Time series analysis

Jan Grandell

Preface

The course *Time series analysis* is based on the book [7] and replaces our previous course *Stationary stochastic processes* which was based on [6]. The books, and by that the courses, differ in many respects, the most obvious is that [7] is more applied that [6]. The word "applied" is partly a fine word for "elementary". A very important part of the course consists in looking at and thinking through the examples in [7] and to play around with the enclosed computer package ITSM or with MATLAB. Jan Enger has constructed a number of M-files in order to facilitate use of MATLAB within this course. Those who want to use MATLAB later in connection with time series can use the toolbox *System Identification* by Lennart Ljung, which contains an extensive library of stochastic models related to time series and control theory.

The main reason for the change in the courses is that half of our intermediate course *Probability theory* treats stationary processes from a theoretical point of view. A second reason is that a course in time series analysis is useful also for students more interested in applications than in the underlying theory. There are many references to [6] in [7] and the best recommendation to give a student interested in the subject also from a more theoretical point of view is to buy both books. However, due to the price of books, this recommendation might be unrealistic. A "cheaper" recommendation to those students is to read this lecture notes, where many parts from our previous course, i.e. in reality from [6], are included. These parts are given in the Appendix and in inserted paragraphs. The inserted paragraphs are written in this style.

I am most grateful for all kind of criticism, from serious mathematical mistakes to trivial misprints and language errors.

Jan Grandell

Contents

Lecture 1

1.1 Introduction

A *time series* is a set of observations x_t, each one being recorded at a specific time t.

Definition 1.1 *A* time series model *for the observed data $\{x_t\}$ is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization.*

In reality we can only observe the time series at a finite number of times, and in that case the underlying sequence of random variables $(X_1, X_2, \ldots, X_n)$ is just a an n-dimensional random variable (or random vector). Often, however, it is convenient to allow the number of observations to be infinite. In that case $\{X_t,\ t = 1, 2, \ldots\}$ is called a *stochastic process*. In order to specify its statistical properties we then need to consider all n-dimensional distributions

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] \quad \text{for all } n = 1, 2, \ldots,$$

cf. Section A.1 on page 111 for details.

Example 1.1 (A binary process) A very simple example of a stochastic process is the binary process $\{X_t,\ t = 1, 2, \ldots\}$ of independent random variables with

$$P(X_t = 1) = P(X_t = -1) = \frac{1}{2}.$$

In this case we have

$$P(X_1 = i_1,\ X_2 = i_2,\ \ldots, X_n = i_n) = 2^{-n}$$

where $i_k = 1$ or -1. In Example A.2 on page 113 it is shown that the binary process is "well-defined". □

Definition 1.2 (IID noise) *A process $\{X_t,\ t \in \mathbb{Z}\}$ is said to be an IID noise with mean 0 and variance σ^2, written*

$$\{X_t\} \sim \text{IID}(0, \sigma^2),$$

if the random variables X_t are independent and identically distributed with $EX_t = 0$ and $\text{Var}(X_t) = \sigma^2$.

The binary process is obviously an IID$(0, 1)$-noise.

In most situations to be considered in this course, we will not need the "full" specification of the underlying stochastic process. The methods will generally rely only on its means and covariances and – sometimes – on some more or less general assumptions.

Consider a stochastic process $\{X_t,\, t \in T\}$, where T is called the *index* or parameter set. Important examples of index sets are
$\mathbb{Z} = \{0,\, \pm 1,\, \pm 2, \dots\}$, $\{0,\, 1,\, 2, \dots\}$, $(-\infty,\, \infty)$ and $[0,\, \infty)$.

A stochastic process with $T \subset \mathbb{Z}$ is often called a *time series*.

Definition 1.3 *Let $\{X_t,\, t \in T\}$ be a stochastic process with* $\mathrm{Var}(X_t) < \infty$. *The* mean function *of $\{X_t\}$ is*

$$\mu_X(t) \stackrel{\text{def}}{=} E(X_t), \quad t \in T.$$

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

$$\gamma_X(r, s) = \mathrm{Cov}(X_r, X_s), \quad r, s \in T.$$

Example 1.2 (Standard Brownian motion) A standard Brownian motion, or a standard Wiener process, $\{B(t),\, t \geq 0\}$ is a stochastic process with $B(0) = 0$, independent increments, and $B(t) - B(s) \sim N(0, t - s)$ for $t \geq s$, see Definition A.5 on page 114 for details. The notation $N(0, t-s)$ means, contrary to the notation used in [1], that the *variance* is $t - s$. We have, for $r \leq s$

$$\gamma_B(r, s) = \mathrm{Cov}(B(r), B(s)) = \mathrm{Cov}(B(r), B(s) - B(r) + B(r))$$

$$= \mathrm{Cov}(B(r), B(s) - B(r)) + \mathrm{Cov}(B(r), B(r)) = 0 + r = r$$

and thus, if nothing is said about the relation between r and s

$$\gamma_B(r, s) = \min(r, s).$$

□

1.2 Stationarity

Loosely speaking, a stochastic process is stationary, if its statistical properties do not change with time. Since, as mentioned, we will generally rely only on properties defined by the means and covariances, we are led to the following definition.

Definition 1.4 *The time series $\{X_t,\, t \in \mathbb{Z}\}$ is said to be* (weakly) stationary *if*

(i) $\mathrm{Var}(X_t) < \infty$ *for all* $t \in \mathbb{Z}$,

(ii) $\mu_X(t) = \mu$ *for all* $t \in \mathbb{Z}$,

(iii) $\gamma_X(r, s) = \gamma_X(r + t, s + t)$ *for all* $r, s, t \in \mathbb{Z}$.

(iii) implies that $\gamma_X(r,s)$ is a function of $r-s$, and it is convenient to define

$$\gamma_X(h) \stackrel{\text{def}}{=} \gamma_X(h,0).$$

The value "h" is referred to as the "lag".

Definition 1.5 *Let $\{X_t,\ t\in\mathbb{Z}\}$ be a stationary time series. The* autocovariance function (ACVF) *of $\{X_t\}$ is*

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

The autocorrelation function (ACF) *is*

$$\rho_X(h) \stackrel{\text{def}}{=} \frac{\gamma_X(h)}{\gamma_X(0)}.$$

A simple example of a stationary process is the white noise, which may be looked a upon as the correspondence to the IID noise when only the means and the covariances are taken into account.

Definition 1.6 (White noise) *A process $\{X_t,\ t\in\mathbb{Z}\}$ is said to be a white noise with mean μ and variance σ^2, written*

$$\{X_t\} \sim \mathrm{WN}(\mu,\sigma^2),$$

if $EX_t = \mu$ *and* $\gamma(h) = \begin{cases}\sigma^2 & \text{if } h=0,\\ 0 & \text{if } h\neq 0.\end{cases}$

Warning: In some literature white noise means IID. □

1.3 Trends and Seasonal Components

Consider the "classical decomposition" model

$$X_t = m_t + s_t + Y_t,$$

where

m_t is a slowly changing function (the "trend component");

s_t is a function with known period d (the "seasonal component");

Y_t is a stationary time series.

Our aim is to estimate and extract the deterministic components m_t and s_t in hope that the residual component Y_t will turn out to be a stationary time series.

1.3.1 No Seasonal Component

Assume that

$$X_t = m_t + Y_t, \quad t = 1, \ldots, n$$

where, without loss of generality, $EY_t = 0$.

Method 1 (Least Squares estimation of m_t)

If we assume that $m_t = a_0 + a_1 t + a_2 t^2$ we choose $\widehat{a}_k$ to minimize

$$\sum_{t=1}^{n} (x_t - a_0 - a_1 t - a_2 t^2)^2.$$

Method 2 (Smoothing by means of a moving average)

Let q be a non-negative integer and consider

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}, \quad q+1 \le t \le n-q.$$

Then

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t+j} \approx m_t,$$

provided

q is so small that m_t is approximately linear over $[t-q, t+q]$

and

q is so large that $\frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t+j} \approx 0$.

For $t \le q$ and $t > n-q$ some modification is necessary, e.g.

$$\widehat{m}_t = \sum_{j=0}^{n-t} \alpha(1-\alpha)^j X_{t+j} \quad \text{for } t = 1, \ldots, q$$

and

$$\widehat{m}_t = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t-j} \quad \text{for } t = n-q+1, \ldots, n.$$

The two requirements on q may be difficult to fulfill in the same time. Let us therefore consider a linear filter

$$\widehat{m}_t = \sum a_j X_{t+j},$$

where $\sum a_j = 1$ and $a_j = a_{-j}$. Such a filter will allow a linear trend to pass without distortion since

$$\sum a_j (a + b(t+j)) = (a+bt) \sum a_j + b \sum a_j j = (a+bt) \cdot 1 + 0.$$

In the above example we have

$$a_j = \begin{cases} \frac{1}{2q+1} & \text{for } |j| \le q, \\ 0 & \text{for } |j| > q. \end{cases}$$

It is possible to choose the weights $\{a_j\}$ so that a larger class of trend functions pass without distortion. Such an example is the Spencer 15-point moving average where

$$[a_0, a_{\pm1}, \dots, a_{\pm7}] = \frac{1}{320}[74, 67, 46, 21, 3, -5, -6, -3] \text{ and } a_j = 0 \text{ for } |j| > 7.$$

Applied to $m_t = at^3 + bt^2 + ct + d$ we get

$$\widehat{m}_t = \sum a_j X_{t+j} = \sum a_j m_{t+j} + \sum a_j Y_{t+j}$$

$$\approx \sum a_j m_{t+j} = \text{see problem 1.12 in [7]} = m_t.$$

Method 3 (Differencing to generate stationarity)

Define the difference operator ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

where B is the backward *shift operator*, i.e. $(BX)_t = X_{t-1}$, and its powers

$$\nabla^k X_t = \nabla(\nabla^{k-1} X)_t.$$

As an example we get

$$\nabla^2 X_t = \nabla X_t - \nabla X_{t-1} =$$

$$(X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}.$$

As an illustration of "the calculus of operators" we give a different "proof":

$$\nabla^2 X_t = (1 - B)^2 X_t = (1 - 2B + B^2)X_t = X_t - 2X_{t-1} + X_{t-2}.$$

If $m_t = a + bt$ we get

$$\nabla X_t = \nabla m_t + \nabla Y_t = a + bt - a - b(t - 1) + \nabla Y_t = b + \nabla Y_t$$

and

$$\text{Cov}[\nabla Y_t, \nabla Y_s] =$$

$$\text{Cov}[Y_t, Y_s] - \text{Cov}[Y_{t-1}, Y_s] - \text{Cov}[Y_t, Y_{s-1}] + \text{Cov}[Y_{t-1}, Y_{s-1}]$$

$$= \gamma_Y(t - s) - \gamma_Y(t - s - 1) - \gamma_Y(t - s + 1) + \gamma_Y(t - s)$$

$$= 2\gamma_Y(t - s) - \gamma_Y(t - s + 1) - \gamma_Y(t - s - 1).$$

Thus ∇X_t is stationary.

Generally, if $m_t = \sum_{j=0}^{k} c_j t^j$ we get

$$\nabla^k X_t = k! c_k + \nabla^k Y_t,$$

which is stationary, cf. problem 1.10 in [7].

Thus is tempting to try to get stationarity by differencing. In practice often $k = 1$ or 2 is enough.

1.3.2 Trend and Seasonality

Let us go back to

$$X_t = m_t + s_t + Y_t,$$

where $EY_t = 0$, $s_{t+d} = s_t$ and $\sum_{k=1}^{d} s_k = 0$. For simplicity we assume that n/d is an integer.

Typical values of d are:

- 24 for period: day and time-unit: hours;
- 7 for period: week and time-unit: days;
- 12 for period: year and time-unit: months.

Sometimes it is convenient to index the data by period and time-unit

$$x_{j,k} = x_{k+d(j-1)}, \quad k = 1, \ldots, d, \ j = 1, \ldots, \frac{n}{d},$$

i.e. $x_{j,k}$ is the observation at the k:th time-unit of the j:th period.

Method S1 (Small trends)

It is natural to regard the trend as constant during each period, which means that we consider the model

$$X_{j,k} = m_j + s_k + Y_{j,k}.$$

Natural estimates are

$$\widehat{m}_j = \frac{1}{d}\sum_{k=1}^{d} x_{j,k} \quad \text{and} \quad \widehat{s}_k = \frac{d}{n}\sum_{j=1}^{n/d}(x_{j,k} - \widehat{m}_j).$$

Method S2 (Moving average estimation)

First we apply a moving average in order to eliminate the seasonal variation and to dampen the noise. For d even we use $q = d/2$ and the estimate

$$\widehat{m}_t = \frac{0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5x_{t+q}}{d}$$

and for d odd we use $q = (d-1)/2$ and the estimate

$$\widehat{m}_t = \frac{x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + x_{t+q}}{d}$$

provided $q + 1 \le t \le n - q$.

In order to estimate s_k we first form the "natural" estimates

$$w_k = \frac{1}{\text{number of summands}} \sum_{\frac{q-k}{d} < j \le \frac{n-q-k}{d}} (x_{k+jd} - \widehat{m}_{k+jd}).$$

(Note that it is "only" the end-effects that force us to this formally complicated estimate. What we really are doing is to take the average of those x_{k+jd}:s where the m_{k+jd}:s can be estimated.)

In order to achieve $\sum_{k=1}^{d} \widehat{s}_k = 0$ we form the estimates

$$\widehat{s}_k = w_k - \frac{1}{d}\sum_{i=1}^{d} w_i, \quad k = 1, \ldots, d.$$

Method S3 (Differencing at lag d)

Define the lag-d difference operator ∇_d by

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t.$$

Then

$$\nabla_d X_t = \nabla_d m_t + \nabla_d Y_t$$

which has no seasonal component, and methods 1 – 3 above can be applied.

Lecture 2

2.1 The autocovariance of a stationary time series

Recall from Definition 1.4 on page 2 that a time series $\{X_t,\, t \in \mathbb{Z}\}$ is (weakly) stationary if

(i) $\mathrm{Var}(X_t) < \infty \quad$ for all $t \in \mathbb{Z}$,

(ii) $\mu_X(t) = \mu \quad$ for all $t \in \mathbb{Z}$,

(iii) $\gamma_X(t,s) = \gamma(t-s) \quad$ for all $s,t \in \mathbb{Z}$,

where $\gamma(h)$ is the autocovariance function (ACVF). Notice that we have suppressed the dependence on X in the ACVF, which we will do when there is no risk for misunderstandings.

It is more or less obvious that

$$\gamma(0) \geq 0,$$
$$|\gamma(h)| \leq \gamma(0) \quad \text{for all } h \in \mathbb{Z},$$
$$\gamma(h) = \gamma(-h) \quad \text{for all } h \in \mathbb{Z}.$$

We shall now give a characterization of the autocovariance function.

Definition 2.1 *A function $\kappa : \mathbb{Z} \to \mathbb{R}$ is said to be* non-negative definite*, or* positive semi-definite*, if*

$$\sum_{i,j=1}^{n} a_i \kappa(t_i - t_j) a_j \geq 0$$

for all n and all vectors $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{t} \in \mathbb{Z}^n$.

Theorem 2.1 *A real-valued even function defined on $\mathbb{Z}$ is non-negative definite if and only if it is the autocovariance function of a stationary time series.*

Proof of the "if" part: Let $\gamma(\cdot)$ be the autocovariance function of a stationary time series X_t. It is also the autocovariance function of $Z_t = X_t - m_X$. Then

$$\Gamma_n = \begin{pmatrix} \gamma(t_1 - t_1) & \dots & \gamma(t_1 - t_n) \\ \vdots & & \\ \gamma(t_n - t_1) & \dots & \gamma(t_n - t_n) \end{pmatrix} \text{ is the covariance matrix of } \boldsymbol{Z_t} = \begin{pmatrix} Z_{t_1} \\ \vdots \\ Z_{t_n} \end{pmatrix}$$

and we have

$$0 \leq \mathrm{Var}(\boldsymbol{a}'\boldsymbol{Z}_{\boldsymbol{t}}) = \boldsymbol{a}'E(\boldsymbol{Z}_{\boldsymbol{t}}\boldsymbol{Z}_{\boldsymbol{t}}')\boldsymbol{a} = \boldsymbol{a}'\Gamma_n\boldsymbol{a} = \sum_{i,j=1}^{n} a_i\gamma(t_i - t_j)a_j.$$

□

Before we consider the "only if" if part, we need some supplementary results.

Definition 2.2 *A $n \times n$ matrix Σ is* non-negative definite *if*

$$\boldsymbol{b}'\Sigma\boldsymbol{b} \geq 0 \quad \textit{for all } \boldsymbol{b} \in \mathbb{R}^n.$$

Lemma 2.1 *A symmetric non-negative definite $n \times n$ matrix Σ has non-negative eigenvalues $\lambda_1, \ldots, \lambda_n$.*

Proof of Lemma 2.1: Let $\boldsymbol{e}_k$ be an eigenvector corresponding to the eigenvalue λ_k. Since Σ is symmetric $\boldsymbol{e}_k$ is both a left and a right eigenvector. Then we have

$$\Sigma\boldsymbol{e}_k = \lambda_k\boldsymbol{e}_k \Rightarrow 0 \leq \boldsymbol{e}_k'\Sigma\boldsymbol{e}_k = \lambda_k\boldsymbol{e}_k'\boldsymbol{e}_k \Rightarrow \lambda_k \geq 0.$$

□

Lemma 2.2 *A symmetric non-negative definite $n \times n$ matrix Σ has the representation $\Sigma = BB'$, where B is an $n \times n$ matrix.*

Proof: Let $\boldsymbol{e}_k$ and λ_k be the eigenvectors and the eigenvalues of Σ. It is well known that $\{\boldsymbol{e}_k\}$ can be chosen orthogonal, i.e. such that $\boldsymbol{e}_k'\boldsymbol{e}_j = 0$ for $k \neq j$. Put

$$\boldsymbol{p}_k = \frac{\boldsymbol{e}_k}{\sqrt{\boldsymbol{e}_k'\boldsymbol{e}_k}} \quad \text{and} \quad P = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$$

which implies that P is orthonormal, i.e. $P^{-1} = P'$. Then, for $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$

$$\Sigma P = P\Lambda \Rightarrow \Sigma PP' = P\Lambda P' \Rightarrow \Sigma = P\Lambda P'.$$

The matrices

$$B = P\Lambda^{1/2} \text{ or } B = P\Lambda^{1/2}P', \text{ where } \Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}),$$

work. The matrix $P\Lambda^{1/2}P'$ is the natural definition of $\Sigma^{1/2}$ since it is non-negative definite and symmetric.

□

Thus, for any $\boldsymbol{\mu}$ and any symmetric non-negative definite $n \times n$ matrix Σ there exists a unique normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

Proof of the "only if" part of Theorem 2.1: Let $\kappa : \mathbb{Z} \to \mathbb{R}$ be an even non-negative definite function. For each n and $\boldsymbol{t} \in \mathbb{Z}^n$

$$K_{\boldsymbol{t}} = \begin{pmatrix} \kappa(t_1 - t_1) & \ldots & \kappa(t_1 - t_n) \\ \vdots & & \\ \kappa(t_n - t_1) & \ldots & \kappa(t_n - t_n) \end{pmatrix}$$

is the covariance matrix for a normal variable with, say, mean $\boldsymbol{0}$. Thus

$$\phi_{\boldsymbol{t}}(\boldsymbol{u}) = \exp(-\tfrac{1}{2}\boldsymbol{u}'K_{\boldsymbol{t}}\boldsymbol{u})$$

and it is easy to check that

$$\lim_{u_i \to 0} \phi_{\boldsymbol{t}}(\boldsymbol{u}) = \phi_{\boldsymbol{t}(i)}(\boldsymbol{u}(i)) = \exp(-\tfrac{1}{2}\boldsymbol{u}(i)'K_{\boldsymbol{t}(i)}\boldsymbol{u}(i))$$

holds. Recall that $\boldsymbol{t}(i) = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)'$ and $\boldsymbol{u}(i) = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n)'$.

Thus the theorem follows from Kolmogorov's existence theorem. □

2.1.1 Strict stationarity

Definition 2.3 *The time series $\{X_t,\, t \in \mathbb{Z}\}$ is said to be* strictly stationary *if the distributions of*

$$(X_{t_1}, \ldots, X_{t_k}) \quad \text{and} \quad (X_{t_1+h}, \ldots, X_{t_k+h})$$

are the same for all k, and all $t_1, \ldots, t_k, h \in \mathbb{Z}$.

Let B be the backward shift operator, i.e. $(BX)_t = X_{t-1}$. In the obvious way we define powers of B; $(B^j X)_t = X_{t-j}$. Then strict stationarity means that $B^h X$ has the same distribution for all $h \in \mathbb{Z}$.

A strictly stationary time series $\{X_t,\, t \in \mathbb{Z}\}$ with $\mathrm{Var}(X_t) < \infty$ is stationary.

A stationary time series $\{X_t,\, t \in \mathbb{Z}\}$ does not need to be strictly stationary: $\{X_t\}$ is a sequence of independent variables and

$$X_t \sim \begin{cases} \mathrm{Exp}(1) & \text{if } t \text{ is odd,} \\ N(1,1) & \text{if } t \text{ is even.} \end{cases}$$

Thus $\{X_t\}$ is $\mathrm{WN}(1,1)$ *but not* $\mathrm{IID}(1,1)$.

Definition 2.4 (Gaussian time series) *The time series $\{X_t,\, t \in \mathbb{Z}\}$ is said to be a Gaussian time series if all finite-dimensional distributions are normal.*

A stationary Gaussian time series $\{X_t,\, t \in \mathbb{Z}\}$ is strictly stationary, since the normal distribution is determined by its mean and its covariance.

2.1.2 The spectral density

Suppose that $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$. The function f given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h), \quad -\pi \le \lambda \le \pi, \tag{2.1}$$

is well-defined. We have

$$\int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda = \int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{i(h-k)\lambda} \gamma(k)\, d\lambda$$

$$= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) \int_{-\pi}^{\pi} e^{i(h-k)\lambda}\, d\lambda = \gamma(h).$$

The function f defined by (2.1) on this page is called the *spectral density* of the time series $\{X_t,\, t \in \mathbb{Z}\}$. The spectral density will turn out to be a very useful notion, to which we will return. Here we only notice that

$$f(0) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h).$$

2.2 Time series models

The simplest time series model is certainly the white noise. A first generalization of the white noise is the moving average.

Definition 2.5 (The MA(q) process) *The process $\{X_t,\, t \in \mathbb{Z}\}$ is said to be a* moving average of order q *if*

$$X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2), \tag{2.2}$$

where $\theta_1, \ldots, \theta_q$ are constants.

We will now extend MA(q) processes to linear processes.

Definition 2.6 (Linear processes) *The process $\{X_t,\, t \in \mathbb{Z}\}$ is said to be a* linear process *if it has the representation*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2), \tag{2.3}$$

where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Warning: In some literature "a linear process" means that $\{Z_t,\, t \in \mathbb{Z}\}$ is a sequence of independent and identically distributed variables, and not only WN, cf. Definition 3.3 on page 20. □

Several questions come naturally in connection with the representation (2.3) on this page, like:

How to interpret an infinite sum of random variables?

Is X_t well-defined?

In order to answer these questions we need some facts about convergence of random variables, for which we refer to Section A.2 on page 115. Here we only notice that $X_n \xrightarrow{\text{m.s.}} X$ means that

$$E(X_n - X)^2 \to 0 \text{ as } n \to \infty.$$

" $\xrightarrow{\text{m.s.}}$ " stands for "mean-square convergence" and the notion requires that $X, X_1, X_2, \ldots$ have finite second moment. An important property of mean-square convergence is that Cauchy-sequences do converge. More precisely this means that if $X_1, X_2, \ldots$ have finite second moment and if

$$E(X_n - X_k)^2 \to 0 \text{ as } n, k \to \infty,$$

then there exists a random variable X with finite second moment such that $X_n \xrightarrow{\text{m.s.}} X$. Another way to express this is:

The space of square integrable random variables is complete *under mean-square convergence.*

The existence of the sum in (2.3) on this page is no problem, due to the following lemma.

Lemma 2.3 *If $\{X_t\}$ is any sequence of random variables such that $\sup_t E|X_t| < \infty$, and if $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then the series*

$$\sum_{j=-\infty}^{\infty} \psi_j X_{t-j},$$

converges absolutely with probability one. If in addition $\sup_t E|X_t|^2 < \infty$ the series converges in mean-square to the same limit.

Proof of mean-square convergence: Assume that $\sup_t E|X_t|^2 < \infty$. Then

$$E\left|\sum_{m<|j|<n} \psi_j X_{t-j}\right|^2 = \sum_{m<|j|<n} \sum_{m<|k|<n} \psi_j \overline{\psi_k} E(X_{t-j}\overline{X_{t-k}})$$

$$\leq \sup_t E|X_t|^2 \left(\sum_{m<|j|<n} |\psi_j|\right)^2 \to 0 \quad \text{as } m, n \to \infty.$$

Completeness does the rest! □

Theorem 2.2 *A linear process $\{X_t,\, t \in \mathbb{Z}\}$ with representation (2.3) on the preceding page is stationary with mean 0, autocovariance function*

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma^2, \tag{2.4}$$

and spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} |\psi(e^{-i\lambda})|^2, \tag{2.5}$$

where $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$.

Proof: We have

$$E(X_t) = E\left(\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}\right) = \sum_{j=-\infty}^{\infty} \psi_j E(Z_{t-j}) = 0$$

and

$$\gamma_X(h) = E(X_{t+h}X_t) = E\left[\left(\sum_{j=-\infty}^{\infty} \psi_j Z_{t+h-j}\right)\left(\sum_{k=-\infty}^{\infty} \psi_k Z_{t-k}\right)\right]$$

$$= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k E(Z_{t+h-j}Z_{t-k}) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k E(Z_{h+k-j}Z_0)$$

$$= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_{h+j} \psi_k E(Z_{k-j}Z_0) = \sum_{j=-\infty}^{\infty} \psi_{h+j} \psi_j E(Z_0 Z_0),$$

and (2.4) follows since $E(Z_0Z_0) = \sigma^2$.

Using (2.1) on page 11 we get

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma_h(h) = \frac{\sigma^2}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}$$

$$= \frac{\sigma^2}{2\pi} \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} e^{-ih\lambda} \psi_j \psi_{j+h} = \frac{\sigma^2}{2\pi} \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} e^{ij\lambda} \psi_j e^{-i(j+h)\lambda} \psi_{j+h}$$

$$= \frac{\sigma^2}{2\pi} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} e^{ij\lambda} \psi_j e^{-ik\lambda} \psi_k = \frac{\sigma^2}{2\pi} \psi(e^{i\lambda}) \psi(e^{-i\lambda}) = \frac{\sigma^2}{2\pi} |\psi(e^{-i\lambda})|^2.$$

□

Definition 2.7 (The ARMA(p, q) process) *The process $\{X_t, t \in \mathbb{Z}\}$ is said to be an* ARMA(p, q) process *if it is stationary and if*

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \tag{2.6}$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. We say that $\{X_t\}$ is an ARMA(p, q) *process with mean μ if $\{X_t - \mu\}$ is an* ARMA(p, q) *process.*

Equations (2.6) can be written as

$$\phi(B) X_t = \theta(B) Z_t, \quad t \in \mathbb{Z},$$

where

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p,$$
$$\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q,$$

and B – as usual – is the backward shift operator, i.e. $(B^j X)_t = X_{t-j}$. The polynomials $\phi(\cdot)$ and $\theta(\cdot)$ are called *generating polynomials*.

Warning: In some literature the generating polynomials are somewhat differently defined, which may imply that roots are differently placed in relation to the unit circle. □

If $p = 0$, i.e. $\phi(z) = 1$ we have a MA(q) process. An even more important special case is when $q = 0$, i.e. when $\theta(z) = 1$.

Definition 2.8 (The AR(p) process) *The process $\{X_t, t \in \mathbb{Z}\}$ is said to be an* AR(p) autoregressive process of order p *if it is stationary and if*

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2). \tag{2.7}$$

We say that $\{X_t\}$ is an AR(p) *process with mean μ if $\{X_t - \mu\}$ is an* AR(p) *process.*

Definition 2.9 *An* ARMA(p, q) *process defined by the equations*

$$\phi(B) X_t = \theta(B) Z_t \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

is said to be causal *if there exists constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}. \tag{2.8}$$

Another way to express the notion of causality is to require that

$$\mathrm{Cov}(X_t, Z_{t+j}) = 0 \quad \text{for} \quad j = 1, 2, \ldots \tag{2.9}$$

Causality is not a property of $\{X_t\}$ alone but rather of the relationship between $\{X_t\}$ and $\{Z_t\}$.

Theorem 2.3 *Let $\{X_t\}$ be an ARMA(p, q) for which $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeros. Then $\{X_t\}$ is causal if and only if $\phi(z) \neq 0$ for all $|z| \leq 1$. The coefficients $\{\psi_j\}$ in (2.8) are determined by the relation*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1.$$

Idea of proof: Assume $\phi(z) \neq 0$ if $|z| \leq 1$. Then there exits $\varepsilon > 0$ such that

$$\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \xi_j z^j = \xi(z), \quad |z| < 1 + \varepsilon.$$

Due to this we may apply the operator $\xi(B)$ to both sides of $\phi(B)X_t = \theta(B)Z_t$, and we obtain

$$X_t = \xi(B)\theta(B)Z_t.$$

□

If $\phi(B)X_t = \theta(B)Z_t$ and if $\phi(z) = 0$ for some z with $|z| = 1$ there exists no stationary solution.

Example 2.1 (AR(1) process) Let $\{X_t\}$ be an AR(1) process, i.e.

$$X_t = Z_t + \phi X_{t-1} \quad \text{or} \quad \phi(z) = 1 - \phi z. \tag{2.10}$$

Since $1 - \phi z = 0$ gives $z = 1/\phi$ it follows that X_t is causal if $|\phi| < 1$. In that case

$$X_t = Z_t + \phi X_{t-1} = Z_t + \phi(Z_{t-1} + \phi X_{t-2}) = Z_t + \phi Z_{t-1} + \phi^2 X_{t-2}$$

$$= \ \ldots \ = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \phi^3 Z_{t-3} + \ldots.$$

Using Theorem 2.2 on page 13 we get immediately

$$\gamma_X(h) = \sum_{j=0}^{\infty} \phi^{2j+|h|}\sigma^2 = \frac{\sigma^2\phi^{|h|}}{1 - \phi^2}. \tag{2.11}$$

We will show an alternative derivation of (2.11) which uses (2.9) on this page. We have

$$\gamma_X(0) = E(X_t^2) = E[(Z_t + \phi X_{t-1})^2] = \sigma^2 + \phi^2\gamma_X(0) \Rightarrow \gamma_X(0) = \frac{\sigma^2}{1 - \phi^2}.$$

and, for $h > 0$,

$$\gamma_X(h) = E(X_t X_{t+h}) = E[X_t(Z_{t+h} + \phi X_{t+h-1})] = \phi\gamma_X(h - 1).$$

Thus (2.11) follows.

If $|\phi| > 1$ we can rewrite (2.10) on the previous page:

$$\phi^{-1}X_t = \phi^{-1}Z_t + X_{t-1} \quad \text{or} \quad X_t = -\phi^{-1}Z_{t+1} + \phi^{-1}X_{t+1}.$$

Thus X_t has the representation

$$X_t = -\phi^{-1}Z_{t+1} - \phi^{-2}Z_{t+2} - \phi^{-3}Z_{t+3} - \dots.$$

If $|\phi| = 1$ there exists no stationary solution. □

Definition 2.10 *An* ARMA(p, q) *process defined by the equations*

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2),$$

is said to be invertible *if there exists constants* $\{\pi_j\}$ *such that* $\sum_{j=0}^{\infty} |\pi_j| < \infty$ *and*

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t \in \mathbb{Z}. \tag{2.12}$$

Theorem 2.4 *Let* $\{X_t\}$ *be an* ARMA(p, q) *for which* $\phi(\cdot)$ *and* $\theta(\cdot)$ *have no common zeros. Then* $\{X_t\}$ *is invertible if and only if* $\theta(z) \neq 0$ *for all* $|z| \leq 1$. *The coefficients* $\{\pi_j\}$ *in* (2.12) *are determined by the relation*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1.$$

Example 2.2 (MA(1) process) Let $\{X_t\}$ be a MA(1) process, i.e.

$$X_t = Z_t + \theta Z_{t-1} \quad \text{or} \quad \theta(z) = 1 + \theta z.$$

Since $1 + \theta z = 0$ gives $z = -1/\theta$ it follows that X_t is invertible if $|\theta| < 1$. In that case

$$\begin{aligned} Z_t &= X_t - \theta Z_{t-1} = X_t - \theta(X_{t-1} - \theta Z_{t-2}) \\ &= \dots = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \theta^3 X_{t-3} + \dots. \end{aligned}$$

By (2.4) on page 13, with $\psi_0 = 1$, $\psi_1 = \phi$ and $\psi_j = 0$ for $j \neq 0, 1$, we get

$$\gamma(h) = \begin{cases} (1 + \theta^2)\sigma^2 & \text{if } h = 0, \\ \theta\sigma^2 & \text{if } |h| = 1, \\ 0 & \text{if } |h| > 1. \end{cases} \tag{2.13}$$

□

Example 2.3 (ARMA(1, 1) process) Let $\{X_t\}$ be a ARMA$(1, 1)$ process, i.e.

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1} \quad \text{or} \quad \phi(B)X_t = \theta(B)Z_t,$$

where $\phi(z) = 1 - \phi z$ and $\theta(z) = 1 + \theta z$. Let $|\phi| < 1$ and $|\theta| < 1$ so that X_t is causal and invertible. Then we have $X_t = \psi(B)Z_t$, where

$$\psi(z) = \frac{\theta(z)}{\phi(z)} = \frac{1+\theta z}{1-\phi z} = \sum_{j=0}^{\infty}(1+\theta z)\phi^j z^j = 1 + \sum_{j=1}^{\infty}(\phi+\theta)\phi^{j-1}z^j.$$

By (2.4) on page 13 we get

$$\gamma(0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 = \sigma^2\left(1 + \sum_{j=1}^{\infty}(\phi+\theta)^2\phi^{2(j-1)}\right)$$

$$= \sigma^2\left(1 + (\phi+\theta)^2 \sum_{j=0}^{\infty}\phi^{2j}\right) = \sigma^2\left(1 + \frac{(\phi+\theta)^2}{1-\phi^2}\right)$$

and, for $h > 0$,

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j\psi_{j+h} = \sigma^2\left((\phi+\theta)\phi^{h-1} + \sum_{j=1}^{\infty}(\phi+\theta)^2\phi^{2(j-1)+h}\right)$$

$$= \sigma^2\phi^{h-1}\left(\phi + \theta + \sum_{j=0}^{\infty}(\phi+\theta)^2\phi^{2j}\right) = \sigma^2\phi^{h-1}\left(\phi + \theta + \frac{(\phi+\theta)^2}{1-\phi^2}\right).$$

Naturally we can use the ACVF above together with (2.1) on page 11 in order to find the spectral density. However, it is simpler to use (2.5) on page 13. Then we get

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2} = \frac{\sigma^2}{2\pi}\frac{|1+\theta\cdot e^{-i\lambda}|^2}{|1-\phi\cdot e^{-i\lambda}|^2}$$

$$= \frac{\sigma^2}{2\pi}\,\frac{1+\theta^2+2\theta\cos(\lambda)}{1+\phi^2-2\phi\cos(\lambda)}, \quad -\pi \le \lambda \le \pi. \tag{2.14}$$

□

Lecture 3

3.1 Estimation of the mean and the autocovariance

Let $\{X_t\}$ be a stationary time series with mean μ, autocovariance function $\gamma(\cdot)$, and – when it exists – spectral density $f(\cdot)$.

We will consider estimation of μ, $\gamma(\cdot)$ and $\rho(\cdot) = \gamma(\cdot)/\gamma(0)$ from observations of $X_1, X_2, \ldots, X_n$. However, we will consider this estimation in slightly more details than done in [7] and therefore we first give the following two definitions about *asymptotic normality*.

Definition 3.1 *Let $Y_1, Y_2, \ldots$ be a sequence of random variables.*
$Y_n \sim \mathrm{AN}(\mu_n, \sigma_n^2)$ means that

$$\lim_{n\to\infty} P\left(\frac{Y_n - \mu_n}{\sigma_n} \le x\right) = \Phi(x).$$

Definition 3.1 is, despite the notation "AN", exactly the same as used in the general course in connection with the central limit theorem.

Definition 3.2 *Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ be a sequence of random k-vectors.*
$\boldsymbol{Y}_n \sim \mathrm{AN}(\boldsymbol{\mu}_n, \Sigma_n)$ means that

(a) *$\Sigma_1, \Sigma_2, \ldots$ have no zero diagonal elements;*

(b) *$\boldsymbol{\lambda}'\boldsymbol{Y}_n \sim \mathrm{AN}(\boldsymbol{\lambda}'\boldsymbol{\mu}_n, \boldsymbol{\lambda}'\Sigma_n\boldsymbol{\lambda})$ for every $\boldsymbol{\lambda} \in \mathbb{R}^k$ such that $\boldsymbol{\lambda}'\Sigma_n\boldsymbol{\lambda} > 0$ for all sufficiently large n.*

3.1.1 Estimation of μ

Consider

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j\,,$$

which is a natural unbiased estimate of μ.

Theorem 3.1 *If $\{X_t\}$ is a stationary time series with mean μ and autocovariance function $\gamma(\cdot)$, then as $n \to \infty$,*

$$\mathrm{Var}(\overline{X}_n) = E(\overline{X}_n - \mu)^2 \to 0 \quad \text{if } \gamma(n) \to 0,$$

and

$$n\,\mathrm{Var}(\overline{X}_n) \to \sum_{h=-\infty}^{\infty} \gamma(h) = 2\pi f(0) \quad \textit{if} \quad \sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

Proof: We have

$$n\,\mathrm{Var}(\overline{X}_n) = \frac{1}{n}\sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) = \sum_{|h|<n} \left(1 - \frac{|h|}{n}\right)\gamma(h) \le \sum_{|h|<n} |\gamma(h)|.$$

If $\gamma(n) \to 0$ then $\frac{1}{n}\sum_{|h|<n}\left(1-\frac{|h|}{n}\right)\gamma(h) \to 0$ and thus $\mathrm{Var}(\overline{X}_n) \to 0$.

If $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$ the result follows by dominated convergence, i.e. since

$$\left|\max\left(1-\frac{|h|}{n},\ 0\right)\gamma(h)\right| \le |\gamma(h)|$$

we have

$$\lim_{n\to\infty} n\,\mathrm{Var}(\overline{X}_n) = \lim_{n\to\infty}\sum_{|h|<n}\left(1-\frac{|h|}{n}\right)\gamma(h)$$

$$= \lim_{n\to\infty}\sum_{h=-\infty}^{\infty}\max\left(1-\frac{|h|}{n},\ 0\right)\gamma(h)$$

$$= \sum_{h=-\infty}^{\infty}\lim_{n\to\infty}\max\left(1-\frac{|h|}{n},\ 0\right)\gamma(h) = \sum_{h=-\infty}^{\infty}\gamma(h).$$

Finally, the assumption $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$ implies $\sum_{h=-\infty}^{\infty}\gamma(h) = 2\pi f(0)$. □

Definition 3.3 (Strictly linear time series) *A stationary time series $\{X_t\}$ is called* strictly linear *if it has the representation*

$$X_t = \mu + \sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}, \qquad \{Z_t\} \sim \mathrm{IID}(0,\sigma^2).$$

Note: $\{Z_t,\ t\in\mathbb{Z}\}$ is a sequence of independent and identically distributed variables, and not only WN as was the case for a linear process. □

Theorem 3.2 *If $\{X_t\}$ is a strictly linear time series where $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty}\psi_j \neq 0$, then*

$$\overline{X}_n \sim \mathrm{AN}\left(\mu, \frac{v}{n}\right),$$

where $v = \sum_{h=-\infty}^{\infty}\gamma(h) = \sigma^2\left(\sum_{j=-\infty}^{\infty}\psi_j\right)^2$.

It is natural to try to find a better estimate of μ than $\overline{X}_n$. If we restrict ourselves to linear unbiased estimates $\overline{X}_n$ is, however, (almost) asymptotically effective. Strictly speaking, this means that if $\widehat{\mu}_n$ is the best linear unbiased estimate and if the spectral density is piecewise continuous, then

$$\lim_{n\to\infty} n\,\mathrm{Var}(\widehat{\mu}_n) = \lim_{n\to\infty} n\,\mathrm{Var}(\overline{X}_n).$$

3.1.2 Estimation of $\gamma(\cdot)$ and $\rho(\cdot)$

We will consider the estimates

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n), \quad 0 \le h \le n-1,$$

and

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)},$$

respectively.

The estimates are biased. The reason to use $\frac{1}{n}$, and not $\frac{1}{n-h}$ or $\frac{1}{n-h-1}$ in the definition of $\widehat{\gamma}(h)$, is that the matrix

$$\widehat{\Gamma}_n = \begin{pmatrix} \widehat{\gamma}(0) & \dots & \widehat{\gamma}(n-1) \\ \vdots & & \\ \widehat{\gamma}(n-1) & \dots & \widehat{\gamma}(0) \end{pmatrix}$$

is non-negative definite.

Theorem 3.3 *If $\{X_t\}$ is a strictly linear time series where $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$ and $EZ_t^4 = \eta\sigma^4 < \infty$, then*

$$\begin{pmatrix} \widehat{\gamma}(0) \\ \vdots \\ \widehat{\gamma}(h) \end{pmatrix} \sim \text{AN}\left(\begin{pmatrix} \gamma(0) \\ \vdots \\ \gamma(h) \end{pmatrix}, n^{-1}V \right),$$

where $V = (v_{ij})_{i,j=0,\dots,h}$ is the covariance matrix and

$$v_{ij} = (\eta - 3)\gamma(i)\gamma(j) + \sum_{k=-\infty}^{\infty}\{\gamma(k)\gamma(k-i+j) + \gamma(k+j)\gamma(k-i)\}.$$

Note: If $\{Z_t,\, t \in \mathbb{Z}\}$ is Gaussian, then $\eta = 3$. □

Somewhat surprisingly, $\rho(\cdot)$ has nicer properties, in the sense that η disappears.

Theorem 3.4 *If $\{X_t\}$ is a strictly linear time series where $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$ and $EZ_t^4 < \infty$, then*

$$\begin{pmatrix} \widehat{\rho}(1) \\ \vdots \\ \widehat{\rho}(h) \end{pmatrix} \sim \text{AN}\left(\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(h) \end{pmatrix}, n^{-1}W \right),$$

where $W = (w_{ij})_{i,j=1,\dots,h}$ is the covariance matrix and

$$w_{ij} = \sum_{k=-\infty}^{\infty}\{\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)\}. \tag{3.1}$$

The expression (3.1) is called *Bartlett's formula*. Simple algebra shows that

$$w_{ij} = \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}, \tag{3.2}$$

which is more convenient for computational purposes.

In the following theorem, the assumption $EZ_t^4 < \infty$ is relaxed at the expense of a slightly stronger assumption on the sequence $\{\psi_j\}$.

Theorem 3.5 *If $\{X_t\}$ is a strictly linear time series where $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $\sum_{j=-\infty}^{\infty} \psi_j^2 |j| < \infty$, then*

$$\begin{pmatrix} \widehat{\rho}(1) \\ \vdots \\ \widehat{\rho}(h) \end{pmatrix} \sim \mathrm{AN}\left(\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(h) \end{pmatrix}, n^{-1}W \right),$$

where W is given by the previous theorem.

Further, and this does not follow from the theorems,

$$\lim_{n\to\infty} \mathrm{Corr}(\widehat{\gamma}(i), \widehat{\gamma}(j)) = \frac{v_{ij}}{\sqrt{v_{ii}v_{jj}}} \quad \text{and} \quad \lim_{n\to\infty} \mathrm{Corr}(\widehat{\rho}(i), \widehat{\rho}(j)) = \frac{w_{ij}}{\sqrt{w_{ii}w_{jj}}}.$$

This implies that $\widehat{\gamma}(0), \ldots, \widehat{\gamma}(h)$ and $\widehat{\rho}(1), \ldots, \widehat{\rho}(h)$ may have a "smooth" appearance, which may give a false impression of the precision.

3.2 Prediction

Prediction, or *forecasting*, means that we want to get knowledge of the outcome of some random variable by means of an observation of some other random variables. A typical – and in this course the most important – situation is that we have a stationary time series $\{X_t, t \in \mathbb{Z}\}$ with known mean and ACVF, which we have observed for certain t-values, and that we want to say something about X_t for some future time t. More precisely, assume that we are at time n, and have observed $X_1, \ldots, X_n$, and want to predict X_{n+h} for some $h > 0$. For purely notational reasons it is sometimes better to say that we are standing at time 0, and to consider prediction of X_h in terms of the observations $X_{-n+1}, \ldots, X_0$. A very general formulation of the prediction problem is the following (too) general question:

How to find a function $\widehat{X}_h(\cdot)$ of $X_{-n+1}, \ldots, X_0$ which gives as good information as possible about X_h?

3.2.1 A short course in inference

Before continuing, we notice that this question is rather similar to usual estimation, where we have an unknown parameter θ and where $\widehat{\theta}$ is an estimate of θ. Let us recall some basic facts from the general course. The starting point was to consider observations $x_1, \ldots, x_n$ of (independent) random variables $X_1, \ldots, X_n$ with a (known) distribution depending on the unknown parameter θ. A point estimate (punktskattning) of θ is then the value $\widehat{\theta}(x_1, \ldots, x_n)$. In order to analyze the estimate we have to consider the estimator (stickprovsvariabeln) $\widehat{\theta}(X_1, \ldots, X_n)$. Some nice properties of an estimate, considered in the general course, are the following:

- An estimate $\widehat{\theta}$ of θ is *unbiased* (väntevärdesriktig) if $E(\widehat{\theta}(X_1, \ldots, X_n)) = \theta$ for all θ.
- An estimate $\widehat{\theta}$ of θ is *consistent* if $P(|\widehat{\theta}(X_1, \ldots, X_n) - \theta| > \varepsilon) \to 0$ for $n \to \infty$.
- If $\widehat{\theta}$ and θ^* are unbiased estimates of θ we say that $\widehat{\theta}$ is *more effective* than θ^* if $\mathrm{Var}(\widehat{\theta}(X_1, \ldots, X_n)) \leq \mathrm{Var}(\theta^*(X_1, \ldots, X_n))$ for all θ.

The next question is to find a good estimate is a specific case, which can be compared with the properties given above. Here we will only remind about the maximum likelihood estimate, and for simplicity we only consider the case of discrete variables:
Let $X_1, \ldots, X_n$ be discrete IID random variables with $p(k, \theta) = P(X_j = k)$ when θ is the true value. The *maximum likelihood estimate* $\widehat{\theta}$ is a value that maximizes the likelihood function

$$L(\theta) = \prod_{j=1}^{n} p(x_j, \theta).$$

As an example we consider the Poisson distribution, i.e. when X_j is $\mathrm{Po}(\theta)$-distributed. In that case we have

$$L(\theta) = \prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta}.$$

By the standard method, i.e. by considering

$$\ln L(\theta) = \sum_{j=1}^{n} \{x_j \ln(\theta) - \ln(x_j!) - \theta\}$$

and

$$\frac{\ln L(\theta)}{d\theta} = \sum_{j=1}^{n} \left\{ \frac{x_j}{\theta} - 1 \right\} = n \cdot \left\{ \frac{\bar{x}}{\theta} - 1 \right\},$$

we get $\widehat{\theta} = \bar{x}$. We have $E(\widehat{\theta}) = \theta$ and $\mathrm{Var}(\widehat{\theta}) = \theta/n$, and thus

$$E\big((\widehat{\theta} - \theta)^2\big) = \frac{\theta}{n}. \tag{3.3}$$

A drawback with this standard approach is the there is, at least from a mathematical point of view, a fundamental difference between random variables and an fixed but unknown parameter. In reality one may discuss if the parameter really is *completely unknown*, since often one has some knowledge, or at least some intuition, about what a reasonable parameter value might be. This knowledge may be somewhat similar as the knowledge about the outcome of a random variable when its distribution, or only its mean, is known.

A completely different approach to estimation of parameters is the so called *Bayesian estimation*. The idea is then to regard the "parameter" θ as the outcome of a random variable Θ with a known distribution F_Θ. That distribution is called the *prior distribution* and the idea is that F_Θ describes our subjective opinion or belief about the parameter θ. This approach is highly controversial, since it requires that a, maybe vague, subjective belief is formalized as a distribution. However, suppose that this can be done in some specific situation, thereby avoiding a further discussion about the philosophical questions connected with subjective probabilities and Bayesian statistics. It is natural to define the *Bayes estimate* of θ as the conditional mean of Θ, given the observations. The name Bayesian estimation comes from Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)},$$

since the main idea is the change the order of conditional probabilities.

Let us illustrate Bayesian estimation in the Poisson case and assume that the prior distribution of Θ is continuous with density f_Θ. The conditional distribution of $X_1, \ldots, X_n$ given $\Theta = \theta$ is given by

$$P(X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta) = \prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta}.$$

The distribution of $\Theta, X_1, \ldots, X_n$ is then given by

$$P(X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta) f_\Theta(\theta) = \left(\prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta} \right) f_\Theta(\theta)$$

and thus the distribution of $X_1, \ldots, X_n$ is given by

$$P(X_1 = x_1, \ldots, X_n = x_n) = \int_0^\infty \left(\prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta} \right) f_\Theta(\theta)\, d\theta.$$

Using Bayes' theorem, maybe together with some intuition, we realize that the conditional distribution of Θ given the observations, or the *posterior distribution*, is given by

$$f_{\Theta|X_1,\ldots,X_n}(\theta) = \frac{\left(\prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta} \right) f_\Theta(\theta)}{\int_0^\infty \left(\prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta} \right) f_\Theta(\theta)\, d\theta} = \frac{\theta^{n\bar{x}} e^{-n\theta} f_\Theta(\theta)}{\int_0^\infty \theta^{n\bar{x}} e^{-n\theta} f_\Theta(\theta)\, d\theta}.$$

Notice that $f_{\Theta|X_1,\ldots,X_n}(\theta)$ depends on the observations only via $n\bar{x} = x_1 + \ldots x_n$. The Bayes estimate $\widehat{\theta}_B$ is now given by

$$\widehat{\theta}_B = E(\Theta \mid X_1, \ldots, X_n) = \frac{\int_0^\infty \theta^{n\bar{x}+1} e^{-n\theta} f_\Theta(\theta)\, d\theta}{\int_0^\infty \theta^{n\bar{x}} e^{-n\theta} f_\Theta(\theta)\, d\theta}.$$

Let now the prior distribution be a $\Gamma(\gamma, \lambda)$-*distribution*. Then

$$f_\Theta(\theta) = \frac{\lambda^\gamma}{\Gamma(\gamma)} \theta^{\gamma-1} e^{-\lambda\theta}, \quad \theta \geq 0,$$

where γ and λ are positive parameters and $\Gamma(\gamma)$ is the Γ-function, defined by

$$\Gamma(\gamma) \stackrel{\text{def}}{=} \int_0^\infty x^{\gamma-1} e^{-x}\, dx, \quad \gamma > 0.$$

γ is called the shape parameter and λ the scale parameter. In this case we say that Θ is $\Gamma(\gamma, \lambda)$-distributed.

This distribution is generally not considered in the general course, but it is mentioned in [1]. We may notice that for $\gamma = 1$ the distribution reduces to the $\text{Exp}(1/\lambda)$-distribution. The importance of the Γ-distribution in this connection is merely for mathematical reasons.

The posterior distribution is now a $\Gamma(\gamma + n\bar{x}, n + \lambda)$-distribution, which implies that the Bayes estimate is

$$\widehat{\theta}_B = \frac{\gamma + n\bar{x}}{\lambda + n} = \frac{\lambda}{\lambda + n} E(\Theta) + \frac{n}{\lambda + n}\, \bar{x}. \tag{3.4}$$

It can further be shown that

$$E\big((\widehat{\theta}_B - \Theta)^2\big) = \frac{\gamma}{\lambda^2 + \lambda n} = \frac{E(\Theta)}{\lambda + n}\,. \tag{3.5}$$

3.2.2 Prediction of random variables

Consider any random variables $W_1, W_2, \ldots, W_n$ and Y with finite means and variances. Put $\mu_i = E(W_i), \quad \mu = E(Y)$,

$$\Gamma_n = \begin{pmatrix} \gamma_{1,1} & \ldots & \gamma_{1,n} \\ \vdots & & \\ \gamma_{n,1} & \ldots & \gamma_{n,n} \end{pmatrix} = \begin{pmatrix} \text{Cov}(W_n, W_n) & \ldots & \text{Cov}(W_n, W_1) \\ \vdots & & \\ \text{Cov}(W_1, W_n) & \ldots & \text{Cov}(W_1, W_1) \end{pmatrix}$$

and

$$\boldsymbol{\gamma}_n = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix} = \begin{pmatrix} \text{Cov}(W_n, Y) \\ \vdots \\ \text{Cov}(W_1, Y) \end{pmatrix}.$$

(The "backward" notation will turn out to be natural when we consider stationary time series.)

Our aim is now to find a predictor $\widehat{Y}$ of Y in terms of $W_1, W_2, \ldots, W_n$. Since we have only assumed the means and the covariances to be known, we have to restrict ourselves to *linear prediction*, i.e. to predictors of the form

$$\widehat{Y} = a + a_1 W_n + \ldots + a_n W_1 \tag{3.6}$$

or, which just is for convenience,

$$\widehat{Y} - \mu = a_0 + a_1(W_n - \mu_n) + \ldots + a_n(W_1 - \mu_1).$$

Put $S(a_0, \ldots, a_n) = E\big((Y - \widehat{Y})^2\big)$. The idea is now to choose $a_0 \ldots, a_n$ so that $S(a_0, \ldots, a_n)$ is minimized. We have

$$\begin{aligned} S(a_0, \ldots, a_n) &= E\big([(Y - \mu) - a_0 - a_1(W_n - \mu_n) - \ldots - a_n(W_1 - \mu_1)]^2\big) \\ &= a_0^2 + E\big([(Y - \mu) - a_1(W_n - \mu_n) - \ldots - a_n(W_1 - \mu_1)]^2\big). \end{aligned}$$

Thus we get $a_0 = 0$ and

$$\begin{aligned} \frac{\partial S}{\partial a_i} &= -2E\big((W_{n-i+1} - \mu_{n-i+1})[(Y - \mu) - a_1(W_n - \mu_n) - \ldots - a_n(W_1 - \mu_1)]\big) \\ &= -2\Big(\gamma_i - \sum_{j=1}^{n} a_j \gamma_{i,j}\Big), \quad i = 1, \ldots, n. \end{aligned}$$

From the properties of S it follows that the predictor must satisfy

$$\frac{\partial S}{\partial a_i} = 0 \quad \Leftrightarrow \quad \gamma_i = \sum_{j=1}^{n} a_j \gamma_{i,j}, \quad i = 1, \ldots, n. \tag{3.7}$$

Let

$$\boldsymbol{a}_n = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}.$$

Then (3.7) can be written on the form

$$\boldsymbol{\gamma}_n = \Gamma_n \boldsymbol{a}_n \quad \text{or, if } \Gamma_n \text{ is non-singular, } \ \boldsymbol{a}_n = \Gamma_n^{-1} \boldsymbol{\gamma}_n. \tag{3.8}$$

It remains to show the $\widehat{Y}$ is uniquely determined also when Γ_n is singular. Assume, in that case, that $\boldsymbol{a}_n^{(1)}$ and $\boldsymbol{a}_n^{(2)}$ both satisfy $\boldsymbol{\gamma}_n = \Gamma_n \boldsymbol{a}_n^{(i)}$ and let $\widehat{Y}^{(i)}$ be the corresponding predictors. Then

$$\mathrm{Var}(\widehat{Y}^{(1)} - \widehat{Y}^{(2)}) = (\boldsymbol{a}^{(1)} - \boldsymbol{a}^{(2)})' \Gamma_n (\boldsymbol{a}^{(1)} - \boldsymbol{a}^{(2)}) = (\boldsymbol{a}^{(1)} - \boldsymbol{a}^{(2)})' (\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n) = 0,$$

and hence $\widehat{Y}^{(1)} = \widehat{Y}^{(2)}$. Notice that we cannot draw the conclusion that $\boldsymbol{a}^{(1)} = \boldsymbol{a}^{(2)}$, which in fact need not to be true.

From the calculations above, we may draw two conclusions:

- *There is no restriction to assume all means to be* 0.

- *The predictor* $\widehat{Y}$ *of* Y *is determined by*

$$E[(\textit{Error}) \times (\textit{Predictor variable})] = 0,$$

or, more precisely, by

$$\mathrm{Cov}(\widehat{Y} - Y, W_i) = 0, \quad \text{for } i = 1, \ldots, n. \tag{3.9}$$

Assume now that $\mu, \mu_1, \ldots \mu_n = 0$ and consider the *mean-square prediction error*

$$v_n \stackrel{\text{def}}{=} E\big((\widehat{Y} - Y)^2\big) = \mathrm{Var}(\widehat{Y} - Y).$$

Assume that Γ_n is non-singular. Using (3.9) we get

$$\mathrm{Var}(Y) = \mathrm{Var}(Y - \widehat{Y} + \widehat{Y}) = \mathrm{Var}(Y - \widehat{Y}) + \mathrm{Var}(\widehat{Y})$$

or

$$\begin{aligned} v_n &= \mathrm{Var}(Y - \widehat{Y}) = \mathrm{Var}(Y) - \mathrm{Var}(\widehat{Y}) \\ &= \mathrm{Var}(Y) - \boldsymbol{a}_n' \Gamma_n \boldsymbol{a}_n = \mathrm{Var}(Y) - \boldsymbol{\gamma}_n' \Gamma_n^{-1} \boldsymbol{\gamma}_n. \end{aligned} \tag{3.10}$$

Lecture 4

4.1 Prediction

4.1.1 Prediction of random variables

Let us, as an example, consider Bayesian estimation of Θ in the Poisson distribution, as was discussed in section 3.2.1. However, for that we need a lemma, which has an independent interest.

Lemma 4.1 *Let Y and W be two random variables. We have*

$$\mathrm{Var}(W) = \mathrm{Var}(E(W \mid Y)) + E(\mathrm{Var}(W \mid Y)).$$

Proof: We will in the proof use $E(W) = E(E(W \mid Y))$ which is "the law of total expectation". We have

$$\begin{aligned}\mathrm{Var}(W) &= E\big((W - E(W))^2\big) = E\big((W - E(W \mid Y) + E(W \mid Y) - E(W))^2\big) \\ &= E\big((W - E(W \mid Y))^2\big) + E\big(E(W \mid Y) - E(W))^2\big) \\ &= E(\mathrm{Var}(W \mid Y)) + \mathrm{Var}(E(W \mid Y)),\end{aligned}$$

which is the lemma. □

Example 4.1 (Prediction of the mean in the Poisson distribution) Let Θ and $X_1, \ldots, X_n$ be as in section 3.2.1, i.e. $X_1, \ldots, X_n$ given $\Theta = \theta$ are independent and Poisson distributed with common mean θ. We will now consider linear prediction, or linear Bayesian estimation, of Θ. Let

$$Y = \Theta \quad \text{and} \quad W_i = X_i, \quad i = 1, \ldots, n.$$

By well-known properties of the Poisson distribution it follows that

$$E(W_i \mid Y) = \mathrm{Var}(W_i \mid Y) = Y.$$

By Lemma 4.1 we get

$$E(W_i) = E(Y) = \mu \quad \text{and} \quad \gamma_{i,i} = \mathrm{Var}(W_i) = \mathrm{Var}(Y) + \mu.$$

By a slight extension of the proof Lemma 4.1 it follows that

$$\gamma_{i,j} = \mathrm{Cov}(W_i, W_j) = \mathrm{Var}(Y), \quad i \neq j,$$

and that

$$\gamma_i = \mathrm{Cov}(W_i, Y) = \mathrm{Var}(Y).$$

From (3.8) we then get

$$\mathrm{Var}(Y) = a_i\mu + \mathrm{Var}(Y)\sum_{j=1}^{n} a_j$$

which implies that $a_i = a$ for all i and $\mathrm{Var}(Y) = a\mu + \mathrm{Var}(Y)na$, or

$$a = \frac{\mathrm{Var}(Y)}{\mu + n\,\mathrm{Var}(Y)}.$$

Thus we have

$$\begin{aligned}\widehat{Y} = \mu + a(n\overline{W} - n\mu) &= \frac{\mu^2 + \mu n\,\mathrm{Var}(Y) + \mathrm{Var}(Y)n\overline{W} - \mathrm{Var}(Y)n\mu}{\mu + n\,\mathrm{Var}(Y)} \\ &= \frac{\mu^2 + \mathrm{Var}(Y)n\overline{W}}{\mu + n\,\mathrm{Var}(Y)}. \end{aligned} \tag{4.1}$$

Using (3.10) we get

$$v_n = \mathrm{Var}(Y) - \boldsymbol{a}_n'\Gamma_n\boldsymbol{a}_n = \mathrm{Var}(Y) - a^2(n^2\,\mathrm{Var}(Y) + n\mu) = \frac{\mu\,\mathrm{Var}(Y)}{\mu + n\,\mathrm{Var}(Y)}$$

Assume now that Y is $\Gamma(\gamma,\, \lambda)$-distributed, which implies that

$$\mu = \frac{\gamma}{\lambda} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\gamma}{\lambda^2}.$$

Then we get

$$\widehat{Y} = \frac{\gamma + n\overline{W}}{\lambda + n} \quad \text{and} \quad v_n = \frac{\gamma}{\lambda^2 + \lambda n}. \tag{4.2}$$

We see that this agrees with the corresponding results for $\widehat{\theta}_B$ given in (3.4) and (3.5) on page 25. This is, however, very natural, since $\widehat{\theta}_B$ was the Bayes estimate. Since it is linear it is furthermore the linear Bayes estimate. Notice that we by that have proved (3.5). It can be shown that $\widehat{\theta}_B$ is linear if and only if the prior distribution of Θ is a Γ-distribution. □

Let us go back and consider the general problem to predict Y linearly in terms of $W_1, W_2, \ldots, W_n$. Since we now know that there is no restriction to let all means be 0, we do so. Thus we consider, cf. (3.6) on page 26, predictors of the form

$$\widehat{Y} = a_1W_n + \ldots + a_nW_1, \tag{4.3}$$

where $\widehat{Y}$ is determined by (3.9) on page 27.

Let us consider the sets or, better expressed, the spaces of random variables

$$\mathcal{M} = a_1W_n + \ldots + a_nW_1, \quad \boldsymbol{a}_n \in \mathbb{R}^n.$$

and

$$\mathcal{H} = bY + a_1 W_n + \ldots + a_n W_1, \quad b \in \mathbb{R},\ \boldsymbol{a}_n \in \mathbb{R}^n.$$

We may regard $\mathcal{H}$ as our "working space" and $\mathcal{M}$ as our "observed space". The predictor $\widehat{Y}$ may then be looked upon as the point in $\mathcal{M}$ being closest to Y, provided distances in $\mathcal{H}$ are measured in terms of variances. We can further introduce a "geometry" in $\mathcal{H}$ by regarding uncorrelated random variables as orthogonal. Let X and Z be random variables in $\mathcal{H}$ and assume that $X \perp Z$, which is just the geometric way of saying that $\mathrm{Cov}(X, Z) = 0$. Then

$$\mathrm{Var}(X + Z) = \mathrm{Var}(X) + \mathrm{Var}(Z),$$

which may be regarded as the Pythagorean theorem for random variables.

In geometrical terminology (3.9) states that $\widehat{Y}$ is determined by $Y - \widehat{Y} \perp \mathcal{M}$, and therefore (3.9) is often called *the Projection theorem*. This geometric interpretation is not just an illustrative way of looking upon prediction, but really the mathematically satisfying approach to prediction. In such an approach $\mathcal{H}$ and $\mathcal{M}$ are assumed to be Hilbert spaces, cf. Section A.2. The general version of the Projection theorem is formulated in Theorem A.2 on page 116. In applications there are often convenient to use the notation

$$\overline{\mathrm{sp}}\{W_1, \ldots, W_n\}, \quad P_{\overline{\mathrm{sp}}\{W_1,\ldots,W_n\}}Y \quad \text{and} \quad \|Y - \widehat{Y}\|^2$$

for the "observed" Hilbert space $\mathcal{M}$, $\widehat{Y}$, and $E(|Y - \widehat{Y}|^2) = \mathrm{Var}(Y - \widehat{Y})$ respectively. The Hilbert space $\overline{\mathrm{sp}}\{W_1, \ldots, W_n\}$ is called the *Hilbert space spanned by* $W_1, \ldots, W_n$, cf. Definition A.8 on page 117. When we consider a Hilbert space spanned by finitely many variables, as above, nothing is really gained by using the Hilbert space approach. The situation is quite different when we consider a Hilbert space spanned by infinitely many variables.

Before continuing this discussion, we consider Example 4.1 again. It follows from (4.2) that $\widehat{Y} \xrightarrow{\text{m.s.}} Y$ when $n \to \infty$. Thus it is natural to regard Y as the "predictor of itself" if infinitely many Ws are observed.

The Hilbert space spanned by a finite or infinite number of random variables may be looked upon as the set of all possible linear predictors in terms of that set. The advantage of the Hilbert space approach is, as indicated, a satisfying mathematical framework; the cost is a higher level of abstraction. In this course we will not rely on Hilbert spaces, but we will try to use notations which simplify given results to be interpreted more generally.

As an example of a Hilbert space formulation, we give the following theorem, which is just a "translation" of the conclusion

- *There is no restriction to assume all means to be* 0.

Theorem 4.1 *Let Y and $\{W_t,\, t \in \mathbb{Z}\}$ be random variables. Then*

$$P_{\overline{\mathrm{sp}}\{1,\,(W_t,\,t\in\mathbb{Z})\}}Y = E[Y] + P_{\overline{\mathrm{sp}}\{(W_t - E[W_t]),\ t\in\mathbb{Z}\}}(Y - E[Y]).$$

Proof: From the prediction equations, see Remark A.1 on page 117, it follows that $P_{\overline{\mathrm{sp}}\{1,\,(W_t,\,t\in\mathbb{Z})\}}Y$ is determined by the equation

$$\begin{aligned} E\big[(Y - P_{\overline{\mathrm{sp}}\{1,\,(W_t,\,t\in\mathbb{Z})\}}Y)\cdot 1\big] &= 0 \\ E\big[(Y - P_{\overline{\mathrm{sp}}\{1,\,(W_t,\,t\in\mathbb{Z})\}}Y)\cdot W_s\big] &= 0 \quad \text{for } s \in \mathbb{Z}. \end{aligned}$$

Thus it is enough to show that

$$E\big[(Y - E[Y] - P_{\overline{\text{sp}}\{(W_t - E[W_t]),\, t\in\mathbb{Z}\}}(Y - E[Y])) \cdot 1\big] = 0 \tag{4.4}$$

$$E\big[(Y - E[Y] - P_{\overline{\text{sp}}\{(W_t - E[W_t]),\, t\in\mathbb{Z}\}}(Y - E[Y])) \cdot (W_s - E[W_s])\big] = 0 \quad \text{for } s \in \mathbb{Z}. \tag{4.5}$$

(4.4) holds since all elements in $\overline{\text{sp}}\{(W_t - E[W_t]),\, t \in \mathbb{Z}\}$ have mean zero. (4.5) is the prediction equations which determine $P_{\overline{\text{sp}}\{(W_t - E[W_t]),\, t\in\mathbb{Z}\}}(Y - E[Y])$. Thus the desired result is proved. □

4.1.2 Prediction for stationary time series

Finitely many observations

Let $\{X_t\}$ be a stationary time series with mean 0 and autocovariance function $\gamma(\cdot)$ and consider $Y = X_{n+1}$. Then

$$\gamma_{i,j} = \gamma(i - j) \quad \text{and} \quad \gamma_i = \gamma(i),$$

so the notation is now quite natural. It can be shown that $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$ implies that Γ_n is non-singular. Thus we have the following theorem.

Theorem 4.2 *If $\{X_t\}$ is a zero-mean stationary time series such that $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, the best linear predictor $\widehat{X}_{n+1}$ of X_{n+1} in terms of $X_1, X_2, \ldots, X_n$ is*

$$\widehat{X}_{n+1} = \sum_{i=1}^{n} \phi_{n,i} X_{n+1-i}, \quad n = 1, 2, \ldots,$$

where

$$\boldsymbol{\phi}_n = \begin{pmatrix} \phi_{n,1} \\ \vdots \\ \phi_{n,n} \end{pmatrix} = \Gamma_n^{-1} \boldsymbol{\gamma}_n, \quad \boldsymbol{\gamma}_n = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(n) \end{pmatrix} \quad \textit{and}$$

$$\Gamma_n = \begin{pmatrix} \gamma(1-1) & \ldots & \gamma(1-n) \\ \vdots & & \\ \gamma(n-1) & \ldots & \gamma(n-n) \end{pmatrix}.$$

The mean-squared error is $v_n = \gamma(0) - \boldsymbol{\gamma}_n' \Gamma_n^{-1} \boldsymbol{\gamma}_n$.

If n is small, this works well. If n is very large, one may consider approximations. If n is of "moderate" size, computations of Γ_n^{-1} may be rather difficult.

We shall now consider recursive methods, i.e. at time $n - 1$ we know $X_1, X_2, \ldots, X_{n-1}$ and have computed $\widehat{X}_n$. When we then get information of X_n we want to compute $\widehat{X}_{n+1}$ based on $X_1, X_2, \ldots, X_n$. The "real" new information is $X_n - \widehat{X}_n$ rather than X_n.

Assume now that we have computed $\widehat{X}_n$, which really means that we know $\boldsymbol{\phi}_{n-1}$ and v_{n-1}. We have the following algorithm.

Theorem 4.3 (The Durbin–Levinson Algorithm) *If $\{X_t\}$ is a zero-mean stationary time series such that $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, then $\phi_{1,1} = \gamma(1)/\gamma(0)$, $v_0 = \gamma(0)$,*

$$\phi_{n,n} = \left[\gamma(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma(n-j)\right] v_{n-1}^{-1}$$

$$\begin{pmatrix} \phi_{n,1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{n,n} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}$$

and

$$v_n = v_{n-1}[1 - \phi_{n,n}^2].$$

Proof: Consider the two orthogonal subspaces $\mathcal{K}_1 = \overline{\text{sp}}\{X_2, \ldots, X_n\}$ and $\mathcal{K}_2 = \overline{\text{sp}}\{X_1 - P_{\mathcal{K}_1}X_1\}$. Then we have

$$\widehat{X}_{n+1} = P_{\mathcal{K}_1}X_{n+1} + P_{\mathcal{K}_2}X_{n+1} = P_{\mathcal{K}_1}X_{n+1} + a(X_1 - P_{\mathcal{K}_1}X_1). \tag{4.6}$$

From the orthogonality and from the projection theorem we get

$$\langle X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1\rangle = \langle \widehat{X}_{n+1}, X_1 - P_{\mathcal{K}_1}X_1\rangle = 0 + a\|X_1 - P_{\mathcal{K}_1}X_1\|^2. \tag{4.7}$$

From (4.6) and the fact that a time-reversed stationary time series has the same autocovariance we get

$$\phi_{n,j} = \begin{cases} \phi_{n-1,j} - \phi_{n,n}\phi_{n-1,n-j} & \text{if } j = 1, \ldots, n-1, \\ a & \text{if } j = n. \end{cases}$$

The explicit form of a follows from (4.7) when $X_1 - P_{\mathcal{K}_1}X_1$ is explicitly written and from $v_{n-1} = \|X_1 - P_{\mathcal{K}_1}X_1\|^2$. Further

$$v_n = \|X_{n+1} - P_{\mathcal{K}_1}X_{n+1} - P_{\mathcal{K}_2}X_{n+1}\|^2$$

$$= \|X_{n+1} - P_{\mathcal{K}_1}X_{n+1}\|^2 + \|P_{\mathcal{K}_2}X_{n+1}\|^2 - 2\langle X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, P_{\mathcal{K}_2}X_{n+1}\rangle$$

$$= \text{orthogonality} = v_{n-1} + a^2 v_{n-1} - 2a\langle X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1\rangle$$

$$= \text{see (4.7)} = v_{n-1} + a^2 v_{n-1} - 2a^2 v_{n-1} = v_{n-1}[1 - \phi_{n,n}^2].$$

□

Since the new information at time n is $X_n - \widehat{X}_n$ rather than X_n it might be natural to consider predictors which are linear combinations of the *innovations* $X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n$. Formally this is no difference, since

$$\overline{\text{sp}}\{X_1, X_2, \ldots, X_n\} = \overline{\text{sp}}\{X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n\}.$$

In this case we do not need to assume stationarity, and we have the following algorithm:

Theorem 4.4 (The Innovations Algorithm) *If $\{X_t\}$ has zero-mean and $E(X_iX_j) = \kappa(i,j)$, where the matrix $\begin{pmatrix} \kappa(1,1) & \dots & \kappa(1,n) \\ \vdots & & \\ \kappa(n,1) & \dots & \kappa(n,n) \end{pmatrix}$ is non-singular, we have*

$$\widehat{X}_{n+1} = \begin{cases} 0 & \text{if } n = 0, \\ \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j}) & \text{if } n \geq 1, \end{cases} \tag{4.8}$$

and

$$\begin{aligned} v_0 &= \kappa(1,1), \\ \theta_{n,n-k} &= v_k^{-1}\left(\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad k = 0,\dots,n-1, \\ v_n &= \kappa(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j. \end{aligned}$$

Proof: Taking the inner product of (4.8) with $X_{k+1} - \widehat{X}_{k+1}$ and using orthogonality we get

$$\langle \widehat{X}_{n+1}, X_{k+1} - \widehat{X}_{k+1} \rangle = \theta_{n,n-k} v_k,$$

and thus

$$\theta_{n,n-k} = v_k^{-1} \cdot \langle X_{n+1}, X_{k+1} - \widehat{X}_{k+1} \rangle.$$

From (4.8), with n replaced by k, we obtain

$$\begin{aligned} \theta_{n,n-k} &= v_k^{-1}\left(\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\langle X_{n+1}, X_{j+1} - \widehat{X}_{j+1}\rangle\right) \\ &= v_k^{-1}\left(\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right). \end{aligned}$$

The form of v_n follows immediately from $\|X_n - \widehat{X}_n\|^2 = \|X_n\|^2 - \|\widehat{X}_n\|^2$. □

Infinitely many observations

Assume now that we at time 0 have observed $X_{-n+1}, \dots, X_0$ and want to predict X_1 or, more generally X_h for $h \geq 1$. None of the Theorems 4.2 – 4.4 are quite satisfying if n if (very) large. In such situations it may be better to consider the predictor of X_h based of X_k for $k \leq 0$ or, more formally, the predictor

$$\widehat{X}_h = P_{\overline{\text{sp}}\{X_k,\, k\leq 0\}} X_h.$$

Let us now *assume* that $\widehat{X}_h$ can be expressed on the form

$$\widehat{X}_h = \sum_{j=0}^{\infty} \alpha_j X_{-j}. \tag{4.9}$$

From Example 4.1, or really from the discussion after the example, we know that this is not always the case.

However, under the assumption that (4.9) holds, it follows from (3.9) on page 27 that $\widehat{X}_h$ is determined by

$$\mathrm{Cov}(\widehat{X}_h, X_{-i}) = \mathrm{Cov}(X_h, X_{-i}), \quad i = 0, 1, \ldots,$$

or

$$\sum_{j=0}^{\infty} \gamma_X(i-j)\alpha_j = \gamma_X(h+i), \quad i = 0, 1, \ldots. \tag{4.10}$$

This set of equations determines $\widehat{X}_h$ *provided the resulting series converge.* Here we have assumed that such a solution exists.

Example 4.2 (MA(1) process) Let $\{X_t\}$ be a MA(1) process, i.e.

$$X_t = Z_t + \theta Z_{t-1} \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

Recall from (2.13) on page 16 that

$$\gamma_X(h) = \begin{cases} (1+\theta^2)\sigma^2 & \text{if } h = 0, \\ \theta\sigma^2 & \text{if } |h| = 1, \\ 0 & \text{if } |h| > 1. \end{cases}$$

We can now apply (4.10). If $h \geq 2$ we see that $\alpha_j \equiv 0$ satisfy (4.10) and $\widehat{X}_h = 0$ follows. This is quite natural, since X_h is uncorrelated with all observations.

Consider now $h = 1$. Then equations (4.10) reduce to

$$\alpha_0(1+\theta^2) + \alpha_1 = \theta \tag{4.11}$$

$$\alpha_{i-1}\theta + \alpha_i(1+\theta^2) + \alpha_{i+1}\theta = 0, \quad i = 1, 2, \ldots \tag{4.12}$$

If we notice that (4.12) can be written as

$$(\alpha_{i-1}\theta + \alpha_i) + \theta(\alpha_i\theta + \alpha_{i+1}) = 0, \quad i = 1, 2, \ldots$$

we get a potential solution $\alpha_i = \alpha_0(-\theta)^i$. Using (4.11) we get

$$\alpha_0(1+\theta^2) + \alpha_0(-\theta)\theta = \alpha_0 = \theta,$$

and thus

$$\alpha_i = -(-\theta)^{i+1} \quad \text{and} \quad \widehat{X}_1 = \sum_{j=0}^{\infty} -(-\theta)^{j+1} X_{-j}.$$

Assume now that $|\theta| < 1$. Then the sum converges and therefore we have found the right predictor. The reader may – and shall – be irritated of the above derivation of the predictor, since it contains "guessing" and a lot of "good luck" and has a smell of "after construction". This is also the case! In fact, we used that $\{X_t\}$ is invertible if $|\theta| < 1$, cf. Example 2.2 on page 16. We have $X_1 = Z_1 + \theta Z_0$ and thus

$$\widehat{X}_1 = P_{\overline{\mathrm{sp}}\{X_0, X_{-1}, \ldots\}} Z_1 + \theta P_{\overline{\mathrm{sp}}\{X_0, X_{-1}, \ldots\}} Z_0 = \theta P_{\overline{\mathrm{sp}}\{X_0, X_{-1}, \ldots\}} Z_0,$$

where the last equality follows since $Z_1 \perp \overline{\text{sp}}\{X_0, X_{-1}, \ldots\}$. Further it was shown in Example 2.2 that

$$Z_0 = X_0 - \theta X_{-1} + \theta^2 X_{-2} - \theta^3 X_{-3} + \ldots.$$

Thus $\widehat{X}_1 = \theta Z_0$.

Assume now that $|\theta| > 1$. Then we write (4.12) as

$$\theta(\alpha_{i-1} + \alpha_i \theta) + (\alpha_i + \alpha_{i+1}\theta) = 0, \quad i = 1, 2, \ldots$$

and get the potential solution $\alpha_i = \alpha_0(-\theta)^{-i}$. Using (4.11) we now get

$$\alpha_0(1 + \theta^2) + \alpha_0(-\theta)^{-1}\theta = \alpha_0\theta^2 = \theta,$$

and thus

$$\alpha_i = -(-\theta)^{-(i+1)} \quad \text{and} \quad \widehat{X}_1 = \sum_{j=0}^{\infty} -(-\theta)^{-(j+1)} X_{-j},$$

which converges.

Instead of "guessing" we may apply a general method of solving difference equations. As we did, we then consider first the homogeneous equations (4.12). The general method of solving (4.12) is related to the solution of differential equations and goes as follows:
Consider

$$x^2\theta + x(1 + \theta^2) + \theta = 0,$$

which has the solutions

$$x_1,\, x_2 = -\frac{1+\theta^2}{2\theta} \pm \sqrt{\left(\frac{1+\theta^2}{2\theta}\right)^2 - 1} = -\frac{1+\theta^2}{2\theta} \pm \frac{1-\theta^2}{2\theta}$$

or $x_1 = -\theta$ and $x_2 = -\theta^{-1}$. In the general case we have three possibilities:

- x_1 and x_2 are real and distinct. In our case this corresponds to $|\theta| \neq 1$. Then $\alpha_i = ax_1^i + bx_2^i$.

- x_1 and x_2 are real and equal, i.e. $x_1 = x_2 = x$. In our case this corresponds to $|\theta| = 1$. Then $\alpha_i = (a + bi)x^i$.

- $x_1 = \overline{x_2}$. This cannot happen in our case. (The general solution is then $\alpha_i = cx_1^i + \overline{c}\,\overline{x_1}^i$.)

If $|\theta| \neq 1$ it is easy to see that we get the given predictor by the requirement that the sum must converge.

If $|\theta| = 1$ it is easy to see that we cannot find a predictor of the assumed form. Nevertheless there exists a predictor! This is no contradiction; it "only" means that the predictor is not of the assumed form. □

From the last comments in the example a natural question arise:

When does there exists a predictor of the assumed form?

We cannot give an answer to that question, but the following theorem by Rozanov holds.

Theorem 4.5 *Let $\{X_t\}$ be a stationary time series with spectral density $f(\cdot)$. Any $Y \in \overline{\text{sp}}\{X_t,\, t \in \mathbb{Z}\}$ can be expressed on the form $Y = \sum_{-\infty}^{\infty} \psi_t X_t$ if and only if*

$$0 < c_1 \leq f(\lambda) \leq c_2 < \infty \quad \textit{for (almost) all } \lambda \in [-\pi, \pi].$$

For an MA(1) process we have, see (2.14) on page 17,

$$f(\lambda) = \frac{\sigma^2\,(1 + \theta^2 + 2\theta\cos(\lambda))}{2\pi}, \quad -\pi \leq \lambda \leq \pi.$$

Thus $f(0) = 0$ if $\theta = -1$ and $f(\pi) = f(-\pi) = 0$ if $\theta = 1$, while

$$f(\lambda) \geq \frac{\sigma^2\,(1 + \theta^2 - 2|\theta|)}{2\pi} = \frac{\sigma^2\,(1 - |\theta|)^2}{2\pi} > 0, \quad -\pi \leq \lambda \leq \pi$$

for all $|\theta| \neq 1$.

In general, predictors based on infinitely many observations are best expressed in terms of spectral properties of the underlying time series. We will return to this in section 6.2 on page 55. Here we will only give a preliminary version of Kolmogorov's formula for the *one-step mean-square prediction error*

$$v_\infty = E(\widehat{X}_{n+1} - X_{n+1})^2 \text{ where } \widehat{X}_{n+1} = P_{\overline{\text{sp}}\{X_t,\, t \leq n\}} X_{n+1}.$$

Theorem 4.6 (Kolmogorov's formula) *Let $\{X_t\}$ be a zero-mean stationary time series with spectral distribution density f. The one-step mean-square prediction error is*

$$v_\infty = 2\pi e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\lambda)\, d\lambda}.$$

Lecture 5

5.1 The Wold decomposition

Let, as usual, $\{X_t\}$ be a zero-mean stationary time series. Put

$$\mathcal{M}_n = \overline{\text{sp}}\{X_t,\, t \le n\}, \quad \sigma^2 = E|X_{n+1} - P_{\mathcal{M}_n} X_{n+1}|^2$$

and

$$\mathcal{M}_{-\infty} = \bigcap_{n=-\infty}^{\infty} \mathcal{M}_n.$$

The Hilbert space $\mathcal{M}_{-\infty}$ is called *"the infinite past"*.

Definition 5.1 *The process* $\{X_t\}$ *is called* deterministic *if* $\sigma^2 = 0$*, or equivalently if* $X_t \in \mathcal{M}_{-\infty}$ *for each* t*.*

Definition 5.2 *The process* $\{X_t\}$ *is called* purely non-deterministic *if*

$$\mathcal{M}_{-\infty} = \{0\}.$$

Theorem 5.1 (The Wold decomposition) *If* $\sigma^2 > 0$ *then* X_t *can be expressed as*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t,$$

where

(i) $\psi_0 = 1$ *and* $\sum_{j=0}^{\infty} \psi_j^2 < \infty$*;*

(ii) $\{Z_t\} \sim \text{WN}(0, \sigma^2)$*;*

(iii) $Z_t \in \mathcal{M}_t$ *for each* $t \in \mathbb{Z}$*;*

(iv) $E(Z_t V_s) = 0$ *for all* $s, t \in \mathbb{Z}$*;*

(v) $V_t \in \mathcal{M}_{-\infty}$ *for each* $t \in \mathbb{Z}$*;*

(vi) $\{V_t\}$ *is deterministic.*

(*Note that* (v) *and* (vi) *are not the same, since* $\mathcal{M}_{-\infty}$ *is defined in terms of* $\{X_t\}$*, not* $\{V_t\}$.)

The sequences $\{\psi_j\}$*,* $\{Z_t\}$ *and* $\{V_t\}$ *are uniquely determined by* $\{X_t\}$ *and conditions* (i) *to* (vi)*.*

We have chosen to give the Wold decomposition in this general, but somewhat abstract, form. A more "popular" version is to be found in [7].

5.2 Partial correlation

Let Y_1 and Y_2 be two random variables. The "relation" between them is often measured by the correlation coefficient

$$\rho(Y_1, Y_2) \stackrel{\text{def}}{=} \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\text{Var}(Y_1)\,\text{Var}(Y_2)}},$$

as is well-known from [1].

A value of $\rho(Y_1, Y_2)$ close to ± 1 is then taken as an indication of a relation between Y_1 and Y_2. Although often misinterpreted, the correlation coefficient only tells about the variation of Y_1 and Y_2 and nothing whether this variation is due to a "direct" influence between the variables. In many cases of so called *false correlation* there exist other variables, let us say $W_1, \ldots, W_k$, which explains most of the variation of Y_1 and Y_2. In a real situation the difficult problem is to find the variables which may explain the variation.

Consider $\widehat{Y}_1 = P_{\overline{\text{sp}}\{1, W_1, \ldots, W_k\}} Y_1$ and $\widehat{Y}_2 = P_{\overline{\text{sp}}\{1, W_1, \ldots, W_k\}} Y_2$.

Definition 5.3 *Let Y_1 and $W_1, \ldots, W_k$ be random variables. The* multiple correlation coefficient *between Y_1 and $W_1, \ldots, W_k$ is defined by $\rho(Y_1, \widehat{Y}_1)$.*

Definition 5.4 *Let Y_1, Y_2 and $W_1, \ldots, W_k$ be random variables. The* partial correlation coefficient *of Y_1 and Y_2 with respect to $W_1, \ldots, W_k$ is defined by*

$$\alpha(Y_1, Y_2) \stackrel{\text{def}}{=} \rho(Y_1 - \widehat{Y}_1, Y_2 - \widehat{Y}_2).$$

In the special case $k = 1$, i.e. when we try to explain the variation with only one variable W, we have

$$\alpha(Y_1, Y_2) = \frac{\rho(Y_1, Y_2) - \rho(Y_1, W)\rho(Y_2, W)}{\sqrt{(1 - \rho(Y_1, W)^2)(1 - \rho(Y_2, W)^2)}},$$

if $|\rho(Y_k, W)| < 1$.

Example 5.1 Let, as a simple – but rather natural – example

$$Y_1 = W + \widetilde{W}_1 \quad \text{and} \quad Y_2 = W + \widetilde{W}_2,$$

where $W, \widetilde{W}_1, \widetilde{W}_2$ are independent random variables with means 0. Then

$$\widehat{Y}_k = P_{\overline{\text{sp}}\{W\}} Y_k = P_{\overline{\text{sp}}\{W\}}(W + \widetilde{W}_k) = P_{\overline{\text{sp}}\{W\}} W + P_{\overline{\text{sp}}\{W\}} \widetilde{W}_k = W + 0 = W,$$

and thus $\alpha(Y_1, Y_2) = \rho(\widetilde{W}_1, \widetilde{W}_2) = 0$. □

Remark 5.1 The partial correlation coefficient is not always easy to interpret, which may be natural since it is a correlation coefficient. Let Y_1 and Y_2 be independent with, for simplicity, means 0 and the same variance and let $W = Y_1 + Y_2$. Then, cf. (3.8) on page 26, we get $\widehat{Y}_k = W/2 = (Y_1 + Y_2)/2$ and thus

$$Y_1 - \widehat{Y}_1 = \frac{Y_1 - Y_2}{2} = -(Y_2 - \widehat{Y}_2),$$

which implies that $\alpha(Y_1, Y_2) = -1$. Thus uncorrelated variables can be "completely partially correlated". □

5.2.1 Partial autocorrelation

Definition 5.5 *Let* $\{X_t,\, t \in \mathbb{Z}\}$ *be a zero-mean stationary time series. The* partial autocorrelation function (PACF) *of* $\{X_t\}$ *is defined by*

$$\alpha(0) = 1,$$

$$\alpha(1) = \rho(1),$$

$$\alpha(h) = \rho(X_{h+1} - P_{\overline{\text{sp}}\{X_2,\ldots,X_h\}}X_{h+1}, X_1 - P_{\overline{\text{sp}}\{X_2,\ldots,X_h\}}X_1), \quad h \geq 2.$$

The PACF $\alpha(h)$ may be regarded as the correlation between X_1 and X_{h+1}, or by stationarity as the correlation between X_t and X_{t+h}, adjusted for the intervening observations. Recall that $P_{\overline{\text{sp}}\{X_2,\ldots,X_h\}}X_{h+1}$, just is short notation for $\sum_{i=1}^{h-1}\phi_{h-1,i}X_{h+1-i}$. The reader is at this point recommended to have a look at Theorem 4.3 on page 33.

Definition 5.5 is not the definition given in [7], but in our opinion Definition 5.5 is more illustrative. The following theorem does, however, show that the two definitions are equivalent.

Theorem 5.2 *Under the assumptions of Theorem 4.3* $\alpha(h) = \phi_{h,h}$ *for* $h \geq 1$.

Proof: We will rely on the proof of Theorem 4.3, and for notational reasons we prove that $\alpha(n) = \phi_{n,n}$. Since $a = \phi_{n,n}$ we get from (4.7) that

$$\phi_{n,n} = \frac{\langle X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1\rangle}{\|X_1 - P_{\mathcal{K}_1}X_1\|^2}.$$

By the projection theorem we have $X_1 - P_{\mathcal{K}_1}X_1 \perp P_{\mathcal{K}_1}X_{n+1}$ and thus

$$\phi_{n,n} = \frac{\langle X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1\rangle}{\|X_1 - P_{\mathcal{K}_1}X_1\|^2} = \alpha(n).$$

□

5.3 ARMA processes

We will now continue the discussion about ARMA processes, which were introduced in Lecture 1.3.2. Recall from Definition 2.7 on page 14 that a zero-mean time series $\{X_t,\, t \in \mathbb{Z}\}$ is called an ARMA(p,q) process if it is stationary and if

$$\phi(B)X_t = \theta(B)Z_t, \quad t \in \mathbb{Z}, \text{ and } \{Z_t\} \sim \text{WN}(0,\sigma^2),$$

where

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p,$$

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

The polynomials $\phi(\cdot)$ and $\theta(\cdot)$ are called generating polynomials. More explicitly this means that

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{WN}(0,\sigma^2).$$

A stationary solution exists if and only if $\phi(z) \neq 0$ for all $|z| = 1$. If $p = 0$, i.e. $\phi(z) = 1$ we have a MA(q) process and if $q = 0$, i.e. $\theta(z) = 1$, we have an AR(p) process.

An ARMA(p, q) process is called causal, cf. Definition 2.9 on page 14 if there exists constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}.$$

It called invertible, cf. Definition 2.10 on page 16 if there exists constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t \in \mathbb{Z}.$$

Let $\{X_t\}$ be an ARMA(p, q) for which $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeros. Causality holds if and only if if and only if $\phi(z) \neq 0$ for all $|z| \leq 1$. The coefficients $\{\psi_j\}$ are determined by the relation

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1.$$

It is invertible if and only if $\theta(z) \neq 0$ for all $|z| \leq 1$. The coefficients $\{\pi_j\}$ are determined by the relation

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1.$$

5.3.1 Calculation of the ACVF

Let $\{X_t\}$ be a causal ARMA(p, q) process

First method

Using $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ it follows from Theorem 2.2 on page 13 that

$$\gamma_X(h) = \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

This method was used in Example 2.1 on page 15 and in Example 2.2 on page 16 for the AR(1) and MA(1) processes.

Second method

If we multiply each side of the equations

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q},$$

by X_{t-k} and take expectations we get

$$\gamma(k) - \phi_1\gamma(k-1) - \ldots - \phi_p\gamma(k-p) = \sigma^2 \sum_{j=0}^{\infty} \theta_{k+j}\psi_j, \tag{5.1}$$

for $k = 0, \ldots, m-1$, and

$$\gamma(k) - \phi_1\gamma(k-1) - \ldots - \phi_p\gamma(k-p) = 0 \quad k \geq m, \tag{5.2}$$

where $m = \max(p, q+1)$, $\theta_0 = 1$, and $\theta_j = 0$ for $j > q$. Sometimes these equations may be explicitly solved. In Example 4.2 on page 35 we discussed a little how difference equations can be solved. Let us here just mention that (5.2) has the general solution

$$\gamma(h) = \alpha_1\xi_1^{-h} + \alpha_2\xi_2^{-h} + \ldots + \alpha_p\xi_p^{-h}, \quad h \geq m-p$$

where $\xi_1, \ldots, \xi_p$ are the roots of the equation $\phi(z) = 0$ and $\alpha_1, \ldots, \alpha_p$ are arbitrary constants, *provided that the roots are distinct.* The constants $\alpha_1, \ldots, \alpha_p$ and the $m-p$ covariances $\gamma(h)$ for $h = 0, \ldots, m-p$ are determined by (5.1).

Example 5.2 (The Yule-Walker equations) Let $\{X_t\}$ be a causal AR(p) process. Them $m = p$ and (5.1) reduces to

$$\gamma(k) - \phi_1\gamma(k-1) - \ldots - \phi_p\gamma(k-p) = \begin{cases} 0, & k = 1, \ldots, p, \\ \sigma^2, & k = 0, \end{cases}$$

which are the *Yule-Walker equations.* For $p = 1$ we get the further reduction

$$\gamma(k) - \phi\gamma(k-1) = \begin{cases} 0, & k = 1, \ldots, \\ \sigma^2, & k = 0. \end{cases}$$

The solution ξ of $\phi(z) = 0$ is $\xi = 1/\phi$ and thus we get $\gamma(h) = \alpha \cdot \phi^h$. The constant α is determined by

$$\sigma^2 = \gamma(0) - \phi\gamma(-1) = \gamma(0) - \phi\gamma(-1) = \alpha - \phi\alpha\phi = \alpha(1-\phi^2),$$

and thus

$$\gamma_X(h) = \frac{\sigma^2\phi^{|h|}}{1-\phi^2},$$

which – of course – agrees with (2.11) on page 15.

□

Third method

This method may be regarded as a numerical version of the second method. The idea is to solve $\gamma(0), \ldots, \gamma(p)$ numerically from the $p+1$ first equations of (5.1) and (5.2) and to use the following equations to determine $\gamma(h)$ for $h = p+1, \ldots$

For an AR(1) process this means that we first consider the system

$$\begin{aligned}\gamma(0) - \phi\gamma(1) &= \sigma^2\\ \gamma(1) - \phi\gamma(0) &= 0\end{aligned}$$

which has the solution $\gamma(0) = \sigma^2/(1-\phi^2)$. Thus we get

$$\gamma(1) = \frac{\sigma^2\phi}{1-\phi^2}, \quad \gamma(2) = \frac{\sigma^2\phi^2}{1-\phi^2}, \quad \gamma(3) = \frac{\sigma^2\phi^3}{1-\phi^2}, \ \ldots$$

5.3.2 Prediction of an ARMA Process

The innovations algorithm can of course be applied directly to a causal ARMA process,

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

The calculations are, however, much simplified if we consider the transformed process

$$\begin{cases} W_t = \sigma^{-1}X_t, & \text{if } t = 1, \ldots, m,\\ W_t = \sigma^{-1}\phi(B)X_t, & \text{if } t > m,\end{cases}$$

where $m = \max(p, q)$.

Note that

$$\mathcal{H}_n = \overline{\mathrm{sp}}\{X_1, X_2, \ldots, X_n\} = \overline{\mathrm{sp}}\{W_1, W_2, \ldots, W_n\}$$

and put, as usual, $\widehat{X}_{n+1} = P_{\mathcal{H}_n}X_{n+1}$ and $\widehat{W}_{n+1} = P_{\mathcal{H}_n}W_{n+1}$. It easy to realize that

$$\begin{cases} \widehat{W}_t = \sigma^{-1}\widehat{X}_t, & \text{if } t = 1, \ldots, m,\\ \widehat{W}_t = \sigma^{-1}[\widehat{X}_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}], & \text{if } t > m,\end{cases}$$

or

$$X_t - \widehat{X}_t = \sigma[W_t - \widehat{W}_t] \quad \text{for all } t \geq 1.$$

The idea is now to apply the innovations algorithm to $\{W_t\}$. It can be shown that

$$\kappa_W(i,j) =$$

$$\begin{cases} \sigma^{-2}\gamma_X(i-j), & 1 \leq i, j \leq m,\\ \sigma^{-2}\left[\gamma_X(i-j) - \sum_{r=1}^{p}\phi_r\gamma_X(r-|i-j|)\right], & \min(i,j) \leq m < \max(i,j) \leq 2m\\ \sum_{r=0}^{q}\theta_r\theta_{r+|i-j|}, & \max(i,j) > m,\\ 0, & \text{otherwise,}\end{cases}$$

where we have adopted the convention $\theta_j = 0$ for $j > q$.

Thus, if we use the innovations algorithm to obtain θ_{nj}, we finally get

$$\widehat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j}) & \text{if } 1 \le n < m, \\ \phi_1 X_n + \cdots + \phi_p X_{n+1-p} & \\ \quad + \sum_{j=1}^{q} \theta_{n,j}(X_{n+1-j} - \widehat{X}_{n+1-j}) & \text{if } n \ge m. \end{cases}$$

Example 5.3 (AR(p) process) It follows immediately that

$$\widehat{X}_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}, \quad \text{if } n \ge p,$$

which is quite natural. It further follows from Theorem 5.2 on page 41 that the partial autocorrelation function $\alpha(h)$ is equal to 0 for $|h| > p$. □

Lecture 6

6.1 Spectral analysis

In Section 2.1.2 on page 11 we defined the *spectral density* $f(\cdot)$ of a stationary time series $\{X_t,\, t \in \mathbb{Z}\}$ by, see (2.1),

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma(h), \quad -\pi \leq \lambda \leq \pi$$

and showed that

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda. \tag{6.1}$$

This definition works fine if $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$.

A first question may be:

Which functions $f(\cdot)$ are spectral densities, i.e. for which functions does (6.1) *define an autocovariance function?*

The answer is simple, and we have the following theorem.

Theorem 6.1 *A real-valued function $f(\cdot)$ on $(-\pi, \pi]$ is the spectral density of a stationary time series if and only if*

(i) $f(\lambda) = f(-\lambda)$,

(ii) $f(\lambda) \geq 0$,

(iii) $\int_{-\pi}^{\pi} f(\lambda)\, d\lambda < \infty$.

A natural second question is:

Do all autocovariance functions have the representation (6.1)*?*

The answer is "No"!

Example 6.1 (A sinusoid process) Let A and B be two uncorrelated random variables with means 0 and variances σ^2 and consider the time series

$$X_t = A\cos(\omega t) + B\sin(\omega t), \quad t \in \mathbb{Z}, \tag{6.2}$$

where $-\pi < \omega < \pi$. Since obviously $E[X_t] = 0$ we get

$$\begin{aligned}
&\gamma_X(h) = E[X_{t+h}X_t] \\
&= EA^2 \cdot \cos(\omega(t+h))\cos(\omega t) + EB^2 \cdot \sin(\omega(t+h))\sin(\omega t) + E[AB] \cdot (\cdots) \\
&= \sigma^2 \cdot (\cos(\omega(t+h))\cos(\omega t) + \sin(\omega(t+h))\sin(\omega t)) + 0 \cdot (\cdots) \\
&= \sigma^2 \cos(\omega h) = \frac{\sigma^2}{2} e^{-ih\omega} + \frac{\sigma^2}{2} e^{ih\omega}. \qquad (6.3)
\end{aligned}$$

□

Formula (6.3) is a spectral representation, but it is not of the form (6.1). However, if we let

$$F(\lambda) = \begin{cases} 0, & \pi < \lambda < -\omega, \\ \sigma^2/2, & -\omega \le \lambda < \omega, \\ \sigma^2, & \omega \le \lambda \le \pi, \end{cases}$$

we can write (6.3) as

$$\gamma(h) = \int_{(-\pi,\pi]} e^{ih\lambda}\, dF(\lambda).$$

This is in fact the general form of the spectral representation. We will consider this representation in slightly more details than done in [7], since it is so important. It is, however, convenient to allow complex-valued time series, although the generalization in itself may be of limited interest.

6.1.1 The spectral distribution

Let X be a complex-valued stochastic variable. This means that $X = \mathrm{Re}\, X + i\mathrm{Im}\, X$ where $\mathrm{Re}\, X$ and $\mathrm{Im}\, X$ are stochastic variables. We define $E(X) = E(\mathrm{Re}\, X) + iE(\mathrm{Im}\, X)$ and $\mathrm{Var}(X) = E[(X - EX)(\overline{X - EX})]$.

X is called normally distributed if $(\mathrm{Re}\, X, \mathrm{Im}\, X)$ is normally distributed.

Definition 6.1 *The complex-valued time series $\{X_t,\, t \in \mathbb{Z}\}$ is said to be stationary if*

(i) $E|X_t|^2 < \infty$ *for all $t \in \mathbb{Z}$,*

(ii) EX_t *is independent of t for all $t \in \mathbb{Z}$,*

(iii) $E[X_{t+h}\overline{X_t}]$ *is independent of t for all $t \in \mathbb{Z}$.*

Definition 6.2 *The autocovariance function $\gamma(\cdot)$ of a complex-valued stationary time series $\{X_t\}$ is*

$$\gamma(h) = E[X_{t+h}\overline{X_t}] - EX_{t+h}E\overline{X_t}.$$

It is more or less obvious that

$$\begin{aligned} \gamma(0) &\geq 0, \\ |\gamma(h)| &\leq \gamma(0) \quad \text{for all } h \in \mathbb{Z}, \\ \gamma(\cdot) \text{ is Hermitian, i.e. } \gamma(h) &= \overline{\gamma(-h)} \quad \text{for all } h \in \mathbb{Z}. \end{aligned}$$

We shall now give a characterization of the autocovariance function, which is a natural extension of Theorem 2.1 on page 9, which holds in the real-valued case.

Theorem 6.2 *A function $K(\cdot)$ defined on $\mathbb{Z}$ is the autocovariance function of a (possibly complex-valued) stationary time series if and only if it is non-negative definite, i.e. if and only if*

$$\sum_{i,j=1}^{n} a_i K(i-j)\overline{a_j} \geq 0$$

for all n and all vectors $\boldsymbol{a} \in \mathbb{C}^n$.

Warning: In the complex case it does not hold that the normal distribution is determined by its mean and its covariance:

Let $X \sim N(0,1)$ be real-valued. Consider iX. Obviously $E[iX] = i \cdot 0 = 0$ and $\mathrm{Var}[iX] = E[iX\overline{iX}] = E[X^2] = 1$.

A stationary Gaussian time series $\{X_t,\, t \in \mathbb{Z}\}$ is not necessarily strictly stationary:

Let $\{X_t\}$ be a real-valued stationary Gaussian time series with mean 0. Consider $e^{it}X_t$ which is not strictly stationary. However,

$$E\left[e^{i(t+h)}X_{t+h}\overline{e^{it}X_t}\right] = e^{ih}E\left[X_{t+h}\overline{X_t}\right] = e^{ih}E[X_{t+h}X_t]\,,$$

which is independent of t! □

Theorem 6.3 (Herglotz's theorem) *A complex-valued function $\gamma(\cdot)$ defined on $\mathbb{Z}$ is non-negative definite if and only if*

$$\gamma(h) = \int_{(-\pi,\pi]} e^{ih\nu}\, dF(\nu) \quad \textit{for all } h \in \mathbb{Z},$$

where $F(\cdot)$ is a right-continuous, non-decreasing, bounded function on $[-\pi,\pi]$ and $F(-\pi) = 0$.

Idea of proof: If $\gamma(h) = \int_{(-\pi,\pi]} e^{ih\nu}\, dF(\nu)$ it is easily seen that γ is non-negative definite.

Assume that γ is non-negative definite. Then,

$$f_N(\nu) \stackrel{\text{def}}{=} \frac{1}{2\pi N}\sum_{r,s=1}^{N} e^{-ir\nu}\gamma(r-s)e^{is\nu}$$

$$= \frac{1}{2\pi N} \sum_{|m|<N} (N - |m|)\gamma(m)e^{-im\nu} \geq 0 \quad \text{for all } \nu \in [-\pi, \pi].$$

Put

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(\nu)\, d\nu.$$

Then

$$\int_{(-\pi,\pi]} e^{ih\nu}\, dF_N(\nu) = \frac{1}{2\pi} \sum_{|m|<N} \left(1 - \frac{|m|}{N}\right) \gamma(m) \int_{-\pi}^{\pi} e^{i(h-m)\nu}\, d\nu$$

$$= \begin{cases} \left(1 - \frac{|h|}{N}\right)\gamma(h), & |h| < N, \\ 0, & \text{otherwise.} \end{cases}$$

There exists a distribution function F and a subsequence $\{N_k\}$ such that

$$\int_{(-\pi,\pi]} g(\nu)\, dF_{N_k}(\nu) \to \int_{(-\pi,\pi]} g(\nu)\, dF(\nu) \quad \text{as } k \to \infty,$$

for all continuous and bounded functions g. Use, for each h, the function $g(\nu) = e^{ih\nu}$. □

Let $\{X_t\}$ be a complex-valued stationary time series with autocovariance function $\gamma_X(\cdot)$. If follows immediately from Theorems 6.2 and 6.3 that $\gamma_X(\cdot)$ has the spectral representation

$$\gamma_X(h) = \int_{(-\pi,\pi]} e^{ih\nu}\, dF_X(\nu).$$

The function F is called *the spectral distribution function* of γ. If $F(\lambda) = \int_{-\pi}^{\lambda} f(\nu)\, d\nu$, then f is the spectral density of γ.

If $\sum_{-\infty}^{\infty} |\gamma_X(h)| < \infty$ we have $F_X(\lambda) = \int_{-\pi}^{\lambda} f_X(\nu)\, d\nu$, and thus

$$\gamma_X(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f_X(\lambda)\, d\lambda.$$

In that case we have, as for real-valued time series,

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda}\gamma_X(n).$$

In the real-valued case the spectral distribution is symmetric, i.e. in general, $F_X(\lambda) = F_X(\pi^-) - F_X(-\lambda^-)$. This is easiest realized if we look upon $dF_X(\cdot)$ as defined on the unit circle with "$-\pi = \pi$".

In case of a density, $f_X(\lambda) = f_X(-\lambda)$, and thus

$$\gamma_X(h) = \int_0^{\pi} \left(e^{ih\lambda} + e^{-ih\lambda}\right) f_X(\lambda)\, d\lambda = 2\int_0^{\pi} \cos(h\lambda) f_X(\lambda)\, d\lambda.$$

6.1.2 Spectral representation of a time series

We begin by continuing Example 6.1 on page 47.

Example 6.2 (A sinusoid process) Consider the real-valued time series

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

where A and B are two uncorrelated random variables with means 0 and variances σ^2 and $-\pi < \omega < \pi$. We can write X_t on the form

$$X_t = \frac{A}{2}\cdot(e^{-it\omega} + e^{it\omega}) + \frac{iB}{2}\cdot(e^{-it\omega} - e^{it\omega})$$

$$= \frac{A+iB}{2}\,e^{-it\omega} + \frac{A-iB}{2}\,e^{it\omega} = Z_{-\omega}e^{-it\omega} + Z_\omega e^{it\omega},$$

where

$$Z_{-\omega} = \frac{A+iB}{2} \quad \text{and} \quad Z_\omega = \frac{A-iB}{2}.$$

The variables $Z_{-\omega}$ and Z_ω are obviously complex-valued and we have

$$E[Z_{-\omega}] = E[Z_\omega] = 0,$$

$$\mathrm{Var}[Z_{-\omega}] = \mathrm{Var}[Z_\omega] = E\left[\frac{A+iB}{2}\frac{A-iB}{2}\right] = E\left[\frac{A^2+B^2}{4}\right] = \frac{\sigma^2}{2}$$

and

$$\mathrm{Cov}[Z_{-\omega}, Z_\omega] = E[Z_{-\omega}\overline{Z_\omega}] = E\left[\frac{A+iB}{2}\frac{A+iB}{2}\right]$$

$$= E\left[\frac{A^2-B^2+2iAB}{4}\right] = E\left[\frac{A^2-B^2}{4}\right] = 0.$$

Thus $Z_{-\omega}$ and Z_ω are uncorrelated although they are "highly" dependent.

□

The time series itself has a spectral representation

$$X_t = \int_{(-\pi,\pi]} e^{it\nu}\,dZ(\nu) \tag{6.4}$$

where $\{Z(\lambda),\ \lambda \in [-\pi,\pi]\}$ is an orthogonal-increment process.

In order to discuss (6.4) in a mathematically satisfying way we must first

define orthogonal-increment processes;

define the integral with respect to an orthogonal-increment process.

To do this is, however, beyond the scope of this course. We may notice that the representation in Example 6.2 is of the form (6.4) if we let

$$Z(\lambda) = \begin{cases} 0, & \pi < \lambda < -\omega, \\ Z_{-\omega}, & -\omega \le \lambda < \omega, \\ Z_{-\omega} + Z_\omega, & \omega \le \lambda \le \pi. \end{cases}$$

In spite of the limitations of the course, we give a rather detailed discussion of the spectral representation, since it is so important.

Definition 6.3 (Orthogonal-increment process) *An orthogonal-increment process on* $[-\pi, \pi]$ *is a complex-valued process* $\{Z(\lambda)\}$ *such that*

$$\langle Z(\lambda), Z(\lambda)\rangle < \infty, \quad -\pi \leq \lambda \leq \pi, \tag{6.5}$$

$$\langle Z(\lambda), 1\rangle = 0, \quad -\pi \leq \lambda \leq \pi, \tag{6.6}$$

and

$$\langle Z(\lambda_4) - Z(\lambda_3), Z(\lambda_2) - Z(\lambda_1)\rangle = 0, \quad \text{if } (\lambda_1, \lambda_2] \cap (\lambda_3, \lambda_4] = \emptyset \tag{6.7}$$

where $\langle X, Y\rangle = EX\overline{Y}$.

The process $\{Z(\lambda)\}$ will be assumed to be *right-continuous*, i.e.

$$Z(\lambda + \delta) \xrightarrow{\text{m.s.}} Z(\lambda) \quad \text{as } \delta \downarrow 0.$$

Theorem 6.4 *If* $\{Z(\lambda)\}$ *is an orthogonal-increment process there exists a unique spectral distribution function* F *such that*

$$F(\mu) - F(\lambda) = \|Z(\mu) - Z(\lambda)\|^2, \quad -\pi \leq \lambda \leq \mu \leq \pi. \tag{6.8}$$

A practical shorthand notation for (6.7) and (6.7) is

$$E[dZ(\lambda)\overline{dZ(\mu)}] = \delta_{\lambda,\mu} dF(\lambda).$$

An integral

$$I(f) = \int_{(-\pi,\pi]} f(\nu)\, dZ(\nu) \tag{6.9}$$

is for a "kind" function, roughly speaking, defined in the "usual" way, with the difference that all convergence is interpreted in mean-square.

We will, however, need some more Hilbert space theory.

Definition 6.4 (Hilbert space isomorphisms) *An isomorphism of the Hilbert space* $\mathcal{H}_1$ *onto the Hilbert space* $\mathcal{H}_2$ *is a one to one mapping* T *of* $\mathcal{H}_1$ *onto* $\mathcal{H}_2$ *such that for all* $f_1,\ f_2 \in \mathcal{H}_1$,

$$T(af_1 + bf_2) = aTf_1 + bTf_2 \quad \text{for all scalars } a \text{ and } b$$

and

$$\langle Tf_1, Tf_2\rangle = \langle f_1, f_2\rangle.$$

We say that $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *are isomorphic if there is an isomorphism* T *of* $\mathcal{H}_1$ *onto* $\mathcal{H}_2$. *The inverse mapping* T^{-1} *is then an isomorphism of* $\mathcal{H}_2$ *onto* $\mathcal{H}_1$.

Useful properties are:

- $\|Tx\| = \|x\|$;
- $\|Tx_n - Tx\| \to 0$ if and only if $\|x_n - x\| \to 0$;
- $\{Tx_n\}$ is a Cauchy sequence if and only if $\{x_n\}$ is a Cauchy sequence;
- $TP_{\overline{\text{sp}}\{x_\lambda,\, \lambda\in\Lambda\}}x = P_{\overline{\text{sp}}\{Tx_\lambda,\, \lambda\in\Lambda\}}Tx$.

The idea is to have two isomorphic Hilbert spaces, and to do a desired operation in the one where it is simplest, and then to see what that means in the other one.

Let $\{Z(\lambda)\}$ be an orthogonal-increment process defined on a probability space $(\Omega,\, \mathcal{F},\, P)$ with spectral distribution function F. Consider the two Hilbert spaces $L^2(\Omega,\, \mathcal{F},\, P)$ of all square-integrable random variables defined on $(\Omega,\, \mathcal{F},\, P)$ and $L^2([-\pi, \pi],\, \mathcal{B},\, F) = L^2(F)$ of all functions f such that $\int_{(-\pi,\pi]} |f(\nu)|^2\, dF(\nu) < \infty$. The inner-product in $L^2(F)$ in defined by

$$\langle f, g\rangle = \int_{(-\pi,\pi]} f(\nu)\overline{g}(\nu)\, dF(\nu).$$

Let $\mathcal{D} \subseteq L^2(F)$ be the set of all functions f of the form

$$f(\lambda) = \sum_{i=0}^{n} f_i I_{(\lambda_i, \lambda_{i+1}]}(\lambda), \quad -\pi = \lambda_0 < \ldots < \lambda_{n+1} = \pi.$$

Define, for $f \in \mathcal{D}$ the mapping I by

$$I(f) = \sum_{i=0}^{n} f_i [Z(\lambda_{i+1}) - Z(\lambda_i)].$$

It is easy to realize that I is an isomorphism and it can be extended to an isomorphism of $\overline{\text{sp}}\{\mathcal{D}\}$ onto a subspace of $L^2(\Omega, \mathcal{F}, P)$. Furthermore $\mathcal{D}$ is dense in $L^2(F)$ and thus $\overline{\text{sp}}\{\mathcal{D}\} = L^2(F)$. Thus I is an isomorphism of $L^2(F)$ onto the subspace $I(L^2(F))$ of $L^2(\Omega, \mathcal{F}, P)$. This mapping is our definition of the integral (6.9). Because of the linearity of I the integral (6.9) has the "usual" properties of an integral.

Now we will consider (6.4). Let $\{X_t\}$ be a stationary time series defined on a probability space $(\Omega, \mathcal{F}, P)$ with spectral distribution function F. Consider the two sub-spaces

$$\overline{\mathcal{H}} = \overline{\text{sp}}\{X_t, t \in \mathbb{Z}\} \subset L^2(\Omega, \mathcal{F}, P) \quad \text{and} \quad \overline{\mathcal{K}} = \overline{\text{sp}}\{e^{it\cdot}, t \in \mathbb{Z}\} \subseteq L^2(F).$$

It is well known from Fourier analysis that $\overline{\mathcal{K}} = L^2(F)$.

Let $\mathcal{H}$ and $\mathcal{K}$ denote all finite linear combinations of $\{X_t\}$ and $\{e^{it\cdot}\}$ respectively. The mapping

$$T\left(\sum_{j=1}^{n} a_j X_{t_j}\right) = \sum_{j=1}^{n} a_j e^{it_j \cdot}$$

is an isomorphism between $\mathcal{H}$ and $\mathcal{K}$ since

$$\left\langle T\left(\sum_{j=1}^{n} a_j X_{t_j}\right), T\left(\sum_{k=1}^{m} b_k X_{s_k}\right)\right\rangle = \left\langle \sum_{j=1}^{n} a_j e^{it_j \cdot}, \sum_{k=1}^{m} b_k e^{is_k \cdot}\right\rangle_{L^2(F)}$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{m} a_j \bar{b}_k \left\langle e^{it_j \cdot}, e^{is_k \cdot}\right\rangle_{L^2(F)} = \sum_{j=1}^{n}\sum_{k=1}^{m} a_j \bar{b}_k \int_{(-\pi,\pi]} e^{i(t_j - s_k)\nu}\, dF(\nu)$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{m} a_j \bar{b}_k \left\langle X_{t_j}, X_{s_k}\right\rangle_{L^2(\Omega, \mathcal{F}, P)} = \left\langle \sum_{j=1}^{n} a_j X_{t_j}, \sum_{k=1}^{m} b_k X_{s_k}\right\rangle.$$

T can be extended to an isomorphism of $\overline{\mathcal{H}}$ onto $L^2(F)$.

Now we want to find functions $g_\lambda(\nu) \in L^2(F)$ such that $T^{-1} g_\lambda = Z(\lambda)$ where $\{Z(\lambda)\}$ is an orthogonal-increment process with distribution function F. Thus we want

$$\int_{(-\pi,\pi]} \Big(g_{\lambda_2}(\nu) - g_{\mu_2}(\nu)\Big)\Big(\,\overline{g_{\lambda_1}(\nu) - g_{\mu_1}(\nu)}\,\Big)\, dF(\nu)$$

$$= \begin{cases} 0, & \text{if } \mu_1 < \lambda_1 < \mu_2 < \lambda_2, \\ F(\lambda_1) - F(\mu_2), & \text{if } \mu_1 < \mu_2 < \lambda_1 < \lambda_2. \end{cases}$$

This is obtained if, for $\mu < \lambda$,

$$g_\lambda(\nu) - g_\mu(\nu) = I_{(\mu,\lambda]}(\nu) = I_{(-\pi,\lambda]}(\nu) - I_{(-\pi,\mu]}(\nu).$$

Therefore it is natural to *define*

$$Z(\lambda) = T^{-1} I_{(-\pi,\lambda]}$$

since, obviously $I_{(-\pi,\lambda]} \in L^2(F)$. For any $f \in \mathcal{D}$, i.e. for any f of the form

$$f(\lambda) = \sum_{i=0}^{n} f_i I_{(\lambda_i, \lambda_{i+1}]}(\lambda), \quad -\pi = \lambda_0 < \ldots < \lambda_{n+1} = \pi,$$

we have

$$I(f) = \sum_{i=0}^{n} f_i[Z(\lambda_{i+1}) - Z(\lambda_i)]$$

$$= \sum_{i=0}^{n} f_i T^{-1} I_{(\lambda_i, \lambda_{i+1}]} = T^{-1} f.$$

Since both I and T^{-1} can be extended to $L^2(F)$ we have

$$I = T^{-1} \quad \text{on } L^2(F).$$

Using this $\{Z(\lambda)\}$ in (6.4) we get

$$\left\| X_t - \int_{(-\pi,\pi]} e^{it\nu}\, dZ(\nu) \right\|^2 = \left\| X_t - I(e^{it\cdot}) \right\|^2$$

$$= \left\| TX_t - TI(e^{it\cdot}) \right\|^2_{L^2(F)} = \left\| e^{it\cdot} - e^{it\cdot} \right\|^2_{L^2(F)} = 0$$

and we have proved (6.4).

Remark 6.1 Any $Y \in \overline{\text{sp}}\{X_t,\, t \in \mathbb{Z}\}$ has the representation

$$\int_{(-\pi,\pi]} f(\nu)\, dZ(\nu) \quad \text{for some } f \in L^2(F).$$

This follows from

$$Y = IT(Y) = \int_{(-\pi,\pi]} TY(\nu)\, dZ(\nu) \quad \text{for } f = TY.$$

□

Remark 6.2 We have derived (6.4) only by using Hilbert spaces, i.e. by using "geometric" or covariance properties. Distributional properties follow from the fact that $Z(\lambda) \in \overline{\text{sp}}\{X_t,\, t \in \mathbb{Z}\}$. If, for example, $\{X_t\}$ is Gaussian, then – since linear combinations of (multivariate) normal random variables are normal – also $\{Z(\lambda)\}$ is a Gaussian process. □

Assume that F has a point of discontinuity at λ_0. Then

$$X_t = \int_{(-\pi,\pi]\setminus\{\lambda_0\}} e^{it\nu}\, dZ(\nu) + (Z(\lambda_0) - Z(\lambda_0^-))e^{it\lambda_0}$$

where $\int_{(-\pi,\pi]\setminus\{\lambda_0\}} e^{it\nu}\, dZ(\nu)$ and $(Z(\lambda_0) - Z(\lambda_0^-))e^{it\lambda_0}$ are uncorrelated and

$$\text{Var}[Z(\lambda_0) - Z(\lambda_0^-)] = F(\lambda_0) - F(\lambda_0^-).$$

The process $Y_t = (Z(\lambda_0) - Z(\lambda_0^-))e^{it\lambda_0}$ is said to be deterministic since Y_t is determined for all t if Y_{t_0} is known for some t_0. If X_t is real also $-\lambda_0$ is a point of discontinuity at F, and we have the deterministic component

$$Y_t = Z_1 e^{it\lambda_0} + Z_2 e^{-it\lambda_0}$$

where Z_1 and Z_2 are uncorrelated and $E[|Z_1|^2] = E[|Z_2|^2] = F(\lambda_0) - F(\lambda_0^-)$. Thus we have

$$Y_t = Z_1(\cos(t\lambda_0) + i\sin(t\lambda_0)) + Z_2(\cos(t\lambda_0) - i\sin(t\lambda_0))$$

$$= (\text{Re}\, Z_1 + i\, \text{Im}\, Z_1)(\cos(t\lambda_0) + i\sin(t\lambda_0))$$

$$+(\text{Re}\, Z_2 + i\, \text{Im}\, Z_2)(\cos(t\lambda_0) - i\sin(t\lambda_0))$$

$$= \text{Re}\, Z_1 \cos(t\lambda_0) - \text{Im}\, Z_1 \sin(t\lambda_0) + \text{Re}\, Z_2 \cos(t\lambda_0) + \text{Im}\, Z_2 \sin(t\lambda_0)$$

$$+ i \cdot \big[\mathrm{Im}\, Z_1 \cos(t\lambda_0) + \mathrm{Re}\, Z_1 \sin(t\lambda_0) + \mathrm{Im}\, Z_2 \cos(t\lambda_0) - \mathrm{Re}\, Z_2 \sin(t\lambda_0)\big].$$

Thus we must have

$$\mathrm{Re}\, Z_1 = \mathrm{Re}\, Z_2$$

$$\mathrm{Im}\, Z_1 = -\,\mathrm{Im}\, Z_2$$

which leads to

$$Y_t = 2\mathrm{Re}\, Z_2 \cos(t\lambda_0) - 2\mathrm{Im}\, Z_1 \sin(t\lambda_0).$$

Since $EY_t^2 = 4E[(\mathrm{Re}\, Z_2)^2]\cos^2(t\lambda_0) + 4E[(\mathrm{Im}\, Z_1)^2]\sin^2(t\lambda_0)$ it follows from stationarity that $E[(\mathrm{Re}\, Z_2)^2] = E[(\mathrm{Im}\, Z_1)^2]$. Put

$$\zeta_1 = -\sqrt{2}\cdot \mathrm{Im}\, Z_1 \quad \text{and} \quad \zeta_2 = \sqrt{2}\cdot \mathrm{Re}\, Z_2$$

and we get

$$Y_t = \sqrt{2}\zeta_2 \cos(t\lambda_0) + \sqrt{2}\zeta_1 \sin(t\lambda_0).$$

Note that $E\zeta_1^2 = 2E[(\mathrm{Im}\, Z_1)^2] = E[(\mathrm{Re}\, Z_1)^2] + E[(\mathrm{Im}\, Z_1)^2] = E[|Z_1|^2] = F(\lambda_0) - F(\lambda_0^-)$ and similar for ζ_2.

In general, if $\{X_t\}$ is real-valued, and if $F_X(0) - F_X(0^-) = 0$, we have the representation

$$X_t = \sqrt{2}\bigg(\int_{(-\pi,0]} \cos(t\nu)\, d\zeta(\nu) + \int_{(0,\pi]} \sin(t\nu)\, d\zeta(\nu)\bigg),$$

where $\{\zeta(\lambda)\}$ is a real-valued orthogonal-increment process with $Ed\zeta^2(\lambda) = dF(\lambda)$.

Compare this rather complicated representation with the spectral representation for $\gamma_X(\cdot)$, which in the real-valued case reduces to

$$\gamma_X(h) = 2\int_0^\pi \cos(h\lambda)\, dF_X(\lambda).$$

6.2 Prediction in the frequency domain

In section 4.1.2, starting at page 34, it was mentioned that prediction based on infinitely many observations is best treated in the framework of spectral properties of the underlying time series.

Let us first consider prediction based on finitely many observations. Let, as usual, $\{X_t\}$ be a zero-mean stationary time series and assume that we have observed $X_1, \ldots, X_n$ and want to predict X_{n+h}. Then we know that

$$\widehat{X}_{n+h} = P_{\overline{\mathrm{sp}}\{X_1,\ldots,X_n\}} X_{n+h} = \alpha_0 X_n + \ldots \alpha_{n-1} X_1$$

for some constants $\alpha_0, \ldots \alpha_{n-1}$. Using (6.4) on page 51 we can write

$$\widehat{X}_{n+h} = \int_{(-\pi,\pi]} g(\nu)\, dZ(\nu) \tag{6.10}$$

where $g(\nu) = \sum_{k=0}^{n-1} \alpha_k e^{i(n-k)\nu}$.

Assume now that we have infinitely many observations $\ldots, X_{n-1}, X_n$ to our disposal. In section 4.1.2, see (4.9) on page 34, we did assume that the predictor could be represented as an infinite sum which is not always the case. However, the predictor does always have a spectral representation of the form (6.10), cf. Remark 6.1 on the preceding page. The idea is now to determine

the function $g(\cdot)$. Although the derivation of $g(\cdot)$ is beyond the scope of this course, we will discuss it in some details due to its importance.

Let $\{X_t\}$ be a zero-mean stationary time series with spectral distribution function F and associated orthogonal-increment process $\{Z(\lambda)\}$. Recall from section 6.1.2that the mapping I defined by

$$I(g) = \int_{(-\pi,\pi]} g(\nu)\, dZ(\nu)$$

is an isomorphism of $L^2(F)$ onto the subspace $\overline{\mathcal{H}} = \overline{\text{sp}}\{X_t,\, t \in \mathbb{Z}\}$ such that

$$I(e^{it\cdot}) = X_t.$$

The idea is to compute projections, i.e. predictors, in $L^2(F)$ and then apply I. More precisely:

$$P_{\overline{\text{sp}}\{X_t,\ t\le n\}} X_{n+h} = I\Big(P_{\overline{\text{sp}}\{e^{it\cdot},\ t\le n\}} e^{i(n+h)\cdot}\Big).$$

We will illustrate this for an ARMA process, although it then follows from Theorem 4.5 on page 37 that the predictor has a representation as an infinite sum. Consider a causal invertible ARMA(p,q) process $\{X_t\}$

$$\phi(B)X_t = \theta(B)Z_t, \qquad \{Z_t\} \sim \text{WN}(0,\sigma^2)$$

with spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}\right|^2 = a(\lambda)\overline{a(\lambda)}.$$

Thus

$$a(\lambda) = \frac{\sigma}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\psi_k e^{-ik\lambda} \qquad \text{where} \quad \sum_{k=0}^{\infty}\psi_k z^k = \frac{\theta(z)}{\phi(z)}, \quad |z|\le 1.$$

From the Theorem A.3 on page 117 it follows that $g(\cdot) = P_{\overline{\text{sp}}\{e^{it\cdot},\ t\le n\}} e^{i(n+h)\cdot}$ must fulfill

$$\left\langle e^{i(n+h)\cdot} - g(\cdot), e^{im\cdot}\right\rangle_{L^2(F)} = \int_{-\pi}^{\pi}\Big(e^{i(n+h)\lambda} - g(\lambda)\Big)\,\overline{e^{im\lambda}} f_X(\lambda)\, d\lambda$$

$$= \int_{-\pi}^{\pi}\Big(e^{i(n+h)\lambda} - g(\lambda)\Big)\, e^{-im\lambda} a(\lambda)\overline{a(\lambda)}\, d\lambda = 0 \qquad \text{for } m \le n.$$

Thus

$$\Big(e^{i(n+h)\lambda} - g(\lambda)\Big)\, a(\lambda)\overline{a(\lambda)} \in \mathcal{M}_+ \stackrel{\text{def}}{=} \overline{\text{sp}}\{e^{im\cdot},\ m > n\} \subset L^2(d\lambda).$$

Since $\{X_t\}$ is invertible we have

$$\frac{1}{a(\cdot)} \in \overline{\text{sp}}\{e^{im\cdot},\ m \le 0\} \subset L^2(d\lambda)$$

and thus

$$\frac{1}{\overline{a(\cdot)}} \in \overline{\text{sp}}\{e^{im\cdot},\ m \ge 0\} \subset L^2(d\lambda).$$

Then it follows that

$$\Big(e^{i(n+h)\cdot} - g(\cdot)\Big)\, a(\cdot) = \Big(e^{i(n+h)\cdot} - g(\cdot)\Big)\, a(\cdot)\overline{a(\cdot)}\cdot\frac{1}{\overline{a(\cdot)}} \in \mathcal{M}_+.$$

Let us write

$$e^{i(n+h)\lambda}a(\lambda) = g(\lambda)a(\lambda) + \Big(e^{i(n+h)\lambda} - g(\lambda)\Big)\, a(\lambda).$$

Since $g(\cdot)a(\cdot) \in \overline{\text{sp}}\{e^{im\cdot},\ m \le n\}$, $\overline{\text{sp}}\{e^{im\cdot},\ m \le n\} \perp \mathcal{M}_+$ in $L^2(d\lambda)$ and an element, here $e^{i(n+h)\cdot}a(\cdot)$, has a unique decomposition in two orthogonal Hilbert spaces, we can make the identification

$$g(\lambda)a(\lambda) = \frac{\sigma}{\sqrt{2\pi}} e^{in\lambda} \sum_{k=0}^{\infty} \psi_{k+h} e^{-ik\lambda}.$$

Thus

$$g(\lambda) = e^{in\lambda} \frac{\sum_{k=0}^{\infty} \psi_{k+h} e^{-ik\lambda}}{\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}} \stackrel{\text{def}}{=} \sum_{j=0}^{\infty} \alpha_j e^{i(n-j)\lambda}.$$

Applying the mapping I we get

$$P_{\overline{\text{sp}}\{X_t,\ t \le n\}} X_{n+h} = \sum_{j=0}^{\infty} \alpha_j X_{n-j}.$$

Example 6.3 (AR(1) process) From Example 5.3 on page 45 we know that $\widehat{X}_{n+1} = \phi_1 X_n$ and it is not difficult to realize that $\widehat{X}_{n+h} = \phi_1^h X_n$. We will, however, see how this also follows from the derivation above. We have $\theta(z) = 1$, $\phi(z) = 1 - \phi_1 z$ and $\psi_k = \phi_1^k$, cf. Example 2.1 on page 15. Thus

$$g(\lambda) = e^{in\lambda} \frac{\sum_{k=0}^{\infty} \phi_1^{k+h} e^{-ik\lambda}}{\frac{1}{1-\phi_1 e^{-i\lambda}}} = \phi_1^h e^{in\lambda},$$

and the predictor follows. □

6.2.1 Interpolation and detection

Interpolation

Let $\{X_t,\, t \in \mathbb{Z}\}$ be a real stationary time series with mean 0 and spectral density f, where $f(\lambda) \ge A > 0$ for all $\lambda \in [-\pi, \pi]$. Assume that the entire time series has been observed except at the time point $t = 0$. The *best linear interpolator* $\widehat{X}_0$ of X_0 is defined by

$$\widehat{X}_0 = P_{\overline{\text{sp}}\{X_t,\, t \ne 0\}} X_0.$$

Let X_t have spectral representation $X_t = \int_{(-\pi,\pi]} e^{it\lambda}\, dZ(\lambda)$. Put

$$\mathcal{H}_0 = \overline{\text{sp}}\{e^{it\cdot}, t \ne 0\} \subset L^2(F).$$

Then

$$\widehat{X}_0 = \int_{(-\pi,\pi]} g(\lambda)\, dZ(\lambda),$$

where $g(\cdot) = P_{\mathcal{H}_0} 1$. By Theorem A.3 on page 117 this means that $g \in \mathcal{H}_0$ is the unique solution of

$$E[(X_0 - \widehat{X}_0)\overline{X_t}] = \int_{-\pi}^{\pi} (1 - g(\lambda)) e^{-it\lambda} f(\lambda)\, d\lambda = 0 \quad \text{for } t \ne 0.$$

Any solution of the projection equations must fulfill

$$(1 - g(\lambda))f(\lambda) = k \quad \text{or} \quad g(\lambda) = 1 - \frac{k}{f(\lambda)}.$$

(It is enough to realize that g above *is* a solution.) The problem is to determine k so that $g \in \mathcal{H}_0$. This means that we must have

$$0 = \int_{-\pi}^{\pi} g(\nu)\, d\nu = \int_{-\pi}^{\pi} 1 - \frac{k}{f(\nu)}\, d\nu = 2\pi - \int_{-\pi}^{\pi} \frac{k}{f(\nu)}\, d\nu,$$

from which we get

$$k = \frac{2\pi}{\int_{-\pi}^{\pi} \frac{d\nu}{f(\nu)}}$$

Thus

$$\widehat{X}_0 = \int_{(-\pi,\pi]} \left(1 - \frac{2\pi}{f(\lambda)\int_{-\pi}^{\pi} \frac{d\nu}{f(\nu)}}\right) dZ(\lambda).$$

Consider now the *mean square interpolation error* $E[(\widehat{X}_0 - X_0)^2]$. We get

$$E[(\widehat{X}_0 - X_0)^2] = \int_{-\pi}^{\pi} |1 - g(\lambda)|^2 f(\lambda)\, d\lambda = \int_{-\pi}^{\pi} \frac{|k|^2}{f(\lambda)^2} f(\lambda)\, d\lambda$$

$$= k^2 \cdot \frac{2\pi}{k} = 2\pi k = \frac{4\pi^2}{\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}}.$$

Example 6.4 (AR(1) process) Recall from Example 7.5 on page 71 that

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{1}{1-\phi_1 e^{-i\lambda}}\right|^2 = \frac{\sigma^2}{2\pi}\frac{1}{1 - 2\phi_1\cos\lambda + \phi_1^2}\,.$$

Since

$$\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)} = \frac{2\pi}{\sigma^2}\int_{-\pi}^{\pi} (1 - 2\phi_1\cos(\lambda) + \phi_1^2)\, d\lambda = \frac{4\pi^2}{\sigma^2}(1+\phi_1^2)$$

we get

$$\widehat{X}_0 = \int_{(-\pi,\pi]} \left(1 - \frac{1}{2\pi(1+\phi_1^2)f(\lambda)}\right) dZ(\lambda) = \int_{(-\pi,\pi]} \left(1 - \frac{1-\phi_1(e^{-i\lambda}+e^{i\lambda})}{1+\phi_1^2}\right) dZ(\lambda)$$

$$= \int_{(-\pi,\pi]} (\phi_1 e^{-i\lambda} + \phi_1 e^{i\lambda})\, dZ(\lambda) = \phi_1 X_{-1} + \phi_1 X_1$$

and

$$E[(\widehat{X}_0 - X_0)^2] = \frac{\sigma^2}{1+\phi_1^2}\,.$$

□

Detection

Let the stationary time series $\{X_t,\, t \in \mathbb{Z}\}$ be a disturbed signal, i.e. it is the sum of a signal $\{S_t,\, t \in \mathbb{Z}\}$ and a noise $\{N_t,\, t \in \mathbb{Z}\}$, where the signal and the noise are independent stationary time series with means 0 and spectral densities f_S and f_N respectively. (Note that the noise is not assumed to be white noise.) Assume that the entire time series $X_t = S_t + N_t$ has been observed. The *best linear detector* $\widehat{S}_0$ of S_0 is defined by

$$\widehat{S}_0 = P_{\overline{\text{sp}}\{X_t,\, t\in\mathbb{Z}\}} S_0,$$

where $\overline{\text{sp}}\{X_t,\, t \in \mathbb{Z}\}$ is a Hilbert sub-space of the Hilbert space $\overline{\text{sp}}\{S_t,\ N_t,\, t \in \mathbb{Z}\}$. In Example 7.4 on page 68 we discussed a much simpler situation.

It follows from the Theorem A.3 on page 117 that $\widehat{S}_0$ is the unique solution of

$$E[(S_0 - \widehat{S}_0)\overline{X_t}] = 0 \quad \text{for all } t.$$

Let S_t and N_t have spectral representations

$$S_t = \int_{(-\pi,\pi]} e^{it\lambda}\, dZ_S(\lambda) \quad \text{and} \quad N_t = \int_{(-\pi,\pi]} e^{it\lambda}\, dZ_N(\lambda)$$

respectively. Then X_t has spectral representation

$$X_t = \int_{(-\pi,\pi]} e^{it\lambda}\,(dZ_S(\lambda) + dZ_N(\lambda)),$$

where Z_S and Z_N are independent. Thus

$$\widehat{S}_0 = \int_{(-\pi,\pi]} g(\lambda)\,(dZ_S(\lambda) + dZ_N(\lambda)),$$

for some function $g \in L^2(F_S + F_N)$.

Now we have

$$0 = E[(S_0 - \widehat{S}_0)\overline{X_t}]$$

$$= E\left[\left(\int_{(-\pi,\pi]} dZ_S(\lambda) - \int_{(-\pi,\pi]} g(\lambda)\,(dZ_S(\lambda) + dZ_N(\lambda))\right)\int_{(-\pi,\pi]} e^{-it\lambda}\,\overline{(dZ_S(\lambda) + dZ_N(\lambda))}\right]$$

$$= \int_{(-\pi,\pi]} e^{-it\lambda} f_S(\lambda)\,d\lambda - \int_{(-\pi,\pi]} e^{-it\lambda} g(\lambda)(f_S(\lambda) + f_N(\lambda))\,d\lambda$$

$$= \int_{(-\pi,\pi]} e^{-it\lambda}\,(f_S(\lambda) - g(\lambda)(f_S(\lambda) + f_N(\lambda)))\;d\lambda.$$

Thus

$$f_S(\lambda) - g(\lambda)(f_S(\lambda) + f_N(\lambda)) = 0 \quad \text{or} \quad g(\lambda) = \frac{f_S(\lambda)}{f_S(\lambda) + f_N(\lambda)}\,.$$

From this we get the best linear detector

$$\widehat{S}_0 = \int_{(-\pi,\pi]} \frac{f_S(\lambda)}{f_S(\lambda) + f_N(\lambda)}\,(dZ_S(\lambda) + dZ_N(\lambda)),$$

and

$$E[(S_0 - \widehat{S}_0)^2] = E[S_0^2] - E[\widehat{S}_0^2]$$

$$= \int_{(-\pi,\pi]} f_S(\lambda)\,d\lambda - \int_{(-\pi,\pi]} |g(\lambda)|^2\,(f_S(\lambda) + f_N(\lambda))\;d\lambda$$

$$= \int_{(-\pi,\pi]} f_S(\lambda)\,d\lambda - \int_{(-\pi,\pi]} \left|\frac{f_S(\lambda)}{f_S(\lambda) + f_N(\lambda)}\right|^2 (f_S(\lambda) + f_N(\lambda))\,d\lambda$$

$$= \int_{(-\pi,\pi]} \left(f_S(\lambda) - \frac{f_S^2(\lambda)}{f_S(\lambda) + f_N(\lambda)}\right)\,d\lambda = \int_{-\pi}^{\pi} \frac{f_S(\lambda) f_N(\lambda)}{f_S(\lambda) + f_N(\lambda)}\,d\lambda.$$

6.3 The Itô integral

If $\{Z(\lambda)\}$ is a standard Wiener process it is rather obvious, cf. (6.4) on page 51, that

$$X_t = \int_{-\pi}^{\pi} e^{it\nu}\,dB(\nu)$$

is a (complex-valued) Gaussian WN with $\sigma^2 = 2\pi$. (The extension of $B(\nu)$ to negative values of ν is of no problem.) Integrals with respect to dB is called *Itô integrals* and are important in e.g. control theory and financial mathematics. Financial mathematics is of course a very broad subject. Here we have especially models for asset pricing on financial markets in mind. These aspects of financial mathematics are discussed in the course "Stochastic Calculus and the Theory of Capital Markets" (Stokastisk kalkyl och kapitalmarknadsteori).

If we consider an integral

$$I(f) = \int_0^t f(\nu)\, dB(\nu)$$

we can define it as above, or more explicitly as a "mean-square Riemann integral". Since B has unbounded variation, we can not define the integral "for each realization". We are often interested in replacing f with a random process. A typical example is to consider

$$I(B) = \int_0^t B(\nu)\, dB(\nu).$$

If we integrate by parts and forget about the unbounded variation we get

$$I(B) = \big[B(\nu)B(\nu)\big]_0^t - I(B) = B^2(t) - I(B) \quad \text{or} \quad I(B) = \frac{B^2(t)}{2}.$$

Is this correct?

If we try to define the integral by approximations we get

$$I(B) = \lim_{n\to\infty} \sum_{k=1}^{n} B(\theta_k)\big(B(t_k) - B(t_{k-1})\big)$$

where $0 = t_0 < t_1 < \ldots < t_n = t$ and $\theta_k \in [t_{k-1}, t_k]$. It turns out that the result depends on the choice of θ_k. The reason is that $dB(\nu)^2 = d\nu$ (due to the unbounded variation) and not $= 0$. It is convenient to choose $\theta_k = t_{k-1}$, since then $B(\theta_k)$ and $B(t_k) - B(t_{k-1})$ are independent. Put $t_k = t \cdot \frac{k}{n}$ and $B(t_k) = B_k$. Then

$$I(B) = \lim_{n\to\infty} \sum_{k=1}^{n} B_{k-1}(B_k - B_{k-1}) = \lim_{n\to\infty} \sum_{k=1}^{n} (B_{k-1} - B_k + B_k)(B_k - B_{k-1})$$

$$= -\lim_{n\to\infty} \sum_{k=1}^{n} (B_k - B_{k-1})^2 + \lim_{n\to\infty} \sum_{k=1}^{n} B_k(B_k - B_{k-1}).$$

For $Y \sim N(0, \sigma^2)$ we have $EY^2 = \sigma^2$ and $EY^4 = 3\sigma^4$ and thus $\mathrm{Var}(Y^2) = 2\sigma^4$. Thus

$$E\left(\lim_{n\to\infty} \sum_{k=1}^{n} (B_k - B_{k-1})^2\right) = \lim_{n\to\infty} \sum_{k=1}^{n} t \cdot \frac{1}{n} = t$$

and

$$\mathrm{Var}\left(\lim_{n\to\infty} \sum_{k=1}^{n} (B_k - B_{k-1})^2\right) = \lim_{n\to\infty} \sum_{k=1}^{n} 2 \cdot t^2 \cdot \frac{1}{n^2} = 0$$

and we get

$$I(B) = -t + \lim_{n\to\infty} \sum_{k=1}^{n} B_k(B_k - B_{k-1}) = -t + \lim_{n\to\infty} \sum_{k=1}^{n} (B_k^2 - B_{k-1}B_k)$$

$$= - t + B_n^2 + \lim_{n\to\infty} \sum_{k=1}^{n} (B_{k-1}^2 - B_{k-1}B_k)$$

$$= - t + B^2(t) + \lim_{n\to\infty} \sum_{k=1}^{n} B_{k-1}(B_{k-1} - B_k) = - t + B^2(t) - I(B)$$

and thus

$$I(B) = \frac{B^2(t)}{2} - \frac{t}{2} \; !$$

This shows that properties of stochastic integrals are by no means obvious.

Lecture 7

7.1 Estimation of the spectral density

Recall that $\gamma(\cdot)$ has the spectral representation $\gamma(h) = \int_{(-\pi,\pi]} e^{ih\nu}\, dF(\nu)$. If $\sum_{-\infty}^{\infty} |\gamma(h)| < \infty$ we have $F(\lambda) = \int_{-\pi}^{\lambda} f(\nu)\, d\nu$. Then

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda$$

and

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda}\gamma(n).$$

Let $\{X_t\}$ be a stationary time series with mean μ and with absolutely summable covariance, i.e. $\sum_{-\infty}^{\infty} |\gamma(h)| < \infty$. As usual we observe $X_1, \ldots, X_n$.

7.1.1 The periodogram

The *Fourier frequencies* are given by

$$\omega_j = \frac{2\pi j}{n}, \qquad -\pi < \omega_j \le \pi.$$

Put

$$F_n \stackrel{\text{def}}{=} \{j \in \mathbb{Z},\ -\pi < \omega_j \le \pi\} = \left\{ -\left[\frac{n-1}{2}\right], \ldots, \left[\frac{n}{2}\right] \right\},$$

where $[x]$ denotes the integer part of x.

Definition 7.1 *The* periodogram $I_n(\cdot)$ *of* $\{X_1, \ldots, X_n\}$ *is defined by*

$$I_n(\omega_j) = \frac{1}{n}\left| \sum_{t=1}^{n} X_t e^{-it\omega_j} \right|^2, \quad j \in F_n.$$

The following proposition shows that the periodogram is related to spectral estimation.

Proposition 7.1 *We have*

$$I_n(\omega_j) = \begin{cases} n|\overline{X}|^2 & \text{if } \omega_j = 0, \\ \sum\limits_{|k|<n} \widehat{\gamma}(k) e^{-ik\omega_j} & \text{if } \omega_j \neq 0, \end{cases}$$

where $\widehat{\gamma}(k) = \frac{1}{n}\sum_{t=1}^{n-|k|}(X_t - \overline{X})(X_{t+|k|} - \overline{X})$ *and* $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$.

Proof: The result is obvious when $\omega_j = 0$, so we consider $\omega_j \neq 0$. We have

$$I_n(\omega_j) = \frac{1}{n}\left|\sum_{t=1}^{n} X_t e^{-it\omega_j}\right|^2 = \frac{1}{n}\sum_{s=1}^{n} X_s e^{-is\omega_j}\sum_{t=1}^{n} X_t e^{it\omega_j}.$$

Now $\sum_{s=1}^{n} e^{-is\omega_j} = \sum_{t=1}^{n} e^{it\omega_j} = 0$, and hence

$$I_n(\omega_j) = \frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n}(X_s - \overline{X})e^{-is\omega_j}(X_t - \overline{X})e^{it\omega_j} = \sum_{|k|<n}\widehat{\gamma}(k)e^{-ik\omega_j}.$$

The last equality follows easily if we observe that each term

$$(X_s - \overline{X})e^{-is\omega_j}(X_t - \overline{X})e^{it\omega_j}$$

exists exactly once in both forms. □

Definition 7.2 (Extension of the periodogram) *For any $\omega \in [-\pi, \pi]$ we define*

$$I_n(\omega) = \begin{cases} I_n(\omega_k) & \text{if } \omega_k - \pi/n < \omega \leq \omega_k + \pi/n \text{ and } 0 \leq \omega \leq \pi, \\ I_n(-\omega) & \text{if } \omega \in [-\pi, 0). \end{cases}$$

It is sometimes comfortable to let $g(n, \omega)$, for $\omega \in [0, \pi]$, denote the multiple of $2\pi/n$ closest to ω (the smaller one if there are two) and, for $\omega \in [-\pi, 0)$, to let $g(n, \omega) = g(n, -\omega)$. Then

$$I_n(\omega) = I_n(g(n, \omega)).$$

Theorem 7.1 *We have*

$$EI_n(0) - n\mu^2 \to 2\pi f(0) \quad \text{as } n \to \infty$$

and

$$EI_n(\omega) \to 2\pi f(\omega) \quad \text{as } n \to \infty \text{ if } \omega \neq 0.$$

(If $\mu = 0$ then $I_n(\omega)$ converges uniformly to $2\pi f(\omega)$ on $[-\pi, \pi)$.)

If $\{X_t\}$ is a strictly linear time series with mean 0, i.e. if

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \qquad \{Z_t\} \sim \text{IID}(0, \sigma^2)$$

rather much can be said about the statistical behaviour of the periodogram. The following theorem gives the – for us – most important properties.

Theorem 7.2 *Let $\{X_t\}$ be a strictly linear time series with*

$$\mu = 0, \quad \sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2} < \infty \quad \text{and} \quad EZ^4 < \infty.$$

Then

$$\text{Cov}(I_n(\omega_j), I_n(\omega_k)) = \begin{cases} 2(2\pi)^2 f^2(\omega_j) + O(n^{-1/2}) & \text{if } \omega_j = \omega_k = 0 \text{ or } \pi, \\ (2\pi)^2 f^2(\omega_j) + O(n^{-1/2}) & \text{if } 0 < \omega_j = \omega_k < \pi, \\ O(n^{-1}) & \text{if } \omega_j \neq \omega_k. \end{cases}$$

Recall from the estimation of $\rho(\cdot)$ that $\widehat{\rho}(h) \sim \text{AN}(\rho(h), n^{-1}w_{hh})$ while

$$\text{Corr}(\widehat{\rho}(i), \widehat{\rho}(j)) \approx \frac{w_{ij}}{\sqrt{w_{ii}w_{jj}}}.$$

Here the situation is the "opposite", and that we will use.

7.1.2 Smoothing the periodogram

The reason why the periodogram does not work is that we estimate the same number of parameters, i.e. $\gamma(0), \ldots, \gamma(n-1)$, as we have observations. A first attempt may be to consider

$$\frac{1}{2\pi} \sum_{|k| \le m} \frac{1}{2m+1} I_n(\omega_{j+k}).$$

More generally we may consider the following class of estimators.

Definition 7.3 *The estimator* $\widehat{f}(\omega) = \widehat{f}(g(n,\omega))$ *with*

$$\widehat{f}(\omega_j) = \frac{1}{2\pi} \sum_{|k| \le m_n} W_n(k) I_n(\omega_{j+k}),$$

where

$$m_n \to \infty \quad \textit{and} \quad m_n/n \to 0 \quad \textit{as} \quad n \to \infty,$$
$$W_n(k) = W_n(-k), \quad W_n(k) \ge 0, \quad \textit{for all } k,$$
$$\sum_{|k| \le m_n} W_n(k) = 1,$$

and

$$\sum_{|k| \le m_n} W_n^2(k) \to 0 \quad \textit{as} \quad n \to \infty,$$

is called a discrete spectral average estimator *of* $f(\omega)$.
(*If* $\omega_{j+k} \notin [-\pi, \pi]$ *the term* $I_n(\omega_{j+k})$ *is evaluated by defining* I_n *to have period* 2π.)

Theorem 7.3 *Let* $\{X_t\}$ *be a strictly linear time series with*

$$\mu = 0, \quad \sum_{j=-\infty}^{\infty} |\psi_j||j|^{1/2} < \infty \quad \textit{and} \quad EZ^4 < \infty.$$

Then

$$\lim_{n\to\infty} E\widehat{f}(\omega) = f(\omega)$$

and

$$\lim_{n\to\infty} \frac{1}{\sum_{|k| \le m_n} W_n^2(k)} \text{Cov}(\widehat{f}(\omega), \widehat{f}(\lambda)) = \begin{cases} 2f^2(\omega) & \textit{if } \omega = \lambda = 0 \textit{ or } \pi, \\ f^2(\omega) & \textit{if } 0 < \omega = \lambda < \pi, \\ 0 & \textit{if } \omega \ne \lambda. \end{cases}$$

Remark 7.1 If $\mu \neq 0$ we ignore $I_n(0)$. Thus we can use

$$\widehat{f}(0) = \frac{1}{2\pi}\left(W_n(0)I_n(\omega_1) + 2\sum_{k=1}^{m_n} W_n(k)I_n(\omega_{k+1})\right).$$

Moreover, whenever $I_n(0)$ appears in $\widehat{f}(\omega_j)$ we replace it with $\widehat{f}(0)$. □

Example 7.1 For the "first attempt" we have

$$W_n(k) = \begin{cases} 1/(2m_n+1) & \text{if } |k| \leq m_n, \\ 0 & \text{if } |k| > m_n, \end{cases}$$

and

$$\mathrm{Var}(\widehat{f}(\omega)) \sim \begin{cases} \frac{1}{m_n} f^2(\omega) & \text{if } \omega = 0 \text{ or } \pi, \\ \frac{1}{m_n}\frac{f^2(\omega)}{2} & \text{if } 0 < \omega < \pi. \end{cases}$$

□

Discrete spectral average estimators may be somewhat unpractical, since the complete periodogram has to be computed for n frequencies. Thus a straightforward approach requires n^2 "operations". By using FFT the number of operations can, however, be reduced. If $n = 2^p$ the number of operations can be reduced to $2n\log_2 n$. In spite of this, it is natural to consider estimators of the following form. These estimates are, however, not discussed in [7].

Definition 7.4 *An estimator $\widehat{f}_L(\omega)$ of the form*

$$\widehat{f}_L(\omega) = \frac{1}{2\pi}\sum_{|h|\leq r_n} w(h/r_n)\widehat{\gamma}(h)e^{-ih\omega}$$

where

$$r_n \to \infty \quad \textit{and} \quad r_n/n \to 0 \quad \textit{as} \quad n \to \infty,$$
$$w(x) = w(-x), \quad w(0) = 1,$$
$$|w(x)| \leq 1, \quad \textit{for all } x,$$

and

$$w(x) = 0, \quad \textit{for } |x| > 1,$$

is called a lag window spectral estimator *of $f(\omega)$.*

We shall now show that discrete spectral average estimators and lag window spectral estimator are – essentially – the same.

Define the *spectral window*

$$W(\omega) = \frac{1}{2\pi}\sum_{|h|\leq r_n} w(h/r_n)e^{-ih\omega}$$

and the (slightly different) extension of the periodogram

$$\tilde{I}_n(\omega) = \sum_{|h|\leq n} \widehat{\gamma}(h)e^{-ih\omega}.$$

Note that

$$\widehat{\gamma}(h) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ih\lambda}\tilde{I}_n(\lambda)\,d\lambda.$$

Then we have

$$\begin{aligned}
\widehat{f}_L(\omega) &= \frac{1}{(2\pi)^2}\sum_{|h|\le r_n} w(h/r_n)\int_{-\pi}^{\pi} e^{-ih(\omega-\lambda)}\tilde{I}_n(\lambda)\,d\lambda \\
&= \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\left(\sum_{|h|\le r_n} w(h/r_n)e^{-ih(\omega-\lambda)}\right)\tilde{I}_n(\lambda)\,d\lambda \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} W(\omega-\lambda)\tilde{I}_n(\lambda)\,d\lambda \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} W(\lambda)\tilde{I}_n(\omega+\lambda)\,d\lambda \\
&\approx \frac{1}{2\pi}\sum_{|j|\le[n/2]} W(\omega_j)\tilde{I}_n(\omega+\omega_j)\,\frac{2\pi}{n} \\
&\approx \frac{1}{2\pi}\sum_{|j|\le[n/2]} W(\omega_j)I_n(g(n,\omega)+\omega_j)\,\frac{2\pi}{n}
\end{aligned}$$

Thus $\widehat{f}_L(\omega)$ is approximated by a discrete spectral average estimator with weights

$$W_n(j) = 2\pi W(\omega_j)/n, \quad |j| \le [n/2].$$

It is easy to show that

$$\sum_{|j|\le[n/2]} W_n^2(j) \approx \frac{r_n}{n}\int_{-1}^{1} w^2(x)\,dx.$$

It is thus not very surprising that the following theorem holds. (It is, however, not quite obvious since the approximating spectral average does not satisfy the conditions of the above definition of a discrete spectral average estimator.)

Theorem 7.4 *Let $\{X_t\}$ be a strictly linear time series with $\mu = 0$, $\sum_{j=-\infty}^{\infty}|\psi_j||j|^{1/2} < \infty$ and $EZ^4 < \infty$. Then*

$$\lim_{n\to\infty} E\widehat{f}_L(\omega) = f(\omega)$$

and

$$\mathrm{Var}(\widehat{f}_L(\omega)) \sim \begin{cases} \frac{r_n}{n}2f^2(\omega)\int_{-1}^{1} w^2(x)\,dx & \text{if } \omega = 0 \text{ or } \pi \\ \frac{r_n}{n}f^2(\omega)\int_{-1}^{1} w^2(x)\,dx & \text{if } 0 < \omega < \pi. \end{cases}$$

Example 7.2 (The Rectangular or Truncated Window) For this window we have

$$w(x) = \begin{cases} 1 & \text{if } |x| \le 1, \\ 0 & \text{if } |x| > 1, \end{cases}$$

and

$$\mathrm{Var}(\widehat{f}_L(\omega)) \sim \frac{2r_n}{n}f^2(\omega) \quad \text{for } 0 < \omega < \pi.$$

Example 7.3 (The Blackman-Tukey Window) For this window we have

$$w(x) = \begin{cases} 1 - 2a + 2a\cos x & \text{if } |x| \le 1, \\ 0 & \text{if } |x| > 1, \end{cases}$$

and

$$\mathrm{Var}(\widehat{f}_L(\omega)) \sim \frac{2r_n}{n}(1 - 4a + 6a^2)f^2(\omega) \quad \text{for } 0 < \omega < \pi.$$

Note that $\widehat{f}_L(\omega) = a\widehat{f}_T(\omega - \pi/r_n) + (1-2a)\widehat{f}_T(\omega) + a\widehat{f}_T(\omega + \pi/r_n)$ where $\widehat{f}_T$ is the truncated estimate. Thus this estimate is easy to compute. Usual choices of a are

.23 (The Tukey-Hamming estimate)

or

.25 (The Tukey-Hanning estimate). □

7.2 Linear filters

Let $\{X_t\}$ be a time series. A *filter* is an operation on a time series in order to obtain a new time series $\{Y_t\}$. $\{X_t\}$ is called the *input* and $\{Y_t\}$ the *output*.

A typical filter $\mathcal{C}$ is the following operation

$$Y_t = \sum_{k=-\infty}^{\infty} c_{t,k} X_k. \tag{7.1}$$

Here we assume that Y_t is well-defined. This filter is a *linear filter*. We will only be interested in linear filtering when $EX_t^2 < \infty$ and $EY_t^2 < \infty$.

A formal way to express this is to require that

$$(\mathcal{C}X)_t = Y_t \in \overline{\mathrm{sp}}\{\dots, X_{t-1}, X_t, X_{t+1}, \dots\},$$

and this can be taken as a definition of a linear filter. There do exist linear filters, which are not of the form (7.1).

Example 7.4 Let S be a random variable with $\mathrm{Var}(S) < \infty$ and $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. Consider $\{X_t\}$ given by

$$X_t = S + Z_t \quad t \in \mathbb{Z}.$$

Since $\frac{1}{n}\sum_{t=1}^n X_t \in \overline{\mathrm{sp}}\{\dots, X_{t-1}, X_t, X_{t+1}, \dots\}$ and since $\frac{1}{n}\sum_{t=1}^n X_t \xrightarrow{\text{m.s.}} S$ it follows that $S \in \overline{\mathrm{sp}}\{\dots, X_{t-1}, X_t, X_{t+1}, \dots\}$. Thus it is natural to regard $\mathcal{C}X = S$ as a linear filter.

This is a simple example of *signal detection*, where S is the signal and $\{Z_t\}$ the noise. With this interpretation one usually assumes that $S \perp \overline{\mathrm{sp}}\{Z_t\}$. Let us go back to Example 4.1 on page 29, where $X_1, \dots, X_n$ given $\Theta = \theta$ were independent and Poisson distributed with common mean θ. Let

$$X_t = \Theta + (X_t - \Theta) = S + Z_t.$$

We have

$$\mathrm{Cov}(X_t - \Theta, \Theta) = E[(X_t - \Theta)\,\Theta] = E[E[(X_t - \Theta)\,\Theta \mid \Theta]] = 0,$$

and, cf. Example 4.1,

$$\mathrm{Cov}(X_t - \Theta, X_s - \Theta) = \mathrm{Cov}(X_t, X_s) - 2\,\mathrm{Cov}(X_t, \Theta) + \mathrm{Var}(\Theta)$$

$$= \mathrm{Cov}(X_t, X_s) - \mathrm{Var}(\Theta) = \begin{cases} \mathrm{Var}(\Theta) + E(\Theta) - \mathrm{Var}(\Theta) = E(\Theta), & t = s, \\ \mathrm{Var}(\Theta) - \mathrm{Var}(\Theta) = 0, & t \neq s. \end{cases}$$

We may interpret Θ as a "signal" and $X_t - \Theta$ as "noise", where the noise is $\mathrm{WN}(0, E(\Theta))$. Notice, however, that

$$\mathrm{Var}(X_t - \Theta \mid \Theta) = \Theta,$$

which implies that Θ and $X_t - \Theta$ are not independent. □

Let us go back to (7.1). The filter $\mathcal{C}$ is called *time-invariant* if $c_{t,k}$ depends only on $t - k$, i.e. if

$$c_{t,k} = h_{t-k}.$$

Generally time-invariance may be defined as

$$\mathcal{C}(BX) = B\mathcal{C}(X) \quad \text{or even more abstractly as } \mathcal{C}B = B\mathcal{C}.$$

A time-invariant linear filter (TLF) is said to by *causal* if

$$h_j = 0 \ \text{ for } j < 0,$$

or generally if

$$(\mathcal{C}X)_t = Y_t \in \overline{\mathrm{sp}}\{\ldots, X_{t-1}, X_t\}.$$

TLF:s are especially interesting when $\{X_t\}$ is stationary, since then $\{Y_t\}$ is also stationary.

Remark 7.2 Let $\{X_t\}$ be a stationary time series with spectral density $f(\cdot)$. Theorem 4.5 on page 37 states that all linear filters are of the form (7.1) if and only if

$$0 < c_1 \leq f(\lambda) \leq c_2 < \infty \quad \text{for (almost) all } \lambda \in [-\pi, \pi].$$

Definition 7.5 *A TLF of the form (7.1) is called* stable *if* $\sum_{k=-\infty}^{\infty} |h_k| < \infty$.

From now on, when nothing else is said, we consider stable TLF:s with stationary input, i.e. we consider filters

$$Y_t = \sum_{k=-\infty}^{\infty} h_{t-k} X_k \qquad \text{where } \sum_{k=-\infty}^{\infty} |h_k| < \infty.$$

Put $h(z) = \sum_{k=-\infty}^{\infty} h_k z^k$. Then $\mathcal{C} = h(B)$.

The function $h(e^{-i\lambda})$ is called the *transfer function* (överföringsfunktion eller frekvenssvarsfunktion).

The function $|h(e^{-i\lambda})|^2$ is called the *power transfer function.*

Consider a filter $h(B)$ and the testfunctions $e^{i\lambda t}$ as input. Then

$$(h(B)e^{i\lambda\cdot})_t = \sum_{k=-\infty}^{\infty} h_{t-k}e^{i\lambda k} = \sum_{j=-\infty}^{\infty} h_j e^{i\lambda(t-j)}$$

$$= e^{i\lambda t}\sum_{j=-\infty}^{\infty} h_j e^{-i\lambda j} = e^{i\lambda t}h(e^{-i\lambda}).$$

Theorem 7.5 *Let $\{X_t\}$ be a possibly complex-valued stationary input in a stable TLF $h(B)$ and let $\{Y_t\}$ be the output, i.e. $Y = h(B)X$. Then*

(a) *$EY_t = h(1)EX_t$;*

(b) *Y_t is stationary;*

(c) *$F_Y(\lambda) = \int_{(-\pi,\lambda]} |h(e^{-i\nu})|^2\, dF_X(\nu)$.*

Proof:
(a) We have

$$EY_t = \sum_{k=-\infty}^{\infty} h_{t-k}EX_k = EX_k\sum_{k=-\infty}^{\infty} h_{t-k} = EX_k h(1).$$

(b) and (c) We have

$$\mathrm{Cov}(Y_{t+h}, Y_t) = \mathrm{Cov}\Big(\sum_{j=-\infty}^{\infty} h_j X_{t+h-j}, \sum_{k=-\infty}^{\infty} h_k X_{t-k}\Big)$$

$$= \sum_{j,k=-\infty}^{\infty} h_j\overline{h_k}\,\mathrm{Cov}(X_{t+h-j}, X_{t-k})$$

$$= \sum_{j,k=-\infty}^{\infty} h_j\overline{h_k}\gamma_X(h-j+k) = \text{thus (b) follows}$$

$$= \sum_{j,k=-\infty}^{\infty} h_j\overline{h_k}\int_{(-\pi,\pi]} e^{i\lambda(h-j+k)}\, dF_X(\lambda)$$

$$= \int_{(-\pi,\pi]} e^{i\lambda h}\Big(\sum_{j=-\infty}^{\infty} h_j e^{-i\lambda j}\Big)\Big(\overline{\sum_{k=-\infty}^{\infty} h_k e^{-i\lambda k}}\Big)\, dF_X(\lambda)$$

$$= \int_{(-\pi,\pi]} e^{i\lambda h}h(e^{-i\lambda})\overline{h(e^{-i\lambda})}\, dF_X(\lambda) = \int_{(-\pi,\pi]} e^{i\lambda h}|h(e^{-i\lambda})|^2\, dF_X(\lambda).$$

□

7.2.1 ARMA processes

Consider an ARMA(p, q) process $\{X_t\}$ given by

$$\phi(B)X_t = \theta(B)Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

and recall that

- $\{X_t\}$ is causal if and only if $\phi(z) \neq 0$ for all $|z| \leq 1$;
- $\{X_t\}$ is invertible if and only if $\theta(z) \neq 0$ for all $|z| \leq 1$;
- there exists no stationary solution if $\phi(z) = 0$ for some z with $|z| = 1$.

It is easily seen that $f_Z(\lambda) = \frac{\sigma^2}{2\pi}$. Using testfunctions we get

$$\phi(e^{-i\lambda})h(e^{-i\lambda}) = \theta(e^{-i\lambda}) \quad \text{or} \quad h(e^{-i\lambda}) = \frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}$$

and thus

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{\theta(e^{-i\lambda})}{\phi(e^{-i\lambda})}\right|^2,$$

which was derived in (2.14) on page 17 in a different way.

Example 7.5 (AR(1) process) Let $\{X_t\}$ be an AR(1) process, i.e.

$$X_t = Z_t + \phi_1 X_{t-1} \quad \text{or} \quad \phi(z) = 1 - \phi_1 z \text{ and } \theta(z) = 1.$$

Thus

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{1}{1-\phi_1 e^{-i\lambda}}\right|^2 = \frac{\sigma^2}{2\pi}\frac{1}{1-2\phi_1\cos\lambda+\phi_1^2}.$$

Assume now that $\{X_t\}$ is *not* causal, i.e. that $|\phi_1| > 1$. Then, cf. Example 2.1 on page 15,

$$\phi_1^{-1}X_t = \phi_1^{-1}Z_t + X_{t-1} \quad \text{or} \quad X_t = -\phi_1^{-1}Z_{t+1} + \phi_1^{-1}X_{t+1}.$$

Let us now introduce the time reversed processes $X_t^* = X_{-t}$ and $Z_t^* = -Z_{-t}$, where obviously also Z_t^* is a WN. Then

$$X_{-t}^* = \phi_1^{-1}Z_{-t-1}^* + \phi_1^{-1}X_{-t-1}^*.$$

Replacing $-t$ with t we get

$$X_t^* = \phi_1^{-1}Z_{t-1}^* + \phi_1^{-1}X_{t-1}^*,$$

and thus X_t^* is a causal AR(1) process. (The fact that "actual" WN is replaced by "one time unit backward" WN does not destroy the causality.) Since time reversal does not change the autocovariance function we have found a "causal representation", but we shall note that the variance of the WN is different.

We can discuss also this in a slightly different way. Put $a_1 = 1/\phi_1$, so that a_1 is the solution of $\phi(z) = 0$. Now we allow a_1 to be complex and note that $|a_1| < 1$. Really, we do only use $|\phi(z)|^2$ for $z = e^{-i\lambda}$, i.e. for z such that $|z| = 1$. Now

$$|\phi(z)|^2 = \left(1-\frac{z}{a_1}\right)\left(\overline{1-\frac{z}{a_1}}\right) = \left(1-\frac{z}{a_1}\right)\left(1-\frac{\overline{z}}{\overline{a}_1}\right)$$

$$= \left(1-\frac{z}{a_1}\right)\left(1-\frac{1}{z\overline{a}_1}\right)$$

where the last equality holds on $|z| = 1$. Since $1 - \frac{1}{z\overline{a}_1} = 0$ gives $z = 1/\overline{a}_1$ it is natural to consider

$$\tilde{\phi}(z) = 1 - \overline{a}_1 z,$$

and thus an AR(1) process $\{\widetilde{X}_t\}$ defined by

$$\tilde{\phi}(B)\widetilde{X}_t = Z_t \quad \text{or} \quad \widetilde{X}_t = Z_t + \overline{a}_1 \widetilde{X}_{t-1} \left(= Z_t + \overline{\phi}_1^{-1} \widetilde{X}_{t-1} \right).$$

(The fact that we consider complex-valued ϕ_1 is for later purposes.)

Thus

$$f_{\widetilde{X}}(\lambda) = \frac{\sigma^2}{2\pi} \left| \frac{1}{1 - \tilde{\phi}_1 e^{-i\lambda}} \right|^2 = \frac{\sigma^2}{2\pi} \left| \frac{1}{1 - \overline{\phi}_1^{-1} e^{-i\lambda}} \right|^2$$

$$= \frac{\sigma^2 |\phi_1|^2}{2\pi} \left| \frac{1}{\overline{\phi}_1 - e^{-i\lambda}} \right|^2 = \frac{\sigma^2 |\phi_1|^2}{2\pi} \left| \frac{1}{\overline{\phi}_1 e^{i\lambda} - 1} \right|^2 = \frac{\sigma^2 |\phi_1|^2}{2\pi} \left| \frac{1}{1 - \phi_1 e^{-i\lambda}} \right|^2.$$

Thus the AR(1) process $\{X_t^+\}$ defined by

$$\tilde{\phi}(B)X_t^+ = \widetilde{Z}_t \quad \text{or} \quad X_t^+ = \widetilde{Z}_t + \overline{a}_1 X_{t-1}^+ \left(= \widetilde{Z}_t + \overline{\phi}_1^{-1} X_{t-1}^+ \right),$$

where

$$\{\widetilde{Z}\} \sim \mathrm{WN}(0, \sigma^2 |a_1|^2) = \mathrm{WN}(0, \sigma^2 / |\phi_1|^2),$$

has the same second-order properties as $\{X_t\}$, i.e. $f_{X^+}(\lambda) = f_X(\lambda)$.

In fact, the process $\{X_t\}$ itself has the causal representation

$$\tilde{\phi}(B)X_t = Z_t^* \quad \text{where } \{Z^*\} \sim \mathrm{WN}(0, \sigma^2 |a_1|^2) = \mathrm{WN}(0, \sigma^2 / |\phi_1|^2). \tag{7.2}$$

This follows by using (7.2) as definition of $\{Z^*\}$ since

$$f_{Z^*}(\lambda) = |\phi(e^{-i\lambda})|^2 f_X(\lambda) = |\phi(e^{-i\lambda})|^2 \frac{\sigma^2}{2\pi} \left| \frac{1}{1 - \phi_1 e^{-i\lambda}} \right|^2 = \frac{\sigma^2}{2\pi}.$$

□

The methods in an AR(1) is easily transferred to an ARMA(p, q) process, where some of zeros lie "wrongly", i.e. inside the unique circle. More precisely, consider the polynomials $\phi(z)$ and $\theta(z)$ and let $a_1, \ldots, a_p$ and $b_1, \ldots, b_q$ denote the zeros.

Thus, by factorization, we get

$$\phi(z) = \prod_{j=1}^{p} (1 - a_j^{-1} z) \quad \text{and} \quad \theta(z) = \prod_{j=1}^{q} (1 - b_j^{-1} z).$$

Assume that

$$|a_j| > 1, \quad 1 \le j \le r \qquad |a_j| < 1, \quad r < j \le p$$

and

$$|b_j| > 1, \quad 1 \le j \le s \qquad |b_j| < 1, \quad s < j \le p.$$

Define

$$\tilde{\phi}(z) = \prod_{j=1}^{r} (1 - a_j^{-1} z) \prod_{j=r+1}^{p} (1 - \overline{a}_j z)$$

and

$$\tilde{\theta}(z) = \prod_{j=1}^{s} (1 - b_j^{-1} z) \prod_{j=s+1}^{q} (1 - \overline{b}_j z).$$

Then $\{X_t\}$, originally having the representation

$$\phi(B)X_t = \theta(B)Z_t, \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

has the causal and invertible representation

$$\tilde{\phi}(B)X_t = \tilde{\theta}(B)Z_t^*, \qquad \{Z_t^*\} \sim \mathrm{WN}\left(0, \sigma^2 \frac{\prod_{j=r+1}^{p} |a_j|^2}{\prod_{j=s+1}^{q} |b_j|^2}\right).$$

One reason to consider "parametric" models is that "kind" processes can be approximated in some sense by them.

Consider a (real-valued) stationary time series $\{X_t\}$ with continuous spectral density f_X.

Theorem 7.6 *If f_X is a symmetric continuous spectral density and $\varepsilon > 0$ then there exist an invertible* MA(q) *process $\{Y_t\}$ and a causal* AR(p) *process $\{U_t\}$ such that*

$$|f_Y(\lambda) - f_X(\lambda)| < \varepsilon \quad \textit{for all } \lambda \in [-\pi, \pi]$$

and

$$|f_U(\lambda) - f_X(\lambda)| < \varepsilon \quad \textit{for all } \lambda \in [-\pi, \pi]$$

If ε is small, one may often have to choose q and p rather large. In practice often is possible to find an ARMA(p', q') process such that $p' + q'$ is smaller that q or p. Some discussion about this is found in [5], which is the "Bible" of ARMA processes.

Lecture 8

8.1 Estimation for ARMA models

The determination of an appropriate ARMA(p, q) model

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2),$$

requires generally first an order selection, i.e. a choice of p and q, and then an estimation of remaining parameters, i.e. the mean (which is already assumed to have been removed above),

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_q \end{pmatrix} \quad \text{and} \quad \sigma^2.$$

As usual we assume that $X_1, \ldots, X_n$ are observed. Notice that we have assumed $\{Z_t\} \sim \text{IID}(0, \sigma^2)$; the reason is that we will give some asymptotic results slightly more precise than given in [7].

8.1.1 Yule-Walker estimation

Consider a causal zero-mean AR(p) process $\{X_t\}$:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2),$$

and recall the Yule-Walker equations

$$\gamma(j) - \phi_1 \gamma(j-1) - \ldots - \phi_p \gamma(j-p) = \begin{cases} 0, & j = 1, \ldots, p, \\ \sigma^2, & j = 0, \end{cases}$$

discussed in Example 5.2 on page 43. If we write these equations on the form

$$\phi_1 \gamma(j-1) + \ldots + \phi_p \gamma(j-p) = \begin{cases} \gamma(j), & j = 1, \ldots, p, \\ \gamma(0) - \sigma^2, & j = 0, \end{cases}$$

we get

$$\Gamma_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p$$

and

$$\boldsymbol{\phi}' \boldsymbol{\gamma}_p = \gamma(0) - \sigma^2 \quad \text{or} \quad \sigma^2 = \gamma(0) - \boldsymbol{\phi}' \boldsymbol{\gamma}_p,$$

where

$$\Gamma_p = \begin{pmatrix} \gamma(0) & \dots & \gamma(p-1) \\ \vdots & & \\ \gamma(p-1) & \dots & \gamma(0) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\gamma}_p = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(p) \end{pmatrix}.$$

Often the Yule-Walker equations are used to determine $\gamma(\cdot)$ from σ^2 and $\boldsymbol{\phi}$, as was done in Example 5.2.

If we, on the other hand, replace Γ_p and $\boldsymbol{\gamma}_p$ with the estimates $\widehat{\Gamma}_p$ and $\widehat{\boldsymbol{\gamma}}_p$ we obtain the following equations for the *Yule-Walker estimates*

$$\widehat{\Gamma}_p \widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\gamma}}_p \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0) - \widehat{\boldsymbol{\phi}}' \widehat{\boldsymbol{\gamma}}_p,$$

where, of course,

$$\widehat{\Gamma}_p = \begin{pmatrix} \widehat{\gamma}(0) & \dots & \widehat{\gamma}(p-1) \\ \vdots & & \\ \widehat{\gamma}(p-1) & \dots & \widehat{\gamma}(0) \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\gamma}}_p = \begin{pmatrix} \widehat{\gamma}(1) \\ \vdots \\ \widehat{\gamma}(p) \end{pmatrix}.$$

(It may seem unnatural to use $\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n)$ instead of $\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h} X_t X_{t+h}$ when we "know" that $\mu = 0$. In practice we have removed the mean, which means that $\overline{X}_n = 0$.)

Recall that $\widehat{\rho}(\cdot)$ has nicer properties than $\widehat{\gamma}(\cdot)$. In this case we can just divide the equations with $\widehat{\gamma}(0)$, and thus we get

$$\widehat{R}_p \widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{\rho}}_p \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0)[1 - \widehat{\boldsymbol{\phi}}' \widehat{\boldsymbol{\rho}}_p],$$

where, of course, $\widehat{R}_p = \frac{1}{\widehat{\gamma}(0)}\widehat{\Gamma}_p$ and $\widehat{\boldsymbol{\rho}}_p = \frac{1}{\widehat{\gamma}(0)}\widehat{\boldsymbol{\gamma}}_p$. Finally we get

$$\widehat{\boldsymbol{\phi}} = \widehat{R}_p^{-1} \widehat{\boldsymbol{\rho}}_p \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0)[1 - \widehat{\boldsymbol{\rho}}_p' \widehat{R}_p^{-1} \widehat{\boldsymbol{\rho}}_p].$$

We have the following theorem.

Theorem 8.1 *If $\{X_t\}$ is a causal AR(p) process with $\{Z_t\} \sim \text{IID}(0, \sigma^2)$, and $\widehat{\boldsymbol{\phi}}$ is the Yule-Walker estimate of $\boldsymbol{\phi}$, then*

$$\widehat{\boldsymbol{\phi}} \sim \text{AN}\left(\boldsymbol{\phi}, \frac{\sigma^2 \Gamma_p^{-1}}{n}\right), \quad \textit{for large values of } n.$$

Moreover,

$$\widehat{\sigma}^2 \xrightarrow{P} \sigma^2.$$

Assume now that $q > 0$, i.e. that we have an ARMA(p, q) process. Instead of the using the Yule-Walker equations we use (5.1) and (5.2) on page 43, from which in fact the Yule-Walker equations were derived. The resulting estimates $\widehat{\boldsymbol{\phi}}$ and $\widehat{\boldsymbol{\theta}}$ may be regarded as obtained by *the method of moments*. We will illustrate this in the MA(1) case.

Example 8.1 (MA(1) process) Let $\{X_t\}$ be a MA(1) process:

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2).$$

where $|\theta| < 1$. In this case (5.1) reduces to

$$\gamma(0) = \sigma^2(1 + \theta^2) \quad \gamma(1) = \sigma^2\theta$$

and (5.2) to $\gamma(k) = 0$ for $k \geq 2$, i.e. we get the autocovariance function, just as we ought to. Using

$$\rho(1) = \frac{\gamma(1)}{\gamma(0)} = \frac{\theta}{1 + \theta^2},$$

it is natural to estimate θ by the method of moments, i.e. to use

$$\widehat{\rho}(1) = \frac{\widehat{\gamma}(1)}{\widehat{\gamma}(0)} = \frac{\widehat{\theta}_n^{(1)}}{1 + \left(\widehat{\theta}_n^{(1)}\right)^2}.$$

This equation has a solution for $|\widehat{\rho}(1)| < \frac{1}{2}$ and it is natural to put

$$\widehat{\theta}_n^{(1)} = \begin{cases} -1 & \text{if } \widehat{\rho}(1) < -\frac{1}{2}\,, \\ \dfrac{1 - \sqrt{1 - 4\widehat{\rho}(1)^2}}{2\widehat{\rho}(1)} & \text{if } |\widehat{\rho}(1)| < \frac{1}{2}\,, \\ 1 & \text{if } \widehat{\rho}(1) > \frac{1}{2}\,. \end{cases}$$

The estimate $\widehat{\theta}_n^{(1)}$ is consistent and further it can be shown that

$$\widehat{\theta}_n^{(1)} \sim \text{AN}(\theta, n^{-1}\sigma_1^2(\theta)), \quad \text{for large values of } n,$$

where

$$\sigma_1^2(\theta) = \frac{1 + \theta^2 + 4\theta^4 + \theta^6 + \theta^8}{(1 - \theta^2)^2}.$$

□

These estimates are known to be good in the AR case, but less good when $q > 0$.

Consider again an AR(p) process. Up to now we have argued as if p was known. A usual way to proceed is as if $\{X_t\}$ was an AR(m) process for $m = 1, 2, \ldots$ until we believe that $m \geq p$. Put, for any fixed $m > p$,

$$\boldsymbol{\phi}_m = \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Theorem 8.1 does still hold (with p replaced by m and $\boldsymbol{\phi}$ by $\boldsymbol{\phi}_m$).

Let us for a moment go back to prediction. Recall that the best linear predictor $\widehat{X}_{n+1}$ of X_{n+1} in terms of $X_1, X_2, \ldots, X_n$ is

$$\widehat{X}_{n+1} = \sum_{i=1}^{n} \phi_{n,i} X_{n+1-i}, \quad n = 1, 2, \ldots,$$

where $\boldsymbol{\phi}_n = \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n$. It is easy to realize, and intuitively quite natural, that $\boldsymbol{\phi}_n = \begin{pmatrix} \boldsymbol{\phi} \\ \mathbf{0} \end{pmatrix}$ when $n > p$ for a causal AR(p) process.

Thus if we try as $\{X_t\}$ was an AR(m) process for $m = 1, 2, \ldots$ we can use the Durbin-Levinson algorithm, see Theorem 4.3 on page 33, with $\gamma(\cdot)$ replaced by $\widehat{\gamma}(\cdot)$.

8.1.2 Burg's algorithm

There exists in the literature several – at least two – algorithms due to Burg. One such algorithm – and not the one to be considered here – is the so called *maximum entropy spectral analysis*. The main idea in that method is to regard $\widehat{\gamma}(h)$ for $h = 0, \ldots, p$ as the "true" covariances and to "estimate" the spectral density with the corresponding spectral density of an AR(p) process. As a parallel we may recall the truncated window, discussed in Example 7.2 on page 67, where the spectral density is estimated with the corresponding spectral density of an MA(r_n) process.

The algorithm to be considered here may also be said to rely on AR-processes, although not as explicitly as in the one mentioned. Assume as usual that $x_1, \ldots, x_n$ are the observations. The idea is to consider one observation after the other and to "predict" it both by forward and backward data. It seems, to the best of our understanding, as if there is some misprints in (at least in the first printing) [7, pp. 145–146] and the algorithm ought to be as follows:

Burg's algorithm:

$$d(1) = \tfrac{1}{2}x_1^2 + x_2^2 + \ldots + x_{n-1}^2 + \tfrac{1}{2}x_n^2 \tag{8.1}$$

$$\phi_{ii}^{(B)} = \frac{1}{d(i)} \sum_{t=i+1}^{n} v_{i-1}(t) u_{i-1}(t-1) \tag{8.2}$$

$$\sigma_i^{(B)2} = \frac{d(i)\big(1 - \phi_{ii}^{(B)2}\big)}{n-i} \tag{8.3}$$

$$d(i+1) = d(i)\big(1 - \phi_{ii}^{(B)2}\big) - \tfrac{1}{2}v_i^2(i+1) - \tfrac{1}{2}u_i^2(n). \tag{8.4}$$

Only (8.2) is in agreement with the algorithm given in the first printing of [7, p. 146]. Further it seems as (5.1.20) in [7, p. 145] ought to be

$$v_i(t) = v_{i-1}(t) - \phi_{ii} u_{i-1}(t-1). \tag{8.5}$$

We will therefore consider the algorithm in some detail.

Consider an observation x_k and its forward and backward predictors (based on i observations)

$$\widehat{x}_k^{(f,i)} = \phi_{i1}x_{k-1} + \ldots + \phi_{ii}x_{k-i}, \quad k = i+1, \ldots, n,$$

and

$$\widehat{x}_k^{(b,i)} = \phi_{i1}x_{k+1} + \ldots + \phi_{ii}x_{k+i}, \quad k = 1, \ldots, n-i.$$

The forward predictor is the "usual" predictor. We will not use the predictors but merely the forward and backward prediction errors defined by

$$u_i(t) = x_{n+1+i-t} - \widehat{x}_{n+1+i-t}^{(f,i)}, \quad t = i+1, \ldots n,\ 0 \le i < n,$$

and

$$v_i(t) = x_{n+1-t} - \widehat{x}_{n+1-t}^{(b,i)}, \quad t = i+1, \ldots n,\ 0 \le i < n.$$

The indices seem rather horrible, but will turn out in the end to be convenient.

Let us see what this means:

$i = 0$.

We have

$$u_0(t) = x_{n+1-t} - \widehat{x}_{n+1-t}^{(f,0)} = x_{n+1-t} - 0 = x_{n+1-t}, \quad t = 1, \ldots n,$$

and

$$v_0(t) = x_{n+1-t} - \widehat{x}_{n+1-t}^{(b,0)} = x_{n+1-t}, \quad t = 1, \ldots n.$$

$i = 1$.

We have, for $t = 2, \ldots n$,

$$u_1(t) = x_{n+2-t} - \phi_{11}x_{n+1-t} = u_0(t-1) - \phi_{11}v_0(t),$$

which is in agreement with (5.1.19) in [7], and

$$v_1(t) = x_{n+1-t} - \phi_{11}x_{n+2-t} = v_0(t) - \phi_{11}u_0(t-1),$$

which is *not* in agreement with (5.1.20) in [7, p. 145], but with (8.5).

General i.

Let us consider $v_i(t)$. In order to relate $v_i(t)$ with prediction errors based on $i-1$ observations we use the Durbin-Levinson algorithm, see Theorem 4.3 on page 33. Then we get

$$\begin{aligned} v_i(t) = x_{n+1-t} - \widehat{x}_{n+1-t}^{(b,i)} &= x_{n+1-t} - \big(\widehat{x}_{n+1-t}^{(b,i-1)} - \phi_{ii}\widehat{x}_{n+1+i-t}^{(f,i-1)}\big) - \phi_{ii}x_{n+1+i-t} \\ &= x_{n+1-t} - \widehat{x}_{n+1-t}^{(b,i-1)} - \phi_{ii}\big(x_{n+1+i-t} - \widehat{x}_{n+1+i-t}^{(f,i-1)}\big) \\ &= v_{i-1}(t) - \phi_{ii}u_{i-1}(t-1), \end{aligned}$$

which is (8.5).

Suppose now that we know $\phi_{i-1,k}$ for $k = 1, \ldots, i-1$ and ϕ_{ii}. Then $\phi_{i,k}$ for $k = 1, \ldots, i-1$ may be obtained by the Durbin-Levinson algorithm. Thus the main problem is to obtain an algorithm for calculating ϕ_{ii} for $i = 1, 2, \ldots$

The *Burg estimate* $\phi_{ii}^{(B)}$ of ϕ_{ii} is obtained by minimizing

$$\sigma_i^2 \stackrel{\text{def}}{=} \frac{1}{2(n-i)} \sum_{t=i+1}^{n} \big[u_i^2(t) + v_i^2(t)\big]$$

with respect to ϕ_{ii}. Using (5.1.19) in [7] and (8.5) we get

$$\sigma_i^2 = \frac{1}{2(n-i)} \sum_{t=i+1}^{n} \big[\big(u_{i-1}(t-1) - \phi_{ii}v_{i-1}(t)\big)^2 + \big(v_{i-1}(t) - \phi_{ii}u_{i-1}(t-1)\big)^2\big]. \qquad (8.6)$$

Differentiation with respect to ϕ_{ii} yields

$$\frac{\sigma_i^2}{d\phi_{ii}} = -\frac{1}{(n-i)}\sum_{t=i+1}^{n}\big[\big(u_{i-1}(t-1)-\phi_{ii}v_{i-1}(t)\big)v_{i-1}(t)+\big(v_{i-1}(t)-\phi_{ii}u_{i-1}(t-1)\big)u_{i-1}(t-1)\big]$$

$$= -\frac{1}{(n-i)}\sum_{t=i+1}^{n}\big[2\big(u_{i-1}(t-1)v_{i-1}(t)-\phi_{ii}\cdot\big(v_{i-1}^2(t)+u_{i-1}^2(t-1)\big)\big)\big].$$

Putting $\frac{\sigma_i^2}{d\phi_{ii}} = 0$ leads to the Burg estimate

$$\phi_{ii}^{(B)} = \frac{2\sum_{t=i+1}^{n} v_{i-1}(t)u_{i-1}(t-1)}{\sum_{t=i+1}^{n}\big[v_{i-1}^2(t)+u_{i-1}^2(t-1)\big]}\,.$$

Let

$$d(i) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{t=i+1}^{n}\big[v_{i-1}^2(t)+u_{i-1}^2(t-1)\big], \tag{8.7}$$

which especially means that

$$d(1) = \frac{1}{2}\sum_{t=2}^{n}\big[v_0^2(t)+u_0^2(t-1)\big] = \frac{1}{2}\sum_{t=2}^{n}\big[x_{n+1-t}^2+x_{n+2-t}^2\big] = \frac{1}{2}\sum_{t=2}^{n}\big[x_{t-1}^2+x_t^2\big]$$

$$= \tfrac{1}{2}x_1^2+x_2^2+\ldots+x_{n-1}^2+\tfrac{1}{2}x_n^2,$$

which is (8.1).

Using (8.7) we get

$$\phi_{ii}^{(B)} = \frac{1}{d(i)}\sum_{t=i+1}^{n} v_{i-1}(t)u_{i-1}(t-1),$$

which is (8.2).

The Burg estimate $\sigma_i^{(B)2}$ of σ_i^2 is the minimum value of (8.6), i.e.

$$2(n-i)\sigma_i^{(B)2} = \sum_{t=i+1}^{n}\big[\big(u_{i-1}(t-1)-\phi_{ii}^{(B)}v_{i-1}(t)\big)^2+\big(v_{i-1}(t)-\phi_{ii}^{(B)}u_{i-1}(t-1)\big)^2\big]$$

$$= \sum_{t=i+1}^{n}\big[\big(u_{i-1}^2(t-1)+v_{i-1}^2(t)\big)\big(1+\phi_{ii}^{(B)2}\big)-4\phi_{ii}^{(B)}u_{i-1}(t-1)v_{i-1}(t)\big]$$

$$= 2d(i)\big(1+\phi_{ii}^{(B)2}\big)-4\phi_{ii}^{(B)}d(i)\phi_{ii}^{(B)} = 2d(i)\big(1-\phi_{ii}^{(B)2}\big),$$

or

$$\sigma_i^{(B)2} = \frac{d(i)\big(1-\phi_{ii}^{(B)2}\big)}{n-i}$$

which is (8.3).

The next step in the algorithm is to express $d(i+1)$ in a convenient way. In order to do so we combine the definition of σ_i^2 with the above expression. Then we get

$$d(i+1) = \frac{1}{2}\sum_{t=i+2}^{n}\big[v_i^2(t)+u_i^2(t-1)\big] = \frac{1}{2}\sum_{t=i+1}^{n}\big[v_i^2(t)+u_i^2(t)\big]-\tfrac{1}{2}v_i^2(i+1)-\tfrac{1}{2}u_i^2(n)$$

$$= d(i)\big(1-\phi_{ii}^{(B)2}\big)-\tfrac{1}{2}v_i^2(i+1)-\tfrac{1}{2}u_i^2(n),$$

which is (8.4).

The Burg estimates for an AR(p) have the same statistical properties for large values of n as the Yule-Walker estimate, i.e. Theorem 8.1 on page 76 holds.

8.1.3 The innovations algorithm

Since an MA(q) process

$$X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2),$$

has, by definition, an innovation representation, it is natural to use the innovations algorithm for prediction in a similar way as the Durbin-Levinson algorithm was used. Since, generally, q is unknown, we can try to fit MA models

$$X_t = Z_t + \widehat{\theta}_{m1} Z_{t-1} + \ldots + \widehat{\theta}_{mm} Z_{t-m}, \quad \{Z_t\} \sim \text{IID}(0, \widehat{v}_m),$$

of orders $m = 1, 2, \ldots$, by means of the innovations algorithm.

Definition 8.1 (Innovations estimates of MA parameters) *If $\widehat{\gamma}(0) > 0$ we define the innovations estimates*

$$\widehat{\boldsymbol{\theta}}_m = \begin{pmatrix} \widehat{\theta}_{m1} \\ \vdots \\ \widehat{\theta}_{mm} \end{pmatrix} \quad \text{and} \quad \widehat{v}_m, \quad m = 1, 2, \ldots, n-1,$$

by the recursion relations

$$\begin{cases} \widehat{v}_0 = \widehat{\gamma}(0), \\ \widehat{\theta}_{m,m-k} = \widehat{v}_k^{-1} \left(\widehat{\gamma}(m-k) - \sum_{j=0}^{k-1} \widehat{\theta}_{m,m-j} \widehat{\theta}_{k,k-j} \widehat{v}_j \right), \quad k = 0, \ldots, m-1, \\ \widehat{v}_m = \widehat{\gamma}(0) - \sum_{j=0}^{m-1} \widehat{\theta}_{m,m-j}^2 \widehat{v}_j. \end{cases}$$

This method, as we will see, works also for causal invertible ARMA processes. The following theorem gives asymptotic statistical properties of the innovations estimates.

Theorem 8.2 *Let $\{X_t\}$ be the causal invertible* ARMA *process $\phi(B)X_t = \theta(B)Z_t$, $\{Z_t\} \sim \text{IID}(0, \sigma^2)$, $EZ_t^4 < \infty$, and let $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, |z| \leq 1$ (with $\psi_0 = 1$ and $\psi_j = 0$ for $j < 0$). Then for any sequence of positive integers $\{m(n),\ n = 1, 2, \ldots\}$ such that $m \to \infty$ and $m = o(n^{1/3})$ as $n \to \infty$, we have for each fixed k,*

$$\begin{pmatrix} \widehat{\theta}_{m1} \\ \vdots \\ \widehat{\theta}_{mk} \end{pmatrix} \sim \text{AN}\left(\begin{pmatrix} \psi_1 \\ \vdots \\ \psi_k \end{pmatrix}, n^{-1} A \right),$$

where $A = (a_{ij})_{i,j=1,\ldots,k}$ and

$$a_{ij} = \sum_{r=1}^{\min(i,j)} \psi_{i-r} \psi_{j-r}.$$

Moreover,

$$\widehat{v}_m \xrightarrow{P} \sigma^2.$$

Before discussing the theorem, we consider the simplest example.

Example 8.2 (MA(1) process) Let $\{X_t\}$ be given by

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2),$$

where $|\theta| < 1$. Then

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2 & \text{if } h = 0, \\ \theta\sigma^2 & \text{if } |h| = 1, \\ 0 & \text{if } |h| > 1. \end{cases}$$

Consider the innovations algorithm.

$m = 1$

$k = 0$

$$\widehat{\theta}_{11} = \widehat{v}_0^{-1}\left(\widehat{\gamma}(1)\right) = \frac{\widehat{\gamma}(1)}{\widehat{\gamma}(0)} \xrightarrow{P} \frac{\theta}{1+\theta^2} \quad \text{as } n \to \infty,$$

$$\widehat{v}_1 = \widehat{\gamma}(0) - \widehat{\theta}_{11}^2 \widehat{v}_0 = \frac{\widehat{\gamma}^2(0) - \widehat{\gamma}^2(1)}{\widehat{\gamma}(0)} \xrightarrow{P} \frac{(1+\theta^2)^2 - \theta^2}{1+\theta^2}\,\sigma^2 \quad \text{as } n \to \infty$$

$m = 2$

$k = 0$

$$\widehat{\theta}_{22} = \widehat{v}_0^{-1}\left(\widehat{\gamma}(2)\right) \xrightarrow{P} 0 \quad \text{as } n \to \infty$$

$k = 1$

$$\widehat{\theta}_{21} = \widehat{v}_1^{-1}\left(\widehat{\gamma}(1) - \widehat{\theta}_{22}\widehat{\theta}_{11}\widehat{v}_0\right) = \widehat{v}_1^{-1}\left(\widehat{\gamma}(1) - \frac{\widehat{\gamma}(1)\widehat{\gamma}(2)}{\widehat{\gamma}(0)}\right)$$

$$\xrightarrow{P} \frac{\theta(1+\theta^2)}{(1+\theta^2)^2 - \theta^2} = \frac{\theta}{(1+\theta^2) - \frac{\theta^2}{1+\theta^2}} \quad \text{as } n \to \infty$$

$\vdots$

$\vdots$

General m

$k = 0$

$$\widehat{\theta}_{mm} = \widehat{v}_0^{-1}\left(\widehat{\gamma}(m)\right) \xrightarrow{P} 0 \quad \text{as } n \to \infty$$

$k = 1$

$$\widehat{\theta}_{m,m-1} = \widehat{v}_1^{-1}\left(\widehat{\gamma}(m-1) - \widehat{\theta}_{mm}\widehat{\theta}_{11}\widehat{v}_0\right) \xrightarrow{P} 0 \quad \text{as } n \to \infty$$

$\vdots$

$k = m - 1$

$$\widehat{\theta}_{m1} = \widehat{v}_{m-1}^{-1}\left(\widehat{\gamma}(1) - \sum_{j=0}^{m-2} \dots\right) \xrightarrow{P} \frac{\gamma(1)}{v_{m-1}} \quad \text{as } n \to \infty$$

at least if $\widehat{v}_{m-1} \xrightarrow{P} v_{m-1}$ as $n \to \infty$. Then

$$\widehat{v}_m \xrightarrow{P} \gamma(0) - \left(\frac{\gamma(1)}{v_{m-1}}\right)^2 v_{m-1} = (1+\theta^2)\sigma^2 - \frac{\theta^2\sigma^4}{v_{m-1}}.$$

If, further, $v_m \approx v_{m-1}$ $(= v)$ we get, compare Example 5.2.1 and Problem 5.5

$$v = (1+\theta^2)\sigma^2 - \frac{\theta^2\sigma^4}{v} \quad \text{or} \quad v = \sigma^2.$$

If we now apply the theorem we have

$$\psi_0 = 1, \quad \psi_1 = \theta \quad \text{and} \quad \psi_j = 0 \ \text{ for } j > 1.$$

Thus

$$A = \begin{pmatrix} 1 & \theta & 0 & 0 & 0 & \dots & 0 & 0 \\ \theta & 1+\theta^2 & \theta & 0 & 0 & \dots & 0 & 0 \\ 0 & \theta & 1+\theta^2 & \theta & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & & & & & \\ 0 & 0 & \dots & & 0 & \dots & 1+\theta^2 & \theta \\ 0 & 0 & \dots & & 0 & \dots & \theta & 1+\theta^2 \end{pmatrix}$$

and especially it follows that $\widehat{\theta}_{m1} \sim \mathrm{AN}(\theta, n^{-1})$ when n and m are large and $m/n^{1/3}$ is small. □

Since $\widehat{\theta}_{m1} \sim \mathrm{AN}(\theta, n^{-1})$ and since $\sigma_1^2(\theta) > 1$ it follows that $\widehat{\theta}_n^{(1)}$ is asymptotically less effective than $\widehat{\theta}_{m1}$.

For an AR(p) the Yule-Walker estimate $\widehat{\boldsymbol{\phi}}_p$ is consistent, i.e. $\widehat{\boldsymbol{\phi}}_p \xrightarrow{P} \boldsymbol{\phi}_p$ as $n \to \infty$. However, for an MA(q) process the estimator $\widehat{\boldsymbol{\theta}}_q$ is not consistent. In order to get consistent estimates we must consider estimates $\begin{pmatrix} \widehat{\theta}_{m1} \\ \vdots \\ \widehat{\theta}_{mq} \end{pmatrix}$ where m and n have the above relation. The requirement "$m \to \infty$" gives the consistency, while "$m = o(n^{1/3})$" guarantees that the number of parameters is enough less than the number of observations.

Information about q are given by both the estimates of θ_m and $\gamma(m)$ since $\theta_m = \gamma(m) = 0$ for $m > q$.

8.1.4 The Hannan–Rissanen algorithm

In a causal AR(p) model it is natural to use *least square estimation*. We will consider this case in some detail.

The defining equation

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t$$

of a causal zero-mean AR(p) can be written on the form

$$\boldsymbol{Y} = X\boldsymbol{\phi} + \boldsymbol{Z} \quad \text{or} \quad \boldsymbol{Z} = \boldsymbol{Y} - X\boldsymbol{\phi},$$

where

$$\boldsymbol{Y} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad \boldsymbol{x}_k = \begin{pmatrix} X_{-k} \\ \vdots \\ X_{n-1-k} \end{pmatrix} \quad \text{for } k = 0, \ldots, p-1,$$

$$X = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{p-1}) = \begin{pmatrix} X_0 & X_{-1} & \ldots & X_{1-p} \\ \vdots & \vdots & & \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} \end{pmatrix} \quad \text{and} \quad \boldsymbol{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}.$$

The idea in least square estimation is to consider the X_k:s as fixed and to minimize $\boldsymbol{Z}'\boldsymbol{Z}$ with respect to $\boldsymbol{\phi}$. We assume that $X_{-p+1}, \ldots, X_n$ are observed. Let $\boldsymbol{\phi}^*$ denote the least square estimate, i.e. the value of $\boldsymbol{\phi}$ which minimizes

$$S(\boldsymbol{\phi}) = \boldsymbol{Z}'\boldsymbol{Z} = \|\boldsymbol{Z}\|^2 = \|\boldsymbol{Y} - X\boldsymbol{\phi}\|^2.$$

Consider the Hilbert spaces

$$\mathcal{H} = \overline{\text{sp}}\{\boldsymbol{Y}, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{p-1}\} \quad \text{and} \quad \mathcal{M} = \overline{\text{sp}}\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{p-1}\}.$$

Since any element in $\mathcal{M}$ has the representation $X\boldsymbol{\phi}$ for some vector $\boldsymbol{\phi}$, it follows from the projection theorem that

$$P_{\mathcal{M}}\boldsymbol{Y} = X\boldsymbol{\phi}^*.$$

Thus we have

$$\langle \boldsymbol{x}_k, X\boldsymbol{\phi}^* \rangle = \langle \boldsymbol{x}_k, \boldsymbol{Y} \rangle \quad \text{for } k = 0, \ldots, p-1$$
$$\Updownarrow$$
$$X'X\boldsymbol{\phi}^* = X'\boldsymbol{Y}$$
$$\Updownarrow$$
$$\boldsymbol{\phi}^* = (X'X)^{-1}X'\boldsymbol{Y} \text{ provided } X'X \text{ is non-singular.}$$

It is easy to realize that $\widehat{\boldsymbol{\phi}} = \boldsymbol{\phi}^*$ if we put the "extra" observations $X_{-p+1}, \ldots, X_0 = 0$ (and $\overline{X}_n = 0$). Thus, since generally $p << n$, it it not too surprising that $\widehat{\boldsymbol{\phi}}$ has nice statistical properties.

Let now $\{X_t\}$ be a general ARMA(p, q) process with $q > 0$:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2).$$

The problem is that X_t is regressed not only onto $X_{t-1}, \ldots, X_{t-p}$ but also on the unobserved quantities $Z_{t-1}, \ldots, Z_{t-q}$. The main idea in the *Hannan–Rissanen algorithm* is to first replace $Z_{t-1}, \ldots, Z_{t-q}$ with their estimates $\widehat{Z}_{t-1}, \ldots, \widehat{Z}_{t-q}$ and then to estimate

$$\boldsymbol{\beta} \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{\phi} \\ \boldsymbol{\theta} \end{pmatrix}$$

by regressing X_t onto $X_{t-1}, \ldots, X_{t-p}, \widehat{Z}_{t-1}, \ldots, \widehat{Z}_{t-q}$. We will consider these ideas in some detail.

Step 1

A high order AR(m) model (with $m > \max(p, q)$) is fitted to the data by Yule-Walker estimation. If $\widehat{\phi}_{m1}, \ldots, \widehat{\phi}_{mm}$ are the estimated coefficients, then Z_t is estimated by

$$\widehat{Z}_t = X_t - \widehat{\phi}_{m1} X_{t-1} - \ldots - \widehat{\phi}_{mm} X_{t-m}, \quad t = m+1, \ldots, n.$$

Step 2

The vector $\boldsymbol{\beta}$ is estimated by least square regression of X_t onto

$$X_{t-1}, \ldots, X_{t-p}, \widehat{Z}_{t-1}, \ldots, \widehat{Z}_{t-q},$$

i.e. by minimizing

$$S(\boldsymbol{\beta}) = \sum_{t=m+1}^{n} (X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} - \theta_1 \widehat{Z}_{t-1} - \ldots - \theta_q \widehat{Z}_{t-q})^2$$

with respect to $\boldsymbol{\beta}$. This gives the Hannan–Rissanen estimator

$$\widehat{\boldsymbol{\beta}} = (Z'Z)^{-1} Z' \boldsymbol{X}_n \text{ provided } Z'Z \text{ is non-singular},$$

where

$$\boldsymbol{X}_n = \begin{pmatrix} X_{m+1} \\ \vdots \\ X_n \end{pmatrix}$$

and

$$Z = \begin{pmatrix} X_m & X_{m-1} & \ldots & X_{m-p+1} & \widehat{Z}_m & \widehat{Z}_{m-1} & \ldots & \widehat{Z}_{m-q+1} \\ \vdots & \vdots & & & & & & \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} & \widehat{Z}_{n-1} & \widehat{Z}_{n-2} & \ldots & \widehat{Z}_{n-q} \end{pmatrix}.$$

The Hannan–Rissanen estimate of the white noise variance σ^2 is

$$\widehat{\sigma}^2_{\mathrm{HR}} = \frac{S(\widehat{\boldsymbol{\beta}})}{n-m}.$$

8.1.5 Maximum Likelihood and Least Square estimation

It is possible to obtain better estimates by the maximum likelihood method (under the assumption of Gaussian processes) or by the least square method. In the least square method we minimize

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{j=1}^{n} \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}},$$

where $r_{j-1} = v_{j-1}/\sigma^2$, with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The estimates has to be obtained by recursive methods, and the estimates discussed are natural starting values. The least square estimate of σ^2 is

$$\widehat{\sigma}^2_{\mathrm{LS}} = \frac{S(\widehat{\boldsymbol{\phi}}_{\mathrm{LS}}, \widehat{\boldsymbol{\theta}}_{\mathrm{LS}})}{n-p-q},$$

where – of course – $(\widehat{\boldsymbol{\phi}}_{\mathrm{LS}}, \widehat{\boldsymbol{\theta}}_{\mathrm{LS}})$ is the estimate obtained by minimizing $S(\boldsymbol{\phi}, \boldsymbol{\theta})$.

Example 8.3 (MA(1) process) We have

$$X_1 = Z_1 + \theta Z_0 \quad \text{or} \quad Z_1 = X_1 - \theta Z_0$$

$$X_2 = Z_2 + \theta Z_1 \quad \text{or} \quad Z_2 = X_2 - \theta Z_1$$

$$\vdots$$

$$X_n = Z_n + \theta Z_{n-1} \quad \text{or} \quad Z_n = X_n - \theta Z_{n-1}$$

If we "know" that $Z_0 = 0$ we can calculate $Z_1, \ldots, Z_n$ for given θ. Since $\widehat{X}_k = \theta Z_{k-1}$ we have $r_j = \sigma^2$ and we can numerically minimize

$$\sum_{j=1}^{n} Z_j^2$$

with respect to θ. Denote the estimate by $\widehat{\theta}_n^{(2)}$. In this case we have

$$\widehat{\theta}_n^{(2)} \sim \text{AN}(\theta, (1-\theta^2)/n), \quad \text{for large values of } n.$$

□

In the general ARMA case we may recursively compute the the $\widehat{X}_j$s by the innovations algorithm discussed in Section 5.3.2 on page 44.

Let us now assume, or at least act as if, the process is Gaussian. Then, for any fixed values of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2, the innovations $X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n$ are independent and normally distributed with zero means and variances $v_0 = \sigma^2 r_0 = \gamma_X(0)$, $v_1 = \sigma^2 r_1, \ldots, v_{n-1} = \sigma^2 r_{n-1}$. Thus the density of $X_j - \widehat{X}_j$ is

$$f_{X_j - \widehat{X}_j}(x) = \frac{1}{\sqrt{2\pi\sigma^2 r_{j-1}}} \exp\left\{ -\frac{x^2}{2\sigma^2 r_{j-1}} \right\}.$$

The likelihood function, see Section 3.2.1 on page 23, is thus

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = \prod_{j=1}^{n} f_{X_j - \widehat{X}_j}(X_j - \widehat{X}_j)$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \cdots r_{n-1}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{(X_j - \widehat{X}_j)^2}{r_{j-1}} \right\}$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \cdots r_{n-1}}} \exp\left\{ -\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma^2} \right\}.$$

(Strictly speaking, we ought to write $x_j - \widehat{x}_j$ instead of $X_j - \widehat{X}_j$ in the formula above.)

Proceeding "in the usual way" we get

$$\ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = -\frac{1}{2} \ln((2\pi\sigma^2)^n r_0 \cdots r_{n-1}) - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma^2}.$$

Obviously $r_0, \ldots, r_{n-1}$ depend on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ but they do not depend on σ^2. For fixed values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ we get

$$\frac{\partial \ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2(\sigma^2)^2},$$

$\ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ is maximized by $\sigma^2 = n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta})$. Thus we get

$$\begin{aligned}
\ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta})) &= -\frac{1}{2} \ln((2\pi n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta}))^n r_0 \cdots r_{n-1}) - \frac{n}{2} \\
&= -\frac{1}{2} \left(n \ln(2\pi) + n \ln(n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta})) + \ln r_0 + \ldots + \ln r_{n-1} \right) - \frac{n}{2} \\
&= -\frac{n}{2} \left(\ln(n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta})) + n^{-1} \sum_{j=1}^{n} \ln r_{j-1} \right) + \text{constant}.
\end{aligned}$$

Thus to maximize $\ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ is the same as to minimize

$$\ell(\boldsymbol{\phi}, \boldsymbol{\theta}) = \ln(n^{-1} S(\boldsymbol{\phi}, \boldsymbol{\theta})) + n^{-1} \sum_{j=1}^{n} \ln r_{j-1},$$

which has to be done numerically.

In the causal and invertible case $r_n \to 1$ and therefore $n^{-1} \sum_{j=1}^{n} \ln r_{j-1}$ is asymptotically negligible compared with $\ln S(\boldsymbol{\phi}, \boldsymbol{\theta})$. Thus both methods – least square and maximum likelihood – give asymptotically the same result in that case.

8.1.6 Order selection

Let us assume that we have a situation where reality really is described by an ARMA(p, q) process. In order to make the discussion simple we further assume that $q = 0$, i.e. that reality is described by an AR(p) process. This fact we regard as "known", but the order is unknown. A natural approach would be to try with AR(m) models for increasing values of m. For each m we may calculate $S(\widehat{\boldsymbol{\phi}})$ or $L(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}^2)$ or some other measure which tells about the realism of the model. In this situation one would expect $S(\widehat{\boldsymbol{\phi}})$ to decrease with m as long as $m \leq p$ and then to remain more or less constant. Similarly $L(\widehat{\boldsymbol{\phi}}, \widehat{\sigma}^2)$ ought to increase for $m \leq p$. However, the situation described is by no means realistic. In (almost) every situation reality is more complex than any simple parametric model.

Assume now that we, in a somewhat more realistic situation than the one described above, want to fit an ARMA(p, q) process to real data, i.e. we want to estimate p, q, $(\boldsymbol{\phi}, \boldsymbol{\theta})$, and σ^2. We restrict ourselves to maximum likelihood estimation. Then we maximize $L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$, or – which is the same – minimize $-2 \ln L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$, where L is regarded as a function also of p and q. Most probably we will get very high values of p and q. Such a model will probably fit the given data very well, but it is more or less useless as a mathematical

model, since it will probably not be lead to reasonable predictors nor describe a different data set well. It is therefore natural to introduce a "penalty factor" to discourage the fitting of models with too many parameters. Instead of maximum likelihood estimation we may apply the *AICC Criterion*:

Choose p, q, and $(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q)$, to minimize

$$\text{AICC} = -2 \ln L(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q, S(\boldsymbol{\phi}_p, \boldsymbol{\theta}_q)/n) + 2(p+q+1)n/(n-p-q-2).$$

(The letters AIC stand for "Akaike's Information Criterion" and the last C for "biased-Corrected".)

The AICC Criterion has certain nice properties, but also its drawbacks. If data really are described by an ARMA(p, q) process, one would like the resulting estimates $\widehat{p}$ and $\widehat{q}$ to be at least be consistent, i.e. that

$$\widehat{p} \xrightarrow{\text{a.s.}} p \text{ and } \widehat{q} \xrightarrow{\text{a.s.}} q \text{ as } n \to \infty.$$

(The notation "$\xrightarrow{\text{a.s.}}$" means "almost sure convergence" or "convergence with probability one"; a notion discussed in Section A.2 on page 115.) This is, however, not the case for estimates obtained by the AICC Criterion. There certain other criteria discussed in [7], as for instance the BIC, which is a consistent order selection criterion.

In general one may say the order selection is genuinely difficult. Many criteria have been proposed, but there exists no canonical criterion. We will here only mention Rissanen's *minimum description length* (MDL) criterion, which seems be rather much used. That criterion states that a model should be sought that allows the shortest possible description of the observed data.

Lecture 9

9.1 Unit roots

We will discuss some questions related to the existence of roots of the generating polynomials on or near the unit circle. The discussion is based on Sections 6.1, 6.3, and 10.5 in [7].

Recall from Section 1.3.1 on page 4 that differencing is a way to generate stationarity.

Let $\{X_t,\, t \in \mathbb{Z}\}$ be a time series and consider

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1},$$

where B is the backward shift operator, i.e. $(BX)_t = X_{t-1}$, or more generally

$$\nabla^d X_t = (1 - B)^d X_t.$$

Assume that $\nabla^d X_t$ is not only stationary, but in fact a causal ARMA(p, q) process.

Definition 9.1 (The ARIMA(p, d, q) process) *Let d be a non-negative integer. The process $\{X_t,\, t \in \mathbb{Z}\}$ is said to be an* ARIMA(p, d, q) process *if $\nabla^d X_t$ is a causal* ARMA(p, q) *process.*

Definition 9.1 means that $\{X_t\}$ satisfies

$$\phi^*(B)X_t = \phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2), \tag{9.1}$$

where $\phi(z) \neq 0$ for all $|z| \leq 1$, see Theorem 2.3 on page 15. For $d \geq 1$ there exists no stationary solution of (9.1) and further neither the mean nor the covariance function are determined by (9.1). If X_t is an ARIMA$(p, 1, q)$ process it satisfies the same difference equations as the process $X_t + Y$, since the random variable Y disappears by differencing. Since Y may have any mean, variance, and covariance relation to $\{X_t\}$ it follows that (9.1) does not determine the mean and the covariance function. This is no problem for the estimation of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2 but for prediction additional assumptions are needed.

Recall from Example 2.1 on page 15 that a causal AR(1) process has autocovariance function (ACVF)

$$\gamma(h) = \frac{\sigma^2 \phi^{|h|}}{1 - \phi^2}, \quad |\phi| < 1.$$

This ACVF decreases (rather) slowly for ϕ close to one. Similarly it holds for any ARMA process that its ACVF decreases slowly if some of the roots of $\phi(z) = 0$ are near the unit circle. In practice, i.e. from a sample of finite length, it is very difficult to distinguish between an ARIMA$(p, 1, q)$ process and an ARMA$(p+1, q)$ with a root of $\phi(z) = 0$ near the unit circle. Therefore a slowly decreasing (estimated) ACVF indicates that *differencing might be advisable*.

Assume now that X_t in fact is a causal and invertible ARMA(p, q) process, i.e.

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

where $\theta(z) \neq 0$ for all $|z| \leq 1$, see Theorem 2.4 on page 16. For some reason this process is differenced. Since

$$\phi(B)\nabla X_t = \phi(B)(1-B)X_t = \theta(B)(1-B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2),$$

it follows that ∇X_t is a causal but non-invertible ARMA$(p, q+1)$ process. Thus a unit root in the moving average polynomial indicates that X_t has been *overdifferenced*.

Up to now we have regarded a slowly decreasing ACVF as an indication of non-stationarity, but naturally we may have a situation where a slowly decreasing ACVF really indicates "long memory" merely than non-stationarity. One may then be tempted to use an ARMA process with roots near the unit circle. However, it can be shown that the ACVF of an ARMA process is geometrically bounded, i.e. that

$$|\gamma(h)| \leq Cr^{|h|}, \quad \text{for all } h,$$

where $C > 0$ and $0 < r < 1$, cf. the ACVF for an AR(1) process. A first idea might be to – in some way – let a root tend to unity, but since we cannot allow for roots on the unit circle this idea seems difficult to transfer to mathematics. In principle the idea is not bad, and we are led to consider "fractionally integrated ARMA processes".

Definition 9.2 (The FARIMA(p, d, q) process) *Let $0 < |d| < 0.5$. The process $\{X_t, t \in \mathbb{Z}\}$ is said to be a* fractionally integrated ARMA process *or a* FARIMA(p, d, q) process *if $\{X_t\}$ is stationary and satisfies*

$$\phi(B)(1-B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

The operator $(1-B)^d$ is defined by the binomial expansion

$$(1-B)^d = \sum_{j=0}^{\infty} \binom{d}{j} (-1)^j B^j,$$

where

$$\binom{d}{0} = 1 \quad \text{and} \quad \binom{d}{j} = \prod_{k=1}^{j} \frac{d-k+1}{k}, \quad j = 1, 2, \ldots$$

(In [7] a FARIMA(p, d, q) process is called an ARIMA(p, d, q) process. Both terms can be found in the literature, and we prefer the used term since there is less risk for misunderstandings.)

It can be shown that the ACVF of a FARIMA process has the property

$$\gamma(h)h^{1-2d} \to c, \quad h \to \infty,$$

where $c > 0$, provided that X_t is a causal and invertible. Due to this, a FARIMA model said to be a *long memory model.*

9.2 Multivariate time series

Not too seldom it is natural to consider several time series at the same time in the same way as it often is natural to consider several random variables simultaneously. In that case we talk about *multivariate time series*. Let

$$\boldsymbol{X}_t \stackrel{\text{def}}{=} \begin{pmatrix} X_{t1} \\ \vdots \\ X_{tm} \end{pmatrix}, \quad t \in \mathbb{Z},$$

where each component is a time series. We will here only give the basic definitions; a discussion based on Sections 7.2 and 7.4 in [7]. The second-order properties of $\{\boldsymbol{X}_t\}$ are specified by the mean vector

$$\boldsymbol{\mu}_t \stackrel{\text{def}}{=} E\boldsymbol{X}_t = \begin{pmatrix} \mu_{t1} \\ \vdots \\ \mu_{tm} \end{pmatrix} = \begin{pmatrix} EX_{t1} \\ \vdots \\ EX_{tm} \end{pmatrix}, \quad t \in \mathbb{Z},$$

and the covariance matrices

$$\Gamma(t+h,t) \stackrel{\text{def}}{=} E[(\boldsymbol{X}_{t+h}-\boldsymbol{\mu}_{t+h})(\boldsymbol{X}_t-\boldsymbol{\mu}_t)'] = \begin{pmatrix} \gamma_{11}(t+h,t) & \dots & \gamma_{1m}(t+h,t) \\ \vdots & & \\ \gamma_{m1}(t+h,t) & \dots & \gamma_{mm}(t+h,t) \end{pmatrix}$$

where $\gamma_{ij}(t+h,t) \stackrel{\text{def}}{=} \text{Cov}(X_{t+h,i}, X_{t,j})$.

Most definitions in the univariate (usual) case have their natural counterparts in the multivariate case.

The following definition of stationarity is almost word for word the same as Definition 1.4 on page 2.

Definition 9.3 *The m-variate time series $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ is said to be* (weakly) stationary *if*

(i) $\boldsymbol{\mu}_t = \boldsymbol{\mu}$ *for all* $t \in \mathbb{Z}$,

(ii) $\Gamma(r,s) = \Gamma(r+t, s+t)$ *for all* $r, s, t \in \mathbb{Z}$.

Item (ii) implies that $\Gamma(r,s)$ is a function of $r-s$, and it is convenient to define

$$\Gamma(h) \stackrel{\text{def}}{=} \Gamma(h,0).$$

The multivariate ACVF has the following properties, cf. Section 2.1 on page 9:

(i) $\Gamma(h) = \Gamma'(-h)$,

(ii) $|\gamma_{ij}(h)| \leq \sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}$.

(iii) $\gamma_{ii}(\cdot)$ is an (univariate) ACVF,

(iv) $\sum_{i,j=1}^{n} \boldsymbol{a}_i'\Gamma(i-j)\boldsymbol{a}_j \geq 0$ for all n and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^n$.

Property (i) implies that $\gamma_{ij}(h) = \gamma_{ji}(-h)$.

Let $\boldsymbol{X} = (X_{t1}, X_{t2})'$ be stationary and consider the complex-valued time series $X_t = X_{t1} + iX_{t2}$. Certainly X_t is stationary in the sense of Definition 6.1 on page 48. However, stationarity in the sense of Definition 6.1 does not imply that $(\operatorname{Re} X, \operatorname{Im} X)'$ is a 2-variate stationary time series. Let Y be a random variable with $E[Y] = 0$ and $\operatorname{Var}[Y] < \infty$ and consider $X_t = e^{it}Y$ which is stationary in the sense of Definition 6.1. We have

$$X_t = \cos(t)Y + i\sin(t)Y,$$

where neither $\cos(t)Y$ nor $\sin(t)Y$ are stationary.

The following definition is the correspondence of Definition 1.6 on page 3.

Definition 9.4 (Multivariate white noise) *An m-variate process*

$$\{\boldsymbol{Z}_t,\, t \in \mathbb{Z}\}$$

is said to be a white noise with mean $\boldsymbol{\mu}$ and covariance matrix Σ, written

$$\{\boldsymbol{Z}_t\} \sim \mathrm{WN}(\boldsymbol{\mu}, \Sigma),$$

if $E\boldsymbol{Z}_t = \boldsymbol{\mu}$ *and* $\Gamma(h) = \begin{cases} \Sigma & \text{if } h = 0, \\ 0 & \text{if } h \neq 0. \end{cases}$

Having Definition 9.4 in mind, the following correspondence to Definition 2.7 on page 14 is hardly surprising.

Definition 9.5 (The ARMA(p,q) process) *The process $\{\boldsymbol{X}_t,\, t \in \mathbb{Z}\}$ is said to be an* ARMA(p,q) process *if it is stationary and if*

$$\boldsymbol{X}_t - \Phi_1\boldsymbol{X}_{t-1} - \ldots - \Phi_p\boldsymbol{X}_{t-p} = \boldsymbol{Z}_t + \Theta_1\boldsymbol{Z}_{t-1} + \ldots + \Theta_q\boldsymbol{Z}_{t-q}, \tag{9.2}$$

where $\{\boldsymbol{Z}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \Sigma)$. We say that $\{\boldsymbol{X}_t\}$ is an ARMA(p,q) *process with mean $\boldsymbol{\mu}$ if $\{\boldsymbol{X}_t - \boldsymbol{\mu}\}$ is an* ARMA(p,q) *process.*

Equations (9.2) can be written as

$$\Phi(B)\boldsymbol{X}_t = \Theta(B)\boldsymbol{Z}_t, \quad t \in \mathbb{Z},$$

where

$$\Phi(z) = I - \Phi_1 z - \ldots - \Phi_p z^p,$$

$$\Theta(z) = I + \Theta_1 z + \ldots + \Theta_q z^q,$$

are matrix-valued polynomials.

A little less obvious is perhaps how causality and invertibility are characterized in terms of the generating polynomials:

Causality: $\boldsymbol{X}_t$ is causal if $\det \Phi(z) \neq 0$ for all $|z| \leq 1$;

Invertibility: $\boldsymbol{X}_t$ is invertible if $\det \Theta(z) \neq 0$ for all $|z| \leq 1$.

Now we will consider spectral properties. Like in the univariate case we will do this in some more details than done in [7].

Assume first that

$$\sum_{h=-\infty}^{\infty} |\gamma_{ij}(h)| < \infty, \quad i, j = 1, \ldots, m. \tag{9.3}$$

Definition 9.6 (The cross spectrum) *Let $\{\boldsymbol{X}_t, t \in \mathbb{Z}\}$ be an m-variate stationary time series whose ACVF satisfies* (9.3)*. The function*

$$f_{jk}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda} \gamma_{jk}(h), \quad -\pi \leq \lambda \leq \pi, \quad j \neq k,$$

is called the cross spectrum *or* cross spectral density *of $\{X_{tj}\}$ and $\{X_{tk}\}$. The matrix*

$$f(\lambda) = \begin{pmatrix} f_{11}(\lambda) & \ldots & f_{1m}(\lambda) \\ \vdots & & \\ f_{m1}(\lambda) & \ldots & f_{mm}(\lambda) \end{pmatrix}$$

is called the spectrum *or* spectral density matrix *of $\{\boldsymbol{X}_t\}$.*

By direct calculations, cf. Section 2.1.2 on page 11, it follows that

$$\Gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda.$$

The function $f_{jj}(\lambda)$ is the spectral density of $\{X_{tj}\}$ and therefore non-negative and symmetric about zero. However, since $\gamma_{ij}(\cdot)$, $i \neq j$, is not in general symmetric about zero, the cross spectral density is typically complex-valued. The spectral density matrix $f(\lambda)$ is non-negative definite for all $\lambda \in [-\pi, \pi]$.

We will now consider the spectral representation of $\Gamma(\cdot)$ when $\sum_{h=-\infty}^{\infty} |\gamma_{ij}(h)| < \infty$ is not assumed and the spectral representation of $\boldsymbol{X}_t$ itself, corresponding to (6.4) on page 51.

Theorem 9.1 *$\Gamma(\cdot)$ it the ACVF of an m-variate stationary time series $\{\boldsymbol{X}_t,\, t \in \mathbb{Z}\}$ if and only if*

$$\Gamma(h) = \int_{(-\pi,\pi]} e^{ih\lambda}\, dF(\lambda), \quad h \in \mathbb{Z},$$

where $F(\cdot)$ is an $m \times m$ matrix distribution, i.e. $F_{jk}(-\pi) = 0$, $F_{jk}(\cdot)$ is right-continuous and $F(\mu) - F(\lambda)$ is non-negative definite for all $\lambda \leq \mu$.

Similarly $\{\boldsymbol{X}_t - \boldsymbol{\mu}\}$ has the representation, cf. (6.4),

$$\boldsymbol{X}_t - \boldsymbol{\mu} = \int_{(-\pi,\pi]} e^{it\lambda}\, d\boldsymbol{Z}(\lambda)$$

where $\{\boldsymbol{Z}(\lambda),\, \lambda \in [-\pi, \pi]\}$ is an m-variate process whose components are complex-valued satisfying

$$E[dZ_j(\lambda)\overline{dZ_k(\nu)}] = \begin{cases} dF_{jk}(\lambda) & \text{if } \lambda = \nu, \\ 0 & \text{if } \lambda \neq \nu. \end{cases}$$

Lecture 10

10.1 Financial time series

Financial time series data, like the relative return of a stock (avkastning av en aktie) or a portfolio of stocks, often consist of periods of "calm" behaviour alternating with periods of very wild fluctuations. One way to express this is the following quotation, taking from [11]:

> *... large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes ...*

In general, the fluctuations or the difficulty to predict a future value of a stock or some other asset is a measure of how risky the asset is. In financial terms this is called the *volatility* of the asset. The celebrated *Black-Scholes formula* for option pricing is partly based on the volatility. An option is a contract giving the right or demand to sell or buy a certain asset at a future time to a specified price.

In this section we will discuss financial time series in some more detail than done in [7]. The reason is mainly that we believe this field to be of interest to students here, as shown by the success of the course "Stochastic Calculus and the Theory of Capital Markets" (Stokastisk kalkyl och kapitalmarknadsteori). A second reason is that it seems as if a person with a sound mathematical and probabilistic background can make important contributions to this field. Therefore we will give more references here than in other sections.

Let $\{X_t,\, t \in \mathbb{Z}\}$ be a "financial time series" and assume that the mean and a possible trend already is withdrawn, so that we may assume that $\{X_t\}$ is stationary. Often it seems as if $\{X_t\}$ is almost WN, but it is, cf. the discussion above, far from IID. A popular way of modeling these kind of processes is by

$$X_t = \sigma_t Z_t, \qquad \{Z_t\} \sim \text{IID}\, N(0,1), \tag{10.1}$$

where the "stochastic volatility" σ_t is a function of $X_{t-1}, X_{t-2}, \ldots$ and $\{Z_t\}$ is a Gaussian white noise. We further assume that Z_t and $X_{t-1}, X_{t-2}, \ldots$ are independent for all t. This may be regarded as an assumption of causality.

Assume that, and this is not at all sure, there exists a time series fulfilling (10.1). Then we have $E[X_t] = E[\sigma_t]E[Z_t] = 0$, provided that $E[\sigma_t] < \infty$, and

$$\text{Cov}[X_s, X_t] = \begin{cases} \text{Var}[X_t] = E[X_t^2] = E[\sigma_t^2]E[Z_t^2] = E[\sigma_t^2], & \text{if } s = t, \\ E[X_s X_t] = E[\sigma_s Z_s \sigma_t Z_t] = E[\sigma_s Z_s \sigma_t]E[Z_t] = 0, & \text{if } s < t. \end{cases}$$

Thus $\{X_t\}$ is WN provided that $E[\sigma_t] < \infty$, which need not to hold.

Consider now the time series $\{X_t^2\}$. In this general setting it is not so easy to say much about $\{X_t^2\}$ other than it is far from a white noise. Let

$$\widetilde{Z}_t = X_t^2 - \sigma_t^2 = \sigma_t^2 \cdot (Z_t^2 - 1). \tag{10.2}$$

Using (A.2) on page 114 we get

$$E\exp(iaZ_t) = 1 + iaE[Z_t] + \ldots + \frac{a^4}{4!}E[Z_t^4] + \ldots = 1 - \frac{a^2}{2 \cdot 1!} + \frac{a^4}{2^2 \cdot 2!} + \ldots$$

and thus $E[Z_t^4] = \frac{4!}{2^2 \cdot 2!} = 3$ follows. Assume that $E[\sigma_t^4] < \infty$. Then we get

$$E[\widetilde{Z}_t] = E[\sigma_t^2]E[Z_t^2 - 1] = E[\sigma_t^2] \cdot (1 - 1) = 0$$

and

$$\mathrm{Cov}[\widetilde{Z}_s, \widetilde{Z}_t] = \begin{cases} E[\sigma_t^4]E[(Z_t^2 - 1)^2] = E[\sigma_t^4]E[Z_t^4 - 2Z_t^2 + 1] = 2E[\sigma_t^4] & \text{if } s = t, \\ E[\sigma_s^2\sigma_t^2(Z_s^2 - 1)]E[Z_t^2 - 1] = E[\ldots] \cdot 0 = 0 & \text{if } s < t. \end{cases}$$

Thus $\{\widetilde{Z}_t\}$ is a white noise.

10.1.1 ARCH processes

The first model of stochastic volatility was the ARCH(p) process, introduced in [9]. The shortening ARCH stands for *autoregressive conditional heteroscedasticity*.

Definition 10.1 (The ARCH(p) process) *The process $\{X_t,\, t \in \mathbb{Z}\}$ is said to be an* ARCH(p) process *if it is stationary and if*

$$X_t = \sigma_t Z_t, \qquad \{Z_t\} \sim \text{IID } N(0,1),$$

where

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 \tag{10.3}$$

and $\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, p$, and if Z_t and $X_{t-1}, X_{t-2}, \ldots$ are independent for all t.

The requirements $\alpha_0 > 0$ and $\alpha_j \geq 0$ guarantee that $\sigma_t > 0$. It is, however, not at all easy to find conditions on α_0 and α_j which ascertain that there really exists an ARCH(p) process.

Example 10.1 (The ARCH(1) process) Let $\{X_t\}$ be an ARCH(1) process with $0 \leq \alpha_1 < 1$, i.e. we have

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2. \tag{10.4}$$

Using $E[X_t^2] = E[\sigma_t^2]$ and stationarity we get

$$E[X_t^2] = \alpha_0 + \alpha_1 E[X_t^2] \Rightarrow E[X_t^2] = \frac{\alpha_0}{1-\alpha_1}.$$

Combining (10.2) and (10.4) we get

$$X_t^2 - \widetilde{Z}_t = \alpha_0 + \alpha_1 X_{t-1}^2,$$

and thus

$$X_t^2 - \frac{\alpha_0}{1-\alpha_1} = \alpha_1 \cdot \left(X_{t-1}^2 - \frac{\alpha_0}{1-\alpha_1} \right) + \widetilde{Z}_t.$$

This implies that $\{X_t^2\}$ is a causal AR(1) process with mean $\alpha_0/(1-\alpha_1)$, cf. Definition 2.8 on page 14, provided that $E[\widetilde{Z}_t^2] < \infty$.

It can be shown that $E[\widetilde{Z}_t^2] < \infty$ if and only if $\alpha_1^2 < 1/3$. In order to indicate this we notice that

$$E[X_t^4] = E[(\alpha_0 + \alpha_1 X_{t-1}^2)^2] E[Z_t^4] = 3E[\alpha_0^2 + 2\alpha_0\alpha_1 X_t^2 + \alpha_1^2 X_t^4]$$

$$= 3\left(\alpha_0^2 + \frac{2\alpha_0^2\alpha_1}{1-\alpha_1} \right) + 3\alpha_1^2 E[X_t^4] = \frac{3\alpha_0^2(1+\alpha_1)}{1-\alpha_1} + 3\alpha_1^2 E[X_t^4].$$

If $3\alpha_1^2 \geq 1$ it follows that $E[X_t^4] = \infty$. For $3\alpha_1^2 < 1$ a solution is

$$E[X_t^4] = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

For $\alpha_1 = 0$ the solution above reduces to $3\alpha_0^2$, which obviously is the correct value of $E[X_t^4]$ in that case. Although we have not at all proved that, it seems reasonable that the solution gives the correct value for all $\alpha_1 < 1/\sqrt{3}$. Then

$$E[\widetilde{Z}_t^2] = 2E[\sigma_t^4] = 2E[\alpha_0^2 + 2\alpha_0\alpha_1 X_t^2 + \alpha_1^2 X_t^4]$$

$$= \frac{2\alpha_0^2(1+\alpha_1)}{1-\alpha_1} + \frac{6\alpha_0^2\alpha_1^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)} = \frac{2\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

This relation between the noise and the process does of course also follow from (2.11) on page 15.

However, in reality we are interested in $\{X_t\}$ rather than in $\{X_t^2\}$. It can be shown that $\{X_t\}$ is weakly stationary if and only if $\alpha_1 < 1$.

A somewhat surprising result may be that $\{X_t\}$ is strictly stationary if and only if $E[\ln(\alpha_1 Z_t^2)] < 0$. The condition $E[\ln(\alpha_1 Z_t^2)] < 0$ is equivalent with

$$\alpha_1 < 2e^{\gamma} \approx 3.56856.$$

The constant γ in the formula above is the Euler constant. As a consequence it follows that

- for $\alpha_1 = 0$, $\{X_t\}$ is Gaussian white noise;
- for $0 < \alpha_1 < 1$, $\{X_t\}$ is stationary with finite variance;

- for $1 \leq \alpha_1 < 2e^{\gamma}$, $\{X_t\}$ is stationary with infinite variance.

Recall that $\{X_t\}$ is WN also for $0 < \alpha_1 < 1$, but not IID. (We have here restricted us to the case with finite variance in order to avoid discussing whether a WN may have infinite variance.) One might have believed that a strictly stationary WN must be IID, but that is thus *not* the case. □

Consider now an ARCH(p) process and the polynomial

$$\alpha(z) = \alpha_1 z + \ldots + \alpha_p z^p.$$

Notice that α_0 is not involved in $\alpha(z)$.

Thus (10.3) may be written on the form

$$\sigma_t^2 = \alpha_0 + \alpha(B)X_t^2.$$

Using, as in Example 10.1, $E[X_t^2] = E[\sigma_t^2]$ and stationarity we get

$$E[X_t^2] = \alpha_0 + \alpha(1)E[X_t^2] \Rightarrow E[X_t^2] = \frac{\alpha_0}{1 - \alpha(1)}\,.$$

It can be shown that the X_t^2, like in the ARCH(1) case, is an AR process. We will not discuss ARCH(p) processes any further, but instead consider a generalization of them. The reason for that generalization is that in practice the order p has to be rather large.

10.1.2 GARCH processes

Numerous parametric specifications for the conditional variance have been proposed. The most important extension of the ARCH process is certainly the *generalized* ARCH, or GARCH, process proposed in [2].

Definition 10.2 (The GARCH(p, q) process) *The process $\{X_t,\, t \in \mathbb{Z}\}$ is said to be an* GARCH(p, q) process *if it is stationary and if*

$$X_t = \sigma_t Z_t, \qquad \{Z_t\} \sim \text{IID } N(0, 1),$$

where

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2 \tag{10.5}$$

and $\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, p$, $\beta_k \geq 0$ for $k = 1, \ldots, q$, and if Z_t and $X_{t-1}, X_{t-2}, \ldots$ are independent for all t.

It seems as the GARCH(1, 1) process often is regarded to be a reasonably realistic model. In spite of that we will shortly consider the general case.

Equation (10.5) can be written as

$$\sigma_t^2 = \alpha_0 + \alpha(B)X_t^2 + \beta(B)\sigma_t^2 \tag{10.6}$$

where

$$\alpha(z) = \alpha_1 z + \ldots + \alpha_p z^p,$$

$$\beta(z) = \beta_1 z + \ldots + \beta_q z^q.$$

Using, as twice before, $E[X_t^2] = E[\sigma_t^2]$ and stationarity we get

$$E[X_t^2] = \alpha_0 + (\alpha(1) + \beta(1))E[X_t^2] \Rightarrow E[X_t^2] = \frac{\alpha_0}{1 - \alpha(1) - \beta(1)}\,.$$

Assume that $E[\sigma_t^4] < \infty$ and recall (10.2) on page 96. We get

$$X_t^2 - \widetilde{Z}_t = \alpha_0 + \alpha(B)X_t^2 + \beta(B)(X_t^2 - \widetilde{Z}_t)$$

or

$$X_t^2 - (\alpha(B) + \beta(B))X_t^2 = \alpha_0 + (1 - \beta(B))\widetilde{Z}_t.$$

Since

$$\frac{\alpha_0}{1 - \alpha(1) - \beta(1)} - \frac{\alpha_0(\alpha(1) + \beta(1))}{1 - \alpha(1) - \beta(1)} = \alpha_0$$

it follows that $\{X_t^2\}$ is an ARMA$(\max(p, q), q)$ process with generating polynomials, cf. Definition 2.7 on page 14,

$$\phi(z) = 1 - \alpha(z) - \beta(z) \quad \text{and} \quad \theta(z) = 1 - \beta(z)$$

and mean $\alpha_0/(1 - \alpha(1) - \beta(1))$.

10.1.3 Further extensions of the ARCH process

As mentioned, numerous parametric specifications for the conditional variance have been proposed. We will here list some of them, without too much comments. Sometimes the shortening for the models seems as creative as the models themselves.

Non-linear ARCH (NARCH) processes

$$\sigma_t^\gamma = \alpha_0 + \alpha(B)|X_t|^\gamma + \beta(B)\sigma_t^\gamma$$

The natural choices are $\gamma = 2$, yielding GARCH, and $\gamma = 1$.

We may let this generalization illustrate a "standard" approach: Consider some sets of financial data. For each set the hypothesis

H_0 : the data is described by a GARCH process, i.e. $\gamma = 2$,

against the alternative

H_1 : the data is described by a NARCH process, but not by GARCH, i.e. $\gamma \neq 2$.

These kind of tests are often called "specification" tests, and the result is generally that H_0 is rejected. The creator of the generalization is then happy, and some papers are written. The typical situation is that we have some further parameters to play with in $H_0 \cup H_1$. Since most – all – models are simplifications of reality, an H_0 which is much smaller than $H_0 \cup H_1$ is (almost) always rejected if the data set is large enough.

ARCH-t processes

A slightly different kind of extension is to let the underlying noise $\{Z_t\}$ be IID $t(f)$, i.e. to let Z_t be $t(f)$-distributed. Notice that "$t(\infty) = N(0,1)$". We then have the degree of freedom f to play with. The idea is that the t-distribution has heavier tails than the normal distribution.

Asymmetric ARCH processes

One drawback with the above models is that positive and negative past values have a symmetric effect on the volatility. Many financial time series are, however, strongly asymmetric. Negative returns are followed by larger increases in the volatility than equally large positive returns. Typical examples of this may be prices of petrol or the interest of a loan. Let

$$X_t^+ \stackrel{\text{def}}{=} \max(X_t, 0) \quad \text{and} \quad X_t^- \stackrel{\text{def}}{=} \min(X_t, 0)$$

and notice that $X_t = X_t^+ + X_t^-$ and $|X_t| = X_t^+ - X_t^-$.

For simplicity we consider extensions of the GARCH$(1,1)$ process. In all models below we assume that $\{Z_t\}$ be IID $N(0,1)$.

Exponential GARCH, EGARCH

The first proposed asymmetric model was the EGARCH model:

$$\ln \sigma_t^2 = \alpha_0 + \beta \ln \sigma_{t-1}^2 + \lambda Z_{t-1} + \varphi \cdot (|Z_{t-1}| - E[|Z_{t-1}|]).$$

Notice that $Z_t = X_t/\sigma_t$ and that $E[|Z_t|] = \sqrt{2/\pi}$.

Quadratic GARCH, QGARCH

$$\sigma_t^2 = \alpha_0 + \zeta X_{t-1} + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2.$$

The GJR model

$$\sigma_t^2 = \alpha_0 + \alpha X_{t-1}^2 - \omega \cdot (X_{t-1}^-)^2 + \beta \sigma_{t-1}^2.$$

The shortening GJR stands for Glosten, Jagannathan and Runkle who proposed the model.

Threshold GARCH, TGARCH

$$\sigma_t = \alpha_0 + \alpha^+ X_{t-1}^+ - \alpha^- X_{t-1}^- + \beta \sigma_{t-1}.$$

Logistic smooth transition GARCH, LSTGARCH

$$\sigma_t^2 = \alpha_0 + \big(\alpha_1 + \alpha_2 F(X_t)\big) X_{t-1}^2 + \beta \sigma_{t-1}^2,$$

where

$$F(x) = \frac{1}{1+e^{-\theta x}} - \frac{1}{2}, \quad \theta > 0.$$

Some variants of asymmetric models

There have been many further models proposed; to some of them we have not even found any shortening or name. Naturally the GJR and the TGARCH models may be, and have been, generalized to

$$\sigma_t^\gamma = \alpha_0 + \alpha^+(X_{t-1}^+)^\gamma - \alpha^-(X_{t-1}^-)^\gamma + \beta\sigma_{t-1}^\gamma.$$

All asymmetric models listed allow for different reactions of the volatility for "good" or "bad" news, but maintains the assertion that the minimum volatility will result when there are no news. The following modification of the NARCH model allows the minimum volatility to occur more generally:

$$\sigma_t^\gamma = \alpha_0 + \alpha|X_{t-1} - \kappa|^\gamma + \beta\sigma_{t-1}^\gamma.$$

10.1.4 Literature about financial time series

Since the appearance of [9] about 200 papers have been devoted to ARCH processes and its extensions. The interested reader is recommended to have a look at the survey papers [3], [4], and [12]. The LSTGARCH model is studied in [10].

Looking in the reference lists of the survey papers mentioned above, it is obvious that the development of financial time series has lived its own life. One hardly finds any references to the "standard" time series literature, and in that literature financial time series are hardly mentioned. The book [7] is in fact an exception. However, we do believe that this is going to be changed. It is further difficult to avoid the feeling that the number of papers written about financial time series and the number of genuine new ideas differ very much.

A nice discussion about financial time series is, however, to be found in [8], which is primarily not a book about time series analysis. We conclude the discussion with the following quotation from that book:[1]

> *None of the above models is really testable. Thus all of them have a right of existence as long as they explain certain phenomena of the real world. By now none of the discrete time models has been accepted as* the model *for stochastic finance. It would also be a great surprise if one single equation or a system of equations could explain the complicated nature of the financial world. This calls for further research and deeper methods.*

[1] Strictly speaking, the quotation is taken from a preliminary version of the book and is not to be found in the printed version. According to one of the authors it was taken away for "diplomatic" reasons.

Lecture 11

11.1 Kalman filtering

The intension with this lecture is to give some glimpses about Kalman filtering, which is a very fruitful approach in for instance control theory. Kalman filtering is treated in the course "Mathematical System Theory" (Matematisk systemteori). One of the important features of that approach is that it does not require stationarity. Our discussion is based on Sections 8.1, 8.3, and 8.4 in [7].

11.1.1 State-Space representations

Like for ARMA and ARIMA processes the processes are driven by white noise. The processes to be considered will be allowed to be multivariate. We will use the notation

$$\{\boldsymbol{Z}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{\Sigma_t\}),$$

to indicate that the process $\{\boldsymbol{Z}_t\}$ has mean $\boldsymbol{0}$ and that

$$E\boldsymbol{Z}_s\boldsymbol{Z}_t' = \begin{cases} \Sigma_t & \text{if } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that this definition is an extension of Definition 9.4 on page 92 in order to allow for non-stationarity.

In control theory the main interest is focused around the *state equation*

$$\boldsymbol{X}_{t+1} = F_t\boldsymbol{X}_t + \boldsymbol{V}_t, \quad t = 1, 2, \ldots, \tag{11.1}$$

where $\{\boldsymbol{X}_t\}$ is a v-variate process describing the state of some system, $\{\boldsymbol{V}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{Q_t\})$, and $\{F_t\}$ is a sequence of $v \times v$ matrices.

Often a system is complicated and the state of it cannot be exactly observed. The observations are described by the *observation equation*

$$\boldsymbol{Y}_t = G_t\boldsymbol{X}_t + \boldsymbol{W}_t, \quad t = 1, 2, \ldots, \tag{11.2}$$

where $\{\boldsymbol{Y}_t\}$ is a w-variate process describing the observed state of some system, $\{\boldsymbol{W}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{R_t\})$, and $\{G_t\}$ is a sequence of $w \times v$ matrices. Further $\{\boldsymbol{W}_t\}$ and $\{\boldsymbol{V}_t\}$ are uncorrelated. To complete the specification it is assumed that the initial state $\boldsymbol{X}_1$ is uncorrelated with $\{\boldsymbol{W}_t\}$ and $\{\boldsymbol{V}_t\}$.

In a "control situation" the observations are often of reduced dimension, i.e. $w < v$. In such a situation (11.1) might be extended to

$$\boldsymbol{X}_{t+1} = H_t \boldsymbol{u}_t + F_t \boldsymbol{X}_t + \boldsymbol{V}_t, \quad t = 1, 2, \ldots,$$

where $H_t \boldsymbol{u}_t$ represents the effect of a "control" $\boldsymbol{u}_t$.

We will, however, not treat control theory at all, but we want to apply the success of state-space models and Kalman filtering to time series.

Definition 11.1 (State-space representation) *A time series* $\{\boldsymbol{Y}_t\}$ *has a* state-space representation *if there exists a state-space model for* $\{\boldsymbol{Y}_t\}$ *as specified by equations* (11.1) *and* (11.2).

Example 11.1 (AR(p) process) Let us first consider a causal AR(1) process, i.e.

$$Y_t = \phi Y_{t-1} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

In this case we don't "need" the observation equation, i.e. we put $G_t = 1$ and $R_t = 0$ in (11.2) so that $Y_t = X_t$. The state equation (11.1) with $F_t = \phi$ and $Q_t = \sigma^2$ together with $X_1 = Y_1 = \sum_{j=0}^{\infty} \phi^j Z_{1-j}$ and $V_t = Z_t$ yield the desired AR(1) model.

Now we consider a causal AR(p) process, i.e.

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

The idea is to increase the dimension in the state-space representation in to allow for the order p. Those readers acquainted with Markov processes recognize this idea as the "standard" way to Markovize a process. Let

$$\boldsymbol{X}_t = \begin{pmatrix} Y_{t-p+1} \\ Y_{t-p+2} \\ \vdots \\ Y_t \end{pmatrix}, \quad t = 1, 2, \ldots$$

and thus we have the observation equation

$$Y_t = (0, \ldots, 0, 1) \boldsymbol{X}_t.$$

Now consider the state equation for $\boldsymbol{X}_{t+1}$. We want the last component to give the AR(p) model and the other just to cause no trouble. This can be done by considering the state equation

$$\boldsymbol{X}_{t+1} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 1 \\ \phi_p & \phi_{p-1} & \phi_{p-2} & \ldots & \phi_1 \end{pmatrix} \boldsymbol{X}_t + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} Z_{t+1}.$$

These equations have the required form with $\boldsymbol{W}_t = \boldsymbol{0}$ and

$$\boldsymbol{V}_t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ Z_{t+1} \end{pmatrix}.$$

It only remains to specify $\boldsymbol{X}_1$, but that is easily done similarly as for $p = 1$. However, it is somewhat more elegant to consider the state-space representation for $t = 0, \pm 1, \ldots$, as done in [7]. □

Example 11.2 (ARMA(p, q) process) Let $\{Y_t\}$ be a causal ARMA(p, q) process satisfying

$$\phi(B)Y_t = \theta(B)Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

The idea is to let the state equation take care of the AR part as in Example 11.1 and the observation equation of the MA part. The trick is to consider an ARMA(r, r) model for $r = \max(p, q + 1)$ and to notice that if U_t is an AR(p) process satisfying $\phi(B)U_t = Z_t$, then $Y_t = \theta(B)U_t$ since

$$\phi(B)Y_t = \phi(B)\theta(B)U_t = \theta(B)\phi(B)U_t = \theta(B)Z_t.$$

For details we refer to [7]. □

Kalman filtering deals with (recursive) best linear estimation of $\boldsymbol{X}_t$ in terms of observations of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ and a random vector $\boldsymbol{Y}_0$ which is uncorrelated with $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ for all $t \geq 1$. Before going into this, we will consider estimation (prediction) of multivariate random variables.

11.1.2 Prediction of multivariate random variables

Let us first recall some basic facts about prediction of (univariate) random variables from Section 3.2.2 on page 25. Notice that we have changed the notation a little.

Consider any random variables $Y_1, Y_2, \ldots, Y_w$ and X with finite means and variances. Put $\mu_i = E(Y_i)$, $\mu = E(X)$,

$$\Gamma_w = \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,w} \\ \vdots & & \\ \gamma_{w,1} & \cdots & \gamma_{w,w} \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(Y_1, Y_1) & \cdots & \mathrm{Cov}(Y_1, Y_w) \\ \vdots & & \\ \mathrm{Cov}(Y_w, Y_1) & \cdots & \mathrm{Cov}(Y_w, Y_w) \end{pmatrix}$$

and

$$\boldsymbol{\gamma}_w = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_w \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(Y_1, X) \\ \vdots \\ \mathrm{Cov}(Y_w, X) \end{pmatrix}.$$

Let for convenience $\mu = 0$ and $\mu_i = 0$; otherwise we replace X with $X - \mu$ and Y_i with $Y_i - \mu_i$. The best linear predictor $\widehat{X}$ of X in terms of $Y_1, Y_2, \ldots, Y_w$ is given by

$$\widehat{X} = \boldsymbol{a}_w' \boldsymbol{Y},$$

where, of course,

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_w \end{pmatrix} \quad \text{and} \quad \boldsymbol{a}_w = \begin{pmatrix} a_1 \\ \vdots \\ a_w \end{pmatrix}$$

and, cf. (3.8) on page 26,

$$\boldsymbol{\gamma}_w = \Gamma_w \boldsymbol{a}_w \quad \text{or, if } \Gamma_w \text{ is non-singular, } \boldsymbol{a}_w' = \boldsymbol{\gamma}_w' \Gamma_w^{-1}.$$

Recall further that $\widehat{X}$ is uniquely determined also when Γ_w is singular although $\boldsymbol{a}_w$ is in that case not uniquely determined. The "matrix way" to express this is that Γ_w^{-1} may be any *generalized inverse* of Γ_w. A generalized inverse of a matrix S is a matrix S^{-1} such that $SS^{-1}S = S$. Every matrix has at least one. From now on we use the notation S^{-1} for the inverse of S if S is non-singular and for any generalized inverse otherwise.

Let

$$P(X \mid \boldsymbol{Y}) \stackrel{\text{def}}{=} \widehat{X}, \quad \boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_v \end{pmatrix} \quad \text{and} \quad P(\boldsymbol{X} \mid \boldsymbol{Y}) \stackrel{\text{def}}{=} \begin{pmatrix} P(X_1 \mid \boldsymbol{Y}) \\ \vdots \\ P(X_v \mid \boldsymbol{Y}) \end{pmatrix} = M\boldsymbol{Y},$$

where M is a $v \times w$ matrix given by $M = E(\boldsymbol{X}\boldsymbol{Y}')[E(\boldsymbol{Y}\boldsymbol{Y}')]^{-1}$. In connection with Kalman filtering the observations are $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t$ and – possibly – a random variable $\boldsymbol{Y}_0$ which is uncorrelated with $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ for all $T \geq 1$. In many case $\boldsymbol{Y}_0$ will be the constant vector $(1, \ldots, 1)$. In that case we use the notation

$$P_t(\boldsymbol{X}) \stackrel{\text{def}}{=} P(\boldsymbol{X} \mid \boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t),$$

i.e. the vector of best linear predictors of $X_1, \ldots, X_v$ in terms of all components of $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t$. Thus we have

$$P_t(\boldsymbol{X}) = A_0 \boldsymbol{Y}_0 + \ldots + A_t \boldsymbol{Y}_t$$

with $v \times w$ matrices $A_0, \ldots, A_t$ such that, cf. (3.9) on page 27,

$$[\boldsymbol{X} - P_t(\boldsymbol{X})] \perp \boldsymbol{Y}_s, \quad s = 0, \ldots, t. \tag{11.3}$$

Recall from the discussion on page 31 that uncorrelated random variables may be regarded as orthogonal. The equations (11.3) are a "matrix version" of Theorem A.2 on page 116. Since projections are linear operators we have

$$P_t(B_1 \boldsymbol{X}_1 + B_2 \boldsymbol{X}_2) = B_1 P_t(\boldsymbol{X}_1) + B_2 P_t(\boldsymbol{X}_2).$$

11.1.3 The Kalman recursions

We will now – at last – consider the Kalman filter or better expressed the *Kalman recursions*. Recall the state-space model defined by (11.1) and (11.2):

$$\boldsymbol{X}_{t+1} = F_t \boldsymbol{X}_t + \boldsymbol{V}_t, \quad \{\boldsymbol{V}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{Q_t\}),$$
$$\boldsymbol{Y}_t = G_t \boldsymbol{X}_t + \boldsymbol{W}_t, \quad \{\boldsymbol{W}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{R_t\}).$$

Linear estimation of $\boldsymbol{X}_t$ in terms of

- $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{t-1}$ defines the *prediction problem*;
- $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t$ defines the *filtering problem*;
- $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_n, \ n > t,$ defines the *smoothing problem*.

Theorem 11.1 (Kalman Prediction) *The predictors* $\widehat{\boldsymbol{X}}_t \stackrel{\text{def}}{=} P_{t-1}(\boldsymbol{X}_t)$ *and the error covariance matrices*

$$\Omega_t \stackrel{\text{def}}{=} E[(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)']$$

are uniquely determined by the initial conditions

$$\widehat{\boldsymbol{X}}_1 = P(\boldsymbol{X}_1 \mid \boldsymbol{Y}_0), \quad \Omega_1 \stackrel{\text{def}}{=} E[(\boldsymbol{X}_1 - \widehat{\boldsymbol{X}}_1)(\boldsymbol{X}_1 - \widehat{\boldsymbol{X}}_1)']$$

and the recursions, for $t = 1, \ldots,$

$$\widehat{\boldsymbol{X}}_{t+1} = F_t \widehat{\boldsymbol{X}}_t + \Theta_t \Delta_t^{-1} (\boldsymbol{Y}_t - G_t \widehat{\boldsymbol{X}}_t) \tag{11.4}$$
$$\Omega_{t+1} = F_t \Omega_t F_t' + Q_t - \Theta_t \Delta_t^{-1} \Theta_t', \tag{11.5}$$

where

$$\Delta_t = G_t \Omega_t G_t' + R_t,$$
$$\Theta_t = F_t \Omega_t G_t'.$$

The matrix $\Theta_t \Delta_t^{-1}$ *is called the* Kalman gain.

Proof: We shall make use of the *innovations*, $\boldsymbol{I}_t$, defined by $\boldsymbol{I}_0 \stackrel{\text{def}}{=} \boldsymbol{Y}_0$ and

$$\boldsymbol{I}_t \stackrel{\text{def}}{=} \boldsymbol{Y}_t - P_{t-1}(\boldsymbol{Y}_t) = \boldsymbol{Y}_t - G_t \widehat{\boldsymbol{X}}_t = G_t(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t) + \boldsymbol{W}_t, t = 1, 2, \ldots$$

It follows from (11.3) that $\{\boldsymbol{I}_t\}$ is orthogonal. Notice that $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_t$ and $\boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{t-1}, \boldsymbol{I}_t$ contain the same information, or differently expressed, span the same Hilbert space. Thus

$$P_t(\boldsymbol{X}) = P(\boldsymbol{X} \mid \boldsymbol{Y}_0, \ldots, \boldsymbol{Y}_{t-1}, \boldsymbol{I}_t) = P_{t-1}(\boldsymbol{X}) + P(\boldsymbol{X} \mid \boldsymbol{I}_t),$$

where the last equality follows from (11.3). Thus we get

$$\widehat{\boldsymbol{X}}_{t+1} = P_t(\boldsymbol{X}_{t+1}) = P_{t-1}(\boldsymbol{X}_{t+1}) + P(\boldsymbol{X}_{t+1} \mid \boldsymbol{I}_t)$$

$$= P_{t-1}(F_t\boldsymbol{X}_t + \boldsymbol{V}_t) + E(\boldsymbol{X}_{t+1}\boldsymbol{I}_t')[E(\boldsymbol{I}_t\boldsymbol{I}_t')]^{-1}\boldsymbol{I}_t$$
$$= P_{t-1}(F_t\boldsymbol{X}_t) + E(\boldsymbol{X}_{t+1}\boldsymbol{I}_t')[E(\boldsymbol{I}_t\boldsymbol{I}_t')]^{-1}\boldsymbol{I}_t.$$

So we get

$$\Delta_t \stackrel{\text{def}}{=} E(\boldsymbol{I}_t\boldsymbol{I}_t') = E[G_t(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)'G_t'] + E(\boldsymbol{W}_t\boldsymbol{W}_t') = G_t\Omega_t G_t' + R_t$$

and

$$\Theta_t \stackrel{\text{def}}{=} E(\boldsymbol{X}_{t+1}\boldsymbol{I}_t') = E[(F_t\boldsymbol{X}_t + \boldsymbol{V}_t)((\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)'G_t' + \boldsymbol{W}_t')]$$
$$= E[(F_t\boldsymbol{X}_t)(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)'G_t'] = F_t\Omega_t G_t'.$$

In order to verify (11.5) we note that

$$\Omega_{t+1} = E(\boldsymbol{X}_{t+1}\boldsymbol{X}_{t+1}') - E(\widehat{\boldsymbol{X}}_{t+1}\widehat{\boldsymbol{X}}_{t+1}').$$

We have

$$E(\boldsymbol{X}_{t+1}\boldsymbol{X}_{t+1}') = E[(F_t\boldsymbol{X}_t + \boldsymbol{V})(F_t\boldsymbol{X}_t + \boldsymbol{V})'] = F_tE(\boldsymbol{X}_t\boldsymbol{X}_t')F_t' + Q_t$$

and

$$E(\widehat{\boldsymbol{X}}_{t+1}\widehat{\boldsymbol{X}}_{t+1}') = E[(F_t\widehat{\boldsymbol{X}}_t + \Theta_t\Delta_t^{-1}\boldsymbol{I}_t)(F_t\widehat{\boldsymbol{X}}_t + \Theta_t\Delta_t^{-1}\boldsymbol{I}_t)']$$
$$= F_tE(\widehat{\boldsymbol{X}}_t\widehat{\boldsymbol{X}}_t')F_t' + \Theta_t\Delta_t^{-1}\Delta_t\Delta_t^{-1}\Theta_t = F_tE(\widehat{\boldsymbol{X}}_t\widehat{\boldsymbol{X}}_t')F_t' + \Theta_t\Delta_t^{-1}\Theta_t.$$

Thus we finally get

$$\Omega_{t+1} = F_t[(E(\boldsymbol{X}_t\boldsymbol{X}_t') - E(\widehat{\boldsymbol{X}}_t\widehat{\boldsymbol{X}}_t')]F_t' + Q_t - \Theta_t\Delta_t^{-1}\Theta_t = F_t\Omega_t F_t' + Q_t - \Theta_t\Delta_t^{-1}\Theta_t,$$

which is the desired result. □

In connection with control theory there is generally more emphasis put on the Kalman gain than in the formulation given in Theorem 11.1. This is for instance true in the presentation given in the course "Mathematical System Theory" (Matematisk systemteori). In that course the notation is also rather different, and in order to make a comparison simpler we will use notation more similar to those used there.

Consider the state-space model:

$$\boldsymbol{X}_{t+1} = A_t\boldsymbol{X}_t + B_t\boldsymbol{V}_t, \quad \{\boldsymbol{V}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{I\}),$$
$$\boldsymbol{Y}_t = C_t\boldsymbol{X}_t + D_t\boldsymbol{W}_t, \quad \{\boldsymbol{W}_t\} \sim \mathrm{WN}(\boldsymbol{0}, \{I\}).$$

Except for the trivial changes $F_t = A_t$ and $G_t = C_t$ we get $Q_t = B_tB_t'$ and $R_t = D_tD_t'$. Further the Kalman gain is denoted by K_t and Ω_t by P_t. Since we have used P_t for projections we prefer to let $\Omega_t = \Pi_t$. A routine reformulation of Theorem 11.1 goes as follows:

The predictors $\widehat{\boldsymbol{X}}_t \stackrel{\text{def}}{=} P_{t-1}(\boldsymbol{X}_t)$ *and the error covariance matrices*

$$\Pi_t \stackrel{\text{def}}{=} E[(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t)']$$

are uniquely determined by the initial conditions

$$\widehat{\boldsymbol{X}}_1 = P(\boldsymbol{X}_1 \mid \boldsymbol{Y}_0), \quad \Pi_1 \stackrel{\text{def}}{=} E[(\boldsymbol{X}_1 - \widehat{\boldsymbol{X}}_1)(\boldsymbol{X}_1 - \widehat{\boldsymbol{X}}_1)']$$

and the recursions, for $t = 1, \ldots,$

$$\widehat{\boldsymbol{X}}_{t+1} = A_t \widehat{\boldsymbol{X}}_t + \Theta_t \Delta_t^{-1} (\boldsymbol{Y}_t - C_t \widehat{\boldsymbol{X}}_t) \tag{11.6}$$

$$\Pi_{t+1} = A_t \Pi_t A_t' + B_t B_t' - \Theta_t \Delta_t^{-1} \Theta_t', \tag{11.7}$$

where

$$\Delta_t = C_t \Pi_t C_t' + D_t D_t',$$
$$\Theta_t = A_t \Pi_t C_t'.$$

The matrix $\Theta_t \Delta_t^{-1}$ *is called the* Kalman gain.

Now we let $K_t = \Theta_t \Delta_t^{-1}$ and so we get

$$\widehat{\boldsymbol{X}}_{t+1} = A_t \widehat{\boldsymbol{X}}_t + K_t (\boldsymbol{Y}_t - C_t \widehat{\boldsymbol{X}}_t)$$

and

$$K_t = A_t \Pi_t C_t' [C_t \Pi_t C_t' + D_t D_t']^{-1}$$

which seem to be the "standard form". The form for Π_{t+1} seems already to be on the "standard form", provided Θ_t and Ω_t are replaced by their definitions.

Theorem 11.2 (Kalman Filtering) *The filtered estimates* $\boldsymbol{X}_{t|t} \stackrel{\text{def}}{=} P_t(\boldsymbol{X}_t)$ *and the error covariance matrices*

$$\Omega_{t|t} \stackrel{\text{def}}{=} E[(\boldsymbol{X}_t - \boldsymbol{X}_{t|t})(\boldsymbol{X}_t - \boldsymbol{X}_{t|t})']$$

are determined by the relations

$$X_{t|t} = P_{t-1}(\boldsymbol{X}_t) + \Omega_t G_t' \Delta_t^{-1} (\boldsymbol{Y}_t - G_t \widehat{\boldsymbol{X}}_t)$$

and

$$\Omega_{t|t+1} = \Omega_t - \Omega_t G_t' \Delta_t^{-1} G_t \Omega_t'.$$

Theorem 11.3 (Kalman Fixed Point Smoothing) *The smoothed estimates* $\boldsymbol{X}_{t|n} \stackrel{\text{def}}{=} P_n(\boldsymbol{X}_t)$ *and the error covariance matrices*

$$\Omega_{t|n} \stackrel{\text{def}}{=} E[(\boldsymbol{X}_t - \boldsymbol{X}_{t|n})(\boldsymbol{X}_t - \boldsymbol{X}_{t|n})']$$

are determined for fixed t *by the recursions, which can be solved successively for* $n = t, t+1, \ldots$:

$$P_n(\boldsymbol{X}_t) = P_{n-1}(\boldsymbol{X}_t) + \Omega_{t.n} G_n' \Delta_n^{-1} (\boldsymbol{Y}_n - G_n \widehat{\boldsymbol{X}}_n),$$
$$\Omega_{t.n+1} = \Omega_{t.n} [F_n - \Theta_n \Delta_n^{-1} G_n]',$$
$$\Omega_{t|n} = \Omega_{t|n-1} - \Omega_{t.n} G_n' \Delta_n^{-1} G_n \Omega_{t.n}',$$

with initial conditions $P_{t-1}(\boldsymbol{X}_t) = \widehat{\boldsymbol{X}}_t$ *and* $\Omega_{t.t} = \Omega_{t|t-1} = \Omega_t$ *found from Kalman prediction.*

For the proof of the two last theorems we refer to [7].

Appendix A

A.1 Stochastic processes

Definition A.1 (Stochastic process) *A* stochastic process *is a family of random variables* $\{X_t,\, t \in T\}$ *defined on a probability space* $(\Omega, \mathcal{F}, P)$.

The *sample space* Ω is the set of all possible outcomes of an experiment.

$\mathcal{F}$ is a σ-*field* (or a σ-algebra), i.e.

(a) $\emptyset \in \mathcal{F}$;

(b) if $A_1, A_2, \cdots \in \mathcal{F}$ then $\bigcup_1^\infty A_i \in \mathcal{F}$;

(c) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

A *probability measure* P is a function $\mathcal{F} \to [0, 1]$ satisfying

(a) $P(\Omega) = 1$;

(b) $P(A) = 1 - P(A^c)$;

(c) if $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint, then

$$P\left(\bigcup_1^\infty A_i\right) = \sum_1^\infty P(A_i).$$

A *random variable* X defined on $(\Omega, \mathcal{F}, P)$ is a function $\Omega \to \mathbb{R}$ such that $\{\omega \in \Omega :\ X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.
T is called the *index* or parameter set. Important examples of index sets are $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$, $\{0, 1, 2, \dots\}$, $(-\infty, \infty)$ and $[0, \infty)$.

A stochastic process with $T \subset \mathbb{Z}$ is often called a *time series.*

Definition A.2 *The functions* $\{X_t(\omega),\, \omega \in \Omega\}$ *on* T *are called* realizations *or sample-paths of the process* $\{X_t,\, t \in T\}$ *on* $(\Omega, \mathcal{F}, P)$.

We will allow ourselves to use the term time series for both a process and a realization of it. The distribution function $F_X(x)$ of a random variable X is defined by $F_X(x) = P(X \leq x)$ for $x \in \mathbb{R}$.

The distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ of an n-dimensional random variable (or random vector) $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ is defined by

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \text{ for } \boldsymbol{x} \in \mathbb{R}^n.$$

Definition A.3 (The distribution of a stochastic process) *Put*

$$\mathcal{T} = \{\boldsymbol{t} \in T^n : t_1 < t_2 < \cdots < t_n,\ n = 1, 2, \ldots\}.$$

The (finite-dimensional) distribution functions are the family $\{F_{\boldsymbol{t}}(\cdot),\ \boldsymbol{t} \in \mathcal{T}\}$ *defined by*

$$F_{\boldsymbol{t}}(\boldsymbol{x}) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n), \quad \boldsymbol{t} \in T^n,\ \boldsymbol{x} \in \mathbb{R}^n.$$

When we talk about "the distribution of $\{X_t,\ t \in T \subset \mathbb{R}\}$" we mean the family $\{F_{\boldsymbol{t}}(\cdot),\ \boldsymbol{t} \in \mathcal{T}\}$.

Theorem A.1 (Kolmogorov's existence theorem) *The family*

$$\{F_{\boldsymbol{t}}(\cdot),\ \boldsymbol{t} \in \mathcal{T}\}$$

are the distribution functions of some stochastic process if and only if for any n, $\boldsymbol{t} = (t_1, \ldots, t_n) \in \mathcal{T}$, $\boldsymbol{x} \in \mathbb{R}^n$ *and* $1 \leq k \leq n$

$$\lim_{x_k \to \infty} F_{\boldsymbol{t}}(\boldsymbol{x}) = F_{\boldsymbol{t}(k)}(\boldsymbol{x}(k)) \tag{A.1}$$

where

$$\boldsymbol{t}(k) = (t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_n)' \quad \textit{and} \quad \boldsymbol{x}(k) = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)'.$$

Condition (A.1) on the current page is very natural. It just means that the two ways of removing the restriction $\{X_{t_k} \leq x_k\}$ shall be equivalent.

It shall be observed that Kolmogorov's existence theorem says that there exists a process $\{X_t,\ t \in T\}$ defined on the probability space $(\Omega, \mathcal{F}, P)$, where $\Omega = \mathbb{R}^T$ and where $\mathcal{F}$ is the σ-field generated by (i.e. the smallest σ-field containing) the sets $\{X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n\}$. Thus Kolmogorov's existence theorem says nothing about regularity of the realizations. Further a natural "event" as $\{\sup_{t \in \mathbb{R}} X_t \leq x\} \notin \mathcal{F}$. Since we shall mainly consider $T \in \{0, \pm 1, \pm 2, \ldots\}$ this is not so important.

Example A.1 Let A and Θ be two independent random variables with $A \geq 0$ and $\Theta \sim R[0, 2\pi]$. Consider the process

$$X_t = r^{-1} A \cos(\nu t + \Theta), \quad t \in \mathbb{R},$$

where $\nu \geq 0$ and $r > 0$ are given. Here the existence is no problem, since for any outcome a and θ of A and Θ the realization is just the function

$$x(t) = r^{-1} a \cos(\nu t + \theta), \quad t \in \mathbb{R}.$$

□

Example A.2 (A binary process) We will now show how the existence of a sequence $\{X_t,\, t = 1, 2, \dots\}$ of independent random variables with

$$P(X_t = 1) = P(X_t = -1) = \frac{1}{2}$$

follows from Kolmogorov's existence theorem. Note that the word "independent" implies that $X_1, X_2, \dots$ are defined on the same probability space. We want

$$P(X_1 = i_1,\, X_2 = i_2, \dots,\, X_n = i_n) = 2^{-n}$$

where $i_k = 1$ or -1. Since

$$P(X_1 = i_1, \dots,\, X_{k-1} = i_{k-1},\, X_k = 1,\, X_{k+1} = i_{k+1}, \dots, X_n = i_n) +$$

$$P(X_1 = i_1, \dots,\, X_{k-1} = i_{k-1},\, X_k = -1,\, X_{k+1} = i_{k+1}, \dots, X_n = i_n)$$

$$= 2^{-n} + 2^{-n} = 2^{-(n-1)}$$

$$= P(X_1 = i_1, \dots,\, X_{k-1} = i_{k-1},\, X_{k+1} = i_{k+1}, \dots, X_n = i_n)$$

the existence follows. □

Sometimes it is comfortable to express (A.1) on the facing page in terms of characteristic functions. Put

$$\phi_{\boldsymbol{t}}(\boldsymbol{u}) = \int_{\mathbb{R}^n} e^{i\boldsymbol{u}'\boldsymbol{x}} F_{\boldsymbol{t}}(dx_1, \dots, dx_n),$$

then (A.1) is equivalent with

$$\lim_{u_i \to 0} \phi_{\boldsymbol{t}}(\boldsymbol{u}) = \phi_{\boldsymbol{t}(i)}(\boldsymbol{u}(i)).$$

Definition A.4 *A random vector* $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$ *is (multivariate)* normally distributed *if there exists a vector* $\boldsymbol{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$, *a matrix* $B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & & \\ b_{m1} & \dots & b_{mn} \end{pmatrix}$ *and a random vector* $\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ *with independent and* $N(0,1)$*-distributed components, such that*

$$\boldsymbol{Y} = \boldsymbol{a} + B\boldsymbol{X}.$$

We have

$$\boldsymbol{\mu}_{\boldsymbol{Y}} \stackrel{\text{def}}{=} E\boldsymbol{Y} \stackrel{\text{def}}{=} \begin{pmatrix} EY_1 \\ \vdots \\ EY_m \end{pmatrix} = E(\boldsymbol{a} + B\boldsymbol{X}) = \boldsymbol{a} + B\boldsymbol{0} = \boldsymbol{a},$$

$$\Sigma_{\boldsymbol{Y}\boldsymbol{Y}} \stackrel{\text{def}}{=} \operatorname{Cov}(\boldsymbol{Y}, \boldsymbol{Y}) \stackrel{\text{def}}{=} E[\boldsymbol{Y} - E(\boldsymbol{Y})][\boldsymbol{Y} - E(\boldsymbol{Y})]'$$

$$= E[\boldsymbol{a} + B\boldsymbol{X} - \boldsymbol{a}][\boldsymbol{a} + B\boldsymbol{X} - \boldsymbol{a}]' = E[B\boldsymbol{X}][B\boldsymbol{X}]'$$

$$= E[B\boldsymbol{X}\boldsymbol{X}'B'] = BE[\boldsymbol{X}\boldsymbol{X}']B' = BB'$$

and

$$\phi_{\boldsymbol{Y}}(\boldsymbol{u}) = E\exp(i\boldsymbol{u}'\boldsymbol{Y}) = E\exp(i\boldsymbol{u}'(\boldsymbol{a}+B\boldsymbol{X}))$$
$$= \exp(i\boldsymbol{u}'\boldsymbol{a})\prod_{k=1}^{m} E\exp(i(\boldsymbol{u}'B)_k X_k)$$

where $(\boldsymbol{u}'B)_k$ is the k^{th} component of the vector $\boldsymbol{u}'B$. Since

$$E\exp(iaX_i) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(iax)\exp(-x^2/2)\,dx = \exp(-a^2/2) \qquad \text{(A.2)}$$

we get

$$\phi_{\boldsymbol{Y}}(\boldsymbol{u}) = \exp(i\boldsymbol{u}'\boldsymbol{a})\prod_{k=1}^{m}\exp(-(\boldsymbol{u}'B)_k^2/2) = \exp(i\boldsymbol{u}'\boldsymbol{a})\exp(-\boldsymbol{u}'BB'\boldsymbol{u}/2)$$
$$= \exp(i\boldsymbol{u}'\boldsymbol{a} - \tfrac{1}{2}\boldsymbol{u}'\Sigma_{\boldsymbol{YY}}\boldsymbol{u}).$$

Note that the normal distribution depends on B only via $\Sigma_{\boldsymbol{YY}} = BB'$.

Now we can consider a more interesting application of Kolmogorov's existence theorem.

Definition A.5 (Standard Brownian motion) *A* standard Brownian motion*, or a standard Wiener process* $\{B(t),\, t \geq 0\}$ *is a stochastic process satisfying*

(a) $B(0) = 0$*;*

(b) *for every* $\boldsymbol{t} = (t_0, t_1, \ldots, t_n)$ *with* $0 = t_0 < t_1 < \cdots < t_n$ *the random variables* $\Delta_1 = B(t_1) - B(t_0), \ldots, \Delta_n = B(t_n) - B(t_{n-1})$ *are independent;*

(c) $B(t) - B(s) \sim N(0, t-s)$ *for* $t \geq s$.

In order to establish the existence we consider

$$\phi_{\boldsymbol{t}}(\boldsymbol{u}) = E\exp[iu_1B(t_1) + \cdots + iu_nB(t_n)]$$
$$= E\exp[iu_1\Delta_1 + iu_2(\Delta_1+\Delta_2) + \cdots + iu_n(\Delta_1+\cdots+\Delta_n)]$$
$$= E\exp[i\Delta_1(u_1+\cdots+u_n) + i\Delta_2(u_2+\cdots+u_n) + \cdots + i\Delta_n u_n]$$
$$= \exp[-\tfrac{1}{2}(u_1+\cdots+u_n)^2(t_1-t_0) - \tfrac{1}{2}(u_2+\cdots+u_n)^2(t_2-t_1) - \cdots - \tfrac{1}{2}u_n^2(t_n-t_{n-1})].$$

If we put $u_k = 0$ it is seen that

$$(u_k+\cdots+u_n)^2(t_k-t_{k-1}) + (u_{k+1}+\cdots+u_n)^2(t_{k+1}-t_k)$$

is replaced by

$$(u_{k+1}+\cdots+u_n)^2(t_k-t_{k-1}) + (u_{k+1}+\cdots+u_n)^2(t_{k+1}-t_k)$$
$$= (u_{k+1}+\cdots+u_n)^2(t_{k+1}-t_{k-1})$$

which is in agreement with (A.1) on page 112.

Definition A.6 (Poisson process) *A* Poisson process $\{N(t),\, t \geq 0\}$ *with mean rate (or intensity)* λ *is a stochastic process satisfying*

(a) $N(0) = 0$*;*

(b) *for every* $\boldsymbol{t} = (t_0, t_1, \ldots, t_n)$ *with* $0 = t_0 < t_1 < \cdots < t_n$ *the random variables* $\Delta_1 = N(t_1) - N(t_0), \ldots, \Delta_n = N(t_n) - N(t_{n-1})$ *are independent;*

(c) $N(t) - N(s) \sim \mathrm{Po}(\lambda(t-s))$ *for* $t \geq s$.

The proof of existence follows as in the Brownian motion. However,

$$E\exp(iu\Delta_j) = \sum_{k=1}^{\infty} e^{iuk} \cdot \frac{[\lambda(t_j - t_{j-1})]^k}{k!} e^{-\lambda(t_j - t_{j-1})}$$

$$= e^{-\lambda(t_j - t_{j-1})} \sum_{k=1}^{\infty} \frac{[\lambda(t_j - t_{j-1})e^{iu}]^k}{k!}$$

$$= e^{-\lambda(t_j - t_{j-1})} e^{\lambda(t_j - t_{j-1})e^{iu}} = e^{-\lambda(t_j - t_{j-1})(1 - e^{iu})}.$$

A.2 Hilbert spaces

Definition A.7 (Hilbert space) *A space* $\mathcal{H}$ *is a* (complex) Hilbert space *if:*

I) $\mathcal{H}$ *is a vector space, i.e.*

(a) *addition is defined:* $x + y \in \mathcal{H}$ *for all* $x, y \in \mathcal{H}$*;*

(b) *scalar multiplication is defined:* $cx \in \mathcal{H}$ *for all* $x \in \mathcal{H}$ *and* $c \in \mathbb{C}$.

II) $\mathcal{H}$ *is an inner-product space, i.e. for all* $x, y, z \in \mathcal{H}$ *there exists* $\langle x, y\rangle \in \mathbb{C}$ *such that*

(a) $\langle x, y\rangle = \overline{\langle y, x\rangle}$*;*

(b) $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$ *for all* $a, b \in \mathbb{C}$.

(c) $\|x\| = \sqrt{\langle x, x\rangle}$ *is a norm, i.e.* $\|x\| \geq 0$ *and* $\|x\| = 0$ *if and only if* $x = 0$.

III) $\mathcal{H}$ *is complete, i.e. if* $x_1, x_2, \ldots \in \mathcal{H}$ *and* $\lim_{n,m\to\infty} \|x_n - x_m\| = 0$ *there exists* $x \in \mathcal{H}$ *such that* $\lim_{n\to\infty} \|x_n - x\| = 0$.

If $\mathbb{C}$ is replaced with $\mathbb{R}$ we talk about a *real Hilbert space.*

$\mathbb{R}^3$ is a Hilbert space and, roughly speaking, the geometry in a general Hilbert space is that of $\mathbb{R}^3$.

Consider now the space $L^2(\Omega, \mathcal{F}, P)$ of all random variables X defined on a probability space $(\Omega, \mathcal{F}, P)$ and satisfying $EX^2 < \infty$. Put $\langle X, Y\rangle = E(XY)$. It is easy to check that $L^2(\Omega, \mathcal{F}, P)$ fulfills condition I) and II), while the completeness is difficult to prove. In order to discuss this, we need some facts about convergence of random variables.

$X_n \xrightarrow{\text{m.s.}} X$ means that $\|X_n - X\| \to 0$ as $n \to \infty$. " $\xrightarrow{\text{m.s.}}$ " stands for "mean-square convergence" and the notion requires that $X, X_1, X_2, \cdots \in L^2$.

$X_n \xrightarrow{P} X$ means that $P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$ for all $\varepsilon > 0$. " $\xrightarrow{P}$ " stands for "convergence in probability".

$X_n \xrightarrow{\text{a.s.}} X$ means that $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for all $\omega \in \Omega \backslash E$ where $P(E) = 0$. " $\xrightarrow{\text{a.s.}}$ " stands for "almost sure convergence" or "convergence with probability one".

We have the following (but no other) relations

$$\begin{array}{l} X_n \xrightarrow{\text{m.s.}} X \Rightarrow \\ \qquad\qquad\qquad X_n \xrightarrow{P} X. \\ X_n \xrightarrow{\text{a.s.}} X \Rightarrow \end{array}$$

If $X_n \xrightarrow{P} X$ but $X_n \xrightarrow{\text{a.s.}} X$ does not hold, then a typical realization contains more and more sparse "exceptional points". It is therefore not too surprising that if $X_n \xrightarrow{P} X$, then there exists $n_1, n_2, \ldots$ such that $X_{n_k} \xrightarrow{\text{a.s.}} X$. Further, if $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{\text{a.s.}} Y$, then $X = Y$ a.s.

The idea in the proof of completeness is to first choose $n_1, n_2, \ldots$ such that X_{n_k} forms a Cauchy-sequence a.s. That means that there exists an $E \in \mathcal{F}$ with $P(E) = 0$ such that $X_{n_k}(\omega)$ forms a Cauchy-sequence for all $\omega \in \Omega \backslash E$. From the completeness of $\mathbb{R}$ it then follows that there exists X such that $X_n \xrightarrow{\text{a.s.}} X$. By Fatou's lemma we get

$$\|X_n - X\|^2 \leq \liminf_{k \to \infty} \|X_n - X_{n_k}\|^2$$

which can be made arbitrarily small since $\{X_n\}$ is a Cauchy-sequence in L^2-norm.

Let $\mathcal{M}$ be a Hilbert sub-space of $\mathcal{H}$. This means that $\mathcal{M} \subset \mathcal{H}$ and that $\mathcal{M}$ is a Hilbert space.

Let $\mathcal{M}$ be any subset of $\mathcal{H}$, i.e. $\mathcal{M}$ does not need to be a Hilbert sub-space of $\mathcal{H}$. The orthogonal complement $\mathcal{M}^\perp$, defined by

$$\mathcal{M}^\perp = \{y : \langle y, x \rangle = 0,\, x \in \mathcal{M}\}$$

is a Hilbert sub-space.

Theorem A.2 (The Projection theorem) *If $\mathcal{M}$ is a Hilbert sub-space of $\mathcal{H}$ and $x \in \mathcal{H}$, then*

(i) *there is a unique element $\widehat{x} \in \mathcal{M}$ such that*

$$\|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|,$$

(ii) *$\widehat{x} \in \mathcal{M}$ and $\|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\widehat{x} \in \mathcal{M}$ and $x - \widehat{x} \in \mathcal{M}^\perp$.*

[The element $\widehat{x}$ is called the (orthogonal) projection of x onto $\mathcal{M}$ and sometimes denoted by $P_{\mathcal{M}} x$.]

The requirement $x - \widehat{x} \in \mathcal{M}^\perp$ means that $\langle x - \widehat{x}, y\rangle = 0 \quad$ for all $y \in \mathcal{M}$.

Idea of proof:
(i) From the completeness it follows that there exists $\widehat{x} \in \mathcal{M}$ such that

$$\|x - \widehat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|.$$

The uniqueness follows from the parallelogram law:

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

(ii) If $\widehat{x} \in \mathcal{M}$ and $x - \widehat{x} \in \mathcal{M}^\perp$ then $\|x - y\| \geq \|x - \widehat{x}\|$ for any $y \in \mathcal{M}$.

If $\widehat{x} \in \mathcal{M}$ and $x - \widehat{x} \notin \mathcal{M}^\perp$ then $\widehat{x}$ is *not* the element of $\mathcal{M}$ closest to x, since, for any $y \in \mathcal{M}$ such that $\langle x - \widehat{x}, y\rangle \neq 0$,

$$\tilde{x} = \widehat{x} + \frac{\langle x - \widehat{x}, y\rangle y}{\|y\|^2}$$

is closer, i.e. $\|\tilde{x} - x\| < \|\widehat{x} - x\|$. □

We will now consider a class of important Hilbert sub-spaces.

Definition A.8 (Closed span) *The* closed span $\overline{\text{sp}}\{x_t,\, t \in T\}$ *of any subset* $\{x_t,\, t \in T\}$ *of a Hilbert space* $\mathcal{H}$ *is defined to be the smallest Hilbert sub-space which contains* $\{x_t,\, t \in T\}$.

If T is finite, say $\{x_1, x_2, \ldots, x_n\}$, then

$$\overline{\text{sp}}\{x_1, x_2, \ldots, x_n\} = \alpha_1 x_1 + \cdots + \alpha_n x_n \quad \alpha_k \in \mathbb{C} \text{ or } \mathbb{R}.$$

If T is infinite then $\overline{\text{sp}}\{x_t,\, t \in T\}$ is the set of all finite linear combinations

$$\alpha_{n_1} x_{n_1} + \cdots + \alpha_{n_k} x_{n_k} \quad \alpha_{n_j} \in \mathbb{C} \text{ or } \mathbb{R}$$

and limits of such combinations.

Remark A.1 If, in the projection theorem, $\mathcal{M} = \overline{\text{sp}}\{x_t,\, t \in T\}$ the *prediction*, or normal, *equations*

$$\langle x - \widehat{x}, y\rangle = 0 \quad \text{for all } y \in \mathcal{M}$$

are reduced to

$$\langle x - \widehat{x}, x_t\rangle = 0 \quad \text{for all } t \in T.$$

Theorem A.3 (Properties of projections) *Let* $\mathcal{H}$ *be a Hilbert space and let* $P_\mathcal{M}$ *denote the projection onto a Hilbert sub-space* $\mathcal{M}$ *of* $\mathcal{H}$. *Then*

(i) $P_\mathcal{M}(\alpha x + \beta y) = \alpha P_\mathcal{M} x + \beta P_\mathcal{M} y, \quad x, y \in \mathcal{H}, \quad \alpha, \beta \in \mathbb{C}$ *or* $\mathbb{R}$;

(ii) $\|x\|^2 = \|P_\mathcal{M} x\|^2 + \|(I - P_\mathcal{M})x\|^2$, *where* I *is the identity mapping;*

(iii) *each* $x \in \mathcal{H}$ *has a unique representation as* $x = P_\mathcal{M} x + (I - P_\mathcal{M})x$;

(iv) $P_\mathcal{M} x_n \xrightarrow{\text{m.s.}} P_\mathcal{M} x$ *if* $x_n \xrightarrow{\text{m.s.}} x$;

(v) $x \in \mathcal{M}$ *if and only if* $P_\mathcal{M} x = x$;

(vi) $x \in \mathcal{M}^\perp$ *if and only if* $P_\mathcal{M} x = 0$;

(vi) $\mathcal{M}_1 \subseteq \mathcal{M}_2$ *if and only if* $P_{\mathcal{M}_1} P_{\mathcal{M}_2} x = P_{\mathcal{M}_1} x$ *for all* $x \in \mathcal{H}$.

References

[1] Blom, G., Enger, J., Englund, G., Grandell, J. och Holst, L. (2005). *Sannolikhetsteori och statistikteori med tillämpningar*, Studentlitteratur, Lund.

[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics, 31*, 307–327.

[3] Bollerslev, T., Chou, R. Y. and Kroner, K. F. (1986). ARCH modeling in finance *Journal of Econometrics, 52*, 5–59.

[4] Bollerslev, T., Engle, R. F. and Nelson, D. B. (1994). ARCH models. In *Handbook of Econometrics, Vol. IV*. Ed. by Engle, R. F. and McFadden, D. L., Elsevier, Amsterdam, 2959–3038.

[5] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control, Revised Edition*, Holden-Day, San Francisco.

[6] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods, 2nd Edition*, Springer-Verlag, New York.

[7] Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*, Springer-Verlag, New York.

[8] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag, Berlin.

[9] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica, 50*, 987–1007.

[10] Hagerud, G. E. (1997). A New Non-Linear GARCH Model. Thesis, Stockholm School of Economics, Stockholm.

[11] Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *Journal of Business, 36*, 394–419.

[12] Palm, F. C. (1996). GARCH Models of Volatility. In *Handbook of Statistics, Vol. 14*. Ed. by Maddala, G. S. and Rao, C. R., Elsevier, Amsterdam, 209–240.

Index