; we assume that thy move together, but we do not what the pass-through rate is.

3. Neither the rank nor the cointegration matrix are known *a priori*.

The last case is of course the most general one: even if you have some prior ideas on how things work, it can be scientifically interesting to check whether they hold empirically. The standard way of dealing with it is by using an approach devised by Søren Johansen, that I'll illustrate in the next subsection.

[7] And rightly so (remember the discussion on the Babylonians I introduced regarding unit root tests in subsection 4.4.7?). However, you should be aware that in certain contexts it makes perfect sense to describe the movement in interest rates as something superimposed on a linear trend. Take for example the yields on US Treasury Bills between 1980 and 2010: the trend is rather clear, and it's due to exogenous macroeconomic factors (declining inflation, the 2007 financial crisis, etc.). If we do not wish to model the underlying macroeconomic framework, we can simply include a trend and be done with it. Nevertheless, with longer samples, this issue does not usually arise.

6.5.1 The Johansen method

A full account of the Johansen procedure would be rather complex,[8] so I shall provide a very cursory description; now more than ever, I refer the reader to the literature.

In this procedure, it is assumed that the cointegrated system can be represented as a finite-order VAR with Gaussian errors. The starting point is to reparameterise the system in ECM form

$$\Delta y_t = d_t + \Pi y_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t; \tag{6.14}$$

implicitly, we assume that the order of the VAR (p) is known. However, finding the a suitable VAR order does not imply any particular problems, because standard methods can be used (most commonly, information criteria).

Under these premises, the estimation problem can be set up within a maximum likelihood framework. Normally, the maximum likelihood estimator of the parameters of a linear regression model with Gaussian errors is none other than OLS. In this case, however, it is also necessary to take into account the constraints associated with the presence of cointegration: in fact, the rank of the matrix Π is equal to the cointegration rank r, so we want the estimation method to guarantee an estimate of Π with reduced rank (which OLS does not produce).

To get an initial estimate of Π, the Johansen method uses the so called "Frisch-Waugh" theorem (described in Section 6.A.2 for those needing help). This gives you $\hat{\Pi}$, an initial consistent estimate of Π. You now need a way to recover α and β from $\hat{\Pi}$.

The problem is that, under cointegration, Π has rank r, whereas $\hat{\Pi}$ has full rank n with probability one. Therefore, we need a way to establish r, the cointegration rank, from the data, so that we can devise with a modified estimator $\hat{\Pi}$ whose rank is guaranteed to be equal to r.

Determining the cointegration rank

The determination of the rank of Π hinges on the matrix algebra fact that for a positive semi-definite matrix the number of positive eigenvalues is equal to its rank, and the others are zero. In the Johansen procedure, two tests are defined on this basis.

These tests work as follows: first, a matrix M is defined (the details of which I shall omit) which by construction is positive semi-definite and has the same rank as Π. The advantage of working with M rather than Π is that the semi-definiteness of M ensures that all its eigenvalues are non-negative real numbers.

Given $\hat{\Pi}$, a consistent estimate $\hat{M}$ of this matrix is available, with the consequence that the eigenvalues of $\hat{M}$ (call them $\hat{\lambda}$) are consistent estimators of

[8] Conversely, estimating a cointegrated system via the Johansen method is actually rather simple, since that the Johansen method is available in all econometric software packages.

the n eigenvalues of M. These are sorted from the largest $\hat{\lambda}_1$ to the smallest $\hat{\lambda}_n$, and then a test is set up to determine whether *the smallest one* is zero.

This test for $H_0 : \lambda_n = 0$ has, asymptotically, a non-standard distribution: fortunately, however, its details were worked out by Johansen first and then refined by subsequent literature, so performing the test is easy.

If the null is rejected, stop: if λ_n is positive, then all of them are positive, so Π has full rank and the system is stationary. Otherwise, one proceeds to consider λ_{n-1}. Here, you have two possibilities:

1. perform a test for $H_0 : \lambda_{n-1} = 0$, taking it for granted that $r < n$ and thus testing the hypothesis $r < n - 1$; this is the so-called λ-max test;

2. alternatively, perform a test for $H_0 : \lambda_n = \lambda_{n-1} = 0$, that is, a *joint* test that the two smallest eigenvalues are both zero, where $r < n$ is not assumed; this is the so-called trace test.

Again, if H_0 is rejected, you stop. Otherwise, you proceed with $\hat{\lambda}_{n-2}$, and so on. By stopping at the first rejection, you obtain an estimate of r. If, on the other hand, the null is never rejected, it means that $r = 0$ and there is no cointegration.

Note that the univariate equivalent to equation (6.14) is exactly the one used for the Augmented Dickey–Fuller test,

$$\Delta y_t = d_t + \rho y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \epsilon_t;$$

so the Johansen test can be seen as the multivariate generalisation of the ADF test. The only difference is that, in the ADF test, the coefficient ρ is a scalar and therefore is either 0 or invertible; in the case of cointegrated systems, however Π may be singular without being zero.

The extreme cases are of interest: failing to reject $H_0 : \lambda_1 = 0$ means that in fact $\Pi = 0$, so there is no cointegration and a system is an ordinary $I(1)$ process. On the other hand, rejecting $H_0 : \lambda_n = 0$, means that Π is invertible, and therefore the system is not even $I(1)$, but rather stationary. This is what I meant in the statement I made a few pages ago (namely at the end of subsection 4.4.7), that the Johansen tests can replace in some cases univariate unit-root tests.

Given the parallelism between the Johansen tests and ADF, it is not surprising that they have non-standard asymptotic distributions. Furthermore, just as occurs in the ADF test, these limiting distributions are not invariant to the deterministic component of the VAR. This may make matters quite intricate, in the light of the remarks made in section 6.4.3.

This is because one must decide the most appropriate degree of the polynomial in t to include in the deterministic component: the usual choice, in univariate unit-root tests, is between a constant or a constant plus trend. However, as we saw in the previous section, it cannot be ruled out that some of the parameter vectors of the deterministic component may disappear once multiplied by $\alpha_\perp$: this leads to the five different possibilities that I mentioned earlier, for each of which there is a different asymptotic distribution.

Estimation of cointegration vectors

Once the cointegration rank that offers a statistically sustainable description of the data has been found, and a decision on r has been made, β can be estimated.

The main complication arising here is that — as I pointed out in section 6.2 — the matrix β is not identified: if β is a cointegration matrix, so is $b = \beta K$, where K is any non-singular $(r \times r)$ matrix. Consequently, there exists an infinite number of $n \times r$ matrices that are observationally equivalent. This argument can also be viewed more explicitly: assume that we know the "true" matrix Π. Now, this matrix can be represented both as

$$\Pi = \alpha\beta'$$

and as

$$\Pi = \alpha K^{-1} K\beta' = ab'$$

and the two representations are completely equivalent. There is no way of using the data to decide which one to use: the classic under-identification problem.

Using some linear algebra concepts, the under-identification problem can also be viewed from a geometric perspective: indeed, it can be shown that the cointegration vectors (the columns of β) form a basis for the r-dimensional hyperspace that is orthogonal to that of the attractor. As is well known, the choice of a basis is arbitrary.

The classic way out from under-identification is by setting appropriate restrictions so as to reduce the number of admissible representation to one. This is a very general point in econometrics, but the reader may recall we discussed some problem at some length in Section 5.6.

It can be shown that the minimum number of restrictions required to identify the parameters is equal to r^2. In the procedure originally proposed by Johansen, identification is achieved by stipulating some constraints on a certain quadratic form in β, that makes perfect sense algebraically, but lacks any meaningful economic interpretation. An alternative approach to identification gives rise to the so-called **Phillips representation** (after P. C. B. Phillips), were it is assumed that the first r rows of β form an identity matrix. Therefore, we can write

$$\tilde{\beta} = \begin{bmatrix} I \\ -\tilde{\beta}_2 \end{bmatrix} \tag{6.15}$$

where $\tilde{\beta}_2$ is, instead, unrestricted.[9]

In general, ML estimation may require iterative algorithms, but fortunately in this case they are not required: the matrix β can be easily recovered from the eigenvectors of the matrix $\hat{M}$ that I mentioned in relation to the trace and λ-max tests, associated to the r largest eigenvalues. From the standpoint of asymptotic theory, the estimator of β has several unusual characteristics: the most remarkable is what is known as **super-consistency**.

[9] Nothing prevents further restrictions from being imposed on $\tilde{\beta}_2$, of course. In this case, however, the mechanics of estimation are more complex. More on this later.

Allow me a brief review of asymptotic theory: the distribution of a consistent estimator collapses to a single point as $T \to \infty$, which is briefly expressed by $\hat{\theta} \xrightarrow{\text{p}} \theta_0$. Thus, $(\hat{\theta} - \theta_0) \xrightarrow{\text{p}} 0$. In standard cases, multiplying this quantity by $\sqrt{T}$ yields something that neither collapses nor diverges, but stabilises into a limiting distribution, most often Gaussian (which is what we call asymptotic normality). And in fact, in the world of stationarity, $\sqrt{T}(\hat{\theta} - \theta_0)$ becomes, as T increases, a random variable that is increasingly similar to a normal distribution with zero mean.

In the case we consider here, on the contrary, the difference $(\hat{\beta} - \beta)$ must be multiplied by T instead of $\sqrt{T}$ to obtain a limiting distribution. In practice, the speed at which $\hat{\beta}$ collapses towards β is much greater; in other words, the dispersion of the estimator is proportional to T^{-1} rather than $T^{-1/2}$. This property, however, does not imply that in finite samples this estimator should be particularly precise: its bias has been found to be rather severe in some cases, and with the sample sizes usually available, a faster convergence rate is only cold comfort.

Once the estimate of β is obtained, it can be used for computing the "disequilibrium" series z_t and plug it into the VECM equation (6.4). Believe it or not, the ML estimates of all the remaining parameters are obtained by simply using OLS. The entire bundle of resulting estimators is consistent (at the rate $\sqrt{T}$), asymptotically efficient, and allows, in many cases, for hypothesis testing via the the standard χ^2 distribution.

Bear in mind that software packages usually take care of the whole procedure automatically, so that once

1. the VAR order,
2. the deterministic component,
3. the cointegration rank

have been decided on, the estimates all come out in one go. That said, doing everything "by hand" is in my opinion rather instructive.

Testing restrictions on the cointegration matrices

Sometimes, one may want to estimate the cointegration matrix by imposing more restrictions greater than those strictly necessary to achieve identification. Naturally, this leads to having a restricted and an unrestricted estimate, so that the two can be compared on the basis of an appropriate test statistic.

There are many hypotheses that can be tested, but the ones most commonly of interest are setting elements of β to fixed values, often 0 or ± 1. Using again example 6.2.1, suppose that unrestricted estimation of a bivariate VAR containing log GDP (y_t) and log money supply (m_t) yielded the following estimate of the cointegration vector:[10]

$$\hat{\beta}' = [1, -0.3]'$$

[10] For the sake of the example, let's that we're actually certain that the cointegration rank is equal to 1.

This result would imply a long-run equilibrium relationship like

$$y_t = 0.3m_t + z_t,$$

which is *not* the one predicted by economic theory. Of course, one could argue that 0.3 is just an estimate: what theory predicts is that the *true* value of β is

$$\hat{\beta}' = [1, -1]'.$$

Of course, hypotheses of this kind can be tested via a likelihood-ratio test. Finding the maximum of the constrained log-likelihood may be more difficult in some cases, because the solution may involve more elaborate procedures than simply computing eigenvectors, but modern algorithm are quite good at this, so the usual LR machine can be employed.

In the light of all the possible inferential complications one may encounter with unit root processes, the fact that these tests have a placid χ^2 asymptotic distribution should come as a relief.

Finally, I'll just mention the possibility of running tests known, in this literature, as **exogeneity** tests. These amount to setting to 0 entire rows of the matrix α, so that the corresponding variables are not affected by the disequilibria terms z_t. These tests are also χ^2 distributed, and I'll leave more a extensive description to the literature, as usual.

A simple thumb rule is: hypothesis tests that involve β may have non-standard distributions, but if you're testing hypotheses on other parameters, you should be OK with standard inference methods.

A notable consequence of this state of affairs is that testing for Granger-causality (which do involve β) tests in cointegrated systems is a non-trivial matter, and it took a few years before satisfactory solutions were found.

6.6 Alternative approaches

If the matrix β were known, there would be no special inferential problems: all the time series appearing in a model of the form

$$\Delta y_t = d_t + \alpha z_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t \tag{6.16}$$

are stationary, so OLS can be used with no strings attached as an estimator with all the standard desirable properties.

In practice, β is often unknown, and therefore the time series z_t is unobservable. However, we can consider obtaining an estimate of the cointegration matrix and using it as if it were the true value of β. This is exactly what we do in the Johansen procedure, where β is estimated super-consistently by maximum likelihood. The fact that $\hat{\beta}$ in the Johansen procedure converges at rate T, rather than $\sqrt{T}$, is very important, because the asymptotic distributions for the estimators of all remaining parameters are the same you would get as if the true β was known.

There are, however, alternative ways to estimate β, which can sometimes be convenient. Since we are moving within the field of non-stationarity, one might conjecture that this is a complicated business. For once, however, the inferential peculiarities of $I(1)$ processes come to our rescue. Indeed, it can be shown that the matrix $\tilde{\beta}_2$ can be estimated using that very same static regression that we criticised so heavily when discussing spurious regression.

Let us consider equation (6.15). Naturally, $\tilde{\beta}$ is a cointegration matrix, and therefore $\tilde{\beta}' y_t$ is stationary. This implies that we can write

$$\tilde{\beta}' y_t = y_{1t} - \tilde{\beta}_2' y_{2t} = z_t$$

and therefore

$$y_{1t} = \tilde{\beta}_2' y_{2t} + z_t,$$

where y_{1t} contains the first r elements of y_t and y_{2t} contains the remaining $n - r$ ones; as usual, z_t is some $I(0)$ process. It can be shown that OLS regression of y_{1t} on y_{2t} yields, too, a super-consistent estimator of $\tilde{\beta}_2$.

Furthermore, its consistency is not jeopardised — as in the stationary case — by a potential 'simultaneous equations' effect. Finally, it should be noted that the distribution of this estimator is not asymptotically normal, meaning that hypothesis testing cannot be performed on the elements of the cointegration matrix estimated in this manner; this is not a problem, however, as long as a point estimate is what we need.

At this point, the reader feel a little trouble: in Section 4.5, I suggested that running OLS on integrated variables can produce rubbish? Why does it work now?

A precise answer requires a significant deployment of asymptotic theory. The informal style of these lecture notes, however, allows me to provide a heuristic argument that will, hopefully, convince most readers: consider the case where y_t and x_t are cointegrated, and therefore $z_t = y_t - \beta x_t$ is $I(0)$. Now consider the function

$$S(\theta) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \theta x_t)^2.$$

Evaluating this function at $\theta = \beta$ yields a statistic that converges in probability to a finite number (the variance of z_t). Conversely, choosing any other $\theta \neq \beta$, the difference $(y_t - \theta x_t)$ becomes $I(1)$, with the consequence that the second moment does not exist and the statistic diverges. Since the purpose of OLS is to choose θ such that $S(\theta)$ is as small as possible, it is clear that OLS will tend to choose a value of θ close to β.

This reasoning can also be exemplified graphically: compare panel (b) of figures 6.2 (p. 125) and 6.3 (p. 125). Given that the OLS statistic inherently tends to "go through the points", it is visually apparent that while in the first case something dubious is occurring, in the second case it is perfectly reasonable that OLS tends to reproduce what we now know to be the attractor.

These considerations suggest that β can first be estimated via regression in levels, and then the short-run dynamics can be modelled by replacing z_{t-1} in equation (6.16) with the residuals (lagged by one period) from the regression in levels. It can be shown that using $\hat{\beta}$ instead of the true cointegration vector β makes no difference (asymptotically speaking). If one then wishes to check that the $\hat{z}_t$ are indeed $I(0)$, an ADF test can be used. Beware, however, that in this case the tables of critical values are slightly different from those for the standard ADF test; fortunately, these have also been computed by means of simulations.

This approach, known as the **Engle-Granger** approach, is extremely simple and does not require computing resources other than those for OLS regression. Essentially, two problems arise:

1. the estimators of the parameters β are not efficient;

2. it is not always possible to perform standard hypothesis tests on the coefficients estimated in this manner, as distributions other than χ^2 may emerge.

Problems of this nature can be overcome by using the Johansen procedure or alternative procedures; these, in essence, introduce some corrections to the first stage of the Engle-Granger procedure. In practice, modified regressions are used to estimate the cointegration vector (or vectors) such that the estimator is efficient and hypothesis tests can be conducted on the elements of β using standard distributions.

The most ingenious ideas, among the ones in common usage are the so-called **Fully Modified OLS** estimator (or FM-OLS for brevity) by Phillips and Hansen, and the **Dynamic OLS** (or DOLS), invented independently by Pentti Saikkonen and by the Stock & Watson duo. For the usual reasons of brevity, however, I do not illustrate them here and instead provide the customary reference to the literature.

6.7 An example

Let us consider an example: Figure 6.4 depicts the monthly spot price of gasoline (g_t), Diesel (d_t) and crude oil (w_t, West Texas Intermediate, or WTI for short). Data go from January 2007 to February 2022. Of course, economic theory suggests that the three series should "go together"; more precisely, one would expect that the long-run relationships between the variables in logs should look like

$$\begin{aligned} g_t &= \beta_0 + \beta_1 w_t \\ d_t &= \gamma_0 + \gamma_1 w_t, \end{aligned}$$

where β_1 and γ_1 are the two elasticities of the two types of fuel to oil. These two numbers should be positive, perhaps equal to 1, but not necessarily. If the series are $I(1)$, there should be a fair chance that they cointegrate. Therefore, we would expect a cointegration rank equal to 2. In terms of the Stock-Watson representation, we would say that the system has one common $I(1)$ trend ($n - r = 3 - 2 = 1$), driving the whole system. Again, this fully agrees with economic intuition. The intercepts β_0 and γ_0 may be non-zero because the units are different (US$ per gallon for g_t and d_t, US$ per barrel for w_t).

Even a cursory glance reveals that all series look very persistent, and have indeed very similar paths. Moreover, the two scatterplots of g_t and d_t versus w_t (Figure 6.5) is very suggestive: the point align along a line, that looks very much like the attractor in Figure 6.3. The two least-squares lines (that should

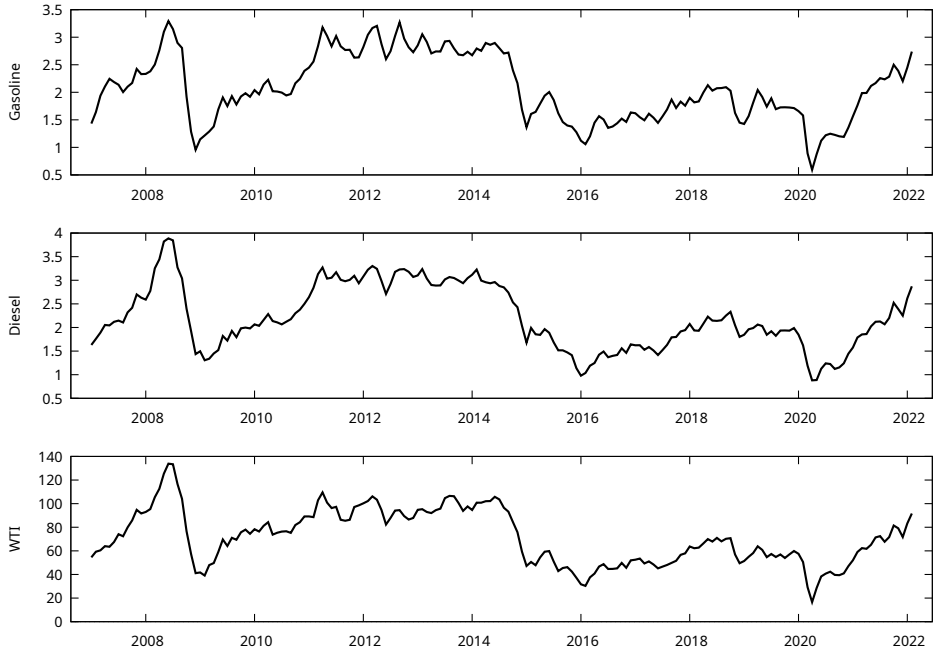


Figure 6.4: Oil product spot prices

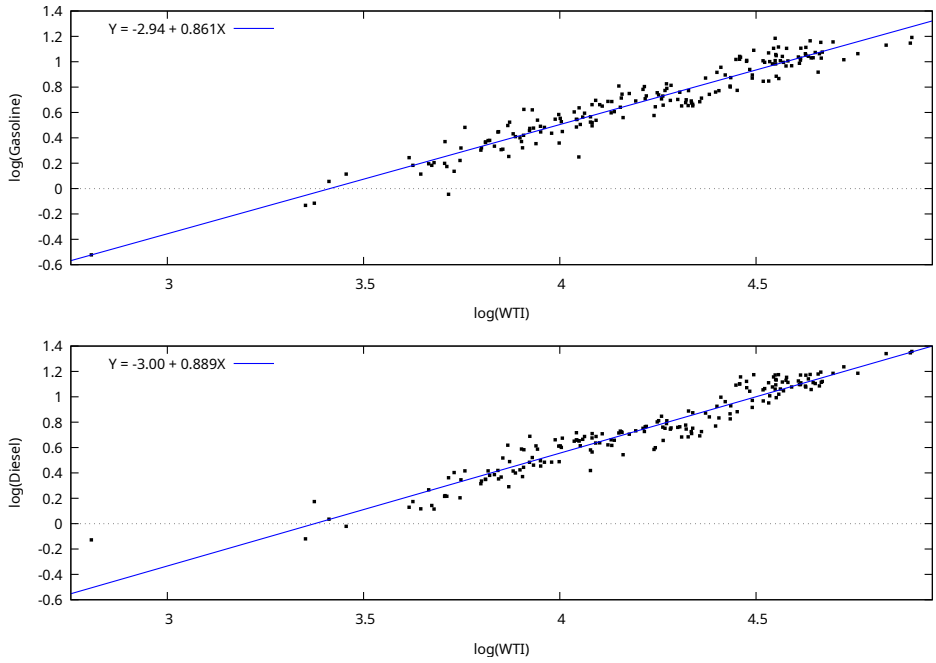


Figure 6.5: Oil product spot prices — scatterplots

give you a super-consistent estimate of the two elasticities) display values fully consistent with what would have expected (both slightly below 1).

An additional remark worth doing here is that the three series show no appreciable trend. Therefore, the best candidate for representing is a VECM with a restricted intercept (case 2 in the taxonomy illustrated in Section 6.5.1):

$$\Delta y_t = \alpha z_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t$$

where we can find p via the usual information criteria on the VAR in levels and

$$z_{t-1} = \begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} = \beta' y_{t-1} = \begin{bmatrix} 1 & 0 & -\beta_1 & -\beta_0 \\ 0 & 1 & -\gamma_1 & -\gamma_0 \end{bmatrix} \begin{bmatrix} g_{t-1} \\ d_{t-1} \\ w_{t-1} \\ 1 \end{bmatrix} \tag{6.17}$$

After choosing $p = 3$ (I'll omit the details as to why), one finds that the rank tests give the following table:

Rank	Eigenvalue	Trace	p-value	λ-max	p-value
0	0.2026	61.021	0.0000	41.198	0.0000
1	0.0649	19.824	0.0559	12.221	0.1788
2	0.0409	7.603	0.0999	7.603	0.0998

With the proviso that I'll focus on the trace test (the λ-max test is much more seldom used), here's how you read it: you start from the bottom and ask yourself: is the smallest eigenvalue equal to 0? If this hypothesis was rejected, the rank of Π would be 3 and, as we know, that would imply that there's no cointegration because the series are in fact $I(0)$. However, the smallest eigenvalue is 0.0409, which translates into a p-value for the trace test if 9.99%, so $H_0 : \lambda_2 = 0$ cannot be rejected. OK. We have at least one unit root here, and the rank of Π is at most 2.

Next, we proceed further up. For the joint hypothesis $H_0 : \lambda_1 = \lambda_2 = 0$ we have a p-value of 5.59%. A blind follower of the p-value religion would now say "OK, the rank is at most 1, let's go ahead". I object. Economic theory and statistical common sense suggest that we're close enough to the 5% threshold to label what happens here as one of the cases when the test lacks power and the hypothesis is in fact false, even though the p-value is (marginally) above the usual threshold. Therefore, I decide that $r = 2$.

On the basis of this (rather dictatorial) decision, let's have a look at the two eigenvectors:

$$\tilde{\beta} = \begin{bmatrix} -11.713 & 10.531 \\ 14.781 & 2.3074 \\ -4.6740 & -12.257 \\ 16.709 & 42.618 \end{bmatrix}$$

these two columns are cointegration vectors, but are very hard to interpret.

However, we can use the Phillips normalisation to yield

$$\hat{\beta} = \begin{bmatrix} -11.713 & 10.531 \\ 14.781 & 2.3074 \\ -4.6740 & -12.257 \\ 16.709 & 42.618 \end{bmatrix} \begin{bmatrix} -11.713 & 10.531 \\ 14.781 & 2.3074 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -0.93266 & -1.0553 \\ 3.2372 & 3.6957 \end{bmatrix}$$

which is exactly in the same format as equation (6.17), so for example our estimate of the long-run elasticity of gasoline to oil would be 0.93266.

Since both elasticities have the "right" sign and reasonable magnitude (they are both close to 1), the idea of testing if they're both equal to one is rather tempting. The constrained estimate of β turns out to be

$$\tilde{\beta} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ 3.5124 & 3.4579 \end{bmatrix},$$

with a corresponding value of the log-likelihood of 798.52578. Since the log-likelihood for the unrestricted model was 801.83601, the LR test is easily computed as

$$LR = 2(801.83601 - 798.52578) = 6.62046.$$

With two constraints, this number needs to be matched against the χ^2_2 distribution, which yields a p-value of 3.65%, only slightly below the 5% threshold. So, should this hypothesis be rejected? A p-value monkey would say "yes"; someone willing to believe in the complete pass-through hypothesis on the grounds of economic theory would label this result as a type I error and say "no". I'll let readers make up their own minds.

6.A Assorted results

6.A.1 Outline of the proof of the Granger Representation Theorem

Let us consider an n-dimensional $I(1)$ stochastic process y_t. Since Δy_t is stationary by hypothesis, it must have a Wold representation

$$\Delta y_t = C(L)\epsilon_t. \tag{6.18}$$

We apply the Beveridge-Nelson decomposition to y_t and obtain

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t. \tag{6.19}$$

This decomposition exists for any multivariate $I(1)$ process; if, however, the system is cointegrated with cointegration rank r, there must exist an $n \times r$ matrix β such that $\beta' y_t$ is stationary. Pre-multiplying equation (6.19) by β', it is immediately apparent that the only condition under which this is possible is that

$$\beta' C(1) = 0.$$

Since $C(1)$ is an $n \times n$ matrix, this can only occur if $C(1)$ is singular; more specifically, its rank must not be greater than $n - r$. In our case, the condition $\beta' C(1) = 0$ is ensured only if $C(1)$ "begins with" $\beta_\perp$, and therefore $C(1)$ can be written as

$$C(1) = \beta_\perp \Lambda'. \tag{6.20}$$

The matrix Λ is something we still know nothing about, except that it is an $n \times (n - r)$ matrix, but we will get to that soon.

Suppose now that a VAR representation of y_t exists, and that it is therefore possible to write $A(L)y_t = \epsilon_t$; our objective now is to prove that $A(1)$ can be written as $\alpha\beta'$, and thus the ECM representation holds. To show this, consider equations (6.6) and (6.7); it is clear that, if both the VAR and VMA representations are valid, it must hold that

$$A(z)C(z) = C(z)A(z) = I \cdot (1 - z). \tag{6.21}$$

By setting $z = 1$ in equation (6.21), it is easy to deduce that

$$A(1)C(1) = C(1)A(1) = 0. \tag{6.22}$$

By combining equations (6.20) and (6.22), we can state about $A(1)$ that, if $C(1)$ "begins with" $\beta_\perp$, then $A(1)$ must "end with" β'; by the same reasoning, it should be equally clear that $A(1)$ must "begin with" $\Lambda_\perp$. In summary, we can say that $A(1)$ must be of the form

$$A(1) = \Lambda_\perp K \beta',$$

where K is some invertible $r \times r$ matrix. And that is it. By defining $\alpha = \Lambda_\perp K$, $A(1)$ can be written as $\alpha\beta'$, as required.

The last remaining unresolved matter is this matrix Λ about which we know nothing. We can try to see how it relates to α. We know that $\alpha = \Lambda_\perp K$, and therefore $\Lambda'\alpha = 0$; it follows that $C(1)$ must be of the form

$$C(1) = \beta_\perp H \alpha'_\perp, \tag{6.23}$$

where H is an $(n - r) \times (n - r)$ matrix whose elements are functions of the parameters of the autoregressive representation.

Done. Although it is not obvious at first glance, these steps have produced a very important result: the fact that the matrix $C(1)$ can be mapped back to the parameters of the ECM, based solely on the fact that $C(1)$ must be singular, ensures that the ECM representation of a cointegrated system *always* exists.

The matrix H appearing in equation (6.23) does not have an immediate interpretation, but, for those who enjoy this sort of thing, let us see how it is structured: if we differentiate equation (6.21) with respect to z, we obtain

$$A(z)\dot{C}(z) + \dot{A}(z)C(z) = -I. \tag{6.24}$$

Note that we are not doing anything particularly esoteric here, but simply differentiating a polynomial:

$$\dot{A}(z) = A_1 + 2A_2 z + 3A_3 z^2 + \ldots + pA_p z^{p-1} = \sum_{i=1}^{p} iA_i z^i$$

and an analogous expression holds for $\dot{C}(z)$. Equation (6.24), evaluated at $z = 1$, yields

$$\alpha\beta'\dot{C}(1) + \dot{A}(1)\beta_\perp H\alpha'_\perp = -I, \tag{6.25}$$

where, naturally, $\dot{A}(1) = \sum_{i=1}^{p} iA_i$, which we assume to be of full rank. The matrix in the previous expression can be pre-multiplied by $\alpha'_\perp$ and post-multiplied by $\dot{A}(1)\beta_\perp$, yielding

$$\left[\alpha'_\perp \dot{A}(1)\beta_\perp\right] H \left[\alpha'_\perp \dot{A}(1)\beta_\perp\right] = -\left[\alpha'_\perp \dot{A}(1)\beta_\perp\right].$$

If the matrix to the right of the equals sign is invertible, we evidently have $H = -\left[\alpha'_\perp \dot{A}(1)\beta_\perp\right]^{-1}$.

I will only mention that this inverse exists if the system is indeed $I(1)$. In processes with a higher order of integration, the matrix is not invertible and the theorem can be generalised, but everything becomes a little esoteric.

6.A.2 The Frisch-Waugh theorem

The Frisch-Waugh theorem is ubiquitous in econometrics. A full proof would need a hefty investment in matrix algebra and the definition of projection operators, but the gist of the theorem is rather simple: suppose that you want to estimate to the following equation:

$$y_t = x_t\beta + z_t\gamma + \epsilon_t$$

but you don't really care for the parameter γ; all you're interested in is β. Of course, the easiest way is to apply OLS to the model above and then just use β and ignore γ.

In some cases, however, it can be useful to have a way to recover the value of β you would get from the procedure above without estimating the full model. This is how you do it:

1. First you run an OLS regression of y_t on z_t alone and save the residuals as u_t^y;
2. then you run an OLS regression of x_t on x_t alone and save the residuals as u_t^x;
3. finally, you regress u_t^y on u_t^x; the coefficient you get is the same β you would have obtained by running OLS on the full model.

Chapter 7

Persistent volatility

In the preceding chapters, we addressed persistence in time series from a somewhat reductive point of view. In fact, we defined a process as persistent when

$$f(x_t) \neq f\left(x_t \mid \mathcal{F}_{t-1}\right), \tag{7.1}$$

which implies the non-independence of a single element of the process, x_t, from its past history. Later, however, we did not actually address persistence according to the definition just given, but we focused on rather a specific case, namely

$$E(x_t) \neq E\left(x_t \mid \mathcal{F}_{t-1}\right). \tag{7.2}$$

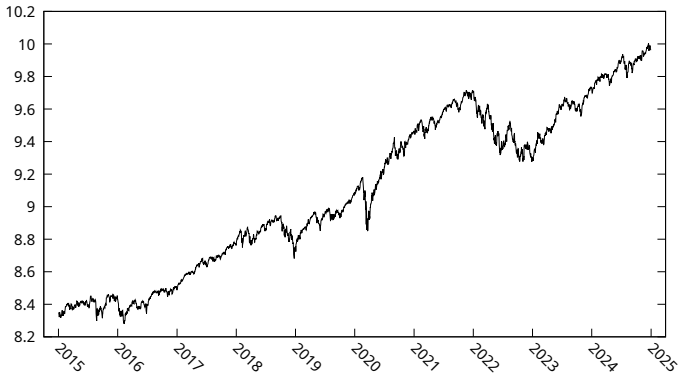
In doing so, we implicitly assumed that, once persistence in the first moments was accounted for, everything interesting was satisfactorily dealt with. In an ARMA model, the conditional mean of y_t does depend on $\mathcal{F}_{t-1}$, but all the remaining distributional properties are independent of time: for instance, if we compare the marginal and conditional variance of an AR(1) process, they are indeed different ($\frac{\sigma^2}{1-\varphi^2} \neq \sigma^2$), yet their difference does not depend in any way on $\mathcal{F}_{t-1}$.

In general, however, equation (7.2) is just a special case of (7.1). There may be persistence features worthy of attention that have nothing to do with first moments, but rather with other distributional characteristics. In this chapter, we will discuss processes where persistence is mainly a second-moment phenomenon, and we have good reasons for doing so: these processes are very good at replicating certain features that are typical of financial time series.

7.1 Stylised facts

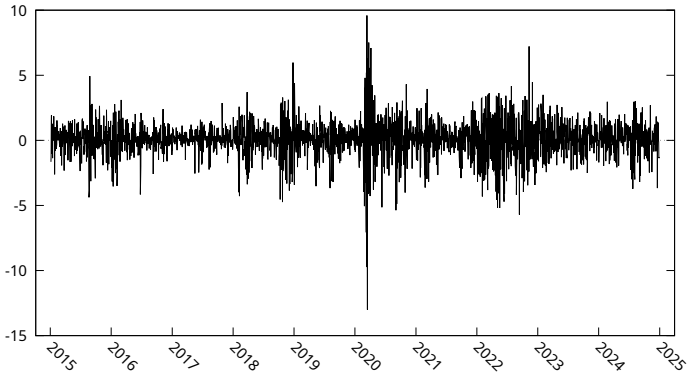
In Figure 7.1, you see the Nasdaq stock index (in logs), recorded daily between 2015 and 2024. As is evident from the plot, and confirmed by any unit root test you can think of, considering this series as a realisation of an $I(0)$ process is untenable. This result that could be at least partly anticipated: in an efficient market, the return on a financial asset must be a martingale difference process, or at least closely resemble one.

Figure 7.1: Nasdaq composite index – logarithms



In fact, there are excellent theoretical reasons why the price of a financial asset may not necessarily be a martingale difference process. Apart from the obvious observation that markets are inevitably imperfect, it must also be noted that this outcome derives from various debatable assumptions.
These include the absence of information asymmetries, the existence of a "representative agent" who acts as a sort of average of all other market participants, and the assumption that this representative agent is risk-neutral rather than risk-averse, and some more. This claim is not by a rabid commie, as some would say; it is by the prince of conservative economists himself, the Nobel laureate Robert Lucas.
Nevertheless, no one has yet discovered a mechanical rule capable of predicting returns (although, if anyone had found one, they would hardly share it with you and me). It must be said, however, that with *very* high frequency data, such as tick-by-tick data, where each and every transaction is recorded separately, these imperfections become very important. There's an enormous amount of literature on "microstructure".

Figure 7.2: Nasdaq composite index – daily returns

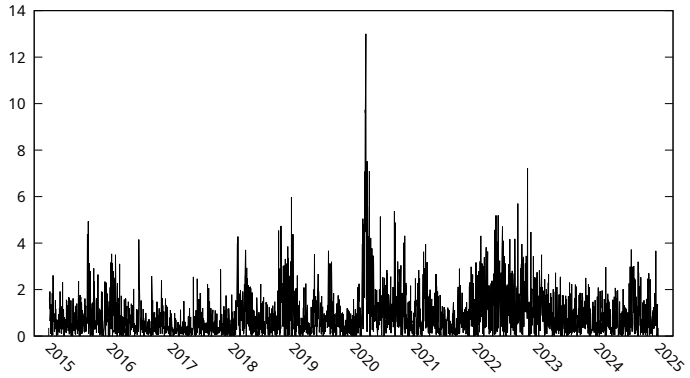


By taking differences, we obtain the returns, shown in Figure 7.2 (exactly the same as Figure 1.5 on page 10). As expected, the returns do not display significant correlations (see page 10, again). Notice, however, that the amplitude of the fluctuations varies over time, giving rise to something labelled

"volatility clustering", where periods of low volatility alternate with periods when volatility is higher. This simple fact suggests that, although there is no persistence in the mean, there might well be in the variance, or more generally in volatility.

Volatility is very important in applied finance, because it is tightly linked to risk: if persistence can be used for making forecasts on volatility, that would be extremely useful in any situation when you have to deal with financial risk.

Figure 7.3: Nasdaq composite index – absolute returns



Of course, we have a very precise word in econometrics to describe sequences of random variables with different variances: heteroskedasticity. The object of this chapter is precisely to illustrate models that make heteroskedasticity a consequence of a (very special kind of) persistence.

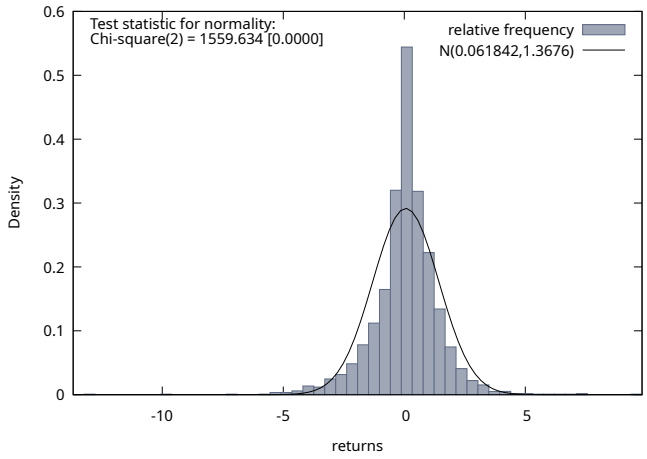
The fact that volatility is indeed persistent can be clearly seen in Figure 7.3 (also presented earlier in Chapter 1), that reports the absolute value of returns — a proxy for volatility. This series is subject to a certain degree of persistence: its autocorrelations are reported in Table 7.1, and they are all statistically significant. This means not only that we have heteroskedasticity, but also (most importantly) that heteroskedasticity seems to follow some time pattern.

Table 7.1: Autocorrelations of absolute returns

Lag	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}$	0.284	0.309	0.289	0.283	0.265	0.280	0.236	0.231	0.234	0.220

Another highly common feature observed in high-frequency financial time series has to do with the shape of their marginal distribution: in an ARMA process of the form $A(L)y_t = C(L)\epsilon_t$, it is easily shown that if the joint distribution of the ϵ_t is a multivariate normal, then the distribution of y_t is likewise normal. The marginal distribution of the Nasdaq index returns time series reveals, by contrast, clear evidence of non-normality (see Figure 7.4), driven primarily by excess kurtosis: the kurtosis of a random variable is its third

Figure 7.4: Nasdaq returns – marginal distribution



moment (centred and standardised). If x is a zero-mean variable,

$$\kappa_x = \frac{E(x^4)}{E(x^2)^2}.$$

If x is normally distributed, is kurtosis is 3. Since the normal distribution serves as a benchmark, excess kurtosis is simply kurtosis minus 3. A distribution with a kurtosis greater than 3 is said to be **leptokurtic** and exhibits "fat tails": even after standardising the distribution to have unit variance, the probability of extreme events (or "outliers", if you prefer) is higher than the one implied by the normal distribution.

In the case of the Nasdaq returns, the sample excess kurtosis is 7.524; Figure 7.4 also reports a normality test (the Doornik-Hansen test) equal to 1559.63; this number should come from a χ^2_2 density under the null, so these findings suggests, at the very least, that if we wished to fit an ARMA model to the observed series, then Gaussianity would be a rather bold assumption; we would rather need to select another type of distribution, preferably with thicker tails. In fact, as we shall see, this feature goes hand in hand with the previous one, since high kurtosis may be, at least in part, a by-product of heteroskedasticity.

7.2 ARCH and GARCH processes

The class of GARCH stochastic processes has enjoyed immense popularity because it manages to replicate the characteristics outlined in the previous section in a simple, natural and elegant way by focusing on **conditional heteroskedasticity**: that is, the variability in the conditional variance of the process, where the conditioning is on the past information set:

$$h_t = V(y_t|\mathcal{F}_{t-1}).$$

Just as ARMA models are a way to describe persistence in the conditional mean, GARCH process do the same with the conditional variance. We begin ARCH processes, a special case that started the whole subsequent literature.

7.2.1 ARCH processes

Generally speaking, a zero-mean conditionally heteroskedastic process (which for the moment we assume to be observable) can always be written as:

$$\epsilon_t = u_t \sqrt{h_t}, \tag{7.3}$$

where u_t is a zero-mean process with unit variance (possibly a persistent process, although we'll assume it's not, for the moment) and h_t is a sequence, which can be deterministic or, more interestingly, can be a stochastic process itself, which we shall always assume to be independent of u_t. From the definition of variance, it clearly follows that $\mathrm{Var}(\epsilon_t) = E(h_t u_t^2)$; since u_t and h_t are independent, the expectation of the product is the product of the expected values, and thus $\mathrm{Var}(\epsilon_t) = E(h_t)E(u_t^2) = E(h_t)$. Note that the variance is constant (and therefore the process is homoskedastic) only if $E(h_t)$ is a constant sequence. Otherwise, we have heteroskedasticity.

Even at this stage, it is interesting to note that the kurtosis of ϵ_t is greater than that of u_t, simply as a consequence of heteroskedasticity. The proof is essentially a straightforward application of Jensen's inequality:[1] the kurtosis coefficient of ϵ_t can be written as

$$\kappa_\epsilon = \frac{E(\epsilon_t^4)}{E(\epsilon_t^2)^2} = \frac{E(h_t^2 u_t^4)}{E(h_t u_t^2)^2},$$

but since u_t and h_t are independent, it follows that

$$\kappa_\epsilon = \frac{E(h_t^2)E(u_t^4)}{E(h_t)^2 E(u_t^2)^2} = \frac{E(h_t^2)}{E(h_t)^2}\kappa_u;$$

since (by Jensen's inequality) $E(h_t^2) > E(h_t)^2$ for any non-degenerate distribution, we deduce that $\kappa_\epsilon > \kappa_u$.

In a conditionally heteroskedastic process of this type, therefore, if u_t is normal (and thus its kurtosis is 3), ϵ_t will certainly be leptokurtic. Note that we have just found a link between two of the main stylised facts discussed earlier.

Clearly, statistical modelling of processes of this type requires the specification of a functional form for h_t, and of course, the simpler this form is, the better. ARCH processes arise when h_t is a linear function of past squared values of ϵ_t, namely:

$$h_t = c + \sum_{i=1}^{p} a_i \epsilon_{t-i}^2,$$

[1] For those in need: Jensen's inequality states that if $g(\cdot)$ is a convex function, then $E[g(X)] \geq g[E(X)]$. For example, $E(X^2)$ is always greater than or equal to $E(X)^2$ (which is why the variance is always non-negative). It is immediate to show that if the function is concave rather than convex, the direction of the inequality is reversed.

where h_t is a deterministic function of variables belonging to $\mathcal{F}_{t-1}$ (note that the summation starts from 1, not from 0). Consequently,

$$V(\epsilon_t|\mathcal{F}_{t-1}) = E(h_t|\mathcal{F}_{t-1}) = h_t.$$

An ARCH(p) model, therefore, describes a process ϵ_t (which, for simplicity, at the moment we assume to be uncorrelated and with zero mean) where the conditional variance is given by:

$$h_t = \text{Var}(\epsilon_t \mid \mathcal{F}_{t-1}) = c + A(L)\epsilon_{t-1}^2, \tag{7.4}$$

where $A(L)$ is a lag polynomial of order $q-1$ (note the "-1").

To analyse the properties of the process, note that equation (7.4) implies that the *squares* of the observed series are just an AR(q) process: define the difference between ϵ_t^2 and its conditional mean as

$$\eta_t = \epsilon_t^2 - h_t. \tag{7.5}$$

Clearly, this definition implies that η_t is a martingale difference sequence:

$$E\left[\eta_t|\mathcal{F}_{t-1}\right] = E\left[\epsilon_{t-1}^2|\mathcal{F}_{t-1}\right] - E\left[h_t|\mathcal{F}_{t-1}\right] = h_t - h_t = 0.$$

If we were willing to assume that the second moments of η_t exist, then all its autocorrelations would be 0, so we could even say that η_t is white noise.[2]

From the definition of η_t, it obviously follows that[3]

$$\epsilon_t^2 = c + A(L)\epsilon_{t-1}^2 + \eta_t \Longrightarrow [1 - A(L)L]\epsilon_t^2 = c + \eta_t \tag{7.6}$$

so ϵ_t^2 is an AR(p) process. This is why the acronym ARCH (AutoRegressive Conditional Heteroskedasticity) is used: what we have here is a conditionally heteroskedastic process where heteroskedasticity is described by an autoregressive mechanism.

If the polynomial $[1 - A(L)L]$ has no unit or explosive roots, then ϵ_t^2 has a finite unconditional expected value equal to

$$E(\epsilon_t^2) = \frac{c}{1 - A(1)} = E(h_t).$$

Since $E(\epsilon_t) = 0$ by assumption, the expected value of its square is also its variance. In high-volatility periods, we would expect ϵ_t to be large in absolute value, and so its square. But then, the autoregressive mechanism implied by (7.6) implies that ϵ_{t+1}^2 will be likely to be a large number too, so volatility will be persistent.

Note that, unlike the ARMA case, special parameter constraints are needed make an ARCH model reasonable: all the coefficients in the polynomial $A(L)$ in equation (7.4) must be positive. Otherwise, an event could potentially occur that drives the conditional variance to be negative (which, you will admit, would be rather embarrassing). Moreover, if we require the process to also possess a marginal variance, we must rule out unit roots as well somehow.

[2] Of course, it would be a rather peculiar white noise process: for instance, while it is true that its conditional mean is zero, this does not imply that η_t is *independent* of $\mathcal{F}_{t-1}$. For example, its support (the set of values over which its density is defined) is obviously bounded from below, given that for $\eta_t < -h_t$ one would obtain a negative squared value (?) in equation (7.5).

[3] I hope readers will not find it insulting if I point out that $A(L)\epsilon_{t-1} = A(L)L\epsilon_t$.

Example 7.2.1 (ARCH(1)) *An example of an ARCH(1) process is given by ϵ_t, where*

$$h_t = V(\epsilon_t \mid \mathcal{F}_{t-1}) = 0.1 + 0.8\epsilon_{t-1}^2.$$

The conditional variance of the process is therefore time-varying. However, in this case, we can write

$$[1 - 0.8L]\epsilon_t^2 = 0.1 + \eta_t$$

and it is easily seen that its unconditional variance of ϵ_t is constant and equal to

$$E\left[\epsilon_{t-1}^2\right] = E(h_t) = V(\epsilon_t) = \frac{0.1}{1 - 0.8} = 0.5.$$

In conclusion: ϵ_t is a process that features a finite and constant variance if we consider its marginal distribution; however, the variance *conditional on the information set* $\mathcal{F}_{t-1}$ is none other than h_t, which varies over t. This is precisely why we speak of conditional heteroskedasticity.

7.2.2 Testing for volatility persistence

In the context of ARMA processes, the simple structure of the law of motion for the conditional variance makes it possible to devise a natural heteroskedasticity test, starting from the widely used Breusch-Pagan test. In the Breusch-Pagan test, it is assumed that h_t can be represented as $h_t = H(z_t'\gamma)$ where $H(\cdot)$ is some differentiable function (no need to specify a precise form). It can be shown, through a rather ingenious procedure, that all you need is an auxiliary regression of the form

$$e_t^2 = \phi_0 + z_t'\phi_1 + \text{residuals}.$$

The test statistic, based on the LM principle, is just the R^2 index, multiplied by the number of observations. Under appropriate hypotheses, this statistic follows an asymptotic χ^2 distribution with degrees of freedom equal to the number of elements in ϕ_1. In our case, the function $H(\cdot)$ is given by equation (7.4); consequently, to test for ARCH effects, it is sufficient to run an OLS regression of the squared residuals from a conditional mean model on their own lags. This test is precisely known as the **LM-ARCH test**.

An ARCH process, therefore, has several characteristics that make it highly appealing for modelling financial time series. However, the persistence structure of the volatility is such that, almost always, polynomials of very high order are needed to replicate it. Consider Table 7.1: it is evident that the autocorrelations are all "small"; nevertheless, a low-order autoregressive model is not very suitable, since the autocorrelation structure is highly persistent.

Using many parameter is never nice, but in the present context it may become problematic on account of the parameter constraints I described earlier, so there is a possibility that the order of the polynomial $A(\cdot)$ may be too high to be manageable in practice. By applying the same logic that led us to the specification of ARMA models, we could alleviate (if not outright eliminate) the problem by introducing a lag polynomial on the right-hand side of the equal sign in (7.6) as well: doing so gives us GARCH processes.

7.2.3 GARCH processes

GARCH processes are a generalisation of ARCH processes (GARCH stands for *Generalised ARCH*), much in the same way that ARMA processes are a generalisation of AR models: in GARCH processes, the conditional variance depends not only on past values of ϵ_t^2 (as in ARCH models) but also on its own past values:

$$h_t = c + A(L)\epsilon_{t-1}^2 + B(L)h_{t-1}, \tag{7.7}$$

where the polynomials $A(\cdot)$ and $B(\cdot)$ are of order $q-1$ and $p-1$, respectively.

Using the same logic applied in the preceding subsection, we can show that a GARCH(p,q) process can also be written as an ARMA process. Start again from the definition of η_t:

$$\eta_t = \epsilon_t^2 - h_t;$$

by combining the definition above with (7.7), you get

$$\epsilon_t^2 = c + A(L)\epsilon_{t-1}^2 + B(L)\left[\epsilon_{t-1}^2 - \eta_{t-1}\right] + \eta_t,$$

which, in turn, yields

$$\{1 - [A(L) + B(L)]\,L\}\,\epsilon_t^2 = c + (1 - B(L)L)\eta_t \tag{7.8}$$

and therefore ϵ_t^2 is an ARMA$(\max(p,q),p)$ process.[4]

Interestingly, if the $[1 - B(L)L]$ polynomial is invertible, a GARCH process can be re-expressed using the Wold representation for ϵ_t^2:

$$\epsilon_t^2 = \frac{c}{1 - B(1)} + \frac{A(L)}{1 - B(L)L}\epsilon_{t-1}^2 + \eta_t = \mu + C(L)\epsilon_{t-1}^2 + \eta_t,$$

which is a MA(∞) process, equivalent to an ARCH(∞) specification.

Example 7.2.2 (GARCH(1,1)) *Let us consider an example with a GARCH(1,1) model. Suppose that:*

$$h_t = 0.5 + 0.3\epsilon_{t-1}^2 + 0.6h_{t-1}.$$

Repeating the earlier reasoning, we can write:

$$\epsilon_t^2 = 0.5 + 0.3\epsilon_{t-1}^2 + 0.6\left[\epsilon_{t-1}^2 - \eta_{t-1}\right] + \eta_t,$$

which yields

$$[1 - 0.9L]\,\epsilon_t^2 = 0.5 + (1 - 0.6L)\eta_t,$$

namely, an ARMA(1,1) process. In this case as well, the unconditional variance is constant:

$$E\left[\epsilon_{t-1}^2\right] = E(h_t) = V(\epsilon_t) = \frac{0.5}{1 - 0.9} = 5.$$

[4] An algebraic detail: $B(L)$ can be equal to 0, which would yield a simple ARCH process, but if $A(L)$ were equal to 0, the ARMA representation in equation (7.8) would contain a common factor, meaning that $B(L)$ would not be identified. See section 3.2.

As with ARCH processes, to ensure the model is valid, some parameter restrictions are necessary: the α_i parameters must be strictly positive and the β_i parameters are non-negative.[5]

7.2.4 Estimation

How do we estimate GARCH models? One might consider equation (7.8) and think: if ϵ_t^2 is an ARMA process, what's the problem? Just compute the square of ϵ_t and go ahead.

However, taking this approach raises several issues. To begin with, ϵ_t may be unobservable. In fact, the models we estimate are almost always of the form

$$y_t = \mu_t + \epsilon_t,$$

where there is a conditional mean component (which can naturally include a deterministic core, exogenous variables, lagged values of y_t, etc.) whose parameters are unknown, meaning that ϵ_t is not directly observable.

OK, I hear you say: let's estimate the conditional mean first via standard techniques (such as OLS or similar methods) and then work with the squared residuals. Yes, in principle you could. However, this procedure is somewhat rudimentary, although it can be shown to yield consistent estimators in many cases. Moreover, this estimator would be inefficient. This efficiency loss comes essentially from two aspects: first, we are estimating the parameters sequentially rather than jointly. Second, the one-step-ahead forecast error in equation (7.8), η_t, is evidently non-normal; for instance, as noted earlier, its support is bounded. Furthermore, the variance of η_t is not guaranteed to exist (this can be proven, though it is not essential here), which may seriously compromise the asymptotic properties of the ARMA estimators, which are typically based on a Gaussian likelihood.

Direct application of maximum likelihood is much preferable: once again, sequential factorisation (see Section 3.3) comes to our aid. Assuming that the conditional distribution of y_t is Gaussian:

$$y_t | F_{t-1} \sim N\left[\mu_t, h_t\right],$$

if we take y_0 as fixed, the log-likelihood can be written as

$$L(\theta) = \sum_{t=1}^{T} \ell_t, \tag{7.9}$$

where

$$\ell_t = \log\left[\frac{1}{\sqrt{h_t(\theta)}}\phi\left(\frac{y_t - \mu_t(\theta)}{\sqrt{h_t(\theta)}}\right)\right] = -\frac{1}{2}\left[\log(h_t(\theta)) + \frac{(y_t - \mu_t(\theta))^2}{h_t(\theta)}\right]$$

[5] As elsewhere, I have simplified things a bit. In reality, the model is well-specified only for those parameter values for which the ARCH(∞) representation yields exclusively positive coefficients. In the GARCH(1,1) case, a necessary and sufficient condition is indeed that both coefficients are positive and that their sum is strictly less than 1. For higher-order cases, more intricate conditions, known as the Nelson-Cao conditions, apply.

and θ is a vector containing the parameters of both the conditional mean and conditional variance (as usual, I employ the notation $\phi(\cdot)$ to denote the standard normal density).

The maximisation process is carried out using the standard iterative methods discussed in section 3.1. In these cases, however, convergence is often more elusive than in ARMA models; consequently, convergence failures are uncommon, but not impossible.

In most cases, this is due to the parameter constraints discussed earlier. It can happen that the maximum of the likelihood function lies outside the set of admissible values for the parameters (for example, $\alpha + \beta < 1$ for a GARCH(1,1)). This can occur when the model is misspecified, or when the sample size is too small.[6]

Typically, estimation routines incorporate specific checks to prevent the log-likelihood from being evaluated at parameter values outside the admissible range. Since the iterative methods mentioned in section 3.1 are designed to operate mechanically, one sometimes witnesses the sad spectacle of the algorithm acting like a fly against a window pane: it throws itself against the glass, hits it, bounces back, throws itself again, and so on *ad libitum*. In such cases, there is little to do except try to improve the model specification and invoke a deity of your choice.

Note that, in these cases, the problem stems from the data, not from the software used for estimation. Sometimes, however, the problem is in fact the software.
A few years ago, I saw with my own eyes a well-known commercial package yield a GARCH(1,1) estimate with negative α. That package had the surreptitious habit of presenting the estimates as though convergence had been achieved, even when, in reality, the algorithm had stopped simply because the maximum number of iterations had been reached. Wow.
I don't use that package, so I cannot check, but I am told that in the meantime this debatable behaviour has been fixed. Phew.

7.2.5 An example

Let us take the time series shown in Figure 7.2 and apply a GARCH model, coupled with an AR(1) specification for the conditional mean. We estimate the parameters in the two equations

$$y_t = \mu + \varphi y_{t-1} + \epsilon_t \tag{7.10}$$
$$h_t = c + a\epsilon_{t-1}^2 + bh_{t-1}. \tag{7.11}$$

Of course, we here assume that the observable process is y_t, rather than ϵ_t. The heteroskedastic component lies in the definition of the GARCH process ϵ_t, the "disturbance term" of equation (7.10). A believer in the efficient markets hypothesis could argue that the autoregressive term in the mean equation (7.10) should not be necessary, but you never know, and it does no harm.

Moreover, one may ask: why $p = q = 1$, instead of a more general choice for equation 7.11? First: this is just an example, ok? Second: most practitioners

[6] It should be noted that, in this context, 100 observations still constitute a *very* small sample. Fortunately, in financial applications, data scarcity is almost never an issue.

adopt this same choice routinely. It is quite unusual (although not impossible) to find higher-order specifications in applied practice.

Table 7.2: GARCH (QML) estimates

Parameter	Coefficient	Std. error	t-statistic	p-value
μ	0.1039	0.0212	4.892	9.99E-07
φ	-0.0402	0.0223	-1.801	0.0717
c	0.0478	0.0139	3.429	0.0006
a	0.1346	0.0228	5.917	3.27E-09
b	0.8408	0.0248	33.94	1.98E-252

Let us have a look at the estimates. The constant μ is significantly positive is no surprise: it just means that, on average, daily returns are slightly positive (the unconditional mean of y_t is 0.062%) or, in other words, that series plotted in Figure 7.1 goes up.

As we could have expected, there is very little persistence in the conditional mean of the returns: the φ parameter is not very significant, and besides, it is very small in absolute value. You could read it as: suppose that on day t the Nasdaq index gained 2% (quite a lot). What should we expect for the next day? Our forecast would be $\hat{y}_{T+1} = 0.1039 - 0.0402 \times 2\% = 0.0235\%$. What if the Nasdaq index had gone *down* by 2% (again, quite a lot)? In that case, $\hat{y}_{T+1} = 0.1039 + 0.0402 \times 2\% = 0.1843\%$. As you can see, the difference is quite small and certainly cannot be used for a fire-proof trading strategy. The weak evidence we have for a very moderate negative persistence is just the consequence of the well-known tendency of markets to "rebound" after a particularly bad day. So, in short, there is very little persistence in the conditional mean, if any at all.

On the other hand, there is clear evidence of persistence in the variance: the parameters a and b are both well away from 0, thus indicating a substantial GARCH effect. More to the point: the estimates imply that the ARMA representation for ϵ_t^2 is given by:

$$[1 - (a+b)L]\,\epsilon_t^2 = c + \eta_t - b\eta_{t-1},$$

that is

$$(1 - 0.9754L)\epsilon_t^2 = 0.0478 + \eta_t - 0.8408\eta_{t-1}.$$

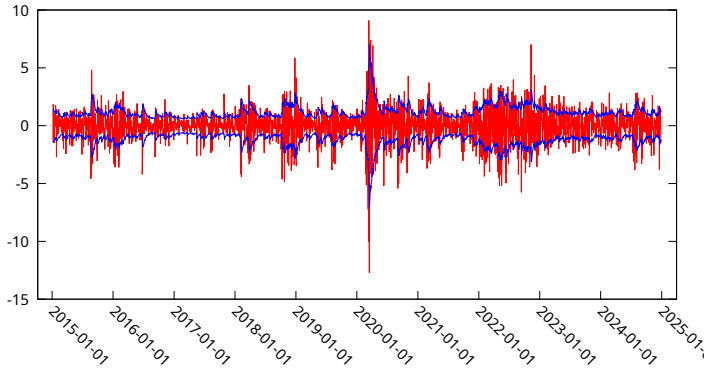
It can easily seen that ϵ_t^2 is a highly persistent process, with an AR coefficient suspiciously close to 1. This is no coincidence: the vast majority of daily financial time series exhibit the same behaviour. Indeed, there are models (the so-called **IGARCH** models) in which a unit root on ϵ_t^2 is *assumed*.

From the estimates above, the unconditional variance of y_t can be consistently estimated as

$$V(y_t) = \frac{c}{1 - a - b} = 1.943;$$

the order of magnitude of the unconditional variance could have also been estimated using the sample variance of y_t (also a consistent estimator), which is indeed 1.8704.

Figure 7.5: Nasdaq index returns – residuals and estimated standard deviation



The h_t series can be reconstructed *ex post* from the forecast errors e_t and the parameter estimates as follows: we set $\hat{h}_0$ to a "plausible" value (any estimate of the unconditional variance will do perfectly well), after which we can compute $\hat{h}_t$ as

$$\hat{h}_t = \hat{c} + \hat{a}e_{t-1}^2 + \hat{b}\hat{h}_{t-1};$$

a fairly common way to present these results is in a plot such as the one in Figure 7.5. The two series drawn in blue are $\pm\sqrt{\hat{h}_t}$, namely the estimated conditional standard deviation and its negative.[7] As one would expect, the residuals — very similar to the series itself, given that the conditional mean is close to zero — are "contained" within these standard deviation bands a high proportion of the time. Note also how, whenever a substantial change in the index occurs (and thus a time series value very far from 0), the estimated value of h_t increases immediately afterwards, only to decline gradually until the next shock. Looking at some descriptive statistics for $\hat{h}_t$, several interesting details emerge: the mean is 1.8697 (again, a consistent estimate of the unconditional variance), while the minimum and maximum are 0.3662 and 49.694, respectively (0.6051 and 7.049 in terms of standard deviation). These numbers clearly convey how volatility fluctuates within the sample under consideration which, it is worth recalling, includes both the COVID pandemic and the start of the Ukraine war. Clearly, once the estimates are obtained, we can also set up a forecasting model: not so much for the conditional mean (as there is very little exploitable persistence), but rather for the volatility. Indeed, it is sufficient to take equation (7.11) and replace the true parameters with their estimated counterparts.

It is also interesting to note that the model does a decent job at accounting for the leptokurtic marginal distribution of the returns. We compute the

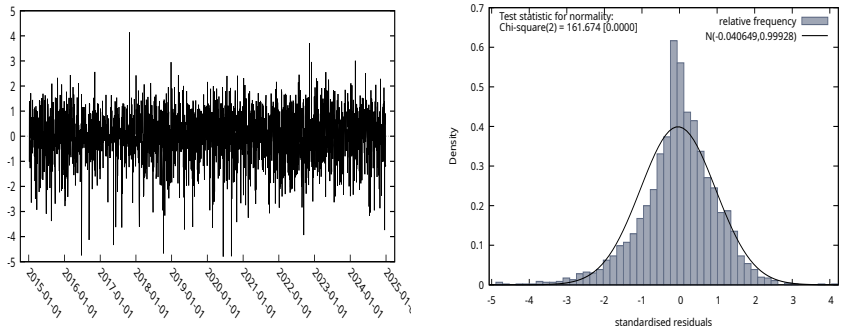
[7] If ϵ_t is normal, the red line should be between the blue lines with 68.3% probability.

"standardised" returns as

$$u_t = \frac{y_t}{\sqrt{\hat{h}_t}}$$

and obtain the time series shown in Figure 7.6 (left-hand pane). It is visually apparent that heteroskedasticity has disappeared, or nearly so. More interestingly, the excess kurtosis of the transformed time series reduces to 1.678, while the Doornik-Hansen test statistic equals 161.674. Non-normality is certainly still there, but to a much lesser extent. It can be deduced that non-normality of the marginal distribution of y_t is largely, although not entirely, attributable to the GARCH effect.

Figure 7.6: Nasdaq returns standardised residuals



Time series plot Marginal distribution

7.3 Extensions

7.3.1 Non-normal distributions

In the example of the previous section, we saw that volatility persistence provided a partial, but not comprehensive explanation for excess kurtosis: even after normalising the time series of residuals by the estimated standard deviation, outliers remain such that the normality hypothesis becomes untenable.

This finding is very common. Therefore, one often uses distributions whose kurtosis can take a value other than 3 are sometimes employed, rather than the normal density; the two most widely used are the Student's t or the *Generalised Error Distribution*, or GED for short.

These two distributions feature density functions that depend on an additional parameter, ν, which must also be estimated alongside those of the mean and variance. For the t-distribution, the parameter ν (not necessarily an integer) is always positive and is termed the "degrees of freedom" parameter; for any value of this parameter, the corresponding density has fat tails, but it converges to the normal distribution as $\nu \to \infty$. For the GED case, ν is also a positive real number, but the excess kurtosis is positive for $\nu < 2$ and negative

for $\nu > 2$. The value 2 is special because the GED density with parameter 2 coincides with the normal distribution (in other words, the normal distribution is a special case of the GED). The analytical expressions for these two densities omitted here, but one gets an intuitive feel by considering Figure 7.7, which displays the densities of the t and GED distributions (normalised to have unit variance) for various values of ν.

Therefore, a viable empirical strategy may involve estimating a GARCH model under an alternative distributional assumption, which amounts to employing a different definition of ℓ_t in equation (7.9). If the GED distribution is used, one also benefits from the advantage of being able to test the normality hypothesis by means of the restriction $\nu = 2$. However, the issue encountered in these cases is that the numerical maximisation of the log-likelihood is considerably more demanding from a computational perspective: given that, as previously noted, the non-convergence of a GARCH estimate is an uncommon but by no means impossible event, this further challenge risks introducing additional complications.

More generally, however, you can never be sure that the density function you choose (be it normal, t or GED) is the "true" one. This why an estimation method known as *quasi-maximum likelihood* (QML) is very often used. The underlying idea is that, under certain regularity conditions (which are omitted here, but are not particularly stringent), the parameter estimates obtained by maximising a certain likelihood remain consistent, albeit inefficient, even if the "true" density is different. What is *not* consistent is the estimator of the parameter covariance matrix. However, there's a remedy: a robust estimator, known as "sandwich" estimator, is available in all econometric software packages. The one typically used in this context is the so-called Bollerslev–Wooldridge estimator. The estimates presented earlier in Table 7.2 were in fact obtained using robust standard errors.[8]

Finally, it must be mentioned that sometimes it could be useful to employ "skewed" variants of the densities mentioned above, to account for the fact that the density of unconditional return may not be symmetric (look at Figures 7.4 and 7.6). Although fat tails can account for large outliers, they cannot account that negative outliers can often be much larger than positive outliers: in fact, we call them "market crashes".

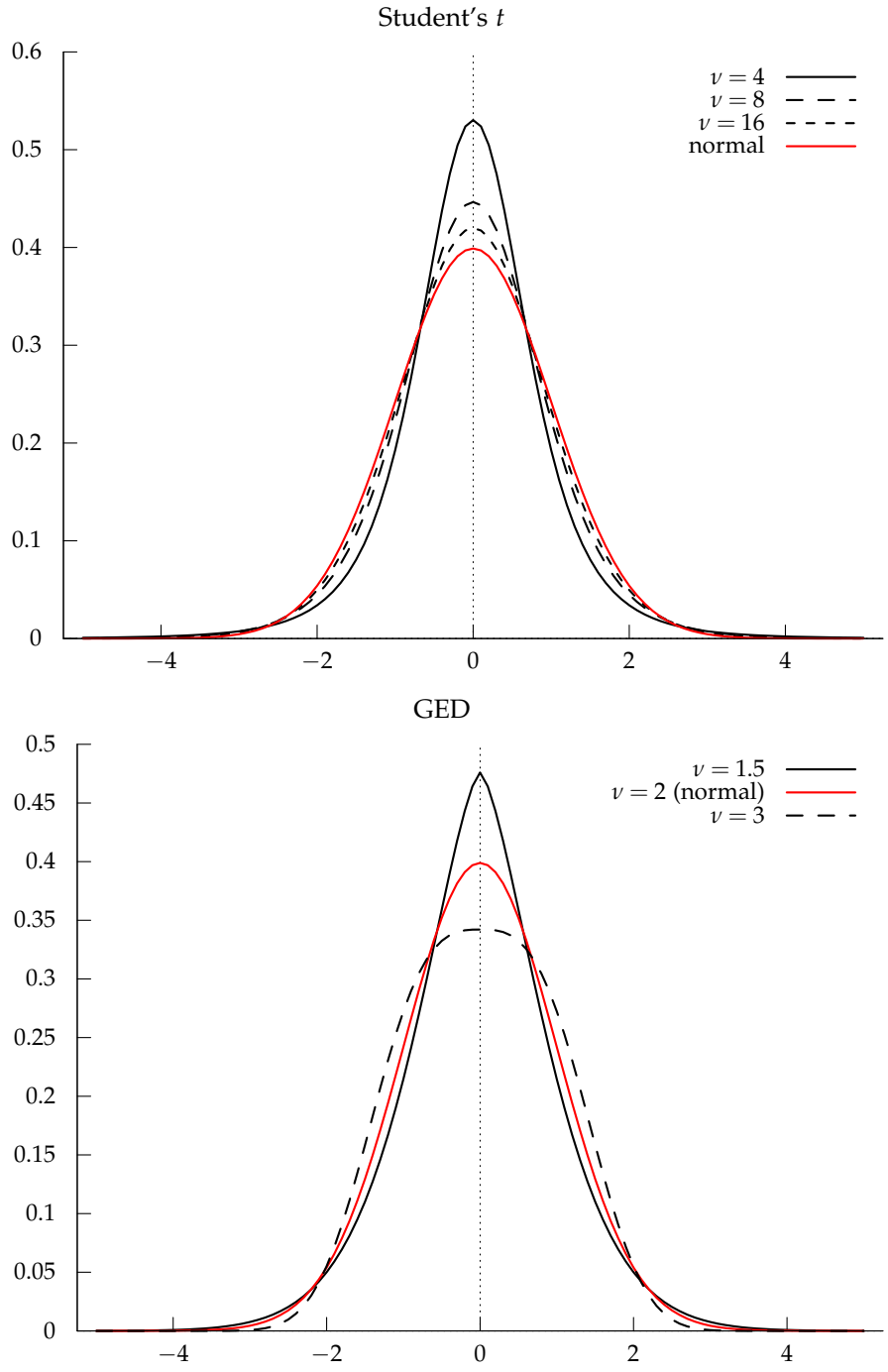
7.3.2 The GARCH zoo

The ARCH and GARCH models I described in the previous sections were introduced in 1982 (by Robert Engle) and in 1986 (by Tim Bollerslev), respectively. I suppose most readers have no personal memory of those years, but I am old enough to remember them very well. Those were the years of

[8] The logic is identical to the one which motivates, for instance, the usage of White's covariance estimator in a cross-sectional regression: in that case, we keep OLS as our estimator of choice because it is inefficient but consistent. We accept the inherent efficiency loss (not necessarily huge) in return for robust inference.

The similarity between the two scenarios is no coincidence: the entire underlying theory was developed through the analysis of misspecified models, pioneered, as it happens, by none other than Halbert White.

Figure 7.7: Alternative Distributions



Madonna, *Back to the Future*, shoulder pads, the first PCs and unfettered optimism about capitalism.

Any econometrician who could come up with a new shiny idea could aspire to academic success, and possibly accumulate considerable wealth in the process. As you can imagine, there was some proliferation of GARCH variants. Needless to say, only few of them stood the test of time. In this section, I will review a few variations on the GARCH theme that have been put forward and are still in use today.

Asymmetric effects

A common stylised fact I haven't mentioned so far goes under the name of **leverage effect**. To see what it is, consider equation (7.7): by the mechanism I described in section 7.2.1, a large shock at time $t-1$ raises variance at time t, which in turn makes large shocks more likely, and so on.

However, a commonly observed empirical phenomenon is that large *negative* shocks seem to increase volatility more than positive ones. Therefore, it may be desirable to incorporate this asymmetry in the equation for h_t. The most widely used GARCH variant to incorporate asymmetric effects is the so-called GJR model (after its inventors Glosten, Jagannathan and Runkle). The law of motion for the GJR model can be specified in two equivalent ways:[9]

$$h_t = c + A(L)\left[\epsilon_{t-1} + \gamma|\epsilon_{t-1}|\right]^2 + B(L)h_{t-1}. \tag{7.12}$$

$$h_t = c + A(L)\left(\epsilon_{t-1}^2 + \gamma d_{t-1}\epsilon_{t-1}^2\right) + B(L)h_{t-1}. \tag{7.13}$$

where $d_t = 1$ if $\epsilon_t < 0$ and 0 otherwise. In both formulations, the parameter γ measures the strength of the asymmetry effect: if $\gamma = 0$, you go back to an ordinary GARCH model.

Note: the asymmetry effect we're dealing with here is different from the asymmetry in the density I discussed at the end of Section 7.3.1. Asymmetry in the density affect possible differences in the probability of observing large positive vs negative outliers. Here, we're considering the effect of positive vs negative shocks on the conditional volatility.

For example: Table 7.3 reports the estimates for the same data as Table 7.2, with two differences: (a) a GJR model was used (see eq. (7.12)) and (b) I used a GED density for ϵ_t. The most important results are:

- the autoregressive parameter φ is no longer significant.
- The γ parameter is very significant; therefore, there is strong evidence of leverage effect.
- A standard test for $H_0 : \nu = 2$ is strongly rejected; therefore, the distribution tails are indeed quite thick and ϵ_t is, without doubt, non-normal.

Finally: the GJR model admits the GARCH model as a nested case. Following this logic, a model called APARCH has been devised, which contains quite a few members of the GARCH family as special cases.

[9] Where proving the equivalence is, as usual, left to an exercise for the reader.

Table 7.3: GJR/GED estimates

Parameter	Coefficient	Std. error	t-statistic	p-value
μ	0.0912	0.0240	3.802	0.0001
φ	-0.0240	0.0217	-1.105	0.269
c	0.0349	0.0104	3.366	0.0008
a	0.0898	0.0204	4.412	1.02E-05
γ	0.4955	0.1413	3.507	0.0005
b	0.8678	0.0203	42.79	0
ν	1.2048	0.0525	22.96	1.15E-116

EGARCH

The EGARCH model was proposed by Daniel Nelson in 1991, and is based on a simple idea: h_t is a variance, so it must be positive. This, as we saw earlier, makes it necessary to enforce certain restrictions on the parameters for its law of motion (7.7). By considering its logarithm instead, these problems disappear. In the EGARCH (Exponential GARCH) model, equations (7.7) or (7.12) are replaced by something like:[10]

$$\ln h_t = c + A(L)\left[u_{t-1} + \gamma|u_{t-1}|\right]^2 + B(L)\ln h_{t-1}. \tag{7.14}$$

The EGARCH model is quite popular, and has pros and cons. Its main advantages are that it incorporates the concept of asymmetric effects in a very natural way. Moreover, there is no need to impose non-negativity constraints on the parameters. Thus, the convergence problems that occasionally affect standard GARCH models do not arise.

On the other hand, deriving analytical derivatives for the EGARCH model is complicated, and almost no software package that I am aware of does so; for the most part, routines rely on numerical differentiation. Consequently, convergence and/or numerical precision issues may arise. As for forecasting, if a forecast for h_t is required, determining how to compute it is far from straightforward. Recall that, typically, the conditional mean is used as the predictor, but the EGARCH specification provides, at most, the expected value of $\ln h_t$. As is well known:

$$E(e^X) \neq e^{E(X)}$$

and therefore simply exponentiating $\widehat{\ln h_t}$ yields a predictor with properties that are not entirely clear. Almost invariably, practitioners use the bootstrap.

GARCH-in-mean

Financial theory predicts the existence of something called "risk premium": for a risk-adverse investor, a risky asset can be considered equivalent to a risk-free asset only if it gives higher return (think stocks vs bonds vs cash). If y_t

[10] This formulation is slightly less general than the "true" one, but it is the one most frequently adopted

is the return of an asset and h_t is its volatility, then one may expect a positive relationship between the two.

This is the base intuition behind **GARCH-in-mean** models, where h_t enters the conditional mean equation:

$$y_t = x_t'\beta + \psi h_t + \epsilon_t \tag{7.15}$$
$$h_t = V(\epsilon_t|\mathcal{F}_{t-1}). \tag{7.16}$$

In some cases, equation (7.15) is modified to use the conditional standard deviation $\sqrt{h_t}$, but the rationale is the same.

In such a context, therefore, the parameter ψ is easily interpreted as a measure of the risk premium: it indicates by how much the return must change as its own volatility increases. In practice, however, these models have become somewhat obsolete, because the "risk-premium term" ψh_t frequently captures other effects omitted from the observable information set rather than the risk premium itself, often leading to estimates of the parameter ψ that are incompatible with reasonable values. To quote Tim Bollerslev, the inventor of GARCH:

> *[...] estimates based on GARCH-in-mean models, or simple regressions of returns on lagged measures of volatility and/or surprises therein, often fail to establish a significant risk-return tradeoff relationship, and sometimes even suggest that expected returns and volatility are negatively related.* (in Bollerslev, 2022)

Work in progress from now on

MEM

In some cases, you want to model things that are (apparently) only vaguely related with volatitlity, such as durations or volumes, that are positive. ... MEM ...

Oddly enough, in its most basic form a MEM model is just a GARCH in disguise...

GAS

GAS stands for "Generalized Autoregressive Score": ...GARCH is a special case under Gaussianity ...

7.4 Stochastic volatility

Unobservable components ... $\log h_t$ is an unobservable AR(1) process ... note that this is *not* the same as an ARCH/EGARCH process, since $\log h_t$ is *not* driven by ϵ_t, but follows an independent path ...

7.5 Realised volatility

The need for estimating h_t comes from the fact that the variance of a random variable is unobservable; therefore, we estimate it by using some parametric assumption. Therefore, you apply GARCH models to daily data to estimate the daily conditional variance of y_t.

However, if one had intra-period data (such as 15-minute or 5-minute data), then the volatility for a given day would be directly estimable. These datasets have been available since the beginning of the XXI century, so considerable work and effort have been devoted to this topic.

Suppose we have 5-minute observations for the log price of an asset p_t: a whole day's worth of observation can therefore be represented as

$$p_{t,0}, p_{t,1}, p_{t,2}, \dots p_{t,(h-1)}, p_{t,h}$$

where the second index identifies the 5-minute interval, so you have $h+1$ such intervals in one day. By defining

$$r_{t,j} = p_{t,j} - p_{t,(j-1)},$$

the high-frequency returns, you could compute "realised volatility" as

$$RV_t = \frac{1}{h} \sum_{j=1}^{h} r_{t,j}^2$$

- Is this a good measure? The theory you need involves the analysis of continuous-time stochastic processes (too advanced);
- there are literally hundreds of alternatives, but apparently the 5-minute naive thing is hard to beat.
- Are parametric models avalaible for forecasting volatility starting from RV? Yes.

...

7.6 Multivariate models

Problems:

1. very large number of unknown parameters,
2. definitness of the matrix Σ_t

- CCC/DCC (many different varieties)
- BEKK

In both cases, to reduce the computational complexity of the problem, you use "variance targeting"...

Chapter 8

Further reading (not really updated very much)

8.1 In general

This text does not assume more than a basic knowledge of the fundamentals of statistics and econometrics. In many instances, this has led to rather drastic simplifications. To begin with, the expert reader will have noticed a sovereign disregard for the highly technical aspects of a probabilistic and inferential nature, which, especially in the analysis of non-stationary processes, can be remarkably demanding. If you are genuinely fond of theory (like the gut who wants to know what ergodicity and mixing *really* are), then take a look at Davidson (1994), which is marvellous (though it requires a bit of effort). McCabe and Tremayne (1993) is somewhat more accessible. For the asymptotics of ML estimators, it's hard not to recommend Duffy (2018). Martin et al. (2012) devote nearly half of their book to methodology, so you may want to check that out too.

In fact, what you need to know to follow up on the many occasions where I suggest consulting the relevant literature can be found in any econometrics textbook, at any level. Greene (2018) or Verbeek (2017) are perfectly OK; I personally adore Davidson and MacKinnon (2021), which is more advanced but absolutely recommended for an in-depth study. If you want a book of inferior quality, but whose style and notation are very consistent with the ones used here, go for Lucchetti (2026).

It should also be noted that a topic which is conspicuous by its absence in these notes is spectral analysis. The great classic on spectral analysis is Brillinger (2001). Shumway and Stoffer (2014) is also excellent (although not specifically tailored to economists) and contains a thorough account of nearly everything you con find here. Brockwell and Davis (1991) is splendid, although sadly a bit outdated now.

8.2 Univariate Stochastic Processes

On univariate stationary time series, a few *very* old sources are worth giving a look: for example, Granger and Newbold (1986), Hamilton (1994) and Brockwell and Davis (1991). A textbook with a very distinctive approach, which one may either love or hate (I love it), is Harvey (1993). Also of great interest is Sargent (1987), primarily because it demonstrates how certain time series analysis concepts (which, moreover, it explains exceptionally well) are applicable within an economic-theoretical framework rather than a purely statistical one.

A more recent book you should definitely get hold of is, again, Shumway and Stoffer (2014).

8.3 VAR processes

Hamilton (1994) is a great classic but, unfortunately, quite outdated now. I recommend Lütkepohl (2005) or Kilian and Lütkepohl (2017), which is more heavily focussed on structural VARs. For those interested in the historical development of econometric ideas, reading the article that initiated the entire VAR literature, namely Sims (1980) is almost mandatory; it also contains interesting reflections on the use of time series analysis in economics.

8.4 I(1) processes and cointegration

Here we enter a somewhat more technical field, and it is difficult to recommend works that are not somewhat advanced. Hamilton (1994) is an excellent departure point, even though I consider it somewhat convoluted on certain topics, and it shows its age here and there. On the other hand, Maddala and Kim (1998) is a textbook with a completely different framework: it features no lengthy theorem proofs, but offers a monstrously comprehensive review of the relevant literature, coupled with assessments that are often sharp yet always insightful. Also recent and, in some respects, ingenious is Davidson (2000). If you really wnaty to know all you'll ever need on univariate unit roots, go for Choi (2015).

A highly interesting textbook that, in my view, balances theory and practice remarkably well is Banerjee et al. (1993). At a more accessible level, I recommend Enders (1995) and Cuthberson et al. (1992). From a theoretical standpoint, the properties of integrated series have been extensively studied, most notably by P.C.B. Phillips, who expressed himself at his best in a series of papers, among which I particularly recommend Phillips (1986) and Phillips and Durlauf (1986). Also highly interesting and more closely tailored to the applied economist is Campbell and Perron (1991).

Regarding non-stationary VARs, a recent and very pleasant reference is Mills (1998), which contains a readable overview of topics usually considered rather esoteric, such as Granger-causality tests in cointegrated VARs. For a more extensive reference on the interpretation of cointegrated models, and

especially on their link with ECMs, one cannot fail to recommend the father of Error Correction Models himself, David Hendry: among his vast body of work, it is worth highlighting Hendry (1995), alongside the previously mentioned Banerjee et al. (1993).

As for Johansen's procedure, the primary source is naturally Johansen himself in Johansen (1995). This book is somewhat demanding; thus, those wishing to merely get an overview can comfortably rely on Hamilton (1994), which also explains the Fully-Modified OLS technique reasonably well, or on Johansen (2000), which is recent and concise. Juselius (2006) is clear, comprehensive and full of applied examples.

Boswijk and Doornik (2003) is a well-crafted paper that moreover has the merit of being downloadable from the Internet (for the time being). Also worth noting is Hargreaves (1994), which provides a very relaxed, albeit somewhat dated, examination of the main estimation methods for cointegrating vectors. At any rate, so many textbooks on cointegration have been published that it is difficult even to keep track of them; one highly introductory text that I consider particularly effective from a pedagogical perspective is Harris (1995).

8.5 Conditional heteroskedastic processes

Since you will have undoubtedly photocopied or downloaded it by now (or better yet, bought it), Hamilton (1994) contains a far from superficial discussion of ARCH/GARCH models. Nevertheless, I am obliged to direct anyone wishing to delve deeper to Bollerslev et al. (1994), which remains the absolute Bible on the subject, even if it is no longer entirely up to date. On the other hand, Mikosch et al. (2009) is very detailed, introductory, and very well reasoned.

It is true, however, that with the explosion of GARCH applications in finance, the bibliographical references have become countless. In fact, an independent literary genre has emerged, which masterfully blends finance and econometrics. Along these lines, the classic reference is Campbell et al. (1997).

Bibliography

Banerjee A, Dolado J, Galbraith J, Hendry D. 1993. *Co-Integration, Error Correction and the Econometric Analysis of Non-Stationary Data*. Oxford University Press.

Bollerslev T. 2022. Realized semi(co)variation: Signs that all volatilities are not created equal. *Journal of Financial Econometrics* **20**: 219–252.

Bollerslev T, Engle RF, Wooldridge J. 1994. ARCH models. In Engle RF, McFadden DL (eds.) *Handbook of Econometrics*, volume 4. Elsevier, 2959–3031.

Boswijk HP, Doornik J. 2003. Identifying, estimating and testing restricted cointegrated systems: An overview. Technical report, Economics Group, Nuffield College, University of Oxford.

Brillinger DR. 2001. *Time series: data analysis and theory*. SIAM.

Brockwell PJ, Davis RA. 1991. *Time Series: Theory and Methods*. Springer-Verlag, 2nd edition.

Campbell JY, Lo AW, Mckinley AC. 1997. *The Econometrics of Financial Markets*. Princeton University Press.

Campbell JY, Perron P. 1991. Pitfalls and opportunities: What macroeconomists should know about unit roots. In Blanchard OJ, Fischer S (eds.) *NBER Macroeconomics Annual 1991*. MIT Press, 141–201.

Choi I. 2015. *Almost All about Unit Roots*. Cambridge University Press.

Cuthberson K, Hall SG, Taylor MP. 1992. *Applied Econometric Techniques*. Philip Allan.

Davidson J. 1994. *Stochastic Limit Theory*. Cambridge University Press.

Davidson J. 2000. *Econometric Theory*. Blackwell.

Davidson R, MacKinnon JG. 2021. *Econometric theory and methods*. Oxford University Press New York.
URL http://qed.econ.queensu.ca/ETM/

Duffy JA. 2018. Asymptotics of extremum estimators. Lecture notes.
URL https://sites.google.com/site/jamesaduffy/lectures/asymptotics-for-extremum-estimators-hilary-2018

Enders W. 1995. *Applied Economic Time Series Analysis*. John Wiley & Sons.

Granger CW, Newbold P. 1986. *Forecasting Economic Time Series*. Academic Press.

Greene W. 2018. *Econometric Analysis*. Pearson, 8th edition.

Hamilton J. 1994. *Time Series Analysis*. Princeton University Press.

Hargreaves CP. 1994. A review of methods of estimating cointegrating relationships. In Hargreaves CP (ed.) *Nonstationary Time Series Analysis and Cointegration*. Oxford University Press, 87–131.

Harris R. 1995. *Using Cointegration Analysis in Econometric Modelling*. Prentice-Hall.

Harvey AC. 1993. *Time Series Models*. Harvester Wheatsheaf, 2nd edition.

Hendry DE. 1995. *Dynamic Econometrics*. Oxford University Press.

Johansen S. 1995. *Maximum Likelihood Inference in Co-Integrated Vector Autoregressive Processes*. Oxford University Press.

Johansen S. 2000. Modelling of cointegration in the vector autoregressive model. *Economic Modelling* **17**: 359–373.

Juselius K. 2006. *The Cointegrated VAR Model: Methodology and Applications*. Advanced Texts in Econometrics. OUP Oxford. ISBN 9780199285662.

Kilian L, Lütkepohl H. 2017. *Structural vector autoregressive analysis*. Themes in modern econometrics. Cambridge, United Kingdom: Cambridge University Press. ISBN 9781316647332.

Lucchetti RJ. 2026. Basic econometrics.
URL http://www2.econ.univpm.it/servizi/hpp/lucchetti/didattica/basic.pdf

Lütkepohl H. 2005. *New introduction to multiple time series analysis*. Springer-Verlag.

Maddala G, Kim I. 1998. *Unit Roots, Cointegration and Structural Change*. Cambridge University Press.

Martin V, Hurn S, Harris D. 2012. *Econometric Modelling with Time Series: Specification, Estimation and Testing*. Themes in Modern Econometrics. Cambridge University Press. ISBN 0521139813; 9780521139816.

McCabe B, Tremayne A. 1993. *Elements of modern asymptotic theory with statistical applications*. Manchester University Press.

Mikosch T, Kreiß JP, Davis RA, Andersen TG (eds.). 2009. *Practical Issues in the Analysis of Univariate GARCH Models*. Berlin, Heidelberg: Springer. ISBN 978-3-540-71297-8, 113–155.

Mills TC. 1998. Recent developments in modelling nonstationary vector autoregressions. *Journal of Economic Surveys* **12**: 279–312.

Pagan AR. 1989. 20 years after: Econometrics 1966–1986. In Cornet B, Tulkens H (eds.) *Contributions to Operations Research and Econometrics: The XXth Anniversary of CORE*. Cambridge, MA: MIT Press, 319–383.

Phillips PC. 1986. Understanding spurious regression in econometrics. *Journal of Econometrics* **33**: 311–40.

Phillips PC, Durlauf S. 1986. Multiple time series regression with integrated processes. *Review of Economic Studies* **53**: 473–95.

Sargent TJ. 1987. *Macroeconomic Theory*. Academic Press, 2nd edition.

Shumway RH, Stoffer DS. 2014. *Time series analysis and its applications: with R examples*. Springer, 4th edition.

Sims CA. 1980. Macroeconomics and reality. *Econometrica* **48**: 1–48.

Sims CA, Stock J, Watson M. 1990. Inference in linear time series models with some unit roots. *Econometrica* **58**: 113–44.

Verbeek M. 2017. *A guide to modern econometrics*. John Wiley & Sons Chichester, 5th edition.

Index

Autocorrelation
    partial, 41
Autocorrelations, 5
Autocovariance, 5

Beveridge-Nelson decomposition, 69
    in cointegrated systems, 143

Cholesky decomposition, 106
Cointegration
    attractor, 124
    definition, 119
    vectors, 120
Common factors (COMFAC), 39
Companion form, 89
Correlogram, 7

Deterministic component
    in cointegrated systems, 130
    in unit root tests, 74
DOLS, 140

ECM (Error Correction Mechanism), 122
Engle-Granger, method, 140
Ergodicity, 4
Explosive root, 23

filter, 13
FM-OLS, 140
Forecast function, 45

Granger
    causality, 101

Identification
    in the Box-Jenkins sense, 41
    in the econometric sense, 40
Impulse response function
    in multivariate processes, 104
    in univariate processes, 49
information criteria, 41
Information criteria, 42
Information set, 5

Johansen, method, 134

Lag operator, 13
Leverage effect, 162
Likelihood, 37
    sequential factorisation, 43

Martingale difference, 17
Mixing, 4

Operator "$\perp$", 127

Persistence, 2

Random Walk, 64
Representation theorem
    Granger, 126
    Wold, 21

Spurious regression, 80
Stationarity, 3
    of AR(1) processes, 22
    of ARMA processes, 27
    of AR processes, 25
    of VAR processes, 88
    Strong, 3
    Weak (or in covariance), 3
Stocastic process
    with persistent volatility, 147
Stochastic process, 2
    $I(1)$, 63
    ARCH, 152
    ARMA, 13, 27
    AR (autoRegressive), 22
    Conditionally heteroskedastic, 12

DS (Difference-Stationary), 63
EGARCH, 163
GARCH, 154
MA (moving average), 17
mean-reverting, 65
Multiplicative ARMA, 28
multivariate, 85
Seasonal ARMA, 28
TS (Trend-Stationary), 62
VAR (vector autoregressive), 87
Structural shocks, 105

Test
Ljung-Box, 42
LM-ARCH, 153
unit root, 71
Trend
common, 129
deterministic, 61, 66
stochastic, 66

Unit root, 22, 64
Unit root test
ADF-GLS, 75, 76
Augmented Dickey-Fuller (ADF), 74
Dickey-Fuller (DF), 73
KPSS, 76
Phillips-Perron (PP), 75

VAR
Structural, 104
VARMA, 87

White noise, 16
multivariate, 86