

# Applied Time Series Analysis — Part I

Robert M. Kunst  
University of Vienna  
and  
Institute for Advanced Studies Vienna

September 28, 2011

## 1 Introduction and overview

### 1.1 What is ‘econometric time-series analysis’?

Time-series analysis is a field of statistics. There are, however, indications that a specifically econometric time-series analysis is developing. Good examples are the books by HAMILTON (1994) and by LÜTKEPOHL (2005). This development reflects the typical features of economics data as well as a certain tradition of approaches that is specific to the economic science.

Economic textbooks often maintain that economics is a non-experimental science, as it typically considers observational data sets that cannot be generated by repeated experiments, i.e. samples cannot be extended. If samples are seen as results of experiments, this means that the experiments of economics are non-repeatable. Compared with experimental sciences, economics therefore lacks an important tool of model evaluation. However, economics is not the only non-experimental science. Most examples shown in time-series texts are non-experimental, or at least the repetition of experiments would incur high costs (for example, sunspots, accidental deaths, glacier river water).

Typical macroeconomic time series are short and come in annual, quarterly, or monthly measurements. Many variables show characteristic growth trends. The search for an adequate treatment of such *trends* has inspired the development of time-series techniques that are tuned to trending data, such as *unit roots* and *co-integration*. Some of these methods are almost exclusively applied to economics data.

In finance, time-series analysis plays an important role. Typical financial time series come at high frequencies—for example, daily or intra-daily—and have large sample size. This permits the application of methods that emphasize the identification of higher moments, of conditional heteroskedasticity (ARCH), nonlinear features, and statistical distributions. Some of these methods are also applied in technical sciences, while they are not well suited to macroeconomic variables. These points concern the nature of the data. Additionally, each science has its own tradition in its perception of data information.

Firstly, it is a tradition of economics to make up for a lack of data information by an increased emphasis on theory. In other words, economics tends to acknowledge the primacy of theory. This attitude implies that the role of data information is generally secondary and that the typical model is theory-driven rather than data-driven. Also, economists and statisticians use different concepts of a *model*. An economic model is a ‘global’ concept that targets an explanation of all observed effects inclusive of the motivation of economic agents. Such a global concept is seen as a valid model even when it fails to specify all functional relations. By contrast, a statistical (and also a time-series analytical) model is a statistical, dynamic, parametric mechanism for the generation of data, without any need to depict or justify the motivation for agents’ action and reaction.

Second, economics is typically convinced that all structures are ever-changing and never time-constant. For this reason, some economists even reject the usage of any data analysis as a sensible means of reaching substance-matter conclusions. If one followed this radical route, time-series analysis would become useless for economics. However, any assumption of time constancy in a time-series model is to be seen as a working hypothesis only. Time constancy of the statistical model does not necessarily preclude non-constancy of structures in the economic model.

## 1.2 What is time-series analysis in general?

Time-series analysis investigates sequences of observations on a variable that have been measured at regularly recurring time points. It aims at finding dynamic regularities that may enable forecasting future observations (beyond the sample end) or even controlling the variable (compare the title of the book by BOX & JENKINS). Sometimes, by comparing time-series structures and theoretical models, time-series analysis helps in establishing evidence on theories. In the jargon of time-series analysis, such a structure is called a ‘data-generating process’ (DGP). The correct interpretation of this term is difficult. It is not guaranteed and it even appears unlikely that any simple

stochastic process exists that indeed *has generated* the data. If such a process exists, then it must be a very complex object, far beyond our available means. On the other hand, one may imagine myriads of processes that *could have generated* a finite sequence of numbers. Thus, that would be a very weak concept. It is maybe preferable to view the DGP as a process, which is *conceivable* as an actual data-generating mechanism, which describes essential structures in the observed data, and which can be utilized for prediction, for example. Some authors recommend restricting the usage of the term ‘DGP’ to situations, where data are really simulated according to an experimental design from a known process. On the whole, we will follow this recommendation.

Examples for economic time series are: the Austrian GDP, the Dow-Jones index, the Swedish unemployment rate. Examples for non-economic time series are: the winning team in the Austrian football league (one annual measurement, value is a member of a set of football teams), the air temperature at Hohe Warte Vienna, the wind speed at the south pole, the vote share of a political party in elections for the Vienna municipality (sometimes irregular time intervals due to early elections); daily personal data: start/end of breakfast, number of cigarettes smoked etc.

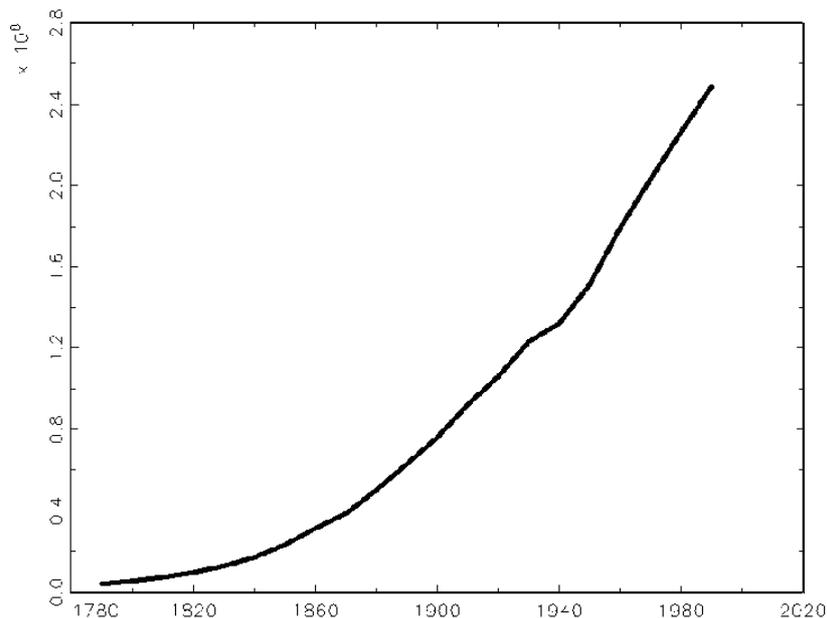


Figure 1: Population of the USA at ten-year intervals 1790–1990 (from BROCKWELL&DAVIS, p.3).

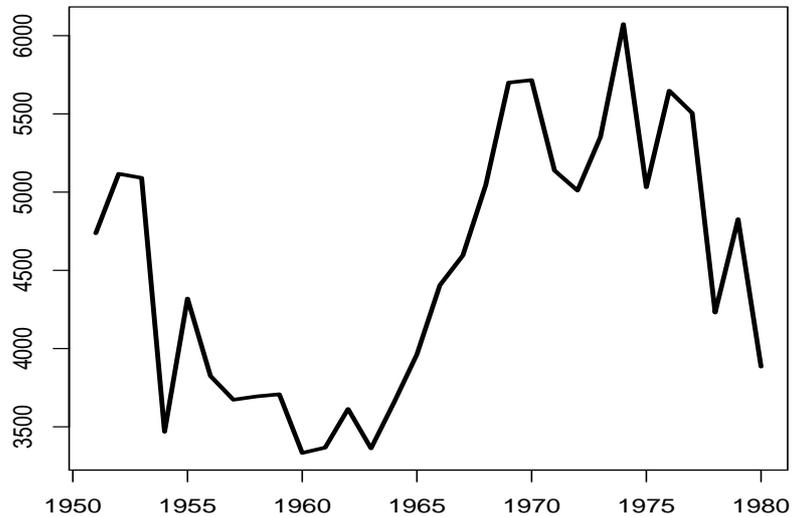


Figure 2: Strikes in the USA, 1951–1980 (from BROCKWELL&DAVIS, p.4).

The textbook by BROCKWELL & DAVIS (1987, B&D; also 2002) provides some typical examples together with their graphic representation. Most of them are non-economic variables. In detail we see: an alternating electrical current measured at a resistor in a pure sine wave; observations on the U.S. population measured at ten-year intervals 1790–1980 (our Figure 1); the number of strikes in the U.S. as annual data 1951–1980 (our Figure 2); the victorious league in an annual contest between the champions of the two major U.S. baseball leagues 1933–1980 with discrete values (either National League or American League); the famous cyclical series of sunspots as annual data 1770–1869, which was compiled by the Swiss astronomers Wolf and Wölfer (our Figure 3); a monthly series on accidental deaths in the USA 1973–1978 (our Figure 4). Figure 7 shows a typical quarterly time series from the Austrian economy, the index of industrial production.

We name three approaches to time-series analysis (*time series analysis*):

1. pre-classical and model-free: fitting of curves to data series, *exponential smoothing* (extrapolation); concept should not be discarded too easily as simplistic, as it was used in the famous studies by Malthus or by Meadows, for example;

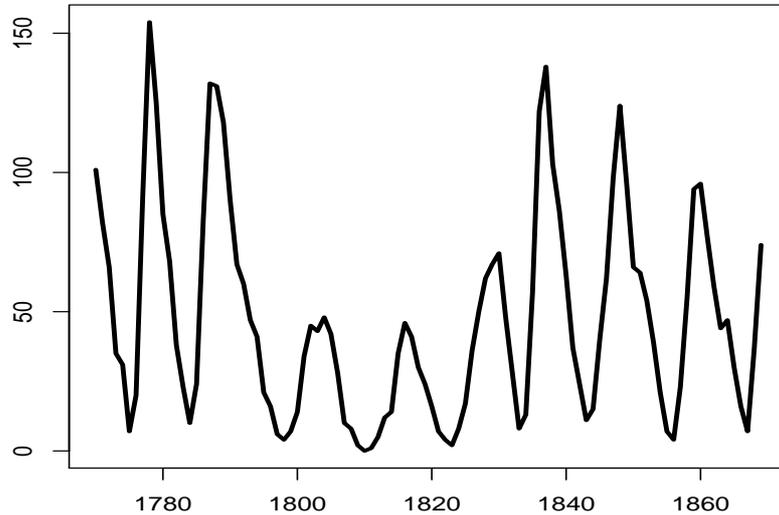


Figure 3: The Wolf&Wölfer sunspot numbers, 1770–1869 (from BROCKWELL&DAVIS, p.6).

2. linear stochastic models, based on the sample autocovariances  $\text{cov}(X_t, X_{t-j})$ : the methods of BOX & JENKINS (1976, re-edited as BOX, JENKINS & REINSEL, 1994); forecasts are model-based extrapolations of the past;
3. focus on cycles and periodicities (e.g. *business cycle*): spectral analysis, Fourier approximation, frequency domain; often this seemingly different approach yields models that are very similar to 2].

In the following, the focus will be on approach 2], which dominates the other paradigms in current academic time-series analysis, as it is represented in its main journal *Journal of Time Series Analysis*. A particularly important recent extension is nonlinear time-series analysis (see GRANGER & TERÄSVIRTA, 1993, PRIESTLEY, 1987, TONG, 1990, FAN & YAO, 2003).

GOURIEROUX AND MONFORT (1997) provide an alternative classification. They distinguish the statistical and time-series analytic approach (they call it *autopredictive*), regression analysis (*explanatory*), and decomposition (*adjustment*) methods. The main characteristic of the first class is its aim to explain the present by the past (of one variable, i.e. *univariate*, or of several

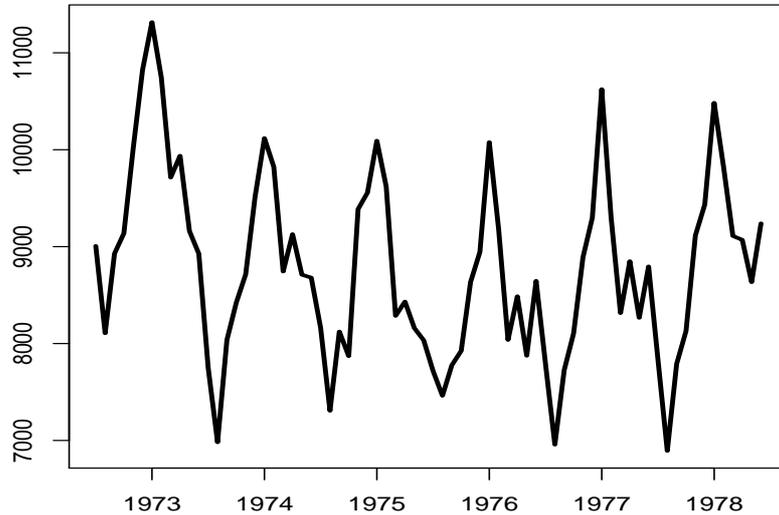


Figure 4: Monthly accidental deaths in the USA, 1973–1978 (from BROCKWELL&DAVIS, p.7).

variables, i.e. *multivariate*). By contrast, regression-analytic procedures use one or more conditioning reference variables, which may be controllable ‘exogenously’ and which influence the dependent variable(s). Then, traditional time-series analysis may be invoked for the exogenous variables. The third class has its separate, very active community (for example, see HARVEY, 1992). Technically, decomposition methods are often equivalent to the first class.

History: ‘The analysis of time series is one of the oldest activities of scientific man’ (WAYNE FULLER, 1996, p.1) [Homo sapiens ssp. scientificus]

### 1.3 Stochastic processes

A (discrete) stochastic process is a sequence of random variables  $(X_1, X_2, \dots)$  that is defined on a common probability space. It becomes a time-series process if the index sequence  $1, 2, \dots$  can be identified with—typically equidistant—time points. The term ‘time series’ denotes a specific realiza-

tion of the process, i.e. a ‘trajectory’<sup>1</sup>, or more commonly a finite segment of a trajectory. If  $(X_t)$  denotes a process, then  $(X_t(\omega))$  for fixed  $\omega$  defines a time series. Hence, a time series is—in simple terms—nothing more than a few (say,  $T$ ) numbers, which can be plotted against the time axis and which can be joined by a curve. By contrast, a process is an object that can be represented symbolically only (for example, by a sequence of density curves).

There is a formal distinction between the finitely indexed time series and the theoretically infinite trajectory. It is a customary abstraction, however, that any stochastic process stretches into the distant future ( $T \rightarrow \infty$ ), even if by definition one may never observe more than a limited time segment of a variable. For example, it appears difficult to concatenate macroeconomic variables such as the German unemployment rate, observed from the end of World War II to the German re-unification, with earlier or later observations. Certainly, however, all observed time series can possibly exist only until the heat death of the universe that is predicted by some physicists.

Some authors insist that a stochastic process is not merely a ‘sequence of random numbers that is written down one by one’. In order to understand the concept of a process, it is of particular importance that all joint distributions of  $n$ -tuples like  $(X_t, X_{t+j}, X_{t+k})$  are always properly defined.

A stochastic process belongs in the realm of time-series analysis, whenever the sequence index can be interpreted as ‘time’. If it has an interpretation as a spatial coordinate, it is a *spatial process*. Even mixed forms of space-time processes have been often considered recently (inspired by observed *panel data*, for example annual data for tourism in Austrian provinces 1960–2000).

If time—at least conceptually—can be condensed such that the process becomes observable continually, then  $(X_t)$  becomes a time-continuous process. This is a sensible abstraction for few economic variables only (for example, some stock-market prices). *Continuous time* is a lively field of research but it is outside of our scope here (for example, see BERGSTROM, 1990). A typical problem is the approximation of a theoretical time-continuous process by an observable time-discrete process and the possible equivalence of properties across the limit operation.

Formally, a stochastic process is defined as measurable mapping from a probability space  $(\Omega, \mathcal{A}, P)$  into a sequence space. Note that the sequence space itself is formally defined as a set of mappings from an index set (for example, natural numbers) into a reasonable codomain (for example, real numbers or a discrete set of characteristics). Anyway, this definition may

---

<sup>1</sup>Statistics differentiates between a random variable (‘the principle of rolling a fair die’) and a specific realization (‘4’). In the case of a process, the realization is a sequence of numbers.

not be immediately accessible to many.

Examples of stochastic processes:

1. rolling a die  $T$  times. The observed  $T$ -sequence (e.g.,  $1, 2, 4, 2, 2, 6, \dots, 3$ ) is a time series. The index set of the sequence is  $\{1, \dots, T\}$ , the probability space is rich enough to contain all events (for example, there are well-defined probabilities that all  $T$  numbers are even or that no 6 appears in the time series). Domain of values (codomain) are the numbers from 1 to 6.
2. tossing a coin  $T$  times. Like 1] but the codomain is (*heads, tails*) or (0,1). In 1] and 2], members of the sequence are independent.
3. numbers generated by a random-number generator. If the generator is known, this is in principle not a stochastic process, as specific sequences have probability one for a given starting value and all thus generated time series are identical. In practice, however, the time series behaves almost like independent random numbers, i.e. like 1] or 2]. The codomain of a generator for normal random numbers is in principle finite (a set of machine numbers with finite precision), although it is usually identified with the real numbers. Usually, the probability space contains all ‘Borel’ sets. This requires that events such as ‘ $2 < x_t < 3.45$  holds exactly 10 times in the sequence’ must have a  $P$ , and also all intersections, unions etc.
4. The problem of the drunk man wandering aimlessly down a street with two bars, who could be seen as taking a step to the right if a tossed coin shows *tails* and a step to the left if it shows *heads*. The drunk’s location can be seen as a stochastic process. Of course, the present location depends on the preceding realization. This is an example of a so-called *random walk* (RW), which for a finite distance to the bars becomes a stopped RW. The codomain are all multiples of step sizes, measured from the point of departure.

In analogy to common practice in econometrics, the observed realization of the time series is viewed as one of (typically infinitely) many virtual realities. This concept is necessary in order to create a statistical framework and to formalize the problem of prediction. An alternative paradigm would be the concept of chaos, i.e. a deterministically generated time series, whose DGP is so complex that it behaves like the trajectory of a stochastic process<sup>2</sup>.

---

<sup>2</sup>A process or a component of a process is called *deterministic* if its trajectory is known (determined) exactly at the time point when it is started. Sometimes, the terms ‘deterministic’ and ‘non-stochastic’ are equated, which is slightly sloppy and misleading.

Actually, such a DGP appears in example 3]. Another paradigm, Bayesian statistics, rejects the idea of a true DGP and views all statistical models as created by continuous interaction of the observer's constructions and the data. Often, Bayesian methods yield results that are similar to traditional ones, while the utilized procedures are considerably more complex.

## 1.4 Stationarity

Even if the stochastic process followed a normal distribution, still  $T$  observations of a time series would require estimating  $T + T(T + 1)/2$  parameters. This is impossible. However, if distributions are assumed as time-constant, the number of free parameters declines palpably to  $T + 1$ . The assumption of further plausible restrictions allows another reduction of the parameter dimension, such that finally efficient estimation becomes feasible. The time-constancy of the statistical distribution (not only of the marginal univariate, but also of all multivariate and hence of all conditional distributions) is called *stationarity*.

Examples 1],2] describe stationary processes, 3] at least conceptually, but not 4]. 1] and 2] are rather special cases, however, due to their independent realizations. An example of a stationary, temporally correlated process is the following:

**Example:** Rolling a die with a bonus. Write down the number of dots on the top face. If it is a 6, the number that comes up at the next roll is doubled, if it is another 6, it is tripled etc. It is easily seen that all multiples of the numbers 1 to 6 can occur. In the univariate marginal distribution, however, the number 18 will be much less probable than 2, for example. Hoping that the marginal distribution does not degenerate (i.e. that it does not 'escape' to  $\infty$ ), our generating law yields a stationary process, which is no more temporally independent.  $\square$

In the example 'rolling a die with a bonus', we of course do not have a normal distribution. The normal distribution has the technical advantage that it is described completely by its first two moments, the mean and the variance (in multivariate applications the mean vector and the variance matrix). We may consider simplifying the work by postulating:

*Postulate A:* The process is stationary and normally distributed

Unfortunately, this is a very restrictive assumption. It is more acceptable to work with a similar concept that also allows restricting attention to the first and second moments. That concept is rooted in the theory of  $L^2$  spaces. By just narrowing our vision to the information of the first two moments, we can build up a powerful theory of time series that admits potential changes of higher moments over time.

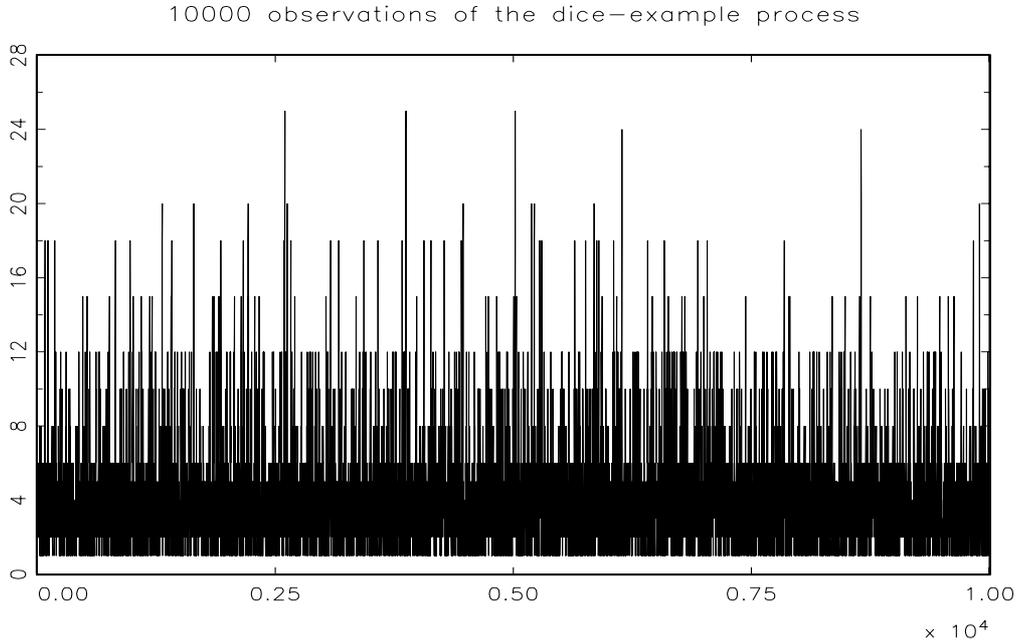


Figure 5: Simulated time series for the example ‘rolling a die with a bonus’.

*Postulate B:* The process has time-constant first and second moments. Postulate B defines the so-called **covariance stationarity** or weak stationarity. For most results and techniques of time-series analysis, this concept suffices. It is so important that we repeat it in formal terms:

$$\begin{aligned}
 EX_t &= \mu \quad \forall t \\
 \text{var}X_t &= \sigma^2 \quad \forall t \\
 \text{cov}(X_t, X_{t-h}) &= \text{cov}(X_s, X_{s-h}) \quad \forall t, s, h
 \end{aligned} \tag{1}$$

An exception to the ubiquitous dominance of covariance stationarity is the analysis of financial markets that requires the modelling of higher moments.

**Example.** ‘alternating nuisance’:  $(X_t)$  is generated by drawing for all even  $t$  from standard Gaussian  $N(0, 1)$  random numbers, for all odd  $t$  from Laplace random numbers scaled to unit variance. The process is not stationary but weakly stationary (mean and variance are constant, all covariances are zero).

**Example.** ‘absurd exception’:  $(X_t)$  is drawn from Cauchy random numbers. The process is certainly stationary but not weakly stationary, as first and second moments do not exist.

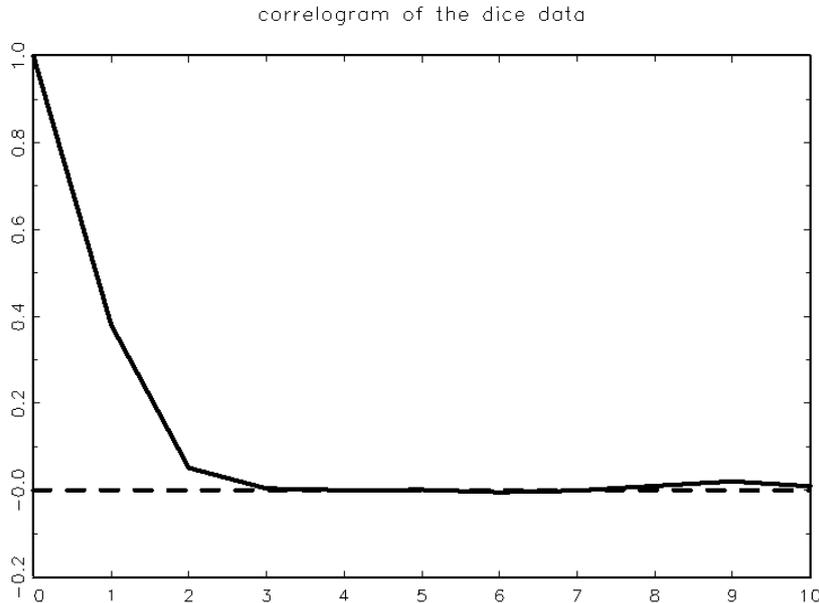


Figure 6: Empirical ACF (correlogram) for the data shown as Figure 5.

Usually, however, stationarity implies weak stationarity. Stationarity *per se* is also called ‘strict stationarity’ to differentiate it from covariance stationarity.

**Summary:** Working with weakly stationary processes can be interpreted as viewing all processes as strictly stationary but restricting focus to the covariance structure; or as viewing actual processes as non-stationary with time-changing higher moments. In a normally distributed world both definitions and interpretations coincide.

**Example:** Sequential rolls of a die or normally distributed (true) random numbers yield processes, whose variance  $\sigma^2$  is constant, whose mean is constant (for normally distributed random numbers typically 0, for the rolled die 3.5), and whose temporal (serial) correlation is zero ( $\text{corr}(X_t, X_{t+i}) = 0 \quad \forall i \neq 0$ ). This very special stationary process, in a sense the most primitive, most elementary and most important of all covariance-stationary processes, is called *white noise* (*wn*). In the following,  $\varepsilon_t$  always denotes a white noise.

**Example:** The definition of a ‘white noise’ determines only the first- and second-moment properties. It is customary to require  $E\varepsilon_t = 0$ . The following law defines a *wn*. Start from an arbitrary  $X_0$ . If we observe  $X_t > 0$ , draw  $X_{t+1}$  from a normal distribution  $N(0, 1)$ . If we observe  $X_t < 0$ , draw  $X_{t+1}$

from a uniform distribution  $U(-1/\sqrt{3}, 1/\sqrt{3})$ . This process is even strictly stationary, while it is not temporally independent. It is easily seen that it is a *wn*. A temporally independent process with time-constant distribution is often indicated by the abbreviation *iid* (*independent identically distributed*).

**Repetition:** *white noise* is a process  $(\varepsilon_t)$  that obeys  $E\varepsilon_t = 0, E\varepsilon_t^2 = \sigma^2$  (positive and finite),  $E\varepsilon_t\varepsilon_{t+i} = 0 \quad \forall i \neq 0$ . Such a process is not necessarily strictly stationary and not necessarily serially independent. *iid* is a process of serially independent draws from the same probability distribution. An *iid*-process is always strictly stationary but it does not necessarily have finite variance. Assumptions regarding the mean or location of random variables (such as  $E\varepsilon_t = 0$ ) are in principle arbitrary.

Examples ‘alternating nuisance’ and ‘absurd exception’ revisited: alternating nuisance is a *wn*-process, certainly not *iid*. Absurd exception is *iid*, though not white noise. Note the subtle differences. Alternating nuisance is *wn* but not strictly stationary, hence it cannot be *iid*. The example with conditional drawings defines a *wn* that is even strictly stationary, whose dependence via higher moments violate the *iid*-property.

In the following, the word ‘stationary’ will always refer to covariance stationarity.

## 1.5 Autocorrelation function ACF

An important tool for the analysis of covariance-stationary processes is the autocorrelation function (ACF) or, alternatively, the autocovariance function (with some authors ACVF). The autocorrelation function is more common. Using the notation  $EX_t = \mu$ , it is defined by

$$\rho(h) = \frac{E\{(X_t - \mu)(X_{t+h} - \mu)\}}{E\{(X_t - \mu)^2\}}, \quad (2)$$

which for  $\mu = 0$  simplifies to

$$\rho(h) = \frac{E(X_t X_{t+h})}{EX_t^2}.$$

The numerator of (2),  $\text{cov}(X_t, X_{t+h})$  as a function of  $h$ , is the autocovariance function and it is commonly denoted by the symbol  $\gamma(h)$ . Because for stationary processes the denominator is constant, the only difference of  $\rho(\cdot)$  and  $\gamma(\cdot)$  is a factor  $\gamma(0)$  that is just the variance of the process.

Note that the formula for  $\rho(h)$  really yields the correlation of  $X_t$  and  $X_{t+h}$ , as the traditional denominator  $\sqrt{E(X_t^2 X_{t+h}^2)}$  is simplified by the assumption of a time-constant variance. Thus, this property is sensitive to deviations

from stationarity. In principle, it makes sense to define the autocovariance function for non-stationary processes by

$$\rho(s, t) = \frac{\mathbb{E}\{(X_s - \mathbb{E}X_s)(X_t - \mathbb{E}X_t)\}}{\sqrt{\mathbb{E}\{(X_s - \mathbb{E}X_s)^2\}\mathbb{E}\{(X_t - \mathbb{E}X_t)^2\}}}$$

as a function of two arguments (e.g., in BROCKWELL & DAVIS), but such generalizations require the correct traditional denominator.

Definition (2) is tuned to stationary processes, as  $\rho(\cdot)$  must not depend on  $t$  in the stationary case. The ACF has several other attractive properties. First we observe:

(FULLER's **Theorem** 1.4.2)  $\rho(-h) = \rho(h)$  and  $\gamma(-h) = \gamma(h)$  for covariance-stationary real-valued processes.

**Remark:** This property does not hold for complex-valued processes. The theorem implies that it is meaningless to evaluate the ACF for negative and positive  $h$  separately.

( $\rho(0) = 1, \rho(1), \rho(2), \dots$ ) defines a sequence of real numbers. Not every combination of numbers can occur in this sequence. The characterization of all possible sequences is achieved by (FULLER's **Theorem** 1.4.1), which states that the ACF is positive semi-definite (or non-negative definite). This means that, for any finite sequence of real numbers  $(a_1, \dots, a_n)$  and for any corresponding sequence of natural numbers  $(i_1, \dots, i_n)$

$$\sum_{j=1}^n \sum_{k=1}^n a_j a_k \rho(i_j - i_k) \geq 0. \quad (3)$$

Setting  $n = 2$  immediately yields the important property that  $|\rho(h)| \leq 1$ , which corresponds to the interpretation as a correlation (proof as an exercise or in FULLER, p.8). For  $n > 2$  it is less easy to interpret this condition. It is recommended to view it simply as an algebraic condition, which guarantees that all constructible variances are non-negative and which limits the selection of possible shapes for the ACF.

An important theoretical result is the **synthesis theorem**, which states that all positive semi-definite functions (with  $\rho(0) = 1$ ) can occur as ACF of a stationary process. To any positive definite function one can find/construct a process that has the given function as its ACF.

Even more important is the fundamental property—whose proof relies on representation theorems—that the ACF (together with the mean  $\mu$  and the variance) contains all essential information on the covariance-stationary process. In other words, the sequence  $\rho(\cdot)$  (& mean & variance) ‘characterizes’ the process. For example, the important problem of selecting the

parametric representation of the process can be based on the ACF alone. No further information that would be extractable from the data helps with this task. In a sense, processes with identical ACF are equivalent.

For example, the process ‘rolling a die with a bonus’ appears to have the ACF that is shown in Figure 6. It is easy to construct a theoretical process (even with normally distributed random errors) that has exactly the same ACF. Both processes—‘Rolling a die with a bonus’ and the synthetical process—are equivalent, although they differ with respect to many statistical properties. As long as the boundary of the so-called  $L^2$  theory of covariance-stationary processes is not transgressed, will such differences play no role, as they cannot be ‘perceived’.

## 1.6 Trend and seasonality

Time-series analysis is primarily concerned with time series (and/or corresponding processes, according to definition) that have a time-constant covariance structure and time-constant mean and volatility. These are called stationary. Approach # 3 in the introduction indicates that time-series analysis may also be envisaged as being concerned with cyclical stochastic processes, whose cycles have finite amplitude and frequency but are never exactly constant. Even this latter view leads to the covariance-stationary processes. Many economics data series, however, do not conform to these assumptions and thus resist a direct treatment by time-series analysis.

### Examples:

1. B&D give as their Example 1.1 a time series of measurements of an electric current at a resistor. This could be a pure deterministic cycle. In such a process, any future observation can be predicted exactly. Also in economic examples can a specific cycle (e.g. the seasonal cycle) dominate to such an extent that all remaining structure becomes entirely unrecognizable.
2. Their Example 1.2 (here Figure 1) shows the population of the USA in 10-year measurements. This is a trending time series. Its mean is clearly not time-constant.
3. Their Example 1.6 (here Figure 4) shows monthly measurements of accidental deaths. This is another seasonal series. Its mean may be constant in the long run, while within the sample the accidental deaths in July appear to exceed those in February. Hence, the mean experiences cyclical oscillations that may have a deterministic component. If

$EX_t$  depends on the remainder class of  $t$ , the assumption of stationarity is violated.

4. Their Example 1.5 (here Figure 3) are the legendary sunspots. From 1800 to 1830, their volatility appears to have been lower than at other times. If this were true,  $\sigma^2$  would be not constant and stationarity would be violated. This interpretation, however, is not coercive. The analysis of financial time series utilizes so-called ARCH processes that are covariance stationary and whose observed volatility changes are explained by cross moments of order four.
5. Their Example 1.3 (here Figure 2) is an annual series of strikes in the USA, which appears to be extremely irregular. If mean and/or volatility change at specific break points, economists call these time points ‘structural breaks’. The adequate treatment of this feature is uncertain, even the role-model example B&D 1.3 would be diagnosed and treated differently by different time-series analysts.

In summary, one should distinguish problematic deviations from the assumptions (volatility changes, structural breaks) and standard deviations with adequate routine treatment (trend and seasonality).

For the trend feature, we have two basic strategies, each of which has its advocates:

1. The time series is regressed on a simple function of time, such that the residual can be viewed as ‘stationary’. Fitting curves to data used to be quite common in the history of econometric time-series analysis, today the choice of time functions is restricted to:
  - (a) a linear trend function, i.e. a straight line;
  - (b) a quadratic trend function, i.e. a parabolic curve, but only if quadratic growth can be justified by plausibility arguments, which excludes unemployment rates, for example;
  - (c) an exponential trend function that conforms to case (a) after transforming the data by logarithms.

A disadvantage of *curve fitting* is that it often yields excellent results *in sample* but leads to nonsensical results if extrapolated *out of sample*.

2. Transforming the raw data so that the assumption of stationarity becomes plausible, primarily by

- (a) *first differences*  $x_t = X_t - X_{t-1}$ , which are usually denoted by  $\Delta X$  or by  $\nabla X$ ;
- (b) logarithmic differences  $x_t = \log(X_t/X_{t-1})$  are first differences of the logarithmized data;
- (c) *growth rates*  $100(X_t - X_{t-1})/X_{t-1}$  have properties similar to case (b) and are therefore often applied to the same data sets (even though there are important conceptual differences between the two methods);
- (d) *seasonal differences*  $x_t = X_t - X_{t-S}$  ( $S = 4$  for quarterly data,  $S = 12$  for monthly data) are also denoted by  $\Delta_S$ . These seasonal differences remove trend and seasonality simultaneously.

The application of first differences (also called *differencing*) has been enjoying growing popularity in econometric time-series analysis since the publication of the book by BOX & JENKINS. Compared to the more traditional *trend fitting*, it has different implications for the dynamic equilibria of econ(etr)ic models and for the long-run behavior of variables. These differences dominated the discussion on persistence and *unit roots* in the 1980s.

3. In particular, the empirical sections of contributions from the field of *real business cycles* have seen an increasing intensity of applications of the de-trending method by HODRICK&PRESCOTT. This procedure uses frequency-domain properties of time-series models.

For the seasonality feature, a wide variety of potential procedures is available. It is recommended to consult the monographs by HYLLEBERG (1986) and by GHYSELS & OSBORN (2001). We mention the following procedures:

1. *Seasonal dummies* are probably the oldest technique<sup>3</sup>. The seasonal cycle is modelled as a strictly periodic deterministic process, whose specification requires 4 or 12 coefficients (without the constant, 3 or 11). This modelling concept is fully supported by most econometric software products. Following an auxiliary regression, the residuals are either viewed as stationary or the seasonal dummy variables are integrated into the modelling process. A seasonal dummy is specified as 1 in a specific month (or quarter) and is 0 otherwise.

Disadvantage: inflexible seasonal pattern, unclear interaction with trend modelling. These problems motivated the development of

---

<sup>3</sup>GOURIEROUX&MONFORT cite a study by BUYS-BALLOT dated 1847.

2. *Seasonal adjustment*, which today commonly relies on the legendary routine ‘Census X-11’ or on one of its descendants X-12 and TRAMO-SEATS. A sophisticated sequence of smoothing and de-trending filters together with outlier adjustment and time-series model steps succeeds in ‘adjusting’ the data, i.e. in purging them from any seasonal characteristics, while other information—such as business cycles or trend behavior—is retained. Particularly the outlier correction and other non-linear components of the procedure are irreversible, meaning that the original data cannot be recovered from the adjusted series.

Disadvantage: Destruction of information, which may be dangerous if the true dynamics of seasonality does not correspond to the model that is implicitly assumed by the Census X-11 procedure; aliasing effects spreading to non-seasonal frequencies; insufficient theoretical basis for all operations; adjusted data may reflect arbitrary manipulations. For these reasons, many time-series analysts (unlike some empirical economists; see also BOX & JENKINS, 1976) recommend:

3. seasonal differences, see item (d) of ‘trend’. Seasonal differencing treats trend and seasonality alike. This method can also be justified on statistical grounds by theoretical models and hypothesis tests (among others, see HYLLEBERG, ENGLE, GRANGER, YOO, 1990, for the so-called HEGY test).

Disadvantage: seasonal variation may not be sufficiently restricted in the HEGY model. The implied stochastic seasonal cycle admits reversals of summer/winter patterns, even in cases where this may appear implausible (e.g. temperature). It should be noted, however, that summer-becomes-winter actually did occur in economic time series, for example in regional tourism, where demand does not peak any more in summer but in winter due to winter sports, or in energy demand, where air conditioning causes a formerly unknown peak in summer.

4. the *seasonal moving average* (SMA) is actually one component of the seasonal difference operator, or in other words: SMA + first differencing = seasonal differencing.

Disadvantage: phase shift by asymmetric filtering (which also appears in seasonal differences); phase correction may be achieved by other methods of weighted averaging, see e.g. B&D methods S1 and S2. Generally, method 4] is to be preferred to 3] if data do not have trends.

Figure 7 shows a typical macroeconomic time series, Austrian industrial production 1962–2002. Trend and seasonality are clearly visible. In Figure

9 the operator  $\Delta$  has eliminated the trend, the seasonal cycle is clearly discernible. Figure 8 shows the series after smoothing by the SMA filter, Figure 10 shows seasonal differences. It is obvious that the operator  $\Delta_4$  eliminates trend as well as seasonality.

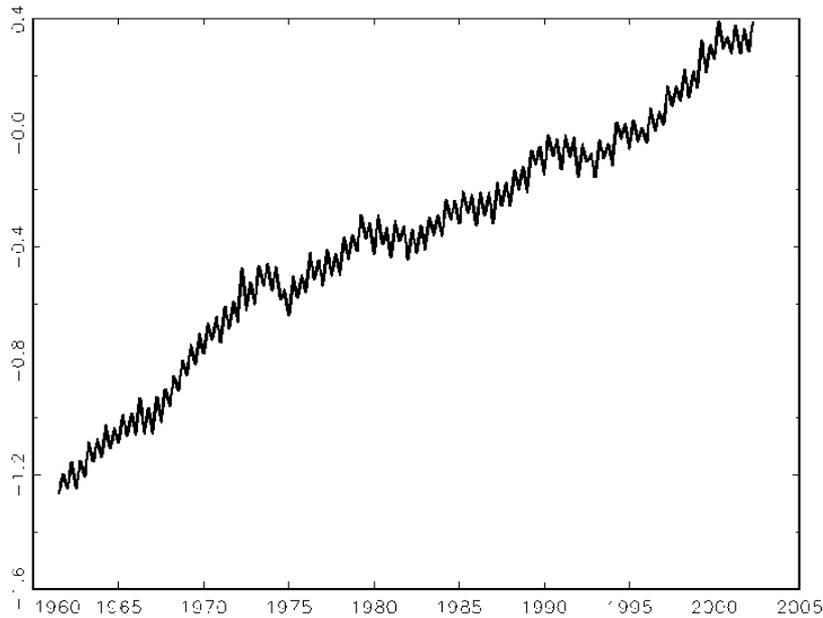


Figure 7: Austrian industrial production 1962–2002. Quarterly data provided by OECD, after logarithmic transformation.

If forecasting is the aim of the analysis, another distinctive feature of the methods may become crucial. Some procedures admit predicting the original (untransformed, ‘unfiltered’) process, while others can only be used for predicting the transformed, ideally stationary process. In particular the B&D methods S1 and S2, X-11/12, the HP filter do not allow forecasting the unadjusted variable. The B&D methods explain trends only locally (by non-parametric smoothing), the trend remains unexplained outside the sample. HP and Census destroy sample information in specific frequency bands (in the long-run band for trends and in an area around the seasonal frequencies, respectively). With regard to X-11/12, this destruction may be justified by the popular view that seasonally adjusted data are (closer to) true data. Under this assumption, of course, forecasting the seasonally adjusted data appears sufficient.

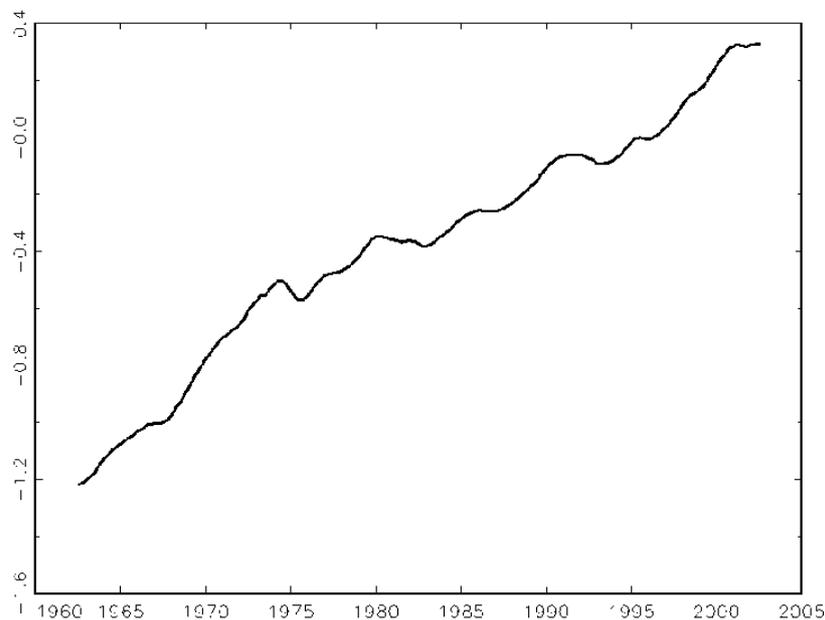


Figure 8: Seasonal moving averages of industrial production. Data filtered by  $1 + B + B^2 + B^3$ .

## 2 ARMA processes

It was already outlined that one of the (more or less equivalent) approaches of time-series analysis is based on the idea of describing observed time series by classes of comparatively simple processes. In other words, this modelling aims at ‘parameterizing’ the ACF, that is to find a model that approximates the empirically estimated ACF as well as possible. It can be shown (Wold’s Theorem) that every covariance-stationary process can be approximated arbitrarily well by so-called moving-average processes (MA). Similarly, almost every covariance-stationary process can be approximated arbitrarily well by autoregressive processes (AR). Note that ‘approximation’ always refers to a matching of theoretical and empirical ACF (second moments), while nothing is stated on higher moments or other stochastic properties. Finally, the ARMA process is an amalgam of its two basic elements AR and MA.

### 2.1 The moving-average (MA) process

The MA process directly generalizes white noise. Assume that  $(\varepsilon_t)$  is *wn* with variance  $\sigma^2$ . In the following, we will use  $\sigma^2$  for the variance of  $(\varepsilon_t)$ ,

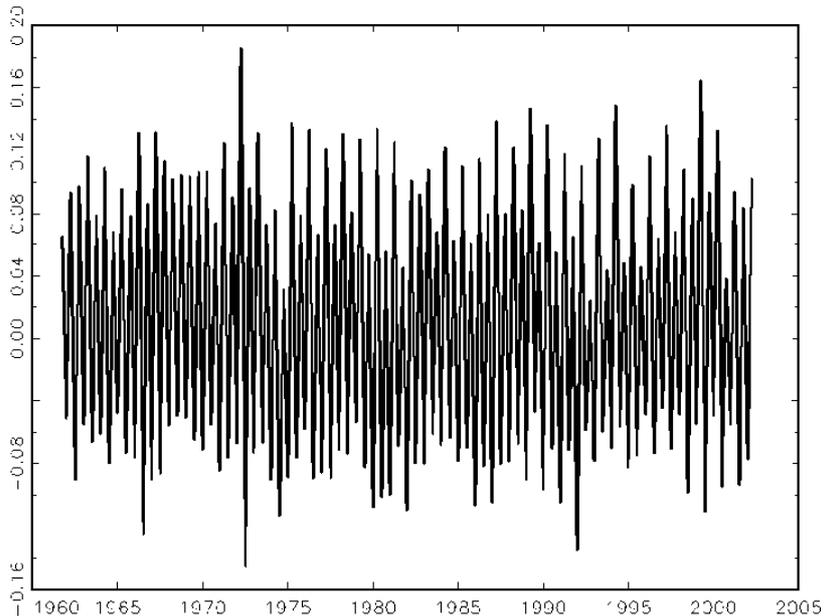


Figure 9: First differences of the logarithmized industrial production data (Growth rates relative to previous quarter).

while otherwise we use the detailed notation of, for example,  $\sigma_X^2$  to denote the variance of  $(X_t)$ . Then,

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1} \quad (4)$$

defines an MA process of order one (a ‘first-order MA process’), symbolically denoted by MA(1). The parameter  $\theta$  may be any real number. This process has the following attractive property:

**Property 1:** The MA(1) process is covariance stationary.

To show this, we just determine the first and second moments. By using the linearity of the expectation operator and the properties of the *wn*  $(\varepsilon_t)$ , it is immediate that  $EX_t = 0$ , that  $\text{var}X_t = (1 + \theta^2)\sigma^2$  and that  $EX_tX_{t-1} = \theta\sigma^2$ . This implies that

$$\rho_1 = \frac{\theta}{1 + \theta^2} \quad , \quad (5)$$

with the shorter notation  $\rho_h$  instead of  $\rho(h)$ , which we will use in the following. All  $\rho_h$  with  $h > 1$  are clearly 0, and hence the thus defined process is stationary. If we replace  $\theta$  by  $1/\theta$  in (4) and (5), we note to our surprise that this yields the same value for  $\rho_1$ . Thus, for example the MA(1)

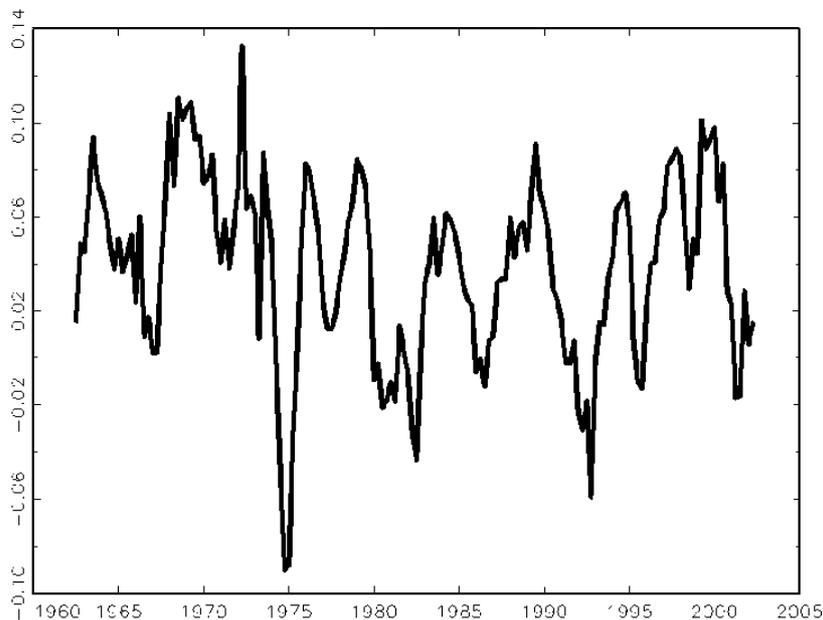


Figure 10: Seasonal differences of the logarithmized industrial production data. Growth rates.

processes with  $\theta = 0.5$  and  $\theta = 2$  have the same ACF and therefore they are ‘equivalent’. Their variance may be different, but the variance of  $\varepsilon_t$  is unobserved. For observed data, the error variance results from  $\theta$  and  $\text{var}X$  via  $\sigma^2 = \text{var}X / (1 + \theta^2)$ . We have established:

**Property 2:** MA(1) processes with coefficients  $\theta$  and with  $1/\theta$  are equivalent. Therefore, one may restrict attention to the case  $|\theta| \leq 1$ .

This means we cannot tell which  $\theta$  out of the two has actually generated the data. If estimation yields a value  $\theta > 1$ , we can replace it by  $1/\theta$  and obtain an equivalent estimate. No substance conclusions can be drawn from the fact that an estimated  $\theta$  exceeds one.

What do these MA(1) processes look like? Formula (5) reveals that  $\rho_1$  is negative for  $\theta < 0$  and positive for  $\theta > 0$ . The boundary case  $\theta = 0$  is of course just the *wn* process. This degenerate case would not be classified as a MA(1) process but as MA(0). Starting from  $\theta = 0$  in either direction, let us increase  $\theta$  in its absolute value. At first,  $\rho_1$  will also increase in its absolute value. However, autocorrelation reaches its maximum of 0.5 for  $\theta = 1$ . Property 2 states that further increases in  $|\theta|$  imply lower values of  $\rho_1$ . This observation implies the following important property:

**Property 3:** The class of MA(1) processes describes all covariance-

stationary processes, for which  $\rho_1 \leq 0.5$  and  $\rho_i = 0$  for all  $i > 1$ .

Therefore, if an observed process has an estimated  $\rho_1 > 0.5$  and if we can recognize several non-zero  $\rho_i$  with  $i > 1$  that are either positive or negative, then MA(1) cannot be a good model for the data-generating process.

Obviously, the next point should be MA processes of higher order. It is straight forward to define the MA(2) process by

$$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}, \quad (6)$$

and more generally the MA( $q$ ) process by

$$X_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \dots + \theta_q\varepsilon_{t-q}. \quad (7)$$

Simple application of the expectations operator and exploiting its linearity as well as the definition of a *wn* yields

**Property 1:** MA( $q$ ) processes are always stationary.  $\rho_h = 0$  for all  $h > q$ . The autocovariances  $\gamma_h$  for  $h \leq q$  can be expressed in the coefficients  $\theta_i$ , in detail:

$$\begin{aligned} \gamma_1 &= \sigma^2(\theta_1 + \theta_1\theta_2 + \dots + \theta_{q-1}\theta_q) \\ \gamma_2 &= \sigma^2(\theta_2 + \theta_1\theta_3 + \dots + \theta_{q-2}\theta_q) \\ \gamma_3 &= \sigma^2(\theta_3 + \theta_1\theta_4 + \dots + \theta_{q-3}\theta_q) \\ &\dots \end{aligned}$$

The structure of the formulae is clearly recognizable. It is even better recognizable if we set formally  $\theta_0 = 1$ . Then, we have generally:

$$\gamma_h = \sigma^2 \sum_{i=0}^{q-h} \theta_i \theta_{i+h}, \quad 0 \leq h \leq q. \quad (8)$$

For  $h = 0$  this yields the variance formula

$$\text{var}X_t = \gamma_0 = \sigma^2 \sum_{i=0}^q \theta_i^2. \quad (9)$$

From these formulae, the ACF evolves as

$$\rho_j = \frac{\sum_{i=0}^{q-j} \theta_i \theta_{i+j}}{\sum_{i=0}^q \theta_i^2}, \quad j \leq q.$$

And what about Property 2 for the general MA( $q$ ) model?

**Property 2:** Generally, there are  $2^q$  different combinations of  $q$  parameters  $\{\theta_1, \dots, \theta_q\}$ , all of which imply the same ACF and therefore define equivalent MA( $q$ ) processes.

In order to obtain a unique representation, we need the concept of a characteristic equation. According to an immediately obvious principle, we attach to any MA model

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

the characteristic polynomial, formally written in the complex variable  $z$ ,

$$\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q.$$

For  $z \in \mathbb{C}$ , this polynomial has exactly  $q$  zeros or ‘roots’  $\zeta_1, \dots, \zeta_q$ . One can show that the equivalent MA processes correspond to those polynomials that are created by replacing one or more of these zeros  $\zeta_j$  by  $\zeta_j^{-1}$ . This implies that there is exactly one MA process, for which *all zeros are larger or equal to one*, according to their absolute value or ‘modulus’. If we convene to always choose this process, then the representation is unique.

**Remark.** These properties are strictly valid only if we ignore the possibility of multiple and unit roots. If we account for this possibility, we should say that there are *at most*  $2^q$  different equivalent representations.

There is a reason why we do not, for example, choose the representation with all zeros *less than one*. The model with large zeros is the one with small coefficients and thus embeds *white noise*. Furthermore, one can show that an MA process can be represented as a convergent series

$$\sum_{j=0}^{\infty} \phi_j X_{t-j} = \varepsilon_t,$$

if and only if the roots of the characteristic polynomial are strictly greater than one. This ‘infinite AR representation’ is convenient for prediction. If it exists, the MA process is called *invertible*.

Hence, it is always possible to select a specific solution that fulfils the identifying restriction. Computer programs could first—utilizing non-linear optimization or comparable techniques—estimate a parameter vector, then replace it by an equivalent one that fulfils the uniqueness condition, and print out this final estimate.

For plausibility reasons, attention typically focuses exclusively on MA processes that are constructed from past  $\varepsilon_t$ . Such processes are also called *causal*, as it is usually assumed that causes precede their effects. The opposite would be an *anticipative* process that explains the past by the future.

Occasionally, it makes sense to consider mixed forms. One may distinguish *one-sided* and *two-sided* MA. A *two-sided* MA-process is, for example, given by

$$X_t = \sum_{j=-q_1}^{q_2} \theta_j \varepsilon_{t-j}.$$

All of these extensions are, of course, stationary.

What do the higher-order MA processes look like? One can show that every covariance-stationary process can be ‘represented’ as an infinite-order(!) MA process (WOLD’s Theorem). The higher the MA order, the more processes can be modeled. Although according to Property 1 ‘the ACF must cut off’, this cut-off can only limit the process behavior in quite large samples ( $T \gg q$ ). Even limitations according to the definiteness requirement of the ACVF—such as  $\rho_1 \leq 0.5$  for MA(1)—lose their importance with increasing order  $p$ .

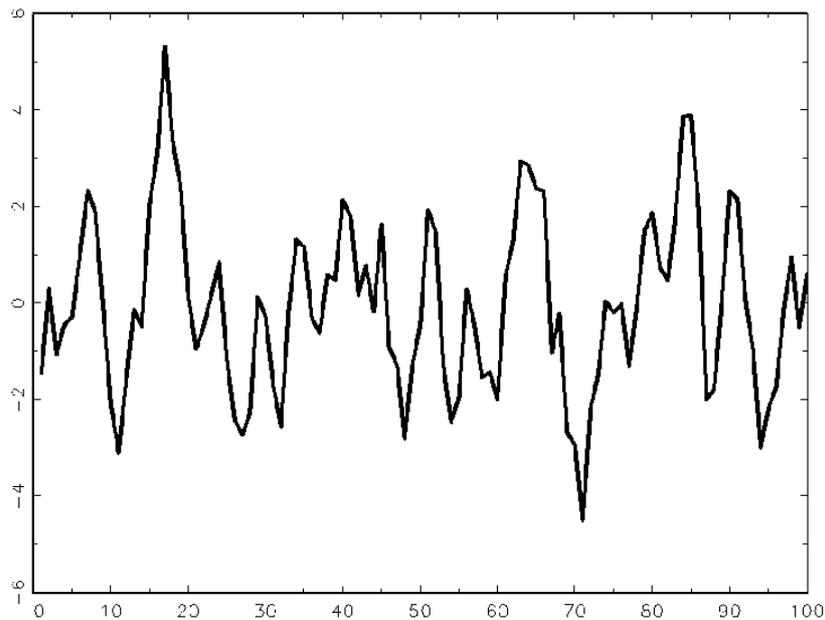


Figure 11: 100 observations of the process  $X_t = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2}$ .

Figure 11 shows a segment of a trajectory from an MA(2) process with  $\theta_1 = \theta_2 = 1$ , which even casual eyeballing may be able to tell apart from a *wn*. Such examples demonstrate that higher-order MA processes tend to ‘oscillate’, which MA(1) processes cannot do. It appears difficult, however,

to discern the cycles visually, the modelling of strong cyclical behavior is more easily done by AR models.

We note that WOLD's Theorem contains a special feature, as the meaning of an infinite-order MA series

$$\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} \tag{10}$$

is not immediately obvious. Such expressions must converge to become meaningful. One can show that 'quadratic convergence' of the series of coefficients suffices for this purpose, i.e.

$$\sum_{j=0}^{\infty} \theta_j^2 < \infty. \tag{11}$$

In order to guarantee convergence of infinite sums of the form  $\sum \theta_j u_{t-j}$  for autocorrelated input—with  $u$  stationary but not  $wn$ —it is helpful to demand the stronger condition of 'absolute convergence' for the coefficients series, i.e.

$$\sum_{j=0}^{\infty} |\theta_j| < \infty. \tag{12}$$

It is easy to see that absolute convergence is stronger than quadratic convergence of series. Consider for example the sequence  $1/n$ . If summed up to a series, it diverges, while the sum of the squared  $1/n^2$  converges to  $\pi^2/6$ .

Within the boundaries of ARMA processes and of these lecture notes, there is no difference between absolute and quadratic convergence. However, there is a cost involved in limiting attention to absolutely convergent MA( $\infty$ ) processes, as this excludes some important stationary cases. Processes without absolute convergence of their Wold coefficients series are said to have *long memory*. It is doubtful whether they play a role in economics, although there is some evidence for long memory in interest rates and other financial series.

**Summary:** An MA process, constructed from a quadratically convergent coefficients sequence ( $\theta_j$ ) and white noise, is covariance stationary. If we consider an infinite sum of weighted observations of a stationary process instead of a weighted sum of  $wn$  terms, then the thus defined process will be stationary if its sequence of coefficients is absolutely summable.

## 2.2 Autoregressive processes (AR)

Whereas the MA-process is formally easier to handle, to many empirical researchers the autoregressive process is more intuitive and easier to interpret.

A dynamic generating law such as

$$X_t = \mu + \phi X_{t-1} + \varepsilon_t \quad (13)$$

can be found in many places, not only in economics. (13) defines the first-order autoregressive process, in short AR(1). We again set  $\mu = 0$ . The properties of the process are determined by two parameters: the coefficient  $\phi$  and the error variance  $\sigma^2$ . For  $\phi = 0$  will the process degenerate to a *wn*.

A generating law such as the difference equation (13) may have diverse interpretations. Some see a recursion starting from a given and fixed value for  $X_1$  or  $X_0$ , a so-called starting value. While drawing repeatedly uncorrelated random variables  $\varepsilon_t$ , the recursion generates the time-series process. The thus defined process will usually not be stationary, as the first observation follows a different distribution from the tenth, say, which still feels the starting value strongly, and this tenth observation will again follow a different law from the 1000th. Under certain conditions, however, will the process approximate a stationary one gradually. Some call this gradual approximation to a stationary steady state ‘asymptotic stationarity’.

A different interpretation is that the process ( $X_t$ ) has already been active for an infinite time span and that we just start observing it at some time point. Such a process may actually be stationary. In detail, one has the following theorem:

**Property 1:** The AR(1) equation (13) has a stationary solution, if and only if  $|\phi| < 1$ .

To proof and to motivate this result, we may consider some calculations. First we may substitute the right-hand side of (13) repeatedly into the left-hand side, which yields after an  $\infty$  number of substitutions:

$$X_t = \sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}. \quad (14)$$

This is an MA( $\infty$ )–process with geometric and therefore (as a series) absolutely convergent sequence of coefficients, which certainly is stationary. By contrast, if  $|\phi| \geq 1$ , then summing up will not be possible. The case  $\phi = 1$  is called the *random walk*, i.e. a non-stationary process. The case  $|\phi| > 1$  is called an *explosive* process. If  $\phi = -1$ , some authors name it a *random jump*.

Another calculation exercise evolves from the attempt to determine the process variance  $\sigma_X^2$  assuming stationarity. If ( $X_t$ ) is stationary, then we have  $\text{var}(X_t) = \text{var}(X_{t-1})$  and hence:

$$E(X_t^2) = \phi^2 E(X_{t-1}^2) + \sigma^2,$$

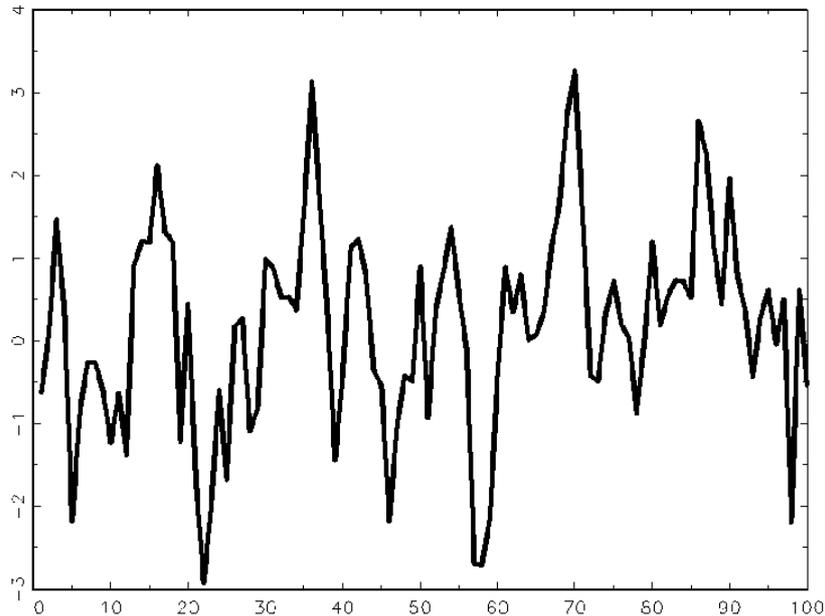


Figure 12: Trajectory of an AR(1) process  $X_t = 0.5X_{t-1} + \varepsilon_t$ . A pre-sample of 50 data points guarantees that the shown process is approximately stationary.

$$E(X_t^2) = \frac{\sigma^2}{1 - \phi^2}. \quad (15)$$

Obviously, a sensible solution is possible only if  $|\phi| < 1$ . Thus, whereas all MA processes are stationary, some AR processes are stationary, while others are not. We note for completeness that the stochastic difference equation (13) possesses stationary solutions even for  $|\phi| > 1$ , but those are anticipative and are seen as implausible. This peculiarity explains why explosive and stationary AR processes have identical ACF in finite samples (i.e. the estimated ACF or correlogram, while the theoretical ACF is undefined for explosive processes). For example,  $\phi = 0.5$  and  $\phi = 2$  in (13) yield identical correlograms. Therefore, the correlogram cannot be used to determine whether the observations follow an explosive scheme.

In the following, a process will be called *stable* if its scheme has a stationary solution, such that ‘the difference equation is stable’. Thus, an AR(1) process is stable for  $|\phi| < 1$  but it may not be stationary. Others call such processes ‘asymptotically stationary’.

**Property 2:** The first-order autocorrelation of a stationary AR(1) process is simply  $\rho_1 = \phi$ . Generally, we have  $\rho_h = \phi^h$ . This is easy to show. It follows that the ACF of a stable AR(1) decays

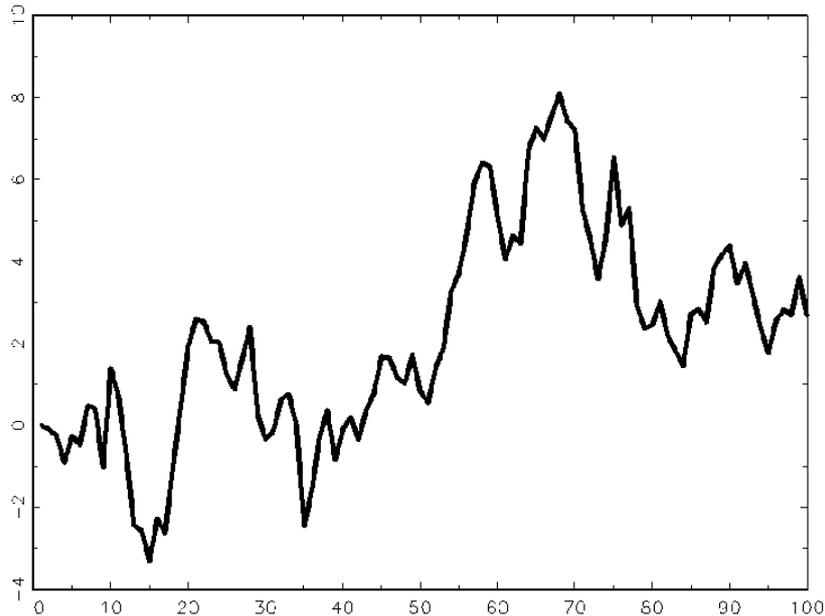


Figure 13: Trajectory of the random walk  $X_t = X_{t-1} + \varepsilon_t$ , started from  $X_1 = 0$ .

geometrically, in an alternating pattern if  $\phi < 0$ . The empirical ACF (the correlogram) reflects these properties only approximately.

This geometric decay (in contrast to the cut-off in the ACF of an MA process) is characteristic for all AR processes. In consequence, if my time-series data yield an ACF with cut-off after a specific  $p$  or a hump-shaped ACF—which is often observed for quarterly data, where  $\rho_4$  is more important than  $\rho_3$ —then AR(1) is probably an inadequate model.

An AR(1) process does not oscillate. Thus, it is not suited for the description of business cycles. Even when  $\phi < 0$ , it does not really oscillate but it alternates or jumps with fixed periodicity, i.e. one observation is positive, the next one is negative, and so on. The properties of difference equations imply that even the most immediate generalization to second order can generate cycles of (almost) every periodicity. Processes of the form

$$X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t \quad (16)$$

are called second-order autoregressive processes or in short AR(2). Again we set  $\mu = 0$ . Then, the process—apart from  $\sigma^2$ —depends on two parameters, the coefficients  $\phi_1$  and  $\phi_2$ . To us, this appears to be the most reasonable parameterization. For the determination of stationarity properties,

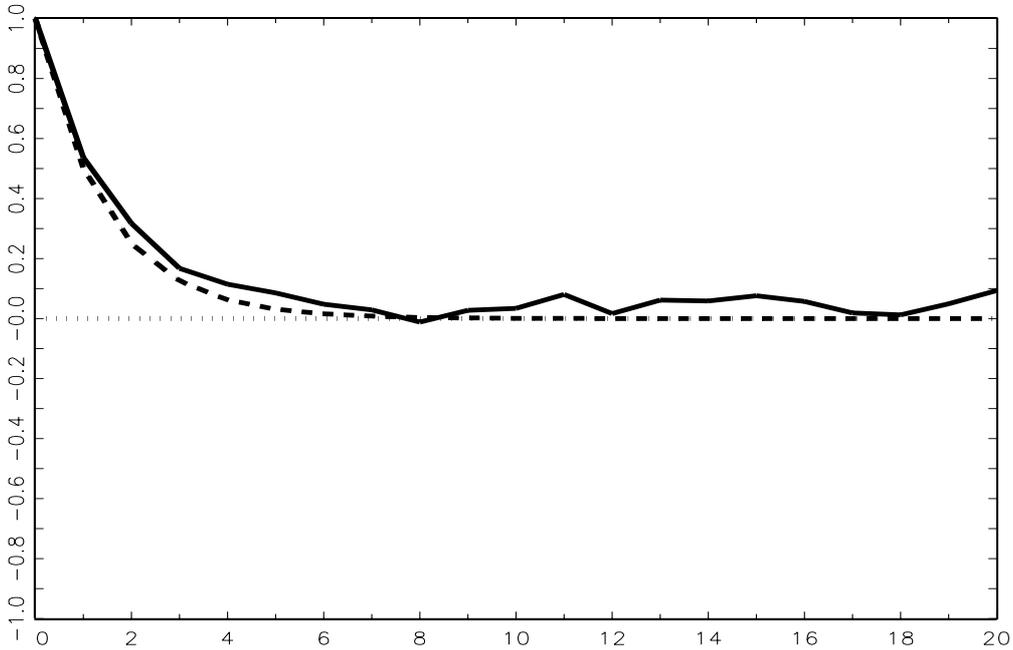


Figure 14: Empirical (solid) and theoretical (dashed) ACF of the process  $X_t = 0.5X_{t-1} + \varepsilon_t$ .

however,  $\phi_1$  and  $\phi_2$  are not the ‘natural’ parameters of the model class. In consequence, the stability or stationarity conditions become quite unwieldy for AR(2) processes. The situation deteriorates further for AR( $p$ )–processes with large  $p$ .

**Property 1:** The AR(2) process is stable if and only if

$$\begin{aligned}
 \text{(a)} \quad & \phi_1 + \phi_2 < 1, \\
 \text{(b)} \quad & \phi_1 - \phi_2 > -1, \\
 \text{(c)} \quad & \phi_2 > -1.
 \end{aligned} \tag{17}$$

These three conditions must be fulfilled jointly and they represent the boundaries of a triangle in the  $(\phi_1, \phi_2)$ –plain. It is possible to derive the property directly but this is uninspiring. It makes more sense to consider a general principle that may be used to calculate corresponding restrictions for all AR processes of arbitrary lag order. This principle is based on the properties of linear difference equations, particularly on the so-called *characteristic equations* or companion polynomials.

The idea starts by re-writing (16), such that we only have  $\varepsilon_t$  on the right-

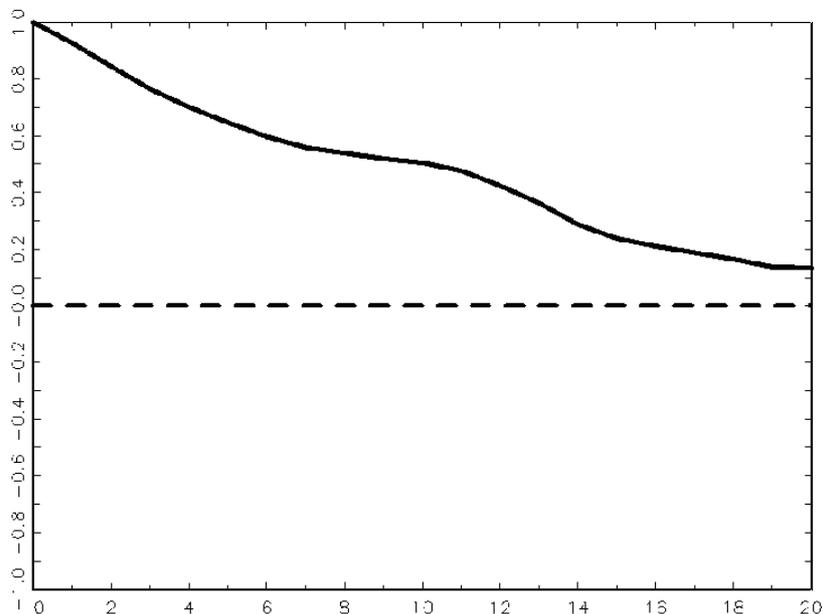


Figure 15: Empirical ACF of a random walk.

hand side:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = \varepsilon_t. \quad (18)$$

Then, according to an obvious scheme we replace the  $j$ th lag by a polynomial argument  $z^j$  and  $\varepsilon_t$  by 0:

$$z^0 - \phi_1 z^1 - \phi_2 z^2 = 0 \quad (19)$$

This is a quadratic equation, high-school math, and it has two zeros (roots) that are either both real or complex conjugates:

$$\zeta_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \hat{=} \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \quad (20)$$

Now we can use the fundamental property that these roots are larger than 1 in their modulus (absolute value) if and only if the corresponding AR process has a covariance-stationary solution. This proposition can be generalized to AR processes of any order. Its proof (for example, in BROCKWELL & DAVIS) appears to be comparatively short and innocuous but it uses some deeper knowledge of the analytical properties of polynomials etc.

**Theorem** (Part of Theorem 3.1.1 in B&D) : The AR( $p$ ) process

$$X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t \quad (21)$$

has a covariance-stationary causal solution (is stable) if and only if all roots  $\zeta_i$  of the polynomial

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p$$

fulfil the condition  $|\zeta_i| > 1$ , i.e. they are all situated ‘outside the unit circle’.

**Remark 1:** Some use instead of  $\phi(\cdot)$  the inverted polynomial  $z^p - \phi_1 z^{p-1} - \dots - \phi_p$ . Then, stability is equivalent to ‘all roots are smaller than 1’. Some prefer this variant, where stability corresponds to a bounded area. Again others invert the signs of the coefficients  $\phi_i$ , which yields nicer coefficients but the interpretation of the AR model as a ‘regression’ is then lost.

**Remark 2:** The theorem does not save use from the work to determine the polynomial zeros for higher-order AR( $p$ ) processes. A result from algebra states that this can be done analytically up to  $p = 4$ . For  $p > 4$ , numerical approximation has to be invoked, i.e. ‘the computer’.

**Remark 3:** The fundamental theorem of algebra states that a  $p$ th-degree polynomial has  $p$  zeros. These ‘roots’ are either complex conjugates or real, and they may be ‘multiple’. These properties have certain effects on the behavior of the corresponding process. Some details will follow at a later point.

Some insight into the condition is obtained by representing the AR models as infinite-order MA models (‘MA( $\infty$ )’). This is trivial for the AR(1) process

$$X_t = \phi X_{t-1} + \varepsilon_t,$$

as iterative substitution yields

$$X_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \phi^3 \varepsilon_{t-3} + \dots$$

The right-hand side consists of an infinite sum of  $wn$  variables. We know that this sum ‘converges’ if the series of coefficients converges quadratically. For  $|\phi| < 1$ , we obtain an absolutely convergent geometric series, because of  $\phi^2 < |\phi| < 1$  it also converges quadratically. For  $|\phi| \geq 1$  the series diverges. In summary, we recover the known stability condition for AR(1) processes. What about AR(2) processes?

Assume that the characteristic equation has two real roots  $\zeta_1, \zeta_2$ . Then, we may conceive the AR(2) model as a sequence of two transformations:

$$\begin{aligned} X_t &= (1/\zeta_1)X_{t-1} + \eta_t, \\ \eta_t &= (1/\zeta_2)\eta_{t-1} + \varepsilon_t. \end{aligned} \tag{22}$$

The second equation yields a convergent MA( $\infty$ ) process iff  $|\zeta_2| > 1$ . The so defined  $(\eta_t)$  is not *wn*, but from the theorem mentioned above we know that the first equation defines a stationary process even if  $(\eta_t)$  is only covariance-stationary, given that the coefficients series is absolute convergent (for the second equation, quadratic convergence would have sufficed). Thus, the conditions for the AR(2) model evolve, and one could proceed similarly for any AR( $p$ ) process.

So what about complex conjugates? Of course, the same method still works but it now requires considering complex-valued processes. Complex-valued MA and AR processes are not attractive, particularly as the data are usually never complex, but these processes share most of their properties with their real-valued siblings.

It is yet to be shown that the sequence of transformations really generates any given AR(2) model. To do so, we insert the first into the second equation in (22):

$$X_t - (1/\zeta_1)X_{t-1} = (1/\zeta_2)\{X_{t-1} - (1/\zeta_1)X_{t-2}\} + \varepsilon_t,$$

which by equating coefficients yields

$$\begin{aligned}\phi_1 &= 1/\zeta_1 + 1/\zeta_2, \\ \phi_2 &= -1/(\zeta_1\zeta_2).\end{aligned}$$

The solution formula for quadratic equations implies that  $\zeta_1$  and  $\zeta_2$  are the roots of the characteristic equation, *q.e.d.* Proofs for higher orders are straight forward, for example using the principle of complete induction.

An empirically important property is the cut-off property of the so-called partial ACF (*PACF*) for all AR processes. The PACF can be introduced by either one of two equivalent definitions.

In the first definition, autoregressive models of increasing order are fitted to the data, for example using a variant of least squares:

$$\begin{aligned}X_t &= \phi_1^{(1)}X_{t-1} + \varepsilon_t^{(1)}, \\ X_t &= \phi_1^{(2)}X_{t-1} + \phi_2^{(2)}X_{t-2} + \varepsilon_t^{(2)}, \\ X_t &= \phi_1^{(3)}X_{t-1} + \phi_2^{(3)}X_{t-2} + \phi_3^{(3)}X_{t-3} + \varepsilon_t^{(3)}, \\ &\dots \\ X_t &= \phi_1^{(p)}X_{t-1} + \phi_2^{(p)}X_{t-2} + \dots + \phi_p^{(p)}X_{t-p} + \varepsilon_t^{(p)}.\end{aligned}$$

If the data have really been generated by an AR( $p$ ) process, then the error terms  $\varepsilon_t^{(k)}$  for  $k < p$  are not *wn*, while they are *wn* for  $k \geq p$ . Furthermore,  $\phi_p^{(p)}$

delivers a good and consistent estimate for the true  $\phi_p$ .  $\phi_p \neq 0$ , as otherwise we would have a process of the order  $p - 1$  or less. The coefficients  $\phi_k^{(k)}$  for  $k < p$  estimate something unclear in misspecified models, while  $\phi_k^{(k)} = 0$  for  $k > p$ . Thus, if we plot  $PACF(k) = \phi_k^{(k)}$  against  $k$ , then this function must cut off at  $k = p$ . This property is characteristic for  $AR(p)$  processes. For a rigorous derivation, one would have to distinguish between the estimates  $\hat{\phi}_k^{(k)}$  of the empirical PACF and their limits for  $T \rightarrow \infty$  for the theoretical PACF.

In the second definition,  $\phi_k^{(k)}$  is equated directly to a conditional correlation

$$PACF(k) = \text{corr}(X_t, X_{t-k} | X_{t-k+1}, \dots, X_{t-1}).$$

This partial correlation explains the name PACF. It is not too demanding to prove the equivalence of both definitions.

**Summary.** The PACF of an  $AR(p)$  process has values different from 0 for  $k = p$  and usually also for  $k < p$ , whereas it is zero for  $k > p$ . Thus, it cuts off at  $k = p$ . By contrast, the PACF of an MA process does not cut off, it decays smoothly to zero as  $k \rightarrow \infty$ . The PACF for AR and MA processes shows a behavior that is exactly the opposite to the ACF, which breaks off for MA processes but decays smoothly to zero for AR processes.

As shown in Figure 16, the empirical pattern may differ from the theoretical cutting-off property. A common criterion for the ‘significance’ of single values for ACF or PACF is  $2/\sqrt{T}$ . With 100 observations, this would yield  $\pm 0.2$ . Values within the band  $[-0.2, 0.2]$  can be regarded as insignificantly different from zero.

## 2.3 Autoregressive moving-average (ARMA) processes

Replacing the *wn* errors of an AR model by an MA term defines the ‘mixed’ ARMA model (*autoregressive moving-average*), which is an important generalization of the AR and MA classes:

$$X_t = \mu + \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}. \quad (23)$$

The thus defined process is formally denoted as  $ARMA(p, q)$ . It inherits its properties from its two components.

**Property.** The ARMA process (23) is asymptotically stationary (stable), if all zeros of the AR polynomial  $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$  have modulus larger than one. It is uniquely defined, if all zeros of the MA polynomial  $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$  have modulus  $\geq 1$  and if the polynomials  $\phi(z)$  and  $\theta(z)$  have no common zeros.

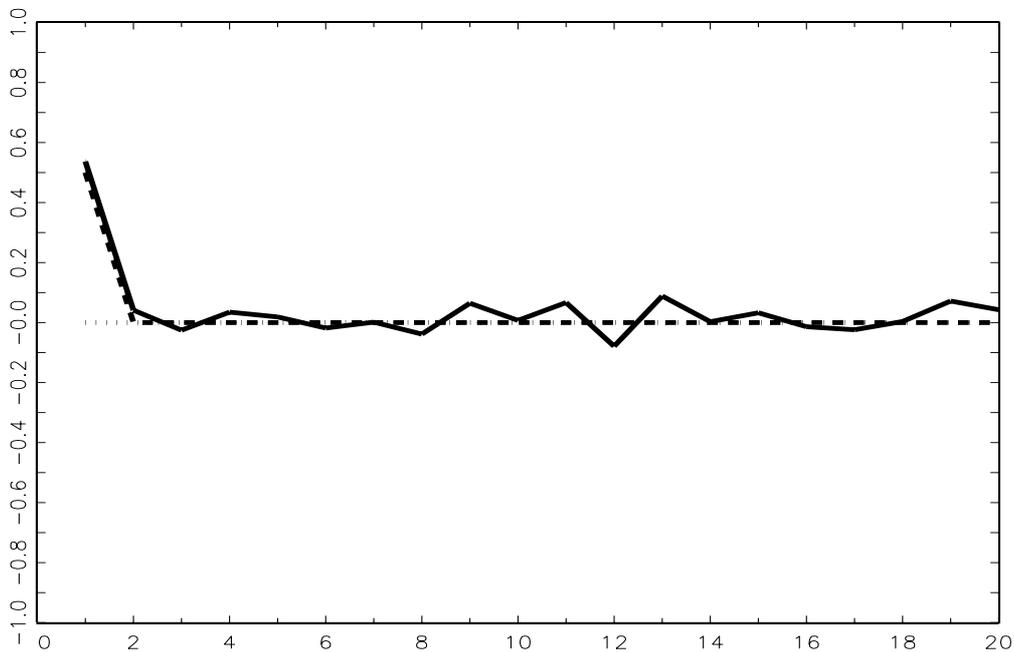


Figure 16: Empirical (solid) and theoretical (dashed) PACF for the AR process  $X_t = 0.5X_{t-1} + \varepsilon_t$ .

**Example.** The ARMA(1,1) process  $X_t = \phi X_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$  has the polynomials  $\phi(z) = 1 - \phi z$  and  $\theta(z) = 1 + \theta z$ . For  $\phi + \theta = 0$  there is a common root at  $1/\phi$ . It is easily seen that the process is simply *wn* in this case. Therefore, neither  $\phi$  nor  $\theta$  can be determined empirically. Problems in estimation, particularly failure of convergence in non-linear optimization steps, of ARMA models are often rooted in nearly ‘cancelling’ roots.

The ACF of an ARMA process is not as easily determined as the ACF of an MA process. As an exercise, we recommend doing it for the ARMA(1,1). It follows directly from their definition and construction that neither ACF nor PACF can cut off. Because for stable processes the *series* of (Wold) MA coefficients must converge quadratically, it is easily seen that both the ACF and the PACF must decrease to zero for large values. This suggests to classify time series as mixed ARMA, if neither the ACF nor the PACF cuts off. In finite samples, however, the distinction often becomes blurred. Furthermore, it is quite difficult to determine the correct orders  $p$  and  $q$  from the correlogram for the mixed ARMA model.

For the identification and estimation of ARMA models, various suggestions are available. According to BOX&JENKINS, one may try the following

pattern.

**Example.** The data were actually generated by the ARMA process  $X_t = 0.5X_{t-1} + \varepsilon_t + 0.8\varepsilon_{t-1}$ . Estimation yields the correlogram and partial correlogram in Figures 17 and 18.

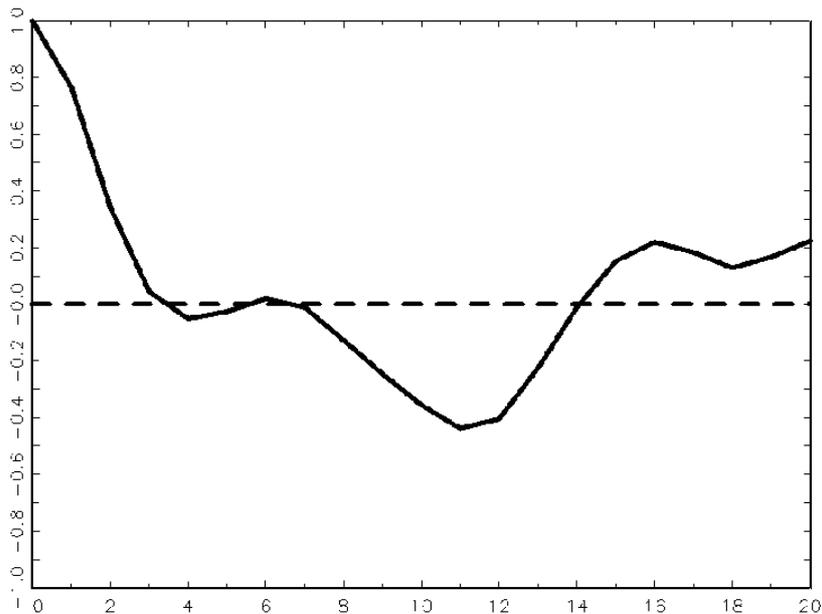


Figure 17: Correlogram of a time series with 100 observations that was generated from the process  $X_t = 0.5X_{t-1} + \varepsilon_t + 0.8\varepsilon_{t-1}$ .

Visual inspection of the plots may suggest an AR(2) or maybe an AR(3) model, as the third partial autocorrelation is close to the significance boundary. The ACF decays relatively fast without cutting off sharply, such that an MA model is not plausible. Thus, one may now estimate the two AR model specifications and inspect the correlograms of their residuals. Doing so, we note that the residuals of an AR(1) model still show significant ACF and PACF values, while these are marginally significant for AR(2) and insignificant for AR(3). Additionally, all 3 coefficients of an AR(3) model are well established on the basis of their  $t$ -statistics. Thus, AR(3) is an acceptable model for the data.

Next, we may try ARMA(3,  $q$ ) models, with  $q \leq 3$ . Indeed estimation of ARMA(3, 1) yields a highly significant MA term and insignificant coefficients  $\phi_2$  and  $\phi_3$ . When these are eliminated and an ARMA(1, 1) model is estimate, we obtain  $\hat{\theta}_1 = 0.85$  and  $\hat{\phi}_1 = 0.54$ , which are good approximations to the—assumedly unknown—true values. The ARMA(1, 1) attains a slightly higher

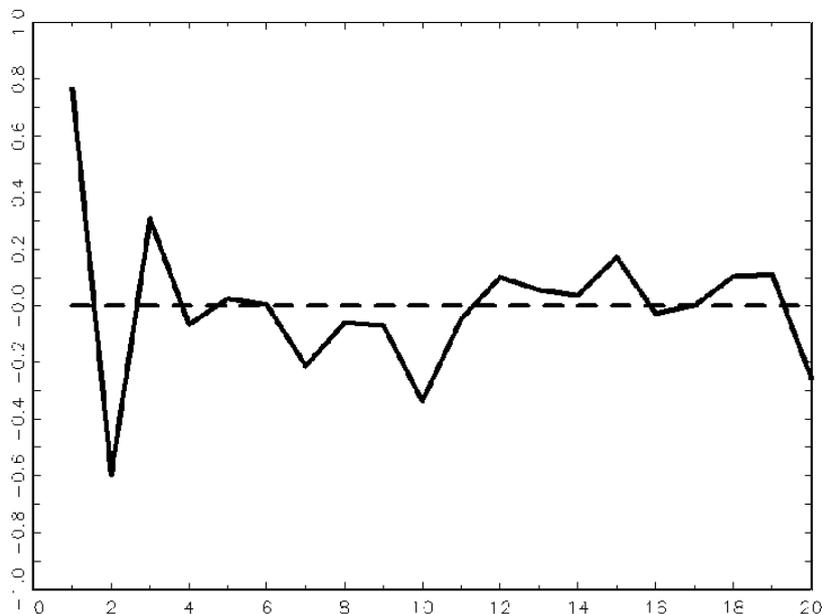


Figure 18: Empirical PACF of a time series with 100 observations that was generated from the process  $X_t = 0.5X_{t-1} + \varepsilon_t + 0.8\varepsilon_{t-1}$ .

$R^2$  using one parameter less than AR(3) and is to be preferred for that reason, although both models give an acceptable description of the data.  $\square$

It is obvious that this *trial-and-error*-procedure is cumbersome. Therefore, many current time-series analysts prefer more mechanical techniques. For example, one may estimate a sizeable set of different ARMA models for  $p = 0, \dots, P$  and  $q = 0, \dots, Q$ . Some criterion is evaluated for each estimated model, and the accordingly best model is then used. Some researchers recommend to inspect the two or three best models more closely and to subject them to specification tests. The most customary criteria are the *information criteria* AIC (by AKAIKE) and BIC (by SCHWARZ). They are defined by

$$\begin{aligned} AIC &= \log \hat{\sigma}^2(p, q) + \frac{2}{T} (p + q), \\ BIC &= \log \hat{\sigma}^2(p, q) + \frac{\log T}{T} (p + q), \end{aligned}$$

or alternatively via the likelihood.  $\hat{\sigma}^2(p, q)$  is the error variance estimated from an ARMA( $p, q$ ) model. If we increase the lag orders  $p$  and  $q$ , the first term  $\hat{\sigma}^2$  will decrease due to a closer fit, while the second term grows, a penalty for complexity. The minimum of the criteria is supposed to indicate

a good model. AIC tends to select slightly larger orders than BIC. Whereas BIC yields the true  $p$  and  $q$  for  $T \rightarrow \infty$  under certain conditions (it is a consistent selection criterion), AIC will over-estimate lag orders asymptotically. Some maintain that AIC has better properties in intermediate samples, while some authors suggest modifications in very small samples (B&D support the AICC, a modified AIC).

The textbook by HAMILTON ignores these criteria and recommends sequences of statistical hypothesis tests, for example a  $t$ -test for the coefficient  $\phi_p$  in an AR( $p$ ) model. If the null hypothesis  $\phi_p = 0$  is accepted, we prefer the AR( $p - 1$ ) model. Because of the danger of common zeros in the ARMA model, this procedure can become unreliable in some cases.

## 2.4 Estimation and testing in ARMA models

Gradually, exact descriptions of algorithms for the estimation of the parameters of ARMA models have disappeared from introductory textbooks, as modern computer software is supposed to yield reliable estimates. Most algorithms are some compromise between *least-squares* and *exact maximum likelihood* (EML). All of them assume normally distributed errors but are still consistent if this assumption is violated. There are no unbiased estimates for finite  $T$  in time-series analysis. Even the best estimation procedures have a *bias* that disappears for  $T \rightarrow \infty$ . SHAMAN & STINE derive expressions for this bias in the case of pure AR models, which possibly could be used for bias corrections. Yet, such corrected estimators are not very common.

*Least squares* directly minimizes the residual sum of squares. With AR models, the usual OLS regression can be used. MA and ARMA require iterative algorithms or nonlinear optimization, as in those models the ‘regressors’ would contain the unobserved  $\varepsilon_{t-j}$ , whose coefficients  $\theta_j$  must be estimated. Simple OLS is inconsistent in the ARMA model, even for the AR coefficients  $\phi_j$ ! In the AR model, *least squares* is easy to implement and it is efficient, if starting values are assumed as fixed. For non-stationary AR processes, this is the only sensible assumption.

EML maximizes the Gaussian likelihood under the assumption that starting values follow the stationary law according to the model. EML is more complex in its implementation and it is often only used in an approximative version. The properties of EML for non-stationary AR processes are unclear, while EML is efficient for stationary Gaussian processes.

*Yule-Walker* is a method of moments that estimates the coefficients analytically from the correlogram. For a long time, this procedure was widespread for AR models, while it is reportedly inefficient for MA and ARMA models and it breaks down completely for (nearly) non-stationary

AR processes. For stationary AR processes, *Yule-Walker* and OLS are very similar.

CML (*conditional maximum likelihood*) is another expression for *least squares*. The likelihood is maximized conditional on starting values .

All routines deliver, apart from point estimates for parameters, also estimates of standard errors and (via division)  $t$ -statistics. These  $t$ -statistics can be used for tests on the hypotheses  $\theta_j = 0$  or  $\phi_j = 0$ , but these tests should not use the  $t$ -distribution. In small samples, the distribution under the null will not be a standard distribution, while it quickly approximates the standard normal  $N(0, 1)$  distribution in large samples. This asymptotic approximation should be the basis for all tests, including the popular  $p$ -values.

On demand, also  $F$ -type tests will be offered by software packages, which may test hypotheses such as, e.g.,  $\phi_p = \theta_q = 0$ . Again, the  $F$ -type restriction statistics will not be  $F$ -distributed under their null, their analysis should rely on asymptotic  $\chi^2$ -distributions.

With regard to specification tests, the following statistics are commonly seen in computer outputs: the Jarque-Bera statistic; the Durbin-Watson statistic; portmanteau or Q statistics. Jarque-Bera tests for normality of errors and follows a  $\chi^2(2)$ -distribution under its null. The DW-statistic is not useful, as its properties in time-series models are unclear. The Q-statistic is a summary representation of the correlogram of the residuals. For stronger deviations from the null hypothesis of uncorrelated errors, it surpasses certain significance points of  $\chi^2$ -distributions. Then, this indicates the invalidity of the fitted ARMA model, as its residuals show systematic dynamic structure, which may recommend trying a higher-order ARMA model. The Q-statistic is simple and convenient, but the thus defined test has rather low power and should not be used as the main criterion in a specification search for  $p$  and  $q$ .

## References

- [1] BERGSTROM, A.R. (1990). *Continuous-Time Econometric Modelling*. Oxford University Press.
- [2] BOX, G.E.P., and JENKINS, G.C. (1976). *Time Series Analysis, Forecasting and Control*. Revised edition. Holden-Day.
- [3] BOX, G.E.P., JENKINS, G.M., and REINSEL, G.C. (1994). *Time Series Analysis, Forecasting and Control*. 3rd edition, Prentice-Hall.

- [4] BROCKWELL, P.J., and DAVIS, R.A. (1991). *Time Series: Theory and Methods*. 2nd Edition, Springer Verlag.
- [5] BROCKWELL, P.J., and DAVIS, R.A. (2002). *Introduction to Time Series and Forecasting*. 2nd Edition, Springer Verlag.
- [6] FAN, J., and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Verlag.
- [7] FULLER, W.A. (1996). *Introduction to Statistical Time Series*. 2nd Edition, Wiley.
- [8] GHYSELS, E., and OSBORN, D.R. (2001). *The Econometric Analysis of Seasonal Time Series*. Cambridge University Press.
- [9] GOURIEROUX, C., and MONFORT, A. (1997) *Time series and dynamic models*. Cambridge University Press.
- [10] GRANGER, C.W.J. and TERÄSVIRTA, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press.
- [11] HAMILTON, J.D. (1994). *Time Series Analysis*. Princeton University Press.
- [12] HARVEY, A.C. (1992). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.
- [13] HYLLEBERG, S. (1986). *Seasonality in Regression*. Academic Press.
- [14] HYLLEBERG, S., ENGLE, R.F., GRANGER, C.W.J., and YOO, B.S. (1990). Seasonal Integration and Cointegration. *Journal of Econometrics* 44, 215-238.
- [15] LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.
- [16] PRIESTLEY, M.B. (1988). *Non-Linear and Non-Stationary Time Series Analysis*. Academic Press.
- [17] SHAMAN, P., and STINE, R.A. (1988) ‘The bias of autoregressive coefficient estimators,’ *Journal of the American Statistical Association* **83**, 842–848.
- [18] TONG, H. (1990). *Nonlinear Time Series: A Dynamical Systems Approach*. Oxford University Press.