LECTURE NOTES ON TIME SERIES

Matteo Barigozzi*

This version December 13, 2019

Contents

*m.barigozzi@lse.ac.uk.

Chapters 6 and 8 are partly taken from Marco Lippi lecture notes.
`http://www.lippi.ws/SC/SC.htm#Lecturenotes`
Chapters 10 and 11 are partly taken from Jack Lucchetti lecture notes (in Italian).
`http://www2.econ.univpm.it/servizi/hpp/lucchetti/didattica/matvario/procstoc.pdf`.

1 Introduction

1.1 Motivation

Time series are measurements of a quantity x_t, taken repeatedly over a certain period of time.

- The quantity x_t can be a scalar, but it can also be a vector, or a more complex object such as an image or a network.
- The time index t can be continuous (when x_t is observed continuously), discrete and equally spaced (when x_t is measured at discrete time intervals, e.g. every day or every month), or have a more complex form (think of an experiment which needs close supervision at the beginning, but can later be observed less frequently).

Time series arise in many sciences, or more generally in many "domains of human endeavour". We first look at some examples of time series, before moving on to describe the branch of statistics called Time Series Analysis.

1.2 Examples of time series

Note that plotting the values of a scalar-valued time series is often the most natural way of visualising datasets. The transformations used are explained below in section 1.3.

1. Finance. Standard & Poor's 500 index, daily data, source: `http://research.stlouisfed.org/fred2/`. Description: "The S&P 500 is regarded as a gauge of the large cap U.S. equities market. The index includes 500 leading companies in leading industries of the U.S. economy, which are publicly held on either the NYSE or NASDAQ, and covers 75% of U.S. equities. Since this is a price index and not a total return index, the S&P 500 index here does not contain dividends."
   (a) Figure 1 shows the data in "levels", i.e. not transformed. The values do not change much from day to day, but over time, clear trends are formed. We notice a general upward trend but also a steep downward trend during the 2007–2008 financial crisis. The mean of this series changes in time.
   (b) Figure 2 shows the time series of the daily percentage increments of the Standard & Poor's 500 index, over the same time period. Contrary to the previous series, there are no clear trends in the mean, which oscillates around zero. However, its variance changes over time.
2. Finance. U.S. / U.K. Foreign Exchange Rate, daily data, source: `http://research.stlouisfed.org/fred2/`. Compare Figures 3 and 4 with Figures 1 and 2.
3. Macroeconomy. U.S. Consumer Price Index (CPI) for all urban consumers and all items, monthly data, source `http://research.stlouisfed.org/fred2/`. Description: "The Consumer Price Index for All Urban Consumers: All Items (CPIAUCSL) is a measure of the average monthly change in the price for goods and services paid by urban consumers between any two time periods. It can also represent the buying habits of urban consumers. This particular index includes roughly 88 percent of the total population, accounting for

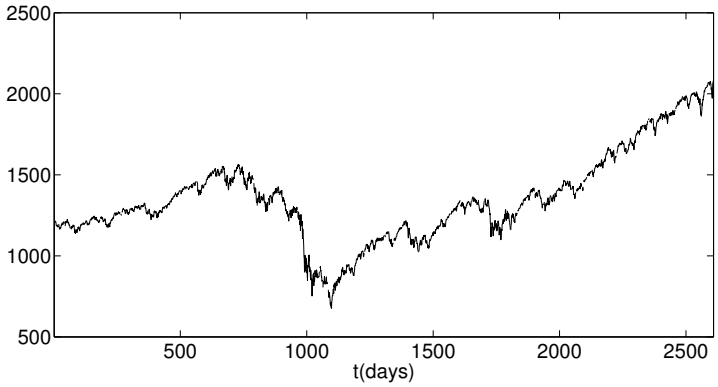


Figure 1: Daily Standard &Poor's 500 Index, from 2004-12-23 to 2014-12-23. Not Seasonally Adjusted

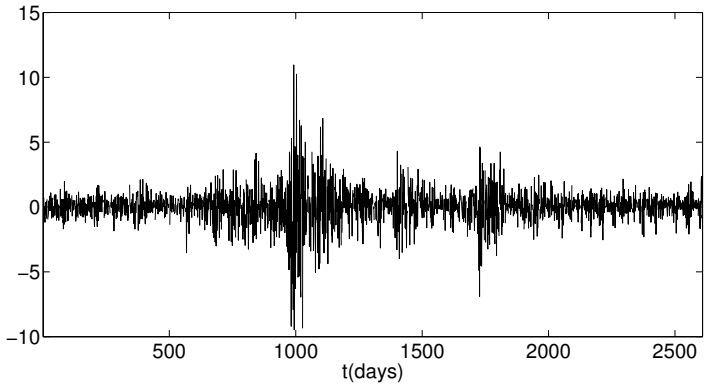


Figure 2: Daily Standard &Poor's 500 percentage returns, from 2004-12-23 to 2014-12-23.

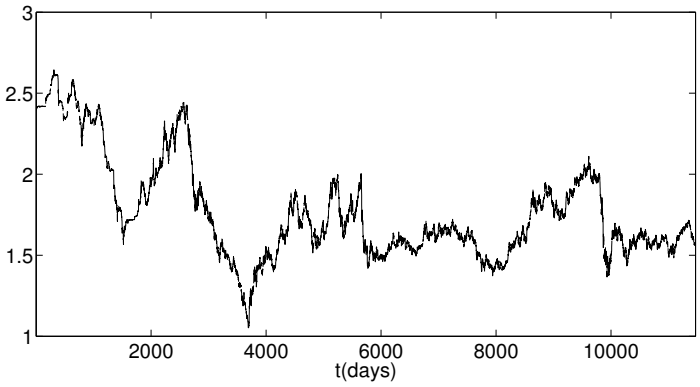


Figure 3: U.S. / U.K. Foreign Exchange Rate U.S. Dollars to One British Pound, from 1971-01-04 to 2014-12-19. Not Seasonally Adjusted

> wage earners, clerical workers, technical workers, self-employed, short-term workers, unemployed, retirees, and those not in the labor force...The CPI can be used to recognize periods of inflation and deflation. Significant increases in the CPI within a short time frame might indicate a period of inflation, and significant decreases in CPI within a short time frame might indicate a period of deflation. "

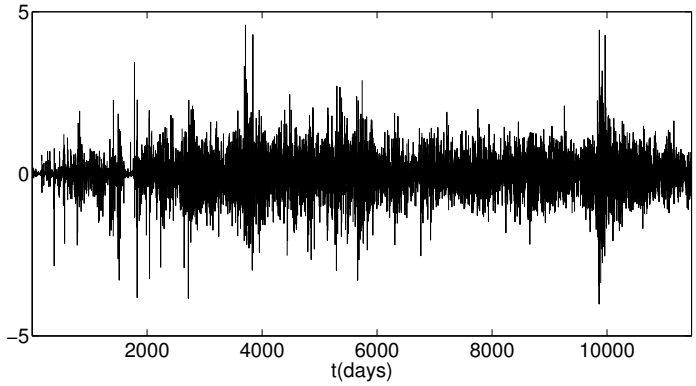


Figure 4: U.S. / U.K. Foreign Exchange Rate percentage returns, from 1971-01-04 to 2014-12-19.

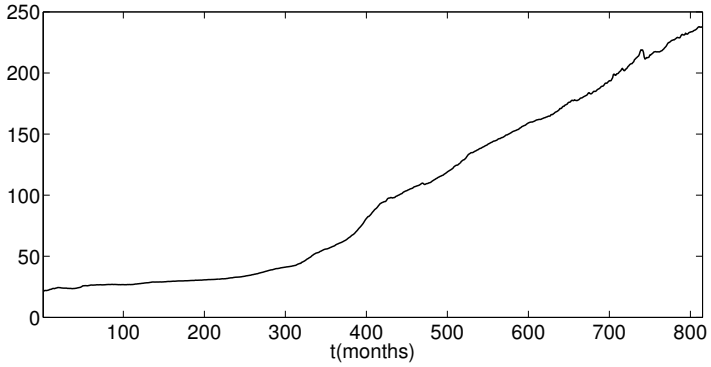


Figure 5: CPI, from 1947-01-01 to 2014-11-01. Index 1982-84=100. Seasonally Adjusted.

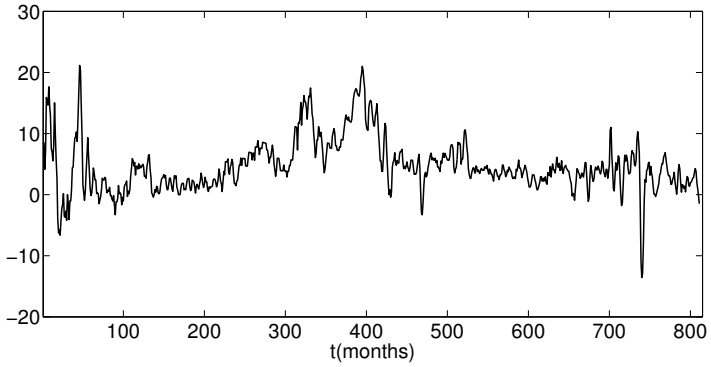


Figure 6: CPI yearly percentage changes, from 1947-01-01 to 2014-11-01.

(a) Figure 5 shows data in "levels", i.e. not transformed. The series is shorter than the previous one (monthly vs. daily), the values oscillate around an increasing trend.

(b) Figure 6 shows the yearly percentage changes in CPI, i.e. the yearly inflation rate. Over the whole period, the data oscillates around an almost flat trend, but in shorter periods we notice increasing and decreasing trends. High inflation in the 1970s and early 1980s, low inflation in the 1990s and 2000s. A sudden drop (deflation) in recent years. Variance changes in time.

(c) Figure 7 shows changes in the yearly percentage changes in CPI, i.e. changes in infla-

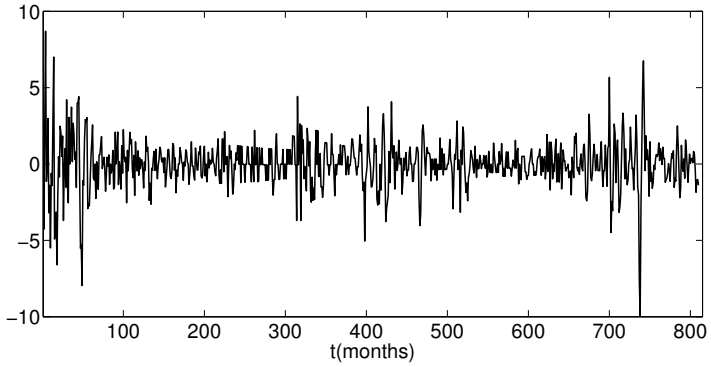


Figure 7: CPI changes of yearly percentage changes, from 1947-01-01 to 2014-11-01.

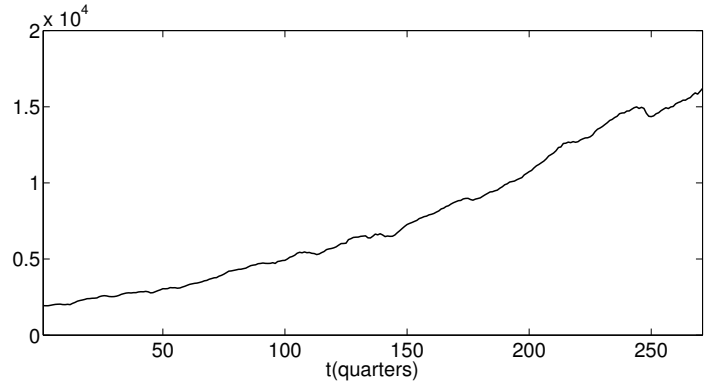


Figure 8: GDP, Billions of Chained 2009 Dollars, from 1947-01-01 to 2014-07-01. Seasonally Adjusted.

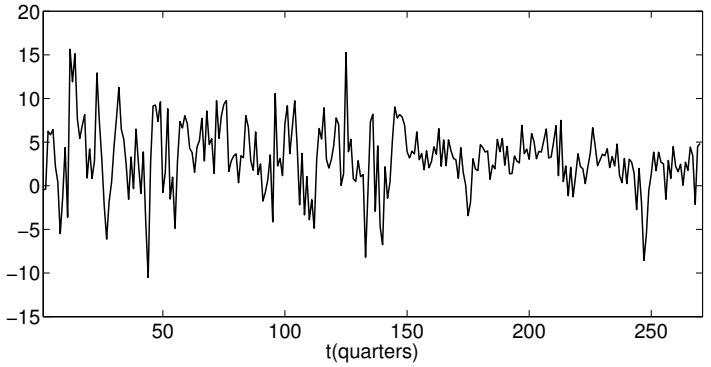


Figure 9: GDP yearly percentage changes, 1947-01-01 to 2014-07-01.

tion. Now data are more stable around a constant (flat) trend.

4. Macroeconomy. U.S. Real Gross Domestic Product (GDP), quarterly data, source `http://research.stlouisfed.org/fred2/`. Compare Figures 8 and 9 with Figures 5 and 6.

5. Weather. Mean maximum temperatures, recorded in Heathrow, monthly data, source `http://www.metoffice.gov.uk/climate/uk/stationdata/index.html`. See Figure 10. The yearly periodicity is very pronounced, as expected. Might there be a slight

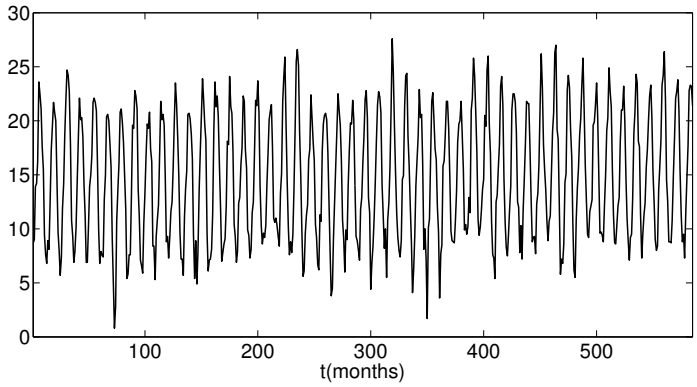


Figure 10: Mean maximum temperatures, recorded in Heathrow, 1948-01-01 to 2014-11-01.

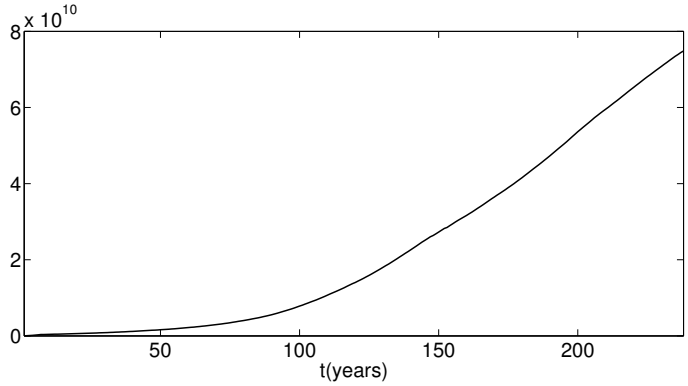


Figure 11: CO2 emissions from fossil-fuels, metric tons, from 1751-01-01 to 2011-12-01.

upward trend towards the end of the series? Anything to do with the "global warming"?

6. Environment. CO2 emissions from fossil-fuels, yearly, source `http://www.gapminder.org/data/`.

   (a) Figure 11 shows data in "levels", i.e. not transformed. The data oscillates around an increasing non-linear trend, maybe quadratic.

   (b) Figure 12 shows yearly differences of data. The data oscillates around an increasing trend which is more linear than before. At the end of the sample the trend is decreasing.

   (c) Figure 13 shows yearly differences of yearly differences (2nd differences). The data now oscillates around a flat trend.

7. Health. Infant mortality rate in Sweden, yearly, source `http://www.gapminder.org/data/`.

   (a) Figure 14 shows data in "levels", i.e. not transformed. The data oscillates around a decreasing non-linear trend, maybe quadratic.

   (b) Figure 15 shows yearly differences of data. The data oscillates around a flat trend with decreasing variance until the variance goes to zero.

8. Engineering. Speech signal (digitised acoustic sound wave) representing the word "Matteo" (my first name) recorded using the `audiorecorder` command in Matlab. Plot in Figure

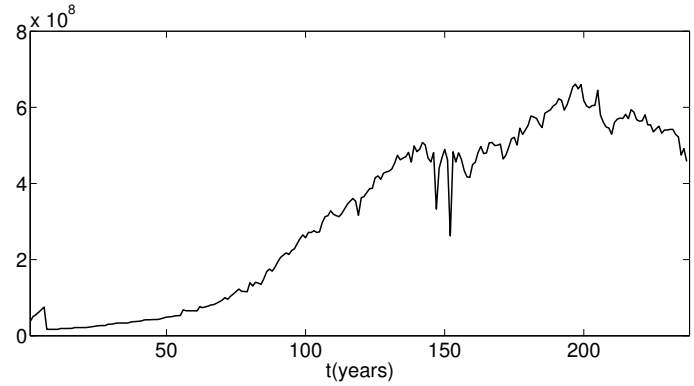


Figure 12: CO2 emissions yearly changes, from 1751-01-01 to 2011-12-12-01.

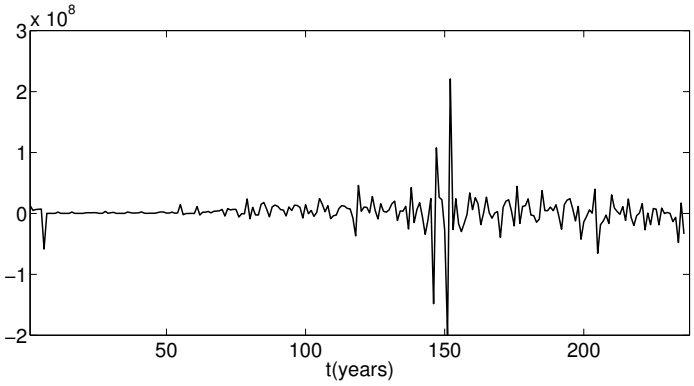


Figure 13: CO2 emissions yearly changes of yearly changes, from 1751-01-01 to 2011-12-01.

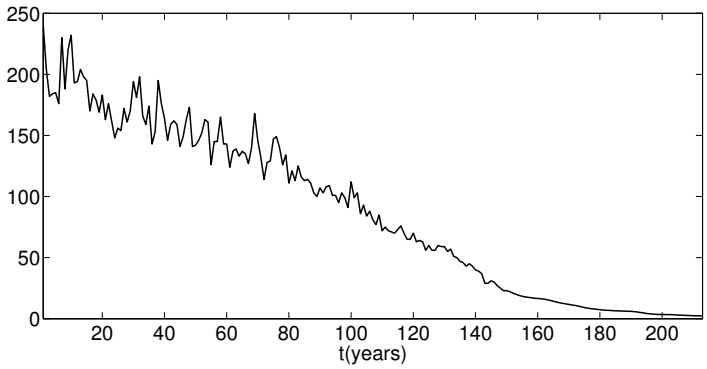


Figure 14: Infant mortality rate, from 1800-01-01 to 2012-12-01.

16. Both the amplitude and the frequency (number of oscillations per second) of the signal change over time.

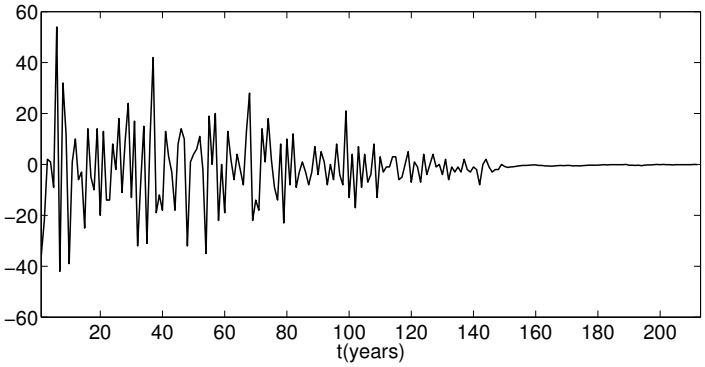


Figure 15: Infant mortality rate yearly changes, from 1800-01-01 to 2012-12-01.

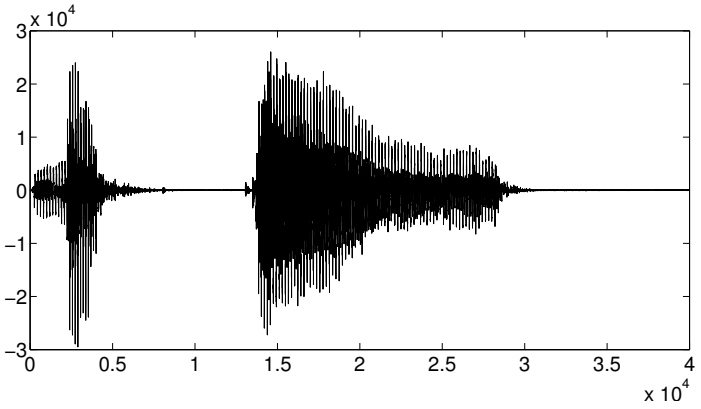


Figure 16: The word “Matteo”.

1.3 Transformation of data

When dealing with a signal x_t sometimes it is useful to transform it (see the examples before) to another signal y_t. We use the symbol Δ to indicate differences, i.e. $\Delta x_t = x_t - x_{t-1}$.

1. data in levels, just take the data as it is $y_t = x_t$;
2. data in logs, take logs of the data, $y_t = \log x_t$, this transformation reduces the variance of the signal;
3. first differences, or changes, $y_t = x_t - x_{t-1} = \Delta x_t$;
4. second differences $y_t = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = \Delta\Delta x_t$ not to be confused with $y_t = x_t - x_{t-2}$;
5. growth rates, there are two possibilities:
   (a) $y_t = \frac{x_t - x_{t-1}}{x_{t-1}}$;
   (b) $y_t = \log \frac{x_t}{x_{t-1}} = \log x_t - \log x_{t-1} = \Delta \log x_t$;

   the two are equivalent for small changes indeed:

$$\frac{\mathrm{d} \log x_t}{\mathrm{d}t} = \frac{1}{x_t} \frac{\mathrm{d}x_t}{\mathrm{d}t}$$

   or, using Taylor approximation in a neighbourhood of x_{t-1},

$$\log x_t = \log x_{t-1} + \frac{1}{x_{t-1}}(x_t - x_{t-1}) + o((x_t - x_{t-1})),$$

   where the last term goes to zero faster than $(x_t - x_{t-1})$ when $x_t - x_{t-1} \to 0$. Transformation (b) is usually preferred, then $100\Delta \log x_t$ is the percentage change from time $t-1$ to time t. This is the way in which we computed the returns of Figure 2 and the percentage changes of other series.

More on the need for these transformations later.

1.4 Statistical time series analysis

Scientists and analysts are interested in a variety of different questions/issues when faced with time series data.

There are at least three kind of questions that can be answered by analysing time series.

1. Forecasting future values in finance and economics/econometrics. For example, for the purpose of potential gain (e.g. in hedge funds or investment banks) or planning for the future (e.g. when should I buy a house?) or for policy makers in setting future interest rates hoping to the improve the state of the economy.

2. Summarise time series data. For example, in the analysis of Electroencephalography (EEG) recordings, how can we decide whether the subject is "healthy" or not? Or how can we decide if the economy in US (see the examples above) has been evolving "significantly differently" from that, say, in China? Or, more generally, given a time series, how can we describe and summarise its evolution?

3. Control the evolution of a time series. This is not quite the same as forecasting, where we do not intervene in the process in any way. As an example of the control problem, consider the global temperature data: is "global warming" really happening, and if so, what impacts the temperature and how can we eliminate or suppress those factors?

As expected, there are often a variety of ways in which those questions can be answered, and many of them do not formally involve statistics at all: for example, people often debate expected trends in house prices and investment opportunities, express their informal views about global warming, or use techniques originating from computer science (e.g. pattern recognition) to aid medical diagnosis in neuroscience. So do we need statistics in time series analysis?

The answer is not necessarily, but there are good arguments why the statistical approach may often be very useful.

1. Firstly, even those informal approaches to time series analysis are in fact often statistical in nature, sometimes in a "hidden" way: for example, people's subjective views about time series can often be formally formulated as priors in Bayesian statistics, and informal forecasts which we often encounter in the media are in fact often instances of simple statistical forecasting procedures, such as trend extrapolation. Also, frequently, techniques originating in computer science (such as: neural networks, machine learning, pattern recognition, artificial intelligence) often have their counterparts in statistics, which do exactly the same thing but are named differently.

2. The above-mentioned tasks: forecasting, understanding the structure of time series, and time series control, have inherent uncertainty about them, which makes probability and statistics a natural tool for describing them.

   (a) Accurate forecasting is impossible. For example, rather than saying "tomorrow's value will be exactly 2.745" it often makes more sense to say "tomorrow's value will be around 2.745", but then there is a chance that we will still be wrong, so our forecasts, even those informal ones, will often be of the form "tomorrow's value will probably be around 2.745", which is already in the territory of probability, since it contains a natural statement of uncertainty.

   (b) It is often impossible to build exact deterministic models describing their structure. Indeed, if we had a correct deterministic model, we would be able to predict the evolution of the time series exactly, but since we are not able to do that, it means that we do not have the exact model. Often, probabilistic models make more sense: for example, "tomorrow's value is about a half of today's value plus a term which is best described as random, i.e. there is no clear pattern in its values from one day to another".

   (c) In the issue of time series control, one natural task that often needs to be performed is to understand what factors affect the evolution of the series. But this is often impossible to specify exactly: it is unlikely that any one factor (out of the ones we are considering), or indeed their combination, is fully responsible for the evolution of the time series.

Therefore, again, a statistical approach, where we permit uncertainty by building a statistical model, might be of use. For example, if we suspect that there is a link between pollution and global warming, it might be helpful to build a statistical model in which we will be able to test this hypothesis, and answer questions like: "how sure are we that there is correspondence between pollution and global warming?", or "what is the strength of this relationship?".

We will find that probability and statistics provide a natural and simple language to express forecasts and their associated uncertainty. In this way, we have a simple model for the evolution of the time series, but in the territory of randomness, i.e. probability and statistics.

1.5 Definition of a stochastic process

A stochastic process is:

1. a correspondence associating a random variable with each integer t:

$$X : t \mapsto X_t$$

2. for any integer s and $t_1 \leq t_2 \leq \ldots \leq t_s$, the joint probability distribution function

$$F_{t_1,t_2,\ldots,t_s}(a_1, a_2, \ldots, a_s) = P(X_{t_1} \leq a_1, X_{t_2} \leq a_2, \ldots, X_{t_s} \leq a_s)$$

Saying that a stochastic process is given means that not only the random variables X_t are given, but also the joint probability measure, for all s-tuples $X_{t_1}, X_{t_2}, \ldots, X_{t_s}$.

We use the notation $\{X_t, t \in \mathbb{Z}\}$ for the stochastic process.

A specification of the process $\{X_t\}$ based on its joint probability distribution function is far too complicated as in general it will depend on too many parameters to be estimated from data. We will specify only the second order moments of the joint distribution, i.e.

$$\mathrm{E}[X_t], \qquad \mathrm{E}[X_t X_{t+h}] \text{ for } t, h \in \mathbb{Z}.$$

These are enough to specify the whole distribution if the data were Gaussian, otherwise some information is lost but all the theory that is developed later in this course is based only on second moments.

In principle there can be infinite realisations of a stochastic process $\{X_t\}$. Typically, we observe only one realization which we denote as the sequence $\{x_t, t \in \mathbb{Z}\}$. In particular, x_t is the realization of the stochastic variable X_t associated with t. We also use, for short, x_t, meaning the process, instead of $\{X_t, t \in \mathbb{Z}\}$.

1.6 Examples of stochastic processes

1. $x_t = A$ for any t is a constant and deterministic process (all its realisations are identical);
2. $x_t = (-1)^t A$ is also a deterministic process;
3. $\{X_t\}$ are i.i.d. zero mean and Gaussian random variables, two possible realisations are in Figure 17;

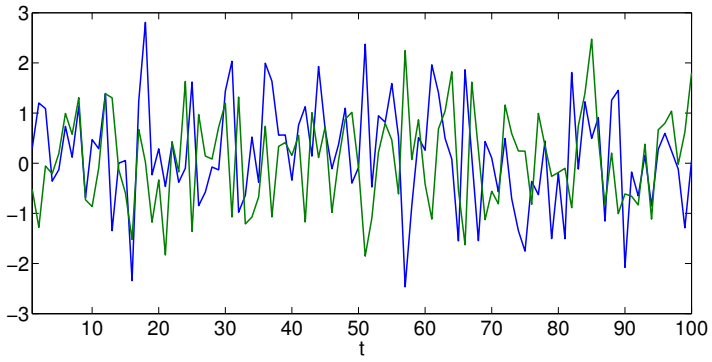


Figure 17: Two realisations of an i.i.d. Gaussian process.

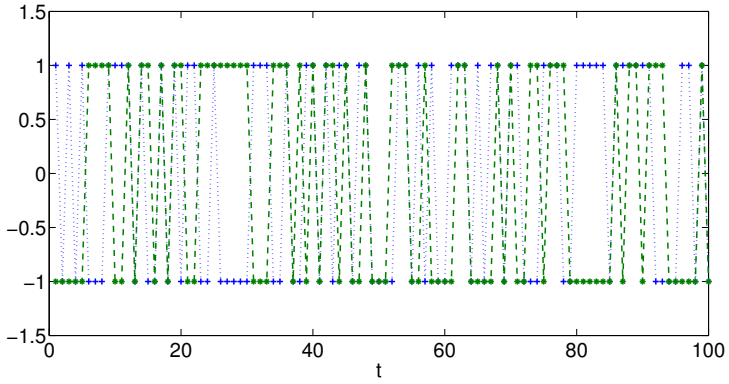


Figure 18: Two realisations of a binary process.

4. $\{X_t\}$ is a binary process, i.e. such that

$$P(X_t = 1) = p, \qquad P(X_t = -1) = 1 - p,$$

two possible realisations are in Figure 18;

5. $\{X_t\}$ is a random walk with zero mean and starting in zero

$$x_0 = 0, \qquad x_t = u_1 + u_2 + \ldots + u_t = \sum_{s=1}^{t} u_s,$$

where u_t are realisations of an i.i.d. Gaussian process as in example 3, four possible realisations are in Figure 19;

6. $\{X_t\}$ is a random walk with drift and starting in zero

$$x_0 = 0, \qquad x_t = t + u_1 + u_2 + \ldots + u_t = t + \sum_{s=1}^{t} u_s,$$

where u_t are realisations of an i.i.d. Gaussian process as in example 3, four possible realisations are in Figure 20.

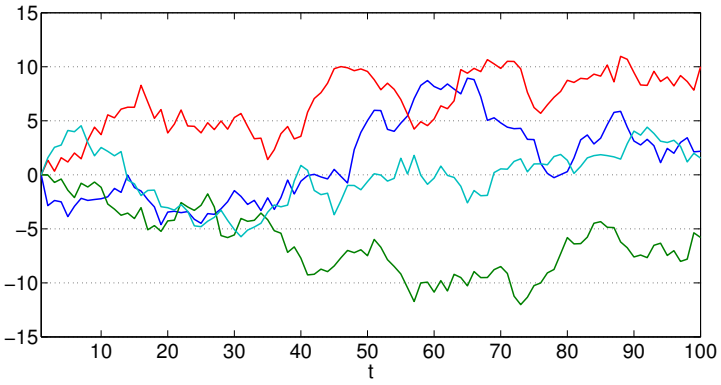


Figure 19: Four realisations of a random walk.

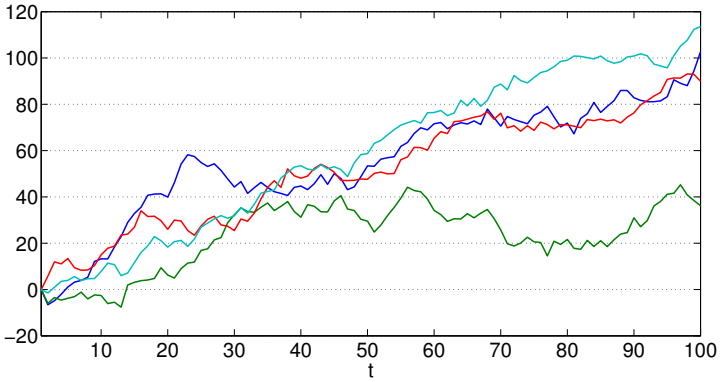


Figure 20: Four realisations of a random walk with drift.

Notice that while in the other cases the process reverts to its mean (which is zero) in the random walk cases this does not happen, and the more the time goes by the more likely the process is to be away from zero. More precisely the uncertainty (variance) increases with time in both cases, while in the case with drift also the mean changes with time (it follows a linear trend).

1.7 Trend and seasonality - part 1

From the examples above it is clear that many time series are made of different components. In general, a process $\{X_t\}$ is made of three components:

$$X_t = T_t + S_t + C_t$$

where T_t is the so called trend, S_t is the seasonal component, and C_t is called cycle and represents fluctuations around the other two components.

Time series analysis proceeds as follows:

1. Check for trends, these can be deterministic functions of time (linear, quadratic...) or also due to some properties of the second order moments of the process. In all cases trends result in time varying mean and/or variance.

2. Remove trends, for example:

(a) if the fluctuations grow linearly with the level of the series, take logs (data must be positive!);

(b) if there are deterministic trends these can be eliminated by means of linear regressions or first differences;

(c) if the are non-deterministic trends (as time varying variances) then first differences should be computed;

(d) taking growth rates accommodates most cases.

in all cases the aim is to obtain a process which is mean reverting (i.e. it is stationary as defined in the next Chapter).

3. Detect and remove seasonal components, as for example by taking other differences of the data or by regressing data on periodic deterministic components.

4. Fit linear models based on the second moments of the residuals of the two steps above. These models are used for forecasting and the results are then combined with the inverse of the above transformations in order to have forecasts of $\{X_t\}$.

Moreover, data can have breaks, i.e. sharp changes in the behaviour of the series, and outliers, i.e. anomalous data points. Both aspects should be taken into account before the analysis.

In the following we will start with models for data with no trend, no seasonality, no breaks, and no outliers. Some of these other aspects will be treated later in the course.

1.8 Autocorrelation

One of the main characteristics of time series data is dependence between observations at different lags: i.e. often, there is a relationship between observations separated by a lag k. Thus given a time series x_t it is natural to study how the signal today depends on its past values. As we said above, classical analysis of time series is based on second order moments and therefore the dependence we consider is only the linear dependence which is measured by means of correlation coefficients.

Suppose then that a linear relationship holds approximately between x_{t+k} and x_t for all integers k, i.e.,

$$x_{t+k} = \alpha_k + \beta_k x_t + \epsilon_{t+k} \tag{1}$$

where ϵ_{t+k} is an error term (cfr this model with linear regressions).

The Pearson product moment correlation coefficient is a summary statistic which measures the strength of the linear relationship between two variables $\{y_t\}$ and $\{z_t\}$ say,

$$\hat{\rho} = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(z_t - \bar{z})}{\sqrt{\sum_{t=1}^{T}(y_t - \bar{y})^2(z_t - \bar{z})^2}}$$

where T is the number of observations we have and $\bar{y}$ and $\bar{z}$ are sample means, thus $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$. (The asymptotic properties of $\hat{\rho}$ are considered later.)

Then for model (1) we have (set $y_t = x_{t+k}$ and $z_t = x_t$) the lag k sample autocorrelation for a time series:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{T-k}(x_{t+k} - \bar{x})(x_t - \bar{x})}{\sqrt{\sum_{t=1}^{T-k} t(x_{t+k} - \bar{x})^2(x_t - \bar{x})^2}}$$

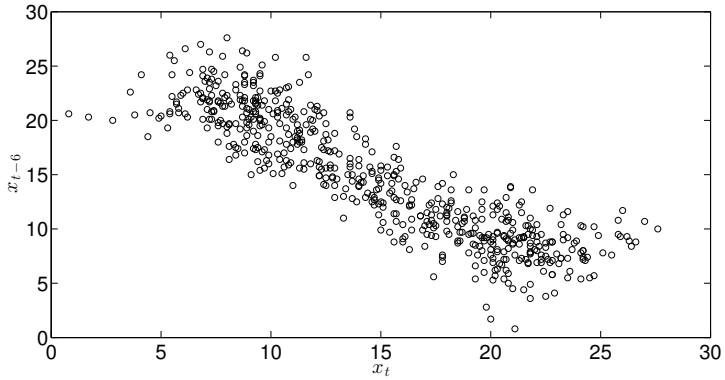


Figure 21: Scatter plot of x_{t+6} against x_t.

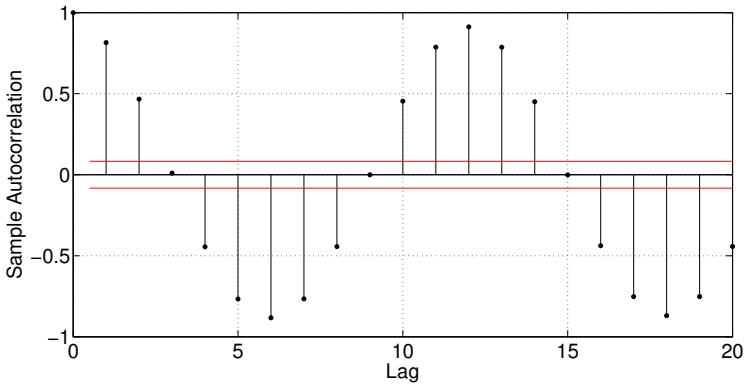


Figure 22: Sample autocorrelation $\hat{\rho}_k$ for x_t.

and $\hat{\rho}_0 = 1$. The sequence $\{\hat{\rho}_k\}$ is called the sample autocorrelation sequence (sample acs) of the time series.

Denote by x_t the Heathrow temperature series from section 1.2. For data, we can illustrate dependence using scatter plots. As an example, consider a scatter plot of x_{t+6} against x_t, as t varies from 1 to $T - 6$, where T is the length of x_t (see Figure 21). As expected, there is a clear negative dependence between temperatures separated by 6 months, due to seasonality and $\hat{\rho}_6$ is the slope of a regression line estimated for this cloud of points. Similar scatter plots could be created for $k = 1, 2, 3, 4, 5, 7, ...,$.

In Figure 22 we show the sequence $\hat{\rho}_k$ for the Heathrow temperature series. This is the way to show autocorrelations for a time series. Notice e.g., that x_t and x_{t+6} are negatively correlated, while x_t and x_{t+12} are positively correlated (consistent with the yearly temperature cycle). When we regard $x_1, \ldots, x_T$ as a realization of the corresponding random variables $X_1, \ldots, X_T$. The quantity $\hat{\rho}_k$ is an estimate of a corresponding population quantity called the lag k theoretical autocorrelation, defined as

$$\rho_k = \frac{\mathrm{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2}$$

where $\mu = \mathrm{E}[X_t]$ is the population mean, and $\sigma^2 = \mathrm{E}[(X_t - \mu)^2]$ is the corresponding population variance. (Note that ρ_k, μ and σ^2 do not depend on t here. As we shall see soon, models for which this is true play a central role in time series analysis and are called stationary).

2 Stationary stochastic processes

We have seen that a process $\{X_t, t \in \mathbb{Z}\}$, is a collection of random variables, i.e. for fixed t, X_t is a random variable (r.v.), and hence there is an associated cumulative probability distribution function (cdf):

$$F_t(a) = P(X_t \leq a)$$

and we can define its mean and variance

$$\mathrm{E}[X_t] = \int_{-\infty}^{\infty} x \mathrm{d}F_t(x) \equiv \mu_t \qquad \mathrm{Var}(X_t) = \int_{-\infty}^{\infty} (x - \mu_t)^2 \mathrm{d}F_t(x) \equiv \sigma_t^2$$

But in time series analysis we are interested in the relationships between the various r.v.s that form the process. For example, for any t_1 and $t_2 \in \mathbb{Z}$,

$$F_{t_1,t_2}(a_1, a_2) = P(X_{t_1} \leq a_1, X_{t_2} \leq a_2).$$

gives the bivariate cdf.

More generally for any integer $s \geq 1$ and $t_1 \leq t_2 \leq \ldots \leq t_s$, the joint probability distribution function is

$$F_{t_1,t_2,\ldots,t_s}(a_1, a_2, \ldots, a_s) = P(X_{t_1} \leq a_1, X_{t_2} \leq a_2, \ldots, X_{t_s} \leq a_s).$$

Notice that in this course we consider only discrete time stochastic processes, i.e. processes for which the time index t is an integer number. In general we can consider cases where $t \in \mathcal{T}$ for some set of indexes $\mathcal{T} \subset \mathbb{R}$.

The class of all stochastic processes is too large to work with in practice. In the rest of the course we consider only the subclass of stationary processes (later we will discuss also some subclasses of non-stationary processes). We have seen that in practice we will use only second moments to describe a stochastic process and these are related with the notion of weak stationarity. However, we start with a more general definition of stationarity, i.e. strong stationarity. The distinction between strong and weakly stationary processes will be useful when considering models for financial time series.

2.1 Strong stationarity

The process $\{X_t\}$ is said to be strongly stationary if, for any $s \geq 1$, and $t_1 \leq t_2 \leq \ldots \leq t_s$, and all integers k the joint cdf of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_s}\}$ is the same as that of $\{X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_s+k}\}$ i.e.,

$$F_{t_1,t_2,\ldots,t_s}(a_1, a_2, \ldots, a_s) = F_{t_1+k,t_2+k,\ldots,t_s+k}(a_1, a_2, \ldots, a_s),$$

or

$$F_{t_1,t_2,\ldots,t_s}(a_1, a_2, \ldots, a_s) = F_{\tau_1,\tau_2,\ldots,\tau_s}(a_1, a_2, \ldots, a_s),$$

with $\tau_j = t_j + k$.

For example, the probability that X_1 lies between 1 and 2, AND X_2 lies between -3 and 3, and the probability that X_{11} lies between 1 and 2, AND X_{12} lies between -3 and 3, are the same. So that the probabilistic structure of a completely stationary process is invariant under a shift in time.

A strongly stationary process is also said to be completely or strictly stationary.

2.2 Weak stationarity

The process $\{X_t\}$ is said to be weakly stationary if, for any integer $s \geq 1$, and $t_1 \leq t_2 \leq \ldots \leq t_s$, and all integers k, all the joint moments of orders 1 and 2 of $\{X_{t_1}, X_{t_2}, \ldots, X_{t_s}\}$ exist, are finite, and equal to the corresponding joint moments of $\{X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_s+k}\}$, i.e. for any t such that $t_1 \leq t \leq t_s$,

$$\mathrm{E}[X_t] \equiv \mu$$

is a constant independent of t, and for any t_i and t_j such that $t_1 \leq t_i, t_j \leq t_s$

$$\mathrm{E}[X_{t_i} X_{t_j}] = \mathrm{E}[X_{t_i+k} X_{t_j+k}].$$

Hence, if we take $t_i = t_j = t$, we have that

$$\mathrm{Var}(X_t) = \mathrm{E}[X_t^2] - \mu^2 \equiv \sigma^2$$

is a constant independent of t.

Moreover, if we let $k = -t_1$,

$$\mathrm{E}[X_{t_1} X_{t_2}] = \mathrm{E}[X_{t_1+k} X_{t_2+k}] = \mathrm{E}[X_0 X_{t_2-t_1}]$$

and with $k = -t_2$,

$$\mathrm{E}[X_{t_1} X_{t_2}] = \mathrm{E}[X_{t_1+k} X_{t_2+k}] = \mathrm{E}[X_{t_1-t_2} X_0]$$

Hence, $\mathrm{E}[X_{t_1} X_{t_2}]$ is a function of the absolute difference $|t_2 - t_1|$ only. Similarly, for the covariance between X_{t_1} and X_{t_2} we have

$$\mathrm{Cov}(X_{t_1}, X_{t_2}) = \mathrm{E}[(X_{t_1} - \mu)(X_{t_2} - \mu)] = \mathrm{E}[X_{t_1} X_{t_2}] - \mu^2,$$

and therefore is a function of the absolute difference $|t_2 - t_1|$ only.

For a discrete time weakly stationary process $\{X_t\}$ we define the autocovariance sequence (acvs) by

$$\gamma_k \equiv \mathrm{Cov}(X_t, X_{t+k}) = \mathrm{Cov}(X_0, X_k),$$

where

1. k is called the lag;
2. $\gamma_0 = \sigma^2$ and $\gamma_{-k} = \gamma_k$;
3. the autocorrelation sequence (acs) is given by
$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\sigma^2};$$
4. by Cauchy-Schwarz inequality, $|\gamma_k| \leq \gamma_0$ and $|\rho_k| \leq 1$;
5. the variance-covariance matrix of the vector of equispaced X's, $(X_1, X_2, \ldots, X_T)^T$
$$\mathbf{V} = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-2} & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{T-3} & \gamma_{T-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{T-4} & \gamma_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{T-2} & \gamma_{T-3} & \gamma_{T-4} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}$$
has the form which is known as a symmetric Toeplitz matrix - all elements on a diagonal are the same and the matrix has only T unique elements, $\gamma_0, \gamma_1, \ldots, \gamma_{T-1}$;

6. the generic i,j-th element of $\mathbf{V}$ is $\text{Cov}(X_{t_i}, X_{t_j}) = \gamma_{|t_i - t_j|}$;

7. The sequence $\{\gamma_k\}$ is positive semidefinite, i.e., for all integers $s \geq 1$, for any $t_1 \leq t_2 \leq , \ldots, \leq t_s$ and for any set of nonzero real numbers $a_1, a_2, \ldots, a_s$

$$\sum_{i=1}^{s}\sum_{j=1}^{s} \gamma_{|t_i - t_j|} a_i a_j \geq 0. \tag{2}$$

   Let $\mathbf{a} = (a_1, a_2, \ldots, a_s)^T$, $\mathbf{X} = (X_{t_1}, X_{t_2}, \ldots, X_{t_s})^T$ and let $\mathbf{V}$ be the covariance matrix of $\mathbf{X}$ with i,j-th element given by $\gamma_{|t_i - t_j|}$. Define

$$w = \sum_{i=1}^{s} a_i X_{t_i} = \mathbf{a}^T \mathbf{X},$$

   then

$$0 \leq \text{Var}(w) = \text{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T \text{Var}(\mathbf{X})\mathbf{a} = \mathbf{a}^T\mathbf{V}\mathbf{a} = \sum_{i=1}^{s}\sum_{j=1}^{s} \gamma_{|t_i - t_j|} a_i a_j,$$

   which proves (2) and implies that the matrix $\mathbf{V}$ is positive semidefinite.

Other important remarks on stationarity:

1. If $\{X_t\}$ is strongly stationary and has finite second moments, then $\{X_t\}$ is weakly stationary. Of course a weakly stationary process is not necessarily strongly stationary. As an example, let $\gamma_k = 0$ for $k \neq 0$ and $\gamma_0 = 1$ then $\{X_t\}$ is weakly stationary. Moreover, assume that $\{X_t\}$ is Gaussian for $t \neq 0$ and uniform for $t = 0$, then $\{X_t\}$ is not strictly stationary.

2. A stochastic process $\{X_t\}$ is called Gaussian if, for all integers $s \geq 1$ and for any $t_1 \leq t_2 \leq \ldots \leq t_s$, the joint cdf of $X_{t_1}, X_{t_2}, \ldots, X_{t_s}$ is multivariate Gaussian. If $\{X_t\}$ is weakly stationary and Gaussian then it is strongly stationary (since a multivariate Gaussian distribution is completely characterized by 1st and 2nd moments). In this case, the vector $\mathbf{X} = (X_1, X_2, \ldots, X_T)^T$ has joint pdf which reads

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^T \det \mathbf{V}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

   where $\mathbf{x} = (x_1, \ldots, x_T)^T$ are the realizations of $\mathbf{X}$, the matrix $\mathbf{V}$ is defined above (notice that it must be positive definite as we need its determinant and its inverse), and the vector $\boldsymbol{\mu}$ is the vector of means with all entries equal to $\mu = \text{E}[X_i]$ by stationarity.

3. A nonlinear function of a strict stationary variable is still strict stationary, but this is not true for weak stationary. For example, the square of a covariance stationary process may not have finite variance (see the ARCH case below and in Chapter 7).

Unless explicitly stated, stationary processes will be weakly stationary. A weakly stationary process is also said to be second-order or covariance stationary.

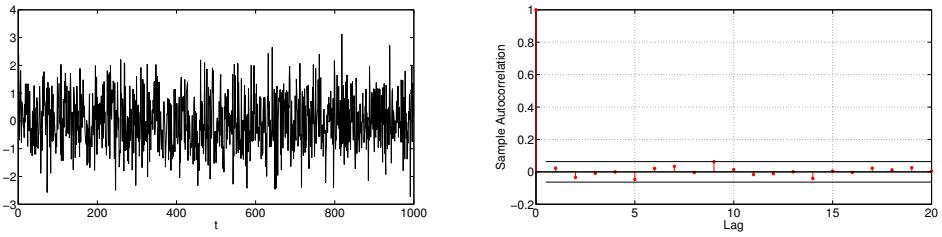


Figure 23: Gaussian white noise.

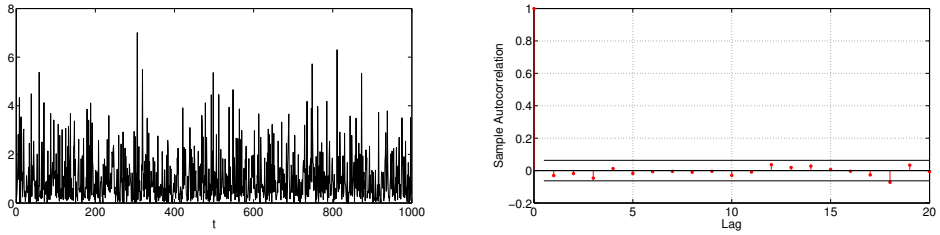


Figure 24: Exponential white noise.

2.3 Examples of stationary processes

In the next examples we will notice that for the weakly stationary processes we have seen the acs decrease quickly (at an exponential rate). In this case we say the process has short memory: a shock to the process has an effect which lasts few periods, in other words we see mean reversion.

1. **White noise process.**
   Also known as a purely random process. Let $\{X_t\}$ be a sequence of uncorrelated r.v.s such that
   $$\mathrm{E}[X_t] = \mu \qquad \mathrm{Var}(X_t) = \sigma^2 \qquad \forall t$$
   and
   $$\gamma_k = \begin{cases} \sigma^2 & \text{if} \quad k = 0 \\ 0 & \text{if} \quad k \neq 0 \end{cases}$$
   This process is a basic building block in time series analysis. An i.i.d. process is of course a white noise. A short way to write that $\{X_t\}$ is white noise is $X_t \sim w.n.(\mu, \sigma^2)$. This notation does not tell us which is the distribution of $\{X_t\}$ and indeed very different realizations of white noise can be obtained for different distributions of $\{X_t\}$. Examples are given in Figures 23 and 24 with their acs.

2. **Moving Average of order q, MA(q).**
   $\{X_t\}$ has realisations given in the form
   $$x_t = \mu + \theta_0 u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q} = \mu + \sum_{j=0}^{q} \theta_j u_{t-j}$$
   where μ and θ_j's are constants (usually we set $\theta_0 \equiv 1$ and $\theta_q \neq 0$), and $\{u_t\}$ is a zero-mean white noise process with variance σ_u^2. Without loss of generality we can assume $\mathrm{E}[X_t] = \mu = 0$. Then $\mathrm{Cov}(X_t, X_{t+k}) = \mathrm{E}[X_t X_{t+k}]$.

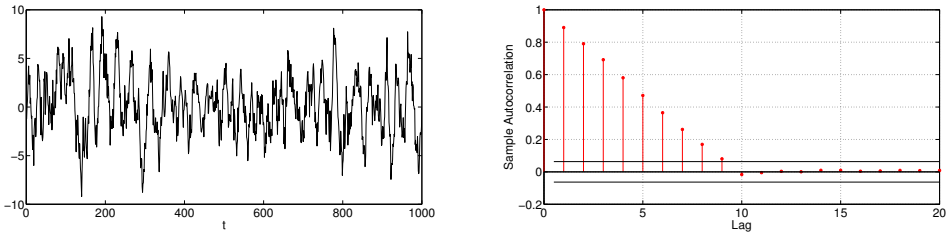


Figure 25: MA(9) with coefficients $\theta_j = 1$, $j = 1, \ldots, 9$.

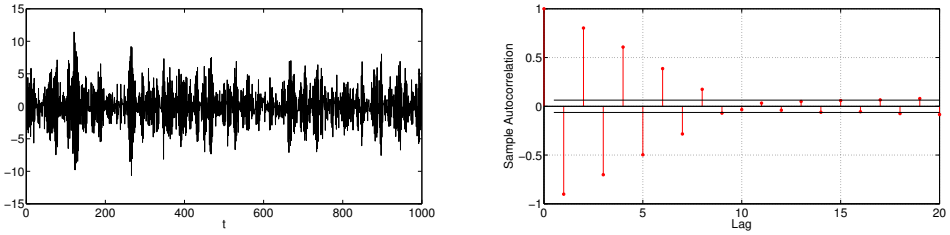


Figure 26: MA(9) with coefficients $\theta_j = (-1)^j$, $j = 1, \ldots, 9$.

Since $\mathrm{E}[u_t u_{t+k}] = 0$ for all $k \neq 0$ we have for $k \geq 0$

$$\gamma_k^x = \mathrm{Cov}(X_t, X_{t+k}) = \sum_{i=0}^{q} \sum_{j=0}^{q} \theta_i \theta_j \mathrm{E}[u_{t-i} u_{t+k-j}] = \sigma_u^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k}$$

as the only nonzero terms are when $j = i + k$. Therefore, γ_k^x does not depend on t. Since the above must hold also for $k < 0$, that is $\gamma_k^x = \gamma_{-k}^x$, then we must use $|k|$ and $\{X_t\}$ is a stationary process with acvs given by

$$\gamma_k^x = \begin{cases} \sigma_u^2 \sum_{i=0}^{q-|k|} \theta_i \theta_{i+|k|} & \text{if} \quad |k| \leq q \\ 0 & \text{if} \quad |k| > q \end{cases}$$

Notice that no restrictions are placed on the θ_j's to ensure stationarity except obviously, $|\theta_j| < \infty$ for all j. Examples are given in Figures 25 and 26 with their acs.

Consider an MA(1)

$$x_t = u_t + \theta_1 u_{t-1}$$

the acvs are

$$\gamma_k^x = \sigma_u^2 \sum_{j=0}^{1-|k|} \theta_j \theta_{j+|k|} \qquad |k| \leq 1$$

and (notice that we usually set $\theta_0 = 1$)

$$\begin{aligned} \gamma_0^x &= \sigma_u^2(\theta_0^2 + \theta_1^2) = \sigma_u^2(1 + \theta_1^2) \\ \gamma_1^x &= \sigma_u^2 \theta_0 \theta_1 = \sigma_u^2 \theta_1 \end{aligned}$$

while the acs are

$$\rho_0^x = 1 \qquad \rho_1^x = \frac{\theta_1}{1 + \theta_1^2}$$

a) if $\theta_1 = 1$ and $\sigma_u^2 = 1$ we have

$$\begin{array}{lll} \gamma_0^x = 2 & \gamma_1^x = 1 & \gamma_2^x = \gamma_3^x = \ldots = 0 \\ \rho_0^x = 1 & \rho_1^x = 0.5 & \rho_2^x = \rho_3^x = \ldots = 0 \end{array}$$

b) if $\theta_1 = -1$ and $\sigma_u^2 = 1$ we have

$$\begin{array}{lll} \gamma_0^x = 2 & \gamma_1^x = -1 & \gamma_2^x = \gamma_3^x = \ldots = 0 \\ \rho_0^x = 1 & \rho_1^x = -0.5 & \rho_2^x = \rho_3^x = \ldots = 0 \end{array}$$

Finally, notice that if we replace θ_1 by θ_1^{-1} the model becomes

$$x_t = u_t + \frac{1}{\theta_1} u_{t-1}$$

and the acs is

$$\rho_1^x = \frac{\frac{1}{\theta_1}}{1 + \frac{1}{\theta_1^2}} = \frac{\theta_1}{1 + \theta_1^2}$$

hence it is unchanged. We cannot identify an MA(1) from its acs.

We can construct a moving average with any process $\{X_t\}$:

$$y_t = \sum_{j=0}^{q} \theta_j x_{t-j}.$$

If $\{X_t\}$ is stationary then $\{Y_t\}$ is stationary. Indeed, $\mathrm{E}[Y_t] = \mathrm{E}[X_t] \sum_{j=0}^{q} \theta_j$, moreover the acvs of $\{Y_t\}$ can be computed from the Toepliz matrix of the acvs of $\{X_t\}$

$$\gamma_k^y = \mathrm{E}[Y_t Y_{t+k}] = \boldsymbol{\theta}^T \begin{bmatrix} \gamma_k^x & \gamma_{k-1}^x & \gamma_{k-2}^x & \cdots & \gamma_{k-q+1}^x & \gamma_{k-q}^x \\ \gamma_{k+1}^x & \gamma_k^x & \gamma_{k-1}^x & \cdots & \gamma_{k-q+2}^x & \gamma_{k-q+1}^x \\ \gamma_{k+2}^x & \gamma_{k+1}^x & \gamma_k^x & \cdots & \gamma_{k-q+3}^x & \gamma_{k-q+2}^x \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{k-q+1}^x & \gamma_{k-q+2}^x & \gamma_{k-q+3}^x & \cdots & \gamma_k^x & \gamma_{k-1}^x \\ \gamma_{k-q}^x & \gamma_{k-q+1}^x & \gamma_{k-q+2}^x & \cdots & \gamma_{k+1}^x & \gamma_k^x \end{bmatrix} \boldsymbol{\theta}$$

which holds for $|k| \leq q$ and $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_q)^T$.

3. **AutoRegression of order p, AR(p).**
$\{X_t\}$ has realisations given in the form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + u_t = \sum_{j=1}^{p} \phi_j x_{t-j} + u_t$$

where ϕ_j's are constants ($\phi_p \neq 0$), and $\{u_t\}$ is a zero-mean white noise process with variance σ_u^2. In contrast to the parameters of an MA(q) process, the $\{\phi_j\}$s must satisfy certain conditions for $\{X_t\}$ to be a stationary process, i.e., not all AR(p) processes are stationary. We discuss later the conditions for stationarity and the general form of acvs while here we consider one example. Examples of a stationary AR(2) are given in Figures 13 and 28 with their acs.

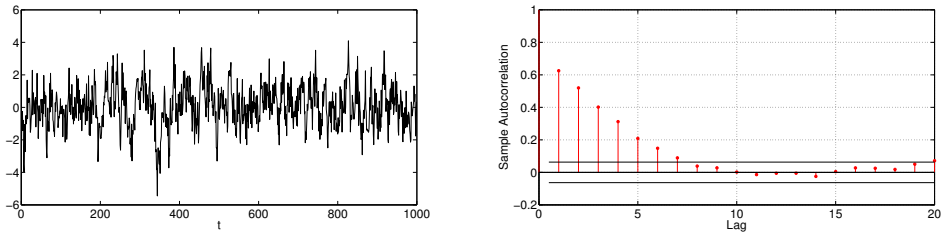


Figure 27: AR(2) with coefficients $\phi_1 = 0.5$ and $\phi_2 = 0.2$.

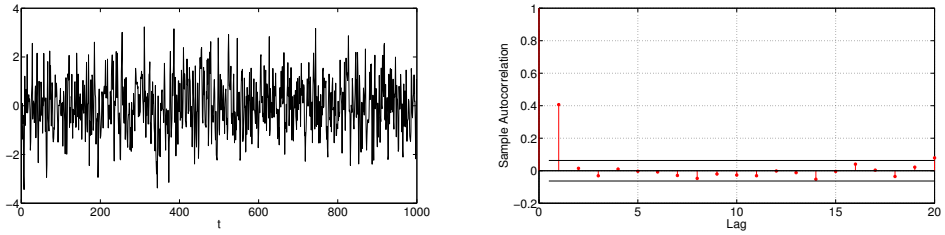


Figure 28: AR(2) with coefficients $\phi_1 = 0.5$ and $\phi_2 = -0.2$.

Consider an AR(1)

$$\begin{aligned} x_t &= \phi_1 x_{t-1} + u_t \\ &= \phi_1(\phi_1 x_{t-2} + u_{t-1}) + u_t \\ &= \phi_1^2 x_{t-2} + \phi_1 u_{t-1} + u_t \\ &= \phi_1^3 x_{t-3} + \phi_1^2 u_{t-2} + \phi_1 u_{t-1} + u_t \\ &\vdots \\ &= \sum_{k=0}^{\infty} \phi_1^k u_{t-k} \end{aligned}$$

with initial condition $x_{-T} = 0$ and we let $T \to \infty$. If we find conditions under which this process is stationary then it is an MA(∞) (MA process are always stationary).

We have (recall that u_t is white noise)

$$\begin{aligned} \mathrm{E}[X_t] &= 0 \\ \mathrm{Var}(X_t) &= \mathrm{Var}\left(\sum_{k=0}^{\infty} \phi_1^k u_{t-k}\right) = \sum_{k=0}^{\infty} \mathrm{Var}\left(\phi_1^k u_{t-k}\right) = \sigma_u^2 \sum_{k=0}^{\infty} \phi_1^{2k} \end{aligned}$$

In order to have $\mathrm{Var}(X_t) < \infty$ we must have $\phi_1^2 < 1$ or, equivalently, $|\phi_1| < 1$, in which case (see the geometric series)

$$\mathrm{Var}(X_t) = \frac{\sigma_u^2}{1 - \phi_1^2}$$

and the process is then stationary and can be written as an MA(∞).[1]

To find the form of the acvs, we notice that for $k > 0$, X_{t-k} is a linear function of $u_{t-k}, u_{t-k-1}, \ldots$ therefore it is uncorrelated with u_t (white noise):

$$\mathrm{E}[X_{t-k} u_t] = 0,$$

[1] We can ask ourselves if an MA(1) can be written as an AR(∞). The answer is yes provided that the MA parameter θ_1 is such that $|\theta_1| < 1$. This is the condition for invertibility of an MA(1), if violated the MA(1) is still stationary but it cannot be written as an AR model. The proof is shown in Chapter 4.

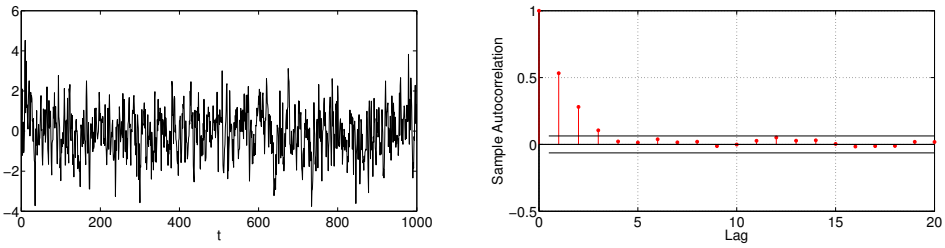


Figure 29: AR(1) with coefficient $\phi_1 = 0.5$.

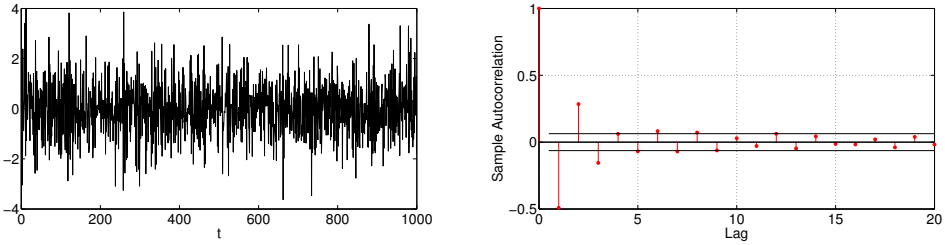


Figure 30: AR(1) with coefficient $\phi_1 = -0.5$.

so, assuming stationarity, multiplying by X_{t-k} and taking expectations we have

$$\mathrm{E}[X_t X_{t-k}] = \phi_1 \mathrm{E}[X_{t-1} X_{t-k}] + \mathrm{E}[u_t X_{t-k}] = \phi_1 \mathrm{E}[X_{t-1} X_{t-k}]$$

i.e.

$$\gamma_k^x = \phi_1 \gamma_{k-1}^x = \phi_1^2 \gamma_{k-2}^x = \ldots = \phi_1^k \gamma_0^x$$

and the acs is

$$\rho_k^x = \phi_1^k.$$

But ρ_k^x must be an even function of k, i.e. the above must hold also for $k < 0$, so

$$\rho_k^x = \phi_1^{|k|} \qquad k = 0, \pm 1, \pm 2, \pm 3, \ldots$$

it has an exponential decay (see Figures 12 and 30).

4. **AutoRegression Moving Average of orders p, q, ARMA(p, q).**
$\{X_t\}$ has realisations given in the form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q}$$

where the ϕ_j's and the θ_j's are all constants ($\phi_p \neq 0, \theta_q \neq 0$) and $\{u_t\}$ is a zero-mean white noise process with variance σ_u^2. Once again we usually set $\theta_0 = 1$.

The ARMA class is important as many data sets may be approximated in a more parsimonious way (meaning fewer parameters are needed) by a mixed ARMA model than by a pure AR or MA process. In brief, time series analysts like MA and AR models for different reasons. MA models are appealing because they are easy to manipulate mathematically, e.g. as we saw, no restrictions on parameter values are needed to ensure stationarity. On the other hand, AR models are more convenient for forecasting, as we will see later. Obviously, the main criterion for whether a model is or isn't useful is whether it performs well at our desired task, which will often be: modelling (or understanding) the data, forecasting, or control. The ARMA model shares the best, and the worst, features of the AR and MA classes and will be the subject of next Chapters.

5. **Moving Average of infinite order, MA(∞).**
We can in principle consider a moving average of a white noise with an infinite number of terms involving both future and past values:

$$x_t = \ldots + a_{-2}u_{t+2} + a_{-1}u_{t+1} + a_0 u_t + a_1 u_{t-1} + a_2 u_{t-2} + \ldots = \sum_{j=-\infty}^{\infty} a_j u_{t-j}$$

which has finite variance and acvs provided that

$$\sum_{j=-\infty}^{\infty} a_j^2 < \infty. \tag{3}$$

Indeed, for an MA we have

$$\gamma_0^x = \sigma_u^2 \sum_{j=-\infty}^{\infty} a_j^2$$

which is finite if condition (3) holds and the lag k acvs is

$$\gamma_k^x = \sigma_u^2 \sum_{j=-\infty}^{\infty} a_j a_{j-k}.$$

By Cauchy Schwarz inequality[2]

$$|\gamma_k^x| \leq \gamma_0^x$$

which is finite if condition (3) holds.

If for example we take

$$\begin{aligned} a_0 &= 1 \\ a_j &= 0 \qquad j < 0 \\ a_{-\infty} &= 0 \\ a_j &= \phi_1^j \qquad j > 0 \end{aligned}$$

Then

$$x_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \ldots$$

which is the MA(∞) representation of an AR(1) seen above. Condition (3) for stationarity becomes

$$\sum_{j=0}^{\infty} \phi_1^{2j} < \infty$$

which implies $|\phi_1| < 1$ as expected for an AR(1) process.

6. **AutoRegressive Conditionally Heteroscedastic model of order p, ARCH(p).**
Assume we have a time series $\{Y_t\}$ that is (approximately) uncorrelated (as a white noise), is stationary, but has a multiplicative component σ_t that changes through time,

$$y_t = \sigma_t u_t \tag{4}$$

[2]The Cauchy Schwarz inequality is

$$\left|\sum_j c_j d_j\right| \leq \left(\sqrt{\sum_j c_j^2}\right)\left(\sqrt{\sum_j d_j^2}\right)$$

then choose $c_j = a_j$ and $d_j = a_{j-k}$ and we have the result.

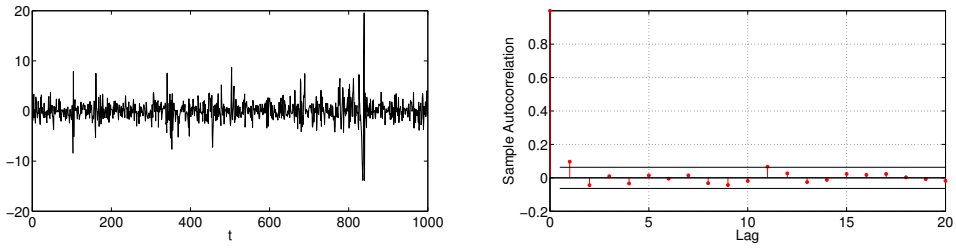


Figure 31: ARCH(1) with coefficient $\alpha_1 = 0.8$.

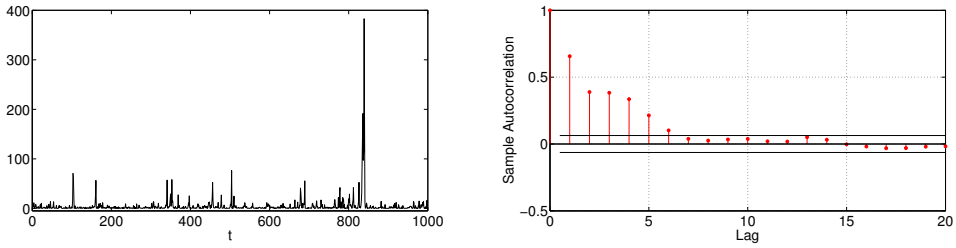


Figure 32: Squares of ARCH(1) with coefficient $\alpha_1 = 0.8$.

where $\{u_t\}$ is a white noise sequence with zero-mean and unit variance. Here, σ_t represents the local conditional standard deviation of the process and is called volatility and denotes changes in the scale of the process at time t. Note that σ_t is not observable directly.

$\{Y_t\}$ is ARCH(p) if its realisations satisfy equation (4) and

$$\sigma_t^2 = \omega + \alpha_1 y_{t-1}^2 + \ldots + \alpha_p y_{t-p}^2 \tag{5}$$

where $\omega > 0$ and $\alpha_j \geq 0$ (to ensure the variance remains positive), and y_{t-1} is the observed value of the time series at time $(t-1)$.

Notice that:

(a) there is no error term in equation (5);

(b) unconstrained estimation often leads to violation of the non-negativity constraints that are needed to ensure positive variance;

(c) quadratic form (i.e. modelling σ_t^2) prevents modelling of asymmetry in volatility (i.e. volatility tends to be higher after a decrease than after an equal increase of y_t and ARCH cannot account for this).

Consider an ARCH(1)

$$\sigma_t^2 = \omega + \alpha_1 y_{t-1}^2$$

define $v_t = y_t^2 - \sigma_t^2$ then $\sigma_t^2 = y_t^2 - v_t$. Then the model can also be written:

$$y_t^2 = \omega + \alpha_1 y_{t-1}^2 + v_t$$

i.e. an AR(1) model for $\{Y_t^2\}$ where the errors, $\{v_t\}$, have zero-mean, but since $v_t = \sigma_t^2(u_t^2 - 1)$ the errors are heteroscedastic (changing variance). An example of an ARCH(1) is in Figures 31 and 32 for $\{Y_t\}$ and $\{Y_t^2\}$ with their acs. This shows that $\{Y_t\}$ is a white noise with observations which are uncorrelated but not independent since their squares are autocorrelated, i.e. in general $\mathrm{E}[Y_t^2 Y_{t-k}^2] \neq 0$.

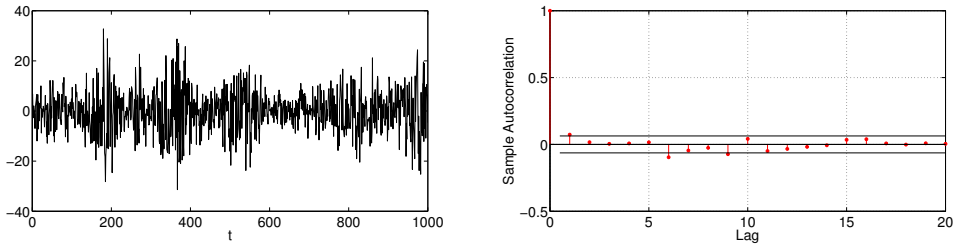


Figure 33: GARCH(1,1) with coefficients $\alpha_1 = 0.1$ and $\beta_1 = 0.89$.

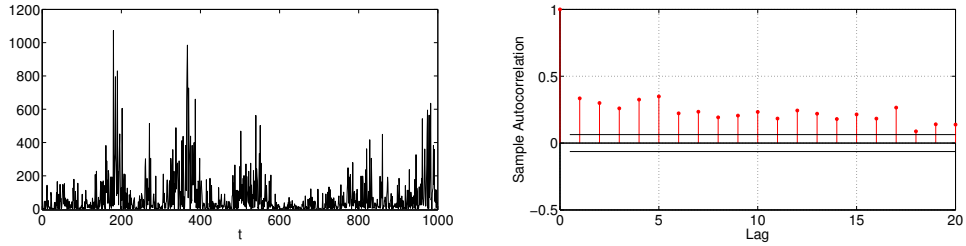


Figure 34: Squares of GARCH(1,1) with coefficients $\alpha_1 = 0.1$ and $\beta_1 = 0.89$.

7. **Generalised AutoRegressive Conditionally Heteroscedastic model of orders p, q, GARCH(p, q).** $\{Y_t\}$ has realizations satisfying

$$\begin{aligned} y_t &= \sigma_t u_t \\ \sigma_t^2 &= \omega + \alpha_1 y_{t-1}^2 + \ldots + \alpha_p y_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2 \end{aligned}$$

where the parameters are chosen to ensure positive variance, that is $\omega > 0$, $\alpha_j + \beta_j < 1$, and $\alpha_j \geq 0$, $\beta_j \geq 0$, and $\{u_t\}$ is a zero-mean white noise with unit variance and therefore also $\{Y_t\}$ is white noise.

GARCH models were introduced because it was observed that the ARCH class does not account sufficiently well for the persistence of volatility in financial time series data; i.e. according to the ARCH model, the process $\{Y_t^2\}$ often has less (theoretical) autocorrelation than real data tend to have in practice.

Consider a GARCH(1,1)

$$y_t = \sigma_t u_t, \qquad \sigma_t^2 = \omega + \alpha_1 y_{t-1}^2 + \sigma_{t-1}^2$$

An important measure in this model is the persistence which is given by $\alpha_1 + \beta_1$. A GARCH with high persistence might seem a non-stationary process, but in fact it is a weakly stationary process, indeed its variance is (we will prove this later)

$$\text{Var}(Y_t) = \frac{\omega}{1 - \alpha_1 - \beta_1}$$

which is positive and finite if and only if $\alpha + \beta < 1$. An example of an GARCH(1,1) is in Figures 33 and 34 for $\{Y_t\}$ and $\{Y_t^2\}$ with their acs.

2.4 The Lag or Backshift operator

If $\{X_t\}$ is a stochastic process, we define the lag operator, denoted by L, by

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

Sometimes it is used also the notation $Bx_t = x_{t-1}$.

Moreover, for $k > 1$,

$$L^k x_t = L(L^{k-1}x_t) = x_{t-k}$$

for example $L^2 x_t = x_{t-2}$. Moreover, $L^0 = 1$ where 1 here means the identity operator: $1x_t = x_t$.

Lastly define $F \equiv L^{-1}$ then

$$Fx_t = L^{-1}x_t = x_{t+1}.$$

indeed $FLx_t = x_t$.

We also define polynomials of L,

$$a(L) = a_{-m}L^{-m} + \ldots + a_{-1}L^{-1} + a_0 + a_1 L + \ldots + a_m L^m = \sum_{j=-m}^{m} a_j L^j$$

Moving averages can then be rewritten as:

$$x_t = a(L)u_t = \left(\sum_{j=-m}^{m} a_j L^j \right) u_t = a_{-m}u_{t+m} + \ldots + a_{-1}u_{t+1} + a_0 u_t + a_1 u_{t-1} + \ldots + a_m u_{t-m}.$$

An ARMA can be written as having realizations which solve the stochastic difference equation

$$\Phi(L)x_t = \Theta(L)u_t,$$

Where

$$\begin{aligned} \Phi(L) &= 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p \\ \Theta(L) &= 1 + \theta_1 L + \theta_2 L^2 - \ldots - \theta_q L^q \end{aligned}$$

are known as the associated or characteristic polynomials.

Notice also that first differences can be written using this notation:

$$\Delta x_t = x_t - x_{t-1} = (1 - L)x_t$$

therefore $\Delta \equiv (1 - L)$. Second differences are then easily computed

$$\Delta\Delta x_t = \Delta^2 x_t = (1-L)^2 x_t = (1+L^2-2L)x_t = x_t + x_{t-2} - 2x_{t-1} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}).$$

Notice that this is different from the difference among two non consecutive periods as e.g. when computing quarterly differences which are given by (if the frequency of the data is monthly)

$$\Delta_4 x_t = (1 - L^4)x_t = x_t - x_{t-4}.$$

Further, we can generalize the class of ARMA models to include differencing to account for certain types of non-stationarity (see below), namely, $\{X_t\}$ is called ARIMA(p, d, q) if it has realizations such that

$$\begin{aligned} \Phi(L)(1 - L)^d x_t &= \Theta(L)u_t, \\ \Phi(L)\Delta^d x_t &= \Theta(L)u_t. \end{aligned}$$

2.5 Trend removal and seasonal adjustment - part 2

We have seen that, in general, a process $\{X_t\}$ is made of three components:

$$X_t = T_t + S_t + C_t$$

the trend is a tendency to increase or decrease slowly steadily over time while seasonaity is given by periodic fluctuations due to seasonal effects (e.g. sales in turistic cities are higher during holidays). The trend T_t is a non-stationary component and can usually be due to mean and/or variance which change over time (higher moments are of less interest for us).

The simplest case which we consider here is the case of a trend which results in changing mean, then

$$X_t = \mu_t + Y_t$$

where $T_t \equiv \mu_t$ is the time dependent mean and $Y_t \equiv S_t + C_t$ is a zero-mean stationary process. In this case the process $\{X_t\}$ is called Trend Stationary. Later we will consider another type of non-stationarity where $\{X_t\}$ is Difference Stationary, i.e. it is stationary once we take its first differences. Such processes have also time varying variance (see the example of the random walk in Chapter 1).

As an example consider the temperature data, last 30 years. The data are plotted in Chapter 1 Figure 10 and the model suggested by plot is: $X_t = \alpha + \beta_t + Y_t$, where Y_t has a seasonal component of period 12 months.

2.5.1 Trend adjustment

Consider the case of a linear trend plus a seasonal component

$$X_t = \alpha + \beta t + S_t + C_t$$

where C_t is a Gaussian white noise with zero-mean and unit variance, and $S_t = A\cos(\pi t/2)$, with $A \sim N(0,1)$. The simulated data is in Figure 35.

There are at least two possible approaches for controlling for the trend.

1. Estimate α and β by least squares, and work with the residuals

   $$\hat{Y}_t = X_t - \hat{\alpha} - \hat{\beta} t.$$

   For the simulated data these are shown in Figure 36 left.

2. Take first differences:

   $$\Delta X_t = X_t - X_{t-1} = \alpha + \beta t + Y_t - (\alpha + \beta(t-1) + Y_{t-1}) = \beta + Y_t - Y_{t-1}.$$

   Notice that if $\{Y_t\}$ is stationary so is $\{\Delta X_t\}$. As the name suggests this approach controls also for the other kind of non-stationarity we will consider later, i.e. the case of Difference Stationary processes. Indeed, it is enough to have $\{\Delta Y_t\}$ stationary for having $\{\Delta X_t\}$ stationary, so even if $\{Y_t\}$ is non-stationary we might obtain a stationary process by taking first differences. In the case of a linear trend, first differences leave the constant β, there are two ways to remove it.

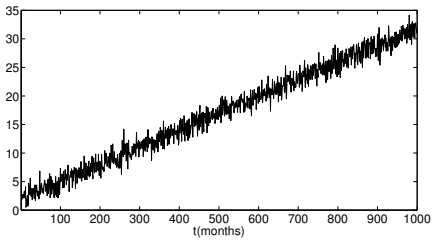


Figure 35: Simulated data from the model $X_t = 2 + 0.03t + A\cos(\pi t/2) + C_t$, with $A \sim N(0,1)$ and $C_t \sim w.n.N(0,1)$.

(a) If we difference twice:

$$(1-L)^2 X_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = (\beta + Y_t - Y_{t-1}) - (\beta + Y_{t-1} - Y_{t-2}) = (1-L)^2 Y_t$$

so that the effect of μ_t has been completely removed. In general, if μ_t is a polynomial of degree $(d-1)$ in t, then d-th differences of μ_t will be zero ($d = 2$ for linear trend). Then,

$$(1-L)^d X_t = \Delta^d X_t = \Delta^d Y_t.$$

(b) Alternatively, β can be removed by demeaning $\{\Delta X_t\}$, i.e. by subtracting its mean

$$\Delta X_t - \mathrm{E}[\Delta X_t] = \beta + Y_t - Y_{t-1} - \beta - \mathrm{E}[Y_t] + \mathrm{E}[Y_{t-1}] = Y_t - Y_{t-1} = \Delta Y_t,$$

since $\{Y_t\}$ is stationary and therefore its mean does not change over time. In practice, we have to remove the sample mean of $\{\Delta X_t\}$

$$\Delta X_t - \overline{\Delta X}_t = \beta + Y_t - Y_{t-1} - \beta - \overline{Y}_t + \overline{Y}_{t-1} = \widehat{\Delta Y}_t.$$

It can be proved that asymptotically, this approach is a consistent one, that is $\{\widehat{\Delta Y}_t\}$ is close to $\{\Delta Y_t\}$ provided that we have a sample large enough. For the simulated data $\{\widehat{\Delta Y}_t\}$ is shown in Figure 36 right.

Taking first differences and then demeaning using the sample mean or taking second differences or detrending via OLS are all equivalent ways for removing completely the effects of a linear trend. However, these methods are not equivalent in terms of the process we are left with, as when detrending with OLS we are left with the process $\{\hat{Y}_t\}$, when taking first differences and then demeaning we are left with $\{\widehat{\Delta Y}_t\}$, while when taking second differences we are left with $\{\Delta^2 Y_t\}$. The latter approach has the advantage that no estimation is required, however the more "differences" we take the more information we lose. Notice also that by comparing the acs of $\{\hat{Y}_t\}$ and $\{\widehat{\Delta Y}_t\}$ (see bottom panel of Figure 36) we see that in the second case autocorrelation at lag 1 is induced.

If we are sure that the process is Trend Stationary then OLS should be preferred. On the other hand as we will see later it might be hard to determine if a process is Trend Stationary or Difference Stationary, in which case taking first differences is the right thing to do. Once first differences are taken, taking second differences is in general not a good idea (unless what we are left with is still Difference Stationary) and it is preferable to demean the differenced process $\{\Delta X_t\}$.

In particular, if we use OLS we immediately have the detrended component of $\{X_t\}$ as $\hat{Y}_t$, while if we detrended via first differences and demeaning such component cannot be recovered.

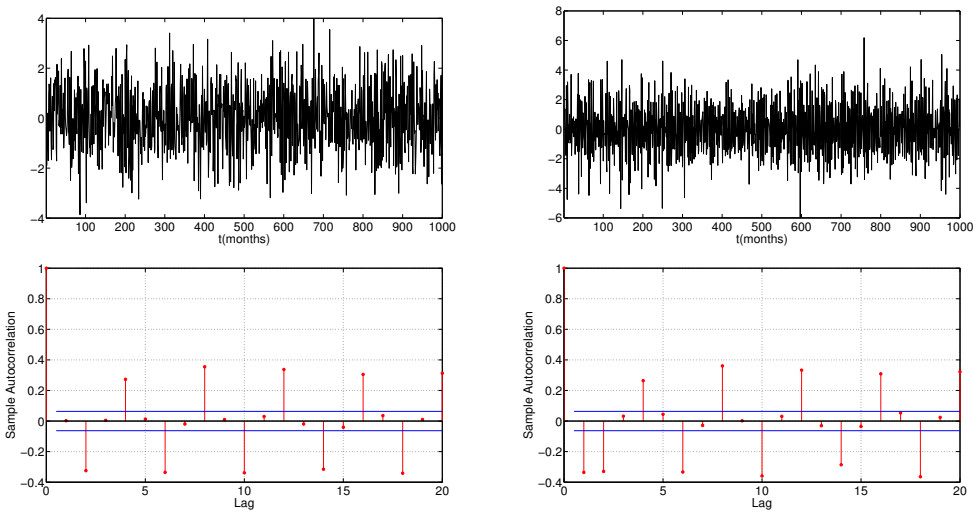


Figure 36: Detrended simulated data. Top left: OLS detrending. Top right: first differences demeaned. Bottom panel: corresponding acs.

Indeed, if we try to cumulate first differences (as integration is the inverse of differentiation)

$$\tilde{Y}_t = \sum_{s=1}^{t} \widehat{\Delta Y}_s$$

we are artificially introducing a new linear trend in the process $\{\tilde{Y}_t\}$, i.e. a line connecting the last to the first observation of $\{Y_t\}$, therefore $\{\tilde{Y}_t\}$ will have a time varying mean due to the linear trend and it is more similar to $\{X_t\}$ rather than to $\{\hat{Y}_t\}$. A similar reasoning applies if we used second differences to detrend and then cumulate twice

$$\check{Y}_t = \sum_{s=1}^{t} \sum_{\tau=1}^{s} \Delta^2 Y_\tau.$$

2.5.2 Seasonal adjustment

Once the trend is removed the model is modified to

$$Y_t = S_t + C_t,$$

where Y_t can be obtained by means of OLS detrending, S_t is the seasonal component, C_t is the zero-mean stationary process (sometime called cycle). If we use first differences the model to be considered after detrending is

$$\Delta X_t = \beta + \Delta Y_t = \beta + \Delta S_t + \Delta C_t,$$

where β is nothing else but the mean of ΔX_t. From the acs in the bottom panel of Figure 36 we see that in both cases we have a seasonal component with period 4.[3] Notice that after removing the

[3]Clearly, if we used second differences to detrend then the model to consider is

$$\Delta^2 X_t = \Delta^2 Y_t = \Delta^2 S_t + \Delta^2 C_t.$$

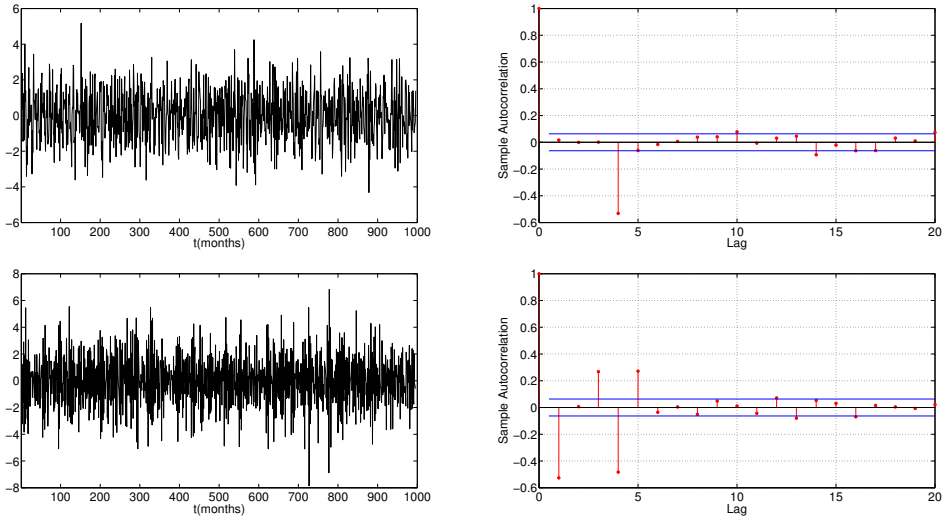


Figure 37: Detrended and deseasonlized simulated data. Top: OLS detrending. Bottom: first differences demeaned.

trend $\{Y_t\}$ is already stationary if S_t is stationary (as in our simulated model) but if S_t is a purely deterministic function of time then it has a time varying mean (it can also be seen that in any case the acvs of S_t do not decrease to zero, therefore they are not summable, a condition needed for estimation, see Chapter 3).

Presuming that the seasonal component maintains a constant pattern over time with period s, i.e. $S_t = S_{t-s}$, there are again several approaches for removing S_t. A popular approach is to use the operator $(1 - L^s)$:

$$(1 - L^s)Y_t = \Delta_s Y_t = Y_t - Y_{t-s} = (S_t + C_t) - (S_{t-s} + C_{t-s}) = C_t - C_{t-s}.$$

This holds if we use OLS to detrend. If we opt for first differences then we remove trend and seasonality by applying the transformation

$$(1 - L^s)(1 - L)X_t = (1 - L^s)(\beta + \Delta S_t + \Delta C_t) = \Delta C_t - \Delta C_{t-s}.$$

Notice that this removes the effect of the linear trend without the need of second differences.

An example with the simulated data for both ways of detrending and $s = 4$ is in Figure 37. Notice that the model was simulated in such a way that C_t is a white noise. However, even when we use OLS for detrending we do not recover $\{C_t\}$ but $\{\Delta_4 C_t\}$ which is autocorrelated at lag 4 by construction. This is clear from inspection of the top left panel in Figure 37.

An alternative way to deseasonalize that allows to recover a white noise is by means of OLS once again. We can regress the process $\{\hat{Y}_t\}$ on dummy variables $d_{1t}, d_{2t}, d_{3t}, d_{4t}$ such that $d_{it} = d_{it+4} = 1$ and $\sum_{i=1}^{4} d_{it} = 1$ for any t.

$$\hat{Y}_t = \gamma_1 d_{1t} + \gamma_2 d_{2t} + \gamma_3 d_{3t} + \gamma_4 d_{4t} + e_t$$

since $\{D_t = (d_{1t}\, d_{2t}\, d_{3t}\, d_{4t})\}$ has the same periodicity as S_t then the residual $\{e_t\}$ is highly correlated with the non-periodic component $\{C_t\}$. The residual process obtained from this regression is then white noise as shown in Figure 38. This approach is valid also if we start from $\{\Delta Y_t\}$ but of course we should not get a white noise as result, since $\{\Delta C_t\}$ is not a white noise.

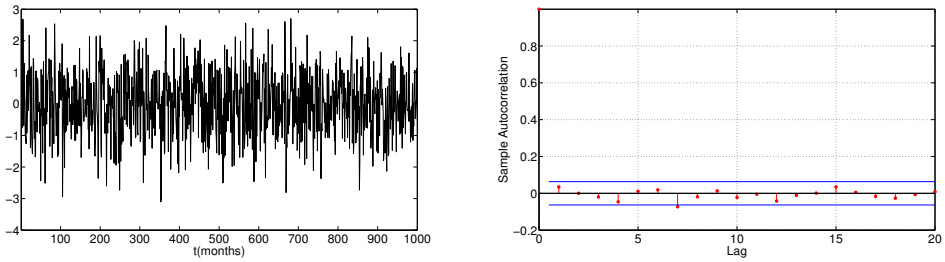


Figure 38: Detrended and deseasonlized simulated data. OLS detrending and dummy deseasonalization.

3 The theory of linear processes

We have seen stationary processes, in particular weakly stationary processes and the MA(q) process

$$x_t = u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q}$$

where u_t is a zero mean white noise process with variance σ_u^2. We know that $\{X_t\}$ is stationary and that its acvs γ_k^x is zero whenever $|k| > q$. The converse is also true, i.e. if $\{X_t\}$ is a process with $\gamma_q^x \neq 0$ but with $\gamma_k^x = 0$ if $|k| > q$, then $\{X_t\}$ is stationary and can be represented as an MA(q) process.

We now generalise and prove this result for generic stationary processes and the MA(∞) case. We say that the process $\{X_t\}$ is a linear process if it can be written as

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j u_{t-j} = \Psi(L) u_t$$

where u_t is a zero mean white noise process with variance σ_u^2 and $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$. Thus $\{X_t\}$ is written as an MA(∞), and $\Psi(L)$ is a linear filter.[4]

In particular, we require square-summability for two reasons

1. The infinite sum converges with probability 1 to $\{X_t\}$. Indeed, we can always write

$$x_t = \sum_{j=-h}^{h} \psi_j u_{t-j} + \sum_{j=-\infty}^{h-1} \psi_j u_{t-j} + \sum_{j=h+1}^{\infty} \psi_j u_{t-j}$$

Moreover, a necessary condition for square-summability is $\psi_j \to 0$ as $j \to \infty$. Then,

$$\lim_{h\to\infty} \mathrm{E}\left[\left(X_t - \sum_{j=-h}^{h} \psi_j u_{t-j}\right)^2\right] = \lim_{h\to\infty} 2 \sum_{j=h+1}^{\infty} \psi_j^2 \sigma_u^2 = 0.$$

Therefore, we have convergence in mean-square which implies convergence in probability by Chebychev's inequality.

[4] A linear filter is an operator like $\Psi(L)$ that transforms linearly a given time series into a new (filtered) one. Then in the definition above the realizations x_t are obtained by filtering u_t.

2. For any k the acvs are

$$|\gamma_k^x| \leq \gamma_0^x = \sigma_u^2 \sum_{i=-\infty}^{\infty} \psi_i^2 < \infty,$$

which ensures stationarity.

Thus linear processes are stationary. Indeed, a linear filter when applied to any stationary time series produces a stationary time series (see the case of moving averages of stationary processes in Section 2.3).

In this Chapter, we will prove that every weakly stationary process is either a linear process or can be transformed to a linear process by subtracting a deterministic component. This is the Wold representation theorem proved below.

3.1 ARMA as linear processes

First, let us consider stationary ARMA. The following hold.

1. They can always be written as MA(∞)

$$x_t = \sum_{j=0}^{\infty} \psi_j u_{t-j} = \Psi(L) u_t$$

with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ (for the MA(q) this is trivial while for the AR(p) it requires more calculations, an example is the AR(1) case with parameter ϕ such that $|\phi| < 1$, then $\psi_j = \phi^j$, see also the next chapter for the general ARMA).

2. Since to have absolute summability a necessary condition is $\psi_j \to 0$ as $j \to \infty$, then

$$\sum_{j=-\infty}^{\infty} \psi_j^2 < \sum_{j=-\infty}^{\infty} |\psi_j| < \infty,$$

i.e. absolute summability implies also square summability of the coefficients.

3. We can prove again that the infinite sum converges with probability 1 to $\{X_t\}$ but without using square summability. Indeed, since[5] $\mathrm{E}[|u_t|] \leq \sigma_u$

$$\mathrm{E}[|X_t|] = \mathrm{E}\left[\left|\sum_{j=-\infty}^{\infty} \psi_j u_{t-j}\right|\right] \leq \sum_{j=-\infty}^{\infty} |\psi_j| \mathrm{E}[|u_{t-j}|] \leq \left(\sum_{j=-\infty}^{\infty} |\psi_j|\right) \sigma_u < \infty$$

then by Markov's inequality, for any finite constant $M > 0$ we have

$$\mathrm{Prob}(|X_t| > M) \leq \frac{\mathrm{E}[|X_t|]}{M} < \infty.$$

Therefore with probability 1 the series does not diverge.

[5] Use the Cauchy-Schwarz inequality

$$\mathrm{E}[|XY|] \leq (\mathrm{E}[X^2])^{1/2} (\mathrm{E}[Y^2])^{1/2}$$

Because of the previous two properties, stationary ARMA processes are a subclass of linear processes.[6] In particular in ARMA we use only past values of u_t and the coefficients decrease at a faster rate (exponential) than a generic linear process.

By comparing the properties of stationary ARMA with those of a generic linear process, we see that while square-summability is necessary and sufficient for stationarity, absolute summability is just sufficient but not necessary.

3.2 Linear prediction

A prediction is formulated as a rule, i.e. function, that associates a predicted value with observed values of the of a process, e.g. temperature. In general a predictor of x_t is[7]

$$\hat{x}_t^f = f(x_{t-1}, x_{t-2}, \ldots)$$

that is $\hat{x}_t^f$ is a stochastic process which is a function of the past values of the process x_t.

Thus in principle we have as many predictors of x_t as many functions. Our task is to select a predictor that is optimal. But to define optimality we need a criterion. For example:

1. minimise the expected value of the absolute prediction error

$$\mathrm{E}[|\hat{x}_t^f - x_t|]$$

2. minimise the expected value of the squared prediction error

$$\mathrm{E}[(\hat{x}_t^f - x_t)^2]$$

Our criterion will be the second:

$$\min_f \mathrm{E}[(\hat{x}_t^f - x_t)^2]$$

which being a convex and differentiable function is easier to minimise than the absolute deviation.[8] So we are seeking an element in the set of all functions, such that the expected squared error is minimum. This is a huge set to explore!

Since we limit ourselves to considering only second moments then we are interested in linear dependences only. Thus, we simplify the problem by restricting the set of functions to linear functions:

$$f(x_{t-1}, x_{t-2}, \ldots) = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots$$

Now the minimisation problem becomes

$$\min_{a_0, a_1, a_2 \ldots} \mathrm{E}[(x_t - (a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots)^2]$$

This can be restated like this:

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots] + e_t$$

[6] There are processes as long memory processes which might be stationary hence linear (because of Wold) but with coefficients that are square summable but not absolute summable (see section 3.3).

[7] Hereafter to simplify notation we denote both the random variables and their realisations as x_t.

[8] Minimising absolute deviations gives results that are more robust to large errors, indeed for large x (actually for $x > 1$) we always have $|x| < x^2$.

We look for the coefficients a_j such that $\mathrm{E}[e_t^2]$ is minimum and this looks very much like a linear regression of x_t on its lags.

If we simplify to the s lags specifications we have

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_s x_{t-s}] + e_t$$

Define the vectors

$$\mathbf{a} = (a_0, a_1, a_2, \ldots, a_s)' \qquad \mathbf{z}_t = (1, x_{t-1}, x_{t-2}, \ldots, x_{t-s})'.$$

then the linear model becomes

$$x_t = \mathbf{a}'\mathbf{z}_t + e_t$$

and the coefficients $\mathbf{a}^*$ that minimise $\mathrm{E}[e_t^2]$ are given by the $s+1$-dimensional vector [9]

$$\mathbf{a}^* = (\mathrm{E}[\mathbf{z}_t \mathbf{z}_t'])^{-1} \mathrm{E}[\mathbf{z}_t x_t] \tag{6}$$

these coefficients satisfy the condition $\mathrm{E}[e_t^* x_{t-k}] = 0$ for any $k \neq 0$, and also $\mathrm{E}[e_t^* 1] = \mathrm{E}[e_t] = 0$, where

$$e_t^* = x_t - (\mathbf{a}^*)'\mathbf{z}_t$$

Notice that definition (6) makes sense only if x_t is a stationary process.

To see that minimising the squared distance between regressors and x_t is equivalent to require orthogonality of the regressors and the residuals consider the case of two stochastic variables y and z and the two related problems.

1. We want the best linear approximation of y by means of z, that is

   $$y = az + e$$

   where according to our optimality criterion a is such that $\mathrm{E}[e^2] = \mathrm{E}[(y - az)^2]$ is minimum. Set to zero the derivative with respect to a we have

   $$\frac{\mathrm{d}}{\mathrm{d}a}\mathrm{E}[(y - az)^2] = 2a\mathrm{E}[z^2] - 2\mathrm{E}[yz] = 0$$

   and you obtain $a = \mathrm{E}[yz]/\mathrm{E}[z^2]$.

2. We want to find the number b such that $e = y - bz$ is orthogonal to z, orthogonality between the stochastic variables w_1 and w_2 meaning that the moment $\mathrm{E}[w_1 w_2]$ is equal to zero.[10] Then we have $\mathrm{E}[ez] = \mathrm{E}[yz] - b\mathrm{E}[z^2] = 0$, which implies $b = \mathrm{E}[yz]/\mathrm{E}[z^2]$, which is equal to a above.

Thus in the best linear approximation of x_t with its past the errors are orthogonal to the regressors and

$$P_{t-1}^s x_t = (\mathbf{a}^*)'\mathbf{z}_t$$

is the linear projection of x_t onto the space spanned by its lagged values. The notation P_{t-1}^s tells us that the projector is computed using s lags and that use observations up to time $t-1$.

[9] Try to solve to using just a_1 and by taking the value that sets to zero the first derivative of $\mathrm{E}[e_t^2]$ with respect to a_1.

[10] This is because we are working in a Hilbert space of functions on a probability space (random variables) which are square integrable (finite variance). Such space has a natural inner product given by the covariance operator. Hence two elements are orthogonal if and only if their inner product is zero, that is their covariance is zero.

As seen from (6) in order to compute the coefficients we need sample acvs and once we replace expectations with these quantities we have

$$\hat{\mathbf{a}}_T = \left(\frac{1}{T} \sum_{t=1}^{T-s} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T-s} \mathbf{z}_t x_t \right).$$

For example if we have

$$x_t = a_1 x_{t-1} + e_t$$

we obtain the usual OLS estimator

$$\hat{a}_{1,T} = \frac{\sum_{t=1}^{T-1} x_t x_{t-1}}{\sum_{t=1}^{T-1} x_{t-1}^2}$$

Note that we are using coefficients that are independent of t. But the coefficients depend on the acvs. Thus, assuming that the coefficients are time-invariant requires that the covariances are time-invariant, i.e. that x_t is weakly stationary. Estimation is discussed in detail in Chapter 5.

Back to the general case with infinite lags we have

$$x_t = [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots] + e_t$$

The best linear predictor (obtained with s lags) is then[11]

$$P_{t-1}^s x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_s x_{t-s}.$$

In the most general case we have to consider the predictor based on infinitely many past observations

$$P_{t-1} x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots.$$

Then, it is possible to prove that

$$P_{t-1} x_t = \lim_{s \to \infty} P_{t-1}^s x_t.$$

Of course in empirical situations, in which only a sample for $t = 1, 2, \ldots, T$ is available, we will estimate a regression on a finite number s of lags, with s determined by some information criterion (see later for more).

The best linear predictor of x_t, based on its past, is then the projection $P_{t-1} x_t$ such that:

$$x_t = P_{t-1} x_t + e_t,$$

this is the prediction equation for x_t and notice that if $\mathrm{E}[e_t | x_{t-1}, x_{t-2}, \ldots] = 0$ then the best linear predictor is also the conditional expectation of x_t given its past. This assumption is the conditional mean independence assumption. It can be shown that it is automatically satisfied if e_t is an independent process (or if e_t is Gaussian) since then e_t is independent of $(x_{t-1}, x_{t-2}, \ldots)$ because each x_{t-k} can be written as an infinite sum of lagged values of e_t (see the Wold representation below). Notice that independence is not strictly necessary to have conditional mean independence, which is indeed a weaker requirement. All the following derivation does not require conditional mean independence, since it is based only on linear predictors (projections).

[11] Hereafter, we assume to be working to the true model, i.e. with coefficients such that they minimise the mean squared error. Therefore, we set $a_j \equiv a_j^*$ and we denote the resulting error as e_t and not as e_t^* for simplicity of notation.

The process e_t, is the one step ahead prediction error, is also called the innovation of the process x_t. Looking at the projection equation, the term innovation seems quite appropriate. The only reason why the process x_t is not completely determined by its past values is the presence of the term e_t.

A very important result is that the process e_t is a white noise. Indeed, we have

$$e_t = x_t - [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots]$$

thus e_t is weakly stationary because is a function of a weakly stationary process. Moreover, $\mathrm{E}[e_t x_{t-h}] = 0$ for any $h > 0$ and $\mathrm{E}[e_t] = 0$ and

$$e_{t-1} = x_{t-1} - [a_0 + a_1 x_{t-2} + a_2 x_{t-3} + \ldots]$$

so that $\mathrm{E}[e_t e_{t-1}] = 0$ etc. Suppose that e_t were not a white noise. For example, the autocovariance $\gamma_1^e \neq 0$. Then in the projection

$$e_t = \alpha e_{t-1} + \epsilon_t,$$

the coefficient α is not zero, this implying that

$$\mathrm{E}[e_t^2] = \alpha^2 \mathrm{E}[e_{t-1}^2] + \mathrm{E}[\epsilon_t^2] > \mathrm{E}[\epsilon_t^2]$$

(recall that in a projection we have $\mathrm{E}[\epsilon_t e_{t-1}] = 0$). Now

$$\begin{aligned} x_t &= P_{t-1} x_t + e_t \\ &= [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots] + \alpha e_{t-1} + \epsilon_t \\ &= [a_0(1-\alpha) + (a_1 + \alpha) x_{t-1} + (a_2 - \alpha a_1) x_{t-2} + \ldots] + \epsilon_t \end{aligned}$$

But this contradicts the assumption that e_t is the residual of the projection of x_t on its past. So e_t is a white noise. This result provides the foundation for defining the AR models as models where x_t is driven by its lags plus a white noise process.

On the other hand, it is also possible to prove that $x_t = e_t$ if and only if x_t is a white noise. Therefore a white noise is unpredictable. Better, we can say that stationary processes are predictable because their pattern of autocorrelation is constant through time. A white noise is the least predictable among stationary processes. Processes whose autocorrelation is not regular through time are also unpredictable.

To conclude this part notice that an empirical rule emerges for choosing the number of lags s to include in a linear model. In order to have the best linear predictor we must add as many lags as necessary to make the error e_t a white noise, in this case we can then use least squares to estimate the model (see Chapter 5 on estimation for details).

Examples.

1. $x_t = A$, where A is a constant. In this case the projection equation is

   $$x_t = x_{t-1} + 0,$$

   but also $x_t = x_{t-2} + 0$, etc. Thus the innovation is zero.

2. $x_t = (-1)^t A$. Same as in the previous case, only that here the projection is $x_t = -x_{t-1} + 0 = x_{t-2} + 0$, etc. The innovation is zero.

3. The AR(1) process,
$$x_t = \phi_1 x_{t-1} + u_t, \qquad |\phi_1| < 1$$
This means that the best linear prediction of x_t is $\phi_1 x_{t-1}$.

4. The MA(1) process
$$x_t = u_t + \theta_1 u_{t-1}$$
then using $u_t = x_t - \theta_1 u_{t-1}$ and $u_{t-1} = x_{t-1} - \theta_1 u_{t-2}$
$$\begin{aligned} x_t &= u_t + \theta_1 u_{t-1} \\ &= u_t + \theta_1 x_{t-1} - \theta_1^2 u_{t-2} \\ &= u_t + \sum_{k=1}^{\infty} (-1)^{k-1} \theta_1^k x_{t-k} \end{aligned}$$
which requires $|\theta_1| < 1$ in order to converge. Thus with the constraint $|\theta_1| < 1$ an MA(1) can be written as an AR(∞) and the best linear prediction of x_t is $\theta_1[x_{t-1} - \theta_1 x_{t-2} + \ldots]$.

3.3 Wold representation theorem

Based on the recursive arguments used in the AR(1) and MA(1) cases of previous section we can rewrite the projection equation as an MA. We have
$$\begin{aligned} x_t &= [a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \ldots] + e_t \\ x_{t-1} &= [a_0 + a_1 x_{t-2} + a_2 x_{t-3} + \ldots] + e_{t-1} \end{aligned}$$
by substituting the second into the first we have
$$\begin{aligned} x_t &= a_0 + a_1([a_0 + a_1 x_{t-2} + a_2 x_{t-3} + \ldots] + e_{t-1}) + a_2 x_{t-2} + e_t \\ &= e_t + a_1 e_{t-1} + (a_0 + a_1 a_0) + (a_1^2 + a_2) x_{t-2} + \ldots \\ &= e_t + b e_{t-1} + [c_0 + c_1 x_{t-2} + c_2 x_{t-3} + \ldots] \end{aligned}$$
We may hope that iterating the procedure we obtain a result like the one obtained in the AR(1) case:
$$x_t = b + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \ldots \tag{7}$$
This is not true in general, as the example $x_t = A$ shows, indeed A is deterministic while (7) is stochastic. Therefore, an AR(∞) cannot be written as a purely MA(∞).

We define a deterministic process as a process such that $P_{t-1} x_t = x_t$ which means that it has zero innovations (see the examples above). Notice that according to this definition also the case $x_t = Z$ with Z being a random variable not depending on time can be considered as deterministic. Then we have the following representation result.

WOLD REPRESENTATION THEOREM
If $\{X_t\}$ is a non-deterministic stationary process, then
$$x_t = \sum_{j=0}^{\infty} b_j e_{t-j} + d_t,$$
where

1. $b_0 = 1$ and $\sum_{j=0}^{\infty} b_j^2 < \infty$;
2. e_t is a white noise with zero mean and variance σ_e^2 and $e_t = P_t e_t$ for all t;
3. d_t is deterministic and $d_t = P_s d_t$ for all s, t;
4. $\text{Cov}(e_s, d_t) = 0$ for all s, t;

and the decomposition is unique. Moreover we have

$$\begin{aligned} e_t &= x_t - P_{t-1} x_t \\ b_j &= \frac{\text{E}[x_t e_{t-j}]}{\text{E}[e_t^2]} \\ d_t &= x_t - \sum_{j=0}^{\infty} b_j e_{t-j}. \end{aligned}$$

Note that the expression for b_j is the result of orthogonal projection of $\{X_t\}$ onto the lags of $\{e_t\}$ but is not a definition which can be used in practice for estimation since $\{e_t\}$ is unknown. When $d_t = 0$ then we say that $\{X_t\}$ is purely non-deterministic. ARMA processes are purely non-deterministic.

In conclusion, a weakly stationary process is the sum of a linear process which has the form of an infinite backward moving average of the innovation, which is a white noise, plus a linearly deterministic process. The two components are orthogonal at all leads and lags. Therefore, we have shown that a weakly stationary process is either a linear process or it can be transformed to a linear process by subtracting a deterministic process.

Notice that we have in general only square summability of the coefficients $\{b_j\}$ and that is a necessary and sufficient condition enough to ensure stationarity as seen at the beginning of the Chapter. We also know that square summable coefficients constitute a broader class than the class of absolutely summable coefficients.[12] In ARMA models we always have geometric (exponential) decay of coefficients of the MA(∞) (think of the AR(1) case where $b_j = \phi^j$ for a given $|\phi| < 1$), then both kind of convergence hold. Hence the Wold theorem contains ARMA as special cases but it is more general since it includes all stationary processes with square summable coefficients. ARMA can be seen as particular stationary processes with memory shorter than the one given by the theorem. Fractionally integrated ARMA have instead coefficients decaying as $b_j \sim 1/j^{1-d}$ which are not absolutely summable but are square summable as long as $0 \leq d < 1/2$, these are called long memory processes.

In practice, given a sample of a weakly stationary time series, finding the Wold representation requires the estimation of an infinite number of parameters $(b_1, b_2, ...)$ using the data, which is clearly not possible as said above. In practice, one typically has to make some assumptions about $(b_1, b_2, ...)$. A common approach is to assume that

$$\sum_{j=0}^{\infty} b_j L^j = \frac{\Theta(L)}{\Phi(L)} = \frac{1 + \sum_{m=1}^{p} \theta_m L^m}{1 - \sum_{m=1}^{q} \phi_m L^m},$$

where $p, q < \infty$, namely one approximates an unfeasible infinite polynomial by the ratio of finite-order polynomials. This again takes us to the definition of an ARMA(p, q) process.

[12] We have the sequence spaces inclusion $\ell_1 \subset \ell_2$, so for example $1/j$ is a sequence which is square summable but not absolutely summable.

Finally, in many cases it is reasonable to assume also conditional mean independence, i.e. that $\mathrm{E}[e_t|x_{t-1}, x_{t-2}, \ldots] = 0$, in which case we say that e_t is a martingale difference sequence with respect to x_t.[13] For example if we restrict ourselves to linear dependencies and we are not interested in any other dependence, then $\mathrm{E}[e_t|x_{t-1}, x_{t-2}, \ldots] = 0$ might be reasonable since we already know that being a white noise e_t is not correlated with $(x_{t-1}, x_{t-2}, \ldots)$. Gaussian innovations are also independent and therefore satisfy conditional mean independence. As said above conditional mean independence of the innovations with respect to the past tells us the the best linear predictor $P_{t-1}x_t$ is actually also the best predictor which is always given by the conditional mean $\mathrm{E}[X_t|X_{t-1}, X_{t-1}, \ldots]$.

3.4 Forecasting

If we assume to have observations only until time T we can use the same approach as before to define h steps ahead forecasts of a time series. We can write the forecast at $T + h$ as

$$x_{T+h} = [a_0 + a_1 x_T + a_2 x_{T-1} + \ldots + a_m x_{T-m}] + e_{T+h} \qquad h > 0.$$

where m can be any positive number (depending how far in the past we go). The best linear predictor (obtained with m lags) is then

$$P_T^m x_{T+h} = a_0 + a_1 x_T + a_2 x_{T-1} + \ldots + a_m x_{T-m}.$$

If we consider the predictor based on infinitely many past observations

$$P_T x_{T+h} = a_0 + a_1 x_T + a_2 x_{T-1} + \ldots.$$

it is possible to prove that

$$P_T x_{T+h} = \lim_{m \to \infty} P_T^m x_{T+h}.$$

Of course in empirical situations, in which only a sample for $t = 1, 2, \ldots, T$ is available, we will estimate a regression on a finite number m of lags, with m determined by some information criterion (see later for more) and we must have $m < T$.

Both for prediction (which is in sample and for forecasting which is out of sample), the choice of the projector to be used of course depends on the model chosen for the time series we are studying. The class of model we will consider in the next Chapter are ARMA processes.

3.5 Sample mean and sample autocorrelation

A stationary process is characterised by its mean μ and its acvs γ_k. Therefore given observations $x_1, x_2, \ldots, x_T$, we should be able to estimate these moments if we want to analyse data and make inference on order to build the most appropriate model for the data. The estimators we build are called sample mean, denoted as $\bar{x}_T$ and sample acs, denoted as $\hat{\rho}_{k,T}$.[14]

[13] A martingale sequence is instead such that $\mathrm{E}[x_t|x_{t-1}] = x_{t-1}$ so a random walk with i.i.d. errors or with errors, that are a martingale difference sequence, is a martingale.

[14] Only in this section we highlight the role of T in estimation by using it as an index.

3.5.1 Ergodicity

Methods we shall look at for estimating quantities such as the mean and the acvs will use observations from a single realisation of the process. Such methods are based on the strategy of re-placing ensemble averages by their corresponding time averages and if this is possible the process is called ergodic. More precisely, if we have N i.i.d. realisations of a stochastic process $\{X_t\}$ denoted as $\{x_t^k\}$, with $k = 1, \ldots, N$, then for any fixed point in time t_0 we define the ensemble average as

$$\bar{x}_{t_0}^N = \frac{1}{N} \sum_{k=1}^{N} x_{t_0}^k$$

Since each realisation of the process is independent of the others, then the above can be seen as a sum of i.i.d. random variables and therefore we know that the Weak Law of Large Numbers applies (for any fixed t_0)

$$\bar{x}_{t_0}^N \xrightarrow{p} \mathrm{E}[X_{t_0}].$$

If we could use the ensemble average then we would not need to require stationarity.

Consider instead the time average based on T observations

$$\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t,$$

this quantity does not depend on time. We say that $\{X_t\}$ is ergodic for the mean, if, as $N, T \to \infty$, the ensemble average tends to the same limit as the sample average of one given realisation, that is, for any t_0,

$$p \lim_{T \to \infty} \bar{x}_T = p \lim_{N \to \infty} \bar{x}_{t_0}^N.$$

Now since the left hand side does not depend on time we must have that the sample mean converges to the mean which must then be independent of time (and of the realisation observed)

$$\bar{x}_T \xrightarrow{p} \mathrm{E}[X_t].$$

Using an analogous definition of sample covariances (see below) we can define ergodicity also for the second (and higher) moments. The interpretation of this property is that a long series of observations yields enough information to conduct inference on the moments of the process itself.

The general definition of ergodicity is related to the whole distribution of $\{X_t\}$ and it is highly technical. We know that ergodicity implies strong stationarity and therefore also weak stationarity. Loosely speaking, a stochastic process $\{X_t\}$ is ergodic if any two collections of random variables partitioned far apart in the sequence are almost independently distributed.[15]

For us stationarity and ergodicity will always be considered jointly. However, notice that in general stationary processes are not necessarily ergodic. For example consider a process $\{Y_t\}$ which has mean μ and variance σ_y^2 and is i.i.d. and a r.v. $X \sim N(0, 1)$ independent of $\{Y_t\}$. Then

$$Z_t = Y_t + X$$

[15] Asymptotic independence (as the distance in time increases) is captured also by other stronger properties not considered here such as mixing. Notice that: (1) mixing does not imply stationarity; (2) a mixing and strongly stationary process is ergodic.

is stationary as it has mean $\mathrm{E}[Z_t] = \mu$ and acvs $\gamma_0^z = \sigma_y^2 + 1$ and $\gamma_h^z = 1$ for $h \neq 0$ since $\gamma_h^x = 1$ for any h. However, it is not ergodic since

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \xrightarrow{p} \mu + X$$

which is not the mean of Z_t and it will be different for any different realisation of the process. Intuitively, the reason why we don't have ergodicity is because Z_t and Z_{t+h} will never become independent as their distance increases, because $\gamma_h^z = 1$ even as $h \to \infty$. With the language of previous section the component X represents the deterministic part in the Wold decomposition of Z_t. This is the reason why stationary process might not be ergodic. An example might be given by a seasonal component which is periodic in time, but deterministic.

If we restrict ourselves to stationary processes with summable acvs, then we can prove that we have ergodicity for the mean. Notice that in the example before the acvs are not summable. This is what we do next.

3.5.2 Mean estimation

Consider a stationary process and its sample mean which is defined as

$$\bar{x}_T = \frac{1}{T}(x_1 + x_2 + \ldots + x_T) = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

Let us study the properties of this estimator.

It is unbiased

$$\mathrm{E}[\bar{x}_T] = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[x_t] = \frac{1}{T}T\mu = \mu$$

Its variance is

$$\begin{aligned}
\mathrm{Var}(\bar{x}_T) &= \frac{1}{T^2}\mathrm{Var}\left(\sum_{t=1}^{T} x_t\right) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\mathrm{Var}(x_t) + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1;s\neq t}^{T}\mathrm{Cov}(x_t, x_s) \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_{t-s}^x \qquad (8)
\end{aligned}$$

We then make a change of variable from s, t to s and $h = (t-s)$. Notice that the double summation above extends over the entries of the the Toepliz matrix seen in Chapter 2. If we use t as an index for rows and s as the index for columns we have

$$\mathbf{V} = \begin{bmatrix}
\mathrm{Cov}(x_1, x_1) & \mathrm{Cov}(x_1, x_2) & \mathrm{Cov}(x_1, x_3) & \ldots & \mathrm{Cov}(x_1, x_{T-1}) & \mathrm{Cov}(x_1, x_T) \\
\mathrm{Cov}(x_2, x_1) & \mathrm{Cov}(x_2, x_2) & \mathrm{Cov}(x_2, x_3) & \ldots & \mathrm{Cov}(x_2, x_{T-1}) & \mathrm{Cov}(x_2, x_T) \\
\mathrm{Cov}(x_3, x_1) & \mathrm{Cov}(x_3, x_2) & \mathrm{Cov}(x_3, x_3) & \ldots & \mathrm{Cov}(x_3, x_{T-1}) & \mathrm{Cov}(x_3, x_T) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\mathrm{Cov}(x_{T-1}, x_1) & \mathrm{Cov}(x_{T-1}, x_2) & \mathrm{Cov}(x_{T-1}, x_3) & \ldots & \mathrm{Cov}(x_{T-1}, x_{T-1}) & \mathrm{Cov}(x_{T-1}, x_T) \\
\mathrm{Cov}(x_T, x_1) & \mathrm{Cov}(x_T, x_2) & \mathrm{Cov}(x_T, x_3) & \ldots & \mathrm{Cov}(x_T, x_{T-1}) & \mathrm{Cov}(x_T, x_T)
\end{bmatrix}$$

While in (8) we sum first over rows and then we add the row sums, we can now sum over diagonals and add the diagonal sums together. To see how h, t, and s are related consider again $\mathbf{V}$ as a function of h

$$\mathbf{V} = \begin{bmatrix} \gamma_0 & \gamma_{-1} & \gamma_{-2} & \cdots & \gamma_{-(T-2)} & \gamma_{-(T-1)} \\ \gamma_1 & \gamma_0 & \gamma_{-1} & \cdots & \gamma_{-(T-3)} & \gamma_{-(T-2)} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{-(T-4)} & \gamma_{-(T-3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{T-2} & \gamma_{T-3} & \gamma_{T-4} & \cdots & \gamma_0 & \gamma_{-1} \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}$$

The sum over the diagonals are indexed by h which goes from $-(T-1)$ to $(T-1)$ (this is because the observations must be at a distance less than $T-1$ to have non zero covariance). Then for $h > 0$ (diagonals below the main diagonal), the index s goes from 1 to $(T-h)$, while for $h < 0$ (diagonals above the main diagonal), s goes from $-h+1$ to T. Since the summand in (8) depends only on h we have that the sum on s gives always $T - |h|$ hence

$$\begin{aligned} \text{Var}(\bar{x}_T) &= \frac{1}{T^2} \sum_{h=-(T-1)}^{T-1} (T - |h|)\, \gamma_h^x \\ &= \frac{1}{T^2} \sum_{h=-(T-1)}^{T-1} T \left(1 - \frac{|h|}{T}\right) \gamma_h^x \\ &= \frac{1}{T} \sum_{h=-(T-1)}^{T-1} \left(1 - \frac{|h|}{T}\right) \gamma_h^x \end{aligned}$$

Compare this result with the case of i.i.d. r.v. where all covariances are zero. We make the assumption that $\sum_{h=-\infty}^{\infty} |\gamma_h^x| < \infty$. Then, by the Cesáro summability theorem, if $\sum_{h=-(T-1)}^{T-1} \gamma_h^x$ converges to a limit as $T \to \infty$[16] then, also $\sum_{h=-(T-1)}^{T-1} \left(1 - \frac{|h|}{T}\right) \gamma_h^x$ converges to the same limit. We can thus conclude that,

$$\lim_{T\to\infty} T\ \text{Var}(\bar{x}_T) = \lim_{T\to\infty} \sum_{h=-(T-1)}^{T-1} \left(1 - \frac{|h|}{T}\right) \gamma_h^x = \sum_{h=-\infty}^{\infty} \gamma_h^x < \infty$$

Therefore as long as $\sum_{h=-\infty}^{\infty} |\gamma_h^x| < \infty$ we have $\text{Var}(\bar{x}_T) \to 0$ when $T \to \infty$. Thus we have convergence in mean square

$$\text{E}[(\bar{x}_T - \mu)^2] = \text{Var}(\bar{x}_T) \to 0, \quad \text{as } T \to \infty,$$

and by Chebychev's inequality also convergence in probability

$$\text{P}(|\bar{x}_T - \mu| > \epsilon) \to 0 \qquad \forall \epsilon > 0, \quad \text{as } T \to \infty,$$

or

$$\bar{x}_T \xrightarrow{p} \mu, \quad \text{as } T \to \infty.$$

[16] It must since

$$\left| \sum_{h=-(T-1)}^{T-1} \gamma_h^x \right| \leq \sum_{h=-(T-1)}^{T-1} |\gamma_h^x| < \infty \qquad \forall T.$$

This is the Weak Law of Large Numbers for stationary time series also known as Ergodic Theorem, since if this result holds then the process is ergodic.[17] Therefore, if the process is stationary and $\sum_{h=-\infty}^{\infty} |\gamma_h^x| < \infty$ then it is ergodic for the mean. Moreover, it can be proved that if the process is ergodic and stationary then $\bar{x}_T \xrightarrow{p} \mu$, as $T \to \infty$.

Consider the Wold decomposition of a purely non-deterministic stationary process, that is the linear process, or MA(∞),

$$x_t = \sum_{j=0}^{\infty} \psi_j u_{t-j}, \quad u_t \sim w.n.(0, \sigma_u^2),$$

but with absolute summability of the coefficients $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such are the ARMA models. Then this condition implies absolute summability of the covariances

$$\begin{aligned}
\sum_{h=-\infty}^{\infty} |\gamma_h^x| &= 2\left\{\sum_{h=0}^{\infty} |\gamma_h^x|\right\} - \gamma_0^x \\
&\leq 2\sigma_u^2 \sum_{h=0}^{\infty} \left| \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \right| \\
&\leq 2\sigma_u^2 \sum_{j=0}^{\infty} |\psi_j| \sum_{h=0}^{\infty} |\psi_{j+h}| \\
&\leq 2\sigma_u^2 \sum_{j=0}^{\infty} |\psi_j| \sum_{h=0}^{\infty} |\psi_h| \\
&\leq 2\sigma_u^2 \left(\sum_{j=0}^{\infty} |\psi_j| \right)^2 < \infty.
\end{aligned}$$

Therefore, any purely non-deterministic stationary process as ARMA (which then has absolutely summable coefficients) satisfy the necessary condition for the Law of Large Numbers to hold. Actually it is enough to have $\gamma_h \to 0$ as $h \to \infty$ to have ergodicity, so a purely non-deterministic stationary process (which then has an MA(∞) representation with square summable coefficients, is ergodic.

Convergence in probability implies also convergence in distribution, and more precisely from the results above we have seen that $T\text{Var}(\bar{x}_T)$ must be a finite number as $T \to \infty$, thus we have a Central Limit Theorem for the sample average of time series that states

$$\sqrt{T}\left(\bar{x}_T - \mu\right) \xrightarrow{d} N\left(0, \sum_{h=-\infty}^{\infty} \gamma_h^x\right), \quad \text{as } T \to \infty.$$

Using this result we can build asymptotic confidence intervals for μ. Thus an approximate (i.e. for $T \to \infty$) 95% confidence interval is

$$\left(\bar{x}_T - 1.96\frac{v^{1/2}}{\sqrt{T}}, \bar{x}_T + 1.96\frac{v^{1/2}}{\sqrt{T}}\right)$$

where $v = \sum_{h=-\infty}^{\infty} \gamma_h^x$.

[17] Indeed, as seen above the ensemble average converges to the same limit.

Now, v is unknown and must be in turn estimated. Usually we use

$$\hat{v}_T = \sum_{h=-\sqrt{T}}^{\sqrt{T}} \left(1 - \frac{|h|}{T}\right) \hat{\gamma}^x_{h,T},$$

where $\hat{\gamma}^x_{h,T}$ is the sample acvs which we now introduce.

3.5.3 Estimation of the autocovariance function

The lag h acvs is defined as

$$\gamma^x_h = \mathrm{E}[(x_{t+|h|} - \mu)(x_t - \mu)].$$

Thus a natural estimator is

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

If we replace $\bar{x}_T$ with μ we have

$$\mathrm{E}[\tilde{\gamma}^x_{h,T}] = \frac{1}{T - |h|} \sum_{t=1}^{T-|h|} \mathrm{E}[(x_{t+|h|} - \mu)(x_t - \mu)] = \frac{1}{T - |h|} \sum_{t=1}^{T-|h|} \gamma^x_h = \gamma^x_h,$$

therefore, $\tilde{\gamma}^x_{h,T}$ is unbiased if we know μ. Most texts refer to $\tilde{\gamma}^x_{h,T}$ as unbiased, however, if μ is estimated by $\bar{x}_T$, $\tilde{\gamma}^x_{h,T}$ is typically a biased estimator of γ^x_h.

Another possible and preferred estimator is

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

If we replace $\bar{x}_T$ with μ we have

$$\mathrm{E}[\hat{\gamma}^x_{h,T}] = \frac{1}{T} \sum_{t=1}^{T-|h|} \mathrm{E}[(x_{t+|h|} - \mu)(x_t - \mu)] = \frac{1}{T} \sum_{t=1}^{T-|h|} \gamma^x_h = \left(1 - \frac{|h|}{T}\right) \gamma^x_h,$$

is a biased estimator, and the magnitude of its bias increases as $|h|$ increases. Most texts refer to $\hat{\gamma}^x_{h,T}$ as biased.

Why should we prefer the "biased" estimator $\hat{\gamma}^x_{h,T}$ to the "unbiased" estimator $\tilde{\gamma}^x_{h,T}$?

1. For many stationary processes of practical interest we have

   $$MSE(\hat{\gamma}^x_{h,T}) < MSE(\tilde{\gamma}^x_{h,T}),$$

   where MSE is the Mean Squared Error defined as

   $$\begin{aligned} MSE(\hat{\gamma}^x_{h,T}) &= \mathrm{E}[(\hat{\gamma}^x_{h,T} - \gamma^x_h)^2] \\ &= \mathrm{E}[(\hat{\gamma}^x_{h,T})^2] - 2\gamma^x_h \mathrm{E}[\hat{\gamma}^x_{h,T}] + (\gamma^x_h)^2 \\ &= \left(\mathrm{E}[(\hat{\gamma}^x_{h,T})^2] - \mathrm{E}[\hat{\gamma}^x_{h,T}]^2\right) + \mathrm{E}[\hat{\gamma}^x_{h,T}]^2 - 2\gamma^x_h \mathrm{E}[\hat{\gamma}^x_{h,T}] + (\gamma^x_h)^2 \\ &= \mathrm{Var}(\hat{\gamma}^x_{h,T}) + \left(\hat{\gamma}^x_{h,T} - \mathrm{E}[\hat{\gamma}^x_{h,T}]\right)^2 \\ &= \text{variance} + (\text{bias})^2 \end{aligned}$$

2. We know that the acvs must be positive semidefinite and the sequence $\{\hat{\gamma}^x_{h,T}\}$ has this property, whereas the sequence $\{\tilde{\gamma}^x_{h,T}\}$ may not. That is to say that the the sample variance-covariance matrix of the equispaced vector $(x_1, x_2, \ldots, x_T)'$ is non-negative definite. This is the estimated Toepliz matrix $\mathbf{V}$ in Section 2.2 where are all acvs are replaced by the sample ones.

Once we choose an estimator of the acvs, the sample acs is then

$$\hat{\rho}^x_{h,T} = \frac{\hat{\gamma}^x_{h,T}}{\hat{\gamma}^x_{0,T}}$$

Notice that given T observations we can in principle compute acvs only up to lag $T-1$. However, in practice we have to stop well before $T-1$ as for large h estimates of acvs are computed using too few values (just one if $h = T-1$). A useful rule of thumb is to compute acvs only up to lag $h \leq T/4$ when $T > 50$.

The sample acvs is needed for

1. computing confidence intervals for the mean;
2. finding which are the relevant acvs for a given process, in order to select the best model to forecast and analyse the data;
3. estimating the parameters of a given model.

In order to challenge part 2 we need a distribution for sample acvs and acs. This is given again by a Central Limit Theorem and it is then valid only when T is large

$$\sqrt{T}\left(\hat{\rho}^x_{h,T} - \rho^x_h\right) \xrightarrow{d} N\left(0, w\right), \quad \text{as } T \to \infty,$$

where w has a long formula (called Bartlett's formula) which we don't give here for the general case but it is the one used to plot confidence intervals of acvs in all the figures of these notes. A further approximation is given by setting $w = 1$.

4 ARMA processes

4.1 MA(1) and AR(1)

We have already seen the MA(1) and the AR(1) processes. Recall that an MA(1) is defined as

$$x_t = u_t + \theta u_{t-1} = (1 + \theta L)u_t, \qquad u_t \sim wn(0, \sigma^2)$$

This process is always stationary and depends only on past values of the white noise (below we will say that is causal), thus this is already the Wold representation of x_t. We can write an MA(1) as an AR(∞):

$$x_t = (1 + \theta L)u_t = u_t + \sum_{k=1}^{\infty} (-1)^{k-1}\theta^k x_{t-k}.$$

For the above series to converge we need $|\theta| < 1$ and then we can use an MA(1) to make predictions, since by definition here u_t is the innovation of x_t (see Chapter 3). Summing up:

1. if $|\theta| < 1$ the MA(1) process is invertible using past values of x_t, i.e. allows us to write an MA(1) as an AR(∞), and is stationary;

2. if $|\theta| > 1$ the MA(1) process is invertible using future values of x_t and we will have an MA(1) with parameter $1/\theta$, clearly the process is still stationary (see below the examples in section 4.2);

3. if $|\theta| = 1$ the MA(1) process is never invertible, but still stationary.

An AR(1) is defined as

$$x_t - \phi x_{t-1} = (1 - \phi L)x_t = u_t, \qquad u_t \sim wn(0, \sigma^2).$$

This equation is already in the form of a linear prediction equation since u_t is the innovation of x_t (see Chapter 3). We can write an AR(1) as an MA(∞):

$$x_t = \phi x_{t-1} + u_t = \sum_{k=0}^{\infty} \phi^k u_{t-k}.$$

Then, in order for the above series to converge we need $|\phi| < 1$ and then the process is stationary and depends only on past values of u_t.[18] Thus, we say that is also causal and the above is also the Wold representation of x_t. Summing up:

1. if $|\phi| < 1$ the AR(1) process is stationary and causal, i.e. allows us to write an AR(1) as an MA(∞) using past values of u_t (see Chapter 2);

2. if $|\phi| > 1$ the AR(1) process is stationary but not causal, i.e. can be written as

$$x_t = -\frac{1}{\phi} u_{t+1} + \frac{1}{\phi} x_{t+1} = -\sum_{k=0}^{\infty} \frac{1}{\phi^k} u_{t+k},$$

   which is a stationary process (using the same argument as in the previous case) but here x_t is correlated with future values of u_t this is a feasible representation but it is unnatural, moreover it is unstable as the initial condition should now be set into the future as for example $x_T = 0$ as $T \to \infty$ which is not likely to be the case;

3. if $|\phi| = 1$ we have a non-stationary process (called random walk when $\phi = 1$) and we say that this process has a unit root. Unit root processes are non-stationary not because of the presence of a linear deterministic trend but because they are driven by a stochastic trend which makes their variance time dependent and are called Difference Stationary processes as opposed to Trend Stationary processes and will be considered later.

[18] To prove convergence in mean square, consider

$$x_t = \sum_{k=0}^{h-1} \phi^k u_{t-k} + \phi^h x_{t-h},$$

then

$$\lim_{h\to\infty} \mathsf{E}\left[\left(x_t - \sum_{k=0}^{h-1} \phi^k u_{t-k}\right)^2\right] = \lim_{h\to\infty} |\phi|^{2h} \mathsf{E}[x_{t-h}^2] = 0,$$

if and only if $|\phi| < 1$ and x_t has finite variance, which is again the case only for $|\phi| < 1$. A similar reasoning applies for the MA(1) case.

In both examples above we have an AR representation which is what we need for prediction, and we say that x_t is a solution of the difference equation implied by the AR if it can be written as an MA, i.e. it has a Wold representation, thus it has to be (i) stationary and causal with (ii) square summable MA coefficients. In the MA case this is trivial (it's already an MA with a finite number of coefficients), but in the AR case we need $|\phi| < 1$ for stationarity and causality which implies that $\sum_{k=0}^{\infty} |\phi^k| < \infty$ and therefore we have also square summability of the MA coefficients which is a more general condition.

4.2 The AR(p) process

4.2.1 Stationary solutions

An AR(p) has realisations which follow the equation

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) x_t = u_t, \qquad u_t \sim wn(0, \sigma_u^2).$$

This can be seen as a stochastic difference equation (compare it with the stochastic differential equations in continuos time), hence we look for a solution of this equation, which means looking for a linear process (i.e. a moving average of u_t) that solves the equation. For $p = 1$ we know that if $|\phi_1| < 1$ the solution is

$$x_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \ldots$$

A quick way to derive it is by computing the inverse of $(1 - \phi_1 L)$ as: [19] [20]

$$\frac{1}{(1 - \phi_1 z)} = (1 - \phi_1 z)^{-1} = 1 + \phi_1 z + \phi_1^2 z^2 + \ldots$$

for any $z \in \mathbb{C}$ and then write

$$x_t = (1 - \phi_1 L)^{-1} u_t = (1 + \phi_1 L + \phi_1^2 L^2 + \ldots) u_t.$$

The above expression converges if and only if $|\phi_1| < 1$, in which case we have a stationary solution. We know that if $|\phi_1| > 1$ there is also another stationary solution but we need future values of u_t, while no stationary solution exists if $|\phi_1| = 1$ since in that case the variance will not exist. We know want to generalise this result to $p > 1$.

We solve the equation of the AR(p) by factorizing the polynomial into first order factors. Hereafter, when considering polynomials we write it as a function of a generic complex number z, i.e. $z \in \mathbb{C}$ since we know that roots of polynomial are complex numbers. The AR polynomial of

[19] The Taylor expansion we are using is

$$\frac{1}{1 - x} = 1 + x + x^2 + \ldots$$

which converges only if $|x| < 1$. In order to see that this applies also to the operator L applied to a stationary process $\{X_t\}$ one has to recognise that since $\mathrm{E}[L X_t X_t] = \mathrm{E}[X_t L^{-1} X_t]$ (because of stationarity $\gamma_{-1}^x = \gamma_1^x$) then the adjoint of L is $L^* = L^{-1}$ and therefore L is a unitary operator and has spectrum on the unit circle and modulus the identity $|L| = (L L^*)^{1/2} = I$, thus in the AR polynomial $|\phi_k L^k| \leq |\phi_k|$.

[20] Notice that this is indeed the inverse since

$$(1 - \phi_1 z)(1 - \phi_1 z)^{-1} = (1 - \phi_1 z)(1 + \phi_1 z + \phi_1^2 z^2 + \phi_1^3 z^3 + \ldots) = (1 - \phi_1^2 z^2 + \phi_1^2 z^2 - \phi_1^3 z^3 + \phi_1^3 z^3 + \ldots) = 1$$

order p has coefficients $\phi_j \in \mathbb{R}$ and can always be written as (this is known as the "characteristic" polynomial)

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = (1 - \beta_1 z)(1 - \beta_2 z) \ldots (1 - \beta_p z) \tag{9}$$

for some coefficient $\beta_j \in \mathbb{C}$. Then if we denote the p roots as $\alpha_1 \ldots \alpha_p$ (not necessarily distinct) we have

$$\begin{aligned} \Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p &= (1 - \beta_1 z)(1 - \beta_2 z) \ldots (1 - \beta_p z) \\ &= (-1)^p \frac{(z - \alpha_1)(z - \alpha_2) \ldots (z - \alpha_p)}{\alpha_1 \alpha_2 \ldots \alpha_p}, \end{aligned} \tag{10}$$

this guarantees that the coefficient of order zero is equal to one also on the right hand side. Therefore, by comparing (9) with (10) we have $\beta_j = 1/\alpha_j$.

We can then invert the AR(p) polynomial by inverting each $(1 - \beta_j z)$ polynomial. Now recall that, if $|\beta_j| < 1$ we can write

$$\frac{1}{(1 - \beta_j L)} = \sum_{k=0}^{\infty} \beta_j^k L^k. \tag{11}$$

And we have seen in the AR(1) case that this series converges if and only if $|\beta_j| < 1$ which is equivalent to ask for the roots of the polynomial $\Phi(z)$, that is the α's, to be such that $|\alpha_j| > 1$. This is our condition for stationarity and causality for an AR(p).

In general the roots might be complex numbers.[21] Thus we require the roots to be outside the unit circle. So we require the polynomial $\Phi(z)$ to have roots in the complex plane and such that $|z| > 1$, i.e. such that $\Phi(z) = 0$ only if $|z| > 1$. If some of the roots α_j are in modulus smaller than 1 then $\Phi(z)$ can still be inverted but not using only the past, so that x_t in this case will be a two-sided moving average of u_t.

But if at least one of the roots α_j has unit modulus, in that case we say that the AR(p) polynomial has unit roots, then the autoregressive equation has no stationary solution.

Hereafter, we assume that the AR(p) polynomial has no root inside or on the unit circle, that is $|\alpha_j| > 1$ for all j or that $|\beta_j| < 1$ for all j. In this case we say that the causality and stationarity condition is fulfilled.

Finally, notice that we can either find the roots of the characteristic AR polynomial $\Phi(z) = 0$ and have them outside the unit circle or we can consider roots of the reciprocal polynomial defined as (basically this means inverting the order of the coefficients)

$$\Phi^*(z) = z^p \Phi(z^{-1}) = z^p - \phi_1 z^{p-1} - \ldots - \phi_p.$$

Then from (10) we have

$$\Phi^*(z) = z^p \Phi(z^{-1}) = z^p (1 - \beta_1 z^{-1})(1 - \beta_2 z^{-1}) \ldots (1 - \beta_p z^{-1}) = (z - \beta_1)(z - \beta_2) \ldots (z - \beta_p)$$

which shows that now the roots are $|\beta_j| < 1$, hence the condition for stationarity becomes that the roots of the reciprocal of the characteristic polynomial are inside the unit circle.

[21] A polynomial of order p has p roots in the complex plane and if $z \in \mathbb{C} \setminus \mathbb{R}$ then $\bar{z}$ is also a root.

Take as an example an AR(2), then the MA(∞) representation reads

$$\begin{aligned} x_t = (1-\phi_1 L - \phi_2 L^2)^{-1} u_t &= (1-\beta_1 L)^{-1}(1-\beta_2 L)^{-1} u_t \\ &= (1+\beta_1 L + \beta_1^2 L^2 + \ldots)(1+\beta_2 L + \beta_2^2 L^2 + \ldots) u_t \\ &= \left(1 + (\beta_1+\beta_2)L + (\beta_1^2 + \beta_1\beta_2 + \beta_2^2)L^2 + \ldots\right) u_t \\ &= \sum_{k=0}^{\infty} \underbrace{\left(\sum_{j=0}^{k} \beta_1^j \beta_2^{k-j} \right)}_{\psi_k} L^k u_t \end{aligned}$$

which is stationary if and only if the coefficients ψ_k are square summable. In particular, we have

$$\psi_1 = \beta_1 + \beta_2, \quad \psi_2 = \beta_1^2 + \beta_2^2 + \beta_1\beta_2, \quad \psi_3 = \beta_1^3 + \beta_2^3 + \beta_1^2\beta_2 + \beta_1\beta_2^2$$

and in general we have $|\psi_k| \leq (|\beta_1| + |\beta_2|)^k$. Then, the condition $|\beta_1| + |\beta_2| < 1$ is sufficient to have a stationary and causal process since the MA(∞) decay exponentially and are therefore absolute summable which implies they are also square summable. Similarly for any AR(p) the condition $\sum_{j=1}^{p} |\phi_j| < 1$ is sufficient for stationary.

Still about the AR(2) we can write

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 = (1-\beta_1 z)(1-\beta_2 z) = 1 - (\beta_1+\beta_2)z + \beta_1\beta_2 z^2$$

which shows that $\phi_1 = \beta_1 + \beta_2$ and $\phi_2 = -\beta_1\beta_2$ and therefore we have the necessary conditions for stationarity and causality $|\phi_2| \leq |\beta_1|\,|\beta_2| < 1$. Moreover, consider the reciprocal of the characteristic polynomial in this case,

$$\Phi^*(z) = z^2 \Phi(z) = z^2 - \phi_1 z - \phi_2$$

then the solutions are

$$z_{1,2} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2}$$

Now if the roots are real $\phi_1^2 + 4\phi_2 \geq 0$ and to have stationarity we must have

$$\begin{aligned} z_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} < 1 &\Rightarrow \phi_1 + \phi_2 < 1 \\ z_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} > -1 &\Rightarrow \phi_2 - \phi_1 < 1 \end{aligned}$$

which are other two necessary conditions for stationarity.[22]

4.2.2 Autocovariance function

The autocovariance function of an AR(p) process can be obtained similarly to the AR(1) case. Assume that $p = 2$:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + u_t, \qquad u_t \sim wn(0, \sigma_u^2), \tag{12}$$

[22] If the roots are complex then $z_1 = \bar{z}_2$ and we must have

$$|z_1|^2 = z_1 \bar{z}_1 = \frac{\phi_1 + i\sqrt{\phi_1^2 + 4\phi_2}}{2} \frac{\phi_1 - i\sqrt{\phi_1^2 + 4\phi_2}}{2} = \frac{\phi_1^2}{2} + \phi_2 < 1$$

or

$$x_t = u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + \ldots, \qquad u_t \sim wn(0, \sigma_u^2). \tag{13}$$

Multiplying both sides of (12) by $x_{t-h} = u_{t-h} + \psi_1 u_{t-h-1} + \ldots$, for $h \geq 1$, taking expectations, using (13) and the fact that u_t is a white noise (thus $\mathrm{E}[u_t x_{t-h}] = 0$ for $h \geq 1$), we obtain

$$\gamma_h^x = \phi_1 \gamma_{h-1}^x + \phi_2 \gamma_{h-2}^x$$

which are the Yule-Walker equations and look like (12). The first two equations, for $h = 1$ and $h = 2$, are

$$\begin{aligned} \gamma_1^x &= \phi_1 \gamma_0^x + \phi_2 \gamma_{-1}^x = \phi_1 \gamma_0^x + \phi_2 \gamma_1^x \\ \gamma_2^x &= \phi_1 \gamma_1^x + \phi_2 \gamma_0^x \end{aligned}$$

Moreover, multiplying both sides of (12) by x_t, taking expectations, and using (13), we have (we have $\mathrm{E}[u_t x_t] = \sigma_u^2$)

$$\gamma_0^x = \phi_1 \gamma_1^x + \phi_2 \gamma_2^x + \sigma_u^2$$

Then we have a system of three equations and three unknowns which can be used to determine $\gamma_0^x, \gamma_1^x, \gamma_2^x$ (see next Chapter). We can also continue to compute acvs for $h \geq 3$:

$$\gamma_3^x = \phi_1 \gamma_2^x + \phi_2 \gamma_1^x$$

and so on.

For an AR(1) we have (see Chapter 2)

$$\begin{aligned} \gamma_0^x &= \phi_1 \gamma_1^x + \sigma_u^2 \\ \gamma_1^x &= \phi_1 \gamma_0^x \\ \gamma_h^x &= \phi_1^h \gamma_0^x \end{aligned}$$

which gives

$$\gamma_0^x = \frac{\sigma_u^2}{1 - \phi_1^2}, \qquad \gamma_h^x = \frac{\phi_1^h \sigma_u^2}{1 - \phi_1^2}$$

which shows that the acvs decay exponentially. This is always the case even when we have an AR(p). Compare it with the acvs of an MA(q) which is exactly zero at lags $|h| > q$.

There is an easy way to compute acvs of stationary AR processes. Let's consider an example with an AR(2). Consider

$$x_t = 1.3 x_{t-1} - 0.4 x_{t-2} + u_t$$

which in L notation is

$$(1 - 1.3L + 0.4L^2) x_t = u_t$$

the roots of the polynomial are both outside the unite circle and are equal to $z = 2$ and $z = 5/4$ (see example 2 at the end of next section) and we can factorize it as

$$(1 - 1.3L + 0.4L^2) x_t = \left(1 - \frac{L}{2}\right)\left(1 - \frac{4}{5}L\right) x_t = u_t$$

and we can invert to write

$$x_t = \frac{1}{\left(1 - \frac{L}{2}\right)\left(1 - \frac{4}{5}L\right)} u_t$$

which gives

$$x_t = \left(\sum_{k=0}^{\infty} \frac{L^k}{2^k}\right)\left(\sum_{k=0}^{\infty} \frac{4^k}{5^k} L^k\right) u_t$$

Collecting the terms, we get

$$\begin{aligned} x_t &= \left(1 + \frac{13}{10}L + \frac{129}{100}L^2 + \ldots\right) u_t \\ &= u_t + \frac{13}{10}u_{t-1} + \frac{129}{100}u_{t-2} + \ldots \end{aligned}$$

Truncating this expansion at a sufficiently large lag, we then proceed in the same way as in calculating the acvs of an MA process.

4.3 The ARMA(p, q) process

Let

$$\begin{aligned} \Phi(L) &= 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p \\ \Theta(L) &= 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q \end{aligned}$$

then an ARMA(p, q) process has realisations such that

$$\Phi(L)x_t = \Theta(L)u_t, \qquad u_t \sim w.n.(0, \sigma_u^2).$$

We assume without loss of generality that the polynomials $\Phi(z)$ and $\Theta(z)$ have no common roots.[23]

Moreover, if $\Phi(z)$ has no roots inside the unit circle, we know that we can invert the AR polynomial

$$x_t = \Phi(L)^{-1}\Theta(L)u_t = \left[\Phi(L)^{-1} + \theta_1\Phi(L)^{-1}L + \theta_2\Phi(L)^{-1}L^2 + \ldots + \theta_q\Phi(L)^{-1}L^q\right] u_t$$

which is MA of infinite order since in general $\Phi(L)^{-1}$ has an infinite order (cfr with the AR(p) case). Thus, we have a MA representation which is stationary when the AR polynomial has no roots inside the unit circle, which is then the condition also for stationarity and causality, we then say that the ARMA is causal. With this definition it should be clear that $\mathrm{E}[X_t] = 0$. If instead $\mathrm{E}[X_t] = \mu$ then we write

$$x_t = \Phi(L)^{-1}\Theta(L)u_t + \mu.$$

Everything we say about ARMA is for the zero mean case, otherwise everything hold for the process $\{X_t - \mu\}$.

Notice that no requirement is made for the roots of the MA polynomial. Take for example the two models

$$x_t = (1 + 0.5L)u_t \qquad x_t = (1 + 2L)u_t$$

[23] Assume that there is a common root in say $1/\delta$, then we can write $\Phi(L) = (1 - \delta L)\Phi^*(L)$ where $\Phi^*(L)$ has degree $p - 1$, and similarly $\Theta(L) = (1 - \delta L)\Theta^*(L)$ where $\Theta^*(L)$ has degree $q - 1$. The ARMA model then reads

$$(1 - \delta L)\Phi^*(L)x_t = (1 - \delta L)\Theta^*(L)u_t$$

and the terms $(1 - \delta L)$ cancel out, so we in fact have an ARMA$(p - 1, q - 1)$.

These two polynomials have roots in -2 and -0.5 respectively, but both are stationary and we know that have the same acvs. In general if all the roots of $\Theta(z)$ are outside the unit circle we say that the MA satisfies the invertibility condition, i.e there exists the inverse polynomial $\Theta(z)^{-1}$ such that

$$\Theta(L)^{-1}\Phi(L)x_t = u_t$$

which is the infinite AR representation. We then say that the ARMA is invertible. To be more precise we say that the MA is invertible in the past as the polynomial $\Theta(L)$ will have only positive power of L. When the roots of $\Theta(z)$ are inside the unit circle an inverse exists but will involve negative powers of L thus we have invertibility but in the future a case which we will not consider as we usually do not consider stationary but non-causal processes. When $\Theta(z)$ has a root on the unit circle the MA is never invertible.

Summing up we have the following.

	AR(p)	MA(q)	ARMA(p, q)
Stationarity and causality	roots of $\Phi(z)$ outside $\|z\| \leq 1$	always	roots of $\Phi(z)$ outside $\|z\| \leq 1$
Invertibility in the past	always	roots of $\Theta(z)$ outside $\|z\| \leq 1$	roots of $\Theta(z)$ outside $\|z\| \leq 1$

Why are stationarity and invertibility desirable?

1. Stationarity — because it assumes that the joint distribution of x_t (or second–order properties) does not change over time. Therefore, we can average over time to obtain better estimates of the characteristics of the process.

2. Invertibility — because it permits us to represent our process as AR, and AR processes are often "easy" to estimate and forecast.

From the ARMA(p, q) if it is causal we can write

$$x_t = \Phi(L)^{-1}\Theta(L)u_t = \Psi(L)u_t, \qquad u_t \sim wn(0, \sigma_u^2), \tag{14}$$

where $\Psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$ and it can be proved that since the ARMA is causal then $\sum_{j=0}^{\infty} |\psi_j| < \infty$. This is because the coefficients have a geometric decrease (think of the AR(1) case where $\psi_j = \phi_1^j$ and see also the ARMA(1,1) below). Moreover, if the ARMA is invertible then u_t is the innovation of $\{X_t\}$ and the MA(∞) is the Wold decomposition of an ARMA.

The coefficients ψ_j can be computed as follows. Start from $\Psi(z) = \Phi(z)^{-1}\Theta(z)$ then

$$(1 - \phi_1 z - \ldots - \phi_p z^p)(\psi_0 + \psi_1 z + \ldots) = (1 + \theta_1 z + \ldots + \theta_q z^q)$$

which shows that

$$\begin{aligned} 1 &= \psi_0 \\ \theta_1 &= \psi_1 - \psi_0\phi_1 \\ \theta_2 &= \psi_2 - \psi_1\phi_1 - \psi_0\phi_2 \\ \dots & \end{aligned}$$

or equivalently

$$\theta_j = \psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} \qquad j = 0, 1, \dots$$

where $\theta_0 = 1$, $\theta_j = 0$ if $j > q$ and $\psi_j = 0$ if $j < 0$. Thus, we have written a causal ARMA(p, q) as an MA(∞).

For an ARMA(1,1) we have

$$(1 - \phi_1 z)(\psi_0 + \psi_1 z + \dots) = (1 + \theta_1 z)$$

which shows that[24]

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 &= \phi_1 + \theta_1 \\ \psi_2 &= (\phi_1 + \theta_1)\phi_1 \\ \dots & \\ \psi_j &= (\phi_1 + \theta_1)\phi_1^{j-1} \end{aligned} \tag{15}$$

which implies

$$x_t = u_t + (\phi_1 + \theta_1) \sum_{j=1}^{\infty} \phi_1^{j-1} u_{t-j} \tag{16}$$

In this case causality implies $|\phi_1| < 1$ and therefore $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

Below are some examples to study invertibility and stationarity of ARMA

1. Consider the following process

   $$x_t = u_t - 1.3u_{t-1} + 0.4u_{t-2}$$

   this is an MA(2) thus it is stationary. Writing this in L notation:

   $$x_t = (1 - 1.3L + 0.4L^2)u_t = \Theta(L)u_t$$

   to check if invertible, find roots of $\Theta(z) = 1 - 1.3z + 0.4z^2$,

   $$\begin{aligned} 1 - 1.3z + 0.4z^2 &= 0 \\ 4z^2 - 13z + 10 &= 0 \\ (4z - 5)(z - 2) &= 0 \end{aligned}$$

[24] Use the fact that $(1 - \phi_1 z)^{-1} = 1 + \phi_1 z + \phi_1^2 z^2 + \dots$

the roots of $\Theta(z)$ are $z = 2$ and $z = 5/4$, which are both outside the unit circle hence the process is invertible.[25]

2. Determine whether the following model is stationary and/or invertible,

$$x_t = 1.3x_{t-1} - 0.4x_{t-2} + u_t - 1.5u_{t-1}.$$

Writing in L notation

$$(1 - 1.3L + 0.4L^2)x_t = (1 - 1.5L)u_t$$

we have

$$\Phi(z) = 1 - 1.3z + 0.4z^2$$

with roots $z = 2$ and $5/4$ (from previous example), so the roots of $\Phi(z) = 0$ both lie outside the unit circle, therefore model is stationary, and

$$\Theta(z) = 1 - 1.5z = 1 - \frac{3}{2}z,$$

so the root of $\Theta(z) = 0$ is given by $z = 2/3$ which lies inside the unit circle and the model is not invertible (at least in the past). Notice that there is an inverse of $\Theta(z)$ also in this case

$$\Theta(z)^{-1} = \frac{1}{1 - \frac{3}{2}z} = \frac{-\frac{2}{3}z^{-1}}{1 - \frac{2}{3}z^{-1}} = -\frac{2}{3}z^{-1}\left(1 + \frac{2}{3}z^{-1} + \left(\frac{2}{3}\right)^2 z^{-2} + \dots\right)$$

and when applied to x_t (assume for simplicity that the AR part is not present) we have (substitute z with L)

$$\Theta(L)^{-1}x_t = -\frac{2}{3}L^{-1}\left(1 + \frac{2}{3}L^{-1} + \left(\frac{2}{3}\right)^2 L^{-2} + \dots\right)x_t = -\frac{2}{3}\left(x_{t+1} + \frac{2}{3}x_{t+2} + \left(\frac{2}{3}\right)^2 x_{t+3} + \dots\right)$$

the inverse of an MA exists but requires future values of x_t.

4.3.1 Autocovariance function

The following are three ways of computing acvs of an ARMA(p, q).

1. Start from the MA(∞) representation (14). Then, using the formula for acvs of MA we have

$$\gamma_h^x = \mathrm{E}[X_{t+h}X_t] = \sigma_u^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$$

Consider an ARMA(1,1)

$$x_t - \phi_1 x_{t-1} = u_t + \theta_1 u_{t-1}, \qquad u_t \sim wn(0, \sigma_u^2),$$

[25] The roots of a polynomial of order 2 as $az^2 + bz + c$ are given by

$$z = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

with $|\phi_1| < 1$. Then, using (16),

$$\begin{aligned}
\gamma_0^x &= \sigma_u^2 \sum_{j=0}^{\infty} \psi_j^2 \\
&= \sigma_u^2 \left\{ 1 + \sum_{j=1}^{\infty} \left[(\phi_1 + \theta_1)\phi_1^{j-1} \right]^2 \right\} \\
&= \sigma_u^2 \left\{ 1 + (\phi_1 + \theta_1)^2 \sum_{j=1}^{\infty} \phi_1^{2j-2} \right\} \\
&= \sigma_u^2 \left\{ 1 + (\phi_1 + \theta_1)^2 \sum_{j=0}^{\infty} \phi_1^{2j} \right\} \\
&= \sigma_u^2 \left\{ 1 + \frac{(\phi_1 + \theta_1)^2}{1 - \phi_1^2} \right\}
\end{aligned}$$

$$\begin{aligned}
\gamma_1^x &= \sigma_u^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} \\
&= \sigma_u^2 \left\{ \psi_1 + \sum_{j=1}^{\infty} \psi_j \psi_{j+1} \right\} \\
&= \sigma_u^2 \left\{ \phi_1 + \theta_1 + \sum_{j=1}^{\infty} (\phi_1 + \theta_1)^2 \phi_1^{j-1} \phi_1^j \right\} \\
&= \sigma_u^2 \left\{ \phi_1 + \theta_1 + (\phi_1 + \theta_1)^2 (\phi_1 + \phi_1 \phi_1^2 + \phi_1^2 \phi_1^3 + \ldots) \right\} \\
&= \sigma_u^2 \left\{ \phi_1 + \theta_1 + (\phi_1 + \theta_1)^2 \phi_1 \sum_{j=0}^{\infty} \phi_1^{2j} \right\} \\
&= \sigma_u^2 \left\{ \phi_1 + \theta_1 + \frac{(\phi_1 + \theta_1)^2 \phi_1}{1 - \phi_1^2} \right\}
\end{aligned}$$

$$
\begin{aligned}
\gamma_2^x &= \sigma_u^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+2} \\
&= \sigma_u^2 \left\{ \psi_2 + \sum_{j=1}^{\infty} \psi_j \psi_{j+2} \right\} \\
&= \sigma_u^2 \left\{ (\phi_1+\theta_1)\phi_1 + \sum_{j=1}^{\infty} (\phi_1+\theta_1)^2 \phi_1^{j+1} \phi_1^j \right\} \\
&= \sigma_u^2 \left\{ (\phi_1+\theta_1)\phi_1 + (\phi_1+\theta_1)^2 (\phi_1^2\phi_1 + \phi_1^3\phi_1^2 + \phi_1^4\phi_1^3 + \ldots) \right\} \\
&= \sigma_u^2 \left\{ (\phi_1+\theta_1)\phi_1 + (\phi_1+\theta_1)^2 \phi_1^2 \sum_{j=0}^{\infty} \phi_1^{2j} \right\} \\
&= \sigma_u^2 \left\{ (\phi_1+\theta_1)\phi_1 + \frac{(\phi_1+\theta_1)^2\phi_1^2}{1-\phi_1^2} \right\} \\
&= \phi_1 \gamma_1^x
\end{aligned}
$$

and in general $\gamma_h^x = \phi_1^{h-1}\gamma_1^x$ for $h \geq 2$.

2. Multiply each side of the ARMA equation by x_{t-h} and take expectations. For an ARMA(1,1) we have

$$\mathrm{E}[x_{t-h}(x_t - \phi_1 x_{t-1})] = \mathrm{E}[x_{t-h}(u_t + \theta_1 u_{t-1})]$$

which implies

$$
\begin{aligned}
\gamma_0^x - \phi_1\gamma_1^x &= \mathrm{E}[\Psi(L)u_t(u_t + \theta_1 u_{t-1})] \\
&= \mathrm{E}[(u_t + \psi_1 u_{t-1} + \ldots)(u_t + \theta_1 u_{t-1})] \\
&= \sigma_u^2 + \sigma_u^2 \psi_1 \theta_1 \\
&= \sigma_u^2(1 + (\theta_1 + \phi_1)\theta_1)
\end{aligned}
$$

and

$$
\begin{aligned}
\gamma_1^x - \phi_1\gamma_0^x &= \mathrm{E}[\Psi(L)u_{t-1}(u_t + \theta_1 u_{t-1})] \\
&= \mathrm{E}[(u_{t-1} + \psi_1 u_{t-2} + \ldots)(u_t + \theta_1 u_{t-1})] \\
&= \sigma_u^2 \theta_1
\end{aligned}
$$

The previous expression are found using these two relations.

Consider as an example the ARMA(2,1)

$$x_t = \phi x_{t-2} + \theta u_{t-1} + u_t, \quad u_t \sim w.n.(0, \sigma^2)$$

Then we can compute the acvs by multiplying each side of the ARMA equation by x_{t-h} and taking expectations. Notice that

$$\mathrm{E}[x_t u_t] = \sigma^2, \qquad \mathrm{E}[x_t u_{t-1}] = \theta\sigma^2, \qquad \mathrm{E}[x_{t-h}u_t] = 0 \ \text{ for } \ h \neq 0.$$

Therefore we have the equations

$$
\begin{aligned}
\mathrm{E}[x_t x_t] &= \gamma_0 = \phi\gamma_2 + \theta^2\sigma^2 + \sigma^2 \\
\mathrm{E}[x_t x_{t-1}] &= \gamma_1 = \phi\gamma_1 + \theta\sigma^2 \\
\mathrm{E}[x_t x_{t-2}] &= \gamma_2 = \phi\gamma_0 \\
\mathrm{E}[x_t x_{t-3}] &= \gamma_3 = \phi\gamma_1
\end{aligned}
$$

Then by using the first three equations we obtain

$$\gamma_0 = \sigma^2 \frac{1+\theta^2}{1-\phi^2}$$
$$\gamma_1 = \sigma^2 \frac{\theta\sigma^2}{1-\phi}$$
$$\gamma_2 = \phi\gamma_0,$$

and the acs are $\rho_0 = 1$, $\rho_1 = \frac{\theta(1+\phi)}{1+\theta^2}$, $\rho_2 = \phi$, and $\rho_3 = \phi\rho_1$ and so on.

3. The third way uses the roots of the AR polynomial. Let us consider the example with AR(2) before and let's add also an invertible MA component

$$x_t = 1.3x_{t-1} - 0.4x_{t-2} + u_t - 0.25u_{t-1}.$$

The AR part is stationary with roots in $z = 2$ and $z = 5/4$ and the MA part is invertible with root in $z = 4$ then

$$x_t = \frac{\left(1 - \frac{L}{4}\right)}{\left(1 - \frac{L}{2}\right)\left(1 - \frac{4}{5}L\right)} u_t$$

and using Taylor expansion and collecting terms we have

$$\begin{aligned} x_t &= \left(1 + \frac{13}{10}L + \frac{129}{100}L^2 + \ldots\right)\left(1 - \frac{L}{4}\right) u_t \\ &= \left[1 + \left(\frac{13}{10} - \frac{1}{4}\right)L + \left(\frac{129}{100} - \frac{13}{40}L^2 + \ldots\right)\right] u_t \end{aligned}$$

Truncating this expansion at a sufficiently large lag, we then proceed in the same way as in calculating the acvs of an MA process.

4.4 Partial autocorrelation

In the section on MA processes, we saw that the acvs sequence cuts off after some lag h. Also, we saw that for an AR(1) process, it never cut off to zero, but decayed exponentially. The same result can be proved for a general AR process. For purposes of model identification, it would be useful to have a quantity which did cut off to zero for a general autoregressive process AR(p). One such quantity is the partial autocorrelation function (or sequence), pacf. For a process x_t, it is defined by

$$\begin{aligned} \pi_1^x &= \text{Corr}(X_2, X_1) \\ \pi_2^x &= \text{Corr}(X_3 - \text{E}[X_3|X_2], X_1 - \text{E}[X_1|X_2]) \\ \pi_3^x &= \text{Corr}(X_4 - \text{E}[X_4|X_3, X_2], X_1 - \text{E}[X_1|X_3, X_2]) \\ &\ldots \end{aligned}$$

The interpretation is that $\text{E}[X_4|X_3, X_2]$ is the part of X_4 that is explained by X_3, X_2 (or more formally, it is the prediction of X_4 based on X_3, X_2). Thus $X_4 - \text{E}[X_4|X_3, X_2]$ is the part of X_4 that is unexplained (un–predicted) by X_2, X_3. Thus the partial autocorrelation at lag h

$$\pi_h^x = \text{Corr}(X_{t+h} - \text{E}[X_{t+h}|X_{t+h-1}, X_{t+h-2}, \ldots, X_{t+1}], X_t - \text{E}[X_t|X_{t+h-1}, X_{t+h-2}, \ldots, X_{t+1}])$$

is the correlation of those portions of X_t, X_{t+h} which are unexplained by the intermediate variables $X_{t+1}, \ldots, X_{t+h-1}$.

In a non time series setting, a partial correlation can be seen in a linear regression framework. Assume to have n random variables variables $X_1 \dots X_n$ all with the same variance and zero mean and to have m observations for each. Consider the regression

$$x_{1i} = \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_n x_{ni} + v_{1i}, \qquad i = 1, \ldots, m$$

where we also assume conditional mean independence $\mathrm{E}[v_1|X_k] = 0$ for $k \neq 1$. Then the partial correlation of X_1 with X_n given all the other variables is β_n. Indeed, take the case $n = 3$, we have

$$X_1 - \mathrm{E}[X_1|X_2] = X_1 - \beta_2 X_2 - \beta_3 \mathrm{E}[X_3|X_2] = \beta_3(X_3 - \mathrm{E}[X_3|X_2]) + v_1,$$

then using the fact that $\mathrm{E}[v_1|X_2] = \mathrm{E}[v_1|X_3] = 0$, the partial correlation of X_1 with X_3 is given by

$$\mathrm{Corr}(X_1 - \mathrm{E}[X_1|X_2], X_3 - \mathrm{E}[X_3|X_2]) = \beta_3 \mathrm{Corr}(X_3 - \mathrm{E}[X_3|X_2], X_3 - \mathrm{E}[X_3|X_2]) = \beta_3.$$

In a time series the pacf sequence is defined as $\pi_0^x = 1$ and $\pi_h^x = \phi_{hh}$, for $h \geq 1$, where ϕ_{hh} is the last coefficient in the regression

$$x_t = \phi_{h1} x_{t-1} + \phi_{h2} x_{t-2} + \ldots + \phi_{hh} x_{t-h} + e_t. \tag{17}$$

Therefore, for an AR(p)

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + u_t, \quad u_t \sim w.n.(0, \sigma_u^2),$$

from (19) we have $\pi_p^x = \phi_p$. And we immediately see that for $h > p$ the pacf is $\pi_h^x = 0$. However, at lags $h < p$ in general $\pi_h^p \neq \phi_h$. If however, the u's are i.i.d. Gaussian, then it can be shown that $\pi_h^x = \phi_h$ for any $h \leq p$.

Usually in AR(p) models we say that the error is white noise, therefore we always have $\mathrm{Cov}(u_t, X_{t-h}) = 0$ for $h > 0$. However, in the above definition of pacf we are implicitly assuming that for ARMA processes we have $\mathrm{E}[u_t|X_{t-1}, X_{t-2} \ldots] = 0$ which a stronger requirement than no-correlation but weaker than independence. Intuitively this is a reasonable assumption since u_t is the innovation so it should not contain relevant information about x_t (at least from a linear point of view).

If the process is an MA or an ARMA and we want an analytical formula for the pacfs, it is not an easy task in general. A simple case is an MA(1) for which we have

$$x_t = u_t + \theta x_{t-1}, \qquad u_t \sim wn(0, \sigma_u^2),$$

we find

$$\pi_h^x = \frac{-(-\theta)^h}{1 + \theta^2 + \ldots + \theta^{2h}}.$$

See the next Chapter for a proof.

5 Estimation of ARMA processes

We have seen in Chapter 3 the sample acvs estimator. Now we will use these to estimate the coefficients of a generic ARMA(p, q) model. Before however we focus on the AR(p) model which is easier to estimate and general enough to cover most cases:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + u_t, \qquad u_t \sim wn(0, \sigma_u^2)$$

It is possible to prove using spectral theory that any time series sufficiently regular can be approximated well by an AR(p) model if p is large enough. Later we use the estimator of an AR as a pre-estimator for an ARMA.

We always suppose that $\mathrm{E}[X_t] = 0$ unless stated otherwise. If $\mathrm{E}[X_t] = \mu \neq 0$ then everything that follows applies to $\{X_t - \mu\}$ or if we deal with estimated quantities estimation must be applied to $\{X_t - \bar{x}\}$ where $\bar{x}$ is the sample mean studied in Chapter 3.

We also assume to observe a time series for T periods thus we observe $(x_1, x_2, \ldots, x_T)$.

5.1 Yule Walker estimator

We start by multiplying the defining equation of an AR(p) by x_{t-h}

$$x_t x_{t-h} = \sum_{j=1}^{p} \phi_j x_{t-j} x_{t-h} + u_t x_{t-h}$$

Taking expectations, for $h > 0$ and recalling $\mathrm{E}[u_t X_{t-h}] = 0$:

$$\gamma_h^x = \sum_{j=1}^{p} \phi_j \gamma_{h-j}^x$$

Let $h = 1, 2, \ldots, p$ and remember that $\gamma_{-h}^x = \gamma_h^x$ then

$$\begin{aligned}
\gamma_1^x &= \phi_1 \gamma_0^x + \phi_2 \gamma_1^x + \ldots + \phi_p \gamma_{p-1}^x \\
\gamma_2^x &= \phi_1 \gamma_1^x + \phi_2 \gamma_0^x + \ldots + \phi_p \gamma_{p-2}^x \\
&\vdots \\
\gamma_p^x &= \phi_1 \gamma_{p-1}^x + \phi_2 \gamma_{p-2}^x + \ldots + \phi_p \gamma_0^x
\end{aligned}$$

these are the Yule Walker equations and in matrix notation they read

$$\boldsymbol{\gamma}_p^x = \boldsymbol{\Gamma}_p^x \boldsymbol{\phi}$$

where $\boldsymbol{\gamma}_p^x = (\gamma_1^x \, \gamma_2^x \ldots \gamma_p^x)'$, $\boldsymbol{\phi} = (\phi_1 \, \phi_2 \ldots \phi_p)'$ and

$$\boldsymbol{\Gamma}_p^x = \begin{pmatrix} \gamma_0^x & \gamma_1^x & \cdots & \gamma_{p-1}^x \\ \gamma_1^x & \gamma_0^x & \cdots & \gamma_{p-2}^x \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1}^x & \gamma_{p-2}^x & \cdots & \gamma_0^x \end{pmatrix}$$

Notice that this is a symmetric Toeplitz matrix which we have met already (see Chapter 2). All elements on a given diagonal are the same.

Given that x_t has zero mean we have the estimated acvs, for $h = 0, 1, \ldots, p$

$$\widehat{\gamma}_h^x = \frac{1}{T} \sum_{t=1}^{T-|h|} x_t x_{t+|h|} \tag{18}$$

and substitute these for the γ_h^x, $\boldsymbol{\gamma}_p^x$ and $\boldsymbol{\Gamma}_p^x$ to obtain $\widehat{\boldsymbol{\gamma}}_p^x$ and $\widehat{\boldsymbol{\Gamma}}_p^x$ from which we estimate $\boldsymbol{\phi}$ as

$$\widehat{\boldsymbol{\phi}} = \left(\widehat{\boldsymbol{\Gamma}}_p^x\right)^{-1} \widehat{\boldsymbol{\gamma}}_p^x$$

Notice that existence of the inverse matrix is crucial therefore we must define an estimator of acvs such that $\widehat{\mathbf{\Gamma}}_p^x$ is positive definite, which is indeed the case for the one used here (see the discussion in Chapter 3). If we estimate the Yule Walker equations for any time series and we truncate at lag h we are actually estimating the pacf at lag h (see Chapter 4). So in the AR(p) case we have $\widehat{\phi}_p = \widehat{\pi}_p^x$.

We have the following central limit theorem for $\widehat{\boldsymbol{\phi}}$

$$\sqrt{T}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{d} N(\mathbf{0}, \sigma_u^2(\mathbf{\Gamma}_p^x)^{-1}), \qquad \text{as } T \to \infty.$$

So for example, for an AR(1) we have just one parameter and

$$\sqrt{T}(\widehat{\phi}_1 - \phi_1) \xrightarrow{d} N(0, \sigma_u^2(\gamma_0^x)^{-1}), \qquad \text{as } T \to \infty,$$

which gives the asymptotic variance of $\widehat{\phi}_1$ as

$$\text{AVar}(\widehat{\phi}_1) = \frac{\sigma_u^2(\gamma_0^x)^{-1}}{T} = \frac{\sigma_u^2(1-\phi_1^2)}{\sigma_u^2 T} = \frac{(1-\phi_1^2)}{T}.$$

A 95% confidence interval for the AR(1) parameter is then given by

$$\widehat{\phi}_1 \pm 1.96 \sqrt{\frac{(1-\phi_1^2)}{T}}.$$

In general, however, in order to compute confidence intervals we then need to estimate the asymptotic variance in particular $(\mathbf{\Gamma}_p^x)^{-1}$ (see (18) above) and the variance of the innovations σ_u^2. To estimate the latter, we multiply the defining equation by x_t and take expectations to obtain

$$\gamma_0^x = \sum_{j=1}^{p} \phi_j \gamma_j^x + \text{E}[u_t X_t] = \sum_{j=1}^{p} \phi_j \gamma_j^x + \sigma_u^2$$

so that as an estimator for σ_u^2 we take

$$\widehat{\sigma}_u^2 = \widehat{\gamma}_0^x - \sum_{j=1}^{p} \widehat{\phi}_j \widehat{\gamma}_j^x$$

The estimators $\widehat{\boldsymbol{\phi}}$ and $\widehat{\sigma}_u^2$ are called Yule Walker estimators.[26]

5.1.1 Yule Walker estimator for partial autocorrelation functions

From the previous Chapter we know that, for a stationary process, given the linear predictor

$$x_t = \phi_{h1} x_{t-1} + \phi_{h2} x_{t-2} + \ldots + \phi_{hh} x_{t-h} + e_t, \tag{19}$$

the partial autocorrelation at lag h is $\pi_h^x = \phi_{hh}$. Since (19) looks like an AR(h) model it can be estimated by solving the Yule Walker equations up to lag h to get the estimated $\widehat{\phi}_{hh}$ which is the last entry of the vector $\widehat{\boldsymbol{\phi}}_h$ solution of the equation

$$\widehat{\boldsymbol{\phi}}_h = \left(\widehat{\mathbf{\Gamma}}_h^x\right)^{-1} \widehat{\boldsymbol{\gamma}}_h^x. \tag{20}$$

Thus once we estimate an AR(p) via Yule Waker equations the last estimated coefficient is also the lag p estimated pacf: $\widehat{\pi}_p^x = \widehat{\phi}_p$. However, unless x_t is Gaussian, at lags $h < p$ in general $\pi_h^x \neq \phi_h$ and we have to solve (20) for every lag.[27]

[26] Alternative methods to estimate the coefficients of an AR are given by the Durbin Levinson algorithm which is a recursion bypassing the inversion of a Toeplitz matrix and the Burg algorithm which estimates the pacf from which the AR coefficients can be computed (see below).

[27] For pacf sometimes the equation (20) is written using acs, i.e. by multiplying and dividing the right hand side by the variance γ_0^x.

If we have to estimate pacf for a generic ARMA, we still have to solve equation (20) for any h using the acvs of the process written as functions of the parameters of the model and then we express the solutions, i.e. the pacf, as function of the parameters. Given estimates of the ARMA coefficients (obtained using the methods outlined below), we have estimates of the pacf. In the last Chapter we saw a formula for the MA(1) case. As an example let us compute the pacf of an MA(1) at lag 3. We have $\gamma_0^x = \sigma_u^2(1+\theta^2)$ and $\gamma_1 = \sigma_u^2\theta$. Then we have to solve (no need to invert the matrix)

$$\boldsymbol{\Gamma}_3^x \boldsymbol{\phi}_3 = \boldsymbol{\gamma}_3^x$$

which is equivalent to

$$\sigma_u^2 \begin{pmatrix} 1+\theta^2 & \theta & 0 \\ \theta & 1+\theta^2 & \theta \\ 0 & \theta & 1+\theta^2 \end{pmatrix} \begin{pmatrix} \phi_{31} \\ \phi_{32} \\ \phi_{33} \end{pmatrix} = \sigma_u^2 \begin{pmatrix} \theta \\ 0 \\ 0 \end{pmatrix}$$

this is a system of three equations with three unknowns and by solving it we get

$$\pi_3^x = \phi_{33} = \frac{\theta}{1+\theta^2+\theta^4+\theta^6}$$

But notice that for the pacf at lag 2 $\pi_2^x \neq \phi_{32}$ and we have to solve another system of equations.

5.2 Least Squares estimator

Start from the defining equation

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + u_t, \qquad u_t \sim wn(0, \sigma_u^2)$$

then we can formulate an appropriate least squares model in terms of data $x_1, x_2, \ldots, x_T$ as follows:

$$\mathbf{x} = \mathbf{F}\boldsymbol{\phi} + \mathbf{u},$$

where $\mathbf{x} = (x_{p+1}\, x_{p+2} \ldots x_T)'$, $\boldsymbol{\phi} = (\phi_1\, \phi_2 \ldots \phi_p)'$, $\mathbf{u} = (u_{p+1}\, u_{p+2} \ldots u_T)'$ and

$$\mathbf{F} = \begin{pmatrix} x_p & x_{p-1} & \ldots & x_1 \\ x_{p+1} & x_p & \ldots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{T-1} & x_{T-2} & \ldots & x_{T-p} \end{pmatrix}$$

which is a $(T-p) \times p$ matrix, thus we use $T-p$ observations and p regressors, so once again we need $p << T$. Notice that if $\mathrm{E}[X_t] = \mu \neq 0$ we have to add a constant to the regression, thus we add one parameter and a regressor equal to 1, i.e. we add a column of ones to the matrix $\mathbf{F}$.

We can then estimate $\boldsymbol{\phi}$ by finding that $\boldsymbol{\phi}$ such that it minimises (least squares estimation)

$$SS(\boldsymbol{\phi}) = \sum_{t=p+1}^{T} \left(x_t - \sum_{j=1}^{p} \phi_j x_{t-j} \right)^2 = (\mathbf{x} - \mathbf{F}\boldsymbol{\phi})'(\mathbf{x} - \mathbf{F}\boldsymbol{\phi}).$$

Notice that $SS(\boldsymbol{\phi}) = \sum_{t=p+1}^{T} u_t^2$ is the sum of squared residuals. If we denote the vector that minimizes $SS(\boldsymbol{\phi})$ as $\widehat{\boldsymbol{\phi}}$, standard least squares theory tells us that it is given by

$$\widehat{\boldsymbol{\phi}} = \left(\mathbf{F}'\mathbf{F}\right)^{-1} \mathbf{F}'\mathbf{x}.$$

As an example in the AR(1) case, i.e. $p = 1$, if we make explicit the matrix products above we have

$$\widehat{\phi}_1 = \left(\sum_{t=2}^{T} x_{t-1}^2 \right)^{-1} \left(\sum_{t=2}^{T} x_{t-1} x_t \right)$$

and by rewriting as

$$\widehat{\phi}_1 = \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1}^2 \right)^{-1} \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1} x_t \right) = \frac{\widehat{\gamma}_1^x}{\widehat{\gamma}_0^x}$$

so once again to compute estimators we need to estimate acvs. We also see that as $T \to \infty$ by the Law of Large Numbers we have

$$\widehat{\phi}_1 \xrightarrow{p} \frac{\gamma_1^x}{\gamma_0^x}.$$

Notice however that $\widehat{\phi}_1$ is biased. Indeed,

$$\begin{aligned}
\widehat{\phi}_1 &= \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1}^2 \right)^{-1} \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1} x_t \right) = \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1}^2 \right)^{-1} \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1} (\phi_1 x_{t-1} + u_t) \right) \\
&= \phi_1 + \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1}^2 \right)^{-1} \left(\frac{1}{T} \sum_{t=2}^{T} x_{t-1} u_t \right) \\
&= \phi_1 + \sum_{t=2}^{T} \frac{x_{t-1}}{\sum_{t=2}^{T} x_{t-1}^2} u_t.
\end{aligned}$$

Now, while u_t is not correlated to x_{t-1} because it is a white noise and therefore $\mathrm{E}[x_{t-1} u_t] = 0$, in general u_t is not uncorrelated with $\sum_{t=2}^{T} x_{t-1}^2$ since this term contains $(x_1 \dots x_{T-1})$ which is correlated with u_t for $t = 2, \dots, T$.[28] Indeed, if ϕ_1 is positive, then a positive shock to u_t raises current and future values of x_t , all of which are in the $\sum_{t=2}^{T} x_{t-1}^2$. This means there is a negative correlation between u_t and $\sum_{t=2}^{T} \frac{x_{t-1}}{\sum_{t=2}^{T} x_{t-1}^2}$ so $\mathrm{E}[\widehat{\phi}_1] < \phi_1$.

The estimated coefficients are also asymptotically normal, and in order to show this we need to use a particular Central Limit Theorem for martingale difference sequences and we must assume that u_t is a martingale difference sequence: $\mathrm{E}[u_t | x_{t-1}, x_{t-2}, \dots] = 0$.[29] For example, in the AR(1) case $z_t = u_t x_{t-1}$ is such that

$$\mathrm{E}[z_t | x_{t-1}] = \mathrm{E}[u_t x_{t-1} | x_{t-1}] = x_{t-1} \mathrm{E}[u_t | x_{t-1}] = 0.$$

Then, it is possible to prove that

$$\frac{1}{T} \sum_{t=2}^{T} x_{t-1}^2 \xrightarrow{p} \mathrm{E}[x_{t-1}^2] = \Omega, \qquad \text{as } T \to \infty,$$

[28] For proving unbiasedness we should compute

$$\begin{aligned}
\mathrm{E}[\widehat{\phi}_1 - \phi_1] &= \mathrm{E}\left[\sum_{t=2}^{T} \frac{x_{t-1}}{\sum_{t=2}^{T} x_{t-1}^2} u_t \right] = \mathrm{E}\left[\sum_{t=2}^{T} \mathrm{E}\left[\frac{x_{t-1}}{\sum_{t=2}^{T} x_{t-1}^2} u_t \middle| x_{T-1} \dots x_1 \right] \right] \\
&= \mathrm{E}\left[\sum_{t=2}^{T} \frac{x_{t-1} \mathrm{E}[u_t | x_{T-1} \dots x_1]}{\sum_{t=2}^{T} x_{t-1}^2} \right] \neq 0
\end{aligned} \tag{21}$$

since $\mathrm{E}[u_t | x_{T-1} \dots x_1] \neq 0$ in an AR model. To have the condition $\mathrm{E}[u_t | x_{T-1} \dots x_1] = 0$ we would need strong exogeneity of the regressors, which is impossible in time series. Notice that strong exogeneity is stronger than asking for the martingale difference property $\mathrm{E}[u_t | x_{t-1} \dots x_1] = 0$ which is called weak exogeneity and could be assumed.

[29] If this is true then we are actually estimating the conditional mean of x_t as $\mathrm{E}[x_t | x_{t-1}] = \phi_1 x_{t-1}$.

because x_t is ergodic since its MA(∞) representation has summable coefficients, and that

$$\frac{1}{\sqrt{T}}\sum_{t=2}^{T} x_{t-1}u_t \xrightarrow{p} N(0,V) \qquad \text{as } T\to\infty,$$

with $V = \mathrm{E}[x_{t-1}^2 u_t^2]$, because of a Central Limit Theorem for martingale difference sequences and since

$$\frac{1}{T}\sum_{t=2}^{T} x_{t-1}u_t \xrightarrow{p} 0, \qquad \text{as } T\to\infty,$$

again because of ergodicity and since $\mathrm{E}[x_{t-1}u_t] = 0$. For these results we also need $\mathrm{E}[x_t^4] < \infty$ to use the ergodic theorem for the squares.

Therefore,

$$\sqrt{T}(\widehat{\phi}_1 - \phi_1) \xrightarrow{d} N(0, \Omega^{-2}V), \qquad \text{as } T\to\infty,$$

by Slutsky's theorem.[30] Consistency follows since for any $Z \sim N(0,1)$, we can apply Slutsky's theorem again and

$$(\widehat{\phi}_1 - \phi_1) = \frac{1}{\sqrt{T\Omega^{-2}V}}\sqrt{T\Omega^{-2}V}(\widehat{\phi}_1 - \phi_1) \xrightarrow{d} \lim_{T\to\infty}\frac{1}{\sqrt{T\Omega^{-2}V}}Z = 0,$$

and convergence in distribution to a point is equivalent to convergence in probability.

Moreover, if we also assume independence of u_t and x_{t-1} then $V = \mathrm{E}[x_{t-1}^2 u_t^2] = \mathrm{E}[x_{t-1}^2]\mathrm{E}[u_t^2] = \Omega\sigma^2$ and we have that the asymptotic variance is the usual Gauss-Markov lower bound found in OLS, i.e. $\mathrm{AVar}(\widehat{\phi}_1) = \sigma_u^2/(\Omega T)$. Everything can be generalised to an AR(p).

Summing up, the OLS estimator of AR models is biased, but consistent, is asymptotically normal only if the errors are a martingale difference sequence and is efficient (lowest possible variance) if the errors are an independent sequence. We can estimate σ_u^2 using the sample variance of the residuals of the regression

$$\widehat{\sigma}_u^2 = \frac{(\mathbf{x} - \mathbf{F}\widehat{\boldsymbol{\phi}})'(\mathbf{x} - \mathbf{F}\widehat{\boldsymbol{\phi}})}{T-2p} = \frac{\mathbf{u}'\mathbf{u}}{T-2p} = \frac{1}{T-2p}\sum_{t=p+1}^{T}\widehat{u}_t^2,$$

where for $t = p+1,\ldots,T$ the residuals are given by

$$\widehat{u}_t = x_t - \widehat{\phi}_1 x_{t-1} - \widehat{\phi}_2 x_{t-2} - \ldots - \widehat{\phi}_p x_{t-p}.$$

Notice the degrees of freedom correction at the denominator which is given by the number of observations $T-p$ minus the number of estimated parameters p. However, as $T\to\infty$ also dividing by T yields consistency.

The estimator of $\boldsymbol{\phi}$ obtained by Least Squares is the Maximum Likelihood estimator of an AR(p) process when the innovations u_t are Normally distributed.

If we have an ARMA process the Least Squares estimator can still be used if we have a preliminary estimate of the innovations. The following is the Hannan Rissanen algorithm for estimating an ARMA(p,q) by Least Squares.

1. Fit a high order AR(m) with $m > \max(p,q)$ using Yule Walker method and obtain the estimated vector $\widehat{\boldsymbol{\phi}}_m = (\widehat{\phi}_{1m}\ldots\widehat{\phi}_{mm})'$.

[30] If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$ as $n\to\infty$ then $X_nY_n \xrightarrow{d} aX$.

2. Compute the residuals of the AR(m) model estimated

$$\widehat{v}_t = x_t - \widehat{\phi}_{1m}x_{t-1} - \ldots - \widehat{\phi}_{mm}x_{t-m}, \qquad t = m+1, \ldots, T.$$

3. The vector of all parameters of the ARMA(p, q) is estimated by Least Squares regression

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \theta_1 \widehat{v}_{t-1} + \ldots + \theta_q \widehat{v}_{t-q} + u_t$$

this is done in the way described above but now we have $p + q$ regressors. Call these parameters $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}})$.

4. The residuals of the regression are

$$\widehat{u}_t = x_t - \widehat{\phi}_1 x_{t-1} - \ldots - \widehat{\phi}_p x_{t-p} - \widehat{\theta}_1 \widehat{v}_{t-1} - \ldots - \widehat{\theta}_q \widehat{v}_{t-q}$$

and their variance is estimated as

$$\widehat{\sigma}_u^2 = \frac{1}{T - m - q} \sum_{t=m+1+q}^{T} \widehat{u}_t^2.$$

Notice that it can shown that while for Yule-Walker the resulting estimated AR polynomial $\widehat{\Phi}(z)$ with coefficients $\widehat{\boldsymbol{\phi}}$ is stable in the sense that all its roots are outside the unit circle, this is general not guaranteed for OLS estimates. However the Yule-Walker estimator is also biased and typically performs worse than the OLS.

5.3 Maximum Likelihood estimator

Consider an ARMA(p, q) model

$$\Phi(L)x_t = \Theta(L)u_t, \qquad u_t \sim wn(0, \sigma_u^2)$$

and suppose the vector $\mathbf{x}_T = (x_1 \ldots x_T)'$ is a multivariate Gaussian vector with zero mean and variance covariance matrix given by the Toepliz matrix $\boldsymbol{\Gamma}_T^x \equiv \mathrm{E}[\mathbf{X}_T \mathbf{X}_T']$ which is $T \times T$ and contain in each diagonal the same elements which are the acvs. If the data generating process is an ARMA(p, q) then the matrix is function of the parameters $\boldsymbol{\phi} = (\phi_1 \ldots \phi_p)'$, $\boldsymbol{\theta} = (\theta_1 \ldots \theta_q)'$, i.e.

$$\boldsymbol{\Gamma}_T^x = \boldsymbol{\Gamma}_T^x(\boldsymbol{\phi}, \boldsymbol{\theta}).$$

Under Gaussianity the likelihood of the data is

$$f(\mathbf{x}_T; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_u^2) = \frac{1}{\sqrt{\det(\boldsymbol{\Gamma}_T^x(\boldsymbol{\phi}, \boldsymbol{\theta}))}(2\pi)^{T/2}} \exp\left[-\frac{1}{2}\mathbf{x}_T' \boldsymbol{\Gamma}_T^x(\boldsymbol{\phi}, \boldsymbol{\theta})^{-1} \mathbf{x}_T\right].$$

Once a sample $\mathbf{x}_T = (x_1 \ldots x_T)'$ is observed this function depends only on the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Typically we work with the log-likelihood

$$L_T(\mathbf{x}_T; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_u^2) = \log f(\mathbf{x}_T; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_u^2) \tag{22}$$

and the Maximum Likelihood estimator of the parameters is the value of the parameters that maximizes this function:

$$(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}_u^2) = \arg\max L_T(\mathbf{x}_T; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_u^2).$$

It can be proved that the estimator is consistent

$$(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}) \xrightarrow{p} (\boldsymbol{\phi}, \boldsymbol{\theta}), \qquad \text{as } T \to \infty$$

and is asymptotically Normal

$$\sqrt{T}\left[(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}) - (\boldsymbol{\phi}, \boldsymbol{\theta})\right] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \text{as } T \to \infty,$$

where the form of $\boldsymbol{\Sigma}$ is complex and depends on the derivatives of L_T and on the parameters too.

Such maximisation is in general not easy (unless we have an AR model ($q = 0$) in which case the solution is the Least Squares estimator). The first order conditions are non-linear in the parameters and they have no analytical solutions. Second when T is large writing the $T \times T$ matrix $\boldsymbol{\Gamma}_T^x$ as function of the parameters is not easy (and we need also its inverse an its determinant). Moreover, the maximisation often depends on the starting values of the parameters used as starting point for the maximisation algorithm. For this reason we need preliminary estimators of the parameters. We have seen the Yule Walker as an example for $\boldsymbol{\phi}$ (and we mentioned other).[31] For the MA part there are also preliminary estimators which are given for example by the Innovation algorithm which starts from the Wold representation of the process. Or the Hannan Rissanen algorithm which provides preliminary estimates for all parameters.

Any preliminary estimator is in general less efficient than the Maximum Likelihood estimator, therefore once we have an initial estimate of the parameters we can plug it into the maximisation of the log-likelihood to obtain a more efficient estimator. Saying that an estimator is more efficient means that its asymptotic variance is smaller.

The most efficient estimator is the Maximum Likelihood estimator if the data are Gaussian. But if data are not Gaussian is Maximum Likelihood still a "good" estimator?

It can be proved that if $\mathbf{x}_t$ is distributed according to a distribution of the Exponential family[32] and the ARMA(p, q) is the true underlying model then the Maximum Likelihood estimator is still consistent and asymptotically Normal, but it will have a larger variance, i.e. is no more the most efficient estimator.

5.4 The relation between OLS and ML

For an AR(p) model there is a clear relation between OLS and ML estimation. This can be seen by rewriting the likelihood of $(x_1 \dots x_T)$ as a product of conditional likelihoods, where at each point in time we condition on the past:[33]

$$f(\mathbf{x}_T) = \prod_{t=1}^{T} f_{X_t|X_1 \dots X_{t-1}}(x_t|x_1, \dots, x_{t-1}). \tag{23}$$

[31] It can be generalised for ARMA models too as the Hannan Rissanen generalises the Least Squares procedure.

[32] Distributions of this family are the Gaussian, Exponential, Gamma, Chi-squared, and even Binomial or Poisson, but for example the Student-t does not belong to this family.

[33] For a pdf of two random variables we have

$$f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x) = f_{X|Y}(x|y) f_Y(y)$$

and for three random variables we have

$$f_{XYZ}(x, y, z) = f_{X|YZ}(x|y, z) f_{YZ}(y, z) = f_{Y|XZ}(y|x, z) f_{XZ}(x, z) = f_{Z|XY}(z|x, y) f_{XY}(x, y).$$

Now for any ARMA model with errors that are also Gaussian we can write the linear prediction equation as (see Chapter 3)

$$x_t = P_{t-1}x_t + e_t = \mathrm{E}[x_t|x_{t-1}\dots x_1] + e_t$$

therefore the conditional distribution of x_t given $(x_1 \dots x_{t-1})$ is the same as the distribution of the one-step-ahead prediction error e_t (notice that due to Gaussianity e_t do not depend on the past values and therefore we can use their unconditional distribution). In particular, for an AR(p) we have

$$x_t = c + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + u_t,$$

where we assume u_t to be a Gaussian zero mean white noise with variance σ_u^2, then

$$x_t|x_1 \dots x_{t-1} \sim N(c + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p}, \sigma_u^2).$$

Notice that actually the conditioning set is just given by $x_{t-p} \dots x_{t-1}$. Now consider an AR(1), then the parameters are c, ϕ_1 and σ_u^2. Using (23) we can write the joint log-likelihood (22) as

$$\begin{aligned}
L_T(\mathbf{x}_T; c, \phi_1, \sigma_u^2) &= \log f_{X_1}(x_1) + \sum_{t=2}^{T} \log f_{X_t|X_{t-1}}(x_t|x_{t-1}) \\
&= \log f_{X_1}(x_1) + \sum_{t=1}^{T} \log f_{u_t}(u_t) \\
&= \log f_{X_1}(x_1) - \frac{1}{2}\sum_{t=1}^{T}\left[\log(2\pi) + \log\sigma_u^2 + \frac{(x_t - c - \phi_1 x_{t-1})^2}{\sigma_u^2}\right] \\
&= \log f_{X_1}(x_1) - \frac{1}{2}\sum_{t=1}^{T}\left[\log(2\pi) + \log\sigma_u^2 + \frac{u_t^2}{\sigma_u^2}\right].
\end{aligned}$$

Notice that we cannot write the first term as function of X_0 since we don't have any realisation at $t = 0$. If the first term were zero then the above would be a Gaussian log-likelihood for a linear model with dependent variable x_t and explanatory variable x_{t-1} plus an intercept and residual u_t. We know that for those models the ML estimator coincides with the OLS, therefore we could simply use OLS to estimate the parameters if it were not for the first term. However, for large T the weight of the first term become negligible. Then OLS estimates tend to ML estimates as $T \to \infty$ and they inherit the asymptotic properties.[34] The same reasoning applies for AR(p) models where the first term will be $\log f_{X_1 \dots X_p}(x_1 \dots x_p)$.

If we have an invertible MA process we still know that the white noise leading the model u_t can be expressed as $x_t - P_{t-1}x_t$, i.e. it is also the one-step-ahead prediction error. Thus if we assume that $x_t|x_1 \dots x_{t-1}$ has a Gaussian distribution, then at each point in time we can compute the prediction of x_t given its past and compute the prediction error as $u_t = x_t - P_{t-1}x_t$ given by the model (see also next Chapter for forecasting MA models). Then the log-likelihood can be written using the sequence of prediction errors computed in this way.

5.5 Order selection

Once we have transformed a given process to a stationary one, how do we determine the AR and MA orders?

[34] Moreover, if u_t is a martingale difference sequence, the OLS estimator is consistent even if the residual u_t is not Gaussian and in this case it is not correct to see the OLS as asymptotically equivalent to the Gaussian ML estimator. We can of course specify other distributions and the maximising the corresponding likelihood.

It might seem that adding more lags can approximate better the AR(∞) or MA(∞) representations, thus reducing the variance of the white noise. However, the more parameters we include in the model the more likely we are to make estimation errors. Moreover, we know that the number of parameters $p+q$ must be smaller than the number of observations used which is $T-p-q$ (cfr. the linear regression case).

There are two main approaches.

1. Information criteria. The original one was proposed by Akaike. The basic idea is to compare the true likelihood of the model (which is unknown) with the likelihood of the data. We should use a the distance between these two distributions and minimise it. As a result we have to select p and q such that they minimise the Akaike Information Criterion

$$AIC(p,q) = -2L_T(\mathbf{x}_T; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_u^2) + 2(p+q+1) = \sum_{t=1}^{T} \widehat{u}_t^2 + 2(p+q+1).$$

   The first term decreases as p, q increase, but the second term is the one that avoids over-parametrization as indeed it grows when p, q increase, for this reason this term is called penalty. Alternatively we can minimise the Bayes (or Schwarz) Information Criterion

$$BIC(p,q) = -2L_T(\mathbf{x}_T; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_u^2) + 2(p+q+1)\log T = \sum_{t=1}^{T} \widehat{u}_t^2 + 2(p+q+1)\log T.$$

   Thus for each choice of p and q we have to estimate the model, compute the log-likelihood and then choose the couple (p, q) that minimises AIC or BIC. The BIC is a consistent criterion in the sense that if the data were truly generated by an ARMA(p, q) the estimated orders that minimise BIC will converge with probability 1 to the true orders as $T \to \infty$. On the other hand minimising AIC for AR processes selects a model with the smallest one step ahead prediction error among all possible AR models. In general BIC is more parsimonious than AIC.

2. Box and Jenkins methodology. By inspection of the autocorrelation (acs) and partial autocorrelation (pacs) we can make assumptions of the AR and MA orders. The acs of an MA(q) is zero at lags $h > q$. For an AR(1) the acs is exponentially decreasing, but for an AR(p) the acs decreases in a less regular way. However, the pacs of an AR(p) is zero at lags $h > p$. With real data the acs and pacs will not become exactly zero but we can plot them together with their confidence intervals (see Chapter 3 for acs). Approximately the 95% confidence interval for acs and pacs is $\pm 1.96/\sqrt{T}$ (cfr the acs plots in Chapter 2).

5.6 Diagnostics

The main steps of model building for a time series are four:

1. check for trends and seasonality and remove them by means of suitable differences;
2. model specification: formulate an ARMA(p, q) with suitable lags;
3. estimation: find values and standard errors for the parameters of the model;
4. checking: verify that the specified model provides an adequate description of the data.

The first three steps have been already considered. Here we focus on model checking or diagnostic checking. This is done by analysing the residuals of an estimated ARMA(p, q).

5.6.1 Residuals of an AR process

We know that from the Wold representation (which is an MA(∞)) we can always write an AR(∞) model and that we can approximate it as an ARMA(p, q). Here we simplify the model to an AR(p) as indeed AR models are the most useful for forecasting which is our ultimate aim. After estimating an AR(p) (via Yule Walker or Least Squares) we can use the estimated parameters to build the predicted value

$$\widehat{x}_t = \widehat{\phi}_1 x_{t-1} + \widehat{\phi}_2 x_{t-2} + \ldots + \widehat{\phi}_p x_{t-p}. \tag{24}$$

We then define the residuals of the model as

$$\widehat{u}_t = x_t - (\widehat{\phi}_1 x_{t-1} + \widehat{\phi}_2 x_{t-2} + \ldots + \widehat{\phi}_p x_{t-p}).$$

Using the notation of Chapter 3 we can write

$$\widehat{x}_t = \widehat{P}^p_{t-1} x_t, \qquad \widehat{u}_t = \widehat{x}_t - \widehat{P}^p_{t-1} x_t,$$

where $\widehat{P}^p_{t-1}$ is the estimated predictor (or projector) of x_t when using estimated coefficients, p lags, and data up to time $t-1$. In the same way we can write the innovations of the AR(p) process as

$$u_t = x_t - (\phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p}) = x_t - P^p_{t-1} x_t.$$

Because of consistency of the estimated coefficients we have

$$\mathrm{E}[(\widehat{P}^p_{t-1} x_t - P^p_{t-1} x_t)^2] \to 0, \qquad T \to \infty$$

and therefore

$$\widehat{P}^p_{t-1} x_t \xrightarrow{p} P^p_{t-1} x_t, \qquad T \to \infty.$$

If model (24) is well specified, the residuals $\widehat{u}_t$ should approximate the errors of an AR(p) which by definition are white noise, or more rigorously we would like to have

$$\mathrm{E}[(\widehat{u}_t - u_t)^2] \to 0, \qquad T \to \infty.$$

This, is a goodness of fit measure which however we cannot compute since we do not know u_t. So a good measure is the Root Mean Squared Prediction Error (RMSPE)

$$RMSPE(p) = \sqrt{\frac{1}{T} \sum_{t=p+1}^{T} \widehat{u}_t^2}$$

which can be computed for any AR(p) model and the smaller it is the better is the fit of the model. Notice however, that by definition the model with the largest p has the smallest $RMSPE$ thus a penalization as in the Information criteria above is necessary and this is the rationale behind those criteria.

5.6.2 Testing for white noise

Depending on the hypothesis we made on u_t before estimation, we might need to check for

1. uncorrelated residuals, if $u_t \sim wn(0, \sigma_u^2)$;

2. independent residuals, if $u_t \sim iid(0, \sigma_u^2)$;

3. normality of residuals, if $u_t \sim N(0, \sigma_u^2)$.

Notice that if 1 and 3 hold then 2 is automatically satisfied. For a general AR model to be correctly specified 1 is enough, but if we are using Maximum Likelihood with a Gaussian distribution then we might want to check also for 3. Here, we discuss only testing for white noise.

In order to test for white noise we can do three things.

1. Plot the residuals and check for outliers which deviate from a purely random behaviour (this is done also in regression analysis);

2. Compute the sample acs of $\widehat{u}_t$ and verify that they are zero. We denote the acs as $\widehat{\rho}_h^{\hat{u}}$. Here we have the errors coming from the estimation of the model, i.e. of $\widehat{u}_t$, and the estimation of the acs that play an important role. Indeed, while the true acs of the true ARMA error would be exactly zero, $\widehat{\rho}_h^{\hat{u}}$ is not exactly zero even if the model were correctly specified. We then need to compute confidence intervals for $\widehat{\rho}_h^{\hat{u}}$. A general rule (which is exact for $h > p$ where p is the AR order) is that the 95% confidence interval is given by $\pm 1.96/\sqrt{T}$. Although for $h < p$ the interval is smaller, it is often approximated using the same value. If we find values of $\widehat{\rho}_h^{\hat{u}}$ which are outside the above interval then we reject at 5% level the null hypothesis that the acs is zero and we say that the model is misspecified.[35]

3. We can also test for all acs being jointly zero up to a given lag K. This is done by computing the portmanteau-test statistics

$$Q = T \sum_{h=1}^{K} (\widehat{\rho}_h^{\hat{u}})^2$$

where we typically choose K between 15 and 30. If the model is correctly specified, i.e. all acs are zero, then $Q \sim \chi^2_{K-p}$. A corrected version for small T is the Ljung Box statistics

$$Q = T(T+2) \sum_{h=1}^{K} \frac{(\widehat{\rho}_h^{\hat{u}})^2}{T-h}.$$

Another possible test is the one by Durbin and Watson

$$d = \frac{\sum_{t=2}^{T} (\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_{t=1}^{T} \widehat{u}_t^2} \simeq 2(1 - \widehat{\rho}_1^{\hat{u}})$$

which tests only for the first acs to be zero, and if this is the case we expect to have $d \simeq 2$.

5.6.3 Residuals and innovations

If we compare the above expression for the predicted value $\widehat{P}_{t-1}^p x_t$ with the AR(∞) given for the general linear predictor in Chapter 2 and we recall that if p is large enough the predictor with p lags is a good approximation of the one with infinite lags, then

$$\mathrm{E}[(\widehat{P}_{t-1}^p x_t - P_{t-1} x_t)^2] \to 0, \qquad T, p \to \infty.$$

[35] Equivalently we reject the null hypothesis if $\{0\} \notin \{\widehat{\rho}_h^{\hat{u}} \pm 1.96/\sqrt{T}\}$.

and this implies that for p large enough u_t is the innovation of x_t which we denoted as e_t and the residuals $\widehat{u}_t$ should approximate in mean square the innovations. Now from Chapter 3 we have

$$e_t = x_t - P_{t-1}x_t,$$

where e_t is the innovation of x_t, which is the unpredictable part of x_t and it is a white noise. Thus, once again we would like to have the residuals $\widehat{u}_t$ to be white noise. That is to say that if the model is correctly specified then the prediction of x_t is the best we can do (among all linear models) since what is left is a white noise, thus it is unpredictable as it contains no useful information about future values of x_t.

In general for ARMA a similar reasoning can be followed once we notice that u_t is the innovation of x_t also in this case. Indeed, consider a stationary causal and invertible ARMA(p, q)

$$\Phi(L)x_t = \Theta(L)u_t, \qquad u_t \sim wn(0, \sigma_u^2)$$

then we can write it as MA(∞)

$$x_t = \Phi(L)^{-1}\Theta(L)u_t = u_t + \psi_1 u_{t-1} + \psi_2 u_{t-2} + \dots$$

so that x_t is a linear combination of $u_t, u_{t-1}, \dots$. This implies that for $k \geq 1$, $\mathrm{E}[u_t X_{t-k}] = 0$ because u_t is a white noise. Under invertibility we have

$$u_t = \Theta(L)^{-1}\Phi(L)x_t$$

so that u_t is a linear combination of $x_t, x_{t-1}, \dots$. Thus both u_t and its past and x_t and its past are in the same space. In that case the projection equation defined in Chapter 3 is

$$x_t = P_{t-1}x_t + e_t = [\phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}] + u_t.$$

In conclusion, if the stationarity and invertibility conditions are satisfied, u_t is the innovation of the ARMA process as it is white noise and uncorrelated with all predictors.

6 Forecasting of ARMA processes

6.1 One-step-ahead forecast

In Chapter 3, we have seen the general definition of linear predictor $P_{t-1}x_t$ (now we set $a_0 = 0$ for simplicity, i.e. $\mathrm{E}[X_t] = 0$)

$$x_t = a_1 x_{t-1} + \dots + e_t = P_{t-1}x_t + e_t \tag{25}$$

where e_t is a zero mean white noise and it is called innovation of x_t. If we forecast a series at time $T + 1$ and we have just T observations we have the prediction equation

$$x_{T+1} = P_T x_{T+1} + e_{T+1}.$$

We have seen that for a stationary and invertible ARMA(p, q) process we have

$$x_t = P_{t-1}x_t + e_t = [\phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}] + u_t,$$

and then u_t is the innovation of x_t and we can write a forecast equation also in this case

$$x_{T+1} = \underbrace{P_T x_{T+1}}_{\text{forecast}} + \underbrace{u_{T+1}}_{\text{forecast error}} = \phi_1 x_T + \ldots + \phi_p x_{T-p+1} + \theta_1 u_T + \ldots + \theta_q u_{T-q+1} + u_{T+1}, \quad (26)$$

where we see that in practice only a finite number of lags p and q can be used as we must ensure that $T - p + 1 > 0$ and $T - q + 1 > 0$ since we have a finite number of observations. The first term on the right hand side of (26) is the one-step-ahead forecast and we denote it as

$$x_{T+1|T} \equiv P_T x_{T+1}.$$

The second term on the right hand side of (26) is the one-step-ahead forecast error defined as

$$\epsilon_{T+1|T} \equiv u_{T+1} = x_T - P_T x_{T+1},$$

which is a white noise process by construction. In other words the best linear predictor is such that the one-step-ahead forecast error (or prediction error) is a white noise (see the derivation in Chapter 3). Both for the forecast and its error we denote in the index the fact that they are obtained when using information up to time T.

6.2 Two-step-ahead forecast

If in (26) we replace x_T with its definition we have an equation for x_{T+1} given observations only until $T - 1$:

$$\begin{aligned} x_{T+1} &= \phi_1(\phi_1 x_{T-1} + \ldots + \phi_p x_{T-p} + u_T + \theta_1 u_{T-1} + \ldots + \theta_q u_{T-q}) + \ldots + \phi_p x_{T-p+1} \\ &\quad + \theta_1 u_T + \ldots + \theta_q u_{T-q+1} + u_{T+1} \\ &= (\phi_1^2 + \phi_2) x_{T-1} + \ldots \phi_1 \phi_p x_{T-p} + (\phi_1 \theta_1 + \theta_2) u_{T-1} + \ldots + \phi_1 \theta_q u_{T-q+1} \\ &\quad + (u_{T+1} + (\phi_1 + \theta_1) u_T). \end{aligned}$$

Since the forecast is made using only data until x_{T-1} the last term is the two-step-ahead forecast error

$$\epsilon_{T+1|T-1} = u_{T+1} + (\phi_1 + \theta_1) u_T.$$

By shifting one step ahead the previous two equations we get

$$\epsilon_{T+2|T} = u_{T+2} + (\phi_1 + \theta_1) u_{T+1},$$

where it is clear that this part is unpredictable if we have observations only up to time T. We can then write

$$x_{T+2} = x_{T+2|T} + \epsilon_{T+2|T}.$$

Notice that the two-step-ahead forecast error is no longer a white noise, indeed for example its lag 1 autocovariance is

$$\mathrm{E}[\epsilon_{T+2|T} \epsilon_{T+1|T}] = \mathrm{E}[(u_{T+2} + (\phi_1 + \theta_1) u_{T+1}) u_{T+1}] = \sigma_u^2 (\phi_1 + \theta_1).$$

6.3 The h-step-ahead forecast

By iterating the previous reasoning we can compute a generic h-step-ahead forecast error defined as $\epsilon_{T+h|T}$ and in general for the h-step-ahead case we have

$$x_{T+h} = x_{T+h|T} + \epsilon_{T+h|T}.$$

How do we derive the properties of the forecast and the forecast error in this case?

Let us consider the MA(∞) representation, which is valid for a stationary and invertible ARMA (it is the Wold representation in this case as there are no deterministic components)

$$x_t = \sum_{k=0}^{\infty} \psi_k u_{t-k} = \Psi(L) u_t, \qquad \psi_0 = 1.$$

Then,

$$x_{T+h} = \sum_{k=0}^{\infty} \psi_k u_{T+h-k} = \Psi(L) u_{T+h}. \tag{27}$$

Now from (27) we have a decomposition of the observation at time $T+h$ into an unpredictable and a predictable part:

$$\begin{aligned} x_{T+h} &= \sum_{k=0}^{\infty} \psi_k u_{T+h-k} \\ &= \sum_{k=0}^{h-1} \psi_k u_{T+h-k} + \sum_{k=h}^{\infty} \psi_k u_{T+h-k} \\ &= \epsilon_{T+h|T} + x_{T+h|T}, \end{aligned} \tag{28}$$

where the first term on the right hand side is the h-step-ahead forecast error which is unpredictable since it depends on observations at time after T and is not a white noise. The second term is predictable as it depends only on information up to time T, this is the h-step-ahead forecast.

On the other hand, if we start from the AR(∞) representation (25), which in the case of a stationary and invertible ARMA implies $e_t \equiv u_t$, we have

$$x_{T+h|T} = \sum_{k=1}^{\infty} a_k x_{T-k+1}. \tag{29}$$

Therefore, from (27) and (29), the h-step-ahead forecast is given by

$$\begin{aligned} x_{T+h|T} &= \sum_{k=1}^{\infty} a_k x_{T-k+1} = \sum_{k=1}^{\infty} a_k \Psi(L) u_{T-k+1} = A(L)\Psi(L) u_T \\ &= B(L) u_T = \sum_{j=0}^{\infty} b_j u_{T-j}, \end{aligned} \tag{30}$$

which shows again that the h-step-ahead forecast can be written as a MA(∞) using data only up to time T. We now want to find the coefficients b_j.

As we have seen in Chapter 3, the best linear forecast is obtained by minimising the one-step-ahead mean squared forecast error (the distance between the true value and the forecasted one). Thus, following the same approach for the best linear h-step-ahead forecast, we have to minimise

$$\begin{aligned}
\mathrm{E}[(X_{T+h} - X_{T+h|T})^2] &= \mathrm{E}\left[\left(\sum_{k=0}^{h-1}\psi_k u_{T+h-k} + \sum_{k=h}^{\infty}\psi_k u_{T+h-k} - \sum_{k=0}^{\infty} b_k u_{T-k}\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{k=0}^{h-1}\psi_k u_{T+h-k} + \sum_{j=0}^{\infty}(\psi_{j+h} - b_j) u_{T-j}\right)^2\right] \\
&= \sigma_u^2\left[\sum_{k=0}^{h-1}\psi_k^2 + \sum_{j=0}^{\infty}(\psi_{j+h} - b_j)^2\right]
\end{aligned}$$

where we used the fact that u_t is a white noise. Notice that we are minimising the variance of the forecast error. The first term is independent of the choice of the b_j's and the second term is clearly minimised by choosing $b_j = \psi_{j+h}$, $j = 0, 1, 2, \ldots$. Therefore, from (30) by setting $k = j + h$, we have that the forecast is

$$x_{T+h|T} = \sum_{j=0}^{\infty} b_j u_{T-j} = \sum_{k=h}^{\infty}\psi_k u_{T+h-k} = \sum_{j=0}^{\infty}\psi_{j+h} u_{T-j}.$$

As a consequence the h-step-forecast error is

$$\epsilon_{T+h|T} = x_{T+h} - x_{T+h|T} = \sum_{k=0}^{h-1}\psi_k u_{T+h-k},$$

and these are the same expressions derived in (28) for a stationary invertible ARMA. In other words, given an ARMA the forecasts computed using the intuitive formulas (28) are such that they minimise the variance of the forecast error.

6.4 Variance of the forecast error

Now, since $\mathrm{E}[\epsilon_{T+h|T}] = 0$, the minimised mean squared h-step-ahead forecast error is the variance of the h-step-ahead forecast error which we denote by $\sigma^2_{T+h|T}$. Indeed,

$$\sigma^2_{T+h|T} \equiv \mathrm{E}[(X_{T+h} - X_{T+h|T})^2] = \mathrm{E}[\epsilon^2_{T+h|T}] = \sigma_u^2\sum_{k=0}^{h-1}\psi_k^2.$$

That is to say that computing the best linear forecast is equivalent to minimising the variance of the forecast error. Of course there might be non-linear forecasts for which the variance of the forecast error can be even smaller (see Section 6.6).

For example, when $h = 1$ we have $b_j = \psi_{j+1}$ and the forecast is

$$x_{T+1|T} = \psi_1 u_T + \psi_2 u_{T-1} + \ldots$$

while

$$x_{T+1} = \psi_0 u_{T+1} + \psi_1 u_T + \psi_2 u_{T-1} + \ldots$$

and, since $\psi_0 = 1$, the one-step-ahead forecast error is

$$\epsilon_{T+1|T} = x_{T+1} - x_{T+1|T} = \psi_0 u_{T+1} = u_{T+1},$$

as derived before and $\sigma^2_{T+1|T} = \text{Var}(u_{T+1}) = \sigma^2_u$.

Recall that

$$x_{T+h} = \sum_{k=0}^{h-1} \psi_k u_{T+h-k} + \sum_{k=h}^{\infty} \psi_k u_{T+h-k} = \epsilon_{T+h|T} + x_{T+h|T}$$

and we see that as $h \to \infty$ the second term disappears while the first dominates. More rigorously, since $\text{E}[\epsilon_{T+h|T} X_{T+h|T}] = 0$ (because u_t is white noise) we can write

$$\text{E}[X^2_{T+h}] = \text{E}[X^2_{T+h|T}] + \text{E}[\epsilon^2_{T+h|T}]. \tag{31}$$

By stationarity, the above result depends only on h (it holds for any t) and we can also write (everything has zero mean)

$$\text{Var}(X_t) = \text{Var}(X_{t+h|t}) + \text{Var}(\epsilon_{t+h|t}).$$

Then, as $h \to \infty$, we can prove two main results.

1. The variance of the forecast tends to zero, and therefore the forecast tends to the mean of the process.

   Indeed, from the definition of h-step-ahead forecast we have

   $$\text{Var}(X_{T+h|T}) = \text{E}[X^2_{T+h|T}] = \sigma^2_u \sum_{k=h}^{\infty} \psi^2_k \to 0, \qquad h \to \infty,$$

   and since the mean of the forecast is zero we have also that (Chebychev inequality)

   $$X_{T+h|T} \xrightarrow{p} 0, \qquad h \to \infty,$$

   i.e. in the long run the best prediction is the mean (which is zero in this case). If $\text{E}[X_t] = \mu \neq 0$, then we would have $X_{T+h|T} \xrightarrow{p} \mu$ as $h \to \infty$.

2. The variance of the forecast error tends to the variance of the process which is its upper bound.

   Indeed, we have

   $$\sigma^2_{T+h|T} \equiv \text{E}[\epsilon^2_{T+h|T}] = \sigma^2_u \sum_{k=0}^{h-1} \psi^2_k \to \sigma^2_u \sum_{k=0}^{\infty} \psi^2_k = \text{Var}(X_t), \qquad h \to \infty.$$

   Moreover, since for any h we have

   $$\sum_{k=0}^{h-1} \psi^2_k \leq \sum_{k=0}^{\infty} \psi^2_k$$

   then $\sigma^2_{T+h|T} \leq \text{Var}(X_t)$, and the variance of the process is the upper bound of the variance of the forecast error. Therefore, as h increases the variance of the forecast error increases: the longer the forecasting horizon, the higher the uncertainty (see the examples at the end of the Chapter).

Finally, we have seen that the h-step-ahead forecast error is not a white noise. We can compute its lag m acvs

$$\begin{aligned}\gamma_m^{\epsilon_h} &= \text{Cov}(\epsilon_{T+h|T}, \epsilon_{T+h+m|T}) = \text{E}[\epsilon_{T+h|T}\epsilon_{T+h+m|T}] \\ &= \text{E}\left[\left(\sum_{k=0}^{h-1}\psi_k u_{T+h-k}\right)\left(\sum_{j=0}^{h+m-1}\psi_j u_{T+h+m-j}\right)\right] \\ &= \sigma_u^2\sum_{k=0}^{h-1}\psi_k\psi_{k+m}.\end{aligned}$$

For example the acvs of the one-step-ahead forecast error is

$$\gamma_m^{\epsilon_1} = \sigma_u^2\psi_0\psi_m = \sigma_u^2\psi_m.$$

This could be quite large and should the forecast for a series wander off target, it is possible for it to remain there in the short run since forecast errors can be quite highly correlated. Hence, when x_{T+1} becomes available we should update the forecast.

6.5 Computing forecasts

Some examples.

1. AR(1)
$$x_t = \phi_1 x_{t-1} + u_t, \qquad u_t \sim wn(0, \sigma_u^2), \quad |\phi_1| < 1.$$
There are three main ways to compute the forecast.

   (a) Using the MA(∞) representation
   $$x_t = \frac{1}{1-\phi_1 L}u_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \dots$$
   so
   $$\Psi(z) = 1 + \phi_1 z + \phi_1^2 z^2 + \dots = \psi_0 + \psi_1 z + \psi_2 z^2 + \dots$$
   which implies $\psi_k = \phi_1^k$. Hence, the h-step-ahead forecast is
   $$\begin{aligned}x_{T+h|T} &= \sum_{k=0}^{\infty} b_k u_{T-k} = \sum_{k=0}^{\infty}\psi_{k+h}u_{T-k} \\ &= \sum_{k=0}^{\infty}\phi_1^{k+h}u_{T-k} = \phi_1^h\sum_{k=0}^{\infty}\phi_1^k u_{T-k} \\ &= \phi_1^h x_T.\end{aligned}$$

   (b) We can write the h-step-ahead forecast as function of past values of the process. We have
   $$x_{T+h|T} = \sum_{j=0}^{\infty}\psi_{j+h}u_{T-j} = \Psi^{(h)}(L)u_T$$
   Moreover,
   $$x_t = \Psi(L)u_t$$

and if the polynomial is invertible we have

$$u_t = \Psi^{-1}(L)x_t$$

and thus

$$x_{T+h|T} = \Psi^{(h)}(L)u_T = \Psi^{(h)}(L)\Psi^{-1}(L)x_T = G^{(h)}(L)x_T,$$

where $G^{(h)}(L)$ is a new polynomial. And for h fixed we have defined a new process of all h-step-ahead forecasts $x_{t+h|t} = G^{(h)}(L)x_t$.

Using the polynomial $G^{(h)}(L)$. From the previous approach we see that $\Psi^{(h)}(z) = \sum_{k=0}^{\infty} \phi_1^{k+h} z^k$. Alternatively using $G^{(h)}(L) = \Psi^{(h)}(L)\Psi^{-1}(L)$ we have

$$x_{T+h|T} = G^{(h)}(L)x_T = \left(\sum_{k=0}^{\infty} \phi_1^{k+h} L^k\right)(1-\phi_1 L)x_T = \phi_1^h x_T$$

as before.

(c) We can consider the AR equations in the future when setting the innovations to zero

$$\begin{aligned} x_{T+1|T} &= \phi_1 x_T + 0 \\ x_{T+2|T} &= \phi_1 x_{T+1|T} + 0 \\ &\vdots \\ x_{T+h|T} &= \phi_1 x_{T+h-1|T} + 0 \end{aligned}$$

so that again $x_{T+h|T} = \phi_1^h x_T$.

Using the MA approach, the variance of the forecast error is then

$$\begin{aligned} \sigma^2_{T+h|T} &= \sigma_u^2 \sum_{k=0}^{h-1} \psi_k^2 = \sigma_u^2 \sum_{k=0}^{h-1} \phi_1^{2k} \\ &= \sigma_u^2 \frac{1-\phi_1^{2h}}{1-\phi_1^2} \to \sigma_u^2 \frac{1}{1-\phi_1^2} = \text{Var}(X_t), \qquad h \to \infty \end{aligned}$$

Using the explicit expression for the forecast we have the same result

$$\begin{aligned} \sigma^2_{T+h|T} &= \text{E}[(X_{T+h} - X_{T+h|T})^2] = \text{E}[(X_{T+h} - \phi_1^h X_T)^2] \\ &= \text{E}[X^2_{T+h}] + \phi_1^{2h}\text{E}[X_T^2] - 2\phi^h \text{E}[X_{T+h}X_T] \\ &= \text{Var}(X_t) + \phi_1^{2h}\text{Var}(X_t) - 2\phi^h \gamma_h^x \to \text{Var}(X_t), \qquad h \to \infty. \end{aligned}$$

We have demonstrated that for the AR(1) model the linear least squares predictor of $x_{T+h|T}$ depends only on the most recent observation, x_T, and does not involve $x_{T-1}, x_{T-2}, \ldots$, in this case we say that x_t is a Markov process. An example is in Figure 39.

2. AR(p). The forecast $x_{T+h|T}$ depends only on the last p observed values of x_t, and may be obtained by solving the AR(p) difference equation with the future u_t set to zero. For example for an AR(p) process and $h = 1$,

$$x_{T+1|T} = \phi_1 x_T + \ldots + \phi_p x_{T-p+1}.$$

while for an AR(2) and generic h we have the iterative formulas

$$\begin{aligned} x_{T+1|T} &= \phi_1 x_T + \phi_2 x_{T-1} \\ x_{T+2|T} &= \phi_1 x_{T+1|T} + \phi_2 x_T \\ x_{T+h|T} &= \phi_1 x_{T+h-1|T} + \phi_2 x_{T+h-2|T}, \quad h > 2. \end{aligned}$$

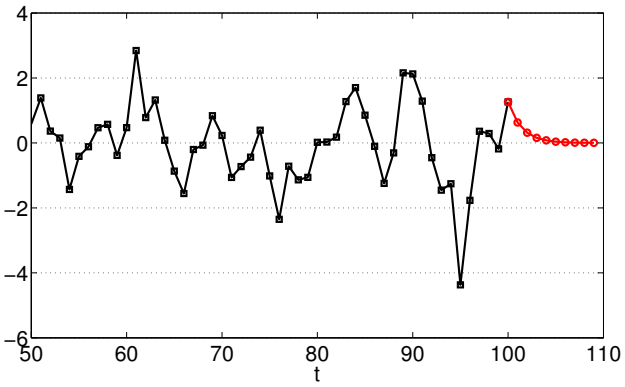


Figure 39: Forecast of an AR(1) model with $\phi_1 = 0.5$ up to lag 10.

3. ARMA(1,1)

$$(1 - \phi_1 L)x_t = (1 - \theta_1 L)u_t, \qquad u_t \sim wn(0, \sigma_u^2) \quad |\phi_1| < 1 \quad |\theta_1| < 1$$

Once again we can compute the forecast by taking the ARMA(1,1) equation at $T + h$

$$x_{T+h} = \phi x_{T+h-1} - \theta u_{T+h-1} - u_{T+h}$$

and setting future innovations to zero, i.e. $u_{T+h} = 0$ for $h > 0$ and replacing forecasts for the unknown values of x_t. Hence

$$\begin{aligned} x_{T+1|T} &= \phi x_T - \theta u_T \\ x_{T+2|T} &= \phi x_{T+1|T} \\ &\vdots \\ x_{T+h|T} &= \phi x_{T+h-1|T} \end{aligned}$$

The MA coefficient is useful only for forecasting one-step-ahead. This is intuitive as the MA component is responsible only for acvs up to lag 1. An example is in Figure 40.

Alternatively, take $\phi_1 = \phi$ and $\theta_1 = \theta$ then

$$x_t = \frac{1 - \theta L}{1 - \phi L} u_t = \Psi(L) u_t$$

so

$$\begin{aligned} \Psi(z) &= (1 - \theta z)(1 + \phi z + \phi^2 z^2 + \ldots) \\ &= 1 + (\phi - \theta)z + \phi(\phi - \theta)z^2 + \ldots + \phi^{h-1}(\phi - \theta)z^h + \ldots \\ &= \psi_0 + \psi_1 z + \psi_2 z^2 + \ldots \end{aligned}$$

so $\psi_0 = 1$ and $\psi_h = \phi^{h-1}(\phi - \theta)$ for $h \geq 1$. The variance of the forecast error is then

$$\begin{aligned} \sigma^2_{T+h|T} &= \sigma_u^2 \sum_{k=0}^{h-1} \psi_k^2 = \sigma_u^2 \left(1 + \sum_{k=1}^{h-1} \psi_k^2 \right) \\ &= \sigma_u^2 \left(1 + (\phi - \theta)^2 \sum_{k=1}^{h-1} \phi^{2k-2} \right) \\ &= \sigma_u^2 \left(1 + (\phi - \theta)^2 \frac{1 - \phi^{2h-2}}{1 - \phi^2} \right) \end{aligned}$$

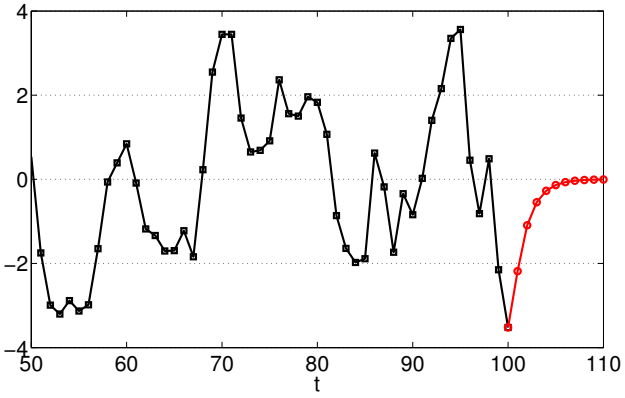


Figure 40: Forecast of an ARMA(1,1) model with $\phi_1 = 0.5$ and $\theta_1 = -0.6$ up to lag 10.

Now to find the forecast we need the polynomial $G^{(h)}(L) = \Psi^{(h)}(L)\Psi^{-1}(L)$. So

$$\begin{aligned}\Psi^{(h)}(z) &= \sum_{k=0}^{\infty} \psi_{k+h} z^k \\ &= \phi^{h-1}(\phi-\theta)\sum_{k=0}^{\infty}\phi^k z^k \\ &= \frac{\phi^{h-1}(\phi-\theta)}{1-\phi z}\end{aligned}$$

and

$$\Psi^{-1}(z) = \frac{1-\phi z}{1-\theta z}$$

which give

$$x_{T+h|T} = G^{(h)}(L)x_T = \frac{\phi^{h-1}(\phi-\theta)}{1-\theta L}x_T$$

Take $h = 1$

$$\begin{aligned}x_{T+1|T} &= \frac{(\phi-\theta)}{1-\theta L}x_T \\ &= (\phi-\theta)(1+\theta L+\theta^2 L^2+\ldots)x_T \\ &= (\phi-\theta)x_T + (\phi-\theta)\theta x_{T-1} + (\phi-\theta)\theta^2 x_{T-2} + \ldots \\ &= \phi x_T - \theta\left(x_T - (\phi-\theta)x_{T-1} - (\phi-\theta)\theta x_{T-2} - \ldots\right)\end{aligned}$$

but

$$u_T = \Psi^{-1}(L)x_T = \frac{1-\phi L}{1-\theta L}x_T = x_T - (\phi-\theta)x_{T-1} - (\phi-\theta)\theta x_{T-2} - \ldots$$

therefore

$$x_{T+1|T} = \phi x_T - \theta u_T$$

4. MA(1)

$$x_t = (1-\theta_1 L)u_t, \qquad u_t \sim wn(0,\sigma_u^2) \quad |\theta_1| < 1$$

then at $T+h$ we have

$$x_{T+h} = u_{T+h} - \theta u_{T+h-1}$$

and again by setting to zero future innovations we see that the only non-zero forecast is the one-step-ahead ($h=1$) when we have

$$x_{T+1|T} = -\theta u_T$$

So in general for an MA(q) we non zero forecast only up to $h=q$. Take the MA(2), we have

$$\begin{aligned} x_{T+1|T} &= -\theta_1 u_T - \theta_2 u_{T-1} \\ x_{T+2|T} &= -\theta_2 u_T \\ x_{T+h|T} &= 0, \quad h > 2 \end{aligned}$$

An example is in Figure 41.

Alternatively we have the MA(∞) representation of an MA(1)

$$\Psi(z) = 1 - \theta_1 z = \psi_0 + \psi_1 z + \psi_2 z^2 + \dots$$

hence $\psi_0 = 1$ and $\psi_1 = -\theta_1$, all other coefficients being zero. The variance of the forecast error is then

$$\sigma^2_{T+h|T} = \sigma^2_u \sum_{k=0}^{h-1} \psi_k^2 = \begin{cases} \sigma^2_u & h=1 \\ \sigma^2_u \left(1+\theta_1^2\right) & h \geq 2 \end{cases}$$

The h-step-ahead forecast is

$$\begin{aligned} x_{T+h|T} &= \sum_{k=0}^{\infty} \psi_{k+h} u_{T-k} \\ &= \psi_h u_T + \psi_{h+1} u_{T-1} + \dots \end{aligned}$$

Now to find the forecast we need the polynomial $G^{(h)}(L) = \Psi^{(h)}(L)\Psi^{-1}(L)$. So

$$\Psi^{(h)}(z) = \sum_{k=0}^{\infty} \psi_{k+h} z^k = \psi_h = \begin{cases} -\theta_1 & h=1 \\ 0 & h \geq 2 \end{cases}$$

and

$$G^{(h)}(z) = \Psi^{(h)}(z)\Psi^{-1}(z) = \begin{cases} \frac{-\theta_1}{1-\theta_1 z} & h=1 \\ 0 & h \geq 2 \end{cases}$$

which shows that for the MA(1) the only non-zero forecast is the one-step-ahead forecast (consistently with what we found in the ARMA case):

$$x_{T+1|T} = G^{(1)}(L)x_T = -\theta_1(1 + \theta_1 L + \theta_1^2 L^2 + \dots)x_T = -\sum_{k=0}^{\infty} \theta_1^{k+1} x_{T-k}.$$

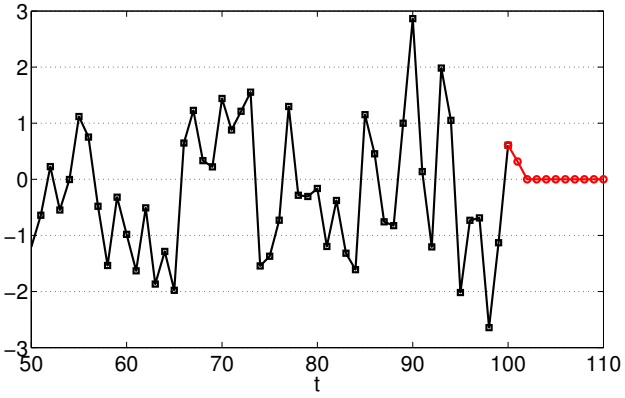


Figure 41: Forecast of an MA(1) model with $\theta_1 = -0.6$ up to lag 10.

6.6 Forecasting and conditional expectations

In general a forecast is a function of $X_1 \dots X_T$. In particular, the function $f(X_1 \dots X_T)$ that minimises the variance of the forecast error or (mean squared forecast error) $\mathrm{E}[(X_{T+h} - f(X_1 \dots X_T))^2]$ is always the conditional expectation of X_{T+h} given $X_1 \dots X_T$. For simplicity of notation denote by $\mathcal{I}_T$ the past history $X_1 \dots X_T$. Then, we have

$$\mathrm{E}[(X_{T+h} - f(\mathcal{I}_T))^2 | \mathcal{I}_T] \quad = \quad \mathrm{E}[X_{T+h}^2 | \mathcal{I}_T] - 2\mathrm{E}[X_{T+h}|\mathcal{I}_T] f(\mathcal{I}_T) + f^2(\mathcal{I}_T),$$

by denoting by $c = f(\mathcal{I}_T)$ and minimising with respect to c we find that the previous equation attains a minimum for $f(\mathcal{I}_T) = \mathrm{E}[X_{T+h}|\mathcal{I}_T]$. By using the law of iterated expectations we have

$$\mathrm{E}\left[\mathrm{E}[(X_{T+h} - f(\mathcal{I}_T))^2 | \mathcal{I}_T]\right] = \mathrm{E}[(X_{T+h} - f(\mathcal{I}_T))^2].$$

Then, also $\mathrm{E}[(X_{T+h} - f(\mathcal{I}_T))^2]$ is minimised by $f(\mathcal{I}_T) = \mathrm{E}[X_{T+h}|\mathcal{I}_T]$.[36]

As a consequence, the best h-step-ahead forecast is the conditional expectation of X_{T+h}

$$X_{T+h|T} = \mathrm{E}[X_{T+h}|X_1 \dots X_T].$$

Thus given an observed history of the process $x_1 \dots x_T$, i.e. a realisation of $\{X_t,\ t = 1, \dots, T\}$, we have

$$x_{T+h|T} = \mathrm{E}[X_{T+h}|X_1 = x_1 \dots X_T = x_T].$$

When we restrict to linear functions such an AR(p) model

$$x_{T+1} = \phi_1 x_T + \dots + \phi_p x_{T-p+1} + u_T,$$

the the best forecast, which is the conditional mean, coincides with the best linear forecast computed in the previous section only if we make the assumption that u_t is a martingale difference sequence, that is $\mathrm{E}[u_t|x_{t-1}, \dots x_{t-p}] = 0$ for any t, or alternatively if we assume Gaussian or independent errors u_t which implies $\mathrm{E}[u_t|x_{t-1} \dots x_{t-p}] = \mathrm{E}[u_t] = 0$ for any t.

[36] Given the random variable $c = f(\mathcal{I}_T)$, then define $g(c) = \mathrm{E}[(X_{T+h} - c)^2|\mathcal{I}_T]$ and we have to minimize $\mathrm{E}[g(c)]$

$$\frac{\mathrm{d}}{\mathrm{d}c}\mathrm{E}[g(c)] = \frac{\mathrm{d}}{\mathrm{d}c}\int g(c)\mathrm{d}F_c(c) = \int \frac{\mathrm{d}}{\mathrm{d}c} g(c)\mathrm{d}F_c(c) = 0$$

where $F_c(c)$ is the cdf of c and the function in the integral is zero when $c = \mathrm{E}[X_{T+h}|\mathcal{I}_T]$, then since the function $g(c)\mathrm{d}F_c(c)$ is always positive also the integral is minimised for $c = \mathrm{E}[X_{T+h}|\mathcal{I}_T]$.

We can then use conditional expectations to compute forecasts, which is a useful approach for the AR case. So for example in the AR(1) case and assuming $\mathrm{E}[u_{T+h}|X_T, \dots X_1] = 0$ we have

$$\begin{aligned} x_{T+h|T} &= \mathrm{E}[X_{T+h}|X_T = x_T, \dots X_1 = x_1] \\ &= \mathrm{E}[\phi_1 X_{T+h-1} + u_{T+h}|X_T = x_T, \dots X_1 = x_1] \\ &= \phi_1 \mathrm{E}[X_{T+h-1}|X_T = x_T, \dots X_1 = x_1] \\ &= \phi_1 \mathrm{E}[\phi_1 X_{T+h-2} + u_{T+h-1}|X_T = x_T, \dots X_1 = x_1] \\ &= \phi_1^2 x_{T+h-2|T} \\ &\vdots \\ &= \phi_1^h x_{T|T} = \phi_1^h x_T. \end{aligned}$$

6.7 Forecast intervals

In practice the coefficients of an ARMA model must be estimated (see previous Chapter) and so the h-step-ahead forecast will contain also the estimation error. We denote the forecast obtained with estimated coefficients as $\widehat{x}_{T+h|T}$. So for example for an AR(1) we have an estimated coefficient $\widehat{\phi}_1$ and the forecast is

$$\widehat{x}_{T+h|T} = \widehat{\phi}_1^h x_T.$$

The forecast error is now defined as

$$\widehat{\epsilon}_{T+h|T} = x_{T+h} - \widehat{x}_{T+h|T}$$

and will also depend on the error made in estimating the parameters. For example in the AR(1) case the one-step-ahead forecast error is

$$\begin{aligned} \widehat{\epsilon}_{T+1|T} &= x_{T+1} - \widehat{\phi}_1 x_T \\ &= \phi_1 x_T + u_{T+1} - \widehat{\phi}_1 x_T \\ &= u_{T+1} + (\phi_1 - \widehat{\phi}_1) x_T. \end{aligned}$$

While the h-step-ahead forecast error of the AR(1) is

$$\begin{aligned} \widehat{\epsilon}_{T+h|T} &= x_{T+h} - \widehat{\phi}_1^h x_T \\ &= \epsilon_{T+h|T} + x_{T+h|T} - \widehat{\phi}_1^h x_T \\ &= \epsilon_{T+h|T} + (\phi_1^h - \widehat{\phi}_1^h) x_T. \end{aligned}$$

The variance of the h-step-ahead estimated forecast error is then

$$\begin{aligned} \widehat{\sigma}^2_{T+h|T} \equiv \mathrm{E}[\widehat{\epsilon}^2_{T+h|T}] &= \mathrm{E}[(x_{T+h} - \widehat{\phi}_1^h x_T)^2] \\ &= \mathrm{E}[\epsilon^2_{T+h|T}] + \mathrm{E}[(\phi_1^h - \widehat{\phi}_1^h)^2 x_T^2] \\ &= \underbrace{\sigma^2_{T+h|T}}_{\text{forecast error variance}} + \underbrace{\mathrm{E}[(\phi_1^h - \widehat{\phi}_1^h)^2 x_T^2]}_{\simeq \text{estimation error variance}} \end{aligned} \tag{32}$$

The quantity above is also called h-step-ahead Mean Squared Forecast Error (MSFE). As we see as $T \to \infty$ the estimation error variance (second term) tends to zero and the MSFE approaches

the forecast error variance defined in previous sections.[37] Moreover, the MSFE has the variance of the process as its upper limit

$$\widehat{\sigma}^2_{T+h|T} \to \text{Var}(X_t), \qquad T, h \to \infty.$$

The quantity $\sqrt{\widehat{\sigma}^2_{T+h|T}}$ is of particular interest and it is called Root Mean Squared Forecast Error (RMSFE).

6.7.1 Analytical formulas

For a given ARMA model, we usually have an analytical expression of the forecast error, hence, if data are Gaussian, we can compute the 95% forecast interval defined as

$$\widehat{x}_{T+h|T} \pm 1.96\sqrt{\widehat{\sigma}^2_{T+h|T}}.$$

A 68% forecast interval is given by

$$\widehat{x}_{T+h|T} \pm \sqrt{\widehat{\sigma}^2_{T+h|T}}.$$

This is a measure of the uncertainty associated with the forecast of a given model. Some examples follow.

1. AR(1). From (32), we have for $h = 1$

   $$\widehat{\sigma}^2_{T+1|T} = \text{E}[\widehat{\epsilon}^2_{T+1|T}] = \text{E}[(x_{T+1} - \widehat{\phi} x_T)^2] = \text{E}[\widehat{u}^2_T] \simeq \widehat{\sigma}^2_u,$$

   where $\widehat{u}_T$ is the residual of the estimated AR(1). The last equality is not exact as we have estimation errors, but if T is large we have

   $$\widehat{\sigma}^2_u = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}^2_T \xrightarrow{p} \text{E}[\widehat{u}^2_T], \quad T \to \infty,$$

   and $\widehat{\phi} \xrightarrow{p} \phi$, as $T \to \infty$.

   For $h = 2$ we have the estimated model

   $$x_{T+2} = \widehat{\phi} x_{T+1} + \widehat{u}_{T+2} = \widehat{\phi}^2 x_T + \widehat{\phi}\widehat{u}_{T+1} + \widehat{u}_{T+2}$$

   thus the estimated forecast error is

   $$\widehat{\epsilon}_{T+2|T} = \widehat{\phi} u_{T+1} + \widehat{u}_{T+2}$$

   which gives the MSFE

   $$\widehat{\sigma}^2_{T+2|T} = \text{E}[\widehat{\epsilon}^2_{T+2|T}] = \text{E}[(\widehat{\phi}\widehat{u}_{T+1} + \widehat{u}_{T+2})^2] \simeq (1 + \widehat{\phi}^2)\widehat{\sigma}^2_u,$$

   where we assumed that $\widehat{u}_t$ is white noise (it should if the model is correctly specified).

[37] Clearly since the estimated coefficients are mean square consistent then we have (using Cauchy Schwarz inequality)

$$\text{E}[(\widehat{x}_{T+h|T} - x_{T+h|T})^2] = \text{E}[(\widehat{\phi}^h_1 - \phi^h_1)^2 x^2_T] \leq \sqrt{\text{E}[(\widehat{\phi}^h_1 - \phi^h_1)^2]\text{E}[x^2_T]} \to 0, \qquad T \to \infty.$$

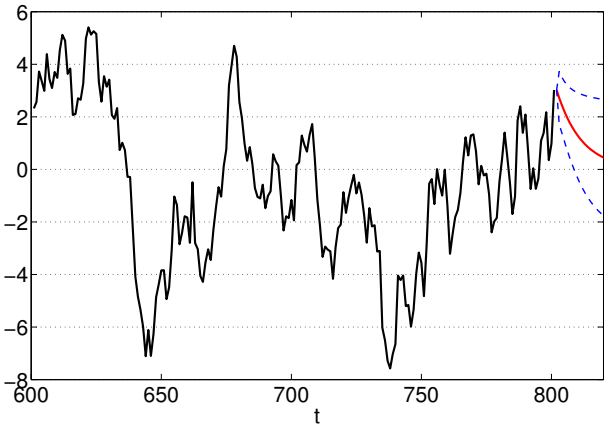


Figure 42: Forecasts of an AR(1) model with $\phi_1 = 0.9$ up to lag 20 with 68% confidence intervals.

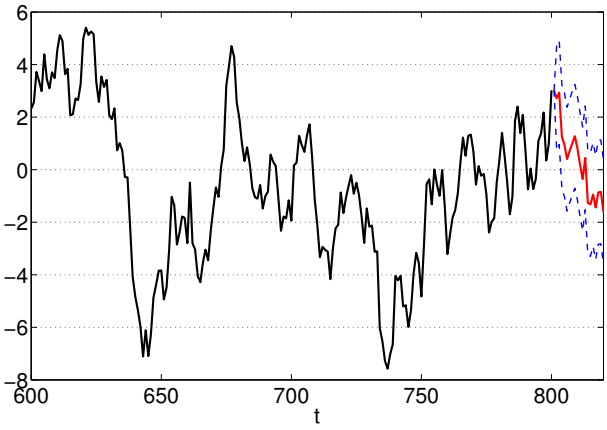


Figure 43: Series of $\tau = 20$ one-step-ahead forecast of an AR(1) model with $\phi_1 = 0.9$ with 68% confidence intervals.

Iteratively we can compute $\widehat{\sigma}^2_{T+h|T}$ for any h. But when h is large enough we can also use the limit result that is we can approximate the h-step-ahead MSFE with the estimated variance of the process

$$\widehat{\sigma}^2_{T+h|T} \simeq \widehat{\sigma}^2_x = \frac{\widehat{\sigma}^2_u}{1-\widehat{\phi}^2}.$$

One example is in Figure 42 when using the analytical formulas above. We compute one forecast for each $h = 1, \ldots, 20$ and the relative confidence interval. We see how the confidence interval increases as h increases, as the longer is the forecasting horizon, the more the uncertainty. Indeed, the variance of the forecast error is always smaller than the variance of the process thus reaching its limit from below which means it increases as h increases.

We often also report a series of forecasts with relative errors, when fixing an horizon h. An example is in Figure 43 where we fix $h = 1$ and we compute 20 one-step-ahead forecasts with their confidence intervals.

2. AR(2). For $h = 1$ we have the estimated model

$$x_{T+1} = \widehat{\phi}_1 x_T + \widehat{\phi}_2 x_{T-1} + \widehat{u}_{T+1}$$

therefore the MSFE is

$$\widehat{\sigma}^2_{T+1|T} = \mathrm{E}[\widehat{u}^2_{T+1}] \simeq \widehat{\sigma}^2_u$$

For $h = 2$ we have the estimated model

$$\begin{aligned} x_{T+2} &= \widehat{\phi}_1 x_{T+1} + \widehat{\phi}_2 x_T + \widehat{u}_{T+2} \\ &= \widehat{\phi}_1 x_T + \widehat{\phi}_1\widehat{\phi}_2 x_{T-1} + \widehat{\phi}_2 x_T + \widehat{\phi}_1\widehat{u}_{T+1} + \widehat{u}_{T+2} \end{aligned} \tag{33}$$

hence the MSFE is

$$\widehat{\sigma}^2_{T+2|T} = \mathrm{E}[(\widehat{\phi} u_{T+1} + \widehat{u}_{T+2})^2] \simeq (1 + \widehat{\phi}^2_1)\widehat{\sigma}^2_u,$$

where we assumed that $\widehat{u}_t$ is white noise (it should if the model is correctly specified).

Iteratively we can compute $\widehat{\sigma}^2_{T+h|T}$ for any h. But when h is large enough we can also use the limit result that is we can approximate the h-step-ahead MSFE with the estimated variance of the process

$$\widehat{\sigma}^2_{T+h|T} \simeq \widehat{\sigma}^2_x = \frac{\widehat{\sigma}^2_u}{1 - \widehat{\phi}^2_1 - \widehat{\phi}^2_2}.$$

6.7.2 Numerical approach

In practice models can be evaluated according to their performance in forecasting real data rather than by computing analytically the forecast intervals. In this case, we should have a record of observed values x_{T+h} to compare our forecasts with. This is typically done by not using all available data to estimate the model and leaving the last observations as pseudo out of sample values. For example, suppose we are able to compute a series of length τ of h-step-ahead forecasts $\widehat{x}_{T+t+h|T+t}$ for $t = 0 \dots \tau - 1$, then we define the h-step-ahead estimated MSFE as

$$\widehat{S}^2(h) = \frac{1}{\tau}\sum_{t=0}^{\tau-1}(\widehat{x}_{T+h+t|T+t} - x_{T+h+t})^2,$$

where the forecast $\widehat{x}_{T+h+t|T+t}$ is computed by estimating the model using all data up to time $T+t$.

For example, for an AR(1) we have

$$\begin{aligned} \widehat{S}^2(1) &= \frac{1}{\tau}\sum_{t=0}^{\tau-1}(\widehat{\phi}_1 x_{T+t} - x_{T+1+t})^2, \\ \widehat{S}^2(h) &= \frac{1}{\tau}\sum_{t=0}^{\tau-1}(\widehat{\phi}^h_1 x_{T+t} - x_{T+h+t})^2. \end{aligned}$$

Notice that we have

$$\widehat{S}^2(h) \xrightarrow{p} \mathrm{E}[(\widehat{x}_{T+h|T} - x_{T+h})^2] = \mathrm{E}[\widehat{\epsilon}^2_{T+h|T}] = \widehat{\sigma}^2_{T+h|T}, \qquad \tau \to \infty.$$

So in this sense $\widehat{S}^2(h)$ is an approximated (estimated) measure of the MSFE and for large τ is an alternative estimator of the MSFE. This measure is built using observed data and comparing them with realised data without computing the MSFE analytically. Moreover, if also $T \to \infty$ and $h \to \infty$ we know that $\widehat{\sigma}^2_{T+h|T} \to \mathrm{Var}(X_t)$ (see previous section). Hence,

$$\widehat{S}^2(h) \xrightarrow{p} \mathrm{Var}(X_t), \qquad \tau, h, T \to \infty.$$

Under the assumption that the forecast error is normally distributed, and therefore the innovations u_t are also normally distributed, a 95% forecast interval is given by

$$\widehat{x}_{T+h+t|T+t} \pm 1.96 \sqrt{\widehat{S}^2(h)}, \qquad t = 0 \dots \tau - 1.$$

Sometimes also the 68% confidence interval is used

$$\widehat{x}_{T+h+t|T+t} \pm \sqrt{\widehat{S}^2(h)}, \qquad t = 0 \dots \tau - 1.$$

Notice that for a series of forecasts, indexed by t, we have a series of upper and lower bounds for any fixed h as in Figure 43.

7 Models for heteroscedastic time series

7.1 Financial returns

The main variable of interest in financial analysis of time series is an asset return. Since stock prices usually exhibit trends and are therefore are non-stationary we work with first differences which are the so-called returns. Given a price time series p_t, there are two types of return definitions: simple and log returns (we have seen these already in Chapter 1).

1. Simple returns
$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}.$$
In other words, r_t is the gross return generated by holding the asset for one period.

2. Log returns.
$$\epsilon_t = \log p_t - \log p_{t-1} = \log \frac{p_t}{p_{t-1}} = \log(1 + r_t).$$
Notice that $\epsilon_t \in (-\infty, +\infty)$, thus it is possible to have negative returns lower than -100%.

If price variations are small, then log and simple returns are very close to each other. This can be seen via a Taylor expansion. If r_t is close to 0, then

$$\varepsilon_t = \log(1 + r_t) \approx \log 1 + \left.\frac{1}{1 + r_t}\right|_{r_t=0} r_t = r_t.$$

In both cases we are taking the first difference of a non-stationary variable and we then have a stationary variable. In practice, it is customary to work with log returns and to express them as a percentage:

$$\varepsilon_t = 100 \times (\log p_t - \log p_{t-1}).$$

7.2 Financial data

The models we are going to study here are used for the analysis of

1. (US) Common Stocks: IBM, Apple, General Motors, Goldman Sachs, ...

2. Stock Indices: S&P 500 index, Dow Jones Industrial average, FTSE 100, CAC 40, DAX, ...

3. ETFs: Index ETFs, Commodity ETFs, ...
   Exchange Traded Funds are investment fund which can be traded on an exchange. ETFs are designed to track the performance of specific types of investment such as investing on an index, a country, a commodity and so forth

4. ...but the models considered are also useful for: exchange rates, inflation, interest rates

Typically financial time series are recorded at high frequency as monthly, daily, or even intradaily. In the latter case observations might not even be equally spaced.

7.3 Stylized Facts

Mandelbrot (1956) lists the following styled facts for financial time series:

1. non-stationarity of prices p_t (random walk), stationarity of returns r_t or ε_t;
2. absence of autocorrelation of returns ε_t, i.e. they can be modelled as a white noise process;
3. non-zero autocorrelation of ε_t^2 or $|\varepsilon_t|$, i.e. returns are not independent;
4. volatility clustering, i.e. large returns (in absolute value) tend to be followed by large returns (in absolute value), and vice versa;
5. fat–tailed distribution of returns with kurtosis $\kappa_\varepsilon > 3$, i.e. non–Gaussian (leptokurtic);
6. leverage effects, i.e. negative returns (decrease in prices) tend to increase volatility by a larger amount than positive returns.

In Figures 44, 45 and 46 we see that returns appear to have weak or no serial dependence but squared returns appear to have strong serial dependence.

These stylised facts have been documented starting from at least the 1960's but the first models able to capture volatility clustering were proposed starting from the 1980's. We are going to analyse volatility clustering and introduce non-linear dynamic models called Generalised Autoregressive Conditional Heteroscedastic (GARCH) models which are able to capture most of the above facts. We will not consider here models to capture leverage effects that can be obtained by generalising the GARCH.

7.4 Volatility Models

The strong evidence of serial dependence in absolute and square returns suggest that the scale of returns changes in time. In other words, the variance of the process is time varying. In order to capture volatility clustering, we need to introduce appropriate time series processes able to model this behavior.

Consider a stationary process $\{X_t\}$, i.e. with $\mathrm{E}[X_t] = 0$ and $\mathrm{Var}[X_t] = \sigma_x^2$ which do not depend on time. We are interested in the conditional moments of $\{X_t\}$ when we condition on the set $\mathcal{I}_{t-1} = (X_{t-1}, X_{t-2}, \ldots)$. We define conditional mean μ_t of the process as

$$\mu_t = \mathrm{E}[X_t|\mathcal{I}_{t-1}]$$

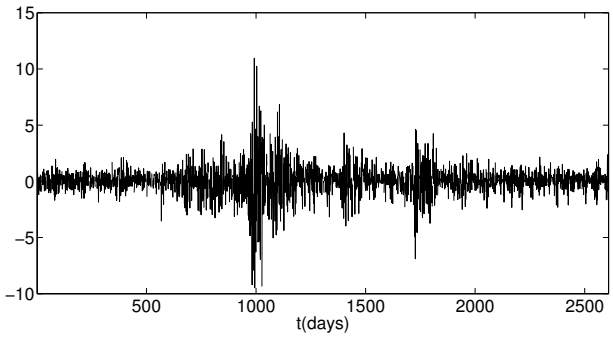


Figure 44: Series of daily returns of S&P500.

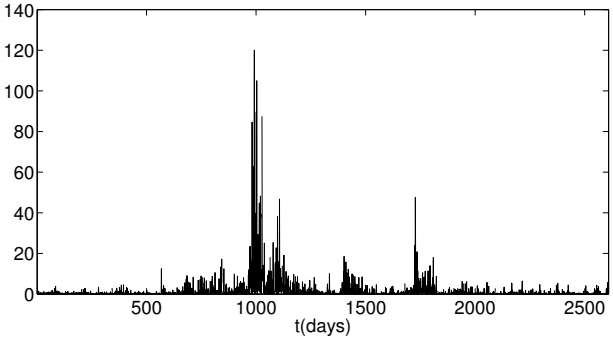


Figure 45: Series of daily squared returns of S&P500.

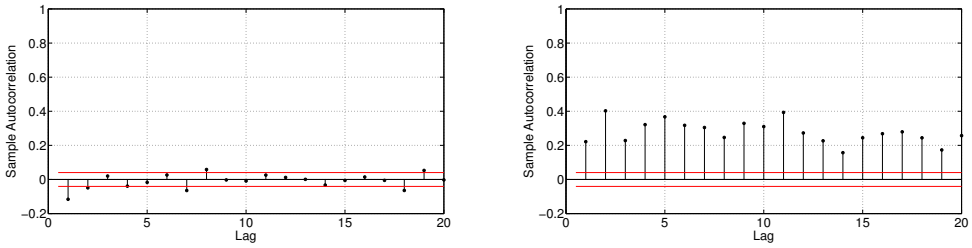


Figure 46: Autocorrelations of returns (left) and squared returns (right) of S&P500.

and conditional variance σ_t^2 as

$$\sigma_t^2 = \text{Var}[y_t|\mathcal{I}_{t-1}] = \text{E}[(y_t - \mu_t)^2|\mathcal{I}_{t-1}]$$

Hereafter, we use the shorthand notation $\text{E}_{t-1}[\cdot]$ in place of $\text{E}[\cdot|\mathcal{I}_{t-1}]$ for simplicity.

Consider a stationary and causal ARMA(1,1) model (notice that now the errors are independent not just uncorrelated)

$$x_t = \phi_1 x_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \qquad \varepsilon_t \sim i.i.d.(0, \sigma_\varepsilon^2)$$

We then have

$$\mu_t = \text{E}_{t-1}[X_t] = \phi_1 x_{t-1} + \theta_1 \varepsilon_{t-1}$$

and

$$\begin{aligned}\sigma_t^2 &= \text{Var}_{t-1}[X_t] \\ &= \text{E}_{t-1}[(X_t - \text{E}_{t-1}[X_t])^2] \\ &= \text{E}_{t-1}[(X_t - \mu_t)^2] \\ &= \text{E}_{t-1}[(X_t - (\phi_1 x_{t-1} + \theta_1 \varepsilon_{t-1}))^2] \\ &= \text{E}_{t-1}[\varepsilon_t^2] = \sigma_\varepsilon^2\end{aligned}$$

A similar result holds for a generic ARMA(p, q). Thus, while the conditional mean of an ARMA is time varying, the conditional variance of an ARMA is constant. In general, an ARMA(p, q) is not able to capture time varying conditional variance. This is true if the errors are i.i.d. or if they are white noise but Normal.

Otherwise ε_t could in principle depend non-linearly on $\mathcal{I}_{t-1}$. For example ε_t^2 could depend on x_{t-1}. So a more general model could be an ARMA(1,1) but with a different specification for the errors

$$\begin{aligned}x_t &= \phi_1 x_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \qquad \varepsilon_t \sim wn(0, \sigma_\varepsilon^2) \\ \varepsilon_t &= \sigma_t z_t, \qquad z_t \sim i.i.d.(0, 1),\end{aligned}$$

where the volatility $\sigma_t > 0$ is a measurable function of $\mathcal{I}_{t-1}$, i.e. $\sigma_t = f(x_{t-1}, x_{t-2}, \ldots)$ and $\{Z_t\}$ is the process of innovations of $\{X_t\}$ which are independent of $\mathcal{I}_{t-1}$. In this model we have

$$\text{Var}_{t-1}[X_t] = \text{E}_{t-1}[\varepsilon_t^2] = \text{E}_{t-1}[\sigma_t^2 Z_t^2] = \text{E}_{t-1}[\sigma_t^2]\text{E}[Z_t^2] = \sigma_t^2,$$

which is time varying.

Since financial returns have (nearly) zero autocorrelation we can model them as white noise, i.e. we can assume $\mu_t = 0$. Then, the general model that is able to capture a time varying conditional variance (called conditional heteroscedastic model) is given by

$$x_t \equiv \varepsilon_t = \sigma_t z_t, \qquad z_t \sim i.i.d.(0, 1). \tag{34}$$

The innovations are assumed to be independent (which is a stronger assumption than being just white noise), i.e. for any $h \neq 0$,

$$\text{E}[g_1(z_t) g_2(z_{t-h})] = \text{E}[g_1(z_t)]\text{E}[g_2(z_{t-h})],$$

and are also independent of past returns, i.e. for any $h > 0$

$$\text{E}[g_1(z_t) g_2(\varepsilon_{t-h})] = \text{E}[g_1(z_t)]\text{E}[g_2(\varepsilon_{t-h})],$$

for any function g_1, g_2.

In particular, the consequences of the above assumptions on the innovations are that

$$\begin{aligned}\text{E}[z_t^2 z_{t-h}^2] &= \text{E}[z_t^2]\text{E}[z_{t-h}^2], \qquad h \neq 0 \\ \text{E}[z_t^2 \varepsilon_{t-h}^2] &= \text{E}[z_t^2]\text{E}[\varepsilon_{t-h}^2], \qquad h > 0 \\ \text{E}[z_t \varepsilon_{t-h}] &= \text{E}[z_t]\text{E}[\varepsilon_{t-h}], \qquad h > 0 \\ \text{E}[z_t \sigma_t] &= \text{E}[z_t]\text{E}[\sigma_t], \\ \text{E}[z_t^2 \sigma_t^2] &= \text{E}[z_t^2]\text{E}[\sigma_t^2].\end{aligned}$$

Using the above assumptions, we can compute the moments of a generic conditional heteroskedastic model as (34).

1. Mean
$$\mu = \mathrm{E}[\varepsilon_t] = \mathrm{E}[\sigma_t z_t] = \mathrm{E}[\sigma_t]\mathrm{E}[z_t] = 0.$$

2. Variance
$$\sigma_\varepsilon^2 = \mathrm{E}[\varepsilon_t^2] = \mathrm{E}[\sigma_t^2 z_t^2] = \mathrm{E}[\sigma_t^2]\mathrm{E}[z_t^2] = \mathrm{E}[\sigma_t^2].$$

3. Conditional mean
$$\mathrm{E}_{t-1}[\varepsilon_t] = \mathrm{E}_{t-1}[\sigma_t z_t] = \sigma_t \mathrm{E}_{t-1}[z_t] = 0.$$

4. Conditional variance
$$\mathrm{E}_{t-1}[\varepsilon_t^2] = \mathrm{E}_{t-1}[\sigma_t^2 z_t^2] = \mathrm{E}_{t-1}[\sigma_t^2]\mathrm{E}_{t-1}[z_t^2] = \sigma_t^2$$

5. Autocovariance
$$\begin{aligned} \gamma_h^\varepsilon = \mathrm{Cov}(\varepsilon_t \varepsilon_{t-h}) &= \mathrm{E}[\varepsilon_t \varepsilon_{t-h}] - \mathrm{E}[\varepsilon_t]\mathrm{E}[\varepsilon_{t-h}] = \mathrm{E}[\varepsilon_t \varepsilon_{t-h}] \\ &= \mathrm{E}[\sigma_t z_t \varepsilon_{t-h}] = \mathrm{E}[z_t]\mathrm{E}[\sigma_t \varepsilon_{t-h}] = 0, \end{aligned}$$
thus ε_t is a white noise process. However, in general
$$\begin{aligned} \gamma_h^{\varepsilon^2} = \mathrm{Cov}(\varepsilon_t^2 \varepsilon_{t-h}^2) &= \mathrm{E}[\varepsilon_t^2 \varepsilon_{t-h}^2] - \mathrm{E}[\varepsilon_t^2]\mathrm{E}[\varepsilon_{t-h}^2] \\ &= \mathrm{E}[\sigma_t^2 z_t^2 \varepsilon_{t-h}^2] - \sigma_\varepsilon^4 = \mathrm{E}[z_t^2]\mathrm{E}[\sigma_t^2 \varepsilon_{t-h}^2] - \sigma_\varepsilon^4 \neq 0, \end{aligned}$$
thus ε_t is not an independent process.

6. Kurtosis
$$\begin{aligned} \kappa_\varepsilon &= \frac{\mathrm{E}[\varepsilon_t^4]}{(\mathrm{E}[\varepsilon_t^2])^2} = \frac{\mathrm{E}[\sigma_t^4 z_t^4]}{(\mathrm{E}[\sigma_t^2 z_t^2])^2} \\ &= \frac{\mathrm{E}[\sigma_t^4]\mathrm{E}[z_t^4]}{(\mathrm{E}[\sigma_t^2])^2(\mathrm{E}[z_t^2])^2} = \kappa_z \frac{\mathrm{E}(\sigma_t^4)}{(\mathrm{E}[\sigma_t^2])^2} \\ &= \kappa_z \left[1 + \frac{\mathrm{Var}(\sigma_t^2)}{(\mathrm{E}[\sigma_t^2])^2}\right] \end{aligned}$$
therefore $\kappa_\varepsilon > \kappa_z$ so even if $z_t \sim N(0,1)$, i.e. $\kappa_z = 3$ we still have that ε_t has kurtosis $\kappa_\varepsilon > 3$ thus with tails that are fatter than the Gaussian case. This reflects the fact that extreme events are more likely to happen in financial data.

In order to model volatility, the literature has proposed specific types of time series models. There are two main approaches in modelling the conditional variance σ_t^2:

1. ARCH Approach: σ_t^2 is a deterministic equation
2. Stochastic Volatility Approach: σ_t^2 is a stochastic equation

In practice, the ARCH approach is more popular while Stochastic Volatility models are typically harder to work with and we are not considering them here.

7.5 The ARCH model

In order to capture volatility clustering, in 1982 Robert Engle proposed the AutoRegressive Conditional Heteroskedasticity (ARCH) model. This simple model has started the literature on nonlinear quantitative modelling of financial time series.

The ARCH(1) model is defined as

$$\begin{aligned} \varepsilon_t &= \sqrt{\sigma_t^2} z_t \quad z_t \sim i.i.d.(0,1) \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 \end{aligned}$$

where $\omega > 0$ and $\alpha \geq 0$ in order to have always $\sigma_t > 0$. The current conditional variance of returns is proportional to the past squared return.

We can check for stationarity of ε_t. Since ε_t is a white noise, the process is stationary if its variance is finite and independent of time. We have

$$\begin{aligned} \sigma_\varepsilon^2 = \text{Var}(\varepsilon_t) &= \text{E}[\varepsilon_t^2] = \text{E}[\sigma_t^2] \\ &= \text{E}[\omega + \alpha \varepsilon_{t-1}^2] = \omega + \alpha \text{E}[\varepsilon_{t-1}^2] \end{aligned}$$

If ε_t is stationary then we must have $\text{E}[\varepsilon_t^2] = \text{E}[\varepsilon_{t-1}^2] = \sigma_\varepsilon^2$ and by substituting above we get

$$\sigma_\varepsilon^2 = \frac{\omega}{1-\alpha}$$

Since the variance must be positive this requires the constraint $\alpha < 1$.

The kurtosis can be computed from the formula in previous section

$$\kappa_\varepsilon = \kappa_z \left[1 + \frac{\text{Var}(\sigma_t^2)}{(\text{E}[\sigma_t^2])^2} \right] = \kappa_z \left[\frac{\text{E}(\sigma_t^4)}{(\text{E}[\sigma_t^2])^2} \right].$$

where $\kappa_z = \text{E}[z_t^4]/(\text{E}[z_t^2])^2 = \text{E}[z_t^4]$. Then, we have

$$\begin{aligned} \text{E}[\sigma_t^4] &= \omega^2 + \alpha^2 \text{E}[\varepsilon_{t-1}^4] + 2\omega\alpha \text{E}[\sigma_{t-1}^2] \\ &= \omega^2 + \alpha^2 \text{E}[\sigma_{t-1}^4] \kappa_z + 2\omega\alpha \text{E}[\sigma_{t-1}^2] \end{aligned}$$

If we assume strong stationarity of ε_t then $\text{E}[\sigma_t^4] = \text{E}[\sigma_{t-1}^4]$ and we have

$$\text{E}[\sigma_t^4] = \frac{\omega^2 + 2\omega\alpha \text{E}[\sigma_{t-1}^2]}{1 - \alpha^2 \kappa_z}$$

and

$$\kappa_\varepsilon = \frac{\text{E}[\sigma_t^4]}{\left\{\text{E}[\sigma_t^2]\right\}^2} \kappa_z = \frac{\omega^2 + 2\omega\alpha \text{E}[\sigma_{t-1}^2]}{(1 - \alpha^2 \kappa_z) \left\{\text{E}[\sigma_t^2]\right\}^2} \kappa_z$$

Using $\text{E}[\sigma_t^2] = \text{E}[\epsilon_t^2] = \sigma_\varepsilon^2 = \omega/(1-\alpha)$ we get

$$\kappa_\varepsilon = \frac{1 - \alpha^2}{1 - \alpha^2 \kappa_z} \kappa_z.$$

If the innovations are Gaussian, $z_t \sim i.i.d.N(0,1)$ then $\kappa_z = 3$ and κ_ε is defined only for $0 \leq \alpha^2 < 1/3$.

More generally, we can define an ARCH(q) model as

$$\begin{aligned}\varepsilon_t &= \sqrt{\sigma_t^2} z_t \quad z_t \sim i.i.d.(0,1)\\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2\end{aligned}$$

where $\omega > 0$ and $\alpha_i \geq 0$ in order to have always $\sigma_t > 0$. Following the same reasoning as before we can prove that the variance is

$$\sigma_\varepsilon^2 = \frac{\omega}{1 - \alpha_1 - \ldots - \alpha_q}$$

and in order to have a well defined variance we need $\sum_{i=1}^q \alpha_i < 1$.

An ARCH(q) is equivalent to an AR(q) for the squared returns. Indeed (if the model is stationary) we can define the innovations of squared returns as usual as the observed series minus its conditional expectation:

$$\nu_t = \varepsilon_t^2 - \mathrm{E}_{t-1}[\varepsilon_t^2] = \varepsilon_t^2 - \sigma_t^2,$$

then $\sigma_t^2 = \varepsilon_t^2 - \nu_t$ and the ARCH(q) becomes

$$\varepsilon_t^2 - \nu_t = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2,$$

which is an AR(q)

$$\varepsilon_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \nu_t.$$

Now ν_t is a white noise as we can write it as

$$\nu_t = \sigma_t^2 z_t^2 - \sigma_t^2 = (z_t^2 - 1)\sigma_t^2,$$

and its acs is

$$\begin{aligned}\gamma_h^\nu &= \mathrm{Cov}(\nu_t, \nu_{t-h}) = \mathrm{E}[\nu_t \nu_{t-h}]\\ &= \mathrm{E}[\nu_t (z_{t-h}^2 - 1)\sigma_{t-h}^2]\\ &= \mathrm{E}[\nu_t \sigma_{t-h}^2]\mathrm{E}[z_{t-h}^2 - 1] = 0,\end{aligned}$$

since z_{t-h} is independent of σ_{t-h}^2 and of ν_t, and $\mathrm{E}[z_{t-h}^2] = 1$. However, ν_t is not an i.i.d. process.

The AR representation allows for computing easily the acs of ε_t^2. Recall that for an AR(1)

$$x_t = \phi x_{t-1} + u_t$$

we have $\rho_h^x = \phi^h$. An ARCH(1) is equivalent to an AR(1) for the squared returns, thus

$$\rho_h^{\varepsilon^2} = \mathrm{Corr}(\varepsilon_t^2, \varepsilon_{t-h}^2) = \alpha^h.$$

Notice that $\rho_h^{\varepsilon^2} > 0$ for any h.

For higher order ARCH we have to solve Yule Walker difference equations

$$\gamma_h^{\varepsilon^2} = \sum_{i=1}^q \alpha_i \gamma_{h-i}^{\varepsilon^2}.$$

So for example, for an ARCH(2) we have

$$\begin{aligned}\rho_2^{\varepsilon^2} &= \alpha_1\rho_1^{\varepsilon^2} + \alpha_2\rho_0^{\varepsilon^2}, \\ \rho_1^{\varepsilon^2} &= \alpha_1\rho_0^{\varepsilon^2} + \alpha_2\rho_{-1}^{\varepsilon^2} = \alpha_1\rho_0^{\varepsilon^2} + \alpha_2\rho_1^{\varepsilon^2}.\end{aligned}$$

Then,

$$\frac{\rho_2^{\varepsilon^2}}{\rho_1^{\varepsilon^2}} = \alpha_1 + \alpha_2\frac{\rho_0^{\varepsilon^2}}{\rho_1^{\varepsilon^2}} = \alpha_1 + \frac{\alpha_2}{\rho_1^{\varepsilon^2}}$$

Thus in general we do not have $\rho_2^{\varepsilon^2} < \rho_1^{\varepsilon^2}$ and the acs do not decrease to zero monotonically.

An alternative parameterization, sometimes useful for forecasting, is the exponential weighted moving average. Using the result of the variance of the ARCH(1) we can reparameterize the conditional variance equation as

$$\begin{aligned}\sigma_t^2 &= \frac{(1-\alpha)(\omega + \alpha\varepsilon_{t-1}^2)}{1-\alpha} \\ &= \frac{\omega}{1-\alpha} + \frac{-\alpha\omega + (1-\alpha)\alpha\varepsilon_{t-1}^2}{1-\alpha} \\ &= \frac{\omega}{1-\alpha} + \frac{-\alpha\omega}{1-\alpha} + \alpha\varepsilon_{t-1}^2 \\ &= \sigma_\varepsilon^2 - \alpha\sigma_\varepsilon^2 + \alpha\varepsilon_{t-1}^2 \\ &= \sigma_\varepsilon^2 + \alpha(\varepsilon_{t-1}^2 - \sigma_\varepsilon^2) \\ &= (1-\alpha)\sigma_\varepsilon^2 + \alpha\varepsilon_{t-1}^2\end{aligned}$$

Therefore, the ARCH(1) specification shows that the conditional variance is made of two components: its mean which is $\mathrm{E}[\sigma_t^2] = \mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2$, and an error that is mean reverting since $\alpha < 1$. Equivalently we see that in the ARCH(1) model the conditional variance is a weighted average of the unconditional variance and the squared returns.

7.5.1 Forecasting with ARCH

Using the AR representation, forecast formulas of the variance of the ARCH(1) are analogous to those of the AR(1). Suppose to have observations $\varepsilon_1 \ldots \varepsilon_T$. We define the one-step-ahead forecast of the conditional variance as the conditional expectation of future values of σ_t^2 given its past[38]

$$\sigma_{T+1|T}^2 = \mathrm{E}_T[\sigma_{T+1}^2] = \mathrm{E}_T[\omega + \alpha\varepsilon_T^2] = \omega + \alpha\varepsilon_T^2.$$

The 2-step-ahead forecast of the conditional variance is

$$\begin{aligned}\sigma_{T+2|T}^2 &= \mathrm{E}_T[\omega + \alpha\varepsilon_{T+1}^2] = \omega + \alpha\mathrm{E}_T[\varepsilon_{T+1}^2] = \omega + \alpha\mathrm{E}_T[\sigma_{T+1}^2] \\ &= \omega + \alpha(\omega + \alpha\varepsilon_T^2) = \omega(1+\alpha) + \alpha^2\varepsilon_T^2.\end{aligned}$$

[38] Notice that here the theory of linear prediction does not apply as the model is non-linear.

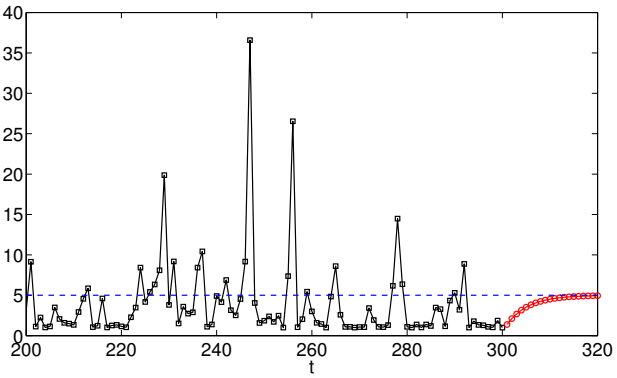


Figure 47: Forecast of an ARCH(1) model with $\alpha = 0.8$ up to lag 20. The blue line is the unconditional variance σ_ε^2.

The h-step-ahead forecast of the variance is

$$\begin{aligned}
\sigma^2_{T+h|T} &= \mathrm{E}_T[\omega + \alpha\varepsilon^2_{T+h-1}] \\
&= \omega + \alpha\mathrm{E}_T[\varepsilon^2_{T+h-1}] \\
&= \omega + \alpha\mathrm{E}_T[\sigma^2_{T+h-1}z^2_{T+h-1}] \\
&= \omega + \alpha\mathrm{E}_T[\sigma^2_{T+h-1}]\mathrm{E}_T[z^2_{T+h-1}] \\
&= \omega + \alpha\mathrm{E}_T[\sigma^2_{T+h-1}] \\
&= \omega + \alpha\mathrm{E}_T[\omega + \alpha\varepsilon^2_{T+h-2}] \\
&= \omega + \alpha\omega + \alpha^2\mathrm{E}_T[\varepsilon^2_{T+h-2}] \\
&\vdots \\
&= \omega\left(\sum_{k=0}^{h-1}\alpha^k\right) + \alpha^h\varepsilon^2_T \\
&= \frac{\omega(1-\alpha^h)}{1-\alpha} + \alpha^h\varepsilon^2_T,
\end{aligned}$$

where the last step comes from the sums of geometric series. Thus, the h-steps-ahead forecast of an ARCH(1) depends only on the returns at time T which is the Markovian property of an AR(1) model. Notice also that if ω were zero then we would have exactly the formulas we had for the AR(1) of a zero mean process. Generalisations to the ARCH(q) case will be based on the AR(q) forecasts.

Moreover, as $h \to \infty$, we have (recall that $\alpha < 1$)

$$\sigma^2_{T+h|T} = \omega\left(\sum_{k=0}^{h-1}\alpha^k\right) + \alpha^h\varepsilon^2_T \to \omega\sum_{k=0}^{\infty}\alpha^k = \frac{\omega}{1-\alpha} = \sigma^2_\varepsilon.$$

As in the ARMA case, the h-steps-ahead forecasts converges to its expected value which in the ARCH case is the unconditional variance of the process σ^2_ε. An example is in Figure 47.

Now let us consider the forecast error denoted as $v_{T+h|T}$ and its variance. In the simple case $h = 1$ we have

$$v_{T+1|T} = \sigma^2_{T+1} - \sigma^2_{T+1|T} = \omega + \alpha\varepsilon^2_T - \omega - \alpha\varepsilon^2_T = 0$$

indeed the ARCH equation has no error in it, thus if we know the observation at time T we know everything about the conditional variance at time $T+1$. If $h=2$ we have

$$v_{T+2|T} = \sigma^2_{T+2} - \sigma^2_{T+2|T} = \alpha\left(\varepsilon^2_{T+1} - (\omega + \alpha\varepsilon^2_T)\right) = \alpha\left(\sigma^2_{T+1}(z^2_{T+1}-1)\right) = \alpha\nu_{T+1}$$

which is a white noise (see above). If $h \to \infty$ we find the same results as for ARMA, i.e. the variance of the forecast error tends to the variance of the process we are forecasting which in this case is $\text{Var}(\sigma^2_t)$. Indeed, the h-step-ahead forecast error is

$$v_{T+h|T} = \sigma^2_{T+h} - \sigma^2_{T+h|T} = \omega + \alpha\varepsilon^2_{T+h-1} - \left(\omega\left(\sum_{k=0}^{h-1}\alpha^k\right) + \alpha^h\varepsilon^2_T\right)$$

hence

$$\text{E}[v_{T+h|T}] = \omega + \alpha\sigma^2_\varepsilon - \left(\omega\left(\sum_{k=0}^{h-1}\alpha^k\right) + \alpha^h\sigma^2_\varepsilon\right).$$

Notice that the expectation of the forecast error is not zero. Then the variance of the forecast error is

$$\begin{aligned}
\text{Var}(v_{T+h|T}) &= \text{E}\left[(v_{T+h|T} - \text{E}[v_{T+h|T}])^2\right] \\
&= \text{E}[(\alpha^h\varepsilon^2_T - \alpha^h\sigma^2_\varepsilon - \alpha\varepsilon^2_{T+h-1} + \alpha\sigma^2_\varepsilon)^2] \\
&= \text{E}[(\alpha^h(\varepsilon^2_T - \sigma^2_\varepsilon) - \alpha(\varepsilon^2_{T+h-1} - \sigma^2_\varepsilon))^2] \\
&= \text{E}[\alpha^{2h}(\varepsilon^2_T - \sigma^2_\varepsilon)^2] + \text{E}[\alpha^2(\varepsilon^2_{T+h-1} - \sigma^2_\varepsilon)^2] \\
&\quad -2\alpha^{h+1}\text{E}[(\varepsilon^2_T - \sigma^2_\varepsilon)(\varepsilon^2_{T+h-1} - \sigma^2_\varepsilon)] \\
&= \alpha^{2h}\text{Var}(\varepsilon^2_T) + \alpha^2\text{Var}(\varepsilon^2_{T+h-1}) - 2\alpha^{h+1}\gamma^{\varepsilon^2}_{h-1}.
\end{aligned}$$

Moreover,

$$\text{Var}(\sigma^2_t) = \text{Var}(\omega + \alpha\varepsilon^2_{t-1}) = \alpha^2\text{Var}(\varepsilon^2_t).$$

Therefore, if ε^2_t is stationary, from the expression above we have, as $h \to \infty$, and since $\alpha < 1$,

$$\text{Var}(v_{T+h|T}) \to \alpha^2\text{Var}(\varepsilon^2_t) = \text{Var}(\sigma^2_t).$$

The same results can be derived using the exponential weighted average parametrisation of previous Section.

7.5.2 Detecting ARCH effects

We can use a simple test to detect the presence of ARCH effects using the AR representation. First, estimate the coefficients of the following autoregression by Least Squares

$$\varepsilon^2_t = \alpha_0 + \alpha_1\varepsilon^2_{t-1} + \ldots + \alpha_q\varepsilon^2_{t-q} + \nu_t,$$

Then the null hypothesis of no ARCH effects is formulated as $H_0 : \alpha_1 = 0, \alpha_2 = 0, \ldots, \alpha_q = 0$. The test statistic for H_0 is then[39]

$$LM = T \cdot R^2$$

where R^2 is the usual "R-squared" coefficient of the linear regression. Under the null of no ARCH effects the test statistic LM is asymptotically distributed as a χ^2_q.

[39] This is an example of Lagrange Multiplier LM test.

7.6 GARCH

In practice, only rather rich ARCH(q) parameterizations, i.e. with large q, are able to fit financial series adequately. However, largely parameterised models can be unstable in forecasting and hard to estimate because we need to estimate many parameters. In order to overcome the shortcomings of the ARCH, Bollerslev (1986) proposed the generalised ARCH model called GARCH. The model allows to fit financial returns adequately while keeping the number of parameters small. The GARCH model is indeed one of the most successfully employed volatility models.

The GARCH(1,1) model is defined as

$$\begin{aligned} \varepsilon_t &= \sqrt{\sigma_t^2} z_t, \qquad z_t \sim i.i.d.(0,1) \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned}$$

where $\omega > 0$, $\alpha \geq 0$ and $\beta > 0$ in order to have a positive variance. If we compute the variance of the process (as for the ARCH case) we have

$$\begin{aligned} \sigma_\varepsilon^2 = \text{Var}(\varepsilon_t) &= \text{E}[\varepsilon_t^2] = \text{E}[\sigma_t^2] \\ &= \text{E}[\omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2] = \omega + \alpha \text{E}[\varepsilon_{t-1}^2] + \beta \text{E}[\sigma_{t-1}^2] \\ &= \omega + \alpha \text{E}[\varepsilon_{t-1}^2] + \beta \text{E}[\varepsilon_{t-1}^2] \end{aligned}$$

where the last step is because $\text{E}[\varepsilon_t^2] = \text{E}[\sigma_t^2] = \sigma_\varepsilon^2$. If ε_t is stationary then we must have $\text{E}[\varepsilon_t^2] = \text{E}[\varepsilon_{t-1}^2] = \sigma_\varepsilon^2$ and by substituting above we get

$$\sigma_\varepsilon^2 = \frac{\omega}{1 - \alpha - \beta}.$$

Since the variance must be positive this requires the constraint $\alpha + \beta < 1$.

The GARCH(p, q) model is[40]

$$\begin{aligned} \varepsilon_t &= \sqrt{\sigma_t^2} z_t, \qquad z_t \sim i.i.d.(0,1) \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_p \sigma_{t-p}^2, \end{aligned}$$

where $\omega > 0$, $\alpha_i \geq 0$ and $\beta_i > 0$ and by competing the variance and following the same reasoning as before we have the constraint

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1.$$

This condition is necessary and sufficient for having weak stationarity of ε_t.[41] Therefore, for a GARCH(p, q) the variance is given by

$$\sigma_\varepsilon^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}.$$

The sum of all GARCH coefficients α_is and β_is is called persistence of the volatility process. Typically, we get very high values for this measure, i.e. very close to one, which implies high variance.

[40] Notice that an ARCH(q) is a GARCH($0, q$) but a GARCH($p, 0$) does not make sense as the equation of the conditional variance will not depend on any observed variable.

[41] Checking for strong stationarity of GARCH(1,1) is more complicated and we do not do it here.

Again by assuming strong stationarity, it can be proved that the kurtosis for GARCH(1,1) is

$$\kappa_\varepsilon = \frac{1-(\alpha+\beta)^2}{1-(\alpha+\beta)^2-\alpha^2(\kappa_z-1)}\kappa_z$$

thus even if $z_\sim i.i.d.N(0,1)$, i.e. $\kappa_z = 3$ we have

$$\kappa_\varepsilon = \frac{3(1-(\alpha+\beta)^2)}{1-(\alpha+\beta)^2-2\alpha^2} > 3$$

Therefore, the GARCH model can capture fat tails, i.e. tails that have more mass than those of a Gaussian distribution.

As for ARCH we have two alternative parameterisations. The exponential weighted moving average representation is based on representing a GARCH as an ARCH(∞) model. So for example, for a GARCH(1,1) we have

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta[\omega + \alpha\varepsilon_{t-2}^2 + \beta(\omega + \alpha\varepsilon_{t-3}^2 + \beta\sigma_{t-3}^2)] \\
&= \omega\left(\sum_{i=0}^{\infty}\beta^i\right) + \alpha\sum_{i=0}^{\infty}\beta^i\varepsilon_{t-i-1}^2 \\
&= \frac{\omega}{1-\beta} + \alpha\sum_{i=0}^{\infty}\beta^i\varepsilon_{t-i-1}^2
\end{aligned}$$

This representation shows how a GARCH(1,1) is a parsimonious way of characterizing ARCH dynamics. The conditional variance of a GARCH(1,1) can be seen as a weighted average of recent returns such that the weight given to past information decreases exponentially fast, since $\beta < 1$.

Equivalently the GARCH(p, q) can be seen as an ARMA for squared returns. As before, we define the innovations $\nu_t = \varepsilon_t^2 - \sigma_t^2$. Then, in case of the GARCH(1,1) we get:

$$\varepsilon_t^2 = \omega + (\alpha+\beta)\varepsilon_{t-1}^2 + \nu_t - \beta\nu_{t-1},$$

which is an ARMA(1,1). In general for a GARCH(p, q) we get:

$$\varepsilon_t^2 = \omega + \sum_{i=1}^{r}(\alpha_i+\beta_i)\varepsilon_{t-i}^2 + \nu_t - \sum_{j=1}^{p}\beta_j\nu_{t-j}$$

where $r = \max(p, q)$ and this is an ARMA(r, p). As before we can prove that ν_t is a white noise but it is not independent.

The ARMA representation allows for computing the acvs of ε_t^2. For a GARCH(1,1) we can use the results for ARMA(1,1) for the squared returns (see Chapter 3)

$$\rho_h^{\varepsilon^2} = \frac{\text{Cov}(\varepsilon_t^2, \varepsilon_{t-h}^2)}{\text{Var}(\varepsilon_t^2)} = \rho_1^{\varepsilon^2}(\alpha+\beta)^{h-1}$$

where

$$\rho_1^{\varepsilon^2} = \frac{\alpha[1-\beta(\alpha+\beta)]}{1-(\alpha+\beta)^2+\alpha^2}$$

Notice that $\rho_h^{\varepsilon^2} > 0$ always and in this case acvs are decreasing monotonically.

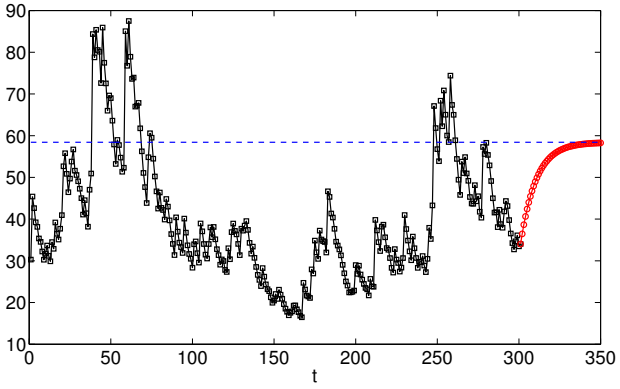


Figure 48: Forecast of a GARCH(1,1) model with $\alpha = 0.1$ and $\beta = 0.8$ up to lag 100. The blue line is the unconditional variance σ_ε^2.

7.6.1 Forecasting with GARCH

The one-step-ahead forecast of the conditional variance in a GARCH(1,1) model is

$$\sigma_{T+1|T}^2 = \mathrm{E}_T[\sigma_{T+1}^2] = \omega + \alpha\varepsilon_T^2 + \beta\sigma_T^2$$

which shows that the one-step-ahead forecast error is zero. The 2-step-ahead forecast is

$$\begin{aligned}\sigma_{T+2|T}^2 &= \mathrm{E}_T[\sigma_{T+2}^2] = \omega + \alpha\mathrm{E}_T[\varepsilon_{T+1}^2] + \beta\sigma_{T+1|T}^2 \\ &= \omega + \alpha\sigma_{T+1|T}^2 + \beta(\omega + \alpha\varepsilon_T^2 + \beta\sigma_T^2) \\ &= \omega + (\alpha+\beta)\sigma_{T+1|T}^2 = \sigma_\varepsilon^2 + (\alpha+\beta)(\sigma_{T+1|T}^2 - \sigma_\varepsilon^2)\end{aligned}$$

since $\omega = \sigma_\varepsilon^2 - \sigma_\varepsilon^2(\alpha+\beta)$. In the same way we can find the h-step-ahead forecast as

$$\sigma_{T+h|T}^2 = \sigma_\varepsilon^2 + (\alpha+\beta)^{h-1}(\sigma_{T+1|T}^2 - \sigma_\varepsilon^2),$$

and we immediately see that since $\alpha + \beta < 1$, as $h \to \infty$, we have $\sigma_{T+h|T}^2 \to \sigma_\varepsilon^2$. An example is in Figure 48 where we see that the reversion to the mean is very slow as in this case $\alpha + \beta = 0.9$.

7.7 Limitations of GARCH

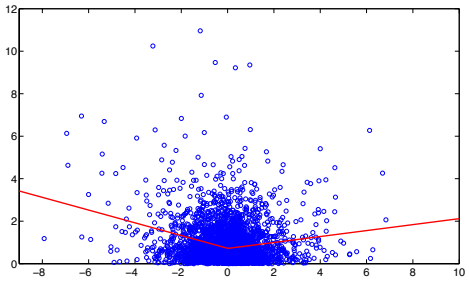
The simple GARCH(p, q) has some limitations

The most important is that it cannot take into account the dependence between volatility and the sign of past returns

- Standard GARCH models assume that positive and negative error terms have a symmetric effect on the volatility, i.e. good and bad news have the same effect on the volatility
- In many real situations the volatility reacts asymmetrically to the sign of the shocks. In particular negative past returns have a bigger effect on σ_t^2 than positive returns of the same size

- This dependence is due to the leverage effect, i.e. a negative shock to returns would increase the debt to equity ratio which in turn will increase uncertainty of future returns

S&P500 absolute returns $|\varepsilon_t|$ vs lagged returns ε_{t-1}



- Consider the model for returns

$$\begin{aligned} \epsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \end{aligned}$$

- Then we can re-write the volatility as

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i z_{t-i}^2 \sigma_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

  which is invariant to changes in the sign of z_t

In GARCH models the volatility σ_t is an even function of ϵ_{t-i}, $i > 0$. For example in a simple ARCH(1) we can write $\sigma_t = h(\epsilon_{t-1}^2)$ and, if the density of z_t is symmetric, we have

$$\text{Cov}(\sigma_t, \epsilon_{t-1}) = \text{E}[\sigma_t \epsilon_{t-1}] = \int h(\epsilon_{t-1}^2) \epsilon_{t-1} f_z(z_t) \text{d}z_t = 0$$

and in general it can be proved that for GARCH models

$$Cov(\sigma_t, \epsilon_{t-h}) = 0, \quad h > 0.$$

which is equivalent to

$$Cov(\epsilon_t^+, \epsilon_{t-h}) = Cov(\epsilon_t^-, \epsilon_{t-h}) = 0, \quad h > 0,$$

with

$$\epsilon_t^+ = \max(\epsilon_t, 0), \quad \epsilon_t^- = \min(\epsilon_t, 0).$$

- Empirically we find that

$$\text{Corr}(\epsilon_t, \epsilon_{t-h}) \simeq 0, \quad \text{Corr}(\epsilon_t^2, \epsilon_{t-h}^2) > 0, \quad \text{Corr}(|\epsilon_t|, |\epsilon_{t-h}|) > 0,$$

  which are properties reproduced by GARCH models

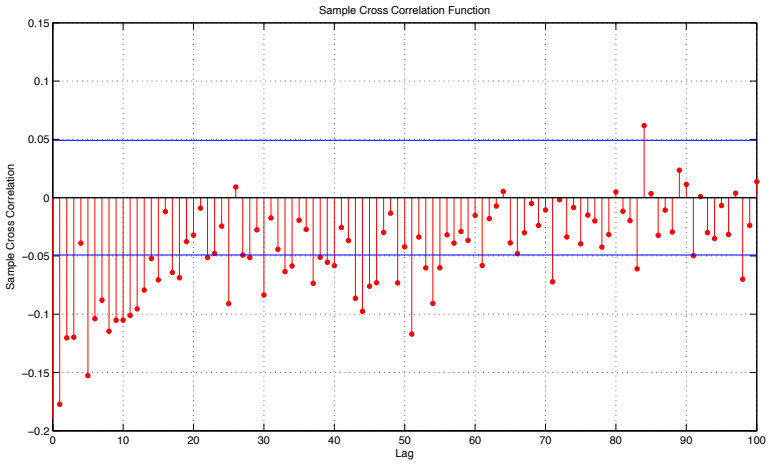
- But we also find $\text{Corr}(\epsilon_t^+, \epsilon_{t-h}) < 0$. When $\epsilon_t = \sigma_t z_t$, with σ_t a positive function of ϵ_t as in GARCH, we have

$$\text{Corr}(\epsilon_t^+, \epsilon_{t-h}) = K\text{Cov}(\sigma_t, \epsilon_{t-h}) < 0,$$

  for some constant $K > 0$

- This is the leverage effect that GARCH models cannot reproduce

- GARCH models with leverage (asymmetric) effects are the subject of next lectures



$\text{Corr}(\epsilon_t^+, \epsilon_{t-h})$

7.8 Estimation of GARCH models

ARCH models are typically estimated by Maximum Likelihood and the estimator has no closed form expression and needs to be found numerically. Moreover, if we have a general model with a conditional mean specified as an ARMA, then the ARMA parametes also depend on the ARCH parameters

Consider the ARMA$(1,1)$-GARCH$(1,1)$ model for a zero mean time series

$$\begin{aligned} x_t &= \mu_t + \epsilon_t, \\ \mu_t &= \phi x_{t-1} + \theta \epsilon_{t-1}, \\ \epsilon_t &= \sqrt{\sigma_t^2} z_t, \qquad z_t \sim i.i.d.(0,1) \\ \sigma_t^2 &= \omega + \alpha_i \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2, \end{aligned}$$

where $\omega_0 > 0$, $\alpha \geq 0$, $\beta \geq 0$ and $|\phi| < 1$, $\alpha + \beta < 1$. We then have the vectors of parameters to be estimated $\boldsymbol{m} = (\phi\, \theta)'$, $\boldsymbol{s} = (\omega\, \alpha\, \beta)'$.

Assume to observe T realizations $(x_1 \dots x_T)$ of $\{X_t\}$. We know that the likelihood is given by the joint density of the observations

$$f(x_1 \dots x_T) = f(x_T | x_1 \dots x_{T-1}) f(x_1 \dots x_{T-1}) = f(x_0) \prod_{t=1}^{T} f(x_t | x_1 \dots x_{t-1}) = f(x_0) \prod_{t=1}^{T} f(x_t | \mathcal{I}_{t-1}).$$

So we need to find $f(x_t|\mathcal{I}_{t-1})$ which is the conditional density. For the model above we have

$$z_t = \frac{\varepsilon_t}{\sigma_t} = \frac{x_t - \mu_t}{\sigma_t} = \frac{x_t - \phi x_{t-1} - \theta\varepsilon_{t-1}}{\sqrt{\omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2}}$$

thus z_t and its density will depend on the parameters of the model. If the pdf of z_t is know then we can write

$$f(x_t|\mathcal{I}_{t-1}) = f_z(z_t)\left|\frac{\partial z_t}{\partial x_t}\right| = f_z(z_t)\frac{1}{\sigma_t}$$

notice that the density of z_t is not conditional as the process is i.i.d.

Usually we assume that $z_t \sim i.i.d.N(0,1)$ and we have the likelihood for the observed data[42]

$$f(x_t|\mathcal{I}_{t-1}) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp\left(-\frac{z_t^2}{2}\right) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp\left(-\frac{(x_t-\mu_t)^2}{2\sigma_t^2}\right)$$

from which we see that the conditional distribution of x_t given information up to time $t-1$ is Gaussian with mean the conditional mean given by the ARMA part and variance the conditional variance given by the GARCH part.

By taking logs we have the log-likelihood for all observations

$$L_T(\mathbf{x}_T, \boldsymbol{m}, \boldsymbol{s}) = \sum_{t=1}^{T} \log f(x_t, \boldsymbol{m}, \boldsymbol{s}|\mathcal{I}_{t-1})$$

and the Maximum Likelihood (ML) estimator of the parameters $\boldsymbol{m}$ and $\boldsymbol{s}$ is defined as

$$(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{s}}) = \arg\max_{\boldsymbol{m},\, s} \; L_T(\mathbf{x}_T, \boldsymbol{m}, \boldsymbol{s}).$$

As for all ML estimators this estimator is consistent

$$(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{s}}) \xrightarrow{p} (\boldsymbol{m}, \boldsymbol{s}), \qquad T \to \infty$$

and is asymptotically normal

$$\sqrt{T}\left[(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{s}}) - (\boldsymbol{m}, \boldsymbol{s})\right] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}), \qquad T \to \infty.$$

The form of $\boldsymbol{\Sigma}$ is complex and depends on derivatives of L_T and on the parameters too. The diagonal elements of the matrix $\boldsymbol{\Sigma}$ can be used to build confidence intervals for the parameters of the model. For an ARMA(1,1)-GARCH(1,1) the matrix is 5×5. If the distribution of z_t is symmetric (as in the Gaussian case) we have that the matrix has a block diagonal structure

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{m}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{s}} \end{pmatrix}.$$

We have asymptotic independence between estimated ARMA and GARCH coefficients. However, the distribution of the estimators of the ARMA coefficients depends on the GARCH coefficients, while on the other hand the asymptotic accuracy in the estimated GARCH coefficients is not affected by the ARMA part. Therefore if we are interested only in the GARCH parameters but we have also an ARMA component, we can first estimate an ARMA with heteroskedastic errors and then take the residuals and estimate a GARCH. This two step procedure is simpler and fast.

[42] Gaussian distribution is the typical choice (see the discussion in Chapter 4 for ARMA) but in financial data also the Student-t distribution is a popular model.

7.9 Diagnostics

Once we have estimated a GARCH(1,1) model we can check for model adequacy by inspection of the so called "standardized residuals", defined as

$$\widehat{z}_t = \frac{\varepsilon_t}{\widehat{\sigma}_t} = \frac{\varepsilon_t}{\sqrt{\widehat{\omega} + \widehat{\alpha}\varepsilon_{t-1}^2 + \widehat{\beta}\widehat{\sigma}_{t-1}^2}}$$

where $\widehat{\omega}$ and $\widehat{\alpha}$, and $\widehat{\beta}$, are obtained by ML under a given distribution D (e.g. a Gaussian). If the specification is correct, standardised residuals should

1. be approximately distributed according to distribution D;
2. do not exhibit dependence in levels, absolute levels, square levels, etc... i.e. they must be i.i.d. We can test for acvs of $\widehat{z}_t$ and $\widehat{z}_t^2$ to be zero as we did for the errors of an ARMA model.

8 Non-stationary processes

8.1 Trend and difference stationary processes

We have seen two kinds of non-stationarity.

1. The presence of a linear trend
$$y_t = a + bt + x_t$$
where $\{X_t\}$ is stationary. The linear trend makes the mean time varying. By taking first differences we have a stationary process
$$(1 - L)y_t = b + (1 - L)x_t.$$
Or alternatively we can regress $\{Y_t\}$ on a constant and a linear trend and the residual would be $\{X_t\}$ which is stationary. In this case $\{Y_t\}$ is called Trend Stationary (TS). Generalizations to trends of higher order (as quadratic trend) are also possible (see Chapter 2).

2. The presence of a unit root in the AR polynomial of an ARMA. So for example the AR(1) with unit coefficient (also known as random walk)
$$y_t = y_{t-1} + u_t, \qquad u_t \sim wn(0, \sigma_u^2)$$
or equivalently using the MA representation
$$y_t = \sum_{k=0}^{\infty} u_{t-k}, \qquad u_t \sim wn(0, \sigma_u^2)$$
or
$$y_t = y_0 + \sum_{k=0}^{t-1} u_{t-k},$$

for which the mean is zero but the variance is infinite

$$\text{Var}(Y_t) = \text{Var}\left(\sum_{k=0}^{\infty} u_{t-k}\right) = \infty.$$

or if we treat y_0 as given

$$\text{Var}(Y_t) = \text{Var}\left(y_0 + \sum_{k=0}^{t-1} u_{t-k}\right) = t\sigma_u^2 \to \infty.$$

By taking first differences we have a stationary process

$$(1-L)y_t = y_t - y_{t-1} = u_t,$$

which is a white noise by assumption. In this case $\{Y_t\}$ is Difference Stationary (DS) and the general case is considered below.

Let us consider the two cases more in detail.[43]

8.2 Trend stationary processes

Consider the process

$$y_t = a + bt + x_t$$

where $\{X_t\}$ is zero mean and stationary. Analysis of the TS process is elementary and we also refer to Chapter 2. Here the source of non stationarity is a deterministic function of time which has no relationship with the stationary component. The process has time varying mean $\text{E}[Y_t] = a + bt$ but the variance is

$$\text{Var}(Y_t) = \text{Var}(X_t)$$

which is not time varying since $\{X_t\}$ is stationary.

The prediction of the TS process is

$$y_{T+h|T} = a + b(T+h) + x_{T+h|T}.$$

Since we know that $x_{T+h|T} \to \text{E}[X_t] = 0$ as $h \to \infty$, the long-run prediction of y_{T+h} is just the trend, i.e. for h large we have

$$y_{T+h|T} \simeq a + b(T+h).$$

In the TS case the influence of y_T on the predicted values tends to zero as h increases and exactly as in the stationary case we have mean reversion, where now the mean is a linear deterministic trend.

As an example consider the case in which $\{X_t\}$ is an AR(1)

$$x_t = \phi x_{t-1} + u_t, \qquad u_t \sim wn(0, \sigma_u^2).$$

[43] Remember that non-stationarity does not mean necessarily that the process has a trend or a unit root. So the process

$$y_t = \begin{cases} x_t & \text{if} \quad t \leq \tau \\ 1 + ax_t & \text{if} \quad t > \tau \end{cases}$$

represents a regime-change (the mean and the variance suddenly change at $t = \tau$).

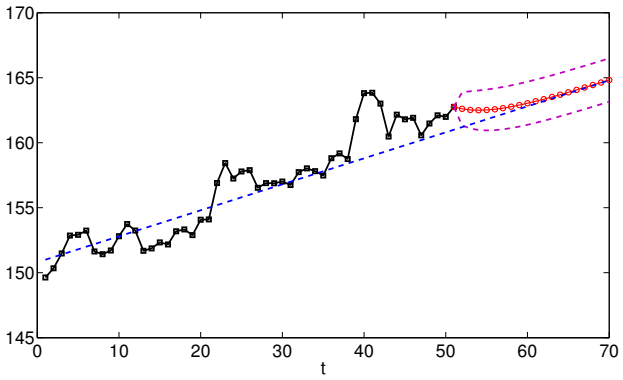


Figure 49: Forecast of a TS process with trend $(1+0.2t)$ (dashed blue line) and underlying AR(1) with parameter $\phi = 0.8$ joint with 68% confidence interval.

Then,

$$x_t = \sum_{k=0}^{\infty} \phi^k u_{t-k} = u_t + \phi u_{t-1} + \phi^2 u_{t-2} + \dots$$

and the h-step-ahead forecast error is

$$\epsilon_{T+h|T} = y_{T+h} - y_{T+h|T} = x_{T+h} - \phi^h x_T = u_{T+h} + \phi u_{T+h-1} + \ldots + \phi^{h-1} u_{T+1}$$

which has variance (i.e. mean squared forecast error)

$$MSFE(h) = \mathrm{E}[\epsilon^2_{T+h|T}] = \sigma_u^2(1 + \phi^2 + \ldots + \phi^{2(h-1)}) \to \frac{\sigma_u^2}{1-\phi^2}$$

i.e. it converges to the variance of x_t as $h \to \infty$, as in the stationary case.

In Figure 49 we see that the forecast of $\{Y_t\}$ tends to its trend and is driven by the AR(1) forecast with 68% confidence bands are given by

$$y_{T+h|T} \pm \sqrt{\sigma_u^2(1 + \phi^2 + \ldots + \phi^{2(h-1)})}$$

Hence the intervals revert to

$$a + b(T+h) \pm \sqrt{\sigma_u^2/(1-\phi^2)}$$

as h increases.

8.3 Difference stationary processes - Random walk with drift

The DS case is much more complicated. Here the source of non-stationarity is not a deterministic component, we say there is a stochastic trend. In the simple case we have an AR(1) model with a unit root and we say that the process is a random walk. If we include a linear deterministic trend, we have the random walk

$$y_t = b + y_{t-1} + u_t, \qquad u_t \sim wn(0, \sigma_u^2)$$

or using the MA representation

$$y_t = y_0 + bt + u_1 + \ldots + u_t$$

Usually we assume that the process starts at $t = 1$ and that y_0 is a given non-stochastic value. Then, conditional on y_0, we have

$$\mathrm{E}[Y_t] = y_0 + bt, \qquad \mathrm{Var}(Y_t) = \mathrm{Var}(u_1 + \ldots + u_t) = t\sigma_u^2$$

Thus both mean and variance are time varying and in particular the variance explodes (unlike the TS case). The deterministic trend $y_0 + bt$ has little importance. Indeed if we had $b = 0$, we would have a simple random walk, and the variance would still be exploding.

Compare the random walk with drift with a simple stationary AR(1) with non-zero mean

$$y_t = b + \phi y_{t-1} + u_t$$

where $|\phi| < 1$. If we write the MA representation of this process (as we did for the random walk) we get

$$y_t = (\phi^t y_0 + b(1 + \phi + \ldots + \phi^{t-1})) + (u_t + \phi u_{t-1} + \ldots + \phi^{t-1} u_1)$$

while the random walk will explode when $t \to \infty$, in this case we have

$$y_t \xrightarrow{p} \frac{b}{1-\phi} + \sum_{k=0}^{\infty} \phi^k u_{t-k}.$$

The second term is known, while the first is due to the non-zero mean of the AR(1) we are considering. We also know that the variance of a stationary AR(1) is finite $\mathrm{Var}(Y_t) = \sigma_u^2/(1-\phi^2)$.

The $T + h$ observation of a random walk with drift can be written as (imagine that the initial condition is not at time 0 but at time T and use the MA representation)

$$y_{T+h} = y_T + bh + u_{T+1} + \ldots u_{T+h}.$$

Then the h-step-ahead forecast and the related forecast error and its variance are

$$\begin{aligned} y_{T+h|T} &= y_T + bh \\ \epsilon_{T+h|T} &= u_{T+1} + \ldots + u_{T+h} \\ \sigma^2_{T+h|T} &= \mathrm{E}[\epsilon^2_{T+h|T}] = h\sigma_u^2 \end{aligned}$$

In the TS case the influence of y_T on the predicted values tends to zero as h increases exactly and as in the stationary case we had mean reversion. Now for a DS process, the effect of y_T on the predicted values never vanishes. Moreover, the prediction error variance tends to infinity as $h \to \infty$. The properties we had for stationary processes do not hold here. Figure 50 shows forecasts for a random walk with drift. By comparing with the TS case in Figure 49, we see that the size of the confidence interval increases with $\sqrt{h}$. Finally, in Figure 51 we consider two forecasts of the same random walk with drift but made at T and at $T - 10$. We see that the effect of the last used observation is persistent and never vanishes, the forecast being a linear extrapolation from last observation Y_T or Y_{T-10} never reverting to anything.

8.4 Difference stationary processes - General case with autocorrelated errors (ARIMA)

The general model is

$$y_t = b + y_{t-1} + x_t$$

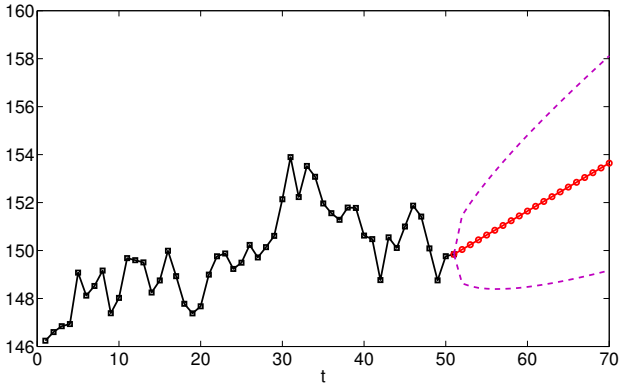


Figure 50: Forecast of a DS process (random walk with drift) with trend $b = 0.2$ joint with 68% confidence interval.

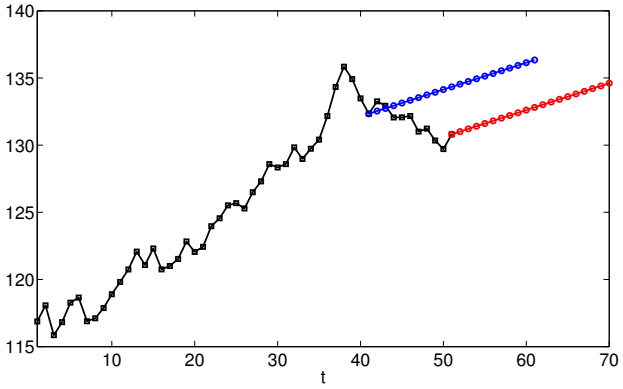


Figure 51: Two forecasts of a DS process (random walk with drift) with trend $b = 0.2$ when forecasting at T (red) and at $T - 10$ (blue).

where $\{X_t\}$ is zero mean stationary with non-zero autocorrelations, as an ARMA process. First of all notice that $\{\Delta Y_t\}$ is stationary, therefore also in this case $\{Y_t\}$ is DS. We have

$$\begin{aligned} y_1 &= b + y_0 + x_1 \\ y_2 &= b + y_1 + x_2 \\ &= b + b + y_0 + x_1 + x_2 \\ \dots & \\ y_t &= \underbrace{b + \ldots + b}_{t \text{ times}} + y_0 + \underbrace{x_1 + x_2 + \ldots + x_t}_{t-1 \text{ terms}} \end{aligned}$$

Then[44]

$$y_t = y_0 + bt + (1 + L + \ldots + L^{t-1})x_t = y_0 + bt + \frac{1 - L^t}{1 - L}x_t.$$

The Wold representation of an ARMA stationary process is such that

$$x_t = c(L)u_t, \qquad u_t \sim wn(0, \sigma_u^2),$$

[44]Using $\frac{1}{1-q} = \sum_{k=0}^{\infty} q^k$ and $\sum_{k=0}^{t-1} q^t = \frac{1-q^t}{1-q}$.

where $\sum_{k=0}^{\infty} c_k^2 < \infty$ and $c_0 = 1$. Notice that this requires $c_k \to 0$ as $k \to \infty$. The coefficient c_k contains the effect of u_{t-k} on x_t, which is therefore vanishing in the long-run, u_t is a transitory shock for x_t.

By substituting, we have

$$y_t = y_0 + bt + \frac{1 - L^t}{1 - L} c(L) u_t \tag{35}$$

Now, suppose first that $c(L) = (1 - L)d(L)$, which implies that $c(1) = \sum_{k=0}^{\infty} c_k = 0$ and we must have also $\sum_{k=0}^{\infty} d_k^2 < \infty$ which implies $d_k \to 0$ as $k \to \infty$. Then, from (35) we have

$$\begin{aligned} y_t &= y_0 + bt + \frac{1 - L^t}{1 - L} c(L) u_t \\ &= y_0 + bt + (1 - L^t) d(L) u_t \\ &= (y_0 - d(L)u_0) + bt + d(L)u_t \\ &= a + bt + d(L)u_t \end{aligned}$$

with $d(L)u_t$ stationary since it is an MA(∞) with square summable coefficients. But then $\{Y_t\}$ would be TS. And the effect of u_{t-k} on y_t is given by d_k, thus as k grows the effect becomes smaller (it tends to zero as $k \to \infty$) and we say that the effect is only temporary. This proves that the effect of innovations on stationary or TS processes is just transitory.

However, since we assumed that $\{Y_t\}$ is DS then we have that for a DS process we must always have $c(1) \neq 0$ and from (35) we have

$$\begin{aligned} y_t &= y_0 + bt + \frac{1 - L^t}{1 - L} c(L) u_t \\ &= \left(y_0 - \frac{1}{1 - L} u_0 \right) + bt + \frac{c(L)}{1 - L} u_t \\ &= a + bt + c(L) \sum_{k=0}^{\infty} u_{t-k} \\ &= a + bt + \sum_{h=0}^{\infty} c_h L^h \left(\sum_{k=0}^{\infty} u_{t-k} \right) \\ &= a + bt + \sum_{h=0}^{\infty} c_h \left(\sum_{k=0}^{\infty} u_{t-k-h} \right) \\ &= a + bt + \sum_{h=0}^{\infty} c_h \left(u_{t-h} + u_{t-h-1} + u_{t-h-2} + \ldots \right) \end{aligned}$$

where we set $c_0 = 1$ as usual. Then in this case the effect of u_{t-k} on y_t is $\sum_{h=0}^{k} c_h$, thus it never vanishes (it tends to $c(1)$ as $k \to \infty$) and we say the effect is permanent. The simple random walk case corresponds to $c(L) = 1$ which implies $c(1) = 1$.

The forecasts of y_{T+h} given T and given $T-1$ are (use the forecast of an MA process)

$$\begin{aligned}
y_{T+h|T} &= bh + y_T + x_{T+1|T} + \ldots + x_{T+h|T} \\
&= bh + y_T + \sum_{k=1}^{\infty} c_k u_{T+1-k} + \ldots + \sum_{k=h}^{\infty} c_k u_{T+h-k} \\
&= bh + y_T + (c_1 + c_2 + \ldots + c_h)u_T + (c_2 + c_3 + \ldots + c_{h+1})u_{T-1} + \ldots \\
y_{T+h|T-1} &= b(h+1) + y_{T-1} + x_{T|T-1} + \ldots + x_{T+h|T-1} \\
&= b(h+1) + y_{T-1} + \sum_{k=1}^{\infty} c_k u_{T-k} + \ldots + \sum_{k=h+1}^{\infty} c_k u_{T+h-k} \\
&= bh + y_T + (c_1 + c_2 + \ldots + c_{h+1})u_{T-1} + (c_2 + c_3 + \ldots + c_{h+2})u_{T-2} + \ldots
\end{aligned}$$

The difference between the two forecasts is then

$$\begin{aligned}
y_{T+h|T} - y_{T+h|T-1} &= y_T - y_{T-1} - b + (c_1 + c_2 + \ldots + c_h)u_T - c_1 u_{T-1} - c_2 u_{T-2} - \ldots \\
&= x_T + (c_1 + c_2 + \ldots + c_h)u_T - c_1 u_{T-1} - c_2 u_{T-2} - \ldots \\
&= \sum_{k=0}^{\infty} c_k u_{T-k} + (c_1 + c_2 + \ldots + c_h)u_T - c_1 u_{T-1} - c_2 u_{T-2} - \ldots \\
&= (1 + c_1 + c_2 + \ldots + c_h)u_T
\end{aligned}$$

and in the limit $h \to \infty$ we have

$$y_{T+h|T} - y_{T+h|T-1} \to \left(\sum_{k=0}^{\infty} c_k\right) u_T = c(1)u_T$$

and the quantity $c(1)$ is called measure of persistence of the process $\{Y_t\}$. It is the change in the long-run prediction due to a shock u_T, divided by u_T. Equivalently it is the change in long-run prediction due to a shock u_T of unitary value. Since it is a constant non-zero number the effect of a shock at time T is persistent at any future point in time. In Figure 52 we show the forecasts of the same random walk with drift made at T and at $T-1$ and with

$$x_t = 0.8x_{t-1} + u_t, \qquad u_t \sim w.n.(0,1)$$

i.e. with a unitary shock, i.e. $\sigma_u^2 = 1$. In this case $c_k = 0.8^k$ and the distance between the blue and red lines is then the persistence

$$c(1) = \sum_{k=0}^{\infty} c_k = \sum_{k=0}^{\infty} 0.8^k = \frac{1}{1-0.8} = 5.$$

Notice that the persistence of a TS process is zero. That is, the long-run prediction does not change with the last observation used. If y_t is TS, $y_t = a + bt + x_t$, with $\{X_t\}$ stationary and zero mean, then as $h \to \infty$

$$\begin{aligned}
y_{T+h|T} - y_{T+h|T-1} &= (a + b(T+h) + x_{T+h|T}) - (a + b(T+h) + x_{T+h|T-1}) \\
&= x_{T+h|T} - x_{T+h|T-1} \\
&= \sum_{k=h}^{\infty} c_k u_{T+h-k} - \sum_{k=h+1}^{\infty} c_k u_{T+h-k} = c_h u_T \to 0
\end{aligned}$$

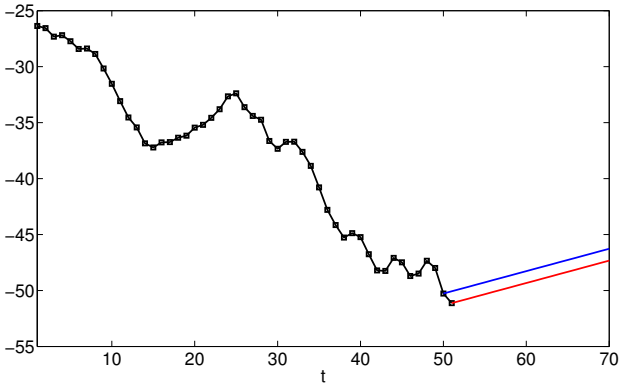


Figure 52: Two forecasts of a DS process (random walk with drift) with trend $b = 0.2$ and x_t is an AR(1) with parameter $\phi = 0.8$, when forecasting at T (red) and at $T - 1$ (blue).

since to have $c(1) = 0$ we must have $c_h \to 0$ as $h \to \infty$. For example if $\{X_t\}$ is a stationary AR(1) we have $c_h = \phi^h \to 0$ as $h \to \infty$. Obviously, in a TS process if $b = 0$ then $\{Y_t\}$ is stationary and again the persistence is zero.

Summing up, while the shock u_t of a DS process has a permanent effect, namely the change in the long-run prediction, the shock of a stationary or TS process has only a transitory effect.

A DS process $\{Y_t\}$ has a unit root, indeed

$$\begin{aligned}(1-L)y_t &= b + (1-L^t)c(L)u_t \\ &= (b - c(L)u_0) + c(L)u_t \\ &= k + c(L)u_t\end{aligned}$$

which is an ARMA with AR polynomial given by $(1 - L)$ that is with a root in $z = 1$. We also say that $\{Y_t\}$ is an integrated process of order 1 denoted as $Y_t \sim I(1)$, that is we must differentiate it once to have a stationary process. As a consequence we have that $\Delta Y_t \sim I(0)$.

Now since the last term is the Wold representation of $x_t = c(L)u_t$, we can assume a causal and invertible ARMA(p, q) model for $x_t = c(L)u_t$, such that we have $\Phi(L)x_t = \Theta(L)u_t$, then

$$(1 - L)\Phi(L)y_t = k + \Theta(L)u_t, \quad u_t \sim w.n.(0, \sigma_u^2)$$

where $\Phi(L)$ is of order p (with no roots inside or on the unit circle) and $\Theta(L)$ is of order q (with no roots inside or on the unit circle) then we say that $\{Y_t\}$ is a casual and invertible ARIMA$(p, 1, q)$ to highlight the unit root.[45] The constant term k accounts for the possible non-zero mean of $\{Y_t\}$.

8.5 Beveridge-Nelson decomposition

Now, let us assume that $b = 0$ as we have seen that its effect is just to introduce a linear trend. We now show that any $I(1)$ process can be written as a random walk process $\{P_t\}$ which is a permanent component of $\{Y_t\}$, i.e. it is $I(1)$, plus a zero-mean stationary process $\{T_t\}$ which is the transitory component of $\{Y_t\}$ and therefore is $I(0)$:

$$y_t = P_t + T_t$$

[45] If we need d differences of $\{Y_t\}$ to have a stationary process then $Y_t \sim I(d)$ and if $\{X_t\}$ is an ARMA(p, q) then $\{Y_t\}$ is an ARIMA(p, d, q).

This decomposition is called Beveridge-Nelson decomposition (BN).

Notice first that for any polynomial $c(z)$ of order q we can always find a polynomial $c^*(z)$ of order $q-1$ such that[46]

$$c(z) = c(1) + c^*(z)(1-z).$$

Now, since by definition we have that $\{\Delta Y_t\}$ is stationary then it has a Wold representation

$$\Delta y_t = c(L)u_t, \qquad u_t \sim wn(0, \sigma_u^2),$$

where $\sum_{k=0}^{\infty} c_k^2 < \infty$ and $c_0 = 1$. Recall from above that for an $I(1)$ process we always have $c(1) \neq 0$. Using the above decomposition of $c(L)$ we have

$$\Delta y_t = c(1)u_t + c^*(L)(1-L)u_t$$

and we see that we must have $\sum_{k=0}^{\infty} c_k^{*2} < \infty$ because Δy_t is stationary, so that the long run effect of u_t on y_t is only through $c(1)$.

Then, define the random walk $\mu_t = \mu_{t-1} + u_t$ or equivalently $\Delta\mu_t = u_t$, then

$$y_t = c(1)\mu_t + c^*(L)u_t = c(1)\left[\mu_0 + \sum_{k=0}^{t-1} u_{t-k}\right] + c^*(L)u_t$$

and therefore $P_t = c(1)\mu_t$ which is a random walk, and $T_t = c^*(L)u_t$ which is a stationary process (it is an MA(∞) with square summable coefficients). Given a time series we can decompose it into a cycle and a trend by first fitting an ARMA on first differences and then computing the BN decomposition using the estimated parameters. This is not the only possibility of decomposing a series into a permanent and transitory component and a common critique is that the permanent component might not be just a pure random walk.

Example: consider $Y_t \sim I(1)$ such that

$$\Delta y_t = u_t + 0.5u_{t-1} = (1 + 0.5L)u_t, \qquad u_t \sim w.n.(0,1).$$

Then, $c(1) = 1.5$ and $c^*(L) = -0.5$, indeed

$$[c(1) + c^*(L)(1-L)]u_t = 1.5u_t - 0.5u_t + 0.5u_{t-1} = u_t + 0.5u_{t-1} = c(L)u_t$$

and $P_t = 1.5\mu_t = 1.5\sum_{k=0}^{t-1} u_{t-k}$ and $T_t = -0.5u_t$.

8.6 Testing for unit roots

There are many ways to test for the presence of a unit root, here we just go through the main one and its underlying idea. Consider the AR(1)

$$y_t = \phi y_{t-1} + u_t,$$

[46] The new polynomial is the remainder of the quotient $c^*(L) = \frac{c(L)-c(1)}{1-L}$ such that

$$c^*(z) = \sum_{j=0}^{\infty} c_j^* z^j, \qquad c_j^* = -\sum_{i=j+1}^{q} c_i.$$

where $u_t \sim wn(0, \sigma^2)$. Then consider its first differences

$$\Delta y_t = y_t - y_{t-1} = (\phi - 1)y_{t-1} + u_t = \rho y_{t-1} + u_t,$$

where $\rho = (\phi - 1)$. Thus $\{Y_t\}$ has a unit root when $\rho = 0$. We might hope then to test this hypothesis by means of a t-test as in usual linear regressions. Given the least squares estimator of the above AR model we have the statistics

$$t = \frac{\hat{\rho}}{\sqrt{\text{Var}(\hat{\rho})}}$$

where $\text{Var}(\hat{\rho})$ is the variance of the estimator (see Chapter 5). That is indeed correct however there are some observations that we must make

1. The asymptotic distribution of t is not a Student-t nor asymptotically is Gaussian. It has a particular shape and its called Dickey-Fuller distribution and its critical value have been computed numerically. The statistic has zero mean under the null of unit root. When testing we have to look at its left tail only (negative values) because under the alternative of no unit root we want stationarity thus $|\phi| < 1$, i.e. $\rho < 0$.

2. In the above model we are testing for $\{Y_t\}$ to be a random walk (when $\phi = 1$) but that might not be the case as u_t might not be a white noise but an ARMA process. For example

   $$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

   where $\epsilon_t \sim wn(0, \sigma^2)$. Using the fact that $y_{t-k} = y_{t-1} - \sum_{i=1}^{k-1} \Delta y_{t-i}$ we get

   $$\Delta y_t = \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{p-1} \Delta y_{t-p+1} + \epsilon_t$$

   where $\rho = (\phi_1 + \ldots + \phi_p) - 1$ e $\gamma_i = -(\phi_{i+1} + \ldots + \phi_p)$. Also in this case we have a unit root if $\rho = 0$, indeed in that case we have

   $$\Delta y_t = \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{p-1} \Delta y_{t-p+1} + \epsilon_t$$

   which shows that Δy_t is stationary and therefore $y_t \sim I(1)$. A test based on the least squares estimator of ρ can be run as before. This is called the Augmented Dickey-Fuller test. The right number of lags to be included can be determined as usual by means of information criteria. The rule is that we must add as many lags as those necessary to have ϵ_t white noise, i.e. no serial dependence is left.

3. We might have as a starting model a random walk with drift, then the model to be considered for the test is

   $$\Delta y_t = a + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_p \Delta y_{t-p+1} + \epsilon_t$$

   and yet another distribution for the test $\rho = 0$ has to be used. Notice that under the null a is the slope of a linear trend.

An alternative derivation of these formulation are obtained using the polynomials in L, then for the Augmented Dickey-Fuller we have

$$\Phi(L) y_t = \epsilon_t$$

thus we have a unit root if $\Phi(1) = 0$. Then, we can write $\Phi(L) = 1 - B(L)L$ where $B(L)L = \sum_{k=1}^{p} \phi_k L^k$, therefore when we have a unit root we must have $B(1) = 1$. Now, the following are all equivalent

$$\begin{aligned}
\Phi(L)y_t &= \epsilon_t \\
\Phi(L)y_t - y_{t-1} &= -y_{t-1} + \epsilon_t \\
(1 - B(L)L)y_t - y_{t-1} &= -y_{t-1} + \epsilon_t \\
\Delta y_t - B(L)y_{t-1} &= -y_{t-1} + \epsilon_t \\
\Delta y_t &= [B(L) - 1]y_{t-1} + \epsilon_t
\end{aligned}$$

by applying the Beveridge-Nelson decomposition to $H(L) = B(L) - 1$ we have

$$\Delta y_t = [B(1) - 1]y_{t-1} + (1 - L)B^*(L)y_{t-1} + u_t,$$

where $B(1) = \sum_{k=1}^{p} \phi_k$, and $B^*(L) = \sum_{k=1}^{p-1} \gamma_k L^k$ with $\gamma_k = -\sum_{j=k+1}^{p} \phi_k$. We see that when we have unit root in $\{Y_t\}$ the first term disappears.

8.7 Spurious regression

Consider two $I(1)$ processes such that

$$\begin{aligned}
y_t &= y_{t-1} + \eta_t \\
x_t &= x_{t-1} + e_t
\end{aligned}$$

where η_t and e_t are two white noises mutually independent. The two processes are random walk and they are not dependent on each other. Therefore we might be tempted to consider the regression

$$y_t = \alpha + \beta x_t + u_t$$

and test if $\beta = 0$. It can however be proved that in this case the usual asymptotic theory of OLS does not apply and indeed we reject the null-hypothesis too often. As a consequence one might think that actually y_t and x_t are related to each other even if they are not, when this happens we are in presence of a spurious regression.

As an explanation consider the case in which $\beta = 0$, then we have $y_t = \alpha + u_t$ and therefore if $y_t \sim I(1)$ either $u_t \sim I(0)$ but then we have a contradiction or if $u_t \sim I(1)$ then the whole asymptotic theory of linear regression does not apply, since we would have autocorrelated residuals.[47] In other words, if y_t is not related to x_t then only u_t can take into account the unit root in y_t.

A possible solution is to add lags of y_t and of x_t to the regression as for example

$$y_t = \alpha + \phi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

in such a way that the lags of y_t capture the persistence in y_t thus making the errors less autocorrelated (they should be white noise). A case in which we can test for $\beta = 0$ in the usual way is in presence of cointegration when indeed by hypothesis y_t and x_t are correlated and u_t is a white noise (see Chapter 11).

[47] In this case we would have

$$\widehat{\beta} - \beta = \frac{\sum_{t=1}^{T} u_t x_t}{\sum_{t=1}^{T} x_t^2}$$

but when $u_t \sim I(1)$ as $T \to \infty$ both terms grow as T^2 and we cannot have consistency.

9 Spectral analysis of time series

The aim of this Chapter is to be able to represent a stationary stochastic process $\{X_t\}$ as a sum (possibly infinite) of sinusoidal components, in order to highlight the periodicities present in the data.

9.1 Fourier analysis

Let us start with a periodic deterministic real function of time $g(t)$ with period 2π, i.e. $g(t) = g(t + 2\pi)$, then due to periodicity such function can be considered only for $t \in [-\pi, \pi]$ and we assume $g \in L^2([-\pi, \pi])$, i.e. such that

$$\int_{-\pi}^{\pi} |g(\theta)|^2 \mathrm{d}\theta < \infty.$$

Then $g(t)$ can be represented as a sum of (possibly infinite) sinusoids. Define the sum of k sinusoids as

$$g_k(t) = \sum_{j=0}^{k} (a_j \cos(jt) + b_j \sin(jt)),$$

where $a_j = \frac{1}{\pi}\int_{-\pi}^{\pi} g(t)\cos(jt)\mathrm{d}t$ and $b_j = \frac{1}{\pi}\int_{-\pi}^{\pi} g(t)\sin(jt)\mathrm{d}t$ are the Fourier coefficients obtained by projecting $g(t)$ onto the orthonormal basis of the space made of cosine and sine functions, and obviously $b_0 = 0$. Then we can prove that

$$\int_{-\pi}^{\pi} [g_k(t) - g(t)]^2 \mathrm{d}t \to 0 \quad k \to \infty.$$

So every function of period 2π which is also in L^2 can be written with an infinite Fourier series.

Assume now the periodicity of $g(t)$ is $2T$ for some T therefore $g \in L^2([-T, T])$, then define a new function $h(t) = g(tT/\pi)$, then $h(t)$ is periodic with period 2π and has again a Fourier series representation. In this way we can map any periodic fucntion to functions on the space $L^2([-\pi, \pi])$ where now for $g(t)$ we use the basis $\cos(\pi jt/T)$ and $\sin(\pi jt/T)$, that is functions with period $2T/j$.

A period of a signal is also called wavelength and its inverse is called frequency which is measured in number of cycles per unit of time. Therefore, a period of $2T/j$ corresponds to a frequency $f_j = j/(2T)$ therefore the basis functions read also as $\cos(2\pi f_j t)$ and $\sin(2\pi f_j t)$. Another possibility is to use the angular frequency defined as $\theta_j = 2\pi f_j$.

If we then consider non-periodic functions, for an arbitrary T we can define a new function

$$g^*(t) = g(t), \qquad g^*(t + 2kT) = g^*(t) \text{ for } k = 1, 2, 3, \ldots$$

which is certainly periodic, therefore, it has a Fourier series

$$g^*(t) = \sum_{j=-\infty}^{\infty} (a_j \cos(2\pi f_j t) + b_j \sin(2\pi f_j t)), \qquad b_0 = 0,$$

and using the orthonormal basis $\{e^{i2\pi f_j t}\}$ in $L^2([-T, T])$ it can be written also as

$$g^*(t) = \sum_{j=-\infty}^{\infty} c_j e^{i2\pi f_j t}$$

where i is the imaginary unit with

$$c_j = \frac{1}{2T}\int_{-T}^{T} g^*(t)e^{-i2\pi f_j t}\mathrm{d}t$$

are known as the Fourier coefficients. Since $g(t) = g^*(t)$ over the interval $[-T, T]$ we also have (recall $f_j = j/2T$ hence $f_j - f_{j-1} = 1/2T$)

$$g(t) = g^*(t) = \sum_{j=-\infty}^{\infty}\left[\int_{-T}^{T} g^*(t)e^{-i2\pi f_j t}\mathrm{d}t\right] e^{i2\pi f_j t}(f_j - f_{j-1})$$

by letting $T \to \infty$ (that is allowing for no periodicities), we have the Fourier integral (or Inverse Fourier transform)

$$g(t) = \int_{-\infty}^{\infty} \tilde{g}(f)e^{i2\pi ft}\mathrm{d}f$$

and the Fourier transform

$$\tilde{g}(f) = \int_{-\infty}^{\infty} g(t)e^{-i2\pi ft}\mathrm{d}t.$$

We say that $g(t)$ and $\tilde{g}(f)$ are a Fourier pair. Notice that we have written $g(t)$ as a sum of infinite sinusoids defined over a continuum of frequencies.

Now notice that if $t \in \mathbb{Z}$, i.e. g is defined on a discrete time scale (as it will be for our time series), then the maximum sampling frequency, is $f = 1/2$, indeed in one interval of time Δt we can have maximum $1/2$ cycle (otherwise the process will not change). This is called Nyquist frequency and in real unit of measures is given by $f_N = 1/(2\Delta t)$ where Δt is the smallest interval of time over which we collect observations. In this case the Fourier integral of our signal will be defined as

$$g(t) = \int_{-1/2}^{1/2} \tilde{g}(f)e^{i2\pi ft}\mathrm{d}f, \qquad t \in \mathbb{Z},$$

while the Fourier transform becomes

$$\tilde{g}(f) = \sum_{t=-\infty}^{\infty} g(t)e^{-i2\pi ft}, \qquad t \in \mathbb{Z},$$

which is called Discrete Fourier transform. This is the formula we want now to generalise for the case in which $g(t)$ is the stochastic time series $\{X_t\}$.

9.2 Spectral representation

9.2.1 Real valued processes

Start considering a stationary real valued time series process $\{X_t\}$ which contains a periodic sinusoidal component with a known frequency f, then

$$x_t = R\cos(2\pi ft + \phi) + z_t$$

where $\{Z_t\}$ is a stationary process (random), R is called amplitude and ϕ is called phase. The angle $(2\pi ft + \phi)$ is measured in radians such that $\pi = 180^o$. Then $\theta = 2\pi f$ is called the angular frequency, while f is the number of cycles per unit of time. The period of a signal, called also wavelength, is given by $1/f = 2\pi/\theta$.

In practice variation in time can be caused by k different frequencies, so we write

$$x_t = \sum_{j=1}^{k} R_j \cos(2\pi f_j t + \phi_j) + z_t \tag{36}$$

For example sales or signals with seasonality components of period 4 months have always a component with frequency $1/4$ superimposed on the variation given by the maximum sampling frequency, that is $f = 1/2$.

Notice that (36) is not stationary since it has a time varying mean (all parameters are constant). So it is customary to model time series using random coefficients R_j such that they are uncorrelated and have zero mean. In this way $\{X_t\}$ is stationary (but in general not ergodic). Using the fact that $\cos(2\pi f_j t + \phi_j) = \cos(2\pi f_j)\cos(\phi_j) + \sin(2\pi f_j)\sin(\phi_j)$ we write (36) as

$$x_t = \sum_{j=1}^{k} (a_j \cos(2\pi f_j)) + b_j \sin(2\pi f_j)) + z_t$$

where $a_j = R_j \cos(\phi_j)$ and $b_j = R_j \sin(\phi_j)$ are new random coefficients with zero mean and still uncorrelated.

Actually as we have seen in the deterministic case the number of frequencies in $\{X_t\}$ might be infinite and they might be continuos (and that would be the case of a generic stationary process where z_t is capturing the non purely periodic component). By letting $k \to \infty$ it can be proved that any discrete time real valued stationary process can be written as

$$x_t = \int_0^{1/2} \cos(2\pi f t)\mathrm{d}u(f) + \int_0^{1/2} \sin(2\pi f t)\mathrm{d}v(f),$$

where $u(f)$ and $v(f)$ are uncorrelated random variables defined on $[0, 1/2]$ with values in $\mathbb{R}$ and with random increments (see below). This is the spectral representation of the process $\{X_t\}$ and it involves stochastic integrals. What is important here is to recognize that $\{X_t\}$ can be written as linear combination of infinite orthogonal sinusoids defined on a continuum of frequencies.

9.2.2 Complex valued processes

The above result is a particular case of the general case obtained for complex valued stochastic process. Recall that for two complex valued random variables X, Y we have

$$\mathrm{Cov}(X, Y) = \mathrm{E}[X\overline{Y}] - \mathrm{E}[X]\overline{\mathrm{E}[Y]}$$

where $\overline{Y}$ is the complex conjugate of Y.

Start with $\{Z(f)\}$ which is a stochastic process indexed by $f \in [-1/2, 1/2]$ and with values in $\mathbb{C}$. Consider infinitely small jumps of $\{Z(f)\}$, defined as

$$\mathrm{d}Z(f) = \begin{cases} Z(f + \mathrm{d}f) - Z(f) & f < 1/2 \\ 0 & f = 1/2. \end{cases}$$

The properties of $\{Z(f)\}$ are the following

1. $\mathrm{E}[\mathrm{d}Z(f)] = 0$ for any $|f| \leq 1/2$;

2. for any two frequencies $f, f' \in [-1/2, 1/2]$ such that $f \neq f'$ $\text{Cov}(\mathrm{d}Z(f), \mathrm{d}Z(f')) = \mathrm{E}[\mathrm{d}Z(f)\overline{\mathrm{d}Z(f')}] = 0$;

3. define $\mathrm{d}S^I(f) = \mathrm{E}[|\mathrm{d}Z(f)|^2] = \mathrm{E}[\mathrm{d}Z(f)\overline{\mathrm{d}Z(f)}] > 0$, this is called integrated spectrum.

Then if the intervals $[f, f + \mathrm{d}f]$ and $[f', f' + \mathrm{d}f']$ are non-overlapping subintervals of $[-1/2, 1/2]$ by property (2) above the process $Z(f)$ is said to have orthogonal increments. Notice that $\mathrm{d}Z(f)$ is a stochastic process that makes sense only if integrated.

Let $\{X_t\}$ be a stationary zero mean (possibly complex-valued) stochastic process with index $t \in \mathbb{Z}$. The spectral representation theorem states that there exists a process $\{Z(f)\}$ with orthogonal increments defined on $[-1/2, 1/2]$ such that

$$x_t = \int_{-1/2}^{1/2} e^{i2\pi ft} \mathrm{d}Z(f), \qquad t \in \mathbb{Z}. \tag{37}$$

Using the polar form of a complex number $\mathrm{d}Z(f) = |\mathrm{d}Z(f)|e^{i \arg(\mathrm{d}Z(f))}$, this means that we can represent any discrete complex-valued stationary process as an infinite sum of complex exponentials at frequencies f with associated random amplitudes $|\mathrm{d}Z(f)|$ and random phases $\arg(\mathrm{d}Z(f))$:

$$\begin{aligned} x_t &= \int_{-1/2}^{1/2} e^{i2\pi ft} |\mathrm{d}Z(f)| e^{i \arg(\mathrm{d}Z(f))} \\ &= \int_{-1/2}^{1/2} \cos(2\pi ft + \arg(\mathrm{d}Z(f)))|\mathrm{d}Z(f)| + i \int_{-1/2}^{1/2} \sin(2\pi ft + \arg(\mathrm{d}Z(f)))|\mathrm{d}Z(f)|. \end{aligned}$$

9.3 Spectral density

Now in practical situation instead of working with $u(f)$ and $v(f)$ or $Z(f)$ we introduce a new function $S^I(f)$, called the power spectral distribution (or integrated spectrum). Then, we can compute the acvs (recall that in general $\{X_t\}$ is complex valued)

$$\begin{aligned} \gamma_h = \mathrm{E}[X_t \overline{X_{t-h}}] &= \mathrm{E}\left[\left(\int_{-1/2}^{1/2} e^{i2\pi ft} \mathrm{d}Z(f) \right) \left(\int_{-1/2}^{1/2} e^{-i2\pi f'(t-h)} \overline{\mathrm{d}Z(f')} \right) \right] \\ &= \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} e^{i2\pi(f-f')t} e^{i2\pi f'h} \mathrm{E}[\mathrm{d}Z(f)\overline{\mathrm{d}Z(f')}] \\ &= \int_{-1/2}^{1/2} e^{i2\pi fh} \mathrm{E}[\mathrm{d}Z(f)\overline{\mathrm{d}Z(f)}] \\ &= \int_{-1/2}^{1/2} e^{i2\pi fh} \mathrm{d}S^I(f), \qquad h \in \mathbb{Z}. \end{aligned}$$

Let us assume that $S^I(f)$ is differentiable everywhere then there exists a function $S(f)$ such that

$$\mathrm{E}[|\mathrm{d}Z(f)|^2] = \mathrm{d}S^I(f) = S(f)\mathrm{d}f$$

notice that this implies that $S(f) \geq 0$. Equivalently

$$S(f) = \frac{\mathrm{d}S^I(f)}{\mathrm{d}f},$$

which is called the spectral density function or simply spectrum. Hence the acvs are given by

$$\gamma_h = \int_{-1/2}^{1/2} e^{i2\pi fh} S(f) \mathrm{d}f, \qquad h \in \mathbb{Z},$$

and since γ_h is deterministic (and discrete) we have that its Discrete Fourier transform is exactly the spectral density

$$S(f) = \sum_{h=-\infty}^{\infty} \gamma_h e^{-i2\pi fh}, \quad f \in [-1/2, 1/2].$$

The acvs and the spectral density are two equivalent ways of describing a stationary process. Notice that for this definition to make sense we must have $\sum_{h=-\infty}^{\infty} |\gamma_h| < \infty$.

If the spectral density exists, then $S(f)\mathrm{d}f$ represents the contribution to the total variance of $\{X_t\}$ given by the components with frequency in $[f, f + \mathrm{d}f]$. Indeed,

$$\mathrm{Var}(X_t) = \gamma_0 = \int_{-1/2}^{1/2} S(f) \mathrm{d}f,$$

which is also called the power.

The following properties of the integrated spectrum can be derived using the spectral density

1. $S^I(f) = \int_{-1/2}^{f} S(f')\mathrm{d}f'$;
2. $0 \leq S^I(f) \leq \gamma_0$ since $S(f) \geq 0$;
3. $S^I(-1/2) = 0$, $S^I(1/2) = \gamma_0$;
4. if $f < f'$, then $S^I(f) \leq S^I(f')$.

Given these properties it is clear that $S^I(f)$ represents the contribution to the total variance of $\{X_t\}$ given by the components with frequency smaller than f. Moreover, $S^I(f)$ has all the properties of a cdf (up to a scale of γ_0), thus it is also called spectral distribution and for this reason $S(f)$ is called spectral density (see the relation between cdf and pdf). Finally

$$\begin{aligned}
S(-f) &= \sum_{h=-\infty}^{\infty} \gamma_h e^{-i2\pi(-f)h} \\
&= \sum_{h=-\infty}^{\infty} \gamma_h (\cos(-2\pi fh) - i\sin(-2\pi fh)), \quad \text{set } k = -h \\
&= \sum_{k=-\infty}^{\infty} \gamma_{-k} (\cos(2\pi fk) - i\sin(2\pi fk)) \\
&= \sum_{k=-\infty}^{\infty} \gamma_k (\cos(2\pi fk) - i\sin(2\pi fk)) \\
&= \sum_{k=-\infty}^{\infty} \gamma_k e^{-i2\pi fk} = S(f)
\end{aligned}$$

Therefore, the spectral density is symmetric about $f = 0$.

Any $S^I(f)$ is usually decomposed as

$$S^I(f) = S_1^I(f) + S_2^I(f)$$

where $S_1^I(f)$ is a non-decreasing continuos function, and $S_2^I(f)$ is a non-decreasing discrete step function. This decomposition corresponds to the Wold decomposition of $\{X_t\}$ where $S_1(f)$ is related to the purely non-deterministic components, while $S_2^I(f)$ is related to the deterministic components. In particular,

1. $S_1^I(f)$ is absolutely continuous, i.e. its derivative exists for almost all f and is equal to the spectral density. For a purely non-deterministic process we then have $S^I(f) = S_1^I(f)$ and the density $S(f)$ is absolutely integrable, thus, we can prove that $\gamma_h \to 0$ as $|h| \to \infty$ (this is an application of the Riemann-Lebesgue lemma). This result is a mixing condition for stationary process with purely continuos spectrum.

2. $S_2^I(f)$ is a step function, as an example consider the case of one component of $\{X_t\}$ with frequency $f_0 > 0$

   $$x_t = R\cos(2\pi f_0 t)$$

   where R has zero mean and variance σ^2.[48] Then,

   $$\begin{aligned}\gamma_h &= \sigma^2 \cos(2\pi f_0 h) \\ &= \sigma^2 \frac{\cos(2\pi f_0 h) + i\sin(2\pi f_0 h) + \cos(-2\pi f_0 h) + i\sin(-2\pi f_0 h)}{2} \\ &= \sum_{j=-1,1} \frac{\sigma_j^2}{2} e^{i2\pi f_j h}, \qquad h \in \mathbb{Z}.\end{aligned}$$

   where $\sigma_j^2 = \sigma_{-j}^2 = \sigma^2$ and $f_1 = -f_{-1} = f_0$. Then, since

   $$\gamma_h = \int_{-1/2}^{1/2} e^{i2\pi f h} \mathrm{d}S^I(f) = \sum_{j=-1,1} \frac{\sigma_j^2}{2} e^{i2\pi f_j h}$$

   we must have

   $$S^I(f) = \begin{cases} 0 & f \in [-1/2, -f_0) \\ \frac{\sigma_{-1}^2}{2} & f \in [-f_0, f_0) \\ \frac{\sigma_{-1}^2}{2} + \frac{\sigma_1^2}{2} & f \in [f_0, 1/2] \end{cases}$$

   $S^I(f)$ has two jumps in $-f_0$ and f_0 which summed give the expected squared amplitude.

Hereafter we consider only the purely non-deterministic case $S_2^I(f) = 0$.

As an example. Consider a speech signal of length $T = 8000$ as in Chapter 1 and shown in Figure 53 corresponding to the syllable "ma" and sampled at a frequency of Hertz (Hz). Therefore, since $1\text{Hz} = 1\text{sec}^{-1}$, the signal has a duration of 1 second. We see that is composed of many wavelets each corresponding to a different periodicity. If we plot the acs of the signal we have a periodic behavior with the main period of roughly $h = 25$, that is corresponding to a frequency $f = 1/25 = 0.04$. An estimator of the spectral density (the scaled periodogram introduced below) has indeed a peak at $f \simeq 0.04$.

We can also compute the spectral density of known processes.

[48] Recall that a deterministic component is such that $P_s d_t = d_t$ for any s, t and this is the case for terms like $R\cos(2\pi f_0 t)$, even if R is random.

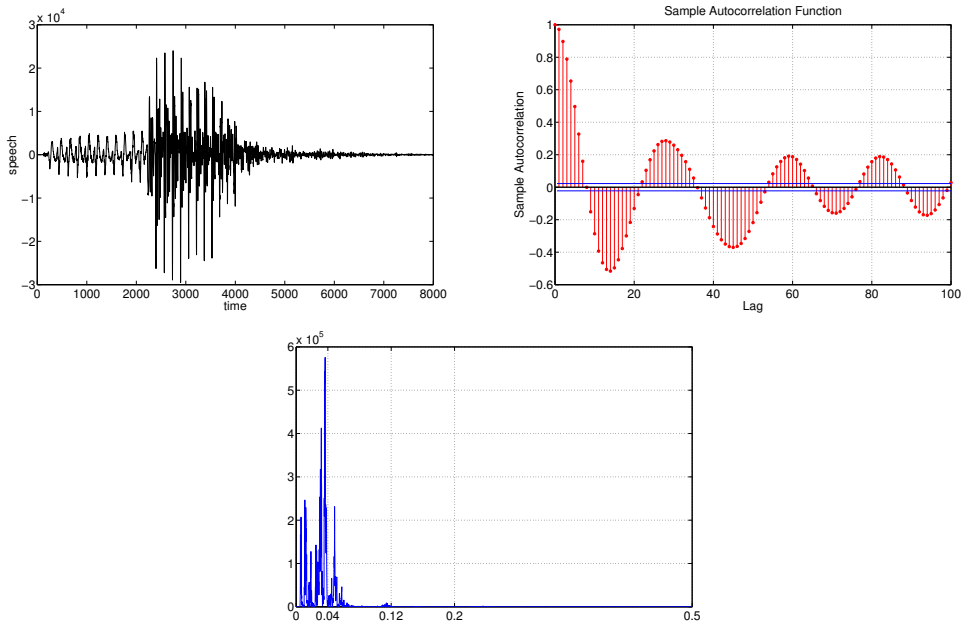


Figure 53: Top left: speech signal; top right: autocorrelation: bottom: scaled periodogram.

1. White noise $\{u_t\}$. Since $\gamma_h^u = 0$ for $|h| \neq 0$ we immediately have

$$S_u(f) = \sum_{h=-\infty}^{\infty} \gamma_h^u e^{-i2\pi fh} = \gamma_0^u,$$

a white noise has a constant spectrum and for this reason is called white (the white light is the sum of all colours, i.e. of all frequencies of light). The spectrum of a white noise is therefore not informative.

2. MA(1) process

$$x_t = u_t + \theta u_{t-1}$$

where u_t is a zero mean white noise with variance σ_u^2. Then, we have $\gamma_1^x = \gamma_{-1}^x = \theta\sigma_u^2$ and $\gamma_0^x = (1+\theta^2)\sigma_u^2$, and $\gamma_h = 0$ otherwise, then

$$\begin{aligned} S_x(f) &= \sum_{h=-\infty}^{\infty} \gamma_h^x e^{-i2\pi fh} \\ &= (1+\theta^2)\sigma_u^2 + \theta\sigma_u^2(e^{-i2\pi f} + e^{i2\pi f}) \\ &= (1+\theta^2+2\theta\cos(2\pi f))\sigma_u^2, \quad f \in [-1/2, 1/2] \end{aligned}$$

Therefore, if $\theta < 0$ the spectral density is minimum at $f = 0$ and maximum at $f = 1/2$, viceversa if $\theta > 0$ the spectral density is maximum at $f = 0$ and minimum at $f = 1/2$ (see Figure 54). Notice that the true spectral density and the one estimated via the periodogram differ substantially, we need a better estimation method (see next).

9.4 Linear filters

We need a general rule to compute spectra of ARMA processes. In order to this we use the notion of linear digital filter $\mathcal{L}$. A filter which transforms and input sequence $\{x_t\}$ into an output sequence $\{y_t\}$ is called linear time-invariant (LTI) digital filter of it satisfies the properties

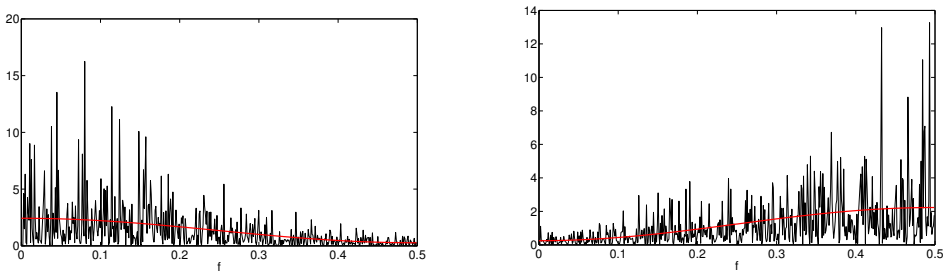


Figure 54: Left: MA(1) with $\theta = 0.5$; right: MA(1) with $\theta = -0.5$. Red: true spectral density; black: periodogram.

1. Scale preservation: $\mathcal{L}(\{\alpha x_t\}) = \alpha\mathcal{L}(\{x_t\})$;
2. Superposition: $\mathcal{L}(\{x_{1t} + x_{2t}\}) = \mathcal{L}(\{x_{1t}\}) + \mathcal{L}(\{x_{2t}\})$;
3. Time invariance: if $\mathcal{L}(\{x_t\}) = \{y_t\}$ then $\mathcal{L}(\{x_{t+h}\}) = \{y_{t+h}\}$ for any $h \in \mathbb{Z}$ and $\{x_{t+h}\}$ is a sequence with t-th element x_{t+h}.

Suppose now that the input sequence is $\{\xi_{f,t}\} = \{e^{i2\pi ft}\}$ and let $\{y_{f,t}\} = \mathcal{L}(\{\xi_{f,t}\})$ for $f \in [-1/2, 1/2]$ (a frequency), then by properties (1) and (3) above

$$\{y_{f,t+h}\} = \mathcal{L}(\{\xi_{f,t+h}\}) = \mathcal{L}(\{e^{i2\pi fh}\xi_{f,t}\}) = e^{i2\pi fh}\{y_{f,t}\}, \qquad h \in \mathbb{Z}.$$

In particular, for $t = 0$ we have $y_{f,h} = e^{i2\pi fh}y_{f,0}$, the first element of the output sequence and if we set $h = t$, we have

$$y_{f,t} = e^{i2\pi ft}y_{f,0}, \qquad t \in \mathbb{Z}$$

which is a generic element of the output sequence. So when $\{\xi_{f,t}\} = \{e^{i2\pi ft}\}$ is the input of an LTI digital filter, the output is the same function $\{\xi_{f,t}\}$ but multiplied by a constant $y_{f,0}$ which is independent of time but depends on the frequency f.

The function $G(f) = y_{f,0}$ is called transfer function or frequency response of $\mathcal{L}$. Since this is in general a complex valued function we can write it using its polar form as

$$G(f) = |G(f)|e^{i\theta(f)}, \qquad \theta(f) = \arg(G(f)).$$

The absolute value of the transfer function, $|G(f)|$ is also called gain. Moreover, $G(f)$ is a deterministic function and it is the Discrete Fourier transform of a discrete deterministic signal $\{g_u\}$ (see Section 9.1) such that

$$G(f) = \sum_{u=-\infty}^{\infty} g_u e^{-i2\pi fu}, \qquad g_u = \int_{-1/2}^{1/2} G(f)e^{i2\pi fu}\mathrm{d}f, \qquad u \in \mathbb{Z},\ f \in [-1/2, 1/2], \quad (38)$$

So $G(f)$ and g_u are a Fourier pair. Then we have

$$\mathcal{L}(\{e^{i2\pi ft}\}) = \{e^{i2\pi ft}\}G(f) = \sum_{u=-\infty}^{\infty} g_u\{e^{i2\pi f(t-u)}\}. \quad (39)$$

The result in (39) can be generalised to any stochastic sequence $\{X_t\}$ and any LTI digital filter thus we can write

$$\mathcal{L}(\{X_t\}) = \sum_{u=-\infty}^{\infty} g_u\{X_{t-u}\}.$$

In this case the deterministic sequence $\{g_u\}$ is called impulse response sequence and its elements are the coefficients of a polynomial in the lag operator

$$\underline{G}(L) = \sum_{u=-\infty}^{\infty} g_u L^u, \tag{40}$$

for this reason we sometime call these polynomials filters (see Chapter 3.1).

Now, for a generic LTI digital filter define the elements of the output sequence as

$$Y_t = \sum_{u=-\infty}^{\infty} g_u X_{t-u} = \underline{G}(L) X_t \tag{41}$$

and recall the spectral representation theorem

$$X_t = \int_{-1/2}^{1/2} e^{i2\pi ft} \mathrm{d}Z_X(f), \qquad Y_t = \int_{-1/2}^{1/2} e^{i2\pi ft} \mathrm{d}Z_Y(f),$$

then from (39) and (41) we have

$$\begin{aligned} Y_t &= \int_{-1/2}^{1/2} e^{i2\pi ft} \mathrm{d}Z_Y(f) = \sum_{u=-\infty}^{\infty} g_u \int_{-1/2}^{1/2} e^{i2\pi f(t-u)} \mathrm{d}Z_X(f) \\ &= \int_{-1/2}^{1/2} e^{i2\pi ft} G(f) \mathrm{d}Z_X(f) \end{aligned} \tag{42}$$

so that

$$\mathrm{d}Z_Y(f) = G(f) \mathrm{d}Z_X(f)$$

and the integrated spectrum of $\{Y_t\}$ is

$$\mathrm{E}[|\mathrm{d}Z_Y(f)|^2] = |G(f)|^2 \mathrm{E}[|\mathrm{d}Z_X(f)|^2] \quad \Longleftrightarrow \quad S_Y^I(f) = |G(f)|^2 S_X^I(f),$$

and if the spectral density of $\{X_t\}$ and $\{Y_t\}$ exists then we also have

$$S_Y(f) = |G(f)|^2 S_X(f).$$

This relation gives us a rule to compute spectral densities of ARMA processes: once we have the impulse response sequence, we can compute the gain (absolute value of the Discrete Fourier transform of the impulse response) and we immediately have the spectral density of the filtered process. Notice that $G(f)$ is computed straightforwardly using (38) and comparing it with (40). Indeed once we have the polynomial $\underline{G}(L)$ we just have to replace L with $e^{-i2\pi f}$, i.e.

$$G(f) = \underline{G}(e^{-i2\pi f}). \tag{43}$$

Some examples follow.

1. MA(q)

   $$x_t = u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q},$$

   where u_t is zero mean white noise with variance σ_u^2. Notice that the MA is obtained as a LTI digital filter acting on $\{u_t\}$:

   $$\{X_t\} = \mathcal{L}_{MA}(\{u_t\}) = \sum_{k=0}^{q} \theta_k \{u_{t-k}\}$$

Then $G(f)$ is given by applying the filter $\mathcal{L}_{MA}$ to $\{e^{i2\pi ft}\}$ (see (38))

$$\mathcal{L}_{MA}(\{e^{i2\pi ft}\}) = \sum_{k=0}^{q} \theta_k \{e^{i2\pi f(t-k)}\} = \{e^{i2\pi ft}\} \underbrace{\sum_{k=0}^{q} \theta_k e^{-i2\pi fk}}_{G(f)}$$

Or more easily define

$$\underline{\Theta}(L) = \sum_{k=0}^{q} \theta_k L^k$$

such that $x_t = \underline{\Theta}(L)u_t$, then by using (43) we have

$$G(f) = \underline{\Theta}(e^{-i2\pi f}) = \sum_{k=0}^{q} \theta_k e^{-i2\pi fk}.$$

Then, since u_t is a white noise $S_u(f) = \sigma_u^2$ and the spectral density of $\{X_t\}$ is given by

$$S_X(f) = |G(f)|^2 S_u(f) = \sigma_u^2 \left| \sum_{k=0}^{q} \theta_k e^{-i2\pi fk} \right|^2.$$

Define $z = e^{-i2\pi f}$, then $\bar{z} = z^{-1}$ and[49]

$$|G(f)|^2 = \underline{\Theta}(z)\underline{\Theta}(z^{-1}) = \underline{\Theta}(e^{-i2\pi f})\underline{\Theta}(e^{i2\pi f}).$$

So for example when $q = 1$ we have the same result as in the previous section:

$$\begin{aligned}
|G(f)|^2 &= (1 + \theta_1 z)(1 + \theta_1 z^{-1}) \\
&= 1 + \theta_1^2 + \theta_1(z + z^{-1}) \\
&= 1 + \theta_1^2 + \theta_1(e^{-i2\pi f} + e^{i2\pi f}) \\
&= 1 + \theta_1^2 + \theta_1 \left(\cos(2\pi f) + i\sin(2\pi f) + \cos(2\pi f) - i\sin(2\pi f)\right) \\
&= 1 + \theta_1^2 + 2\theta_1 \cos(2\pi f).
\end{aligned}$$

Moreover, notice that if α is a root of $\underline{\Theta}(z)$ then α^{-1} is a root of $\underline{\Theta}(z^{-1})$. Thus for an invertible (in the past) MA we must have $|\alpha| > 1$, but there exists also a non invertible MA with polynomial $\underline{\Theta}(z^{-1})$ such that it has the same gain and the same spectral density. Take as an example the MAs

$$x_t = (1 + \frac{1}{2}L)u_t = \underline{\Theta}_1(L)u_t, \qquad x_t = (1 + 2L)\frac{1}{2}u_t = \underline{\Theta}_2(L)v_t,$$

where $\text{Var}(u_t) = \sigma_u^2$ and therefore $\text{Var}(v_t) = \frac{\sigma_u^2}{4}$. The spectral densities are identical, indeed

$$\begin{aligned}
S_{X_1}(f) = \sigma_u^2 |G_1(f)|^2 &= \sigma_u^2 \underline{\Theta}_1(z)\underline{\Theta}_1(z^{-1}) \\
&= \sigma_u^2 (1 + \frac{1}{2}z)(1 + \frac{1}{2}z^{-1}) = \sigma_u^2 (1 + \frac{1}{4} + \frac{1}{2}(z + z^{-1})) \\
S_{X_2}(f) = \frac{\sigma_u^2}{4} |G_2(f)|^2 &= \frac{\sigma_u^2}{4} \underline{\Theta}_2(z)\underline{\Theta}_2(z^{-1}) \\
&= \sigma_u^2 (\frac{1}{2} + z)(\frac{1}{2} + z^{-1}) = \sigma_u^2 (\frac{1}{4} + 1 + \frac{1}{2}(z + z^{-1}))
\end{aligned}$$

We cannot distinguish from the spectrum if an MA is invertible or not, the same result we had for acvs.

[49] Recall that for a complex number $z \in \mathbb{C}$ we have $|z|^2 = z\bar{z}$.

2. AR(p)

$$x_t = u_t + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p},$$

where u_t is zero mean white noise with variance σ_u^2. Notice that the AR is obtained as a LTI digital filter acting on $\{X_t\}$:

$$\{u_t\} = \mathcal{L}_{AR}(\{X_t\}) = \{X_t\} - \sum_{k=1}^{p} \phi_k \{X_{t-k}\}$$

Then $G(f)$ is given by applying the filter $\mathcal{L}_{AR}$ to $\{e^{i2\pi ft}\}$ (see (38))

$$\mathcal{L}_{AR}(\{e^{i2\pi ft}\}) = \{e^{i2\pi ft}\} - \sum_{k=1}^{p} \phi_k \{e^{i2\pi f(t-k)}\} = \{e^{i2\pi ft}\} \underbrace{\left(1 - \sum_{k=1}^{p} \phi_k e^{-i2\pi fk}\right)}_{G(f)}$$

Equivalently, define the polynomial

$$\underline{\Phi}(L) = 1 - \sum_{k=1}^{p} \phi_k L^k$$

such that $\underline{\Phi}(L)x_t = u_t$. Then, the associated $G(f)$ is (using (43))

$$G(f) = \underline{\Phi}(e^{-i2\pi f}) = 1 - \sum_{k=1}^{p} \phi_k e^{-i2\pi fk}.$$

and the spectral density is

$$S_X(f) = \frac{\sigma_u^2}{|G(f)|^2} = \frac{\sigma_u^2}{\left|1 - \sum_{k=1}^{p} \phi_k e^{-i2\pi fk}\right|^2}.$$

Consider the case $p = 1$ and set $z = e^{-i2\pi f}$, then

$$\begin{aligned} |G(f)|^2 &= (1 - \phi_1 z)(1 - \phi_1 z^{-1}) \\ &= 1 + \phi_1^2 - \phi_1(z + z^{-1}) \\ &= 1 + \phi_1^2 - \phi_1(e^{-i2\pi f} + e^{i2\pi f}) \\ &= 1 + \phi_1^2 - 2\phi_1 \cos(2\pi f). \end{aligned}$$

Notice that since $|z| = 1$ the inverse of the previous expression is always well defined unless $|\phi| = 1$, i.e in case of uint root, when the process is non-stationary (causality being a time related concept is not required here) see Figure 55. The spectral density is

$$S_X(f) = \frac{\sigma_u^2}{1 + \phi_1^2 - 2\phi_1 \cos(2\pi f)}.$$

Consider the an AR(2) with two complex roots and for a causal process which in polar form can be written as

$$z_{1,2} = \frac{1}{r} e^{\pm i2\pi f'}, \qquad 0 < r < 1.$$

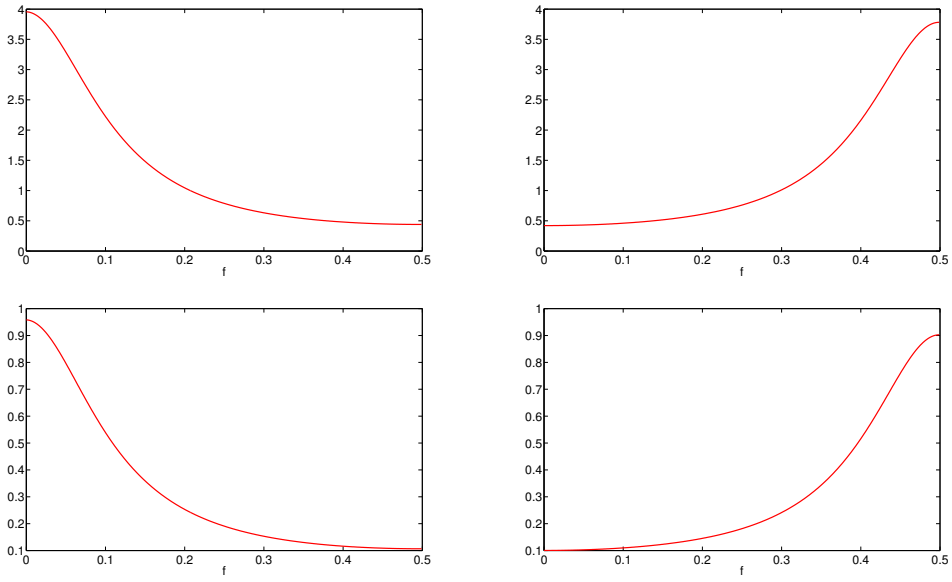


Figure 55: Top left: AR(1) with $\phi = 0.5$; top right: AR(1) with $\phi = -0.5$; bottom left: AR(1) with $\phi = 2$; bottom right: AR(1) with $\phi = -2$.

The characteristic polynomial of the AR(2) is written in terms of these roots as

$$\begin{aligned} 1 - \phi_1 z - \phi_2 z^2 &= \left(\frac{z}{z_1} - 1\right)\left(\frac{z}{z_2} - 1\right) = (rz - e^{-i2\pi f'})(rz - e^{i2\pi f'}) \\ &= r^2 z^2 - zr(e^{-i2\pi f'} + e^{i2\pi f'}) + 1 \\ &= r^2 z^2 - 2zr\cos(2\pi f') + 1. \end{aligned} \tag{44}$$

We can then write the AR(2) model as (replace z with L)

$$x_t = 2r\cos(2\pi f')x_{t-1} - r^2 x_{t-2} + u_t$$

In this case $\{X_t\}$ has a periodic component of frequency f' (this can happen only for complex roots which come in pairs, therefore not for an AR(1) which has only one real root). The spectrum is (replace z with $e^{-i2\pi f}$)

$$S_X(f) = \frac{\sigma_u^2}{|1 - \phi_1 e^{-i2\pi f} - \phi_2 e^{-i4\pi f}|^2} = \frac{\sigma_u^2}{|re^{-i2\pi f} - e^{-i2\pi f'}|^2 \, |re^{-i2\pi f} - e^{i2\pi f'}|^2}$$

Then

$$|re^{-i2\pi f} - e^{-i2\pi f'}|^2 = (r - e^{i2\pi(f'-f)})(r - e^{-i2\pi(f'-f)}) = r^2 - 2r\cos(2\pi(f' - f)) + 1$$

and this gives

$$S_X(f) = \frac{\sigma_u^2}{(r^2 - 2r\cos(2\pi(f' - f)) + 1)(r^2 - 2r\cos(2\pi(f' + f)) + 1)}$$

The maximum of the spectrum is when the denominator is minimum, which for r close to 1 occurs when $f = \pm f'$, the spectrum becoming larger as $r \to 1$ (from below) an example is in Figure 56. Generally speaking complex roots, which correspond to an oscillatory (or cyclical or periodic) behaviour, induce a peak in the spectrum indicating the frequency f' of

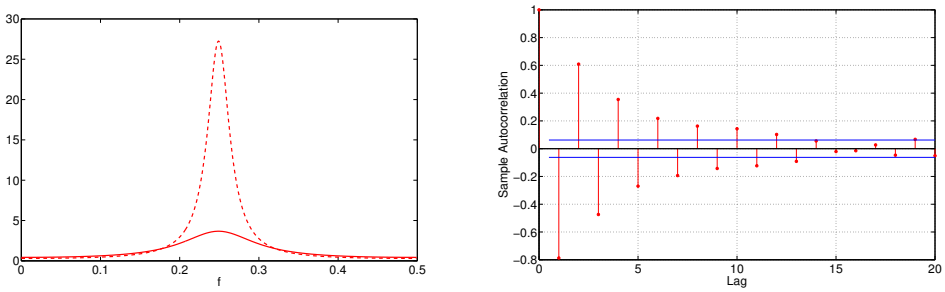


Figure 56: Left: AR(2) spectra, solid: with $r = 0.7$ and $f' = 1/4$; dashed: with $r = 0.9$ and $f' = 1/4$; Right: autocorrelation for the case $r = 0.9$ and $f' = 1/4$.

the cycle. The larger is r (the more persistent is the process, as is closer to non-stationarity) the more dominant becomes the cycle. This is a pseudo-cyclical behaviour since a purely deterministic behaviour would have a sharp spike, i.e. a line in the spectrum and acvs would never decline to zero.

3. ARMA(p, q)

$$\underline{\Phi}(L)x_t = \underline{\Theta}(L)u_t,$$

where u_t is zero mean white noise with variance σ_u^2. Define $y_t = \underline{\Theta}(L)u_t$, then

$$|G_\phi(f)|^2 S_X(f) = S_Y(f)$$

and

$$S_Y(f) = |G_\theta(f)|^2 \sigma_u^2$$

which gives

$$S_X(f) = \sigma_u^2 \frac{|G_\theta(f)|^2}{|G_\phi(f)|^2} = \sigma_u^2 \frac{\left|\sum_{k=0}^q \theta_k e^{-i2\pi fk}\right|^2}{\left|1 - \sum_{k=1}^p \phi_k e^{-i2\pi fk}\right|^2}$$

4. Differencing. Let $\{X_t\}$ be stationary with spectral density $S_X(f)$, and let $Y_t = X_t - X_{t-1}$ which defines a LTI digital filter $\mathcal{L}(\{X_t\}) = \{Y_t\}$, such that

$$\mathcal{L}(\{e^{i2\pi ft}\}) = \{e^{i2\pi ft} - e^{i2\pi f(t-1)}\} = \{e^{i2\pi ft}\}(1 - e^{-i2\pi f}) = \{e^{i2\pi ft}\}G(f)$$

which is obviously an AR(1) filter with parameter $\phi = 1$. The associated gain is (recall that $|e^{i\pi fk}| = 1$ for any $k \in \mathbb{Z}$)

$$|G(f)|^2 = |1 - e^{-i2\pi f}|^2 = |e^{-i\pi f}(e^{i\pi f} - e^{-i\pi f})|^2 = |e^{-i\pi f} 2i \sin(\pi f)|^2 = 4\sin^2(\pi f).$$

The spectral density of $\{Y_t\}$ is then

$$S_Y(f) = |G(f)|^2 S_X(f) = 4\sin^2(\pi f) S_X(f).$$

As an example consider the case in which $\{X_t\}$ is a white noise with mean zero and variance σ^2, then

$$S_Y(f) = 4\sin^2(\pi f)\sigma^2.$$

which is the spectral density of an MA(1) with parameter $\theta = -1$. Indeed, from the general MA(1) formula we have[50]

$$S_Y(f) = 1 + \theta^2 + 2\theta\cos(2\pi f) = 2(1 - \cos(2\pi f)) = 2\sin^2(2\pi f) = 4\sin^2(\pi f).$$

[50] Use the relation $\cos(2a) = \cos^2(a) - \sin^2(a) = 1 - 2\sin^2(a)$ since $\sin^2(a) + \cos^2(a) = 1$.

9.5 Estimation

9.5.1 Finding periodicity in the data

10 Multivariate stochastic processes

We extend all basic definitions and results of stationarity and of ARMA processes to vector processes.

10.1 Vector stochastic process

Given a probability space $(\Omega, \mathcal{F}, P)$ an n-dimensional random vector $\mathbf{X}$ is a function

$$\mathbf{X} : \Omega \to \mathcal{S}.$$

where $\mathcal{S}$ is the state space, i.e. the space of the values taken by the random variables. For us it will always be $\mathcal{S} = \mathbb{R}^n$. The choice of state space is specified by the physical situation being described.

An n-dimensional stochastic process associates a random vector to each $t \in \mathcal{T}$

$$\mathbf{X} : t \mapsto \mathbf{X}_t.$$

where $\mathcal{T}$ is the set of values of the time index, and for us it will always be $\mathcal{T} = \mathbb{Z}$. We use the notation $\{\mathbf{X}_t, t \in \mathbb{Z}\}$ for the n-dimensional stochastic process. This process has n components given by the processes $\{X_{it}, t \in \mathbb{Z}\}$, for $i = 1, \ldots, n$.

In principle for any integer s and $t_1 \leq t_2 \leq \ldots \leq t_s$, we need to specify the joint probability distribution function of $\mathbf{X}_{t_1} \ldots \mathbf{X}_{t_s}$ which takes values on a space of dimension $t_s \times n$. As for the case $n = 1$ we will just consider a theory based only on second moments which in this case are a more rich set of moments. In particular, we have the n-dimensional mean vector

$$\mathrm{E}[\mathbf{X}_t] = \begin{pmatrix} \mathrm{E}[X_{1t}] \\ \vdots \\ \mathrm{E}[X_{nt}] \end{pmatrix}, \qquad t \in \mathbb{Z},$$

and the $n \times n$ matrices of the second moments

$$\mathrm{E}[\mathbf{X}_t \mathbf{X}'_{t+h}] = \begin{pmatrix} \mathrm{E}[X_{1t}X_{1t+h}] & \ldots & \mathrm{E}[X_{1t}X_{nt+h}] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[X_{nt}X_{1t+h}] & \ldots & \mathrm{E}[X_{nt}X_{nt+h}] \end{pmatrix}, \qquad t, h \in \mathbb{Z},$$

so out of the diagonal we have the dependencies among the different components of $\{\mathbf{X}_t\}$. The matrix

$$\mathbf{\Gamma}_h = \mathrm{Cov}(\mathbf{X}_t, \mathbf{X}_{t+h}) = \mathrm{E}[\mathbf{X}_t \mathbf{X}'_{t+h}] - \mathrm{E}[\mathbf{X}_t]\,\mathrm{E}[\mathbf{X}'_{t+h}]$$

is the lag h autocovariance matrix of $\{\mathbf{X}_t\}$.

10.2 Weak stationarity

The n-dimensional vector process $\{\mathbf{X}_t\}$ is weakly stationary if for any integer s and $t_1 \leq t_2 \leq \ldots \leq t_s$, and all integers k, all the joint moments of order 1 and 2 of $\{\mathbf{X}_{t_1} \ldots \mathbf{X}_{t_s}\}$ exist, are finite, and equal to the corresponding joint moments of $\{\mathbf{X}_{t_1+k} \ldots \mathbf{X}_{t_s+k}\}$. Thus,

1. $\mathrm{E}[\mathbf{X}_t] = \boldsymbol{\mu}$ does not depend on t;
2. $\mathrm{E}[\mathbf{X}_t \mathbf{X}'_{t+k}]$ depends only on $|k|$, therefore

$$\mathrm{E}[\mathbf{X}_t \mathbf{X}'_{t+k}] = \mathrm{E}[\mathbf{X}_{t-k} \mathbf{X}'_t] = \mathrm{E}[\mathbf{X}_t \mathbf{X}'_{t-k}]',$$

and then $\boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}'_{-k}$.

The stationarity of each of the components of $\{\mathbf{X}_t\}$ does not imply stationarity of the vector $\{\mathbf{X}_t\}$. On the other hand stationarity of the vector $\{\mathbf{X}_t\}$ requires that the components of the vector are also stationary and co-stationary.

Example of co-stationarity. Consider the two processes $\{X_t\}$ and $\{Y_t\}$, formed from the sum and difference of two successive values of a white noise process:

$$x_t = w_t + w_{t-1}, \quad y_t = w_t - w_{t-1}, \quad w_t \sim w.n.(0, 1).$$

We have, $\mathrm{E}[X_t] = \mathrm{E}[Y_t] = 0$, then

$$\gamma_0^x = \mathrm{E}[(w_t + w_{t-1})^2] = 2 = \gamma_0^y.$$

Moreover

$$\begin{aligned} \gamma_1^x &= \mathrm{Cov}(X_t, X_{t+1}) = \mathrm{E}[X_t X_{t+1}] \\ &= \mathrm{E}[(w_t + w_{t-1})(w_{t+1} + w_t)] = \mathrm{E}[w_t^2] = 1 = \gamma_{-1}^x. \end{aligned}$$

Similarly, we can show that $\gamma_1^y = -1 = \gamma_{-1}^y$. Also, we can define the cross-covariance at lag 1 as

$$\gamma_1^{xy} = \mathrm{Cov}(X_{t+1}, Y_t) = \mathrm{E}[(w_{t+1} + w_t)(w_t - w_{t-1})] = \mathrm{E}[w_t^2] = 1$$

and

$$\gamma_{-1}^{xy} = \mathrm{Cov}(X_{t-1}, Y_t) = \mathrm{E}[(w_{t-1} + w_{t-2})(w_t - w_{t-1})] = -\mathrm{E}[w_{t-1}^2] = -1$$

which illustrates the non-symmetric behavior about zero of the cross-autocovariance. Moreover

$$\gamma_0^{xy} = \mathrm{Cov}(X_t, Y_t) = \mathrm{E}[(w_t + w_{t-1})(w_t - w_{t-1})] = 0.$$

So, we finally obtain:

$$\rho_k^{xy} = \begin{cases} -1/2 & \text{if} \quad k = -1 \\ +1/2 & \text{if} \quad k = +1 \\ 0 & \text{otherwise.} \end{cases}$$

And for the 2-dimensional process $\mathbf{Z}_t = (X_t, Y_t)'$ we have

$$\boldsymbol{\Gamma}_0^Z = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \boldsymbol{\Gamma}_1^Z = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \boldsymbol{\Gamma}_{-1}^{Z'}$$

10.3 Vector white noise

The n-dimensional stationary process $\{\mathbf{X}_t\}$ is a white noise if $\mathrm{E}[\mathbf{X}_t] = \boldsymbol{\mu}$ which does not depend on time and $\boldsymbol{\Gamma}_k = \mathbf{0}_{n\times n}$ for all $k \neq 0$. We denote the process as $\mathbf{X}_t \sim w.n.(\boldsymbol{\mu}, \boldsymbol{\Gamma}_0)$.

A vector whose components are white noise is not necessarily a white noise. For example, let u_t be a scalar white noise with zero-mean and variance σ_u^2 as in previous chapters. Then consider the bidimensional process $\mathbf{X}_t = (u_t\ u_{t-1})'$. We have

$$\boldsymbol{\Gamma}_0 = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}, \qquad \boldsymbol{\Gamma}_1 = \begin{pmatrix} 0 & 0 \\ \sigma_u^2 & 0 \end{pmatrix},$$

which shows that $\{\mathbf{X}_t\}$ is not a white noise.

Consider for $n = 2$ the covariance matrix of a generic process $\{\mathbf{X}_t\}$

$$\boldsymbol{\Gamma}_0 = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

then the definition of vector white noise has no implications for this matrix since we are not requiring that $\sigma_{12} = 0$.

The definition of vector white noise does not imply that $\boldsymbol{\Gamma}_0$ has maximum rank either. This however will be assumed in this chapter, if $\{\mathbf{X}_t\}$ is a white noise then $\boldsymbol{\Gamma}_0$ is non-singular.

10.4 Vector moving average

Given an n-dimensional zero-mean vector white noise $\mathbf{u}_t = (u_{1t} \dots u_{nt})'$ with covariance matrix $\boldsymbol{\Gamma}_0^u$, then a moving average of $\mathbf{u}_t$ of order m is a process $\{\mathbf{X}_t\}$ with realisations

$$\mathbf{x}_t = \sum_{k=0}^{m} \mathbf{A}_k \mathbf{u}_{t-k} = \mathbf{A}_0\mathbf{u}_t + \mathbf{A}_1\mathbf{u}_{t-1} + \ldots + \mathbf{A}_m\mathbf{u}_{t-m}$$

where $\mathbf{A}_k$ are $n \times n$ matrices of coefficients.[51] Therefore, each component of $\{\mathbf{X}_t\}$ depends on all the components of $\mathbf{u}_t$.

We usually assume that $\mathbf{A}_0$ is non-singular and as a consequence the moving average can be rewritten as

$$\mathbf{x}_t = \mathbf{v}_t + \sum_{k=1}^{m} \mathbf{B}_k \mathbf{v}_{t-k} = \mathbf{v}_t + \mathbf{B}_1\mathbf{u}_{t-1} + \ldots + \mathbf{B}_m\mathbf{u}_{t-m}$$

where $\mathbf{v}_t = \mathbf{A}_0\mathbf{u}_t$ and $\mathbf{B}_k = \mathbf{A}_k\mathbf{A}_0^{-1}$. We can then prove that $\mathbf{v}_t$ is also a vector white noise since

$$\mathrm{E}[\mathbf{v}_t\mathbf{v}_{t-k}'] = \mathbf{A}_0\mathrm{E}[\mathbf{u}_t\mathbf{u}_{t-k}']\mathbf{A}_0'$$

which is zero unless $k = 0$ because $\mathbf{u}_t$ is a white noise.

Consider the lagged MA

$$\mathbf{x}_{t-1} = \mathbf{A}_0\mathbf{u}_{t-1} + \mathbf{A}_1\mathbf{u}_{t-2} + \ldots + \mathbf{A}_m\mathbf{u}_{t-m-1}$$

[51] Two-sided vector moving averages are also possible: $x_t = \sum_{k=-m}^{m} \mathbf{A}_k\mathbf{u}_{t-k}$.

Since by construction $\mathrm{E}[\mathbf{X}_t] = \mathbf{0}_n$, for example the lag 1 acvs of $\{\mathbf{X}_t\}$ is then computed as

$$\mathbf{\Gamma}_1 = \mathrm{E}[\mathbf{X}_t \mathbf{X}'_{t-1}] = \mathbf{A}_m \mathbf{\Gamma}_0^u \mathbf{A}'_{m-1} + \ldots + \mathbf{A}_2 \mathbf{\Gamma}_0^u \mathbf{A}'_1 + \mathbf{A}_1 \mathbf{\Gamma}_0^u \mathbf{A}'_0.$$

As in the scalar case vector moving averages are always stationary.

Consider now a more general case of infinite vector moving averages

$$\mathbf{x}_t = \sum_{k=0}^{\infty} \mathbf{A}_k \mathbf{u}_{t-k} \tag{45}$$

then each component of $\{\mathbf{X}_t\}$ has realisations

$$x_{jt} = \sum_{k=0}^{\infty} \sum_{h=1}^{n} a_{jh,k} u_{h,t-k}, \qquad j = 1, \ldots, n,$$

and in order for this process to have finite variance

$$\mathrm{Var}(x_{jt}) = \sum_{k=0}^{\infty} \sum_{h=1}^{n} \sum_{\ell=1}^{n} a_{jh,k} a_{j\ell,k} \mathrm{Cov}(u_{ht}, u_{\ell t})$$

we need for any $j = 1, \ldots, n$

$$\sum_{k=0}^{\infty} \left(\sum_{h=1}^{n} a_{jh,k} \right)^2 \leq \sum_{k=-\infty}^{\infty} \sum_{h=1}^{n} a_{jh,k}^2 < \infty.$$

which is equivalent to asking

$$\sum_{h=0}^{\infty} \|\mathbf{A}_k\|^2 < \infty.$$

10.5 Vector autoregression - VAR

10.5.1 VAR(1)

As in the scalar case the infinite vector moving average is equivalent to a vector autoregression of order 1, VAR(1) which has realisations

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t,$$

where $\mathbf{u}_t$ is an n-dimensional zero-mean vector white noise with covariance matrix $\mathbf{\Gamma}_0^u$. In the case $n = 2$ the two components of $\{\mathbf{X}_t\}$ have realisations

$$\begin{aligned} x_{1t} &= a_{11} x_{1t-1} + a_{12} x_{2t-1} + u_{1t} \\ x_{2t} &= a_{21} x_{1t-1} + a_{22} x_{2t-1} + u_{2t}, \end{aligned}$$

where a_{ij} is the generic i, j-th entry of $\mathbf{A}$. By iterating we have

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t \\ &= \mathbf{A}[\mathbf{A}\mathbf{x}_{t-2} + \mathbf{u}_{t-1}] + \mathbf{u}_t \\ &= \mathbf{A}^2 \mathbf{x}_{t-2} + \mathbf{A}\mathbf{u}_{t-1} + \mathbf{u}_t \\ &\vdots \\ &= \mathbf{u}_t + \mathbf{A}\mathbf{u}_{t-1} + \mathbf{A}^2 \mathbf{u}_{t-2} + \ldots \\ &= \sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{u}_{t-k} \end{aligned}$$

We now need to find the conditions for the above process to have a finite variance.

Given the square $n \times n$ matrix $\mathbf{A}$, a vector $\mathbf{v} \neq \mathbf{0}_n$ is called an eigenvector of $\mathbf{A}$ if $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ where λ is a real number, i.e. if $\mathbf{v}$ and the transformed vector $\mathbf{A}\mathbf{v}$ are parallel. The number λ, which is uniquely determined by $\mathbf{v}$ is the eigenvalue associated with $\mathbf{v}$. Alternatively, λ is an eigenvalue of $\mathbf{A}$ if

$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0. \tag{46}$$

Since the matrix $\mathbf{A} - \lambda\mathbf{I}_n$ is singular (it has zero determinant), there exists $\mathbf{v} \neq \mathbf{0}_n$ such that $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v} = \mathbf{0}_n$, i.e. such that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Equation (46) is an algebraic equation of degree n, hence the matrix $\mathbf{A}$ has n eigenvalues and eigenvectors (they can be complex numbers).

Thus, for $j = 1, \ldots, n$ we have

$$\mathbf{A}\mathbf{v}_j = \lambda_j\mathbf{v}_j,$$

which can be rewritten as

$$\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}, \qquad \mathbf{V} = (\mathbf{v}_1 \ldots \mathbf{v}_n),' \qquad \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \ldots 0 & \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix}$$

that is

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}.$$

This is known as diagonalization of a matrix. Every matrix $\mathbf{A}$ (we assume here that the eigenvalues are distinct) is equivalent to a diagonal matrix, with the eigenvalues of $\mathbf{A}$ on the diagonal. Notice that

$$\mathbf{A}^s = \mathbf{V}\mathbf{\Lambda}^s\mathbf{V}^{-1}, \qquad \mathbf{\Lambda}^s = \begin{pmatrix} \lambda_1^s & \ldots 0 & \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n^s \end{pmatrix}.$$

Now the infinite vector moving average can be written as

$$\mathbf{x}_t = \mathbf{u}_t + \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}\mathbf{u}_{t-1} + \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^{-1}\mathbf{u}_{t-2} + \ldots$$

or defining $\mathbf{y}_t = \mathbf{V}^{-1}\mathbf{x}_t$ and $\mathbf{v}_t = \mathbf{V}^{-1}\mathbf{u}_t$ we have

$$\mathbf{y}_t = \mathbf{v}_t + \mathbf{\Lambda}\mathbf{v}_{t-1} + \mathbf{\Lambda}^2\mathbf{v}_{t-2} + \ldots$$

and each component of $\{\mathbf{Y}_t\}$ is not dependent on the others indeed since $\mathbf{\Lambda}$ is diagonal we have

$$y_{it} = v_{it} + \lambda_i v_{it-1} + \lambda_i^2 v_{it-2} + \ldots = \sum_{k=0}^{\infty} \lambda_i^k v_{it-k} = \lambda_i y_{it-1} + v_{it},$$

which is an AR(1) process and it is stationary and causal as long as $|\lambda_i| < 1$. Moreover, $\{\mathbf{X}_t\}$ is stationary if and only if $\{\mathbf{Y}_t\}$ is stationary.[52] Then we must have that all eigenvalues of the matrix $\mathbf{A}$ must be inside the unit circle, i.e. such that $|\lambda_i| < 1$. Diagonalization of $\mathbf{A}$ transforms a vector problem into a collection of scalar problems.

Alternatively we can start from the equation

$$\mathbf{A}(L)\mathbf{x}_t = (\mathbf{I}_n - \mathbf{A}L)\mathbf{x}_t = \mathbf{u}_t.$$

[52] Indeed for any integer k we have $\mathbf{\Gamma}_k^x = \mathbf{V}\mathbf{\Gamma}_k^y\mathbf{V}'$ since $\mathbf{x}_t = \mathbf{V}\mathbf{y}_t$.

If we can invert the matrix $\mathbf{A}(L)$ then the process would be written as a Moving Average and therefore it would be stationary. To $\mathbf{A}(L)$ we can associate $\det \mathbf{A}(L)$ and the matrix $\mathbf{A}_{ad}(L)$ is the transposed of the cofactor matrix which in turn has in entry i, j the determinant of $\mathbf{A}(L)$ when removing row i and column j multiplied by $(-1)^{i+j}$ such that[53]

$$\mathbf{A}(L)\mathbf{A}_{ad}(L) = [\det \mathbf{A}(L)]\mathbf{I}_n.$$

Therefore, (48) becomes

$$[\det \mathbf{A}(L)]\mathbf{x}_t = \mathbf{A}_{ad}(L)\mathbf{u}_t. \tag{47}$$

Consider the example with $n = 2$ and $p = 1$

$$\begin{pmatrix} 1 - a_{11}L & -a_{12}L \\ -a_{21}L & 1 - a_{22}L \end{pmatrix} \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

which becomes

$$[\det \mathbf{A}(L)] \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} 1 - a_{22}L & a_{12}L \\ a_{21}L & 1 - a_{11}L \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix},$$

where $\det \mathbf{A}(L) = (1 - a_{11}L)(1 - a_{22}L) - a_{12}a_{21}L^2$. That is we have two equations

$$\begin{aligned} [\det \mathbf{A}(L)]x_{1t} &= u_{1t} - a_{22}u_{1,t-1} + a_{12}u_{2,t-1} \\ [\det \mathbf{A}(L)]x_{2t} &= u_{2t} - a_{11}u_{2,t-1} + a_{21}u_{1,t-1} \end{aligned}$$

Again, the autoregressive matrix $\mathbf{A}(L)$ has been transformed into $n = 2$ autoregressive polynomials $\det \mathbf{A}(L)$ of order $n = 2$, therefore they have two roots.

Back to the general case, we know from Chapter 4 that a polynomial in $z \in \mathbb{C}$ is invertible if all its roots are outside the unit circle. Therefore, if all the roots of $\det \mathbf{A}(z) = 0$ lie outside the unit circle, then $\det \mathbf{A}(L)$ can be inverted backwards and (47) or equivalently (48) has solution

$$\mathbf{x}_t = [\det \mathbf{A}(L)]^{-1}\mathbf{A}_{ad}(L)\mathbf{u}_t.$$

For $p = 1$, we have n roots of $\det \mathbf{A}(z) = \det(\mathbf{I}_n - \mathbf{A}z) = 0$ which are the reciprocals of the roots of $\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0$, that is the eigenvalues of $\mathbf{A}$. The latter lie within the unit circle if and only if the former lie without, which is the result obtained before (the results obtained with the different approaches are consistent).

10.5.2 VAR(p)

Consider now the generic VAR of order p

$$\mathbf{x}_t = \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_2\mathbf{x}_{t-2} + \ldots + \mathbf{A}_p\mathbf{x}_{t-p} + \mathbf{u}_t$$

where $\mathbf{u}_t$ is an n-dimensional zero-mean vector white noise with covariance matrix $\mathbf{\Gamma}_0^u$. In order to study stationarity we study the solutions of the equation

$$(\mathbf{I}_n - \mathbf{A}_1L - \mathbf{A}_2L^2 - \ldots - \mathbf{A}_pL^p)\mathbf{x}_t = \mathbf{A}(L)\mathbf{x}_t = \mathbf{u}_t \tag{48}$$

[53] In more detail, define as A_{ij} the (i, j) minor of $\mathbf{A}$, i.e. the determinant of the $(n-1) \times (n-1)$ matrix that results from deleting row i and column j of $\mathbf{A}$. Then the cofactor matrix of $\mathbf{A}$ is the $n \times n$ matrix $\mathbf{C}$ whose (i, j) entry is $\mathbf{C}_{ij} = (-1)^{i+j}A_{ij}$. Then, $(\mathbf{A}_{ad})_{ij} = \mathbf{C}_{ji} = (-1)^{i+j}A_{ji}$.

The entries of the polynomial matrix $\mathbf{A}(L)$ are polynomials in L, for example

$$A_{11}(L) = 1 - a_{11,1}L - a_{11,2}L^2 - \ldots - a_{11,p}L^p, \quad A_{12}(L) = -a_{12,1}L - a_{12,2}L^2 - \ldots - a_{12,p}L^p.$$

We can always rewrite this as a VAR(1) of dimension np.

Start with an example. Suppose that $n = p = 2$

$$\mathbf{x}_t = \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_2\mathbf{x}_{t-2} + \mathbf{u}_t$$

Define $\mathbf{y}_t = \mathbf{x}_{t-1}$. Then we have the system of two equations

$$\begin{aligned}\mathbf{x}_t &= \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-1} + \mathbf{u}_t \\ \mathbf{y}_t &= \mathbf{x}_{t-1}\end{aligned}$$

If we define also

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{I}_2 & \mathbf{0}_2 \end{pmatrix} \qquad \mathbf{z}_t = (\mathbf{x}_t' \; \mathbf{y}_t'), \qquad \mathbf{v}_t = (\mathbf{u}_t' \; 0 \; 0)'$$

we have the new VAR(1)

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{v}_t$$

where now $p = 1$ but $n = 4$. Moreover, notice that we have

$$\det(\mathbf{I}_4 - \mathbf{A}L) = \det\left(\begin{pmatrix} \mathbf{I}_2 - \mathbf{A}_1L & -\mathbf{A}_2L \\ -\mathbf{I}_2L & \mathbf{I}_2 \end{pmatrix}\right) = \det(\mathbf{I}_2 - \mathbf{A}_1L - \mathbf{A}_2L^2)$$

By studying the roots of the above equation we can study stationarity. Also we have that the eigenvalues of $\mathbf{A}$ satisfy

$$\det(\mathbf{A} - \lambda\mathbf{I}_4) = \det\left(\begin{pmatrix} \mathbf{A}_1 - \lambda\mathbf{I}_2 & \mathbf{A}_2 \\ \mathbf{I}_2 & -\lambda\mathbf{I}_2 \end{pmatrix}\right) = \det(\mathbf{I}_2\lambda^2 - \mathbf{A}_1\lambda - \mathbf{A}_2) = 0.$$

Hence roots of $\det(\mathbf{I}_2 - \mathbf{A}_1L - \mathbf{A}_2L^2) = 0$ are the reciprocals of eigenvalues.

In general given the VAR(p) in (48), we can define $\mathbf{y}_{1t} = \mathbf{x}_{t-1}$, $\mathbf{y}_{2t} = \mathbf{y}_{1,t-1} = \mathbf{x}_{t-2}$, $\mathbf{y}_{3t} = \mathbf{y}_{2,t-1} = \mathbf{x}_{t-3}$, and so on until $\mathbf{y}_{p-1,t} = \mathbf{y}_{p-2,t-1} = \mathbf{x}_{t-p+1}$ and we obtain

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{v}_t$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \ldots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I}_n & \mathbf{0}_n & \ldots & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{I}_n & \ldots & \mathbf{0}_n & \mathbf{0}_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_n & \mathbf{0}_n & \ldots & \mathbf{I}_n & \mathbf{0}_n \end{pmatrix} \qquad \mathbf{z}_t = (\mathbf{x}_t' \; \mathbf{y}_{1t}' \ldots \mathbf{y}_{p-1,t}'), \qquad \mathbf{v}_t = (\mathbf{u}_t' \; \underbrace{0 \; 0 \ldots 0}_{n(p-1)\text{times}})'$$

In conclusion, (48), which has dimension n and order p, can be transformed into an equation of dimension np and order 1. The VAR(1) equation in $\mathbf{z}_t$ is called the companion equation and we then know how to study stationarity of a VAR(1) by looking at the eigenvalues of the companion matrix $\mathbf{A}$. Or equivalently, as for the case $n = 2$ and $p = 2$ above, we can prove that $\det(\mathbf{I}_{np} - \mathbf{A}L) = \det(\mathbf{I}_n - \mathbf{A}_1L - \ldots - \mathbf{A}_pL^p)$ therefore we can also look at the roots of $\det(\mathbf{I}_{np} - \mathbf{A}z) =$

0. Notice also that $\mathbf{v}_t$ is singular (many of its components are zero) but this does not affect stationarity.

Summing up, as in the VAR(1) and VAR(2) cases, also for the VAR(p) the roots of

$$\det(\mathbf{I}_{np} - \mathbf{A}z) = \det(\mathbf{I}_n - \mathbf{A}_1 z - \ldots - \mathbf{A}_{p-1}z^{p-1} - \mathbf{A}_p z^p) = 0$$

are the reciprocals of the eigenvalues of the companion matrix defined by

$$\det(\mathbf{A} - \lambda \mathbf{I}_{np}) = \det(\mathbf{I}_n \lambda^p - \mathbf{A}_1 \lambda^{p-1} - \ldots - \mathbf{A}_{p-1}\lambda - \mathbf{A}_p) = 0.$$

Example. Obviously an AR(2) can be written as a VAR(1) with $n = 2$. Take

$$y_t = 1.3y_{t-1} - 0.4y_{t-2} + u_t,$$

then we must find the roots of $A(z) = 1 - 1.3z + 0.4z^2$, this gives $z_1 = 2$ and $z_2 = 1.25$, therefore the model is stationary. Using the companion form we have

$$\begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1.3 & -0.4 \\ 1 & 0 \end{pmatrix}}_{\mathbf{A}} \begin{pmatrix} y_{t-1} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \end{pmatrix}.$$

We can then compute the eigenvalues of $\mathbf{A}$ by solving $\det(\mathbf{A} - \mathbf{I}_2\lambda) = 0$, which gives

$$-\lambda(1.3 - \lambda) + 0.4 = 0,$$

hence we have the solutions $\lambda_1 = 0.8$ and $\lambda_2 = 0.5$, which implies that the process is stationary. Notice that $\lambda_1 = z_1^{-1}$ and $\lambda_2 = z_2^{-1}$.

10.6 VARMA

Vector ARMA processes (VARMA) are the stationary solutions of equations of the form

$$(\mathbf{I}_n - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \ldots - \mathbf{A}_p L^p)\mathbf{x}_t = (\mathbf{I}_n + \mathbf{B}_1 L + \mathbf{B}_2 L^2 + \ldots + \mathbf{B}_q L^q)\mathbf{u}_t$$

Under the assumptions that the roots of $\det(\mathbf{I}_n - \mathbf{A}_1 z - \mathbf{A}_2 z^2 - \ldots - \mathbf{A}_p z^p) = 0$ lie outside of the unit circle, the stationary solution is

$$\mathbf{x}_t = \mathbf{C}(L)\mathbf{u}_t = \mathbf{A}(L)^{-1}\mathbf{B}(L)\mathbf{u}_t = [\det \mathbf{A}(L)]^{-1}\mathbf{A}_{ad}(L)\mathbf{B}(L)\mathbf{u}_t$$

Here $\mathbf{C}(L) = \sum_{k=0}^{\infty} \mathbf{C}_k L^k$ and the coefficients of $\mathbf{C}(L)$ tend to zero (they must to have stationarity and therefore the previous sum well defined) with a speed determined by the worst root of $\det \mathbf{A}(L)$, precisely like in the scalar case (see the example of an AR(1)).[54] So, for a VAR(1) the MA(∞) is given by $\mathbf{C}(L) = \sum_{k=0}^{\infty} \mathbf{A}^k L^k$.

If also the roots of $\det(\mathbf{I}_n + \mathbf{B}_1 z + \mathbf{B}_2 z^2 + \ldots + \mathbf{B}_q z^q) = 0$ are outside the unit circle we say that the VARMA is invertible in the past. In that case

$$\mathbf{u}_t = \mathbf{B}(L)^{-1}\mathbf{A}(L)\mathbf{x}_t$$

so that $\mathbf{u}_t$ is a linear combination of present and past values of $\mathbf{x}_t$. This implies, that $\mathbf{u}_t$ is the innovation of $\mathbf{x}_t$ (see next section).

[54] More precisely we have $\sum_{k=0}^{\infty} \|\mathbf{C}_k\|^2 < \infty$, where $\|\mathbf{C}_k\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}[\mathbf{C}_k]_{ij}^2$. This also implies $\|\mathbf{C}_k\| \to 0$ as $k \to \infty$.

10.7 Prediction and the Wold decomposition

Now consider the vector $(x_{1t}x_{2t})'$. We want to predict each of the variables by using the past of both variables. We consider only linear prediction. Given what we have discussed in the scalar $n = 1$ case, the solution of this problem is given by projecting x_{jt}, $j = 1, 2$, on 1 and $x_{j,t-k}$, $j = 1, 2$, $k > 0$

$$x_{1t} = [a_{10} + a_{11,1}x_{1,t-1} + a_{12,1}x_{2,t-1} + a_{11,2}x_{1,t-2} + a_{12,2}x_{2,t-2} + \ldots] + e_{1t}$$
$$x_{2t} = [a_{20} + a_{21,1}x_{1,t-1} + a_{22,1}x_{2,t-1} + a_{21,2}x_{1,t-2} + a_{22,2}x_{2,t-2} + \ldots] + e_{2t}$$

In vector notation

$$\mathbf{x}_t = [\mathbf{A}_0 + \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_2\mathbf{x}_{t-2} + \ldots] + \mathbf{e}_t$$

where $\mathbf{A}_0$ is a 2×1 vector while $\mathbf{A}_k$, for $k \geq 1$, is 2×2.

Using the same argument employed in the scalar case we obtain the result that $\mathbf{e}_t$ is a vector white noise, that is e_{1t} is orthogonal to past values of both e_{1t} and e_{2t}, and the same for e_{2t}. For, e_{1t} is orthogonal to 1, $x_{1,t-1}$, $x_{2,t-1}$,.... But $e_{1,t-1}$ and $e_{2,t-1}$ are linear combinations of 1, $x_{1,t-1}$, $x_{2,t-1}$,....

In general, if $\mathbf{x}_t$ is n-dimensional the best linear prediction of the components of $\mathbf{x}_t$ is obtained by projecting each of them on 1 and past values of all the components of $\mathbf{x}_t$:

$$\mathbf{x}_t = \mathbf{A}_0 + \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_2\mathbf{x}_{t-2} + \ldots + \mathbf{e}_t$$

where $\mathbf{A}_0$ is $n \times 1$, $\mathbf{A}_k$ is $n \times n$ for $k \geq 1$. The result that $\mathbf{e}_t$ is an n-dimensional white noise is obtained by an obvious generalization of the argument used for $n = 2$. As in the scalar case we then have the prediction equation for $\mathbf{x}_t$

$$\mathbf{x}_t = P_{t-1}\mathbf{x}_t + \mathbf{e}_t$$

where $P_{t-1}\mathbf{x}_t$ is the projection of $\mathbf{x}_t$ on its past and $\mathbf{e}_t$ since it is a white noise is the innovation or one step ahead prediction error. The only reason why the process $\mathbf{x}_t$ is not completely determined by its past values is the presence of the term $\mathbf{e}_t$.

Now project each component of $\mathbf{x}_t$ on 1 and all components of $\mathbf{e}_{t-k}$, $k \geq 0$,

$$\mathbf{x}_t = [\mathbf{B} + \mathbf{e}_t + \mathbf{B}_1\mathbf{e}_{t-1} + \mathbf{B}_2\mathbf{e}_{t-2} + \ldots] + \mathbf{D}_t$$

where $\mathbf{B}$ is $n \times 1$, $\mathbf{B}_k$ is $n \times n$ for $k > 0$. Moreover, the residual of the projection $\mathbf{D}_t$ is an n-dimensional vector that is predictable without error given its past, that is $P_{t-1}\mathbf{D}_t = \mathbf{D}_t$ is a deterministic process.

Examples.

1. Stationary VAR(1)

   $$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t$$

   the best linear predictor is $\mathbf{A}\mathbf{x}_{t-1}$ and $\mathbf{u}_t$ is the innovation.

2. A VMA(1) invertible in the past

   $$\mathbf{x}_t = \mathbf{u}_t + \mathbf{B}\mathbf{u}_{t-1}$$

   this can be re-written as a VAR(∞)

   $$\mathbf{x}_t = \mathbf{u}_t + \sum_{k=1}^{\infty}(-1)^{k-1}\mathbf{B}^k\mathbf{x}_{t-k}$$

   and therefore the best linear predictor is $\mathbf{B}[\mathbf{x}_{t-1} - \mathbf{B}\mathbf{x}_{t-2} + \ldots]$ and $\mathbf{u}_t$ is the innovation.

3. VARMA(p, q) stationary and invertible in the past

$$\mathbf{x}_t = [\mathbf{A}_1\mathbf{x}_{t-1} + \ldots + \mathbf{A}_p\mathbf{x}_{t-p}] + [\mathbf{B}_1\mathbf{u}_{t-1} + \ldots + \mathbf{B}_q\mathbf{u}_{t-p}] + \mathbf{u}_t$$

by using the arguments in the two previous examples we see that the innovation is $\mathbf{u}_t$.

Forecasting of VARs then follows straightforwardly as for the scalar AR case.

10.8 VAR estimation

A VAR can be estimated by OLS. In the scalar case we have seen that under Gaussianity this is also asymptotically equivalent to Maximum Likelihood estimation. Consider for example a VAR(1),

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t$$

where $\mathbf{u}_t$ is a white noise with zero mean and covariance $\mathbf{\Gamma}_u$. We have the OLS estimator

$$\widehat{\mathbf{A}} = \left(\sum_{t=2}^{T} \mathbf{x}_{t-1}\mathbf{x}'_{t-1}\right)^{-1} \left(\sum_{t=2}^{T} \mathbf{x}_{t-1}\mathbf{x}'_{t}\right) \tag{49}$$

when seen as an ML estimator consistency and asymptotic normality are immediately given. In general, if no Gaussianity is assumed, this estimator is consistent and asymptotically normal provided that $\mathbf{u}_t$ is such that $\mathrm{E}[\mathbf{u}_t|\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots] = \mathbf{0}$ (it is a martingale difference sequence).

The above estimator (49) corresponds to estimation of each of the n VAR equations by OLS, and by looking at the first n equations of the companion form we see that any VAR(p) model can also be estimated by OLS equation-by-equation. Consider a VAR(p) and the companion form

$$\mathbf{z}_t = \widetilde{\mathbf{A}}\mathbf{z}_{t-1} + \mathbf{v}_t$$

where $\mathbf{z}_{t-1} = (\mathbf{x}'_{t-1}\ \mathbf{x}'_{t-2} \ldots \mathbf{x}'_{t-p})$. Then, for equation j we have the OLS estimator:

$$\widehat{\widetilde{\mathbf{a}}}_j = \left(\sum_{t=2}^{T} \mathbf{z}_{t-1}\mathbf{z}'_{t-1}\right)^{-1} \left(\sum_{t=2}^{T} \mathbf{z}_{t-1}x'_{jt}\right), \quad j = 1, \ldots n.$$

Then each $\widehat{\widetilde{\mathbf{a}}}_j$ contains the np estimated coefficient of the j-th equation of the VAR(p) that is of

$$x_{jt} = a_{j1,1}x_{1,t-1} \ldots a_{jn,1}x_{n,t-1} + \ldots + a_{j1,p}x_{1,t-p} + \ldots + a_{jn,p}x_{n,t-p} + u_{jt}, \quad j = 1, \ldots n.$$

So $\widehat{\widetilde{\mathbf{a}}}_1 \ldots \widehat{\widetilde{\mathbf{a}}}_n$ contain estimates of all parameters $\mathbf{A}_1 \ldots \mathbf{A}_p$. The above is a system of seemingly unrelated regression equations and OLS is in general consistent but not efficient even under Gaussianity. However, since in a VAR the regressors are the same in each equation it can be proved that OLS is also efficient provided Gaussianity or serial independence of $\mathbf{u}_t$ are assumed (this is a consequence of Kruskal's theorem).[55]

The covariance matrix of $\mathbf{u}_t$ in a VAR (p) is then estimated as

$$\widehat{\mathbf{\Gamma}}_0^u = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{u}}_t\widehat{\mathbf{u}}'_t,$$

where $\widehat{\mathbf{u}}_t = \mathbf{x}_t - \sum_{k=1}^{p} \widehat{\mathbf{A}}_k\mathbf{x}_{t-k}$. It can be proved that also $\widehat{\mathbf{\Gamma}}_0^u$ is consistent when fourth moments of $\mathbf{u}_t$ exist.

[55] If that were not the case then generalised least squares should be preferred using an estimator of the covariance of the $\mathbf{u}_t$'s obtained by a first OLS estimation.

10.9 Granger causality

We say that a process $\{X_t\}$ Granger Causes (GC) a process $\{Y_t\}$ if

$$\mathrm{E}[Y_t|Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots] \neq \mathrm{E}[Y_t|Y_{t-1}, Y_{t-2}, \ldots]$$

that is to say that the past observations of $\{X_t\}$ help in predicting $\{Y_t\}$.

As an example consider a VAR with $n = 2$

$$\begin{pmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{pmatrix} \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \mathbf{u}_t$$

then $\{X_t\}$ does not Granger Causes $\{Y_t\}$ if and only if $A_{12}(L) = 0$. Therefore if the VAR is stationary and we estimate it by OLS then we can just use the ordinary F-test to test for Granger causality. In the example above we have the first equation of the VAR

$$y_t = a_{1,11}y_{t-1} + a_{1,12}x_{t-1} + \ldots + a_{p,11}y_{t-p} + a_{p,12}x_{t-p} + u_{1t}$$

then $\{X_t\}$ does not Granger Causes $\{Y_t\}$ if and only if $a_{k,12} = 0$ for any $k = 1, \ldots, p$, which is a set of linear restriction that can be tested by means of the F-test.

10.10 Systems of simultaneous equations and impulse response functions

A VAR is a system of simultaneous equations. These systems were usually employed by macroeconomists to study jointly many economic indicators. In general the models were written as

$$\mathbf{\Gamma}\mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \mathbf{u}_t$$

where $\mathbf{y}_t$ are the n endogenous variables, $\mathbf{x}_t$ are k exogenous variables, and $\mathbf{u}_t$ is a white noise with mean zero and covariance the identity matrix, such that the elements of $\mathbf{u}_t$ are uncorrelated and are interpreted by macroeconomists as "structural shocks" hitting the economy. In order to have an interpretation of the model we would like to estimate $\mathbf{\Gamma}$ and $\mathbf{B}$ but these cannot be consistently estimated via OLS. Hence we reduce ourselves to the model

$$\mathbf{y}_t = \mathbf{\Pi}\mathbf{x}_t + \mathbf{w}_t$$

where $\mathbf{w}_t$ is still a white noise with mean zero but covariance $\mathbf{\Sigma}$ and $\mathbf{\Pi} = \mathbf{\Gamma}^{-1}\mathbf{B}$. This model can be estimated by OLS but the coefficients in $\mathbf{\Pi}$ do not have an economic interpretation. In order to identify $\mathbf{\Gamma}$ and $\mathbf{B}$ separately we should solve $\mathbf{\Gamma\Pi} = \mathbf{B}$ but there are more parameters ($n^2 + nk$) than equations (n) thus no unique solution exists. Some economic restriction should then be imposed on those matrices in order to achieve identification.

A VAR model can then also be seen as the reduced form of a structural model in which there are no exogenous regressors but the regressors are the lags of the endogenous variables and we have seen that OLS estimation is possible. The question is how do we then identify the parameters? Traditional Structural VAR literature has focused on identifying the dynamic dependencies among macroeconomic variables by identifying the so called impulse response functions (IRF) of the variables to unexpected orthogonal (non-correlated) shocks which represent the new information entering the economy at each point in time as the usual innovation process.

Given the VAR process

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{e}_t,$$

where $\mathbf{e}_t$ is a white noise with zero mean and covariance matrix $\mathbf{\Gamma}_e$, the definition of IRF comes from the equivalent VMA representation[56]

$$\mathbf{x}_t = \sum_{k=0}^{\infty} \mathbf{C}_k \mathbf{e}_{t-k} = \mathbf{C}(L)\mathbf{e}_t,$$

as usual we have $\mathbf{C}_0 = \mathbf{I}_n$. We define the IRF as

$$h(i,j,k) = \frac{\partial x_{it}}{\partial e_{j,t-k}} = (\mathbf{C}_k)_{ij}$$

this is the response of the ith variable to the j-th shock after k periods. Notice that $\mathbf{e}_t$, is the one-step-ahead prediction error which is the innovation in the sense that it is unpredictable and belongs to the space of present and past values of $\mathbf{x}_t$. However, with respect to the scalar case we have n prediction errors one for each series and we can try to disentangle the shocks trying to find the prediction error for each single series. This can be done by working on their covariance matrix. Formally, we assume that the prediction errors are a linear relation of the structural shocks, $\mathbf{u}_t$ which are also a white noise with zero mean but with covariance the identity matrix.

$$\mathbf{e}_t = \mathbf{B}\mathbf{u}_t$$

and we assume $\mathbf{B}$ to be $n \times n$ invertible, that is we have n distinct structural shocks.[57] While $\mathbf{e}_t$ can be estimated as the residual of a VAR since it is an innovation, $\mathbf{u}_t$ and $\mathbf{B}$ cannot be estimated directly.

Let us start by assuming we knew $\mathbf{B}$, then we would have the VAR

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{B}\mathbf{u}_t,$$

which once inverted would give

$$\mathbf{x}_t = \sum_{k=0}^{\infty} \mathbf{C}_k \mathbf{B}\mathbf{u}_{t-k}$$

and the structural IRF are defined as

$$IRF(i,j,k) = \frac{\partial x_{it}}{\partial u_{j,t-k}} = (\mathbf{C}_k\mathbf{B})_{ij}$$

Therefore, we have identification if we find a way to obtain $\mathbf{B}$ starting from the coefficient of an estimated VAR $\mathbf{A}(L)$ and its innovations $\mathbf{e}_t$. In particular this is possible under the assumption that the structural shocks $\mathbf{u}_t$ like the innovations $\mathbf{e}_t$ belong to the space of present and past values of $\mathbf{x}_t$. In this case we say that $\mathbf{u}_t$ is fundamental for $\mathbf{x}_t$. Economic models with non-fundamental structural shocks are possible but cannot be identified using VAR (see the end of the section).

For ease of notation we will denote as $\mathbf{D}(L) = \mathbf{C}(L)\mathbf{B}$ the polynomial of structural IRF. There are no general rules to identify IRF that and we review two examples.

[56] In order to invert a VAR(p) we can use the companion form and notice that for a VAR(1)

$$(\mathbf{I}_n - \mathbf{A}L)\mathbf{x}_t = \mathbf{e}_t$$

the inverse of the VAR polynomial is given by

$$(\mathbf{I}_n - \mathbf{A}L)^{-1} = \mathbf{I}_n + \mathbf{A}L + \mathbf{A}^2L^2 \ldots = \sum_{k=0}^{\infty} \mathbf{A}^k L^k,$$

thus $\mathbf{C}_k = \mathbf{A}^k$.

[57] It could also be singular if there were fewer structural shocks than series.

1. Recursive or Choleski identification. This is based on the covariance matrix of the VAR residuals, $\mathbf{e}_t$, denoted as $\mathbf{\Gamma}_e$. Notice that since $\mathbf{u}_t$ have as covariance $\mathbf{I}_n$, we must have $\mathbf{\Gamma}_e = \mathbf{B}\mathbf{B}'$. If we impose $\mathbf{B}$ to be lower triangular then the solution is given by the Choleski factor of $\mathbf{\Gamma}_e$ and its unique. Saying that $\mathbf{B}$ is lower triangular means that only the first shock has a contemporaneous effect on all variables x_{it}, indeed take for example $n = 2$ then at $k = 0$ the structural IRF would be (recall that $\mathbf{C}_0 = \mathbf{I}_n$ always)

$$\mathbf{C}_0\mathbf{B} = \mathbf{B} = \begin{pmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{pmatrix}$$

hence since $b_{12} = 0$ we have $IRF(1, 2, 0) = 0$ that is the second shock does not have a contemporaneous effect on the first variable. Or equivalently we have that the prediction errors are given by

$$\begin{pmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{nt} \end{pmatrix} = \begin{pmatrix} b_{11} & 0 & \dots & 0 \\ b_{21} & b_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{nt} \end{pmatrix} = \begin{pmatrix} b_{11}u_{1t} \\ b_{21}u_{1t} + b_{22}u_{2t} \\ \vdots \\ \sum_{k=1}^{n} b_{nk}u_{kt} \end{pmatrix}$$

The first prediction error depends only on the first structural shock, the second prediction error depends only on the first two structural shocks, and in general the ith prediction error depends only on the first i structural shocks. This implies that the ordering of the elements of $\mathbf{x}_t$ is not arbitrary. The structural shocks are then identified recursively from the prediction errors.

As an example consider the three series, GDP growth rate, CPI growth rate (inflation), and Federal Funds Rate. They are all stationary. A monetary policy shock is identified as a shock that has contemporaneous effect only on the Federal Funds Rate so we can use a Choleski identification such that the effect of shocks on variables is given by

$$\begin{pmatrix} \Delta \log GDP_t \\ \Delta \log CPI_t \\ FFR_t \end{pmatrix} = \mathbf{C}_0\mathbf{B}\mathbf{u}_t + \sum_{k=1}^{\infty} \mathbf{C}_k\mathbf{B} = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix} + \sum_{k=1}^{\infty} \mathbf{C}_k\mathbf{B}$$

so we see that at impact ($k = 0$) the third shock has effect only on the third variable and we can define it as the monetary policy shock. The impulse responses implied by this identification and using US data from 1954q3 to 2006q2 are in are in Figure 59.

2. Long run restrictions. These are based on the assumption that in the long run only some shocks have an effect on the variables. First, notice that by definition of IRF we have that on a stationary variable the lag k effect of a shock is given by $\mathbf{C}_k$, thus the long run effect is given by $\lim_{k\to\infty} \mathbf{C}_k = 0$, since because of stationarity we must have square summable IRF $\lim_{k\to\infty} \mathbf{C}_k = 0$.[58] So this identification makes sense only if we work with $I(1)$ variables and we look at their response in the long run. Take $\mathbf{y}_t \sim I(1)$, so that $\Delta\mathbf{y}_t \sim I(0)$ and we the structural VAR in first differences

$$\mathbf{A}(L)\Delta\mathbf{y}_t = \mathbf{B}\mathbf{u}_t$$

and then the structural MA representation reads

$$\Delta\mathbf{y}_t = \sum_{k=0}^{\infty} \mathbf{C}_k\mathbf{B}\mathbf{u}_{t-k}. \tag{50}$$

[58] For example take the stationary VAR(1) process then $\sum_{k=0}^{\infty} \mathbf{A}^k = \sum_{k=0}^{\infty} \mathbf{C}_k$ converges, or in other words the MA representation is well defined, and in order to have that we must have $\lim_{k\to\infty} \mathbf{A}^k = 0$ which is indeed the case since because of stationarity $\mathbf{A}$ has the largest eigenvalue inside the unit circle.

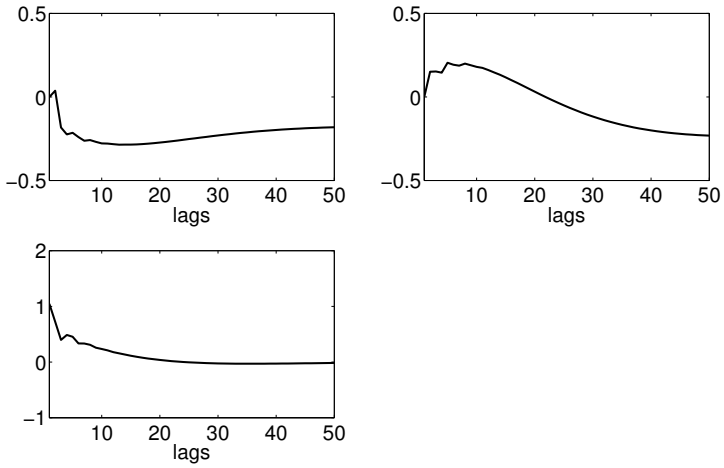


Figure 57: IRF to a monetary policy shock. Top left: GDP; top right: CPI; bottom left: Federal Funds Rate.

If we are interested on the effect of u_{jt-k} on y_{it} we must compute the cumulated structural IRF

$$CIRF(i,j,k) = \frac{\partial y_{it}}{\partial u_{jt-k}} = \frac{\partial y_{it+k}}{\partial u_{jt}} = \frac{\partial \sum_{\ell=0}^{k} \Delta y_{it+\ell}}{\partial u_{jt}} = \sum_{\ell=0}^{k} IRF(i,j,\ell) = \sum_{\ell=0}^{k} (\mathbf{C}_\ell \mathbf{B})_{ij}\,.$$

The long run effect is then the cumulative response after infinite periods: $\sum_{k=0}^{\infty} IRF(i,j,k)$, which is given by $\mathbf{C}(1)\mathbf{B} = \sum_{k=0}^{\infty} \mathbf{C}_k \mathbf{B}$. This series converges since $\Delta \mathbf{y}_t$ is stationary but it is in general not zero since $\mathbf{y}_t \sim I(1)$ (see Chapter 8).

Example. Assume that Δy_{1t} is the GDP growth rate and Δy_{2t} is the unemployment rate. Then $n = 2$ we have

$$\begin{aligned}\mathbf{C}(1)\mathbf{B} &= \begin{pmatrix} \sum_{k=0}^{\infty} c_{k,11} & \sum_{k=0}^{\infty} c_{k,12} \\ \sum_{k=0}^{\infty} c_{k,21} & \sum_{k=0}^{\infty} c_{k,22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ &= \begin{pmatrix} C_{11}(1)b_{11} + C_{12}(1)b_{21} & C_{11}(1)b_{12} + C_{12}(1)b_{22} \\ C_{21}(1)b_{11} + C_{22}(1)b_{21} & C_{21}(1)b_{12} + C_{22}(1)b_{22} \end{pmatrix}\end{aligned}$$

Blanchard and Quah say that the second shock is a demand shock and therefore does not have permanent effect on the first variable which is GDP. Then we must have $C_{11}(1)b_{12} + C_{12}(1)b_{22} = 0$, or by setting $\theta = -\frac{C_{12}(1)}{C_{11}(1)}$ we have

$$\mathbf{B}\mathbf{B}' = \begin{pmatrix} b_{11}^2 + \theta^2 b_{22}^2 & b_{11}b_{21} + \theta b_{22} \\ b_{11}b_{21} + \theta b_{22} & b_{21}^2 + b_{22}^2 \end{pmatrix} = \mathbf{\Gamma}_e = \begin{pmatrix} \sigma_{e_1}^2 & \sigma_{e_1 e_2} \\ \sigma_{e_1 e_2} & \sigma_{e_2}^2 \end{pmatrix}$$

which implies a system of three equations in three unknowns (b_{11}, b_{21}, b_{22}). The IRF implied by this identification and using US data from 1950q1 to 2010q4 are in Figure 58 while when using the sample 1950q1 to 1989q4 are in Figure 59. The first permanent shock is called a supply shock and the second non-permanent is called a demand shock.[59]

[59]The same result is found by considering a rotation of the innovations such that $\mathbf{R}\mathbf{e}_t = \mathbf{u}_t$ with $\mathbf{R}\mathbf{R}' = \mathbf{I}$, which in the case $n = 2$ has the simple form

$$\mathbf{R} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \qquad \varphi \in [0, 2\pi]$$

solving for φ we get the same identification with $\varphi = \tan^{-1}\theta$.

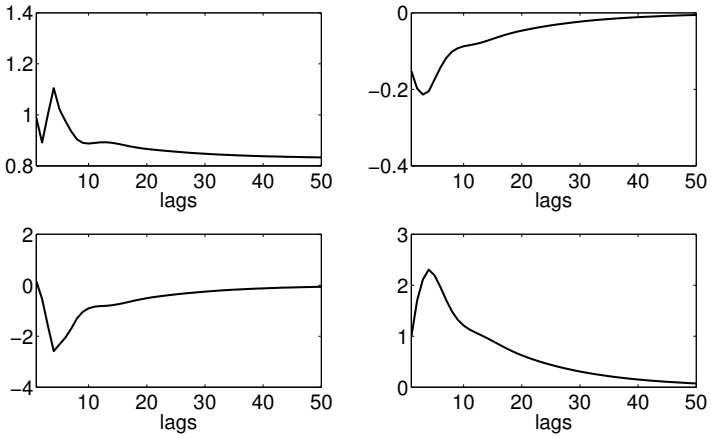


Figure 58: IRF to a demand and supply shocks. Top left: GDP to supply; top right: GDP to demand; bottom left: unemployment to supply; bottom right: unemployment to demand.

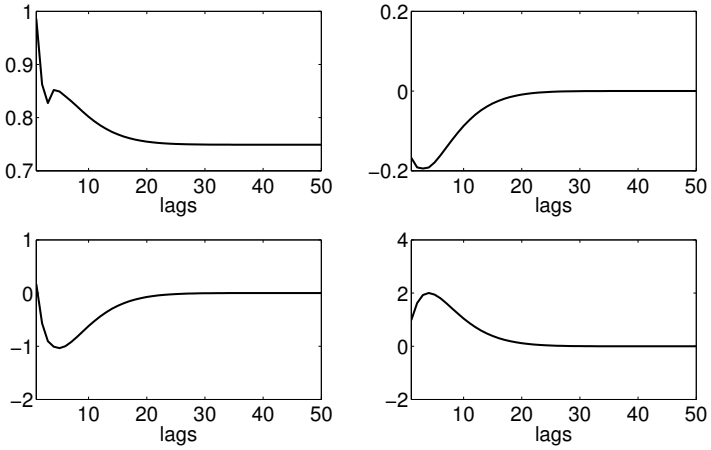


Figure 59: IRF to a demand and supply shocks. Top left: GDP to supply; top right: GDP to demand; bottom left: unemployment to supply; bottom right: unemployment to demand.

As we said the above methods are possible only under the assumption that the structural shocks $\mathbf{u}_t$ like the innovations $\mathbf{e}_t$ belong to the space of present and past values of $\mathbf{x}_t$, that is when $\mathbf{u}_t$ is fundamental for $\mathbf{x}_t$. Notice that in practice economic theory starts from an MA model where structural IRF are defined. Then the structural shocks are a linear combination of the VAR innovations (which are those that we estimate) only if the MA can be written as a VAR that is only if the MA is invertible in the past. Consider a VAR(p) then

$$\mathbf{e}_t = \mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \ldots - \mathbf{A}_p\mathbf{x}_{t-p}$$

hence the innovations are always fundamental. There are however economic examples in which the MA is invertible only in the future and we say in that case that the shocks are non-fundamental.[60]

As an example consider the the permanent income Friedman-Muth model. Income y_t is decomposed in a permanent part y_{1t} and a transitory part y_{0t} which are independently affected by

[60]When the MA is not invertible because it has unit roots then we might still have fundamental shocks.

uncorrelated shocks

$$(1-L)y_{1t} = u_{1t}, \qquad y_{0t} = u_{0t}.$$

If consumption c_t follows the permanent income hypothesis, we have: $(1-L)c_t = u_{1t}+(1-\beta)u_{0t}$ where $\beta \in (0,1)$ is the agent discount factor. Therefore, we have the structural MA model

$$\mathbf{x}_t = \begin{pmatrix} (1-L)y_t \\ (1-L)c_t \end{pmatrix} = \begin{pmatrix} 1 & 1-L \\ 1 & 1-\beta \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{0t} \end{pmatrix} = \mathbf{C}(L)\mathbf{u}_t.$$

In this case $\det \mathbf{C}(z) = (z-\beta)$ and hence it has the only root in $z = \beta$, which by definition lies inside the unit circle. The above econometric model representing the permanent income model is nonfundamental. Permanent and transitory components of income are not recoverable by considering only present and past values of income and consumption and VAR cannot be used as an estimation tool.

11 Multivariate unit root processes

11.1 VAR for $I(1)$ processes

We generalise the case of unit root processes to a process of dimension n. Consider a VAR(p) model

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{u}_t, \quad \mathbf{u}_t \sim w.n.(\mathbf{0}_n, \mathbf{\Gamma}_0^u),$$

which is stationary if the equation $\det(\mathbf{A}(z)) = 0$ has roots outside the unit circle. We now consider the case in which some of the roots are such that $|z| = 1$. Notice that even in the case $p = 1$ there are n roots to be considered, and therefore in general there are np roots to be considered.[61] So an important issue is not only whether there are unit roots but also knowing how many unit roots are present.

Start with a VAR(2) model

$$(\mathbf{I}_n - \mathbf{A}_1 L - \mathbf{A}_2 L^2)\mathbf{x}_t = \mathbf{u}_t, \quad \mathbf{u}_t \sim w.n.(\mathbf{0}_n, \mathbf{\Gamma}_0^u),$$

then the companion matrix is

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{I}_n & \mathbf{0}_n \end{pmatrix}$$

such that it has eigenvalues defined by

$$\det(\tilde{\mathbf{A}} - \lambda \mathbf{I}_{np}) = \det(\lambda^2 \mathbf{I}_n - \lambda \mathbf{A}_1 - \mathbf{A}_2) = 0.$$

If we have a unit root then by definition $\det(\mathbf{A}(1)) = 0$ where $\mathbf{A}(1) = \mathbf{I}_n - \mathbf{A}_1 - \mathbf{A}_2$. But notice also that

$$\det(\tilde{\mathbf{A}} - \mathbf{I}_{np}) = \det(\mathbf{I}_n - \mathbf{A}_1 - \mathbf{A}_2) = \det(\mathbf{A}(1)) = 0,$$

which shows that when we have unit roots at least one eigenvalue of the companion matrix is $\lambda = 1$.

There are two possible cases that give $\det(\mathbf{A}(1)) = 0$:

1. $\mathbf{A}(1) = \mathbf{0}_n$, that is rank$(\mathbf{A}(1)) = 0$;

[61] Recall that the condition for stationarity can also be restated by looking at the eigenvalues of the companion matrix which has dimension $np \times np$.

2. rank$(\mathbf{A}(1)) = r < n$.

Let us define rank$(\mathbf{A}(1)) = k$ with $0 \leq k < n$, then to study the two cases it is useful to write the matrix $\mathbf{A}(L)$ using its Smith McMillan form

$$\mathbf{A}(L) = \mathbf{U}(L) \begin{pmatrix} \mathbf{I}_{n-k}(1-L) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix} \mathbf{V}(L) \tag{51}$$

where $\mathbf{U}(L)$ and $\mathbf{V}(L)$ are $n \times n$ and have no unit root hence are invertible. Notice that if we had $k = n$ then the VAR is invertible since there are no unit roots.

1. $\mathbf{A}(1) = \mathbf{0}_n$ then $k = 0$ and we immediately see from (51) that there are n unit roots, all components of $\mathbf{x}_t$ are non-stationary. It is useful then to consider the Beveridge-Nelson decompositon for a polynomial matrix (see Chapter 8.5) which is given by[62]

$$\mathbf{A}(z) = \mathbf{A}(1) + \mathbf{A}^*(z)(1-z)$$

which, since $\mathbf{A}(1) = \mathbf{0}_n$, implies that

$$\mathbf{A}^*(L)(1-L)\mathbf{x}_t = \mathbf{u}_t$$

where from (51) we have $\mathbf{A}^*(L) = \mathbf{U}(L)\mathbf{V}(L)$. Now $\mathbf{A}^*(L)$ has no unit root, thus $\Delta\mathbf{x}_t$ is stationary and the process $\mathbf{x}_t$ is $I(1)$.[63] In this case we must carry on the empirical analysis on $\Delta\mathbf{x}_t$ rather than on $\mathbf{x}_t$. It can also be proved that the number of unitary eigenvalues of the companion matrix is n (but the viceversa is in general not true). Moreover there is no Wold representation for $\mathbf{x}_t$ since it is non-stationary. However we have the Wold decomposition for the first differences

$$\Delta\mathbf{x}_t = \mathbf{C}(L)\mathbf{u}_t$$

where $\mathbf{C}(L) = (\mathbf{A}^*(L))^{-1} = \mathbf{V}^{-1}(L)\mathbf{U}^{-1}(L)$ (invertible since $\mathbf{A}^*(L)$ has no unit root). Thus, the IRFs are $\mathbf{D}(L)$ such that $(1-L)\mathbf{D}(L) = \mathbf{C}(L)$ and

$$[\mathbf{D}_k]_{i,j} = \sum_{h=0}^{k} \frac{\partial \Delta x_{it}}{\partial u_{j,t-h}} = \sum_{h=0}^{k} (\mathbf{C}_h)_{ij}$$

and we see that in this case $\mathbf{D}_k \not\to \mathbf{0}$ hence are not square-summable. Identification can then proceed as shown in the previous Chapter (see for example the case of long run restrictions).

As an example take again the VAR(2), then $\mathbf{A}^*(L)$ must be of order 1 such that

$$\mathbf{A}^*(L)(1-L)\mathbf{x}_t = (\mathbf{I}_n - \mathbf{A}_1^* L)(1-L)\mathbf{x}_t = (\mathbf{I}_n - (\mathbf{I}_n + \mathbf{A}_1^*)L + \mathbf{A}_1^* L^2)\mathbf{x}_t = \mathbf{u}_t$$

Then $\mathbf{A}(1) = \mathbf{0}_n$ as expected and the companion matrix is

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{I}_n + \mathbf{A}_1^* & -\mathbf{A}_1^* \\ \mathbf{I}_n & \mathbf{0}_n \end{pmatrix}$$

such that it has eigenvalues that satisfy $\det(\tilde{\mathbf{A}} - \lambda\mathbf{I}_{2n}) = \det(\lambda^2\mathbf{I}_n - (\mathbf{I}_n + \mathbf{A}_1^*)\lambda + \mathbf{A}_1^*) = 0$. This last equation has a solution for $\lambda = 1$. Now there are n unit roots for $\det(\mathbf{A}(L)) = 0$

[62] If $\mathbf{A}(L)$ is of order p then $\mathbf{A}^*(L)$ is of order $p-1$ with coefficients $\mathbf{A}_k = -\sum_{j=k+1}^{p} \mathbf{A}_j$.

[63] If $\mathbf{A}^*(L)$ had a unit root then $\mathbf{x}_t$ would be $I(2)$ and we have to take differences twice, we do not consider this case further.

which are also roots of $\det(\tilde{\mathbf{A}}(L)) = 0$. Hence since the eigenvalues of $\tilde{\mathbf{A}}$ are the reciprocal of the roots of $\det(\tilde{\mathbf{A}}(L)) = 0$ we have n eigenvalues equal 1.

As it is seen the impulse response functions (top left block of the inverse companion matrix $\sum_{k=0}^{\infty} \tilde{\mathbf{A}}^k L^k$) are of the type $\mathbf{D}_k = (\mathbf{I}_n + \mathbf{A}_1^*)^k$ hence they never decrease to zero.[64]

2. $\mathbf{A}(1)$ has reduced rank, say $\text{rank}(\mathbf{A}(1)) = r < n$, that is in (51) we have $k = r$, then we still have $\det(\mathbf{A}(1)) = 0$. From (51) we see that there are $n - r$ unit roots which imply $n - r$ unit eigenvalues in the companion matrix. In this case we have cointegration (see the next Section).

11.2 Cointegration

We say that $\mathbf{y}_t$ is cointegrated if there exists at least one vector $\boldsymbol{\beta}$ such that the linear combination $z_t = \boldsymbol{\beta}'\mathbf{y}_t = \beta_1 y_{1t} + \ldots + \beta_n y_{nt}$ is such that $z_t \sim I(0)$. Then we say that $\boldsymbol{\beta}$ is a cointegration vector. There might be more than one vector with this property and then all those vectors form the columns of a cointegration matrix. The number of linearly independent cointegration vectors (the rank of the cointegration matrix) is called cointegration rank. For a vector of dimension n the maximum cointegration rank is $n - 1$.[65] For macroeconomic variables the cointegration relations $\boldsymbol{\beta}'\mathbf{y}_t = \mathbf{0}_r$ give the long run equilibrium conditions.

Example. Take $\mathbf{y}_t = (x_{1t}, x_{2t})'$ such that

$$\begin{aligned} x_{1t} &= x_{1t-1} + e_t \\ x_{2t} &= x_{1t} + u_t \end{aligned}$$

where e_t and u_t are white noise with zero mean and variance one. Then $x_{1t} \sim I(1)$ and $x_{2t} \sim I(1)$ but $z_t = x_{2t} - x_{1t} = u_t \sim I(1)$, therefore the cointegration rank is 1 and the cointegration vector is $(-1, 1)'$.

Now consider a VAR(1)

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{u}_t, \qquad \mathbf{u}_t \sim w.n.(\mathbf{0}, \boldsymbol{\Gamma}_u)$$

which can be rewritten as (define $\boldsymbol{\Pi} = \mathbf{A} - \mathbf{I}_n$)

$$\begin{aligned} \mathbf{x}_t - \mathbf{x}_{t-1} &= (\mathbf{A} - \mathbf{I}_n)\mathbf{x}_{t-1} + \mathbf{u}_t \\ \Delta\mathbf{x}_t &= \boldsymbol{\Pi}\mathbf{x}_{t-1} + \mathbf{u}_t \end{aligned} \tag{52}$$

Notice that $\boldsymbol{\Pi} = -\mathbf{A}(1)$. Now we already considered the case in which $\boldsymbol{\Pi} = \mathbf{0}$, in which case there are n unit roots and we must use the model in first differences because all components of $\mathbf{x}_t$ are $I(1)$ (see previous Section). In particular we see that (52) is an equation of an n-dimensional random walk. If on the other hand $\det(\boldsymbol{\Pi}) \neq 0$ then $\det(\mathbf{A}(1)) \neq 0$ and there are no unit roots, therefore $\mathbf{x}_t \sim I(0)$.[66]

[64] The eigenvalues of $\mathbf{D}_k$ are always greater than 1, this means that the norm of $\mathbf{D}_k$ which is given by its largest eigenvalue is always greater than 1.

[65] If we had n cointegration vectors then we would have a full-rank matrix $\mathbf{B}$ such that $\mathbf{z}_t = \mathbf{B}\mathbf{y}_t \sim I(0)$ but since $\mathbf{B}$ is invertible then also $\mathbf{y}_t = \mathbf{B}^{-1}\mathbf{z}_t \sim I(0)$ which is a contradiction.

[66] In this case $\boldsymbol{\Pi}$ is invertible and (52) is written also as

$$\boldsymbol{\Pi}^{-1}\Delta\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{u}_t$$

and since the left hand side is $I(0)$ so must be the right hand side and in particular $\mathbf{x}_{t-1} \sim I(0)$.

However, we also know that if $\text{rank}(\mathbf{A}(1)) = \text{rank}(\mathbf{\Pi}) = r$ for some $0 < r < n$ we have $n - r$ unit roots in the VAR(1), then we still have $\det(\mathbf{\Pi}) = \det(\mathbf{A} - \mathbf{I}_n) = 0$. In this case $\mathbf{\Pi}$ is not zero but has reduced rank and we can always write $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $n \times r$:

$$\Delta \mathbf{x}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{x}_{t-1} + \mathbf{u}_t$$

Now since the left hand side is stationary then $\mathbf{z}_{t-1} = \boldsymbol{\beta}' \mathbf{x}_{t-1}$ must also be stationary and therefore $\boldsymbol{\beta}$ is the cointegration matrix having as columns the r cointegration vectors, so the cointegration rank is nothing else but the rank of $\mathbf{A}(1)$. Cointegration arises when we have fewer unit roots than processes.

The process $\mathbf{z}_t$ represents the deviation from the cointegration relations defined by the equilibrium $\boldsymbol{\beta}' \mathbf{x}_{t-1} = \mathbf{0}_r$. Then the dynamics of a cointegrated vector is described by two components: the first one is the white noise $\mathbf{u}_t$ (or more in general is random vector of shocks), the other one is the magnitude at $t-1$ of the deviation $\mathbf{z}_{t-1}$ from the long-run equilibrium. The matrix $\boldsymbol{\alpha}$ is called loading of $\mathbf{z}_t$ and describes how the dynamic adjusts to revert to the long run equilibrium. This mechanism of correction given by $\boldsymbol{\alpha}\boldsymbol{\beta}' \mathbf{x}_{t-1}$ is called error correction mechanism (ECM) and (52) is called Vector Error Correction Model (VECM).

Let us generalise the previous result to the VAR(p)

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{u}_t, \qquad \mathbf{u}_t \sim w.n.(\mathbf{0}, \mathbf{\Gamma}_u)$$

then define $\mathbf{B}(L)L = (\mathbf{I}_n - \mathbf{A}(L))$ such that

$$\mathbf{x}_t = \mathbf{B}(L)\mathbf{x}_{t-1} + \mathbf{u}_t$$

By using the Beveridge-Nelson decomposition of $\mathbf{B}(L)$, we have

$$\mathbf{x}_t = [\mathbf{B}(1) + \mathbf{\Gamma}(L)(1-L)]\mathbf{x}_{t-1} + \mathbf{u}_t$$

and by subtracting $\mathbf{x}_{t-1}$ from both sides we get (notice that $\mathbf{B}(1) = \mathbf{I}_n - \mathbf{A}(1)$)

$$\begin{aligned}
\Delta \mathbf{x}_t &= [\mathbf{B}(1) - \mathbf{I}_n]\mathbf{x}_{t-1} + \mathbf{\Gamma}(L)\Delta \mathbf{x}_{t-1} + \mathbf{u}_t \\
&= -\mathbf{A}(1)\mathbf{x}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \Delta \mathbf{x}_{t-i} + \mathbf{u}_t \\
&= \mathbf{\Pi}\mathbf{x}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \Delta \mathbf{x}_{t-i} + \mathbf{u}_t \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \Delta \mathbf{x}_{t-i} + \mathbf{u}_t
\end{aligned} \tag{53}$$

As for the VAR(1) case the cointegration rank is nothing else but the rank of $\mathbf{A}(1)$. Notice that (53) is equivalent to the VAR(p) from which we started our derivation

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{u}_t \tag{54}$$

where $\mathbf{A}_0 = \mathbf{I}$, $\mathbf{A}_1 = (\mathbf{\Gamma}_1 - \boldsymbol{\alpha}\boldsymbol{\beta}' + \mathbf{I})$, $\mathbf{A}_2 = \mathbf{\Gamma}_2 - \mathbf{\Gamma}_1$, and $\mathbf{A}_p = -\mathbf{\Gamma}_{p-1}$.

Now, consider the matrices

$$\mathbf{M}(L) = \begin{pmatrix} \mathbf{I}_{n-r}(1-L) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_r \end{pmatrix}, \qquad \bar{\mathbf{M}}(L) = \begin{pmatrix} \mathbf{I}_{n-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_r(1-L) \end{pmatrix}$$

then $\bar{\mathbf{M}}(L)\mathbf{M}(L) = \mathbf{I}_n(1-L)$. Then, from (51) we can invert the VAR as follows:

$$(1-L)\mathbf{x}_t = \mathbf{V}^{-1}(L)\bar{\mathbf{M}}(L)\mathbf{U}^{-1}(L)\mathbf{u}_t = \mathbf{V}^{-1}(L)\begin{pmatrix} \mathbf{I}_{n-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_r(1-L) \end{pmatrix}\mathbf{U}^{-1}(L)\mathbf{u}_t = \mathbf{C}(L)\mathbf{u}_t$$

From this MA representation we see that $\text{rank}(\mathbf{C}(1)) = n - r < n$. As before, the IRFs are defined by $\mathbf{D}(L)$ such that $(1-L)\mathbf{D}(L) = \mathbf{C}(L)$. Hence, for $n-r$ shocks we have that the IRFs are not summable that is the shocks have a permanent effect. While for the remaining r shocks the IRFs are square summable so the long run effect is zero. This justifies the search for Permanent Transitory decompositions of $\mathbf{x}_t$ when there is cointegration. Notice also that if $\mathbf{x}_t$ were stationary (no unit roots) then $r = n$ and $\text{rank}(\mathbf{C}(1)) = 0$ and square summable IRFs equal to $\mathbf{D}(L) = \mathbf{V}^{-1}(L)\mathbf{U}^{-1}(L)$. On the other hand if we had n unit roots then $r = 0$ and $\text{rank}(\mathbf{C}(1)) = n$ and the IRFs $\mathbf{D}(L)$ of all shocks are never square summable (see previous section).

When we have cointegration it means that we have $n-r$ trends which are common to all series. Since $\Delta\mathbf{x}_t$ is stationary, then it admits a Wold decomposition which using the Beveridge Nelson decomposition reads

$$\Delta\mathbf{x}_t = \mathbf{C}(L)\mathbf{u}_t = [\mathbf{C}(1) + (1-L)\mathbf{C}^*(L)]\mathbf{u}_t$$

or equivalently

$$\mathbf{x}_t = \mathbf{C}(1)\sum_{k=0}^{\infty}\mathbf{u}_{t-k} + \mathbf{C}^*(L)\mathbf{u}_t$$

therefore the first component is driven by multivariate random walk which we denote as $\boldsymbol{\mu}_t$. Moreover, under the hypothesis of cointegration we have shown that $\mathbf{C}(1)$ has a reduced rank $n-r$, thus we can always write

$$\mathbf{x}_t = \boldsymbol{\psi}\boldsymbol{\eta}'\boldsymbol{\mu}_t + \mathbf{C}^*(L)\mathbf{u}_t \tag{55}$$

where $\boldsymbol{\psi}$ and $\boldsymbol{\eta}$ are $n \times (n-r)$. This is the common trend representation by Stock and Watson which is nothing else but the multivariate Beveridge Nelson decomposition. It is also called the common trend representation. In particular since $\boldsymbol{\beta}'\mathbf{x}_t \sim I(0)$ it must happen that $\boldsymbol{\beta}'\mathbf{C}(1) = \mathbf{0}$ to achieve stationarity in (55). Moreover, the vector $\boldsymbol{\tau}_t = \boldsymbol{\eta}'\boldsymbol{\mu}_t$ is a random walk of dimension $n-r$, so we can write (55) as

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\psi}\boldsymbol{\tau}_t + \mathbf{C}^*(L)\mathbf{u}_t \\ \boldsymbol{\tau}_t &= \boldsymbol{\tau}_{t-1} + \mathbf{u}_t, \end{aligned} \tag{56}$$

and the process is made of two components: $\boldsymbol{\tau}_t$ which are the so called common trends and are the non-stationary component of $\mathbf{x}_t$ and $\mathbf{C}^*(L)\mathbf{u}_t$ which is stationary by definition.

The Granger Representation Theorem. A cointegrated vector can be expressed in two equivalent ways, corresponding to the autoregressive and moving average representations. The representation derived from a VAR is the VECM in (53), while the representation derived from the moving average is the common trend representation in (55).

A last remark is about identification of the cointegration matrix. If $\boldsymbol{\beta}'\mathbf{x}_t \sim I(0)$ then also $\mathbf{R}\boldsymbol{\beta}'\mathbf{x}_t \sim I(0)$ for any non-singular matrix $\mathbf{R}$. If we call $\mathbf{T} = \mathbf{R}\boldsymbol{\beta}'$ then we can always choose $\mathbf{R}$ such that $\mathbf{T}\mathbf{x}_t$ represents some economic meaningful long-run relations (e.g. money demand, purchasing power parity, long-run consumption or investment as function of GDP). Typically we set one entry of each column of $\boldsymbol{\beta}$ to be one. Or we can also choose $\mathbf{T}$ such that $\mathbf{T}'\mathbf{T} = \mathbf{I}$.

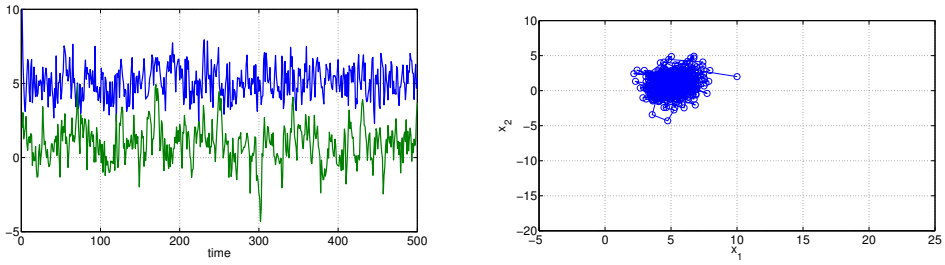


Figure 60: Stationary VAR(1). Left: time series; right: phase diagram (x_{1t}, x_{2t}).

As an example to have an intuition of the meaning of cointegration consider a VAR(1) with $n = 2$ and $\mathrm{E}[\mathbf{x}_t] = (3, 0)$ and with initial condition $(x_{10}, x_{20}) = (10, 2)$. In the case of stationarity, the system quickly converges to its mean and it oscillates around it, this is seen in the phase diagram too (see Figure 60). The consider the case of 2 unit roots, which implies that $\mathbf{x}_t$ is a random walk with mean equal to its initial condition, however the process never converge but keeps fluctuating with very high dispersion around the mean, compare the dispersion of the phase diagram of this case with the previous one. Notice that if in this case we run the regression

$$x_{2t} = \beta_1 x_{1t} + w_t$$

we get $\beta_1 \simeq 4$ (quite different from zero) and we have a spurious regression indeed the acvs of w_t show that these are clearly non-stationary (see Figure 61).

Last consider the case of one unit root, thus of cointegration with rank 1, then the two processes are random walks that co-move (one common trend), in the phase diagram this comovement is around a line given by the cointegration vector $\boldsymbol{\beta}'\mathbf{x}_t = \beta_1 x_{1t} + \beta_2 x_{2t} = 0$. The elements of $\boldsymbol{\beta}$ can be found for example by fixing $\beta_2 = 1$ and by considering the linear regression

$$x_{2t} = \beta_1 x_{1t} + e_t$$

and in this case we get $\beta_1 \simeq 5$. Notice that from the acvs of e_t we see that this is now a stationary process. In this sense we say there is an equilibrium condition (see Figure 62).

11.3 Estimation of cointegrated systems

In terms of estimation the matrices $\mathbf{A}_k$ of the VAR the model in (54) can be consistently estimated by OLS even when the data have unit roots, that is without accounting for the ECM, this is shown by Sims, Stock, and Watson. However, the long run matrix $\mathbf{A}(1)$ would not be estimated consistently and actually it would converge to a random variable (the errors of the OLS estimation cumulate and do not vanish) as proved by Phillips. Indeed unless we add an ECM we cannot get an estimate of $\mathbf{A}(1)$ with reduced rank and as a consequence the long run impulse response would be estimated as having full-rank which is not correct. Therefore, the long run impulse response functions entailed by the VECM cannot be recovered unless we take into account explicitly the ECM, that is unless we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

As for $\boldsymbol{\beta}$ the most common way is to estimate it via Maximum Likelihood as proposed by Johansen (unless $r = 1$ and then we can use OLS). Once an estimator of $\boldsymbol{\beta}$ is given then we can estimate the other VECM parameters by simple OLS. More precisely, if we have the VECM

$$\Delta\mathbf{x}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1} + \mathbf{u}_t$$

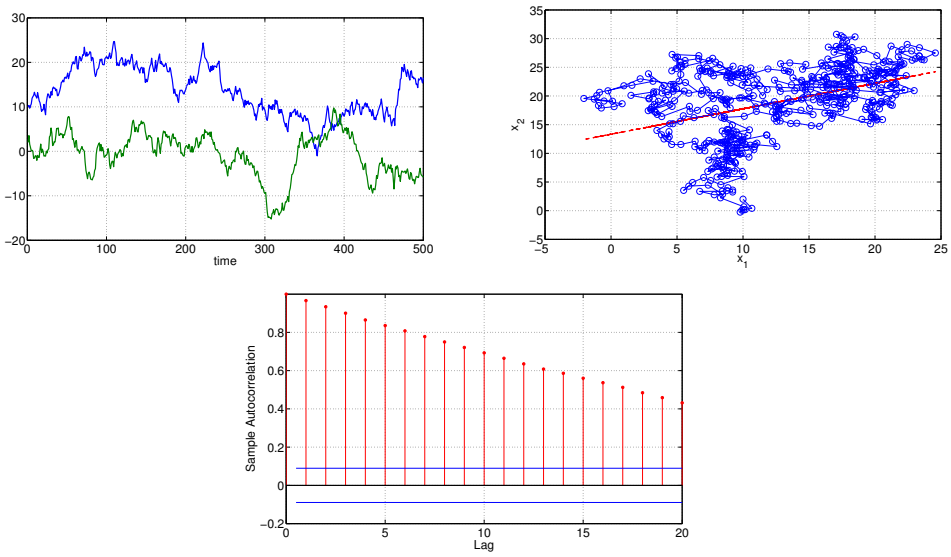


Figure 61: Random walk. Left: time series; right: phase diagram (x_{1t}, x_{2t}) with regression line in red; bottom: acvs of the regression errors.

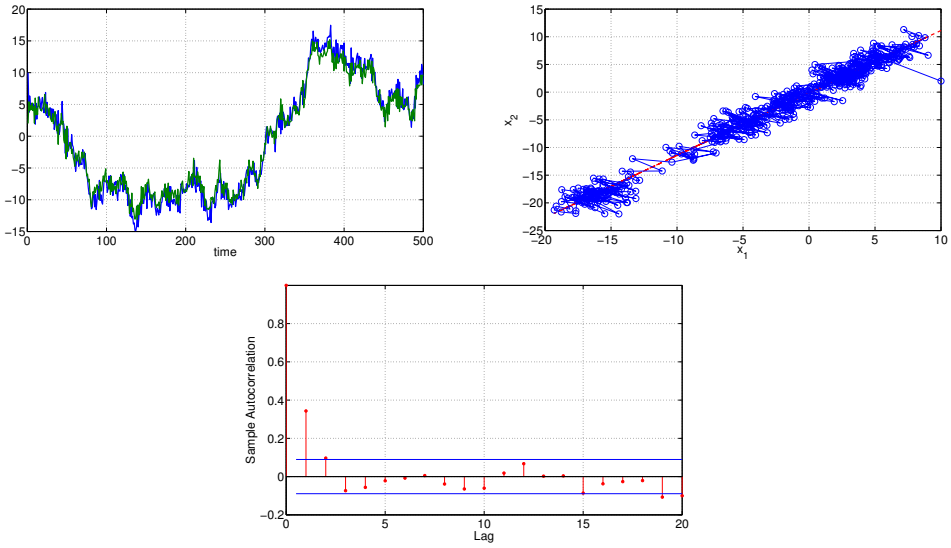


Figure 62: Cointegration. Left: time series; right: phase diagram (x_{1t}, x_{2t}) with regression line with slope β_1 in red; bottom: acvs of the regression errors.

Then, define[67]

$$\mathbf{M}_{11} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \quad \mathbf{M}_{10} = \mathbf{M}_{01}' = \frac{1}{T}\sum_{t=2}^{T} \mathbf{x}_{t-1} \Delta \mathbf{x}_t' \quad \mathbf{M}_{00} = \frac{1}{T}\sum_{t=1}^{T} \Delta \mathbf{x}_t \Delta \mathbf{x}_t'.$$

The cointegration matrix is estimated as

$$(\mathbf{M}_{11} - \mathbf{M}_{10}\mathbf{M}_{00}^{-1}\mathbf{M}_{01})\widehat{\boldsymbol{\beta}}_j = \mu_j \widehat{\boldsymbol{\beta}}_j,$$

[67] If there are lags of $\Delta \mathbf{x}_t$ in the VECM these matrices should become conditional covariances with respect to those lags.

where μ_j are the r-largest eigenvalues of the above matrix, that is the estimated cointegration vectors $\widehat{\boldsymbol{\beta}}_j$ are the corresponding eigenvectors. Consistency of this estimator can be proved and it is with rate T, that is the estimator is super-consistent, thus justifying the second step in estimation of the VECM.

11.4 Permanent and transitory decompositions

The Beveridge Nelson is just one possible PT decomposition. In this section we review the most famous PT decompositions based on the parameter of the VECM. In what follows we adopt the conventional assumption that $\mathbf{u}_s = 0$ for $s \leq 0$ and we allow $\mathbf{x}_t$ to have a non-random initial value, which by our assumptions is $\mathbf{x}_0 = \mathbf{0}$.

The simplest and most used PT decomposition is the common trends representation by Stock and Watson (1988) derived above. Define $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \mathbf{u}_t$, a $n-r$-dimensional random walk. Then

$$\mathbf{x}_t = \boldsymbol{\psi}\left[\boldsymbol{\eta}'\boldsymbol{\mu}_t\right] + \mathbf{C}^*(L)\mathbf{u}_t, \tag{57}$$

where $\boldsymbol{\eta}$ is $n \times n-r$ and $\boldsymbol{\psi}$ is $n \times n-r$ such that $\mathbf{C}(1) = \boldsymbol{\psi}\boldsymbol{\eta}'$, while $\mathbf{C}^*(L) = (1-L)^{-1}(\mathbf{C}(L) - \mathbf{C}(1))$ is a $n \times n$ infinite matrix polynomial with square summable coefficients.[68] As a consequence we must also have that $\boldsymbol{\beta}'\boldsymbol{\psi} = \mathbf{0}_{r \times n-r}$, by definition of cointegration matrix. The first term on the right hand side of (57) is driven by $n-r$ common trends, $\boldsymbol{\eta}'\boldsymbol{\mu}_t$, which are all random walks, while the second term on the right hand side is instead stationary by construction. Note that (57) is the multivariate generalisation of the Beveridge Nelson (1981) decomposition in the case of cointegrated processes. For an autoregressive representation of (57) see Proietti, 1997.

As pointed out by many authors (Lippi and Reichlin, 1994), the main limitation of (57) is that the long-run component is made only of random walks, while permanent shocks might have an effect which is only gradually absorbed into the economy, thus are associated to a more general MA-type dynamics (e.g. technological diffusion). Depending on the assumed identification constraints on such dynamics, different decompositions with orthogonal shocks can be derived from (57) (Lippi and Reichlin, 1994, Gonzalo and Ng, 2001). Any such decomposition can be written starting from a form like

$$\mathbf{x}_t = \boldsymbol{\psi}\left[\boldsymbol{\eta}'\boldsymbol{\mu}_t\right] + \mathbf{C}^*(L)\boldsymbol{\eta}(\boldsymbol{\eta}'\boldsymbol{\eta})^{-1}\left[\boldsymbol{\eta}'\mathbf{u}_t\right] + \mathbf{C}^*(L)\boldsymbol{\eta}_\perp(\boldsymbol{\eta}_\perp'\boldsymbol{\eta}_\perp')^{-1}\left[\boldsymbol{\eta}_\perp'\mathbf{u}_t\right]. \tag{58}$$

Any identifying restriction can then be imposed in (58), by multiplication by suitable invertible matrices.[69] The shocks $\boldsymbol{\eta}'\mathbf{u}_t$ generate the common trends of (57), therefore are called permanent and drive the first and second term on the right hand side of (58). Those two terms define the long-run component, which is more than just a random walk. The third term, driven only by the transitory shocks $\boldsymbol{\eta}_\perp'\mathbf{u}_t$, generates instead the short-run dynamics.

Based on the same concept of permanent and transitory components, Gonzalo and Granger (1995) propose the decomposition[70]

$$\mathbf{x}_t = \boldsymbol{\beta}_\perp(\boldsymbol{\alpha}_\perp'\boldsymbol{\beta}_\perp)^{-1}\left[\boldsymbol{\alpha}_\perp'\mathbf{x}_t\right] + \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\alpha})^{-1}\left[\boldsymbol{\beta}'\mathbf{x}_t\right]. \tag{59}$$

Here, the long-run component, which is the first term on the right hand side of (59), is defined as everything that is not the Error Correction term in the VECM, that is the only shocks that can

[68] If $\mathbf{C}_k$ are the coefficient of $\mathbf{C}(L)$ and $\mathbf{C}_k^*$ those of $\mathbf{C}^*(L)$, then $\mathbf{C}_k^* = -\sum_{j=k+1}^{\infty}\mathbf{C}_j$.

[69] For example, transitory shocks can be identifed by means of an orthogonal $d \times d$ matrix $\mathcal{R}$ such that the identified implies responses are $\mathbf{C}^*(L)\boldsymbol{\eta}_\perp(\boldsymbol{\eta}_\perp'\boldsymbol{\eta}_\perp')^{-1}\mathcal{R}$, while the identified shocks are $\mathcal{R}'\boldsymbol{\eta}_\perp'\mathbf{u}_t$.

[70] This is obtained by inverting $(\boldsymbol{\alpha}_\perp'\ \boldsymbol{\beta}')'\mathbf{x}_t$ in order to get $\mathbf{x}_t$ as the sum of two components.

affect long run forecasts are the permanent ones since for these there is no mean reversion (in the VECM there is mean reversion because of the ECM). As a consequence, the second term is driven only by the cointegration relations. To illustrate the relation between (59) and (57), let us consider the VECM in when $p = 2$. Then, the long-run component of (59) reads

$$\boldsymbol{\alpha}'_{\perp}\mathbf{x}_t - \boldsymbol{\alpha}'_{\perp}\mathbf{x}_{t-1} = \boldsymbol{\alpha}'_{\perp}\mathbf{u}_t + \boldsymbol{\alpha}'_{\perp}\boldsymbol{\Gamma}_1\Delta\mathbf{x}_{t-1}. \tag{60}$$

First, notice that if $p = 1$ the last term on the right hand side of (60) disappears and then the long-run component looks exactly like the one in (57), i.e. it is generated by the $n - r$-dimensional random walk $\boldsymbol{\alpha}'_{\perp}\mathbf{x}_t$. However, when $p = 2$, and in general when $p > 1$, the long-run component, has also a stationary part, and (59) is similar to (58), although in this case the two components are in general not orthogonal.

Finally, an alternative orthogonal decomposition, is proposed by Johansen (1991) and Kasa (1992):

$$\mathbf{x}_t = \boldsymbol{\beta}_{\perp}(\boldsymbol{\beta}'_{\perp}\boldsymbol{\beta}_{\perp})^{-1}\left[\boldsymbol{\beta}'_{\perp}\mathbf{x}_t\right] + \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}\left[\boldsymbol{\beta}'\mathbf{x}_t\right]. \tag{61}$$

Contrary to (59), here firstly the stationary component is defined as the part driven by the cointegration relations only, then the non-stationary term is obtained by simply considering the orthogonal complement. Such a decomposition is not based on any economic assumption, but is based on a purely geometric argument. In particular, it has to be noticed that shocks to the stationary term might have permanent effects on the non-stationary one.

11.5 Cointegration and common factors

A cointegrated system always admits a factor representation, as shown, for example, by Escribano and Peña (1994):

$$\mathbf{x}_t = \boldsymbol{\Psi}_1\boldsymbol{\tau}_t + \boldsymbol{\Psi}_0\boldsymbol{c}_t, \tag{62}$$

where $\boldsymbol{\Psi}_1$ is $n \times n - r$ and $\boldsymbol{\Psi}_0$ is $n \times r$, so that $\boldsymbol{\tau}_t$ is $n - r \times 1$ and $\boldsymbol{c}_t$ is $r \times 1$. Model (62) is such that $\boldsymbol{\tau}_t$ has $\tau_{jt} \sim I(1)$ for any $j = 1, \ldots n - r$, while $\boldsymbol{c}_t \sim I(0)$. A PT decomposition can then be obtained by specifying $\boldsymbol{\Psi}_1$, $\boldsymbol{\tau}_t$, $\boldsymbol{\Psi}_0$, and $\boldsymbol{c}_t$. The most common choices are based on the VECM representation and where outlined in the previous section.

A non-parametric PT decomposition instead makes use only of unconditional second moments. Peña and Poncela (1997, 2006) show that the space spanned by the columns of $\boldsymbol{\Psi}_1$ coincides with the space spanned by the eigenvectors corresponding to the $n - r$ non-zero eigenvalues of the $n \times n$ random matrix $\mathbf{C}(1)\left(\int_0^1 \mathbf{W}(u)\mathbf{W}(u)'\mathrm{d}u\right)\mathbf{C}(1)'$, where $\mathbf{W}(\cdot)$ is the $n - r$-dimensional Brownian motion.[71] The matrix $\boldsymbol{\Psi}_0$ is then given by the remaining r eigenvectors. The PT decomposition obtained in this way has the following properties.

1. It is based just on the long-run second moments of $\mathbf{x}_t$, hence it is completely agnostic in terms of economic assumptions and does not require to specify and estimate a VECM.

2. Since $(\boldsymbol{\Psi}_1\boldsymbol{\Psi}_0)$ are an orthonormal basis, the two components in (62) are orthogonal and $\boldsymbol{\Psi}_0 = \boldsymbol{\Psi}_{1\perp}$.

3. From orthonormality of eigenvectors we have

$$\boldsymbol{\Psi}'_1\mathbf{x}_t = \boldsymbol{\Psi}'_1\boldsymbol{\Psi}_1\boldsymbol{\tau}_t = \boldsymbol{\tau}_t, \qquad \boldsymbol{\Psi}'_0\mathbf{x}_t = \boldsymbol{\Psi}'_0\boldsymbol{\Psi}_0\boldsymbol{c}_t = \boldsymbol{c}_t. \tag{63}$$

[71] This matrix is estimated by means of $\frac{1}{T^2}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}'_t$.

hence the PT reads

$$\mathbf{x}_t = \mathbf{\Psi}_1 \mathbf{\Psi}_1' \mathbf{x}_t + \mathbf{\Psi}_0 \mathbf{\Psi}_0' \mathbf{x}_t$$

4. The $n - r$ non-stationary processes $\boldsymbol{\tau}_t$ are driven by $n - r$ common trends in the sense of Stock and Watson (1988), but $\boldsymbol{\tau}_t$ are not necessarily pure random walks. They generate the trend or permanent component of $\mathbf{x}_t$.

5. The stationary processes $\boldsymbol{c}_t$ are r common cycles in the sense of Vahid and Engle (1993). They generate the cycle or transitory component of $\mathbf{x}_t$.

6. The cointegration matrix is given by the columns of $\mathbf{\Psi}_0$.

7. $\mathbf{x}_t$ is driven by $n-r$ common factors which are $I(1)$ and r common factors which are $I(0)$.

Such a geometric approach is similar to the one in (61) and as said above cannot be used if the goal is identifying permanent and transitory shocks.

12 Unobserved component models and signal extraction INCOMPLETE

The PT decompositions are based on the idea of a latent process driving the dynamics of observed data. This is a particular case of the general problem of extracting a signal from a noisy observations. Assume to observe $\mathbf{x}_t$ which is n dimensional and assume that it is driven by r unknown, hence latent, signals (or states or factors) which we call $\mathbf{F}_t$. If $\mathbf{x}_t$ is observed with noise then on general we have

$$\mathbf{x}_t = \mathbf{\Lambda}_t(\mathbf{F}_t, \mathbf{e}_t), \qquad \mathbf{e}_t \sim wn(\mathbf{0}, \mathbf{R}) \tag{64}$$

where $\mathbf{\Lambda}_t$ is a generic time varying possibly non-linear function. We also assume that the factors at time $t = 0$ are observed (possibly with a covariance matrix $\mathbf{P}_0$), while the successive observations are contaminated by a noise process:

$$\mathbf{F}_t = \mathbf{A}_t(\mathbf{F}_{t-1}, \mathbf{u}_t, \mathbf{P}_0), \qquad \mathbf{u}_t \sim wn(\mathbf{0}, \mathbf{Q}). \tag{65}$$

The system (64)-(65) is a state-space model where (64) is called measurement equation and (65) is called state equation.

12.1 Linear time invariant state space models

We now make the assumptions that the function in (64)-(65) are linear and time invariant and the noise is Gaussian,

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim wn\, N(\mathbf{0}, \mathbf{R}), \tag{66}$$

$$\mathbf{F}_t = \mathbf{A}\mathbf{F}_{t-1} + \mathbf{H}\mathbf{u}_t, \qquad \mathbf{u}_t \sim wn\, N(\mathbf{0}, \mathbf{Q}). \tag{67}$$

we also assume $\mathrm{E}[e_{it}u_{js}] = \mathbf{0}$ for any i, j, s, t. Notice that in general the number of shocks $\mathbf{u}_t$ can be different from the number of states $\mathbf{F}_t$. This model is very flexible and can accommodate many models seen up to this point.

Examples of stationary ARMA models in state-space form.

1. AR(2). $x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$ with $w_t \sim wnN(0, \sigma^2)$, this is equivalent to

$$x_t = \delta^* + \underbrace{[1 \;\; 0]}_{\mathbf{\Lambda}} \mathbf{F}_t$$

$$\mathbf{F}_t = \begin{pmatrix} x_t - \delta^* \\ x_{t-1} - \delta^* \end{pmatrix} = \underbrace{\begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} x_{t-1} - \delta^* \\ x_{t-2} - \delta^* \end{pmatrix}}_{\mathbf{F}_{t-1}} + \underbrace{\begin{pmatrix} w_t \\ 0 \end{pmatrix}}_{\mathbf{u}_t}, \qquad \mathbf{Q} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix}$$

where $\delta^* = \frac{\delta}{1-\phi_1-\phi_2}$ and therefore $\mathbf{R} = \mathbf{0}_2$ and $\mathbf{H} = \mathbf{I}_2$.

2. MA(1). $x_t = w_t + \theta w_{t-1}$ with $w_t \sim wnN(0, \sigma^2)$, this is equivalent to

$$x_t = \underbrace{[1 \;\; 0]}_{\mathbf{\Lambda}} \mathbf{F}_t$$

$$\mathbf{F}_t = \begin{pmatrix} \theta w_{t-1} + w_t \\ \theta w_t \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} \theta w_{t-2} + w_{t-1} \\ \theta w_{t-1} \end{pmatrix}}_{\mathbf{F}_{t-1}} + \underbrace{\begin{pmatrix} 1 \\ \theta \end{pmatrix}}_{\mathbf{H}} w_t$$

and therefore $\mathbf{R} = \mathbf{0}_2$ and $\mathbf{Q} = \sigma^2$. Or alternatively we can write

$$x_t = \theta F_t + w_t$$
$$F_t = w_{t-1}$$

which has the advantage of having a smaller state vector but the disadvantage of having the errors of both equations correlated.

3. ARMA(2,1). $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t + \theta w_{t-1}$ with $w_t \sim wnN(0, \sigma^2)$, this is equivalent to

$$x_t = \underbrace{[1 \;\; \theta]}_{\mathbf{\Lambda}} \mathbf{F}_t$$

$$\mathbf{F}_t = \underbrace{\begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}}_{\mathbf{A}} \mathbf{F}_{t-1} + \underbrace{\begin{pmatrix} w_t \\ 0 \end{pmatrix}}_{\mathbf{u}_t}, \qquad \mathbf{Q} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix}$$

and therefore $\mathbf{R} = \mathbf{0}_2$ and $\mathbf{H} = \mathbf{I}_2$. Note that the state equations give

$$F_{1t} = (1 - \phi_1 L - \phi_2 L^2)^{-1} w_t$$
$$F_{2t} = F_{1t-1},$$

while the observation equation gives

$$x_t = (1 + \theta L) F_{1t} = (1 + \theta L)(1 - \phi_1 L - \phi_2 L^2)^{-1} w_t$$

which is an ARMA(2,1). This is Hamilton (1994) formulation. Alternatively, following Harvey (1989) we can write

$$x_t = \underbrace{[1 \;\; 0]}_{\mathbf{\Lambda}} \mathbf{F}_t$$

$$\mathbf{F}_t = \underbrace{\begin{pmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{pmatrix}}_{\mathbf{A}} \mathbf{F}_{t-1} + \underbrace{\begin{pmatrix} 1 \\ \theta \end{pmatrix}}_{\mathbf{H}} w_t \tag{68}$$

and therefore $\mathbf{R} = \mathbf{0}_2$ and $\mathbf{Q} = \sigma^2$. Note that the state equations now give

$$\begin{aligned} F_{2t} &= \phi_2 F_{1t-1} + \theta w_t \\ F_{1t} &= \phi_1 F_{1t-1} + F_{2t-1} + w_t \\ &= \phi_1 F_{1t-1} + \phi_2 F_{1t-2} + \theta w_{t-1} + w_t \end{aligned}$$

and the observation equation gives $x_t = F_{1t}$ which is an ARMA(2,1).

In general any Gaussian ARMA or VARMA model can be written in state-space form although such form is in general not unique. However, the state-space formulation allows also for models with more structure than ARMA where we can impose some a priori assumed structure on the dynamics of the system. These are the structural time series models.

Examples of structural time series models.

1. Trend cycle. Let $y_t = y_{1t} + y_{2t}$ with $y_{1t} = \delta + y_{1t-1} + e_{1t}$ being the trend component (a random walk) and $y_{2t} = \phi_1 y_{2t-1} + \phi_2 y_{2t-2} + e_{2t}$ being the cycle component (a stationary AR(2)). Let $e_{it} \sim wnN(0, \sigma_i^2)$ for $i = 1, 2$ and $\mathrm{E}[e_{1t} e_{2s}] = 0$ for any t, s. There are three ways of writing this model in state-space form:

   (a) only y_{2t} is an unobserved state, then

$$\begin{aligned} \Delta y_t &= \delta + \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{pmatrix} y_{2t} \\ y_{2t-1} \end{pmatrix} + e_{1t} \\ \begin{pmatrix} y_{2t} \\ y_{2t-1} \end{pmatrix} &= \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_{2t-1} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} e_{2t} \\ 0 \end{pmatrix}; \end{aligned}$$

   (b) only y_{2t} is an unobserved state, then

$$\begin{aligned} y_t &= \begin{bmatrix} 1 & -\phi_1 & -\phi_2 \end{bmatrix} \begin{pmatrix} y_{1t} \\ y_{1t-1} \\ y_{1t-2} \end{pmatrix} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_{2t} \\ \begin{pmatrix} y_{1t} \\ y_{1t-1} \\ y_{1t-2} \end{pmatrix} &= \begin{pmatrix} \delta \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{1t-2} \\ y_{1t-3} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ 0 \\ 0 \end{pmatrix}; \end{aligned}$$

   (c) both y_{1t} and y_{2t} are unobserved states, then

$$\begin{aligned} y_t &= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \\ y_{2t-1} \end{pmatrix} \\ \begin{pmatrix} y_{1t} \\ y_{2t} \\ y_{2t-1} \end{pmatrix} &= \begin{pmatrix} \delta \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & \phi_1 & \phi_2 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ e_{2t} \\ 0 \end{pmatrix}. \end{aligned}$$

2. Local linear trend.

3. Trigonometric cycle.

4. Trend plus trigonometric cycle.

5. Cointegrated systems. If $\mathbf{x}_t \sim I(1)$ then the first equation is a PT decomposition and $\mathbf{F}_t$ are the common trends (see previous chapter) where under the assumption that rank$(\mathbf{\Lambda}) = r$, the cointegration matrix is given by $\mathbf{\Lambda}_\perp$ such that $\mathbf{\Lambda}_\perp \mathbf{\Lambda} = \mathbf{0}_{n-r\times r}$. For convenience of notation in this chapter we change the role of r thus we have $n-r$ cointegration relations and r common trends.

6. Also seasonal components can be modeled using (66)-(67).

12.2 Forward filter - Kalman filter

Assume that all parameters of the model $(\mathbf{\Lambda}, \mathbf{R}, \mathbf{A}, \mathbf{H}, \mathbf{Q})$ are known. Then the signals can be extracted from the data using the so called Kalman filter. We consider two ways for deriving the algorithm that characterises the Kalman filter.

First, assume that $\mathbf{F}_0$ is observed then (assume for simplicity $\mathbf{H} = \mathbf{I}_r$)

$$\mathbf{F}_1 = \mathbf{A}\mathbf{F}_0 + \mathbf{H}\mathbf{u}_1, \qquad \mathbf{u}_1 \sim wn(\mathbf{0}, \mathbf{Q})$$

Then, since $\mathbf{u}_1$ is a white noise the distribution of $\mathbf{F}_1$ is the same as the one of $\mathbf{u}_1$ with mean $\mathbf{A}\mathbf{F}_0$ and covariance $\mathbf{H}\mathbf{Q}\mathbf{H}'$. The best guess (one-step ahead prediciton) for $\mathbf{F}_1$ is its mean:

$$\mathbf{F}_{1|0} = \mathbf{A}\mathbf{F}_0 \tag{69}$$

Moreover, given the observed value $\mathbf{x}_1$ we also have

$$\mathbf{x}_1 = \mathbf{\Lambda}\mathbf{F}_1 + \mathbf{e}_1, \qquad \mathbf{e}_1 \sim wn(\mathbf{0}, \mathbf{R}). \tag{70}$$

and by combining (69) and (70), we have

$$\mathbf{x}_1 - \mathbf{\Lambda}\mathbf{F}_{1|0} = \mathbf{v}_{1|0} \tag{71}$$

where $\mathbf{v}_{1|0}$ is the one-step-ahead prediction error.

At time $t = 1$ we know both $\mathbf{F}_{1|0}$ and $\mathbf{x}_1$, which follow the model

$$\begin{bmatrix} \mathbf{A}\mathbf{F}_0 \\ \mathbf{x}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{\Lambda} \end{bmatrix} \mathbf{F}_1 + \begin{bmatrix} -\mathbf{u}_1 \\ \mathbf{e}_1 \end{bmatrix}$$

The unknown is here $\mathbf{F}_1$ and can be estimated by GLS.[72] We define the GLS estimator of $\mathbf{F}_1$ at time $t = 1$ given information up to time $t = 1$ as $\mathbf{F}_{1|1}$ hence

$$\begin{aligned}
\mathbf{F}_{1|1} &= \left\{ \begin{bmatrix} \mathbf{I}_r & \mathbf{\Lambda}' \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}_r \\ \mathbf{\Lambda} \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{I}_r & \mathbf{\Lambda}' \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}\mathbf{F}_0 \\ \mathbf{x}_1 \end{bmatrix} \\
&= \left\{ \mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \left\{ \mathbf{Q}^{-1}\mathbf{A}\mathbf{F}_0 + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{x}_1 \right\} \\
&= \left\{ \mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{Q}^{-1}\mathbf{A}\mathbf{F}_0 + \left\{ \mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{x}_1
\end{aligned} \tag{72}$$

which is a weighted average of the two known vectors $\mathbf{F}_0$ and $\mathbf{x}_1$.[73] Moreover, using (71) we have

$$\begin{aligned}
\mathbf{F}_{1|1} &= \left\{ \mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{Q}^{-1}\mathbf{A}\mathbf{F}_0 + \left\{ \mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{\Lambda}'\mathbf{R}^{-1} \left\{ \mathbf{\Lambda}\mathbf{A}\mathbf{F}_0 + \mathbf{v}_{1|0} \right\} \\
&= \mathbf{A}\mathbf{F}_0 + \left\{ \mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{v}_{1|0}
\end{aligned} \tag{73}$$

[72] Given the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim iid(\mathbf{0}, \mathbf{G})$ then the GLS estimator is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y}$.

[73] If we had $\mathbf{H}$ such that $\mathbf{H}\mathbf{H}' = \mathbf{I}$ then the formula would be

$$\mathbf{F}_{1|1} = \left\{ \mathbf{H}'\mathbf{Q}^{-1}\mathbf{H} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{H}'\mathbf{Q}^{-1}\mathbf{H}\mathbf{A}\mathbf{F}_0 + \left\{ \mathbf{H}'\mathbf{Q}^{-1}\mathbf{H} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda} \right\}^{-1} \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{x}_1.$$

and the second term on the right hand side is the correction due to the prediction error when predicting $\mathbf{F}_1$ at time $t = 0$. Notice that Gaussianity is not necessary for deriving this formulas. In general, GLS is always consistent but under Gaussianity we know that GLS is also efficient. By iterating at $t \geq 1$ we obtain the whole path $\mathbf{F}_{t|t}$. This can be done provided we take into account that at each step we also have uncertainty in the estimate of $\mathbf{F}_t$, that is in general the covariance of $\mathbf{F}_{t|t}$ is not zero and must be updated at each point in time. We do not do this using the above approach but we turn to a different approach.

The general iterations of the Kalman filter allowing also for a random initial condition are easily derived under Gaussianity. Assume that at $t = 0$ we have

$$\mathbf{F}_0 \sim N(\mathbf{F}_{0|0}, \mathbf{P}_{0|0}).$$

where $\mathbf{F}_{0|0}$ and $\mathbf{P}_{0|0}$ are known (notice that the previous case we had $\mathbf{P}_{0|0} = \mathbf{0}$, thus $\mathbf{F}_0 = \mathbf{F}_{0|0}$). Then at $t = 1$ we have

$$\mathbf{F}_1 = \mathbf{A}\mathbf{F}_0 + \mathbf{H}\mathbf{u}_1, \qquad u_1 \sim wnN(\mathbf{0}, \mathbf{Q})$$

Then since $\mathbf{u}_1$ is white noise and Gaussian then is also independent of $\mathbf{F}_0$ and the sum of two independent Gaussians is also Gaussian:

$$\mathbf{F}_1 \sim N(\mathbf{A}\mathbf{F}_{0|0}, \mathbf{A}\mathbf{P}_{0|0}\mathbf{A}' + \mathbf{H}\mathbf{Q}\mathbf{H}')$$

We denote the one-step-ahead prediction of $\mathbf{F}_1$ as $\mathbf{F}_{1|0}$ and we define it as its mean

$$\mathbf{F}_{1|0} = \mathbf{A}\mathbf{F}_{0|0}. \tag{74}$$

Then, the associated one-step-ahead mean squared error is the variance of $\mathbf{F}_1$

$$\mathbf{P}_{1|0} = \mathrm{E}[(\mathbf{F}_1 - \mathbf{F}_{1|0})^2] = \mathbf{A}\mathbf{P}_{0|0}\mathbf{A}' + \mathbf{H}\mathbf{Q}\mathbf{H}'. \tag{75}$$

Thus,

$$\mathbf{F}_1 \sim N(\mathbf{F}_{1|0}, \mathbf{P}_{1|0}).$$

Now assume to observe $\mathbf{x}_1$ and we look for the distribution of $\mathbf{F}_1$ given $\mathbf{x}_1$. We have

$$\begin{aligned}
\mathbf{F}_1 &= \mathbf{F}_{1|0} + (\mathbf{F}_1 - \mathbf{F}_{1|0}) \\
\mathbf{x}_1 &= \mathbf{\Lambda}\mathbf{F}_1 + \mathbf{e}_1, \qquad e_1 \sim wnN(\mathbf{0}, \mathbf{R}) \\
&= \mathbf{\Lambda}\mathbf{F}_{1|0} + \mathbf{\Lambda}(\mathbf{F}_1 - \mathbf{F}_{1|0}) + \mathbf{e}_1
\end{aligned}$$

Then, since $\mathbf{u}_1$ and $\mathbf{e}_1$ are independent because of Gaussianity, we have the joint Gaussian distribution[74]

$$\begin{bmatrix} \mathbf{F}_1 \\ \mathbf{x}_1 \end{bmatrix} = N\left(\begin{bmatrix} \mathbf{F}_{1|0} \\ \mathbf{\Lambda}\mathbf{F}_{1|0} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{1|0} & \mathbf{P}_{1|0}\mathbf{\Lambda}' \\ \mathbf{\Lambda}\mathbf{P}_{1|0} & \mathbf{\Lambda}\mathbf{P}_{1|0}\mathbf{\Lambda}' + \mathbf{R} \end{bmatrix} \right) \tag{76}$$

Then, the conditional mean and variance of $\mathbf{F}_1$ given information up to time $t = 1$ are given by[75]

$$\begin{aligned}
\mathbf{F}_{1|1} = \mathrm{E}[\mathbf{F}_1|\mathbf{x}_1] &= \mathbf{F}_{1|0} + \mathbf{P}_{1|0}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{1|0}\mathbf{\Lambda}' + \mathbf{R})^{-1}(\mathbf{x}_1 - \mathbf{\Lambda}\mathbf{F}_{1|0}) \\
&= \mathbf{F}_{1|0} + \mathbf{P}_{1|0}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{1|0}\mathbf{\Lambda}' + \mathbf{R})^{-1}\mathbf{v}_{1|0} \qquad (77) \\
\mathbf{P}_{1|1} = \mathrm{Cov}(\mathbf{F}_1|\mathbf{x}_1) &= \mathbf{P}_{1|0} - \mathbf{P}_{1|0}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{1|0}\mathbf{\Lambda}' + \mathbf{R})^{-1}\mathbf{\Lambda}\mathbf{P}_{1|0}. \qquad (78)
\end{aligned}$$

[74] Notice that $\mathrm{E}[(\mathbf{F}_1 - \mathbf{F}_{1|0})] = \mathbf{0}$ and also $\mathrm{E}[(\mathbf{F}_1 - \mathbf{F}_{1|0})\mathbf{F}_{1|0}] = \mathbf{0}$ because $\mathbf{F}_{1|0}$ is known.

[75] Given $(\mathbf{X}, \mathbf{Y})$ jointly normal we have

$$\mathrm{E}[\mathbf{X}|\mathbf{Y}] = \mathrm{E}[\mathbf{X}] + \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}]), \qquad \mathrm{Cov}(\mathbf{X}|\mathbf{Y}) = \mathbf{\Sigma}_{XX} - \mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}$$

By generalising (74), (75), (77), and (78) to any $t \geq 1$ we have

$$\mathbf{F}_{t|t-1} = \mathbf{A}\mathbf{F}_{t-1|t-1} \tag{79}$$
$$\mathbf{P}_{t|t-1} = \mathbf{A}\mathbf{P}_{t-1|t-1}\mathbf{A}' + \mathbf{H}\mathbf{Q}\mathbf{H}' \tag{80}$$

$$\mathbf{F}_{t|t} = \mathbf{F}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{t|t-1}\mathbf{\Lambda}' + \mathbf{R})^{-1}(\mathbf{x}_t - \mathbf{\Lambda}\mathbf{F}_{t|t-1}) \tag{81}$$
$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{t|t-1}\mathbf{\Lambda}' + \mathbf{R})^{-1}\mathbf{\Lambda}\mathbf{P}_{t|t-1} \tag{82}$$

The equations (79) and (80) are called prediction equations and the equations (81) and (82) are called update equation and together they form the Kalman filter. Notice that the algorithm is optimal since under Gaussianity estimates the signal as the conditional mean

$$\mathbf{F}_{t|t} = \mathrm{E}[\mathbf{F}_t|\mathbf{x}_t]$$

which is the best predictor of $\mathbf{F}_t$ given information up to time t. The associated mean-squared-error

$$\mathbf{P}_{t|t} = \mathrm{E}[(\mathbf{F}_t - \mathbf{F}_{t|t})(\mathbf{F}_t - \mathbf{F}_{t|t})']$$

which is therefore minimised by definition of conditional mean.

The prediction and update equations can also be written as a unique recursion which is useful for forecasting

$$\begin{aligned}\mathbf{F}_{t+1|t} &= \mathbf{A}\mathbf{F}_{t|t} \\ &= \mathbf{A}\mathbf{F}_{t|t-1} + \mathbf{A}\mathbf{P}_{t|t-1}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{t|t-1}\mathbf{\Lambda}' + \mathbf{R})^{-1}(\mathbf{x}_t - \mathbf{\Lambda}\mathbf{F}_{t|t-1}) \\ &= \mathbf{A}\mathbf{F}_{t|t-1} + \mathbf{K}_t\mathbf{v}_{t|t-1}\end{aligned}$$

where $\mathbf{K}_t = \mathbf{A}\mathbf{P}_{t|t-1}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{t|t-1}\mathbf{\Lambda}' + \mathbf{R})^{-1}$ is the Kalman gain and $\mathbf{v}_{t|t-1} = \mathbf{x}_t - \mathbf{\Lambda}\mathbf{F}_{t|t-1}$ is the one-step-ahead prediction error of $\mathbf{x}_t$ given information up to time $t-1$. We also have the one-step-ahead mean squared prediction error as

$$\begin{aligned}\mathbf{P}_{t+1|t} &= \mathbf{A}\mathbf{P}_{t|t}\mathbf{A}' + \mathbf{H}\mathbf{Q}\mathbf{H}' \\ &= \mathbf{A}\left(\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{P}_{t|t-1}\mathbf{\Lambda}' + \mathbf{R})^{-1}\mathbf{\Lambda}\mathbf{P}_{t|t-1}\right)\mathbf{A}' + \mathbf{H}\mathbf{Q}\mathbf{H}'\end{aligned}$$

which is called Riccati difference equation.

Notice that (77) is equivalent to (73). Indeed when $\mathbf{P}_{0|0} = 0$ and $\mathbf{H} = \mathbf{I}$, we have $\mathbf{F}_{0|0} = \mathbf{F}_0$ and $\mathbf{P}_{1|0} = \mathbf{Q}$ and (73) and (77) give respectively

$$\mathbf{F}_{1|1} = \mathbf{A}\mathbf{F}_0 + \left\{\mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda}\right\}^{-1}\mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{v}_{1|0} \tag{83}$$
$$\mathbf{F}_{1|1} = \mathbf{A}\mathbf{F}_0 + \mathbf{Q}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{Q}\mathbf{\Lambda}' + \mathbf{R})^{-1}\mathbf{v}_{1|0} \tag{84}$$

Now because of the Woodbury formula when $r < n$ (fewer signals than variables observed) then

$$\mathbf{Q}\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{Q}\mathbf{\Lambda}' + \mathbf{R})^{-1} = (\mathbf{Q}^{-1} + \mathbf{\Lambda}'\mathbf{R}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}'\mathbf{R}^{-1}$$

and (83) and (84) are equivalent. Notice however that if $\mathbf{H} \neq \mathbf{I}$ or there are more signals than variables then only the left hand side makes sense. Indeed in this case the GLS derivation cannot be done.

12.3 Backward filter - Kalman smoother

Once all the history of $\mathbf{x}_t$ is available then we could obtain a better estimate of the latent signals by considering the conditional mean when conditioning on $\mathbf{x}_1 \dots \mathbf{x}_T$ defined as

$$\mathbf{F}_{t|T} = \mathrm{E}[\mathbf{F}_t | \mathbf{x}_1 \dots \mathbf{x}_T] \tag{85}$$

by definition this estimator minimises the mean-squared-error

$$\mathbf{P}_{t|T} = \mathrm{E}[(\mathbf{F}_t - \mathbf{F}_{t|T})(\mathbf{F}_t - \mathbf{F}_{t|T})'].$$

Because $\mathbf{F}_{t|T}$ is the conditional mean and therefore minimises the mean-squared-error, then we must have

$$\mathrm{E}[(\mathbf{F}_t - \mathbf{F}_{t|T})\mathbf{x}_s'] = \mathbf{0}_{r\times n}, \quad s = 1 \dots T$$

which implies

$$\mathrm{E}[\mathbf{F}_t \mathbf{x}_s'] = \mathrm{E}[\mathbf{F}_{t|T}\mathbf{x}_s']. \tag{86}$$

Now, under Gaussianity the conditional mean is linear therefore can be written as a weighted average of all available observations

$$\mathbf{F}_{t|T} = \mathbf{H}(L)\mathbf{x}_t = \sum_{k=t-1}^{T} \mathbf{H}_k \mathbf{x}_{t-k} = \sum_{\tau=1}^{T} \mathbf{H}_{t-\tau}\mathbf{x}_\tau \tag{87}$$

notice that this is a two-sided filter. By substituting (87) into (86) we have

$$\mathrm{E}[\mathbf{F}_t \mathbf{x}_s'] = \sum_{k=t-1}^{T} \mathbf{H}_{t-\tau}\mathrm{E}[\mathbf{x}_\tau \mathbf{x}_s']$$

This is the Weiner-Hopf equation. Moreover using (66) we can write

$$\mathrm{E}[\mathbf{F}_t \mathbf{F}_s']\mathbf{\Lambda}' = \sum_{k=t-1}^{T} \mathbf{H}_{t-\tau}\mathrm{E}[\mathbf{x}_\tau \mathbf{x}_s'] \tag{88}$$

where we used the fact that since $\mathrm{E}[u_{it}e_{js}] = 0$ for any i, j, s, t, then $\mathrm{E}[\mathbf{F}_t \mathbf{e}_s'] = \mathbf{0}_{r\times n}$.

Now if $\mathbf{x}_t \sim I(0)$ we can take the Fourier transform of both sides of (88) and we have

$$\mathbf{S}_F(\theta)\mathbf{\Lambda}' = \mathbf{H}(e^{-i\theta})\mathbf{S}_x(\theta)$$

and by solving for the filter we have

$$\begin{aligned} \mathbf{H}(e^{-i\theta}) &= \mathbf{S}_F(\theta)\mathbf{\Lambda}'\mathbf{S}_x^{-1}(\theta) \\ &= \mathbf{S}_F(\theta)\mathbf{\Lambda}'(\mathbf{\Lambda}\mathbf{S}_F(\theta)\mathbf{\Lambda}' + \mathbf{S}_e(\theta))^{-1} \end{aligned} \tag{89}$$

which defines the Wiener-Kolmogorov filter.

A Complex numbers

Definition

The complex numbers are an extension of the real numbers containing all roots of quadratic equations. If we define i to be a solution of the equation $x = \sqrt{-1}$, then the set $\mathbb{C}$ of complex numbers is represented in standard form as

$$\mathbb{C} = \{a + bi | a, b \in \mathbb{R}\}.$$

The complex number i is called "imaginary unit". We often use the variable $z = a + bi$ to represent a complex number. The number a is called the real part of z ($\Re z$ or Rez) while b is called the imaginary part of z ($\Im z$ or Imz). Two complex numbers are equal if and only if their real parts are equal and their imaginary parts are equal.

From the definition of $\mathbb{C}$ we see that there is a one-to-one correspondence between the complex set $\mathbb{C}$ and the set $\mathbb{R}^2 = \mathbb{R}\times\mathbb{R}$. We represent complex numbers graphically by associating $z = a+bi$ with the point (a, b) on the plane $\mathbb{R}^2$. This plane is also called the complex plane or the Argand Plane. On the Argand Plane the horizontal axis is called the real axis and the vertical axis is called the imaginary axis.

Basic operations

First notice that by definition $i = \sqrt{-1}$, so for any $c > 0$ we have $\sqrt{-c} = i\sqrt{c}$. The number i has no real part, $\Re i = 0$. Moreover, $i^2 = -1$, $i^3 = -i$, $i^4 = 1$ and so on...

The sum and difference of complex numbers is defined by adding or subtracting their real components i.e.:

$$(a + bi) + (c + di) = (a + c) + (b + d)i,$$
$$(a + bi) - (c + di) = (a - c) + (b - d)i.$$

The communitive and distributive properties hold for the product of complex numbers i.e.:

$$(a + bi)(c + di) = ac + adi + bci + bdi^2 = (ac - bd) + (bc + ad)i.$$

Conjugates and absolute values

If we have a complex number defined as $z = a + bi$ then the conjuate would be $\bar{z} = a - bi$. The geometric interpretation of a complex conjugate is the reflection along the real axis. Properties of conjugates:

1. $\bar{\bar{z}} = z$;
2. $\overline{z + w} = \bar{z} + \bar{w}$;
3. $\overline{zw} = \bar{z}\bar{w}$;
4. $\overline{z^n} = \bar{z}^n$;
5. if $z \neq 0$, then $\overline{w/z} = \bar{w}/\bar{z}$;

6. $z \in \mathbb{R}$ if and only if $\bar{z} = z$.

The distance from the origin of the plane to any complex number is the absolute value or modulus. Pythagoras' Theorem gives us a formula to calculate the absolute value of a complex number $z = a + bi$

$$|z| = |a + bi| = \sqrt{a^2 + b^2},$$

moreover

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

Thus by multiplying a complex number times its conjugate we always get a real number. Properties of absolute values:

1. $|z| = 0$ if and only if $z = 0$, i.e. $\Re z = \Im z = 0$;
2. $|z| = |\bar{z}|$;
3. $|zw| = |z|\ |w|$;
4. if $z \neq 0$, then $|1/z| = 1/|z|$;
5. $|z + w| \leq |z| + |w|$.

Fundamental Theorem of Algebra

Every polynomial function equation

$$f(x) = a_k x^k + a_{k-1}x^{k-1} + \ldots a_1 x + a_0 = 0, \qquad a_k \neq 0, \quad k \geq 1, \tag{90}$$

has at least one complex root. Keep in mind that complex numbers include real numbers. Moreover, in polynomial function equations, non-real complex roots always occur in conjugate pairs. In other words, if a complex number with an imaginary part is a root of a polynomial function equation, then its conjugate is also a root of that same function. That is, if for $z \in \mathbb{C}$

$$f(z) = 0 \Leftrightarrow f(\bar{z}) = 0.$$

Notice that if $z \in \mathbb{R}$ then the above is still trivially true since for real numbers $z = \bar{z}$.

The Linear Factorization Theorem Consider (90), then we can always write

$$f(x) = a_k(x - c_1)(x - c_2)\ldots(x - c_k)$$

where $c_1, c_2, \ldots, c_k$ are complex numbers (possibly real and not necessarily distinct). In other words, a polynomial function of degree k, where $k > 0$, can be factored into k (not necessarily distinct) linear factors over the complex number field.

Example: find the roots of $x^4 - 8x^2 - 33 = 0$. There must be four complex roots. We have

$$\begin{aligned} x^4 - 8x^2 - 33 &= (x^2 - 11)(x^2 + 3) \\ &= (x + \sqrt{11})(x - \sqrt{11})(x^2 + 3) \end{aligned}$$

This gives us two real roots equal to $\pm\sqrt{11}$. Then we must find the roots of $x^2 + 3 = 0$. We can either compute them using the rule for quadratic equations, which gives the two roots

$$x_{1,2} = \frac{0 \pm \sqrt{-4 \cdot 3}}{2} = \pm\sqrt{-3} = \pm i\sqrt{3},$$

or use the rule seen above $(a+bi)(a-bi) = a^2+b^2$ where here $a = x$ and $b = \sqrt{3}$, then

$$x^2 + 3 = (x - i\sqrt{3})(x + i\sqrt{3}).$$

So the four roots we are looking for are the two real $\pm\sqrt{11}$ and the two complex conjugates $\pm i\sqrt{3}$.

Example: find the roots of $z^3 = 1$ or equivalently $z^3 - 1 = 0$. There must be three complex roots. A first one is the real number 1. Then we have

$$z^3 - 1 = (z-1)(z^2 + z + 1)$$

then the other two roots are given by $z^2 + z + 1 = 0$, thus they are

$$z_{1,2} = \frac{-1 \pm \sqrt{1-4}}{2} = \frac{-1 \pm i\sqrt{3}}{2} = -\frac{1}{2} \pm i0.866$$

(notice that these are complex conjugates so they come in pairs).

Polar form

Another useful way to write complex numbers is by using polar coordinates. Then, if $z = a + bi$ we have

$$a = r\cos\theta, \qquad b = r\sin\theta, \quad r \geq 0, \quad \theta \in [0, 2\pi].$$

Where r is the absolute value or modulus of z and θ is called argument of z, it's the angle between the real axis and the vector linking the origin to the point z in the complex plane. Indeed, we have

$$r = \sqrt{a^2+b^2} = |z|, \qquad \tan\theta = \frac{b}{a}.$$

Then, we can write, using Euler's formula ($e^{i\theta} = \cos\theta + i\sin\theta$)

$$z = r\cos\theta + ir\sin\theta = re^{i\theta}.$$

So for example we have

$$\begin{aligned}
e^{i\pi} &= \cos\pi + i\sin\pi = -1 \\
3e^{i\pi/2} &= 3\cos\pi/2 + i3\sin\pi/2 = 3i \\
(1+i)^8 &= \{[\sqrt{1^2+1^2}]\ \exp[i\tan^{-1}(1/1)]\}^8 \\
&= \{\sqrt{2}\ \exp[i\pi/4]\}^8 \\
&= 16\exp[i2\pi] \\
&= 16\cos(2\pi) + 16i\sin(2\pi) = 16
\end{aligned}$$

Roots to unity

The polar form can be useful for solving equations as $z^n = 1$. The solutions have all modulus 1, so $r = 1$ and we can write $z = e^{i\theta}$. Then,

$$\begin{aligned}
z^n &= 1 \\
e^{i\theta n} &= 1 \\
e^{i\theta n} &= e^{i2\pi k}, \quad k = 0, \pm 1, \pm 2 \ldots \\
\theta &= \frac{2\pi k}{n}
\end{aligned}$$

So for $k = 0, 1, 2, \ldots n$ there are n distinct roots evenly distributed on the unit circle and then they repeat.

For $n = 1$ we have $z = e^{i\theta} = e^{i2\pi k} = 1$.
For $n = 2$ we have $z^2 = e^{2i\theta} = e^{i2\pi k} = 1$ thus $z = e^{i\pi k} = \cos(\pi k) + i\sin(\pi k)$ so when $k = 0$ we have $z = 1$ and when $k = 1$ we have $z = -1$.
For $n = 3$ we have $z^3 = e^{3i\theta} = e^{i2\pi k} = 1$ thus $z = e^{i2\pi k/3} = \cos(2\pi k/3) + i\sin(2\pi k/3)$ so when $k = 0$ we have $z = 1$, when $k = 1$ we have $z = -1/2 + i0.866$, when $k = 2$ we have $z = -1/2 - i0.866$ (cf. with the result above).
For $n = 4$ we have $z^4 = e^{4i\theta} = e^{i2\pi k} = 1$ thus $z = e^{i\pi k/2} = \cos(\pi k/2) + i\sin(\pi k/2)$ so when $k = 0$ we have $z = 1$, when $k = 1$ we have $z = i$, when $k = 2$ we have $z = -1$, when $k = 3$ we have $z = -i$.

B Matrix algebra

Properties of matrices

- rectangular matrices $\mathbf{A}$ of dimension $m \times n$, with rows $\mathbf{A}_{i:}$ and columns $\mathbf{A}_{:j}$;
- the product between matrices is rows times columns, thus if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times k$, then $\mathbf{AB}$ is $m \times k$ and
$$(\mathbf{AB})_{il} = \sum_{j=1}^{n} a_{ij} b_{jl},$$
or
$$\mathbf{AB}_{i:} = \mathbf{A}_{i:}\mathbf{B}, \qquad \mathbf{AB}_{:j} = \mathbf{AB}_{:j};$$
- the product is non-commutative $\mathbf{AB} \neq \mathbf{BA}$;
- squared matrices $\mathbf{A}$ of dimension $n \times n$:
  - identity matrix
$$\mathbf{I}_n = \begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \end{pmatrix};$$
  - diagonal matrix $\mathbf{A} = \text{diag}(a_{11}, \ldots, a_{nn})$ if
$$\mathbf{A} = \begin{pmatrix} a_{11} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & a_{nn} \end{pmatrix};$$
  - transposed matrix $\mathbf{A}^T$ such that $(\mathbf{A}^T)_{ij} = (\mathbf{A})_{ji}$;
- vectors are particular matrices: $\mathbf{a}$ is a column vector of $\mathbb{R}^n$ with dimension $n \times 1$, $\mathbf{a}^T$ (transposed) is a row vector of dimension $1 \times n$, both have components $(a_1, \ldots a_n)$;
- squared length of a vector (norm)
$$||\mathbf{a}||^2 = \sum_{i=1}^{n} a_i^2 = \mathbf{a}^T\mathbf{a};$$

- scalar product
$$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a} = \sum_{i=1}^{n} a_i b_i;$$
- if two vectors are orthogonal, $\mathbf{a} \perp \mathbf{b}$ then $\mathbf{a}^T\mathbf{b} = 0$;
- outer product between two vectors $\mathbf{a}$ of dimension $m \times 1$ and $\mathbf{b}$ of dimension $n \times 1$, then $\mathbf{a}\mathbf{b}^T$ is a matrix $m \times n$;
- symmetric matrix such that $\mathbf{A}0\mathbf{A}^T$;
- for any matrix $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are squared symmetric matrices;
- properties of transposed matrices
$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T, \qquad (\mathbf{A}\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T;$$
- determinant of a squared matrix $\mathbf{A}$ of dimension n: if $n = 2$,
$$\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21},$$
in general
$$\det \mathbf{A} = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} M_{i,j},$$
where $M_{i,j}$ is the determinant of the matrix that results from $\mathbf{A}$ by removing the i-th row and the j-th column, we call $(-1)^{i+j}M_{i,j}$ cofactor;
- properties of the determinant:
  - if $\mathbf{B}$ results from $\mathbf{A}$ by interchanging two rows or two columns, then $\det \mathbf{B} = -\det \mathbf{A}$,
  - if $\mathbf{B}$ results from $\mathbf{A}$ by multiplying one row or column with a number c, then $\det \mathbf{B} = c \det \mathbf{A}$,
  - if two rows or two columns of $\mathbf{A}$ are equal, then $\det \mathbf{A} = 0$,
  - if a row or column of $\mathbf{A}$ is a linear combination of two rows or columns of the same matrix, then $\det \mathbf{A} = 0$,
  - $\det(\mathbf{A} + \mathbf{B}) \neq \det \mathbf{A} + \det \mathbf{B}$,
  - $\det(c\mathbf{A}) = c^n \det \mathbf{A}$,
  - $\det \mathbf{A}^T = \det \mathbf{A}$,
  - if $\mathbf{A}$ is triangular or diagonal, then
$$\det \mathbf{A} = \prod_{i=1}^{n} a_{ii},$$
  - $\det(\mathbf{A}\mathbf{B}) = \det \mathbf{A} \det \mathbf{B}$ (Binet's theorem);
- inverse of a squared matrix is $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$;
- alternative necessary and sufficient conditions for the existence of the inverse matrix:
  - $\det \mathbf{A} \neq 0$ so that $\det \mathbf{A}^{-1} = (\det \mathbf{A})^{-1}$,

  - the columns of $\mathbf{A}$ are linearly independent thus form a basis of $\mathbb{R}^n$,
  - the linear system $\mathbf{Ax} = \mathbf{b}$ has a unique solution,
  - the linear system $\mathbf{Ax} = \mathbf{0}$ has the unique trivial solution $\mathbf{x} = \mathbf{0}$;
- singular matrix if $\det \mathbf{A} = 0$;
- adjoint matrix is the matrix $\mathbf{A}_{adj}$ such that

$$(\mathbf{A}_{adj})_{ij} = (-1)^{i+j} M_{j,i},$$

  is the transposed of the matrix of cofactors;
- the inverse of a matrix $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \frac{\mathbf{A}_{adj}}{\det \mathbf{A}};$$

- properties of the inverse matrix:
  - the inverse matrix is unique,
  - $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$,
  - $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$,
  - if $\mathbf{A} = \text{diag}(a_{11}, \ldots, a_{nn})$, then $\mathbf{A}^{-1} = \text{diag}(a_{11}^{-1}, \ldots, a_{nn}^{-1})$;
- rank of a matrix is the maximum number of linearly independent rows or columns;
- if $\mathbf{A}$ is $n \times k$ with $k < n$, then rank$\mathbf{A} \leq k$ and if rank$\mathbf{A} = k$ then $\mathbf{A}$ is full-rank;
- a non-singular squared matrix has full-rank;
- properties of the rank:
  - rank$(\mathbf{A}^T\mathbf{A}) =$ rank$(\mathbf{AA}^T) =$ rank$\mathbf{A}$,
  - rank$(\mathbf{AB}) \leq$ rank$\mathbf{A}$ and rank$(\mathbf{AB}) \leq$ rank$\mathbf{B}$,
  - if $\mathbf{B}$ is a squared non-singular matrix then rank$(\mathbf{AB}) =$ rank$\mathbf{A}$,
- trace of a squared matrix $\mathbf{A}$ is the sum of the diagonal elements

$$\text{trace}\mathbf{A} = \sum_{i=1}^{n} a_{ii},$$

  and trace$\mathbf{AB} =$ trace$\mathbf{BA}$;
- orthogonal matrix (or rotation in $\mathbb{R}^n$) $\mathbf{Q}$ is such that $\mathbf{QQ}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$ or equivalently $\mathbf{Q}^T = \mathbf{Q}^{-1}$ and $\det \mathbf{Q} = \pm 1$;
- idempotent matrix is such that $\mathbf{A}^2 = \mathbf{A}$, trivial examples are $\mathbf{0}$, $\mathbf{I}_n$;
- other idempotent matrices:
  - projector on the space spanned by the columns of $\mathbf{X}$ defined as

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T;$$

    provided that $\mathbf{X}$ is full-rank so that $\mathbf{X}^T\mathbf{X}$ exists,

  - projector on the space orthogonal to the space spanned by the columns of $\mathbf{X}$ defined as
$$\mathbf{M}_{\mathbf{X}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}};$$
  - deviation from the average defined as
$$\mathbf{A} = \mathbf{I}_n - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^T}{n},$$
    where $\boldsymbol{\iota} = (1, \ldots, 1)^T$, such that
$$\mathbf{A}\mathbf{y} = \mathbf{y} - \frac{1}{n}\sum_{i=1}^{n} y_i;$$

- quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$, for a $n \times 1$ vector $\mathbf{x}$ and a squared matrix $\mathbf{A}$ of dimension n;
- $\mathbf{A}$ is positive definite if
$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \qquad \forall\, \mathbf{x} \in \mathbb{R}^n,$$
  is positive semidefinite if
$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0, \qquad \forall\, \mathbf{x} \in \mathbb{R}^n;$$
- properties of positive definite matrices:
  - given two squared matrices $\mathbf{A}$ and $\mathbf{B}$ both of dimension n we say that $\mathbf{A} > \mathbf{B}$ if $(\mathbf{A} - \mathbf{B})$ is positive semidefinite,
  - the inverse of a positive definite matrix is positive definite,
  - if $\mathbf{P}$ is $m \times n$ with $n < m$ and rank$\mathbf{P} = n$, then $\mathbf{P}^T\mathbf{P}$ is positive definite,
  - if $\mathbf{A}$ is positive definite and symmetric, then $\mathbf{P}^T\mathbf{A}\mathbf{P}$ is positive definite with $\mathbf{P}$ defined above,
  - in particular, if $\mathbf{P}$ is squared and non-singular, then $\mathbf{P}^{\mathbf{P}}$ and $\mathbf{P}\mathbf{P}^T$ are positive definite,
  - if $\mathbf{A}$ is symmetric and positive definite, then it always exists at least one squared matrix $\mathbf{P}$ with full-rank such that $\mathbf{A} = \mathbf{P}^T\mathbf{P}$;

Eigenvalues and eigenvectors

Given a squared matrix $\mathbf{A}$ of dimension n such that
$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$
for some $n \times 1$ vector $\mathbf{v}$ and some scalar λ, we call $\mathbf{v}$ eigenvector and λ the associated eigenvalue.

We find the eigenvalues by solving the characteristic equation
$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0,$$
and the solution can be real or complex (couples of complex conjugate roots).

Once an eigenvalue is determined (call it λ^*) the corresponding eigenvector is the solution of the linear system
$$(\mathbf{A} - \lambda^*\mathbf{I}_n)\mathbf{v} = 0$$
which is non-trivial because the determinant of the matrix is by definition null.

Properties of eigenvalues and eigenvectors:

- to real eigenvalues correspond real eigenvectors,
- eigenvectors are defined up to a scale (length),
- eigenvectors of multiple eigenvalues are defined also up to a rotation (i.e. an orthogonal matrix);

Properties of a squared symmetric matrix $\mathbf{A}$ of dimension n:

- has all n eigenvalues real, i.e. $\lambda \in \mathbb{R}$,
- has n orthogonal eigenvectors, of we rescale them in such a way that $||\mathbf{v}|| = 1$, then we have n orthonormal eigenvectors, i.e. for two eigenvectors $\mathbf{v}_i$ and $mathbfv_j$ we have

$$\mathbf{v}_i^T \mathbf{v}_j = \delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases},$$

- the orthonormal eigenvectors can be collected in the column of a squared orthogonal matrix $\mathbf{Q}$, also the rows of $\mathbf{Q}$ are orthonormal,
- the matrix of eigenvectors $\mathbf{Q}$ is such that

$$\mathbf{Q}^T \mathbf{A} \mathbf{Q} = \mathbf{L} = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

  where λ_i are the eigenvalues of $\mathbf{A}$,
- determinant and trace

$$\det \mathbf{A} = \prod_{i=1}^{n} \lambda_i, \qquad \text{trace}\mathbf{A} = \sum_{i=1}^{n} \lambda_i,$$

- the rank of a squared symmetric matrix is the number of non-zero eigenvalues,
- the matrix $\mathbf{A}^2 = \mathbf{A}\mathbf{A}$ has eigenvalues λ_i^2 but has the same eigenvectors,
- if $\mathbf{A}$ is non-singular $\mathbf{A}^{-1}$ has eigenvalues λ_i^{-1} but has the same eigenvectors,
- if $\mathbf{A}$ is idempotent, then $\lambda_i = 0$ or 1 and $\text{rank}\mathbf{A} = \text{trace}\mathbf{A}$,
- if $\mathbf{A}$ is positive definite, then $\lambda_i > 0$ for any i, indeed $\mathbf{Q}^T \mathbf{A} \mathbf{Q} = \mathbf{L}$, thus, for any i,

$$\lambda_i = \mathbf{v}_i^T \mathbf{A} \mathbf{v}_i > 0,$$

  since $\mathbf{A}$ is positive definite.

Derivatives

1. **Gradient** $\nabla_{\mathbf{x}}$: the vector of first derivatives of a scalar function with respect to the vector of variables.

   Define a function $f : \mathbb{R}^n \to \mathbb{R}$, and a vector $\mathbf{x} \in \mathbb{R}^n$, then

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

Example: the derivative of the scalar product:

$$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \begin{pmatrix} \frac{\partial(\mathbf{x}^T\mathbf{y})}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial(\mathbf{x}^T\mathbf{y})}{\partial x_n}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{y}.$$

and analogously $\nabla_{\mathbf{y}}(\mathbf{x}^T\mathbf{y}) = \mathbf{x}$.

Example: the derivative of the quadratic form:

$$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}.$$

2. **Jacobian $\mathbf{J}_{\mathbf{x}}$**: the matrix of first derivatives of a vector valued function with respect to the vector of variables.

   Define a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, and a vector $\mathbf{x} \in \mathbb{R}^n$, then

$$\mathbf{J}_{\mathbf{x}}\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

   Example: given a matrix $\mathbf{B}$ of dimension $m \times n$

$$\mathbf{J}_{\mathbf{x}}(\mathbf{B}\mathbf{x}) = \mathbf{B},$$

   matrices are linear vector valued functions.

3. **Hessian $\mathbf{H}_{\mathbf{x}}$**: the matrix of second derivatives of a scalar function with respect to the vector of variables.

   Define a function $f : \mathbb{R}^n \to \mathbb{R}$, and a vector $\mathbf{x} \in \mathbb{R}^n$, then

$$\mathbf{H}_{\mathbf{x}}\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{x}) \end{pmatrix}.$$

   Example: second derivative of the quadratic form

$$\mathbf{H}_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{A} + \mathbf{A}^T.$$

4. the maximum $\mathbf{x}_0$ of a scalar function f is such that

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0) = 0, \qquad \mathbf{v}^T\mathbf{H}_{\mathbf{x}} f(\mathbf{x})\mathbf{v} < 0, \ \forall\, \mathbf{v} \in \mathbb{R}^n,$$

   the minimum $\mathbf{x}_0$ of a scalar function f is such that

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0) = 0, \qquad \mathbf{v}^T\mathbf{H}_{\mathbf{x}} f(\mathbf{x}_0)\mathbf{v} > 0, \ \forall\, \mathbf{v} \in \mathbb{R}^n.$$