TIME SERIES

A.W. van der Vaart

Vrije Universiteit Amsterdam

PREFACE

These are lecture notes for the courses “Tijdreeksen”, “Time Series” and “Financial Time Series”. The material is more than can be treated in a one-semester course. See next section for the exam requirements.

Parts marked by an asterisk “*” do not belong to the exam requirements.

Exercises marked by a single asterisk “*” are either hard or to be considered of secondary importance. Exercises marked by a double asterisk “**” are questions to which I do not know the solution.

Amsterdam, 1995–2004 (revisions, extensions),

A.W. van der Vaart

LITERATURE

The following list is a small selection of books on time series analysis. Azencott/Dacunha-Castelle and Brockwell/Davis are close to the core material treated in these notes. The first book by Brockwell/Davis is a standard book for graduate courses for statisticians. Their second book is prettier, because it lacks the overload of formulas and computations of the first, but is of a lower level.

Chatfield is less mathematical, but perhaps of interest from a data-analysis point of view. Hannan and Deistler is tough reading, and on systems, which overlaps with time series analysis, but is not focused on statistics. Hamilton is a standard work used by econometricians; be aware, it has the existence results for ARMA processes wrong. Brillinger's book is old, but contains some material that is not covered in the later works. Rosenblatt's book is new, and also original in its choice of subjects. Harvey is a proponent of using system theory and the Kalman filter for a statistical time series analysis. His book is not very mathematical, and a good background to state space modelling.

Most books lack a treatment of developments of the last 10–15 years, such as GARCH models, stochastic volatility models, or cointegration. Mills and Gourieroux fill this gap to some extent. The first contains a lot of material, including examples fitting models to economic time series, but little mathematics. The second appears to be written for a more mathematical audience, but is not completely satisfying. For instance, its discussion of existence and stationarity of GARCH processes is incomplete, and the presentation is mathematically imprecise at many places. An alternative to these books are several review papers on volatility models, such as Bollerslev et al., Ghysels et al., and Shepard. Besides introductory discussion, also inclusing empirical evidence, these have extensive lists of references for further reading.

The book by Taniguchi and Kakizawa is unique in its emphasis on asymptotic theory, including some results on local asymptotic normality. It is valuable as a resource.

[1] Azencott, R. and Dacunha-Castelle, D., (1984). *Séries d'Observations Irrégulières.* Masson, Paris.

[2] Brillinger, D.R., (1981). *Time Series Analysis: Data Analysis and Theory.* Holt, Rinehart & Winston.

[3] Bollerslev, T., Chou, Y.C. and Kroner, K., (1992). ARCH modelling in finance: a selective review of the theory and empirical evidence. *J. Econometrics* **52**, 201–224.

[4] Bollerslev, R., Engle, R. and Nelson, D., (ARCH models). *Handbook of Econometrics IV* (eds: RF Engle and D McFadden). North Holland, Amsterdam.

[5] Brockwell, P.J. and Davis, R.A., (1991). *Time Series: Theory and Methods.* Springer.

[6] Chatfield, C., (1984). *The Analysis of Time Series: An Introduction.* Chapman and Hall.

[7] Gourieroux, C., (1997). *ARCH Models and Financial Applications.* Springer.

[8] Hall, P. and Heyde, C.C., (1980). *Martingale Limit Theory and Its Applications.* Academic Press, New York.

[9] Hamilton, J.D., (1994). *Time Series Analysis*. Princeton.

[10] Hannan, E.J. and Deistler, M., (1988). *The Statistical Theory of Linear Systems*. John Wiley, New York.

[11] Harvey, A.C., (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

[12] Mills, T.C., (1999). *The Econometric Modelling of Financial Times Series*. Cambridge University Press.

[13] Rio, A., (2002). *Mixing??*. Springer Verlag??.

[14] Rosenblatt, M., (2000). *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. Springer, New York.

[15] Taniguchi, M. and Kakizawa, Y., (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer.

CONTENTS

1
Introduction

1.1 Basic Definitions

In this course a stochastic *time series* is a doubly infinite sequence

$$\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$$

of random variables or random vectors. (Oddly enough a time series is a mathematical sequence, not a series.) We refer to the index t of X_t as *time* and think of X_t as the state or output of a stochastic system at time t, even though this is unimportant for the mathematical theory that we develop. Unless stated otherwise, the variable X_t is assumed to be real-valued, but we shall also consider series of random vectors and complex-valued variables. We write "the time series X_t" rather than using the more complete $(X_t\colon t \in \mathbb{Z})$. Instead of "time series" we may also use "process" or "stochastic process".

Of course, the set of random variables X_t, and other variables that we may introduce, are defined as measurable maps on some underlying probability space. We only make this more formal if otherwise there could be confusion, and then denote this probability space by $(\Omega, \mathcal{U}, \mathrm{P})$, with ω a typical element of Ω.

Time series theory is a mixture of probabilistic and statistical concepts. The probabilistic part is to study and characterize probability distributions of sets of variables X_t that will typically be dependent. The statistical problem is to characterize the probability distribution of the time series given observations $X_1, \ldots, X_n$ at times $1, 2, \ldots, n$. The resulting stochastic model can be used in two ways:

- understanding the stochastic system;
- predicting the "future", i.e. $X_{n+1}, X_{n+2}, \ldots,$.

In order to have any chance of success it is necessary to assume some a-priori structure of the time series. Indeed, if the X_t could be completely arbitrary random variables, then $(X_1, \ldots, X_n)$ would constitute a single observation from a completely unknown distribution on $\mathbb{R}^n$. Conclusions about this distribution would be impossible, let alone about the distribution of the future values $X_{n+1}, X_{n+2}, \ldots$.

A basic type of structure is stationarity. This comes in two forms.

1.1 Definition. *The time series X_t is strictly stationary if the distribution (on $\mathbb{R}^{h+1}$) of the vector $(X_t, X_{t+1}, \ldots, X_{t+h})$ is independent of t, for every $h \in \mathbb{N}$.*

1.2 Definition. *The time series X_t is stationary (or more precisely second order stationary) if $\mathrm{E}X_t$ and $\mathrm{E}X_{t+h}X_t$ exist and are finite and do not depend on t, for every $h \in \mathbb{N}$.*

It is clear that a strictly stationary time series with finite second moments is also stationary. For a stationary time series the *auto-covariance* and *auto-correlation* at *lag* $h \in \mathbb{Z}$ are defined by

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

$$\rho_X(h) = \rho(X_{t+h}, X_t) = \frac{\gamma_X(h)}{\gamma_X(0)}.$$

The auto-covariance and auto-correlation are functions on $\mathbb{Z}$ that together with the mean $\mu = \mathrm{E}X_t$ determine the first and second moments of the stationary time series. Note that $\gamma_X(0) = \operatorname{var} X_t$ is the variance of X_t and $\rho_X(0) = 1$.

1.3 Example (White noise). A doubly infinite sequence of independent, identically distributed random variables X_t is a strictly stationary time series. Its auto-covariance function is, with $\sigma^2 = \operatorname{var} X_t$,

$$\gamma_X(h) = \begin{cases} \sigma^2, & \text{if } h = 0, \\ 0, & \text{if } h \neq 0. \end{cases}$$

Any time series X_t with mean zero and covariance function of this type is called a *white noise* series. Thus any mean-zero i.i.d. sequence with finite variances is a white noise series. The converse is not true: there exist white noise series' that are not strictly stationary.

The name "noise" should be intuitively clear. We shall see why it is called "white" when discussing spectral theory of time series in Chapter 6.

White noise series are important building blocks to construct other series, but from the point of view of time series analysis they are not so interesting. More interesting are series where the random variables are dependent, so that, to a certain extent, the future can be predicted from the past. □

1.4 EXERCISE. Construct a white noise sequence that is not strictly stationary.

1.5 Example (Deterministic trigonometric series). Let A and B be given, uncorrelated random variables with mean zero and variance σ^2, and let λ be a given number. Then

$$X_t = A\cos(t\lambda) + B\sin(t\lambda)$$

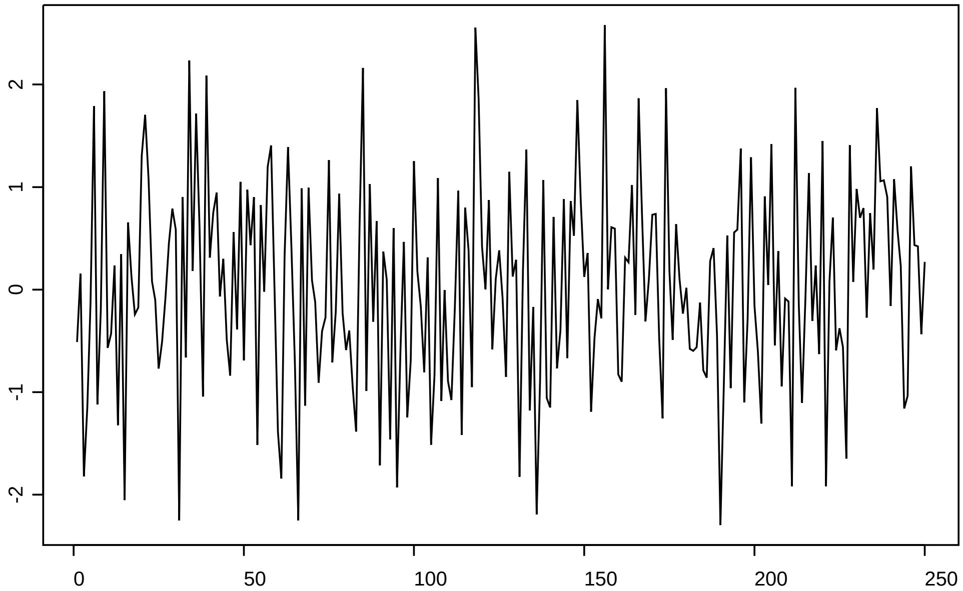

Figure 1.1. Realization of a Gaussian white noise series of length 250.

defines a stationary time series. Indeed, $\mathrm{E}X_t = 0$ and

$$\begin{aligned}\gamma_X(h) &= \operatorname{cov}(X_{t+h}, X_t)\\ &= \cos\bigl((t+h)\lambda\bigr)\cos(t\lambda)\operatorname{var} A + \sin\bigl((t+h)\lambda\bigr)\sin(t\lambda)\operatorname{var} B\\ &= \sigma^2\cos(h\lambda).\end{aligned}$$

Even though A and B are random variables, this type of time series is called *deterministic* in time series theory. Once A and B have been determined (at time $-\infty$ say), the process behaves as a deterministic trigonometric function. This type of time series is an important building block to model cyclic events in a system, but it is not the typical example of a statistical time series that we study in this course. Predicting the future is too easy in this case. □

1.6 Example (Moving average). Given a white noise series Z_t with variance σ^2 and a number θ set

$$X_t = Z_t + \theta Z_{t-1}.$$

This is called a *moving average* of order 1. The series is stationary with $\mathrm{E}X_t = 0$ and

$$\gamma_X(h) = \mathrm{cov}(Z_{t+h} + \theta Z_{t+h-1}, Z_t + \theta Z_{t-1}) = \begin{cases} (1+\theta^2)\sigma^2, & \text{if } h = 0, \\ \theta\sigma^2, & \text{if } h = \pm 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus X_s and X_t are uncorrelated whenever s and t are two or more time instants apart. We speak of *short range dependence* and say that the time series has *short memory.*

If the Z_t are an i.i.d. sequence, then the moving average is strictly stationary.

A natural generalization are higher order moving averages of the form $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$. □

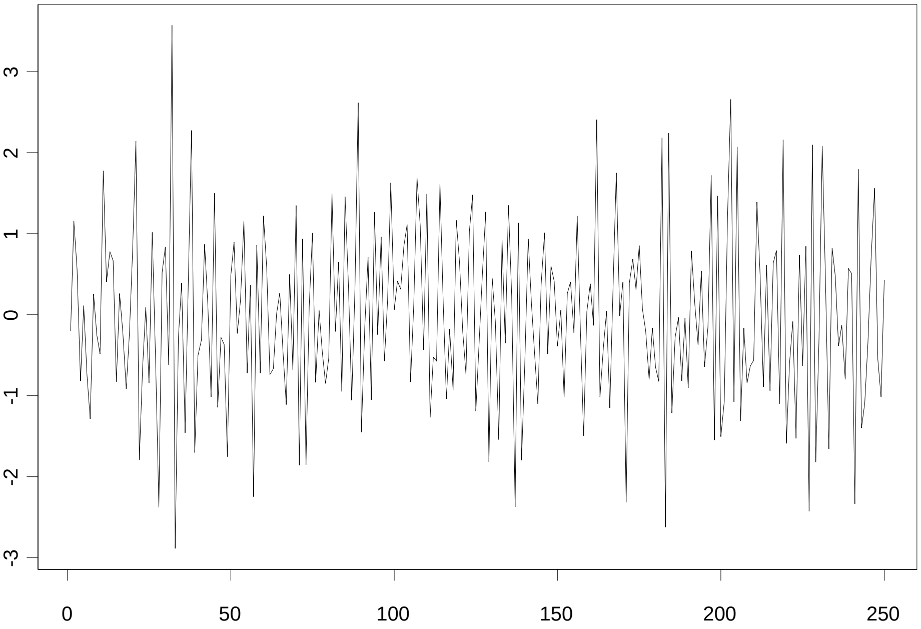

Figure 1.2. Realization of length 250 of the moving average series $X_t = Z_t - 0.5Z_{t-1}$ for Gaussian white noise Z_t.

1.7 EXERCISE. Prove that the series X_t in Example 1.6 are strictly stationary if Z_t is a strictly stationary sequence.

1.8 Example (Autoregression). Given a white noise series Z_t with variance σ^2 consider the equations

$$X_t = \theta X_{t-1} + Z_t, \qquad t \in \mathbb{Z}.$$

The white noise series Z_t is defined on some probability space $(\Omega, \mathcal{U}, \mathrm{P})$ and we consider the equation as "pointwise in ω". This equation does not define X_t, but in general has many solutions. Indeed, we can define the sequence Z_t and the variable X_0 in some arbitrary way on the given probability space and next define the remaining variables X_t for $t \in \mathbb{Z} \setminus \{0\}$ by the equation. However, suppose that we are only interested in stationary solutions. Then there is either no solution or a unique solution, depending on the value of θ, as we shall now prove.

Suppose first that $|\theta| < 1$. By iteration we find that

$$\begin{aligned} X_t &= \theta(\theta X_{t-2} + Z_{t-1}) + Z_t = \cdots \\ &= \theta^k X_{t-k} + \theta^{k-1} Z_{t-k+1} + \cdots + \theta Z_{t-1} + Z_t. \end{aligned}$$

For a stationary sequence X_t we have that $\mathrm{E}(\theta^k X_{t-k})^2 = \theta^{2k} \mathrm{E} X_0^2 \to 0$ as $k \to \infty$. This suggests that a solution of the equation is given by the infinite series

$$X_t = Z_t + \theta Z_{t-1} + \theta^2 Z_{t-2} + \cdots = \sum_{j=0}^{\infty} \theta^j Z_{t-j}.$$

We show below in Lemma 1.28 that the series on the right side converges almost surely, so that the preceding display indeed defines some random variable X_t. This is a moving average of infinite order. We can check directly, by substitution in the equation, that X_t satisfies the auto-regressive relation. (For every ω for which the series converges; hence only almost surely. We shall consider this to be good enough.)

If we are allowed to change expectations and infinite sums, then we see that

$$\begin{aligned} \mathrm{E} X_t &= \sum_{j=0}^{\infty} \theta^j \mathrm{E} Z_{t-j} = 0, \\ \gamma_X(h) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \theta^i \theta^j \mathrm{E} Z_{t+h-i} Z_{t-j} = \sum_{j=0}^{\infty} \theta^{h+j} \theta^j \sigma^2 = \frac{\theta^{|h|}}{1-\theta^2} \sigma^2. \end{aligned}$$

We prove the validity of these formulas in Lemma 1.28. It follows that X_t is indeed a stationary time series. In this case $\gamma_X(h) \neq 0$ for every h, so that every pair X_s and X_t are dependent. However, because $\gamma_X(h) \to 0$ at exponential speed as $h \to \infty$, this series is still considered to be short-range dependent. Note that $\gamma_X(h)$ oscillates if $\theta < 0$ and decreases monotonely if $\theta > 0$.

For $\theta = 1$ the situation is very different: no stationary solution exists. To see this note that the equation obtained before by iteration now takes the form, for $k = t$,

$$X_t = X_0 + Z_1 + \cdots + Z_t.$$

This implies that $\mathrm{var}(X_t - X_0) = t\sigma^2 \to \infty$ as $t \to \infty$. However, by the triangle inequality we have that

$$\mathrm{sd}(X_t - X_0) \leq \mathrm{sd}\, X_t + \mathrm{sd}\, X_0 = 2\, \mathrm{sd}\, X_0,$$

for a stationary sequence X_t. Hence no stationary solution exists. The situation for $\theta = 1$ is characterized as *explosive*: the randomness increases significantly as $t \to \infty$ due to the introduction of a new Z_t for every t.

The cases $\theta = -1$ and $|\theta| > 1$ are left as an exercise.

The auto-regressive time series of order one generalizes naturally to auto-regressive series of the form $X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p} + Z_t$. The existence of stationary solutions X_t to this equation is discussed in Chapter 7. □

1.9 EXERCISE. Consider the cases $\theta = -1$ and $|\theta| > 1$. Show that in the first case there is no stationary solution and in the second case there is a unique stationary solution. (For $|\theta| > 1$ mimic the argument for $|\theta| < 1$, but with time reversed: iterate $X_{t-1} = (1/\theta)X_t - Z_t/\theta$.)

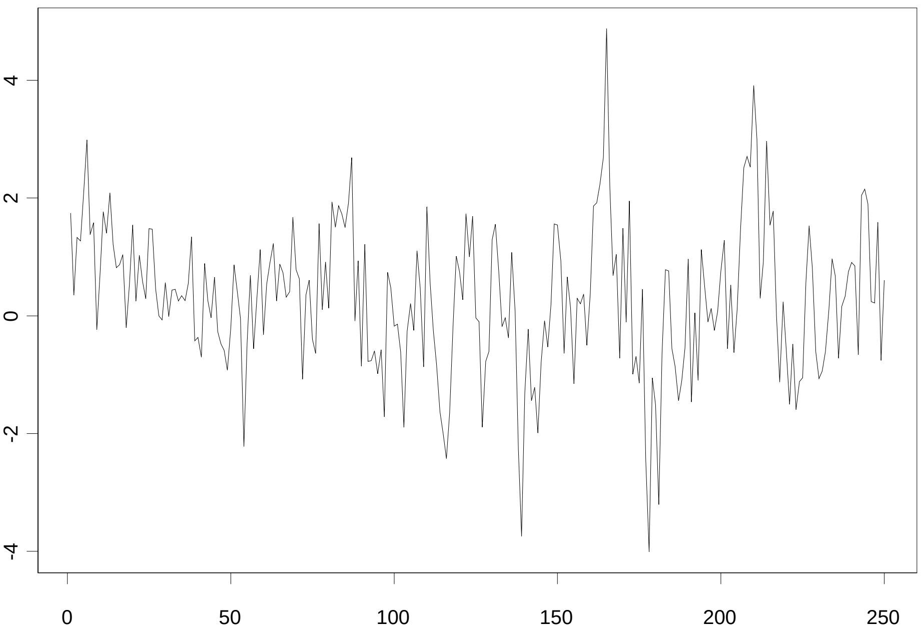

Figure 1.3. Realization of length 250 of the stationary solution to the equation $X_t = 0.5X_{t-1} + 0.2X_{t-2} + Z_t$ for Z_t Gaussian white noise.

1.10 Example (GARCH). A time series X_t is called a *GARCH(1, 1) process* if, for given nonnegative constants α, θ and ϕ, and a given i.i.d. sequence Z_t with mean zero and unit variance, it satisfies a system of equations of the form

$$\begin{aligned} \sigma_t^2 &= \alpha + \phi\sigma_{t-1}^2 + \theta X_{t-1}^2, \\ X_t &= \sigma_t Z_t. \end{aligned}$$

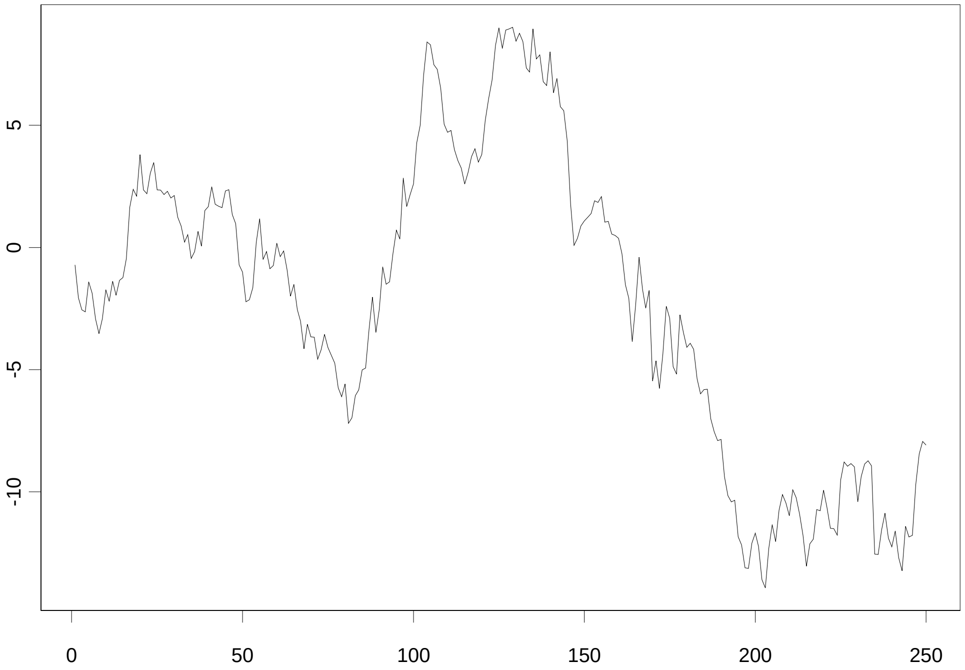

Figure 1.4. Realization of a random walk $X_t = Z_t + \cdots + Z_0$ of length 250 for Z_t Gaussian white noise.

We shall see below that for $0 \leq \theta + \phi < 1$ there exists a unique stationary solution (X_t, σ_t) to these equations and this has the further properties that σ_t^2 is a measurable function of $X_{t-1}, X_{t-2}, \ldots$, and that Z_t is independent of these variables. The latter two properties are usually also included explicitly in the requirements for a GARCH series. They imply that

$$\begin{aligned} \mathrm{E}X_t &= \mathrm{E}\sigma_t \mathrm{E}Z_t = 0, \\ \mathrm{E}X_s X_t &= \mathrm{E}(X_s \sigma_t)\mathrm{E}Z_t = 0, \qquad (s < t). \end{aligned}$$

Therefore, a stationary GARCH process with $\theta + \phi \in [0, 1)$ is a white noise process. However, it is not an i.i.d. process, unless $\theta + \phi = 0$. Because Z_t is independent of $X_{t-1}, X_{t-2}, \ldots$, and σ_t a measurable function of these variables,

$$\begin{aligned} \mathrm{E}(X_t \mid X_{t-1}, X_{t-2}, \ldots) &= \sigma_t \mathrm{E}Z_t = 0, \\ \mathrm{E}(X_t^2 \mid X_{t-1}, X_{t-2}, \ldots) &= \sigma_t^2 \mathrm{E}Z_t^2 = \sigma_t^2. \end{aligned}$$

The first equation shows that X_t is a "martingale difference series". The second exhibits σ_t^2 as the conditional variance of X_t given the past. By assumption σ_t^2 is dependent on X_{t-1} and hence the time series X_t is not i.i.d..

The abbreviation GARCH is for "generalized auto-regressive conditional heteroscedasticity": the conditional variances are not i.i.d., and depend on the past through an auto-regressive scheme. Typically, the conditional variances σ_t^2 are not directly observed, but must be inferred from the observed sequence X_t.

Because the conditional mean of X_t given the past is zero, a GARCH process will fluctuate around the value 0. A large deviation $|X_{t-1}|$ from 0 at time $t-1$ will cause a large conditional variance $\sigma_t^2 = \alpha + \theta X_{t-1}^2 + \phi\sigma_{t-1}^2$ at time t, and then the deviation of $X_t = \sigma_t Z_t$ from 0 will tend to be large as well. Similarly, small deviations from 0 will tend to be followed by other small deviations. Thus a GARCH process will alternate between periods of big fluctuations and periods of small fluctuations. This is also expressed by saying that a GARCH process exhibits *volatility clustering*, a process being "volatile" if it fluctuates a lot. Volatility clustering is commonly observed in time series of stock returns. The GARCH$(1,1)$ process has become a popular model for such time series.

The signs of the X_t are equal to the signs of the Z_t and hence will be independent over time.

Being a white noise process, a GARCH process can itself be used as input in another scheme, such as an auto-regressive or a moving average series. There are many generalizations of the GARCH process as introduced here. In a GARCH(p,q) process σ_t^2 is allowed to depend on $\sigma_{t-1}^2,\ldots,\sigma_{t-p}^2$ and $X_{t-1}^2,\ldots,X_{t-q}^2$. A GARCH $(0,q)$ process is also called an ARCH process. The rationale of using the squares X_t^2 appears to be mostly that these are nonnegative and simple; there are many variations using other functions.

As in the case of the auto-regressive relation, the two GARCH equations do not define the time series X_t, but must be complemented with an initial value, for instance σ_0^2 if we are only interested in the process for $t \geq 0$. Alternatively, we may "define" this initial value implicitly by requiring that the series X_t be stationary. We shall now show that a stationary solution exists, and is unique given the sequence Z_t.

By iterating the GARCH relation we find that, for every $n \geq 0$,

$$\sigma_t^2 = \alpha + (\phi + \theta Z_{t-1}^2)\sigma_{t-1}^2 = \alpha + \alpha\sum_{j=1}^n (\phi + \theta Z_{t-1}^2)\cdots(\phi + \theta Z_{t-j}^2) \\ + (\phi + \theta Z_{t-1}^2)\cdots(\phi + \theta Z_{t-n-1}^2)\sigma_{t-n-1}^2.$$

The sequence $\bigl((\phi+\theta Z_{t-1}^2)\cdots(\phi+\theta Z_{t-n-1}^2)\bigr)_{n=1}^\infty$, which consists of nonnegative variables with means $(\phi+\theta)^{n+1}$, converges in probability to zero if $\theta+\phi<1$. If the time series σ_t^2 is stationary, then the term on the far right converges to zero in probability as $n \to \infty$. Thus for a stationary solution (X_t, σ_t) we must have

$$\sigma_t^2 = \alpha + \alpha\sum_{j=1}^\infty (\phi + \theta Z_{t-1}^2)\cdots(\phi + \theta Z_{t-j}^2). \tag{1.1}$$

Because the series Z_t is assumed i.i.d., the variable Z_t is independent of $\sigma_{t-1}^2, \sigma_{t-2}^2, \ldots$ and also of $X_{t-1} = \sigma_{t-1}Z_{t-1}, X_{t-2} = \sigma_{t-2}Z_{t-2}, \ldots$. In addition it follows that the time series $X_t = \sigma_t Z_t$ is strictly stationary, being a fixed measurable transformation of $(Z_t, Z_{t-1}, \ldots)$ for every t.

The infinite sum in (1.1) converges in mean if $\theta+\phi<1$ (Cf. Lemma 1.26). Given the series Z_t we can define a process X_t by first defining the conditional variance σ_t^2 by

(1.1), and next setting $X_t = \sigma_t Z_t$. It can be verified by substitution that the process X_t solves the GARCH relationship and hence a stationary solution to the GARCH equations exists if $\phi + \theta < 1$.

By iterating the auto-regressive relation $\sigma_t^2 = \phi\sigma_{t-1}^2 + W_t$, with $W_t = \alpha + \theta X_{t-1}^2$, in the same way as in Example 1.8, we also find that for the stationary solution $\sigma_t^2 = \sum_{j=0}^\infty \phi^j W_{t-j}$. Hence σ_t is $\sigma(X_{t-1}, X_{t-2}, \ldots)$-measurable.

An inspection of the preceding argument shows that a strictly stationary solution exists under a weaker condition than $\phi+\theta < 1$. If the sequence $(\phi+\theta Z_{t-1}^2)\cdots(\phi+\theta Z_{t-n}^2)$ converges to zero in in probability as $n \to \infty$ and X_t is a solution of the GARCH relation such that σ_t^2 is bounded in probability as $t \to -\infty$, then the same argument shows that σ_t^2 must relate to the Z_t as given. Furthermore, if the series on the right side of (1.1) converges in probability, then X_t may be defined as before. It can be shown that this is the case under the condition that $\mathrm{E}\log(\phi + \theta Z_t^2) < 0$. (See Exercise 1.14 or Chapter 8.) For instance, for standard normal variables Z_t and $\phi = 0$ this reduces to $\theta < 2e^\gamma \approx 3.56$. On the other hand, the condition $\phi + \theta < 1$ is necessary for the GARCH process to have finite second moments. □

1.11 EXERCISE. Let $\theta + \phi \in [0, 1)$ and $1 - \overline{\kappa}\theta^2 - \phi^2 - 2\theta\phi > 0$, where $\overline{\kappa} = \mathrm{E}Z_t^4$. Show that the second and fourth (marginal) moments of a stationary GARCH process are given by $\alpha/(1 - \theta - \phi)$ and $\overline{\kappa}\alpha^2(1 + \theta + \phi)/(1 - \overline{\kappa}\theta^2 - \phi^2 - 2\theta\phi)(1 - \theta - \phi)$. From this compute the kurtosis of the GARCH process with standard normal Z_t. [You can use (1.1), but it is easier to use the GARCH relations.]

1.12 EXERCISE. Show that $\mathrm{E}X_t^4 = \infty$ if $1 - \overline{\kappa}\theta^2 - \phi^2 - 2\theta\phi = 0$.

1.13 EXERCISE. Suppose that the process X_t is square-integrable and satisfies the GARCH relation for an i.i.d. sequence Z_t such that Z_t is independent of $X_{t-1}, X_{t-2}, \ldots$ and such that $\sigma_t^2 = \mathrm{E}(X_t^2|\, X_{t-1}, X_{t-2}, \ldots)$, for every t, and some $\alpha, \phi, \theta > 0$. Show that $\phi + \theta < 1$. [Derive that $\mathrm{E}X_t^2 = \alpha + \alpha\sum_{j=1}^n (\phi + \theta)^j + (\phi + \theta)^{n+1}\mathrm{E}X_{t-n-1}^2$.]

1.14 EXERCISE. Let Z_t be an i.i.d. sequence with $\mathrm{E}\log(Z_t^2) < 0$. Show that $\sum_{j=0}^\infty Z_t^2 Z_{t-1}^2 \cdots Z_{t-j}^2 < \infty$ almost surely. [By the law of large numbers there exists for almost every realization of Z_t a number N such that $n^{-1}\sum_{j=1}^n \log Z_j^2 < c < 0$ for every $n \geq N$. Show that this implies that $\sum_{n\geq N} Z_t^2 Z_{t-1}^2 \cdots Z_{t-j}^2 < \infty$ almost surely.]

1.15 Example (Stochastic volatility). A general approach to obtain a time series with volatility clustering is to define $X_t = \sigma_t Z_t$ for an i.i.d. sequence Z_t and a process σ_t that depends "positively on its past". A GARCH model fits this scheme, but a simpler way to achieve the same aim is to let σ_t depend only on its own past and independent noise. Because σ_t is to have an interpretation as a scale parameter, we restrain it to be positive. One way to combine these requirements is to set

$$\begin{aligned} h_t &= \theta h_{t-1} + W_t, \\ \sigma_t^2 &= e^{h_t}, \\ X_t &= \sigma_t Z_t. \end{aligned}$$

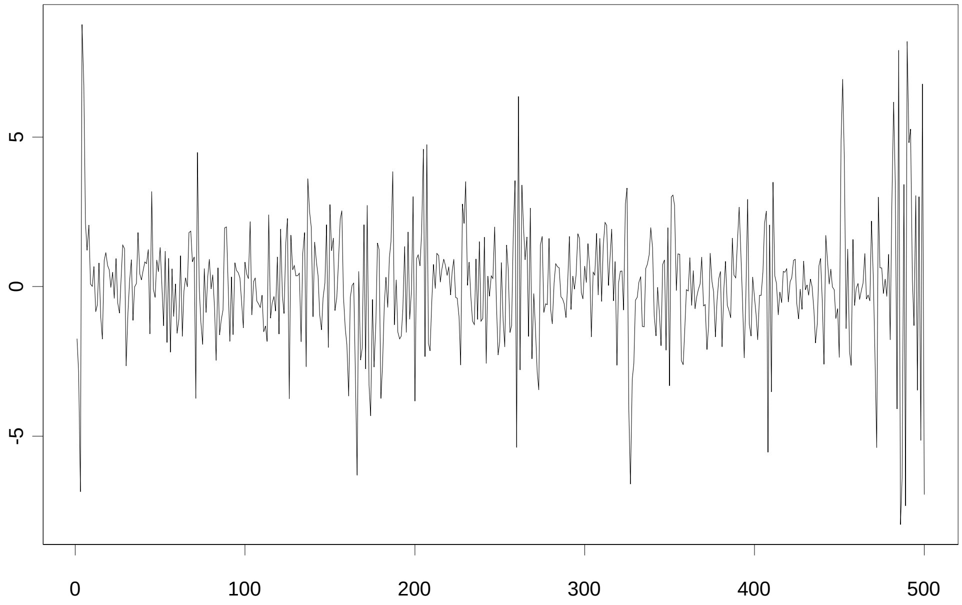

Figure 1.5. Realization of the GARCH$(1,1)$ process with $\alpha = 0.1$, $\phi = 0$ and $\theta = 0.8$ of length 500 for Z_t Gaussian white noise.

Here W_t is a white noise sequence, h_t is a (stationary) solution to the auto-regressive equation, and the process Z_t is i.i.d. and independent of the process W_t. If $\theta > 0$ and $\sigma_{t-1} = e^{h_{t-1}/2}$ is large, then $\sigma_t = e^{h_t/2}$ will tend to be large as well, and hence the process X_t will exhibit volatility clustering.

The process h_t will typically not be observed and for that reason is sometimes called *latent*. A "stochastic volatility process" of this type is an example of a (nonlinear) state space model, discussed in Chapter 9. Rather than defining σ_t by an auto-regression in the exponent, we may choose a different scheme. For instance, an EGARCH$(p, 0)$ model postulates the relationship

$$\log \sigma_t = \alpha + \sum_{j=1}^{p} \phi_j \log \sigma_{t-j}.$$

This is not a stochastic volatility model, because it does not include a random disturbance. The symmetric EGARCH (p, q) model repairs this by adding terms depending on the past of the observed series $X_t = \sigma_t Z_t$, giving

$$\log \sigma_t = \alpha + \sum_{j=1}^{q} \theta_j |Z_{t-j}| + \sum_{j=1}^{p} \phi_j \log \sigma_{t-j}.$$

In this sense GARCH processes and their variants are much related to stochastic volatility models. In view of the recursive nature of the definitions of σ_t and X_t, they are perhaps more complicated. □

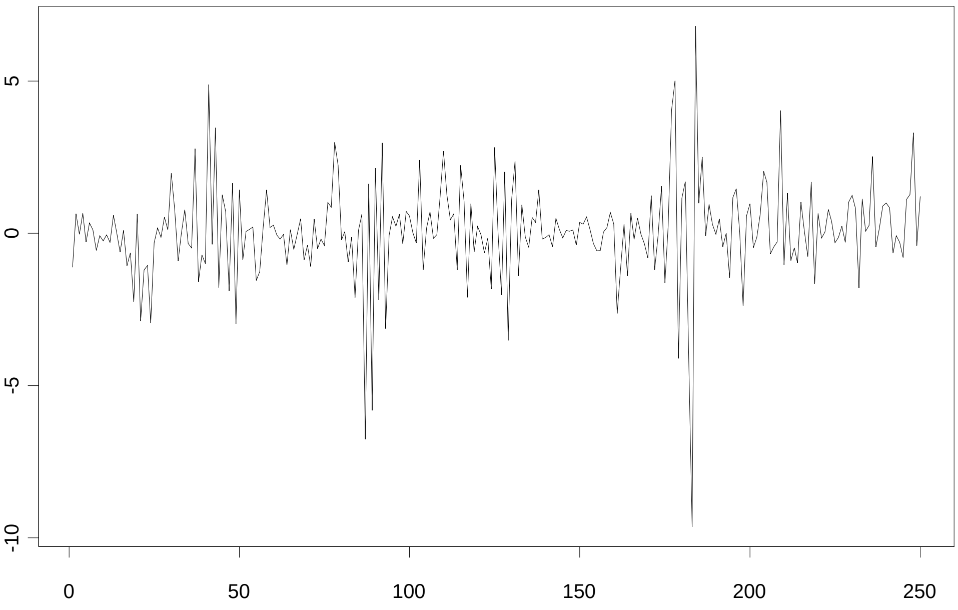

Figure 1.6. Realization of length 250 of the stochastic volatility model $X_t = e^{h_t/2}Z_t$ for a standard Gaussian i.i.d. process Z_t and a stationary auto-regressive process $h_t = 0.8h_{t-1} + W_t$ for a standard Gaussian i.i.d. process W_t.

1.2 Filters

Many time series in real life are not stationary. Rather than modelling a nonstationary sequence, such a sequence is often transformed in one or more time series that are (assumed to be) stationary. The statistical analysis next focuses on the transformed series.

Two important deviations from stationarity are *trend* and *seasonality.* A trend is a long term, steady increase or decrease in the general level of the time series. A seasonal component is a cyclic change in the level, the cycle length being for instance a year or a week. Even though Example 1.5 shows that a perfectly cyclic series can be modelled as a stationary series, it is often considered wise to remove such perfect cycles from a given series before applying statistical techniques.

There are many ways in which a given time series can be transformed in a series that is easier to analyse. Transforming individual variables X_t into variables $f(X_t)$ by a fixed function f (such as the logarithm) is a common technique as is *detrending* by substracting a "best fitting polynomial in t" of some fixed degree. This is commonly found by the method of least squares: given a nonstationary time series X_t we determine constants $\beta_0, \ldots, \beta_p$ by minimizing

$$(\beta_0, \ldots, \beta_p) \mapsto \sum_{t=1}^{n} \bigl(X_t - \beta_0 - \beta_1 t - \cdots - \beta_p t^p\bigr)^2.$$

Next the time series $X_t - \beta_0 - \beta_1 t - \cdots - \beta_p t^p$, for the minimizing coefficients $\beta_0, \ldots, \beta_p$, is assumed to be stationary.

A standard transformation for financial time series is to *(log) returns*, given by

$$\log \frac{X_t}{X_{t-1}}, \qquad \text{or} \qquad \frac{X_t}{X_{t-1}} - 1.$$

If X_t/X_{t-1} is close to unity for all t, then these transformations are similar, as $\log x \approx x - 1$ for $x \approx 1$. Because $\log(e^{ct}/e^{c(t-1)}) = c$, a log return can be intuitively interpreted as the exponent of exponential growth. Many financial time series exhibit an exponential trend.

A general method to transform a nonstationary sequence in a stationary one, advocated with much success in a famous book by Box and Jenkins, is filtering.

1.16 Definition. *The (linear) filter with filter coefficients ψ_j for $j \in \mathbb{Z}$ is the operation that transforms a given time series X_t into the time series $Y_t = \sum_{j\in\mathbb{Z}} \psi_j X_{t-j}$.*

A linear filter is a moving average of infinite order. In Lemma 1.28 we give conditions for the infinite series to be well defined. All filters used in practice are finite filters: only finitely many coefficients are nonzero. Important examples are the *difference filter* $\nabla X_t = X_t - X_{t-1}$, its repetitions $\nabla^k X_t = \nabla\nabla^{k-1} X_t$ defined recursely for $k = 2, 3, \ldots$, and the *seasonal difference filter* $\nabla_k X_t = X_t - X_{t-k}$.

1.17 Example (Polynomial trend). A *linear trend model* could take the form $X_t = at + Z_t$ for a strictly stationary time series Z_t. If $a \neq 0$, then the time series X_t is not stationary in the mean. However, the differenced series $\nabla X_t = a + Z_t - Z_{t-1}$ is stationary.

Thus differencing can be used to remove a linear trend. Similarly, a polynomial trend can be removed by repeated differencing: a polynomial trend of degree k is removed by applying ∇^k. □

1.18 EXERCISE. Check this for a series of the form $X_t = at + bt^2 + Z_t$.

1.19 EXERCISE. Does a (repeated) seasonal filter also remove polynomial trend?

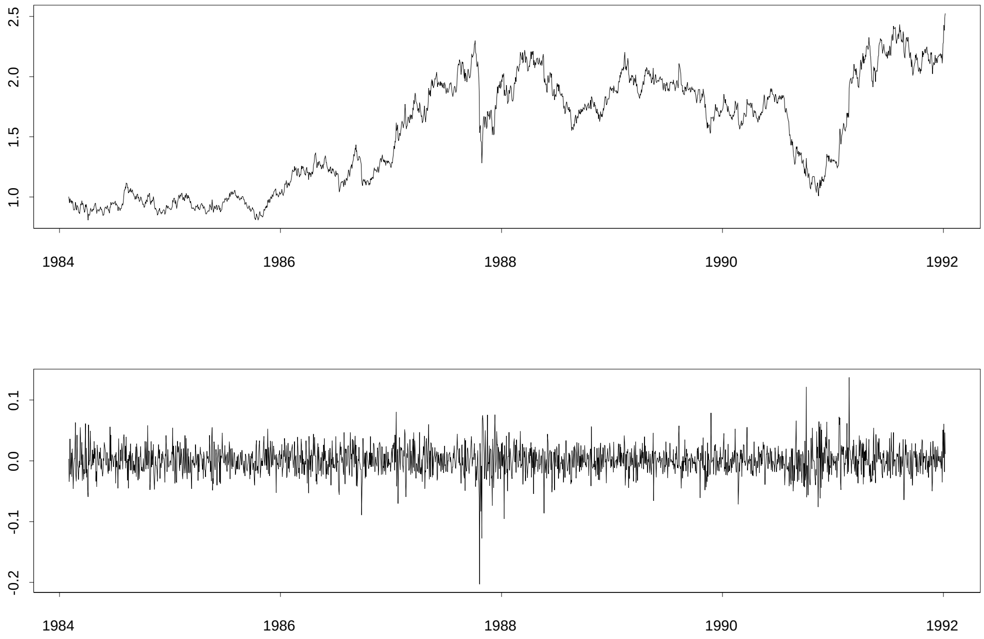

Figure 1.7. Prices of Hewlett Packard on New York Stock Exchange and corresponding log returns.

1.20 Example (Random walk). A *random walk* is defined as the sequence of partial sums $X_t = Z_1 + Z_2 + \cdots + Z_t$ of an i.i.d. sequence Z_t. A random walk is not stationary, but the differenced series $\nabla X_t = Z_t$ certainly is. □

1.21 Example (Monthly cycle). If X_t is the value of a system in month t, then $\nabla_{12} X_t$ is the change in the system during the past year. For seasonable variables without trend this series might be modelled as stationary. For series that contain both yearly seasonality and trend, the series $\nabla^k \nabla_{12} X_t$ might be stationary. □

1.22 Example (Weekly cycle). If X_t is the value of a system at day t, then $Y_t = (1/7)\sum_{j=0}^{6} X_{t-j}$ is the average value over the last week. This series might show trend, but should not show seasonality due to day-of-the-week. We could study seasonality by considering the time series $X_t - Y_t$, which results from filtering the series X_t with coefficients $(\psi_0, \ldots, \psi_6) = (6/7, -1/7, \ldots, -1/7)$. □

1.23 Example (Exponential smoothing). An ad-hoc method for predicting the future is to equate the future to the present or, more generally, to the average of the last k observed values of a time series. When averaging past values it is natural to gives more weight to the most recent values. Exponentially decreasing weights appear to have some

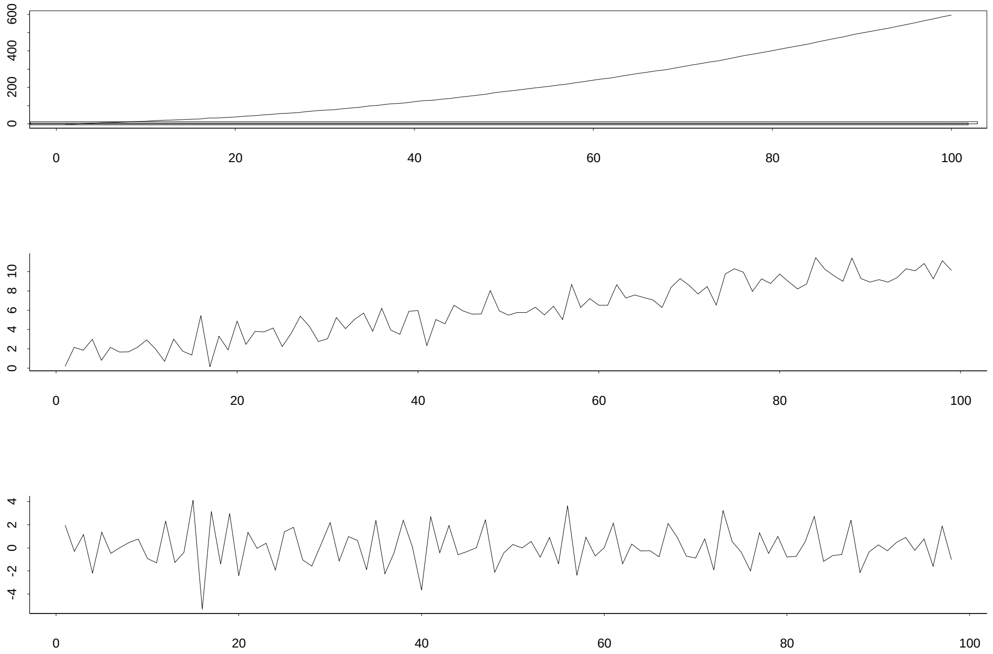

Figure 1.8. Realization of the time series $t + 0.05t^2 + X_t$ for the stationary auto-regressive process X_t satisfying $X_t - 0.8X_{t-1} = Z_t$ for Gaussian white noise Z_t, and the same series after once and twice differencing.

popularity. This corresponds to predicting a future value of a time series X_t by the weighted average $\sum_{j=0}^{\infty} \theta^j/(1-\theta)X_{t-j}$ for some $\theta \in (0,1)$. □

1.24 EXERCISE. Show that the result of two filters with coefficients α_j and β_j applied in turn (if well defined) is the filter with coefficients γ_j given by $\gamma_k = \sum_j \alpha_j \beta_{k-j}$. This is called the *convolution* of the two filters. Infer that filtering is commutative.

1.25 Definition. *A filter with coefficients ψ_j is causal if $\psi_j = 0$ for every $j < 0$.*

For a causal filter the variable $Y_t = \sum_j \psi_j X_{t-j}$ depends only on the values $X_t, X_{t-1}, \ldots$ of the original time series in the present and past, not the future. This is important for prediction. Given X_t up to some time t, we can calculate Y_t up to time t. If Y_t is stationary, we can use results for stationary time series to predict the future value Y_{t+1}. Next we predict the future value X_{t+1} by $X_{t+1} = \psi_0^{-1}(Y_{t+1} - \sum_{j>0} \psi_j X_{t+1-j})$.

In order to derive conditions that guarantee that an infinite filter is well defined, we start with a lemma concerning series' of random variables. Recall that a series $\sum_t x_t$ of nonnegative numbers is always well defined (although possibly ∞), where the order of summation is irrelevant. Furthermore, for general numbers x_t the absolute convergence

$\sum_t |x_t| < \infty$ implies that $\sum_t x_t$ exists as a finite number, where the order of summation is again irrelevant. We shall be concerned with series indexed by $t \in \mathbb{N}$, $t \in \mathbb{Z}$, $t \in \mathbb{Z}^2$, or t contained in some other countable set T. It follows from the preceding that $\sum_{t\in T} x_t$ is well defined as a limit as $n \to \infty$ of partial sums $\sum_{t \in T_n} x_t$, for any increasing sequence of finite subsets $T_n \subset T$ with union T, if either every x_t is nonnegative or $\sum_t |x_t| < \infty$. For instance, in the case that the index set T is equal to $\mathbb{Z}$, we can choose the sets $T_n = \{t \in \mathbb{Z}: |t| \le n\}$.

1.26 Lemma. *Let $(X_t: t \in T)$ be an arbitrary countable set of random variables.*

(i) If $X_t \ge 0$ for every t, then $\mathrm{E}\sum_t X_t = \sum_t \mathrm{E}X_t$ (possibly $+\infty$);

(ii) If $\sum_t \mathrm{E}|X_t| < \infty$, then the series $\sum_t X_t$ converges absolutely almost surely and $\mathrm{E}\sum_t X_t = \sum_t \mathrm{E}X_t$.

Proof. Suppose $T = \cup_j T_j$ for an increasing sequence $T_1 \subset T_2 \subset \cdots$ of finite subsets of T. Assertion (i) follows from the monotone convergence theorem applied to the variables $Y_j = \sum_{t \in T_j} X_t$. The second part of assertion (ii) follows from the dominated convergence theorem applied to the same variables Y_j. These are dominated by $\sum_t |X_t|$, which is integrable because its expectation can be computed as $\sum_t \mathrm{E}|X_t|$ by (i). The first assertion of (ii) follows because $\mathrm{E}\sum_t |X_t| < \infty$. ■

The dominated convergence theorem in the proof of (ii) actually gives a better result, namely: if $\sum_t \mathrm{E}|X_t| < \infty$, then

$$\mathrm{E}\Big|\sum_{t\in T} X_t - \sum_{t \in T_j} X_t\Big| \to 0, \qquad \text{if } T_1 \subset T_2 \subset \cdots \uparrow T.$$

This is called the *convergence in mean* of the series $\sum_t X_t$. The analogous convergence of the second moment is called the *convergence in second mean.* Alternatively, we speak of "convergence in quadratic mean" or "convergence in L_1 or L_2".

1.27 EXERCISE. Suppose that $\mathrm{E}|X_n - X|^p \to 0$ and $\mathrm{E}|X|^p < \infty$ for some $p \ge 1$. Show that $\mathrm{E}X_n^k \to \mathrm{E}X^k$ for every $0 < k \le p$.

1.28 Lemma. *Let $(Z_t: t \in \mathbb{Z})$ be an arbitrary time series and let $\sum_j |\psi_j| < \infty$.*

(i) If $\sup_t \mathrm{E}|Z_t| < \infty$, then $\sum_j \psi_j Z_{t-j}$ converges absolutely, almost surely and in mean.

(ii) If $\sup_t \mathrm{E}|Z_t|^2 < \infty$, then $\sum_j \psi_j Z_{t-j}$ converges in second mean as well.

(iii) If the series Z_t is stationary, then so is the series $X_t = \sum_j \psi_j Z_{t-j}$ and $\gamma_X(h) = \sum_l \sum_j \psi_j \psi_{j+l-h} \gamma_Z(l)$.

Proof. (i). Because $\sum_t \mathrm{E}|\psi_j Z_{t-j}| \le \sup_t \mathrm{E}|Z_t| \sum_j |\psi_j| < \infty$, it follows by (ii) of the preceding lemma that the series $\sum_j \psi_j Z_t$ is absolutely convergent, almost surely. The convergence in mean follows as in the remark following the lemma.

(ii). By (i) the series is well defined almost surely, and $\sum_j \psi_j Z_{t-j} - \sum_{|j|\le k} \psi_j Z_{t-j} = \sum_{|j|>k} \psi_j Z_{t-j}$. By the triangle inequality we have

$$\Big|\sum_{|j|>k} \psi_j Z_{t-j}\Big|^2 \le \Big(\sum_{|j|>k} |\psi_j Z_{t-j}|\Big)^2 = \sum_{|j|>k}\sum_{|i|>k} |\psi_j||\psi_i||Z_{t-j}||Z_{t-i}|.$$

By the Cauchy-Schwarz inequality $\mathrm{E}|Z_{t-j}||Z_{t-i}| \leq \bigl(\mathrm{E}|Z_{t-j}|^2|\mathrm{E}Z_{t-i}|^2\bigr)^{1/2}$, which is bounded by $\sup_t \mathrm{E}|Z_t|^2$. Therefore, in view of (i) of the preceding lemma the expectation of the left side of the preceding display is bounded above by

$$\sum_{|j|>k}\sum_{|i|>k} |\psi_j||\psi_i| \sup_t \mathrm{E}|Z_t|^2 = \Bigl(\sum_{|j|>k} |\psi_j|\Bigr)^2 \sup_t \mathrm{E}|Z_t|^2.$$

This converges to zero as $k \to \infty$.

(iii). By (i) the series $\sum_j \psi_j Z_{t-j}$ converges in mean. Therefore, $\mathrm{E}\sum_j \psi_j Z_{t-j} = \sum_j \psi_j \mathrm{E}Z_t$, which is independent of t. Using arguments as before, we see that we can also justify the interchange of the order of expectations (hidden in the covariance) and double sums in

$$\begin{aligned}\gamma_X(h) &= \mathrm{cov}\Bigl(\sum_j \psi_j X_{t+h-j}, \sum_i \psi_i X_{t-i}\Bigr)\\ &= \sum_j\sum_i \psi_j\psi_i \,\mathrm{cov}(Z_{t+h-j}, Z_{t-i}) = \sum_j\sum_i \psi_j\psi_i\gamma_Z(h-j+i).\end{aligned}$$

This can be written in the form given by the lemma by the change of variables $(j,i) \mapsto (j, l-h+j)$. ■

1.29 EXERCISE. Suppose that the series Z_t in (iii) is strictly stationary. Show that the series X_t is strictly stationary whenever it is well defined.

* **1.30** EXERCISE. For a white noise series Z_t, part (ii) of the preceding lemma can be improved: Suppose that Z_t is a white noise sequence and $\sum_j \psi_j^2 < \infty$. Show that $\sum_j \psi_j Z_{t-j}$ converges in second mean. (For this exercise you need some of the material of Chapter 2.)

1.3 Complex Random Variables

Even though no real-life time series is complex valued, the use of complex numbers is notationally convenient to develop the mathematical theory. In this section we discuss complex-valued random variables.

A *complex random variable* Z is a map from some probability space into the field of complex numbers whose real and imaginary parts are random variables. For complex random variables $Z = X + iY$, Z_1 and Z_2, we define

$$\begin{aligned}\mathrm{E}Z &= \mathrm{E}X + i\mathrm{E}Y,\\ \mathrm{var}\, Z &= \mathrm{E}|Z - \mathrm{E}Z|^2,\\ \mathrm{cov}(Z_1, Z_2) &= \mathrm{E}(Z_1 - \mathrm{E}Z_1)\overline{(Z_2 - \mathrm{E}Z_2)}.\end{aligned}$$

Some simple properties are, for $\alpha, \beta \in \mathbb{C}$,

$$\mathrm{E}\alpha Z = \alpha \mathrm{E}Z, \qquad \mathrm{E}\overline{Z} = \overline{\mathrm{E}Z},$$
$$\operatorname{var} Z = \mathrm{E}|Z|^2 - |\mathrm{E}Z|^2 = \operatorname{var} X + \operatorname{var} Y = \operatorname{cov}(Z, Z),$$
$$\operatorname{var}(\alpha Z) = |\alpha|^2 \operatorname{var} Z,$$
$$\operatorname{cov}(\alpha Z_1, \beta Z_2) = \alpha\overline{\beta} \operatorname{cov}(Z_1, Z_2),$$
$$\operatorname{cov}(Z_1, Z_2) = \overline{\operatorname{cov}(Z_2, Z_1)} = \mathrm{E}Z_1\overline{Z}_2 - \mathrm{E}Z_1\mathrm{E}\overline{Z}_2.$$

1.31 EXERCISE. Prove the preceding identities.

The definitions given for real time series apply equally well to complex time series. Lemma 1.28 also extends to complex time series Z_t, where in (iii) we must read $\gamma_X(h) = \sum_l \sum_j \psi_j \overline{\psi}_{j+l-h} \gamma_Z(l)$.

1.32 EXERCISE. Show that the auto-covariance function of a complex stationary time series Z_t is conjugate symmetric: $\gamma_Z(-h) = \overline{\gamma_Z(h)}$ for every $h \in \mathbb{Z}$.

1.4 Multivariate Time Series

In many applications the interest is in the time evolution of several variables jointly. This can be modelled through vector-valued time series. The definition of a stationary time series applies without changes to vector-valued series $X_t = (X_{t,1}, \ldots, X_{t,d})$. Here the mean $\mathrm{E}X_t$ is understood to be the vector $(\mathrm{E}X_{t,1}, \ldots, X_{t,d})$ of means of the coordinates and the auto-covariance function is defined to be the matrix

$$\gamma_X(h) = \Bigl(\operatorname{cov}(X_{t+h,i}, X_{t,j})\Bigr)_{i,j=1,\ldots,d} = \mathrm{E}(X_{t+h} - \mathrm{E}X_{t+h})(\overline{X_t - \mathrm{E}X_t})^T.$$

The auto-correlation at lag h is defined as

$$\rho_X(h) = \Bigl(\rho(X_{t+h,i}, X_{t,j})\Bigr)_{i,j=1,\ldots,d} = \Bigl(\frac{\gamma_X(h)_{i,j}}{\sqrt{\gamma_X(0)_{i,i}\gamma_X(0)_{j,j}}}\Bigr)_{i,j=1,\ldots,d}.$$

The study of properties of multivariate time series can often be reduced to the study of univariate time series by taking linear combinations $a^T X_t$ of the coordinates. The first and second moments satisfy

$$\mathrm{E}a^T X_t = a^T \mathrm{E}X_t, \qquad \gamma_{a^T X}(h) = a^T \gamma_X(h)\overline{a}.$$

1.33 EXERCISE. What is the relationship between $\gamma_X(h)$ and $\gamma_X(-h)$?

2
Hilbert Spaces and Prediction

In this chapter we first recall definitions and basic facts concerning Hilbert spaces. Next we apply these to solve the *prediction problem*: finding the "best" predictor of X_{n+1} based on observations $X_1, \ldots, X_n$.

2.1 Hilbert Spaces and Projections

Given a measure space $(\Omega, \mathcal{U}, \mu)$ define $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ as the set of all measurable functions $f: \Omega \to \mathbb{C}$ such that $\int |f|^2\, d\mu < \infty$. (Alternatively, all measurable functions with values in $\mathbb{R}$ with this property.) Here a complex-valued function is said to be measurable if both its real and imaginary parts are measurable functions, and its integral is by definition $\int f\, d\mu = \int \operatorname{Re} f\, d\mu + i \int \operatorname{Im} f\, d\mu$, provided the two integrals on the right are defined and finite. Define

$$\langle f_1, f_2 \rangle = \int f_1 \overline{f}_2\, d\mu,$$

$$\|f\| = \sqrt{\int |f|^2\, d\mu},$$

$$d(f_1, f_2) = \|f_1 - f_2\| = \sqrt{\int |f_1 - f_2|^2\, d\mu}.$$

These define a semi-inner product, a semi-norm, and a semi-metric, respectively. The first is a *semi-inner product* in view of the properties:

$$\begin{aligned}
\langle f_1 + f_2, f_3 \rangle &= \langle f_1, f_3 \rangle + \langle f_2, f_3 \rangle, \\
\langle \alpha f_1, \beta f_2 \rangle &= \alpha \overline{\beta} \langle f_1, f_2 \rangle, \\
\langle f_2, f_1 \rangle &= \overline{\langle f_1, f_2 \rangle}, \\
\langle f, f \rangle &\geq 0, \quad \text{with equality iff } f = 0, \text{ a.e..}
\end{aligned}$$

The second is a *semi-norm* because it has the properties:

$$\begin{aligned}\|f_1 + f_2\| &\le \|f_1\| + \|f_2\|,\\ \|\alpha f\| &= |\alpha|\|f\|,\\ \|f\| &= 0 \quad \text{iff } f = 0, \text{a.e.}.\end{aligned}$$

Here the first line, the *triangle inequality* is not immediate, but it can be proved with the help of the Cauchy-Schwarz inequality, given below. The other properties are more obvious. The third is a *semi-distance*, in view of the relations:

$$\begin{aligned}&d(f_1, f_3) \le d(f_1, f_2) + d(f_2, f_3),\\ &d(f_1, f_2) = d(f_2, f_1),\\ &d(f_1, f_2) = 0 \text{ iff } f_1 = f_2, \text{a.e.}.\end{aligned}$$

Immediate consequences of the definitions and the properties of the inner product are

$$\begin{aligned}\|f+g\|^2 = \langle f+g, f+g\rangle &= \|f\|^2 + \langle f, g\rangle + \langle g, f\rangle + \|g\|^2,\\ \|f+g\|^2 &= \|f\|^2 + \|g\|^2, \qquad \text{if } \langle f, g\rangle = 0.\end{aligned}$$

The last equality is known as the *Pythagorean rule*. In the complex case this is true, more generally, if $\mathrm{Re}\langle f, g\rangle = 0$.

2.1 Lemma (Cauchy-Schwarz). *Any pair f, g in $L_2(\Omega, \mathcal{U}, \mu)$ satisfies $\bigl|\langle f, g\rangle\bigr| \le \|f\|\|g\|$.*

Proof. For real-valued functions this follows upon working out the inequality $\|f - \lambda g\|^2 \ge 0$ for $\lambda = \langle f, g\rangle/\|g\|^2$. In the complex case we write $\langle f, g\rangle = \bigl|\langle f, g\rangle\bigr| e^{i\theta}$ for some $\theta \in \mathbb{R}$ and work out $\|f - \lambda e^{i\theta} g\|^2 \ge 0$ for the same choice of λ. ∎

Now the triangle inequality for the norm follows from the preceding decomposition of $\|f+g\|^2$ and the Cauchy-Schwarz inequality, which, when combined, yield

$$\|f+g\|^2 \le \|f\|^2 + 2\|f\|\|g\| + \|g\|^2 = \bigl(\|f\| + \|g\|\bigr)^2.$$

Another consequence of the Cauchy-Schwarz inequality is the continuity of the inner product:

$$f_n \to f, g_n \to g \text{ implies that } \langle f_n, g_n\rangle \to \langle f, g\rangle.$$

2.2 EXERCISE. Prove this.

2.3 EXERCISE. Prove that $\bigl|\|f\| - \|g\|\bigr| \le \|f - g\|$.

2.4 EXERCISE. Derive the *parallellogram rule*: $\|f+g\|^2 + \|f-g\|^2 = 2\|f\|^2 + 2\|g\|^2$.

2.5 EXERCISE. Prove that $\|f + ig\|^2 = \|f\|^2 + \|g\|^2$ for every pair f, g of real functions in $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$.

2.6 EXERCISE. Let $\Omega = \{1, 2, \ldots, k\}$, $\mathcal{U} = 2^\Omega$ the power set of Ω and μ the counting measure on Ω. Show that $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ is exactly $\mathbb{C}^k$ (or $\mathbb{R}^k$ in the real case).

We attached the qualifier "semi" to the inner product, norm and distance defined previously, because in every of the three cases, the last property involves a null set. For instance $\|f\| = 0$ does not imply that $f = 0$, but only that $f = 0$ almost everywhere. If we think of two functions that are equal almost everywere as the same "function", then we obtain a true inner product, norm and distance. We define $L_2(\Omega, \mathcal{U}, \mu)$ as the set of all equivalence classes in $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ under the equivalence relation "$f \equiv g$ if and only if $f = g$ almost everywhere". It is a common abuse of terminology, which we adopt as well, to refer to the equivalence classes as "functions".

2.7 Proposition. *The metric space $L_2(\Omega, \mathcal{U}, \mu)$ is complete under the metric d.*

We shall need this proposition only occasionally, and do not provide a proof. (See e.g. Rudin, Theorem 3.11.) The proposition asserts that for every sequence f_n of functions in $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ such that $\int |f_n - f_m|^2 \, d\mu \to$ as $m, n \to \infty$ (a *Cauchy sequence*), there exists a function $f \in \mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ such that $\int |f_n - f|^2 \, d\mu \to 0$ as $n \to \infty$.

A *Hilbert space* is a general inner product space that is metrically complete. The space $L_2(\Omega, \mathcal{U}, \mu)$ is an example, and the only example we need. (In fact, this is not a great loss of generality, because it can be proved that any Hilbert space is (isometrically) isomorphic to a space $L_2(\Omega, \mathcal{U}, \mu)$ for some $(\Omega, \mathcal{U}, \mu)$.)

2.8 Definition. *Two elements f, g of $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ are orthogonal if $\langle f, g \rangle = 0$. This is denoted $f \perp g$. Two subsets $\mathcal{F}, \mathcal{G}$ of $\mathcal{L}_2(\Omega, \mathcal{U}, \mu)$ are orthogonal if $f \perp g$ for every $f \in \mathcal{F}$ and $g \in \mathcal{G}$. This is denoted $\mathcal{F} \perp \mathcal{G}$.*

<u>2.9</u> EXERCISE. If $f \perp \mathcal{G}$ for some subset $\mathcal{G} \subset \mathcal{L}_2(\Omega, \mathcal{U}, \mathrm{P})$, show that $f \perp \overline{\mathrm{lin}}\, \mathcal{G}$, where $\overline{\mathrm{lin}}\, \mathcal{G}$ is the closure of the linear span of $\mathcal{G}$.

2.10 Theorem (Projection theorem). *Let $L \subset L_2(\Omega, \mathcal{U}, \mu)$ be a closed linear subspace. For every $f \in L_2(\Omega, \mathcal{U}, \mu)$ there exists a unique element $\Pi f \in L$ that minimizes $\|f - l\|^2$ over $l \in L$. This element is uniquely determined by the requirements $\Pi f \in L$ and $f - \Pi f \perp L$.*

Proof. Let $d = \inf_{l \in L} \|f - l\|$ be the "minimal" distance of f to L. This is finite, because $0 \in L$. Let l_n be a sequence in L such that $\|f - l_n\|^2 \to d$. By the parallellogram law

$$\begin{aligned} \bigl\|(l_m - f) + (f - l_n)\bigr\|^2 &= 2\|l_m - f\|^2 + 2\|f - l_n\|^2 - \bigl\|(l_m - f) - (f - l_n)\bigr\|^2 \\ &= 2\|l_m - f\|^2 + 2\|f - l_n\|^2 - 4\bigl\|\tfrac{1}{2}(l_m + l_n) - f\bigr\|^2. \end{aligned}$$

Because $(l_m + l_n)/2 \in L$, the last term on the right is bounded above by $-4d^2$. The two first terms on the far right both converge to $2d^2$ as $m, n \to \infty$. We conclude that the left side, which is $\|l_m - l_n\|^2$, is bounded above by $4d^2 + o(1) - 4d^2$ and hence, being nonnegative, converges to zero. Thus the sequence l_n is Cauchy and has a limit l by the completeness of $L_2(\Omega,\mathcal{U},\mu)$. The limit is in L, because L is closed. By the continuity of the norm $\|f-l\| = \lim\|f-l_n\| = d$. Thus the limit l qualifies as Πf.

If both $\Pi_1 f$ and $\Pi_2 f$ are candidates for Πf, then we can take the sequence $l_1, l_2, l_3, \ldots$ in the preceding argument equal to the sequence $\Pi_1 f, \Pi_2 f, \Pi_1 f, \ldots$. It then follows that this sequence is a Cauchy-sequence and hence converges to a limit. The latter is possibly only if $\Pi_1 f = \Pi_2 f$.

Finally, we consider the orthogonality relation. For every real number a and $l \in L$, we have

$$\bigl\|f-(\Pi f + al)\bigr\|^2 = \|f-\Pi f\|^2 - 2a\,\mathrm{Re}\langle f-\Pi f, l\rangle + a^2\|l\|^2.$$

By definition of Πf this is minimal as a function of a at the value $a=0$, whence the given parabola (in a) must have its bottom at zero, which is the case if and only if $\mathrm{Re}\langle f-\Pi f, l\rangle = 0$. A similar argument with ia instead of a shows that $\mathrm{Im}\langle f-\Pi f, l\rangle = 0$ as well. Thus $f-\Pi f \perp L$.

Conversely, if $\langle f-\Pi f, l\rangle = 0$ for every $l \in L$ and $\Pi f \in L$, then $\Pi f - l \in L$ for every $l \in L$ and by Pythagoras' rule

$$\|f-l\|^2 = \bigl\|(f-\Pi f)+(\Pi f - l)\bigr\|^2 = \|f-\Pi f\|^2 + \|\Pi f - l\|^2 \geq \|f-\Pi f\|^2.$$

This proves that Πf minimizes $l \mapsto \|f-l\|^2$ over $l \in L$. ■

The function Πf given in the preceding theorem is called the (orthogonal) *projection* of f onto L. From the orthogonality characterization of Πf, we can see that the map $f \mapsto \Pi f$ is linear and decreases norm:

$$\begin{aligned}\Pi(f+g) &= \Pi f + \Pi g,\\ \Pi(\alpha f) &= \alpha \Pi f,\\ \|\Pi f\| &\leq \|f\|.\end{aligned}$$

A further important property relates to repeated projections. If $\Pi_L f$ denotes the projection of f onto L, then

$$\Pi_{L_1}\Pi_{L_2} f = \Pi_{L_1} f, \qquad \text{if } L_1 \subset L_2.$$

Thus, we can find a projection in steps, by projecting a projection onto a bigger space a second time on the smaller space. This, again, is best proved using the orthogonality relations.

2.11 EXERCISE. Prove the relations in the two preceding displays.

The projection $\Pi_{L_1+L_2}$ onto the sum $L_1+L_2 = \{l_1+l_2: l_i \in L_i\}$ of two closed linear spaces is not necessarily the sum $\Pi_{L_1}+\Pi_{L_2}$ of the projections. (It is also not true that the sum of two closed linear subspaces is necessarily closed, so that $\Pi_{L_1+L_2}$ may not even be well defined.) However, this *is* true if the spaces L_1 and L_2 are orthogonal:

$$\Pi_{L_1+L_2} f = \Pi_{L_1} f + \Pi_{L_2} f, \qquad \text{if } L_1 \perp L_2.$$

2.12 EXERCISE.
(i) Show by counterexample that the condition $L_1 \perp L_2$ cannot be omitted.
(ii) Show that $L_1 + L_2$ is closed if $L_1 \perp L_2$ and both L_1 and L_2 are closed subspaces.
(iii) Show that $L_1 \perp L_2$ is sufficient in (i).

[Hint for (ii): It must be shown that if $z_n = x_n + y_n$ with $x_n \in L_1$, $y_n \in L_2$ for every n and $z_n \to z$, then $z = x + y$ for some $x \in L_1$ and $y \in L_2$. How can you find x_n and y_n from z_n?]

2.13 EXERCISE. Find the projection $\Pi_L f$ for L the one-dimensional space $\{\lambda l_0 \colon \lambda \in \mathbb{C}\}$.

* **2.14** EXERCISE. Suppose that the set L has the form $L = L_1 + iL_2$ for two closed, linear spaces L_1, L_2 of real functions. Show that the minimizer of $l \mapsto \|f - l\|$ over $l \in L$ for a real function f is the same as the minimizer of $l \mapsto \|f - l\|$ over L_1. Does this imply that $f - \Pi f \perp L_2$? Why is the preceding projection theorem of no use?

2.2 Square-integrable Random Variables

For $(\Omega, \mathcal{U}, \mathrm{P})$ a probability space the space $\mathcal{L}_2(\Omega, \mathcal{U}, \mathrm{P})$ is exactly the set of all complex (or real) random variables X with finite second moment $\mathrm{E}|X|^2$. The inner product is the product expectation $\langle X, Y\rangle = \mathrm{E}X\overline{Y}$, and the inner product between centered variables is the covariance:

$$\langle X - \mathrm{E}X, Y - \mathrm{E}Y\rangle = \mathrm{cov}(X, Y).$$

The Cauchy-Schwarz inequality takes the form

$$|\mathrm{E}X\overline{Y}|^2 \le \mathrm{E}|X|^2 \mathrm{E}|Y|^2.$$

When combined the preceding displays imply that $\big|\mathrm{cov}(X, Y)\big|^2 \le \mathrm{var}\, X\, \mathrm{var}\, Y$. Convergence $X_n \to X$ relative to the norm means that $\mathrm{E}|X_n - X|^2 \to 0$ and is referred to as *convergence in second mean*. This implies the convergence in mean $\mathrm{E}|X_n - X| \to 0$, because $\mathrm{E}|X| \le \sqrt{\mathrm{E}|X|^2}$ by the Cauchy-Schwarz inequality. The continuity of the inner product gives that:

$$\mathrm{E}|X_n - X|^2 \to 0, \mathrm{E}|Y_n - Y|^2 \to 0 \quad \text{implies} \ \mathrm{cov}(X_n, Y_n) \to \mathrm{cov}(X, Y).$$

2.15 EXERCISE. How can you apply this rule to prove equalities of the type $\mathrm{cov}(\sum \alpha_j X_{t-j}, \sum \beta_j Y_{t-j}) = \sum_i \sum_j \alpha_i \overline{\beta_j}\, \mathrm{cov}(X_{t-i}, Y_{t-j})$, such as in Lemma 1.28?

2.16 EXERCISE. Show that $\mathrm{sd}(X+Y) \le \mathrm{sd}(X) + \mathrm{sd}(Y)$ for any pair of random variables X and Y.

2.2.1 Conditional Expectation

Let $\mathcal{U}_0 \subset \mathcal{U}$ be a sub σ-field of $\mathcal{U}$. The collection L of all $\mathcal{U}_0$-measurable variables $Y \in L_2(\Omega, \mathcal{U}, \mathrm{P})$ is a closed, linear subspace of $L_2(\Omega, \mathcal{U}, \mathrm{P})$ (which can be identified with $L_2(\Omega, \mathcal{U}_0, \mathrm{P})$). By the projection theorem every square-integrable random variable X possesses a projection onto L. This particular projection is important enough to derive a number of special properties.

2.17 Definition. *The projection of $X \in L_2(\Omega, \mathcal{U}, \mathrm{P})$ onto the the set of all $\mathcal{U}_0$-measurable square-integrable random variables is called the conditional expectation of X given $\mathcal{U}_0$. It is denoted by* $\mathrm{E}(X|\,\mathcal{U}_0)$.

The name “conditional expectation” suggests that there exists another, more intuitive interpretation of this projection. An alternative definition of a conditional expectation is as follows.

2.18 Definition. *The conditional expectation given $\mathcal{U}_0$ of a random variable X which is either nonnegative or integrable is defined as a $\mathcal{U}_0$-measurable variable X' such that* $\mathrm{E}X1_A = \mathrm{E}X'1_A$ *for every $A \in \mathcal{U}_0$.*

It is clear from the definition that any other $\mathcal{U}_0$-measurable map X'' such that $X'' = X'$ almost surely is also a conditional expectation. Apart from this indeterminacy on null sets, a conditional expectation as in the second definition can be shown to be unique. Its existence can be proved using the Radon-Nikodym theorem. We shall not give proofs of these facts here.

Because a variable $X \in L_2(\Omega, \mathcal{U}, \mathrm{P})$ is automatically integrable, the second definition defines a conditional expectation for a larger class of variables. If $\mathrm{E}|X|^2 < \infty$, so that both definitions apply, then they agree. To see this it suffices to show that a projection $\mathrm{E}(X|\,\mathcal{U}_0)$ as in the first definition is the conditional expectation X' of the second definition. Now $\mathrm{E}(X|\,\mathcal{U}_0)$ is $\mathcal{U}_0$-measurable by definition and satisfies the equality $\mathrm{E}\bigl(X - \mathrm{E}(X|\,\mathcal{U}_0)\bigr)1_A = 0$ for every $A \in \mathcal{U}_0$, by the orthogonality relationship of a projection. Thus $X' = \mathrm{E}(X|\,\mathcal{U}_0)$ satisfies the requirements of Definition 2.18.

Definition 2.18 does show that a conditional expectation has to do with expectations, but is not very intuitive. Some examples help to gain more insight in conditional expectations.

2.19 Example (Ordinary expectation). The expectation $\mathrm{E}X$ of a random variable X is a number, and as such can be viewed as a degenerate random variable. It is also the conditional expectation relative to the trivial σ-field $\mathcal{U}_0 = \{\emptyset, \Omega\}$. More generally, we have that $\mathrm{E}(X|\,\mathcal{U}_0) = \mathrm{E}X$ if X and $\mathcal{U}_0$ are independent. In this case $\mathcal{U}_0$ gives “no information” about X and hence the expectation given $\mathcal{U}_0$ is the “unconditional” expectation.

To see this note that $\mathrm{E}(\mathrm{E}X)1_F = \mathrm{E}X\mathrm{E}1_F = \mathrm{E}X1_F$ for every measurable set F such that X and F are independent. □

2.20 Example. At the other extreme we have that $\mathrm{E}(X|\,\mathcal{U}_0) = X$ if X itself is $\mathcal{U}_0$-measurable. This is immediate from the definition. “Given $\mathcal{U}_0$ we then know X exactly.” □

A measurable map $Y{:}\,\Omega \to \mathbb{D}$ with values in some measurable space $(\mathbb{D}, \mathcal{D})$ generates a σ-field $\sigma(Y)$. The notation $\mathrm{E}(X|\,Y)$ is an abbreviation of $\mathrm{E}(X|\,\sigma(Y))$.

2.21 Example. Let $(X,Y){:}\,\Omega \to \mathbb{R} \times \mathbb{R}^k$ be measurable and possess a density $f(x,y)$ relative to a σ-finite product measure $\mu \times \nu$ on $\mathbb{R} \times \mathbb{R}^k$ (for instance, the Lebesgue measure on $\mathbb{R}^{k+1}$). Then it is customary to define a *conditional density* of X given $Y = y$ by

$$f(x|\,y) = \frac{f(x,y)}{\int f(x,y)\,d\mu(x)}.$$

This is well defined for every y for which the denominator is positive, i.e. for all y in a set of measure one under the distribution of Y.

We now have that the conditional expection is given by the "usual formula"

$$\mathrm{E}(X|\,Y) = \int x f(x|\,Y)\,d\mu(x),$$

where we may define the right hand zero as zero if the expression is not well defined.

That this formula is the conditional expectation according to Definition 2.18 follows by a number of applications of Fubini's theorem. Note that, to begin with, it is a part of the statement of Fubini's theorem that the function on the right is a measurable function of Y. □

2.22 Lemma (Properties).
(i) $\mathrm{E}\mathrm{E}(X|\,\mathcal{U}_0) = \mathrm{E}X$.
(ii) *If Z is $\mathcal{U}_0$-measurable, then* $\mathrm{E}(ZX|\,\mathcal{U}_0) = Z\mathrm{E}(X|\,\mathcal{U}_0)$ a.s.. *(Here require that $X \in L_p(\Omega, \mathcal{U}, \mathrm{P})$ and $Z \in L_q(\Omega, \mathcal{U}, \mathrm{P})$ for $1 \leq p \leq \infty$ and $p^{-1} + q^{-1} = 1$.)*
(iii) *(linearity)* $\mathrm{E}(\alpha X + \beta Y|\,\mathcal{U}_0) = \alpha\mathrm{E}(X|\,\mathcal{U}_0) + \beta\mathrm{E}(Y|\,\mathcal{U}_0)$ a.s..
(iv) *(positivity) If* $X \geq 0$ a.s., *then* $\mathrm{E}(X|\,\mathcal{U}_0) \geq 0$ a.s..
(v) *(towering property) If* $\mathcal{U}_0 \subset \mathcal{U}_1 \subset \mathcal{U}$, *then* $\mathrm{E}\bigl(\mathrm{E}(X|\,\mathcal{U}_1)|\,\mathcal{U}_0\bigr) = \mathrm{E}(X|\,\mathcal{U}_0)$ a.s..

The conditional expectation $\mathrm{E}(X|\,Y)$ given a random vector Y is by definition a $\sigma(Y)$-measurable function. For most Y, this means that it is a measurable function $g(Y)$ of Y. (See the following lemma.) The value $g(y)$ is often denoted by $\mathrm{E}(X|\,Y = y)$.

Warning. Unless $\mathrm{P}(Y = y) > 0$ it is not right to give a meaning to $\mathrm{E}(X|\,Y = y)$ for a fixed, single y, even though the interpretation as an expectation given "that we know that $Y = y$" often makes this tempting. We may only think of a conditional expectation as a function $y \mapsto \mathrm{E}(X|\,Y = y)$ and this is only determined up to null sets.

2.23 Lemma. *Let $\{Y_\alpha{:}\,\alpha \in A\}$ be random variables on Ω and let X be a $\sigma(Y_\alpha{:}\,\alpha \in A)$-measurable random variable.*
(i) *If $A = \{1, 2, \ldots, k\}$, then there exists a measurable map $g{:}\,\mathbb{R}^k \to \mathbb{R}$ such that $X = g(Y_1, \ldots, Y_k)$.*
(ii) *If $|A| = \infty$, then there exists a countable subset $\{\alpha_n\}_{n=1}^\infty \subset A$ and a measurable map $g{:}\,\mathbb{R}^\infty \to \mathbb{R}$ such that $X = g(Y_{\alpha_1}, Y_{\alpha_2}, \ldots)$.*

Proof. For the proof of (i), see e.g. Dudley Theorem 4.28. ■

2.3 Linear Prediction

Suppose that we observe the values $X_1, \ldots, X_n$ from a stationary, mean zero time series X_t.

2.24 Definition. *Suppose that* $\mathrm{E}X_t \equiv 0$*. The best linear predictor of* X_{n+1} *is the linear combination* $\phi_1 X_n + \phi_2 X_{n-1} + \cdots + \phi_n X_1$ *that minimizes* $\mathrm{E}|X_{n+1} - Y|^2$ *over all linear combinations* Y *of* $X_1, \ldots, X_n$*. The minimal value* $\mathrm{E}|X_{n+1} - \phi_1 X_n - \cdots - \phi_n X_1|^2$ *is called the square prediction error.*

In the terminology of the preceding section, the best linear predictor of X_{n+1} is the projection of X_{n+1} onto the linear subspace $\mathrm{lin}\,(X_1, \ldots, X_n)$ spanned by $X_1, \ldots, X_n$. A common notation is $\Pi_n X_{n+1}$, for Π_n the projection onto $\mathrm{lin}\,(X_1, \ldots, X_n)$. Best linear predictors of other random variables are defined similarly.

Warning. The coefficients $\phi_1, \ldots, \phi_n$ in the formula $\Pi_n X_{n+1} = \phi_1 X_n + \cdots + \phi_n X_1$ depend on n, even though we shall often suppress this dependence in the notation.

By Theorem 2.10 the best linear predictor can be found from the *prediction equations*

$$\langle X_{n+1} - \phi_1 X_n - \cdots - \phi_n X_1, X_t \rangle = 0, \qquad t = 1, \ldots, n.$$

For a stationary time series X_t this system can be written in the form

$$\begin{pmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(n-1) \\ \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_X(n-1) & \gamma_X(n-2) & \cdots & \gamma_X(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_n \end{pmatrix} = \begin{pmatrix} \gamma_X(1) \\ \vdots \\ \gamma_X(n) \end{pmatrix}. \tag{2.1}$$

If the $(n \times n)$-matrix on the left is nonsingular, then $\phi_1, \ldots, \phi_n$ can be solved uniquely. Otherwise there are more solutions for the vector $(\phi_1, \ldots, \phi_n)$, but any solution will give the best linear predictor $\Pi_n X_{n+1} = \phi_1 X_n + \cdots + \phi_n X_1$. The equations express $\phi_1, \ldots, \phi_n$ in the auto-covariance function γ_X. In practice, we do not know this function, but estimate it from the data. (We consider this estimation problem later on.) Then we use the corresponding estimates for $\phi_1, \ldots, \phi_n$ to calculate the predictor.

The square prediction error can be expressed in the coefficients using Pythagoras' rule, which gives, for a stationary time series X_t,

$$\begin{aligned} \mathrm{E}|X_{n+1} - \Pi_n X_{n+1}|^2 &= \mathrm{E}|X_{n+1}|^2 - \mathrm{E}|\Pi_n X_{n+1}|^2 \\ &= \gamma_X(0) - (\phi_1, \ldots, \phi_n)^T \Gamma_n (\phi_1, \ldots, \phi_n), \end{aligned} \tag{2.2}$$

for Γ_n the covariance matrix of the vector $(X_1, \ldots, X_n)$, i.e. the matrix on the left left side of (2.1).

Similar arguments apply to predicting X_{n+h} for $h > 1$. If we wish to predict the future values at many time lags $h = 1, 2, \ldots$, then solving a n-dimensional linear system for every h separately can be computer-intensive, as n may be large. Several more efficient, recursive algorithms use the predictions at earlier times to calculate the next prediction. We omit a discussion.

2.25 Example (Autoregression). Prediction is extremely simple for the stationary auto-regressive time series satisfying $X_t = \phi X_{t-1} + Z_t$ for a white noise sequence Z_t and $|\phi| < 1$. The best linear predictor of X_{n+1} given $X_1, \ldots, X_n$ is simply ϕX_n (for $n \geq 1$). Thus we predict $X_{n+1} = \phi X_n + Z_{n+1}$ by simply setting the unknown Z_{n+1} equal to its mean, zero. The interpretation is that the Z_t are external noise factors that are completely unpredictable based on the past. The square prediction error $\mathrm{E}|X_{n+1} - \phi X_n|^2 = \mathrm{E}Z_{n+1}^2$ is equal to the variance of this noise variable.

The claim is not obvious, as is proved by the fact that it is wrong in the case that $|\phi| > 1$. To prove the claim we recall from Example 1.8 that the unique stationary solution to the auto-regressive equation in the case that $|\phi| < 1$ is given by $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. Thus X_t depends only on Z_s from the past and the present. Because Z_t is a white noise sequence, it follows that X_t is uncorrelated with the variables $Z_{t+1}, Z_{t+2}, \ldots$. Therefore $\langle X_{n+1} - \phi X_n, X_t \rangle = \langle Z_{n+1}, X_t \rangle = 0$ for $t = 1, 2, \ldots, n$. This verifies the orthogonality relationship; it is obvious that ϕX_n is contained in the linear span of $X_1, \ldots, X_n$. □

2.26 EXERCISE. There is a hidden use of the continuity of the inner product in the preceding example. Can you see where?

2.27 Example (Deterministic trigonometric series). For the process $X_t = A\cos(\lambda t) + B\sin(\lambda t)$, considered in Example 1.5, the best linear predictor of X_{n+1} given $X_1, \ldots, X_n$ is given by $2(\cos\lambda)X_n - X_{n-1}$, for $n \geq 2$. The prediction error is equal to 0! This underscores that this type of time series is deterministic in character: if we know it at two time instants, then we know the time series at all other time instants. The explanation is that from the values of X_t at two time instants we can recover the values A and B.

These assertions follow by explicit calculations, solving the prediction equations. It suffices to do this for $n = 2$: if X_3 can be predicted without error by $2(\cos\lambda)X_2 - X_1$, then, by stationarity, X_{n+1} can be predicted without error by $2(\cos\lambda)X_n - X_{n-1}$. □

2.28 EXERCISE.

(i) Prove the assertions in the preceding example.

(ii) Are the coefficients $2\cos\lambda, -1, 0, \ldots, 0$ in this example unique?

If a given time series X_t is not centered at 0, then it is natural to allow a constant term in the predictor. Write 1 for the random variable that is equal to 1 almost surely.

2.29 Definition. *The best linear predictor of X_{n+1} based on $X_1, \ldots, X_n$ is the projection of X_{n+1} onto the linear space spanned by $1, X_1, \ldots, X_n$.*

If the time series X_t does have mean zero, then the introduction of the constant term 1 does not help. Indeed, the relation $\mathrm{E}X_t = 0$ is equivalent to $X_t \perp 1$, which implies both that $1 \perp \mathrm{lin}\,(X_1, \ldots, X_n)$ and that the projection of X_{n+1} onto lin 1 is zero. By the orthogonality the projection of X_{n+1} onto $\mathrm{lin}\,(1, X_1, \ldots, X_n)$ is the sum of the projections of X_{n+1} onto lin 1 and $\mathrm{lin}\,(X_1, \ldots, X_n)$, which is the projection on $\mathrm{lin}\,(X_1, \ldots, X_n)$, the first projection being 0.

By a similar argument we see that for a time series with mean $\mu = \mathrm{E}X_t$ possibly nonzero,

$$\Pi_{\mathrm{lin}\,(1,X_1,\ldots,X_n)}X_{n+1} = \mu + \Pi_{\mathrm{lin}\,(X_1-\mu,\ldots,X_n-\mu)}(X_{n+1}-\mu). \tag{2.3}$$

Thus the recipe for prediction with uncentered time series is: substract the mean from every X_t, calculate the projection for the centered time series $X_t - \mu$, and finally add the mean. Because the auto-covariance function γ_X gives the inner produts of the centered process, the coefficients $\phi_1, \ldots, \phi_n$ of $X_n - \mu, \ldots, X_1 - \mu$ are still given by the prediction equations (2.1).

2.30 EXERCISE. Prove formula (2.3), noting that $\mathrm{E}X_t = \mu$ is equivalent to $X_t - \mu \perp 1$.

2.4 Nonlinear Prediction

The method of linear prediction is commonly used in time series analysis. Its main advantage is simplicity: the linear predictor depends on the mean and auto-covariance function only, and in a simple fashion. On the other hand, if we allow general functions $f(X_1, \ldots, X_n)$ of the observations as predictors, then we might be able to decrease the prediction error.

2.31 Definition. *The best predictor of X_{n+1} based on $X_1, \ldots, X_n$ is the function $f_n(X_1, \ldots, X_n)$ that minimizes $\mathrm{E}\bigl|X_{n+1} - f(X_1, \ldots, X_n)\bigr|^2$ over all measurable functions $f: \mathbb{R}^n \to \mathbb{R}$.*

In view of the discussion in Section 2.2.1 the best predictor is the conditional expectation $\mathrm{E}(X_{n+1}|\, X_1, \ldots, X_n)$ of X_{n+1} given the variables $X_1, \ldots, X_n$. Best predictors of other variables are defined similarly as conditional expectations.

The difference between linear and nonlinear prediction can be substantial. In "classical" time series theory linear models with Gaussian errors were predominant and for those models the two predictors coincide. Given nonlinear models, or non-Gaussian distributions, nonlinear prediction should be the method of choice, if feasible.

2.32 Example (GARCH). In the GARCH model of Example 1.10 the variable X_{n+1} is given as $\sigma_{n+1}Z_{n+1}$, where σ_{n+1} is a function of $X_n, X_{n-1}, \ldots$ and Z_{n+1} is independent of these variables. It follows that the best predictor of X_{n+1} given the infinite past $X_n, X_{n-1}, \ldots$ is given by $\sigma_{n+1}\mathrm{E}(Z_{n+1}|\, X_n, X_{n-1}, \ldots) = 0$. We can find the best predictor given $X_n, \ldots, X_1$ by projecting this predictor further onto the space of all measurable functions of $X_n, \ldots, X_1$. By the linearity of the projection we again find 0.

We conclude that a GARCH model does not allow a "true prediction" of the future, if "true" refers to predicting the values of the time series itself.

On the other hand, we can predict other quantities of interest. For instance, the uncertainty of the value of X_{n+1} is determined by the size of σ_{n+1}. If σ_{n+1} is close to zero, then we may expect X_{n+1} to be close to zero, and conversely. Given the infinite past $X_n, X_{n-1}, \ldots$ the variable σ_{n+1} is known completely, but in the more realistic situation that we know only $X_n, \ldots, X_1$ some chance component will be left. (For large n the difference between these two situations will be small.) The dependence of σ_{n+1} on $X_n, X_{n-1}, \ldots$ is given in Example 1.10 as $\sigma_{n+1}^2 = \sum_{j=0}^\infty \phi^j(\alpha + \theta X_{n-j}^2)$ and is nonlinear. For large n this is close to $\sum_{j=0}^{n-1} \phi^j(\alpha + \theta X_{n-j}^2)$, which is a function of $X_1, \ldots, X_n$. By definition the best predictor $\hat\sigma_{n+1}^2$ based on $1, X_1, \ldots, X_n$ is the closest function and hence it satisfies

$$\mathrm{E}\bigl|\hat\sigma_{n+1}^2 - \sigma_{n+1}^2\bigr|^2 \le \mathrm{E}\Bigl|\sum_{j=0}^{n-1} \phi^j(\alpha + \theta X_{n-j}^2) - \sigma_{n+1}^2\Bigr|^2 = \mathrm{E}\Bigl|\sum_{j=n}^{\infty} \phi^j(\alpha + \theta X_{n-j}^2)\Bigr|^2.$$

For small ϕ and large n this will be small if the sequence X_n is sufficiently integrable. Thus nonlinear prediction of σ_{n+1}^2 is feasible. □

2.5 Partial Auto-Correlation

For a given mean-zero stationary time series X_t the *partial auto-correlation* of lag h is defined as the correlation between $X_h - \Pi_{h-1}X_h$ and $X_0 - \Pi_{h-1}X_0$, where Π_h is the projection onto $\mathrm{lin}\,(X_1, \ldots, X_h)$. This is the "correlation between X_h and X_0 with the correlation due to the intermediate variables $X_1, \ldots, X_{h-1}$ removed". We shall denote it by

$$\alpha_X(h) = \rho\bigl(X_h - \Pi_{h-1}X_h, X_0 - \Pi_{h-1}X_0\bigr).$$

For an uncentered stationary time series we set the partial auto-correlation by definition equal to the partial auto-correlation of the centered series $X_t - \mathrm{E}X_t$. A convenient method to compute α_X is given by the prediction equations combined with the following lemma: $\alpha_X(h)$ is the coefficient of X_1 of the best linear predictor of X_{h+1} based on $X_1, \ldots, X_h$.

2.33 Lemma. *Suppose that X_t is a mean-zero stationary time series. If $\phi_1 X_h + \phi_2 X_{h-1} + \cdots + \phi_h X_1$ is the best linear predictor of X_{h+1} based on $X_1, \ldots, X_h$, then $\alpha_X(h) = \phi_h$.*

Proof. Let $\psi_1 X_h + \cdots + \psi_{h-1}X_2 =: \Pi_{2,h}X_1$ be the best linear predictor of X_1 based on $X_2, \ldots, , X_h$. The best linear predictor of X_{h+1} based on $X_1, \ldots, X_h$ can be decomposed as

$$\begin{aligned}\Pi_h X_{h+1} &= \phi_1 X_h + \cdots + \phi_h X_1\\ &= \bigl[(\phi_1 + \phi_h\psi_1)X_h + \cdots + (\phi_{h-1} + \phi_h\psi_{h-1})X_2\bigr] + \phi_h\bigl[(X_1 - \Pi_{2,h}X_1)\bigr].\end{aligned}$$

The two terms in square brackets are orthogonal, because $X_1 - \Pi_{2,h}X_1 \perp \mathrm{lin}\,(X_2, \ldots, X_h)$ by the projection theorem. Therefore, the second term in square brackets is the projection of $\Pi_h X_{h+1}$ onto the one-dimensional subspace $\mathrm{lin}\,(X_1 - \Pi_{2,h} X_1)$. It is also the projection of X_{h+1} onto this one-dimensional subspace, because $\mathrm{lin}\,(X_1 - \Pi_{2,h} X_1) \subset \mathrm{lin}\,(X_1, \ldots, X_h)$ and we can compute projections by first projecting onto a bigger subspace.

The projection of X_{h+1} onto the one-dimensional subspace $\mathrm{lin}\,(X_1 - \Pi_{2,h} X_1)$ is easy to compute directly. It is given by $\alpha(X_1 - \Pi_{2,h} X_1)$ for α given by

$$\alpha = \frac{\langle X_{h+1}, X_1 - \Pi_{2,h} X_1 \rangle}{\| X_1 - \Pi_{2,h} X_1 \|^2} = \frac{\langle X_{h+1} - \Pi_{2,h} X_{h+1}, X_1 - \Pi_{2,h} X_1 \rangle}{\| X_1 - \Pi_{2,h} X_1 \|^2}.$$

Because the prediction problem is symmetric in time, as it depends on the auto-covariance function only, $\| X_1 - \Pi_{2,h} X_1 \| = \| X_{h+1} - \Pi_{2,h} X_1 \|$. Therefore, the right side is exactly $\alpha_X(h)$. In view of the preceding paragraph, we have $\alpha = \phi_h$ and the lemma is proved. ∎

2.34 Example (Autoregression). According to Example 2.25, for the stationary auto-regressive process $X_t = \phi X_{t-1} + Z_t$ with $|\phi| < 1$, the best linear predictor of X_{n+1} based on $X_1, \ldots, X_n$ is ϕX_n, for $n \geq 1$. Thus the partial auto-correlations $\alpha_X(h)$ of lags $h > 1$ are zero and $\alpha_X(1) = \phi$. This is often viewed as the dual of the property that for the moving average sequence of order 1, considered in Example 1.6, the auto-correlations of lags $h > 1$ vanish.

In Chapter 7 we shall see that for higher order stationary auto-regressive processes $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$ the partial auto-correlations of lags $h > p$ are zero under the (standard) assumption that the time series is "causal". □

3
Stochastic Convergence

This chapter provides a review of modes of convergence of sequences of stochastic vectors. In particular, convergence in distribution and in probability. Many proofs are omitted, but can be found in most standard probability books.

3.1 Basic theory

A *random vector* in $\mathbb{R}^k$ is a vector $X = (X_1, \ldots, X_k)$ of real random variables. More formally it is a Borel measurable map from some probability space in $\mathbb{R}^k$. The *distribution function* of X is the map $x \to \mathrm{P}(X \leq x)$.

A sequence of random vectors X_n is said to *converge in distribution* to X if

$$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x),$$

for every x at which the distribution function $x \to \mathrm{P}(X \leq x)$ is continuous. Alternative names are *weak convergence* and *convergence in law*. As the last name suggests, the convergence only depends on the induced laws of the vectors and not on the probability spaces on which they are defined. Weak convergence is denoted by $X_n \rightsquigarrow X$; if X has distribution L or a distribution with a standard code such as $N(0,1)$, then also by $X_n \rightsquigarrow L$ or $X_n \rightsquigarrow N(0,1)$.

Let $d(x,y)$ be any distance function on $\mathbb{R}^k$ that generates the usual topology. For instance

$$d(x,y) = \|x - y\| = \Bigl(\sum_{i=1}^{k} (x_i - y_i)^2\Bigr)^{1/2}.$$

A sequence of random variables X_n is said to *converge in probability* to X if for all $\varepsilon > 0$

$$\mathrm{P}\bigl(d(X_n, X) > \varepsilon\bigr) \to 0.$$

This is denoted by $X_n \xrightarrow{\mathrm{P}} X$. In this notation convergence in probability is the same as $d(X_n, X) \xrightarrow{\mathrm{P}} 0$.

As we shall see convergence in probability is stronger than convergence in distribution. Even stronger modes of convergence are almost sure convergence and convergence in pth mean. The sequence X_n is said to *converge almost surely* to X if $d(X_n, X) \to 0$ with probability one:

$$\mathrm{P}\bigl(\lim d(X_n, X) = 0\bigr) = 1.$$

This is denoted by $X_n \xrightarrow{\mathrm{as}} X$. The sequence X_n is said to *converge in pth mean* to X if

$$\mathrm{E}d(X_n, X)^p \to 0.$$

This is denoted $X_n \xrightarrow{L_p} X$. We already encountered the special cases $p = 1$ or $p = 2$, which are referred to as "convergence in mean" and "convergence in quadratic mean".

Convergence in probability, almost surely, or in mean only make sense if each X_n and X are defined on the same probability space. For convergence in distribution this is not necessary.

The portmanteau lemma gives a number of equivalent descriptions of weak convergence. Most of the characterizations are only useful in proofs. The last one also has intuitive value.

3.1 Lemma (Portmanteau). *For any random vectors X_n and X the following statements are equivalent.*

(i) $\mathrm{P}(X_n \le x) \to \mathrm{P}(X \le x)$ *for all continuity points of* $x \to \mathrm{P}(X \le x)$;

(ii) $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for all bounded, continuous functions f;*

(iii) $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ *for all bounded, Lipschitz*[†] *functions f;*

(iv) $\liminf \mathrm{P}(X_n \in G) \ge \mathrm{P}(X \in G)$ *for every open set G;*

(v) $\limsup \mathrm{P}(X_n \in F) \le \mathrm{P}(X \in F)$ *for every closed set F;*

(vi) $\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$ *for all Borel sets B with* $\mathrm{P}(X \in \delta B) = 0$ *where* $\delta B = \overline{B} - \mathring{B}$ *is the boundary of B.*

The continuous mapping theorem is a simple result, but is extremely useful. If the sequence of random vector X_n converges to X and g is continuous, then $g(X_n)$ converges to $g(X)$. This is true without further conditions for three of our four modes of stochastic convergence.

3.2 Theorem (Continuous mapping). *Let $g: \mathbb{R}^k \to \mathbb{R}^m$ be measurable and continuous at every point of a set C such that* $\mathrm{P}(X \in C) = 1$.

(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$;

(ii) If $X_n \xrightarrow{\mathrm{P}} X$, then $g(X_n) \xrightarrow{\mathrm{P}} g(X)$;

(iii) If $X_n \xrightarrow{\mathrm{as}} X$, then $g(X_n) \xrightarrow{\mathrm{as}} g(X)$.

Any random vector X is *tight*: for every $\varepsilon > 0$ there exists a constant M such that $\mathrm{P}\bigl(\|X\| > M\bigr) < \varepsilon$. A set of random vectors $\{X_\alpha: \alpha \in A\}$ is called *uniformly tight* if M

[†] A function is called *Lipschitz* if there exists a number L such that $|f(x) - f(y)| \le L d(x, y)$ for every x and y. The least such number L is denoted $\|f\|_{Lip}$.

can be chosen the same for every X_α: for every $\varepsilon > 0$ there exists a constant M such that

$$\sup_\alpha \mathrm{P}\bigl(\|X_\alpha\| > M\bigr) < \varepsilon.$$

Thus there exists a compact set to which all X_α give probability almost one. Another name for uniformly tight is *bounded in probability*. It is not hard to see that every weakly converging sequence X_n is uniformly tight. More surprisingly, the converse of this statement is almost true: according to Prohorov's theorem every uniformly tight sequence contains a weakly converging subsequence.

3.3 Theorem (Prohorov's theorem). *Let X_n be random vectors in $\mathbb{R}^k$.*

(i) If $X_n \rightsquigarrow X$ for some X, then $\{X_n : n \in \mathbb{N}\}$ is uniformly tight;

(ii) If X_n is uniformly tight, then there is a subsequence with $X_{n_j} \rightsquigarrow X$ as $j \to \infty$ for some X.

3.4 Example. A sequence X_n of random variables with $\mathrm{E}|X_n| = O(1)$ is uniformly tight. This follows since by Markov's inequality: $\mathrm{P}\bigl(|X_n| > M\bigr) \le \mathrm{E}|X_n|/M$. This can be made arbitrarily small uniformly in n by choosing sufficiently large M.

The first absolute moment could of course be replaced by any other absolute moment.

Since the second moment is the sum of the variance and the square of the mean an alternative sufficient condition for uniform tightness is: $\mathrm{E}X_n = O(1)$ and $\operatorname{var} X_n = O(1)$. □

Consider some of the relationships between the three modes of convergence. Convergence in distribution is weaker than convergence in probability, which is in turn weaker than almost sure convergence and convergence in pth mean.

3.5 Theorem. *Let X_n, X and Y_n be random vectors. Then*

(i) $X_n \xrightarrow{\text{as}} X$ implies $X_n \xrightarrow{\mathrm{P}} X$;

(ii) $X_n \xrightarrow{L_p} X$ implies $X_n \xrightarrow{\mathrm{P}} X$;

(iii) $X_n \xrightarrow{\mathrm{P}} X$ implies $X_n \rightsquigarrow X$;

(iv) $X_n \xrightarrow{\mathrm{P}} c$ for a constant c if and only if $X_n \rightsquigarrow c$;

(v) if $X_n \rightsquigarrow X$ and $d(X_n, Y_n) \xrightarrow{\mathrm{P}} 0$, then $Y_n \rightsquigarrow X$;

(vi) if $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{\mathrm{P}} c$ for a constant c, then $(X_n, Y_n) \rightsquigarrow (X, c)$;

(vii) if $X_n \xrightarrow{\mathrm{P}} X$ and $Y_n \xrightarrow{\mathrm{P}} Y$, then $(X_n, Y_n) \xrightarrow{\mathrm{P}} (X, Y)$.

Proof. (i). The sequence of sets $A_n = \cup_{m \ge n}\bigl\{d(X_m, X) > \varepsilon\bigr\}$ is decreasing for every $\varepsilon > 0$ and decreases to the empty set if $X_n(\omega) \to X(\omega)$ for every ω. If $X_n \xrightarrow{\text{as}} X$, then $\mathrm{P}\bigl(d(X_n, X) > \varepsilon\bigr) \le \mathrm{P}(A_n) \to 0$.

(ii). This is an immediate consequence of Markov's inequality, according to which $\mathrm{P}\bigl(d(X_n, X) > \varepsilon\bigr) \le \varepsilon^{-p}\mathrm{E}d(X_n, X)^p$ for every $\varepsilon > 0$.

(v). For every bounded Lipschitz function f and every $\varepsilon > 0$ we have

$$\begin{aligned}\bigl|\mathrm{E}f(X_n) - \mathrm{E}f(Y_n)\bigr| &\le \varepsilon\|f\|_{Lip}\mathrm{E}1\bigl\{d(X_n, Y_n) \le \varepsilon\bigr\} \\ &\qquad + 2\|f\|_\infty \mathrm{E}1\bigl\{d(X_n, Y_n) > \varepsilon\bigr\}.\end{aligned}$$

The second term on the right converges to zero as $n \to \infty$. The first term can be made arbitrarily small by choice of ε. Conclude that the sequences $\mathrm{E}f(X_n)$ and $\mathrm{E}f(Y_n)$ have the same limit. The result follows from the portmanteau lemma.

(iii). Since $d(X_n, X) \xrightarrow{\mathrm{P}} 0$ and trivially $X \rightsquigarrow X$ it follows that $X_n \rightsquigarrow X$ by (v).

(iv). The 'only if' part is a special case of (iii). For the converse let $\mathrm{ball}(c, \varepsilon)$ be the open ball of radius ε around c. Then $\mathrm{P}\bigl(d(X_n, c) \geq \varepsilon\bigr) = \mathrm{P}\bigl(X_n \in \mathrm{ball}(c, \varepsilon)^c\bigr)$. If $X_n \rightsquigarrow c$, then the lim sup of the last probability is bounded by $\mathrm{P}\bigl(c \in \mathrm{ball}(c, \varepsilon)^c\bigr) = 0$.

(vi). First note that $d\bigl((X_n, Y_n), (X_n, c)\bigr) = d(Y_n, c) \xrightarrow{\mathrm{P}} 0$. Thus according to (v) it suffices to show that $(X_n, c) \rightsquigarrow (X, c)$. For every continuous, bounded function $(x, y) \to f(x, y)$, the function $x \to f(x, c)$ is continuous and bounded. Thus $\mathrm{E}f(X_n, c) \to \mathrm{E}f(X, c)$ if $X_n \rightsquigarrow X$.

(vii). This follows from $d\bigl((x_1, y_1), (x_2, y_2)\bigr) \leq d(x_1, x_2) + d(y_1, y_2)$. ∎

According to the last assertion of the lemma convergence in probability of a sequence of vectors $X_n = (X_{n,1}, \ldots, X_{n,k})$ is equivalent to convergence of every one of the sequences of components $X_{n,i}$ separately. The analogous statement for convergence in distribution is false: convergence in distribution of the sequence X_n is stronger than convergence of every one of the sequences of components $X_{n,i}$. The point is that the distribution of the components $X_{n,i}$ separately does not determine their joint distribution: they might be independent or dependent in many ways. One speaks of *joint convergence* in distribution versus *marginal convergence.*

The one before last assertion of the lemma has some useful consequences. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $(X_n, Y_n) \rightsquigarrow (X, c)$. Consequently, by the continuous mapping theorem $g(X_n, Y_n) \rightsquigarrow g(X, c)$ for every map g that is continuous at the set $\mathbb{R}^k \times \{c\}$ where the vector (X, c) takes its values. Thus for every g such that

$$\lim_{x \to x_0, y \to c} g(x, y) = g(x_0, c), \qquad \text{every } x_0.$$

Some particular applications of this principle are known as Slutsky's lemma.

3.6 Lemma (Slutsky). *Let X_n, X and Y_n be random vectors or variables. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ for a constant c, then*
(i) $X_n + Y_n \rightsquigarrow X + c$;
(ii) $Y_n X_n \rightsquigarrow cX$;
(iii) $X_n / Y_n \rightsquigarrow X/c$ provided $c \neq 0$.

In (i) the "constant" c must be a vector of the same dimension as X, and in (ii) c is probably initially understood to be a scalar. However, (ii) is also true if every Y_n and c are matrices (which can be identified with vectors, for instance by aligning rows, to give a meaning to the convergence $Y_n \rightsquigarrow c$), simply because matrix multiplication $(y, x) \to yx$ is a continuous operation. Another true result in this case is that $X_n Y_n \rightsquigarrow Xc$, if this statement is well defined. Even (iii) is valid for matrices Y_n and c and vectors X_n provided $c \neq 0$ is understood as c being invertible and division is interpreted as (pre)multiplication by the inverse, because taking an inverse is also continuous.

3.7 Example. Let T_n and S_n be statistical estimators satisfying

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, \sigma^2), \qquad\qquad S_n^2 \xrightarrow{\mathrm{P}} \sigma^2,$$

for certain parameters θ and σ^2 depending on the underlying distribution, for every distribution in the model. Then $\theta = T_n \pm S_n/\sqrt{n}\,\xi_\alpha$ is a confidence interval for θ of asymptotic level $1 - 2\alpha$.

This is a consequence of the fact that the sequence $\sqrt{n}(T_n - \theta)/S_n$ is asymptotically standard normal distributed. □

* 3.2 Convergence of Moments

By the portmanteau lemma, weak convergence $X_n \rightsquigarrow X$ implies that $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ for every continuous, bounded function f. The condition that f be bounded is not superfluous: it is not difficult to find examples of a sequence $X_n \rightsquigarrow X$ and an unbounded, continuous function f for which the convergence fails. In particular, in general convergence in distribution does not imply convergence $\mathrm{E}X_n^p \to \mathrm{E}X^p$ of moments. However, in many situations such convergence occurs, but it requires more effort to prove it.

A sequence of random variables Y_n is called *asymptotically uniformly integrable* if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{E}|Y_n| 1\{|Y_n| > M\} = 0.$$

A simple sufficient condition for this is that for some $p > 1$ the sequence $\mathrm{E}|Y_n|^p$ is bounded in n.

Uniform integrability is the missing link between convergence in distribution and convergence of moments.

3.8 Theorem. *Let $f\colon \mathbb{R}^k \to \mathbb{R}$ be measurable and continuous at every point in a set C. Let $X_n \rightsquigarrow X$ where X takes its values in C. Then $\mathrm{E}f(X_n) \to \mathrm{E}f(X)$ if and only if the sequence of random variables $f(X_n)$ is asymptotically uniformly integrable.*

3.9 Example. Suppose X_n is a sequence of random variables such that $X_n \rightsquigarrow X$ and $\limsup \mathrm{E}|X_n|^p < \infty$ for some p. Then all moments of order strictly less than p converge also: $\mathrm{E}X_n^k \to \mathrm{E}X^k$ for every $k < p$.

By the preceding theorem, it suffices to prove that the sequence X_n^k is asymptotically uniformly integrable. By Markov's inequality

$$\mathrm{E}|X_n|^k 1\bigl\{|X_n|^k \geq M\bigr\} \leq M^{1-p/k}\,\mathrm{E}|X_n|^p.$$

The limsup, as $n \to \infty$ followed by $M \to \infty$, of the right side is zero if $k < p$. □

3.3 Arrays

Consider an infinite array $x_{n,l}$ of numbers, indexed by $(n,l)\in\mathbb{N}\times\mathbb{N}$, such that every column has a limit, and the limits x_l themselves converge to a limit along the columns.

$$\begin{array}{cccccc}
x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} & \ldots & \\
x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} & \ldots & \\
x_{3,1} & x_{3,2} & x_{3,3} & x_{3,4} & \ldots & \\
\vdots & \vdots & \vdots & \vdots & \ldots & \\
\downarrow & \downarrow & \downarrow & \downarrow & \ldots & \\
x_1 & x_2 & x_3 & x_4 & \ldots & \rightarrow x
\end{array}$$

Then we can find a "path" x_{n,l_n}, indexed by $n\in\mathbb{N}$ through the array along which $x_{n,l_n}\rightarrow x$ as $n\rightarrow\infty$. (The point is to move to the right slowly in the array while going down, i.e. $l_n\rightarrow\infty$.) A similar property is valid for sequences of random vectors, where the convergence is taken as convergence in distribution.

3.10 Lemma. *For $n,l\in\mathbb{N}$ let $X_{n,l}$ be random vectors such that $X_{n,l}\rightsquigarrow X_l$ as $n\rightarrow\infty$ for every fixed l for random vectors such that $X_l\rightsquigarrow X$ as $l\rightarrow\infty$. Then there exists a sequence $l_n\rightarrow\infty$ such $X_{n,l_n}\rightsquigarrow X$ as $n\rightarrow\infty$.*

Proof. Let $D=\{d_1,d_2,\ldots\}$ be a countable set that is dense in $\mathbb{R}^k$ and that only contains points at which the distribution functions of the limits $X,X_1,X_2,\ldots$ are continuous. Then an arbitrary sequence of random variables Y_n converges in distribution to one of the variables $Y\in\{X,X_1,X_2,\ldots\}$ if and only if $\mathrm{P}(Y_n\le d_i)\rightarrow\mathrm{P}(Y\le d_i)$ for every $d_i\in D$. We can prove this using the monotonicity and right-continuity of distribution functions. In turn $\mathrm{P}(Y_n\le d_i)\rightarrow\mathrm{P}(Y\le d_i)$ as $n\rightarrow\infty$ for every $d_i\in D$ if and only if

$$\sum_{i=1}^{\infty}\bigl|\mathrm{P}(Y_n\le d_i)-\mathrm{P}(Y\le d_i)\bigr|2^{-i}\rightarrow 0.$$

Now define

$$\begin{aligned}
p_{n,l}&=\sum_{i=1}^{\infty}\bigl|\mathrm{P}(X_{n,l}\le d_i)-\mathrm{P}(X_l\le d_i)\bigr|\frac{1}{2^i},\\
p_l&=\sum_{i=1}^{\infty}\bigl|\mathrm{P}(X_l\le d_i)-\mathrm{P}(X\le d_i)\bigr|2^{-i}.
\end{aligned}$$

The assumptions entail that $p_{n,l}\rightarrow 0$ as $n\rightarrow\infty$ for every fixed l, and that $p_l\rightarrow 0$ as $l\rightarrow\infty$. This implies that there exists a sequence $l_n\rightarrow\infty$ such that $p_{n,l_n}\rightarrow 0$. By the triangle inequality

$$\sum_{i=1}^{\infty}\bigl|\mathrm{P}(X_{n,l_n}\le d_i)-\mathrm{P}(X\le d_i)\bigr|2^{-i}\le p_{n,l_n}+p_{l_n}\rightarrow 0.$$

This implies that $X_{n,l_n}\rightsquigarrow X$ as $n\rightarrow\infty$. ∎

3.4 Stochastic o and O symbols

It is convenient to have short expressions for terms that converge in probability to zero or are uniformly tight. The notation $o_P(1)$ ('small "oh-P-one"') is short for a sequence of random vectors that converges to zero in probability. The expression $O_P(1)$ ('big "oh-P-one"') denotes a sequence that is bounded in probability. More generally, for a given sequence of random variables R_n

$$\begin{aligned} X_n = o_P(R_n) &\quad \text{means} \quad X_n = Y_n R_n \text{ and } Y_n \xrightarrow{\mathrm{P}} 0; \\ X_n = O_P(R_n) &\quad \text{means} \quad X_n = Y_n R_n \text{ and } Y_n = O_P(1). \end{aligned}$$

This expresses that the sequence X_n converges in probability to zero or is bounded in probability at 'rate' R_n. For deterministic sequences X_n and R_n the stochastic oh-symbols reduce to the usual o and O from calculus.

There are many rules of calculus with o and O symbols, which will be applied without comment. For instance,

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ O_P(1) o_P(1) &= o_P(1) \\ \bigl(1 + o_P(1)\bigr)^{-1} &= O_P(1) \\ o_P(R_n) &= R_n o_P(1) \\ O_P(R_n) &= R_n O_P(1) \\ o_P\bigl(O_P(1)\bigr) &= o_P(1). \end{aligned}$$

To see the validity of these "rules" it suffices to restate them in terms of explicitly named vectors, where each $o_P(1)$ and $O_P(1)$ should be replaced by a different sequence of vectors that converges to zero or is bounded in probability. In this manner the first rule says: if $X_n \xrightarrow{\mathrm{P}} 0$ and $Y_n \xrightarrow{\mathrm{P}} 0$, then $Z_n = X_n + Y_n \xrightarrow{\mathrm{P}} 0$; this is an example of the continuous mapping theorem. The third rule is short for: if X_n is bounded in probability and $Y_n \xrightarrow{\mathrm{P}} 0$, then $X_n Y_n \xrightarrow{\mathrm{P}} 0$. If X_n would also converge in distribution, then this would be statement (ii) of Slutsky's lemma (with $c = 0$). But by Prohorov's theorem X_n converges in distribution "along subsequences" if it is bounded in probability, so that the third rule can still be deduced from Slutsky's lemma by "arguing along subsequences".

Note that both rules are in fact implications and should be read from left to right, even though they are stated with the help of the equality "=" sign. Similarly, while it is true that $o_P(1) + o_P(1) = 2o_P(1)$, writing down this rule does not reflect understanding of the o_P-symbol.

Two more complicated rules are given by the following lemma.

3.11 Lemma. *Let R be a function defined on a neighbourhood of $0 \in \mathbb{R}^k$ such that $R(0) = 0$. Let X_n be a sequence of random vectors that converges in probability to zero.*
(i) if $R(h) = o(\|h\|)$ as $h \to 0$, then $R(X_n) = o_P(\|X_n\|)$;

(ii) if $R(h) = O(\|h\|)$ as $h \to 0$, then $R(X_n) = O_P(\|X_n\|)$.

Proof. Define $g(h)$ as $g(h) = R(h)/\|h\|$ for $h \neq 0$ and $g(0) = 0$. Then $R(X_n) = g(X_n)\|X_n\|$.

(i). Since the function g is continuous at zero by assumption, $g(X_n) \xrightarrow{\mathrm{P}} g(0) = 0$ by the continuous mapping theorem.

(ii). By assumption there exist M and $\delta > 0$ such that $\bigl|g(h)\bigr| \leq M$ whenever $\|h\| \leq \delta$. Thus $\mathrm{P}\bigl(\bigl|g(X_n)\bigr| > M\bigr) \leq \mathrm{P}\bigl(\|X_n\| > \delta\bigr) \to 0$, and the sequence $g(X_n)$ is tight. ■

It should be noted that the rule expressed by the lemma is not a simple plug-in rule. For instance it is not true that $R(h) = o(\|h\|)$ implies that $R(X_n) = o_P(\|X_n\|)$ for every sequence of random vectors X_n.

3.5 Transforms

It is sometimes possible to show convergence in distribution of a sequence of random vectors directly from the definition. In other cases 'transforms' of probability measures may help. The basic idea is that it suffices to show characterization (ii) of the portmanteau lemma for a small subset of functions f only.

The most important transform is the *characteristic function*

$$t \to \mathrm{E}e^{it^TX}, \qquad t \in \mathbb{R}^k.$$

Each of the functions $x \to e^{it^Tx}$ is continuous and bounded. Thus by the portmanteau lemma $\mathrm{E}e^{it^TX_n} \to \mathrm{E}e^{it^TX}$ for every t if $X_n \rightsquigarrow X$. By Lévy's continuity theorem the converse is also true: pointwise convergence of characteristic functions is equivalent to weak convergence.

3.12 Theorem (Lévy's continuity theorem). *Let X_n and X be random vectors in $\mathbb{R}^k$. Then $X_n \rightsquigarrow X$ if and only if $Ee^{it^TX_n} \to \mathrm{E}e^{it^TX}$ for every $t \in \mathbb{R}^k$. Moreover, if $Ee^{it^TX_n}$ converges pointwise to a function $\phi(t)$ that is continuous at zero, then ϕ is the characteristic function of a random vector X and $X_n \rightsquigarrow X$.*

The following lemma, which gives a variation on Lévy's theorem, is less well known, but will be useful in Chapter 4.

3.13 Lemma. *Let X_n be a sequence of random variables such that $\mathrm{E}|X_n|^2 = O(1)$ and such that $\mathrm{E}(iX_n + vt)e^{itX_n} \to 0$ as $n \to \infty$, for every $t \in \mathbb{R}$ and some $v > 0$. Then $X_n \rightsquigarrow N(0, v)$.*

Proof. By Markov's inequality and the bound on the second moments, the sequence X_n is uniformly tight. In view of Prohorov's theorem it suffices to show that $N(0, v)$ is the only weak limit point.

If $X_n \rightsquigarrow X$ along some sequence of n, then by the boundedness of the second moments and the continuity of the function $x \mapsto (ix+vt)e^{itx}$, we have $\mathrm{E}(iX_n+vt)e^{itX_n} \to \mathrm{E}(iX+vt)e^{itX}$ for every $t \in \mathbb{R}$. (Cf. Theorem 3.8.) Combining this with the assumption, we see that $\mathrm{E}(iX+vt)e^{itX} = 0$. By Fatou's lemma $\mathrm{E}X^2 \leq \liminf \mathrm{E}X_n^2 < \infty$ and hence we can differentiate the the characteristic function $\phi(t) = \mathrm{E}e^{itX}$ under the expectation to find that $\phi'(t) = \mathrm{E}iXe^{itX}$. We conclude that $\phi'(t) = -vt\phi(t)$. This differential equation possesses $\phi(t) = e^{-vt^2/2}$ as the only solution within the class of characteristic functions. Thus X is normally distributed with mean zero and variance v. ■

3.6 Cramér-Wold Device

The characteristic function $t \to \mathrm{E}e^{it^TX}$ of a vector X is determined by the set of all characteristic functions $u \to \mathrm{E}e^{iu(t^TX)}$ of all linear combinations t^TX of the components of X. Therefore the continuity theorem implies that weak convergence of vectors is equivalent to weak convergence of linear combinations:

$$X_n \rightsquigarrow X \qquad \text{if and only if} \qquad t^TX_n \rightsquigarrow t^TX \quad \text{for all} \quad t \in \mathbb{R}^k.$$

This is known as the *Cramér-Wold device*. It allows to reduce all higher dimensional weak convergence problems to the one-dimensional case.

3.14 Example (Multivariate central limit theorem). Let $Y, Y_1, Y_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^k$ with mean vector $\mu = \mathrm{E}Y$ and covariance matrix $\Sigma = \mathrm{E}(Y-\mu)(Y-\mu)^T$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \mu) = \sqrt{n}(\overline{Y}_n - \mu) \rightsquigarrow N_k(0, \Sigma).$$

(The sum is taken coordinatewise.) By the Cramér-Wold device the problem can be reduced to finding the limit distribution of the sequences of real-variables

$$t^T\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \mu)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (t^TY_i - t^T\mu).$$

Since the random variables $t^TY_1 - t^T\mu, t^TY_2 - t^T\mu, \ldots$ are i.i.d. with zero mean and variance $t^T\Sigma t$ this sequence is asymptotically $N_1(0, t^T\Sigma t)$ distributed by the univariate central limit theorem. This is exactly the distribution of t^TX if X possesses a $N_k(0, \Sigma)$ distribution. □

3.7 Delta-method

Let T_n be a sequence of random vectors with values in $\mathbb{R}^k$ and let $\phi: \mathbb{R}^k \to \mathbb{R}^m$ be a given function defined at least on the range of T_n and a neighbourhood of a vector θ. We shall assume that, for given constants $r_n \to \infty$, the sequence $r_n(T_n - \theta)$ converges in distribution, and wish to derive a similar result concerning the sequence $r_n\bigl(\phi(T_n) - \phi(\theta)\bigr)$.

Recall that ϕ is *differentiable* at θ if there exists a linear map (matrix) $\phi'_\theta: \mathbb{R}^k \to \mathbb{R}^m$ such that

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o(\|h\|), \qquad h \to 0.$$

All the expressions in this equation are vectors of length m and $\|h\|$ is the Euclidean norm. The linear map $h \to \phi'_\theta(h)$ is sometimes called a total derivative, as opposed to partial derivatives. A sufficient condition for ϕ to be (totally) differentiable is that all partial derivatives $\partial\phi_j(x)/\partial x_i$ exist for x in a neighbourhood of θ and are continuous at θ. (Just existence of the partial derivatives is not enough.) In any case the total derivative is found from the partial derivatives. If ϕ is differentiable, then it is partially differentiable and the derivative map $h \to \phi'_\theta(h)$ is matrix multiplication by the matrix

$$\phi'_\theta = \begin{pmatrix} \frac{\partial\phi_1}{\partial x_1}(\theta) & \cdots & \frac{\partial\phi_1}{\partial x_k}(\theta) \\ \vdots & & \vdots \\ \frac{\partial\phi_m}{\partial x_1}(\theta) & \cdots & \frac{\partial\phi_m}{\partial x_k}(\theta) \end{pmatrix}.$$

If the dependence of the derivative ϕ'_θ on θ is continuous, then ϕ is called continuously differentiable.

3.15 Theorem. *Let $\phi: \mathbb{R}^k \to \mathbb{R}^m$ be a measurable map defined on a subset of $\mathbb{R}^k$ and differentiable at θ. Let T_n be random vectors taking their values in the domain of ϕ. If $r_n(T_n - \theta) \rightsquigarrow T$ for numbers $r_n \to \infty$, then $r_n\bigl(\phi(T_n) - \phi(\theta)\bigr) \rightsquigarrow \phi'_\theta(T)$. Moreover, the difference between $r_n\bigl(\phi(T_n) - \phi(\theta)\bigr)$ and $\phi'_\theta\bigl(r_n(T_n - \theta)\bigr)$ converges to zero in probability.*

Proof. Because $r_n \to \infty$, we have by Slutsky's lemma $T_n - \theta = (1/r_n)r_n(T_n - \theta) \rightsquigarrow 0T = 0$ and hence $T_n - \theta$ converges to zero in probability. Define a function g by

$$g(0) = 0, \qquad\qquad g(h) = \frac{\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)}{\|h\|}, \qquad \text{if } h \neq 0.$$

Then g is continuous at 0 by the differentiability of ϕ. Therefore, by the continuous mapping theorem, $g(T_n - \theta) \xrightarrow{\mathrm{P}} 0$ and hence, by Slutsky's lemma and again the continuous mapping theorem $r_n\|T_n - \theta\|g(T_n - \theta) \xrightarrow{\mathrm{P}} \|T\|0 = 0$. Consequently,

$$r_n\bigl(\phi(T_n) - \phi(\theta) - \phi'_\theta(T_n - \theta)\bigr) = r_n\|T_n - \theta\|g(T_n - \theta) \xrightarrow{\mathrm{P}} 0.$$

This yields the last statement of the theorem. Since matrix multiplication is continuous, $\phi'_\theta\bigl(r_n(T_n - \theta)\bigr) \rightsquigarrow \phi'_\theta(T)$ by the continuous-mapping theorem. Finally, apply Slutsky's lemma to conclude that the sequence $r_n\bigl(\phi(T_n) - \phi(\theta)\bigr)$ has the same weak limit. ∎

A common situation is that $\sqrt{n}(T_n - \theta)$ converges to a multivariate normal distribution $N_k(\mu, \Sigma)$. Then the conclusion of the theorem is that the sequence $\sqrt{n}\bigl(\phi(T_n) - \phi(\theta)\bigr)$ converges in law to the $N_m\bigl(\phi'_\theta \mu, \phi'_\theta \Sigma (\phi'_\theta)^T\bigr)$ distribution.

3.8 Lindeberg Central Limit Theorem

In this section we state, for later reference, a central limit theorem for independent, but not necessarily identically distributed random vectors.

3.16 Theorem (Lindeberg). *For each $n \in \mathbb{N}$ let $Y_{n,1}, \ldots, Y_{n,n}$ be independent random vectors with finite covariance matrices such that*

$$\frac{1}{n}\sum_{i=1}^n \operatorname{Cov} Y_{n,i} \to \Sigma,$$

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\|Y_{n,i}\|^2 1\bigl\{\|Y_{n,i}\| > \varepsilon\sqrt{n}\bigr\} \to 0, \qquad \textit{for every } \varepsilon > 0.$$

Then the sequence $n^{-1/2}\sum_{i=1}^n (Y_{n,i} - \mathrm{E}Y_{n,i})$ converges in distribution to the normal distribution with mean zero and covariance matrix Σ.

3.9 Minimum Contrast Estimators

Many estimators $\hat\theta_n$ of a parameter θ are defined as the point of minimum (or maximum) of a given stochastic process $\theta \mapsto \mathbb{M}_n(\theta)$. In this section we state basic theorems that give the asymptotic behaviour of such *minimum contrast estimators* or *M-estimators* $\hat\theta_n$ in the case that the contrast function $\mathbb{M}_n$ fluctuates around a deterministic, smooth function.

Let $\mathbb{M}_n$ be a sequence of stochastic processes indexed by a subset Θ of $\mathbb{R}^d$, defined on given probability spaces, and let $\hat\theta_n$ be random vectors defined on the same probability spaces with values in Θ such that $\mathbb{M}_n(\hat\theta_n) \le \mathbb{M}_n(\theta)$ for every $\theta \in \Theta$. Typically it will be true that $\mathbb{M}_n(\theta) \xrightarrow{\mathrm{P}} M(\theta)$ for each θ and a given deterministic function M. Then we may expect that $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$ for θ_0 a point of minimum of the map $\theta \to M(\theta)$. The following theorem gives a sufficient condition for this. It applies to the more general situation that the “limit” function M is actually a random process.

For a sequence of random variables X_n we write $X_n \overset{\mathrm{P}}{\gg} 0$ if $X_n > 0$ for every n and $1/X_n = O_P(1)$.

3.17 Theorem. *Let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a semi-metric space Θ such that, for some $\theta_0 \in \Theta$,*

$$\sup_{\theta\in\Theta}\bigl|\mathbb{M}_n(\theta) - M_n(\theta)\bigr| \xrightarrow{\mathrm{P}} 0,$$

$$\inf_{\theta\in\Theta: d(\theta,\theta_0)>\delta} M_n(\theta) - M_n(\theta_0) \overset{\mathrm{P}}{\gg} 0.$$

If $\hat\theta_n$ are random elements with values in Θ with $\mathbb{M}_n(\hat\theta_n) \geq \mathbb{M}_n(\theta_0) - o_P(1)$, then $d(\hat\theta_n, \theta_0) \xrightarrow{\mathrm{P}} 0$.

Proof. By the uniform convergence to zero of $\mathbb{M}_n - M_n$ and the minimizing property of $\hat\theta_n$, we have $M_n(\hat\theta_n) = \mathbb{M}_n(\hat\theta_n) + o_P(1) \leq \mathbb{M}_n(\theta_0) + o_P(1) = M_n(\theta_0) + o_P(1)$. Write $Z_n(\delta)$ for the left side of the second equation in the display of the theorem. Then $d(\hat\theta_n, \theta_0) > \delta$ implies that $M_n(\hat\theta_n) - M_n(\theta_0) \geq Z_n(\delta)$. Combined with the preceding this implies that $Z_n(\delta) \leq o_P(1)$. By assumption the probability of this event tends to zero. ■

If the limit criterion function $\theta \to M(\theta)$ is smooth and takes its minimum at the point θ_0, then its first derivative must vanish at θ_0, and the second derivative V must be positive definite. Thus it possesses a parabolic approximation $M(\theta) = M(\theta_0) + \frac{1}{2}(\theta-\theta_0)^T V(\theta-\theta_0)$ around θ_0. The random criterion function $\mathbb{M}_n$ can be thought of as the limiting criterion function plus the random perturbation $\mathbb{M}_n - M$ and possesses approximation

$$\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) \approx \tfrac{1}{2}(\theta-\theta_0)^T V(\theta-\theta_0) + \bigl[(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_0)\bigr].$$

We shall assume that the term in square brackets possesses a linear approximation of the form $(\theta-\theta_0)^T Z_n/\sqrt{n}$. If we ignore all the remainder terms and minimize the quadratic form

$$\theta - \theta_0 \mapsto \tfrac{1}{2}(\theta-\theta_0)^T V(\theta-\theta_0) + (\theta-\theta_0)^T Z_n/\sqrt{n}$$

over $\theta - \theta_0$, then we find that the minimum is taken for $\theta - \theta_0 = -V^{-1}Z_n/\sqrt{n}$. Thus we expect that the M-estimator $\hat\theta_n$ satisfies $\sqrt{n}(\hat\theta_n - \theta_0) = -V^{-1}Z_n + o_P(1)$. This derivation is made rigorous in the following theorem.

3.18 Theorem. *Let $\mathbb{M}_n$ be stochastic processes indexed by an open subset Θ of Euclidean space and let $M\colon \Theta \to \mathbb{R}$ be a deterministic function. Assume that $\theta \to M(\theta)$ is twice continuously differentiable at a point of minimum θ_0 with nonsingular second-derivative matrix V.*[‡] *Suppose that*

$$\begin{aligned} &r_n(\mathbb{M}_n - M)(\tilde\theta_n) - r_n(\mathbb{M}_n - M)(\theta_0) \\ &\qquad = (\tilde\theta_n - \theta_0)' Z_n + o_P^*\bigl(\|\tilde\theta_n - \theta_0\| + r_n\|\tilde\theta_n - \theta_0\|^2 + r_n^{-1}\bigr), \end{aligned}$$

[‡] It suffices that a two-term Taylor expansion is valid at θ_0.

for every random sequence $\tilde\theta_n = \theta_0 + o_P^*(1)$ and a uniformly tight sequence of random vectors Z_n. If the sequence $\hat\theta_n$ converges in outer probability to θ_0 and satisfies $\mathbb{M}_n(\hat\theta_n) \le \inf_\theta \mathbb{M}_n(\theta) + o_P(r_n^{-2})$ for every n, then

$$r_n(\hat\theta_n - \theta_0) = -V^{-1}Z_n + o_P^*(1).$$

If it is known that the sequence $r_n(\hat\theta_n - \theta_0)$ is uniformly tight, then the displayed condition needs to be verified for sequences $\tilde\theta_n = \theta_0 + O_P^*(r_n^{-1})$ only.

Proof. The stochastic differentiability condition of the theorem together with the two-times differentiability of the map $\theta \to M(\theta)$ yields for every sequence $\tilde h_n = o_P^*(1)$

$$\begin{aligned}\mathbb{M}_n(\theta_0 + \tilde h_n) - \mathbb{M}_n(\theta_0) &= \tfrac12 \tilde h_n' V \tilde h_n + r_n^{-1}\tilde h_n' Z_n \\ &\quad + o_P^*\bigl(\|\tilde h_n\|^2 + r_n^{-1}\|\tilde h_n\| + r_n^{-2}\bigr).\end{aligned} \tag{3.1}$$

For $\tilde h_n$ chosen equal to $\hat h_n = \hat\theta_n - \theta_0$, the left side (and hence the right side) is at most $o_P(r_n^{-2})$ by the definition of $\hat\theta_n$. In the right side the term $\tilde h_n' V \tilde h_n$ can be bounded below by $c\|\tilde h_n\|^2$ for a positive constant c, since the matrix V is strictly positive definite. Conclude that

$$c\|\hat h_n\|^2 + r_n^{-1}\|\hat h_n\| O_P(1) + o_P\bigl(\|\hat h_n\|^2 + r_n^{-2}\bigr) \le o_P(r_n^{-2}).$$

Complete the square to see that this implies that

$$\bigl(c + o_P(1)\bigr)\Bigl(\|\hat h_n\| - O_P(r_n^{-1})\Bigr)^2 \le O_P(r_n^{-2}).$$

This can be true only if $\|\hat h_n\| = O_P^*(r_n^{-1})$.

For any sequence $\tilde h_n$ of the order $O_P^*(r_n^{-1})$, the three parts of the remainder term in (3.1) are of the order $o_P(r_n^{-2})$. Apply this with the choices $\hat h_n$ and $-r_n^{-1}V^{-1}Z_n$ to conclude that

$$\begin{aligned}\mathbb{M}_n(\theta_0 + \hat h_n) - \mathbb{M}_n(\theta_0) &= \tfrac12 \hat h_n' V \hat h_n + r_n^{-1}\hat h_n' Z_n + o_P^*(r_n^{-2}), \\ \mathbb{M}_n(\theta_0 - r_n^{-1}V^{-1}Z_n) - \mathbb{M}_n(\theta_0) &= -\tfrac12 r_n^{-2} Z_n' V^{-1} Z_n + o_P^*(r_n^{-2}).\end{aligned}$$

The left-hand side of the first equation is smaller than the second, up to an $o_P^*(r_n^{-2})$-term. Subtract the second equation from the first to find that

$$\tfrac12 (\hat h_n + r_n^{-1}V^{-1}Z_n)' V (\hat h_n + r_n^{-1}V^{-1}Z_n) \le o_P(r_n^{-2}).$$

Since V is strictly positive definite, this yields the first assertion of the theorem.

If it is known that the sequence $\hat\theta_n$ is r_n-consistent, then the middle part of the preceding proof is unnecessary and we can proceed to inserting $\hat h_n$ and $-r_n^{-1}V^{-1}Z_n$ in (3.1) immediately. The latter equation is then needed for sequences $\tilde h_n = O_P^*(r_n^{-1})$ only. ∎

4 Central Limit Theorem

The classical central limit theorem asserts that the mean of independent, identically distributed random variables with finite variance is asymptotically normally distributed. In this chapter we extend this to certain dependent and/or nonidentically distributed sequences.

Given a stationary time series X_t let $\overline{X}_n$ be the average of the variables $X_1, \ldots, X_n$. If μ and γ_X are the mean and auto-covariance function of the time series, then, by the usual rules for expectation and variance,

$$\mathrm{E}\overline{X}_n = \mu,$$

$$\mathrm{var}(\sqrt{n}\overline{X}_n) = \frac{1}{n}\sum_{s=1}^n\sum_{t=1}^n \mathrm{cov}(X_s, X_t) = \sum_{h=-n}^n \Bigl(\frac{n-|h|}{n}\Bigr)\gamma_X(h). \tag{4.1}$$

In the expression for the variance every of the terms $\bigl(n-|h|\bigr)/n$ is bounded by 1 and converges to 1 as $n \to \infty$. If $\sum\bigl|\gamma_X(h)\bigr| < \infty$, then we can apply the dominated convergence theorem and obtain that $\mathrm{var}(\sqrt{n}\overline{X}_n) \to \sum_h \gamma_X(h)$. In any case

$$\mathrm{var}\,\sqrt{n}\overline{X}_n \le \sum_h \bigl|\gamma_X(h)\bigr|. \tag{4.2}$$

Hence absolute convergence of the series of auto-covariances implies that the sequence $\sqrt{n}(\overline{X}_n - \mu)$ is uniformly tight. The purpose of the chapter is to give conditions for this sequence to be asymptotically normally distributed with mean zero and variance $\sum_h \gamma_X(h)$.

Such conditions are of two broad types. One possibility is to assume a particular type of dependence of the variables X_t, such as Markov or martingale properties. Second, we might require that the time series is “not too far” from being a sequence of independent variables. We present three sets of sufficient conditions of the second type. These have in common that they all require that the elements X_t and X_{t+h} at large time lags h are approximately independent. We start with the simplest case, that of finitely dependent time series. Next we generalize this in two directions: to linear processes and to mixing time series. The term “mixing” is often used in a vague sense to refer to time series’ whose elements at large time lags are approximately independent. For a central limit theorem we then require that the time series is “sufficiently mixing” and this can be made precise in several ways. In ergodic theory the term “mixing” is also used in a precise sense. We briefly discuss the application to the law of large numbers.

* **4.1** EXERCISE. Suppose that the series $v:=\sum_h \gamma_X(h)$ converges (not necessarily absolutely). Show that var $\sqrt{n}\overline{X}_n \to v$. [Write var $\sqrt{n}\overline{X}_n$ as $\overline{v}_n$ for $v_h=\sum_{|j|<h}\gamma_X(j)$ and apply Cesaro’s lemma: if $v_n\to v$, then $\overline{v}_n \to v$.]

4.1 Finite Dependence

A time series X_t is called *m-dependent* if the random vectors $(\ldots, X_{t-1}, X_t)$ and $(X_{t+m+1}, X_{t+m+2},\ldots)$ are independent for every $t\in\mathbb{Z}$. In other words, “past” and “future” are independent if m “present” variables are left out.

4.2 EXERCISE. Show that the moving average $X_t=Z_t+\theta Z_{t-1}$ considered in Example 1.6 is 1-dependent.

4.3 EXERCISE. Show that “0-dependent” is equivalent to “independent”.

4.4 Theorem. *Let X_t be a strictly stationary, m-dependent time series with mean zero and finite variance. Then the sequence $\sqrt{n}\overline{X}_n$ converges in distribution to a normal distribution with mean 0 and variance $\sum_{h=-m}^m \gamma_X(h)$.*

Proof. Choose a (large) integer l and divide $X_1,\ldots,X_n$ into $r=\lfloor n/l\rfloor$ groups of size l and a remainder group of size $n-rl<l$. Let $A_{1,l},\ldots,A_{r,l}$ and $B_{1,l},\ldots,B_{r,l}$ be the sums of the first $l-m$ and last m of the variables X_i in the r groups. (Cf. Figure 4.1.) Then both $A_{1,l},\ldots,A_{r,l}$ and $B_{1,l},\ldots,B_{r,l}$ are sequences of independent identically distributed random variables (but the two sequences may be dependent) and

$$\sum_{i=1}^n X_i=\sum_{j=1}^r A_{j,l}+\sum_{j=1}^r B_{j,l}+\sum_{i=rl+1}^n X_i. \tag{4.3}$$

For fixed l and $n\to\infty$ (hence $r\to\infty$) the classical central limit theorem applied to the variables $A_{j,l}$ yields

$$\frac{1}{\sqrt{n}}\sum_{j=1}^r A_{j,l}=\sqrt{\frac{r}{n}}\frac{1}{\sqrt{r}}\sum_{j=1}^r A_{j,l}\rightsquigarrow\frac{1}{\sqrt{l}}N(0,\operatorname{var}A_{1,l}).$$

Furthermore, by the triangle inequality, for fixed l as $n \to \infty$,

$$\sigma\Bigl(\frac{1}{\sqrt{n}} \sum_{i=rl+1}^{n} X_i\Bigr) \le \frac{l}{\sqrt{n}} \sigma(X_1) \to 0.$$

Because the mean of the variables $n^{-1/2} \sum_{i=rl+1}^{n} X_i$ is zero, this sequence converges to zero in probability by Chebyshev's inequality. We conclude by Slutsky's lemma that, as $n \to \infty$,

$$S_{n,l} := \frac{1}{\sqrt{n}} \sum_{j=1}^{r} A_{j,l} + \frac{1}{\sqrt{n}} \sum_{i=rl+1}^{n} X_i \rightsquigarrow N\Bigl(0, \frac{1}{l} \operatorname{var} A_{1,l}\Bigr).$$

This is true for every fixed l. If $l \to \infty$, then

$$\frac{1}{l} \operatorname{var} A_{1,l} = \frac{1}{l} \operatorname{var} \sum_{i=1}^{l-m} X_i = \sum_{h=m-l}^{l-m} \frac{l-m-|h|}{l} \gamma_X(h) \to v := \sum_{h=-m}^{m} \gamma_X(h).$$

By Lemma 3.10 there exists a sequence $l_n \to \infty$ such that $S_{n,l_n} \rightsquigarrow N(0,v)$. Let $r_n = \lfloor n/l_n \rfloor$ be the corresponding sequence of values of r_n, so that $r_n/n \to 0$. By the strict stationarity of the series X_t each $B_{j,l}$ is equal in distribution to $X_1 + \cdots + X_m$ and hence is independent of (j,l). Hence $\operatorname{var} B_{j,l}$ is independent of j and l and

$$\mathrm{E}\Bigl(\frac{1}{\sqrt{n}} \sum_{j=1}^{r} B_{j,l_n}\Bigr)^2 = \frac{r}{n} \operatorname{var} B_{1,l_n} \to 0.$$

Thus the sequence of random variables in the left side converges to zero in probablity, by Chebyshev's inequality. In view of (4.3) another application of Slutsky's lemma gives the result. ∎

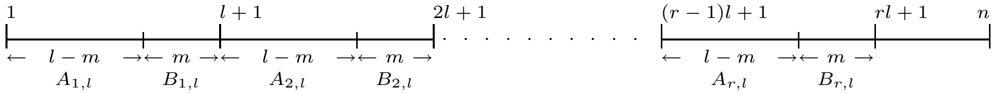
1
$l+1$
$2l+1$
$(r-1)l+1$
$rl+1$
n
$\leftarrow\ l-m\ \rightarrow$
$\leftarrow\ m\ \rightarrow$
$\leftarrow\ l-m\ \rightarrow$
$\leftarrow\ m\ \rightarrow$
$\leftarrow\ l-m\ \rightarrow$
$\leftarrow\ m\ \rightarrow$
$A_{1,l}$
$B_{1,l}$
$A_{2,l}$
$B_{2,l}$
$A_{r,l}$
$B_{r,l}$

Figure 4.1. Blocking of observations in the proof of Theorem 4.4.

4.2 Linear Processes

In this section we extend the central limit theorem from finitely dependent time series to *linear processes*. These are processes that can be represented as infinite moving averages. Given a sequence $\ldots, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ of independent and identically distributed variables with $\mathrm{E}Z_t = 0$, a constant μ, and constants ψ_j with $\sum_j |\psi_j| < \infty$, we assume that

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}. \tag{4.4}$$

This may seem special, but we shall see later that this includes, for instance, the rich class of all ARMA-processes.

By (iii) of Lemma 1.28 the covariance function of a linear process is given by $\gamma_X(h) = \sigma^2 \sum_j \psi_j \psi_{j+h}$, where $\sigma^2 = \operatorname{var} Z_t$, and hence the asymptotic variance of $\sqrt{n}\overline{X}_n$ is given by

$$v{:} = \sum_h \gamma_X(h) = \sigma^2 \sum_h \sum_j \psi_j \psi_{j+h} = \sigma^2 \Bigl(\sum_j \psi_j\Bigr)^2.$$

4.5 Theorem. *Suppose that (4.4) holds for an i.i.d. sequence Z_t with mean zero and finite variance and numbers ψ_j with $\sum_j |\psi_j| < \infty$. Then the sequence $\sqrt{n}(\overline{X}_n - \mu)$ converges in distribution to a normal distribution with mean zero and variance v.*

Proof. We can assume without loss of generality that $\mu = 0$. For a fixed (large) integer m define the time series

$$X_t^m = \sum_{|j|\le m} \psi_j Z_{t-j} = \sum_j \psi_j^m Z_{t-j},$$

where $\psi_j^m = \psi_j$ if $|j| \le m$ and 0 otherwise. Then X_t^m is $(2m+1)$-dependent and strictly stationary. By Theorem 4.4, the sequence $\sqrt{n}\overline{X_n^m}$ converges in distribution to a normal distribution with mean zero and variance

$$v_m{:} = \sum_h \gamma_{X^m}(h) = \sigma^2 \sum_h \sum_j \psi_j^m \psi_{j+h}^m = \sigma^2 \Bigl(\sum_{|j|\le m} \psi_j\Bigr)^2.$$

The first equality follows from (iii) of Lemma 1.28. As $m \to \infty$ this variance converges to v. Because $N(0, v_m) \rightsquigarrow N(0, v)$, Lemma 3.10 guarantees that there exists a sequence $m_n \to \infty$ such that $\sqrt{n}\overline{X_n^{m_n}} \rightsquigarrow N(0, v)$.

In view of Slutsky's lemma the proof will be complete once we have shown that $\sqrt{n}(\overline{X}_n - \overline{X_n^{m_n}}) \xrightarrow{\mathrm{P}} 0$. This concerns the average $\overline{Y_n^{m_n}}$ of the differences $Y_t^m = X_t - X_t^m = \sum_{|j|>m} \psi_j Z_{t-j}$. These satisfy

$$\mathrm{E}\bigl(\sqrt{n}\overline{X}_n - \sqrt{n}\overline{X_n^{m_n}}\bigr)^2 = \operatorname{var} \sqrt{n}\overline{Y_n^{m_n}} \le \sum_h \bigl|\gamma_{Y^{m_n}}(h)\bigr| \le \sigma^2 \Bigl(\sum_{|j|>m_n} |\psi_j|\Bigr)^2.$$

The inequalities follow by (4.1) and Lemma 1.28(iii). The right side converges to zero as $m_n \to \infty$. ∎

4.3 Strong Mixing

The *α-mixing coefficients* (or *strong mixing coefficients*) of a time series X_t are defined by $\alpha(0) = \frac{1}{2}$ and for $k \in \mathbb{N}$[♭]

$$\alpha(h) = 2 \sup_t \sup_{\substack{A \in \sigma(\ldots, X_{t-1}, X_t) \\ B \in \sigma(X_{t+h}, X_{t+h+1}, \ldots)}} \bigl|\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B)\bigr|.$$

The events A and B in this display depend on elements X_t of the "past" and "future" that are h time lags apart. Thus the α-mixing coefficients measure the extent by which events A and B that are separated by h time instants fail to satisfy the equality $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$, which is valid for independent events. If the series X_t is strictly stationary, then the supremum over t is unnecessary, and the mixing coefficient $\alpha(h)$ can be defined using the σ-fields $\sigma(\ldots, X_{-1}, X_0)$ and $\sigma(X_h, X_{h+1}, \ldots)$ only.

It is immediate from their definition that the coefficients $\alpha(1), \alpha(2), \ldots$ are decreasing and nonnegative. Furthermore, if the time series is m-dependent, then $\alpha(h) = 0$ for $h > m$.

4.6 EXERCISE. Show that $\alpha(1) \leq \frac{1}{2} \equiv \alpha(0)$. [Apply the inequality of Cauchy-Schwarz to $\mathrm{P}(A \cap B) - \mathrm{P}(A)\mathrm{P}(B) = \mathrm{cov}(1_A, 1_B)$.]

If $\alpha(h) \to 0$ as $h \to \infty$, then the time series X_t is called *α-mixing* or *strong mixing.* Then events connected to time sets that are far apart are "approximately independent". For a central limit theorem to hold, we also need that the convergence to 0 takes place at a sufficient speed, dependent on the "sizes" of the variables X_t.

A precise formulation can best be given in terms of the inverse function of the mixing coefficients. We can extend α to a function $\alpha\colon [0, \infty) \to [0, 1]$ by defining it to be constant on the intervals $[h, h+1)$ for integers h. This yields a monotone function that decreases in steps from $\alpha(0) = \frac{1}{2}$ to 0 at infinity in the case that the time series is mixing. The generalized inverse $\alpha^{-1}\colon [0, 1] \to [0, \infty)$ is defined by

$$\alpha^{-1}(u) = \inf\bigl\{x \geq 0\colon \alpha(x) \leq u\bigr\} = \sum_{h=0}^{\infty} 1_{u < \alpha(h)}.$$

Similarly, the quantile function F_X^{-1} of a random variable X is the generalized inverse of the distribution function F_X of X, and is given by

$$F_X^{-1}(1-u) = \inf\{x\colon 1 - F_X(x) \leq u\}.$$

4.7 Theorem. *If X_t is a strictly stationary time series with mean zero such that $\int_0^1 \alpha^{-1}(u) F_{|X_0|}^{-1}(1-u)^2\,du < \infty$, then the series $v = \sum_h \gamma_X(h)$ converges absolutely and $\sqrt{n}\overline{X}_n \rightsquigarrow N(0, v)$.*

At first sight the condition of the theorem is complicated. Finiteness of the integral requires that the mixing coefficients converge to zero fast enough, but the rate at which

[♭] We denote by $\sigma(X_t\colon t \in I)$ the σ-field generated by the set of random variables $\{X_t\colon t \in I\}$.

this must happen is also dependent on the tails of the variables X_t. To make this concrete we can derive finiteness of the integral under a combination of a mixing and a moment condition. If $c^r := \mathrm{E}|X_0|^r < \infty$ for some $r > 2$, then $1 - F_{|X_0|}(x) \leq c^r/x^r$ by Markov's inequality and hence $F_{|X_0|}^{-1}(1-u) \leq c/u^{1/r}$. Then we obtain the bound

$$\int_0^1 \alpha^{-1}(u) F_{|X_0|}^{-1}(1-u)^2\,du \leq \sum_{h=0}^\infty \int_0^1 1_{u<\alpha(h)} \frac{c^2}{u^{2/r}}\,du = \frac{c^2 r}{r-2} \sum_{h=0}^\infty \alpha(h)^{1-2/r}.$$

Thus the moment condition $\mathrm{E}|X_t|^r < \infty$ and the mixing condition $\sum_{h=0}^\infty \alpha(h)^{1-2/r} < \infty$ together are sufficient for the central limit theorem. This allows a trade-off between moments and mixing: for larger values of r the moment condition is more restrictive, but the mixing condition is weaker.

4.8 EXERCISE **(Case $r = \infty$).** Show that $\int_0^1 \alpha^{-1}(u) F_{|X_0|}^{-1}(1-u)^2\,du$ is bounded above by $\|X_0\|_\infty^2 \sum_{h=0}^\infty \alpha(h)$. [Note that $F_{|X_0|}^{-1}(1-U)$ is distributed as $|X_0|$ if U is uniformly distributed and hence is bounded by $\|X_0\|_\infty$ almost surely.]

4.9 EXERCISE. Show that $\int_0^1 \alpha^{-1}(u) F_{|X_0|}^{-1}(1-u)^2\,du \leq (m+1)\mathrm{E}X_0^2$ if the time series X_t is m-dependent. Recover Theorem 4.4 from Theorem 4.7.

4.10 EXERCISE. Show that $\int_0^1 \alpha^{-1}(u) F_{|X_0|}^{-1}(1-u)^2\,du < \infty$ implies that $\mathrm{E}|X_0|^2 < \infty$.

The key to the proof of Theorem 4.7 is a lemma that bounds covariances in terms of mixing coefficients. Let $\|X\|_p$ denote the L_p-norm of a random variable X, i.e.

$$\|X\|_p = \bigl(\mathrm{E}|X|^p\bigr)^{1/p}, \quad 1 \leq p < \infty, \qquad \|X\|_\infty = \inf\{M: \mathrm{P}(|X| \leq M) = 1\}.$$

Recall *Hölder's inequality*: for any pair of numbers $p, q > 0$ (possibly infinite) with $p^{-1} + q^{-1} = 1$ and random variables X and Y

$$\mathrm{E}|XY| \leq \|X\|_p \|Y\|_q.$$

For $p = q = 2$ this is precisely the inequality of Cauchy-Schwarz. The other case that will be of interest to us is the case $p = 1, q = \infty$, for which the inequality is easy to prove. By repeated application the inequality can be extended to more than two random variables. For instance, for any numbers $p, q, r > 0$ with $p^{-1} + q^{-1} + r^{-1} = 1$ and random variables X, Y, and Z

$$\mathrm{E}|XYZ| \leq \|X\|_p \|Y\|_q \|Z\|_r.$$

4.11 Lemma (Covariance bound). *Let X_t be a time series with α-mixing coefficients $\alpha(h)$ and let Y and Z be random variables that are measurable relative to $\sigma(\ldots, X_{-1}, X_0)$ and $\sigma(X_h, X_{h+1}, \ldots)$, respectively, for a given $h \geq 0$. Then, for any $p, q, r > 0$ such that $p^{-1} + q^{-1} + r^{-1} = 1$,*

$$\bigl|\mathrm{cov}(Y, Z)\bigr| \leq 2 \int_0^{\alpha(h)} F_{|Y|}^{-1}(1-u) F_{|Z|}^{-1}(1-u)\,du \leq 2\alpha(h)^{1/p} \|Y\|_q \|Z\|_r.$$

Proof. By the definition of the mixing coefficients, we have, for every $y, z > 0$,

$$\bigl|\mathrm{cov}(1_{Y^+>y}, 1_{Z^+>z})\bigr| \le \tfrac{1}{2}\alpha(h).$$

The same inequality is valid with Y^+ and/or Z^+ replaced by Y^- and/or Z^-. It follows that

$$\bigl|\mathrm{cov}(1_{Y^+>y} - 1_{Y^->y}, 1_{Z^+>z} - 1_{Z^->z})\bigr| \le 2\alpha(h).$$

Because $\bigl|\mathrm{cov}(U,V)\bigr| \le 2(\mathrm{E}|U|)\|V\|_\infty$ for any pair of random variables U, V (the simplest Hölder inequality), we obtain that the covariance on the left side of the preceding display is also bounded by $2\bigl(\mathrm{P}(Y^+ > y) + \mathrm{P}(Y^- > y)\bigr)$. Yet another bound for the covariance is obtained by interchanging the roles of Y and Z. Combining the three inequalities, we see that, for any $y, z > 0$,

$$\begin{aligned}\bigl|\mathrm{cov}(1_{Y^+>y} - 1_{Y^->y}, 1_{Z^+>z} - 1_{Z^->z})\bigr| &\le 2\alpha(h) \wedge 2\mathrm{P}(|Y| > y) \wedge 2\mathrm{P}(|Z| > z)\\ &= 2\int_0^{\alpha(h)} 1_{1-F_{|Y|}(y)>u}\, 1_{1-F_{|Z|}(z)>u}\, du.\end{aligned}$$

Next we write $Y = Y^+ - Y^- = \int_0^\infty (1_{Y^+>y} - 1_{Y^->y})\, dy$ and similarly for Z, to obtain, by Fubini's theorem,

$$\begin{aligned}\bigl|\mathrm{cov}(Y,Z)\bigr| &= \Bigl|\int_0^\infty\int_0^\infty \mathrm{cov}(1_{Y^+>y} - 1_{Y^->y}, 1_{Z^+>z} - 1_{Z^->z})\, dy\, dz\Bigr|\\ &\le 2\int_0^\infty\int_0^\infty\int_0^{\alpha(h)} 1_{F_{|Y|}(y)<1-u}\, 1_{F_{|Z|}(z)<1-u}\, du\, dy\, dz.\end{aligned}$$

Any pair of a distribution and a quantile function satisfies $F_X(x) < u$ if and only $x < F_X^{-1}(u)$, for every x and u. We can conclude the proof of the first inequality of the lemma by another application of Fubini's theorem.

The second inequality follows upon noting that $F_{|Y|}^{-1}(1-U)$ is distributed as $|Y|$ if U is uniformly distributed on $[0,1]$, and next applying Hölder's inequality. ■

Proof of Theorem 4.7. As a consequence of Lemma 4.11 we find that

$$\sum_{h\ge0}\bigl|\gamma_X(h)\bigr| \le 2\sum_{h\ge0}\int_0^{\alpha(h)} F_{|X_0|}^{-1}(1-u)^2\, du = 2\int_0^1 \alpha^{-1}(u)\, F_{|X_0|}^{-1}(1-u)^2\, du.$$

This already proves the first assertion of Theorem 4.7. Furthermore, in view of (4.2) and the symmetry of the auto-covariance function,

$$\mathrm{var}\,\sqrt{n}\overline{X}_n \le 4\int_0^1 \alpha^{-1}(u)\, F_{|X_0|}^{-1}(1-u)^2\, du. \tag{4.5}$$

For a given $M > 0$ let $X_t^M = X_t 1\{|X_t| \le M\}$ and let $Y_t^M = X_t - X_t^M$. Because X_t^M is a measurable transformation of X_t, it is immediate from the definition of the mixing coefficients that the series Y_t^M is mixing with smaller mixing coefficients than the series X_t. Therefore, in view of (4.5)

$$\operatorname{var} \sqrt{n}(\overline{X}_n - \overline{X_n^M}) = \operatorname{var} \sqrt{n}\overline{Y_n^M} \leq 4 \int_0^1 \alpha^{-1}(u)\, F^{-1}_{|Y_0^M|}(1-u)^2\, du.$$

Because $Y_0^M = 0$ whenever $|X_0| \leq M$, it follows that $Y_0^M \rightsquigarrow 0$ as $M \to \infty$ and hence $F^{-1}_{|Y_0^M|}(u) \to 0$ for every $u \in (0,1)$. Furthermore, because $|Y_0^M| \leq |X_0|$, its quantile function is bounded above by the quantile function of $|X_0|$. By the dominated convergence theorem the integral in the preceding display converges to zero as $M \to \infty$, and hence the variance in the left side converges to zero as $M \to \infty$, uniformly in n. If we can show that $\sqrt{n}(\overline{X_n^M} - \mathrm{E}X_0^M) \rightsquigarrow N(0, v^M)$ as $n \to \infty$ for $v^M = \lim \operatorname{var} \sqrt{n}\overline{X_n^M}$ and every fixed M, then it follows that $\sqrt{n}(\overline{X}_n - \mathrm{E}X_0) \rightsquigarrow N(0, v)$ for $v = \lim v^M = \lim \operatorname{var} \sqrt{n}\overline{X_n}$, by Lemma 3.10, and the proof is complete.

Thus it suffices to prove the theorem for uniformly bounded variables X_t. Let M be the uniform bound.

Fix some sequence $m_n \to \infty$ such that $\sqrt{n}\alpha(m_n) \to 0$ and $m_n/\sqrt{n} \to 0$. Such a sequence exists. To see this, first note that $\sqrt{n}\alpha(\lfloor\sqrt{n}/k\rfloor) \to 0$ as $n \to \infty$, for every fixed k. (See Problem 4.12). Thus by Lemma 3.10 there exists $k_n \to \infty$ such that $\sqrt{n}\alpha(\lfloor\sqrt{n}/k_n\rfloor) \to 0$ as $k_n \to \infty$. Now set $m_n = \lfloor\sqrt{n}/k_n\rfloor$. For simplicity write m for m_n. Also let it be silently understood that all summation indices are restricted to the integers $1, 2, \ldots, n$, unless indicated otherwise.

Let $S_n = n^{-1/2}\sum_{t=1}^n X_t$ and, for every given t, set $S_n(t) = n^{-1/2}\sum_{|j-t|<m} X_j$. Because $|e^{i\lambda} - 1 - i\lambda| \leq \frac{1}{2}\lambda^2$ for every $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}
\Big|\mathrm{E}\Big[\frac{1}{\sqrt{n}}\sum_{t=1}^n X_t e^{i\lambda S_n}\Big(e^{-i\lambda S_n(t)} - 1 + i\lambda S_n(t)\Big)\Big]\Big| &\leq \frac{\lambda^2 n M}{2\sqrt{n}}\sum_{t=1}^n \mathrm{E}S_n^2(t) \\
&= \frac{\lambda^2 M}{2\sqrt{n}}\sum_{t=1}^n \sum_{|i-t|<m}\sum_{|j-t|<m} \gamma_X(i-j) \\
&\leq \frac{\lambda^2 M}{2\sqrt{n}} m \sum_h \big|\gamma_X(h)\big| \to 0.
\end{aligned}$$

Furthermore, with $A_n(t)$ and $B_n(t)$ defined as $n^{-1/2}$ times the sum of the X_j with $1 \leq j \leq t-m$ and $t+m \leq j \leq n$, respectively, we have $S_n - S_n(t) = A_n(t) + B_n(t)$ and

$$\begin{aligned}
&\Big|\mathrm{E}\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^n X_t e^{i\lambda S_n} e^{-i\lambda S_n(t)}\Big)\Big| \\
&\quad\leq \frac{1}{\sqrt{n}}\sum_{t=1}^n \Big|\operatorname{cov}\big(X_t e^{i\lambda A_n(t)}, e^{i\lambda B_n(t)}\big) + \operatorname{cov}\big(X_t, e^{i\lambda A_n(t)}\big)\mathrm{E}e^{i\lambda B_n(t)}\Big| \\
&\quad\leq 4\sqrt{n}M\alpha(m) \to 0,
\end{aligned}$$

by the second inequality of Lemma 4.11, with $p = 1$ and $q = r = \infty$. Combining the preceding pair of displays we see that

$$\mathrm{E}S_n e^{i\lambda S_n} = \mathrm{E}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t e^{i\lambda S_n} i\lambda S_n(t) + o(1) = i\lambda \mathrm{E}\Bigl(e^{i\lambda S_n}\frac{1}{n}\sum\sum_{|s-t|<m} X_s X_t\Bigr) + o(1).$$

If we can show that $n^{-1}\sum\sum_{|s-t|<m} X_s X_t$ converges in mean to v, then the right side of the last display is asymptotically equivalent to $i\lambda \mathrm{E}e^{i\lambda S_n} v$, and the theorem is proved in view of Lemma 3.13.

In fact, we show that $n^{-1}\sum\sum_{|s-t|<m} X_s X_t \to v$ in second mean. First,

$$\mathrm{E}\frac{1}{n}\sum\sum_{|s-t|<m} X_s X_t = \sum_{|h|<m}\Bigl(\frac{n-|h|}{n}\Bigr)\gamma_X(h) \to v,$$

By the dominated convergence theorem, in view of (4.2). Second,

$$\mathrm{var}\Bigl(\frac{1}{n}\sum\sum_{|s-t|<m} X_s X_t\Bigr) \le \frac{1}{n^2}\sum\sum_{|s-t|<m}\sum\sum_{|i-j|<m}\bigl|\mathrm{cov}(X_s X_t, X_i X_j)\bigr|.$$

The first double sum on the right can be split in the sums over the pairs (s,t) with $s < t$ and $s \ge t$, respectively, and similarly for the second double sum relative to (i,j). By symmetry the right side is bounded by

$$\begin{aligned}
&\frac{4}{n^2}\sum\sum_{\substack{|s-t|<m\\ s\le t}}\sum\sum_{\substack{|i-j|<m\\ i\le j}}\bigl|\mathrm{cov}(X_s X_t, X_i X_j)\bigr| \\
&\quad\le \frac{4}{n^2}\sum_{s=1}^{n}\sum_{t=0}^{m}\sum_{i=1}^{n}\sum_{j=0}^{m}\bigl|\mathrm{cov}(X_s X_{s+t}, X_i X_{i+j})\bigr| \\
&\quad\le \frac{8}{n^2}\sum_{s=1}^{n}\sum_{t=0}^{m}\sum_{i=1}^{n}\sum_{j=0}^{m}\bigl|\mathrm{cov}(X_s X_{s+t}, X_{s+i} X_{s+i+j})\bigr|,
\end{aligned}$$

by the same argument, this time splitting the sums over $s \le i$ and $s > i$ and using symmetry between s and i. If $i \ge t$, then the covariance in the sum is bounded above by $2\alpha(i-t)M^4$, by Lemma 4.11, because there are $i-t$ time instants between $X_s X_{s+t}$ and $X_{s+i}X_{s+i+j}$. If $i < t$, then we rewrite the absolute covariance as

$$\Bigl|\mathrm{cov}(X_s, X_{s+t}X_{s+i}X_{s+i+j}) - \mathrm{cov}(X_s, X_{s+t})\mathrm{E}X_{s+i}X_{s+i+j}\Bigr| \le 4\alpha(i)M^4.$$

Thus the four-fold sum is bounded above by

$$\frac{32}{n^2}\sum_{s=1}^{n}\sum_{t=0}^{m}\sum_{i=1}^{n}\sum_{j=0}^{m}\Bigl(\alpha(i-t)M^4 1_{i\ge t} + \alpha(i)M^4 1_{i<t}\Bigr) \le 64M^4\frac{m^2}{n}\sum_{i\ge 0}\alpha(i).$$

Because $F^{-1}_{|X_0|}$ is bounded away from zero in a neighbourhood of 0, finiteness of the integral $\int_0^1 \alpha^{-1}(u)F^{-1}_{|X_0|}(1-u)^2\,du$ implies that the series on the right converges. This conclude the proof. ■

* **4.12** EXERCISE. Suppose that $\alpha(h)$ is a decreasing sequence of nonnegative numbers $(h = 1, 2, \ldots)$ with $\sum_h \alpha(h) < \infty$. Show that $h\alpha(h) \to 0$ as $h \to \infty$. [First derive, using the monotonicity, that $\sum_h 2^h\alpha(2^h) < \infty$ and conclude from this that $2^h\alpha(2^h) \to 0$. Next use the monotonicity again "to fill the gaps".]

* 4.4 Uniform Mixing

There are several other types of mixing coefficients. The *ϕ-mixing coefficients* or *uniform mixing coefficients* of a strictly stationary time series X_t are defined by

$$\phi(h) = \sup_{\substack{A\in\sigma(\ldots,X_{-1},X_0),\mathrm{P}(A)\neq 0\\ B\in\sigma(X_h,X_{h+1},\ldots)}} \bigl|\mathrm{P}(B|\,A) - \mathrm{P}(B)\bigr|,$$

$$\tilde{\phi}(h) = \sup_{\substack{A\in\sigma(\ldots,X_{-1},X_0)\\ B\in\sigma(X_h,X_{h+1},\ldots),\mathrm{P}(B)\neq 0}} \bigl|\mathrm{P}(A|\,B) - \mathrm{P}(A)\bigr|.$$

It is immediate from the definitions that $\alpha(h) \leq 2\bigl(\phi(h)\wedge\tilde{\phi}(h)\bigr)$. Thus a ϕ-mixing time series is always α-mixing. It appears that conditions in terms of ϕ-mixing are often much more restrictive, even though there is no complete overlap.

4.13 Lemma (Covariance bound). *Let X_t be a strictly stationary time series with ϕ-mixing coefficients $\phi(h)$ and $\tilde{\phi}(h)$ and let Y and Z be random variables that are measurable relative to $\sigma(\ldots,X_{-1},X_0)$ and $\sigma(X_h,X_{h+1},\ldots)$, respectively, for a given $h\geq 0$. Then, for any $p,q>0$ with $p^{-1}+q^{-1}=1$,*

$$\bigl|\mathrm{cov}(Y,Z)\bigr| \leq 2\phi(h)^{1/p}\tilde{\phi}(h)^{1/q}\|Y\|_p\|Z\|_q.$$

Proof. Let Q be the measure $P^{Y,Z} - P^Y\otimes P^Z$ on $\mathbb{R}^2$, and let $|Q|$ be its absolute value. Then

$$\bigl|\mathrm{cov}(Y,Z)\bigr| = \Bigl|\int\!\!\int yz\,dQ(y,z)\Bigr| \leq \Bigl(\int\!\!\int |y|^p\,dQ(y,z)\Bigr)^{1/p}\Bigl(\int\!\!\int |z|^q\,dQ(y,z)\Bigr)^{1/q},$$

by Hölder's inequality. It suffices to show that the first and second marginals of $|Q|$ are bounded above by the measures $2\phi(h)P^Y$ and $2\tilde{\phi}(h)P^Z$, respectively. By symmetry it suffices to consider the first marginal.

By definition we have that

$$|Q|(C) = \sup_D\Bigl(\bigl|Q(C\cap D)\bigr| + \bigl|Q(C\cap D^c)\bigr|\Bigr)$$

for the supremum taken over all Borel sets D in $\mathbb{R}^2$. Equivalently, we can compute the supremum over any algebra that generates the Borel sets. In particular, we can use the algebra consisting of all finite unions of rectangles $A\times B$. Conclude from this that

$$|Q|(C) = \sup\sum_i\sum_j\bigl|Q(C\cap(A_i\times B_j)\bigr|,$$

for the supremum taken over all pairs of partitions $\mathbb{R} = \cup_i A_i$ and $\mathbb{R} = \cup_j B_j$. It follows that

$$\begin{aligned}|Q|(A \times \mathbb{R}) &= \sup \sum_i \sum_j \bigl|Q\bigl((A \cap A_i) \times B_j\bigr)\bigr| \\ &= \sup \sum_i \sum_j \bigl|P^{Z|Y}(B_j |\, A \cap A_i) - P^Z(B_j)\bigr| P^Y(A \cap A_i).\end{aligned}$$

If, for fixed i, B_i^+ consists of the union of all B_j such that $P^{Z|Y}(B_j |\, A \cap A_i) - P^Z(B_j) > 0$ and B_i^- is the union of the remaining B_j, then the double sum can be rewritten

$$\sum_i \Bigl(\bigl|P^{Z|Y}(B_i^+ |\, A \cap A_i) - P^Z(B_i^+)\bigr| + \bigl|P^{Z|Y}(B_i^- |\, A \cap A_i) - P^Z(B_i^-)\bigr|\Bigr) P^Y(A \cap A_i).$$

The sum between round brackets is bounded above by $2\phi(h)$, by the definition of ϕ. Thus the display is bounded above by $2\phi(h)P^Y(A)$. ■

4.14 Theorem. *If X_t is a strictly stationary time series with mean zero such that $\mathrm{E}|X_t|^{p \vee q} < \infty$ and $\sum_h \phi(h)^{1/p} \tilde{\phi}(h)^{1/q} < \infty$ for some $p, q > 0$ with $p^{-1} + q^{-1} = 1$, then the series $v = \sum_h \gamma_X(h)$ converges absolutely and $\sqrt{n}\overline{X}_n \rightsquigarrow N(0, v)$.*

Proof. For a given $M > 0$ let $X_t^M = X_t 1\{|X_t| \le M\}$ and let $Y_t^M = X_t - X_t^M$. Because X_t^M is a measurable transformation of X_t, it is immediate from the definition of the mixing coefficients that Y_t^M is mixing with smaller mixing coefficients than X_t. Therefore, by (4.2) and Lemma 4.13,

$$\operatorname{var} \sqrt{n}(\overline{X}_n - \overline{X_n^M}) \le 2 \sum_h \phi(h)^{1/p} \tilde{\phi}(h)^{1/q} \|Y_0^M\|_p \|Y_0^M\|_q.$$

As $M \to \infty$, the right side converges to zero, and hence the left side converges to zero, uniformly in n. This means that we can reduce the problem to the case of uniformly bounded time series X_t, as in the proof of Theorem 4.7.

Because the α-mixing coefficients are bounded above by the ϕ-mixing coefficients, we have that $\sum_h \alpha(h) < \infty$. Therefore, the second part of the proof of Theorem 4.7 applies without changes. ■

4.5 Law of Large Numbers

The law of large numbers is concerned with the convergence of the sequence $\overline{X}_n$ rather than the sequence $\sqrt{n}(\overline{X}_n - \mu)$. By Slutsky's lemma $\overline{X}_n \to \mu$ in probability if the sequence $\sqrt{n}(\overline{X}_n - \mu)$ is uniformly tight. Thus a central limit theorem implies a weak law of large numbers. However, the latter is valid under much weaker conditions. The weakening not only concerns moments, but also the dependence between the X_t.

The strong law of large numbers for a strictly stationary time series is the central result in *ergodic theory*. In this section we discuss the main facts and some examples. For a nonstationary sequence or a triangular array an alternative is to apply mixing conditions. For the weak law for second order stationary time series also see Example 6.30.

4.5.1 Ergodic Theorem

Given a strictly stationary sequence X_t defined on some probability space $(\Omega, \mathcal{U}, \mathrm{P})$, with values in some measurable space $(\mathcal{X}, \mathcal{A})$ the *invariant σ-field*, denoted $\mathcal{U}_{\text{inv}}$, is the σ-field consisting of all sets A such that $A = (\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)^{-1}(B)$ for all t and some measurable set $B \subset \mathcal{X}^\infty$. Here throughout this section the product space $\mathcal{X}^\infty$ is equipped with the product σ-field $\mathcal{A}^\infty$.

Our notation in the definition of the invariant σ-field is awkward, if not unclear, because we look at two-sided infinite series. The triple X_{t-1}, X_t, X_{t+1} in the definition of A is meant to be centered at a fixed position in $\mathbb{Z}$. We can write this down more precisely using the forward shift function $S\colon \mathcal{X}^\infty \to \mathcal{X}^\infty$ defined by $S(x)_i = x_{i+1}$. The two-sided sequence $(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$ defines a map $X\colon \Omega \to \mathcal{X}^\infty$. With this notation the invariant sets A are the sets such that $A = \{S^t X \in B\}$ for all t and some measurable set $B \subset \mathcal{X}^\infty$. The strict stationarity of the sequence X is identical to the invariance of its induced law P^X on $\mathcal{X}^\infty$ under the shift S.

The inverse images $X^{-1}(B)$ of measurable sets $B \subset \mathcal{X}^\infty$ with $B = SB$ are clearly invariant. Conversely, it can be shown that, up to null sets, all invariant sets take this form. (See Exercise 4.16.) The *symmetric events* are special examples of invariant sets. They are the events that depend symmetrically on the variables X_t. For instance, $\cap_t X_t^{-1}(B)$ for some measurable set $B \subset \mathcal{X}$.

* **4.15** EXERCISE. Call a set $B \subset \mathcal{X}^\infty$ invariant under the shift $S\colon \mathcal{X}^\infty \to \mathcal{X}^\infty$ if $B = SB$. Call it *almost invariant* relative to a measure P^X if $P^X(B \triangle SB) = 0$. Show that a set B is almost invariant if and only if there exists an invariant set $\tilde{B}$ such that $P^X(B \triangle \tilde{B}) = 0$. [Try $\tilde{B} = \cap_t S^t B$.]

* **4.16** EXERCISE. Define the invariant σ-field $\mathcal{B}_{\text{inv}}$ on $\mathcal{X}^\infty$ as the collection of measurable sets that are invariant under the shift operation, and let $\overline{\mathcal{B}}_{\text{inv}}$ be its completion under the measure P^X. Show that $X^{-1}(\mathcal{B}_{\text{inv}}) \subset \mathcal{U}_{\text{inv}} \subset \overline{X^{-1}(\overline{\mathcal{B}}_{\text{inv}})}$, where the long bar on the right denotes completion relative to P. [Note that $\{X \in B\} = \{X \in SB\}$ implies that $P^X(B \triangle SB) = 0$. Use the preceding exercise to replace B by an invariant set $\tilde{B}$.]

4.17 Theorem (Birkhoff). *If X_t is a strictly stationary time series with $\mathrm{E}|X_t| < \infty$, then $\overline{X}_n \to \mathrm{E}(X_0 | \mathcal{U}_{\text{inv}})$ almost surely and in mean.*

Proof. For a given $\alpha \in \mathbb{R}$ define a set $B = \bigl\{x \in \mathbb{R}^\infty\colon \limsup_{n\to\infty} \overline{x}_n > \alpha\bigr\}$. Because

$$\overline{x}_{n+1} = \frac{x_1}{n+1} + \frac{n}{n+1}\frac{1}{n}\sum_{t=2}^{n+1} x_t,$$

a point x is contained in B if and only if $\limsup n^{-1}\sum_{t=2}^{n+1} x_t > \alpha$. Equivalently, $x \in B$ if and only if $Sx \in B$. Thus the set B is invariant under the shift operation $S\colon \mathbb{R}^\infty \to \mathbb{R}^\infty$. We conclude from this that the variable $\limsup_{n\to\infty} \overline{X}_n$ is measurable relative to the invariant σ-field.

Fix some measurable set $B \subset \mathbb{R}^\infty$. For every invariant set $A \in \mathcal{U}_{\mathrm{inv}}$ there exists a measurable set $C \subset \mathbb{R}^\infty$ such that $A = \{S^t X \in C\}$ for every t. By the strict stationarity of X,

$$\mathrm{P}\bigl(\{S^t X \in B\} \cap A\bigr) = \mathrm{P}(S^t X \in B, S^t X \in C) = \mathrm{P}(X \in B, X \in C) = \mathrm{P}\bigl(\{X \in B\} \cap A\bigr).$$

This shows that $\mathrm{P}(S^t X \in B|\,\mathcal{U}_{\mathrm{inv}}) = \mathrm{P}(X \in B|\,\mathcal{U}_{\mathrm{inv}})$ almost surely. We conclude that the conditional laws of $S^t X$ and X given the invariant σ-field are identical.

In particular, the conditional means $\mathrm{E}(X_t|\,\mathcal{U}_{\mathrm{inv}}) = \mathrm{E}(X_1|\,\mathcal{U}_{\mathrm{inv}})$ are identical for every t, almost surely. It also follows that a time series Z_t of the type $Z_t = (X_t, R)$ for $R\colon \Omega \to \mathcal{R}$ a fixed $\mathcal{U}_{\mathrm{inv}}$-measurable variable (for instance with values in $\mathcal{R} = \mathbb{R}^2$) is strictly stationary, the conditional law $B \mapsto \mathrm{P}(X \in B|\, R) = \mathrm{E}\bigl(\mathrm{P}(X \in B|\,\mathcal{U}_{\mathrm{inv}})|\, R\bigr)$ of its first marginal (on $\mathcal{X}^\infty$) being strictly stationary by the preceding paragraph, and the second marginal (on $\mathcal{R}^\infty$) being independent of t.

For the almost sure convergence of the sequence $\overline{X}_n$ it suffices to show that, for every $\varepsilon > 0$, the event

$$A = \Bigl\{\limsup_{n\to\infty} \overline{X}_n > \mathrm{E}(X_1|\,\mathcal{U}_{\mathrm{inv}}) + \varepsilon\Bigr\}$$

and a corresponding event for the lower tail have probably zero. By the preceding the event A is contained in the invariant σ-field. Furthermore, the time series $Y_t = \bigl(X_t - \mathrm{E}(X_1|\,\mathcal{U}_{\mathrm{inv}}) - \varepsilon\bigr)1_A$, being a fixed transformation of the time series $Z_t = \bigl(X_t, \mathrm{E}(X_1|\,\mathcal{U}_{\mathrm{inv}}), 1_A\bigr)$, is strictly stationary. We can write $A = \cup_n A_n$ for $A_n = \cup_{t=1}^n \{\overline{Y}_t > 0\}$. Then $\mathrm{E}Y_1 1_{A_n} \to \mathrm{E}Y_1 1_A$ by the dominated convergence theorem, in view of the assumption that X_t is integrable. If we can show that $\mathrm{E}Y_1 1_{A_n} \geq 0$ for every n, then we can conclude that

$$0 \leq \mathrm{E}Y_1 1_A = \mathrm{E}\bigl(X_1 - \mathrm{E}(X_1|\,\mathcal{U}_{\mathrm{inv}})\bigr)1_A - \varepsilon\mathrm{P}(A) = -\varepsilon\mathrm{P}(A),$$

because $A \in \mathcal{U}_{\mathrm{inv}}$. This implies that $\mathrm{P}(A) = 0$, concluding the proof of almost sure convergence.

The L_1-convergence can next be proved by a truncation argument. We can first show, more generally, but by an identical argument, that $n^{-1}\sum_{t=1}^n f(X_t) \to \mathrm{E}\bigl(f(X_0)|\,\mathcal{U}_{\mathrm{inv}}\bigr)$ almost surely, for every measurable function $f\colon \mathcal{X} \to \mathbb{R}$ with $\mathrm{E}|f(X_t)| < \infty$. We can apply this to the functions $f(x) = x1_{|x|\leq M}$ for given M.

We complete the proof by showing that $\mathrm{E}Y_1 1_{A_n} \geq 0$ for every strictly stationary time series Y_t and every fixed n, and $A_n = \cup_{t=1}^n \{\overline{Y}_t > 0\}$. For every $2 \leq j \leq n$,

$$Y_1 + \cdots + Y_j \leq Y_1 + \max(Y_2, Y_2 + Y_3, \cdots, Y_2 + \cdots + Y_{n+1}).$$

If we add the number 0 in the maximum on the right, then this is also true for $j = 1$. We can rewrite the resulting n inequalities as the single inequality

$$Y_1 \geq \max(Y_1, Y_1 + Y_2, \ldots, Y_1 + \cdots + Y_n) - \max(0, Y_2, Y_2 + Y_3, \cdots, Y_2 + \cdots + Y_{n+1}).$$

The event A_n is precisely the event that the first of the two maxima on the right is positive. Thus on this event the inequality remains true if we add also a zero to the first maximum. It follows that $\mathrm{E}Y_1 1_{A_n}$ is bounded below by

$$\mathrm{E}\Big(\max(0, Y_1, Y_1+Y_2, \ldots, Y_1+\cdots+Y_n) - \max(0, Y_2, Y_2+Y_3, \cdots, Y_2+\cdots+Y_{n+1})\Big)1_{A_n}.$$

Off the event A_n the first maximum is zero, whereas the second maximum is always nonnegative. Thus the expression does not increase if we cancel the indicator 1_{A_n}. The resulting expression is identically zero by the strict stationarity of the series Y_t. ■

Thus a strong law is valid for *every* integrable strictly stationary sequence, without any further conditions on possible dependence of the variables. However, the limit $\mathrm{E}(X_0|\mathcal{U}_{\mathrm{inv}})$ in the preceding theorem will often be a true random variable. Only if the invariant σ-field is trivial, we can be sure that the limit is degenerate. Here "trivial" may be taken to mean that the invariant σ-field consists of sets of probability 0 or 1 only. If this is the case, then the time series X_t is called *ergodic*.

* **4.18** EXERCISE. Suppose that X_t is strictly stationary. Show that X_t is ergodic if and only if every sequence $Y_t = f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$ for a measurable map f that is integrable satisfies the law of large numbers $\overline{Y}_n \to \mathrm{E}Y_1$ almost surely. [Given an invariant set $A = (\ldots, X_{-1}, X_0, X_1, \ldots)^{-1}(B)$ consider $Y_t = 1_B(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$. Then $\overline{Y}_n = 1_A$.]

Checking that the invariant σ-field is trivial may be a nontrivial operation. There are other concepts that imply ergodicity and may be easier to verify. A time series X_t is called *mixing* if, for any measurable sets A and B, as $h \to \infty$,

$$\begin{aligned}&\mathrm{P}\big((\ldots, X_{h-1}, X_h, X_{h+1}, \ldots) \in A, (\ldots, X_{-1}, X_0, X_1, \ldots) \in B\big)\\ &\qquad\qquad \to \mathrm{P}\big((\ldots, X_{h-1}, X_h, X_{h+1}, \ldots) \in A\big)\mathrm{P}\big((\ldots, X_{-1}, X_0, X_1, \ldots) \in B\big).\end{aligned}$$

Every mixing time series is ergodic. This follows because if we take $A = B$ equal to an invariant set, the preceding display reads $P^X(A) \to P^X(A)P^X(A)$, for P^X the law of the infinite series X_t, and hence $P^X(A)$ is 0 or 1.

The present type of mixing is related to the mixing concepts used to obtain central limit theorems, and is weaker.

4.19 Theorem. *Any strictly stationary α-mixing time series is mixing.*

Proof. For t-dimensional cylinder sets A and B in $\mathcal{X}^\infty$ (i.e. sets that depend on finitely many coordinates only) the mixing condition becomes

$$\mathrm{P}\big((X_h, \ldots X_{t+h}) \in A, (X_0, \ldots, X_t) \in B\big) \to \mathrm{P}\big((X_h, \ldots X_{t+h}) \in A\big)\mathrm{P}\big((X_0, \ldots, X_t) \in B\big).$$

For $h > t$ the absolute value of the difference of the two sides of the display is bounded by $\alpha(h-t)$ and hence converges to zero as $h \to \infty$, for each fixed t.

Thus the mixing condition is satisfied by the collection of all cylinder sets. This collection is intersection-stable, i.e. a π-system, and generates the product σ-field on $\mathcal{X}^\infty$. The proof is complete if we can show that the collections of sets A and B for which the mixing condition holds, for a given set B or A, is a σ-field. By the π-λ theorem it suffices to show that these collections of sets are a λ-system.

The mixing property can be written as $P^X(S^{-h}A\cap B)-P^X(A)P^X(B)\to 0$, as $h\to\infty$. Because S is a bijection we have $S^{-h}(A_2-A_1)=S^{-h}A_2-S^{-h}A_1$. If $A_1\subset A_2$, then

$$\begin{aligned}P^X\bigl(S^{-h}(A_2-A_1)\cap B\bigr)&=P^X\bigl(S^{-h}A_2\cap B\bigr)-P^X\bigl(S^{-h}A_1\cap B\bigr),\\ P^X(A_2-A_1)P^X(B)&=P^X(A_2)P^X(B)-P^X(A_1)P^X(B).\end{aligned}$$

If, for a given set B, the sets A_1 and A_2 satisfy the mixing condition, then the right hand sides are asymptotically the same, as $h\to\infty$, and hence so are the left sides. Thus A_2-A_1 satisfies the mixing condition. If $A_n\uparrow A$, then $S^{-h}A_n\uparrow S^{-h}A$ as $n\to\infty$ and hence

$$P^X(S^{-h}A_n\cap B)-P^X(A_n)P^X(B)\to P^X(S^{-h}A\cap B)-P^X(A)P^X(B).$$

The absolute difference of left and right sides is bounded above by $2|P^X(A_n)-P^X(A)|$. Hence the convergence in the display is uniform in h. If every of the sets A_n satisfies the mixing condition, for a given set B, then so does A. Thus the collection of all sets A that satisfies the condition, for a given B, is a λ-system.

We can prove similarly, but more easily, that the collection of all sets B is also a λ-system. ■

4.20 Theorem. *Any strictly stationary time series X_t with trivial tail σ-field is mixing.*

Proof. The tail σ-field is defined as $\cap_{h\in\mathbb{Z}}\sigma(X_h,X_{h+1},\ldots)$.

As in the proof of the preceding theorem we need to verify the mixing condition only for finite cylinder sets A and B. We can write

$$\begin{aligned}&\Bigl|\mathrm{E}1_{X_h,\ldots,X_{t+h}\in A}\bigl(1_{X_0,\ldots,X_t\in B}-\mathrm{P}(X_0,\ldots,X_t\in B)\bigr)\Bigr|\\ &\quad=\Bigl|\mathrm{E}1_{X_h,\ldots,X_{t+h}\in A}\bigl(\mathrm{P}(X_0,\ldots,X_t\in B|\,X_h,X_{h+1},\ldots)-\mathrm{P}(X_0,\ldots,X_t\in B)\bigr)\Bigr|\\ &\quad\le \mathrm{E}\Bigl|\mathrm{P}(X_0,\ldots,X_t\in B|\,X_h,X_{h+1},\ldots)-\mathrm{P}(X_0,\ldots,X_t\in B)\bigr)\Bigr|.\end{aligned}$$

For every integrable variable Y the sequence $\mathrm{E}(Y|\,X_h,X_{h+1},\ldots)$ converges in L_1 to the conditional expectation of Y given the tail σ-field, as $h\to\infty$. Because the tail σ-field is trivial, in the present case this is $\mathrm{E}Y$. Thus the right side of the preceding display converges to zero as $h\to\infty$. ■

* **4.21** EXERCISE. Show that a strictly stationary time series X_t is ergodic if and only if $n^{-1}\sum_{h=1}^n P^X(S^{-h}A\cap B)\to P^X(A)P^X(B)$, as $n\to\infty$, for every measurable subsets A and B of $\mathcal{X}^\infty$. [Use the ergodic theorem on the stationary time series $Y_t=1_{S^tX\in A}$ to see that $n^{-1}\sum 1_{X\in S^{-t}A}1_B\to P^X(A)1_B$ for the proof in one direction.]

* **4.22** EXERCISE. Show that a strictly stationary time series X_t is ergodic if and only if the one-sided time series $X_0, X_1, X_2, \ldots$ is ergodic, in the sense that the "one-sided invariant σ-field", consisting of all sets A such that $A = (X_t, X_{t+1}, \ldots)^{-1}(B)$ for some measurable set B and every $t \geq 0$, is trivial. [Use the preceding exercise.]

The preceding theorems can be used as starting points to construct ergodic sequences. For instance, every i.i.d. sequence is ergodic by the preceding theorems, because its tail σ-field is trivial by Kolmogorov's 0-1 law, or because it is α-mixing. To construct more examples we can combine the theorems with the following stability property. From a given ergodic sequence X_t we construct a process Y_t by transforming the vector $(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$ with a given map f from the product space $\mathcal{X}^\infty$ into some measurable space $(\mathcal{Y}, \mathcal{B})$. As before, the X_t in $(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$ is meant to be at a fixed 0th position in $\mathbb{Z}$, so that the different variables Y_t are obtained by sliding the function f along the sequence $(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$.

4.23 Lemma. *The sequence $Y_t = f(\ldots, X_{t-1}, X_t, X_{t+1}, \ldots)$ obtained by application of a measurable map $f: \mathcal{X}^\infty \to \mathcal{Y}$ to an ergodic sequence X_t is ergodic.*

Proof. Define $\overline{f}: \mathcal{X}^\infty \to \mathcal{Y}^\infty$ by $\overline{f}(x) = \bigl(\cdots, f(S^{-1}x), f(x), f(Sx), \cdots\bigr)$, for S the forward shift on $\mathcal{X}^\infty$. Then $Y = \overline{f}(X)$ and $S'\overline{f}(x) = \overline{f}(Sx)$ if S' is the forward shift on $\mathcal{Y}^\infty$. If $A = \{(S')^t Y \in B\}$ is invariant for the series Y_t, then $A = \{\overline{f}(S^t X) \in B\} = \{S^t X \in \overline{f}^{-1}(B)\}$ for every t, and hence A is also invariant for the series X_t. ∎

4.24 EXERCISE. Let Z_t be an i.i.d. sequence of integrable variables and let $X_t = \sum_j \psi_j Z_{t-j}$ for a sequence ψ_j such that $\sum_j |\psi_j| < \infty$. Show that X_t satisfies the law of large numbers (with degenerate limit).

4.25 EXERCISE. Show that the GARCH$(1,1)$ process defined in Example 1.10 is ergodic.

4.26 Example. Every stationary irreducible Markov chain on a countable state space is ergodic. Conversely, a stationary reducible Markov chain on a countable state space whose initial (or marginal) law is positive everywhere is nonergodic.

To prove the ergodicity note that a stationary irreducible Markov chain is (positively) recurrent (e.g. Durrett, p266). If A is an invariant set of the form $A = (X_0, X_1, \ldots)^{-1}(B)$, then $A \in \sigma(X_h, X_{h-1}, \ldots)$ for all h and hence

$$\begin{aligned}1_A = \mathrm{P}(A|\, X_h, X_{h-1}, \ldots) &= \mathrm{P}\bigl((X_{h+1}, X_{h+2}, \ldots) \in B|\, X_h, X_{h-1}, \ldots\bigr)\\ &= \mathrm{P}\bigl((X_{h+1}, X_{h+2}, \ldots) \in B|\, X_h\bigr).\end{aligned}$$

We can write the right side as $g(X_h)$ for the function $g(x) = \mathrm{P}\bigl(A|\, X_{-1} = x\bigr)$. By recurrence, for almost every ω in the underlying probability space, the right side runs infinitely often through every of the numbers $g(x)$ with x in the state space. Because the left side is 0 or 1 for a fixed ω, the function g and hence 1_A must be constant. Thus every invariant set of this type is trivial, showing the ergodicity of the one-sided sequence

$X_0, X_1, \ldots$. It can be shown that one-sided and two-sided ergodicity are the same. (Cf. Exercise 4.22.)

Conversely, if the Markov chain is reducible, then the state space can be split into two sets $\mathcal{X}_1$ and $\mathcal{X}_2$ such that the chain will remain in $\mathcal{X}_1$ or $\mathcal{X}_2$ once it enters there. If the initial distribution puts positive mass everywhere, then each of the two possibilities occurs with positive probability. The sets $A_i = \{X_0 \in \mathcal{X}_i\}$ are then invariant and nontrivial and hence the chain is not ergodic.

It can also be shown that a stationary irreducible Markov chain is mixing if and only if it is aperiodic. (See e.g. Durrett, p310.) Furthermore, the tail σ-field of any irreducible stationary aperiodic Markov chain is trivial. (See e.g. Durrett, p279.) □

Ergodicity is a powerful, but somewhat complicated concept. If we are only interested in a law of large numbers for a given sequence, then it may be advantageous to use more elementary tools. For instance, the means $\overline{X}_n$ of any stationary time series X_t converge in L_2 to a random variable; this limit is degenerate if and only if the spectral mass of the series X_t at zero is zero. See Example 6.30.

4.5.2 Mixing

In the preceding section it was seen that an α-mixing, strictly stationary time series is ergodic and hence satisfies the law of large numbers if it is integrable. In this section we extend the law of large numbers to possibly nonstationary α-mixing time series.

The key is the bound on the tails of the distribution of the sample mean given in the following lemma.

4.27 Lemma. *For any mean zero time series X_t with α-mixing numbers $\alpha(h)$, every $x > 0$ and every $h \in \mathbb{N}$, with $Q_t = F^{-1}_{|X_t|}$,*

$$\mathrm{P}(\overline{X}_n \geq 2x) \leq \frac{2}{nx^2}\int_0^1 \bigl(\alpha^{-1}(u) \wedge h\bigr)\,\frac{1}{n}\sum_{t=1}^n Q_t^2(1-u)\,du + \frac{2}{x}\int_0^{\alpha(h)} \frac{1}{n}\sum_{t=1}^n Q_t(1-u)\,du.$$

Proof. The quantile function of the variable $|X_t|/(xn)$ is equal to $u \mapsto F^{-1}_{|X_t|}(u)/(nx)$. Therefore, by a rescaling argument we can see that it suffices to bound the probability $\mathrm{P}\bigl(\sum_{t=1}^n X_t \geq 2\bigr)$ by the right side of the lemma, but with the factors $2/(nx^2)$ and $2/x$ replaced by 2 and the factor n^{-1} in front of $\sum Q_t^2$ and $\sum Q_t$ dropped. For ease of notation set $S_0 = 0$ and $S_n = \sum_{t=1}^n X_t$.

Define the function $g\colon \mathbb{R} \to \mathbb{R}$ to be 0 on the interval $(-\infty, 0]$, to be $x \mapsto \frac{1}{2}x^2$ on $[0,1]$, to be $x \mapsto 1 - \frac{1}{2}(x-2)^2$ on $[1,2]$, and to be 1 on $[2,\infty)$. Then g is continuously differentiable with uniformly Lipschitz derivative. By Taylor's theorem it follows that $\bigl|g(x) - g(y) - g'(x)(x-y)\bigr| \leq \frac{1}{2}|x-y|^2$ for every $x, y \in \mathbb{R}$. Because $1_{[2,\infty)} \leq g$ and $S_t - S_{t-1} = X_t$,

$$\mathrm{P}(S_n \geq 2) \leq \mathrm{E}g(S_n) = \sum_{t=1}^n \mathrm{E}\bigl(g(S_t) - g(S_{t-1})\bigr) \leq \sum_{t=1}^n \mathrm{E}\bigl|g'(S_{t-1})X_t\bigr| + \tfrac{1}{2}\sum_{t=1}^n \mathrm{E}X_t^2.$$

The last term on the right can be written $\frac{1}{2}\sum_{t=1}^n\int_0^1 Q_t^2(1-u)\,du$, which is bounded by $\sum_{t=1}^n\int_0^{\alpha(0)} Q_t^2(1-u)\,du$, because $\alpha(0)=\frac{1}{2}$ and $u\mapsto Q_t(1-u)$ is decreasing.

For $i\geq 1$ the variable $g'(S_{t-i})-g'(S_{t-i-1})$ is measurable relative to $\sigma(X_s\colon s\leq t-i)$ and is bounded in absolute value by $|X_{t-i}|$. Therefore, Lemma 4.11 yields the inequality

$$\bigl|\mathrm{E}\bigl(g'(S_{t-i})-g'(S_{t-i-1})\bigr)X_t\bigr|\leq 2\int_0^{\alpha(i)} Q_{t-i}(1-u)Q_t(1-u)\,du.$$

For $t\leq h$ we can write $g'(S_{t-1})=\sum_{i=1}^{t-1}\bigl(g'(S_{t-i})-g(S_{t-i-1})\bigr)$. Substituting this in the left side of the following display and applying the preceding display, we find that

$$\sum_{t=1}^h \mathrm{E}\bigl|g'(S_{t-1})X_t\bigr|\leq 2\sum_{t=1}^h\sum_{i=1}^{t-1}\int_0^{\alpha(i)} Q_{t-i}(1-u)Q_t(1-u)\,du.$$

For $t>h$ we can write $g'(S_{t-1})=g'(S_{t-h})+\sum_{i=1}^{h-1}\bigl(g'(S_{t-i})-g(S_{t-i-1})\bigr)$. By a similar argument, this time also using that the function $|g'|$ is uniformly bounded by 1, we find

$$\sum_{t=h+1}^n \mathrm{E}\bigl|g'(S_{t-1})X_t\bigr|\leq 2\int_0^{\alpha(h)} Q_t(1-u)\,du+2\sum_{t=h+1}^n\sum_{i=1}^{h-1}\int_0^{\alpha(i)} Q_{t-i}(1-u)Q_t(1-u)\,du.$$

Combining the preceding displays we obtain that $\mathrm{P}(S_n\geq 2)$ is bounded above by

$$2\int_0^{\alpha(h)} Q_t(1-u)\,du+2\sum_{t=1}^n\sum_{i=1}^{t\wedge h-1}\int_0^{\alpha(i)} Q_{t-i}(1-u)Q_t(1-u)\,du+\tfrac{1}{2}\sum_{t=1}^n \mathrm{E}X_t^2.$$

In the second term we can bound $2Q_{t-i}(1-u)Q_t(1-u)$ by $Q_{t-i}^2(1-u)+Q_t^2(1-u)$ and next change the order of summation to $\sum_{i=1}^{h-1}\sum_{t=i+1}^n$. Because $\sum_{t=i+1}^n(Q_{t-i}^2+Q_t^2)\leq 2\sum_{t=1}^n Q_t^2$ this term is bounded by $2\sum_{i=1}^{h-1}\int_0^{\alpha(i)}\sum_{t=1}^n Q_t^2(1-u)\,du$. Together with the third term on the right this gives rise to by the first sum on the right of the lemma, as $\sum_{i=0}^{h-1}1_{u\leq\alpha(i)}=\alpha^{-1}(u)\wedge h$. ∎

4.28 Theorem. *For each n let the time series $(X_{n,t}\colon t\in\mathbb{Z})$ be mixing with mixing coefficients $\alpha_n(h)$. If $\sup_n\alpha_n(h)\to 0$ as $h\to\infty$ and $(X_{n,t}\colon t\in\mathbb{Z}, n\in\mathbb{N})$ is uniformly integrable, then the sequence $\overline{X}_n-\mathrm{E}\overline{X}_n$ converges to zero in probability.*

Proof. By the assumption of uniform integrability $n^{-1}\sum_{t=1}^n\mathrm{E}|X_{n,t}|1_{|X_{n,t}|>M}\to 0$ as $M\to\infty$ uniformly in n. Therefore we can assume without loss of generality that $X_{n,t}$ is bounded in absolute value by a constant M. We can also assume that it is centered at mean zero.

Then the quantile function of $|X_{n,t}|$ is bounded by M and the preceding lemma yields the bound

$$\mathrm{P}(|\overline{X}_n|\geq 2\varepsilon)\leq\frac{4hM^2}{n\varepsilon^2}+\frac{4M}{x}\sup_n\alpha_n(h).$$

This converges to zero as $n\to\infty$ followed by $h\to\infty$. ∎

* **4.29** EXERCISE. Relax the conditions in the preceding theorem to, for every $\varepsilon > 0$:

$$n^{-1}\sum_{t=1}^{n} \mathrm{E}|X_{n,t}|1_{|X_{n,t}|>\varepsilon n \wedge F^{-1}_{|X_{n,t}|}(1-\alpha_n(h))} \to 0.$$

[Hint: truncate at the level $n\varepsilon_n$ and note that $\mathrm{E}X1_{X>M} = \int_0^1 Q(1-u)1_{Q(1-u)>M}\,du$ for $Q(u) = F_X^{-1}(u)$.]

* 4.5.3 Subadditive Ergodic Theorem

The subadditive theorem of Kingman can be considered an extension of the ergodic theorem, which gives the almost sure convergence of more general functions of a strictly stationary sequence than the consecutive means. Given a strictly stationary time series X_t with values in some measurable space $(\mathcal{X}, \mathcal{A})$ and defined on some probability space $(\Omega, \mathcal{U}, \mathrm{P})$, write X for the induced map $(\ldots, X_{-1}, X_0, X_1, \ldots)\colon \Omega \to \mathcal{X}^\infty$, and let $S\colon \mathcal{X}^\infty \to \mathcal{X}^\infty$ be the forward shift function, all as before. A family $(T_n\colon n \in \mathbb{N})$ of maps $T_n\colon \mathcal{X}^\infty \to \mathbb{R}$ is called *subadditive* if, for every $m, n \in \mathbb{N}$,

$$T_{m+n}(X) \le T_m(X) + T_n(S^m X).$$

4.30 Theorem (Kingman). *If X is strictly stationary with invariant σ-field $\mathcal{U}_{\mathrm{inv}}$ and the maps $(T_n\colon n \in \mathbb{N})$ are subadditive with finite means $\mathrm{E}T_n(X)$, then $T_n(X)/n \to \gamma\colon= \inf_n n^{-1}\mathrm{E}\bigl(T_n(X)|\,\mathcal{U}_{\mathrm{inv}}\bigr)$ almost surely. Furthermore, the limit γ satisfies $\mathrm{E}\gamma > -\infty$ if and only if $\inf_n \mathrm{E}T_n(X)/n > -\infty$ and in that case the convergence $T_n(X) \to \gamma$ takes also place in mean.*

Because the maps $T_n(X) = \sum_{t=1}^n X_t$ are subadditive, the "ordinary" ergodic theorem by Birkhoff is a special case of Kingman's theorem. If the time series X_t is ergodic, then the limit γ in Kingman's theorem is equal to $\gamma = \inf_n n^{-1}\mathrm{E}T_n(X)$.

4.31 EXERCISE. Show that the normalized means $n^{-1}\mathrm{E}T_n(X)$ of a subadditive map are decreasing in n.

4.32 EXERCISE. Let X_t be a time series with values in the collection of $(d \times d)$ matrices. Show that $T_n(X) = \log \|X_{-1}\cdots X_{-n}\|$ defines subadditive maps.

4.33 EXERCISE. Show that Kingman's theorem remains true if the forward shift operator in the definition of subadditivity is replaced by the backward shift operator.

4.6 Martingale Differences

The partial sums $\sum_{t=1}^n X_t$ of an i.i.d. sequence grow by increments X_t that are independent from the "past". The classical central limit theorem shows that this induces asymptotic normality provided the increments are centered and not too big (finite variance suffices). The mixing central limit theorem relax the independence to near independence of variables at large time lags, which are conditions involving the whole distribution. The martingale central limit theorem given in this section imposes conditions on the conditional first and second moments of the increments given the past, without directly involving other aspects of the distribution. The first moments given the past are assumed zero; the second moments given the past must not be too big.

A *filtration* $\mathcal{F}_t$ is a nondecreasing collection of σ-fields $\cdots \subset \mathcal{F}_{-1} \subset \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$. The σ-field $\mathcal{F}_t$ is to be thought of as the "events that are known" at time t. Often it will be the σ-field generated by variables $X_t, X_{t-1}, X_{t-2}, \ldots$. The corresponding filtration is called the *natural filtration* of the time series X_t, or the filtration generated by this series. A *martingale difference series* relative to a given filtration is a time series X_t such that, for every t,

(i) X_t is $\mathcal{F}_t$-measurable;

(ii) $\mathrm{E}(X_t|\,\mathcal{F}_{t-1}) = 0$.

The second requirement implicitly includes the assumption that $\mathrm{E}|X_t| < \infty$, so that the conditional expectation is well defined; the identity is understood to be in the almost-sure sense.

4.34 EXERCISE. Show that a martingale difference series with finite variances is a white noise series.

4.35 Theorem. *Let X_t be a martingale difference series relative to the filtration $\mathcal{F}_t$ such that $n^{-1}\sum_{t=1}^n \mathrm{E}(X_t^2|\,\mathcal{F}_{t-1}) \xrightarrow{\mathrm{P}} v$ for a positive constant v, and such that $n^{-1}\sum_{t=1}^n \mathrm{E}\bigl(X_t^2 1\{|X_t| > \varepsilon\sqrt{n}\}|\,\mathcal{F}_{t-1}\bigr) \xrightarrow{\mathrm{P}} 0$ for every $\varepsilon > 0$. Then $\sqrt{n}\overline{X}_n \rightsquigarrow N(0, v)$.*

4.36 Corollary. *Let X_t be a strictly stationary, ergodic martingale difference series relative to its natural filtration with mean zero and $v = \mathrm{E}X_t^2 < \infty$. Show that $\sqrt{n}\overline{X}_n \rightsquigarrow N(0, v)$.*

Proof. By strict stationarity there exists a measurable function $g\colon \mathbb{R}^\infty \to \mathbb{R}^\infty$ such that $\mathrm{E}(X_t|\,X_{t-1}, X_{t-2}, \ldots) = g(X_{t-1}, X_{t-2}, \ldots)$ almost surely, for every t. The ergodicity of the series X_t is inherited by the series $Y_t = g(X_{t-1}, X_{t-2}, \ldots)$ and hence $\overline{Y}_n \to \mathrm{E}Y_1 = \mathrm{E}X_1^2$ almost surely. By a similar argument the averages $n^{-1}\sum_{t=1}^n \mathrm{E}\bigl(X_t^2 1_{|X_t|>M}|\,\mathcal{F}_{t-1}\bigr)$ converge almost surely to their expectation, for every fixed M. This expectation can be made arbitrarily small by choosing M large. The sequence $n^{-1}\sum_{t=1}^n \mathrm{E}\bigl(X_t^2 1\{|X_t| > \varepsilon\sqrt{n}\}|\,\mathcal{F}_{t-1}\bigr)$ is bounded by this sequence eventually, for any M, and hence converges almost surely to zero. ∎

* 4.7 Projections

Let X_t be a centered time series and $\mathcal{F}_0=\sigma(X_0,X_{-1},\ldots)$. For a suitably mixing time series the covariance $\mathrm{E}\bigl(X_n\mathrm{E}(X_j|\,\mathcal{F}_0)\bigr)$ between X_n and the best prediction of X_j at time 0 should be small as $n\to\infty$. The following theorem gives a precise and remarkably simple sufficient condition for the central limit theorem in terms of these quantities.

4.37 Theorem. *let X_t be a strictly stationary, mean zero, ergodic time series with $\sum_h\bigl|\gamma_X(h)\bigr|<\infty$ and, as $n\to\infty$,*

$$\sum_{j=0}^{\infty}\bigl|\mathrm{E}\bigl(X_n\mathrm{E}(X_j|\,\mathcal{F}_0)\bigr)\bigr|\to 0.$$

Then $\sqrt{n}\overline{X}_n\rightsquigarrow N(0,v)$, for $v=\sum_h\gamma_X(h)$.

Proof. For a fixed integer m define a time series

$$Y_{t,m}=\sum_{j=t}^{t+m}\bigl(\mathrm{E}(X_j|\,\mathcal{F}_t)-\mathrm{E}(X_j|\,\mathcal{F}_{t-1})\bigr).$$

Then $Y_{t,m}$ is a strictly stationary martingale difference series. By the ergodicity of the series X_t, for fixed m as $n\to\infty$,

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(Y_{t,m}^2|\,\mathcal{F}_{t-1})\to\mathrm{E}Y_{0,m}^2=:v_m,$$

almost surely and in mean. The number v_m is finite, because the series X_t is square-integrable by assumption. By the martingale central limit theorem, Theorem 4.35, we conclude that $\sqrt{n}\overline{Y}_{n,m}\rightsquigarrow N(0,v_m)$ as $n\to\infty$, for every fixed m.

Because $X_t=\mathrm{E}(X_t|\,\mathcal{F}_t)$ we can write

$$\begin{aligned}\sum_{t=1}^{n}(Y_{t,m}-X_t)&=\sum_{t=1}^{n}\sum_{j=t+1}^{t+m}\mathrm{E}(X_j|\,\mathcal{F}_t)-\sum_{t=1}^{n}\sum_{j=t}^{t+m}\mathrm{E}(X_j|\,\mathcal{F}_{t-1})\\&=\sum_{j=n+1}^{n+m}\mathrm{E}(X_j|\,\mathcal{F}_n)-\sum_{j=1}^{m}\mathrm{E}(X_j|\,\mathcal{F}_0)-\sum_{t=1}^{n}\mathrm{E}(X_{t+m}|\,\mathcal{F}_{t-1}).\end{aligned}$$

Write the right side as $Z_{n,m}-Z_{0,m}-R_{n,m}$. Then the time series $Z_{t,m}$ is stationary with

$$\mathrm{E}Z_{0,m}^2=\sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{E}\bigl(\mathrm{E}(X_i|\,\mathcal{F}_0)\mathrm{E}(X_j|\,\mathcal{F}_0)\bigr)\le m^2\mathrm{E}X_0^2.$$

The right side divided by n converges to zero as $n \to \infty$, for every fixed m. Furthermore,

$$\begin{aligned}\mathrm{E}R_{n,m}^2 &= \sum_{s=1}^n \sum_{t=1}^n \mathrm{E}\Big(\mathrm{E}(X_{s+m}|\,\mathcal{F}_{s-1})\mathrm{E}(X_{t+m}|\,\mathcal{F}_{t-1})\Big)\\ &\leq 2 \sum\sum_{1\leq s\leq t\leq n} \mathrm{E}\big(\mathrm{E}(X_{s+m}|\,\mathcal{F}_{s-1})X_{t+m}\big)\\ &\leq 2n \sum_{h=1}^\infty \big|\mathrm{E}\mathrm{E}(X_{m+1}|\,\mathcal{F}_0)X_{h+m}\big| = 2n \sum_{h=m+1}^\infty \big|\mathrm{E}X_{m+1}\mathrm{E}(X_h|\,\mathcal{F}_0)\big|.\end{aligned}$$

The right side divided by n converges to zero as $m \to \infty$. Combining the three preceding displays we see that the sequence $\sqrt{n}(\overline{Y}_{n,m} - \overline{X}_n) = (Z_{n,m} - Z_{0,m} - R_{n,m})/\sqrt{n}$ converges to zero in second mean as $n \to \infty$ followed by $m \to \infty$.

Because $Y_{t,m}$ is a martingale difference series, the variables $Y_{t,m}$ are uncorrelated and hence

$$\mathrm{var}\,\sqrt{n}\overline{Y}_{n,m} = \mathrm{E}Y_{0,m}^2 = v_m.$$

Because, as usual, $\mathrm{var}\,\sqrt{n}\overline{X}_n \to v$ as $n \to \infty$, combination with the preceding paragraph shows that $v_m \to v$ as $m \to \infty$. Consequently, by Lemma 3.10 there exists $m_n \to \infty$ such that $\sqrt{n}\overline{Y}_{n,m_n} \rightsquigarrow N(0,v)$ and $\sqrt{n}(\overline{Y}_{n,m_n} - \overline{X}_n) \rightsquigarrow 0$. This implies the theorem in view of Slutsky's lemma. ∎

4.38 Example. We can use the preceding theorem for an alternative proof of the α-mixing central limit theorem, Theorem 4.7. The absolute convergence of the series $\sum_h \gamma_X(h)$ can be verified under the condition of Theorem 4.7 as in the first lines of the proof of that theorem. We concentrate on the verification of the displayed condition of the preceding theorem. Set $Y_n = \mathrm{E}(X_n|\,\mathcal{F}_0)$ and

$$1 - U_n = F_{|Y_n|}\big(|Y_n|-\big) + V\Delta F_{|Y_n|}\big(|Y_n|\big),$$

where ΔF denotes the jump sizes of a cumulative distribution function and V is a uniform variable independent of the other variables. The latter definition is an extended form of the probability integral transformation, allowing for jumps in the distribution function. The variable U_n is uniformly distributed and $F_{|Y_n|}^{-1}(1 - U_n) = |Y_n|$ almost surely. Because Y_n is $\mathcal{F}_0$-measurable the covariance inequality, Lemma 4.11, gives

$$\begin{aligned}\big|\mathrm{E}\big(\mathrm{E}(X_n|\,\mathcal{F}_0)X_j\big)\big| &\leq 2\int_0^{\alpha_j} F_{|Y_n|}^{-1}(1-u)F_{|X_j|}^{-1}(1-u)\,du\\ &= 2\mathrm{E}Y_n\,\mathrm{sign}(Y_n)F_{|X_j|}^{-1}(1-U_n)1_{U_n<\alpha_j}\\ &= 2\mathrm{E}X_n\,\mathrm{sign}(Y_n)F_{|X_j|}^{-1}(1-U_n)1_{U_n<\alpha_j}\\ &\leq 2\mathrm{E}|X_n|F_{|X_j|}^{-1}(1-U_n)1_{U_n<\alpha_j}\\ &\leq 4\int_0^1 F_{|X_n|}^{-1}(1-u)G^{-1}(1-u)\,du\end{aligned}$$

by a second application of Lemma 4.11, with $\alpha = 1$ and G the distribution function of the random variable $F_{|X_j|}^{-1}(1 - U_n)1_{U_n<\alpha_j}$. The corresponding quantile function $G^{-1}(1 - u)$ vanishes off $[0, \alpha_j]$ and is bounded above by the quantile function of $|X_j|$. Therefore, the expression is further bounded by $\int_0^{\alpha_j} F_{|X_0|}^{-1}(1 - u)^2\, du$. We finish by summing up over j. $\square$

5
Nonparametric Estimation of Mean and Covariance

Suppose we observe the values $X_1, \ldots, X_n$ from the stationary time series X_t with mean $\mu_X = \mathrm{E}X_t$, covariance function $\gamma_X(h)$, and correlation function $\rho_X(h)$. If nothing is known about the distribution of the time series, besides that it is stationary, then "obvious" estimators for these parameters are

$$\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t,$$

$$\hat{\gamma}_n(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \overline{X}_n)(X_t - \overline{X}_n), \qquad (0 \leq h < n),$$

$$\hat{\rho}_n(h) = \frac{\hat{\gamma}_n(h)}{\hat{\gamma}_n(0)}.$$

These estimators are called *nonparametric*, because they are not motivated by a statistical model that restricts the distribution of the time series. The advantage is that they work for (almost) every stationary time series. However, given a statistical model, it might be possible to find better estimators for μ_X, γ_X and ρ_X. We shall see examples of this when discussing ARMA-processes in a later chapter.

* **5.1** EXERCISE. The factor $1/n$ in the definition of $\hat{\gamma}_n(h)$ is sometimes replaced by $1/(n-h)$, because there are $n-h$ terms in the sum. Show that with the present definition of $\hat{\gamma}_n$ the corresponding estimate $\bigl(\hat{\gamma}_n(s-t)\bigr)_{s,t=1,\ldots,h}$ for the covariance matrix of $(X_1, \ldots, X_h)$ is nonnegative-definite. Show by example that this is not true if we use $1/(n-h)$. [Write the matrix as QQ^T for a suitable $(n \times (2n))$ matrix Q.]

The time series X_t is called *Gaussian* if the joint distribution of any finite number of the variables X_t is multivariate-normal. In that case $\overline{X}_n$ is normally distributed. The distributions of $\hat{\gamma}_n(h)$ and $\hat{\rho}_n(h)$ are complicated, even under normality. Distributional statements considering these estimators are therefore usually asymptotic in nature, as

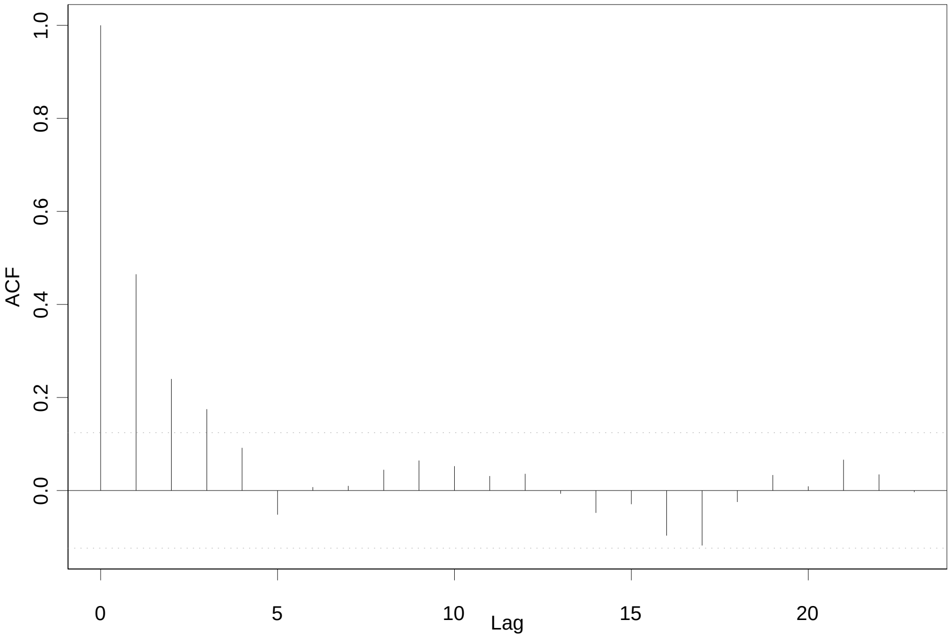

Figure 5.1. Realization of the sample auto-correlation function ($n = 250$) of the stationary time series satisfying $X_{t+1} = 0.5X_t + Z_t$ for standard normal white noise Z_t.

$n \to \infty$. In this chapter we discuss conditions under which each of the three estimators are asymptotically normally distributed. This knowledge can be used to set approximate confidence intervals.

5.1 Sample Mean

The asymptotic normality of the sample mean is the subject of Chapter 4. The statistical significance of the central limit theorem is that the sample mean is an asymptotically consistent estimator of the true mean μ_X, and converges at the rate $1/\sqrt{n}$. The central limit theorem can be used in a preciser way to derive an asymptotic confidence interval for μ_X. This requires an estimate of the (asymptotic) variance of the sample mean, which we discuss in this section.

An approximate confidence interval for μ_X based on the sample mean typically takes the form

$$\Bigl(\overline{X}_n - \frac{\hat\sigma_n}{\sqrt{n}}1.96, \overline{X}_n + \frac{\hat\sigma_n}{\sqrt{n}}1.96\Bigr).$$

If $\sqrt{n}(\overline{X}_n - \mu_X)/\hat{\sigma}_n \rightsquigarrow N(0,1)$ as $n \to \infty$, then the confidence level of this interval converges to 95%. The problem is to find suitable estimators $\hat{\sigma}_n$.

If the sequence $\sqrt{n}(\overline{X}_n - \mu_X)$ is asymptotically normal, as it is under the conditions of the preceding chapter, the procedure works if the $\hat{\sigma}_n$ are consistent estimators of the (asymptotic) variance of $\overline{X}_n$. Unlike in the case of independent, identically distributed variables, the variance of the sample mean depends on characteristics of the joint distribution of $(X_1, \ldots, X_n)$, rather than only on the marginal distributions. (See (4.1).) The limiting variance $\sum_h \gamma_X(h)$ depends even on the joint distribution of the infinite sequence $(X_1, X_2, \ldots)$. With a sufficient number of observations it is possible to estimate the auto-covariances $\gamma_X(h)$ at smaller lags h, but, without further information, this is not true for larger lags $h \approx n$ (let alone $h \geq n$), unless we make special assumptions. Setting a confidence interval is therefore much harder than in the case of independent, identically distributed variables.

If a reliable model is available, expressed in a vector of parameters, then the problem can be solved by a *model-based estimator*. We express the variance of the sample mean in these parameters, and next plug in estimates for these parameters. If there are not too many parameters in the model this should be feasible. (Methods to estimate parameters are discussed in later chapters.)

5.2 EXERCISE.

(i) Calculate the asymptotic variance of the sample mean for the moving average $X_t = Z_t + \theta Z_{t-1}$.

(ii) Same question for the stationary solution of $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$.

However, the use of a model-based estimator is at odds with the theme of this chapter: nonparametric estimation. It is possible to estimate the variance nonparametrically provided the time series is sufficiently mixing. We discuss several methods.

A commonly used method is the method of *batched means*. The total set of observations is split into r blocks $[X_1, \ldots, X_l], [X_{l+1}, \ldots, X_{2l}], \ldots, [X_{(r-1)l+1}, \ldots, X_{rl}]$ of l observations. (Assume that $n = rl$ for simplicity; drop a last batch of fewer than l observations.) If $Y_1, \ldots, Y_r$ are the sample means of the r blocks, then $\overline{Y}_r = \overline{X}_n$ and hence $\operatorname{var} \overline{Y}_r = \operatorname{var} \overline{X}_n$. The hope is that we can ignore the dependence between $Y_1, \ldots, Y_r$ and can simply estimate the variance $\operatorname{var}(\sqrt{r}\overline{Y}_r)$ by the sample variance $S^2_{r,Y}$ of $Y_1, \ldots, Y_r$. If l is "large enough" and the orginal series X_t is sufficiently mixing, then this actually works, to some extent.

Presumably, the method of batched means uses disjoint blocks of X_t in order to achieve the approximate independence of the block means $Y_1, \ldots, Y_r$ used for its motivation. In general the block means are still dependent. This does not cause much (additional) bias in the estimate of the variance, but it may have an effect on the precision. It turns out that it is better to use *all* blocks of l consecutive X_t, even though these may be more dependent. Thus in our second method we consider all blocks $[X_1, \ldots, X_l], [X_2, \ldots, X_{l+1}], \ldots, [X_{n-l+1}, \ldots, X_n]$ of l consecutive observations. We let $Z_1, Z_2, \ldots, Z_{n-l+1}$ be the sample means of the $n - l + 1$ blocks, so that $l \operatorname{var} Z_i = \operatorname{var}(\sqrt{l}\overline{X}_l) \approx \operatorname{var}(\sqrt{n}\overline{X}_n)$, if l is large. This suggests to estimate the variance of $\sqrt{n}\overline{X}_n$ by $lS^2_{n-l+1,Z}$. The following theorem shows that this method works under some conditions, provided that l is chosen dependent on n with $l_n \to \infty$ at a not too fast rate. Because in the theorem l depends on n, so do the block means, and we denote them by $Z_{n,1}, \ldots, Z_{n,n-l_n+1}$. The theorem considers both the sample variance of the block means,

$$S_{n,Z}^2 = \frac{1}{n-l_n+1} \sum_{i=1}^{n-l_n+1} (Z_{n,i} - \overline{Z}_{n-l_n+1})^2,$$

and the centered empirical distribution function of the block means,

$$F_n(x) = \frac{1}{n-l_n+1} \sum_{i=1}^{n-l_n+1} 1\bigl\{\sqrt{l_n}(Z_{n,i} - \overline{X}_n) \le x\bigr\}.$$

5.3 Theorem. *Suppose that the time series X_t is strictly stationary and α-mixing with mixing coefficients satisfying $\sum_h \alpha(h) < \infty$. Let $l_n \to \infty$ such that $l_n/n \to 0$. Furthermore, suppose that $\sqrt{n}(\overline{X}_n - \mu_X) \rightsquigarrow N(0,v)$, for some number v. Then, for every x, the sequence $F_n(x)$ converges in probability to $\Phi(x/\sqrt{v})$. Furthermore, if $v = \sum_h \gamma_X(h)$ and $\sum_h \bigl|\gamma_X(h)\bigr| < \infty$, then the variance $l_n S_{n,Z}^2$ of F_n converges in probabity to v.*

Proof. Let G_n be the distribution function obtained by replacing the average $\overline{X}_n$ in the definition of F_n by μ_X. These functions are related through $F_n(x) = G_n\bigl(x + \sqrt{l_n}(\overline{X}_n - \mu_X)\bigr)$. The sequence $\sqrt{l_n}(\overline{X}_n - \mu_X)$ converges in probabity to zero, by the assumptions that the sequence $\sqrt{n}(\overline{X}_n - \mu_X)$ converges weakly and that $l_n/n \to 0$. In view of the monotonicity of the functions F_n and G_n it suffices to show that $G_n(x) \xrightarrow{\text{P}} \Phi(x/\sqrt{v})$ for every x.

Fix some x and define $Y_t = 1\bigl\{\sqrt{l_n}(Z_{n,t} - \mu_X) \le x\bigr\}$. Then the time series Y_t is strictly stationary and $G_n(x) = \overline{Y}_{n-l_n+1}$. By assumption

$$\mathrm{E}\overline{Y}_{n-l_n+1} = \mathrm{P}\bigl(\sqrt{l_n}(\overline{X}_{l_n} - \mu_X) \le x\bigr) \to \Phi(x/\sqrt{v}).$$

Because the variable Y_t depends only on the variables X_s with $t \le s < t + l_n$, the series Y_t is α-mixing with mixing coefficients bounded above by $\alpha(h - l_n)$ for $h > l_n$. Therefore, by (4.2) followed by Lemma 4.11 (with $q = r = \infty$),

$$\operatorname{var} \overline{Y}_{n-l_n+1} \le \frac{1}{n-l_n+1} \sum_h \bigl|\gamma_Y(h)\bigr| \le \frac{4}{n-l_n+1}\Bigl(\sum_{h \ge l_n} \alpha(h - l_n) + l_n \tfrac{1}{2}\Bigr).$$

This converges to zero as $n \to \infty$. Thus $G_n(x) = \overline{Y}_{n-l_n+1} \to \Phi(x/\sqrt{v})$ in probability by Chebyshev's inequality, and the first assertion of the theorem is proved.

To prove the convergence of the variance of F_n, we first note that the variances of F_n and G_n are the same. Because $G_n \rightsquigarrow N(0,v)$, Theorem 3.8 shows that the variance of G_n converges to v if and only $\int_{|x| \ge M} x^2\, dG_n(x) \xrightarrow{\text{P}} 0$ as $n \to \infty$ followed by $M \to \infty$. Now

$$\begin{aligned}\mathrm{E}\int_{|x| \ge M} x^2\, dG_n(x) &= \mathrm{E}\frac{1}{n-l_n+1} \sum_{i=1}^{n-l_n+1} \bigl|\sqrt{l_n}(Z_{n,i} - \mu_X)\bigr|^2 1\bigl\{\sqrt{l_n}|Z_{n,i} - \mu_X| \ge M\bigr\}\\ &= \mathrm{E}\bigl|\sqrt{l_n}(\overline{X}_{l_n} - \mu_X)\bigr|^2 1\bigl\{\sqrt{l_n}|\overline{X}_{l_n} - \mu_X| \ge M\bigr\}.\end{aligned}$$

By assumption $\sqrt{l_n}(\overline{X}_{l_n} - \mu_X) \rightsquigarrow N(0, v)$, while $\mathrm{E}\bigl|\sqrt{l_n}(\overline{X}_{l_n} - \mu_X)\bigr|^2 \to v$ by (4.1). Thus we can apply Theorem 3.8 in the other direction to conclude that the right side of the display converges to zero as $n \to \infty$ followed by $M \to \infty$. ∎

The usefulness of the estimate F_n goes beyond its variance. because the sequence F_n tends to the same limit distribution as the sequence $\sqrt{n}(\overline{X}_n - \mu_X)$, we can think of it as an estimator of the distribution of the latter variable. In particular, we could use the quantiles of F_n as estimators of the quantiles of $\sqrt{n}(\overline{X}_n - \mu_X)$ and use these to replace the normal quantiles and $\hat{\sigma}_n$ in the construction of a confidence interval. This gives the interval

$$\Bigl[\overline{X}_n - \frac{F_n^{-1}(0.975)}{\sqrt{n}}, \overline{X}_n - \frac{F_n^{-1}(0.025)}{\sqrt{n}}\Bigr].$$

The preceding theorem shows that this interval has asymptotic confidence level 95% for covering μ_X.

Another, related method is the *blockwise bootstrap.* Assume that $n = lr$ for simplicity. Given the same blocks $[X_1, \ldots, X_l], [X_2, \ldots, X_{l+1}], \ldots, [X_{n-l+1}, \ldots, X_n]$, we choose $r = n/l$ blocks at random with replacement and put the r blocks in a row, in any order, but preserving the order of the X_t within the r blocks. We denote the row of $n = rl$ variables obtained in this way by $X_1^*, X_2^*, \ldots, X_n^*$ and let $\overline{X_n^*}$ be their average. The *bootstrap estimate* of the distribution of $\sqrt{n}(\overline{X}_n - \mu_X)$ is by definition the conditional distribution of $\sqrt{n}(\overline{X_n^*} - \overline{X}_n)$ given $X_1, \ldots, X_n$. The corresponding estimate of the variance of $\sqrt{n}(\overline{X}_n - \mu_X)$ is the variance of this conditional distribution.

Another, but equivalent, description of the bootstrap procedure is to choose a random sample with replacement from the block averages $Z_{n,1}, \ldots, Z_{n,n-l_n+1}$. If this sample is denoted by $Z_1^*, \ldots, Z_r^*$, then the average $\overline{X_n^*}$ is also the average $\overline{Z_r^*}$. It follows that the bootstrap estimate of the variance of $\overline{X}_n$ is the conditional variance of the mean of a random sample of size r from the block averages given the values $Z_{n,1}, \ldots, Z_{n,n-l_n+1}$ of these averages. This is simply $(n/r)S^2_{n-l_n+1,Z}$, as before.

Other aspects of the bootstrap estimators of the distribution, for instance quantiles, are hard to calculate explicitly. In practice we perform computer simulation to obtain an approximation of the bootstrap estimate. By repeating the sampling procedure a large number of times (with the same values of $X_1, \ldots, X_n$), and taking the empirical distribution over the realizations, we can, in principle obtain arbitrary precision.

All three methods discussed previously are based on forming blocks of a certain length l. The proper choice of the block length is crucial for their succes: the preceding theorem shows that (one of) the estimators will be consistent provided $l_n \to \infty$ such that $l_n/n \to 0$. Additional calculations show that, under general conditions, the variances of the variance estimators are minimal if l_n is proportional to $n^{1/3}$.[♯]

5.4 EXERCISE. Extend the preceding theorem to the method of batched means. Show that the variance estimator is consistent.

[♯] See Künsch (1989), Annals of Statistics 17, p1217–1241.

5.2 Sample Auto Covariances

Replacing a given time series X_t by the centered time series $X_t - \mu_X$ does not change the auto-covariance function. Therefore, for the study of the asymptotic properties of the sample auto covariance function $\hat\gamma_n(h)$, it is not a loss of generality to assume that $\mu_X = 0$. The sample auto-covariance function can be written as

$$\hat\gamma_n(h) = \frac{1}{n}\sum_{t=1}^{n-h} X_{t+h}X_t - \overline{X}_n\Bigl(\frac{1}{n}\sum_{t=1}^{n-h} X_t\Bigr) - \Bigl(\frac{1}{n}\sum_{t=h+1}^{n} X_t\Bigr)\overline{X}_n + (\overline{X}_n)^2.$$

Under the conditions of Chapter 4 and the assumption $\mu_X = 0$, the sample mean $\overline{X}_n$ is of the order $O_P(1/\sqrt{n})$ and hence the last term on the right is of the order $O_P(1/n)$. For fixed h the second and third term are almost equivalent to $(\overline{X}_n)^2$ and are also of the order $O_P(1/n)$. Thus, under the assumption that $\mu_X = 0$,

$$\hat\gamma_n(h) = \frac{1}{n}\sum_{t=1}^{n-h} X_{t+h}X_t + O_P\Bigl(\frac{1}{n}\Bigr).$$

It follows from this and Slutsky's lemma that the asymptotic behaviour of the sequence $\sqrt{n}\bigl(\hat\gamma_n(h) - \gamma_X(h)\bigr)$ depends only on $n^{-1}\sum_{t=1}^{n-h} X_{t+h}X_t$. Here a change of n by $n-h$ (or $n-h$ by n) is asymptotically negligible, so that, for simplicity of notation, we can equivalently study the averages

$$\hat\gamma_n^*(h) = \frac{1}{n}\sum_{t=1}^{n} X_{t+h}X_t.$$

These are unbiased estimators of $\mathrm{E}X_{t+h}X_t = \gamma_X(h)$, under the condition that $\mu_X = 0$. Their asymptotic distribution can be derived by applying a central limit theorem to the averages $\overline{Y}_n$ of the variables $Y_t = X_{t+h}X_t$.

If the time series X_t is mixing with mixing coefficients $\alpha(k)$, then the time series Y_t is mixing with mixing coefficients bounded above by $\alpha(k-h)$ for $k > h \geq 0$. Because the conditions for a central limit theorem depend only on the speed at which the mixing coefficients converge to zero, this means that in most cases the mixing coefficients of X_t and Y_t are equivalent. By the Cauchy-Schwarz inequality the series Y_t has finite moments of order k if the series X_t has finite moments of order $2k$. This means that the mixing central limit theorems for the sample mean apply without further difficulties to proving the asymptotic normality of the sample auto-covariance function. The asymptotic variance takes the form $\sum_g \gamma_Y(g)$ and in general depends on fourth order moments of the type $\mathrm{E}X_{t+g+h}X_{t+g}X_{t+h}X_t$ as well as on the auto-covariance function of the series X_t. In its generality, its precise form is not of much interest.

5.5 Theorem. *If X_t is a strictly stationary, mixing time series with α-mixing coefficients such that $\int_0^1 \alpha^{-1}(u)F^{-1}_{|X_hX_0|}(1-u)^2\,du < \infty$, then the sequence $\sqrt{n}\bigl(\hat\gamma_n(h) - \gamma_X(h)\bigr)$ converges in distribution to a normal distribution.*

Another approach to central limit theorems is special to linear processes, of the form

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}. \tag{5.1}$$

Here we assume that $\ldots, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ is a sequence of independent and identically distributed variables with $\mathrm{E}Z_t = 0$, and that the constants ψ_j satisfy $\sum_j |\psi_j| < \infty$. The sample auto-covariance function of a linear process is also asymptotically normal, but the proof of this requires additional work. This work is worth while mainly because the limit variance takes a simple form in this case.

Under (5.1) with $\mu = 0$, the auto-covariance function of the series $Y_t = X_{t+h}X_t$ can be calculated as

$$\begin{aligned}\gamma_Y(g) &= \operatorname{cov}\bigl(X_{t+g+h}X_{t+g}, X_{t+h}X_t\bigr)\\ &= \sum_i\sum_j\sum_k\sum_l \psi_{t-i}\psi_{t+h-j}\psi_{t+g-k}\psi_{t+g+h-l}\operatorname{cov}(Z_iZ_j, Z_kZ_l).\end{aligned}$$

Here $\operatorname{cov}(Z_iZ_j, Z_kZ_l)$ is zero whenever one of the indices i, j, k, l occurs only once. For instance $\mathrm{E}Z_1Z_2Z_{10}Z_2 = \mathrm{E}Z_1\mathrm{E}Z_2^2\mathrm{E}Z_{10} = 0$. It also vanishes if $i = j \neq k = l$. The covariance is nonzero only if all four indices are the same, or if the indices occur in the pairs $i = k \neq j = l$ or $i = l \neq j = k$. Thus the preceding display can be rewritten as

$$\begin{aligned}&\operatorname{cov}(Z_1^2, Z_1^2)\sum_i \psi_{t-i}\psi_{t+h-i}\psi_{t+g-i}\psi_{t+g+h-i}\\ &\qquad + \operatorname{cov}(Z_1Z_2, Z_1Z_2)\sum\sum_{i\neq j} \psi_{t-i}\psi_{t+h-j}\psi_{t+g-i}\psi_{t+g+h-j}\\ &\qquad\qquad + \operatorname{cov}(Z_1Z_2, Z_2Z_1)\sum\sum_{i\neq j} \psi_{t-i}\psi_{t+h-j}\psi_{t+g-j}\psi_{t+g+h-i}\\ &= \bigl(\mathrm{E}Z_1^4 - 3(\mathrm{E}Z_1^2)^2\bigr)\sum_i \psi_i\psi_{i+h}\psi_{i+g}\psi_{i+g+h} + \gamma_X(g)^2 + \gamma_X(g+h)\gamma_X(g-h).\end{aligned}$$

In the last step we use Lemma 1.28(iii) twice, after first adding in the diagonal terms $i = j$ into the double sums. Since $\operatorname{cov}(Z_1Z_2, Z_1Z_2) = (\mathrm{E}Z_1^2)^2$, these diagonal terms account for -2 of the -3 times the sum in the first term. The variance of $\hat\gamma_n^*(h) = \overline{Y}_n$ converges to the sum over g of this expression. With $\kappa_4(Z) = \mathrm{E}Z_1^4/(\mathrm{E}Z_1^2)^2 - 3$, the *fourth cumulant* (equal to the *kurtosis* minus 3) of Z_t, this sum can be written as

$$V_{h,h} = \kappa_4(Z)\gamma_X(h)^2 + \sum_g \gamma_X(g)^2 + \sum_g \gamma_X(g+h)\gamma_X(g-h).$$

5.6 Theorem. *Suppose that (5.1) holds for an i.i.d. sequence Z_t with mean zero and $\mathrm{E}Z_t^4 < \infty$ and numbers ψ_j with $\sum_j |\psi_j| < \infty$. Then $\sqrt{n}\bigl(\hat\gamma_n(h) - \gamma_X(h)\bigr) \rightsquigarrow N(0, V_{h,h})$.*

Proof. As explained in the discussion preceding the statement of the theorem, it suffices to show that the sequence $\sqrt{n}\bigl(\hat\gamma_n^*(h) - \gamma_X(h)\bigr)$ has the given asymptotic distribution in the case that $\mu = 0$. Define $Y_t = X_{t+h}X_t$ and, for fixed $m \in \mathbb{N}$,

$$Y_t^m = \sum_{|j|\le m} \psi_j Z_{t+h-j} \sum_{|j|\le m} \psi_j Z_{t-j} = X_{t+h}^m X_t^m.$$

The time series Y_t^m is $(2m+h+1)$-dependent and strictly stationary. By Theorem 4.4 the sequence $\sqrt{n}(\overline{Y_n^m} - \mathrm{E}\overline{Y_n^m})$ is asymptotically normal with mean zero and variance

$$\sigma_m^2 = \sum_g \gamma_{Y_m}(g) = \kappa_4(Z)\gamma_{X^m}(h)^2 + \sum_g \gamma_{X^m}(g)^2 + \sum_g \gamma_{X^m}(g+h)\gamma_{X^m}(g-h),$$

where the second equality follows from the calculations preceding the theorem. For every g, as $m \to \infty$,

$$\gamma_{X^m}(g) = \mathrm{E}Z_1^2 \sum_{j:|j|\le m, |j+g|\le m} \psi_j\psi_{j+g} \to \mathrm{E}Z_1^2 \sum_j \psi_j\psi_{j+g} = \gamma_X(g).$$

Furthermore, the numbers on the left are bounded above by $\mathrm{E}Z_1^2 \sum_j |\psi_j\psi_{j+g}|$, and

$$\sum_g \Bigl(\sum_j |\psi_j\psi_{j+g}|\Bigr)^2 = \sum_g \sum_i \sum_k |\psi_i\psi_k\psi_{i+g}\psi_{k+g}| \le \sup_j |\psi_j| \Bigl(\sum_j |\psi_j|\Bigr)^3 < \infty.$$

Therefore, by the dominated convergence theorem $\sum_g \gamma_{X^m}(g)^2 \to \sum_g \gamma_X(g)^2$ as $m \to \infty$. By a similar argument, we obtain the corresponding property for the third term in the expression defining σ_m^2, whence $\sigma_m^2 \to V_{h,h}$ as $m \to \infty$.

We conclude by Lemma 3.10 that there exists a sequence $m_n \to \infty$ such that $\sqrt{n}(\overline{Y_n^{m_n}} - \mathrm{E}\overline{Y_n^{m_n}}) \rightsquigarrow N(0, V_{h,h})$. The proof of the theorem is complete once we also have shown that the difference between the sequences $\sqrt{n}(\overline{Y_n} - \mathrm{E}\overline{Y_n})$ and $\sqrt{n}(\overline{Y_n^{m_n}} - \mathrm{E}\overline{Y_n^{m_n}})$ converges to zero in probability.

Both sequences are centered at mean zero. In view of Chebyshev's inequality it suffices to show that $n \operatorname{var}(\overline{Y_n} - \overline{Y_n^{m_n}}) \to 0$. We can write

$$Y_t - Y_t^m = X_{t+h}X_t - X_{t+h}^m X_t^m = \sum_i \sum_j \psi_{t-i,t+h-j}^m Z_i Z_j,$$

where $\psi_{i,j}^m = \psi_i\psi_j$ if $|i| > m$ or $|j| > m$ and is 0 otherwise. The variables $\overline{Y_n} - \overline{Y_n^m}$ are the averages of these double sums and hence $\sqrt{n}$ times their variance can be found as

$$\begin{aligned}
&\sum_{g=-n}^{n} \Bigl(\frac{n-|g|}{n}\Bigr)\gamma_{Y-Y^m}(g)\\
&\qquad = \sum_{g=-n}^{n} \Bigl(\frac{n-|g|}{n}\Bigr) \sum_i \sum_j \sum_k \sum_l \psi_{t-i,t+h-j}^m \psi_{t+g-k,t+g+h-l}^m \operatorname{cov}(Z_iZ_j, Z_kZ_l).
\end{aligned}$$

Most terms in this five-fold sum are zero and by similar arguments as before the whole expression can be bounded in absolute value by

$$\operatorname{cov}(Z_1^2, Z_1^2)\sum_g\sum_i|\psi^m_{i,i+h}\psi^m_{g,g+h}|+(\mathrm{E}Z_1^2)^2\sum_g\sum_i\sum_j|\psi^m_{i,j}\psi^m_{i+g,j+g}|$$
$$+(\mathrm{E}Z_1^2)^2\sum_g\sum_i\sum_j|\psi^m_{i,j+h}\psi^m_{j+g,i+g+h}|.$$

We have that $\psi^m_{i,j}\to 0$ as $m\to\infty$ for every fixed (i,j), $|\psi^m_{i,j}|\le|\psi_i\psi_j|$, and $\sup_i|\psi_i|<\infty$. By the dominated convergence theorem the double and triple sums converge to zero as well. ∎

By similar arguments we can also prove the joint asymptotic normality of the sample auto-covariances for a number of lags h simultaneously. By the Cramér-Wold device a sequence of k-dimensional random vectors X_n converges in distribution to a random vector X if and only if $a^TX_n\rightsquigarrow a^TX$ for every $a\in\mathbb{R}^k$. A linear combination of sample auto-covariances can be written as an average, as before. These averages can be shown to be asymptotically normal by the same methods, with only the notation becoming more complex.

5.7 Theorem. *Under the conditions of either Theorem 5.5 or 5.6, for every $h\in\mathbb{N}$ and some $(h+1)\times(h+1)$-matrix V,*

$$\sqrt{n}\left(\begin{pmatrix}\hat\gamma_n(0)\\ \vdots\\ \hat\gamma_n(h)\end{pmatrix}-\begin{pmatrix}\gamma_X(0)\\ \vdots\\ \gamma_X(h)\end{pmatrix}\right)\rightsquigarrow N_{h+1}(0,V).$$

For a linear process X_t the matrix V has (g,h)-element

$$V_{g,h}=\kappa_4(Z)\gamma_X(g)\gamma_X(h)+\sum_k\gamma_X(k+g)\gamma_X(k+h)+\sum_k\gamma_X(k-g)\gamma_X(k+h).$$

5.3 Sample Auto Correlations

The asymptotic distribution of the auto-correlations $\hat\rho_n(h)$ can be obtained from the asymptotic distribution of the auto-covariance function by the Delta-method (Theorem 3.15). We can write

$$\hat\rho_n(h)=\frac{\hat\gamma_n(h)}{\hat\gamma_n(0)}=\phi\bigl(\hat\gamma_n(0),\hat\gamma_n(h)\bigr),$$

for ϕ the function $\phi(u,v)=v/u$. This function has gradient $(-v/u^2,1/u)$. By the Delta-method,

$$\sqrt{n}\bigl(\hat\rho_n(h)-\rho_X(h)\bigr)=-\frac{\gamma_X(h)}{\gamma_X(0)^2}\sqrt{n}\bigl(\hat\gamma_n(0)-\gamma_X(0)\bigr)\\ +\frac{1}{\gamma_X(0)}\sqrt{n}\bigl(\hat\gamma_n(h)-\gamma_X(h)\bigr)+o_P(1).$$

The limit distribution of the right side is the distribution of the random variable $-\gamma_X(h)/\gamma_X(0)^2Y_0+1/\gamma_X(0)Y_h$ for Y a random vector with the $N_{h+1}(0,V)$-distribution given in Theorem 5.7. The joint limit distribution of a vector of auto-correlations is the joint distribution of the corresponding linear combinations of the Y_h. By linearity this is a Gaussian distribution; its mean is zero and its covariance matrix can be expressed in the matrix V by linear algebra.

5.8 Theorem. *Under the conditions of either Theorem 5.5 or 5.6, for every $h\in\mathbb{N}$ and some $h\times h$-matrix W,*

$$\sqrt{n}\left(\begin{pmatrix}\hat\rho_n(1)\\ \vdots\\ \hat\rho_n(h)\end{pmatrix}-\begin{pmatrix}\rho_X(1)\\ \vdots\\ \rho_X(h)\end{pmatrix}\right)\rightsquigarrow N_h(0,W),$$

For a linear process X_t the matrix W has (g,h)-element

$$W_{g,h}=\sum_k\Bigl[\rho_X(k+g)\rho_X(k+h)+\rho_X(k-g)\rho_X(k+h)+2\rho_X(g)\rho_X(h)\rho_X(k)^2\\ -2\rho_X(g)\rho_X(k)\rho_X(k+h)-2\rho_X(h)\rho_X(k)\rho_X(k+g)\Bigr].$$

The expression for the asymptotic covariance matrix W of the auto-correlation coefficients in the case of a linear process is known as *Bartlett's formula*. An interesting fact is that W depends on the auto-correlation function ρ_X only, although the asymptotic covariance matrix V of the sample auto-covariance coefficients depends also on the second and fourth moments of Z_1. We discuss two interesting examples of this formula.

5.9 Example (Iid sequence). For $\psi_0=1$ and $\psi_j=0$ for $j\neq 0$, the linear process X_t given by (5.1) is equal to the i.i.d. sequence $\mu+Z_t$. Then $\rho_X(h)=0$ for every $h\neq 0$ and the matrix W given by Bartlett's formula reduces to the identity matrix. This means that for large n the sample auto-correlations $\hat\rho_n(1),\ldots,\hat\rho_n(h)$ are approximately independent normal variables with mean zero and variance $1/n$.

This can be used to test whether a given sequence of random variables is independent. If the variables are independent and identically distributed, then approximately 95 % of the computed auto-correlations should be in the interval $[-1.96/\sqrt{n},1.96/\sqrt{n}]$. This is often verified graphically, from a plot of the auto-correlation function, on which the given interval is indicated by two horizontal lines. Note that, just as we should expect that 95 % of the sample auto-correlations are inside the two bands in the plot, we should also expect that 5 % of them are not! A more formal test would be to compare the sum of the squared sample auto-correlations to the appropriate chisquare table. The *Ljung-Box* statistic is defined by

$$\sum_{h=1}^{k} \frac{n(n+2)}{n-h} \hat\rho_n(h)^2.$$

By the preceding theorem, for fixed k, this sequence of statistics tends to the χ^2 distribution with k degrees of freedom, as $n \to \infty$. (The coefficients $n(n+2)/(n-h)$ are motivated by a calculation of moments for finite n and are thought to improve the chisquare approximation, but are asymptotically equivalent to n.)

The more auto-correlations we use in a procedure of this type, the more information we extract from the data and hence the better the result. However, the tests are based on the asymptotic distribution of the sample auto-correlations and this was derived under the assumption that the lag h is fixed and $n \to \infty$. We should expect that the convergence to normality is slower for sample auto-correlations $\hat\rho_n(h)$ of larger lags h, since there are fewer terms in the sums defining them. Thus in practice we should not use sample auto-correlations of lags that are large relative to n. □

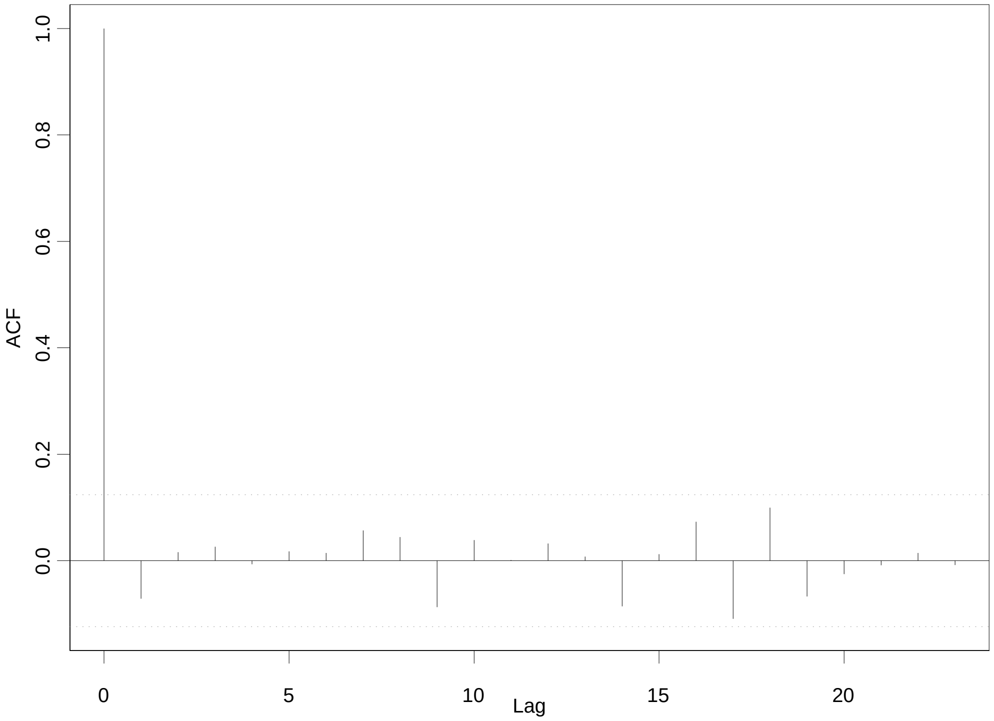

Figure 5.2. Realization of the sample auto-correlation function of a Gaussian white noise series of length 250.

5.10 Example (Moving average). For a moving average $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$ of order q, the auto-correlations $\rho_X(h)$ of lags $h > q$ vanish. By the preceding theorem the sequence $\sqrt{n}\hat\rho_n(h)$ converges for $h > q$ in distribution to a normal distribution with variance

$$W_{h,h} = \sum_k \rho_X(k)^2 = 1 + 2\rho_X(1)^2 + \cdots + 2\rho_X(q)^2, \qquad h > q.$$

This can be used to test whether a moving average of a given order q is an appropriate model for a given observed time series. A plot of the auto-correlation function shows nonzero auto-correlations for lags $1, \ldots, q$, and zero values for lags $h > q$. In practice we plot the sample auto-correlation function. Just as in the preceding example, we should expect that some sample auto-correlations of lags $h > q$ are significantly different from zero, due to the estimation error. The asymptotic variances $W_{h,h}$ are bigger than 1 and hence we should take the confidence bands a bit wider than the intervals $[-1.96/\sqrt{n}, 1.96/\sqrt{n}]$ as in the preceding example. A proper interpretation is more complicated, because the sample auto-correlations are not asymptotically independent. □

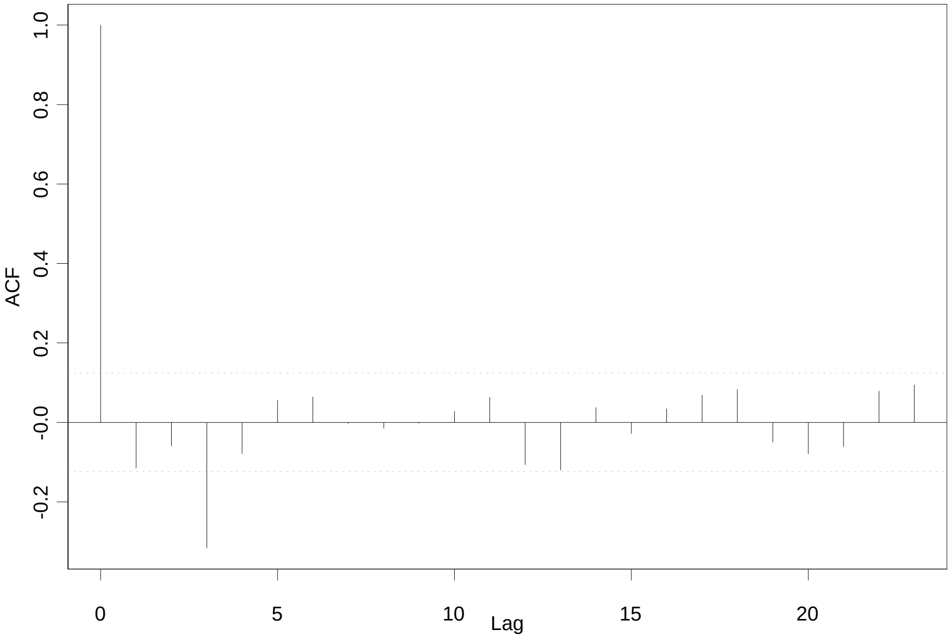

Figure 5.3. Realization ($n = 250$) of the sample auto-correlation function of the moving average process $X_t = 0.5Z_t + 0.2Z_{t-1} + 0.5Z_{t-2}$ for a Gaussian white noise series Z_t.

5.11 EXERCISE. Verify the formula for $W_{h,h}$ in the preceding example.

5.12 EXERCISE. Find $W_{1,1}$ as a function of θ for the process $X_t = Z_t + \theta Z_{t-1}$.

5.13 EXERCISE. Verify Bartlett's formula.

5.4 Sample Partial Auto Correlations

By Lemma 2.33 and the prediction equations the partial auto-correlation $\alpha_X(h)$ is the solution ϕ_h of the system of equations

$$\begin{pmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(h-1) \\ \vdots & \vdots & & \vdots \\ \gamma_X(h-1) & \gamma_X(h-2) & \cdots & \gamma_X(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_h \end{pmatrix} = \begin{pmatrix} \gamma_X(1) \\ \vdots \\ \gamma_X(h) \end{pmatrix}.$$

A nonparametric estimator $\hat\alpha_n(h)$ of $\alpha_X(h)$ is obtained by replacing the auto-covariance function in this linear system by the sample auto-covariance function $\hat\gamma_n$. This yields estimators $\hat\phi_1, \ldots, \hat\phi_h$ of the prediction coefficients satisfying

$$\begin{pmatrix} \hat\gamma_n(0) & \hat\gamma_n(1) & \cdots & \hat\gamma_n(h-1) \\ \vdots & \vdots & & \vdots \\ \hat\gamma_n(h-1) & \hat\gamma_n(h-2) & \cdots & \hat\gamma_n(0) \end{pmatrix} \begin{pmatrix} \hat\phi_1 \\ \vdots \\ \hat\phi_h \end{pmatrix} = \begin{pmatrix} \hat\gamma_n(1) \\ \vdots \\ \hat\gamma_n(h) \end{pmatrix}.$$

Then we define a nonparametric estimator for $\alpha_X(h)$ by $\hat\alpha_n(h) = \hat\phi_h$.

If we write these two systems of equations as $\Gamma\phi = \gamma$ and $\hat\Gamma\hat\phi = \hat\gamma$, respectively, then we obtain that

$$\hat\phi - \phi = \hat\Gamma^{-1}\hat\gamma - \Gamma^{-1}\gamma = \hat\Gamma^{-1}(\hat\gamma - \gamma) - \hat\Gamma^{-1}(\hat\Gamma - \Gamma)\Gamma^{-1}\gamma.$$

The sequences $\sqrt{n}(\hat\gamma - \gamma)$ and $\sqrt{n}(\hat\Gamma - \Gamma)$ are jointly asymptotically normal by Theorem 5.7. With the help of Slutsky's lemma we readily obtain the asymptotic normality of the sequence $\sqrt{n}(\hat\phi - \phi)$ and hence of the sequence $\sqrt{n}\bigl(\hat\alpha_n(h) - \alpha_X(h)\bigr)$. The asymptotic covariance matrix appears to be complicated, in general; we shall not derive it.

5.14 Example (Auto regression). For the stationary solution to $X_t = \phi X_{t-1} + Z_t$ and $|\phi| < 1$, the partial auto-correlations of lags $h \geq 2$ vanish, by Example 2.34. We shall see later that in this case the sequence $\sqrt{n}\hat\alpha_n(h)$ is asymptotically standard normally distributed, for every $h \geq 2$.

This result extends to the "causal" solution of the pth order auto-regressive scheme $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$ and the auto-correlations of lags $h > p$. (The meaning of "causal" is explained in Chapter 7.) This property can be used to find an appropriate order p when fitting an auto-regressive model to a given time series. The order is chosen such that "most" of the sample auto-correlations of lags bigger than p are within the band $[-1.96/\sqrt{n}, 1.96/\sqrt{n}]$. A proper interpretation of "most" requires that the dependence of the $\hat\alpha_n(h)$ is taken into consideration. □

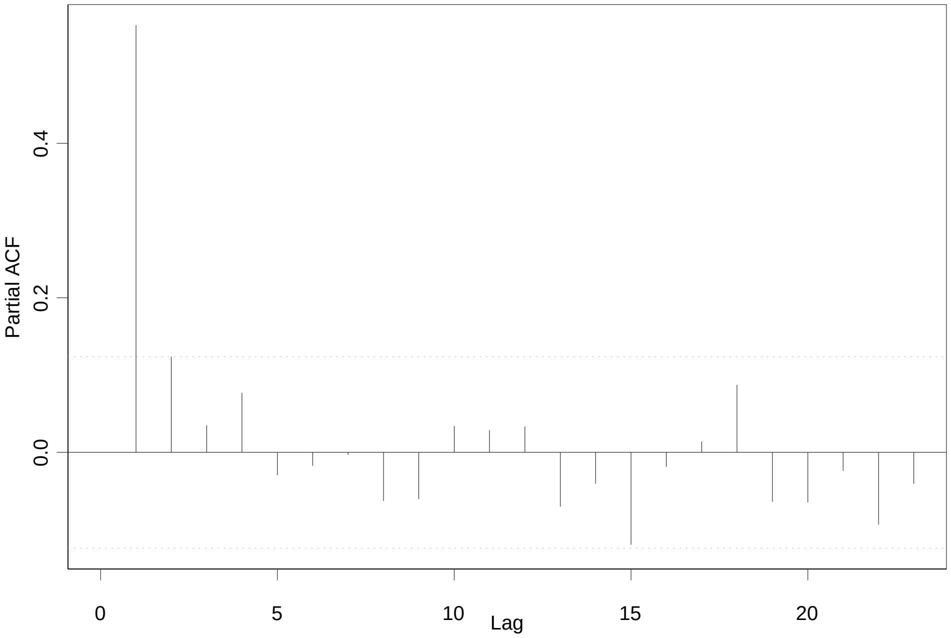

Figure 5.4. Realization ($n = 250$) of the partial auto-correlation function of the stationary solution to $X_t = 0.5X_{t-1} + 0.2X_{t-1} + Z_t$ for a Gaussian white noise series.

6
Spectral Theory

Let X_t be a stationary, possibly complex, time series with auto-covariance function γ_X. If the series $\sum_h |\gamma_X(h)|$ is convergent, then the series

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma_X(h) e^{-ih\lambda}, \tag{6.1}$$

is absolutely convergent, uniformly in $\lambda \in \mathbb{R}$. This function is called the *spectral density* of the time series X_t. Because it is periodic with period 2π it suffices to consider it on an interval of length 2π, which we shall take to be $(-\pi, \pi]$. In the present context the values λ in this interval are often referred to as *frequencies*, for reasons that will become clear. By the uniform convergence, we can exchange the order of sum and integral when computing $\int_{-\pi}^{\pi} e^{ih\lambda}\, f_X(\lambda)\, d\lambda$ and we find that, for every $h \in \mathbb{Z}$,

$$\gamma_X(h) = \int_{-\pi}^{\pi} e^{ih\lambda}\, f_X(\lambda)\, d\lambda.$$

Thus the spectral density f_X determines the auto-covariance function, just as the auto-covariance function determines the spectral density.

6.1 EXERCISE. Prove this *inversion formula*, after first verifying that $\int_{-\pi}^{\pi} e^{ih\lambda}\, d\lambda = 0$ for integers $h \neq 0$ and $\int_{-\pi}^{\pi} e^{ih\lambda}\, d\lambda = 2\pi$ for $h = 0$.

In mathematical analysis the series f_X is called a *Fourier series* and the numbers $\gamma_X(h)$ are called the *Fourier coefficients* of f_X. (The factor $1/(2\pi)$ is sometimes omitted or replaced by another number, and the Fourier series is often defined as $f_X(-\lambda)$ rather than $f_X(\lambda)$, but this is inessential.) A main topic of Fourier analysis is to derive conditions under which a Fourier series converges, in an appropriate sense, and to investigate whether the inversion formula is valid. We have just answered these questions under the assumption that $\sum_h \bigl|\gamma_X(h)\bigr| < \infty$. This condition is more restrictive than necessary, but is sufficient for most of our purposes.

6.1 Spectral Measures

The requirement that the series $\sum_h \gamma_X(h)$ is absolutely convergent means roughly that $\gamma_X(h) \to 0$ as $h \to \pm\infty$ at a "sufficiently fast" rate. In statistical terms it means that variables X_t that are widely separated in time must be approximately uncorrelated. This is not true for every time series, and consequently not every time series possesses a spectral density. However, every stationary time series does have a "spectral measure", by the following theorem.

6.2 Theorem (Herglotz). *For every stationary time series X_t there exists a unique finite measure F_X on $(-\pi, \pi]$ such that*

$$\gamma_X(h) = \int_{(-\pi,\pi]} e^{ih\lambda}\, dF_X(\lambda), \qquad h \in \mathbb{Z}.$$

Proof. Define F_n as the measure on $[-\pi, \pi]$ with Lebesgue density equal to

$$f_n(\lambda) = \frac{1}{2\pi} \sum_{h=-n}^{n} \gamma_X(h)\Bigl(1 - \frac{|h|}{n}\Bigr) e^{-ih\lambda}.$$

It is not immediately clear that this is a real-valued, nonnegative function, but this follows from the fact that

$$0 \le \frac{1}{2\pi n} \operatorname{var}\Bigl(\sum_{t=1}^{n} X_t e^{-it\lambda}\Bigr) = \frac{1}{2\pi n} \sum_{s=1}^{n}\sum_{t=1}^{n} \operatorname{cov}(X_s, X_t) e^{i(t-s)\lambda} = f_n(\lambda).$$

It is clear from the definition of f_n that the numbers $\gamma_X(h)\bigl(1 - |h|/n\bigr)$ are the Fourier coefficients of f_n for $|h| \le n$ (and the remaining Fourier coefficients of f_n are zero). Thus, by the inversion formula,

$$\gamma_X(h)\Bigl(1 - \frac{|h|}{n}\Bigr) = \int_{-\pi}^{\pi} e^{ih\lambda} f_n(\lambda)\, d\lambda = \int_{-\pi}^{\pi} e^{ih\lambda}\, dF_n(\lambda), \qquad |h| \le n.$$

Setting $h = 0$ in this equation, we see that $F_n[-\pi, \pi] = \gamma_X(0)$ for every n. Thus, apart from multiplication by the constant $\gamma_X(0)$, the F_n are probability distributions. Because the interval $[-\pi, \pi]$ is compact, the sequence F_n is uniformly tight. By Prohorov's theorem there exists a subsequence $F_{n'}$ that converges weakly to a distribution F on $[-\pi, \pi]$. Because $\lambda \mapsto e^{ih\lambda}$ is a continuous function, it follows by the portmanteau lemma that

$$\int_{[-\pi,\pi]} e^{ih\lambda}\, dF(\lambda) = \lim_{n'\to\infty} \int_{[-\pi,\pi]} e^{ih\lambda}\, dF_n(\lambda) = \gamma_X(h),$$

by the preceding display. If F puts a positive mass at $-\pi$, we can move this to the point π without affecting this identity, since $e^{-ih\pi} = e^{ih\pi}$ for every $h \in \mathbb{Z}$. The resulting F satisfies the requirements for F_X.

That this F is unique can be proved using the fact that the linear span of the functions $\lambda \mapsto e^{ih\lambda}$ is uniformly dense in the set of continuous, periodic functions (the Césaro sums of the Fourier series of a continuous, periodic function converge uniformly), which, in turn, are dense in $L_1(F)$. We omit the details of this step, which is standard Fourier analysis. ■

The measure F_X is called the *spectral measure* of the time series X_t. If the spectral measure F_X admits a density f_X relative to the Lebesgue measure, then the latter is called the *spectral density*. A sufficient condition for this is that the series $\sum \gamma_X(h)$ is absolutely convergent. Then the spectral density is the Fourier series (6.1) with coefficients $\gamma_X(h)$ introduced previously.[†]

6.3 EXERCISE. Show that the spectral density of a real-valued time series with $\sum_h \bigl|\gamma_X(h)\bigr| < \infty$ is symmetric about zero.

* **6.4** EXERCISE. Show that the spectral measure of a real-valued time series is symmetric about zero, apart from a possible point mass at π. [Hint: Use the uniqueness of a spectral measure.]

6.5 Example (White noise). The covariance function of a white noise sequence X_t is 0 for $h \neq 0$. Thus the Fourier series defining the spectral density has only one term and reduces to

$$f_X(\lambda) = \frac{1}{2\pi}\gamma_X(0).$$

The spectral measure is the uniform measure with total mass $\gamma_X(0)$. Hence "a white noise series contains all possible frequencies in an equal amount". □

6.6 Example (Deterministic trigonometric series). Let $X_t = A\cos(\lambda t) + B\sin(\lambda t)$ for mean-zero, uncorrelated variables A and B of variance σ^2, and $\lambda \in (0, \pi)$. By Example 1.5 the covariance function is given by

$$\gamma_X(h) = \sigma^2 \cos(h\lambda) = \sigma^2 \tfrac{1}{2}(e^{i\lambda h} + e^{-i\lambda h}).$$

It follows that the spectral measure F_X is the discrete 2-point measure with $F_X\{\lambda\} = F_X\{-\lambda\} = \sigma^2/2$.

Because the time series is real, the point mass at $-\lambda$ does not really count: because the spectral measure of a real time series is symmetric, the point $-\lambda$ must be there because λ is there. The form of the spectral measure and the fact that the time series in this example is a trigonometric series of frequency λ, are good motivation for referring to the values λ as "frequencies". □

6.7 EXERCISE.

(i) Show that the spectral measure of the sum $X_t + Y_t$ of two uncorrelated time series is the sum of the spectral measures of X_t and Y_t.

(ii) Construct a time series with spectral measure equal to a symmetric discrete measure on the points $\pm\lambda_1, \pm\lambda_2, \ldots, \pm\lambda_k$ with $0 < \lambda_1 < \cdots < \lambda_k < \pi$.

(iii) Construct a time series with spectral measure the 1-point measure with $F_X\{0\} = \sigma^2$.

[†] This condition is not necessary; if the series $\lambda \mapsto \Sigma_h \gamma_X(h) e^{-ih\lambda}$ converges in $L_2(F_X)$, then this series is a version of the spectral density.

(iv) Same question, but now with $F_X\{\pi\}=\sigma^2$.

* **6.8** EXERCISE. Show that every finite measure on $(-\pi,\pi]$ is the spectral measure of some stationary time series.

The spectrum of a time series is an important theoretical concept, but it is also an important practical tool to gain insight in periodicities in the data. Inference using the spectrum is called *spectral analysis* or analysis in the *frequency domain* as opposed to "ordinary" analysis, which is in the *time domain.* However, we should not have too great expectations of the insight offered by the spectrum. In some situations a spectral analysis leads to clear cut results, but in other situations the interpretation of the spectrum is complicated, or even unclear, due to the fact that all possible frequencies are present to some extent.

The idea of a spectral analysis is to view the consecutive values

$$\ldots, X_{-1}, X_0, X_1, X_2, \ldots$$

of a time series as a random function, from $\mathbb{Z}\subset\mathbb{R}$ to $\mathbb{R}$, and to write this as a weighted sum (or integral) of trigonometric functions $t\mapsto\cos\lambda t$ or $t\mapsto\sin\lambda t$ of different frequencies λ. In simple cases finitely many frequencies suffice, whereas in other situations all frequencies $\lambda\in(-\pi,\pi]$ are needed to give a full description, and the "weighted sum" becomes an integral. Two extreme examples are provided by a deterministic trigonometric series (which incorporates a single frequency) and a white noise series (which has all frequencies in equal amounts). The spectral measure gives the weights of the different frequencies in the sum. Physicists would call a time series a *signal* and refer to the spectrum as the weights at which the frequencies are present in the given signal.

We shall derive the *spectral decomposition*, the theoretical basis for this interpretation, in Section 6.3. Another method to gain insight in the interpretation of a spectrum is to consider the transformation of a spectrum by filtering. The term "filtering" stems from the field of signal processing, where a filter takes the form of an electronic device that filters out certain frequencies from a given electric current. For us, a filter will remain an infinite moving average as defined in Chapter 1. For a given filter with filter coefficients ψ_j the function $\psi(\lambda)=\sum_j\psi_j e^{-ij\lambda}$ is called the *transfer function* of the filter.

6.9 Theorem. *Let X_t be a stationary time series with spectral measure F_X and let $\sum_j|\psi_j|<\infty$. Then $Y_t=\sum_j\psi_jX_{t-j}$ has spectral measure F_Y given by*

$$dF_Y(\lambda)=\bigl|\psi(\lambda)\bigr|^2\,dF_X(\lambda).$$

Proof. According to Lemma 1.28(iii) (if necessary extended to complex-valued filters), the series Y_t is stationary with auto-covariance function

$$\gamma_Y(h)=\sum_k\sum_l\psi_k\overline{\psi}_l\gamma_X(h-k+l)=\sum_k\sum_l\psi_k\overline{\psi}_l\int e^{i(h-k+l)\lambda}\,dF_X(\lambda).$$

By the dominated convergence theorem we are allowed to change the order of (double) summation and integration. Next we can rewrite the right side as $\int\bigl|\psi(\lambda)\bigr|^2e^{ih\lambda}\,dF_X(\lambda)$.

This proves the theorem, in view of Theorem 6.2 and the uniqueness of the spectral measure. ■

6.10 Example (Moving average). A white noise process Z_t has a constant spectral density $\sigma^2/(2\pi)$. By the preceding theorem the moving average $X_t = Z_t + \theta Z_{t-1}$ has spectral density

$$f_X(\lambda) = |1 + \theta e^{-i\lambda}|^2 \,\frac{\sigma^2}{2\pi} = (1 + 2\theta\cos\lambda + \theta^2)\,\frac{\sigma^2}{2\pi}.$$

If $\theta > 0$, then the small frequencies dominate, whereas the bigger frequencies are more important if $\theta < 0$. This suggests that the sample paths of this time series will be more wiggly if $\theta < 0$. However, in both cases all frequencies are present in the signal. □

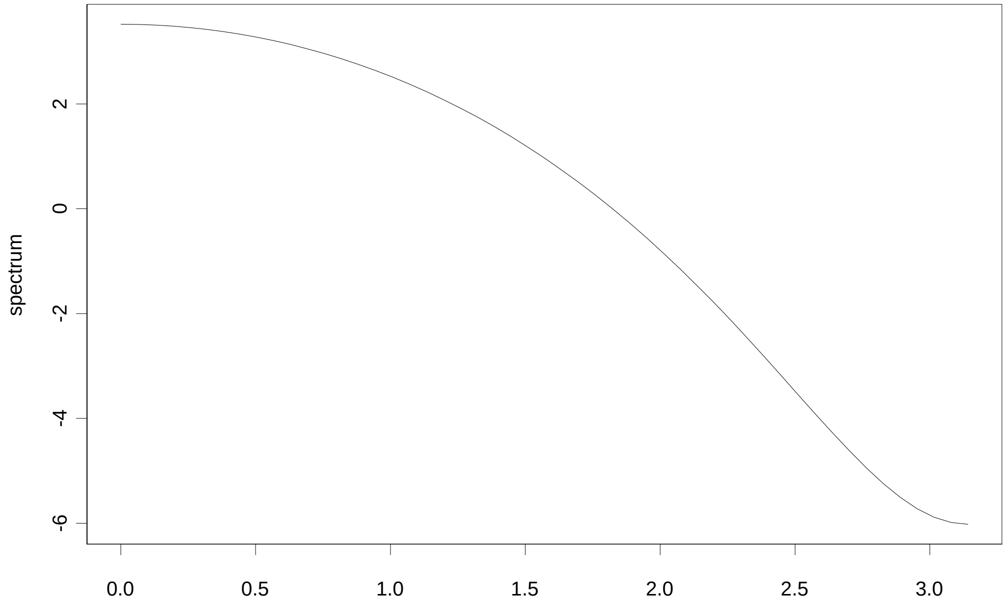

Figure 6.1. Spectral density of the moving average $X_t = Z_t + .5Z_{t-1}$. (Vertical scale in decibels.)

6.11 Example. The process $X_t = Ae^{i\lambda t}$ for a mean zero variable A and $\lambda \in (-\pi, \pi]$ has covariance function

$$\gamma_X(h) = \operatorname{cov}(Ae^{i\lambda(t+h)}, Ae^{i\lambda t}) = e^{ih\lambda}\mathrm{E}|A|^2.$$

The corresponding spectral measure is the 1-point measure F_X with $F_X\{\lambda\} = \mathrm{E}|A|^2$. Therefore, the filtered series $Y_t = \sum_j \psi_j X_{t-j}$ has spectral measure the 1-point measure with $F_Y\{\lambda\} = \bigl|\psi(\lambda)\bigr|^2 \mathrm{E}|A|^2$. By direct calculation we find that

$$Y_t = \sum_j \psi_j A e^{i\lambda(t-j)} = A e^{i\lambda t}\psi(\lambda) = \psi(\lambda) X_t.$$

This suggests an interpretation for the term "transfer function". Filtering a "pure signal" $Ae^{it\lambda}$ of a single frequency apparently yields another signal of the same single frequency, but the amplitude of the signal changes by multiplication with the factor $\psi(\lambda)$. If $\psi(\lambda) = 0$, then the frequency is "not transmitted", whereas values of $\bigl|\psi(\lambda)\bigr|$ bigger or smaller than 1 mean that the frequency λ is amplified or weakened. □

6.12 EXERCISE. Find the spectral measure of $X_t = Ae^{i\lambda t}$ for λ not necessarily belonging to $(-\pi, \pi]$.

To give a further interpretation to the spectral measure consider a *band pass filter*. This is a filter with transfer function of the form

$$\psi(\lambda) = \begin{cases} 0, & \text{if } |\lambda - \lambda_0| > \delta, \\ 1, & \text{if } |\lambda - \lambda_0| \leq \delta, \end{cases}$$

for a fixed frequency λ_0 and fixed *band width* 2δ. According to Example 6.11 this filter "kills" all the signals $Ae^{i\lambda t}$ of frequencies λ outside the interval $[\lambda_0 - \delta, \lambda_0 + \delta]$ and transmits all signals $Ae^{it\lambda}$ for λ inside this range unchanged. The spectral density of the filtered signal $Y_t = \sum_j \psi_j X_{t-j}$ relates to the spectral density of the original signal X_t (if there exists one) as

$$f_Y(\lambda) = \bigl|\psi(\lambda)\bigr|^2 f_X(\lambda) = \begin{cases} 0, & \text{if } |\lambda - \lambda_0| > \delta, \\ f_X(\lambda), & \text{if } |\lambda - \lambda_0| \leq \delta. \end{cases}$$

Now think of X_t as a signal composed of many frequencies. The band pass filter transmits only the subsignals of frequencies in the interval $[\lambda_0 - \delta, \lambda_0 + \delta]$. This explains that the spectral density of the filtered sequence Y_t vanishes outside this interval. For small $\delta > 0$,

$$\operatorname{var} Y_t = \gamma_Y(0) = \int_{-\pi}^{\pi} f_Y(\lambda)\, d\lambda = \int_{\lambda_0-\delta}^{\lambda_0+\delta} f_X(\lambda)\, d\lambda \approx 2\delta f_X(\lambda_0).$$

We interprete this as saying that $f_X(\lambda_0)$ is proportional to the variance of the subsignals in X_t of frequency λ_0. The total variance $\operatorname{var} X_t = \gamma_X(0) = \int_{-\pi}^{\pi} f_X(\lambda)\, d\lambda$ in the signal X_t is the total area under the spectral density. This can be viewed as the sum of the variances of the subsignals of frequencies λ, the area under f_X between $\lambda_0 - \delta$ and $\lambda_0 + \delta$ being the variance of the subsignals of frequencies in this interval.

A band pass filter is a theoretical filter: in practice it is not possible to filter out an exact range of frequencies. Only smooth transfer functions can be implemented on a computer, and only the ones corresponding to finite filters (the ones with only finitely many nonzero filter coefficients ψ_j).

The filter coefficients ψ_j relate to the transfer function $\psi(\lambda)$ in the same way as the auto-covariances $\gamma_X(h)$ relate to the spectral density $f_X(h)$, apart from a factor 2π. Thus, to find the filter coefficients of a given transfer function ψ, it suffices to apply the Fourier inversion formula

$$\psi_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ij\lambda}\psi(\lambda)\,d\lambda.$$

6.13 EXERCISE. Find the filter coefficients of a band pass filter.

6.14 Example (Low frequency and trend). An apparent trend in observed data $X_1,\ldots,X_n$ could be modelled as a real trend in a nonstationary time series, but could alternatively be viewed as the beginning of a long cycle. In practice, where we get to see only a finite stretch of a time series, low frequency cycles and slowly moving trends cannot be discriminated. It was seen in Chapter 1 that differencing $Y_t = X_t - X_{t-1}$ of a time series X_t removes a linear trend, and repeated differencing removes higher order polynomial trends. In view of the preceding observation the differencing filter should remove, to a certain extent, low frequencies.

The differencing filter has transfer function

$$\psi(\lambda) = 1 - e^{-i\lambda} = 2ie^{-i\lambda/2}\sin\frac{\lambda}{2}.$$

The absolute value $\bigl|\psi(\lambda)\bigr|$ of this transfer function increases from 0 at 0 to its maximum value at π. Thus, indeed, it filters away low frequencies, albeit only with partial success. □

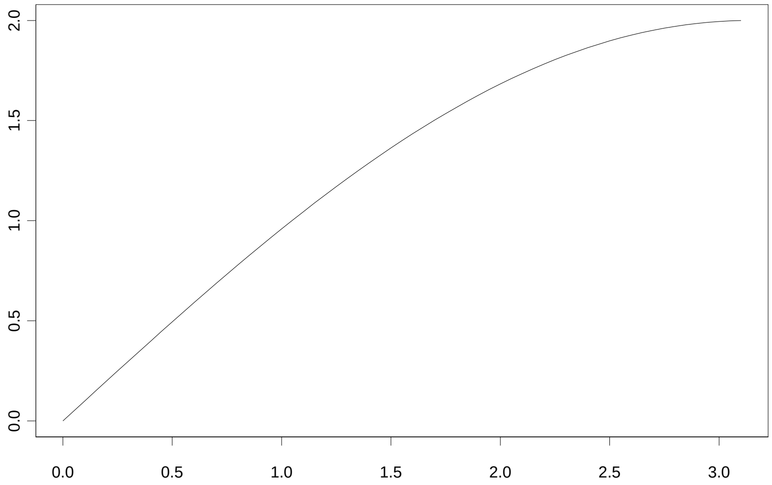

Figure 6.2. Absolute value of the transfer function of the difference filter.

6.15 Example (Averaging). The averaging filter $Y_t = (2M+1)^{-1}\sum_{j=-M}^{M} X_{t-j}$ has transfer function

$$\psi(\lambda) = \frac{1}{2M+1}\sum_{j=-M}^{M} e^{-ij\lambda} = \frac{\sin\bigl((M+\frac12)\lambda\bigr)}{(2M+1)\sin(\frac12\lambda)}.$$

(The expression on the right is defined by continuity, as 1, at $\lambda = 0$.) This function is proportional to the *Dirichlet kernel*, which is the function obtained by replacing the factor $2M+1$ by 2π. From a picture of this kernel we conclude that averaging removes high frequencies to a certain extent (and in an uneven manner depending on M), but retains low frequencies. □

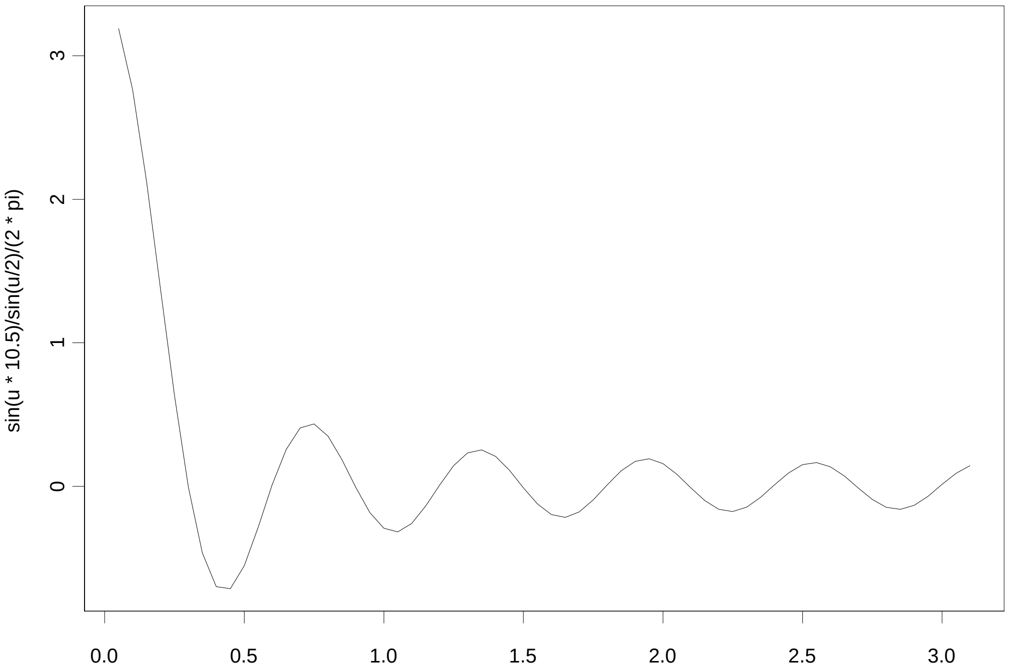

Figure 6.3. Dirichlet kernel of order $M = 10$.

6.16 EXERCISE. Express the variance of Y_t in the preceding example in ψ and the spectral density of the time series X_t (assuming that there is one). What happens if $M \to \infty$? Which conclusion can you draw? Does this remain true if the series X_t does not have a spectral density?

6.17 EXERCISE. Find the transfer function of the filter $Y_t = X_t - X_{t-12}$. Interprete the result.

Instead of in terms of frequencies we can also think in terms of *periods.* A series of the form $t \mapsto e^{i\lambda t}$ repeats itself after $2\pi/\lambda$ instants of time. Therefore, the period is defined as

$$\text{period} = \frac{2\pi}{\text{frequency}}.$$

Most monthly time series (one observation per month) have a period effect of 12 months. If so, this will be visible as a peak in the spectrum at the frequency $2\pi/12 = \pi/6$.[‡] Often the 12-month cycle is not completely regular. This may produce additional (but smaller) peaks at the *harmonic frequencies* $2\pi/6, 3\pi/6, \ldots$, or $\pi/12, \pi/18, \ldots$.

It is surprising at first that the highest possible frequency is π, the so-called *Nyquist frequency.* This is caused by the fact that the series is measured only at discrete time points. Very high fluctuations fall completely between the measurements and hence cannot be observed. The Nyquist frequency π corresponds to a period of $2\pi/\pi = 2$ time instants and this is clearly the smallest period that is observable. For time series that are observed in continuous time a spectrum is defined to contain all frequencies in $\mathbb{R}$.

* 6.2 Nonsummable filters

If given filter coefficients ψ_j satisfy $\sum_j |\psi_j| < \infty$, then the series $\psi(\lambda) = \sum_j \psi_j e^{-ij\lambda}$ converges uniformly on $(-\pi, \pi]$, and the coefficients can be recovered from the transfer function ψ by the Fourier inversion formula $\psi_j = (2\pi)^{-1} \int_{-\pi}^{\pi} e^{ij\lambda} \psi(\pi)\, d\lambda$. (See Problem 6.1.) Unfortunately, not all filters have summable coefficients. An example is the band pass filter considered previously. In fact, if a sequence of filter coefficients is summable, then the corresponding transfer function must be continuous, and $\lambda \mapsto \psi(\lambda) = 1_{[\lambda_0 - \delta, \lambda_0 + \delta]}(\lambda)$ is not. Nevertheless, the series $\sum_j \psi_j e^{-ij\lambda}$ is well defined for the band pass filter and has the function $1_{[\lambda_0 - \delta, \lambda_0 + \delta]}(\lambda)$ as its limit in a certain sense. To handle examples such as this it is worthwhile to generalize Theorem 6.9 (and Lemma 1.28) a little.

6.18 Theorem. *Let X_t be a stationary time series with spectral measure F_X, defined on the probability space $(\Omega, \mathcal{U}, \mathrm{P})$. Then the series $\psi(\lambda) = \sum_j \psi_j e^{-i\lambda j}$ converges in $L_2(F_X)$ if and only if $Y_t = \sum_j \psi_j X_{t-j}$ converges in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ for some t (and then for every $t \in \mathbb{Z}$) and in that case*

$$dF_Y(\lambda) = \bigl|\psi(\lambda)\bigr|^2\, dF_X(\lambda).$$

Proof. For $0 \le m \le n$ let $\psi_j^{m,n}$ be equal to ψ_j for $m \le |j| \le n$ and be 0 otherwise, and define $Y_t^{m,n}$ as the series X_t filtered by the coefficients $\psi_j^{m,n}$. Then certainly $\sum_j |\psi_j^{m,n}| < \infty$ for every fixed pair (m, n) and hence we can apply Lemma 1.28 and Theorem 6.9 to

[‡] That this is a complicated number is an inconvenient consequence of our convention to define the spectrum on the interval $(-\pi, \pi]$. This can be repaired. For instance, the Splus package produces spectral plots with the frequencies rescaled to the interval $(-\frac{1}{2}, \frac{1}{2}]$. Then a 12-month period gives a peak at 1/12.

the series $Y_t^{m,n}$. This yields

$$\begin{aligned}\mathrm{E}\Bigl|\sum_{m\le|j|\le n}\psi_jX_{t-j}\Bigr|^2&=\mathrm{E}|Y_t^{m,n}|^2=\gamma_{Y^{m,n}}(0)\\&=\int_{(-\pi,\pi]}dF_{Y^{m,n}}=\int_{(-\pi,\pi]}\Bigl|\sum_{m\le|j|\le n}\psi_je^{-i\lambda j}\Bigr|^2\,dF_X(\lambda).\end{aligned}$$

The left side converges to zero for $m,n\rightarrow\infty$ if and only if the partial sums of the series $Y_t=\sum_j\psi_jX_{t-j}$ form a Cauchy sequence in $L_2(\Omega,\mathcal{U},\mathrm{P})$. The right side converges to zero if and only if the partial sums of the sequence $\sum_j\psi_je^{-i\lambda j}$ form a Cauchy sequence in $L_2(F_X)$. The first assertion of the theorem now follows, because both spaces are complete.

To prove the second assertion, we first note that, by Theorem 6.9,

$$\operatorname{cov}\Bigl(\sum_{|j|\le n}\psi_jX_{t+h-j},\sum_{|j|\le n}\psi_jX_{t-j}\Bigr)=\gamma_{Y^{0,n}}(h)=\int_{(-\pi,\pi]}\Bigl|\sum_{|j|\le n}\psi_je^{-i\lambda j}\Bigr|^2e^{ih\lambda}\,dF_X(\lambda).$$

We now take limits of the left and right sides as $n\rightarrow\infty$ to find that $\gamma_Y(h)=\int_{(-\pi,\pi]}\bigl|\psi(\lambda)\bigr|^2\,e^{ih\lambda}\,dF_X(\lambda)$, for every h. ■

6.19 Example. If the filter coefficients satisfy $\sum_j|\psi_j|^2<\infty$ (which is weaker than absolute convergence), then the series $\sum_j\psi_je^{-ij\lambda}$ converges in $L_2(\lambda)$ for λ the Lebesgue measure on $(-\pi,\pi]$. This is a central fact in Fourier theory and follows from

$$\int_{-\pi}^{\pi}\Bigl|\sum_{m\le|j|\le n}\psi_je^{-ij\lambda}\Bigr|^2\,d\lambda=\sum_{m\le|k|\le n}\sum_{m\le|l|\le n}\psi_k\overline{\psi}_l\int_{-\pi}^{\pi}e^{i(l-k)\lambda}\,d\lambda=\sum_{m\le|j|\le n}|\psi_j|^2.$$

Consequently, the series also converges in $L_2(F_X)$ for every spectral measure F_X that possesses a bounded density.

Thus, in many cases a sequence of square-summable coefficients defines a valid filter. A particular example is a band pass filter, for which $|\psi_j|=O(1/|j|)$ as $j\rightarrow\pm\infty$. □

* 6.3 Spectral Decomposition

In the preceding section we interpreted the mass $F_X(I)$ that the spectral distribution gives to an interval I as the size of the contribution of the components of frequencies $\lambda\in I$ to the signal $t\mapsto X_t$. In this section we give a precise mathematical meaning to this idea. We show that a given stationary time series X_t can be written as a randomly weighted sum of single frequency signals $e^{i\lambda t}$.

This decomposition is simple in the case of a discrete spectral measure. For given uncorrelated mean zero random variables $Z_1, \ldots, Z_k$ and numbers $\lambda_1, \ldots, \lambda_k \in (-\pi, \pi]$ the process

$$X_t = \sum_{j=1}^{k} Z_j e^{i\lambda_j t}$$

possesses as spectral measure F_X the discrete measure with point masses of sizes $F_X\{\lambda_j\} = \mathrm{E}|Z_j|^2$ at the frequencies $\lambda_1, \ldots, \lambda_k$ (and no other mass). The series X_t is the sum of uncorrelated, single-frequency signals of stochastic amplitudes $|Z_j|$. This is called the *spectral decomposition* of the series X_t. We prove below that this construction can be reversed: given a mean zero, stationary time series X_t with discrete spectral measure as given, there exist mean zero uncorrelated random variables $Z_1, \ldots, Z_k$ with variances $F_X\{\lambda_j\}$ such that the decomposition is valid.

This justifies the interpretation of the spectrum given in the preceding section. The possibility of the decomposition is surprising in that the spectral measure only involves the auto-covariance function of a time series, whereas the spectral decomposition is a decomposition of the sample paths of the time series: if the series X_t is defined on a given probability space $(\Omega, \mathcal{U}, \mathrm{P})$, then so are the random variables Z_j and the preceding spectral decomposition may be understood as being valid for (almost) every $\omega \in \Omega$. This can be true, of course, only if the variables $Z_1, \ldots, Z_k$ also have other properties besides the ones described. The spectral theorem below does not give any information about these further properties. For instance, even though uncorrelated, the Z_j need not be independent. This restricts the usefulness of the spectral decomposition, but we could not expect more. The spectrum only involves the second moment properties of the time series, and thus leaves most of the distribution of the series undescribed. An important exception to this rule is if the series X_t is Gaussian. Then the first and second moments, and hence the mean and the spectral distribution, completely describe the distribution of the series X_t.

The spectral decomposition is not restricted to time series' with discrete spectral measures. However, in general, the spectral decomposition involves a continuum of frequencies and the sum becomes an integral

$$X_t = \int_{(-\pi,\pi]} e^{i\lambda t}\, dZ(\lambda).$$

A technical complication is that such an integral, relative to a "random measure" Z, is not defined in ordinary measure theory. We must first give it a meaning.

6.20 Definition. *A random measure with orthogonal increments Z is a collection $\{Z(B)\colon B \in \mathcal{B}\}$ of mean zero, complex random variables $Z(B)$ indexed by the Borel sets $\mathcal{B}$ in $(-\pi, \pi]$ defined on some probability space $(\Omega, \mathcal{U}, \mathrm{P})$ such that, for some finite Borel measure μ on $(-\pi, \pi]$,*

$$\operatorname{cov}\bigl(Z(B_1), Z(B_2)\bigr) = \mu(B_1 \cap B_2), \qquad \textit{every } B_1, B_2 \in \mathcal{B}.$$

This definition does not appear to include a basic requirement of a measure: that the measure of a countable union of disjoint sets is the sum of the measures of the individual sets. However, we leave it as an exercise to show that this is implied by the covariance property.

6.21 EXERCISE. Let Z be a random measure with orthogonal increments. Show that $Z(\cup_j B_j) = \sum_j Z(B_j)$ in mean square, whenever $B_1, B_2, \ldots$ is a sequence of pairwise disjoint Borel sets.

6.22 EXERCISE. Let Z be a random measure with orthogonal increments and define $Z_\lambda = Z(-\pi, \lambda]$. Show that $(Z_\lambda: \lambda \in (-\pi, \pi])$ is a stochastic process with uncorrelated increments: for $\lambda_1 < \lambda_2 \le \lambda_3 < \lambda_4$ the variables $Z_{\lambda_4} - Z_{\lambda_3}$ and $Z_{\lambda_2} - Z_{\lambda_1}$ are uncorrelated. This explains the phrase "with orthogonal increments".

* **6.23** EXERCISE. Suppose that Z_λ is a mean zero stochastic process with finite second moments and uncorrelated increments. Show that this process corresponds to a random measure with orthogonal increments as in the preceding exercise. [This asks you to reconstruct the random measure Z from the weights $Z_\lambda = Z(-\pi, \lambda]$ it gives to cells, similarly as an ordinary measure can be reconstructed from its distribution function.]

Next we define an "integral" $\int f\, dZ$ for given functions $f: (-\pi, \pi] \to \mathbb{C}$. For an indicator function $f = 1_B$ of a Borel set B, we define, in analogy with an ordinary integral, $\int 1_B\, dZ = Z(B)$. Because we wish the integral to be linear, we are lead to the definition

$$\int \sum_j \alpha_j 1_{B_j}\, dZ = \sum_j \alpha_j Z(B_j),$$

for every finite collections of complex numbers α_j and Borel sets B_j. This determines the integral for many, but not all functions f. We extend its domain by continuity: we require that $\int f_n\, dZ \to \int f\, dZ$ whenever $f_n \to f$ in $L_2(\mu)$. The following lemma shows that these definitions and requirements can be consistently made, and serves as a definition of $\int f\, dZ$.

6.24 Lemma. *For every random measure with orthogonal increments Z there exists a unique map, denoted $f \mapsto \int f\, dZ$, from $L_2(\mu)$ into $L_2(\Omega, \mathcal{U}, \mathrm{P})$ with the properties*
(i) $\int 1_B\, dZ = Z(B)$;
(ii) $\int (\alpha f + \beta g)\, dZ = \alpha \int f\, dZ + \beta \int g\, dZ$;
(iii) $\mathrm{E}\bigl|\int f\, dZ\bigr|^2 = \int |f|^2\, d\mu$.
In other words, the map $f \mapsto \int f\, dZ$ is a linear isometry such that $1_B \mapsto Z(B)$.

Proof. By the defining property of Z, for any complex numbers α_i and Borel sets B_i,

$$\mathrm{E}\Bigl|\sum_{i=1}^k \alpha_i Z(B_i)\Bigr|^2 = \sum_i \sum_j \alpha_i \overline{\alpha}_j \operatorname{cov}\bigl(Z(B_i), Z(B_j)\bigr) = \int \Bigl|\sum_{j=1}^k \alpha_i 1_{B_i}\Bigr|^2\, d\mu.$$

For f a simple function of the form $f = \sum_i \alpha_i 1_{B_i}$, we define $\int f\, dZ$ as $\sum_i \alpha_i Z(B_i)$. This is well defined, for, if f also has the representation $f = \sum_j \beta_j 1_{D_j}$, then

$\sum_i \alpha_i Z(B_i) = \sum_j \beta_j Z(D_j)$ almost surely. This follows by applying the preceding identity to $\sum_i \alpha_i Z(B_i) - \sum_j \beta_j Z(D_j)$.

The "integral" $\int f\, dZ$ that is now defined on the domain of all simple functions f trivially satisfies (i) and (ii), while (iii) is exactly the identity in the preceding display. The proof is complete upon showing that the map $f \mapsto \int f\, dZ$ can be extended from the domain of simple functions to the domain $L_2(\mu)$, meanwhile retaining the properties (i)–(iii).

We extend the map by continuity. For every $f \in L_2(\mu)$ there exists a sequence of simple functions f_n such that $\int |f_n - f|^2\, d\mu \to 0$. We define $\int f\, d\mu$ as the limit of the sequence $\int f_n\, dZ$. This is well defined. First, the limit exists, because, by the linearity of the integral and the identity,

$$\mathrm{E}\Big|\int f_n\, d\mu - \int f_m\, d\mu|^2 = \int |f_n - f_m|^2\, d\mu,$$

since $f_n - f_m$ is a simple function. Because f_n is a Cauchy sequence in $L_2(\mu)$, the right side converges to zero as $m, n \to \infty$. We conclude that $\int f_n\, dZ$ is a Cauchy sequence in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ and hence it has a limit by the completeness of this space. Second, the definition of $\int f\, dZ$ does not depend on the particular sequence $f_n \to f$ we use. This follows, because given another sequence of simple functions $g_n \to f$, we have $\int |f_n - g_n|^2\, d\mu \to 0$ and hence $\mathrm{E}\big|\int f_n\, dZ - \int g_n\, dZ\big|^2 \to 0$.

We conclude the proof by noting that the properties (i)–(iii) are retained under taking limits. ∎

6.25 EXERCISE. Show that a linear isometry $\Phi: \mathbb{H}_1 \to \mathbb{H}_2$ between two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ retains inner products, i.e. $\langle \Phi(f_1), \Phi(f_2)\rangle_2 = \langle f_1, f_2\rangle_1$. Conclude that $\mathrm{cov}\big(\int f\, dZ, \int g\, dZ\big) = \int f\overline{g}\, d\mu$.

We are now ready to derive the spectral decomposition for a general stationary time series X_t. Let $L_2(X_t: t \in \mathbb{Z})$ be the closed, linear span of the elements of the time series in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ (i.e. the closure of the linear span of the set $\{X_t: t \in \mathbb{Z}\}$).

6.26 Theorem. *For any mean zero stationary time series X_t with spectral distribution F_X there exists a random measure Z with orthogonal increments relative to the measure F_X such that $\big\{Z(B): B \in \mathcal{B}\big\} \subset L_2(X_t: t \in \mathbb{Z})$ and such that $X_t = \int e^{i\lambda t}\, dZ(\lambda)$ almost surely for every $t \in \mathbb{Z}$.*

Proof. By the definition of the spectral measure F_X we have, for every finite collections of complex numbers α_j and integers t_j,

$$\mathrm{E}\Big|\sum \alpha_j X_{t_j}\Big|^2 = \sum_i \sum_j \alpha_i \overline{\alpha}_j \gamma_X(t_i - t_j) = \int \Big|\sum \alpha_j e^{it_j\lambda}\Big|^2\, dF_X(\lambda).$$

Now define a map $\Phi: L_2(F_X) \to L_2(X_t: t \in \mathbb{Z})$ as follows. For f of the form $f = \sum_j \alpha_j e^{it_j\lambda}$ define $\Phi(f) = \sum \alpha_j X_{t_j}$. By the preceding identity this is well defined.

(Check!) Furthermore, Φ is a linear isometry. By the same arguments as in the preceding lemma, it can be extended to a linear isometry on the closure of the space of all functions $\sum_j \alpha_j e^{it_j\lambda}$. By Féjer's theorem from Fourier theory, this closure contains at least all Lipschitz periodic functions. By measure theory this collection is dense in $L_2(F_X)$. Thus the closure is all of $L_2(F_X)$. In particular, it contains all indicator functions 1_B of Borel sets B. Define $Z(B) = \Phi(1_B)$. Because Φ is a linear isometry, it retains inner products and hence

$$\operatorname{cov}\bigl(Z(B_1), Z(B_2)\bigr) = \bigl\langle \Phi(1_{B_1}), \Phi(1_{B_2})\bigr\rangle = \int 1_{B_1} 1_{B_2}\, dF_X.$$

This shows that Z is a random measure with orthogonal increments. By definition

$$\int \sum_j \alpha_j 1_{B_j}\, dZ = \sum_j \alpha_j Z(B_j) = \sum \alpha_j \Phi(1_{B_j}) = \Phi\Bigl(\sum_j \alpha_j 1_{B_j}\Bigr).$$

Thus $\int f\, dZ = \Phi(f)$ for every simple function f. Both sides of this identity are linear isometries when seen as functions of $f \in L_2(F_X)$. Hence the identity extends to all $f \in L_2(F_X)$. In particular, we obtain $\int e^{it\lambda}\, dZ(\lambda) = \Phi(e^{it\lambda}) = X_t$ on choosing $f(\lambda) = e^{it\lambda}$. ■

Thus we have managed to give a precise mathematical formulation to the spectral decomposition

$$X_t = \int_{(-\pi,\pi]} e^{it\lambda}\, dZ(\lambda)$$

of a mean zero stationary time series X_t. The definition may seem a bit involved. An insightful interpretation is obtained by approximation through Riemann sums. Given a partition $-\pi = \lambda_{0,k} < \lambda_{1,k} < \cdots < \lambda_{k,k} = \pi$ and a fixed time $t \in \mathbb{Z}$, consider the function $\lambda \mapsto f_k(\lambda)$ that is piecewise constant, and takes the value $e^{it\lambda_{j,k}}$ on the interval $(\lambda_{j-1,k}, \lambda_{j,k}]$. If the partitions are chosen such that the mesh width of the partitions converges to zero as $k \to \infty$, then $\bigl|f_k(\lambda) - e^{it\lambda}\bigr|$ converges to zero, uniformly in $\lambda \in (-\pi, \pi]$, by the uniform continuity of the function $\lambda \mapsto e^{it\lambda}$, and hence $f_k(\lambda) \to e^{it\lambda}$ in $L_2(F_X)$. Because the stochastic integral $f \mapsto \int f\, dZ$ is linear, we have $\int f_k\, dZ = \sum_j e^{it\lambda_{j,k}} Z(\lambda_{j-1,k}\lambda_{j,k}]$ and because it is an isometry, we find

$$\mathrm{E}\Bigl|X_t - \sum_{j=1}^k e^{it\lambda_{j,k}} Z(\lambda_{j-1,k}, \lambda_{j,k}]\Bigr|^2 = \int \bigl|e^{it\lambda} - f_k(\lambda)\bigr|^2\, dF_X(\lambda) \to 0.$$

Because the intervals $(\lambda_{j-1,k}, \lambda_{j,k}]$ are pairwise disjoint, the random variables $Z_j := Z(\lambda_{j-1,k}, \lambda_{j,k}]$ are uncorrelated, by the defining property of an orthogonal random measure. Thus the time series X_t can be approximated by a time series of the form $\sum_j Z_j e^{it\lambda_j}$, as in the introduction of this section. The spectral measure $F_X(\lambda_{j-1,k}, \lambda_{j,k}]$ of the interval $(\lambda_{j-1,k}, \lambda_{j,k}]$ is the variance of the random weight Z_j in this decomposition.

6.27 Example. If the spectral measure F_X is discrete with support points $\lambda_1, \ldots, \lambda_k$, then the integral on the right in the preceding display (with $\lambda_{j,k} = \lambda_j$) is identically zero. In that case $X_t = \sum_j Z_j e^{i\lambda_j t}$ almost surely for every t. □

6.28 Example. If the time series X_t is Gaussian, then all variables in the linear span of the X_t are normally distributed (possibly degenerate) and hence all variables in $L_2(X_t : t \in \mathbb{Z})$ are normally distributed. In that case the variables $Z(B)$ obtained from the random measure Z of the spectral decomposition of X_t are jointly normally distributed. The zero correlation of two variables $Z(B_1)$ and $Z(B_2)$ for disjoint sets B_1 and B_2 now implies independence of these variables. □

Theorem 6.9 shows how a spectral measure changes under filtering. There is a corresponding result for the spectral decomposition.

6.29 Theorem. *Let X_t be a mean zero, stationary time series with spectral measure F_X and associated random measure Z_X, defined on some probability space $(\Omega, \mathcal{U}, \mathrm{P})$. If $\psi(\lambda) = \sum_j \psi_j e^{-i\lambda j}$ converges in $L_2(F_X)$, then $Y_t = \sum_j \psi_j X_{t-j}$ converges in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ and has spectral measure F_Y and associated random measure Z_Y such that, for every $f \in L_2(F_Y)$,*

$$\int f \, dZ_Y = \int f\psi \, dZ_X.$$

Proof. The series Y_t converges by Theorem 6.9, and the spectral measure F_Y has density $\bigl|\psi(\lambda)\bigr|^2$ relative to F_X. By definition,

$$\int e^{it\lambda} \, dZ_Y(\lambda) = Y_t = \sum_j \psi_j \int e^{i(t-j)\lambda} \, dZ_X(\lambda) = \int e^{it\lambda} \psi(\lambda) \, dZ_X(\lambda),$$

where in the last step changing the order of integration and summation is justified by the convergence of the series $\sum_j \psi_j e^{i(t-j)\lambda}$ in $L_2(F_X)$ and the continuity of the stochastic integral $f \mapsto \int f \, dZ_X$. We conclude that the identity of the theorem is satisfied for every f of the form $f(\lambda) = e^{it\lambda}$. Both sides of the identity are linear in f and isometries on the domain $f \in L_2(F_Y)$. Because the linear span of the functions $\lambda \mapsto e^{it\lambda}$ for $t \in \mathbb{Z}$ is dense in $L_2(F_Y)$, the identity extends to all of $L_2(F_Y)$, by linearity and continuity. ■

6.30 Example (Law of large numbers). An interesting application of the spectral decomposition is the following law of large numbers. If X_t is a mean zero, stationary time series with associated random measure Z_X, then $\overline{X}_n \xrightarrow{\mathrm{P}} Z_X\{0\}$ as $n \to \infty$. In particular, if $F_X\{0\} = 0$, then $\overline{X}_n \xrightarrow{\mathrm{P}} 0$.

To see this, we write

$$\overline{X}_n = \frac{1}{n} \sum_{t=1}^n \int e^{it\lambda} \, dZ_X(\lambda) = \int \frac{e^{i\lambda}(1 - e^{i\lambda n})}{n(1 - e^{i\lambda})} \, dZ_X(\lambda).$$

Here the integrand must be read as 1 if $\lambda = 0$. For all other $\lambda \in (-\pi, \pi]$ the integrand converges to zero as $n \to \infty$. It is bounded by 1 for every λ. Hence the integrand converges in second mean to $1_{\{0\}}$ in $L_2(F_X)$ (and every other $L_2(\mu)$-space for a finite measure μ). By the continuity of the integral $f \mapsto \int f\, dZ_X$, we find that $\overline{X}_n$ converges in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ to $\int 1_{\{0\}}\, dZ_X = Z_X\{0\}$. □

* 6.4 Multivariate Spectra

If spectral analysis of univariate time series' is hard, spectral analysis of multivariate time series is an art. It concerns not only "frequencies present in a single signal", but also "dependencies between signals at given frequencies".

This difficulty concerns the interpretation only: the mathematical theory does not pose new challenges. The covariance function γ_X of a vector-valued times series X_t is matrix-valued. If the series $\sum_{h\in\mathbb{Z}} \bigl\|\gamma_X(h)\bigr\|$ is convergent, then the *spectral density* of the series X_t can be defined by exactly the same formula as before:

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \gamma_X(h) e^{-ih\lambda}.$$

The summation is now understood to be entry-wise, and hence $\lambda \mapsto f_X(\lambda)$ maps the interval $(-\pi, \pi]$ into the set of $(d \times d)$-matrices, for d the dimension of the series X_t. Because the covariance function of the univariate series $a^T X_t$ is given by $\gamma_{a^T X} = a^T \gamma_X \overline{a}$, it follows that, for every $a \in \mathbb{C}^k$,

$$a^T f_X(\lambda) \overline{a} = f_{a^T X}(\lambda).$$

In particular, the matrix $f_X(\lambda)$ is nonnegative-definite, for every λ. From the identity $\gamma_X(-h)^T = \overline{\gamma_X}(h)$ it can also be ascertained that it is Hermitian. The diagonal elements are nonnegative, but the off-diagonal elements of $f_X(\lambda)$ are complex valued, in general.

As in the case of univariate time series, not every vector-valued time series possesses a spectral density, but every such series does possess a spectral distribution. This "distribution" is a matrix-valued, complex measure. A *complex Borel measure* on $(-\pi, \pi]$ is a map $B \mapsto F(B)$ on the Borel sets that can be written as $F = F_1 - F_2 + i(F_3 - F_4)$ for finite Borel measures F_1, F_2, F_3, F_4. If the complex part $F_3 - F_4$ is identically zero, then F is a *signed measure*. The spectral measure F_X of a d-dimensional time series X_t is a $(d \times d)$ matrix whose d^2 entries are complex Borel measures on $(-\pi, \pi]$. The diagonal elements are precisely the spectral measures of the coordinate time series' and hence are ordinary measures, but the off-diagonal measures are typically signed or complex measures. The measure F_X is *Hermitian* in the sense that $F_X(B) = \overline{F_X}(B)$ for every Borel set B.

6.31 Theorem (Herglotz). *For every stationary vector-valued time series X_t there exists a unique Hermitian-matrix-valued complex measure F_X on $(-\pi, \pi]$ such that*

$$\gamma_X(h) = \int_{(-\pi,\pi]} e^{ih\lambda}\, dF_X(\lambda), \qquad h \in \mathbb{Z}.$$

Proof. For every $a \in \mathbb{C}^d$ the time series $a^T X_t$ is univariate and possesses a spectral measure $F_{a^T X}$. By Theorem 6.2, for every $h \in \mathbb{Z}$,

$$a^T \gamma_X(h)\overline{a} = \gamma_{a^T X}(h) = \int_{(-\pi,\pi]} e^{ih\lambda}\, dF_{a^T X}(\lambda).$$

We can express any entry of the matrix $\gamma_X(h)$ as a linear combination of the the quadratic form on the left side, evaluated for different vectors a. One possibility is to write, with e_i the ith unit vector in $\mathbb{C}^d$,

$$\begin{aligned} 2\,\mathrm{Re}\,\gamma_X(h)_{i,j} &= e_i^T \gamma_X(h) e_i + e_j^T \gamma_X(h) e_j - (e_i - e_j)^T \gamma_X(h)(e_i - e_j), \\ 2\,\mathrm{Im}\,\gamma_X(h)_{i,j} &= e_i^T \gamma_X(h) e_i - e_j^T \gamma_X(h) e_j - i(e_i - ie_j)^T \gamma_X(h)(e_i + ie_j). \end{aligned}$$

By expressing the right-hand sides in the spectral matrices $F_{a^T X}$, by using the first display, we obtain representations $\gamma_X(h)_{i,j} = \int e^{ih\lambda}\, dF_{i,j}$ for complex-valued measures $F_{i,j}$, for every (i,j). The matrix-valued complex measure $F = (F_{i,j})$ can be chosen Hermitian-valued.

If F is an Hermitian-matrix-valued complex measure with the representing property, then $a^T F \overline{a}$ must be the spectral measure of the time series $a^T X_t$ and hence is uniquely determined. This determines F. ■

Consider in particular a bivariate time series, written as (X_t, Y_t) for univariate times series X_t and Y_t. The spectral density of (X_t, Y_t), if it exists, is a (2×2)-matrix valued function. The diagonal elements are the spectral densities f_X and f_Y of the univariate series X_t and Y_t. The off-diagonal elements are complex conjugates and thus define one function, say f_{XY} for the $(1,2)$-element of the matrix. The following derived functions are often plotted:

$\mathrm{Re}\, f_{XY}$,	co-spectrum,
$\mathrm{Im}\, f_{XY}$,	quadrature,
$\dfrac{\lvert f_{XY}\rvert^2}{f_X f_Y}$,	coherency,
$\lvert f_{XY}\rvert$,	amplitude,
$\arg f_{XY}$,	phase.

It requires some experience to read the plots of these functions appropriately. The coherency is perhaps the easiest to interprete: it is the “correlation between the series' X and Y at the frequency λ”.

7
ARIMA Processes

For many years ARIMA processes were the work horses of time series analysis, "time series analysis" being almost identical to fitting an appropriate ARIMA process. This important class of time series models are defined through linear relations between the observations and noise factors.

7.1 Backshift Calculus

To simplify notation we define the *backshift operator* B through

$$BX_t = X_{t-1}, \qquad B^k X_t = X_{t-k}.$$

This is viewed as operating on a complete time series X_t, transforming this into a new series by a time shift. Even though we use the word "operator" we shall use B only as a notational device. In particular, $BY_t = Y_{t-1}$ for any other time series Y_t.[♭]

For a given polynomial $\psi(z) = \sum_j \psi_j z^j$ we also abbreviate

$$\psi(B)X_t = \sum_j \psi_j X_{t-j}.$$

If the series on the right is well defined, then we even use this notation for infinite *Laurent series* $\sum_{j=-\infty}^{\infty} \psi_j z^j$. Then $\psi(B)X_t$ is simply a short-hand notation for the (infinite) linear filters that we encountered before. By Lemma 1.28 the time series $\psi(B)X_t$ is certainly well defined if $\sum_j |\psi_j| < \infty$ and $\sup_t \mathrm{E}|X_t| < \infty$, in which case the series converges both almost surely and in mean.

[♭] Be aware of the dangers of this notation. For instance, if $Y_t = X_{-t}$, then $BY_t = Y_{t-1} = X_{-(t-1)}$. This is the intended meaning. We could also argue that $BY_t = BX_{-t} = X_{-t-1}$. This is something else. Such inconsistencies can be avoided by defining B as a true operator, for instance a linear operator acting on the linear span of a given time series, possibly depending on the time series.

If $\sum_j |\psi_j| < \infty$, then the Laurent series $\sum_j \psi_j z^j$ converges absolutely on the unit circle $\bigl\{z \in \mathbb{C}: |z| = 1\bigr\}$ in the complex plane and hence defines a function $\psi(z)$. Given two of such series (or functions) $\psi_1(z) = \sum_j \psi_{1,j} z^j$ and $\psi_2(z) = \sum_j \psi_{2,j} z^j$, the product $\psi(z) = \psi_1(z)\psi_2(z)$ is a well-defined function on (at least) the unit circle. By changing the summation indices this can be written as

$$\psi(z) = \psi_1(z)\psi_2(z) = \sum_j \psi_j z^j, \qquad \psi_k = \sum_j \psi_{1,j}\psi_{2,k-j}.$$

The coefficients ψ_j are called the *convolutions* of the coefficients $\psi_{1,j}$ and $\psi_{2,j}$. Under the condition that $\sum_j |\psi_{i,j}| < \infty$, the Laurent series $\sum_k \psi_k z^k$ converges absolutely at least on the unit circle. In fact $\sum_k |\psi_k| < \infty$.

7.1 EXERCISE. Show that $\sum_k |\psi_k| \le \sum_j |\psi_{1,j}| \sum_j |\psi_{2,j}|$.

Having defined the function $\psi(z)$ and verified that it has an absolutely convergent Laurent series representation on the unit circle, we can now also define the time series $\psi(B)X_t$. The following lemma shows that the convolution formula remains valid if z is replaced by B, at least when applied to stationary time series.

7.2 Lemma. *If both $\sum_j |\psi_{1,j}| < \infty$ and $\sum_j |\psi_{2,j}| < \infty$, then, for every time series X_t with $\sup_t \mathrm{E}|X_t| < \infty$,*

$$\psi(B)X_t = \psi_1(B)\bigl[\psi_2(B)X_t\bigr], \qquad \text{a.s..}$$

Proof. The right side is to be read as $\psi_1(B)Y_t$ for $Y_t = \psi_2(B)X_t$. The variable Y_t is well defined almost surely by Lemma 1.28, because $\sum_j |\psi_{2,j}| < \infty$ and $\sup_t \mathrm{E}|X_t| < \infty$. Furthermore,

$$\sup_t \mathrm{E}|Y_t| = \sup_t \mathrm{E}\Bigl|\sum_j \psi_{2,j} X_{t-j}\Bigr| \le \sum_j |\psi_{2,j}| \sup_t \mathrm{E}|X_t| < \infty.$$

Thus the time series $\psi_1(B)Y_t$ is also well defined by Lemma 1.28. Now

$$\mathrm{E}\sum_i\sum_j |\psi_{1,i}||\psi_{2,j}||X_{t-i-j}| \le \sup_t \mathrm{E}|X_t| \sum_i |\psi_{1,i}| \sum_j |\psi_{2,j}| < \infty.$$

This implies that the double series $\sum_i\sum_j \psi_{1,i}\psi_{2,j}X_{t-i-j}$ converges absolutely, almost surely, and hence unconditionally. The latter means that we may sum the terms in an arbitrary order. In particular, by the change of variables $(i,j) \mapsto (i = l, i + j = k)$,

$$\sum_i \psi_{1,i}\Bigl(\sum_j \psi_{2,j}X_{t-i-j}\Bigr) = \sum_k\Bigl(\sum_l \psi_{1,l}\psi_{2,k-l}\Bigr)X_{t-k}, \qquad \text{a.s..}$$

This is the assertion of the lemma, with $\psi_1(B)\bigl[\psi_2(B)X_t\bigr]$ on the left side. ■

The lemma implies that the "operators" $\psi_1(B)$ and $\psi_2(B)$ commute, and in a sense asserts that the "product" $\psi_1(B)\psi_2(B)X_t$ is associative. Thus from now on we may omit the square brackets in $\psi_1(B)\bigl[\psi_2(B)X_t\bigr]$.

7.3 EXERCISE. Verify that the lemma remains valid for any sequences ψ_1 and ψ_2 with $\sum_j |\psi_{i,j}| < \infty$ and every process X_t such that $\sum_i \sum_j |\psi_{1,i}||\psi_{2,j}||X_{t-i-j}| < \infty$ almost surely. In particular, conclude that $\psi_1(B)\psi_2(B)X_t = (\psi_1\psi_2)(B)X_t$ for any polynomials ψ_1 and ψ_2 and every time series X_t.

7.2 ARMA Processes

Linear regression models attempt to explain a variable by the sum of a linear function of explanatory variables and a noise variable. ARMA processes are a time series version of linear regression, where the explanatory variables are the past values of the time series itself and the added noise is a moving average process.

7.4 Definition. *A time series X_t is an ARMA(p, q)-process if there exist polynomials ϕ and θ of degrees p and q, respectively, and a white noise series Z_t such that $\phi(B)X_t = \theta(B)Z_t$.*

The equation $\phi(B)X_t = \theta(B)Z_t$ is to be understood as "pointwise almost surely" on the underlying probability space: the random variables X_t and Z_t are defined on a probability space $(\Omega, \mathcal{U}, \mathrm{P})$ and satisfy $\phi(B)X_t(\omega) = \theta(B)Z_t(\omega)$ for almost every $\omega \in \Omega$.

The polynomials are often[♯] written in the forms $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$. Then the equation $\phi(B)X_t = \theta(B)Z_t$ takes the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

In other words: the value of the time series X_t at time t is the sum of a linear regression on its own past and of a moving average. An ARMA$(p, 0)$-process is also called an *auto-regressive process* and denoted AR(p); an ARMA$(0, q)$-process is also called a *moving average process* and denoted MA(q). Thus an auto-regressive process is a solution X_t to the equation $\phi(B)X_t = Z_t$, and a moving average process is explicitly given by $X_t = \theta(B)Z_t$.

7.5 EXERCISE. Why is it not a loss of generality to assume $\phi_0 = \theta_0 = 1$?

We next investigate for which pairs of polynomials ϕ and θ there exists a corresponding stationary ARMA-process. For given polynomials ϕ and θ there are always many time series X_t and Z_t satisfying the ARMA equation, but there need not be a stationary series X_t. If there exists a stationary solution, then we are also interested in knowing whether this is uniquely determined by the pair (ϕ, θ) and/or the white noise series Z_t, and in what way it depends on the series Z_t.

[♯] A notable exception is the Splus package. Its makers appear to have overdone the cleverness of including minus-signs in the coefficients of ϕ and have included them in the coefficients of θ also.

7.6 Example. The polynomial $\phi(z) = 1 - \phi z$ leads to the auto-regressive equation $X_t = \phi X_{t-1} + Z_t$. In Example 1.8 we have seen that a stationary solution exists if and only if $|\phi| \neq 1$. □

7.7 EXERCISE. Let arbitrary polynomials ϕ and θ, a white noise sequence Z_t and variables $X_1, \ldots, X_p$ be given. Show that there exists a time series X_t that satisfies the equation $\phi(B)X_t = \theta(B)Z_t$ and coincides with the given $X_1, \ldots, X_p$ at times $1, \ldots, p$. What does this imply about existence of solutions if only the Z_t and the polynomials ϕ and θ are given?

In the following theorem we shall see that a stationary solution to the ARMA-equation exists if the polynomial $z \mapsto \phi(z)$ has no roots on the unit circle $\{z \in \mathbb{C}: |z| = 1\}$. To prove this, we need some facts from complex analysis. The function

$$\psi(z) = \frac{\theta(z)}{\phi(z)}$$

is well defined and analytic on the region $\{z \in \mathbb{C}: \phi(z) \neq 0\}$. If ϕ has no roots on the unit circle $\{z: |z| = 1\}$, then since it has at most p different roots, there is an annulus $\{z: r < |z| < R\}$ with $r < 1 < R$ on which it has no roots. On this annulus ψ is an analytic function, and it has a *Laurent series* representation

$$\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j.$$

This series is uniformly and absolutely convergent on every compact subset of the annulus, and the coefficients ψ_j are uniquely determined by the values of ψ on the annulus. In particular, because the unit circle is inside the annulus, we obtain that $\sum_j |\psi_j| < \infty$.

Then we know that $\psi(B)Z_t$ is a well defined, stationary time series. By the following theorem it is the unique stationary solution to the ARMA-equation. (Here by "solution" we mean a time series that solves the equation up to null sets, and the uniqueness is also up to null sets.)

7.8 Theorem. *Let ϕ and θ be polynomials such that ϕ has no roots on the complex unit circle, and let Z_t be a white noise process. Define $\psi = \theta/\phi$. Then $X_t = \psi(B)Z_t$ is the unique stationary solution to the equation $\phi(B)X_t = \theta(B)Z_t$. It is also the only solution that is bounded in L_1.*

Proof. By the rules of calculus justified by Lemma 7.2, $\phi(B)\psi(B)Z_t = \theta(B)Z_t$, because $\phi(z)\psi(z) = \theta(z)$ on an annulus around the unit circle, $\sum_j |phi_j|$ and $\sum_j |\psi_j|$ are finite and the time series Z_t is bounded in absolute mean. This proves that $\psi(B)Z_t$ is a solution to the ARMA-equation. It is stationary by Lemma 1.28.

Let X_t be an arbitrary solution to the ARMA equation that is bounded in L_1, for instance a stationary solution. The function $\tilde{\phi}(z) = 1/\phi(z)$ is analytic on an annulus around the unit circle and hence possesses a unique Laurent series representation $\tilde{\phi}(z) = \sum_j \tilde{\phi}_j z^j$. Because $\sum_j |\tilde{\phi}_j| < \infty$, the infinite series $\tilde{\phi}(B)Y_t$ is well defined for every stationary time series Y_t by Lemma 1.28. By the calculus of Lemma 7.2 $\tilde{\phi}(B)\phi(B)X_t = X_t$ almost surely, because $\tilde{\phi}(z)\phi(z) = 1$, the filter coefficients are summable and the time series X_t is bounded in absolute mean. Therefore, the equation $\phi(B)X_t = \theta(B)Z_t$ implies, after multiplying by $\tilde{\phi}(B)$, that $X_t = \tilde{\phi}(B)\theta(B)Z_t = \psi(B)Z_t$, again by the calculus of Lemma 7.2, because $\tilde{\phi}(z)\theta(z) = \psi(z)$. This proves that $\psi(B)Z_t$ is the unique stationary solution to the ARMA-equation. ■

7.9 EXERCISE. It is certainly not true that $\psi(B)Z_t$ is the only solution to the ARMA-equation. Can you trace where exactly in the preceding proof we use the required stationarity of the solution? Would you agree that the "calculus" of Lemma 7.2 is perhaps more subtle than it appeared to be at first?

Thus the condition that ϕ has no roots on the unit circle is sufficient for the existence of a stationary solution. It is almost necessary. The only point is that it is really the quotient θ/ϕ that counts, not the function ϕ on its own. If ϕ has a zero on the unit circle of the same or smaller multiplicity as θ, then this quotient is still a nice function. Once this possibility is excluded, there can be no stationary solution if $\phi(z) = 0$ for some z with $|z| = 1$.

7.10 Theorem. *Let ϕ and θ be polynomials such that ϕ has a root on the unit circle that is not a root of θ, and let Z_t be a white noise process. Then there exists no stationary solution X_t to the equation $\phi(B)X_t = \theta(B)Z_t$.*

Proof. Suppose that the contrary is true and let X_t be a stationary solution. Then X_t has a spectral distribution F_X, and hence so does the time series $\phi(B)X_t = \theta(B)Z_t$. By Theorem 6.9 and Example 6.5 we must have

$$\left|\phi(e^{-i\lambda})\right|^2 dF_X(\lambda) = \left|\theta(e^{-i\lambda})\right|^2 \frac{\sigma^2}{2\pi}\, d\lambda.$$

Now suppose that $\phi(e^{-i\lambda_0}) = 0$ and $\theta(e^{-i\lambda_0}) \neq 0$ for some $\lambda_0 \in (-\pi, \pi]$. The preceding display is just an equation between densities of measures and should not be interpreted as being valid for every λ, so we cannot immediately conclude that there is a contradiction. By differentiability of ϕ and continuity of θ there exist positive numbers A and B and a neighbourhood of λ_0 on which both $\left|\phi(e^{-i\lambda})\right| \leq A|\lambda - \lambda_0|$ and $\left|\theta(e^{-i\lambda})\right| \geq B$. Combining this with the preceding display, we see that, for all sufficiently small $\varepsilon > 0$,

$$\int_{\lambda_0-\varepsilon}^{\lambda_0+\varepsilon} A^2|\lambda - \lambda_0|^2\, dF_X(\lambda) \geq \int_{\lambda_0-\varepsilon}^{\lambda_0+\varepsilon} B^2 \frac{\sigma^2}{2\pi}\, d\lambda.$$

The left side is bounded above by $A^2\varepsilon^2 F_X(\lambda_0 - \varepsilon, \lambda_0 + \varepsilon)$, whereas the right side is equal to $B^2\sigma^2\varepsilon/\pi$. This shows that $F_X(\lambda_0 - \varepsilon, \lambda_0 + \varepsilon) \to \infty$ as $\varepsilon \to 0$ and contradicts the fact that F_X is a finite measure. ■

7.11 Example. The AR(1)-equation $X_t = \phi X_{t-1} + Z_t$ corresponds to the polynomial $\phi(z) = 1 - \phi z$. This has root ϕ^{-1}. Therefore a stationary solution exists if and only if

$|\phi^{-1}| \neq 1$. In the latter case, the Laurent series expansion of $\psi(z) = 1/(1-\phi z)$ around the unit circle is given by $\psi(z) = \sum_{j=0}^\infty \phi^j z^j$ for $|\phi| < 1$ and is given by $-\sum_{j=1}^\infty \phi^{-j} z^{-j}$ for $|\phi| > 1$. Consequently, the unique stationary solutions in these cases are given by

$$X_t = \begin{cases} \sum_{j=0}^\infty \phi^j Z_{t-j}, & \text{if } |\phi| < 1, \\ -\sum_{j=1}^\infty \frac{1}{\phi^j} Z_{t+j}, & \text{if } |\phi| > 1. \end{cases}$$

This is in agreement, of course, with Example 1.8. □

7.12 EXERCISE. Investigate the existence of stationary solutions to:
(i) $X_t = \frac{1}{2}X_{t-1} + \frac{1}{2}X_{t-2} + Z_t$;
(ii) $X_t = \frac{1}{2}X_{t-1} + \frac{1}{4}X_{t-2} + Z_t + \frac{1}{2}Z_{t-1} + \frac{1}{4}Z_{t-2}$.

Warning. Some authors require by definition that an ARMA process be stationary. Many authors occasionally forget to say explicitly that they are concerned with a stationary ARMA process. Some authors mistakenly believe that stationarity requires that ϕ has no roots inside the unit circle and may fail to recognize that the ARMA equation does not define a process without some sort of initialization.

If given time series' X_t and Z_t satisfy the ARMA-equation $\phi(B)X_t = \theta(B)Z_t$, then they also satisfy $r(B)\phi(B)X_t = r(B)\theta(B)Z_t$, for any polynomial r. From observed data X_t it is impossible to determine whether (ϕ, θ) or $(r\phi, r\theta)$ are the "right" polynomials. To avoid this problem of indeterminacy, we assume from now on that the ARMA-model is always written in its simplest form. This is when ϕ and θ do not have common factors (are relatively prime in the algebraic sense), or equivalently, when ϕ and θ do not have common (complex) roots. Then, in view of the preceding theorems, a stationary solution X_t to the ARMA-equation exists if and only if ϕ has no roots on the unit circle, and this is uniquely given by

$$X_t = \psi(B)Z_t = \sum_j \psi_j Z_{t-j}, \qquad \psi = \frac{\theta}{\phi}.$$

7.13 Definition. *An ARMA-process X_t is called causal if, in the preceding representation, the filter is causal: i.e. $\psi_j = 0$ for every $j < 0$.*

Thus a causal ARMA-process X_t depends on present and past values $Z_t, Z_{t-1}, \ldots$ of the noise sequence only. Intuitively, this is a desirable situation, if time is really time and Z_t is really attached to time t. We come back to this in Section 7.6.

A mathematically equivalent definition of causality is that the function $\psi(z)$ is analytic in a neighbourhood of the unit disc $\{z \in \mathbb{C}: |z| \leq 1\}$. This follows, because the Laurent series $\sum_{j=-\infty}^\infty \psi_j z^j$ is analytic inside the unit disc if and only if the negative powers of z do not occur. Still another description of causality is that all roots of ϕ are outside the unit circle, because only then is the function $\psi = \theta/\phi$ analytic on the unit disc.

The proof of Theorem 7.8 does not use that Z_t is a white noise process, but only that the series Z_t is bounded in L_1. Therefore, the same arguments can be used to invert the ARMA-equation in the other direction. If θ has no roots on the unit circle and X_t is stationary, then $\phi(B)X_t = \theta(B)Z_t$ implies that

$$Z_t = \pi(B)X_t = \sum_j \pi_j X_{t-j}, \qquad \pi = \frac{\phi}{\theta}.$$

7.14 Definition. *An ARMA-process X_t is called invertible if, in the preceding representation, the filter is causal: i.e. $\pi_j = 0$ for every $j < 0$.*

Equivalent mathematical definitions are that $\pi(z)$ is an analytic function on the unit disc or that θ has all its roots outside the unit circle. In the definition of invertibility we implicitly assume that θ has no roots *on* the unit circle. The general situation is more technical and is discussed in the next section.

* 7.3 Invertibility

In this section we discuss the proper definition of invertibility in the case that θ has roots on the unit circle. The intended meaning of "invertibility" is that every Z_t can be written as a linear function of the X_s that are prior or simultaneous to t. Two reasonable ways to make this precise are:

(i) $Z_t = \sum_{j=0}^\infty \pi_j X_{t-j}$ for a sequence π_j such that $\sum_{j=0}^\infty |\pi_j| < \infty$.

(ii) Z_t is contained in the closed linear span of $X_t, X_{t-1}, X_{t-2}, \ldots$ in $L_2(\Omega, \mathcal{U}, \mathrm{P})$.

In both cases we require that X_t depends linearly on the prior X_s, but the second requirement is weaker. It turns out that if X_t is an ARMA process relative to Z_t and (i) holds, then the polynomial θ cannot have roots on the unit circle. In that case the definition of invertibility given in the preceding section is appropriate (and equivalent to (i)). However, the requirement (ii) does not exclude the possibility that θ has zeros on the unit circle. An ARMA process is invertible in the sense of (ii) as soon as θ does not have roots *inside* the unit circle.

7.15 Lemma. *Let X_t be a stationary ARMA process satisfying $\phi(B)X_t = \theta(B)Z_t$ for polynomials ϕ and θ that are relatively prime.*

(i) Then $Z_t = \sum_{j=0}^\infty \pi_j X_{t-j}$ for a sequence π_j such that $\sum_{j=0}^\infty |\pi_j| < \infty$ if and only if θ has no roots on or inside the unit circle.

(ii) If θ has no roots inside the unit circle, then Z_t is contained in the closed linear span of $X_t, X_{t-1}, X_{t-2}, \ldots$.

Proof. (i). If θ has no roots on or inside the unit circle, then the ARMA process is invertible by the arguments given previously. We must argue the other direction. If Z_t has the given given reprentation, then consideration of the spectral measures gives

$$\frac{\sigma^2}{2\pi}\, d\lambda = dF_Z(\lambda) = \left|\pi(e^{-i\lambda})\right|^2 dF_X(\lambda) = \left|\pi(e^{-i\lambda})\right|^2 \frac{\left|\theta(e^{-i\lambda})\right|^2}{\left|\phi(e^{-i\lambda})\right|^2} \frac{\sigma^2}{2\pi} d\lambda.$$

Hence $\left|\pi(e^{-i\lambda})\theta(e^{-i\lambda})\right| = \left|\phi(e^{-i\lambda})\right|$ Lebesgue almost everywhere. If $\sum_j |\pi_j| < \infty$, then the function $\lambda \mapsto \pi(e^{-i\lambda})$ is continuous, as are the functions ϕ and θ, and hence this equality must hold for every λ. Since $\phi(z)$ has no roots on the unit circle, nor can $\theta(z)$.

(ii). Suppose that ζ^{-1} is a zero of θ, so that $|\zeta| \leq 1$ and $\theta(z) = (1 - \zeta z)\theta_1(z)$ for a polynomial θ_1 of degree $q-1$. Define $Y_t = \phi(B)X_t$ and $V_t = \theta_1(B)Z_t$, whence $Y_t = V_t - \zeta V_{t-1}$. It follows that

$$\sum_{j=0}^{k-1} \zeta^j Y_{t-j} = \sum_{j=0}^{k-1} \zeta^j (V_{t-j} - \zeta V_{t-j-1}) = V_t - \zeta^k V_{t-k}.$$

If $|\zeta| < 1$, then the right side converges to V_t in quadratic mean as $k \to \infty$ and hence it follows that V_t is contained in the closed linear span of $Y_t, Y_{t-1}, \ldots$, which is clearly contained in the closed linear span of $X_t, X_{t-1}, \ldots$, because $Y_t = \phi(B)X_t$. If $q = 1$, then V_t and Z_t are equal up to a constant and the proof is complete. If $q > 1$, then we repeat the argument with θ_1 instead of θ and V_t in the place of Y_t and we shall be finished after finitely many recursions.

If $|\zeta| = 1$, then the right side of the preceding display still converges to V_t as $k \to \infty$, but only in the weak sense that $\mathrm{E}(V_t - \zeta^k V_{t-k})W \to \mathrm{E}V_t W$ for every square integrable variable W. This implies that V_t is in the weak closure of $\mathrm{lin}\,(Y_t, Y_{t-1}, \ldots)$, but this is equal to the strong closure by an application of the Hahn-Banach theorem. Thus we arrive at the same conclusion.

To see the weak convergence, note first that the projection of W onto the closed linear span of $\{Z_t : t \in \mathbb{Z}\}$ is given by $\sum_j \psi_j Z_j$ for some sequence ψ_j with $\sum_j |\psi_j|^2 < \infty$. Because $V_{t-k} \in \mathrm{lin}\,(Z_s : s \leq t-k)$, we have $|\mathrm{E}V_{t-k}W| = |\sum_j \psi_j \mathrm{E}V_{t-k}Z_j| \leq \sum_{j \leq t-k} |\psi_j| \,\mathrm{sd}\, V_0 \,\mathrm{sd}\, Z_0 \to 0$ as $k \to \infty$. ■

7.16 Example. The moving average $X_t = Z_t - Z_{t-1}$ is invertible in the sense of (ii), but not in the sense of (i). The moving average $X_t = Z_t - 1.01 Z_{t-1}$ is not invertible.

Thus $X_t = Z_t - Z_{t-1}$ implies that $Z_t \in \overline{\mathrm{lin}}\,(X_t, X_{t-1}, \ldots)$. An unexpected phenomenon is that it is also true that Z_t is contained in $\overline{\mathrm{lin}}\,(X_{t+1}, X_{t+2}, \ldots)$. This follows by time reversal: define $U_t = X_{-t+1}$ and $W_t = -Z_{-t}$ and apply the preceding to the processes $U_t = W_t - W_{t-1}$. Thus it appears that the "opposite" of invertibility is true as well! □

7.17 EXERCISE. Suppose that $X_t = \theta(B)Z_t$ for a polynomial θ of degree q that has all its roots on the unit circle. Show that $Z_t \in \overline{\mathrm{lin}}\,(X_{t+q}, X_{t+q+1}, \ldots)$. [As in (ii) of the preceding proof, it follows that $V_t = \zeta^{-k}(V_{t+k} - \sum_{j=0}^{k-1} \zeta^j X_{t+k+j})$. Here the first term on the right side converges weakly to zero as $k \to \infty$.]

7.4 Prediction

As to be expected from their definitions, causality and invertibility are important for calculating predictions for ARMA processes. For a causal and invertible stationary ARMA process X_t satisfying $\phi(B)X_t = \theta(B)Z_t$ we have

$$\begin{aligned} X_t &\in \overline{\text{lin}}\,(Z_t, Z_{t-1}, \ldots), &\qquad &\text{(causality)},\\ Z_t &\in \overline{\text{lin}}\,(X_t, X_{t-1}, \ldots), &\qquad &\text{(invertibility)}. \end{aligned}$$

Here $\overline{\text{lin}}$, the *closed linear span*, is the operation of first forming all (finite) linear combinations and next taking the metric closure in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ of this linear span. Since Z_t is a white noise process, the variable Z_{t+1} is orthogonal to the linear span of $Z_t, Z_{t-1}, \ldots$. By the continuity of the inner product it is then also orthogonal to the closed linear span of $Z_t, Z_{t-1}, \ldots$ and hence, under causality, it is orthogonal to X_s for every $s \leq t$. This shows that the variable Z_{t+1} is totally (linearly) unpredictable at time t given the observations $X_1, \ldots, X_t$. This is often interpreted in the sense that the variable Z_t is an "external noise variable" that is generated at time t independently of the history of the system before time t.

7.18 EXERCISE. The preceding argument gives that Z_{t+1} is uncorrelated with the system variables $X_t, X_{t-1}, \ldots$ of the past. Show that if the variables Z_t are independent, then Z_{t+1} is independent of the system up to time t, not just uncorrelated.

This general discussion readily gives the structure of the best linear predictor for causal auto-regressive stationary processes. Suppose that

$$X_{t+1} = \phi_1 X_t + \cdots + \phi_p X_{t+1-p} + Z_{t+1}.$$

If $t \geq p$, then $X_t, \ldots, X_{t-p+1}$ are perfectly predictable based on the past variables $X_1, \ldots, X_t$; by themselves. If the series is causal, then Z_{t+1} is totally unpredictable (in the sense that its best prediction is zero), in view of the preceding discussion. Since a best linear predictor is a projection and projections are linear maps, the best linear predictor of X_{t+1} based on $X_1, \ldots, X_t$ is given by

$$\Pi_t X_{t+1} = \phi_1 X_1 + \cdots + \phi_p X_{t+1-p}, \qquad (t \geq p).$$

We should be able to obtain this result also from the prediction equations (2.1) and the explicit form of the auto-covariance function, but that calculation would be more complicated.

7.19 EXERCISE. Find a formula for the best linear predictor of X_{t+2} based on $X_1, \ldots, X_t$, if $t - p \geq 1$.

For moving average and general ARMA processes the situation is more complicated. Here a similar argument works only for computing the best linear predictor $\Pi_{-\infty,t} X_{t+1}$

based on the infinite past $X_t, X_{t-1}, \ldots$ down to time $-\infty$. Assume that X_t is a causal and invertible stationary ARMA process satisfying

$$X_{t+1} = \phi_1 X_t + \cdots + \phi_p X_{t+1-p} + Z_{t+1} + \theta_1 Z_t + \cdots + \theta_q Z_{t+1-q}.$$

By causality the variable Z_{t+1} is completely unpredictable. By invertibility the variable Z_s is perfectly predictable based on $X_s, X_{s-1}, \ldots$ and hence is perfectly predictable based on $X_t, X_{t-1}, \ldots$ for every $s \le t$. Therefore,

$$\Pi_{-\infty,t} X_{t+1} = \phi_1 X_t + \cdots + \phi_p X_{t+1-p} + \theta_1 Z_t + \cdots + \theta_q Z_{t+1-q}.$$

The practical importance of this formula is small, because we never observe the complete past. However, if we observe a long series $X_1, \ldots, X_t$, then the "distant past" $X_0, X_{-1}, \ldots$ will not give much additional information over the "recent past" $X_t, \ldots, X_1$, and $\Pi_{-\infty,t} X_{t+1}$ and $\Pi_t X_{t+1}$ will be close.

* **7.20** EXERCISE. Suppose that ϕ and θ do not have zeros on or inside the unit circle. Show that $\mathrm{E}|\Pi_{-\infty,t} X_{t+1} - \Pi_t X_{t+1}|^2 \to 0$ as $t \to \infty$. [Express Z_t as $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$; show that $|\pi_j|$ decreases exponentially fast. The difference $|\Pi_{-\infty,t} X_{t+1} - \Pi_t X_{t+1}|$ is bounded above by $\sum_{j=1}^{q} |\theta_j| |Z_{t+1-j} - \Pi_t Z_{t+1-j}|$.]

We conclude by remarking that for causal stationary auto-regressive processes the square *prediction error* $\mathrm{E}|X_{t+1} - \Pi_t X_{t+1}|^2$ is equal to $\mathrm{E}Z_{t+1}^2$; for general stationary ARMA-processes this is approximately true for large t; in both cases $\mathrm{E}|X_{t+1} - \Pi_{-\infty,t} X_{t+1}|^2 = \mathrm{E}Z_{t+1}^2$.

7.5 Auto Correlation and Spectrum

In this section we discuss several methods to express the auto-covariance function of a stationary ARMA-process in its parameters and obtain an expression for the spectral density.

The latter is immediate from the representation $X_t = \psi(B) Z_t$ and Theorem 6.9.

7.21 Theorem. *The stationary ARMA process satisfying $\phi(B) X_t = \theta(B) Z_t$ possesses a spectral density given by*

$$f_X(\lambda) = \left| \frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})} \right|^2 \frac{\sigma^2}{2\pi}.$$

Finding a simple expression for the auto-covariance function is harder, except for the special case of moving average processes, for which the auto-covariances can be expressed in the parameters $\theta_1, \ldots, \theta_q$ by a direct computation (cf. Example 1.6 and Lemma 1.28). The auto-covariances of a general stationary ARMA process can be solved from a system

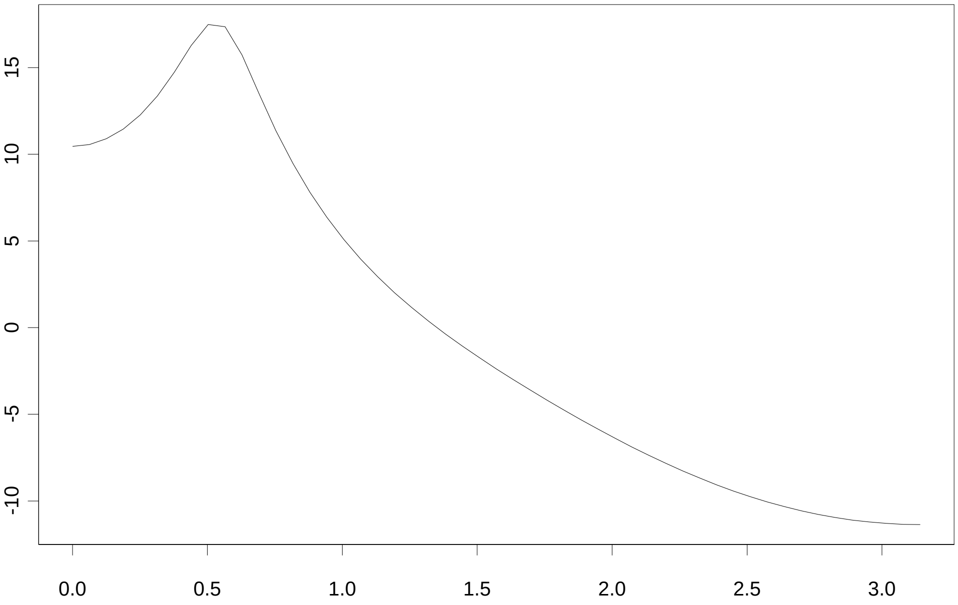

Figure 7.1. Spectral density of the AR series satisfying $X_t - 1.5X_{t-1} + 0.9X_{t-2} - 0.2X_{t-3} + 0.1X_{t-9} = Z_t$. (Vertical axis in *decibels*, i.e. it gives the logarithm of the spectrum.)

of equations. In view of Lemma 1.28(iii), the equation $\phi(B)X_t = \theta(B)Z_t$ leads to the identities, with $\phi(z) = \sum_j \tilde{\phi}_j z^j$ and $\theta(z) = \sum_j \theta_j z^j$,

$$\sum_l \Bigl(\sum_j \tilde{\phi}_j \tilde{\phi}_{j+l-h}\Bigr)\gamma_X(l) = \sigma^2 \sum_j \theta_j \theta_{j+h}, \qquad h \in \mathbb{Z}.$$

In principle this system of equations can be solved for the values $\gamma_X(l)$.

An alternative method to compute the auto-covariance function is to write $X_t = \psi(B)Z_t$ for $\psi = \theta/\phi$, whence, by Lemma 1.28(iii),

$$\gamma_X(h) = \sigma^2 \sum_j \psi_j \psi_{j+h}.$$

This requires the computation of the coefficients ψ_j, which can be expressed in the coefficients of ϕ and θ by comparing coefficients in the power series equation $\phi(z)\psi(z) = \theta(z)$.

7.22 Example. For the AR(1) series $X_t = \phi X_{t-1} + Z_t$ with $|\phi| < 1$ we obtain $\psi(z) = (1-\phi z)^{-1} = \sum_{j=0}^\infty \phi^j z^j$. Therefore, $\gamma_X(h) = \sigma^2 \sum_{j=0}^\infty \phi^j \phi^{j+h} = \sigma^2 \phi^h/(1-\phi^2)$ for $h \geq 0$. □

7.23 EXERCISE. Find $\gamma_X(h)$ for the stationary ARMA$(1,1)$ series $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$ with $|\phi| < 1$.

* **7.24** EXERCISE. Show that the auto-covariance function of a stationary ARMA process decreases exponentially. Give an estimate of the constant in the exponent in terms of the distance of the zeros of ϕ to the unit circle.

A third method to express the auto-covariance function in the coefficients of the polynomials ϕ and θ uses the spectral representation

$$\gamma_X(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f_X(\lambda)\,d\lambda = \frac{\sigma^2}{2\pi i}\int_{|z|=1} z^{h-1}\frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})}\,dz.$$

The second integral is a contour integral along the positively oriented unit circle in the complex plane. We have assumed that the coefficients of the polynomials ϕ and θ are real, so that $\phi(z)\phi(z^{-1}) = \phi(z)\overline{\phi(z)} = |\phi(z)|^2$ for every z in the unit circle, and similarly for θ. The next step is to evaluate the contour integral with the help of the residue theorem from complex function theory. The poles of the integrand are contained in the set consisting of the zeros v_i and their inverses v_i^{-1} of ϕ and possibly the point 0. The auto-covariance function can be written as a function of the residues at these points.

7.25 Example (ARMA$(1,1)$). Consider the stationary ARMA$(1,1)$ series $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$ with $0 < |\phi| < 1$. The corresponding function $\phi(z)\phi(z^{-1})$ has zeros of multiplicity 1 at the points ϕ^{-1} and ϕ. Both points yield a pole of first order for the integrand in the contour integral. The number ϕ^{-1} is outside the unit circle, so we only need to compute the residue at the second point. The function $\theta(z^{-1})/\phi(z^{-1}) = (z+\theta)/(z-\phi)$ is analytic in a neighbourhood of 0 and hence does not contribute other poles, but the term z^{h-1} may contribute a pole at 0. For $h \geq 1$ the integrand has poles at ϕ and ϕ^{-1} only and hence

$$\gamma_X(h) = \sigma^2 \operatorname*{res}_{z=\phi} z^{h-1}\frac{(1+\theta z)(1+\theta z^{-1})}{(1-\phi z)(1-\phi z^{-1})} = \sigma^2\phi^h\frac{(1+\theta\phi)(1+\theta/\phi)}{1-\phi^2}.$$

For $h = 0$ the integrand has an additional pole at $z = 0$ and the integral evaluates to the sum of the residues at the two poles at $z = 0$ and $z = \phi$. The first residue is equal to $-\theta/\phi$. Thus

$$\gamma_X(0) = \sigma^2\Big(\frac{(1+\theta\phi)(1+\theta/\phi)}{1-\phi^2} - \frac{\theta}{\phi}\Big).$$

The values of $\gamma_X(h)$ for $h < 0$ follow by symmetry. □

7.26 EXERCISE. Find the auto-covariance function for a MA(q) process by using the residue theorem. (This is not easier than the direct derivation, but perhaps instructive.)

We do not present an additional method to compute the partial auto-correlation function of an ARMA process. However, we make the important observation that for a causal AR(p) process the partial auto-correlations $\alpha_X(h)$ of lags $h > p$ vanish. This follows by combining Lemma 2.33 and the expression for the best linear predictor found in the preceding section.

7.6 Existence of Causal and Invertible Solutions

In practice we never observe the white noise process Z_t in the definition of an ARMA process. The Z_t are "hidden variables" whose existence is hypothesized to explain the observed series X_t. From this point of view our earlier question of existence of a stationary solution to the ARMA equation is perhaps not the right question, as it took the sequence Z_t as given. In this section we turn this question around and consider an ARMA(p,q) process X_t as given. Then we shall see that there are at least 2^{p+q} white noise processes Z_t such that $\phi(B)X_t = \theta(B)Z_t$ for certain polynomials ϕ and θ of degrees p and q, respectively. (These polynomials depend on the choice of Z_t and hence are not necessarily the ones that are initially given.) Thus the white noise process Z_t is far from being uniquely determined by the observed series X_t. On the other hand, among the multitude of solutions, only one choice yields a representation of X_t as a stationary ARMA process that is both causal and invertible.

7.27 Theorem. *For every stationary ARMA process X_t satisfying $\phi(B)X_t = \theta(B)Z_t$ for polynomials ϕ and θ such that θ has no roots on the unit circle, there exist polynomials ϕ^* and θ^* of the same or smaller degrees as ϕ and θ that have all roots outside the unit disc and a white noise process Z_t^* such that $\phi^*(B)X_t = \theta^*(B)Z_t$ almost surely for every $t \in \mathbb{Z}$.*

Proof. The existence of the stationary ARMA process X_t and our implicit assumption that ϕ and θ are relatively prime imply that ϕ has no roots on the unit circle. Thus all roots of ϕ and θ are either inside or outside the unit circle. We shall show that we can move the roots inside the unit circle to roots outside the unit circle by a filtering procedure. Suppose that

$$\phi(z) = -\phi_p(z - v_1)\cdots(z - v_p), \qquad \theta(z) = \theta_q(z - w_1)\cdots(z - w_q).$$

Consider any zero z_i of ϕ or θ. If $|z_i| < 1$, then we replace the term $(z - z_i)$ in the above products by the term $(1 - \overline{z}_i z)$; otherwise we keep $(z - z_i)$. For $z_i = 0$, this means that we drop the term $z - z_i$ and the degree of the polynomial decreases; otherwise, the degree remains the same. We apply this procedure to all zeros v_i and w_i and denote the resulting polynomials by ϕ^* and θ^*. Because $0 < |z_i| < 1$ implies that $|\overline{z}_i^{-1}| > 1$, the polynomials ϕ^* and θ^* have all zeros outside the unit circle. We have that

$$\frac{\theta(z)}{\phi(z)} = \frac{\theta^*(z)}{\phi^*(z)}\kappa(z), \qquad \kappa(z) = \prod_{i:|v_i|<1} \frac{1 - \overline{v}_i z}{z - v_i} \prod_{i:|w_i|<1} \frac{z - w_i}{1 - \overline{w}_i z}.$$

Because $X_t = (\theta/\phi)(B)Z_t$ and we want that $X_t = (\theta^*/\phi^*)(B)Z_t^*$, we define the process Z_t^* by $Z_t^* = \kappa(B)Z_t$. This is to be understood in the sense that we expand $\kappa(z)$ in its Laurent series $\kappa(z) = \sum_j \kappa_j z^j$ and apply the corresponding linear filter to Z_t.

By construction we now have that $\phi^*(B)X_t = \theta^*(B)Z_t$. If $|z| = 1$, then $|1 - \overline{z}_i z| = |z - z_i|$. In view of the definition of κ this implies that $\bigl|\kappa(z)\bigr| = 1$ for every z on the unit circle and hence the spectral density of Z_t^* satisfies

$$f_{Z^*}(\lambda) = \bigl|\kappa(e^{-i\lambda})\bigr|^2 f_Z(\lambda) = 1 \cdot \frac{\sigma^2}{2\pi}.$$

This shows that Z_t^* is a white noise process, as desired. ■

As are many results in time series analysis, the preceding theorem is a result on second moments only. Even if Z_t is an i.i.d. sequence, then the theorem does not guarantee that Z_t^* is an i.i.d. sequence as well. Only first and second moments are preserved by the filtering procedure in the proof, in general. Nevertheless, the theorem is often interpreted as implying that not much is lost by assuming a-priori that ϕ and θ have all their roots outside the unit circle.

7.28 EXERCISE. Suppose that the time series Z_t is Gaussian. Show that the series Z_t^* constructed in the preceding proof is Gaussian and hence i.i.d..

* 7.7 Stability

Let ϕ and θ be polynomials, with ϕ having no roots on the unit circle. Given initial values $X_1, \ldots, X_p$ and a process Z_t, we can recursively define a solution to the ARMA equation $\phi(B)X_t = \theta(B)Z_t$ by

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \theta(B)Z_t, \qquad t > p, \\ X_{t-p} &= \phi_p^{-1}\bigl(X_t - \phi_1 X_{t-1} - \cdots - \phi_{p-1}X_{t-p+1} - \theta(B)Z_t\bigr), \qquad t - p < 1. \end{aligned} \tag{7.1}$$

In view of Theorem 7.8 the resulting process X_t can only be bounded in L_2 if the initial values $X_1, \ldots, X_p$ are chosen randomly according to the stationary distribution. In particular, the process X_t obtained from deterministic initial values must necessarily be unbounded (on the full time scale $t \in \mathbb{Z}$).

In this section we show that in the causal situation, when ϕ has no zeros on the unit disc, the process X_t tends to stationarity as $t \to \infty$, given arbitrary initial values. Hence in this case the unboundedness occurs as $t \to -\infty$. This is another reason to prefer the case that ϕ has no roots on the unit disc: in this case the effect of initializing the process wears off as time goes by.

Let Z_t be a given white noise process and let $(X_1, \ldots, X_p)$ and $(\tilde X_1, \ldots, \tilde X_p)$ be two possible sets of initial values, consisting of random variables defined on the same probability space.

7.29 Theorem. *Let ϕ and θ be polynomials such that ϕ has no roots on the unit disc. Let X_t and $\tilde X_t$ be the ARMA processes as in defined (7.1) with initial values $(X_1,\ldots,X_p)$ and $(\tilde X_1,\ldots,\tilde X_p)$, respectively. Then $X_t-\tilde X_t\to 0$ almost surely as $t\to\infty$.*

7.30 Corollary. *Let ϕ and θ be polynomials such that ϕ has no roots on the unit disc. If X_t is an ARMA process with arbitrary initial values, then the vector $(X_t,\ldots,X_{t+k})$ converges in distribution to the distribution of the stationary solution to the ARMA equation, as $t\to\infty$, for every fixed k.*

Proofs. For the corollary we take $(\tilde X_1,\ldots,\tilde X_p)$ equal to the values of the stationary solution. Then we can conclude that the difference between X_t and the stationary solution converges almost surely to zero and hence the difference between the distributions tends to zero.

For the proof of the theorem we write the ARMA relationship in the "state space form", for $t>p$,

$$\begin{pmatrix} X_t\\ X_{t-1}\\ \vdots\\ X_{t-p+1}\end{pmatrix}=\begin{pmatrix}\phi_1&\phi_2&\cdots&\phi_{p-1}&\phi_p\\ 1&0&\cdots&0&0\\ \vdots&\vdots&&\vdots&\vdots\\ 0&0&\cdots&1&0\end{pmatrix}\begin{pmatrix}X_{t-1}\\ X_{t-2}\\ \vdots\\ X_{t-p}\end{pmatrix}+\begin{pmatrix}\theta(B)Z_t\\ 0\\ \vdots\\ 0\end{pmatrix}.$$

Denote this system by $Y_t=\Phi Y_{t-1}+B_t$. By some algebra it can be shown that

$$\det(\Phi-zI)=(-1)^p z^p\phi(z^{-1}),\qquad z\neq 0.$$

Thus the assumption that ϕ has no roots on the unit disc implies that the eigenvalues of Φ are all inside the unit circle. In other words, the spectral radius of Φ, the maximum of the moduli of the eigenvalues, is strictly less than 1. Because the sequence $\|\Phi^n\|^{1/n}$ converges to the spectral radius as $n\to\infty$, we can conclude that $\|\Phi^n\|^{1/n}$ is strictly less than 1 for all sufficiently large n, and hence $\|\Phi^n\|\to 0$ as $n\to\infty$.

If $\tilde Y_t$ relates to $\tilde X_t$ as Y_t relates to X_t, then $Y_t-\tilde Y_t=\Phi^{t-p}(Y_p-\tilde Y_p)\to 0$ almost surely as $t\to\infty$. ∎

7.31 EXERCISE. Suppose that $\phi(z)$ has no zeros on the unit circle and at least one zero inside the unit circle. Show that there exist initial values $(X_1,\ldots,X_p)$ such that the resulting process X_t is not bounded in probability as $t\to\infty$. [Let $\tilde X_t$ be the stationary solution and let X_t be the solution given initial values $(X_1,\ldots,X_p)$. Then, with notation as in the preceding proof, $Y_t-\tilde Y_t=\Phi^{t-p}(Y_p-\tilde Y_p)$. Choose an appropriate deterministic vector for $Y_p-\tilde Y_p$.]

7.8 ARIMA Processes

In Chapter 1 differencing is introduced as a method to transform a nonstationary time series in a stationary one. This method is particularly attractive in combination with ARMA modelling: in the notation of the present chapter the differencing filters can be written as

$$\nabla X_t = (1 - B)X_t, \qquad \nabla^d X_t = (1 - B)^d X_t, \qquad \nabla_k X_t = (1 - B^k)X_t.$$

Thus the differencing filters ∇, ∇^d and ∇_k correspond to applying $\phi(B)$ for the polynomials $\phi(z) = 1-z$, $\phi(z) = (1-z)^d$ and $\phi(z) = (1-z^k)$, respectively. These polynomials have in common that all their roots are on the complex unit circle. Thus they were "forbidden" polynomials in our preceding discussion of ARMA processes. In fact, by Theorem 7.10, for the three given polynomials ϕ the series $Y_t = \phi(B)X_t$ cannot be a stationary ARMA process if X_t is already a stationary ARMA process relative to polynomials without zeros on the unit circle.

On the other hand, $Y_t = \phi(B)X_t$ can well be a stationary ARMA process if X_t is a non-stationary time series. Thus we can use polynomials ϕ with roots on the unit circle to extend the domain of ARMA modelling to nonstationary time series.

7.32 Definition. *A time series X_t is an ARIMA(p, d, q) process if $\nabla^d X_t$ is a stationary ARMA(p, q) process.*

In other words, the time series X_t is an ARIMA(p, d, q) process if there exist polynomials ϕ and θ of degrees p and q and a white noise series Z_t such that the time series $\nabla^d X_t$ is stationary and $\phi(B)\nabla^d X_t = \theta(B)Z_t$ almost surely. The additional "I" in ARIMA is for "integrated". If we view taking differences ∇^d as differentiating, then the definition requires that a derivative of X_t is a stationary ARMA process, whence X_t itself is an "integrated ARMA process".

The following definition goes a step further.

7.33 Definition. *A time series X_t is a SARIMA$(p, d, q)(P, D, Q, per)$ process if there exist polynomials ϕ, θ, Φ and Θ of degrees p, q, P and Q and a white noise series Z_t such that the time series $\nabla_{per}^D \nabla^d X_t$ is stationary and $\Phi(B^{per})\phi(B)\nabla_{per}^D \nabla^d X_t = \Theta(B^{per})\theta(B)Z_t$ almost surely.*

The "S" in SARIMA is short for "seasonal". The idea of a seasonal model is that we might only want to use certain powers B^{per} of the backshift operator in our model, because the series is thought to have a certain period. Including the terms $\Phi(B^{per})$ and $\Theta(B^{per})$ does not make the model more general (as these terms could be subsumed in $\phi(B)$ and $\theta(B)$), but reflects our a-priori idea that certain coefficients in the polynomials are zero. This a-priori knowledge will be important when estimating the coefficients from an observed time series.

Modelling an observed time series by an ARIMA, or SARIMA, model has become popular through an influential book by Box and Jenkins. The unified filtering paradigm of a "Box-Jenkins analysis" is indeed attractive. The popularity is probably also due to the compelling manner in which Box and Jenkins explain the reader how he or she must set up the analysis, going through a fixed number of steps. They thus provide the data-analyst with a clear algorithm to carry out an analysis that is intrinsically difficult. It is obvious that the results of such an analysis will not always be good, but an alternative is less obvious.

7.34 EXERCISE. Plot the spectral densities of the following time series:
(i) $X_t = Z_t + 0.9Z_{t-1}$;
(ii) $X_t = Z_t - 0.9Z_{t-1}$;
(iii) $X_t - 0.7X_{t-1} = Z_t$;
(iv) $X_t + 0.7X_{t-1} = Z_t$;
(v) $X_t - 1.5X_{t-1} + 0.9X_{t-2} - 0.2X_{t-3} + 0.1X_{t-9} = Z_t$.

7.35 EXERCISE. Simulate a series of length 200 according to the model $X_t - 1.3X_{t-1} + 0.7X_{t-2} = Z_t + 0.7Z_{t-1}$. Plot the sample auto-correlation and sample partial auto-correlation functions.

* 7.9 VARMA Processes

A *VARMA process* is a vector-valued ARMA process. Given matrices Φ_j and Θ_j and a white noise sequence Z_t of dimension d, a VARMA(p,q) process satisfies the relationship

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + Z_t + \Theta_1 Z_{t-1} + \cdots + \Theta_q Z_{t-q}.$$

The theory for VARMA process closely resembles the theory for ARMA processes. The role of the polynomials ϕ and θ is taken over by the matrix-valued polynomials

$$\Phi(z) = 1 - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p,$$
$$\Theta(z) = 1 + \Theta_1 z + \Theta_2 z^2 + \cdots + \Theta_q z^q.$$

These identities and sums are to be interpreted entry-wise and hence define $(d \times d)$-matrices with entries that are polynomials in $z \in \mathbb{C}$.

Instead of looking at zeros of polynomials we must now look at the values of z for which the matrices $\Phi(z)$ and $\Theta(z)$ are singular. Equivalently, we must look at the zeros of the complex functions $z \mapsto \det \Phi(z)$ and $z \mapsto \det \Theta(z)$. Apart from this difference, the conditions for existence of a stationary solution, causality and invertibility are the same.

7.36 Theorem. *If the matrix-valued polynomial $\Phi(z)$ is invertible for every z in the unit circle, then there exists a unique stationary solution X_t to the VARMA equations. If the matrix-valued polynomial $\Phi(z)$ is invertible for every z on the unit disc, then this can be written in the form $X_t = \sum_{j=0}^\infty \Psi_j Z_{t-j}$ for matrices Ψ_j with $\sum_{j=0}^\infty \|\Psi_j\| < \infty$. If, moreover, the polynomial $\Theta(z)$ is invertible for every z on the unit disc, then we also have that $Z_t = \sum_{j=0}^\infty \Pi_j Z_{t-j}$ for matrices Π_j with $\sum_{j=0}^\infty \|\Pi_j\| < \infty$.*

The norm $\|\cdot\|$ in the preceding may be any matrix norm. The proof of this theorem is the same as the proofs of the corresponding results in the one-dimensional case, in view of the following observations.

A series of the type $\sum_{j=-\infty}^{\infty}\Psi_j Z_{t-j}$ for matrices Ψ_j with $\sum_{j=0}^{\infty}\|\Psi_j\| < \infty$ and a vector-valued process Z_t with $\sup_t \mathrm{E}\|Z_t\| < \infty$ converges almost surely and in mean. Furthermore, the analogue of Lemma 7.2 is true.

The functions $z \mapsto \det \Phi(z)$ and $z \mapsto \det \Theta(z)$ are polynomials. Hence if they are nonzero on the unit circle, then they are nonzero on an open annulus containing the unit circle, and the matrices $\Phi(z)$ and $\Theta(z)$ are invertible for every z in this annulus. Cramer's rule, which expresses the solution of a system of linear equations in determinants, shows that the entries of the inverse matrices $\Phi(z)^{-1}$ and $\Theta(z)^{-1}$ are quotients of polynomials. The denominators are the determinants $\det \Phi(z)$ and $\det \Theta(z)$ and hence are nonzero in a neighbourhood of the unit circle. These matrices may thus be expanded in Laurent series'

$$\Phi(z)^{-1} = \Bigl(\sum_{j=-\infty}^{\infty} (\Psi_j)_{k,l} z^j\Bigr)_{k,l=1,\ldots,d} = \sum_{j=-\infty}^{\infty} \Psi_j z^j,$$

where the Ψ_j are matrices such that $\sum_{j=-\infty}^{\infty}\|\Psi_j\| < \infty$, and similarly for $\Theta(z)^{-1}$.

8 GARCH Processes

White noise processes are basic building blocks for time series models, but can also be of interest on their own. A sequence of i.i.d. variables is an example of a white noise sequence, but is not of great interest as a time series. On the other hand, many financial time series appear to be realizations of white noise series, but are not well described by i.i.d. sequences. This is possible because the white noise property only concerns the second moments of the process, so that the variables of a white noise process may possess many types of dependence. GARCH processes are a class of white noise sequences that have been found useful for modelling certain financial time series.

Figure 8.1 shows a realization of a GARCH process. The striking feature are the "bursts of activity", which alternate with quiet periods of the series. Here the frequency of the movements of the series is constant over time, but their amplitude changes, alternating between "volatile" periods (large amplitude) and quiet periods. This phenomenon is referred to as *volatility clustering.* A look at the auto-correlation function of the realization, Figure 8.2, shows that the alternations are not reflected in the second moments of the series: the series can be modelled as white noise, at least in the sense that the correlations are zero.

Recall that a *white noise* series is any stationary time series whose auto-covariances at nonzero lags vanish. We shall speak of a *heteroscedastic white noise* if the auto-covariances at nonzero lags vanish, but the variances are possibly time-dependent. A related concept is that of a martingale difference series. Recall that a *filtration* $\mathcal{F}_t$ is a nondecreasing collection of σ-fields $\cdots \subset \mathcal{F}_{-1} \subset \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$. A *martingale difference series* relative to the filtration $\mathcal{F}_t$ is a time series X_t such that X_t is $\mathcal{F}_t$-measurable and $\mathrm{E}(X_t|\,\mathcal{F}_{t-1}) = 0$ almost surely for every t. The latter includes the assumption that $\mathrm{E}|X_t| < \infty$, so that the conditional expectation is well defined.

Any martingale difference series X_t with finite second moments is a (possibly heteroscedastic) white noise series. Indeed, the equality $\mathrm{E}(X_t|\,\mathcal{F}_{t-1}) = 0$ is equivalent to X_t being orthogonal to all random variables $Y \in \mathcal{F}_{t-1}$, and this includes the variables $X_s \in \mathcal{F}_s \subset \mathcal{F}_{t-1}$, for every $s < t$, so that $\mathrm{E}X_tX_s = 0$ for every $s < t$. Conversely, not every white noise is a martingale difference series (relative to a natural filtration). This

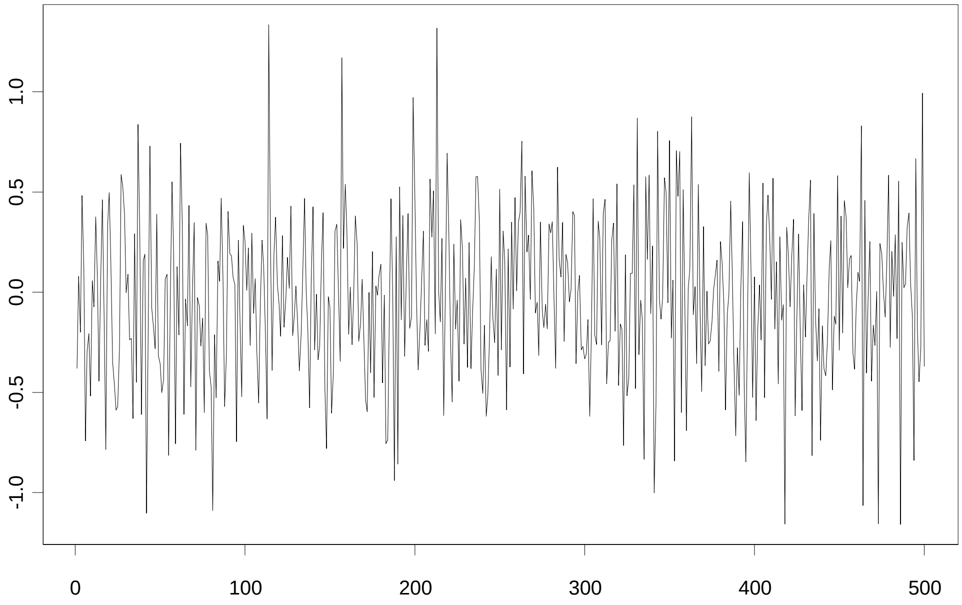

Figure 8.1. Realization of length 500 of the stationary Garch(1, 1) process with $\alpha = 0.15$, $\phi_1 = 0.4$, $\theta_1 = 0.4$ and standard normal variables Z_t.

is because $\mathrm{E}(X|\,Y) = 0$ implies that X is orthogonal to all measurable functions of Y, not orthogonal just to linear functions.

8.1 EXERCISE. If X_t is a martingale difference series, show that $\mathrm{E}(X_{t+k}X_{t+l}|\,\mathcal{F}_t) = 0$ almost surely for every $k \neq l > 0$. Thus "future variables are uncorrelated given the present". Find a white noise series which lacks this property.

A martingale difference sequence has zero first moment given the past. A natural step for further modelling is to postulate a specific form of the conditional second moment. GARCH models are examples, and in that sense are again concerned only with first and second moments of the time series, albeit conditional moments. They turn out to capture many features of observed time series, in particular those in finance, that are not captured by ARMA processes. Besides volatility clustering thes "stylized facts" include leptokurtic (i.e. heavy) tailed marginal distributions and nonzero auto-correlations for the process X_t^2 of squares.

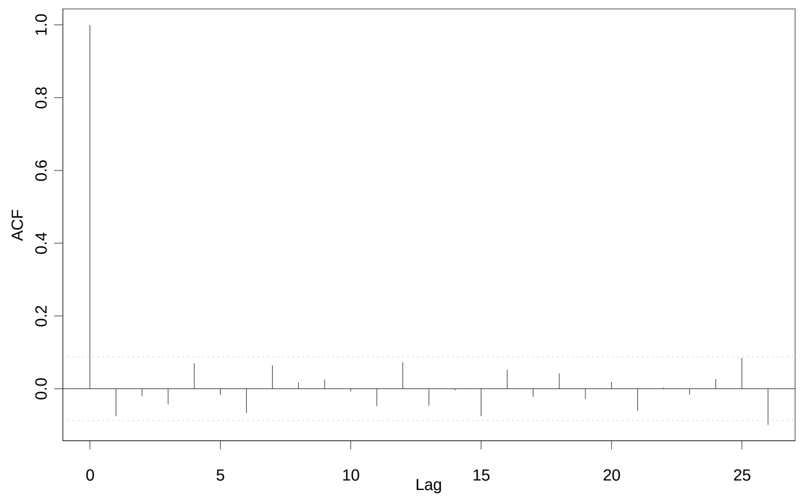

Figure 8.2. Sample auto-covariance function of the time series in Figure 8.1.

8.1 Linear GARCH

There are many types of GARCH processes, of which we discuss a selection in the following sections. Linear GARCH processes were the earliest GARCH processes to be studied, and may be viewed as *the* GARCH processes.

8.2 Definition. *A GARCH (p,q) process is a martingale difference sequence X_t, relative to a given filtration $\mathcal{F}_t$, whose conditional variances $\sigma_t^2 = \mathrm{E}(X_t^2 \,|\, \mathcal{F}_{t-1})$ satisfy, for every $t \in \mathbb{Z}$ and given constants $\alpha, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$,*

$$\sigma_t^2 = \alpha + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_p \sigma_{t-p}^2 + \theta_1 X_{t-1}^2 + \cdots + \theta_q X_{t-q}^2, \qquad \text{a.s..} \tag{8.1}$$

With the usual convention that $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = \theta_1 z + \cdots + \theta_q z^q$, the equation for the conditional variance $\sigma_t^2 = \mathrm{var}(X_t \,|\, \mathcal{F}_{t-1})$ can be abbreviated to

$$\phi(B)\sigma_t^2 = \alpha + \theta(B) X_t^2.$$

Note that the polynomial θ is assumed to have zero intercept. If the coefficients $\phi_1, \ldots, \phi_p$ all vanish, then σ_t^2 is modelled as a linear function of $X_{t-1}^2, \ldots, X_{t-q}^2$. This is called an ARCH (q) model, from “auto-regressive conditional heteroscedastic”. The additional G of GARCH is for the nondescript “generalized”.

If $\sigma_t > 0$, as we shall assume, then we can define $Z_t = X_t/\sigma_t$. The martingale difference property of $X_t = \sigma_t Z_t$ and the definition of σ_t^2 as the conditional variance imply

$$\mathrm{E}(Z_t \,|\, \mathcal{F}_{t-1}) = 0, \qquad \mathrm{E}(Z_t^2 \,|\, \mathcal{F}_{t-1}) = 1. \tag{8.2}$$

Conversely, given an adapted process Z_t satisfying this display (a "scaled martingale difference process") and a process σ_t that is $\mathcal{F}_{t-1}$-measurable we can define a process X_t by $X_t = \sigma_t Z_t$. Then σ_t is the conditional variance of X_t and the process X_t is a GARCH process if (8.1) is valid. It is then often added as an assumption that the variables Z_t are i.i.d. and that Z_t is independent of $\mathcal{F}_{t-1}$. This is equivalent to assuming that the conditional law of the variables $Z_t = X_t/\sigma_t$ given $\mathcal{F}_{t-1}$ is a given distribution, for instance a standard normal distribution. In order to satisfy (8.2) this distribution must have a finite second moment, but this is not strictly necessary for all of the following. The "conditional variances" in Definition 8.2 may be understood in the general sense that does not require that the variances $\mathrm{E}X_t^2$ are finite.

If we substitute $\sigma_t^2 = X_t^2 - W_t$ in (8.1), then we find after rearranging the terms,

$$\begin{aligned} X_t^2 = \alpha + (\phi_1 + \theta_1)X_{t-1}^2 + \cdots + (\phi_r + \theta_r)X_{t-r}^2 \\ + W_t - \phi_1 W_{t-1} - \cdots - \phi_p W_{t-p}, \end{aligned} \tag{8.3}$$

where $r = p \vee q$ and the sequences $\phi_1, \ldots, \phi_p$ or $\theta_1, \ldots, \theta_q$ are padded with zeros to increase their lengths to r, if necessary. We can abbreviate this to

$$(\phi - \theta)(B)X_t^2 = \alpha + \phi(B)W_t, \qquad W_t = X_t^2 - \mathrm{E}(X_t^2 |\, \mathcal{F}_{t-1}).$$

This is the characterizing equation for an ARMA(r, r) process X_t^2 relative to the noise process W_t. The variable $W_t = X_t^2 - \sigma_t^2$ is the prediction error when predicting X_t^2 by its conditional expectation $\sigma_t^2 = \mathrm{E}(X_t^2 |\, \mathcal{F}_{t-1})$ and hence W_t is orthogonal to $\mathcal{F}_{t-1}$. Thus W_t is a martingale difference series and a-fortiori a white noise sequence if its second moments exist and are independent of t. Under this conditions the time series of squares X_t^2 is an ARMA process in the sense of Definition 7.4. A warning against applying the results on ARMA processes unthinkingly to the process X_t^2, for instance to infer results on existence given certain parameter values, is that W_t is defined itself in terms of the process X_t and therefore does not have a simple interpretation as a noise process that drives the process X_t^2. This limits the importance of equation (8.3), although it is can useful to compute the auto-covariance function of the process of squares. (See e.g. Example 8.7.)

* **8.3** EXERCISE. Suppose that X_t and W_t are martingale diffference series' relative to a given filtration such that $\phi(B)X_t^2 = \theta(B)W_t$ for polynomials ϕ and θ of degrees p and q. Show that X_t is a GARCH process. Does strict stationarity of the time series X_t^2 or W_t imply strict stationarity of the time series X_t?

8.4 EXERCISE. Write σ_t^2 as the solution to an ARMA$(p \vee q, q-1)$ equation by substituting $X_t^2 = \sigma_t^2 + W_t$ in (8.3).

Alternatively, we can substitute $X_t = \sigma_t Z_t$ in the GARCH relation (8.1) and obtain

$$\sigma_t^2 = \alpha + (\phi_1 + \theta_1 Z_{t-1}^2)\sigma_{t-1}^2 + \cdots + (\phi_r + \theta_r Z_{t-r}^2)\sigma_{t-r}^2. \tag{8.4}$$

This exhibits the process σ_t^2 as an auto-regressive process "with random coefficients and deterministic innovations". This relation is useful to construct GARCH processes.

In the following theorem we consider given a martingale difference sequence Z_t as in (8.2), defined on a fixed probability space. Next we construct a GARCH process such that $X_t = \sigma_t Z_t$ by first defining the process of squares σ_t^2 in terms of the Z_t. If the coefficients α, ϕ_j, θ_j are nonnegative we obtain a stationary solution if the polynomial $1 - \sum_{j=1}^r (\phi_j + \theta_j) z^j$ possesses no zeros on the unit disc. Under the condition that the coefficients are nonnegative, the second is equivalent to $\sum_j (\phi_j + \theta_j) < 1$.

8.5 EXERCISE. If $p_1, \ldots, p_r$ are nonnegative real numbers, then the polynomial $p(z) = 1 - \sum_{j=1}^r p_j z^j$ possesses no roots on the unit disc if and only if $p(1) > 0$. [Use that $p(0) = 1 > 0$; furthermore, use the triangle inequality.]

8.6 Theorem. *Let $\alpha > 0$, let $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ be nonnegative numbers, and let Z_t be a martingale difference sequence satisfying (8.2) relative to an arbitrary filtration $\mathcal{F}_t$.*

(i) There exists a stationary GARCH process X_t such that $X_t = \sigma_t Z_t$, where $\sigma_t^2 = \mathrm{E}(X_t^2 |\, \mathcal{F}_{t-1})$, if and only if $\sum_j (\phi_j + \theta_j) < 1$.

(ii) This process is unique among the GARCH processes X_t with $X_t = \sigma_t Z_t$ that are bounded in L_2.

(iii) This process satisfies $\sigma(X_t, X_{t-1}, \ldots) = \sigma(Z_t, Z_{t-1}, \ldots)$ for every t, and $\sigma_t^2 = \mathrm{E}(X_t^2 |\, \mathcal{F}_{t-1})$ is $\sigma(X_{t-1}, X_{t-2}, \ldots)$-measurable.

Proof. Assume first that $\sum_j (\phi_j + \theta_j) < 1$. Furthermore, assume that there exists a GARCH process X_t that is bounded in L_2. Then the conditional variance $\sigma_t^2 = \mathrm{E}(X_t^2 |\, \mathcal{F}_{t-1})$ is bounded in L_1 and satisfies, by (8.4),

$$\begin{pmatrix} \sigma_t^2 \\ \sigma_{t-1}^2 \\ \vdots \\ \sigma_{t-r+1}^2 \end{pmatrix} = \begin{pmatrix} \phi_1 + \theta_1 Z_{t-1}^2 & \cdots & \phi_{r-1} + \theta_{r-1} Z_{t-2}^2 & \phi_r + \theta_r Z_{t-r}^2 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \sigma_{t-1}^2 \\ \sigma_{t-2}^2 \\ \vdots \\ \sigma_{t-r}^2 \end{pmatrix} + \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Write this system as $Y_t = A_t Y_{t-1} + b$ and set $A = \mathrm{E} A_t$. With some effort it can be shown that

$$\det(A - zI) = (-1)^r \bigl(z^r - \sum_{j=1}^r (\phi_j + \theta_j) z^{r-j}\bigr).$$

If $\sum_j (\phi_j + \theta_j) < 1$, then the polynomial on the right has all its roots inside the unit circle. (See Exercise 8.5.) Equivalently, the spectral radius (the maximum of the moduli of the eigenvalues) of the operator A is strictly smaller than 1. This implies that $\|A^n\|$ is smaller than 1 for all sufficiently large n and hence $\sum_{n=0}^\infty \|A^n\| < \infty$.

Iterating the equation $Y_t = A_t Y_{t-1} + b$ we find that

$$(8.5) \qquad Y_t = b + A_t b + A_t A_{t-1} b + \cdots + A_t A_{t-1} \cdots A_{t-n+1} b + A_t A_{t-1} \cdots A_{t-n} Y_{t-n-1}.$$

Because $Z_t = X_t / \sigma_t$ is $\mathcal{F}_t$-measurable and $\mathrm{E}(Z_t^2 |\, \mathcal{F}_{t-1}) = 1$ for every t, we have that $\mathrm{E} Z_{t_1}^2 \cdots Z_{t_k}^2 = 1$, for every $t_1 < t_2 < \cdots < t_k$. By some matrix algebra it can be seen that this implies that

$$\mathrm{E} A_t A_{t-1} \cdots A_{t-n} = A^{n+1} \to 0, \qquad n \to \infty.$$

Because the matrices A_t possess nonnegative entries, this implies that the sequence $A_tA_{t-1}\cdots A_{t-n}$ converges to zero in probability. If the process X_t is bounded in L_2, then, in view of its definition, the process Y_t is bounded in L_1. We conclude that $A_tA_{t-1}\cdots A_{t-n}Y_{t-n-1} \to 0$ in probability as $n \to \infty$. Combining this with the expression for Y_t in (8.5), we see that

$$Y_t = b + \sum_{j=1}^{\infty} A_tA_{t-1}\cdots A_{t-j+1}b. \tag{8.6}$$

This implies that $\mathrm{E}Y_t = \sum_{j=0}^{\infty} A^j b$, whence $\mathrm{E}Y_t$ and hence $\mathrm{E}X_t^2$ are independent of t.

Because the matrices A_t are measurable functions of $(Z_{t-1}, Z_{t-2}, \ldots)$, the variable Y_t is a measurable transformation of these variables as well, and hence the variable $X_t = \sigma_t Z_t$ is a measurable transformation of $(Z_t, Z_{t-1}, \ldots)$.

The process $W_t = X_t^2 - \sigma_t^2$ is bounded in L_1 and satisfies the ARMA relation $(\phi - \theta)(B)X_t^2 = \alpha + \phi(B)W_t$ as in (8.3). Because ϕ has no roots on the unit disc, this relation is invertible, whence $W_t = (1/\phi)(B)\bigl((\phi-\theta)(B)X_t^2 - \alpha\bigr)$ is a measurable transformation of $X_t^2, X_{t-1}^2, \ldots$. We conclude that $\sigma_t^2 = W_t + X_t^2$ and hence $Z_t = X_t/\sigma_t$ are $\sigma(X_t, X_{t-1}, \ldots)$-measurable. Since σ_t^2 is $\sigma(Z_{t-1}, Z_{t-2}, \ldots)$-measurable by the preceding paragraph, it follows that it is $\sigma(X_{t-1}, X_{t-2}, \ldots)$-measurable.

We have proved that any GARCH process X_t that is bounded in L_2 defines a conditional variance process σ_t^2 and corresponding process Y_t that satisfies (8.6). Furthermore, we have proved (iii) for this process.

We next construct a GARCH process X_t by reversing the definitions, still assuming that $\sum_j(\phi_j + \theta_j) < 1$. We define matrices A_t in terms of the process Z_t as before. The series on the right of (8.6) converges in L_1 and hence defines a process Y_t. Simple algebra shows that this satisfies $Y_t = A_tY_{t-1} + b$ for every t. All coordinates of Y_t are nonnegative and $\sigma(Z_{t-1}, Z_{t-2}, \ldots)$-measurable.

Given the processes (Z_t, Y_t) we define processes (X_t, σ_t) by

$$\sigma_t = \sqrt{Y_{t,1}}, \qquad X_t = \sigma_t Z_t.$$

Because σ_t is $\sigma(Z_{t-1}, Z_{t-2}, \ldots) \subset \mathcal{F}_{t-1}$-measurable, we have that $\mathrm{E}(X_t|\,\mathcal{F}_{t-1}) = \sigma_t\mathrm{E}(Z_t|\,\mathcal{F}_{t-1}) = 0$ and $\mathrm{E}(X_t^2|\,\mathcal{F}_{t-1}) = \sigma_t^2\mathrm{E}(Z_t^2|\,\mathcal{F}_{t-1}) = \sigma_t^2$. That σ_t^2 satisfies (8.1) is a consequence of the relations $Y_t = A_tY_{t-1} + B_t$, whose first line expresses σ_t^2 into σ_{t-1}^2 and $Y_{t-1,2}, \ldots, Y_{t-1,r}$, and whose other lines permit to reexpress the variable $Y_{t-1,k}$ for $k > 1$ as σ_{t-k}^2 by recursive use of the relations $Y_{t,k} = Y_{t-1,k-1}$, and the definitions $Y_{t-k,1} = \sigma_{t-k}^2$.

This concludes the proof that there exists a stationary solution as soon as $\sum_j(\theta_j + \theta_j) < 1$. Finally, we show that this inequality is necessary. If X_t is a stationary solution, then Y_t in (8.5) is integrable. Taking the expectation of left and right of this equation for $t = 0$ and remembering that all terms are nonnegative, we see that $\sum_{j=0}^{n} A^j b \le \mathrm{E}Y_0$, for every n. This implies that $A^n b \to 0$ as $n \to \infty$, or, equivalently $A^n e_1 \to 0$, where e_i

is the ith unit vector. In view of the definition of A we see, recursively, that

$$\begin{aligned} A^n e_r &= A^{n-1}(\phi_r+\theta_r)e_1 \to 0,\\ A^n e_{r-1} &= A^{n-1}\bigl((\phi_{r-1}+\theta_{r-1})e_1+e_r\bigr) \to 0,\\ &\vdots\\ A^n e_2 &= A^{n-1}\bigl((\phi_2+\theta_2)e_1+e_3\bigr) \to 0. \end{aligned}$$

Therefore, the sequence A^n converges to zero. This can only happen if none of its eigenvalues is on or outside the unit disc. Equivalently, it is necessary that the polynomial $1-\sum_{j=1}^r(\phi_j+\theta_j)z^j$ possesses no roots on or inside the unit disc. ■

Volatility clustering is one of the *stylized facts* of financial time series, and it is captured by GARCH processes: large absolute values of a GARCH series at times $t-1,\ldots,t-q$ lead, through the GARCH equation, to a large conditional variance σ_t^2 at time t, and hence the value $X_t=\sigma_t Z_t$ of the time series at time t tends to be large.

A second stylized fact are the leptokurtic tails of the marginal distribution of a typical financial time series. A distribution on $\mathbb{R}$ is called *leptokurtic* if it has fat tails, for instance fatter than normal tails. A quantitative measure of "fatness" of the tails of the distribution of a random variable X is the *kurtosis* defined as $\overline{\kappa}_4(X)=\mathrm{E}(X-\mathrm{E}X)^4/(\operatorname{var}X)^2$. It is equal to 3 for a normally distributed variable. If $X_t=\sigma_t Z_t$, where σ_t is $\mathcal{F}_{t-1}$-measurable and Z_t is independent of $\mathcal{F}_{t-1}$ with mean zero and variance 1, then

$$\mathrm{E}X_t^4=\mathrm{E}\sigma_t^4\mathrm{E}Z_t^4=\overline{\kappa}_4(Z_t)\mathrm{E}\bigl(\mathrm{E}(X_t^2|\,\mathcal{F}_{t-1})\bigr)^2\ge\overline{\kappa}_4(Z_t)\bigl(\mathrm{E}\mathrm{E}(X_t^2|\,\mathcal{F}_{t-1})\bigr)^2=\overline{\kappa}_4(Z_t)(\mathrm{E}X_t^2)^2.$$

Dividing the left and right sides by $(\mathrm{E}X_t^2)^2$, we see that $\overline{\kappa}_4(X_t)\ge\overline{\kappa}_4(Z_t)$. The difference can be substantial if the variance of the random variable $\mathrm{E}(X_t^2|\,\mathcal{F}_{t-1})$ is large. In fact, taking the difference of the left and right sides of the preceding display yields

$$\kappa_4(X_t)=\kappa_4(Z_t)\Bigl(1+\frac{\operatorname{var}\mathrm{E}(X_t^2|\,\mathcal{F}_{t-1})}{(\mathrm{E}X_t^2)^2}\Bigr).$$

It follows that the GARCH structure is also able to capture some of the observed leptokurtosis of financial time series.

If we use a Gaussian process Z_t, then the kurtosis of the observed series X_t is always bigger than 3. It has been observed that this usually does not go far enough in explaining "excess kurtosis" over the normal distribution. The use of one of Student's t-distributions can often improve the fit of a GARCH process substantially.

A third stylized fact observed in financial time series are positive auto-correlations for the sequence of squares X_t^2. The auto-correlation function of the squares of a GARCH series will exist under appropriate additional conditions on the coefficients and the driving noise process Z_t. The ARMA relation (8.3) for the square process X_t^2 may be used to compute this function, using formulas for the auto-correlation function of an ARMA process. Here we must not forget that the process W_t in (8.3) is defined through X_t and hence its variance depends on the parameters in the GARCH relation.

8.7 Example (GARCH$(1,1)$). The conditional variances of a GARCH(1,1) process satisfy $\sigma_t^2 = \alpha + \phi\sigma_{t-1}^2 + \theta X_{t-1}^2$. If we assume the process X_t to be stationary, then $\mathrm{E}\sigma_t^2 = \mathrm{E}X_t^2$ is independent of t. Taking the expectation across the GARCH equation and rearranging then immediately gives

$$\mathrm{E}\sigma_t^2 = \mathrm{E}X_t^2 = \frac{\alpha}{1-\phi-\theta}.$$

To compute the auto-correlation function of the time series of squares X_t^2, we employ (8.3), which reveals this process as an ARMA(1,1) process with the auto-regressive and moving average polynomials given as $1-(\phi+\theta)z$ and $1-\phi z$, respectively. The calculations in Example 7.25 yield that

$$\gamma_{X^2}(h) = \tau^2(\phi+\theta)^h\frac{(1-\phi(\phi+\theta))(1-\phi/(\phi+\theta))}{1-(\phi+\theta)^2}, \qquad h>0,$$
$$\gamma_{X^2}(0) = \tau^2\Bigl(\frac{(1-\phi(\phi+\theta))(1-\phi/(\phi+\theta))}{1-(\phi+\theta)^2} + \frac{\phi}{\phi+\theta}\Bigr).$$

Here τ^2 is the variance of the process $W_t = X_t^2 - \mathrm{E}(X_t^2|\,\mathcal{F}_{t-1})$, which is also dependent on the parameters θ and ϕ. By squaring the GARCH equation we find

$$\sigma_t^4 = \alpha^2 + \phi^2\sigma_{t-1}^4 + \theta^2 X_{t-1}^4 + 2\alpha\phi\sigma_{t-1}^2 + 2\alpha\theta X_{t-1}^2 + 2\phi\theta\sigma_{t-1}^2X_{t-1}^2.$$

If Z_t is independent of $\mathcal{F}_{t-1}$, then $\mathrm{E}\sigma_t^2X_t^2 = \mathrm{E}\sigma_t^4$ and $\mathrm{E}X_t^4 = \kappa_4(Z_t)\mathrm{E}\sigma_t^4$. If we assume, moreover, that the moments exists and are independent of t, then we can take the expectation across the preceding display and rearrange to find that

$$\mathrm{E}\sigma_t^4(1-\phi^2-2\phi\theta-\kappa_4(Z)\theta^2) = \alpha^2 + (2\alpha\phi+2\alpha\theta)\mathrm{E}\sigma_t^2.$$

Together with the formulas obtained previously, this gives the variance of $W_t = X_t^2 - \mathrm{E}(X_t^2|\,\mathcal{F}_{t-1})$, since $\mathrm{E}W_t = 0$ and $\mathrm{E}W_t^2 = \mathrm{E}X_t^4 - \mathrm{E}\sigma_t^4$, by the Pythagorean identity for projections. □

8.8 EXERCISE. Find the auto-covariance function of the process σ_t^2 for a GARCH$(1,1)$ process.

8.9 EXERCISE. Find an expression for the kurtosis of the marginal distribution in a stationary GARCH$(1,1)$ process as in the preceding example. Can this be made arbitrarily large?

The condition that $\sum_j(\phi_j+\theta_j)<1$ is necessary for existence of a GARCH process with bounded second moments, but stronger than necessary if we are interested in a strictly stationary solution to the GARCH equations with possibly infinite second moments. We can see this from the proof of Theorem 8.6, where the GARCH process is defined from the series in (8.6). If this series converges in an almost sure sense, then a strictly stationary GARCH process exists. The series involves products of random matrices A_t; its convergence depends on the value of their *top Lyapounov exponent*, defined by

$$\gamma = \inf_{n\in\mathbb{N}} \frac{1}{n}\mathrm{E}\log\|A_{-1}A_{-2}\cdots A_{-n}\|.$$

Here $\|\cdot\|$ may be any matrix norm (all matrix norms being equivalent). If the process Z_t is ergodic, for instance i.i.d., then we can apply Kingman's subergodic theorem (e.g. Dudley (1987, Theorem 10.7.1)) to the process $\log\|A_{-1}A_{-2}\cdots A_{-n}\|$ to see that

$$\frac{1}{n}\log\|A_{-1}A_{-2}\cdots A_{-n}\| \to \gamma, \qquad \text{a.s..}$$

This implies that the sequence of matrices $A_{-1}A_{-2}\cdots A_{-n}$ converges to zero almost surely as soon as $\gamma < 0$. The convergence is then exponentially fast and the series in (8.6) will converge.

Thus sufficient conditions for the existence of strictly stationary solutions to the GARCH equations can be given in terms of the top Lyapounov exponent of the random matrices A_t. This exponent is in general difficult to compute explicitly, but it can easily be estimated numerically for a given sequence Z_t.

To obtain conditions that are both sufficient and necessary the preceding proof must be adapted somewhat. The following theorem is in terms of the top Lyapounov exponent of the matrices

$$(8.7)\qquad A_t = \begin{pmatrix} \phi_1+\theta_1 Z_{t-1}^2 & \phi_2 & \cdots & \phi_{p-1} & \phi_p & \theta_2 & \cdots & \theta_{q-1} & \theta_q \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ Z_{t-1}^2 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

These matrices have the advantage of being independent and identically distributed if the process Z_t is i.i.d.. They are motivated by the equation obtained by substituting $X_{t-1} = \sigma_{t-1}Z_{t-1}$ in the GARCH equation (8.1), leaving $X_{t-2},\ldots,X_{t-q}$ untouched:

$$\sigma_t^2 = \alpha + (\phi_1+\theta_1 Z_{t-1}^2)\sigma_{t-1}^2 + \phi_2\sigma_{t-1}^2 + \cdots + \phi_p\sigma_{t-p}^2 + \theta_2 X_{t-2}^2 + \cdots + \theta_q X_{t-q}^2.$$

This equation gives rise to the system of equations $Y_t = A_t Y_{t-1} + b$ for the random vectors $Y_t = (\sigma_t^2,\ldots,\sigma_{t-p+1}^2, X_{t-1}^2,\ldots,X_{t-q+1}^2)^T$ and the vector b equal to α times the first unit vector.

8.10 Theorem. *Let $\alpha > 0$, let $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ be nonnegative numbers, and let Z_t be an i.i.d. sequence with mean zero and unit variance. There exists a strictly stationary GARCH process X_t such that $X_t = \sigma_t Z_t$, where $\sigma_t^2 = \mathrm{E}(X_t^2 | \mathcal{F}_{t-1})$ and $\mathcal{F}_t = \sigma(Z_t, Z_{t-1}, \ldots)$, if and only if the top Lyapounov coefficient of the random matrices A_t given by (8.7) is strictly negative. For this process $\sigma(X_t, X_{t-1}, \ldots) = \sigma(Z_t, Z_{t-1}, \ldots)$.*

Proof. Let $b = \alpha e_1$, where e_i is the ith unit vector in $\mathbb{R}^{p+q-1}$. If γ' is strictly larger than the top Lyapounov exponent γ, then $\|A_t A_{t-1} \cdots A_{t-n+1}\| < e^{\gamma' n}$, eventually as $n \to \infty$, almost surely, and hence, eventually,

$$\big\|A_t A_{t-1} \cdots A_{t-n+1} b\big\| < e^{\gamma' n} \|b\|.$$

If $\gamma < 0$, then we may choose $\gamma' < 0$, and hence $\sum_n \|A_t A_{t-1} \cdots A_{t-n+1} b\| < \infty$ almost surely. Then the series on the right side of (8.6), but with the matrix A_t defined as in (8.7), converges almost surely and defines a process Y_t. We can then define processes σ_t and X_t by setting $\sigma_t = \sqrt{Y_{t,1}}$ and $X_t = \sigma_t Z_t$. That these processes satisfy the GARCH relation follows from the relations $Y_t = A_t Y_{t-1} + b$, as in the proof of Theorem 8.6. Being a fixed measurable transformation of $(Z_t, Z_{t-1}, \ldots)$ for each t, the process (σ_t, X_t) is strictly stationary.

By construction the variable X_t is $\sigma(Z_t, Z_{t-1}, \ldots)$-measurable for every t. To see that, conversely, Z_t is $\sigma(X_t, X_{t-1}, \ldots)$-measurable, we apply a similar argument as in the proof of Theorem 8.6, based on inverting the relation $(\phi - \theta)(B) X_t^2 = \alpha + \phi(B) W_t$, for $W_t = X_t^2 - \sigma_t^2$. Presently, the series' X_t^2 and W_t are not necessarily integrable, but Lemma 8.11 below still allows to conclude that W_t is $\sigma(X_t^2, X_{t-1}^2, \ldots)$-measurable, provided that the polynomial ϕ has no zeros on the unit disc.

The matrix B obtained by replacing the variables Z_{t-1} and the numbers θ_j in the matrix A_t by zero is bounded above by A_t in a coordinatewise sense. By the nonnegativity of the entries this implies that $B^n \leq A_0 A_{-1} \cdots A_{-n+1}$ and hence $B^n \to 0$. This can happen only if all eigenvalues of B are inside the unit circle. Indeed, if z is an eigenvalue of B with $|z| \geq 1$ and $c \neq 0$ a corresponding eigenvector, then $B^n c = z^n c$ does not converge to zero. Now

$$\det(B - zI) = (-1)^{p+q-1} \big(z^{p+q-1} - \sum_{j=1}^{p+q-1} \phi_j z^{p+q-1-j} \big).$$

Thus z is a zero of ϕ if and only if z^{-1} is an eigenvalue of B. We conclude that ϕ has no zeros on the unit disc.

Finally, we show the necessity of the top Lyapounov exponent being negative. If there exists a strictly stationary solution to the GARCH equations, then, by (8.5) and the nonnegativity of the coefficients, $\sum_{j=1}^n A_0 A_{-1} \cdots A_{-n+1} b \leq Y_0$ for every n, and hence $A_0 A_{-1} \cdots A_{-n+1} b \to 0$ as $n \to \infty$, almost surely. By the form of b this is equivalent to $A_0 A_{-1} \cdots A_{-n+1} e_1 \to 0$. Using the structure of the matrices A_t we next see that $A_0 A_{-1} \cdots A_{-n+1} \to 0$ in probability as $n \to \infty$, by an argument similar as in the proof of Theorem 8.6. Because the matrices A_t are independent and the event where $A_0 A_{-1} \cdots A_{-n+1} \to 0$ is a tail event, this event must have probability one. It can be shown that this is possible only if the top Lyapounov exponent of the matrices A_t is negative.[†] ■

8.11 Lemma. *Let ϕ be a polynomial without roots on the unit disc and let X_t be a time series that is bounded in probability. If $Z_t = \phi(B)X_t$ for every t, then X_t is $\sigma(Z_t, Z_{t-1}, \ldots)$-measurable.*

Proof. Because $\phi(0) \neq 0$ by assumption, we can assume without loss of generality that ϕ possesses intercept 1. If ϕ is of degree 0, then $X_t = Z_t$ for every t and the assertion is certainly true. We next proceed by induction on the degree of ϕ. If ϕ is of degree $p \geq 1$, then we can write it as $\phi(z) = (1 - \phi z)\phi_{p-1}(z)$ for a polynomial ϕ_{p-1} of degree $p - 1$ and a complex number ϕ with $|\phi| < 1$. The series $Y_t = (1 - \phi B)X_t$ is bounded in probability and $\phi_{p-1}(B)Y_t = Z_t$, whence Y_t is $\sigma(Z_t, Z_{t-1}, \ldots)$-measurable, by the induction hypothesis. By iterating the relation $X_t = \phi X_{t-1} + Y_t$, we find that $X_t = \phi^n X_{t-n} + \sum_{j=0}^{n-1} \phi^j Y_{t-j}$. Because the sequence X_t is uniformly tight and $\phi^n \to 0$, the sequence $\phi^n X_{t-n}$ converges to zero in probability. Hence X_t is the limit in probability of a sequence that is $\sigma(Y_t, Y_{t-1}, \ldots)$-measurable and hence is $\sigma(Z_t, Z_{t-1}, \ldots)$-measurable. This implies the result. ■

** **8.12** EXERCISE. In the preceding lemma the function $\psi(z) = 1/\phi(z)$ possesses a power series representation $\psi(z) = \sum_{j=0}^\infty \psi_j z^j$ on a neighbourhood of the unit disc. Is it true under the conditions of the lemma that $X_t = \sum_{j=0}^\infty \psi_j Z_{t-j}$, where the series converges (at least) in probability?

8.13 Example. For the GARCH(1, 1) process the random matrices A_t given by (8.7) reduce to the random variables $\phi_1 + \theta_1 Z_{t-1}^2$. The top Lyapounov exponent of these random (1×1) matrices is equal to $\mathrm{E} \log(\phi_1 + \theta_1 Z_t^2)$. This number can be written as an integral relative to the distribution of Z_t, but in general is not easy to compute analytically. □

The proofs of the preceding theorems provide a recipe for generating a GARCH process starting from initial values. Given a centered and scaled i.i.d. process Z_t and an arbitrary random vector Y_0 of dimension $p + q - 1$, we define a process Y_t through the recursions $Y_t = A_t Y_{t-1} + b$, with the matrices A_t given in (8.7) and the vector b equal to α times the first unit vector. Next we set $\sigma_t = \sqrt{Y_{t,1}}$ and $X_t = \sigma_t Z_t$ for $t \geq 1$. Because the stationary solution to the GARCH equation is unique, the initial vector Y_0 must be simulated from a "stationary distribution" in order to obtain a stationary GARCH process. However, the effect of a "nonstationary" initialization wears off as $t \to \infty$ and the process will approach stationarity, provided the coefficients in the GARCH equation are such that a stationary solution exists. This is true both for L_2-stationarity and strict stationarity, under the appropriate conditions on the coefficients.

[†] See Bougerol (), Lemma ?.

8.14 Theorem. *Let $\alpha > 0$, let $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ be nonnegative numbers, and let Z_t be an i.i.d. process with mean zero and unit variance such that Z_t is independent of $\mathcal{F}_{t-1}$ for every $t \in \mathbb{Z}$.*

(i) If $\sum_j (\phi_j + \theta_j) < 1$, then the difference $X_t - \tilde{X}_t$ of any two solutions $X_t = \sigma_t Z_t$ and $\tilde{X}_t = \tilde{\sigma}_t Z_t$ of the GARCH equations that are square-integrable converges to zero in L_2 as $t \to \infty$.

(ii) If the top Lyapounov exponent of the matrices A_t in (8.7) is negative, then the difference $X_t - \tilde{X}_t$ of any two solutions $X_t = \sigma_t Z_t$ and $\tilde{X}_t = \tilde{\sigma}_t Z_t$ of the GARCH equations converges to zero in probability as $t \to \infty$.

Proof. From the two given GARCH processes X_t and $\tilde{X}_t$ define processes Y_t and $\tilde{Y}_t$ as indicated preceding the statement of Theorem 8.10. These processes satisfy (8.5) for the matrices A_t given in (8.7). Choosing $n = t - 1$ and taking differences we see that

$$Y_t - \tilde{Y}_t = A_t A_{t-1} \cdots A_1 (Y_0 - \tilde{Y}_0).$$

If the top Lyapounov exponent of the matrices A_t is negative, then the norm of the right side can be bounded, almost surely for sufficiently large t, by by $e^{\gamma' t} \|Y_0 - \tilde{Y}_0\|$ for some number $\gamma' < 0$. This follows from the subergodic theorem, as before (even though this time the matrix product grows on its left side). This converges to zero as $t \to \infty$, implying that $\sigma_t - \tilde{\sigma}_t \to 0$ almost surely as $t \to \infty$. This in turn implies (ii).

Under the condition of (i), the spectral radius of the matrix $A = \mathrm{E} A_t$ is strictly smaller than 1 and hence $\|A^n\| \to 0$. By the nonnegativity of the entries of the matrices A_t the absolute values of the coordinates of the vectors $Y_t - \tilde{Y}_t$ are bounded above by the coordinates of the vector $A_t A_{t-1} \cdots A_1 Z_0$, for Z_0 the vector obtained by replacing the coordinates of $Y_0 - \tilde{Y}_0$ by their absolute values. By the independence of the matrices A_t and vector Z_0, the expectation of $A_t A_{t-1} \cdots A_1 Z_0$ is bounded by $A^t \mathrm{E} Z_0$, which converges to zero. Because $\sigma_t^2 = Y_{t,1}$ and $X_t = \sigma_t Z_t$, this implies that, as $t \to \infty$,

$$\mathrm{E}|X_t^2 - \tilde{X}_t^2| = \mathrm{E}|\sigma_t^2 - \tilde{\sigma}_t^2| Z_t^2 = \mathrm{E}|\sigma_t^2 - \tilde{\sigma}_t^2| \to 0.$$

For the stationary solution X_t the sequence (X_t^2) is uniformly integrable, because the variables X_t^2 possess a fixed marginal distribution with finite second moment. By the preceding display this is then also true for $\tilde{X}_t$, and hence also for a general $\tilde{X}_t$. The sequence $X_t - \tilde{X}_t$ is then uniformly square-integrable as well. Combining this with the fact that $X_t - \tilde{X}_t \to 0$ in probability, we see that $X_t - \tilde{X}_t$ converges to zero in second mean. ∎

The preceding theorem may seem at odds with a common interpretation of a stationary and stability condition as a condition for "persistence". The condition for L_2-stationarity of a GARCH process is stronger than the condition for strict stationarity, so that it appears as if we have found two different conditions for persistence. Whenever a strictly stationary solution exists, the influence of initial values wears off as time goes to infinity, and hence the initial values are not persistent. This is true independently of the validity of the condition $\sum_j (\phi_j + \theta_j) < 1$ for L_2-stationarity. However, the latter condition, if it holds, does ensure that the process approaches stationarity in the stronger L_2-sense.

The condition $\sum_j(\phi_j+\theta_j) < 1$ is necessary for the strictly stationary solution to have finite second moments. By an appropriate initialization we can ensure that a GARCH process has finite second moments for every t, even if this condition fails. (It will then not be stationary.) However, in this case the variances $\mathrm{E}X_t^2$ must diverge to infinity as $t \to \infty$. This follows by a Fatou type argument, because the process will approach the strictly stationary solution and this has infinite variance.

8.15 EXERCISE. Suppose that the time series $\tilde{X}_t$ is strictly stationary with infinite second moments and $X_t - \tilde{X}_t \to 0$ in probability as $t \to \infty$. Show that $\mathrm{E}X_t^2 \to \infty$.

We can make this more concrete by considering the prediction formula for the conditional variance process σ_t^2. For the GARCH$(1,1)$ process we prove below that

$$\mathrm{E}(X_{t+h}^2|\,\mathcal{F}_t) = \mathrm{E}(\sigma_{t+h}^2|\,\mathcal{F}_t) = (\phi_1+\theta_1)^{h-1}\sigma_{t+1}^2 + \alpha\sum_{j=0}^{h-2}(\phi_1+\theta_1)^j. \tag{8.8}$$

For $\phi_1+\theta_1 < 1$ the first term on the far right converges to zero as $h \to \infty$, indicating that information at the present time t does not help to predict the conditional variance process in the "infinite future". On the other hand, if $\phi_1+\theta_1 \geq 1$ and $\alpha > 0$ then both terms on the far right side contribute positively as $h \to \infty$. If $\phi_1+\theta_1 = 1$, then the relative contribution of the term $(\phi_1+\theta_1)^{h-1}\sigma_t^2$ tends to zero as $h \to \infty$, whereas if $\phi_1+\theta_1 > 1$ the contributions are of the same order. In the last case the value σ_t^2 appears to be particularly "persistent".

The case that $\sum_j(\phi_i+\theta_j) = 1$ is often viewed as having particular interest and is referred to as *integrated GARCH* or IGARCH. Many financial time series yield GARCH fits that are close to IGARCH.

A GARCH process, being a martingale difference, does not allow nontrivial predictions of its mean values. However, it is of interest to predict the conditional variances σ_t^2, or equivalently the process of squares X_t^2. Predictions based on the infinite past $\mathcal{F}_t$ can be obtained using the auto-regressive representation from the proof of Theorem 8.10. Let A_t be the matrix given in (8.7) and let $Y_t = (\sigma_t^2,\ldots,\sigma_{t-p+1}^2, X_{t-1}^2,\ldots,X_{t-q+1}^2)^T$, so that $Y_t = A_tY_{t-1}+b$ for every t. The vector Y_{t-1} is $\mathcal{F}_{t-2}$-measurable, and the matrix A_t depends on Z_{t-1} only, with $A = \mathrm{E}(A_t|\,\mathcal{F}_{t-2})$ independent of t. It follows that

$$\mathrm{E}(Y_t|\,\mathcal{F}_{t-2}) = \mathrm{E}(A_t|\,\mathcal{F}_{t-2})Y_{t-1}+b = AY_{t-1}+b.$$

By iterating this equation we find that, for $h > 1$,

$$\mathrm{E}(Y_t|\,\mathcal{F}_{t-h}) = A^{h-1}Y_{t-h+1} + \sum_{j=0}^{h-2}A^jb.$$

In the case of a GARCH$(1,1)$ process the vector Y_t is equal to σ_t^2 and the matrix A reduces to the number $\phi_1+\theta_1$, whence we obtain the equation (8.8). For a general

GARCH(p,q) process the process σ_t^2 is the first coordinate of Y_t, and the prediction equation takes a more involved form, but is still explicitly given in the preceding display. If $\sum_j(\phi_j+\theta_j)<1$, then the spectral radius of the matrix A is strictly smaller than 1, and both terms on the right converge to zero at an exponential rate, as $h\to\infty$. In this case the potential of predicting the conditional variance process is limited to the very near future.

8.16 EXERCISE. Suppose that $\sum_j(\phi_j+\theta_j)<1$ and let X_t be a stationary Garch process. Show that $\mathrm{E}(X_{t+h}^2|\,\mathcal{F}_t)\to \mathrm{E}X_t^2$ as $h\to\infty$.

* 8.2 Linear GARCH with Leverage and Power GARCH

Fluctuations of foreign exchange rates tend to be symmetric, in view of the two-sided nature of the foreign exchange market. However, it is an empirical finding that for asset prices the current returns and future volatility are negatively correlated. For instance, a crash in the stock market will be followed by large volatility.

A linear GARCH model is not able to capture this type of asymmetric relationship, because it models the volatility as a function of the squares of the past returns. One attempt to allow for asymmetry is to replace the GARCH equation (8.1) by

$$\sigma_t^2=\alpha+\phi_1\sigma_{t-1}^2+\cdots+\phi_p\sigma_{t-p}^2+\theta_1(|X_{t-1}|+\gamma_1X_{t-1})^2+\cdots+\theta_q(|X_{t-q}|+\gamma_qX_{t-q})^2.$$

This reduces to the ordinary GARCH equation if the *leverage coefficients* γ_i are set equal to zero. If these coefficients are negative, then a positive deviation of the process X_t contributes to lower volatility in the near future, and conversely.

A power GARCH model is obtained by replacing the squares in the preceding display by other powers.

* 8.3 Exponential GARCH

The exponential GARCH or EGARCH model is significantly different from the GARCH models described so far. It retains the basic set-up of a process of the form $X_t=\sigma_tZ_t$ for a martingale difference sequence Z_t satisfying (8.2) and an $\mathcal{F}_{t-1}$-adapted process σ_t, but replaces the GARCH equation by

$$\log\sigma_t^2=\alpha+\phi_1\log\sigma_{t-1}^2+\cdots+\phi_p\log\sigma_{t-p}^2+\theta_1(|Z_{t-1}|+\gamma_1Z_{t-1})+\cdots+\theta_q(|Z_{t-q}|+\gamma_qZ_{t-q}).$$

Through the presence of both the variables Z_t and their absolute values and the transformation to the logarithmic scale this can also capture the leverage effect. An advantage

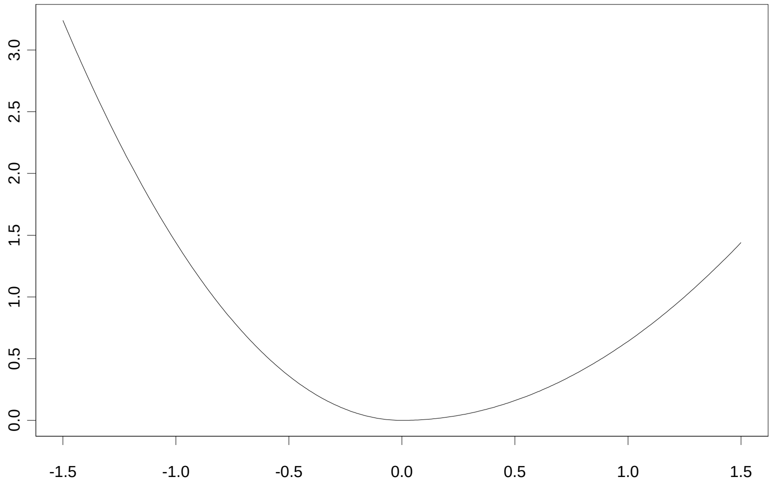

Figure 8.3. The function $x \mapsto (|x| + \gamma x)^2$ for $\gamma = -0.2$.

of modelling the logarithm of the volatility is that the parameters of the model need not be restricted to be positive.

Because the EGARCH model specifies the log volatility directly in terms of the noise process Z_t and its own past, its definition is less recursive than the ordinary GARCH definition, and easier to handle. In particular, for fixed and identical leverage coefficients $\gamma_i = \gamma$ the EGARCH equation describes the log volatility process $\log \sigma_t^2$ as a regular ARMA process driven by the noise process $|Z_t| + \gamma Z_t$, and we may use the theory for ARMA processes to study its properties. In particular, if the roots of the polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ are outside the unit circle, then there exists a stationary solution $\log \sigma_t^2$ that is measurable relative to the σ-field generated by the process Z_{t-1}. If the process Z_t is strictly stationary, then so is the stationary solution $\log \sigma_t^2$ and so is the EGARCH process $X_t = \sigma_t Z_t$.

* 8.4 GARCH in Mean

A GARCH process by its definition is a white noise process, and thus it could be a useful candidate to drive another process. For instance, an observed process Y_t could be assumed to satisfy the ARMA equation

$$\overline{\phi}(B)Y_t = \overline{\theta}(B)X_t,$$

for X_t a GARCH process, relative to other polynomials ϕ and θ (which are unrelated to

$\overline{\phi}$ and $\overline{\theta}$). One then says that Y_t is "ARMA in the mean" and "GARCH in the variance", or that Y_t is an ARMA-GARCH series. Results on ARMA processes that hold for any driving white noise process will clearly also hold in the present case, where the white noise process is a GARCH process.

8.17 EXERCISE. Let X_t be a stationary GARCH process relative to polynomials ϕ and θ and let the time series Y_t be the unique stationary solution to the equation $\overline{\phi}(B)Y_t = \overline{\theta}(B)X_t$, for $\overline{\phi}$ and $\overline{\theta}$ polynomials that have all their roots outside the unit disc. Let $\mathcal{F}_t$ be the filtration generated by Y_t. Show that $\mathrm{var}(Y_t |\, \mathcal{F}_{t-1}) = \mathrm{var}(X_t |\, X_{t-1}, X_{t-2}, \ldots)$ almost surely.

It has been found useful to go a step further and let also the conditional variance of the driving GARCH process appear in the mean model for the process Y_t. Thus given a GARCH process X_t with conditional variance process $\sigma_t^2 = \mathrm{var}(X_t |\, \mathcal{F}_{t-1})$ it is assumed that $Y_t = f(\sigma_t, X_t)$ for a fixed function f. The function f is assumed known up to a number of parameters. For instance,

$$\begin{aligned}\overline{\phi}(B)Y_t &= \psi\sigma_t + \overline{\theta}(B)X_t,\\ \overline{\phi}(B)Y_t &= \psi\sigma_t^2 + \overline{\theta}(B)X_t,\\ \overline{\phi}(B)Y_t &= \psi\log\sigma_t^2 + \overline{\theta}(B)X_t.\end{aligned}$$

These models are known as *GARCH-in-mean*, or GARCH-M models.

9
State Space Models

A causal, stationary AR(1) process with i.i.d. innovations Z_t is a *Markov process*: the conditional distribution of the "future value" $X_{t+1} = \phi X_t + Z_{t+1}$ given the "past values" $X_1, \ldots, X_t$ depends on the "present value" X_t only. Specifically, the conditional density of X_{t+1} is given by

$$p_{X_{t+1}|X_1,\ldots,X_t}(x) = p_Z(x - \phi X_t).$$

(The assumption of causality ensures that Z_{t+1} is independent of $X_1, \ldots, X_t$.) The Markov structure has an obvious practical interpretation and suggests a recursive algorithm to compute predictions. It also allows a simple factorization of the likelihood function. For instance, the likelihood for the causal AR(1) process in the previous paragraph can be written

$$p_{X_1,\ldots,X_n}(X_1, \ldots, X_n) = \prod_{t=2}^{n} p_Z(X_t - \phi X_{t-1}) p_{X_1}(X_1).$$

It would be of interest to have a similar property for more general time series.

Some non-Markovian time series can be forced into Markov form by incorporating enough past information into a "present state". For instance, an $AR(p)$ process with $p \geq 2$ is not Markov, because X_{t+1} depends on p variables in the past. We can remedy this by defining a "present state" to consist of the vector $\vec{X}_t := (X_t, \ldots, X_{t-p+1})$: the process $\vec{X}_t$ is Markov. In general, to induce Markov structure we must define a state in such a way that it incorporates all relevant information for transition to the next state. This is of interest mostly if this is possible using "states" that are of not too high complexity.

A *hidden Markov model* consists of a Markov chain, but rather than the state at time t we observe a transformation of it, up to noise which is independent of the Markov chain. A related structure is the *state space model*. Given an "initial state" X_0, "disturbances" $V_1, W_1, V_2, \ldots$ and functions f_t and g_t, processes X_t and Y_t are defined recursively by

$$\begin{aligned} X_t &= f_t(X_{t-1}, V_t), \\ Y_t &= g_t(X_t, W_t). \end{aligned} \tag{9.1}$$

We refer to X_t as the "state" at time t and to Y_t as the "output". The state process X_t can be viewed as primary and evolving in the background, describing the consecutive states of a system in time. At each time t the system is "measured", producing an output Y_t. If the sequence $X_0, V_1, V_2, \ldots$ consists of independent variables, then the state process X_t is a Markov chain. If the variables $X_0, V_1, W_1, V_2, W_2, V_3, \ldots$ are independent, then for every t given the state X_t the output Y_t is conditionally independent of the states $X_0, X_1, \ldots$ and outputs $Y_1, \ldots, Y_{t-1}$. Under this condition the state space model becomes a hidden Markov model.

9.1 EXERCISE. Formulate the claims and statements in the preceding two sentences precisely, and give proofs.

Typically, the state process X_t is not observed, but instead at time t we only observe the output Y_t. For this reason the process Y_t is also referred to as the "measurement process". The second equation in the display (9.1) is called the "measurement equation", while the first is the "state equation". Inference might be directed at estimating parameters attached to the functions f_t or g_t, to the distribution of the errors or to the initial state, and/or on predicting or reconstructing the states X_t from the observed outputs $Y_1, \ldots, Y_n$. Predicting or reconstructing the state sequence is referred to as "filtering" or "smoothing".

For linear functions f_t and g_t and vector-valued states and outputs the state space model takes the form

$$\begin{aligned} X_t &= F_t X_{t-1} + V_t, \\ Y_t &= G_t X_t + W_t. \end{aligned} \tag{9.2}$$

The matrices F_t and G_t are often postulated to be independent of t. In this linear state space model the analysis usually concerns linear predictions, and then a common assumption is that the vectors $X_0, V_1, W_1, V_2, \ldots$ are uncorrelated. If F_t is independent of t and the vectors V_t form a white noise process, then the series X_t is a VAR(1) process.

Because state space models are easy to handle, it is of interest to represent a given observable time series Y_t as the output of a state space model. This entails finding a state space, a state process X_t, and a corresponding state space model with the given series Y_t as output. It is particularly attractive to find a linear state space model. Such a *state space representation* is definitely not unique. An important issue in systems theory is to find a (linear) state space representation of minimum dimension.

9.2 Example (State space representation ARMA). Let Y_t be a stationary, causal ARMA$(r+1, r)$ process satisfying $\phi(B)Y_t = \theta(B)Z_t$ for an i.i.d. process Z_t. (The choice $p = q + 1$ can always be achieved by padding the set of coefficients of the polynomials ϕ or θ with zeros.) Then the $AR(p)$ process $X_t = (1/\phi)(B)Z_t$ is related to Y_t through

$Y_t = \theta(B)X_t$. Thus

$$Y_t = (\,\theta_0, \ldots, \theta_r\,) \begin{pmatrix} X_t \\ \vdots \\ X_{t-r} \end{pmatrix},$$

$$\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-r} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_r & \phi_{r+1} \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-r-1} \end{pmatrix} + \begin{pmatrix} Z_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

This is a linear state space representation (9.2) with state vector $(X_t, \ldots, X_{t-r})$, and matrices F_t and G_t, that are independent of t. Under causality the innovations $V_t = (Z_t, 0, \ldots, 0)$ are orthogonal to the past X_t and Y_t; the innovations W_t as in (9.2) are defined to be zero. The state vectors are typically unobserved, except when θ is of degree zero. (If the ARMA process is invertible and the coefficients of θ are known, then they can be reconstructed from the infinite past through the relation $X_t = (1/\theta)(B)Y_t$.)

In the present representation the state-dimension of the ARMA(p,q) process is $r + 1 = \max(p, q+1)$. By using a more complicated noise process it is possible to represent an ARMA(p,q) process in dimension $\max(p,q)$, but this difference appears not to be very important.[‡] □

9.3 Example (State space representation ARIMA). Consider a time series Z_t whose differences $Y_t = \nabla Z_t$ satisfy the linear state space model (9.2) for a state sequence X_t. Writing $Z_t = Y_t + Z_{t-1} = G_t X_t + W_t + Z_{t-1}$, we obtain that

$$\begin{pmatrix} X_t \\ Z_{t-1} \end{pmatrix} = \begin{pmatrix} F_t & 0 \\ G_{t-1} & 1 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-2} \end{pmatrix} + \begin{pmatrix} V_t \\ W_{t-1} \end{pmatrix}$$
$$Z_t = (\,G_t \quad 1\,) \begin{pmatrix} X_t \\ Z_{t-1} \end{pmatrix} + W_t.$$

We conclude that the time series Z_t possesses a linear state space representation, with states of one dimension higher than the states of the original series.

A drawback of the preceding representation is that the error vectors (V_t, W_{t-1}, W_t) are not necessarily uncorrelated if the error vectors (V_t, W_t) in the system with outputs Y_t have this property. In the case that Z_t is an ARIMA$(p,1,q)$ process, we may use the state representation of the preceding example for the ARMA(p,q) process Y_t, which has errors $W_t = 0$, and this disadvantage does not arise. Alternatively, we can avoid this problem by using another state space representation. For instance, we can write

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \begin{pmatrix} F_t & 0 \\ G_t F_t & 1 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} V_t \\ G_t V_t + W_t \end{pmatrix}$$
$$Z_t = (\,0 \quad 1\,) \begin{pmatrix} X_t \\ Z_t \end{pmatrix}.$$

[‡] See e.g. Brockwell and Davis, p469–471.

This illustrates that there may be multiple possibilities to represent a time series as the output of a (linear) state space model.

The preceding can be extended to general ARIMA(p,d,q) models. If $Y_t=(1-B)^d Z_t$, then $Z_t=Y_t-\sum_{j=1}^d \binom{d}{j}(-1)^j Z_{t-j}$. If the process Y_t can be represented as the output of a state space model with state vectors X_t, then Z_t can be represented as the output of a state space model with the extended states $(X_t,Z_{t-1},\ldots,Z_{t-d})$, or, alternatively, $(X_t,Z_t,\ldots,Z_{t-d+1})$. □

9.4 Example (Stochastic linear trend). A time series with a linear trend could be modelled as $Y_t=\alpha+\beta t+W_t$ for constants α and β, and a stationary process W_t (for instance an ARMA process). This restricts the nonstationary part of the time series to a deterministic component, which may be unrealistic. An alternative is the *stochastic linear trend model* described by

$$\begin{pmatrix}A_t\\ B_t\end{pmatrix}=\begin{pmatrix}1&1\\0&1\end{pmatrix}\begin{pmatrix}A_{t-1}\\ B_{t-1}\end{pmatrix}+V_t$$
$$Y_t=A_t+W_t.$$

The stochastic processes (A_t,B_t) and noise processes (V_t,W_t) are unobserved. This state space model contains the deterministic linear trend model as the degenerate case where $V_t\equiv 0$, so that $B_t\equiv B_0$ and $A_t\equiv A_0+B_0t$.

The state equations imply that $\nabla A_t=B_{t-1}+V_{t,1}$ and $\nabla B_t=V_{t,2}$, for $V_t=(V_{t,1},V_{t,2})^T$. Taking differences on the output equation $Y_t=A_t+W_t$ twice, we find that

$$\nabla^2 Y_t=\nabla B_{t-1}+\nabla V_{t,1}+\nabla^2 W_t=V_{t,2}+\nabla V_{t,1}+\nabla^2 W_t.$$

If the process (V_t,W_t) is a white noise process, then the auto-correlation function of the process on the right vanishes for lags bigger than 2 (the polynomial $\nabla^2=(1-B)^2$ being of degree 2). Thus the right side is an MA(2) process, whence the process Y_t is an ARIMA(0,2,2) process. □

9.5 Example (Structural models). Besides a trend we may suspect that a given time series shows a seasonal effect. One possible parametrization of a deterministic seasonal effect with S seasons is the function

$$(9.3)\qquad t\mapsto\sum_{s=1}^{\lfloor S/2\rfloor}\gamma_s\cos(\lambda_s t)+\delta_s\sin(\lambda_s t),\qquad \lambda_s=\frac{2\pi s}{S},\quad s=1,\ldots,\lfloor S/2\rfloor.$$

By appropriate choice of the parameters γ_s and δ_s this function is able to adapt to any periodic function on the integers with period S. We could add this deterministic function to a given time series model in order to account for seasonality. Again it may not be realistic to require the seasonality a-priori to be deterministic. An alternative is to replace the fixed function $s\mapsto(\gamma_s,\delta_s)$ by the time series defined by

$$\begin{pmatrix}\gamma_{s,t}\\ \delta_{s,t}\end{pmatrix}=\begin{pmatrix}\cos\lambda_s&\sin\lambda_s\\ \sin\lambda_s&-\cos\lambda_s\end{pmatrix}\begin{pmatrix}\gamma_{s,t-1}\\ \delta_{s,t-1}\end{pmatrix}+V_{s,t}.$$

An observed time series may next have the form

$$Y_t = (\,1 \quad 1 \quad \ldots \quad 1\,) \begin{pmatrix} \gamma_{1,t} \\ \gamma_{2,t} \\ \vdots \\ \gamma_{s,t} \end{pmatrix} + Z_t.$$

Together these equations again constitute a linear state space model. If $V_t = 0$, then this reduces to the deterministic trend model. (Cf. Exercise 9.6.)

A model with both a stochastic linear trend and a stochastic seasonal component is known as a "structural model". □

9.6 EXERCISE. Consider the state space model with state equations $\gamma_t = \gamma_{t-1}\cos\lambda + \delta_{t-1}\sin\lambda + V_{t,1}$ and $\delta_t = \gamma_{t-1}\sin\lambda - \delta_{t-1}\cos\lambda + V_{t,2}$ and output equation $Y_t = \gamma_t + W_t$. What does this model reduce to if $V_t \equiv 0$?

9.7 EXERCISE.
(i) Show that the function (of $t \in \mathbb{Z}$) in (9.3) is periodic with period S.
(ii) Show that any periodic function $f\colon \mathbb{Z} \to \mathbb{R}$ with period S can be written in the form (9.3).
[For (ii) it suffices to show that the any vector $\bigl(f(1), \ldots, f(S)\bigr)$ can be represented as a linear combination of the vectors $\bigl(\cos\lambda_s, \ldots, \cos(S\lambda_s)\bigr)$ and $\bigl(\sin\lambda_s, \ldots, \sin(S\lambda_s)\bigr)$.]

The showpiece of state space modelling is the Kalman filter. This is an algorithm to compute linear predictions (for linear state space models), under the assumption that the parameters of the system are known. Because the formulas for the predictors, which are functions of the parameters and the outputs, can in turn be used to set up estimating equations for the parameters, the Kalman filter is also important for statistical analysis. We start discussing parameter estimation in Chapter 10.

The variables X_t and Y_t in a state space model will typically be random vectors. For two random vectors X and Y of dimensions m and n the *covariance* or "cross-covariance" is the $(m\times n)$ matrix $\mathrm{Cov}(X,Y) = \mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)^T$. The random vectors X and Y are called "uncorrelated" if $\mathrm{Cov}(X,Y) = 0$, or equivalently if $\mathrm{cov}(X_i,Y_j) = 0$ for every pair (i,j). The *linear span* of a set of vectors is defined as the linear span of all their coordinates. Thus this is a space of (univariate) random variables, rather than random vectors! We shall also understand a projection operator Π, which is a map on the space of random variables, to act coordinatewise on vectors: if X is a vector, then ΠX is the vector consisting of the projections of the coordinates of X. As a vector-valued operator a projection Π is still linear, in that $\Pi(FX+Y) = F\Pi X + \Pi Y$, for any matrix F and random vectors X and Y.

9.1 Kalman Filtering

The *Kalman filter* is a recursive algorithm to compute best linear predictions of the states $X_1, X_2, \ldots$ given observations $Y_1, Y_2, \ldots$ in the linear state space model (9.2). The core algorithm allows to compute predictions $\Pi_t X_t$ of the states X_t given observed outputs $Y_1, \ldots, Y_t$. Here by "predictions" we mean Hilbert space projections, but given the time values involved "reconstructions" would perhaps be more appropriate. "Filtering" is the preferred term in systems theory. Given the reconstructions $\Pi_t X_t$, it is easy to compute predictions of future states and future outputs. A next step is "Kalman smoothing", which is the name for the reconstruction (through projections) of the full state sequence $X_1, \ldots, X_n$ given the outputs $Y_1, \ldots, Y_n$.

In the simplest situation the vectors $X_0, V_1, W_1, V_2, W_2, \ldots$ are assumed uncorrelated. We shall first derive the filter under the more general assumption that the vectors $X_0, (V_1, W_1), (V_2, W_2), \ldots$ are uncorrelated, and in Section 9.2.3 we further relax this condition. The matrices F_t and G_t as well as the covariance matrices of the noise variables (V_t, W_t) are assumed known.

By applying (9.2) recursively, we see that the vector X_t is contained in the linear span of the variables $X_0, V_1, \ldots, V_t$. It is immediate from (9.2) that the vector Y_t is contained in the linear span of X_t and W_t. These facts are true for every $t \in \mathbb{N}$. It follows that under our conditions the noise variables V_t and W_t are uncorrelated with all vectors X_s and Y_s with $s < t$.

Let H_0 be a given closed linear subspace of $L_2(\Omega, \mathcal{U}, \mathrm{P})$ that contains the constants, and for $t \geq 0$ let Π_t be the orthogonal projection onto the space $H_t = H_0 + \mathrm{lin}\,(Y_1, \ldots, Y_t)$. The space H_0 may be viewed as our "knowledge" at time 0; it may be $H_0 = \mathrm{lin}\,\{1\}$. We assume that the noise vectors $V_1, W_1, V_2, \ldots$ are orthogonal to H_0. Combined with the preceding this shows that the vector (V_t, W_t) is orthogonal to the space H_{t-1}, for every $t \geq 1$.

The Kalman filter consists of the recursions

$$\cdots \to \begin{pmatrix} \Pi_{t-1} X_{t-1} \\ \mathrm{Cov}(\Pi_{t-1} X_{t-1}) \\ \mathrm{Cov}(X_{t-1}) \end{pmatrix} \overset{(1)}{\to} \begin{pmatrix} \Pi_{t-1} X_t \\ \mathrm{Cov}(\Pi_{t-1} X_t) \\ \mathrm{Cov}(X_t) \end{pmatrix} \overset{(2)}{\to} \begin{pmatrix} \Pi_t X_t \\ \mathrm{Cov}(\Pi_t X_t) \\ \mathrm{Cov}(X_t) \end{pmatrix} \to \cdots$$

Thus the Kalman filter alternates between "updating the current state", step (1), and "updating the prediction space", step (2).

Step (1) is simple. Because $V_t \perp H_0, Y_1, \ldots, Y_{t-1}$ by assumption, we have $\Pi_{t-1} V_t = 0$. Applying Π_t to the state equation $X_t = F_t X_{t-1} + V_t$ we find that, by the linearity of a projection,

$$\begin{aligned} \Pi_{t-1} X_t &= F_t (\Pi_{t-1} X_{t-1}), \\ \mathrm{Cov}(\Pi_{t-1} X_t) &= F_t \,\mathrm{Cov}(\Pi_{t-1} X_{t-1}) F_t^T, \\ \mathrm{Cov}(X_t) &= F_t \,\mathrm{Cov}(X_{t-1}) F_t^T + \mathrm{Cov}(V_t). \end{aligned}$$

This gives a complete description of step (1) of the algorithm.

Step (2) is more involved, but also comes down to simple matrix computations. The vector $\tilde{W}_t = Y_t - \Pi_{t-1} Y_t$ is known as the *innovation* at time t, because it is the part of Y_t that is not explainable at time $t-1$. It is orthogonal to H_{t-1}, and together with this space spans H_t. It follows that H_t can be orthogonally decomposed as $H_t = H_{t-1} + \mathrm{lin}\,\tilde W_t$ and hence the projection onto H_t is the sum of the projections onto the spaces H_{t-1} and $\mathrm{lin}\,\tilde W_t$. At the beginning of step (2) the vector $\tilde W_t$ is known, because we can write, using the measurement equation and the fact that $\Pi_{t-1}W_t = 0$,

$$\tilde W_t = Y_t - \Pi_{t-1}Y_t = Y_t - G_t\Pi_{t-1}X_t = G_t(X_t - \Pi_{t-1}X_t) + W_t. \tag{9.4}$$

The middle expression is easy to compute from the current values at the beginning of step (2). Applying this to projecting the variable X_t, we find

$$\Pi_t X_t = \Pi_{t-1}X_t + \Lambda_t\tilde W_t, \qquad \Lambda_t = \mathrm{Cov}(X_t, \tilde W_t)\,\mathrm{Cov}(\tilde W_t)^{-1}. \tag{9.5}$$

The matrix Λ_t is chosen such that $\Lambda_t\tilde W_t$ is the projection of X_t onto $\mathrm{lin}\,\tilde W_t$. Because $W_t \perp X_{t-1}$ the state equation equation yields $\mathrm{Cov}(X_t, W_t) = \mathrm{Cov}(V_t, W_t)$. By the orthogonality property of projections $\mathrm{Cov}(X_t, X_t - \Pi_{t-1}X_t) = \mathrm{Cov}(X_t - \Pi_{t-1}X_t)$. Combining this and the identity $\tilde W_t = G_t(X_t - \Pi_{t-1}X_t) + W_t$ from (9.4), we compute

$$\begin{aligned} \mathrm{Cov}(X_t, \tilde W_t) &= \mathrm{Cov}(X_t - \Pi_{t-1}X_t)G_t^T + \mathrm{Cov}(V_t, W_t),\\ \mathrm{Cov}(\tilde W_t) &= G_t\,\mathrm{Cov}(X_t - \Pi_{t-1}X_t)G_t^T + G_t\,\mathrm{Cov}(V_t, W_t)\\ &\qquad + \mathrm{Cov}(W_t, V_t)G_t^T + \mathrm{Cov}(W_t),\\ \mathrm{Cov}(X_t - \Pi_{t-1}X_t) &= \mathrm{Cov}(X_t) - \mathrm{Cov}(\Pi_{t-1}X_t). \end{aligned} \tag{9.6}$$

The matrix $\mathrm{Cov}(X_t - \Pi_{t-1}X_t)$ is the *prediction error matrix* at time $t-1$ and the last equation follows by Pythagoras' rule. To complete the recursion of step (2) we compute from (9.5)

$$\mathrm{Cov}(\Pi_t X_t) = \mathrm{Cov}(\Pi_{t-1}X_t) + \Lambda_t\,\mathrm{Cov}(\tilde W_t)\Lambda_t^T. \tag{9.7}$$

Equations (9.5)–(9.7) give a complete description of step (2) of the Kalman recursion.

The Kalman algorithm must be initialized in one of its two steps, for instance by providing Π_0X_1 and its covariance matrix, so that the recursion can start with a step of type (2). It is here where the choice of H_0 plays a role. Choosing $H_0 = \mathrm{lin}\,(1)$ gives predictions using $Y_1, \ldots, Y_t$ as well as an intercept and requires that we know $\Pi_0X_1 = \mathrm{E}X_1$. It may also be desired that $\Pi_{t-1}X_t$ is the projection onto $\mathrm{lin}\,(1, Y_{t-1}, Y_{t-2}, \ldots)$ for a stationary extension of Y_t into the past. Then we set Π_0X_1 equal to the projection of X_1 onto $H_0 = \mathrm{lin}\,(1, Y_0, Y_{-1}, \ldots)$.

9.2 Future States and Outputs

Predictions of future values of the state variable follow easily from Π_tX_t, because $\Pi_tX_{t+h} = F_{t+h}\Pi_tX_{t+h-1}$ for any $h \geq 1$. Given the predicted states, future outputs can be predicted from the measurement equation by $\Pi_tY_{t+h} = G_{t+h}\Pi_tX_{t+h}$.

* 9.2.1 Missing Observations

A considerable attraction of the Kalman filter algorithm is the ease by which missing observations can be accomodated. This can be achieved by simply filling in the missing data points by "external" variables that are independent of the system. Suppose that (X_t, Y_t) follows the linear state space model (9.2) and that we observe a subset $(Y_t)_{t\in T}$ of the variables $Y_1, \ldots, Y_n$. We define a new set of matrices G_t^* and noise variables W_t^* by

$$\begin{aligned} G_t^* &= G_t, \qquad W_t^* = W_t, \qquad t \in T,\\ G_t^* &= 0, \qquad W_t^* = \overline{W}_t, \qquad t \notin T, \end{aligned}$$

for random vectors $\overline{W}_t$ that are independent of the vectors that are already in the system. The choice $\overline{W}_t = 0$ is permitted. Next we set

$$\begin{aligned} X_t &= F_t X_{t-1} + V_t,\\ Y_t^* &= G_t^* X_t + W_t^*. \end{aligned}$$

The variables (X_t, Y_t^*) follow a state space model with the same state vectors X_t. For $t \in T$ the outputs $Y_t^* = Y_t$ are identical to the outputs in the original system, while for $t \notin T$ the output is $Y_t^* = \overline{W}_t$, which is pure noise by assumption. Because the noise variables $\overline{W}_t$ cannot contribute to the prediction of the hidden states X_t, best predictions of states based on the observed outputs $(Y_t)_{t\in T}$ or based on $Y_1^*, \ldots, Y_n^*$ are identical. We can compute the best predictions based on $Y_1^*, \ldots, Y_n^*$ by the Kalman recursions, but with the matrices G_t^* and $\mathrm{Cov}(W_t^*)$ substituted for G_t and $\mathrm{Cov}(W_t)$. Because the Y_t^* with $t \notin T$ will not appear in the projection formula, we can just as well set their "observed values" equal to zero in the computations.

* 9.2.2 Kalman Smoothing

Besides in predicting future states or outputs we may be interested in reconstructing the complete state sequence $X_0, X_1, \ldots, X_n$ from the outputs $Y_1, \ldots, Y_n$. The computation of $\Pi_n X_n$ is known as the *filtering problem*, and is step (2) of our description of the Kalman filter. The computation of $P_n X_t$ for $t = 0, 1, \ldots, n-1$ is known as the *smoothing problem*. For a given t it can be achieved through the recursions, with $\tilde W_n$ as given in (9.4),

$$\begin{pmatrix} \Pi_n X_t\\ \mathrm{Cov}(X_t, \tilde W_n)\\ \mathrm{Cov}(X_t, X_n - \Pi_{n-1}X_n)\end{pmatrix} \to \begin{pmatrix} \Pi_{n+1} X_t\\ \mathrm{Cov}(X_t, \tilde W_{n+1})\\ \mathrm{Cov}(X_t, X_{n+1} - \Pi_n X_{n+1})\end{pmatrix}, \qquad n = t, t+1, \ldots.$$

The initial value at $n = t$ of the recursions and the covariance matrices $\mathrm{Cov}(\tilde W_n)$ of the innovations $\tilde W_n$ are given by (9.6)–(9.7), and hence can be assumed known.

Because H_{n+1} is the sum of the orthogonal spaces H_n and $\mathrm{lin}\, \tilde W_{n+1}$, we have, as in (9.5),

$$\Pi_{n+1} X_t = \Pi_n X_t + \Lambda_{t,n+1} \tilde W_{n+1}, \qquad \Lambda_{t,n+1} = \mathrm{Cov}(X_t, \tilde W_{n+1})\, \mathrm{Cov}(\tilde W_{n+1})^{-1}.$$

The recursion for the first coordinate $\Pi_n X_t$ follows from this and the recursions for the second and third coordinates, the covariance matrices $\mathrm{Cov}(X_t, \tilde W_{n+1})$ and $\mathrm{Cov}(X_t, X_{n+1} - \Pi_n X_{n+1})$.

Using in turn the state equation and equation (9.5), we find

$$\begin{aligned}\mathrm{Cov}(X_t, X_{n+1} - \Pi_n X_{n+1}) &= \mathrm{Cov}\bigl(X_t, F_{n+1}(X_n - \Pi_n X_n) + V_{n+1}\bigr)\\ &= \mathrm{Cov}\bigl(X_t, F_{n+1}(X_n - \Pi_{n-1} X_n + \Lambda_n \tilde W_n)\bigr).\end{aligned}$$

This readily gives the recursion for the third component, the matrix Λ_n being known from (9.5)–(9.6). Next using equation (9.4), we find

$$\mathrm{Cov}(X_t, \tilde W_{n+1}) = \mathrm{Cov}(X_t, X_{n+1} - \Pi_n X_{n+1}) G_{n+1}^T.$$

* 9.2.3 Lagged Correlations

In the preceding we have assumed that the vectors $X_0, (V_1, W_1), (V_2, W_2), \ldots$ are uncorrelated. An alternative assumption is that the vectors $X_0, V_1, (W_1, V_2), (W_2, V_3), \ldots$ are uncorrelated. (The awkward pairing of W_t and V_{t+1} can be avoided by writing the state equation as $X_t = F_t X_{t-1} + V_{t-1}$ and next making the assumption as before.) Under this condition the Kalman filter takes a slightly different form, where for economy of computation it can be useful to combine the steps (1) and (2).

Both possibilities are covered by the assumptions that

- the vectors $X_0, V_1, V_2, \ldots$ are orthogonal.
- the vectors $W_1, W_2, \ldots$ are orthogonal.
- the vectors V_s and W_t are orthogonal for all (s,t) except possibly $s = t$ or $s = t + 1$.
- all vectors are orthogonal to H_0.

Under these assumptions step (2) of the Kalman filter remains valid as described. Step (1) must be adapted, because it is no longer true that $\Pi_{t-1} V_t = 0$.

Because $V_t \perp H_{t-2}$, we can compute $\Pi_{t-1} V_t$ from the innovation decomposition $H_{t-1} = H_{t-2} + \mathrm{lin}\ \tilde W_{t-1}$, as $\Pi_{t-1} V_t = K_{t-1} \tilde W_{t-1}$ for the matrix

$$K_{t-1} = \mathrm{Cov}(V_t, W_{t-1})\,\mathrm{Cov}(\tilde W_{t-1})^{-1}.$$

Note here that $\mathrm{Cov}(V_t, \tilde W_{t-1}) = \mathrm{Cov}(V_t, W_{t-1})$, in view of (9.4). We replace the calculations for step (1) by

$$\begin{aligned}\Pi_{t-1} X_t &= F_t(\Pi_{t-1} X_{t-1}) + K_t \tilde W_{t-1},\\ \mathrm{Cov}(\Pi_{t-1} X_t) &= F_t\,\mathrm{Cov}(\Pi_{t-1} X_{t-1}) F_t^T + K_t\,\mathrm{Cov}(\tilde W_{t-1}) K_t^T,\\ \mathrm{Cov}(X_t) &= F_t\,\mathrm{Cov}(X_{t-1}) F_t^T + \mathrm{Cov}(V_t).\end{aligned}$$

This gives a complete description of step (1) of the algorithm, under the assumption that the vector $\tilde W_{t-1}$, and its covariance matrix are kept in memory after the preceding step (2).

The smoothing algorithm goes through as stated except for the recursion for the matrices $\mathrm{Cov}(X_t, X_n - \Pi_{n-1} X_n)$. Because $\Pi_n V_{n+1}$ may be nonzero, this becomes

$$\begin{aligned}\mathrm{Cov}(X_t, X_{n+1} - \Pi_n X_{n+1}) = \mathrm{Cov}\bigl(X_t, X_n - \Pi_{n-1} X_n) F_{n+1}^T + \mathrm{Cov}(X_t, \tilde W_n) \Lambda_n^T F_{n+1}^T\\ + \mathrm{Cov}(X_t, \tilde W_n) K_n^T.\end{aligned}$$

* 9.3 Nonlinear Filtering

The simplicity of the Kalman filter results both from the simplicity of the linear state space model and the fact that it concerns linear predictions. Together these lead to update formulas expressed in the form of matrix algebra. The principle of recursive predictions can be applied more generally to compute nonlinear predictions in nonlinear state space models, provided the conditional densities of the variables in the system are available and certain integrals involving these densities can be evaluated, analytically, numerically, or by stochastic simulation.

Somewhat abusing notation we write a conditional density of a variable X given another variable Y as $p(x|\,y)$, and a marginal density of X as $p(x)$. Consider the nonlinear state space model (9.1), where we assume that the vectors $X_0, V_1, W_1, V_2, \ldots$ are independent. Then the outputs $Y_1, \ldots, Y_n$ are conditionally independent given the state sequence $X_0, X_1, \ldots, X_n$, and the conditional law of a single output Y_t given the state sequence depends on X_t only. In principle the (conditional) densities $p(x_0), p(x_1|\,x_0), p(x_2|\,x_1), \ldots$ and the conditional densities $p(y_t|\,x_t)$ of the outputs are available from the form of the functions f_t and g_t and the distributions of the noise variables (V_t, W_t). Under the assumption of independent noise vectors the system is a hidden Markov model, and the joint density of states up till time $n+1$ and outputs up till time n can be expressed in these densities as

$$p(x_0)p(x_1|\,x_0)\cdots p(x_{n+1}|\,x_n)p(y_1|\,x_1)p(y_2|\,x_2)\cdots p(y_n|\,x_n). \tag{9.8}$$

The marginal density of the outputs is obtained by integrating this function relative to $(x_0, \ldots, x_{n+1})$. The conditional density of the state sequence $(X_0, \ldots, X_{n+1})$ given the outputs is proportional to the function in the display, the norming constant being the marginal density of the outputs. In principle, this allows the computation of all conditional expectations $\mathrm{E}(X_t|\,Y_1, \ldots, Y_n)$, the (nonlinear) "predictions" of the state. However, because this approach expresses these predictions as a quotient of $n+1$-dimensional integrals, and n may be large, this is unattractive unless the integrals can be evaluated easily.

An alternative for finding predictions is a recursive scheme for calculating conditional densities, of the form

$$\cdots \to p(x_{t-1}|\,y_{t-1}, \ldots, y_1) \overset{(1)}{\to} p(x_t|\,y_{t-1}, \ldots, y_1) \overset{(2)}{\to} p(x_t|\,y_t, \ldots, y_1) \to \cdots.$$

This is completely analogous to the updates of the linear Kalman filter: the recursions alternate between "updating the state", (1), and "updating the prediction space", (2).

Step (1) can be summarized by the formula

$$\begin{aligned} p(x_t|\,y_{t-1}, \ldots, y_1) &= \int p(x_t|\,x_{t-1}, y_{t-1}, \ldots, y_1)p(x_{t-1}|\,y_{t-1}, \ldots, y_1)\,d\mu_{t-1}(x_{t-1}) \\ &= \int p(x_t|\,x_{t-1})p(x_{t-1}|\,y_{t-1}, \ldots, y_1)\,d\mu_{t-1}(x_{t-1}). \end{aligned}$$

The second equality follows from the conditional independence of the vectors X_t and $Y_{t-1}, \ldots, Y_1$ given X_{t-1}. This is a consequence of the form of $X_t = f_t(X_{t-1}, V_t)$

and the independence of V_t and the vectors $X_{t-1}, Y_{t-1}, \ldots, Y_1$ (which are functions of $X_0, V_1, \ldots, V_{t-1}, W_1, \ldots, W_{t-1}$).

To obtain a recursion for step (2) we apply Bayes formula to the conditional density of the pair (X_t, Y_t) given $Y_{t-1}, \ldots, Y_1$ to obtain

$$\begin{aligned} p(x_t|\, y_t, \ldots, y_1) &= \frac{p(y_t|\, x_t, y_{t-1}, \ldots, y_1)p(x_t|\, y_{t-1}, \ldots, y_1)}{\int p(y_t|\, x_t, y_{t-1}, \ldots, y_1)p(x_t|\, y_{t-1}, \ldots, y_1)\, d\mu_t(x_t)} \\ &= \frac{p(y_t|\, x_t)p(x_t|\, y_{t-1}, \ldots, y_1)}{p(y_t|\, y_{t-1}, \ldots, y_1)}. \end{aligned}$$

The second equation is a consequence of the fact that $Y_t = g_t(X_t, W_t)$ is conditionally independent of $Y_{t-1}, \ldots, Y_1$ given X_t. The conditional density $p(y_t|\, y_{t-1}, \ldots, y_1)$ in the denominator is a nuisance, because it will rarely be available explicitly, but acts only as a norming constant.

The preceding formulas are useful only if the integrals can be evaluated. If analytical evaluation is impossible, then perhaps numerical methods or stochastic simulation could be of help.

If stochastic simulation is the method of choice, then it may be attractive to apply *Markov Chain Monte Carlo* for direct evaluation of the joint law, without recursions. The idea is to simulate a sample from the conditional density $p(x_0, \ldots, x_{n+1}|\, y_1, \ldots, y_n)$ of the states given the outputs. The biggest challenge is the dimensionality of this conditional density. The *Gibbs sampler* overcomes this by simulating recursively from the marginal conditional densities $p(x_t|\, x_{-t}, y_1, \ldots, y_n)$ of the single variables X_t given the outputs $Y_1, \ldots, Y_n$ and the vectors $X_{-t} = (X_0, \ldots, X_{t-1}, X_{t+1}, \ldots, X_{n+1})$ of remaining states. We refer to the literature for general discussion of the Gibbs sampler, but shall show that these marginal distributions are relatively easy to obtain for the general state space model (9.1).

Under independence of the vectors $X_0, V_1, W_1, V_2, \ldots$ the joint density of states and outputs takes the hidden Markov form (9.8). The conditional density of X_t given the other vectors is proportional to this expression viewed as function of x_t only. Only three terms of the product depend on x_t and hence we find

$$p(x_t|\, x_{-t}, y_1, \ldots, y_n) \asymp p(x_t|\, x_{t-1})p(x_{t+1}|\, x_t)p(y_t|\, x_t).$$

The norming constant is a function of the conditioning variables $x_{-t}, y_1, \ldots, y_n$ only and can be recovered from the fact that the left side is a probability density as a function of x_t. A closer look will reveal that it is equal to $p(y_t|\, x_{t-1}, x_{t+1})p(x_{t+1}|\, x_{t-1})$. However, many simulation methods, in particular the popular Metropolis-Hastings algorithm, can be implemented without an explicit expression for the proportionality constant. The forms of the three densities on the right side should follow from the specification of the system.

The assumption that the variables $X_0, V_1, W_2, V_2, \ldots$ are independent may be too restrictive, although it is natural to try and construct the state variables so that it is satisfied. Somewhat more complicated formulas can be obtained under more general assumptions. Assumptions that are in the spirit of the preceding derivations in this chapter are:

(i) the vectors $X_0, X_1, X_2, \ldots$ form a Markov chain.
(ii) the vectors $Y_1, \ldots, Y_n$ are conditionally independent given $X_0, X_1, \ldots, X_{n+1}$.
(iii) for each $t \in \{1, \ldots, n\}$ the vector Y_t is conditionally independent of the vector $(X_0, \ldots, X_{t-2}, X_{t+2}, \ldots, X_{n+1})$ given (X_{t-1}, X_t, X_{t+1}).

The first assumption is true if the vectors $X_0, V_1, V_2, \ldots$ are independent. The second and third assumptions are certainly satisfied if all noise vectors $X_0, V_1, W_1, V_2, W_2, V_3, \ldots$ are independent. The exercises below give more general sufficient conditions for (i)–(iii) in terms of the noise variables.

In comparison to the hidden Markov situation considered previously not much changes. The joint density of states and outputs can be written in a product form similar to (9.8), the difference being that each conditional density $p(y_t|\, x_t)$ must be replaced by $p(y_t|\, x_{t-1}, x_t, x_{t+1})$. The variable x_t then occurs in five terms of the product and hence we obtain

$$\begin{aligned} p(x_t|\, x_{-t}, y_1, \ldots, y_n) \asymp\, & p(x_{t+1}|\, x_t) p(x_t|\, x_{t-1}) \times \\ & \times p(y_{t-1}|\, x_{t-2}, x_{t-1}, x_t) p(y_t|\, x_{t-1}, x_t, x_{t+1}) p(y_{t+1}|\, x_t, x_{t+1}, x_{t+2}). \end{aligned}$$

This formula is general enough to cover the case of the ARV model discussed in the next section.

9.8 EXERCISE. Suppose that $X_0, V_1, W_1, V_2, W_2, V_3, \ldots$ are independent, and define states X_t and outputs Y_t by (9.1). Show that (i)–(iii) hold, where in (iii) the vector Y_t is even conditionally independent of $(X_s\colon s \neq t)$ given X_t.

9.9 EXERCISE. Suppose that $X_0, V_1, V_2, \ldots, Z_1, Z_2, \ldots$ are independent, and define states X_t and outputs Y_t through (9.2) with $W_t = h_t(V_t, V_{t+1}, Z_t)$ for measurable functions h_t. Show that (i)–(iii) hold. [Under (9.2) there exists a measurable bijection between the vectors $(X_0, V_1, \ldots, V_t)$ and $(X_0, X_1, \ldots, X_n)$, and also between the vectors (X_t, X_{t-1}, X_{t+1}) and (X_t, V_t, V_{t+1}). Thus conditioning on $(X_0, X_1, \ldots, X_{n+1})$ is the same as conditioning on $(X_0, V_1, \ldots, V_{n+1})$ or on $(X_0, V_1, \ldots, V_n, X_{t-1}, X_t, X_{t+1})$.]

* **9.10** EXERCISE. Show that the condition in the preceding exercise that $W_t = h_t(V_t, V_{t+1}, Z_t)$ for Z_t independent of the other variables is equivalent to the conditional independence of W_t and $X_0, V_1, \ldots, V_n, W_s\colon s \neq t$ given V_t, V_{t+1}.

9.4 Stochastic Volatility Models

The term "volatility", which we have used at multiple occasions to describe the "movability" of a time series, appears to have its origins in the theory of option pricing. The *Black-Scholes model* for pricing an option on a given asset with price S_t is based on a diffusion equation of the type

$$dS_t = \mu_t S_t \, dt + \sigma_t S_t \, dB_t.$$

Here B_t is a Brownian motion process and μ_t and σ_t are stochastic processes, which are usually assumed to be adapted to the filtration generated by the process S_t. In the original Black-Scholes model the process σ_t is assumed constant, and the constant is known as the "volatility" of the process S_t.

The Black-Scholes diffusion equation can also be written in the form

$$\log\frac{S_t}{S_0}=\int_0^t(\mu_s-\tfrac{1}{2}\sigma_s^2)\,ds+\int_0^t\sigma_s\,dB_s.$$

If μ and σ are deterministic processes this shows that the log returns $\log S_t/S_{t-1}$ over the intervals $(t-1,t]$ are independent, normally distributed variables ($t=1,2,\ldots$) with means $\int_{t-1}^t(\mu_s-\frac{1}{2}\sigma_s^2)\,ds$ and variances $\int_{t-1}^t\sigma_s^2\,ds$. In other words, if these means and variances are denoted by $\overline{\mu}_t$ and $\overline{\sigma}_t^2$, then the variables

$$Z_t=\frac{\log S_t/S_{t-1}-\overline{\mu}_t}{\overline{\sigma}_t}$$

are an i.i.d. sample from the standard normal distribution. The standard deviation $\overline{\sigma}_t$ can be viewed as an "average volatility" over the interval $(t-1,t]$. If the processes μ_t and σ_t are not deterministic, then the process Z_t is not necessarily Gaussian. However, if the unit of time is small, so that the intervals $(t-1,t]$ correspond to short time intervals in real time, then it is still believable that the variables Z_t are approximately normally distributed. In that case it is also believable that the processes μ_t and σ_t are approximately constant and hence these processes can replace the averages $\overline{\mu}_t$ and $\overline{\sigma}_t$. Usually, one even assumes that the process $\overline{\mu}_t$ is constant in time. For simplicity of notation we shall take $\overline{\mu}_t$ to be zero in the following, leading to a model of the form

$$\log S_t/S_{t-1}=\sigma_t Z_t,$$

for standard normal variables Z_t and a "volatility" process σ_t. The choice $\overline{\mu}_t=\mu_t-\frac{1}{2}\sigma_t^2=0$ corresponds to modelling under the "risk-free" martingale measure, but is made here only for convenience.

There is ample empirical evidence that models with constant volatility do not fit observed financial time series. In particular, this has been documented through a comparison of the option prices predicted by the Black-Scholes formula to the observed prices on the option market. Because the Black-Scholes price of an option on a given asset depends only on the volatility parameter of the asset price process, a single parameter volatility model would allow to calculate this parameter from the observed price of an option on this asset, by inversion of the Black-Scholes formula. Given a range of options written on a given asset, but with different maturities and/or different strike prices, this inversion process usually leads to a range of "implied volatilities", all connected to the same asset price process. These *implied volatilities* usually vary with the maturity and strike price.

This discrepancy could be taken as proof of the failure of the reasoning behind the Black-Scholes formula, but the more common explanation is that "volatility" is a random process itself. One possible model for this process is a diffusion equation of the type

$$d\sigma_t=\lambda_t\sigma_t\,dt+\gamma_t\sigma_t\,dW_t,$$

where W_t is another Brownian motion process. This leads to a "stochastic volatility model in continuous time". Many different parametric forms for the processes λ_t and γ_t are suggested in the literature. One particular choice is to assume that $\log \sigma_t$ is an *Ornstein-Uhlenbeck process*, i.e. it satisfies

$$d \log \sigma_t = \lambda(\xi - \log \sigma_t)\, dt + \gamma\, dW_t.$$

(An application of Itô's formula show that this corresponds to the choices $\lambda_t = \frac{1}{2}\gamma^2 + \lambda(\xi - \log \sigma_t)$ and $\gamma_t = \gamma$.) The Brownian motions B_t and W_t are often assumed to be dependent, with quadratic variation $\langle B, W\rangle_t = \delta t$ for some parameter $\delta \leq 0$.

A diffusion equation is a stochastic differential equation in continuous time, and does not fit well into our basic set-up, which considers the time variable t to be integer-valued. One approach would be to use continuous time models, but assume that the continuous time processes are observed only at a grid of time points. In view of the importance of the option-pricing paradigm in finance it has been also useful to give a definition of "volatility" directly through discrete time models. These models are usually motivated by an analogy with the continuous time set-up. "Stochastic volatility models" in discrete time are specifically meant to parallel continuous time diffusion models.

The most popular stochastic volatility model in discrete time is the *auto-regressive random variance model* or ARV model. A discrete time analogue of the Ornstein-Uhlenbeck type volatility process σ_t is the specification

$$\log \sigma_t = \alpha + \phi \log \sigma_{t-1} + V_{t-1}. \tag{9.9}$$

For $|\phi| < 1$ and a white noise process V_t this auto-regressive equation possesses a causal stationary solution $\log \sigma_t$. We select this solution in the following. The observed log return process X_t is modelled as

$$X_t = \sigma_t Z_t, \tag{9.10}$$

where it is assumed that the time series (V_t, Z_t) is i.i.d.. The latter implies that Z_t is independent of $V_{t-1}, Z_{t-1}, V_{t-2}, Z_{t-2}, \ldots$ and hence of $X_{t-1}, X_{t-2}, \ldots$, but allows dependence between V_t and Z_t. The volatility process σ_t is not observed.

A dependence between V_t and Z_t allows for a leverage effect, one of the "stylized facts" of financial time series. In particular, if V_t and Z_t are negatively correlated, then a small return X_t, which is indicative of a small value of Z_t, suggests a large value of V_t, and hence a large value of the log volatility $\log \sigma_{t+1}$ at the next time instant. (Note that the time index $t-1$ of V_{t-1} in the auto-regressive equation (9.9) is unusual, because in other situations we would have written V_t. It is meant to support the idea that σ_t is determined at time $t-1$.)

An ARV stochastic volatility process is a nonlinear state space model. It induces a linear state space model for the log volatilities and log absolute log returns of the form

$$\begin{aligned} \log \sigma_t &= \begin{pmatrix} \alpha & \phi \end{pmatrix} \begin{pmatrix} 1 \\ \log \sigma_{t-1} \end{pmatrix} + V_{t-1} \\ \log |X_t| &= \log \sigma_t + \log |Z_t|. \end{aligned}$$

In order to take the logarithm of the observed series X_t it was necessary to take the absolute value $|X_t|$ first. Usually this is not a serious loss of information, because the sign of X_t is equal to the sign of Z_t, and this is a Bernoulli $\frac{1}{2}$ series if Z_t is symmetrically distributed.

The linear state space form allows the application of the Kalman filter to compute best linear projections of the unobserved log volatilities $\log \sigma_t$ based on the observed log absolute log returns $\log |X_t|$. Although this approach is computationally attractive, a disadvantage is that the best predictions of the volatilities σ_t based on the log returns X_t may be much better than the exponentials of the best linear predictions of the log volatilities $\log \sigma_t$ based on the log returns. Forcing the model in linear form is not entirely natural here. However, the computation of best nonlinear predictions is involved. Markov Chain Monte Carlo methods are perhaps the most promising technique, but are highly computer-intensive.

An ARV process X_t is a martingale difference series relative to its natural filtration $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$. To see this we first note that by causality $\sigma_t \in \sigma(V_{t-1}, V_{t-2}, \ldots)$, whence $\mathcal{F}_t$ is contained in the filtration $\mathcal{G}_t = \sigma(V_s, Z_s\colon s \le t)$. The process X_t is actually already a martingale difference relative to this bigger filtration, because by the assumed independence of Z_t from $\mathcal{G}_{t-1}$

$$\mathrm{E}(X_t|\, \mathcal{G}_{t-1}) = \sigma_t \mathrm{E}(Z_t|\, \mathcal{G}_{t-1}) = 0.$$

A fortiori the process X_t is a martingale difference series relative to the filtration $\mathcal{F}_t$.

There is no correspondingly simple expression for the conditional variance process $\mathrm{E}(X_t^2|\, \mathcal{F}_{t-1})$ of an ARV series. By the same argument

$$\mathrm{E}(X_t^2|\, \mathcal{G}_{t-1}) = \sigma_t^2 \mathrm{E} Z_t^2.$$

If $\mathrm{E}Z_t^2 = 1$ it follows that $\mathrm{E}(X_t^2|\, \mathcal{F}_{t-1}) = \mathrm{E}(\sigma_t^2|\, \mathcal{F}_{t-1})$, but this is intractable for further evaluation. In particular, the process σ_t^2 is not the conditional variance process, unlike in the situation of a GARCH process. Correspondingly, in the present context, in which σ_t is considered the “volatility”, the volatility and conditional variance processes do not coincide.

9.11 EXERCISE. One definition of a volatility process σ_t of a time series X_t is a process σ_t such that X_t/σ_t is an i.i.d. standard normal series. Suppose that $X_t = \tilde\sigma_t Z_t$ is a GARCH process with conditional variance process $\tilde\sigma_t^2$ and driven by an i.i.d. process Z_t. If Z_t is standard normal, show that $\tilde\sigma_t$ qualifies as a volatility process. [Trivial.] If Z_t is a t_p-process show that there exists a process S_t^2 with a chisquare distribution with p degrees of freedom such that $\sqrt{p}\, \tilde\sigma_t / S_t$ qualifies as a volatility process.

9.12 EXERCISE. In the ARV model is σ_t measurable relative to the σ-field generated by $X_{t-1}, X_{t-2}, \ldots$? Compare with GARCH models.

In view of the analogy with continuous time diffusion processes the assumption that the variables (V_t, Z_t) in (9.9)–(9.10) are normally distributed could be natural. This assumption certainly helps to compute moments of the series. The stationary solution $\log \sigma_t$ of the auto-regressive equation (9.9) is given by (for $|\phi| < 1$)

$$\log \sigma_t = \sum_{j=0}^{\infty} \phi^j (V_{t-1-j} + \alpha) = \sum_{j=0}^{\infty} \phi^j V_{t-1-j} + \frac{\alpha}{1-\phi}.$$

If the time series V_t is i.i.d. Gaussian with mean zero and variance σ^2, then it follows that the variable $\log \sigma_t$ is normally distributed with mean $\alpha/(1-\phi)$ and variance $\sigma^2/(1-\phi^2)$. The Laplace transform $\mathrm{E}\exp(aZ)$ of a standard normal variable Z is given by $\exp(\frac{1}{2}a^2)$. Therefore, under the normality assumption on the process V_t it is straightforward to compute that, for $p > 0$,

$$\mathrm{E}|X_t|^p = \mathrm{E}e^{p\log \sigma_t}\mathrm{E}|Z_t|^p = \exp\Bigl(\tfrac{1}{2}\frac{\sigma^2 p^2}{1-\phi^2} + \frac{\alpha p}{1-\phi}\Bigr)\mathrm{E}|Z_t|^p.$$

Consequently, the kurtosis of the variables X_t can be computed to be

$$\kappa_4(X) = e^{4\sigma^2/(1-\phi^2)}\kappa_4(Z).$$

If follows that the time series X_t possesses a larger kurtosis than the series Z_t. This is true even for $\phi = 0$, but the effect is more pronounced for values of ϕ that are close to 1, which are commonly found in practice. Thus the ARV model is able to explain leptokurtic tails of an observed time series.

Under the assumption that the variables (V_t, Z_t) are i.i.d. and bivariate normally distributed, it is also possible to compute the auto-correlation function of the squared series X_t^2 explicitly. If $\delta = \rho(V_t, Z_t)$ is the correlation between the variables V_t and Z_t, then the vectors $(\log \sigma_t, \log \sigma_{t+h}, Z_t)$ possess a three-dimensional normal distribution with covariance matrix

$$\begin{pmatrix} \beta^2 & \beta^2\phi^h & 0 \\ \beta^2\phi^h & \beta^2 & \phi^{h-1}\delta\sigma \\ 0 & \phi^{h-1}\delta\sigma & 1 \end{pmatrix}, \qquad \beta^2 = \frac{\sigma^2}{1-\phi^2}.$$

Some calculations show that the auto-correlation function of the square process is given by

$$\rho_{X^2}(h) = \frac{(1 + 4\delta^2\sigma^2\phi^{2h-2})e^{4\sigma^2\phi^h/(1-\phi^2)} - 1}{3e^{4\sigma^2/(1-\phi^2)} - 1}, \qquad h > 0.$$

The auto-correlation is positive at positive lags and decreases exponentially fast to zero, with a rate depending on the proximity of ϕ to 1. For values of ϕ close to 1, the decrease is relatively slow.

9.13 EXERCISE. Derive the formula for the auto-correlation function.

9.14 EXERCISE. Suppose that the variables V_t and Z_t are independent for every t, in addition to independence of the vectors (V_t, Z_t), and assume that the variables V_t (but not necessarily the variables Z_t) are normally distributed. Show that

$$\rho_{X^2}(h) = \frac{e^{4\sigma^2\phi^h/(1-\phi^2)} - 1}{\kappa_4(Z)e^{4\sigma^2/(1-\phi^2)} - 1}, \qquad h > 0.$$

[Factorize $\mathrm{E}\sigma_{t+h}^2\sigma_t^2 Z_{t+h}^2 Z_t^2$ as $\mathrm{E}\sigma_{t+h}^2\sigma_t^2 \mathrm{E}Z_{t+h}^2 Z_t^2$.]

The choice of the logarithmic function in the auto-regressive equation (9.9) has some arbitrariness, and other possibilities, such as a power function, have been explored.

10
Moment and Least Squares Estimators

Suppose that we observe realizations $X_1, \ldots, X_n$ from a time series X_t whose distribution is (partly) described by a parameter $\theta \in \mathbb{R}^d$. For instance, an ARMA process with the parameter $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$, or a GARCH process with parameter $(\alpha, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$, both ranging over a subset of $\mathbb{R}^{p+q+1}$. In this chapter we discuss two methods of estimation of the parameters, based on the observations $X_1, \ldots, X_n$: the "method of moments" and the "least squares method".

When applied in the standard form to auto-regressive processes, the two methods are essentially the same, but for other models the two methods may yield quite different estimators. Depending on the moments used and the underlying model, least squares estimators can be more efficient, although sometimes they are not usable at all. The "generalized method of moments" tries to bridge the efficiency gap, by increasing the number of moments employed.

Moment and least squares estimators are popular in time series analysis, but in general they are less efficient than maximum likelihood and Bayes estimators. The difference in efficiency depends on the model and the true distribution of the time series. Maximum likelihood estimation using a Gaussian model can be viewed as an extension of the method of least squares. We discuss the method of maximum likelihood in Chapter 12.

10.1 Yule-Walker Estimators

Suppose that the time series $X_t - \mu$ is a stationary auto-regressive process of known order p and with unknown parameters $\phi_1, \ldots, \phi_p$ and σ^2. The mean $\mu = \mathrm{E}X_t$ of the series may also be unknown, but we assume that it is estimated by $\overline{X}_n$ and concentrate attention on estimating the remaining parameters.

From Chapter 7 we know that the parameters of an auto-regressive process are not uniquely determined by the series X_t, but can be replaced by others if the white noise process is changed appropriately as well. We shall aim at estimating the parameter under the assumption that the series is causal. This is equivalent to requiring that all roots of the polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ are outside the unit circle.

Under causality the best linear predictor of X_{p+1} based on $1, X_p, \ldots, X_1$ is given by $\Pi_p X_{p+1} = \mu + \phi_1(X_p - \mu) + \cdots + \phi_p(X_1 - \mu)$. (See Section 7.4.) Alternatively, the best linear predictor can be obtained by solving the general prediction equations (2.1). This shows that the parameters $\phi_1, \ldots, \phi_p$ satisfy

$$\begin{pmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(p-1) \\ \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(p-2) \\ \vdots & \vdots & & \vdots \\ \gamma_X(p-1) & \gamma_X(p-2) & \cdots & \gamma_X(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma_X(1) \\ \gamma_X(2) \\ \vdots \\ \gamma_X(p) \end{pmatrix}.$$

We abbreviate this system of equations by $\Gamma_p \vec\phi_p = \vec\gamma_p$. These equations, which are known as the *Yule-Walker equations*, express the parameters into second moments of the observations. The *Yule-Walker estimators* are defined by replacing the true auto-covariances $\gamma_X(h)$ by their sample versions $\hat\gamma_n(h)$ and next solving for $\phi_1, \ldots, \phi_p$. This leads to the estimators

$$\hat{\vec\phi}_p := \begin{pmatrix} \hat\phi_1 \\ \hat\phi_2 \\ \vdots \\ \hat\phi_p \end{pmatrix} = \begin{pmatrix} \hat\gamma_n(0) & \hat\gamma_n(1) & \cdots & \hat\gamma_n(p-1) \\ \hat\gamma_n(1) & \hat\gamma_n(0) & \cdots & \hat\gamma_n(p-2) \\ \vdots & \vdots & & \vdots \\ \hat\gamma_n(p-1) & \hat\gamma_n(p-2) & \cdots & \hat\gamma_n(0) \end{pmatrix}^{-1} \begin{pmatrix} \hat\gamma_n(1) \\ \hat\gamma_n(2) \\ \vdots \\ \hat\gamma_n(p) \end{pmatrix} =: \hat\Gamma_p^{-1} \hat{\vec\gamma}_p.$$

The parameter σ^2 is by definition the variance of Z_{p+1}, which is the prediction error $X_{p+1} - \Pi_p X_{p+1}$ when predicting X_{p+1} by the preceding observations, under the assumption that the time series is causal. By the orthogonality of the prediction error and the predictor $\Pi_p X_{p+1}$ and Pythagoras' rule,

$$\sigma^2 = \mathrm{E}(X_{p+1} - \mu)^2 - \mathrm{E}(\Pi_p X_{p+1} - \mu)^2 = \gamma_X(0) - \vec\phi_p^T \Gamma_p \vec\phi_p. \tag{10.1}$$

We define an estimator $\hat\sigma^2$ by replacing all unknowns by their moment estimators, i.e.

$$\sigma^2 = \hat\gamma_n(0) - \hat{\vec\phi}_p^T \hat\Gamma_p \hat{\vec\phi}_p.$$

10.1 EXERCISE. An alternative method to derive the Yule-Walker equations is to work out the equations $\mathrm{cov}\bigl(\phi(B)(X_t - \mu), X_{t-k} - \mu\bigr) = \mathrm{cov}\bigl(Z_t, \sum_{j\geq 0} \psi_j Z_{t-j-k}\bigr)$ for $k = 0, \ldots, p$. Check this. Do you need causality? What if the time series would not be causal?

10.2 EXERCISE. Show that the matrix Γ_p is invertible for every p. [Suggestion: write $\alpha^T \Gamma_p \alpha$ in terms of the spectral density.]

Another reasonable method to find estimators is to start from the fact that the true values of $\phi_1, \ldots, \phi_p$ minimize the expectation

$$(\beta_1, \ldots, \beta_p) \mapsto \mathrm{E}\bigl(X_t - \mu - \beta_1(X_{t-1} - \mu) - \cdots - \beta_p(X_{t-p} - \mu)\bigr)^2.$$

The *least squares estimators* are defined by replacing this criterion function by an "empirical" (i.e. observable) version of it and next minimizing this. Let $\hat\phi_1,\ldots,\hat\phi_p$ minimize the function

$$(\beta_1,\ldots,\beta_p)\mapsto \frac{1}{n}\sum_{t=p+1}^{n}\bigl(X_t-\overline{X}_n-\beta_1(X_{t-1}-\overline{X}_n)-\cdots-\beta_p(X_{t-p}-\overline{X}_n)\bigr)^2.$$

The minimum value itself is a reasonable estimator of the minimum value of the expectation of this criterion function, which is $\mathrm{E}Z_t^2=\sigma^2$. The least squares estimators $\hat\phi_j$ obtained in this way are not identical to the Yule-Walker estimators, but the difference is small. To see this, we rewrite the least squares estimators as the solution of a system of equations. The approach is the same as for the "ordinary" least squares estimators in the linear regression model. The criterion function that we wish to minimize is the square of the norm $\|Y_n-D_n\vec\beta_p\|$ for $\vec\beta_p=(\beta_1,\ldots,\beta_p)^T$ the vector of parameters and Y_n and D_n the vector and matrix given by

$$Y_n=\begin{pmatrix}X_n-\overline{X}_n\\ X_{n-1}-\overline{X}_n\\ \vdots\\ X_{p+1}-\overline{X}_n\end{pmatrix},\qquad D_n=\begin{pmatrix}X_{n-1}-\overline{X}_n & X_{n-2}-\overline{X}_n & \cdots & X_{n-p}-\overline{X}_n\\ X_{n-2}-\overline{X}_n & X_{n-3}-\overline{X}_n & \cdots & X_{n-p-1}-\overline{X}_n\\ \vdots & \vdots & & \vdots\\ X_p-\overline{X}_n & X_{p-1}-\overline{X}_n & \cdots & X_1-\overline{X}_n\end{pmatrix}.$$

The value $\hat{\vec\phi}_p$ that minimizing the norm $\beta\mapsto\|Y_n-D_n\vec\beta_p\|$ is the vector $\vec\beta_p$ such that $D_n\vec\beta_p$ is the projection of the vector Y_n onto the range of the matrix D_n. By the projection theorem, Theorem 2.10, this is characterized by the relationship that the residuul $Y_n-D_n\vec\beta_p$ is orthogonal to the range of D_n. Algebraically this orthogonality can be expressed as $D_n^T(Y_n-D_n\vec\beta_p)=0$, are relationship that can be solved for β_p to yield that the minimizing vector is given by

$$\hat{\vec\phi}_p=\Bigl(\frac{1}{n}D_n^TD_n\Bigr)^{-1}\frac{1}{n}D_n^T(\vec X_n-\overline{X}_n).$$

At closer inspection this vector is nearly identical to the Yule-Walker estimators. Indeed, for every s,t,

$$\Bigl(\frac{1}{n}D_n^TD_n\Bigr)_{s,t}=\frac{1}{n}\sum_{j=p+1}^{n}(X_{j-s}-\overline{X}_n)(X_{j-t}-\overline{X}_n)\approx\hat\gamma_n(s-t),$$

$$\Bigl(\frac{1}{n}D_n^T(\vec X_n-\overline{X}_n)\Bigr)_t=\frac{1}{n}\sum_{j=p+1}^{n}(X_{j-t}-\overline{X}_n)(X_j-\overline{X}_n)\approx\hat\gamma_n(t).$$

Asymptotically the difference between the Yule-Walker and least squares estimators is negligible. They possess the same (normal) limit distribution.

10.3 Theorem. *Let $(X_t-\mu)$ be a causal AR(p) process relative to an i.i.d. sequence Z_t with finite fourth moments. Then both the Yule-Walker and the least squares estimators satisfy, with Γ_p the covariance matrix of $(X_1,\ldots,X_p)$,*

$$\sqrt{n}(\hat{\vec{\phi}}_p-\vec{\phi}_p)\rightsquigarrow N(0,\sigma^2\Gamma_p^{-1}).$$

Proof. We can assume without loss of generality that $\mu=0$. The AR equations $\phi(B)X_t=Z_t$ for $t=n,n-1,\ldots,p+1$ can be written in the matrix form

$$\begin{pmatrix}X_n\\ X_{n-1}\\ \vdots\\ X_{p+1}\end{pmatrix}=\begin{pmatrix}X_{n-1}&X_{n-2}&\cdots&X_{n-p}\\ X_{n-2}&X_{n-3}&\cdots&X_{n-p-1}\\ \vdots&\vdots&&\vdots\\ X_p&X_{p-1}&\cdots&X_1\end{pmatrix}\begin{pmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_p\end{pmatrix}+\begin{pmatrix}Z_n\\ Z_{n-1}\\ \vdots\\ Z_{p+1}\end{pmatrix}=D_n\vec{\phi}_p+\vec{\tilde{Z}}_n,$$

for $\vec{\tilde{Z}}_n$ the vector with coordinates $Z_t+\overline{X}_n\sum\phi_i$, and D_n the "design matrix" as before. We can solve $\vec{\phi}_p$ from this as

$$\vec{\phi}_p=(D_n^TD_n)^{-1}D_n^T(\vec{X}_n-\vec{\tilde{Z}}_n).$$

Combining this with the analogous representation of the least squares estimators $\hat{\phi}_j$ we find

$$\sqrt{n}(\hat{\vec{\phi}}_p-\vec{\phi}_p)=\Bigl(\frac{1}{n}D_n^TD_n\Bigr)^{-1}\frac{1}{\sqrt{n}}D_n^T\bigl(\vec{Z}_n-\vec{1}\overline{X}_n(1-\sum_i\phi_i)\bigr).$$

Because X_t is an auto-regressive process, it possesses a representation $X_t=\sum_j\psi_jZ_{t-j}$ for a sequence ψ_j with $\sum_j|\psi_j|<\infty$. Therefore, the results of Chapter 7 apply and show that $n^{-1}D_n^TD_n\xrightarrow{\mathrm{P}}\Gamma_p$. (In view of Problem 10.2 this also shows that the matrix $D_n^TD_n$ is invertible, as was assumed implicitly in the preceding.)

In view of Slutsky's lemma it now suffices to show that

$$\frac{1}{\sqrt{n}}D_n^T\vec{Z}_n\rightsquigarrow N(0,\sigma^2\Gamma_p),\qquad \frac{1}{\sqrt{n}}D_n^T1\overline{X}_n\xrightarrow{\mathrm{P}}0.$$

A typical coordinate of the last vector is $(t=1,\ldots,p)$

$$\frac{1}{\sqrt{n}}\sum_{j=p+1}^n(X_{j-t}-\overline{X}_n)\overline{X}_n=\frac{1}{\sqrt{n}}\sum_{j=p+1}^nX_{j-t}\overline{X}_n-\frac{n-p}{\sqrt{n}}\overline{X}_n^2.$$

In view of Theorem 4.5 and the assumption that $\mu=0$, the sequence $\sqrt{n}\overline{X}_n$ converges in distribution and hence both terms on the right side are of the order $O_P(1/\sqrt{n})$.

A typical coordinate of the first vector is $(t=1,\ldots,p)$

$$\frac{1}{\sqrt{n}}\sum_{j=p+1}^n(X_{j-t}-\overline{X}_n)Z_j=\sqrt{\frac{n-p}{n}}\frac{1}{\sqrt{n-p}}\sum_{j=1}^{n-p}Y_j+O_P(1/\sqrt{n}),$$

for $Y_j = X_{p-t+j}Z_{p+j}$. By causality of the series X_t we have $Z_{p+j} \perp X_{p-s+j}$ for $s > 0$ and hence $\mathrm{E}Y_j = \mathrm{E}X_{p-s+j}\mathrm{E}Z_{p+j} = 0$ for every j. The same type of arguments as in Chapter 5 will give us the asymptotic normality of the sequence $\sqrt{n}\overline{Y}_n$, with asymptotic variance

$$\sum_{g=-\infty}^{\infty} \gamma_Y(g) = \sum_{g=-\infty}^{\infty} \mathrm{E}X_{p-t+g}Z_{p+g}X_{p-t}Z_p.$$

In this series all terms with $g > 0$ vanish because, by the assumption of causality and the fact that Z_t is an i.i.d. sequence, Z_{p+g} is independent of $(X_{p-t+g}, X_{p-t}, Z_p)$. All terms with $g < 0$ vanish by symmetry. Thus the series is equal to $\gamma_Y(0) = \mathrm{E}X_{p-t}^2 Z_p^2 = \gamma_X(0)\sigma^2$, which is the diagonal element of $\sigma^2\Gamma_p$. This concludes the proof of the convergence in distribution of all marginals of $n^{-1/2}D_n^T\vec{Z}_n$. The joint convergence is proved in a similar way, using the Cramér-Wold device.

This concludes the proof of the asymptotic normality of the least squares estimators. The Yule-Walker estimators can be proved to be asymptotically equivalent to the least squares estimators, in that the difference is of the order $o_P(1/\sqrt{n})$. Next we apply Slutsky's lemma. ■

10.4 EXERCISE. Show that the time series Y_t in the preceding proof is strictly stationary.

* **10.5** EXERCISE. Give a precise proof of the asymptotic normality of $\sqrt{n}\overline{Y}_n$ as defined in the preceding proof.

10.1.1 Order Selection

In the preceding derivation of the least squares and Yule-Walker estimators the order p of the AR process is assumed known a-priori. Theorem 10.3 is false if $X_t - \mu$ were in reality an AR (p_0) process of order $p_0 > p$. In that case $\hat\phi_1, \ldots \hat\phi_p$ are estimators of the coefficients of the best linear predictor based on p observations, but need not converge to the p_0 coefficients $\phi_1, \ldots, \phi_{p_0}$. On the other hand, Theorem 10.3 remains valid if the series X_t is an auto-regressive process of "true" order p_0 strictly smaller than the order p used to define the estimators. This follows because for $p_0 \le p$ an AR(p_0) process is also an AR(p) process, albeit that $\phi_{p_0+1}, \ldots, \phi_p$ are zero. Theorem 10.3 shows that "overfitting" (choosing too big an order) does not cause great harm: if $\hat\phi_1^{(p)}, \ldots, \hat\phi_j^{(p)}$ are the Yule-Walker estimators when fitting an AR(p) model and the observations are an AR(p_0) process with $p_0 \le p$, then

$$\sqrt{n}\hat\phi_j^{(p)} \rightsquigarrow N\bigl(0, \sigma^2(\Gamma_p^{-1})_{j,j}\bigr), \qquad j = p_0 + 1, \ldots, p.$$

It is recomforting that the estimators of the "unnecessary" coefficients $\phi_{p_0+1}, \ldots, \phi_p$ converge to zero at rate $1/\sqrt{n}$. However, there is also a price to be paid by overfitting.

By Theorem 10.3, if fitting an AR(p)-model, then the estimators of the first p_0 coefficients satisfy

$$\sqrt{n}\left(\begin{pmatrix}\hat\phi_1^{(p)}\\ \vdots\\ \hat\phi_{p_0}^{(p)}\end{pmatrix}-\begin{pmatrix}\phi_1\\ \vdots\\ \phi_{p_0}\end{pmatrix}\right)\rightsquigarrow N\left(0,\sigma^2(\Gamma_p^{-1})_{s,t=1,\ldots,p_0}\right).$$

The covariance matrix in the right side, the $(p_0\times p_0)$ upper principal submatrix of the $(p\times p)$ matrix Γ_p^{-1}, is not equal to $\Gamma_{p_0}^{-1}$, which would have been the asymptotic covariance matrix if we had fitted an AR model of the "correct" order p_0. In fact, it is bigger in that

$$(\Gamma_p^{-1})_{s,t=1,\ldots,p_0}-\Gamma_{p_0}^{-1}\geq 0.$$

(Here we write $A\geq 0$ for a matrix A if A is positive definite.) In particular, the diagonal elements of these matrices, which are the differences of the asymptotic variances of the estimators $\phi_j^{(p)}$ and the estimators $\phi_j^{(p_0)}$, are nonnegative. Thus overfitting leads to more uncertainty in the estimators of both $\phi_1,\ldots,\phi_{p_0}$ and $\phi_{p_0+1},\ldots,\phi_p$. Fitting an auto-regressive process of very high order p increases the chance of having the model fit well to the data, but generally will result in poor estimates of the coefficients, which render the final outcome less useful.

* **10.6** EXERCISE. Prove the assertion that the given matrix is nonnegative definite.

In practice we do not know the correct order to use. A suitable order is often determined by a preliminary data-analysis, such as an inspection of the plot of the sample partial auto-correlation function. More formal methods are discussed within the general context of maximum likelihood estimation in Chapter 12.

10.7 Example. If we fit an AR(1) process to observations of an AR(1) series, then the asymptotic covariance of $\sqrt{n}(\hat\phi_1-\phi_1)$ is equal to $\sigma^2\Gamma_1^{-1}=\sigma^2/\gamma_X(0)$. If to this same process we fit an AR(2) process, then we obtain estimators $(\hat\phi_1^{(2)},\hat\phi_2^{(2)})$ (not related to the earlier $\hat\phi_1$) such that $\sqrt{n}(\hat\phi_1^{(2)}-\phi_1,\hat\phi_2^{(2)}-\phi_2)$ has asymptotic covariance matrix

$$\sigma^2\Gamma_2^{-1}=\sigma^2\begin{pmatrix}\gamma_X(0)&\gamma_X(1)\\ \gamma_X(1)&\gamma_X(0)\end{pmatrix}^{-1}=\frac{\sigma^2}{\gamma_X^2(0)-\gamma_X^2(1)}\begin{pmatrix}\gamma_X(0)&-\gamma_X(1)\\ -\gamma_X(1)&\gamma_X(0)\end{pmatrix}.$$

Thus the asymptotic variance of the sequence $\sqrt{n}(\hat\phi_1^{(2)}-\phi_1)$ is equal to

$$\frac{\sigma^2\gamma_X(0)}{\gamma_X^2(0)-\gamma_X^2(1)}=\frac{\sigma^2}{\gamma_X(0)}\frac{1}{1-\phi_1^2}.$$

(Note that $\phi_1=\gamma_X(1)/\gamma_X(0)$.) Thus overfitting by one degree leads to a loss in efficiency of $1-\phi_1^2$. This is particularly harmful if the true value of $|\phi_1|$ is close to 1, i.e. the time series is close to being a (nonstationary) random walk. □

10.1.2 Partial Auto-Correlations

Recall that the partial auto-correlation coefficient $\alpha_X(h)$ of a centered time series X_t is the coefficient of X_1 in the formula $\beta_1 X_h + \cdots + \beta_h X_1$ for the best linear predictor of X_{h+1} based on $X_1, \ldots, X_h$. In particular, for the causal AR(p) process satisfying $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$ we have $\alpha_X(p) = \phi_p$ and $\alpha_X(h) = 0$ for $h > p$. The sample partial auto-correlation coefficient is defined in Section 5.4 as the Yule-Walker estimator $\hat\phi_h$ when fitting an AR(h) model. This connection provides an alternative method to derive the limit distribution in the special situation of auto-regressive processes. The simplicity of the result makes it worth the effort.

10.8 Corollary. *Let $X_t - \mu$ be a causal stationary AR(p) process relative to an i.i.d. sequence Z_t with finite fourth moments. Then, for every $h > p$,*

$$\sqrt{n}\,\hat\alpha_n(h) \rightsquigarrow N(0,1).$$

Proof. For $h > p$ the time series $X_t - \mu$ is also an AR(h) process and hence we can apply Theorem 10.3 to find that the Yule-Walker estimators $\hat\phi_1^{(h)}, \ldots, \hat\phi_h^{(h)}$ when fitting an AR(h) model satisfy

$$\sqrt{n}(\hat\phi_h - \phi_h) \rightsquigarrow N\big(0, \sigma^2 (\Gamma_h^{-1})_{h,h}\big).$$

The left side is exactly $\sqrt{n}\,\hat\alpha_n(h)$. We show that the variance of the normal distribution on the right side is unity. By Cramér's rule the (h,h)-element of the matrix Γ_h^{-1} can be found as $\det\Gamma_{h-1}/\det\Gamma_h$. By the prediction equations we have for $h \geq p$

$$\begin{pmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(h-1) \\ \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(h-2) \\ \vdots & \vdots & & \vdots \\ \gamma_X(h-1) & \gamma_X(h-2) & \cdots & \gamma_X(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \gamma_X(1) \\ \gamma_X(2) \\ \vdots \\ \gamma_X(h) \end{pmatrix}.$$

This expresses the vector on the right as a linear combination of the first p columns of the matrix Γ_h on the left. We can use this to rewrite $\det\Gamma_{h+1}$ (by a "sweeping" operation) in the form

$$\begin{vmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \gamma_X(h) \\ \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(h-1) \\ \vdots & \vdots & & \vdots \\ \gamma_X(h) & \gamma_X(h-1) & \cdots & \gamma_X(0) \end{vmatrix}$$
$$= \begin{vmatrix} \gamma_X(0) - \phi_1\gamma_X(1) - \cdots - \phi_p\gamma_X(p) & 0 & \cdots & 0 \\ \gamma_X(1) & \gamma_X(0) & \cdots & \gamma_X(h-1) \\ \vdots & \vdots & & \vdots \\ \gamma_X(h) & \gamma_X(h-1) & \cdots & \gamma_X(0) \end{vmatrix}.$$

The $(1,1)$-element in the last determinant is equal to σ^2 by (10.1). Thus this determinant is equal to $\sigma^2 \det\Gamma_h$ and the theorem follows. ∎

This corollary is used often when choosing the order p if fitting an auto-regressive model to a given observed time series. The true partial auto-correlation coefficients of lags higher than the true order p are all zero. When we estimate these coefficients by the sample auto-correlation coefficients, then we should expect that the estimates are inside a band of the type $(-2/\sqrt{n}, 2\sqrt{n})$. Thus we should not choose the order equal to p if $\hat{\alpha}_n(p+k)$ is outside this band for too many $k \geq 1$. Here we should expect a fraction of 5 % of the $\hat{\alpha}_n(p+k)$ for which we perform this "test" to be outside the band in any case.

To turn this procedure in a more formal statistical test we must also take the dependence between the different $\hat{\alpha}_n(p+k)$ into account, but this appears to be complicated.

* **10.9** EXERCISE. Find the asymptotic limit distribution of the sequence $\bigl(\hat{\alpha}_n(h), \hat{\alpha}_n(h+1)\bigr)$ for $h > p$, e.g. in the case that $p = 0$ and $h = 1$.

* 10.1.3 Indirect Estimation

The parameters $\phi_1, \ldots, \phi_p$ of a causal auto-regressive process are exactly the coefficients of the one-step ahead linear predictor using p variables from the past. This makes application of the least squares method to obtain estimators for these parameters particularly straightforward. For an arbitrary stationary time series the best linear predictor of X_{p+1} given $1, X_1, \ldots, X_p$ is the linear combination $\mu + \phi_1(X_p - \mu) + \cdots + \phi_1(X_1 - \mu)$ whose coefficients satisfy the prediction equations (2.1). The Yule-Walker estimators are the solutions to these equations after replacing the true auto-covariances by the sample auto-covariances. It follows that the Yule-Walker estimators can be considered estimators for the prediction coefficients (using p variables from the past) for any stationary time series. The case of auto-regressive processes is special only in that these prediction coefficients are exactly the parameters of the model.

Furthermore, it remains true that the Yule-Walker estimators are $\sqrt{n}$-consistent and asymptotically normal. This does not follow from Theorem 10.3, because this uses the auto-regressive structure explicitly, but it can be inferred from the asymptotic normality of the auto-covariances, given in Theorem 5.7. (The argument is the same as used in Section 5.4. The asymptotic covariance matrix will be different from the one in Theorem 10.3, and more complicated.)

If the prediction coefficients (using a fixed number of past variables) are not the parameters of main interest, then these remarks may seem little useful. However, if the parameter of interest θ is of dimension d, then we may hope that there exists a one-to-one relationship between θ and the prediction coefficients $\phi_1, \ldots, \phi_p$ if we choose $p = d$. (More generally, we can apply this to a subvector of θ and a matching number of ϕ_j's.) Then we can first estimate $\phi_1, \ldots, \phi_d$ by the Yule-Walker estimators and next employ the relationshiop between $\phi_1, \ldots, \phi_p$ to infer an estimate of θ. If the inverse map giving θ as a function of $\phi_1, \ldots, \phi_d$ is differentiable, then it follows by the Delta-method that the resulting estimator for θ is $\sqrt{n}$-consistent and asymptotically normal, and hence we obtain good estimators.

If the relationship between θ and $(\phi_1, \ldots, \phi_d)$ is complicated, then this idea may be hard to implement. One way out of this problem is to determine the prediction coefficients

$\phi_1, \ldots, \phi_d$ for a grid of values of θ, possibly through simulation. The value on the grid that yields the Yule-Walker estimators is the estimator for θ we are looking for.

10.10 EXERCISE. Indicate how you could obtain (approximate) values for $\phi_1, \ldots, \phi_p$ given θ using computer simulation, for instance for a stochastic volatility model.

10.2 Moment Estimators

The Yule-Walker estimators can be viewed as arising from a comparison of sample auto-covariances to true auto-covariances and therefore are examples of *moment estimators*. Moment estimators are defined in general by matching sample moments and population moments. Population moments of a time series X_t are true expectations of functions of the variables X_t, for instance,

$$\mathrm{E}_\theta X_t, \quad \mathrm{E}_\theta X_t^2, \quad \mathrm{E}_\theta X_{t+h}X_t, \quad \mathrm{E}_\theta X_{t+h}^2 X_t^2.$$

In every case, the subscript θ indicates the dependence on the unknown parameter θ: in principle, every of these moments is a function of θ. The principle of the method of moments is to estimate θ by that value $\hat\theta_n$ for which the corresponding population moments coincide with a corresponding sample moment, for instance,

$$\frac{1}{n}\sum_{t=1}^n X_t, \quad \frac{1}{n}\sum_{t=1}^n X_t^2, \quad \frac{1}{n}\sum_{t=1}^n X_{t+h}X_t, \quad \frac{1}{n}\sum_{t=1}^n X_{t+h}^2 X_t^2.$$

From Chapter 5 we know that these sample moments converge, as $n \to \infty$, to the true moments, and hence it is believable that the sequence of moment estimators $\hat\theta_n$ also converges to the true parameter, under some conditions.

Rather than true moments it is often convenient to define moment estimators through derived moments such as an auto-covariance at a fixed lag, or an auto-correlation, which are both functions of moments of degree smaller than 2. These derived moments are then matched by the corresponding sample quantities.

The choice of moments to be used is crucial for the existence and consistency of the moment estimators, and also for their efficiency.

For existence we shall generally need to match as many moments as there are parameters in the model. If not, then we should expect that a moment estimator is not uniquely defined if we use fewer moments, and we should expect to find no solution to the moment equations if we try and match too many moments. Because in general the moments are highly nonlinear functions of the parameters, it is hard to make this statement precise, as it is hard to characterize solutions of systems of nonlinear equations in general. This is illustrated already in the case of moving average processes, where a characterization of the existence of solutions requires effort, and where conditions and restrictions are needed to ensure their uniqueness. (Cf. Section 10.2.1.)

To ensure consistency and improve efficiency it is necessary to use moments that can be estimated well from the data. Auto-covariances at high lags, or moments of high degree should generally be avoided. Besides on the quality of the initial estimates of the population moments, the efficiency of the moment estimators also depends on the inverse map giving the parameter as a function of the moments. To see this we may formalize the method of moments through the scheme

$$\begin{aligned}\phi(\theta) &= \mathrm{E}_\theta f(X_t,\ldots,X_{t+h}),\\ \phi(\hat\theta_n) &= \frac{1}{n}\sum_{t=1}^n f(X_t,\ldots,X_{t+h}).\end{aligned}$$

Here $f\colon \mathbb{R}^{h+1}\to\mathbb{R}^d$ is a given map, which defines the moments used. (For definiteness we allow it to depend on the joint distribution of at most $h+1$ consecutive observations.) We assume that the time series $t\mapsto f(X_t,\ldots,X_{t+h})$ is strictly stationary, so that the mean values $\phi(\theta)$ in the first line do not depend on t, and for simplicity of notation we assume that we observe $X_1,\ldots,X_{n+h}$, so that the right side of the second line is indeed an observable quantity. We shall assume that the map $\phi\colon\Theta\to\mathbb{R}^d$ is one-to-one, so that the second line uniquely defines the estimator $\hat\theta_n$ as the inverse

$$\hat\theta_n = \phi^{-1}\bigl(\hat f_n\bigr),\qquad \hat f_n = \frac{1}{n}\sum_{t=1}^n f(X_t,\ldots,X_{t+h}).$$

We shall generally construct $\hat f_n$ such that it converges in probability to its mean $\phi(\theta)$ as $n\to\infty$. If this is the case and ϕ^{-1} is continuous at $\phi(\theta)$, then we have that $\hat\theta_n\to\phi^{-1}\phi(\theta)=\theta$, in probability as $n\to\infty$, and hence the moment estimator is asymptotically consistent.

Many sample moments converge at $\sqrt{n}$-rate, with a normal limit distribution. This allows to refine the consistency result, in view of the Delta-method, given by Theorem 3.15. If ϕ^{-1} is differentiable at $\phi(\theta)$ and $\sqrt{n}\bigl(\hat f_n-\phi(\theta)\bigr)$ converges in distribution to a normal distribution with mean zero and covariance matrix Σ_θ, then

$$\sqrt{n}(\hat\theta_n-\theta)\rightsquigarrow N\bigl(0,{\phi'_\theta}^{-1}\Sigma_\theta({\phi'_\theta}^{-1})^T\bigr).$$

Here ${\phi'_\theta}^{-1}$ is the derivative of ϕ^{-1} at $\phi(\theta)$, which is the inverse of the derivative of ϕ at θ, assumed to be nonsingular. We conclude that, under these conditions, the moment estimators are $\sqrt{n}$-consistent with a normal limit distribution, a desirable property.

A closer look concerns the size of the asymptotic covariance matrix ${\phi'_\theta}^{-1}\Sigma_\theta({\phi'_\theta}^{-1})^T$. Clearly, it depends both on the accuracy by which the chosen moments can be estimated from the data (through the matrix Σ_θ) and the "smoothness" of the inverse ϕ^{-1}. If the inverse map has a "large" derivative, then extracting the moment estimator $\hat\theta_n$ from the sample moments $\hat f_n$ magnifies the error of $\hat f_n$ as an estimate of $\phi(\theta)$, and the moment estimator will be relatively inefficient. Unfortunately, it is hard to see how a particular implementation of the method of moments works out without doing (part of) the algebra leading to the asymptotic covariance matrix. Furthermore, the outcome may depend on the true value of the parameter, a given moment estimator being relatively efficient for some parameter values, but (very) inefficient for others.

Moment estimators are measurable functions of the sample moments $\hat{f}_n$ and hence cannot be better than the "best" estimator based on $\hat{f}_n$. In most cases summarizing the data through the sample moments $\hat{f}_n$ incurs a loss of information. Only if the sample moments are sufficient (in the statistical sense), moment estimators can be fully efficient for estimating the parameters. This is an exceptional situation. The loss of information can be controlled somewhat by working with the right type of moments, but is usually unavoidable through the restriction of using only as many moments as there are parameters. The reduction of a sample of size n to a "sample" of empirical moments of size d usually entails a loss of information.

This observation motivates the *generalized method of moments.* The idea is to reduce the sample to more "empirical moments" than there are parameters. Given a function $f\colon \mathbb{R}^{h+1} \to \mathbb{R}^e$ for $e > d$ with corresponding mean function $\phi(\theta) = \mathrm{E}_\theta f(X_t, \ldots, X_{t+h})$, there is no hope, in general, to solve an estimator $\hat{\theta}_n$ from the system of equations $\phi(\theta) = \hat{f}_n$, because these are $e > d$ equations in d unknowns. The generalized method of moments overcomes this by defining $\hat{\theta}_n$ as the minimizer of the quadratic form, for a given (possibly random) matrix $\hat{V}_n$,

$$\theta \mapsto \bigl(\phi(\theta) - \hat{f}_n\bigr)^T \hat{V}_n \bigl(\phi(\theta) - \hat{f}_n\bigr). \tag{10.2}$$

Thus a generalized moment estimator tries to solve the system of equations $\phi(\theta) = \hat{f}_n$ as well as possible, where the discrepancy is measured through a certain quadratic form. The matrix $\hat{V}_n$ weighs the influence of the different components of $\hat{f}_n$ on the estimator $\hat{\theta}_n$, and is typically chosen dependent on the data to increase the efficiency of the generalized moment estimator. We assume that $\hat{V}_n$ is symmetric and positive-definite.

As $n \to \infty$ the estimator $\hat{f}_n$ typically converges to its expectation under the true parameter, which we shall denote by θ_0 for clarity. If we replace $\hat{f}_n$ in the criterion function by its expectation $\phi(\theta_0)$, then we can reduce the resulting quadratic form to zero by choosing θ equal to θ_0. This is clearly the minimal value of the quadratic form, and the choice $\theta = \theta_0$ will be unique as soon as the map ϕ is one-to-one. This suggests that the generalized moment estimator $\hat{\theta}_n$ is asymptotically consistent. As for ordinary moment estimators, a rigorous justification of the consistency must take into account the properties of the function ϕ.

The distributional limit properties of a generalized moment estimator can be understood by linearizing the function ϕ around the true parameter. Insertion of the first order Taylor expansion $\phi(\theta) = \phi(\theta_0) + \phi'_{\theta_0}(\theta - \theta_0)$ into the quadratic form yields the approximate criterion

$$\begin{aligned}\theta \mapsto &\bigl(\hat{f}_n - \phi(\theta_0) - \phi'_{\theta_0}(\theta - \theta_0)\bigr)^T \hat{V}_n \bigl(\hat{f}_n - \phi(\theta_0) - \phi'_{\theta_0}(\theta - \theta_0)\bigr)\\ &= \frac{1}{n}\bigl(Z_n - \phi'_{\theta_0}\sqrt{n}(\theta - \theta_0)\bigr)^T \hat{V}_n \bigl(Z_n - \phi'_{\theta_0}\sqrt{n}(\theta - \theta_0)\bigr),\end{aligned}$$

for $Z_n = \sqrt{n}\bigl(\hat{f}_n - \phi(\theta_0)\bigr)$. The sequence Z_n is typically asymptotically normally distributed, with mean zero. Minimization of this approximate criterion over $h = \sqrt{n}(\theta - \theta_0)$

is equivalent to minimizing the quadratic form $h \mapsto (Z_n - \phi'_{\theta_0} h)\hat{V}_n (Z_n - \phi'_{\theta_0} h)$, or equivalently minimizing the norm of the vector $Z_n - \phi'_{\theta_0} h$ over h in the Hilbert space $\mathbb{R}^d$ with inner product defined by $\langle x, y\rangle = x^T \hat{V}_n y$. This comes down to projecting the vector Z_n onto the range of the linear map ϕ'_{θ_0} and hence by the projection theorem, Theorem 2.10, the minimizer $\hat{h} = \sqrt{n}(\hat{\theta} - \theta_0)$ is characterized by the orthogonality of the vector $Z_n - \phi'_{\theta_0}\hat{h}$ to the range of ϕ'_{θ_0}. The algebraic expression of this orthogonality is that $(\phi'_{\theta_0})^T (Z_n - \phi'_{\theta_0}\hat{h}) = 0$, which can be rewritten in the form

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \bigl((\phi'_{\theta_0})^T \hat{V}_n \phi'_{\theta_0}\bigr)^{-1} (\phi'_{\theta_0})^T \hat{V}_n Z_n.$$

This readily gives the asymptotic normality of the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$, with mean zero and a somewhat complicated covariance matrix depending on ϕ'_{θ_0}, $\hat{V}_n$ and the asymptotic covariance matrix of Z_n.

The best nonrandom weight matrices $\hat{V}_n$, in terms of minimizing the asymptotic covariance of $\sqrt{n}(\hat{\theta}_n - \theta)$, is the inverse of the covariance matrix of Z_n. (Cf. Problem 10.11.) For our present situation this suggests to choose the matrix $\hat{V}_n$ to be consistent for the inverse of the asymptotic covariance matrix of the sequence $Z_n = \sqrt{n}\bigl(\hat{f}_n - \phi(\theta_0)\bigr)$. With this choice and the asymptotic covariance matrix denoted by Σ_{θ_0}, we may expect that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\Bigl(0, \bigl((\phi'_{\theta_0})^T \Sigma_{\theta_0}^{-1} \phi'_{\theta_0}\bigr)^{-1}\Bigr).$$

The argument shows that the generalized moment estimator can be viewed as a weighted least squares estimators for regressing $\sqrt{n}\bigl(\hat{f}_n - \phi(\theta_0)\bigr)$ onto ϕ'_{θ_0}. With the optimal weighting matrix it is the best such estimator. If we use more initial moments to define $\hat{f}_n$ and hence $\phi(\theta)$, then we add "observations" and corresponding rows to the design matrix ϕ'_{θ_0}, but keep the same parameter $\sqrt{n}(\theta - \theta_0)$. This suggests that the asymptotic efficiency of the optimally weigthed generalized moment estimator increases if we use a longer vector of initial moments $\hat{f}_n$. In particular, the optimally weigthed generalized moment estimator is more efficient than an ordinary moment estimator based on a subset of d of the initial moments. Thus, the generalized method of moments achieves the aim of using more information contained in the observations.

10.11 EXERCISE. Let Σ be a symmetric, positive-definite matrix and A a given matrix. Show that the matrix $(A^T V A)^{-1} A^T V \Sigma A (A^T V A)^{-1}$ is minimized over nonnegative-definite matrices V (where we say that $V \leq W$ if $W - V$ is nonnegative definite) for $V = \Sigma^{-1}$. [The given matrix is the covariance matrix of $\beta_A = (A^T V A)^{-1} A^T V Z$ for Z a random vector with the normal distribution with covariance matrix Σ. Show that $\mathrm{Cov}(\beta_A - \beta_{\Sigma^{-1}}, \beta_{\Sigma^{-1}}) = 0$.]

These arguments are based on asymptotic approximations. They are reasonably accurate for values of n that are large relative to the values of d and e, but should not be applied if d or e are large. In particular, it is illegal to push the preceding argument to its extreme and infer that is necessarily right to use as many initial moments as possible. Increasing the dimension of the vector $\hat{f}_n$ indefinitely may contribute more "variability" to the criterion (and hence to the estimator) without increasing the information much, depending on the accuracy of the estimator $\hat V_n$.

The implementation of the (generalized) method of moments requires that the expectations $\phi(\theta) = \mathrm{E}_\theta f(X_t,\ldots,X_{t+h})$ are available as functions of θ. In some models, such as AR or MA models, this causes no difficulty, but already in ARMA models the required analytical computations become complicated. Sometimes it is easy to simulate realizations of a time series, but hard to compute explicit formulas for moments. In this case the values $\phi(\theta)$ may be estimated stochastically for a grid of values of θ by simulating realizations of the given time series, taking in turn each of the grid points as the "true" parameter, and next computing the empirical moment for the simulated time series. If the grid is sufficiently dense and the simulations are sufficiently long, then the grid point for which the simulated empirical moment matches the empirical moment of the data is close to the moment estimator. Taking it to be the moment estimator is called the *simulated method of moments*.

In the following theorem we make the preceding informal derivation of the asymptotics of generalized moment estimators rigorous. The theorem is a corollary of Theorems 3.17 and 3.18 on the asymptotics of general minimum contrast estimators. Consider generalized moment estimators as previously, defined as the point of minimum of a quadratic form of the type (10.2). In most cases the function $\phi(\theta)$ will be the expected value of the random vectors $\hat f_n$ under the parameter θ, but this is not necessary. The following theorem is applicable as soon as $\phi(\theta)$ gives a correct "centering" to ensure that the sequenc $\sqrt{n}\bigl(\hat f_n - \phi(\theta)\bigr)$ converges to a limit distribution, and hence may also apply to nonstationary time series.

10.12 Theorem. *Let $\hat V_n$ be random matrices such that $\hat V_n \xrightarrow{\mathrm{P}} V_0$ for some matrix V_0. Assume that $\phi\colon \Theta\subset\mathbb{R}^d\to\mathbb{R}^e$ is differentiable at an inner point θ_0 of Θ with derivative ϕ'_{θ_0} such that the matrix $(\phi'_{\theta_0})^T V_0\phi'_{\theta_0}$ is nonsingular and satisfies, for every $\delta>0$,*

$$\inf_{\theta:\|\theta-\theta_0\|>\delta}\bigl(\phi(\theta)-\phi(\theta_0)\bigr)^T V_0\bigl(\phi(\theta)-\phi(\theta_0)\bigr)>0.$$

Assume either that V_0 is invertible or that the set $\{\phi(\theta)\colon\theta\in\Theta\}$ is bounded. Finally, suppose that the sequence of random vectors $Z_n=\sqrt{n}\bigl(\hat f_n-\phi(\theta_0)\bigr)$ is uniformly tight. If $\hat\theta_n$ are random vectors that minimize the criterion (10.2), then $\sqrt{n}(\hat\theta_n-\theta_0) = -((\phi'_{\theta_0})^T V_0\phi'_{\theta_0})^{-1}V_0 Z_n + o_P(1)$.

Proof. We first prove that $\hat\theta_n\xrightarrow{\mathrm{P}}\theta_0$ using Theorem 3.17, with the criterion functions

$$\begin{aligned}\mathbb{M}_n(\theta)&=\bigl\|\hat V_n^{1/2}\bigl(\hat f_n-\phi(\theta)\bigr)\bigr\|,\\ M_n(\theta)&=\bigl\|\hat V_n^{1/2}\bigl(\phi(\theta)-\phi(\theta_0)\bigr)\bigr\|.\end{aligned}$$

The squares of these functions are the criterion in (10.2) and the quadratic form in the display of the theorem, but with V_0 replaced by $\hat V_n$, respectively. By the triangle inequality $|\mathbb{M}_n(\theta)-M_n(\theta)|\le\bigl\|\hat V_n^{1/2}\bigl(\hat f_n-\phi(\theta_0)\bigr)\bigr\|\to 0$ in probability, uniformly in θ. Thus the first condition of Theorem 3.17 is satisfied. The second condition, that

$\inf\{M_n(\theta)\colon \|\theta-\theta_0\|>\delta\}$ is stochastically bounded away from zero for every $\delta>0$, is satisfied by assumption in the case that $\hat V_n=V_0$ is fixed. Because $\hat V_n \xrightarrow{\mathrm{P}} V_0$, where V_0 is invertible or the set $\{\phi(\theta)\colon \|\theta-\theta_0\|>\delta\}$ is bounded, it is also satisfied in the general case, in view of Exercise 10.13. This concludes the proof of consistency of $\hat\theta_n$.

For the proof of asymptotic normality we use Theorem 3.18 with the criterion functions $\mathbb{M}_n$ and M_n redefined as the squares of the functions $\mathbb{M}_n$ and M_n as used in the consistency proof (so that $\mathbb{M}_n(\theta)$ is the criterion function in (10.2)) and with the centering function M defined by

$$M(\theta)=\bigl(\phi(\theta)-\phi(\theta_0)\bigr)^T V_0\bigl(\phi(\theta)-\phi(\theta_0)\bigr).$$

It follows that, for any random sequence $\tilde\theta_n \xrightarrow{\mathrm{P}} \theta_0$,

$$\begin{aligned}
&n(\mathbb{M}_n-M_n)(\tilde\theta_n)-n(\mathbb{M}_n-M_n)(\theta_0)\\
&\quad=\bigl(Z_n-\sqrt{n}(\phi(\tilde\theta_n)-\phi(\theta_0))\bigr)^T\hat V_n\bigl(Z_n-\sqrt{n}(\phi(\tilde\theta_n)-\phi(\theta_0))\bigr)\\
&\qquad\qquad-\sqrt{n}\bigl(\phi(\tilde\theta_n)-\phi(\theta_0)\bigr)^T\hat V_n\sqrt{n}\bigl(\phi(\tilde\theta_n)-\phi(\theta_0)\bigr)-Z_n^T\hat V_nZ_n,\\
&\quad=-2\sqrt{n}\bigl(\phi(\tilde\theta_n)-\phi(\theta_0)\bigr)^T\hat V_nZ_n,\\
&\quad=-2(\tilde\theta_n-\theta_0)^T(\phi'_{\theta_0})^T\hat V_nZ_n+o_P(\tilde\theta_n-\theta_0),
\end{aligned}$$

by the differentiability of ϕ at θ_0. Together with the convergence of $\hat V_n$ to V_0, the differentiability of ϕ also gives that $M_n(\tilde\theta_n)-M(\tilde\theta_n)=o_P\bigl(\|\tilde\theta_n-\theta_0\|^2\bigr)$ for any sequence $\tilde\theta_n \xrightarrow{\mathrm{P}} \theta_0$. Therefore, we may replace M_n by M in the left side of the preceding display, if we add an $o_P\bigl(\|\tilde\theta_n-\theta_0\|^2\bigr)$-term on the right. By a third application of the differentiability of ϕ, the function M permits the two-term Taylor expansion $M(\theta)=(\theta-\theta_0)^TW(\theta-\theta_0)+o(\theta-\theta_0)^2$, for $W=(\phi'_{\theta_0})^TV_0\phi'_{\theta_0}$. Thus the conditions of Theorem 3.18 are satisfied and the proof of asymptotic normality is complete. ■

10.13 EXERCISE. Let V_n be a sequence of nonnegative-definite matrices such that $V_n\to V$ for a matrix V such that $\inf\{x^TVx\colon x\in C\}>0$ for some set C. Show that:
(i) If V is invertible, then $\liminf\inf\{x^TV_nx\colon x\in C\}>0$.
(ii) If C is bounded, then $\liminf\inf\{x^TV_nx\colon x\in C\}>0$.
(iii) The assertion of (i)-(ii) may fail without some additional assumption.
[Suppose that $x_n^TV_nx_n\to0$. If V is invertible, then it follows that $x_n\to0$. If the sequence x_n is bounded, then $x_n^TVx_n-x_n^TV_nx_n\to0$. As counterexample let V_n be the matrices with eigenvectors propertional to $(n,1)$ and $(-1,n)$ and eigenvalues 1 and 0, let $C=\{x\colon|x_1|>\delta\}$ and let $x_n=\delta(-1,n)$.]

10.2.1 Moving Average Processes

Suppose that $X_t-\mu=\sum_{j=0}^q\theta_jZ_{t-j}$ is a moving average process of order q. For simplicity of notation assume that $1=\theta_0$ and define $\theta_j=0$ for $j<0$ or $j>q$. Then the auto-covariance function of X_t can be written in the form

$$\gamma_X(h)=\sigma^2\sum_j\theta_j\theta_{j+h}.$$

Given observations $X_1, \ldots, X_n$ we can estimate $\gamma_X(h)$ by the sample auto-covariance function and next obtain estimators for $\sigma^2, \theta_1, \ldots, \theta_q$ by solving from the system of equations

$$\hat\gamma_n(h) = \hat\sigma^2 \sum_j \hat\theta_j \hat\theta_{j+h}, \qquad h = 0, 1, \ldots, q.$$

A solution of this system, which has $q+1$ equations with $q+1$ unknowns, does not necessarily exist, or may be nonunique. It cannot be derived in closed form, but must be determined numerically by an iterative method. Thus applying the method of moments for moving average processes is considerably more involved than for auto-regressive processes. The real drawback of this method is, however, that the moment estimators are less efficient than the least squares estimators that we discuss later in this chapter. Moment estimators are therefore at best only used as starting points for numerical procedures to compute other estimators.

10.14 Example (MA(1)). For the moving average process $X_t = Z_t + \theta Z_{t-1}$ the moment equations are

$$\gamma_X(0) = \sigma^2(1+\theta^2), \qquad \gamma_X(1) = \theta\sigma^2.$$

Replacing γ_X by $\hat\gamma_n$ and solving for σ^2 and θ yields the moment estimators

$$\hat\theta_n = \frac{1 \pm \sqrt{1 - 4\hat\rho_n^2(1)}}{2\hat\rho_n(1)}, \qquad \hat\sigma^2 = \frac{\hat\gamma_n(1)}{\hat\theta_n}.$$

We obtain a real solution for $\hat\theta_n$ only if $|\hat\rho_n(1)| \le 1/2$. Because the true auto-correlation $\rho_X(1)$ is contained in the interval $[-1/2, 1/2]$, it is reasonable to truncate the sample auto-correlation $\hat\rho_n(1)$ to this interval and then we always have some solution. If $|\hat\rho_n(1)| < 1/2$, then there are two solutions for $\hat\theta_n$, corresponding to the ± sign. This situation will happen with probability tending to one if the true auto-correlation $\rho_X(1)$ is strictly contained in the interval $(-1/2, 1/2)$. From the two solutions, one solution has $|\hat\theta_n| < 1$ and corresponds to an invertible moving average process; the other solution has $|\hat\theta_n| > 1$. The existence of multiple solutions was to be expected in view of Theorem 7.27.

Assume that the true value $|\theta| < 1$, so that $\rho_X(1) \in (-1/2, 1/2)$ and

$$\theta = \frac{1 - \sqrt{1 - 4\rho_X^2(1)}}{2\rho_X(1)}.$$

Of course, we use the estimator $\hat\theta_n$ defined by the minus sign. Then $\hat\theta_n - \theta$ can be written as $\phi\bigl(\hat\rho_n(1)\bigr) - \phi\bigl(\rho_X(1)\bigr)$ for the function ϕ given by

$$\phi(\rho) = \frac{1 - \sqrt{1 - 4\rho^2}}{2\rho}.$$

This function is differentiable on the interval $(-1/2, 1/2)$. By the Delta-method the limit distribution of the sequence $\sqrt{n}(\hat\theta_n - \theta)$ is the same as the limit distribution of the sequence $\phi'\bigl(\rho_X(1)\bigr)\sqrt{n}\bigl(\hat\rho_n(1)-\rho_X(1)\bigr)$. Using Theorem 5.8 we obtain, after a long calculation, that

$$\sqrt{n}(\hat\theta_n-\theta)\rightsquigarrow N\Bigl(0,\frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{(1-\theta^2)^2}\Bigr).$$

Thus, to a certain extent, the method of moments works: the moment estimator $\hat\theta_n$ converges at a rate of $1/\sqrt{n}$ to the true parameter. However, the asymptotic variance is large, in particular for $\theta\approx 1$. We shall see later that there exist estimators with asymptotic variance $1-\theta^2$, which is smaller for every θ, and is particularly small for $\theta\approx 1$. □

10.15 EXERCISE. Derive the formula for the asymptotic variance, or at least convince yourself that you know how to get it.

The asymptotic behaviour of the moment estimators for moving averages of order higher than 1 can be analysed, as in the preceding example, by the Delta-method as well. Define $\phi:\mathbb{R}^{q+1}\to\mathbb{R}^{q+1}$ by

$$\phi\begin{pmatrix}\sigma^2\\ \theta_1\\ \vdots\\ \theta_q\end{pmatrix}=\sigma^2\begin{pmatrix}\sum_j\theta_j^2\\ \sum_j\theta_j\theta_{j+1}\\ \vdots\\ \sum_j\theta_j\theta_{j+q}\end{pmatrix}.$$

Then the moment estimators and true parameters satisfy

$$\begin{pmatrix}\hat\sigma^2\\ \hat\theta_1\\ \vdots\\ \hat\theta_q\end{pmatrix}=\phi^{-1}\begin{pmatrix}\hat\gamma_n(0)\\ \hat\gamma_X(1)\\ \vdots\\ \hat\gamma_n(q)\end{pmatrix},\qquad\begin{pmatrix}\sigma^2\\ \theta_1\\ \vdots\\ \theta_q\end{pmatrix}=\phi^{-1}\begin{pmatrix}\gamma_X(0)\\ \gamma_X(1)\\ \vdots\\ \gamma_X(q)\end{pmatrix}.$$

The joint limit distribution of the sequences $\sqrt{n}\bigl(\hat\gamma_n(h)-\gamma_X(h)\bigr)$ is known from Theorem 5.7. Therefore, the limit distribution of the moment estimators $\hat\sigma^2,\hat\theta_1,\ldots,\hat\theta_q$ follows by the Delta-method, provided the map ϕ^{-1} is differentiable at $\bigl(\gamma_X(0),\ldots,\gamma_X(q)\bigr)$.

Practical and theoretical complications arise from the fact that the moment equations may have zero or multiple solutions, as illustrated in the preceding example. This difficulty disappears if we insist on an invertible representation of the moving average process, i.e. require that the polynomial $1+\theta_1z+\cdots+\theta_qz^q$ has no roots in the complex unit disc. This follows by the following lemma, whose proof also contains an algorithm to compute the moment estimators numerically.

10.16 Lemma. *Let $\Theta\subset\mathbb{R}^q$ be the set of all vectors $(\theta_1,\ldots,\theta_q)$ such that all roots of $1+\theta_1z+\cdots+\theta_qz^q$ are outside the unit circle. Then the map $\phi:\mathbb{R}^+\times\Theta$ is one-to-one and continuously differentiable. Furthermore, the map ϕ^{-1} is differentiable at every point $\phi(\sigma^2,\theta_1,\ldots,\theta_q)$ for which the roots of $1+\theta_1z+\cdots+\theta_qz^q$ are distinct.*

* **Proof.** Abbreviate $\gamma_h = \gamma_X(h)$. The system of equations $\sigma^2 \sum_j \theta_j\theta_{j+h} = \gamma_h$ for $h = 0, \ldots, q$ implies that

$$\sum_{h=-q}^{q} \gamma_h z^h = \sigma^2 \sum_h \sum_j \theta_j \theta_{j+h} z^h = \sigma^2 \theta(z^{-1})\theta(z).$$

For any $h \geq 0$ the function $z^h + z^{-h}$ can be expressed as a polynomial of degree h in $w = z + z^{-1}$. For instance, $z^2 + z^{-2} = w^2 - 2$ and $z^3 + z^{-3} = w^3 - 3w$. The case of general h can be treated by induction, upon noting that by rearranging Newton's binomial formula

$$z^{h+1} + z^{-h-1} - w^{h+1} = -\binom{h+1}{(h+1)/2} - \sum_{j\neq 0} \binom{h+1}{(h+1-j)/2} (z^j + z^{-j}).$$

Thus the left side of the preceding display can be written in the form

$$\gamma_0 + \sum_{h=1} \gamma_j (z^j + z^{-j}) = a_0 + a_1 w + \cdots + a_q w^q,$$

for certain coefficients $(a_0, \ldots, a_q)$. Let $w_1, \ldots, w_q$ be the zeros of the polynomial on the right, and for each j let η_j and η_j^{-1} be the solutions of the quadratic equation $z + z^{-1} = w_j$. Choose $|\eta_j| \geq 1$. Then we can rewrite the right side of the preceding display as

$$a_q \prod_{j=1}^{q} (z + z^{-1} - w_j) = a_q (z - \eta_j)(\eta_j - z^{-1})\eta_j^{-1}.$$

On comparing this to the first display of the proof, we see that $\eta_1, \ldots, \eta_q$ are the zeros of the polynomial $\theta(z)$. This allows us to construct a map

$$(\gamma_0, \ldots, \gamma_q) \mapsto (a_0, \ldots, a_q) \mapsto (w_1, \ldots, w_q, a_q) \mapsto (\eta_1, \ldots, \eta_q, a_q) \mapsto (\theta_1, \ldots, \theta_q, \sigma^2).$$

If restricted to the range of ϕ this is exactly the map ϕ^{-1}. It is not hard to see that the first and last step in this decomposition of ϕ^{-1} are analytic functions. The two middle steps concern mapping coefficients of polynomials into their zeros.

For $\alpha = (\alpha_0, \ldots, \alpha_q) \in \mathbb{C}^{q+1}$ let $p_\alpha(w) = \alpha_0 + \alpha_1 w + \cdots + \alpha_q w^q$. By the implicit function theorem for functions of several complex variables we can show the following. If for some α the polynomial p_α has a root of order 1 at a point w_α, then there exists neighbourhoods U_α and V_α of α and w_α such that for every $\beta \in U_\alpha$ the polynomial p_β has exactly one zero $w_\beta \in V_\alpha$ and the map $\beta \mapsto w_\beta$ is analytic on U_α. Thus, under the assumption that all roots are or multiplicity one, the roots can be viewed as analytic functions of the coefficients. If θ has distinct roots, then $\eta_1, \ldots, \eta_q$ are of multiplicity one and hence so are $w_1, \ldots, w_q$. In that case the map is analytic. ■

* 10.2.2 Moment Estimators for ARMA Processes

If $X_t - \mu$ is a stationary ARMA process satisfying $\phi(B)(X_t - \mu) = \theta(B)Z_t$, then

$$\operatorname{cov}\bigl(\phi(B)(X_t - \mu), X_{t-k}\bigr) = \mathrm{E}\bigl(\theta(B)Z_t\bigr)X_{t-k}.$$

If $X_t - \mu$ is a causal, stationary ARMA process, then the right side vanishes for $k > q$. Working out the left side, we obtain the eqations

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = 0, \qquad k > q.$$

For $k = q+1, \ldots, q+p$ this leads to the system

$$\begin{pmatrix} \gamma_X(q) & \gamma_X(q-1) & \cdots & \gamma_X(q-p+1) \\ \gamma_X(q+1) & \gamma_X(q) & \cdots & \gamma_X(q-p+2) \\ \vdots & \vdots & & \vdots \\ \gamma_X(q+p-1) & \gamma_X(q+p-2) & \cdots & \gamma_X(q) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma_X(q+1) \\ \gamma_X(q+2) \\ \vdots \\ \gamma_X(q+p) \end{pmatrix}.$$

These are the *Yule-Walker equations* for general stationary ARMA processes and may be used to obtain estimators $\hat\phi_1, \ldots, \hat\phi_p$ of the auto-regressive parameters in the same way as for auto-regressive processes: we replace γ_X by $\hat\gamma_n$ and solve for $\phi_1, \ldots, \phi_p$.

Next we apply the method of moments for moving averages to the time series $Y_t = \theta(B)Z_t$ to obtain estimators for the parameters $\sigma^2, \theta_1, \ldots, \theta_q$. Because also $Y_t = \phi(B)(X_t - \mu)$ we can estimate the covariance function γ_Y from

$$\gamma_Y(h) = \sum_i \sum_j \tilde\phi_i \tilde\phi_j \gamma_X(h + i - j), \qquad \text{if } \phi(z) = \sum_j \tilde\phi_j z^j.$$

Let $\hat\gamma_Y(h)$ be the estimators obtained by replacing the unknown parameters $\tilde\phi_j$ and $\gamma_X(h)$ by their moment estimators and sample moments, respectively. Next we solve $\hat\sigma^2, \hat\theta_1, \ldots, \hat\theta_q$ from the system of equations

$$\hat\gamma_Y(h) = \hat\sigma^2 \sum_j \hat\theta_j \hat\theta_{j+h}, \qquad h = 0, 1, \ldots, q.$$

As is explained in the preceding section, if $X_t - \mu$ is invertible, then the solution is unique, with probability tending to one, if the coefficients $\theta_1, \ldots, \theta_q$ are restricted to give an invertible stationary ARMA process.

The resulting estimators $(\hat\sigma^2, \hat\theta_1, \ldots, \hat\theta_q, \hat\phi_1, \ldots, \hat\phi_p)$ can be written as a function of $\bigl(\hat\gamma_n(0), \ldots, \hat\gamma_n(q+p)\bigr)$. The true values of the parameters can be written as the same function of the vector $\bigl(\gamma_X(0), \ldots, \gamma_X(q+p)\bigr)$. In principle, under some conditions, the limit distribution of the estimators can be obtained by the Delta-method.

10.2.3 Stochastic Volatility Models

In the stochastic volatility model discussed in Section 9.4 an observation X_t is defined as $X_t = \sigma_t Z_t$ for $\log \sigma_t$ a stationary auto-regressive process satisfying $\log \sigma_t = \alpha + \phi \log \sigma_{t-1} + \sigma V_{t-1}$, and (V_t, Z_t) an i.i.d. sequence of bivariate normal vectors with mean zero, unit variances and correlation δ. Thus the model is parameterized by four parameters $\alpha, \phi, \sigma, \delta$.

The series X_t is a white noise series and hence we cannot use the auto-covariances $\gamma_X(h)$ at lags $h \neq 0$ to construct moment estimators. Instead, we might use higher marginal moments or auto-covariances of powers of the series. In particular, it is computed in Section 9.4 that

$$\begin{aligned}
\mathrm{E}|X_t| &= \exp\Bigl(\tfrac{1}{2}\frac{\sigma^2}{1-\phi^2} + \frac{\alpha}{1-\phi}\Bigr)\sqrt{\frac{2}{\pi}},\\
\mathrm{E}|X_t|^2 &= \exp\Bigl(\tfrac{1}{2}\frac{4\sigma^2}{1-\phi^2} + \frac{2\alpha}{1-\phi}\Bigr),\\
\mathrm{E}|X_t|^3 &= \exp\Bigl(\tfrac{1}{2}\frac{9\sigma^2}{1-\phi^2} + \frac{3\alpha}{1-\phi}\Bigr)2\sqrt{\frac{2}{\pi}},\\
\mathrm{E}X_t^4 &= \exp\Bigl(\frac{8\sigma^2}{1-\phi^2} + \frac{4\alpha}{1-\phi}\Bigr)3,\\
\rho_{X^2}(1) &= \frac{(1+4\delta^2\sigma^2)e^{4\sigma^2\phi/(1-\phi^2)} - 1}{3e^{4\sigma^2/(1-\phi^2)} - 1},\\
\rho_{X^2}(2) &= \frac{(1+4\delta^2\sigma^2\phi^2)e^{4\sigma^2\phi^2/(1-\phi^2)} - 1}{3e^{4\sigma^2/(1-\phi^2)} - 1},\\
\rho_{X^2}(3) &= \frac{(1+4\delta^2\sigma^2\phi^4)e^{4\sigma^2\phi^3/(1-\phi^2)} - 1}{3e^{4\sigma^2/(1-\phi^2)} - 1}.
\end{aligned}$$

We can use a selection of these moments to define moment estimators, or use some or all of them to define generalized moments estimators. Because the functions on the right side are complicated, this requires some effort, but it is feasible.[♭]

10.3 Least Squares Estimators

For auto-regressive processes the method of least squares is directly suggested by the structural equation defining the model, but it can also be derived from the prediction problem. The second point of view is deeper and can be applied to general time series.

A *least squares estimator* is based on comparing the predicted value of an observation X_t based on the preceding observations to the actually observed value X_t. Such a prediction $\Pi_{t-1}X_t$ will generally depend on the underlying parameter θ of the model,

[♭] See Taylor (1986).

which we shall make visible in the notation by writing it as $\Pi_{t-1}X_t(\theta)$. The index $t-1$ of Π_{t-1} indicates that $\Pi_{t-1}X_t(\theta)$ is a function of $X_1,\ldots,X_{t-1}$ (and the parameter) only. By convention we define $\Pi_0 X_1 = 0$. A *weighted least squares estimator*, with inverse weights $w_t(\theta)$, is defined as the minimizer, if it exists, of the function

$$\theta \mapsto \sum_{t=1}^n \frac{\bigl(X_t - \Pi_{t-1}X_t(\theta)\bigr)^2}{w_t(\theta)}. \tag{10.3}$$

This expression depends only on the observations $X_1,\ldots,X_n$ and the unknown parameter θ and hence is an "observable criterion function". The idea is that using the "true" parameter should yield the "best" predictions. The weights $w_t(\theta)$ could be chosen equal to one, but are generally chosen to increase the efficiency of the resulting estimator.

This least squares principle is intuitively reasonable for any sense of prediction, in particular both for linear and nonlinear prediction. For nonlinear prediction we set $\Pi_{t-1}X_t(\theta)$ equal to the conditional expectation $\mathrm{E}_\theta(X_t|\,X_1,\ldots,X_{t-1})$, an expression that may or may not be easy to derive analytically.

For linear prediction, if we assume that the the time series X_t is centered at mean zero, we set $\Pi_{t-1}X_t(\theta)$ equal to the linear combination $\beta_1 X_{t-1}+\cdots+\beta_{t-1}X_1$ that minimizes

$$(\beta_1,\ldots,\beta_{t-1}) \mapsto \mathrm{E}_\theta\bigl(X_t - (\beta_1 X_{t-1}+\cdots+\beta_{t-1}X_1)\bigr)^2, \qquad \beta_1,\ldots,\beta_t \in \mathbb{R}.$$

In Chapter 2 the coefficients of the best linear predictor are expressed in the auto-covariance function γ_X by the prediction equations (2.1). Thus the coefficients β_t depend on the parameter θ of the underlying model through the auto-covariance function. Hence the least squares estimators using linear predictors can also be viewed as moment estimators.

The difference $X_t - \Pi_{t-1}X_t(\theta)$ between the true value and its prediction is called *innovation.* Its second moment

$$v_{t-1}(\theta) = \mathrm{E}_\theta\bigl(X_t - \Pi_{t-1}X_t(\theta)\bigr)^2$$

is called the (square) *prediction error* at time $t-1$. The weights $w_t(\theta)$ are often chosen equal to the prediction errors $v_{t-1}(\theta)$ in order to ensure that the terms of the sum of squares contribute "equal" amounts of information.

For both linear and nonlinear predictors the innovations $X_1 - \Pi_0 X_1(\theta), X_2 - \Pi_1 X_2(\theta),\ldots,X_n - \Pi_{n-1}X_n(\theta)$ are uncorrelated random variables. This orthogonality suggests that the terms of the sum contribute "additive information" to the criterion, which should be good. It also shows that there is usually no need to replace the sum of squares by a more general quadratic form, which would be the standard approach in ordinary least squares estimation.

Whether the sum of squares indeed possesses a (unique) point of minimum $\hat\theta$ and whether this constitutes a good estimator of the parameter θ depends on the statistical model for the time series. Moreover, this model determines the feasibility of computing the point of minimum given the data. Auto-regressive and GARCH processes provide a positive and a negative example.

10.17 Example (AR). A mean-zero causal, stationary, auto-regressive process of order p is modelled through the parameter $\theta = (\sigma^2, \phi_1, \ldots, \phi_p)$. For $t \geq p$ the best linear predictor is given by $\Pi_{t-1}X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p}$ and the prediction error is $v_{t-1} = \mathrm{E}Z_t^2 = \sigma^2$. For $t < p$ the formulas are more complicated, but could be obtained in principle.

The weighted sum of squares with weights $w_t = v_{t-1}$ reduces to

$$\sum_{t=1}^{p} \frac{\bigl(X_t - \Pi_{t-1}X_t(\phi_1, \ldots, \phi_p)\bigr)^2}{v_{t-1}(\sigma^2, \phi_1, \ldots, \phi_p)} + \sum_{t=p+1}^{n} \frac{\bigl(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}\bigr)^2}{\sigma^2}.$$

Because the first term, consisting of p of the n terms of the sum of squares, possesses a complicated form, it is often dropped from the sum of squares. Then we obtain exactly the sum of squares considered in Section 10.1, but with $\overline{X}_n$ replaced by 0 and divided by σ^2. For large n the difference between the sums of squares and hence between the two types of least squares estimators should be negligible.

Another popular strategy to simplify the sum of squares is to act as if the "observations" $X_0, X_{-1}, \ldots, X_{-p+1}$ are available and to redefine $\Pi_{t-1}X_t$ for $t = 1, \ldots, p$ accordingly. This is equivalent to dropping the first term and letting the sum in the second term start at $t = 1$ rather than at $t = p + 1$. To implement the estimator we must now choose numerical values for the missing observations $X_0, X_{-1}, \ldots, X_{-p+1}$; zero is a common choice.

The least squares estimators for $\phi_1, \ldots, \phi_p$, being (almost) identical to the Yule-Walker estimators, are $\sqrt{n}$-consistent and asymptotically normal. However, the least squares criterion does not lead to a useful estimator for σ^2: minimization over σ^2 leads to $\sigma^2 = \infty$ and this is obviously not a good estimator. A more honest conclusion is that the least squares criterion as posed originally fails for auto-regressive processes, since minimization over the full parameter $\theta = (\sigma^2, \phi_1, \ldots, \phi_p)$ leads to a zero sum of squares for $\sigma^2 = \infty$ and arbitrary (finite) values of the remaining parameters. The method of least squares works only for the subparameter $(\phi_1, \ldots, \phi_p)$ if we first drop σ^2 from the sum of squares. □

10.18 Example (GARCH). A GARCH process is a martingale difference series and hence the one-step predictions $\Pi_{t-1}X_t(\theta)$ are identically zero. Consequently, the weighted least squares sum, with weights equal to the prediction errors, reduces to

$$\sum_{t=1}^{n} \frac{X_t^2}{v_{t-1}(\theta)}.$$

Minimizing this criterion over θ is equivalent to maximizing the prediction errors $v_{t-1}(\theta)$. It is intuitively clear that this does not lead to reasonable estimators.

One alternative is to apply the least squares method to the squared series X_t^2. This satisfies an ARMA equation in view of (8.3). (Note however that the innovations in that equation are also dependent on the parameter.)

The best fix of the least squares method is to augment the least squares criterion to the Gaussian likelihood, as discussed in Chapter 12. □

So far the discussion in this section has assumed implicitly that the mean value $\mu = \mathrm{E}X_t$ of the time series is zero. If this is not the case, then we apply the preceding discussion to the time series $X_t - \mu$ instead of to X_t, assuming first that μ is known. Then the parameter μ will show up in the least squares criterion. To define estimators we can either replace the unknown value μ by the sample mean $\overline{X}_n$ and minimize the sum of squares with respect to the remaining parameters, or perform a joint minimization over all parameters.

Least squares estimators can rarely be written in closed form, the case of stationary auto-regressive processes being an exception, but iterative algorithms for the approximate calculation are implemented in many computer packages. Newton-type algorithms provide one possibility. The best linear predictions $\Pi_{t-1}X_t$ are often computed recursively in t (for a grid of values θ), for instance with the help of a state space representation of the time series and the Kalman filter. We do not discuss this numerical aspect, but remark that even with modern day computing power, the use of a carefully designed algorithm is advisable.

The method of least squares is closely related to Gaussian likelihood, as discussed in Chapter 12. Gaussian likelihood is perhaps more fundamental than the method of least squares. For this reason we restrict further discussion of the method of least squares to ARMA processes.

10.3.1 ARMA Processes

The method of least squares works well for estimating the regression and moving average parameters $(\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)$ of ARMA processes, if we perform the minimization for a fixed value of the parameter σ^2. In general, if some parameter, such as σ^2 for ARMA processes, enters the covariance function as a multiplicative factor, then the best linear predictor $\Pi_t X_{t+1}$ is free from this parameter, by the prediction equations (2.1). On the other hand, the prediction error $v_{t+1} = \gamma_X(0) - (\beta_1,\ldots,\beta_t)\Gamma_t(\beta_1,\ldots,\beta_t)^T$ (where $\beta_1,\ldots,\beta_t$ are the coefficients of the best linear predictor) contains such a parameter as a multiplicative factor. It follows that the inverse of the parameter will enter the least squares criterion as a multiplicative factor. Thus on the one hand the least squares methods does not yield an estimator for this parameter; on the other hand, we can just omit the parameter and minimize the criterion over the remaining parameters. In particular, in the case of ARMA processes the least squares estimators for $(\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)$ are defined as the minimizers of, for $\tilde v_t = \sigma^{-2}v_t$,

$$\sum_{t=1}^n \frac{\bigl(X_t - \Pi_{t-1}X_t(\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)\bigr)^2}{\tilde v_{t-1}(\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)}.$$

This is a complicated function of the parameters. However, for a fixed value of $(\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)$ it can be computed using the state space representation of an ARMA process and the Kalman filter.

10.19 Theorem. *Let X_t be a causal and invertible stationary ARMA(p,q) process relative to an i.i.d. sequence Z_t with finite fourth moments. Then the least squares estimators* *satisfy*

$$\sqrt{n}\left(\begin{pmatrix}\hat{\vec{\phi}}_p\\ \hat{\vec{\theta}}_q\end{pmatrix}-\begin{pmatrix}\vec{\phi}_p\\ \vec{\theta}_q\end{pmatrix}\right)\rightsquigarrow N(0,\sigma^2 J^{-1}_{\vec{\phi}_p,\vec{\theta}_q}),$$

where $J_{\vec{\phi}_p,\vec{\theta}_q}$ *is the covariance matrix of* $(U_{-1},\ldots,U_{-p},V_{-1},\ldots,V_{-q})$ *for stationary autoregressive processes* U_t *and* V_t *satisfying* $\phi(B)U_t=\theta(B)V_t=Z_t$.

Proof. The proof of this theorem is long and technical. See e.g. Brockwell and Davis (1991), pages 375–396, Theorem 10.8.2. ∎

10.20 Example (MA(1)). The least squares estimator $\hat\theta_n$ for θ in the moving average process $X_t=Z_t+\theta Z_{t-1}$ with $|\theta|<1$ possesses asymptotic variance equal to $\sigma^2/\operatorname{var} V_{-1}$, where V_t is the stationary solution to the equation $\theta(B)V_t=Z_t$. Note that V_t is an autoregressive process of order 1, not a moving average!

As we have seen before the process V_t possesses the representation $V_t=\sum_{j=0}^\infty\theta^j Z_{t-j}$ and hence $\operatorname{var} V_t=\sigma^2/(1-\theta^2)$ for every t.

Thus the sequence $\sqrt{n}(\hat\theta_n-\theta)$ is asymptotically normally distributed with mean zero and variance equal to $1-\theta^2$. This should be compared to the asymptotic distribution of the moment estimator, obtained in Example 10.14. □

10.21 EXERCISE. Find the asymptotic covariance matrix of the sequence $\sqrt{n}(\hat\phi_n-\phi,\hat\theta_n-\theta)$ for $(\hat\phi_n,\hat\theta_n)$ the least squares estimators for the stationary, causal, invertible ARMA process satisfying $X_t=\phi X_{t-1}+Z_t+\theta Z_{t-1}$.

11
Spectral Estimation

In this chapter we study nonparametric estimators of the spectral density and spectral distribution of a stationary time series. As in Chapter 5 "nonparametric" means that no a-priori structure of the series is assumed, apart from stationarity.

If a well-fitting model is available, then an alternative to the methods of this chapter is to use spectral estimators suited to this model. For instance, the spectral density of a stationary ARMA process can be expressed in the parameters $\sigma^2, \phi_1, \ldots \phi_p, \theta_1, \ldots, \theta_q$ of the model. It is natural to use the formula given in Section 7.5 for estimating the spectrum, by simply plugging in estimators for the parameters. If the ARMA model is appropriate, this should lead to better estimators than the nonparametric estimators discussed in this chapter. We do not further discuss this type of estimator.

Let the observations $X_1, \ldots, X_n$ be the values at times $1, \ldots, n$ of a stationary time series X_t, and let $\hat\gamma_n$ be their sample auto-covariance function. In view of the definition of the spectral density $f_X(\lambda)$, a natural estimator is

$$\hat f_{n,r}(\lambda) = \frac{1}{2\pi} \sum_{|h|<r} \hat\gamma_n(h) e^{-ih\lambda}. \tag{11.1}$$

Whereas $f_X(\lambda)$ is defined as an infinite series, the estimator $\hat f_{n,r}$ is truncated at its rth term. Because the estimators $\hat\gamma_n(h)$ are defined only for $|h| < n$ and there is no hope of estimating the auto-covariances $\gamma_X(h)$ for lags $|h| \geq n$, we must choose $r \leq n$. Because the estimators $\hat\gamma_n(h)$ are unreliable for $|h| \approx n$, it may be wise to choose r much smaller than n. We shall see that a good choice of r depends on the smoothness of the spectral density and also on which aspect of the spectrum is of interest. For estimating $f_X(\lambda)$ at a point, values of r such as n^α for some $\alpha \in (0,1)$ may be appropriate, whereas for estimating the spectral distribution function (i.e. areas under f_X) the choice $r = n$ works well.

In any case, since the covariances of lags $|h| \geq n$ can never be estimated from the data, nonparametric estimation of the spectrum is hopeless, unless one is willing to assume that expressions such as $\sum_{|h|\geq n} \bigl|\gamma_X(h)\bigr|$ are small. In Section 11.3 ahead we relate this tail series to the smoothness of the function $\lambda \mapsto f_X(\lambda)$. If the spectral density f_X

is smooth, then the auto-covariance function decreases fast to zero, and nonparametric estimation is feasible.

11.1 Finite Fourier Transform

The finite Fourier transform is a useful tool in spectral analysis, both for theory and practice. The practical use comes from the fact that it can be computed very efficiently by a clever algorithm, the *Fast Fourier Transform* (FFT).[♯]

The *finite Fourier transform* of an arbitrary sequence $x_1, \ldots, x_n$ of complex numbers is defined as the function $\lambda \mapsto d_x(\lambda)$ given by

$$d_x(\lambda) = \frac{1}{\sqrt{n}} \sum_{t=1}^n x_t e^{-i\lambda t}, \qquad \lambda \in (-\pi, \pi].$$

In other words, the function $\sqrt{n/2\pi}\, d_x(\lambda)$ is the Fourier series corresponding to the coefficients $\ldots, 0, 0, x_1, x_2, \ldots, x_n, 0, 0, \ldots$. The inversion formula (or a short calculation) shows that

$$x_t = \frac{\sqrt{n}}{2\pi} \int_{-\pi}^{\pi} e^{it\lambda} d_x(\lambda)\, d\lambda, \qquad t = 1, 2, \ldots, n.$$

Thus there is a one-to-one relationship between the numbers $x_1, \ldots, x_n$ and the function d_x; we may view the function d_x as "encoding" the numbers $x_1, \ldots, x_n$.

Encoding n numbers by a function on the interval $(-\pi, \pi]$ is rather inefficient. At closer inspection the numbers $x_1, \ldots, x_n$ can also be recovered from the values of d_x on the grid

$$\ldots, -\frac{4\pi}{n}, -\frac{2\pi}{n}, 0, \frac{2\pi}{n}, \frac{4\pi}{n}, \ldots \ \subset (-\pi, \pi].$$

These n points are called the *natural frequencies* at time n.

11.1 Lemma. *If d_x is the finite Fourier transform of $x_1, \ldots, x_n \in \mathbb{C}$, then*

$$x_t = \frac{1}{\sqrt{n}} \sum_j d_x(\lambda_j) e^{it\lambda_j}, \qquad t = 1, 2, \ldots, n,$$

where the sum is computed over the natural frequencies $\lambda_j \in (-\pi, \pi]$ at time n.

Proof. For every of the natural frequencies λ_j define a vector

$$e_j = \frac{1}{\sqrt{n}} (e^{i\lambda_j}, e^{i2\lambda_j}, \ldots, e^{in\lambda_j}).$$

It is straightforward to check that the n vectors e_j form an orthonormal set in $\mathbb{C}^n$ and hence a basis. Thus the vector $x = (x_1, \ldots, x_n)$ can be written as $x = \sum_j \langle x, e_j \rangle e_j$. Now $\langle x, e_j \rangle = d_x(\lambda_j)$ and the lemma follows. ∎

[♯] See e.g. Brockwell and Davis, Chapter 10 for a discussion.

The proof of the preceding lemma shows how the numbers $d_x(\lambda_j)$ can be interpreted. View the coordinates of the vector $x=(x_1,\ldots,x_n)$ as the values of a signal at the time instants $1,2,\ldots,n$. Similarly, view the coordinates of the vector e_j as the values of the pure trigonometric signal $t\mapsto n^{-1/2}e^{it\lambda_j}$ at these time instants. By the preceding lemma the signal x can be written as a linear combination of the signals e_j. The value $\bigl|d_x(\lambda_j)\bigr|$ is the weight of signal e_j, and hence of frequency λ_j, in x.

11.2 EXERCISE. How is the weight of frequency 0 expressed in $(x_1,\ldots,x_n)$?

11.3 EXERCISE. Show that $d_{(\mu,\mu,\ldots,\mu)}(\lambda_j)=0$ for every natural frequency λ_j and every $\mu\in\mathbb{C}$. Conclude that $d_{x-1\overline{x}_n}(\lambda_j)=d_x(\lambda_j)$.

11.2 Periodogram

The *periodogram* of a sequence of observations $X_1,\ldots,X_n$ is defined as the function $\lambda\mapsto I_n(\lambda)$ given by

$$I_n(\lambda)=\frac{1}{2\pi}\bigl|d_X(\lambda)\bigr|^2=\frac{1}{2\pi n}\Bigl|\sum_{t=1}^n X_t e^{-it\lambda}\Bigr|^2.$$

We write $I_{n,X}$ if the dependence on $X_1,\ldots,X_n$ needs to be stressed.

In view of the interpretation of the finite Fourier transform in the preceding section $I_n(\lambda)$ is the square of the weight of frequency λ in the signal $X_1,\ldots,X_n$. The spectral density $f_X(\lambda)$ can be interpreted as the variance of the component of frequency λ in the time series X_t. Thus $I_n(\lambda)$ appears to be a reasonable estimator of the spectral density. This is true to a certain extent, but not quite true. While we shall show that the expected value of $I_n(\lambda)$ converges to $f_X(\lambda)$, we shall also show that there are much better estimators than the periodogram. Because these will be derived from the periodogram, it is still of interest to study its properties.

By evaluating the square in its definition and rearranging the resulting double sum, the periodogram can be rewritten in the form (if $x_1,\ldots,x_n$ are real)

$$I_n(\lambda)=\frac{1}{2\pi}\sum_{|h|<n}\Bigl(\frac{1}{n}\sum_{t=1}^{n-|h|}X_{t+|h|}X_t\Bigr)e^{-ih\lambda}. \tag{11.2}$$

For natural frequencies $\lambda_j\neq 0$ we have that $d_{X-1\mu}(\lambda_j)=0$ for every μ, in particular for $\mu=\overline{X}_n$. This implies that $I_{n,X}(\lambda_j)=I_{n,X-\overline{X}_n}(\lambda_j)$ and hence

$$I_n(\lambda_j)=\frac{1}{2\pi}\sum_{|h|<n}\hat\gamma_n(h)e^{-ih\lambda},\qquad \lambda_j\in\frac{2\pi}{n}\mathbb{Z}-\{0\}. \tag{11.3}$$

This is exactly the estimator $\hat{f}_{n,n}(\lambda)$ given in (11.1). As noted before, we should expect this estimator to be unreliable as an estimator of $f_X(\lambda)$, because of the imprecision of the estimators $\hat{\gamma}_n(h)$ for lags $|h|$ close to n.

Assuming that the time series X_t is stationary, we can compute the mean of the periodogram, for $\lambda \neq 0$, as

$$\begin{aligned} \mathrm{E}I_n(\lambda) &= \frac{1}{2\pi}\mathrm{E}\bigl|d_X(\lambda)\bigr|^2 = \frac{1}{2\pi}\operatorname{var} d_X(\lambda) + \frac{1}{2\pi}\bigl|\mathrm{E}d_X(\lambda)\bigr|^2 \\ &= \frac{1}{2\pi n}\sum_{s=1}^n\sum_{t=1}^n \operatorname{cov}(X_s, X_t)e^{i\lambda(s-t)} + \frac{1}{2\pi n}\Bigl|\sum_{t=1}^n \mathrm{E}X_t e^{-i\lambda t}\Bigr|^2 \\ &= \frac{1}{2\pi}\sum_{|h|<n}\Bigl(1 - \frac{|h|}{n}\Bigr)\gamma_X(h)e^{-i\lambda h} + \frac{\mu^2}{2\pi n}\left|\frac{1-e^{-i\lambda n}}{1-e^{-i\lambda}}\right|^2. \end{aligned}$$

The second term on the far right is of the order $O(1/n)$ for every $\lambda \neq 0$ and even vanishes for every natural frequency $\lambda_j \neq 0$. Under the condition that $\sum_h \bigl|\gamma_X(h)\bigr| < \infty$, the first term converges to $f_X(\lambda)$ as $n \to \infty$, by the dominated convergence theorem. We conclude that the periodogram is asymptotically unbiased for estimating the spectral density in that $\mathrm{E}I_n(\lambda) \to f_X(\lambda)$. This is a good property.

However, the periodogram is not a consistent estimator for $f_X(\lambda)$: the following theorem shows that $I_n(\lambda)$ is asymptotically exponentially distributed with mean $f_X(\lambda)$, whence we do *not* have that $I_n(\lambda) \xrightarrow{\mathrm{P}} f_X(\lambda)$. Using the periodogram as an estimator of $f_X(\lambda)$ is, for $n \to \infty$, equivalent to estimating $f_X(\lambda)$ based on one observation with an exponential distribution. This is disappointing, because we should hope that after observing the time series X_t long enough, we would be able to estimate its spectral density with arbitrary precision. The periodogram does not fulfill this hope, as it keeps fluctuating around the target value $f_X(\lambda)$. Apparently, it does not effectively use the information available in the observations $X_1, \ldots, X_n$.

11.4 Theorem. *Let $X_t = \sum \psi_j Z_{t-j}$ for an i.i.d. sequence Z_t with mean zero and finite second moment and coefficients ψ_j with $\sum_j |\psi_j| < \infty$. Then for any values $0 < \mu_1 < \cdots < \mu_k < \pi$ the variables $I_n(\mu_1), \ldots, I_n(\mu_k)$ are asymptotically distributed as independent exponential variables with means $f_X(\mu_1), \ldots, f_X(\mu_k)$, respectively.*

Proof. First consider the case that $X_t = Z_t$ for every t. Then the spectral density $f_X(\lambda)$ is the function $f_Z(\lambda) = \sigma^2/2\pi$, for σ^2 the variance of the white noise sequence. We can write

$$d_Z(\lambda) = \frac{1}{\sqrt{n}}\sum_{t=1}^n Z_t\cos(\lambda t) - i\frac{1}{\sqrt{n}}\sum_{t=1}^n Z_t\sin(\lambda t) =: A_n(\lambda) - iB_n(\lambda).$$

By straightforward calculus we find that, for any $\lambda, \mu \in (0, \pi)$,

$$\operatorname{cov}\bigl(A_n(\lambda), A_n(\mu)\bigr) = \frac{\sigma^2}{n}\sum_{t=1}^n \cos(\lambda t)\cos(\mu t) \to \begin{cases} \sigma^2/2 & \text{if } \lambda = \mu, \\ 0 & \text{if } \lambda \neq \mu, \end{cases}$$

$$\operatorname{cov}\bigl(B_n(\lambda), B_n(\mu)\bigr) = \frac{\sigma^2}{n}\sum_{t=1}^n \sin(\lambda t)\sin(\mu t) \to \begin{cases} \sigma^2/2 & \text{if } \lambda = \mu, \\ 0 & \text{if } \lambda \neq \mu, \end{cases}$$

$$\operatorname{cov}\bigl(A_n(\lambda), B_n(\mu)\bigr) = \frac{\sigma^2}{n}\sum_{t=1}^n \cos(\lambda t)\sin(\mu t) \to 0.$$

By the Lindeberg central limit theorem, Theorem 3.16, we now find that the sequence of vectors $\bigl(A_n(\lambda), B_n(\lambda), A_n(\mu), B_n(\mu)\bigr)$ converges in distribution to a vector (G_1, G_2, G_3, G_4) with the $N_4\bigl(0, (\sigma^2/2)I\bigr)$ distribution. Consequently, by the continuous mapping theorem,

$$\bigl(I_n(\lambda), I_n(\mu)\bigr) = \frac{1}{2\pi}\bigl(A_n^2(\lambda) + B_n^2(\lambda), A_n^2(\mu) + B_n^2(\mu)\bigr) \rightsquigarrow \frac{1}{2\pi}(G_1^2 + G_2^2, G_3^2 + G_4^2).$$

The vector on the right is distributed as $\sigma^2/(4\pi)$ times a vector of two independent χ_2^2 variables. Because the chisquare distribution with two degrees of freedom is identical to the standard exponential distribution with parameter 1/2, this is the same as a vector of two independent exponential variables with means $\sigma^2/(2\pi)$.

This concludes the proof in the special case that $X_t = Z_t$ and for two frequencies λ and μ. The case of k different frequencies $\mu_1, \ldots, \mu_k$ can be treated in exactly the same way, but is notationally more involved.

Now consider the case of a general time series of the form $X_t = \sum_j \psi_j Z_{t-j}$. Then $f_X(\lambda) = \bigl|\psi(\lambda)\bigr|^2 f_Z(\lambda)$, for $\psi(\lambda) = \sum \psi_j e^{-ij\lambda}$ the transfer function of the linear filter. We shall prove the theorem by showing that the periodograms $I_{n,X}$ and $I_{n,Z}$ satisfy a similar relation, approximately. Indeed, rearranging sums we find

$$d_X(\lambda) = \frac{1}{\sqrt{n}}\sum_{t=1}^n \Bigl(\sum_j \psi_j Z_{t-j}\Bigr)e^{-it\lambda} = \sum_j \psi_j e^{-ij\lambda}\Bigl(\frac{1}{\sqrt{n}}\sum_{s=1-j}^{n-j} Z_s e^{-is\lambda}\Bigr).$$

If we replace the sum $\sum_{s=1-j}^{n-j}$ in the right side by the sum $\sum_{s=1}^n$, then the right side of the display becomes $\psi(\lambda)d_Z(\lambda)$. These two sums differ by $2(|j| \wedge n)$ terms, every of the terms $Z_s e^{-i\lambda t}$ having mean zero and variance bounded by σ^2, and the terms being independent. Thus

$$\mathrm{E}\Bigl|\frac{1}{\sqrt{n}}\sum_{s=1-j}^{n-j} Z_s e^{-is\lambda} - d_Z(\lambda)\Bigr|^2 \leq 2\frac{|j| \wedge n}{n}\sigma^2.$$

In view of the inequality $\mathrm{E}|X| \leq \bigl(\mathrm{E}X^2\bigr)^{1/2}$, we can drop the square on the left side if we take a root on the right side. Next combining the two preceding displays and applying the triangle inequality, we find

$$\mathrm{E}\bigl|d_X(\lambda) - \psi(\lambda)d_Z(\lambda)\bigr| \leq \sum_j |\psi_j|\Bigl(2\frac{|j| \wedge n}{n}\Bigr)^{1/2}\sigma.$$

The jth term of the series is bounded by $|\psi_j|\bigl(2|j|/n\bigr)^{1/2}\sigma$ and hence converges to zero as $n \to \infty$, for every fixed j; it is also dominated by $|\psi_j|\sqrt{2}\sigma$. Therefore, the right side of preceding display converges to zero as $n \to \infty$.

By Markov's and Slutsky's lemmas it follows that $d_X(\lambda)$ has the same limit distribution as $\psi(\lambda)d_Z(\lambda)$. By the continuous mapping theorem $I_{n,X}(\lambda)$ has the same limit distribution as $\bigl|\psi(\lambda)\bigr|^2 I_{n,Z}(\lambda)$. This is true for every fixed λ, but also for finite sets of λ jointly. The proof is finished, because the variables $\bigl|\psi(\lambda)\bigr|^2 I_{n,Z}(\lambda)$ are distributed as independent exponential variables with means $\bigl|\psi(\lambda)\bigr|^2 f_Z(\lambda)$, by the first part of the proof. ■

A remarkable aspect of the preceding theorem is that the periodogram values $I_n(\lambda)$ at different frequencies are asymptotically independent. This is well visible already for finite values of n in plots of the periodogram, which typically have a wild and peaky appearance. The theorem says that for large n such a plot should be similar to a plot of independent exponentially distributed variables E_λ with means $f_X(\lambda)$ (on the y-axis) versus λ (on the x-axis).

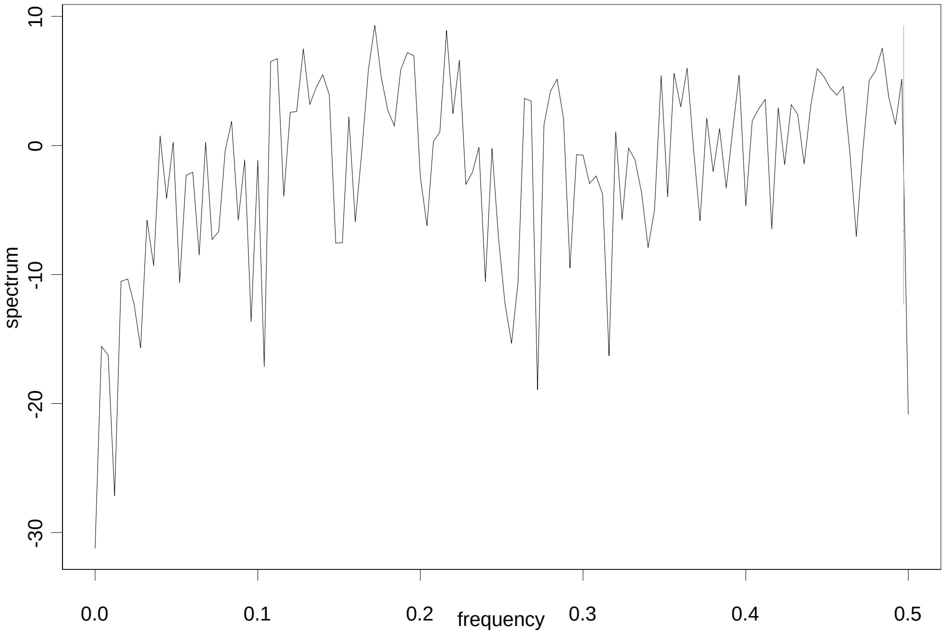

Figure 11.1. Periodogram of a realization of the moving average $X_t = 0.5Z_t + 0.2Z_{t-1} + 0.5Z_{t-2}$ for a Gaussian white noise series. (Vertical scale in decibel, i.e. ${}^{10}\log$.)

The following theorem shows that we even have independence of the periodogram values at natural frequencies that converge to the same value.

11.5 Theorem. *Let $X_t = \sum \psi_j Z_{t-j}$ for an i.i.d. sequence Z_t with finite second moments and coefficients ψ_j with $\sum_j |\psi_j| < \infty$. Let $\lambda_n = (2\pi/n)j_n$ for $j_n \in \mathbb{Z}$ be a sequence of natural frequencies such that $\lambda_n \to \lambda \in (0,\pi)$. Then for any $k \in \mathbb{Z}$ the variables $I_n(\lambda_n - k2\pi/n), I_n\bigl(\lambda_n - (k-1)2\pi/n\bigr), \ldots, I_n(\lambda_n + k2\pi/n)$ are asymptotically distributed as independent exponential variables with mean $f_X(\lambda)$.*

Proof. The second part of the proof of Theorem 11.4 is valid uniformly in λ and hence applies to sequences of frequencies λ_n. For instance, the continuity of $\psi(\lambda)$ and the proof shows that $\bigl|d_X(\mu_n) - \psi(\mu_n)d_Z(\mu_n)\bigr| \xrightarrow{\mathrm{P}} 0$ for any sequence μ_n. It suffices to extend the first part of the proof, which concerns the special case that $X_t = Z_t$.

Here we apply the same method as in the proof of Theorem 11.4. The limits of the covariances are as before, where in the present case we use the fact that we are considering natural frequencies only. For instance,

$$\operatorname{cov}\Bigl(A_n\Bigl(k\frac{2\pi}{n}\Bigr), B_n\Bigl(l\frac{2\pi}{n}\Bigr)\Bigr) = \frac{\sigma^2}{2}\sum_{t=1}^n \cos\Bigl(kt\frac{2\pi}{n}\Bigr)\sin\Bigl(lt\frac{2\pi}{n}\Bigr) = 0,$$

for every integers k, l such that $(k+l)/n$ and $(k-l)/n$ are not contained in $\mathbb{Z}$. An application of the Lindeberg central limit theorem concludes the proof. ∎

The sequences of frequencies $\lambda_n + j(2\pi/n)$ considered in the preceding theorem all converge to the same value λ. That Theorem 11.4 remains valid (it does) if we replace the fixed frequencies μ_j in this theorem by sequences $\mu_{j,n} \to \mu_j$ is not very surprising. More surprising is the asymptotic independence of the periodograms $I_n(\mu_{n,j})$ at different frequencies $\mu_{n,j}$ even if every sequence $\mu_{n,j}$ converges to the same frequency λ. As the proof of the preceding theorem shows, this depends crucially on using natural frequencies $\mu_{n,j}$.

The remarkable independence of the periodogram at frequencies that are very close together is a further explanation of the peaky appearance of the periodogram $I_n(\lambda)$ as a function of λ. It is clear that this function is not a good estimator of the spectral density. However, the independence suggests ways of improving our estimator for $f_X(\lambda)$. The values $I_n(\lambda_n - k2\pi/n), I_n\bigl(\lambda_n - (k-1)2\pi/n\bigr), \ldots, I_n(\lambda_n + k2\pi/n)$ can be viewed as a sample of independent estimators of $f_X(\lambda)$, for any k. Rather than one exponentially distributed veriable, we therefore have many exponentially distributed variables, all with the same (asymptotic) mean. We exploit this in the next section.

In practice the periodogram may have one or a few extremely high peaks that completely dominate its graph. This indicates an important cyclic component in the time series at those frequencies. Cyclic components of smaller amplitude at other frequencies may be hidden. It is practical wisdom that in such a case a fruitful spectral analysis at other frequencies requires that the peak frequencies are first removed from the signal (by a filter with the appropriate transfer function). We next estimate the spectrum of the new time series and, if desired, transform this back to obtain the spectrum of the original series, using the formula given in Theorem 6.9. Because a spectrum without high peaks is similar to the uniform spectrum of a white noise series, this procedure is known as *prewhitening* of the data.

11.3 Estimating a Spectral Density

Given $\lambda \in (0,\pi)$ and n, let λ_n be the natural frequency closest to λ. Then $\lambda_n \to \lambda$ as $n \to \infty$ and Theorem 11.5 shows that for any $k \in \mathbb{Z}$ the variables $I_n(\lambda_n + j2\pi/n)$ for $j = -k, \ldots, k$ are asymptotically distributed as independent exponential variables with mean $f_X(\lambda)$. This suggests to estimate $f_X(\lambda)$ by the average

$$\hat{f}_k(\lambda) = \frac{1}{2k+1} \sum_{|j|\le k} I_n\Bigl(\lambda_n + \frac{2\pi}{n} j\Bigr). \tag{11.4}$$

As a consequence of Theorem 11.5, the variables $(2k+1)\hat{f}_k(\lambda)$ are asymptotically distributed according the gamma distribution with shape parameter $2k+1$ and mean $(2k+1)f_X(\lambda)$. This suggests a confidence interval for $f_X(\lambda)$ of the form, with $\chi^2_{k,\alpha}$ the upper α-quantile of the chisquare distribution with k degrees of freedom,

$$\left(\frac{(4k+2)\hat{f}_k(\lambda)}{\chi^2_{4k+2,\alpha}}, \frac{(4k+2)\hat{f}_k(\lambda)}{\chi^2_{4k+2,1-\alpha}} \right).$$

11.6 EXERCISE. Show that, for every fixed k, this interval is asymptotically of level $1-2\alpha$.

Instead of a simple average we may prefer a weighted average. For given weights W_j such that $\sum_j W_j = 1$, we use

$$\hat{f}_k(\lambda) = \sum_j W_j I_n\Bigl(\lambda_n + \frac{2\pi}{n} j\Bigr). \tag{11.5}$$

This allows to give greater weight to frequencies $\lambda_n + (2\pi/n)j$ that are closer to λ. A disadvantage is that the asymptotic distribution is relatively complicated: it is a weighted sum of independent exponential variables. Because tabulating these types of distributions is complicated, one often approximates it by a scaled chisquare distribution, where the scaling and the degrees of freedom are chosen to match the first two moments: the estimator $c^{-1}\hat{f}_k(\lambda)$ is approximately χ^2_ν distributed for c and ν solving the equations

$$\begin{aligned} \text{asymptotic mean of } \hat{f}_k(\lambda) &= f_X(\lambda) = c\nu, \\ \text{asymptotic variance of } \hat{f}_k(\lambda) &= \sum_j W_j^2 f_X^2(\lambda) = c^2 2\nu. \end{aligned}$$

This yields c proportional to $f_X(\lambda)$ and ν independent of $f_X(\lambda)$, and thus confidence intervals based on this approximation can be derived as before. Rather than using the

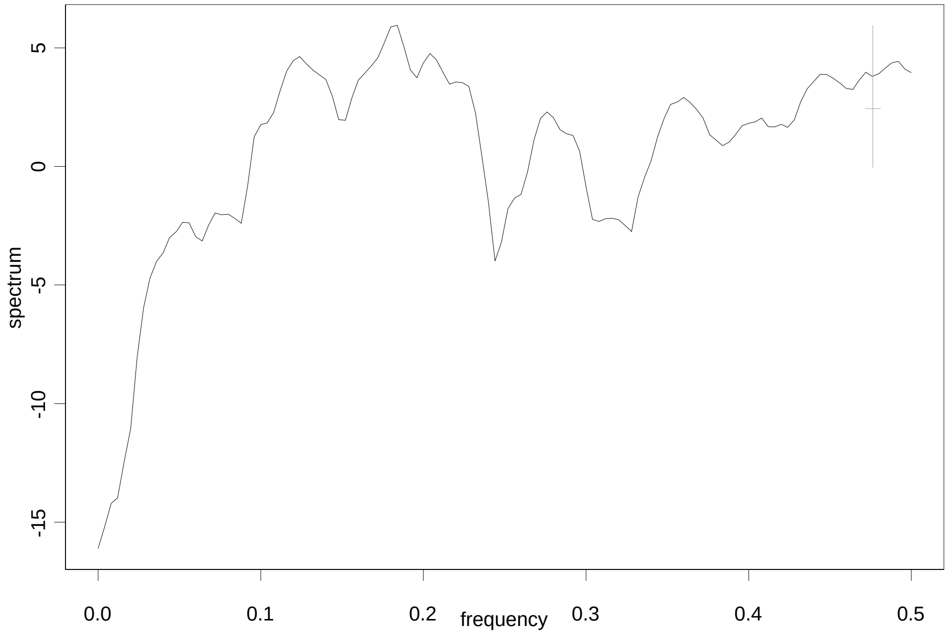

Figure 11.2. Smoothed periodogram of a realization of the moving average $X_t = 0.5Z_t + 0.2Z_{t-1} + 0.5Z_{t-2}$ for a Gaussian white noise series. (Vertical scale in decibel, i.e. ${}^{10}\log$.)

approximation, we could of course determine the desired quantiles by computer simulation.

Because the periodogram is continuous as a function of λ, a discrete average (over natural frequencies) can be closely approximated by a continuous average of the form

$$\hat{f}_W(\lambda) = \int W(\omega) I_{n,X-1\overline{X}}(\lambda-\omega)\,d\omega. \tag{11.6}$$

Here the weight function W is to satisfy $\int W(\omega)\,d\omega = 1$ and would typically concentrate its mass around zero, so that the average is computed over $I_{n,X-1\overline{X}}(\omega)$ for $\omega \approx \lambda$. We use the periodogram of the centered series $X - 1\overline{X}$, because the average involves nonnatural frequencies. In view of (11.3) this estimator can be written in the form

$$\begin{aligned}\hat{f}_W(\lambda) &= \int W(\omega) \frac{1}{2\pi} \sum_{|h|<n} \hat{\gamma}_n(h) e^{-i(\lambda-\omega)h}\,d\omega,\\ &= \frac{1}{2\pi} \sum_{|h|<n} w(h)\hat{\gamma}_n(h) e^{-i\lambda h},\end{aligned}$$

where $w(h) = \int e^{i\omega h} W(\omega)\, d\omega$ are the Fourier coefficients of the weight function. Thus we have arrived at a generalization of the estimator (11.1). If we choose $w(h) = 1$ for $|h| < r$ and $w(h) = 0$ otherwise, then the preceding display exactly gives (11.1). The more general form can be motivated by the same reasoning: the role of the coefficients $w(h)$ is to diminish the influence of the relatively unrealiable estimators $\hat{\gamma}_n(h)$ (for $h \approx n$), when plugging in these sample estimators for the true auto-covariances in the expression for the spectral density. Thus, the weights $w(h)$ are typically chosen to decrease in absolute value from $\bigl|w(0)\bigr| = 1$ to $\bigl|w(n)\bigr| = 0$ if h increases from 0 to n.

The function W is known as the *spectral window*; its Fourier coefficients $w(h)$ are known as the *lag window*, *tapering function* or *convergence factors*. The last name comes from Fourier analysis, where convergence factors were introduced to improve the approximation properties of a Fourier series: it was noted that for suitably chosen weights $w(h)$ the partial sums $\sum_{|h|<n} w(h)\gamma_X(h)e^{-ih\lambda}$ could be much closer to the full series $\sum_h \gamma_X(h)e^{-ih\lambda}$ than the same partial sums with $w \equiv 1$. In our statistical context this is even more so the case, because we introduce additional approximation error by replacing the coefficients $\gamma_X(h)$ by the estimators $\hat{\gamma}_n(h)$.

11.7 Example. The tapering function

$$w(h) = \begin{cases} 0 & \text{if } |h| \le r, \\ 1 & \text{if } |h| > r, \end{cases}$$

corresponds to the Dirichlet kernel

$$W(\lambda) = \frac{1}{2\pi} \sum_{|h| \le r} e^{ih\lambda} = \frac{1}{2\pi} \frac{\sin(r + \frac{1}{2})\lambda}{\sin \frac{1}{2}\lambda}.$$

Therefore, the estimator (11.1) should be compared to the estimators (11.5) and (11.6) with weights W_j chosen according to the Dirichlet kernel. □

11.8 Example. The uniform kernel

$$W(\lambda) = \begin{cases} r/(2\pi) & \text{if } |\lambda| \le \pi/r, \\ 0 & \text{if } |\lambda| > \pi/r, \end{cases}$$

corresponds to the weight function $w(h) = r \sin(\pi h)/(\pi h)$. These choices of spectral and lag windows correspond to the estimator (11.4). □

All estimators for the spectral density considered so far can be viewed as *smoothed periodograms*: the value $\hat{f}(\lambda)$ of the estimator at λ is an average or weighted average of values $I_n(\mu)$ for μ in a neighbourhood of λ. Thus "irregularities" in the periodogram are "smoothed out". The amount of smoothing is crucial for the accuracy of the estimators. This amount, called the *bandwidth*, is determined by the parameter k in (11.4), the weights W_j in (11.5), the kernel W in (11.6), and, more hidden, by the parameter r in (11.1). For instance, a large value of k in (11.4) or a kernel W with a large variance in (11.6) result in a large amount of smoothing (large bandwidth). *Oversmoothing*, choosing a bandwidth that is too large, results in spectral estimators that are too flat and therefore inaccurate, whereas *undersmoothing*, choosing too small a bandwidth, yields spectral estimators that share the bad properties of the periodogram. In practice an "optimal" bandwidth is often determined by plotting the spectral estimators for a number of different bandwidths and next choosing the one that looks "reasonable". An alternative is to use one of several methods of "data-driven" choices of bandwidths, such as cross validation or penalization. We omit a discussion.

Theoretical analysis of the choice of the bandwidth is almost exclusively asymptotical in nature. Given a number of observations tending to infinity, the "optimal" bandwidth decreases to zero. A main concern of an asymptotic analysis is to determine the rate at which the bandwidth should decrease as the number of observations tends to infinity. The key concept is the *bias-variance trade-off.* Because the periodogram is more or less unbiased, little smoothing gives an estimator with small bias. However, as we have seen, the estimator will have a large variance. Much smoothing has the opposite effects. Because accurate estimation requires that both bias and variance are small, we need an intermediate value of the bandwidth.

We shall quantify this bias-variance trade-off for estimators of the type (11.1), where we consider r as the bandwidth parameter. As our objective we take to minimize the mean integrated square error

$$2\pi\,\mathrm{E}\int_{-\pi}^{\pi}\bigl|\hat f_{n,r}(\lambda)-f_X(\lambda)\bigr|^2\,d\lambda.$$

The integrated square error is a global measure of the discrepancy between $\hat f_{n,r}$ and f_X. Because we are interested in f_X as a function, it is more relevant than the distance $\bigl|\hat f_{n,r}(\lambda)-f_X(\lambda)\bigr|$ for any fixed λ.

We shall use Parseval's identity, which says that the space $L_2(-\pi,\pi]$ is isometric to the space ℓ_2.

11.9 Lemma (Parseval's identity). *Let $f\colon(-\pi,\pi]\to\mathbb{C}$ be a measurable function such that $\int|f|^2(\lambda)\,d\lambda<\infty$. Then its Fourier coefficients $f_j=\int_{-\pi}^{\pi}e^{ij\lambda}f(\lambda)\,d\lambda$ satisfy*

$$\int_{-\pi}^{\pi}\bigl|f(\lambda)\bigr|^2\,d\lambda=\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}|f_j|^2.$$

11.10 EXERCISE. Prove this identity. Also show that for a pair of square-integrable, measurable functions $f,g\colon(-\pi,\pi]\to\mathbb{C}$ we have $\int f(\lambda)\overline{g}(\lambda)\,d\lambda=\sum_j f_j\overline{g}_j$.

The function $\hat f_{n,r}-f_X$ possesses the Fourier coefficients $\hat\gamma_n(h)-\gamma_X(h)$ for $|h|<r$ and $-\gamma_X(h)$ for $|h|\ge r$. Thus, Parseval's identity yields that the preceding display is equal to

$$\mathrm{E}\sum_{|h|<r}\bigl|\hat\gamma_n(h)-\gamma_X(h)\bigr|^2+\sum_{|h|\ge r}\bigl|\gamma_X(h)\bigr|^2$$

In a rough sense the two terms in this formula are the "variance" and the "bias" term. A large value of r clearly decreases the second, bias term, but increases the first, variance term. This variance term can itself be split into a bias and variance term and we can reexpress the mean integrated square error as

$$\sum_{|h|<r} \operatorname{var} \hat{\gamma}_n(h) + \sum_{|h|<r} \bigl|\mathrm{E}\hat{\gamma}_n(h) - \gamma_X(h)\bigr|^2 + \sum_{|h|\geq r} \bigl|\gamma_X(h)\bigr|^2.$$

Assume for simplicity that $\mathrm{E}X_t = 0$ and that $X_t = \sum_j \psi_j Z_{t-j}$ for an absolutely converging series $\sum \psi_j$ and i.i.d. sequence Z_t with finite fourth moments. Furthermore, assume that we use the estimator $\hat{\gamma}_n(h) = n^{-1}\sum_{t=1}^{n-h} X_{t+h}X_t$ rather than the true sample autocovariance function. (The results for the general case are similar, but the calculations will be even more involved than they already are. Note that the difference between the present estimator and the usual one is approximately $\overline{X}$ and $\sum_{|h|<r} \mathrm{E}(\overline{X})^4 = O(r/n^2)$. This is negligible in the following.) Then the calculations in Chapter 5 show that the preceding display is equal to

$$\begin{aligned}
&\sum_{|h|<r} \frac{1}{n^2} \sum_{|g|\leq n-|h|} \bigl(n - |h| - |g|\bigr)\Bigl[\kappa_4\sigma^4 \sum_i \psi_i\psi_{i+h}\psi_{i+g}\psi_{i+g+h} + \gamma_X^2(g) \\
&\qquad\qquad + \gamma_X(g+h)\gamma_X(g-h)\Bigr] + \sum_{|h|<r} \Bigl(\frac{|h|}{n}\gamma_X(h)\Bigr)^2 + \sum_{|h|\geq r} \gamma_X^2(h) \\
&\leq \frac{|\kappa_4|\sigma^4}{n} \sum_h \sum_g \sum_i |\psi_i\psi_{i+h}\psi_{i+g}\psi_{i+g+h}| + \frac{2r}{n}\sum_g \gamma_X^2(g) \\
&\qquad\qquad + \frac{1}{n}\sum_h \sum_g \bigl|\gamma_X(g+h)\gamma_X(g-h)\bigr| + \frac{r^2}{n^2}\sum_h \gamma_X^2(h) + \sum_{|h|\geq r} \gamma_X^2(h).
\end{aligned}$$

To ensure that the last term on the right converges to zero as $n \to \infty$ we must choose $r = r_n \to \infty$. Then the second term on the right converges to zero if and only if $r_n/n \to 0$. The first and third term are of the order $O(1/n)$, and the fourth term is of the order $O(r_n^2/n^2)$. Under the requirements $r_n \to \infty$ and $r_n/n \to 0$ these terms are dominated by the other terms, and the whole expression is of the order

$$\frac{r_n}{n} + \sum_{|h|\geq r_n} \gamma_X^2(h).$$

A first conclusion is that the sequence of estimators $\hat{f}_{n,r_n}$ is asymptotically consistent for estimating f_X relative to the L_2-distance whenever $r_n \to \infty$ and $r_n/n \to 0$. A wide range of sequences r_n satisfies these constraints. For an optimal choice we must make assumptions regarding the rate at which the bias term $\sum_{|h|\geq r} \gamma_X^2(h)$ converges to zero as $r \to \infty$. For any constant m we have that

$$\frac{r_n}{n} + \sum_{|h|\geq r_n} \gamma_X(h)^2 \leq \frac{r_n}{n} + \frac{1}{r_n^{2m}} \sum_h \gamma_X^2(h) h^{2m}.$$

Suppose that the series on the far right converges; this means roughly that the autocovariances $\gamma_X(h)$ decrease faster than $|h|^{-m-1/2}$ as $|h| \to \infty$. Then we can make a bias-variance trade-off by balancing the terms r_n/n and $1/r_n^{2m}$. These terms are of equal order for $r_n = n^{1/(2m+1)}$; for this choice of r_n we find that

$$\mathrm{E}\int_{-\pi}^{\pi} \bigl|\hat{f}_{n,r_n}(\lambda) - f_X(\lambda)\bigr|^2\, d\lambda = O\Bigl(n^{-2m/(2m+1)}\Bigr).$$

Large values of m yield the fastest rates of convergence. The rate $n^{-m/(2m+1)}$ is always slower than $n^{-1/2}$, the rate obtained when using parametric spectral estimators, but approaches this rate as $m \to \infty$.

Unfortunately, in practice we do not know $\gamma_X(h)$ and therefore cannot check whether the preceding derivation is valid. So-called *cross-validation* techniques may be used to determine a suitable constant m from the data.

The condition that $\sum_h \gamma_X^2(h)h^{2m} < \infty$ can be interpreted in terms of the smoothness of the spectral density. By differentiating the series $f_X(\lambda) = (2\pi)^{-1}\sum_h \gamma_X(h)e^{-ih\lambda}$ repeatedly we obtain that the mth derivative of f_X is given by

$$f_X^{(m)}(\lambda) = \frac{1}{2\pi}\sum_h \gamma_X(h)(-ih)^m e^{-ih\lambda}.$$

This shows that the numbers $\gamma_X(h)(-ih)^m$ are the Fourier coefficients of $f_X^{(m)}$. Consequently, by Parseval's identity

$$\sum_h \gamma_X^2(h)h^{2m} = \int_{-\pi}^{\pi} f_X^{(m)}(\lambda)^2\, d\lambda.$$

Thus the left side is finite if and only if the mth derivative of f_X exists and is square-integrable. We say that f_X is *m-smooth*. For time series with an m-smooth spectral density, one can estimate the spectral density with an integrated square error of order $O(n^{-2m/(2m+1)})$. This rate is uniform over the set of all time series with spectral densities such that $\int f_X^{(m)}(\lambda)^2\, d\lambda$ is uniformly bounded.

This conclusion is similar to the conclusion in the problem of estimating a density given a random sample from this density, where also m-smooth densities can be estimated with an integrated square error of order $O(n^{-2m/(2m+1)})$. The smoothing methods discussed previously (the estimator (11.6) in particular) are also related to the method of kernel smoothing for density estimation. It is interesting that historically the method of smoothing was first applied to the problem of estimating a spectral density. Here kernel smoothing of the periodogram was a natural extension of taking simple averages as in (11.4), which itself is motivated by the independence property of the periodogram. The method of kernel smoothing for the problem of density estimation based on a random sample from this density was invented later, even though this problem by itself appears to be simpler.

* 11.4 Estimating a Spectral Distribution

In the preceding section it is seen that nonparametric estimation of a spectral density requires smoothing and yields rates of convergence $n^{-\alpha}$ for values of $\alpha < 1/2$. In contrast, a spectral distribution function can be estimated at the "usual" rate of convergence $n^{-1/2}$ and natural estimators are asymptotically normally distributed. We assume X_t is a stationary time series with spectral density f_X.

The spectral distribution function $F_X(\lambda_0) = \int_{-\pi}^{\lambda_0} f_X(\lambda)$ can be written in the form

$$\int_{-\pi}^{\pi} a(\lambda) f_X(\lambda)\, d\lambda$$

for a the indicator function of the interval $(-\pi, \lambda_0]$. We shall consider estimation of a general functional of this type by the estimator

$$I_n(a) := \int_{-\pi}^{\pi} a(\lambda) I_n(\lambda)\, d\lambda.$$

11.11 Theorem. *Suppose that $X_t = \sum_j \psi_j Z_{t-j}$ for an i.i.d. sequence Z_t with finite fourth cumulant κ_4 and constants ψ_j such that $\sum_j |\psi_j| < \infty$. Moreover, assume that $\sum_h |h| \gamma_X^2(h) < \infty$. Then, for any symmetric function a such that $\int_{-\pi}^{\pi} a^2(\lambda)\, d\lambda < \infty$,*

$$\sqrt{n}\Bigl(I_n(a) - \int a f_X\, d\lambda\Bigr) \rightsquigarrow N\Bigl(0, \kappa_4\Bigl(\int a f_X\, d\lambda\Bigr)^2 + 4\pi \int a^2 f_X^2\, d\lambda\Bigr).$$

Proof. We can expand $a(\lambda)$ in its Fourier series $a(\lambda) = \sum_j a_j e^{-ij\lambda}$ (say). By Parseval's identity

$$\int a f_X\, d\lambda = \sum_h \gamma_X(h) a_h.$$

Similarly, by (11.2) and Parseval's identity

$$I_n(a) = \int a I_n\, d\lambda = \sum_{|h|<n} \hat\gamma_n^*(h) a_h.$$

First suppose that $a_h = 0$ for $|h| > m$ and some m. Then $\int a I_n\, d\lambda$ is a linear combination of $\bigl(\hat\gamma_n(0), \ldots, \hat\gamma_n(m)\bigr)$. By Theorem 5.7, as $n \to \infty$,

$$\sqrt{n}\Bigl(\sum_h \hat\gamma_n(h) a_h - \sum_h \gamma_X(h) a_h\Bigr) \rightsquigarrow \sum_h a_h Z_h,$$

where $(Z_{-m}, \ldots, Z_0, Z_1, \ldots, Z_m)$ is a mean zero normally distributed random vector such that $(Z_0, \ldots, Z_m)$ has covariance matrix V as in Theorem 5.7 and $Z_{-h} = Z_h$ for every

h. Thus $\sum_h a_h Z_h$ is normally distributed with mean zero and variance

$$\begin{aligned}\sum_g\sum_h V_{g,h}a_ga_h &= \kappa_4\Bigl(\sum_g a_g\gamma_X(g)\Bigr)^2 + \sum_g\sum_h\Bigl(\sum_k \gamma_X(k+h)\gamma_X(k+g)\\ &\qquad + \sum_k \gamma_X(k+h)\gamma_X(k-g)\Bigr)a_ga_h \\ &= \kappa_4\Bigl(\int a f_X\,d\lambda\Bigr)^2 + 4\pi\int a^2 f_X^2\,d\lambda.\end{aligned} \tag{11.7}$$

The last equality follows after a short calculation, using that $a_h = a_{-h}$. (Note that we have used the expression for $V_{g,h}$ given in Theorem 5.7 also for negative g or h, which is correct, because both $\operatorname{cov}(Z_g, Z_h)$ and the expression in Theorem 5.7 remain the same if g or h is replaced by $-g$ or $-h$.)

This concludes the proof in the case that $a_h = 0$ for $|h| > m$, for some m. The general case is treated with the help of Lemma 3.10. Set $a_m = \sum_{|j|\le m} a_j e^{-i\lambda j}$ and apply the preceding argument to $X_{n,m} := \sqrt{n}\int a_m(I_n - f_X)\,d\lambda)$ to see that $X_{n,m} \rightsquigarrow N(0,\sigma_m^2)$ as $n\to\infty$, for every fixed m. The asymptotic variance σ_m^2 is the expression given in the theorem with a_m instead of a. If $m\to\infty$, then σ_m^2 converges to the expression in the theorem, by the dominated convergence theorem, because a is squared-integrable and f_X is uniformly bounded. Therefore, by Lemma 3.10 it suffices to show that for every $m_n\to\infty$

$$\sqrt{n}\int (a - a_{m_n})(I_n - f_X)\,d\lambda \xrightarrow{\mathrm{P}} 0. \tag{11.8}$$

Set $b = a - a_{m_n}$. The variance of the random variable in (11.8) is the same as the variance of $\int (a - a_{m_n})I_n\,d\lambda$, and can be computed as, in view of Parseval's identity,

$$\begin{aligned}\operatorname{var}\Bigl(\frac{\sqrt{n}}{2\pi}\sum_{|h|<n}\hat\gamma_n(h)b_h\Bigr) &= \frac{n}{4\pi^2}\sum_{|g|<n}\sum_{|h|<n}\operatorname{cov}\bigl(\hat\gamma_n(g),\hat\gamma_n(h)\bigr)b_gb_h\\ &= \frac{n}{4\pi^2}\sum_{|g|<n}\sum_{|h|<n}\frac{1}{n^2}\sum_{s=1}^{n-|g|}\sum_{t=1}^{n-|h|}\operatorname{cov}(X_{s+g}X_s, X_{t+h}X_t)b_gb_h.\end{aligned}$$

Using the same approach as in Section 5.2, we can rewrite this as

$$\begin{aligned}\frac{1}{4\pi^2 n}\sum_{|g|<n}\sum_{|h|<n}\sum_{s=1}^{n-|g|}\sum_{t=1}^{n-|h|}\Bigl(\kappa_4\sigma^4\sum_i \psi_{i+g}\psi_i\psi_{t-s+h+i}\psi_{t-s+i}\\ + \gamma_X(t-s+h-g)\gamma_X(t-s) + \gamma_X(t-s-g)\gamma_X(t-s+h)\Bigr)b_gb_h.\end{aligned}$$

The absolute value of this expression can be bounded above by

$$\begin{aligned}\frac{1}{4\pi^2}\sum_g\sum_h\sum_k\Bigl(\sum_i|\psi_{i+g}\psi_i\psi_{k+h+i}\psi_{k+i}||\kappa_4|\sigma^4+\bigl|\gamma_X(k+h-g)\gamma_X(k)\bigr|\\ +\bigl|\gamma_X(k-g)\gamma_X(k+h)\bigr|\Bigr)|b_gb_h|\\ =\frac{1}{4\pi^2}\biggl(|k_4|\Bigl(\int\underline{b}\underline{\underline{f}}_X\,d\lambda\Bigr)^2+4\pi\underline{b}^2\int\underline{f}_X^2\,d\lambda\biggr),\end{aligned}$$

by the same calculation as in (11.7), where we define

$$\underline{b}=\sum_h|b_h|e^{-i\lambda h},\quad \underline{\underline{f}}_X(\lambda)=\sum_h\sum_i|\psi_i\psi_{i+h}|\sigma^2e^{-i\lambda h},\quad \underline{f}_X(\lambda)=\sum_h\bigl|\gamma_X(h)\bigr|e^{-i\lambda h}.$$

Under our assumptions $\underline{f}_X$ and $\underline{\underline{f}}_X$ are bounded functions. It follows that var $\int b_nI_n\,d\lambda\to 0$ if $\int\underline{b}_n^2\,d\lambda\to 0$. This is true in particular for $b_n=a-a_{m_n}$.

Next the mean of the left side of (11.8) can be computed as

$$\begin{aligned}\sqrt{n}\Bigl(\sum_{|h|<n}\mathrm{E}\hat\gamma_n(h)b_h-\int b\,f_X\,d\lambda\Bigr)&=\sqrt{n}\Bigl(\sum_{|h|<n}\frac{n-|h|}{n}\gamma_X(h)b_h-\sum_h\gamma_X(h)b_h\Bigr)\\ &=-\sum_h\frac{n\wedge|h|}{\sqrt{n}}\gamma_X(h)b_h.\end{aligned}$$

By the Cauchy-Schwarz inequality this is bounded in absolute value by the square root of

$$\sum_h|b_h|^2\sum_h\frac{(n\wedge|h|)^2}{n}\gamma_X^2(h)\le\sum_h|b_h|^2\sum_h|h|\gamma_X^2(h).$$

Under our assumptions this converges to zero as $\int\underline{b}^2\,d\lambda\to 0$. ∎

The preceding theorem is restricted to symmetric functions a, but can easily be extended to general functions, because by the symmetry of the spectral density

$$\int a(\lambda)f_X(\lambda)\,d\lambda=\int\frac{a(\lambda)+a(-\lambda)}{2}f_X(\lambda)\,d\lambda.$$

11.12 EXERCISE. Show that for a possibly nonsymmetric function a the theorem is valid, but with asymptotic variance

$$\kappa_4\Bigl(\int af_X\,d\lambda\Bigr)^2+2\pi\int_{-\pi}^{\pi}a^2f_X^2\,d\lambda+2\pi\int_{-\pi}^{\pi}a(\lambda)a(-\lambda)f_X^2(\lambda)\,d\lambda.$$

11.13 Example. To obtain the limit distribution of the estimator for the spectral distribution function at the point $\lambda_0\in[0,\pi]$, we apply the theorem with the symmetric function $a=\bigl(1_{(-\pi,\lambda_0]}+1_{(-\lambda_0,\pi]}\bigr)$. The asymptotic variance is equal to $\kappa_4F_X(\lambda_0)^2+4\pi\int_{\lambda_0}^{\lambda_0}f_X^2\,d\lambda+2\pi\int_{\lambda_0}^{\pi}f_X^2\,d\lambda$ for $\lambda_0\in[0,\pi]$. □

11.14 Example. The choice $a(\lambda)=\cos(h\lambda)$ yields the estimator (for $0\le h<n$)

$$\int \cos(h\lambda) I_n(\lambda)\,d\lambda = \mathrm{Re}\int e^{ih\lambda} I_n(\lambda)\,d\lambda = \frac{1}{n}\sum_{t=1}^{n-h} X_{t+h}X_t$$

of the auto-covariance function $\gamma_X(h)$ in the case that $\mathrm{E}X_t=0$. Thus the preceding theorem contains Theorem 5.7 as a special case. The present theorem shows how the asymptotic covariance of the sample auto-covariance function can be expressed in the spectral density. □

11.15 EXERCISE. Show that the sequence of bivariate random vectors $\sqrt{n}\bigl(\int a(I_n - f_X)\,d\lambda, \int b(I_n - f_X)\,d\lambda\bigr)$ converges in distribution to a bivariate Gaussian vector (G_a, G_b) with mean zero and $EG_aG_b = \kappa_4 \int af_X\,d\lambda \int bf_X\,d\lambda + 4\pi \int abf_X^2\,d\lambda$.

11.16 EXERCISE. Plot the periodogram of a white noise series of length 200. Does this look like a plot of 200 independent exponential variables?

11.17 EXERCISE. Estimate the spectral density of the simulated time series given in the file `s̃da/Cursusdata/sim2` by a smoothed periodogram. Compare this to the estimate obtained assuming that `sim2` is an AR(3) series.

11.18 EXERCISE. Estimate the spectral density of the Wölfer sunspot numbers (the object `sunspots` in Splus) by
(i) a smoothed periodogram;
(ii) the spectral density of an appropriate AR-model.
Note: the mean is nonzero.

12 Maximum Likelihood

The method of maximum likelihood is one of the unifying principles of statistics, and applies equally well to models for replicated experiments as to time series models. Given observations $X_1, \ldots, X_n$ with a joint probability density $(x_1, \ldots, x_n) \mapsto p_{n,\theta}(x_1, \ldots, x_n)$ that depends on a parameter θ, the *likelihood function* is the stochastic process

$$\theta \mapsto p_{n,\theta}(X_1, \ldots, X_n).$$

The *maximum likelihood estimator* for θ, if it exists, is the value of θ that maximizes the likelihood function.

The likelihood function corresponding to i.i.d. observations $X_1, \ldots, X_n$ is the product of the likelihood functions of the individual observations, which makes likelihood inference relatively easy in this case. For time series models the likelihood function may be a more complicated function of the observations and the parameter. This complicates both the practical implementation of likelihood inference and their theoretical analysis, but in "most" situations the final results are not that different from the more familiar i.i.d. case. In particular, maximum likelihood estimators are typically $\sqrt{n}$-consistent and possess a normal limit distribution, with mean zero and covariance the inverse of a certain Fisher information matrix.

In this chapter we study the maximum likelihood estimator, and some approximations. We also consider the effect of *model misspecification*: using the likelihood for a model that does not contain the "true" distribution of the data. Such misspecification of the model may be unintended, but is sometimes the result of a conscious choice. For instance, the likelihood under the assumption that $X_1, \ldots, X_n$ is part of a stationary Gaussian time series X_t is popular for inference, even if one may not believe that the time series is Gaussian. The corresponding maximum likelihood estimator is closely related to the least squares estimators and turns out to perform well also for a wide range of non-Gaussian time series. Another example is to postulate that the innovations in a GARCH model are Gaussian, even though we may not believe strongly in this assumption. The resulting estimators again work well also for non-Gaussian innovations. A misspecified likelihood is also referred to as a *quasi likelihood* and the resulting estimators as *quasi likelihood estimators*. A misspecified maximum likelihood estimator also falls in the class of *M-estimators*, which are defined as the maximizers of a given stochastic process, sometimes called a *contrast function*.

12.1 General Likelihood

A convenient representation of a likelihood is obtained by repeated conditioning. To alleviate the notation we abuse notation by writing $p(y|\,x)$ for a conditional density of a variable Y given that another variable X takes the value x, and denote the marginal density of X by $p(x)$. Thus we write the likelihood corresponding to the observations $X_1,\ldots,X_n$ as $\theta\mapsto p_\theta(x_1,\ldots,x_n)$, and this can be decomposed as

$$\theta\mapsto p_\theta(x_1,\ldots,x_n)=p_\theta(x_1)p_\theta(x_2|\,x_1)\cdots p_\theta(x_n|\,x_{n-1},\ldots,x_1). \tag{12.1}$$

Clearly we must select appropriate versions of the (conditional) densities, but we shall not worry about technical details in this section.

The decomposition resembles the factorization of the likelihood of an i.i.d. sample of observations, but an important difference is that the n terms on the right may all be of a different form. Even if the time series X_t is strictly stationary, each further term entails conditioning on a bigger past and hence is potentially of a different character than the earlier terms. However, in many examples the "present" X_t is nearly independent of the "distant past" $(X_s{:}\, s\ll t)$ given the "near past" $(X_s{:}\, s<t, s\approx t)$. Then the likelihood does not change much if the conditioning in each term is limited to a fixed number of variables in the past, and most of the terms of the product will take almost a common form. Alternatively, we may augment the conditioning in each term to include the full "infinite past", yielding the *pseudo likelihood*

$$\theta\mapsto p_\theta(x_1|\,x_0,x_{-1},\ldots)p_\theta(x_2|\,x_1,x_0,\ldots)\cdots p_\theta(x_n|\,x_{n-1},x_{n-2},\ldots). \tag{12.2}$$

If the time series X_t is strictly stationary, then the tth term $p_t(x_t|\,x_{t-1},x_{t-2},\ldots)$ in this product is a fixed measurable function, independent of t, applied to the vector $(x_t,x_{t-1},\ldots)$. In particular, the terms of the product form a strictly stationary time series, which will be ergodic if the original time series X_t is ergodic. This is almost as good as the i.i.d. terms obtained in the case of an i.i.d. time series.

The pseudo likelihood (12.2) cannot be used in practice, because the "negative" variables $X_0,X_{-1},\ldots$ are not observed. However, the preceding discussion suggests that the *maximum pseudo likelihood estimator*, defined as the maximizer of the pseudo likelihood, may behave the same as the true maximum likelihood estimator. Moreover, if it is true that the past observations $X_0,X_{-1},\ldots$ do not play an important role in defining the pseudo likelihood, then we could also replace them by arbitrary values, for instance zero, and hence obtain an observable criterion function.

12.1 Example (Markov time series). If the time series X_t is Markov, then the conditioning in each term $p_\theta(x_t|\,x_{t-1},\ldots,x_1)$ or $p_\theta(x_t|\,x_{t-1},x_{t-2},\ldots)$ can be restricted to a single variable, the variable x_{t-1}. In this case the likelihood and the pseudo likelihood differ only in their first terms, which are $p_\theta(x_1)$ and $p_\theta(x_1|\, x_0)$, respectively. This difference should be negligible if n is large.

Similarly, if the time series is *Markov of order* p, i.e. $p(x_t|\, x_{t-1}, x_{t-2}, \ldots)$ depends only on $x_t, x_{t-1}, \ldots, x_{t-p}$, then the two likelihoods differ only in p terms. This should be negligible if n is large relative to p.

A causal auto-regressive time series defined relative to an i.i.d. white noise series is an example of this situation. Maximum likelihood estimators for auto-regressive processes are commonly defined by using the pseudo likelihood with $X_0, \ldots, X_{-p+1}$ set equal to zero. Alternatively, we simply drop the first p terms of the likelihood, and work with the approximate likelihood

$$(\sigma, \phi_1, \ldots, \phi_p) \mapsto \prod_{t=p+1}^{n} p_{Z,\sigma}\bigl(X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}\bigr),$$

for $p_{Z,\sigma}$ the density of the innovations. This can also be considered a *conditional likelihood* given the observations $X_1, \ldots, X_p$. The difference of this likelihood with the true likelihood is precisely the marginal density of the vector $(X_1, \ldots, X_p)$, which is complicated in general, but should have a noticable effect on the maximum likelihood estimator only if p is large relative to n. □

12.2 Example (GARCH). A strictly stationary GARCH process X_t relative to an i.i.d. series Z_t can be written as $X_t = \sigma_t Z_t$, for $\sigma_t^2 = \mathrm{E}(X_t^2|\, \mathcal{F}_{t-1})$ and $\mathcal{F}_t$ the filtration generated by $X_t, X_{t-1}, \ldots$. From Theorem 8.10 it is known that the filtration $\mathcal{F}_t$ is also the natural filtration of the process Z_t and hence the variable Z_t is independent of σ_t^2, which is measurable relative to $\mathcal{F}_{t-1}$. It follows that the conditional distribution of $X_t = \sigma_t Z_t$ given $X_{t-1}, X_{t-2}, \ldots$ is obtained by first calculating σ_t^2 from $X_{t-1}, X_{t-2}, \ldots$ and next multiplying σ_t by an independent variable Z_t. If p_Z is the marginal density of the variables Z_t, then the pseudo likelihood (12.2) takes the form

$$\prod_{t=1}^{n} \frac{1}{\sigma_t} p_Z\Bigl(\frac{X_t}{\sigma_t}\Bigr). \tag{12.3}$$

The parameters $\alpha, \phi_1, \ldots \phi_p, \theta_1, \ldots, \theta_q$ are hidden in the process σ_t, through the GARCH relation (8.1). Formula (12.3) is not the true likelihood, because it depends on the unobservable variables $X_0, X_{-1}, \ldots$ through the σ_t.

For an ARCH(q) process the conditional variances σ_t^2 depend only on the variables $X_{t-1}, \ldots, X_{t-q}$, in the simple form

$$\sigma_t^2 = \alpha + \theta_1 X_{t-1}^2 + \cdots + \theta_q X_{t-q}^2.$$

In this case the true likelihood and the pseudolikelihood differ only in the first q of the n terms. This difference should be negligible. For practical purposes, if n is large relative to q, we could either drop those first q terms, giving a conditional likelihood, or act as if the unobserved variables $X_0^2, \ldots, X_{-q+1}^2$ are zero.

For general GARCH processes the difference between the likelihoods is more substantial, because the conditional variance σ_t^2 depends on $X_{t_1}^2,\ldots,X_{t-q}^2$ as well as on the previous σ_s^2, causing a dependence on the full past X_s^2 with $s<t$ of the process of squares. However, the dependenceof σ_t^2 on lagged variables X_s^2 decreases exponentially fast as $s\to\infty$, at least in the case of a second order stationary GARCH series. That the variables X_s^2 with $s\le 0$ do not play an important role in the definition of the (pseudo) likelihood is also suggested by Theorem 8.14, which shows that a GARCH series defined from arbitrary starting values converges to (strict) stationarity as t grows to infinity (provided that a strictly stationary GARCH process exists). This suggests that we might again use the pseudo likelihood with the unobserved variables X_s^2 replaced by zero. (The stability assertion of Theorem 8.14 certainly does not quarantee that this will work, as it says nothing about the likelihood function.)

A practical implementation is to define $\sigma_0^2,\ldots,\sigma_{-p+1}^2$ and $X_0^2,\ldots,X_{-q+1}^2$ to be zero, and next compute $\sigma_1^2,\sigma_2^2,\ldots$ recursively, using the GARCH relation (8.1) and the observed values $X_1,\ldots,X_n$. By Theorem 8.14 these zero starting values cannot be the true values of the series if the series is strictly stationary, but any other initialization should yield approximately the same likelihood. Given $\sigma_1^2,\sigma_2^2,\ldots$ and $X_1,\ldots,X_n$ we can use (12.3) as a contrast function. □

12.1.1 Consistency

To gain insight in the asymptotic properties of maximum likelihood estimators, we adopt the working hypothesis that these have the same asymptotic properties as the corresponding maximum pseudo likelihood estimators. Furthermore, we assume that the time series X_t is strictly stationary and ergodic. These conditions are certainly too stringent, but they simplify the arguments. The conclusions typically apply to any time series that "approaches stationarity" as $t\to\infty$ and for which averages converge to constants.

Abbreviate $x_t,x_{t-1},\ldots$ to $\vec{x}_t$. The maximum pseudo likelihood estimator maximizes the function

$$\theta\mapsto M_n(\theta)=\frac{1}{n}\sum_{t=1}^n\log p_\theta(X_t|\,\vec{X}_{t-1}). \tag{12.4}$$

If the variables $\log p_\theta(X_t|\,\vec{X}_{t-1})$ are integrable, as we assume, then, by the ergodic theorem, Theorem 4.17, the averages $M_n(\theta)$ converges to their expectation

$$M(\theta)=\mathrm{E}_{\theta_0}\log p_\theta(X_1|\,\vec{X}_0).$$

The expectation is taken under the "true" parameter θ_0 governing the distribution of the time series X_t. The difference of the expected values $M(\theta_0)$ and $M(\theta)$ can also be written as

$$M(\theta_0)-M(\theta)=\mathrm{E}_{\theta_0}\int\Bigl(\log\frac{p_{\theta_0}(x_1|\,\vec{X}_0)}{p_\theta(x_1|\,\vec{X}_0)}\Bigr)p_{\theta_0}(x_1|\,\vec{X}_0)\,d\mu(x_1).$$

The integral inside the expectation is the *Kullback-Leibler divergence* between the (conditional) measures with densities $p_\theta(\cdot|\,\vec{x}_0)$ and $p_{\theta_0}(\cdot|\,\vec{x}_0)$. It is well known that the Kullback-Leibler divergence between two probability measures is nonnegative and is zero if and only if the two measures are the same. Thus $M(\theta) \le M(\theta_0)$ for every θ with equality if and only if the two conditional measures are the same for almost every $\vec{x}_0$. Under the reasonable assumption that each value of θ indexes a different underlying distribution of the time series X_t ("identifiability of θ"), we conclude that the map $\theta \mapsto M(\theta)$ possesses a unique absolute maximum at $\theta = \theta_0$.

The convergence of the criterion functions M_n to M, and the definitions of $\hat\theta_n$ and θ_0 as the points of maxima of these functions suggest that $\hat\theta_n$ converges to θ_0. In other words, we expect the maximum likelihood estimators to be consistent for the "true" value θ_0. This argument can be made mathematically rigorous, for instance by imposing additional conditions that guarantee the uniform convergence of M_n to M. See e.g. Theorem 3.17.

12.1.2 Asymptotic Normality

If it is true that $\hat\theta_n \to \theta_0$, then the question poses itself to characterize the rate of convergence at which the difference $\hat\theta_n - \theta_0$ converges to zero and to find a possible limit distribution for this rescaled difference. For this we assume that the parameter set Θ is a subset of $\mathbb{R}^d$. Under smoothness conditions on the contrast function $\theta \mapsto M_n(\theta)$ we shall establish the asymptotic normality of the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$.

We assume that the gradient $\dot M_n(\theta)$ and second derivative matrix $\ddot M_n(\theta)$ of the map $\theta \mapsto M_n(\theta)$ exist and are continuous. Because $\hat\theta_n$ is a point of maximum of M_n, it satisfies the stationary equation $\dot M_n(\hat\theta_n) = 0$, if $\hat\theta_n$ is an inner point of the parameter set. Because we assume consistency of $\hat\theta_n$, this is the case with probability tending to 1 if θ_0 is an inner point, as we shall assume. By Taylor's theorem there exists a point $\tilde\theta_n$ on the line segment between θ_0 and $\hat\theta_n$ such that

$$0 = \dot M_n(\hat\theta_n) = \dot M_n(\theta_0) + \ddot M_n(\tilde\theta_n)(\hat\theta_n - \theta).$$

By simple algebra this can be rewritten as

$$\sqrt{n}(\hat\theta_n - \theta_0) = -\big(\ddot M_n(\tilde\theta_n)\big)^{-1}\sqrt{n}\dot M_n(\theta_0).$$

The matrices $\ddot M_n(\theta) = n^{-1}\sum_{i=1}^n \partial^2/\partial\theta^2 \log p_q(X_t|\,\vec{X}_{t-1})$ are averages and hence the ergodic theorem guarantees their convergence in probability to a fixed matrix under reasonable conditions. Because $\tilde\theta_n \xrightarrow{\mathrm{P}} \theta_0$ if θ_0 is the true parameter, it is a reasonable assumption that the matrices $\ddot M_n(\tilde\theta_n)$ and $\ddot M_n(\theta_0)$ possess the same limit. If we can also show that the sequence $\sqrt{n}\dot M_n(\theta_0)$ converges in distribution, then we can conclude that the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ converges in distribution, by Slutsky's lemma.

The convergence of the sequence $\sqrt{n}\dot M_n(\theta_0)$ can be established by the martingale central limit theorem, Theorem 4.35. To see this, first differentiate the identify $\int p_\theta(x_1|\,\vec{x}_0)\,d\mu(x_1) = 1$ twice to verify that

$$\int \dot p_\theta(x_1|\,\vec{x}_0)\,d\mu(x_1) = \int \ddot p_\theta(x_1|\,\vec{x}_0)\,d\mu(x_1) = 0.$$

The function $\ell_\theta(x_t|\,\vec{x}_{t-1}) = \log p_\theta(x_t|\,\vec{x}_{t-1})$ possesses partial derivatives relative to θ given by $\dot\ell_\theta = \dot p_\theta/p_\theta$ and $\ddot\ell_\theta = \ddot p_\theta/p_\theta - \dot\ell_\theta\dot\ell_\theta^T$. Combination with the preceding display yields conditional versions of the usual identities "expectation of score function is zero" and "expectation of observed information is minus the Fisher information", showing that

$$\begin{aligned} \mathrm{E}_\theta\Bigl(\dot\ell_\theta(X_1|\,\vec{X}_0)|\,\vec{X}_0\Bigr) &= 0,\\ \mathrm{Cov}_\theta\Bigl(\ddot\ell_\theta(X_1|\,\vec{X}_0)|\,\vec{X}_0\Bigr) &= -\mathrm{E}_\theta\Bigl(\dot\ell_\theta\dot\ell_\theta^T(X_1|\,\vec{X}_0)|\,\vec{X}_0\Bigr) = -\mathrm{E}_\theta\bigl(\ddot\ell_\theta(X_1|\,\vec{X}_0)|\,\vec{X}_0\bigr). \end{aligned}$$

The first identity shows that the sequence $\sqrt{n}\dot M_n(\theta) = n^{-1/2}\sum_{t=1}^n \dot\ell_\theta(X_t|\,\vec{X}_{t-1})$ is a martingale under the true measure specified by the parameter θ. Under reasonable conditions the martingale central limit theorem yields that the sequence $\sqrt{n}\dot M_n(\theta)$ is asymptotically normal with mean zero and covariance matrix

$$I_\theta = \mathrm{Cov}_\theta\bigl(\dot\ell_\theta(X_1|\,\vec{X}_0)\bigr).$$

By the second identity $\mathrm{E}_\theta\ddot M_n(\theta) = -I_\theta$ and hence we may expect that $\ddot M_n(\theta) \xrightarrow{\mathrm{P}} -I_\theta$, by the ergodic theorem. Combining this with Slutsky's lemma as indicated before, we find that, under true parameter θ_0,

$$\sqrt{n}(\hat\theta_n - \theta_0) \rightsquigarrow N(0, I_{\theta_0}^{-1}).$$

The matrix I_θ is known as the *Fisher information matrix*. Typically, it can also be found through the limits

$$\begin{aligned} I_{n,\theta} = \frac{1}{n}\mathrm{E}_\theta\frac{\partial}{\partial\theta}\log p_\theta(X_1,\ldots,X_n)\frac{\partial}{\partial\theta}\log p_\theta(X_1,\ldots,X_n)^T &\to I_\theta,\\ -\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_1,\ldots,X_n)_{|\theta=\hat\theta_n} &\to I_\theta. \end{aligned}$$

The expression on the left in the second line is the second derivative matrix of the likelihood surface at the maximum likelihood estimator, and is known as the *observed information*. It gives an estimate for the inverse of the asymptotic covariance matrix of the sequence $\sqrt{n}(\hat\theta_n - \theta)$. Thus a large observed information indicates that the maximum likelihood estimator has small asymptotic covariance.

The first of the preceding display connects the matrix I_θ to the definition of the Fisher information for arbitrary observations, which appears in the *Cramér-Rao bound* for the variance of unbiased estimators. According to the Cramér-Rao theorem, the covariance matrix of any unbiased estimator T_n of θ satisfies

$$\mathrm{Cov}_\theta\bigl(\sqrt{n}(T_n - \theta)\bigr) \geq (nI_{n,\theta})^{-1}.$$

The preceding informal derivation suggests that the asymptotic covariance matrix of the sequence $\sqrt{n}(\hat\theta_n - \theta)$ is equal to I_θ^{-1}. We interprete this as saying that the maximum likelihood estimator is asymptotically of minimal variance, or *asymptotically efficient*.

It is possible to give a rigorous proof of the asymptotic normality of the maximum likelihood estimator, and also of a precise formulation of its asymptotic efficiency. See ??

12.3 EXERCISE. Compute the conditional maximum likelihood estimator for (θ, σ^2) in a stationary, causal AR(1) model $X_t = \theta X_{t-1} + Z_t$ with Gaussian innovations Z_t. What is its limit distribution? Calculate the Fisher information matrix I_θ.

12.4 EXERCISE. Find the pair of (conditional) likelihood equations $\dot{M}_n(\alpha, \theta) = 0$ for estimating the parameters (α, θ) in an ARCH(1) model. Verify the martingale property of $n\dot{M}_n(\alpha, \theta)$.

12.2 Misspecification

Specification of a correct statistical model for a given time series is generally difficult, and it is typically hard to decide which of two given reasonable models is the better one. This observation is often taken as motivation for modelling a time series as a Gaussian series, Gaussianity being considered as good as any other specification and Gaussian likelihoods being relatively easy to handle. Meanwhile the validity of the Gaussian assumption may not really be accepted. It is therefore important, in time series analysis even more than in statistics for replicated experiments, to consider the behaviour of estimation procedures under misspecification of a model.

Consider an estimator $\hat{\theta}_n$ defined as the point of maximum of a likelihood function of a model that possibly does not contain the true density of the observations. It is again easier to consider the pseudo likelihood (12.2) than the true likelihood. The misspecified maximum pseudo likelihood estimator is still the point of maximum of the map $\theta \mapsto M_n(\theta)$ defined in (12.4). For the asymptotic analysis of $\hat{\theta}_n$ we again apply the ergodic theorem to see that $M_n(\theta) \to M(\theta)$ almost surely, for $M(\theta)$ the expectation of $M_n(\theta)$, defined by

$$M(\theta) = \mathrm{E}\ell_\theta(X_1 |\, \vec{X}_0).$$

The difference with the foregoing is that presently the expectation is taken under the true model for the series X_t, which may or may not be representable through one of the parameters θ. However, the same reasoning suggests that $\hat{\theta}_n$ converges in probability to a value θ_0 that maximizes the map $\theta \mapsto M(\theta)$. Without further specification of the model and the true distribution of the time series, there is little more we can say about this maximizing value than that it gives conditional densities $p_{\theta_0}(\cdot |\, \vec{x}_0)$ that are, on the average, closest to the true conditional densities $p(\cdot |\, \vec{x}_0)$ of the time series in terms of the Kullback-Leibler divergence.

Having ascertained that the sequence $\hat{\theta}_n$ ought to converge to a limit, most of the subsequent arguments to establish asymptotic normality of the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ go through, also under misspecification, provided that

$$\mathrm{E}\bigl(\dot{\ell}_{\theta_0}(X_1 |\, \vec{X}_0) |\, \vec{X}_0\bigr) = 0, \qquad \text{a.s.}. \tag{12.5}$$

In that case the sequence $\sqrt{n}\dot{M}_n(\theta_0)$ is still a martingale, and may be expected to be asymptotically normal by the martingale central limit theorem. By the assumed ergodicity of the series X_t the sequence $\ddot{M}_n(\theta_0)$ will still converge to a fixed matrix, and the same may be expected to be true for the sequence of second derivatives $\ddot{M}_n(\tilde{\theta}_n)$ evaluated at a point between $\hat{\theta}_n$ and θ_0. A difference is that the asymptotic covariance matrix Σ_{θ_0} of the sequence $\sqrt{n}\dot{M}_n(\theta_0)$ and the limit R_{θ_0} of the sequence $\ddot{M}_n(\theta_0)$ may no longer be each other's negatives. The conclusion will therefore take the more complicated form

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\bigl(0, R_{\theta_0}^{-1}\Sigma_{\theta_0}(R_{\theta_0}^{-1})^T\bigr).$$

The asymptotic covariance of the normal limit distribution is referred to as the *sandwich formula*.

Thus under (12.5) we may expect that the sequence $\hat{\theta}_n$ will converge rapidly to a limit θ_0. Then "fitting the wrong model" will be useful as long as the density p_{θ_0} is sufficiently close to the true distribution of the time series.

Condition (12.5) is odd, and it appears that it is not always satisfied. It is certainly satisfied if the point of maximum θ_0 of the map $\theta \mapsto M(\theta)$ is such that for every $\vec{x}_0$ it is also a point of maximum of the map, with p the true conditional density of the time series,

$$\theta \mapsto \int \log p_\theta(x_1|\, \vec{x}_0)\, p(x_1|\, \vec{x}_0)\, d\mu(x_1).$$

This is not necessarily the case, as the points of maxima of the functions in the display may be different for different values of $\vec{x}_0$. The point θ_0 is by definition the point of maximum of the average of these functions over $\vec{x}_0$, weighted by the distribution of $\vec{X}_0$. Failure of (12.5) does not necessarily mean that the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is not asymptotically normally distributed, but it does mean that we cannot apply the martingale central limit theorem, as in the preceding argument.

In the next section we discuss a major example of possible misspecification: estimating a parameter by Gaussian maximum likelihood. The following example concerns GARCH processes, and illustrates that some misspecifications are harmless, whereas others may cause trouble.

12.5 Example (GARCH). As found in Example 12.2, the pseudo likelihood for a GARCH(p, q) process takes the form

$$\prod_{t=1}^n \frac{1}{\sigma_t(\theta)} p_Z\Bigl(\frac{X_t}{\sigma_t(\theta)}\Bigr),$$

where p_Z is the density of the innovations and $\theta = (\alpha, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$. In Chapter 8 it was noted that a t-density p_Z may be appropriate to explain the observed leptokurtic tails of financial time series. However, the Gaussian density $p_Z(z) = \exp(-\frac{1}{2}z^2)/\sqrt{2\pi}$ is more popular for likelihood based inference for GARCH processes. The corresponding log pseudo likelihood is up to additive and multiplicative constants equal to

$$\theta \mapsto -\frac{1}{n}\sum_{t=1}^n \log \sigma_t^2(\theta) - \frac{1}{n}\sum_{t=1}^n \frac{X_t^2}{\sigma_t^2(\theta)}.$$

The expectation of this criterion function can be written as

$$M(\theta) = -\mathrm{E}\Bigl(\log \sigma_1^2(\theta) + \frac{\mathrm{E}(X_1^2|\, \mathcal{F}_0)}{\sigma_1^2(\theta)}\Bigr).$$

Both expectations on the right side are taken relative to the true distribution of the time series. The sequence $\hat\theta_n$ may be expected to converge to the point of maximum of the map $\theta \mapsto M(\theta)$.

Suppose that the GARCH equation (8.1) for the conditional variances is correctly specified, even though the true density of the innovations may not be standard normal. In other words, $\mathrm{E}(X_1^2|\,\mathcal{F}_0) = \sigma_1^2(\theta_0)$ for the true parameter θ_0 and hence

$$M(\theta) = -\mathrm{E}\Bigl(\log \sigma_1^2(\theta) + \frac{\sigma_1^2(\theta_0)}{\sigma_1^2(\theta)}\Bigr).$$

For every fixed σ_0^2, the map $\sigma^2 \mapsto \log\sigma^2 + \sigma_0^2/\sigma^2$ assumes its minimal value on the domain $(0,\infty)$ at $\sigma^2 = \sigma_0^2$. It follows that the map $\theta \mapsto M(\theta)$ is maximized at $\theta = \theta_0$ no matter the distribution of $\vec X_0$ that determines the expectation that defines $M(\theta)$.

We conclude that the use of the Gaussian density for p_Z will lead to consistent estimators $\hat\theta_n$ for the coefficients of the GARCH equation as long as the conditional variance model is correctly specified. In particular, the true density p_Z of the innovations need not be Gaussian. As shown in the preceding arguments, this pleasant fact is the result of the fact that the likelihood based on choosing the normal density for p_Z depends on the observations X_t only through a linear function of the squares X_t^2. For another choice of density p_Z, such as a t-density, this is not true, and we cannot hope to be guarded against misspecification in that case.

The maximizing parameter θ_0 in this model can be seen to yield a conditional density $p_{\theta_0}(\cdot|\, x_0, x_{-1},\ldots)$ that is closest to the true conditional density $p(\cdot|\, x_0, x_{-1},\ldots)$ relative to the Kullback-Leibler divergence, for any given values $x_0, x_{-1},\ldots$. This implies that equation (12.5) is satisfied in this case, and hence we expect the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ to be asymptotically normal, with asymptotic variance given by the sandwich formula. □

12.3 Gaussian Likelihood

A *Gaussian time series* is a time series X_t such that the joint distribution of every finite subvector $(X_{t_1},\ldots,X_{t_n})$ of the series possesses a multivariate normal distribution. In particular, the vector $(X_1,\ldots,X_n)$ is multivariate normally distributed, and hence its distribution is completely specified by a mean vector $\mu_n \in \mathbb{R}^n$ and an $(n\times n)$ covariance matrix Γ_n. If the time series X_t is covariance stationary, then the matrix Γ_n has entries $(\Gamma_n)_{s,t} = \gamma_X(s-t)$, for γ_X the auto-covariance function of X_t. We assume that both the mean μ_n and the covariance function γ_X can be expressed in a parameter θ of fixed dimension, so that we can write $\mu_n = \mu_n(\theta)$ and $\Gamma_n = \Gamma_n(\theta)$.

The likelihood function under the assumption that X_t is a Gaussian time series is the multivariate normal density viewed as function of the parameter and takes the form

$$\theta \mapsto \frac{1}{(2\pi)^{n/2}}\frac{1}{\sqrt{\det\Gamma_n(\theta)}}e^{-\frac12(\vec X_n-\mu_n(\theta))^T\Gamma_n(\theta)^{-1}(\vec X_n-\mu_n(\theta))}. \tag{12.6}$$

We refer to this function as the *Gaussian likelihood*, and to its point of maximum $\hat{\theta}_n$, if it exists, as the *maximum Gaussian likelihood estimator*. The Gaussian likelihood and the corresponding estimator are commonly used, also in the case that the time series X_t is non-Gaussian.

Maximum Gaussian likelihood is closely related to the method of least squares, described in Section 10.3. We can see this using the likelihood factorization (12.1). For a Gaussian process the conditional densities $p_\theta(x_t|\, X_{t-1},\ldots,X_1)$ are univariate normal densities with means $\mathrm{E}_\theta(X_t|\, X_{t-1},\ldots,X_1)$ and variances $v_{t-1}(\theta)$ equal to the prediction errors. (Cf. Exercise 12.6.) Furthermore, the best nonlinear predictor $\mathrm{E}_\theta(X_t|\, X_{t-1},\ldots,X_1)$ is automatically a linear combination of the predicting variables and hence coincides with the best linear predictor $\Pi_{t-1}X_t(\theta)$. This shows that the factorization (12.1) reduces to

$$\prod_{t=1}^n \frac{1}{\sqrt{v_{t-1}(\theta)}}\phi\Bigl(\frac{X_t-\Pi_{t-1}X_t(\theta)}{\sqrt{v_{t-1}(\theta)}}\Bigr).$$

Maximizing this relatively to θ is equivalent to maximizing its logarithm, which can be written in the form

$$\theta\mapsto -\frac{n}{2}\log(2\pi)-\frac{1}{2}\sum_{t=1}^n\log v_{t-1}(\theta)-\frac{1}{2}\sum_{t=1}^n\frac{\bigl(X_t-\Pi_{t-1}X_t(\theta)\bigr)^2}{v_{t-1}(\theta)}. \tag{12.7}$$

This function differs in form from the least squares criterion function (10.3) only in the presence of the function $\theta\mapsto -\frac{1}{2}\sum_{t=1}^n\log v_{t-1}(\theta)$. In situations where this function is almost constant least squares and Gaussian maximum likelihood estimators are almost the same.

12.6 EXERCISE. Suppose that the vector $(X_1,\ldots,X_t)$ possesses a multivariate normal distribution. Show that the conditional distribution of X_t given $(X_1,\ldots,X_{t-1})$ is normal with mean $\mathrm{E}(X_t|\, X_1,\ldots,X_{t-1})$ and variance $\mathrm{var}(X_t|\, X_1,\ldots,X_{t-1})$. [Write $X_t=(X_t-Y)+Y$ for $Y=\mathrm{E}(X_t|\, X_1,\ldots,X_{t-1})$, prove that Y is a linear function of $X_1,\ldots,X_{t-1}$, and conclude that the vector (X_t-Y,Y) is bivariate normal with correlation zero. Conclude that X_t-Y and Y are independent.]

12.7 Example (Auto regression). For causal stationary auto-regressive processes of order p and $t>p$ the best linear predictor of X_t is equal to $\phi_1X_{t-1}+\cdots+\phi_pX_{t-p}$. Thus the innovations $X_t-\Pi_{t-1}X_t$ are equal to the noise input Z_t, and the prediction errors v_t are equal to $\sigma^2=\mathrm{E}Z_{t+1}^2$ for $t\geq p$. Thus the function $\theta\mapsto -\frac{1}{2}\sum_{t=1}^n\log v_{t-1}(\theta)$ in the formula for the Gaussian likelihood is approximately equal to $-\frac{1}{2}n\log\sigma^2$, if n is much bigger than p. The log Gaussian likelihood is approximately equal to

$$-\frac{n}{2}\log(2\pi)-\frac{1}{2}n\log\sigma^2-\frac{1}{2}\sum_{t=p+1}^n\frac{\bigl(X_t-\phi_1X_{t-1}-\cdots-\phi_pX_{t-p}\bigr)^2}{\sigma^2}.$$

For a fixed σ^2 maximization relative to $\phi_1,\ldots,\phi_p$ is equivalent to minimization of the sum of squares and hence gives identical results as the method of least squares discussed in Sections 10.1 and 10.3. Maximization relative to σ^2 gives (almost) the Yule-Walker estimator discussed in Section 10.1. □

12.8 Example (ARMA). In ARMA models the parameter σ^2 enters as a multiplicative factor in the covariance function (cf. Section 10.3). This implies that the log Gaussian likelihood function can be written in the form, with $\theta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$,

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{t=1}^n \log\tilde{v}_{t-1}(\theta) - \frac{1}{2}\sum_{t=1}^n \frac{\big(X_t - \Pi_{t-1}X_t(\theta)\big)^2}{\sigma^2\tilde{v}_{t-1}(\theta)}.$$

Differentiating this with respect to σ^2 we see that for every fixed θ, the Gaussian likelihood is maximized relative to σ^2 by

$$\hat{\sigma}^2(\theta) = \frac{1}{n}\sum_{t=1}^n \frac{\big(X_t - \Pi_{t-1}X_t(\theta)\big)^2}{\tilde{v}_{t-1}(\theta)}.$$

Substituting this expression in the log Gaussian likelihood, we see that the maximum Gaussian likelihood estimator of θ maximizes the function

$$\theta \mapsto -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2(\theta) - \frac{1}{2}\sum_{t=1}^n \log\tilde{v}_{t-1}(\theta) - \frac{n}{2}.$$

The latter function is called the *profile likelihood* for θ, and the process of eliminating the parameter σ^2 is referred to as *concentrating out* this parameter. We can drop the constant terms in the profile likelihood and conclude that the maximum Gaussian likelihood estimator $\hat{\theta}$ for θ minimizes

$$\theta \mapsto \log\frac{1}{n}\sum_{t=1}^n \frac{\big(X_t - \Pi_{t-1}X_t(\theta)\big)^2}{\tilde{v}_{t-1}(\theta)} + \frac{1}{2}\sum_{t=1}^n \log\tilde{v}_{t-1}(\theta). \tag{12.8}$$

The maximum Gaussian likelihood estimator for σ^2 is $\hat{\sigma}^2(\hat{\theta})$.

For causal, invertible stationary ARMA processes the innovations $X_t - \Pi_{t-1}X_t$ are for large t approximately equal to Z_t, whence $\tilde{v}_{t-1}(\theta) \approx \mathrm{E}Z_t^2/\sigma^2 = 1$. (Cf. the discussion in Section 7.4. In fact, it can be shown that $|\tilde{v}_{t-1} - 1| \le c^t$ for some $0 < c < 1$ and sufficiently large t.) This suggests that the criterion function (12.8) does not change much if we drop its second term and retain only the sum of squares. The corresponding approximate maximum Gaussian likelihood estimator is precisely the least squares estimator, discussed in Section 10.3. □

12.9 Example (GARCH). The distribution of a GARCH process $X_t = \sigma_t Z_t$ depends on the distribution of the innovations Z_t, but is rarely (or never?) Gaussian. Nevertheless we may try and apply the method of Gaussian likelihood.

Because a GARCH series is a white noise series, the linear one-step ahead predictors are identically zero, and the prediction variances are equal to the variances $v_{t-1}^2 = \mathrm{E}X_t^2$ of the process. For a stationary GARCH process these are constant and can be expressed in the parameters of the GARCH process. For instance, for the GARCH$(1,1)$ process we have that $\mathrm{E}X_t^2 = \alpha/(1-\phi-\theta)$. Because the predictions are zero, the Gaussian likelihood depends on the parameters of the model only through the prediction variances $v_{t-1}^2 = \mathrm{E}X_t^2$. It follows that the likelihood is constant on sets of constant prediction variance and hence can at best yield good estimators for functions of this variance. The GARCH parameters cannot be recovered from this. For instance, we cannot estimate the parameter (α,ϕ,θ) of a GARCH $(1,1)$ process from a criterion function that depends on these parameters only through $\alpha/(1-\phi-\theta)$.

We conclude that the method of Gaussian likelihood is useless for GARCH processes. (We note that the Gaussian likelihood is similar in form to the likelihood obtained by assuming that the innovations Z_t are Gaussian (cf. Example 12.2), but with the conditional variances σ_t^2 in the latter replaced by their expectations.) □

In the preceding examples we have seen that for AR and ARMA processes the Gaussian maximum likelihood estimators are, asymptotically as $n\to\infty$, close to the least squares estimators. The following theorem shows that the asymptotic behaviour of these estimators is identical to that of the least squares estimators, which is given in Theorem 10.19.

12.10 Theorem. *Let X_t be a causal, invertible stationary ARMA(p,q) process relative to an i.i.d. sequence Z_t. Then the Gaussian maximum likelihood estimator satisfies*

$$\sqrt{n}\left(\begin{pmatrix}\hat{\vec{\phi}}_p\\ \hat{\vec{\theta}}_q\end{pmatrix}-\begin{pmatrix}\vec{\phi}_p\\ \vec{\theta}_q\end{pmatrix}\right)\rightsquigarrow N(0,\sigma^2 J^{-1}_{\vec{\phi}_p,\vec{\theta}_q}),$$

where $J_{\vec{\phi}_p,\vec{\theta}_q}$ is the covariance matrix of $(U_{-1},\ldots,U_{-p},V_{-1},\ldots,V_{-q})$ for stationary autoregressive processes U_t and V_t satisfying $\phi(B)U_t=\theta(B)V_t=Z_t$.

Proof. The proof is long and technical. See Brockwell and Davis (1991), pages 375–396, Theorem 10.8.2. ■

The theorem does not assume that the time series X_t itself is Gaussian; it uses the Gaussianity only as a working hypothesis to define maximum likelihood estimators. Apparently, using “the wrong likelihood” still leads to reasonable estimators. This is plausible, because Gaussian maximum likelihood estimators are asymptotically equivalent to least squares estimators and the method of least squares can be motivated without reference to Gaussianity. Alternatively, it can be explained from a consideration of the Kullback-Leibler divergence, as in Section 12.2.

On the other hand, in the case that the series X_t is not Gaussian the true maximum likelihood estimators (if the true model, i.e. the true distribution of the noise factors Z_t is known) are likely to perform better than the least squares estimators. In this respect time series analysis is not different from the situation for replicated experiments. An important difference is that in practice non-Gaussianity may be difficult to detect, other plausible distributions difficult to motivate, and other likelihoods may yield computational problems. The Gaussian distribution is therefore frequently adopted as a working hypothesis.

12.3.1 Whittle Estimators

Because the Gaussian likelihood function of a mean zero time series depends on the autocovariance function only, it can be helpful to write it in terms of the spectral density. The covariance matrix of a vector $(X_1,\ldots,X_n)$ belonging to a stationary time series X_t with spectral density f_X can be written as $\Gamma_n(f_X)$, for

$$\Gamma_n(f)=\Bigl(\int_{-\pi}^{\pi} e^{i(s-t)}f(\lambda)\,d\lambda\Bigr)_{s,t=1,\ldots,n}.$$

Thus if the time series X_t has spectral density f_θ under the parameter θ and mean zero, then the log Gaussian likelihood can be written in the form

$$-\frac{n}{2}\log(2\pi)-\frac12\log\det\Gamma_n(f_\theta)-\frac12\vec{X}_n^T\Gamma_n(f_\theta)^{-1}\vec{X}_n.$$

Maximizing this expression over θ is equivalent to maximizing the Gaussian likelihood as discussed previously, but gives a different perspective. For instance, to fit an ARMA process we would maximize this expression over all "rational spectral densities" of the form $\sigma^2\bigl|\theta(e^{-i\lambda})\bigr|^2/\bigl|\phi(e^{-i\lambda})\bigr|^2$.

The true advantage of writing the likelihood in spectral notation is that it suggests a convenient approximation. The *Whittle approximation* is defined as

$$-\frac{n}{2}\log(2\pi)-\frac{n}{4\pi}\int_{-\pi}^{\pi}\log f_\theta(\lambda)\,d\lambda-\frac{n}{4\pi}\int_{-\pi}^{\pi}\frac{I_n(\lambda)}{f_\theta(\lambda)}\,d\lambda,$$

where $I_n(\lambda)$ is the periodogram of the time series X_t, as defined in Section 11.2. This approximation results from the following approximations, for a sufficiently regular function f,

$$\Gamma_n(f)^{-1}\approx\Gamma_n\Bigl(\frac1f\Bigr)\frac{1}{4\pi^2},$$

$$\frac1n\log\det\Gamma_n(f)\approx\log(2\pi)+\frac{1}{2\pi}\int_{-\pi}^{\pi}\log f(\lambda)\,d\lambda,$$

combined with the identity

$$\frac1n\vec{X}_n^T\Gamma_n(f)\vec{X}_n=2\pi\int I_n(\lambda)f(\lambda)\,d\lambda.$$

The approximations are made precise in Lemma ?, whereas the identity follows by some algebra.

12.11 EXERCISE. Verify the identity in the preceding display.

The Whittle approximation is both more convenient for numerical manipulation and more readily amenable to theoretical analysis. The point of maximum $\hat\theta_n$ of the Whittle approximation, if it exists, is known as the *Whittle estimator*. Conceptually, this again comes down to a search in the class of spectral densities f_θ defined through the model.

12.12 Example (Auto-regressive process). For an auto-regressive time series of fixed order the Whittle estimators are identical to the Yule-Walker estimators, which are also (almost) identical to the maximum Gaussian likelihood estimators. This can be seen as follows.

The Whittle estimators are defined by maximizing the Whittle approximation over all spectral densities of the form $f_\theta(\lambda)=\sigma^2/\bigl|\phi(e^{-i\lambda})\bigr|^2$, for ϕ the auto-regressive polynomial $\phi(z)=1-\phi_1 z-\cdots-\phi_p z^p$. By the Kolmogorov-Szegö formula (See ?), or direct computation, $\int\log f_\theta(\lambda)\,d\lambda=2\pi\log(\sigma^2/2\pi)$ is independent of the parameters $\phi_1,\ldots,\phi_p$. Thus the stationary equations for maximizing the Whittle approximation with respect to the parameters take the form

$$\begin{aligned}0=\frac{\partial}{\partial\phi_k}\int\frac{I_n(\lambda)}{\sigma^2 f_\theta(\lambda)}\,d\lambda&=\frac{\partial}{\partial\phi_k}\int\phi(e^{i\lambda})\phi(e^{-i\lambda})I_n(\lambda)\,d\lambda\\&=\int\Bigl[-e^{i\lambda k}\phi(e^{-i\lambda})-\phi(e^{i\lambda})e^{-i\lambda k}\Bigr]I_n(\lambda)\,d\lambda\\&=-2\,\mathrm{Re}\int\Bigl[e^{i\lambda k}-\phi_1 e^{-i\lambda(k-1)}-\cdots-\phi_p e^{i\lambda(k-p)}\Bigr]I_n(\lambda)\,d\lambda\\&=-2\,\mathrm{Re}\Bigl[\hat\gamma_n^*(k)-\phi_1\hat\gamma_n^*(k-1)-\cdots-\phi_p\hat\gamma_n^*(k-p)\Bigr],\end{aligned}$$

because $\hat\gamma_n^*(h)=n^{-1}\sum_{t=1}^{n-h}X_{t+h}X_t$ are the Fourier coefficients of the function I_n for $0\le h<n$, by (11.2). Thus the stationary equations are the Yule-Walker equations, apart from the fact that the observations have been centered at mean zero, rather than $\overline{X}_n$. □

12.13 EXERCISE. Derive the Whittle estimator for σ^2 for an autoregressive process.

If we write $I_n(f)$ for $\int I_n(\lambda)f(\lambda)\,d\lambda$, then a Whittle estimator is a point of minimum of the map

$$\theta\mapsto M_n(\theta)=\int_{-\pi}^{\pi}\log f_\theta(\lambda)\,d\lambda+I_n\Bigl(\frac{1}{f_\theta}\Bigr).$$

In Section 11.4 it is shown that the sequence $\sqrt{n}\bigl(I_n(f)-\int f f_X\,d\lambda\bigr)$ is asymptotically normally distributed with mean zero and some variance $\sigma^2(f)$, under some conditions. This implies that the sequence $M_n(\theta)$ converges for every fixed θ in probability to

$$M(\theta)=\int\log f_\theta(\lambda)\,d\lambda+\int\frac{f_X}{f_\theta}\,d\lambda.$$

By reasoning as in Section 12.1 we expect that the Whittle estimators $\hat\theta_n$ will be asymptotically consistent for the parameter θ_0 that minimizes the function $\theta\mapsto M(\theta)$.

If the true spectral density f_X takes the form f_{θ_0} for some parameter θ_0, then this parameter is the minimizing value. Indeed, by the inequality $-\log x+(x-1)\ge\bigl(\sqrt{x}-1\bigr)^2$, valid for every $x\ge0$,

$$M(\theta)-M(\theta_0)=\int\Bigl(\log\frac{f_\theta}{f_{\theta_0}}(\lambda)+\frac{f_{\theta_0}}{f_\theta}-1\Bigr)\,d\lambda\ge\int\Bigl(\sqrt{\frac{f_{\theta_0}}{f_\theta}(\lambda)}-1\Bigr)^2\,d\lambda.$$

This shows that the function $\theta \mapsto M(\theta)$ possesses a minimum value at $\theta = \theta_0$, and this point of minimum is unique as soon as the parameter θ is identifiable from the spectral density.

To derive the form of the limit distribution of the Whittle estimators we replace $\dot M(\theta)$ by its linear approximation, as in Section 12.1, and obtain that

$$\sqrt{n}(\hat\theta_n - \theta_0) = -\bigl(\ddot M_n(\tilde\theta_n)\bigr)^{-1}\sqrt{n}\dot M_n(\theta_0).$$

Denoting the gradient and second order derivative matrix of the function $\theta \mapsto \log f_\theta(\lambda)$ by $\dot\ell_\theta(\lambda)$ and $\ddot\ell_\theta(\lambda)$, we can write

$$\begin{aligned}\sqrt{n}\dot M_n(\theta) &= \int \dot\ell_\theta(\lambda)\,d\lambda - I_n\Bigl(\frac{\dot\ell_\theta}{f_\theta}\Bigr),\\ \ddot M_n(\theta) &= \int \ddot\ell_\theta(\lambda)\,d\lambda + I_n\Bigl(\frac{\dot\ell_\theta\dot\ell_\theta^T - \ddot\ell_\theta}{f_\theta}\Bigr).\end{aligned}$$

By the results of Section 11.4 the sequence $\sqrt{n}\dot M_n(\theta_0)$ converges in distribution to a normal distribution with mean zero and variance $\sigma^2(\dot\ell_{\theta_0}/f_{\theta_0})$, under some conditions. Furthermore, the sequence $\ddot M_n(\theta_0)$ converges in probability to $\int \dot\ell_{\theta_0}\dot\ell_{\theta_0}^T(\lambda)\,d\lambda =: J_{\theta_0}$. If both are satisfied, then we obtain that

$$\sqrt{n}(\hat\theta_n - \theta_0) \rightsquigarrow N\Bigl(0, J_{\theta_0}^{-1}\sigma^2(\dot\ell_{\theta_0}/f_{\theta_0})J_{\theta_0}^{-1}\Bigr).$$

The asymptotic covariance is of the "sandwich form", but reduces to a simpler expression in the case that the time series X_t is Gaussian, and the Whittle likelihood is an approximation for the correctly specified likelihood. In this case,

$$\sigma^2(f) = 4\pi\int ff^T(\lambda)f_X^2(\lambda)\,d\lambda.$$

It follows that in the case, and with $f_X = f_{\theta_0}$, the asymptotic covariance of the sequence $\sqrt{n}\dot M_n(\theta_0)$ reduces to $4\pi J_{\theta_0}$, and the sandwich covariance reduces to $4\pi J_{\theta_0}^{-1}$.

12.14 Example (ARMA). The log spectral density of a stationary, causal, invertible ARMA(p,q) process with parameter vector $\theta = (\sigma^2, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ can be written in the form

$$\log f_\theta(\lambda) = \log\sigma^2 - \log(2\pi) + \log\theta(e^{i\lambda}) + \log\theta(e^{-i\lambda}) - \log\phi(e^{i\lambda}) - \log\phi(e^{-i\lambda}).$$

Straightforward differentiation shows that the gradient of this function is equal to

$$\dot\ell_\theta(\lambda) = \begin{pmatrix} \sigma^{-2}\\ \frac{e^{i\lambda k}}{\phi(e^{i\lambda})} + \frac{e^{-i\lambda k}}{\phi(e^{-i\lambda})}\\ \frac{e^{i\lambda l}}{\theta(e^{i\lambda})} + \frac{e^{-i\lambda l}}{\theta(e^{-i\lambda})}\end{pmatrix}$$

Here the second and third lines of the vector on the right are abbreviations of vectors of length p and q, respectively, obtained by letting k and l range over the values $1,\ldots,p$ and $1,\ldots,q$, respectively. The matrix $J_\theta = \int \dot\ell_\theta \dot\ell_\theta^T(\lambda)\,d\lambda$ takes the form

$$J_\theta = \begin{pmatrix} \frac{2\pi}{\sigma^4} & 0 & 0 \\ 0 & AR & MAAR \\ 0 & MAAR^T & MA \end{pmatrix},$$

where AR, MA, and $MAAR$ are matrices of dimensions $(p\times p)$, $(q\times q)$ and $(p\times q)$, respectively, which are described in more detail in the following. The zeros must be replicated to fulfil the dimension requirements, and result from calculations of the type, for $k \geq 1$,

$$\int \frac{e^{i\lambda k}}{\phi(e^{i\lambda})}\,d\lambda = \frac{1}{i}\int_{|z|=1} \frac{z^{k-1}}{\phi(z)}\,dz = 0,$$

by Cauchy's theorem, because the function $z \mapsto z^{k-1}/\phi(z)$ is analytic on a neighbourhood of the unit disc, by the assumption of causility of the ARMA process.

Using the identity $(f+\overline{f})(g+\overline{g}) = 2\,\mathrm{Re}(fg + f\overline{g})$ we can compute the (k,l)-element of the matrix MA as

$$\begin{aligned} 2\,\mathrm{Re}\int\Big[\frac{e^{i\lambda k}}{\theta(e^{i\lambda})}\frac{e^{i\lambda l}}{\theta(e^{i\lambda})} + \frac{e^{i\lambda k}}{\theta(e^{i\lambda})}\frac{e^{-i\lambda l}}{\theta(e^{-i\lambda})}\Big]\,d\lambda \\ = 2\,\mathrm{Re}\Big(0 + \int \frac{e^{i\lambda(k-l)}}{|\theta(e^{i\lambda})|^2}\,d\lambda\Big) = 2\gamma_V(k-l)2\pi, \end{aligned}$$

where V_t is a stationary auto-regressive process satisfying $\theta(B)V_t = Z_t$ for a white noise process Z_t of unit variance. The matrix AR can be expressed similarly as the covariance matrix of p consecutive elements of an auto-regressive process U_t satisfying $\phi(B)U_t = Z_t$. The (k,l)-element of the matrix $MAAR$ can be written in the form

$$2\,\mathrm{Re}\int\Big[\frac{e^{i\lambda k}}{\phi(e^{i\lambda})}\frac{e^{i\lambda l}}{\theta(e^{i\lambda})} + \frac{e^{i\lambda k}}{\phi(e^{i\lambda})}\frac{e^{-i\lambda l}}{\theta(e^{-i\lambda})}\Big]\,d\lambda = 2\,\mathrm{Re}\Big(0 + 2\pi\int f_{UV}(\lambda)e^{i\lambda(k-l)}\,d\lambda\Big).$$

Here $f_{UV}(\lambda) = 1/\big(2\pi\phi(e^{i\lambda})\theta(e^{-i\lambda})\big)$ is the cross spectral density of the auto-regressive processes U_t and V_t defined previously (using the same white noise process Z_t)(?). Hence the integral on the far left is equal to 2π times the complex conjugate of the cross covariance $\gamma_{UV}(k-l)$.

Taking this all together we see that the matrix resulting from deleting the first row and first column from the matrix $J_\theta/(4\pi)$ results in the matrix $J_{\vec\phi_p,\vec\theta_q}$ that occurs in Theorem 12.10. Thus the Whittle estimators and maximum Gaussian likelihood estimators have asymptotically identical behaviour.

The Whittle estimator for σ^2 is asymptotically independent of the estimators of the remaing parameters. □

12.3.2 Gaussian Time Series

In this section we study the behaviour of the maximum likelihood estimators for general Gaussian time series in more detail. Thus $\hat{\theta}_n$ is the point of maximum of (12.6) (or equivalently (12.7)), and we study the properties of the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ under the assumption that the true density of $(X_1, \ldots, X_n)$ possesses the form (12.6), for some θ. For simplicity we assume that the time series is centered at mean zero, so that the model is completely parametrized by the covariance matrix $\Gamma_n(\theta)$. Equivalently, it is determined by the spectral density f_θ, which is related to the covariance matrix by

$$\bigl(\Gamma_n(\theta)\bigr)_{s,t} = \int_{-\pi}^{\pi} e^{i(s-t)\lambda}\, f_\theta(\lambda)\, d\lambda.$$

It is easier to express conditions and results in terms of the spectral density f_θ, which is fixed, than in terms of the sequence of matrices $\Gamma_n(\theta)$. The asymptotic *Fisher information* for θ is defined as

$$I_\theta = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial \log f_\theta}{\partial\theta}(\lambda) \Bigl(\frac{\partial \log f_\theta}{\partial\theta}(\lambda)\Bigr)^T d\lambda.$$

12.15 Theorem. *Suppose that X_t is a Gaussian time series with zero mean and spectral density f_θ such that the map $\theta \mapsto f_\theta$ is one-to-one and the map $(\theta, \lambda) \mapsto f_\theta(\lambda)$ is three times continuously differentiable and strictly positive. Suppose that θ ranges over a bounded, open subset of $\mathbb{R}^d$. Then the maximum likelihood estimator $\hat{\theta}_n$ based on $X_1, \ldots, X_n$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, I_\theta^{-1})$.*

Proof. See Azencott and Dacunha-Castelle (1984), Chapitre XIII. ∎

The theorem is similar in form to the theorem for maximum likelihood estimators based on replicated experiments. If $p_{n,\theta}$ is the density of $\vec{X}_n = (X_1, \ldots, X_n)$ (given in (12.6)), then it can be shown under the conditions of the theorem that

$$I_{n,\theta} := \frac{1}{n} \mathrm{E}_\theta \frac{\partial}{\partial\theta} \log p_{n,\theta}(\vec{X}_n) \Bigl(\frac{\partial}{\partial\theta} \log p_{n,\theta}(\vec{X}_n)\Bigr)^T \to I_\theta.$$

The left side of this display is the true Fisher information for θ based on $\vec{X}_n$, and this explains the name asymptotic Fisher information for I_θ. With this in mind the analogy with the situation for replicated experiments is perfect.

12.4 Model Selection

In the preceding sections and chapters we have studied estimators for the parameters of ARMA or GARCH processes assuming that the orders p and q are known a-priori. In practice reasonable values of p and q can be chosen from plots of the auto-correlation and the partial auto-correlation functions, followed by diagnostic checking after fitting a particular model. Alternatively (or in addition) we can estimate appropriate values of p and q and the corresponding parameters simultaneously from the data. The maximum likelihood method must then be augmented by penalization.

The value of the likelihood (12.1) depends on the dimension the parameter θ. If models of different dimension are available, then we can make the dependence explicit by denoting the log likelihood as, with d the dimension of the model,

$$\Lambda_n(\theta,d)=\sum_{t=1}^n \log p_{\theta,d}(X_t|\,X_{t-1},\ldots,X_1).$$

A first idea to select a reasonable dimension is to maximize the function Λ_n jointly over (θ,d). This rarely works. The models of interest are typically nested in that a model of dimension d is a submodel of a model of dimension $d+1$. The maximum over (θ,d) is then taken for the largest possible dimension. To counter this preference for large dimension we can introduce a *penalty function*. Instead of Λ_n we maximize

$$(\theta,d)\mapsto \Lambda_n(\theta,d)-\phi_n(d),$$

where ϕ_n is a fixed function that takes large values for large values of its argument. Maximizing this function jointly over (θ,d) must strike a balance between maximizing Λ_n, which leads to big values of d, and minimizing ϕ_n, which leads to small values of d. The choice of penalty function is crucial for this balance to yield good results.

Several penalty functions are in use, each of them motivated by certain considerations. There is no general agreement as to which penalty function works best, partly because there are several reasonable criteria for “best”. Three examples for models of dimension d are

$$\begin{aligned}\mathrm{AIC}(d)&=d,\\ \mathrm{AICC}(d)&=\frac{nd}{n-d-1},\\ \mathrm{BIC}(d)&=\tfrac{1}{2}d\log n.\end{aligned}$$

The abbreviations are for *Akaike’s Information Criterion*, Akaike’s information corrected criterion, and *Bayesian Information Criterion* respectively.

It seems reasonable to choose a penalty function such that as $n\to\infty$ the value $\hat d_n$ that maximizes the penalized likelihood converges to the true value (in probability or almost surely). By the following theorem penalties such that $\phi_n(d)\to\infty$ faster than $\log\log n$ achieve this aim in the case of ARMA processes. Here an ARMA(p,q) process is understood to be exactly of orders p and q, i.e. the leading coefficients of the polynomials ϕ and θ of degrees p and q are nonzero.

12.16 Theorem. *Let X_t be a Gaussian causal, invertible stationary ARMA(p_0,q_0) process and let $(\hat\theta,\hat p,\hat q)$ maximize the penalized likelihood over $\cup_{p+q\le d_0}(\Theta_{p,q},p,q)$, where for each (p,q) the set $\Theta_{p,q}$ is a compact subset of $\mathbb{R}^{p+q+1}$ consisting of parameters of a causal, invertible stationary ARMA(p,q) process and $d_0\ge p_0+q_0$ is fixed. If $\phi_n(d)/n\to 0$ and $\liminf \phi_n(d)/\log\log n$ is sufficiently large for every $d\le d_0$, then $\hat p\to p_0$ and $\hat q\to q_0$ almost surely.*

Proof. See Azencott and Dacunha-Castelle (1984), Chapitre XIV. ■

The condition on the penalty is met by the BIC penalty, but not by Akaike's penalty function. It is observed in practice that the use of Akaike's criterion overestimates the order of the model. The AICC criterion, which puts slightly bigger penalty on big models, is an attempt to correct this.

However, choosing a model of the correct order is perhaps not the most relevant criterion for "good" estimation. A different criterion is the distance of the estimated model, specified by a pair of a dimension $\hat{d}$ and a corresponding parameter $\hat{\theta}$, to the true law of the observations. Depending on the distance used, an "incorrect" estimate $\hat{d}$ together with a good estimate $\hat{\theta}$ of that dimension may well yield a model that is closer than the estimated model of the correct (higher) dimension. This paradox arises because fitting a model of higher dimension requires the estimation of more parameters, which may result in poorer estimators of all parameters. (Cf. Section 10.1.1.) For ARMA processes the AIC criterion performs well in this respect.

The AIC criterion is based on the Kullback-Leibler distance.

12.17 EXERCISE. Repeatedly simulate a MA(1) process with $\theta = .8$ for $n = 50$ or $n = 100$.

(i) Compare the quality of the moment estimator and the maximum likleihood estimator.

(ii) Are the sampling distributions of the estimators approximately normal?

12.18 EXERCISE. Find best fitting AR and ARMA models for the Wölfer sunspot numbers (object `sunspots` in Splus).