united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR: 1170/8

WORKSHOP ON "MODELLING REAL SYSTEMS: A HANDS-ON FIRST ENCOUNTER WITH INDUSTRIAL MATHEMATICS

(27 September - 22 October 1999)

"The Prague Lectures ECONOMETRICS II"

presented by:

Manfred DEISTLER
Institute for Econometrics
Operations Research and System Theory
University of Technolgy
1040 Vienna
Austria

These are preliminary lecture notes, intended only for distribution to participants.

The Prague Lectures
ECONOMETRICS II

M. Deistler and W. Scherrer
Institute for Econometrics, Operations Research
and System Theory
University of Technology, Vienna

Lectures given at CERGE Prague 20/01/92 - 15/05/92

DRAFT April 13, 1994

Table of Contents

0 Introduction **1**

0.1 History . . . 1

0.2 Main uses for TSA . . . 2

0.3 Main areas of application . . . 3

1 Descriptive Analysis of Time Series **4**

1.1 Means, Covariances and Correlations . . . 4

1.2 Trends . . . 6

1.3 Hidden Periodicities . . . 9

1.4 Autoregression and Autoregression with Exogeneous Variables . . . 11

1.5 Choosing the Number of Regressors . . . 12

1.6 Exercises . . . 14

2 Stationary Processes in Time Domain **16**

2.1 Basic Definitions . . . 16

2.2 Examples for stationary processes . . . 20

2.2.1 White Noise Process . . . 20

2.2.2 Moving Average Process . . . 21

2.2.3 Infinite Moving Average Process . . . 21

2.2.4 Autoregressive Process . . . 22

2.2.5 ARMA Process . . . 26

2.2.6 Harmonic Processes . . . 26

2.3 Properties of Covariance Functions . . . 31

2.4 Hilbert-spaces . . . 32

2.4.1 Examples . . . 33

2.5 Exercises . . . 34

3 Stationary Processes in Frequency Domain **38**

3.1 The Spectral Representation of Stationary Processes . . . 38

3.2 Linear Transformations of Stationary Processes . . . 45

3.3 Exercises . . . 54

4 Linear Vector Difference Equations **58**

4.1 Solution of VDE's . . . 58

4.2 Representations of VDE's . . . 63

4.3 Identifiability . . . 69

4.4 Structural Identifiability . . . 74

4.5 ARMAX and state space systems 75
4.6 Exercises 77

5 Prediction and Filtering **78**
5.1 Prediction from a Finite Past 79
5.2 Prediction from the Infinite Past: the ARMA Case 82
5.3 Wold decomposition 83
5.4 Filtering 87
5.5 Exercises 89

6 Estimation of the Mean and of the Covariance Function **91**
6.1 Convergence Concepts 91
6.2 Estimation of the Mean 93
6.3 Estimation of the Covariances and Correlations 97
6.4 Exercises 100

7 Estimation of the spectrum **102**
7.1 Properties of the Periodogram 102
7.2 Smoothed Spectral Estimates 106
7.3 Exercises 111

8 Identification of Linear Systems **113**
8.1 ARIMA processes 113
8.2 Identification of Scalar AR Systems 114
8.3 Identification of Vector ARX systems 123
8.4 Identification of ARMA systems 130

I. TIME SERIES ANALYSIS

0 Introduction

Time series: Observations ordered in time; here in particular

$$x_t \quad ; \quad t = 1, \ldots, T \quad , \quad x_t \in \mathbf{R}^n$$

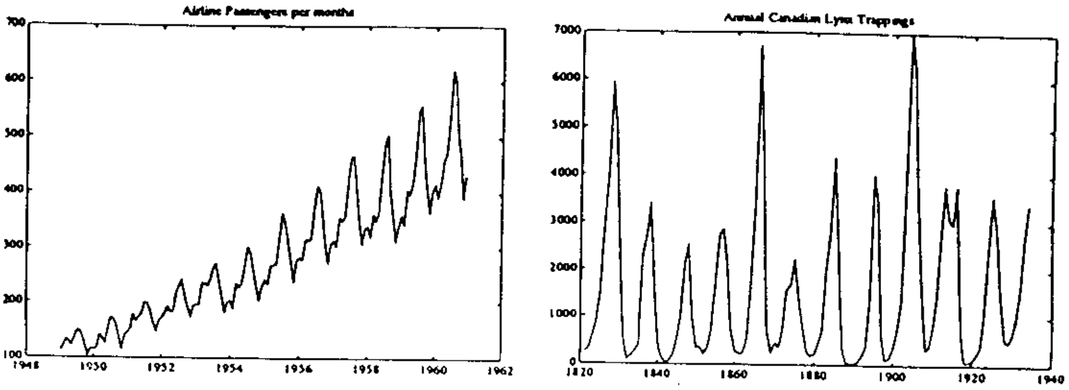


Figure 0.1: International airline passengers per month in thousands of passengers from January 1949 to December 1960 (Box and Jenkins [2]). Annual Canadian lynx trappings from 1821 to 1934 (Brockwell and Davis [3]).

Time series analysis is concerned with extraction of information from time series data.

In time series analysis the *order* of the observations contains important information. This is contrary to the case of "classical" i.i.d statistics, where a permutation of the data leaves the results unchanged.

A common approach to time series analysis (TSA) is to assume an underlying stochastic or/and dynamic model generating the time series. The problem then is to determine such a model from time series data (statistical inference, inverse problem).

0.1 History

(i) Search for hidden periodicities and trends; unobserved components. Astronomy: Search for "secular changes" in the orbit of the planets and the moon; Laplace (1787), Euler, Lagrange.

(ii) Periodogram: Stokes (1879); used by Schuster (1898) for the analysis of sunspot data and by Beveridge (1921, 1922) for economic data.

(iii) Theory of stationary processes: Cramer, Kolmogorov, Wiener, Wold (1930-45). Spectral theory; prediction and filtering. Probability theory; no statistics (in the narrow sense).

(iv) Early econometrics: Cowles Commission, T.W. Anderson, Haavelmo, Klein, Koopmans (1940-50). Simultaneous equation systems; identifiability and (Gaussian) Maximum Likelihood (ML) estimation. Subsequent: Two stage and three stage least squares (Theil, Zellner).

(v) Spectral estimation. Tukey (late 40ies, 50ies); for economic times series: Granger, Hatanaka.

(vi) Asymptotic Theory for AR and ARMA systems (T.W Anderson, Hannan, Walker, 60ies and 70ies).

(vii) Kalman filtering (1960), State space models.

(viii) "Box-Jenkins" approach: "Integrated" approach: Differencing, order estimation, ML-algorithms, 1971.

(ix) Automatic order estimation, information criteria AIC, BIC; Akaike, Rissannen (1969-85).

(x) Causality tests

(xi) Rational expectations

(xii) Cointegration, integrated processes

(xiii) ARCH-processes

0.2 Main uses for TSA

- General analysis: Search for general features such as trends, cycles or linear dependencies.
- Testing of (conflicting) theories and for estimation of parameters (which are theoretically meaningfull).
- Prediction (unconditional or conditional), Simulation.

- Control
- Preprocessing of data such as seasonal adjustment or the construction of stylized facts.

0.3 Main areas of application

- Engineering: Signal processing (speech processing, radar and sonar applications); control engineering (control of missiles, airplanes, ships, chemical or technical plants); modeling of technical parts (e.g. engines, cars, ...) e.g. for simulation; monitoring (e.g. of oil platforms).
- Economics: Forecasting and policy simulations (in particular for macrovariables like GNP, unemployment,...); "Verification" of theories, estimation of "deep" parameters; Analysis and forecasting of finacial data (e.g. stock market prices, exchange rates); Forecasting on a microlevel, (e.g. of sales or inventories of a firm).
- Biology and Medicine: e.g. analysis of EEG data, circumdianic rythms.
- Geophysics: e.g. Exploration for oil, propagation of earth quakes.

1 Descriptive Analysis of Time Series

In this short chapter we introduce and discuss some descriptive measures for time series. Here descriptive means that we have no underlying probability model. As will be seen later, however, if we have an underlying probability model, these sample measures can be interpreted as estimates for their population counterparts. Our focus is on the description of levels, trends, linear dependencies and on hidden periodicities.

1.1 Means, Covariances and Correlations

The most common measure for the level is the *(arithmetic) mean*

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

For the case of two scalar (i.e. $n = 1$) time series $x = (x_1, \dots, x_T)$ and $y = (y_1, \dots, y_T)$ the *covariance* is defined as

$$\widehat{\text{Cov}}(x, y) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})$$

The interpretation of the covariance can be seen from picture 1.1.

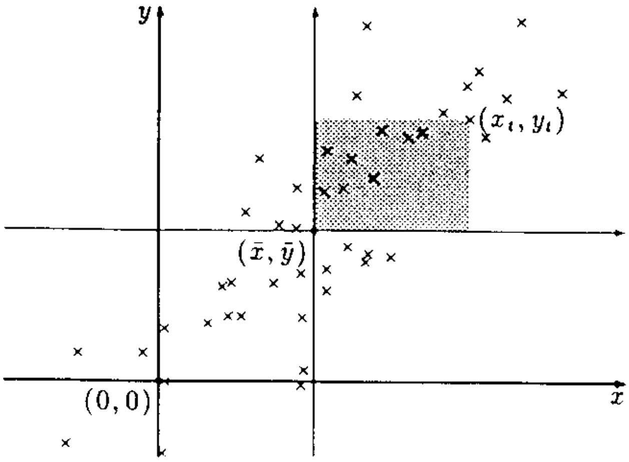


Figure 1.1: Interpretation of the sample covariance.

In particular the *variance* of x, $\widehat{\text{Var}}(x)$ is defined as $\widehat{\text{Cov}}(x, x)$.

The *noncentral covariance* is defined by

$$\frac{1}{T}\sum_{t=1}^{T} x_t y_t$$

and has an analogous interpretation as $\widehat{\mathrm{Cov}}(x,y)$. We have (see exercises)

$$\frac{1}{T}\sum_{t=1}^{T} x_t y_t = \widehat{\mathrm{Cov}}(x,y) + \bar{x}\bar{y}.$$

Covariances are measures of linear dependence. They are scale dependent. For this reason, the *correlation* as a normalized measure of linear dependence is introduced:

$$\widehat{\mathrm{Corr}}(x,y) = \frac{\widehat{\mathrm{Cov}}(x,y)}{\sqrt{\widehat{\mathrm{Var}}(x)\,\widehat{\mathrm{Var}}(y)}}$$

(Here we have tacitly assumed $\widehat{\mathrm{Var}}(x) > 0$, $\widehat{\mathrm{Var}}(y) > 0$.)

Clearly $\widehat{\mathrm{Cov}}$ can be interpreted as the inner product of the two vectors $\tilde{x} = (x_1 - \bar{x}, \ldots, x_T - \bar{x})$ and $\tilde{y} = (y_1 - \bar{y}, \ldots, y_T - \bar{y})$ and $\widehat{\mathrm{Corr}}$ can be interpreted as the cosine of the angle between these two vectors. From the Cauchy-Schwarz inequality (see exercises) for vectors in $\mathbf{R}^n$ we directly obtain

$$-1 \leq \widehat{\mathrm{Corr}}(x,y) \leq 1$$

and $|\widehat{\mathrm{Corr}}(x,y)| = 1$ if and only if $\tilde{y} = a\tilde{x}$ where $a = \frac{\tilde{x}'\tilde{y}}{\tilde{x}'\tilde{x}}$ holds. Analogous results hold for the noncentral correlation

$$\frac{\sum_{t=1}^{T} x_t y_t}{\sqrt{(\sum_{t=1}^{T} x_t^2)(\sum_{t=1}^{T} y_t^2)}}$$

For a given univariate (i.e. scalar) time series $x_1, \ldots, x_T$, the *autocovariance function* $\hat{\gamma} : \mathbf{Z} \to \mathbf{R}$ is defined by

$$\hat{\gamma}(s) = \frac{1}{T}\sum_{t=\max(1,1-s)}^{\min(T,T-s)} (x_{t+s} - \bar{x})(x_t - \bar{x})$$

Note that $\hat{\gamma}(s) = 0$ for $|s| \geq T$, $\hat{\gamma}(0) = \widehat{\mathrm{Var}}(x)$ and $\hat{\gamma}(s) = \hat{\gamma}(-s)$.

The *autocorrelation function* $\hat{\rho} : \mathbf{Z} \to \mathbf{R}$ is defined by

$$\hat{\rho}(s) = \frac{\hat{\gamma}(s)}{\widehat{\mathrm{Var}}(x)} = \frac{\hat{\gamma}(s)}{\hat{\gamma}(0)}$$

The functions $\hat{\gamma}$ and $\hat{\rho}$ are measures for linear dependence in *time*. In a completely analogous way we can define the corresponding noncentral functions.

Despite of the fact that these descriptive measures are defined using practically no assumptions on the "nature" of the time series, one has to be careful with the interpretation if there is no underlying model.

For two scalar time series $x_t, t = 1, \ldots, T$ and $y_t, t = 1, \ldots, T$ we define the *cross-covariance function* $\hat{\gamma}_{xy}$ by

$$\hat{\gamma}_{xy}(s) = \frac{1}{T} \sum_{t=\max(1,1-s)}^{\min(T,T-s)} (x_{t+s} - \bar{x})(y_t - \bar{y})$$

and the *cross-correlation function* $\hat{\rho}_{xy}$ by

$$\hat{\rho}_{xy}(s) = \frac{\hat{\gamma}_{xy}(s)}{\sqrt{\hat{\gamma}_x(0)\,\hat{\gamma}_y(0)}},$$

where $\hat{\gamma}_x$ and $\hat{\gamma}_y$ denote the autocovariance functions of x and y respectively. $\hat{\gamma}_{xy}(s)$ and $\hat{\rho}_{xy}(s)$ are measures for linear dependence between the two time series at lag s.

For a vector time series $x_t, t = 1, \ldots, T, x_t \in \mathbb{R}^n$, the autocovariance function $\hat{\gamma}$, defined by

$$\hat{\gamma}(s) = \frac{1}{T} \sum_{t=\max(1,1-s)}^{\min(T,T-s)} (x_{t+s} - \bar{x})(x_t - \bar{x})',$$

can be considered as an $n \times n$ matrix $\hat{\gamma} = (\hat{\gamma}_{ij})$ of functions $\hat{\gamma}_{ij} : \mathbb{Z} \to \mathbb{R}$, where the $\hat{\gamma}_{ii}$ are the autocovariance functions and the $\hat{\gamma}_{ij}$, $i \neq j$ are the cross-covariance functions of the respective component series.

1.2 Trends

Vaguely speaking, trends represent the long term behaviour of a time series or the dependence of levels on time. The aim may be estimation and/or elimination of trends or the decomposition of the time series in a number of (unobserved) components, one of which represents the trend.

Trendregression using least squares is a common procedure for estimation and extraction of trends. For the case of polynomial trends $m_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p$ we have

$$x_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + u_t \tag{1.1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are the (linear, ordinary) least squares estimates. Check that the matrix

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 4 & \cdots & 2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & T & T^2 & \cdots & T^p \end{pmatrix} \in \mathbf{R}^{T \times p+1}$$

is of full rank $p+1$ for $T > p$.

Clearly also other functions, as e.g. exponentials or logistic functions can be used for trend fitting by least squares; however in most cases this will give nonlinear normal equations.

Differencing is a common procedure for trend elimination: First differences are defined by

$$\Delta x_t = x_t - x_{t-1}; \text{ for } t = 2, \ldots, T$$

and p-th differences are defined by

$$\Delta^p x_t = \Delta(\Delta^{p-1} x_t); \text{ for } t = p+1, \ldots, T.$$

As is easiliy seen, first differences of a linear function give a constant and more generally p-th differences of a polynomial of order p give a constant. It should be noted that by differencing some information concerning the trend is lost (differencing is a non invertible operation, if the "initial values", e.g. x_1 for first differences, are not known).

Differencing may also be used to eliminate seasonal patterns which are very common for economic time series. E.g. for quarterly data we might use

$$\Delta_4 x_t = x_t - x_{t-4} \quad ; \quad t = 5, \ldots, T.$$

Note that for a time series $x_t = \beta_0 + \beta_1 t + s_t$, where s_t is a periodic function of period 4 (i.e. $s_t = s_{t-4}$, for all t) we have $\Delta_4 x_t = 4\beta_1$. To eliminate a quadratic trend plus a seasonal we could use a combination of first and seasonal differences, e.g. $\Delta\Delta_4 x_t$.

In order to estimate trends, *moving averages* are often used. A simple example is of the form

$$m_t = \frac{1}{2h+1} \sum_{j=-h}^{h} x_{t-j} \quad ; \quad t = h+1, \ldots, T-h$$

The idea here is to estimate local means for the time series. In a weighted moving average we have

$$m_t = \sum_{j=-h}^{h} b_j x_{t-j} \quad ; \quad t = h+1, \ldots, T-h \quad ; \quad b_j = b_{-j} \text{ and } \sum_{j=-h}^{h} b_j = 1,$$

where usually the weights will be smaller at the tails of the sum. (See exercises.)

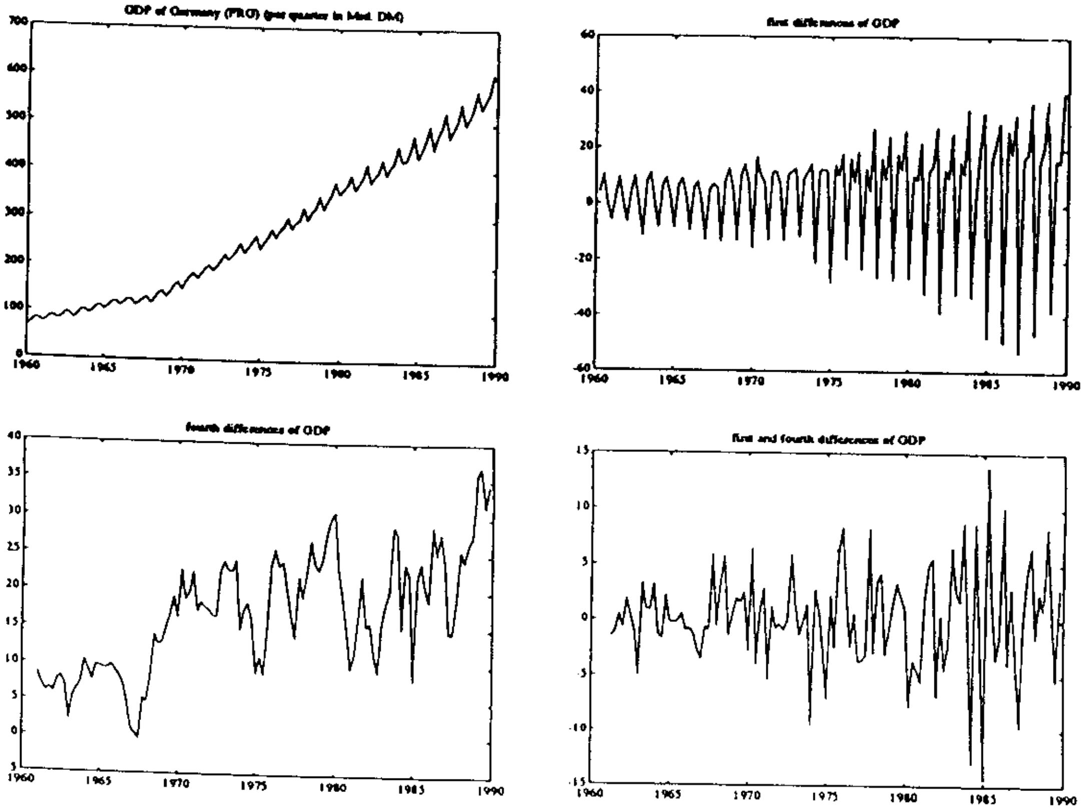


Figure 1.2: (West) German GDP (x_t), first differences ($\Delta x_t = x_t - x_{t-1}$), fourth differences ($\Delta_4 x_t = x_t - x_{t-4}$) and first plus fourth differences ($\Delta\Delta_4 x_t = x_t - x_{t-1} - x_{t-4} + x_{t-5}$) of the German GDP.

1.3 Hidden Periodicities

If we suspect hidden periodicities in a scalar time series one approach is to start from

$$x_t = \sum_{j=1}^{h}(a_j \cos\lambda_j t + b_j \sin\lambda_j t) + u_t \tag{1.2}$$

where $\lambda_j \in [0,\pi]$ (see section 2.2). The unknown parameters in (1.2) are a_j, b_j and λ_j, $j = 1,\ldots,h$. If the angular frequencies λ_j are a priori known (e.g. in the case of seasonal or weekly cycles) the a_j and b_j can be determined by least squares in a linear way.

For special frequencies , namely for the so called *Fourier frequencies*

$$\lambda_j = j\frac{2\pi}{T} \quad ; \quad j = 0,1,\ldots,\lfloor T/2 \rfloor$$

the regressors in (1.2) are even orthogonal. Here $\lfloor x \rfloor$ denotes the largest integer that is less or equal to x. Thus $\lfloor T/2 \rfloor$ is $T/2$ for T even and $(T-1)/2$ for odd T. We have

$$\begin{aligned}
\textstyle\sum_{t=1}^{T} \sin\lambda_j t &= 0 \\
\textstyle\sum_{t=1}^{T} \cos\lambda_j t &= \begin{cases} 0 \text{ for } \lambda_j \neq 0 \\ T \text{ for } \lambda_j = 0 \end{cases} \\
\textstyle\sum_{t=1}^{T} (\sin\lambda_j t)^2 &= \begin{cases} \frac{T}{2} \text{ for } \lambda_j \neq 0,\pi \\ 0 \text{ for } \lambda_j = 0,\pi \end{cases} \\
\textstyle\sum_{t=1}^{T} \sin\lambda_i t \sin\lambda_j t &= 0 \text{ for } \lambda_i \neq \lambda_j \\
\textstyle\sum_{t=1}^{T} (\cos\lambda_j t)^2 &= \begin{cases} \frac{T}{2} \text{ for } \lambda_j \neq 0,\pi \\ T \text{ for } \lambda_j = 0,\pi \end{cases} \\
\textstyle\sum_{t=1}^{T} \cos\lambda_i t \cos\lambda_j t &= 0 \text{ for } \lambda_i \neq \lambda_j \\
\textstyle\sum_{t=1}^{T} \sin\lambda_i t \cos\lambda_j t &= 0
\end{aligned} \tag{1.3}$$

Note that $\sin 0t = \sin \pi t = 0$ for all $t \in \mathbf{Z}$ and therefore we will omit $\sin 0t$ and $\sin \pi t$ from the set of regressors. Note also that $\cos 0t = 1$ and therefore $\cos 0t$ corresponds to the intercept. It is easy to see that with these restrictions we have exactly T canditate regressors, since only for even T the frequency π is a Fourier frequency (i.e. for T even and $j = T/2$ we have $\lambda_j = \pi$). Due to the orthogonality of the regressors, the least squares coefficients in (1.2) are given by

$$\begin{aligned}
\hat{a}_0 &= \bar{x} \\
\hat{b}_0 &= 0 \text{ since } \sin 0t \text{ is omitted} \\
\hat{a}_{T/2} &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T}(x_t - \bar{x})\cos\lambda_{T/2}t && \text{- (for even } T\text{)} \\
\hat{b}_{T/2} &= 0 \text{ since } sin\pi t \text{ is omitted} && \text{- (for even } T\text{)} \\
\hat{a}_j &= \tfrac{2}{T}\textstyle\sum_{t=1}^{T}(x_t - \bar{x})\cos\lambda_j t && \text{- else} \\
\hat{b}_j &= \tfrac{2}{T}\textstyle\sum_{t=1}^{T}(x_t - \bar{x})\sin\lambda_j t && \text{- else.}
\end{aligned}$$

If the angular frequencies of the hidden periodicities are unknown, the covariances between x_t, $t = 1, \dots, T$ and $\sin \lambda t$ or $\cos \lambda t$, $t = 1, \dots, T$:

$$\begin{array}{l} c(\lambda) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})\cos \lambda t \\ s(\lambda) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})\sin \lambda t \end{array} \tag{1.4}$$

are of interest. The idea here is that the linear dependence will be large for strong hidden frequencies. Equation (1.4) can be written (using $e^{i\lambda t} = \cos \lambda t + i \sin \lambda t$) as

$$x(\lambda) = c(\lambda) - is(\lambda) = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})e^{-i\lambda t} \tag{1.5}$$

which is a more compact notation. The periodogram then is defined by

$$I(\lambda) = T|x(\lambda)|^2 = T(c(\lambda)^2 + s(\lambda)^2) \quad ; \quad \lambda \in (-\pi, \pi] \tag{1.6}$$

The periodogram may be interpreted as a (frequency) dependent squared complex covariance. We have

$$I(\lambda) = T(\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})e^{-i\lambda t})(\frac{1}{T}\sum_{s=1}^{T}(x_s - \bar{x})e^{i\lambda s}) = \sum_{s=-T+1}^{T-1}\hat{\gamma}(s)e^{-i\lambda s} \tag{1.7}$$

Thus the periodogram is the Fouriertransform of the covariance function $\hat{\gamma}$. The periodogram has the following properties

$$\begin{array}{rcl} I(\lambda) & \geq & 0 \\ I(\lambda) & = & I(-\lambda) \\ I(0) & = & 0 \\ \int_{-\pi}^{\pi} I(\lambda)d\lambda & = & 2\pi\,\hat{\gamma}(0) = 2\pi\,\widehat{\mathrm{Var}}(x) \\ \int_{-\pi}^{\pi} I(\lambda)e^{i\lambda s}d\lambda & = & 2\pi\,\hat{\gamma}(s) \end{array} \tag{1.8}$$

If we "extend" the Fourier frequencies to the interval $(-\pi, \pi]$, i.e. if we define

$$\lambda_j = \frac{2\pi j}{T} \quad ; \quad j = -\lfloor (T-1)/2 \rfloor, \dots, 0, \dots, \lfloor T/2 \rfloor$$

the following "orthogonality" relations hold:

$$\begin{array}{rcl} \sum_{t=1}^{T} e^{i\lambda_j t}e^{-i\lambda_l t} & = & \begin{cases} T & \text{for } j = l \\ 0 & \text{else.} \end{cases} \\ \sum_{j=-\lfloor (T-1)/2 \rfloor}^{\lfloor T/2 \rfloor} e^{i\lambda_j s}e^{-i\lambda_j t} & = & \begin{cases} T & \text{for } s = t \\ 0 & \text{else.} \end{cases} \end{array} \tag{1.9}$$

The first equations of (1.9) (summation over time points t) are the complex counterpart to the (real) equations (1.3). From these relations we get

$$\begin{aligned} (x_t - \bar{x}) &= \sum_{j=-\lfloor (T-1)/2 \rfloor}^{\lfloor T/2 \rfloor} x(\lambda_j) e^{i\lambda_j t} \\ \hat{\gamma}(0) &= \frac{1}{T} \sum_{j=-\lfloor (T-1)/2 \rfloor}^{\lfloor T/2 \rfloor} I(\lambda_j) \end{aligned} \tag{1.10}$$

Together (1.7) and (1.8) show that the sample autocovariance function $\hat{\gamma}(s)$ and the periodogram are in a one-to-one relation. Both contain the same information, but display it in a different way. According to (1.10) (and (1.9)) the ratio $2I(\lambda_j)/\hat{\gamma}(0)$ (for $\lambda_j \neq 0$) is the coefficient of determination if we regress x_t on $e^{i\lambda_j t}$ and $e^{-i\lambda_j t}$ (respectively on $\cos \lambda_j t$ and $\sin \lambda_j t$). Thus $I(\lambda_j)$ shows how much of the sample variance of x_t (of the "energy" of the signal x_t) we can explain by a harmonic oscillation of frequency λ_j.

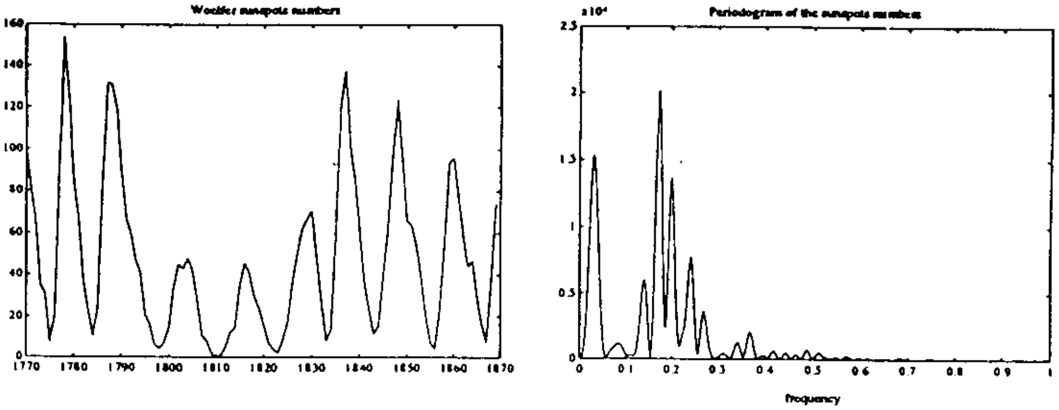


Figure 1.3: Wölfer sunspots numbers from 1770 to 1861 (Brockwell and Davis [3]) and the periodogram of this time series.

1.4 Autoregression and Autoregression with Exogeneous Variables

A rather obvious way for describing dependence in time for a time series is to approximate x_t by its "past until $t - p$", i.e. by $x_{t-1}, \ldots, x_{t-p}$ in the linear least squares sense. Consider the scalar (i.e. $n = 1$) case. Then

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + u_t \tag{1.11}$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)'$ are the OLS estimates satisfying the normal equations

$$(X'X)\alpha = X'y$$

with

$$X = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ x_1 & 0 & & 0 \\ \vdots & & \ddots & \vdots \\ \vdots & & & \vdots \\ x_{T-1} & & & x_{T-p} \end{pmatrix} \quad ; \quad y = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_T \end{pmatrix}.$$

Equation (1.11) may be used e.g. for prediction. Note that x_t is not defined by equation (1.11), since nothing is assumed about the structure of u_t (in particular we have no probability structure structure for (u_t) in contrary to the case of AR systems in section 2.2.). Therefore, in particular, $X'X$ is not necessary nonsingular (although this is very likely). Note that here we have defined the OLS estimator by putting $x_0 = x_{-1} = \cdots = x_{1-p} = 0$; other ways are possible.

For the vector case, $n > 1$, we can procede analogously. An extension of (1.11) is to approximate x_t by its own past up to time $t-p$ $(x_{t-1}, \ldots, x_{t-p})$ and by exogenous variables (inputs) $z_t, \ldots, z_{t-q}$:

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + \beta_0 z_t + \beta_1 z_{t-1} + \cdots + \beta_q z_{t-q} + u_t \qquad (1.12)$$

The OLS estimators $\alpha_1, \ldots, \alpha_p, \beta_0, \ldots, \beta_q$ are defined in an analogous way as before.

1.5 Choosing the Number of Regressors

In many cases in regression, the number of regressors is not given a priori and may be choosen from the data. E.g. we might be interested in determing the order of a polynomial trend in (1.1) or the order of an autoregression (1.11) from data.

Here for simplicity we assume that we have already a fixed list of regressors of candidate regressors which are ordered according to their importance. So choosing the number, p say, of regressors means to choose the first p regressors from the list. We then write

$$x_t = \beta_1 z_{t,1} + \cdots + \beta_p z_{t,p} + u_t$$

clearly there are two dangers involved in choosing p: p may be too small, such that we miss out important features for the description of our time series (*underfitting*); or p may be too large (*overfitting*), such that we might "overinterprete" the noise term and have to estimate too many parameters.

Clearly in such a situation, optimizing criteria of goodness of (miss)fit such as the residual sum of squares are inappropriate for estimating p, since, in general they will lead to overfitting. E.g. for a sample size of T, a polynomial trend of order $T-1$ will give a perfect fit. In general, if $\hat{\sigma}^2(p)$ denotes $\frac{1}{T}\sum u_t^2$ for given p, we will get a picture as shown in figure 1.4.

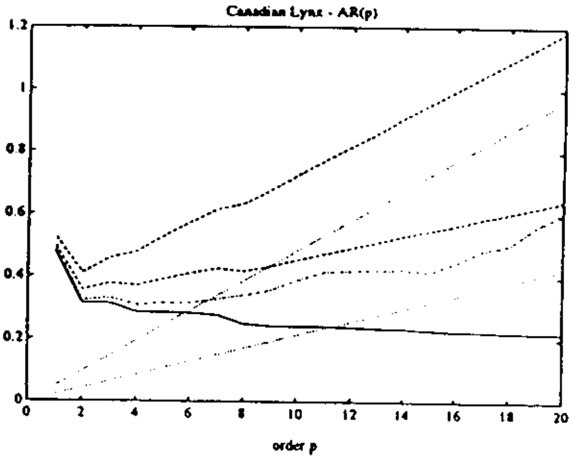


Figure 1.4: For the Canadian Lynx data AR(p) models for orders $p = 1, \ldots, 20$ were fitted by OLS. The figure shows the "in sample fit" $\hat{\sigma}^2(p)$ (solid curve), the prediction error $\sigma^2(p)$ (dashed-dotted curve) and the information criteria AIC(p) and BIC(p) (dotted curve) as a function of p. The two dotted lines are the "penalty terms" $2p/T$ and $\ln Tp/T$ of the AIC and BIC criterion respectively.

One possibility is to compare one-step-ahead prediction error variances for "honest" predictors, i.e. where $\beta_{1|n}, \ldots, \beta_{p|n}$ are the OLS estimates for $t = 1, \ldots, n$ and we define

$$u_{n+1|n}(p) = x_{n+1} - \beta_{1|n} z_{n+1,1} + \cdots + \beta_{p|n} z_{n+1,p} \qquad (1.13)$$

and $\sigma^2(p) = \frac{1}{T}\sum_n u_{n+1|n}^2$. This in general would give a picture like in figure 1.4

Now minimizing $\sigma^2(p)$ as a function of p gives a way to determine p. The basic idea behind this procedure is that in "honest" (i.e. out of sample) prediction, a model which has too many variables, in average will perform worse compared to a model with the right number of parameters (despite the fact the "in sample fit" of the first model is better).

Our problem may be seen as making a decision involving two conflicting goals namely to

- *maximize goodness of fit* of an equation to the data
- *minimize the "complexity"* of the equation used; here we measure complexity by the number of parameters p.

One approach then is to define a criterion which defines a certain trade-off between these goals and to determine p by optimizing such a criterion. Two common criteria are the *AIC criterion*

$$\mathrm{AIC}(p) = \log \hat{\sigma}^2(p) + \frac{2p}{T} \tag{1.14}$$

and the *BIC criterion*

$$\mathrm{BIC}(p) = \log \hat{\sigma}^2(p) + \frac{p \ln T}{T} \tag{1.15}$$

For both criteria, p is determined by minimizing. (See e.g. figure 1.4.) The criteria have been introduced by Akaike (1973, 1978) and they may be justified by stochastic considerations.

1.6 Exercises

(1.1) Prove the Cauchy Schwarz inequality for vectors in $\mathbf{R}^n$, i.e. for two vectors $x, y \in \mathbf{R}^n$ we have

$$|x'y| \leq \|x\|\|y\|$$

and equality holds if and only if x and y are linearly dependent. (Hint: Consider the OLS estimate of regressing y on x.)

This inequality also holds for random variables x and y in the sense that $|\mathrm{E}\,xy| \leq \sqrt{\mathrm{E}\,x^2\,\mathrm{E}\,y^2}$.

(1.2) Let $x_1, \dots, x_T$, $y_1, \dots, y_T$ be two scalar time series. Repeat the definition of $\bar{x}$, $\bar{y}$, $\widehat{\mathrm{Var}}(x)$, $\widehat{\mathrm{Cov}}(x,y)$ and $\widehat{\mathrm{Corr}}(x,y)$. Prove the following statements:

(a) $\widehat{\mathrm{Cov}}(x,y) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \bar{x}\bar{y}$

(b) For $a, b \in \mathbf{R}$ and $a \neq 0$ we have $|\widehat{\mathrm{Corr}}(ax+b, y)| = |\widehat{\mathrm{Corr}}(x,y)|$

(c) If $|\widehat{\mathrm{Corr}}(x,y)| = 1$ then there exists a perfect linear relation between x and y in the sense that there exist $a, b \in \mathbf{R}$ such that $y_t = ax_t + b$ for all $t = 1, \dots, T$.

(1.3) Consider the (scalar, non random) process $x_t = a + bt$, where a, b are two real constants. Show that the sample autocorrelation $\hat{\rho}(s)$ converges to $+1$ if the sample size T converges to infinity.

(1.4) Prove that the matrix $\hat{\Gamma}_T$ is always nonnegative.

$$\hat{\Gamma}_T = \begin{pmatrix} \hat{\gamma}(0) & \cdots & \hat{\gamma}(T-1) \\ \vdots & \ddots & \vdots \\ \hat{\gamma}(-T+1) & \cdots & \hat{\gamma}(0) \end{pmatrix} \geq 0.$$

Note that this is in general not true for the autocovariance function defined by

$$\hat{\hat{\gamma}}(s) = \begin{cases} \frac{1}{T-|s|} \sum_{t=\max(1,1-s)}^{\min(T,T-s)} (x_t - \bar{x})(x_{t+s} - \bar{x})^* & \text{for } |s| < T \\ 0 & \text{else} \end{cases}$$

(1.5) Suppose that $m_t = c_0 + c_1 t + c_2 t^2$, $t \in \mathbf{Z}$. Show that $m_t = \sum_{i=-2}^{2} b_i m_{t-i}$, where $b_2 = b_{-2} = -\frac{3}{35}$, $b_1 = b_{-1} = \frac{12}{35}$ and $b_0 = \frac{17}{35}$.

(1.6) Analyze the "Austrian industrial production" using the programm RATS. Proceed along the following guideline:

- plot of the time series
- basic statistics of data (mean, variance, ...)
- estimate and plot the autocorrelation function
- estimate trend- and seasonal- component by OLS
- plot of time series, trend, season, ...
- plot of residuals
- basic statistics of residuals
- estimate and plot the autocorrelation function of the residuals

(1.7) Given a realization (with 100 observations) of an AR-process. Try to estimate the order of this process by comparing the "in sample fit", the "out of sample one-step-ahead" prediction error and the information criteria AIC and BIC. (Hint: the true order is less than 10!)

2 Stationary Processes in Time Domain

In this chapter we introduce some basic ideas for stochastic processes; thereby emphasis is layed on the concept of (wide sense) stationarity and on the properties of the covariance function. We give a number of examples which describe important classes of stationary processes. The analysis of stochastic processes considerabely benefits from using Hilbert space formulation although this is not necessary for understanding our lecture. Therefore we introduce some basic Hilbert space terminology and results. This section is addressed mainly to mathematically oriented part of the audience. The last section is concerned with probabilities on the space of all trajectories and with strict stationarity.

2.1 Basic Definitions

From now on we will assume that the time series are generated by an underlying stochastic mechanism, by a stochastic process. This makes time series analysis part of inferential statistics and we can evaluate the quality of estimation procedures and tests.

We commence from an underlying probability space $(\Omega, \mathcal{A}, P)$ and we consider random variables $x_t : \Omega \to \mathbf{C}^n$ (i.e. measurable functions) or $x_t : \Omega \to \mathbf{R}^n$; to be more precise for the observed process we restrict ourselves to the real case, the complex notation is choosen for reasons that will become clear later. We write $x_t = (x_t^{(i)})^{i=1,\dots,n}; x_t^{(i)} : \Omega \to \mathbf{C}$.

Definition 2.1 *A* stochastic process *is a family of random variables* $(x_t | t \in \mathbf{T})$ *defined on* $(\Omega, \mathcal{A}, P)$.

For us in most cases, a stochastic process is a model for random phenomena evolving in *time*. Then $\mathbf{T} \subset \mathbf{R}$ is understood as a set of time points. We almost exclusively deal with the *discrete* time case here, where $\mathbf{T} = \mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$ or $\mathbf{T} = \mathrm{N} = \{1, 2, \dots\}$ (equidistant points).

Definition 2.2 *A function* $(x_t(\omega) | t \in \mathbf{T})$ *(for fixed* $\omega \in \Omega$*) is called a* realization *(or trajectory or sample path) of* $(x_t | t \in \mathbf{T})$.

If we assume that a given time series $x_t, t = 1, \dots, T$ is generated by an underlying stochastic process $(x_t | t \in \mathbf{T})$ we have to equate $x_t \in \mathbf{R}^n$ with the realization $x_t(\omega)$ of the random variable x_t.

Unless the contrary is stated explicetely the limit of a sequence of random variables will be understood in the *mean squares* sense which is defined as follows:

Definition 2.3 *Let* $(x_k | k \in \mathrm{N})$ *be a sequence of random variables. We will say that* $(x_k | k \in \mathrm{N})$ *converges to* x_0 *in mean squares sense if*

$$\mathrm{E}\, x_0^* x_0 < \infty$$

and

$$\lim_{k\to\infty} \mathrm{E}(x_k - x_0)^*(x_k - x_0) = 0$$

holds. We then use the symbol

$$x_0 = \operatorname*{l.i.m}_{k\to\infty} x_k$$

Note that the limit x_0 is unique a.e. (i.e. on a subset $\mathcal{M}$ of Ω with $P(\mathcal{M}) = 1$). Proof: Suppose that there are two limits x_0 and x of the sequence x_k. Then

$$\begin{aligned}
&\mathrm{E}(x_0 - x)^*(x_0 - x) = \\
&\mathrm{E}((x_0 - x_k) - (x - x_k))^*((x_0 - x_k) - (x - x_k)) = \\
&\mathrm{E}(x_0 - x_k)^*(x_0 - x_k) - \mathrm{E}(x_0 - x_k)^*(x - x_k) \\
&- \mathrm{E}(x - x_k)^*(x_0 - x_k) + \mathrm{E}(x - x_k)^*(x - x_k)
\end{aligned}$$

Now by assumption and by the Cauchy-Schwarz inequality all terms on the right hand side of the above equation converge to zero for $k \to \infty$. Thus we have $\mathrm{E}(x_0 - x)^*(x_0 - x) = 0$ and therefore $x_0 = x$ a.e.

The expectation is continuous with respect to limits in the mean squares sense i.e. $\mathrm{E}(\operatorname{l.i.m}_{k\to\infty} x_k) = \lim_{k\to\infty}(\mathrm{E}\, x_k)$ if x_k is a convergent sequence of random variables. (See exercises.)

The following theorem gives an important criterion for the existence of a limit in the mean squares sense:

Theorem 2.1 *(Riesz-Fisher Theorem) Given a sequence of random variables x_k with $x_k^* x_k < \infty$. There exist a random variable x_0 such that $x_0 = \operatorname{l.i.m} x_k$ if and only if $\lim \mathrm{E}(x_k - x_l)^*(x_k - x_l) = 0$ for $k, l \to \infty$.*

For the proof see for example Brockwell & Davis [3, p. 68–69]. The theorem above says there exists a limit in the mean squares sense if and only if x_k is a Cauchy sequence, where the distance between x_k and x_l is defined by $\mathrm{E}(x_k - x_l)^*(x_k - x_l)$.

For many cases, important properties of a stochastic process are described by the first and second moments.

Definition 2.4

(i) Let $\mathrm{E}\,|x_t^{(i)}| < \infty$, $\forall i = 1, \ldots, n, t \in \mathbf{T}$, then $(\mathrm{E}\, x_t | t \in \mathbf{T})$ is called the (population) mean function of $(x_t | t \in \mathbf{T})$.

(ii) Let $\mathrm{E}\, x_t^* x_t < \infty$, $t \in \mathbf{T}$; then the function

$$\begin{array}{rccl}
\gamma : & \mathbf{T}\times\mathbf{T} : & \longrightarrow & \mathbb{C}^{n\times n} \\
& (s,t) & \longmapsto & \gamma(s,t) = \mathrm{E}(x_s - \mathrm{E}\, x_s)(x_t - \mathrm{E}\, x_t)^*
\end{array}$$

(where * denotes the conjugate transpose) is called the (population) covariance function *of* $(x_t|t \in \mathbf{T})$.

γ can also be interpreted as a matrix $(\gamma_{ij})_{i,j=1,\ldots,n}$ whose entries are functions γ_{ij} : $\mathbf{T} \times \mathbf{T} \to \mathbf{C}$. (For real valued processes, of course all these entries are realvalued.) The function γ may be interpreted as an "address book" for linear dependence relations; in particular $\gamma_{ij}(s,t) = \mathrm{Cov}(x_s^{(i)}, x_t^{(j)}) = \mathrm{E}(x_s^{(i)} - \mathrm{E}\,x_s^{(i)})\overline{(x_t^{(j)} - \mathrm{E}\,x_t^{(j)})}$ describes the linear dependence between $x_s^{(i)}$ and $x_t^{(j)}$. Clearly $(\mathrm{E}\,x_t|t \in \mathbf{T})$ describes the levels of the process.

Note that $\mathrm{E}\,x_t^* x_t < \infty$ implies that $\mathrm{E}\,x_t$ exists and that γ exists. (See exercises; the existence of γ is a consequence of the Cauchy-Schwarz inequality.)

Unless the contrary has been stated explicetely we let $\mathbf{T} = \mathbf{Z}$; then the process is written as (x_t).

We now consider a special class of stochastic processes, namely those that are stationary (in the wide sense), i.e. whose first and second moments are invariant to translation on the time axis $\mathbf{T}$ ($= \mathbf{Z}$ in our case).

Definition 2.5 *A stochastic process is called (wide sense)* stationary *if*

(i) $\mathrm{E}\,x_t^* x_t < \infty$ *for all* $t \in \mathbf{Z}$

(ii) $\mathrm{E}\,x_t = m = const$ *for all* $t \in \mathbf{Z}$ *and*

(iii) $\gamma(s,t) = \gamma(s+r, t+r)$ *holds for all* $r, s, t \in \mathbf{Z}$.

For a stationary process in particular the covariances only depend on time differences. As γ depends on one argument we write

$$\gamma(s) = \gamma(s,0) = \gamma(s+r, r)$$

and use γ for a function $\mathbf{Z} \to \mathbf{C}^{n \times n}$.

The function $\gamma_{ii} : \mathbf{Z} \to \mathbf{C}$ is called the *autocavariance function* of the i-th component process $(x_t^{(i)}|t \in \mathbf{Z})$. The funtion $\gamma_{ij} : \mathbf{Z} \to \mathbf{C}$, $i \neq j$ is called the *cross-covariance function* between the i-th and the j-th component process.

We may alternative write (x_t) as double infinite vector

$$x = \begin{pmatrix} \vdots \\ x_{-1} \\ x_0 \\ x_1 \\ \vdots \end{pmatrix}$$

and Γ as its (double infinite) covariance matrix

$$\Gamma = \mathrm{E}\, x x^{*} - (\mathrm{E}\, x)(\mathrm{E}\, x)^{*} = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \gamma(0) & \gamma(-1) & \gamma(-2) & \ddots \\ \ddots & \gamma(1) & \gamma(0) & \gamma(-1) & \ddots \\ \ddots & \gamma(2) & \gamma(1) & \gamma(0) & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

Note: $(0,0)$ element has to be assigned. Γ is a *Block-Toeplitz* matrix; i.e. all block entries in the main diagonal are identical, the same holds for parallel diagonals.

Finite parts of this matrix are e.g.

$$\Gamma_T = \mathrm{E} \begin{pmatrix} x_1 - m \\ \vdots \\ x_T - m \end{pmatrix} \begin{pmatrix} x_1 - m \\ \vdots \\ x_T - m \end{pmatrix}^{*} =$$

$$\begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(-T+1) \\ \gamma(1) & \gamma(0) & & \vdots \\ \vdots & & \ddots & \vdots \\ \gamma(T-1) & \cdots & \cdots & \gamma(0) \end{pmatrix} \in \mathbb{C}^{nT \times nT}, \text{ where } m = \mathrm{E}\, x_t .$$

In most cases we assume $\mathrm{E}\, x_t = 0$; i.e. we replace the original process $(\tilde{x}_t)$ by a centered one: $x_t = \tilde{x}_t - \mathrm{E}\, \tilde{x}_t$. The assumption of constant mean for stationary processes turns out to be of no importance for the theory.

Stationary processes occur in stable random "mechanisms" driven with constant energy if they are in steady state.

The theory of stationary processes however is important also for nonstationary time series; such series are often transformed (e.g. by trend regressions or differencing) to stationary ones.

The importance of stationary processes for statistics is that for a wide class of such processes (the class of ergodic processes, see chapter 6) a single trajectory of such a process displays the whole probability law of the process. Thus averaging over time gives the same result as averaging over population; the sample moments converge to their population counterparts. In particular:

$$
\begin{aligned}
\lim \frac{1}{T}\sum_{t=1}^{T} x_t &= \mathrm{E}\, x_t \qquad \text{a.e.} \\
\lim \frac{1}{T}\sum_{t=\max(1,1-s)}^{\min(T,T-s)} (x_{t+s}-\bar{x})(x_t-\bar{x})^* &= \mathrm{E}\, x_s x_0^* \qquad \text{a.e.}
\end{aligned}
\qquad (2.1)
$$

(i.e. on a subset of Ω of probability 1)

Equations (2.1) define invariants for almost all trajectories. Such trajectories will be called typical.

2.2 Examples for stationary processes

2.2.1 White Noise Process

A stochastic process (ϵ_t) satisfying

(i) $\mathrm{E}\,\epsilon_t = 0$

(ii) $\mathrm{E}\,\epsilon_s \epsilon_t^* = \delta_{st}\Sigma \quad (\Sigma \geq 0)$

is called *white noise.*

Clearly (ϵ_t) is stationary. A white noise process has no (linear) "memory" and constant variance.

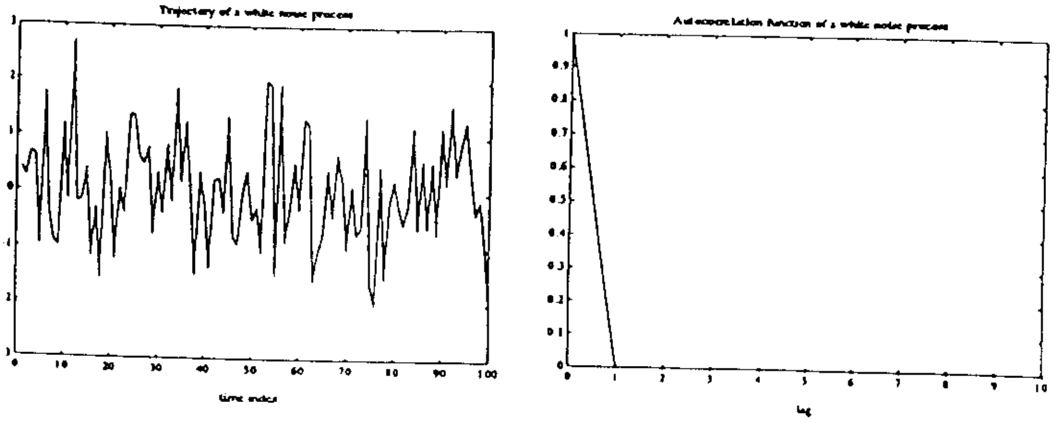


Figure 2.1: "Typical" trajectory and the autocorrelation function of a white noise process

2.2.2 Moving Average Process

Let (ϵ_t) be white noise; a process given by

$$x_t = \sum_{j=0}^{q} b_j \epsilon_{t-j} \quad ; \quad b_j \in \mathbf{R}^{m \times n}, \text{ for all } t \tag{2.2}$$

is called a *moving average (MA) process* (of order q if $b_0 \neq 0$ and $b_q \neq 0$).

We have

$$\mathrm{E}\, x_t^* x_t < \infty \text{ (see exercises)}, \ \mathrm{E}\, x_t = 0$$

and

$$\gamma(s,t) = \mathrm{E}(\sum_{i=0}^{q} b_i \epsilon_{s-i})(\sum_{j=0}^{q} b_j \epsilon_{t-j})^* \underbrace{=}_{\left(\substack{s-t=i-j \\ j=i-s+t}\right)} \begin{cases} \sum_{\substack{0 \le i \le q \\ 0 \le i-s+t \le q}} b_i \Sigma b_{i-s+t}^* \\ 0 \text{ for } |s-t| > q \end{cases}$$

depends on $(s-t)$ only. Thus an MA process is stationary. An MA process has a finite (linear) memory (i.e. $\gamma(s) = 0$ for some q and all $|s| > q$). Thus Γ is a band matrix. Conversely every stationary finite memory process has an MA representation (2.2) with a suitably white noise process (ϵ_t) (needs proof).

2.2.3 Infinite Moving Average Process

Let (ϵ_t) be white noise; a process given by

$$x_t = \sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j} \quad ; \quad b_j \in \mathbf{R}^{m \times n} \text{ for all } t \tag{2.3}$$

is called an *infinite moving average process.*

We interpret the infinite sum (2.3) as the (mean squares sense) limit of the sequence (of partial sums) $(\sum_{j=-N}^{N} b_j \epsilon_{t-j} | N \in \mathrm{N})$ We have for $\underline{N \geq M \geq 0}$

$$\mathrm{E}\left(\sum_{j=-N}^{N} b_j \epsilon_{t-j} - \sum_{j=-M}^{M} b_j \epsilon_{t-j}\right)^{\overbrace{\hphantom{x}}} \left(\sum_{j=-N}^{N} b_j \epsilon_{t-j} - \sum_{j=-M}^{M} b_j \epsilon_{t-j}\right) = \sum_{M \leq |j| \leq N} b_j^* \Sigma b_j$$

and thus, since for a sequence of random variables y_k

$$\mathrm{E}\, y_k^* y_k < \infty \text{ and } \mathrm{E}(y_k - y_l)^*(y_k - y_l) \to 0 \text{ for } k, l \to \infty$$

is equivalent to the existence of a random variable y such that $\mathrm{l.i.m}\, y_k = y$ the infinite sum (2.3) exists iff

$$\sum_{j=-\infty}^{+\infty} b_j^* \Sigma b_j < \infty \tag{2.4}$$

holds.

Prove: If $\Sigma > 0$ then

$$\sum_{j=-\infty}^{+\infty} b_j^* \Sigma b_j < \infty \Longleftrightarrow \sum_{j=-\infty}^{+\infty} b_j^* b_j < \infty.$$

We further have

$$\mathrm{E}\, x_t^* x_t < \infty$$

(by definiton of the infinite sum in (2.3))

$$\mathrm{E}\, x_t = \mathrm{E} \sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j} = \sum_{j=-\infty}^{\infty} b_j\, \mathrm{E}\, \epsilon_{t-j} = 0$$

(by the continuity of the expectation) and

$$\mathrm{E}\, x_s x_t^* = \mathrm{E}(\sum_{i=-\infty}^{\infty} b_i \epsilon_{s-i})(\sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j})^* \underbrace{=}_{\binom{s-i=t-j}{j=t-s+i}} \sum_{i=-\infty}^{\infty} b_i \Sigma b_{t-s+i}^*$$

and thus x_t is stationary.

If $b_j = 0$ for $j < 0$ holds, (x_t) is called a *one sided* (or *causal*) infite moving average of white noise.

Infinite moving averages represent allready a wide class of stationary processes; in general they have infinite (linear) memory, which however fades.

2.2.4 Autoregressive Process

Consider the linear difference equation of the form

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} = \epsilon_t, \tag{2.5}$$

where $a_j \in \mathbf{R}^{n \times n}$ and ϵ_t is white noise. Then (2.5) is called an AR system. A solution on $\mathbf{Z}$ (or on $\mathbf{Z}_+$) of (2.5) (i.e. a process y_t satisfying (2.5) for all $t \in \mathbf{Z}$ (or all $t \in \mathbf{Z}_+$) for given a_j and (ϵ_t)) is called an *autoregressive* (AR) process (of order p if $a_p \neq 0$ holds). The name comes from the fact that y_t depends on his own past. AR processes will be discussed in chapter 3 in detail. We only consider a special case, namely scalar (i.e. n=1) AR processes of order one; these are generated by

$$y_t = a y_{t-1} + \epsilon_t \tag{2.6}$$

where we always assume $\sigma^2 = \mathrm{E}\epsilon_t^2 > 0$. (note that a in (2.6) corresponds to $-a_1$ in equation (2.5).)

First we consider solutions on $\mathbf{Z}_+$, starting from an initial value y_0: From iterative (foreward) substitution we obtain

$$y_t = \sum_{j=0}^{t-1} a^j \epsilon_{t-j} + a^t y_0 \tag{2.7}$$

Assume that y_0 is nonstochastic for the moment. Then

$$\begin{aligned} \mathrm{E}\, y_t &= a^t y_0 \\ \mathrm{Var}\, y_t &= \textstyle\sum_{j=0}^{t-1} a^{2j} \\ \mathrm{Cov}(y_s, y_t) &= \mathrm{E}(\textstyle\sum_{i=0}^{s-1} a^i \epsilon_{s-i})(\sum_{j=0}^{t-1} a^j \epsilon_{t-j}) = \sigma^2 \sum_{i=1}^{\min(t,s)} a^{t+s-2i}. \end{aligned} \tag{2.8}$$

We can distinguish the following cases

(i) If $a = 0$ then clearly $y_t = \epsilon_t$; in all other cases (y_t) is nonstationary.

(ii) For $|a| < 1$ we have

$$\mathrm{E}\, y_t \to 0 \text{ for } t \to \infty$$

and

$$\mathrm{Var}\, y_t = \sum_{j=0}^{t-1} a^{2j} \to \frac{1}{1-a^2}\sigma^2 \text{ for } t \to \infty.$$

For these reasons, this is called the *stable* case.

(iii) For $|a| > 1$ we have

$$|\,\mathrm{E}\, y_t| = |a|^t |y_0| \to \infty \text{ for } t \to \infty$$

and

$$\mathrm{Var}\, y_t = \sum_{j=0}^{t-1} a^{2j} = \frac{1-a^{2t}}{1-a^2}\sigma^2 \to \infty \text{ for } t \to \infty.$$

This is the *exponentially unstable* case.

(iv) For $|a| = 1$ we have

$$|\,\mathrm{E}\, y_t| = |y_0|$$

and

$$\mathrm{Var}\, y_t = \sigma^2 t.$$

This is still *unstable* (the variance is exploding) but not exponentially unstable. In particular for $a = 1$ we get the *random walk*

$$y_t = \sum_{j=0}^{t-1} \epsilon_{t-j} + y_0 \tag{2.9}$$

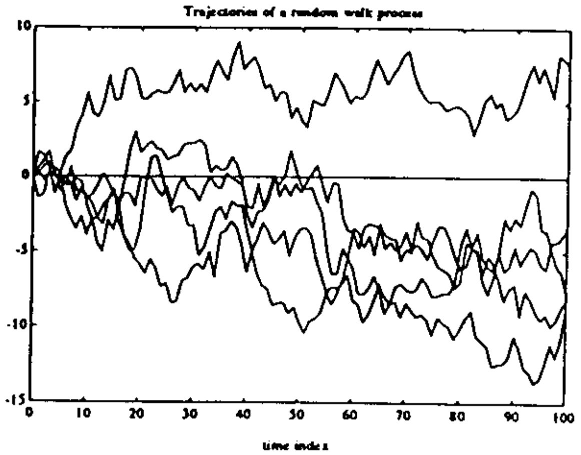


Figure 2.2: Some "typical" trajectories of a random walk process.

Now we do not start the system at time $t = 0$ with y_0, but at time $t = -T$ with initial value y_{-T}. Then for the stable case, in the limit for $T \to \infty$ we will get a solution on $\mathbf{Z}$ of the form

$$y_t = \sum_{j=0}^{\infty} a^j \epsilon_{t-j} \tag{2.10}$$

(Note that sum on the right hand side of (2.10) exists since $\sum_{j=0}^{\infty} a^{2j} < \infty$ holds.) This solution is called the ***steady state solution*** since it is the solution of the stable system "started in the infinite past". It is clear that every y_t (independent of y_0) of the form (2.7), for $|a| < 1$, will converge to y_t defined by (2.10) for $t \to \infty$.

Note that the set of all solutions of (2.5) on $\mathbf{Z}$ can be written as the particular solution (2.7) plus the set of all solutions of the homogenous equation

$$y_t - ay_{t-1} = 0$$

Here, almost exclusevely, we consider the steady state solution (2.10). Note that (2.10) gives a one sided, infinite moving average and thus has mean zero and is stationary.

Positive autocorrelation ($a > 0$) gives "clustering". AR(1) processes have a geometrically fading (linear) memory.

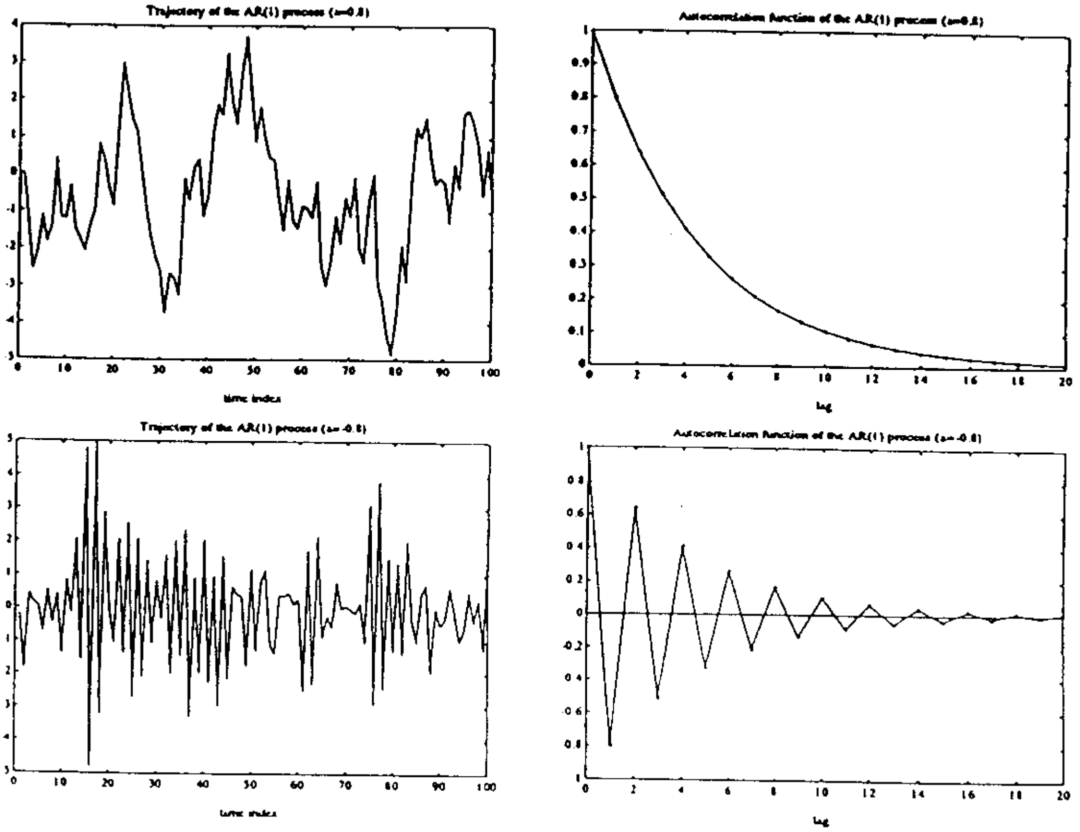


Figure 2.3: "Typical" trajectories and the autocorrelation function of the AR(1) processes, $y_t = 0.8y_{t-1} + \epsilon_t$ and $y_t = -0.8y_{t-1} + \epsilon_t$.

2.2.5 ARMA Process

Consider a linear difference equation of the form

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} = b_0 \epsilon_t + b_1 \epsilon_{t-1} + \cdots + b_q \epsilon_{t-q} \tag{2.11}$$

where $a_j, b_j \in \mathbf{R}^{n \times n}$ and (ϵ_t) is white noise. Then (2.11) is called an ARMA system and a solution of (2.11) is called an *ARMA process*. ARMA systems will be discussed in detail in chapter 3.

2.2.6 Harmonic Processes

Consider a process (x_t) defined by

$$x_t = \sum_{j=1}^{h} e^{i\lambda_j t} z_j \tag{2.12}$$

where $z_j : \Omega \to \mathbf{C}^n$, $j = 1, \ldots, h$ are (genuine complex) n-dimensional random variables satisfying

$$\mathrm{E}\, z_j^* z_j < \infty \quad ; \quad j = 1, \ldots, h$$

W.r.g we assume $\lambda_1 < \lambda_2 < \cdots < \lambda_h$. Processes of this form are called *cyclical or harmonic processes*. They are mainly interesting because of the insights and interpretations they provide, rather than as (complete) models for time series. From $e^{i\lambda_j t} = \cos \lambda_j t + i \sin \lambda_j t$, it is clear that (x_t) is the sum of a finite number of trigonometric time functions (in discrete time) with random weights z_j. The entries of z_j, $z_j^{(l)} : \Omega \to \mathbf{C}$ can be written as

$$z_j^{(l)} = |z_j^{(l)}| e^{i\phi_j^{(l)}}$$

where $|z_j^{(l)}|$ and $\phi_j^{(l)}$ are random variables and describe the amplitude and the phase shifts for the time function $e^{i\lambda_j t}$. The λ_j is called the *angular frequency*; it is related by $T_j = \frac{2\pi}{\lambda_j}$ to the period length T_j. Note that we may restrict ourselves to the interval $(-\pi, \pi]$ for the angular frequencies λ_j. This is a consequence of the fact that we consider time functions on $\mathbf{Z}$ and is easily seen as follows: For arbitrary $\lambda_j \in \mathbf{R}$, we write $\lambda_j = k2\pi + \tilde{\lambda}_j$ where $\tilde{\lambda}_j \in (-\pi, \pi]$ and $k \in \mathbf{Z}$. Then

$$e^{i\lambda_j t} = e^{i(k2\pi + \tilde{\lambda}_j)t} = e^{i\tilde{\lambda}_j t} \text{ for all } t \in \mathbf{Z}$$

as $e^{ik2\pi t} = 1$ for all $t \in \mathbf{Z}$. In other words we could never distinguish between the time function $e^{i\lambda_j t}$ and $e^{i\tilde{\lambda}_j t}$ defined on $\mathbf{Z}$. The frequency π which is the highest frequency which can be observed is called the *Nyquist frequency*. The Nyquist frequency evidently

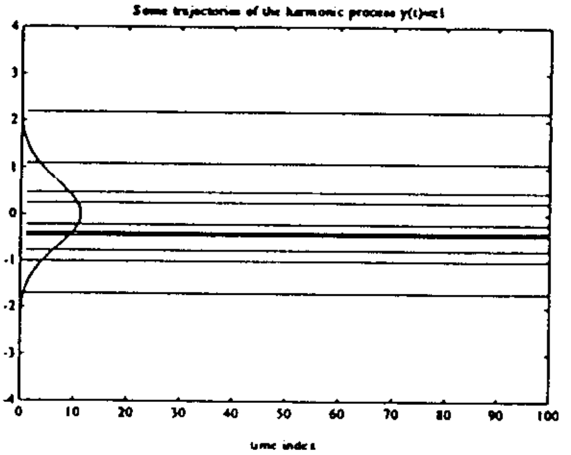


Figure 2.4: "Typical" trajectories of the harmonic process $x_t = z_1$; Clearly such a process is not ergodic for the mean.

corresponds to a period length of 2; thus e.g. for quarterly data we can at best observe sinusoids (or cosines) with a period length of 2 quarters.

Now let us investigate under which conditions (x_t) will be stationary. We have

$$\mathrm{E}\, x_t = \sum_{j=1}^{h} e^{i\lambda_j t}\, \mathrm{E}\, z_j \tag{2.13}$$

Note that the class of functions $e^{i\lambda_j t} : \mathbb{Z} \to \mathbb{C}$, $\lambda_j \in (-\pi, \pi]$ are linearly independent. This can be shown as follows: Consider the linear combination

$$\sum_{j=1}^{h} c_j e^{i\lambda_j t} = 0 \quad ; \quad \lambda_1 < \lambda_2 < \cdots < \lambda_h .$$

This implies

$$0 = \sum_{j=1}^{h} c_j \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} e^{i\lambda_j t} e^{-i\lambda_l t} \tag{2.14}$$

Now since

$$\frac{1}{T} \sum_{t=0}^{T} e^{i\lambda_j t} e^{-i\lambda_l t} = \begin{cases} 1 & \text{for } j = l \\ \frac{1}{T} \frac{1 - e^{i(\lambda_j - \lambda_l)T}}{1 - e^{i(\lambda_j - \lambda_l)}} & \text{for } j \neq l \end{cases}$$

we have

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} e^{i\lambda_j t} e^{-i\lambda_l t} = \begin{cases} 1 \text{ for } j = l \\ 0 \text{ for } j \neq l \end{cases}$$

Therefore, we have from (2.14)

$$c_j = 0 \quad ; \quad j = 1, \ldots, h$$

which proves the statement. But then (2.13) immediateley implies the mean is constant if and only if $\mathrm{E}\, z_j = 0$ holds for all $\lambda_j \neq 0$. We thus have from (2.13):

$$\mathrm{E}\, x_t = \begin{cases} \mathrm{E}\, z_j & \text{for } \lambda_j = 0 \\ 0 & \text{if all } \lambda_j \text{ are unequal to zero.} \end{cases} \tag{2.15}$$

Next, consider the (noncentral) covariance function

$$\mathrm{E}\, x_s x_t^* = \mathrm{E}(\sum_{j=1}^{h} e^{i\lambda_j s} z_j)(\sum_{l=1}^{h} e^{i\lambda_l t} z_l)^* = \sum_{j,l} e^{i(\lambda_j s - \lambda_l t)}\, \mathrm{E}\, z_j z_l^*$$

If we assume in addition

$$\mathrm{E}\, z_j z_l^* = 0 \text{ for all } j \neq l \tag{2.16}$$

then

$$\mathrm{E}\, x_s x_t^* = \sum_{j=1}^{h} e^{i\lambda_j (s-t)}\, \mathrm{E}\, z_j z_j^* \tag{2.17}$$

and thus (x_t) is stationary. Conversely if (2.16) does not hold, then looking at $\mathrm{E}\, x_s x_s^*$, again due to the linear indepency of the $e^{i\lambda_j t}$, we see that (x_t) is not stationary. Note that γ is a periodic function (on $\mathbb{Z}$) for rational $\lambda_j/(2\pi)$ and then in particular we have perfect correlation (i.e. the correlation is either 1 or -1) at certain lags.

In particular a harmonic process has an infinite and in particular not fading memory.

Since we are only interested in real time series, we are only interested in $\mathbb{R}^n$-valued random variables x_t. This implies:

$$0 = \Im\{\sum_j e^{i\lambda_j t} z_j\} = \sum_j (\Im\{z_j\} \cos \lambda_j t + \Re\{z_j\} \sin \lambda_j t)$$

(here $\Im$ and $\Re$ denote the imaginary and real part respectively). This holds if and only if

$$\begin{matrix} \lambda_{1+j} = -\lambda_{h-j} \text{ and} \\ z_{1+j} = \bar{z}_{h-j} \end{matrix} \quad ; \quad \text{for } j = 0, \ldots, \lfloor (h-1)/2 \rfloor . \tag{2.18}$$

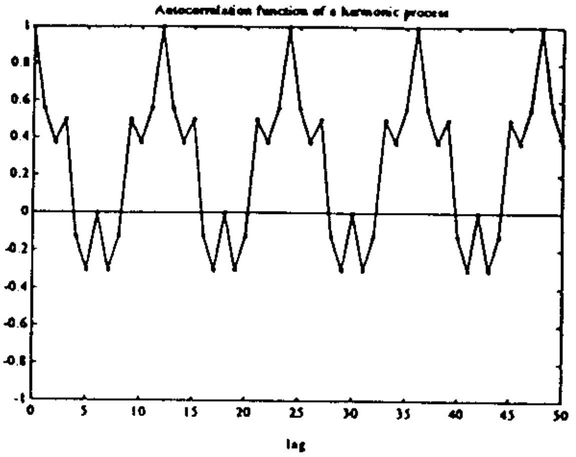


Figure 2.5: Autocorrelation function of the harmonic process $x_t = z_1 e^{i\lambda_1 t} + z_2 e^{i\lambda_2 t} + z_3 + z_4 e^{i\lambda_4 t} + z_5 e^{i\lambda_5 t}$, where $\lambda_5 = -\lambda_1 = 2\pi/3$, $\lambda_4 = -\lambda_2 = \pi/6$, $\mathrm{E}\, z_1 \overline{z_1} = \mathrm{E}\, z_5 \overline{z_5} = 0.5$ and $\mathrm{E}\, z_2 \overline{z_2} = \mathrm{E}\, z_4 \overline{z_4} = \mathrm{E}\, z_3 \overline{z_3} = 1$. The picture shows perfect correlation at lags $12k$, $k \in \mathbb{Z}$.

Thus we may write (2.12) as

$$x_t = \sum_{j=1}^{h} (\Re\{z_j\} \cos \lambda_j t - \Im\{z_j\} \sin \lambda_j t) \tag{2.19}$$

If we define "new" frequencies $\nu_j = \lambda_{j+\lfloor h/2 \rfloor}$, $j = 1, \ldots, \lfloor (h+1)/2 \rfloor$ ($\nu_j \in [0, \pi]$) we may rewrite (2.19) as

$$x_t = \sum_{j=1}^{\lfloor (h+1)/2 \rfloor} a_j \cos \nu_j t + b_j \sin \nu_j t$$

where

$$\begin{aligned} a_1 &= \Re\{z_{(h+1)/2}\} && \text{- for odd } h \\ b_1 &= 0 && \text{- for odd } h \\ a_j &= 2\Re\{z_{j+\lfloor h/2 \rfloor}\} && \text{- for even } h \text{ or } j > 1 \\ b_j &= 2\Im\{z_{j+\lfloor h/2 \rfloor}\} && \text{- for even } h \text{ or } j > 1 \end{aligned}$$

(Note that for odd h, $\lambda_{(h+1)/2} = \nu_1 = 0$ and $z_{(h+1)/2}$ must be real. For h even the frequency 0 is not contained in the set of frequncies.)

Clearly from (2.15) and (2.17) we see that the first and second moments of the z_j, $j = 1, \ldots, h$ uniquely determine the first and second moments of (x_t).

Conversely, from $\mathrm{E}(x_s x_0^*)$, $s \in \mathbf{Z}$ (and $\mathrm{E}\, x_0$) the $F_j = \mathrm{E}\, z_j z_j^*$ (and $\mathrm{E}\, z_j$, $\lambda_j = 0$) are uniquely determined, by (2.17) (and (2.15)); thus there is a one to one correspondance between the non-central covariance function of (x_t) and the function $(F_j | j = 1, \ldots, h)$ or the so called ***spectral distribution function***

$$\begin{array}{rcl} F: \ (-\pi, \pi]: & \longrightarrow & \mathbf{R}^{n \times n} \\ \lambda & \longmapsto & \sum_{j:\lambda_j \leq \lambda} F_j \end{array}$$

Now let us give some additional interpretation for a harmonic process, first for the scalar case ($n = 1$):

From (2.12) it is clear that, that for $\lambda_j \neq 0$, $F_j = \mathrm{Var}\, z_j = \mathrm{E}|z_j|^2$ is a measure for the expected amplitude of the frequency component $e^{i\lambda_j t}$. For $\lambda_j = 0$ we have $F_j = \mathrm{E}|z_j|^2 = \mathrm{Var}\, z_j + |\mathrm{E}\, x_t|^2$ which has an analogous interpretation.

Thus the spectral distribution function $F(\lambda)$, which is in a one-to-one relation with the (noncentral) covariance function and therefore contains the same information, displays this information about the underlying process in a different form. In particular we see that F is a monotonic, nonnegative, right continuous step function; the values at which jumps occur show the frequencies of the underlying process and the step sizes are measures for the expected amplitudes.

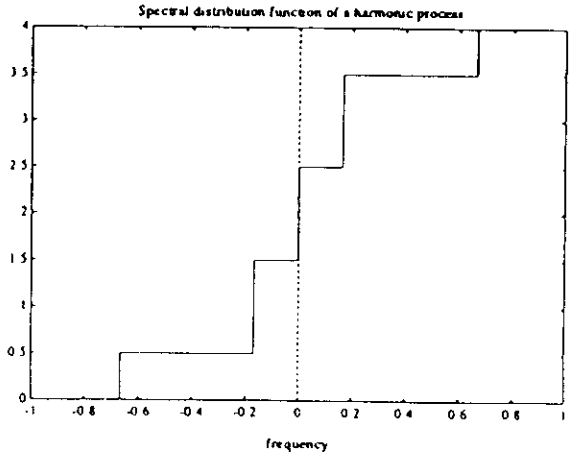


Figure 2.6: Spectral distribution function of the scalar harmonic process defined in figure (2.5)

In the multivariate case (i.e. $n > 1$), the interpretation of the off-diagonal elements of the spectral distribution function F can be seen from the following considerations: The

il-th entry of the complex matrix F_j is given by

$$F_{il,j} = \mathrm{E}\, z_j^{(i)} \bar{z}_j^{(l)}$$

where $z_j^{(i)}$ denotes the i-th component of z_j. It provides a measure of the linear dependence between the i-th and the l-th component at frequency λ_j in terms of its absolute value and an measure of the expected phase shift between the two components in terms of its phase.

2.3 Properties of Covariance Functions

In this short section some properties of covariance functions are given.

Definition 2.6 *A function $\gamma : \mathbf{Z} \to \mathbf{C}^{n \times n}$ is called* nonnegative definite *if*

$$\sum_{p,q=1}^{T} a_p \bar{a}_q \, \gamma_{i_p i_q}(t_p - t_q) \geq 0 \tag{2.20}$$

holds for arbitrary chosen $T \in \mathrm{N}$, $a_1, \ldots, a_T \in \mathbf{C}$, $i_1, \ldots, i_T \in \{1, \ldots, n\}$ and $t_1, \ldots, t_T \in \mathbf{Z}$.

Note that a function γ is nonnegative definite iff the matrices

$$\Gamma_T = \begin{pmatrix} \gamma(0) & \cdots & \gamma(-T+1) \\ \vdots & \ddots & \vdots \\ \gamma(T-1) & \cdots & \gamma(0) \end{pmatrix}$$

are nonnegative definite. Note also that in the complex case (as opposed to the real case) a nonnegative matrix is hermitean, i.e. $\Gamma_T = \Gamma_T^*$ holds.

The following theorem provides a mathematical characterization of a covariance function (and thus also for correlation functions).

Theorem 2.2 *A function $\gamma : \mathbf{Z} \to \mathbf{C}^{n \times n}$ is a covariance function of a stationary process if and only if γ is nonnegative.*

Proof: For the "$\Leftarrow$" part see e.g. Rozanov [5]. We here will only prove the easier "$\Rightarrow$" part; we have

$$\begin{array}{c} \sum_{p,q=1}^{T} a_p \bar{a}_q \, \gamma_{i_p i_q}(t_p - t_q) = \\ \sum_{p,q=1}^{T} a_p \bar{a}_q \, \mathrm{E}(x_{t_p}^{(i_p)} \bar{x}_{t_q}^{(i_q)}) = \\ \mathrm{E}\left|\sum_{p=1}^{T} a_p x_{t_p}^{(i_p)}\right|^2 \geq 0 \end{array}$$

where the last inequality holds since the integrand is a monotonic functional (i.e. $f \geq 0$ implies $\int f \, dP \geq 0$).

Note that the theorem above is a "natural" extension of the fact that a matrix Σ is a covariance matrix iff Σ is symmetric and nonnegative definite. For an arbitrary nonnegative function, the corresponding process may be chosen as Gaussian.

In particular, we have the following properties of the covariance function

(i) $\gamma(0) \geq 0$; $\gamma(0) = 0 \Leftrightarrow x_t = 0$ a.s.

(ii) $\gamma(s) = \mathrm{E}\, x_s x_0^* = \mathrm{E}(x_0 x_s^*)^* = \gamma(-s)^*$

2.4 Hilbert-spaces

Hilbert spaces are important for understanding the geometry of stationary processes and they provide powerful tools in particular in connection with linear approximation problems. The reader should know that the complete lecture could also be understood without the knowledge of Hilbert spaces. However even a basic knowledge of Hilbert spaces make the understanding of a number of things easier.

Definition 2.7 *A mapping* $< \cdot, \cdot >: \mathbf{H} \times \mathbf{H} \rightarrow \mathbb{C}$ *where* $\mathbf{H}$ *is a linear space is an* inner product *if*

(i) $< a_1 x_1 + a_2 x_2, y > = a_1 < x_1, y > + a_2 < x_2, y >$ *for all* $a_1, a_2 \in \mathbb{C}$ *and* $x_1, x_2, y \in \mathbf{H}$.

(ii) $< x, y > = \overline{< y, x >}$

(iii) $< x, x > \geq 0$ *and* $< x, x > = 0 \Longleftrightarrow x = 0$.

Definition 2.8 *A set* $\mathbf{H}$ *is a* Hilbert space *if*

(i) $\mathbf{H}$ *is a linear space*

(ii) with an inner product

(iii) which is complete in the norm defined by the inner product.

Here *complete* means that every *Cauchy sequence* (i.e. a sequence $x_n \in \mathbf{H}$ which satisfies $\lim_{m,n \rightarrow \infty} \|x_n - x_m\| = 0$) converges; i.e. there exists an $x \in \mathbf{H}$ such that $\lim_{n \rightarrow \infty} \|x - x_n\| = 0$ holds.

2.4.1 Examples

(i) $\mathbf{R}^n$ together with the innerproduct defined by $< x, y >= \sum_{i=1}^n x_i y_i$ is a Hilbert space.

(ii) $\mathbf{C}^n$ together with the innerproduct defined by $< x, y >= \sum_{i=1}^n x_i \overline{y_i}$ is a Hilbert space.

(iii) Consider the probability space $(\Omega, \mathcal{A}, P)$ and let $\mathcal{L}_2(\Omega, \mathcal{A}, P)$ be the set of all random variables $x : \Omega \to \mathbf{C}$ defined over this probability space which are square integrable (i.e. $\mathrm{E}\,|x|^2 < \infty$). We define an equivalence relation $x \equiv y$ on $\mathcal{L}_2$ by $x \equiv y$ iff $x = y$ a.s. The set of these equivalence classes is denoted by $\mathbf{L}_2(\Omega, \mathcal{A}, P)$.

It is easy to prove that $\mathbf{L}_2$ is a linear space and that $< x, y >= \mathrm{E}\,x\overline{y}$ is welldefined and an inner product. (It is clear that $< x, y >$ defined over $\mathcal{L}_2$ is not an inner product since $\mathrm{E}\,x^2 = 0$ does not imply that $x = 0$.)

From the Riesz-Fisher Theorem we also know that $\mathbf{L}_2$ is complete and thus is a Hilbert space.

From the definition of the inner product it is clear that $< x, y >$ is the (non central) Covariance of the two random variables and that $\|x\|^2 =< x, x >$ is the (noncentral) variance of the random variable x. The limit in "Hilbert space" sense is nothing else than the limit in the mean squares sense.

Consider a stationary scalar proces (x_t). The stationarity conditions imply that

$$\begin{aligned} \mathrm{E}\,x_t &=< x_t, 1 >= \text{ const} \\ \mathrm{E}\,x_t^2 &=< x_t, x_t >= \|x_t\|^2 = \text{ const} \\ \mathrm{E}\,x_t x_{t-1} &=< x_t, x_{t-1} >= \text{ const} \\ &\vdots \end{aligned}$$

This means that all x_t are vectors in the Hilbert space $\mathbf{L}_2$ of the same length and the angle of x_t and random variable 1 and the angle between x_t and x_{t-1} is constant and does not depend on t and the same holds for the angle between x_t and x_{t-s}.

Therefore there exists a unitary operator U (which is the generalization of an orthonogal matrix in $\mathbf{R}^n$ and corresponds to rotations or reflections) such that U generates the process (x_t) by:

$$x_t = U^t x_0.$$

The next theorem provides an extremly useful tool for the solution of linear approximation problems. For its proof see e.g. Brockwell and Davis [3, p.51f].

Theorem 2.3 *(Projection Theorem) Let* $\mathbf{H}$ *be a Hilbert space and* M *be a subspace (i.e. a subspace which is a Hilbert space of its own). Then for every* $x \in \mathbf{H}$ *there exists unique decomposition*

$$x = \hat{x} + u$$

such that $\hat{x} \in M$ *and* $u \perp M$ *(i.e.* $< u, y >= 0$ *for all* $y \in M$*). In addition* $\hat{x}$ *is the unique element of* M *satisfying*

$$\|x - \hat{x}\| = \min_{y \in M} \|x - y\|$$

2.5 Exercises

(2.1) Discuss the terms: probability space, random variables, expectation, variance, distribution function and probability density function.

(2.2) Let x, y be two scalar (real) random variables. Repeat the definition of $\mathrm{Var}(x)$, $\mathrm{Cov}(x,y)$ and $\mathrm{Corr}(x,y)$. Prove the following statements:

(a) $\mathrm{Cov}(x,y) = \mathrm{E}\,xy - (\mathrm{E}\,x)(\mathrm{E}\,y)$

(b) For $a, b \in \mathbf{R}$ and $a \neq 0$ we have $|\mathrm{Corr}(ax+b, y)| = |\mathrm{Corr}(x,y)|$

(c) If $|\mathrm{Corr}(x,y)| = 1$ then there exists a perfect linear relation between x and y in the sense that there exist $a, b \in \mathbf{R}$ such that $y = ax + b$ a.e.

(2.3) Prove the following statements: (x is an n-dimensional random variable.)

(a) $\mathrm{E}\,x^* x < \infty \Rightarrow \mathrm{E}\,x$ exists and is finite.

(b) $\mathrm{E}\,x^* x < \infty \Rightarrow \mathrm{E}\,x x^* < \infty$.

(2.4) Consider a white noise process (ϵ_t). Prove that the process

$$x_t = \sum_{j=0}^{q} b_j x_{t-j} \quad ; \quad b_j \in \mathbf{R}^{m \times n} \text{ for all } t \in \mathbb{Z}$$

is a stationary process. For the scalar case $\epsilon_t \in \mathbf{R}$, $b_j \in \mathbf{R}$ and $q = 1$

$$x_t = b_0 x_0 + b_1 x_{t-1} \quad ; \text{ for all } t \in \mathbb{Z}$$

compute the autocorrelation function $\rho(s)$ of x_t. What is the possible range of values of $\rho(1)$?

(2.5) Prove the "continuity" of the expectation, i.e. if (x_k) is a sequence of random variables which converges in the mean square sense then

$$\mathrm{E}(\operatorname*{l.i.m}_{k\to\infty} x_k) = \lim_{k\to\infty}(\mathrm{E}\, x_k)$$

holds.

(2.6) Consider a sequence of (vector valued) random variables (x_k) $(x_k \in \mathbf{R}^n)$. Prove that (x_k) converges in mean squares sense if and only if each component $(x_k^{(i)})$ converges in mean squares sense:

$$\operatorname{l.i.m} x_k = x_0 \Longleftrightarrow \operatorname{l.i.m} x_k^{(i)} = x_0^{(i)} \text{ for all } i = 1,\ldots,n$$

Please discuss the stochastic processes defined in the next four examples. Compute their mean- and autocovariance- function. Are they stationary? What does a "typical" trajectory look like?

(2.7) Discuss the (scalar) stochastic process $(w_t \mid t \in \mathrm{N})$ defined by

$$\begin{aligned} w_0 &= 0 \\ w_t &= w_{t-1} + \epsilon_t \text{ for } t > 0 \end{aligned}$$

where (ϵ_t) is a (scalar) white noise process. $(\mathrm{E}\,\epsilon_t^2 = \sigma^2.)$ (w_t) is called a *random walk* process.

(2.8) Given two (scalar, real) random variables x, y with $\mathrm{E}\, x = \mathrm{E}\, y = 0$, $\mathrm{E}\, x^2 = \mathrm{E}\, y^2 = 1$ and $\mathrm{E}\, xy = 0.5$. Consider the process (w_t) defined by

$$w_t = \begin{cases} x & \text{for even } t \\ y & \text{for odd } t \end{cases}$$

(2.9) Given a stationary process (x_t) with $\mathrm{E}\, x_t = m$ and autocovariance function $\gamma(s)$. Consider the process (y_t) defined by

$$y_t = \begin{cases} x_t & \text{for even } t \\ x_t + 1 & \text{for odd } t \end{cases}$$

(2.10) Given two white noise process (ϵ_t) and (μ_t) with $\mathrm{E}\,\epsilon_t^2 = \mathrm{E}\,\mu_t^2 = 1$. Consider the three processes (x_t), (y_t) and (z_t) defined by

$$\begin{aligned} x_t &= \epsilon_t + b_1\epsilon_{t-1} \\ y_t &= \epsilon_{t-1} + b_1\epsilon_{t-2} \quad ; \quad b_1 \in \mathbf{R} \\ z_t &= b_1\mu_t + \mu_{t-1} \end{aligned}$$

(2.11) Consider an MA(1) process $y_t = b_0\epsilon_t + b_1\epsilon_{t-1}$, where (ϵ_t) is an *unobserved* white noise process with $\mathrm{E}\epsilon_t^2 = 1$. Assume that you know the autocovariance function $\gamma(s)$ of (y_t) (or that you have a very good estimate $\hat{\gamma}(s)$ for $\gamma(s)$). What can you say about the parameters b_0, $b_1 \in \mathbf{R}$ of the MA process? (Note that $\mathrm{E}\epsilon_t^2 = 1$ and that you can't observe ϵ_t, so that you have no information about $\mathrm{Cov}(y_s, \epsilon_t)$!)

(2.12) Compute the autocovariance function γ of a scalar AR(1) process (y_t) defined by $y_t + ay_{t-1} = \epsilon_t$, where (ϵ_t) is a white noise process with $\mathrm{E}\epsilon_t^2 = \sigma^2$.

(2.13) Suppose that the autocovariance function $\gamma(s)$ of an AR(1) process is given. Try to determine the parameters a, σ^2 of the AR process from $\gamma(s)$. Can you interpret your results. (Compare the exercise above.)

(2.14) Consider the scalar ARMA(1,1) process defined by $y_t + ay_{t-1} = \epsilon_t + b\epsilon_{t-1}$, where (ϵ_t) is a white noise process with $\mathrm{E}\epsilon_t^2 = \sigma^2$.

- Calculate the steady state solution of the above difference equation. Is the corresponding process (y_t) stationary?
- Calculate its autocovariance function.
- Given the autocovariance function $\gamma(s)$. Try to determine the parameter a.

(2.15) Consider the difference equation $y_t = 5y_{t-1} + \epsilon_t$, where ϵ_t is a white noise process with variance σ^2. Find a solution for this difference equation by *backward substitution* beginning from a terminal value y_0. Does there exist a "steady state" solution, if we start from the "infinite future"?

(2.16) As we have seen in the lecture, a harmonic process

$$y_t = \sum_j e^{i\lambda_j t} z_j$$

is real valued iff for every frequency $\lambda_j \neq 0$ there exists a frequency $\lambda_l = -\lambda_j$ and $z_l = \bar{z}_j$ holds. Is this condition a contradiction to the "stationarity condition": $\mathrm{E}\, z_j z_l^* = 0$ for all $\lambda_j \neq \lambda_l$. (In other words: Are there real valued stationary harmonic processes?)

(2.17) Consider a stochastic process (y_t) defined by $y_t = a\cos(\lambda t + \phi)$, where $\lambda \in (-\pi, \pi]$ is a (fixed) angular frequncy, a, ϕ are independent random variables and ϕ is uniformly distributed in $[0, 2\pi]$. Is (y_t) stationary?

(2.18) Consider two (scalar) stationary processes (x_t), (y_t) which are uncorralated, i.e. $\mathrm{E}\, x_s y_t = 0$ for all t, s, with mean zero and autocovariance functions $\gamma_x(s)$ and $\gamma_y(s)$.

Show that the process $z_t = x_t + y_t$ is stationary and compute its autocovariance function.

(2.19) Is the function

$$\begin{aligned} \gamma : \mathbf{Z} &\longrightarrow \mathbf{R} \\ s &\longmapsto \begin{cases} 1 & \text{for } s = 0 \\ 0.9 & \text{for } s = 1 \\ 0.9 & \text{for } s = -1 \\ 0 & \text{else} \end{cases} \end{aligned}$$

nonnegative definite?

3 Stationary Processes in Frequency Domain

The spectral representation is one of the center-parts of the theory of stationary processes; both from the point of view of methods and the interpretation of the process. From the spectral representation of a stationary process, the spectral representation of the covariance function and the spectral density as the Fourier transform of the covariance function are obtained. The spectral density turns out to be extremly important for the understanding of the process. Finally we consider linear transformations of stationary processes and the corresponding transformation of the second moments, which shows a further advantage of spectral representations.

3.1 The Spectral Representation of Stationary Processes

The main result of this section implies that every stationary process (x_t) can be approximated with arbitrary accuracy by an harmonic process. Or more precisely there is a sequence $((x_{t,n} \mid t \in \mathbf{Z}) \mid n \in \mathbf{N})$ of harmonic processes $(x_{t,n} \mid t \in \mathbf{Z})$ such that

$$\operatorname*{l.i.m}_{n \to \infty} x_{t,n} = x_t \tag{3.1}$$

holds for every t.

In order to state this result we have to introduce an appropriate integral. For this reason we define:

Definition 3.1 *A stochastic process* $(z(\lambda) \mid \lambda \in [-\pi, \pi])$ *with random variables* $z(\lambda) : \Omega \to \mathbb{C}^n$ *is called a* process of orthogonal increments *if the following conditions are satisfied:*

(i) $z(-\pi) = 0$ *a.e. and* $z(\pi) = x_0$ *a.e.*

(ii) $\operatorname{l.i.m}_{\epsilon \downarrow 0} z(\lambda + \epsilon) = z(\lambda)$ *for* $\lambda \in [-\pi, \pi)$. *(right continuity)*

(iii) $\mathrm{E}\, z(\lambda)^* z(\lambda) < \infty$ *for all* $\lambda \in [-\pi, \pi]$

(iv) $\mathrm{E}\,\{(z(\lambda_4) - z(\lambda_3))(z(\lambda_2) - z(\lambda_1))^*\} = 0$ *for all* $\lambda_1 < \lambda_2 \leq \lambda_3 < \lambda_4$.

If we define the function $F : [-\pi, \pi] \to \mathbb{C}^{n \times n}$ by $F(\lambda) = \mathrm{E}\, z(\lambda) z(\lambda)^*$ the following relations hold:

$$\begin{aligned} F(-\pi) &= 0 \\ F(\lambda) &\geq 0 \\ F(\lambda_2) - F(\lambda_1) &= \mathrm{E}\,\{(z(\lambda_2) - z(\lambda_1))(z(\lambda_2) - z(\lambda_1))^*\} \text{ for } \lambda_1 \leq \lambda_2 \end{aligned} \tag{3.2}$$

Thus $F(\lambda)$ is a nondecreasing right continous function on. (Here nondecreasing means that the difference $F(\lambda_2) - F(\lambda_1)$ is a nonnegative definite matrix for all $\lambda_1 \leq \lambda_2$.)

Note that $(z(\lambda) \mid \lambda \in [-\pi, \pi])$ is a stochastic process with a continuous index set $[-\pi, \pi]$ and we will interpret these indices λ not as time points but as (angular) frequencies!

Suppose we have given a (deterministic, scalar) function $g : [-\pi, \pi] \to \mathbf{C}$: $\lambda \mapsto g(\lambda)$ and a partition $-\pi = \lambda_0^n < \lambda_1^n < \cdots < \lambda_n^n = \pi$ of the interval $[-\pi, \pi]$. We then define a (finite) sum

$$\mathrm{I}_n(g) = \sum_{i=0}^{n-1} g(\lambda_i^n) \left(z(\lambda_{i+1}^n) - z(\lambda_i^n)\right).$$

If for all sequences of partitions with $\max_i(\lambda_{i+1}^n - \lambda_i^n) \to 0$ for $n \to \infty$ the limit in mean squares sense of $\mathrm{I}_n(g)$ exists and is the same, then we define

$$\mathrm{I}(g) = \int_{-\pi}^{\pi} g(\lambda) dz(\lambda) = \underset{n \to \infty}{\mathrm{l.i.m}}\, \mathrm{I}_n(g) \tag{3.3}$$

$\mathrm{I}(g)$ is called the *stochastic Integral* of g with respect to the process $z(\lambda)$. Note that this definition is a "natural" extension of the definition of a Riemann or a Riemann-Stieltjes integral. But instead of weighting the value of $g(\lambda_i)$ with the length of the interval $[\lambda_i, \lambda_{i+1}]$ or the measure of this interval with respect to some nonstochastic distribution function, we weight $g(\lambda_i)$ with the increment of the stochastic process $z(\lambda)$ which may be interpreted as a stochastic distribution function. Of course $\mathrm{I}(g)$ is in general a random variable.

If $z(\lambda) = F(\lambda)$ is a (nonstochastic) distribution function defined over $[\pi, \pi]$ then the integral with respect to $z(\lambda)$ is of course the integral of $g(\lambda)$ with respect to the matrix valued measure defined by $F(\lambda)$.

The integral $\mathrm{I}(g)$ defined above has the following properties:

(i)

$$\begin{aligned}
&\mathrm{E}\, \mathrm{I}(g) = \\
&\mathrm{E}\, \mathrm{l.i.m} \textstyle\sum_{i=0}^{n-1} g(\lambda_i^n) \left(z(\lambda_{i+1}^n) - z(\lambda_i^n)\right) = \lim \textstyle\sum_{i=0}^{n-1} g(\lambda_i^n) \left(\mathrm{E}\, z(\lambda_{i+1}^n) - \mathrm{E}\, z(\lambda_i^n)\right) = \\
&\textstyle\int_{-\pi}^{\pi} g(\lambda) d\mathrm{E}(z(\lambda)).
\end{aligned}$$

(ii)

$$\begin{aligned}
&\mathrm{E}\, \mathrm{I}(g)\, \mathrm{I}(h)^* = \\
&\mathrm{E} \left\{ \mathrm{l.i.m} \textstyle\sum_{i=0}^{n-1} g(\lambda_i^n) \left(z(\lambda_{i+1}^n) - z(\lambda_i^n)\right) \textstyle\sum_{j=0}^{n-1} \overline{h(\lambda_j^n)} \left(z(\lambda_{j+1}^n) - z(\lambda_j^n)\right)^* \right\} = \\
&\lim \textstyle\sum_{i,j=0}^{n-1} g(\lambda_i^n) \overline{h(\lambda_j^n)}\, \mathrm{E} \left\{ \left(z(\lambda_{i+1}^n) - z(\lambda_i^n)\right) \left(z(\lambda_{j+1}^n) - z(\lambda_j^n)\right)^* \right\} = \\
&\lim \textstyle\sum_{i=0}^{n-1} g(\lambda_i^n) \overline{h(\lambda_i^n)} \left(F(\lambda_{i+1}^n) - F(\lambda_i^n)\right) = \\
&\textstyle\int_{-\pi}^{\pi} g(\lambda) \overline{h(\lambda)} dF(\lambda),
\end{aligned}$$

where $F(\lambda) = \mathrm{E}\, z(\lambda) z(\lambda)^*$.

Here we have used the continuity and linearity of the expectation and the properties of $z(\lambda)$.

Theorem 3.1 *(Spectral Representation Theorem) For every stationary process (x_t) there exists a process $(z(\lambda) \mid \lambda \in [-\pi, \pi])$ with orthogonal increments such that*

$$x_t = \int_{-\pi}^{\pi} e^{i\lambda t} dz(\lambda) \text{ a.e.} \tag{3.4}$$

holds. The process $(z(\lambda))$ is a.e. uniquely determined from (x_t)

Proof: We do not give a proof of this theorem here. For a proof see e.g. Rozanov [5] or Brockwell and Davis [3]. One way to prove this result is to show that every stationary process is associated with an unitary operator in an Hilbert space and to use the spectral representation of unitary operators.□

As a direct consequence of the theorem above (and of the definition of a stochastic integral) we now see that every stationary process (x_t) can be approximated by harmonic processes:

$$x_t = \int_{-\pi}^{\pi} e^{i\lambda t} dz(\lambda) = \operatorname{l.i.m} \sum_{j=0}^{n-1} e^{i\lambda_j^n t} \left(z(\lambda_{j+1}^n) - z(\lambda_j^n) \right).$$

Note that this is a "pointwise" result holding for every fixed $t \in \mathbf{Z}$ and that in general the convergence cannot be made uniform in t.

If $z(\lambda)$ is the (orthogonal increment) process corresponding to (x_t) we call $F(\lambda) = \mathrm{E}\, z(\lambda) z(\lambda)^*$ the *spectral distribution function* of (x_t). If there exists a function $\mathrm{f} : [-\pi, \pi] \to \mathbf{C}^{n \times n}$ such that

$$F(\lambda) = \int_{-\pi}^{\lambda} \mathrm{f}(\nu) d\nu,$$

where ν denotes the Lebesque measure, then f is called the *spectral density (function)* of (x_t).

If we assume (for simplicity of notation) that $\mathrm{E}\, x_t = 0$ we have

$$\gamma(s) = \mathrm{E}\, x_s x_0^* = \mathrm{E} \int_{-\pi}^{\pi} e^{i\lambda s} dz(\lambda) \int_{-\pi}^{\pi} \overline{e^{i\lambda 0} dz(\lambda)} = \int_{-\pi}^{\pi} e^{i\lambda s} dF(\lambda), \tag{3.5}$$

which is the spectral representation of the autocovariance function. If $\mathrm{f}(\lambda)$ exists, we further get

$$\gamma(s) = \int_{-\pi}^{\pi} e^{i\lambda s} \mathrm{f}(\lambda) d\lambda. \tag{3.6}$$

Note that not for every stationary process the spectral density exists. One condition (which is not the most general) to ensure the existence of the spectral density, is

$$\sum_{s=-\infty}^{\infty} \| \gamma(s) \| < \infty.$$

Under this condition we can represent $f(\lambda)$ (using the inverse Fourier transformation) as

$$f(\lambda) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} e^{-i\lambda s} \gamma(s). \tag{3.7}$$

In the next theorem we give a mathematical characterization of spectral distribution functions and spectral densities. (This Theorem corresponds to the characterization of covariance functions we gave at the end of the last section.)

Theorem 3.2

- $F : [-\pi, \pi] \to \mathbf{C}^{n \times n}$ *is a spectral distribution function if and only if*

$$\begin{array}{lll} F(-\pi) = 0 \quad ; \quad F(\pi) < \infty & & \\ \lim_{\epsilon \downarrow 0} F(\lambda + \epsilon) = F(\lambda) & & \textit{(right continuity)} \\ (F(\lambda_2) - F(\lambda_1)) \geq 0 \textit{ for } \lambda_1 \leq \lambda_2 & & \textit{(F is monotonically non decreasing)} \end{array} \tag{3.8}$$

- $f : [-\pi, \pi] \to \mathbf{C}^{n \times n}$ *is a spectral density function if and only if*

$$\begin{array}{ll} f(\lambda) \geq 0 & \quad (\lambda \textit{ a.e.}) \\ \int_{-\pi}^{\pi} f(\lambda) d\lambda < \infty & \end{array} \tag{3.9}$$

Proof: We only prove the "$\Rightarrow$" - part. If $F(\lambda)$ is a spectral distribution function, we know from (3.2) that all of the above conditions are fullfilled. For a spectral density $f(\lambda)$ we only have to prove that it is nonnegative. This follows from the fact that for every intervall $[\lambda_1, \lambda_2]$ the integral $\int_{\lambda_1}^{\lambda_2} f(\lambda) d\lambda = F(\lambda_2) - F(\lambda_1) \geq 0$ must be a nonnegative matrix. This implies our proposition. □

One way to prove the opposite direction is to consider the function $\gamma(s) = \int e^{i\lambda s} dF(\lambda)$. From (3.2) it is possible to show that $\gamma(s)$ is a nonnegative definite function and therefore there must exist a corresponding stationary process (x_t) such that $\gamma(s)$ is the covariance function of (x_t).

For real processes we have in addition

$$f(\lambda) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} e^{-i\lambda s} \gamma(s) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} e^{i\lambda s} \gamma(-s) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} e^{i\lambda s} \gamma(s)' = f(-\lambda)'$$

and thus in particular in the scalar case $f(\lambda) = f(-\lambda)$. Thus it is sufficient to consider the spectral density on the interval $[0, \pi]$.

Contrary to the simple case of harmonic processes we here in general deal with an uncountable set of frequencies; Correspondingly instead of summing over frequencies we have integrals over frequencies. We therefore consider always intervals of frequencies (frequency-bands) rather than single points. But nevertheless the interpretation is similar to the case of harmonic processes.

Let us consider the scalar case first. From (3.2) we see that the increments of the spectral distribution function over a frequency-band is equal to the (noncentral) variance of the increment of the process $z(\lambda)$ over the same frequency-band. If the spectral density $f(\lambda)$ exists we have from (3.2)

$$\int_{\lambda_1}^{\lambda_2} f(\lambda) d\lambda = E(z(\lambda_2) - z(\lambda_1))\overline{(z(\lambda_2) - z(\lambda_1))} \tag{3.10}$$

and thus the area under f from λ_1 to λ_2 is a measure how much this frequency-band contributes to the process (x_t). In addition as $\int f(\lambda)d\lambda = \text{Var}(x_t)$ the ratio

$$\frac{\int_{\lambda_1}^{\lambda_2} f(\lambda)d\lambda}{\text{Var}(x_t)}$$

is a measure of the relative importance of this frequency-band. In this sense *strong increases in the spectral distribution and peaks in the spectral density indicate the important frequency-bands.*

Consider the following four examples:

- For a harmonic process $x_t = \sum_{j=1}^{h} e^{i\lambda_j t} z_j$ the corresponding orthogonal increment process $z(\lambda)$ is given by $z(\lambda) = \sum_{j:\lambda_j \leq \lambda} z_j$. It is easy to see that the stationarity conditions for $(z_j | j = 1, \ldots, h)$ imply that $z(\lambda)$ is a stochastic process with orthogonal increments. Of course we have $x_t = \sum_{j=1}^{h} e^{i\lambda_j t} z_j = \int e^{i\lambda t} dz(\lambda)$ as in this case the stochastic integral with respect to the "step-process" $z(\lambda)$ is nothing but a finite sum over the frequencies λ_j. We also want to stress the fact, that in this case the spectral density function does not exist!

- For a white noise process (ϵ_t) we have $\sum \|\gamma(s)\| < \infty$ since $\gamma(s) = 0$ for $s \neq 0$ and thus the spectral density function exists and is given by $f(\lambda) = \frac{1}{2\pi} \sum e^{i\lambda s} \gamma(s) = \frac{1}{2\pi}\gamma(0)$. In this case the spectral density is constant and thus frequency-bands of the same length equaly contribute to the process (x_t). Since this property resembles to the property of white light where all colors are contained, these processes have been called white noise. See figure 3.1.

- Now consider a scalar AR(1) process

$$y_t = ay_t + \epsilon_t.$$

For $|a| < 1$, we have $\sum_{j=-\infty}^{\infty} \|\gamma(s)\| < \infty$ and

$$f(\lambda) = \sum_{j=-\infty}^{\infty} e^{i\lambda s}\gamma(s) = \frac{\sigma^2}{2\pi(1 - 2a\cos(\lambda) + a^2)}.$$

(See also next section.) As we can see from figure 3.1, in the case $a > 0$ the spectral density has a peak at $\lambda = 0$ and thus the low frequencies are dominating the behaviour of the process (x_t). This gives "smooth" trajectories. In the case $a < 0$ we have a peak at π and thus the high frequencies dominate and the process shows a very erratic behaviour.

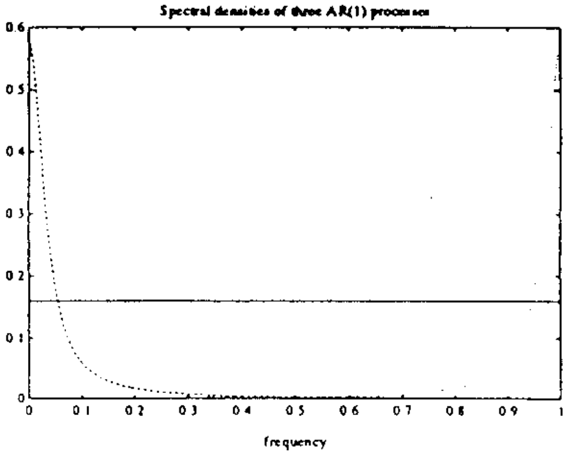


Figure 3.1: Spectral density of a white noise process (solid), of the AR(1) process $y_t = 0.9y_{t-1} + \epsilon_t$ (dashed) and of the AR(1) process $y_t = -0.9y_{t-1} + \epsilon_t$ (dotted). All three processes are scaled such that they have the same variance $\gamma(0) = 1$.

- For an AR(4) process $y_t = 0.95y_{t-4} + \epsilon_t$ we have a spectrum as shown in figure 3.2. Here we have very high peaks at frequency $\lambda = 2\pi/4$ which corresponds to a period of 4 and at frequency $\lambda = 2\pi/2$ which is the first superharmonic. Thus a process of this form could be a good model for quarterly data with a "random" seasonal pattern.

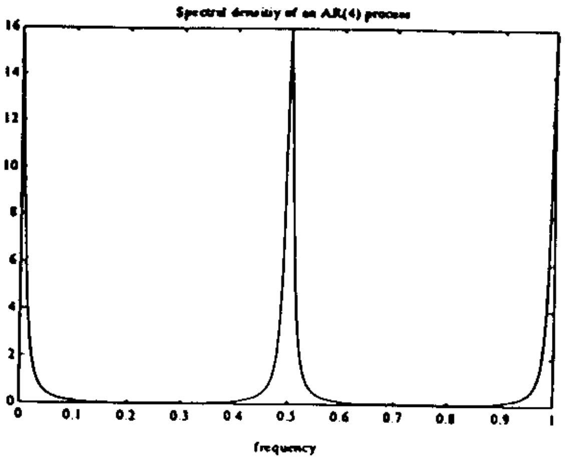


Figure 3.2: Spectral density of the AR(4) process $y_t = 0.95y_{t-4} + \epsilon_t$.

The cross-spectral density f_{xy} between two scalar processes (x_t) and (y_t) is in general a complex valued function also for real processes. In polar coordinates we write $f_{yx}(\lambda) = |f_{yx}(\lambda)|e^{i\phi_{yx}(\lambda)}$. It is easy to see from (3.10) that $|f_{yx}(\lambda)|$ is a measure of the linear dependence between (x_t) and (y_t) at frequency λ. (In order to be completely precise we have to consider frequency-bands again.) The phase $\phi_{yx}(\lambda)$ is a measure for the expected phase shift between the frequency components of (x_t) and (y_t) at frequency λ. As the cross spectrum is scale dependent a normalized measure for linear dependence the so called *coherence*

$$C^2(\lambda) = \frac{|f_{yx}(\lambda)|^2}{f_x(\lambda) f_y(\lambda)} : [-\pi, \pi] \longrightarrow [0, 1]$$

may be considered. The coherence is a frequency specific squared correlation.

Suppose we have given a vector process (z_t) and we split this process into two (vector) components as

$$z_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix} \quad ; \quad x_t \in \mathbf{R}^n, \ y_t \in \mathbf{R}^m$$

where e.g. x_t contains the exogenous variables and y_t the endogenous variables at time t. This induces a corresponding partitioning of the covariance function $\gamma(s)$ and of the spectral density $f(\lambda)$ of the process (z_t):

$$\gamma(s) = \begin{pmatrix} \gamma_x(s) & \gamma_{xy}(s) \\ \gamma_{yx}(s) & \gamma_y(s) \end{pmatrix} \quad \text{and} \quad f(\lambda) = \begin{pmatrix} f_x(\lambda) & f_{xy}(\lambda) \\ f_{yx}(\lambda) & f_y(\lambda) \end{pmatrix}.$$

We will call γ_x (resp. f_x) the autocovariance function (resp. the autospectrum) of the process (x_t) and γ_{xy} (resp. f_{xy}) the cross covariance function (resp. the cross spectrum) between the processes (x_t) and (y_t). Analog notions will apply to γ_y, f_y, γ_{yx} and f_{yx}.

Let us repeat that there is a one-to-one relation between the covariance function and the spectral distribution function and if it exists also for the spectral density function. Therefore these three functions contain the same information about the underlying process. However the information is displayed in a different way and the information may be easier to read from the spectral density function.

3.2 Linear Transformations of Stationary Processes

If (x_t) is a stationary process then

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j} \quad ; \quad a_j \in \mathbf{R}^{m\times n} \tag{3.11}$$

is called a *linear transformation* of (x_t). In order to guarantee the existence of the sum (3.11) for all stationary (x_t) we assume

$$\sum_{j=-\infty}^{\infty} \|a_j\| < \infty, \tag{3.12}$$

where $\|\cdot\|$ is an arbitrary matrix norm. $(a_j | j \in \mathbf{Z})$ is called the *weighting sequence* of the linear transformation.

Theorem 3.3 *If (x_t) is stationary and (3.12) holds then $(y_t', x_t')'$ is stationary.*

Proof: Straightforward. □

We have from the spectral representation theorem

$$\begin{aligned} x_t &= \int_{-\pi}^{\pi} e^{i\lambda t} dz_x(\lambda) \\ y_t &= \sum_{j=-\infty}^{\infty} a_j x_{t-j} = \sum_{j=-\infty}^{\infty} a_j \int_{-\pi}^{\pi} e^{i\lambda(t-j)} dz_x(\lambda) \\ &= \int_{-\pi}^{\pi} e^{i\lambda t} \left(\sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j} \right) dz_x(\lambda), \end{aligned} \tag{3.13}$$

where $z_x(\lambda)$ is the orthogonal increment process associated with (x_t). By assumption (3.12) the infinite sum $\sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}$ exists and we define the *transferfunction* $k : [-\pi, \pi] \to \mathbf{C}^{m\times n}$ corresponding to the linear transformation (3.11) by

$$k(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}. \tag{3.14}$$

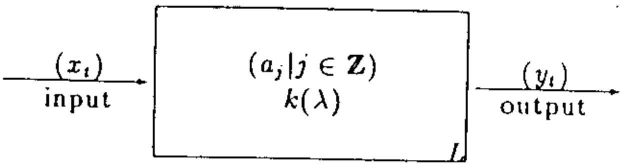
The transferfunction $k(\lambda)$ and the *weighting sequence* $(a_j|j \in \mathbf{Z})$ are in a one-to-one relation as we have $a_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda j} k(\lambda) d\lambda$.

If we replace (3.12) by the more general condition

$$\sum_{j=-\infty}^{\infty} \|a_j\|^2 < \infty \tag{3.15}$$

then the sum on the right hand side of (3.11) does not necessarily converge for all stationary (x_t) but it does converge for example for white noise processes. Whereas under (3.12) the right hand side of (3.14) converges pointwise and uniform in λ under (3.15) the convergence can only be guaranteed in the mean squares sense.

Let us give a *system theoretic* interpretation for linear transformations: A linear transformation which is sometimes also called a *linear filter* can be interpreted as a *linear system* relating the inputs (x_t) to the outputs (y_t). Thereby the inputs (exogenous variables) represent the influence of the "outside world" on the outputs (endogenous variables).



The action of the *system* L is described either by the *weighting sequence* $(a_j|j \in \mathbb{Z})$ or by the transferfunction $k(\lambda)$. Both are equivalent, but (a_j) describes the *filter* in time domain whereas $k(\lambda)$ describes the filter in frequency domain.

Next we want to explain some important notions. Let us write $(y_t) = L(x_t)$ as a short form for (3.11). (Here L denotes an operator which operates on stochastic processes.)

The "system" (3.11) is *linear* since $L(b_1x_{1,t} + b_2x_{2,t}) = b_1L(x_{1,t}) + b_2L(x_{2,t})$, which states that the output of a linear combination of two inputs $(x_{1,t})$ and $(x_{2,t})$ is just the corresponding linear combination of the two outputs $L(x_{1,t})$ and $L(x_{2,t})$.

The system (3.11) is time invariant, i.e. $L(x_{t+s}|t \in \mathbb{Z}) = (y_{t+s}|t \in \mathbb{Z})$ which means that the output of the shifted input is just the shifted output. This is an easy consequence of the fact that the *weighting sequence* $(a_j|j \in \mathbb{Z})$ does not depend on time t.

By our assumption (3.12) bounded inputs generate bounded outputs. This is one possible definition of ***stability***.

The system (3.11) is in general ***dynamic***. A ***static*** system is defined by the property that the present output y_t only depends on the present input x_t whereas for a dynamic system also future and/or past inputs influence y_t. Thus the system (3.11) is dynamic unless $a_j = 0$ for all $j \neq 0$.

A system is called *causal* if the present output y_t depends only on present and past values of the inputs $(x_{t-j}|j \geq 0)$. This means that (3.11) is causal iff $y_t = \sum_{j=0}^{\infty} a_j x_{t-j}$.

We now have from (3.13)

$$\begin{aligned} y_t &= \int_{-\pi}^{\pi} e^{i\lambda t} k(\lambda) dz_x(\lambda) \\ &= \int_{-\pi}^{\pi} e^{i\lambda t} dz_y(\lambda), \end{aligned} \tag{3.16}$$

where z_y is the process with orthogonal increments associated with (y_t) From the formula above we can give an interpretation of the transferfunction k. For simplicity the case $m = n = 1$:

- for $|k(\lambda)| > 1$ the frequency λ is amplified.
- for $|k(\lambda)| < 1$ the frequency λ is diminished.
- for $|k(\lambda)| = 0$ the frequency λ is cancelled.
- If we write $k(\lambda) = |k(\lambda)|e^{i\phi(\lambda)}$ then we see that $\phi(\lambda)$ indicates the phase shift at frequency λ.

Theorem 3.4 *Let (x_t) be stationary with spectral density $\mathfrak{f}_x$. If $y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}$ holds then the spectral density $\mathfrak{f}_y$ of (y_t) and the cross spectral density $\mathfrak{f}_{yx}$ between (y_t) and (x_t) exist and are given by*

$$\begin{aligned} \mathfrak{f}_y &= k(\lambda)\,\mathfrak{f}_x(\lambda) k(\lambda)^* \\ \mathfrak{f}_{yx}(\lambda) &= k(\lambda)\,\mathfrak{f}_x(\lambda) \end{aligned} \tag{3.17}$$

where $k(\lambda) = \sum_{j=-\infty}^{\infty} c^{-i\lambda j} a_j$ is the transferfunction.

Proof: For the first part we will use an intuitive but mathematically not exact proof. From equation (3.16) we see that the "increments" of the orthogonal increment process corresponding to (y_t) are given by

$$dz_y(\lambda) = k(\lambda) dz_x(\lambda).$$

From this we have

$$\mathfrak{f}_y(\lambda)d\lambda = \mathrm{E}\left(dz_y(\lambda)dz_y(\lambda)^*\right) = k(\lambda) \underbrace{\mathrm{E}\left(dz_x(\lambda)dz_x(\lambda)^*\right)}_{\mathfrak{f}_x(\lambda)d\lambda} k(\lambda)^* = k(\lambda)\,\mathfrak{f}_x(\lambda)k(\lambda)^* d\lambda.$$

In order to prove the second part we proceed as follows.

$$\begin{aligned}
\underbrace{\mathrm{E}\, y_t x_0^*}_{\gamma_{yx}(t)} &= \textstyle\sum_{j=-\infty}^{\infty} a_j \underbrace{\mathrm{E}\, x_{t-j} x_0^*}_{\gamma_x(t-j)} \\
\gamma_{yx}(t) &= \textstyle\sum_{j=-\infty}^{\infty} a_j\, \gamma_x(t-j) \\
\textstyle\int_{-\pi}^{\pi} e^{i\lambda t}\, \mathrm{f}_{yx}(\lambda) d\lambda &= \textstyle\sum_{j=-\infty}^{\infty} a_j \int_{-\pi}^{\pi} e^{i\lambda(t-j)}\, \mathrm{f}_x(\lambda) d\lambda = \int_{-\pi}^{\pi} e^{i\lambda t} \underbrace{\sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j}}_{k(\lambda)} \mathrm{f}_x(\lambda) d\lambda
\end{aligned}$$

Here we have used the spectral representation of the cross- and auto- covariance function (3.6). Since the last equality in the last line holds for all t (and the functions $(e^{i\cdot t})$ form an orthogonal basis for the space of all square integrable functions over $[-\pi,\pi]$) we have proven that $\mathrm{f}_{yx} = k(\lambda)\, \mathrm{f}_x(\lambda)$ holds.

Note that the idea of the proof of the first part can be applied also to the second part and vice versa. □

We now give two examples how the transferfunction can be used to control the effect of linear filters:

Example 1: Consider first differences defined by $y_t = \Delta x_t = x_t - x_{t-1}$. We have

$$\begin{aligned}
k(\lambda) &= 1 - e^{-i\lambda} \\
|k(\lambda)|^2 &= (1 - e^{-i\lambda})(1 - e^{i\lambda}) = 2(1 - cos(\lambda)) \\
\phi(\lambda) &= \arctan \tfrac{\sin(\lambda)}{1-\cos(\lambda)} = \arctan \tfrac{2\sin(\lambda/2)\cos(\lambda/2)}{2\sin^2(\lambda/2)} = \arctan \tfrac{\sin((\pi-\lambda)/2)}{\cos((\pi-\lambda)/2)} = \tfrac{\pi-\lambda}{2}
\end{aligned}$$

From figure (3.3) we can see that the zero frequency component is completely cancelled, the low frequencies are attenuated and the high frequencies are amplified. In addition there is a phase shift depending on frequency, e.g. the business cycle component in the original series and the differenced series will have different phase.

Example 2: Consider fourth differences defined by $y_t = \Delta_4 x_t = x_t - x_{t-4}$. We have

$$\begin{aligned}
k(\lambda) &= 1 - e^{-4i\lambda} \\
|k(\lambda)|^2 &= (1 - e^{-4i\lambda})(1 - e^{4i\lambda}) = 2(1 - cos(4\lambda)) \\
\phi(\lambda) &= \arctan \tfrac{\sin(4\lambda)}{1-\cos(4\lambda)} = \tfrac{\pi-4\lambda}{2}
\end{aligned}$$

From figure (3.4) we see that for quarterly data fourth differences eliminate trend, seasonal fluctuation (corresponding to a period of 4 quarters) and the first superharmonic (corresponding to a period of 2 quarters). For theses reasons fourth differences are often used to remove seasonal patterns from time series. Note however that also the other frequencies are affected by the differencing operation and that there is a phase shift.

In a number of applications it is desirable to have no phase shift in the adjusted series. Think for instance of unemployment data, where seasonal adjustment is performed

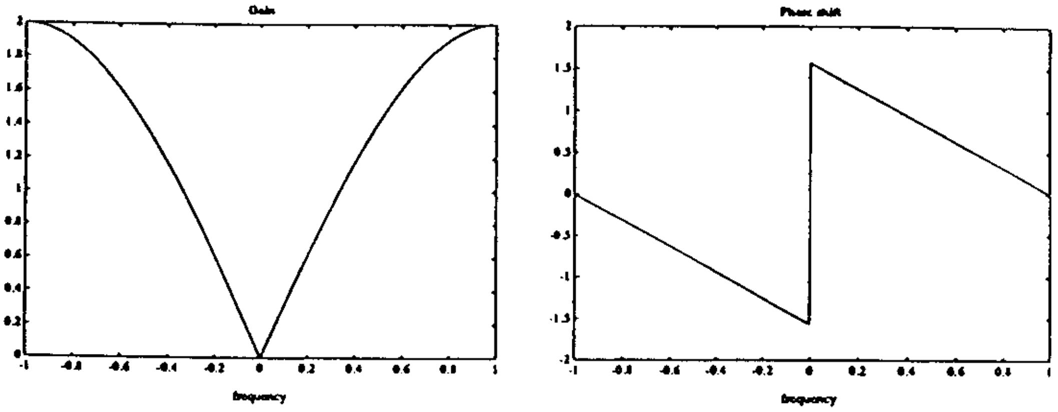


Figure 3.3: Gain and phase shift of the transferfunction of the first difference filter $y_t = \Delta x_t = x_t - x_{t-1}$.

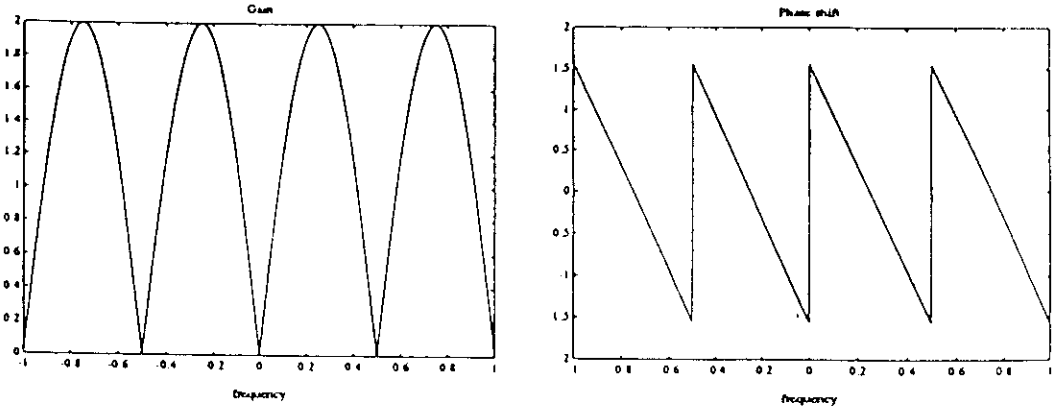


Figure 3.4: Gain and phase shift of the transferfunction of the fourth difference filter $y_t = \Delta_4 x_t = x_t - x_{t-4}$.

to check wether "real" unemployment goes up or down. A linear transformation (3.11) leaves the phases unchanged iff the transferfunction is real for all $\lambda \in [-\pi, \pi]$. One can easily prove that this is the case if and only if the *filter weights* are symmetric, i.e.

$$k(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j} \in \mathbf{R} \quad \forall\lambda \Longleftrightarrow a_j = a_{-j} \quad \forall j \in \mathbf{Z}.$$

Therefore onesided (causal) transformations $y_t = \sum_{j=0}^{\infty} a_j x_{t-j}$ will allways cause phase shifts unless $a_j = 0$ for all $j > 0$. Of course the disadvantage of symmetric filters is that (for $a_j \neq 0$ for at least one $j \neq 0$) they will never provide us with the most recent seasonal adjusted data.

If the spectral density f_x of the input process (x_t) and the cross spectrum f_{yx} between the outputs and the inputs is given we can use equation (3.17) to compute the transferfunction of the system:

$$k(\lambda) = \mathrm{f}_{yx}(\lambda)\mathrm{f}_x^{-1}(\lambda).$$

This formula is called the *Filter Formula* and will be discussed in more detail in section???. Note also that this formula is in a certain sense the generalization of the OLS-formula to an infinite number of regressors.

Note that the spectral approach provides a powerfull tool; both for understanding and calculating the effect of a linear transformation and of the associated transformation of the second moments. This tool can also be used for the design of filters with certain desired properties. For example given (3.17) can be used for the design of seasonal adjustment filters. A "good" procedure for seasonal adjustement should essentially satisfy the following conditions:

(i) the gain $|k(\lambda)|$ should be as close as possible to an ideal gain as depicted in figure 3.5. Thus the filter should cancel the seasonal frequencies on the one hand and on the other hand leave the other frequencies almost unchanged. Note however that such an ideal filter gain as shown in figure 3.5 can not be achieved by a finite filter.

(ii) almost no phase shift. Here we have to make a compromise between the conflicting aims of zero phase shift and the desire to have also the most recent values for the filtered process (i.e. one sided filters).

Formulas (3.17) also provide an elegant way to compute the spectral densities of infinite MA processes $x_t = \sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j}$, where (ϵ_t) is white noise with variance Σ. The transferfunction of the filter $(b_j | j \in \mathbf{Z})$ is of the form $k(\lambda) = \sum_{j=-\infty}^{\infty} b_j e^{-i\lambda j}$ and thus

$$\mathrm{f}_x(\lambda) = \frac{1}{2\pi} k(\lambda)\Sigma k(\lambda)^*$$

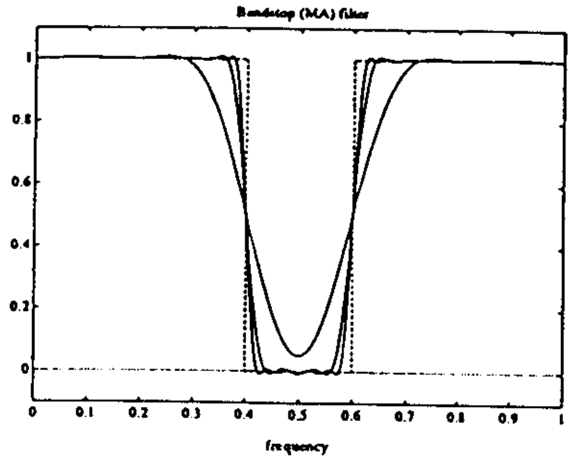
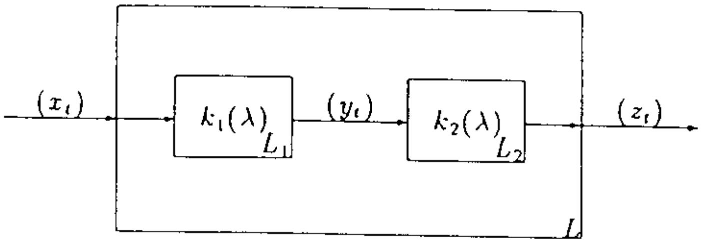


Figure 3.5: Optimal seasonal filter (dashed curve) in comparison with symmetric MA-filters of orders 10, 20 and 50.

since $\frac{1}{2\pi}\Sigma$ is the spectral density of the white noise process (ϵ_t).

Now consider two time invariant linear filters L_1 and L_2 in series i.e. we successively apply two filter to the process (x_t):



Let $y_t = \sum_j a_{1,j}x_{t-j}$ be the first filter and $z_t = \sum_j a_{2,j}y_{t-j}$ the second one. Then it is easy to prove that

$$z_t = \int_{-\pi}^{\pi} e^{i\lambda t}k_2(\lambda)k_1(\lambda)dz_x(\lambda)$$

holds. Here z_x is the orthogonal increment process associated with (x_t) and k_1 and k_2 are the transferfunctions of the two filters respectively. Thus the transferfunction of these two systems in series is

$$k(\lambda) = k_2(\lambda)k_1(\lambda)$$

i.e. the overall transferfunction is the product of the two transferfunctions. In time domain the overall weighting sequence is obtained from the discrete convolution of the two weighting sequences. This shows that in frequency domain the overall transferfunction is the product of the two transferfunctions. The mathematicians among the audience are well aware of the fact that for Fourier transformations convolutions translate to multiplications.

Inverse Systems:

If we have a filter L_1 we might ask if there exists a filter L_2 such that

$$L_2 L_1 = L_1 L_2 = 1 \text{ (identity)}$$

holds. In this case L_2 is called the *inverse system* of L_1. With an inverse system we are able to reconstruct every stationary input of the original system from the outputs. We see that for the case $m = n$ (i.e. the number of outputs equals the number of inputs and therefore k_1 is square) if $k_1(\lambda)$ is nonsingular for all $\lambda \in [-\pi, \pi]$ then $k_2(\lambda)k_1(\lambda) = k_1(\lambda)k_2(\lambda) = I$ holds with $k_2(\lambda) = k_1(\lambda)^{-1}$. Thus

$$z_t = \int_{-\pi}^{\pi} e^{i\lambda t} k_2(\lambda)k_1(\lambda)dz_x(\lambda) = \int_{-\pi}^{\pi} e^{i\lambda t} I dz_x(\lambda) = x_t$$

and the inverse system exists and its transferfunction is given by

$$k_2(\lambda) = k_1^{-1}(\lambda).$$

Let us consider three examples:

For the "differencing filter" $\Delta x_t = x_t - x_{t-1}$ there exists no inverse filter since the transferfunction $k(\lambda) = 1 - e^{-i\lambda}$ is singular for $\lambda = 0$. These means that we can't reconstruct the inputs from the differenced series, since the information about the mean is lost due to differencing.

Let us consider a scalar AR(1)-process defined by $x_t - ax_{t-1} = \epsilon_t$, where $|a| < 1$ and (ϵ_t) is white noise with variance σ^2. In this case the transferfunction $a(\lambda) = 1 - ae^{-i\lambda}$ is nonzero for all λ and thus invertible. The inverse is given by $k(\lambda) = \frac{1}{1-ae^{-i\lambda}}$ and we have

$$x_t = \int_{-\pi}^{\pi} e^{i\lambda t}\frac{1}{1 - ae^{-i\lambda}}dz_\epsilon(\lambda).$$

To compute the corresponding weighting sequence of this inverse transferfunction we substitue $z = e^{-i\lambda}$ and from the formula for geometric series we get

$$\frac{1}{1-az} = \sum_{j=0}^{\infty} a^j z^j = \sum_{j=0}^{\infty} a^j e^{-i\lambda j}$$

which converges since $|az| < 1$. Thus the corresponding weighting sequence is $(a^j | j \in \mathbf{Z}^+)$ and we get

$$x_t = \sum_{j=0}^{\infty} a^j \epsilon_{t-j}$$

which is our well known steady state solution. We also immediately get for the spectral density of (x_t)

$$\mathrm{f}_x(\lambda) = \frac{1}{1 - ae^{-i\lambda}} \frac{\sigma^2}{2\pi} \overline{\left(\frac{1}{1 - ae^{-i\lambda}} \right)} = \frac{\sigma^2}{2\pi |1 - ae^{-i\lambda}|^2} = \frac{\sigma^2}{2\pi (1 - 2a\cos\lambda + a^2)}.$$

Let us consider a scalar (i.e. $n = 1$) AR(p) process of the form

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} = \epsilon_t$$

where (ϵ_t) is white noise and define the transferfunction $a(z) = 1 + a_1 z + a_2 z^2 + \cdots + a_p z^p$ over $\mathbf{C}$ rather than on the unit circle $\{z \,|\, |z| = 1\}$. (Note however that every rational function defined on the unit circle can be uniquely extended from $\{z \,|\, |z| = 1\}$ to $\mathbf{C}$ and thus from an abstract point of view there is no difference between the transferfunction $a(e^{i\lambda})$ and $a(z)$.) If $a(z) \neq 0$ holds for all $|z| = 1$ then by what was said above the inverse system exists and its transferfunction is given by $k(z) = a(z)^{-1}$. Since $a(z)$ is a polynomial and has no zeros on the unit circle there exists an annulus containing the unit circle such that $a(z) \neq 0$ in this annulus. Thus there exists a Laurent-series expansion of $k(z)$ defined on this annulus of the form

$$k(z) = \sum_{j=-\infty}^{\infty} k_j z^j.$$

which corresponds to a weighting sequence $(k_j | j \in \mathbb{Z})$. The steady state solution to the difference equation above is obtained by applying this inverse to both sides of the difference equation which gives

$$\sum_{i=0}^{\infty} k_i (\sum_{j=0}^{p} a_j y_{t-i-j}) = \sum_{l=0}^{\infty} (\sum_{j=0}^{\min(p,l)} k_{l-j} a_j) y_{t-l} = y_t = \sum_{i=0}^{\infty} k_i \epsilon_{t-i}.$$

The Laurent series expansion may be calculated as follows. If $z_1, \ldots, z_p$ are the roots of the polynomial $a(z)$ (i.e. $a(z_i) = 0$) we have $a(z) = c(z - z_1) \cdots (z - z_p)$ where c is some constant. It is now easy to invert each of this factors, using the formula for the geometric series:

$$\frac{1}{z - z_i} = \begin{cases} -\frac{1}{z_i} \frac{1}{1 - z/z_i} = -\sum_{j=0}^{\infty} z_i^{-j-1} z^j & \text{for } |z_i| > 1 \\ \frac{1}{z} \frac{1}{1 - z_i/z} = \sum_{j=-1}^{-\infty} z_i^{-j-1} z^j & \text{for } |z_i| < 1 \end{cases}$$

Note that by assumption the case $|z_i| = 1$ is excluded. The Laurent series for $k(z)$ is now obtained as the product of the inverse of all these factors times $1/c$.

In the case $p = 1$, we have $a(z) = 1 + a_1 z = a_1(z - (-1/a_1))$ and therefore the case $|z_1| > 1$ corresponds to the case $|a_1| < 1$. Thus we have for $|a_1| < 1$

$$x_t = \frac{-1}{a_1}\sum_{j=0}^{\infty}\left(\frac{-1}{a_1}\right)^{-j-1}\epsilon_{t-j} = \sum_{j=0}^{\infty}(-a_1)^j\epsilon_{t-j}$$

which is our well known causal solution.

For the case $|z_1| < 1$ ($|a_1| > 1$) we have

$$x_t = \frac{1}{a_1}\sum_{j=-1}^{-\infty}\left(\frac{-1}{a_1}\right)^{-j-1}\epsilon_{t-j} = -\sum_{j=1}^{\infty}\left(\frac{-1}{a_1}\right)^{j}\epsilon_{t+j}$$

which is a noncausal solution.

In general we have for the case $|z_i| > 1$ for all i, a causal solution and in the case $|z_i| < 1$ for at least one i a noncausal solution. If we have roots inside and outside the unit circle, then the corresponding solution is a *twosided* infinite MA process. In other words if we impose the restriction

$$a(z) \neq 0 \text{ for all } |z| \leq 1 \tag{3.18}$$

we restrict ourselves the *stable causal case*. The assumption (3.18) is often called the *stability* condition.

In the causal case the actual calculation of the weighting sequence (a_j) is done in a different way: From a comparison of coefficients in the equation $a(z)k(z) = 1$ we obtain the following simple recursive system for the coefficients of the inverse $k(z) = \sum_{j=0}^{\infty} k_j z^j$:

$$\begin{array}{lll} z^0: & a_0k_0 = 1 & \Rightarrow k_0 = 1/a_0 = 1 \\ z^1: & a_1k_0 + a_0k_1 = 0 & \Rightarrow k_1 = -a_1/(a_0^2) = -a_1 \\ z^2: & a_2k_0 + a_1k_1 + a_0k_2 = 0 & \Rightarrow k_2 = .. \\ & \vdots & \end{array}$$

(Note that $a(z) = a_0z^0 + a_1z^1 + \cdots + a_pz^p$ is a polynomial and that $a_0 = 1$.)

This method of computing the solution is called the *z-Transform* or *discrete Laplace-Transform*.

3.3 Exercises

(3.1) Given a (real and scalar) harmonic process (x_t) defined by

$$x_t = \sum_{j=1}^{3} e^{i\lambda_j t} z_j,$$

where z_j, $j = 1,2,3$ are three complex random variables. What do the assumptions that (x_t) is real and stationary imply for the frequencies λ_j and the random variables z_j? Give an alternative representation of this process in terms of cosines and sinusoids.

We now define a stochastic process on $[-\pi, \pi]$ by

$$z(\lambda) = \sum_{j:\lambda_j \leq \lambda} z_j$$

Prove that $z(\lambda)$ is a stochastic process with orthogonal increments. Give an interpretation of

$$\int_{-\pi}^{\pi} e^{i\lambda t} dz(\lambda)$$
$$F(\lambda) = \mathrm{E}\, z(\lambda) z(\lambda)^* \text{ and}$$
$$\int_{-\pi}^{\pi} e^{i\lambda t} dF(\lambda).$$

(3.2) Compute the spectral density of the AR(1)-process $y_t = ay_{t-1} + \epsilon_t$, where ϵ_t is white noise with $\mathrm{E}\,\epsilon_t^2 = \sigma^2$. Consider the two special cases $a = 0.9$ and $a = -0.9$ and try to interpret the corresponding spectral densities.

(3.3) Discuss the AR(4)-process defined by $x_t = ax_{t-4} + \epsilon_t$, where $|a| < 1$ and (ϵ_t) is a white noise process. Prove that the steady state solution is given by $x_t = \sum_{j=0}^{\infty} a^j \epsilon_{t-4j}$ and is therefore stationary. Compute the spectral density of this process.

(3.4) Consider a (scalar) filter $y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}$ whose transferfunction is given by $k(\lambda) = \sum_{j=-\infty}^{\infty} a_j e^{-i\lambda j} = |k(\lambda)| e^{i\phi(\lambda)}$. Prove that for a deterministic process $x_t = \cos(\lambda t + \phi)$ the output of this filter is given by $y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j} = |k(\lambda)| \cos(\lambda t + \phi + \phi(\lambda))$. (Hint: What is the effect of this filter to the complex process $x_t = e^{i(\lambda t + \phi)}$?)

(3.5) Consider the filter $y_t = x_{t-1}$. Discuss the properties of its transferfunction.

(3.6) Discuss the properties of the MA-filter $y_t = \frac{1}{8}(x_{t-2} + 2x_{t-1} + 2x_t + 2x_{t+1} + x_{t+2})$. (You can can do this either analytically or by using RATS to make a plot of the transferfunction.) For which type of data is this filter especially useful.

(3.7) Given a (scalar) linear filter $y_t = \sum_{j=0}^{\infty} a_j x_{t-j}$.

(a) What is the output of the filter to an input of the form $x_0 = 1$ and $x_t = 0$ for all $t \neq 0$? (*Impulse reponse*)

(b) What is the output of the filter to an input of the form $x_t = 0$ for $t < 0$ and $x_t = 1$ for $t \geq 0$. (*Step response*)

(3.8) Let $\{x_t\}$ denote the Wölfer sunspot numbers and let $\{y_t\}$ denote the mean corrected series, $y_t = x_t - 46.93, t = 1, \ldots, 100$. The following AR(2) model for $\{y_t\}$ is obtained by OLS regression

$$y_t - 1.31 y_{t-1} + 0.63 y_{t-2} = \epsilon_t,$$

where (ϵ_t) is white noise and has variance 289,3.

Determine the spectral density of the fitted model and find the frequency at which it achieves its maximal value. What is the corresponding period? (You can use RATS to make a plot of the spectral density.)

(3.9) Consider an AR(2) process (x_t) defined by

$$x_t + a_1 x_{t-1} + a_2 x_{t-2} = \epsilon_t \quad ; \quad \mathrm{E}\epsilon_t^2 = 1.$$

where $a(z) = 1 + a_1 z + a_2 z^2 \neq 0$ for all $|z| \leq 1$. Prove that x_t has a representation as an causal infinite MA-process i.e. $x_t = \sum_{j=0}^{\infty} b_j \epsilon_{t-j}$. Prove that the autocovariance function $\gamma(s)$ of (x_t) satisfies the difference equation:

$$\gamma(s) + a_1 \gamma(s-1) + a_2 \gamma(s-2) = 0 \text{ for all } s \geq 2$$

Show that if z_1 and z_2 are the roots of $a(z)$ and $z_1 \neq z_2$ then the general solution of the difference equation $h_t + a_1 h_{t-1} + a_2 h_{t-2} = 0$ is of the form $h_t = a z_1^{-t} + b z_2^{-t}$ where $a, b \in \mathbb{C}$.

What does this result imply for the autocovariance function of AR(2) processes? (Consider the case when $a(z)$ has two real roots and the case when $a(z)$ has a pair of complex conjugated roots.)

(3.10) Consider a bivariate AR(1) model defined by

$$x_t - A x_{t-1} = \epsilon_t \quad ; \quad A \in \mathbf{R}^{2\times 2}, \quad \mathrm{E}\epsilon_t\epsilon_t' = \Sigma \in \mathbf{R}^{2\times 2},$$

where $(\epsilon_t \in \mathbf{R}^2)$ is a white noise process. The (matrix) transferfunction $a(z) : \mathbb{C} \rightarrow \mathbb{C}^{2\times 2}$ is defined by $a(z) = I - Az$, where I denotes the 2×2 identity matrix. In addition we assume that $\det(a(z)) \neq 0$ for all $|z| \leq 1$.

Prove that the inverse transferfunction $k(z) = a^{-1}(z)$ exists for all $|z| = 1$ and that $k(z)$ has a causal Laurent series expansion i.e. $k(z) = \sum_{j=0}^{\infty} B_j z^j$ for all $|z| \leq 1$. (Hint: the inverse of a matrix can be computed as the adjoint of the matrix divided by the determinant of the matrix.) Show that $k(z) = \sum_{j=0}^{\infty} A^j z^j$ by using the identity $a(z)k(z) = I$.

(3.11) ** Compute the spectral density of (x_t) defined in the above example for the numerical values:

$$A = \begin{pmatrix} 0.1 & -0.5 \\ 0.8 & -0.1 \end{pmatrix} \text{ and } \Sigma = 2\pi I.$$

Make a plot of the two autospectra of the two component processes of (x_t), of the cross spectrum (absolute value and phase) and of the coherence (using RATS).

(3.12) Construct an AR model for a quarterly time series that shows a strong seasonal pattern and a business cycle with a period of 5 years.

(3.13) Computer example: (Differencing and Periodogram). Choose one of the time series in one of the RATS data files. Compute the first differences and seasonal differences of this time series. Compute the autocorrelation function and the periodogram of these three series. Investigate the effect of this differencing operations by looking at the plots of the time series, of the autocorrelation functions and of the periodograms.

4 Linear Vector Difference Equations

Linear *vector difference equations* (VDE) are frequently used as models for time series and for the relation between time series. In many cases VDE's arise from the use of a priori theory. The importance of VDE's for modeling is caused by two reasons at least:

(i) Only a finite number of parameters is needed for the description (and thus in estimation we are in the realm of parametric statistics).

(ii) VDE's have good approximation properties for general classes of stationary processes and linear systems.

In this chapter we will deal with the following problems: solutions of VDE's, ARMAX and state space systems and their mutual relation and finaly with the problem of the relation between the second moments of the observed processes and the underlying ARMAX or state space systems.

4.1 Solution of VDE's

Consider the linear VDE

$$a_0 y_t + \cdots + a_p y_{t-p} = b_0 u_t + \cdots + b_q u_{t-q} \tag{4.1}$$

Here (y_t) denotes the n-dimensional *output process* and (u_t) the m-dimensional *input process*. The parameters are the (real) matrices $a_j \in \mathbf{R}^{n \times n}$ and $b_j \in \mathbf{R}^{n \times m}$ and e.g. the two integers p and q.

A *solution* on $\mathbf{Z}$ (or on $\mathbf{N}$) is any process (y_t) satisfying (4.1) for given parameters and inputs.

We define the *backward shift operator* z on $\mathbf{Z}$ by

$$z(y_t | t \in \mathbf{Z}) = (y_{t-1} | t \in \mathbf{Z}).$$

The backward shift operator is *linear* and *bijective*. (Note that the backward shift defined on $\mathbf{N}$ is not bijective.)

Using this operator we now can write (4.1) as

$$a(z)y_t = b(z)u_t \text{ where } \left\{ \begin{array}{l} a(z) = a_0 + a_1 z + \cdots + a_p z^p \text{ and} \\ b(z) = b_0 + b_1 z + \cdots + b_q z^q. \end{array} \right. \tag{4.2}$$

Note that the notation above is a little bit sloppy since we have written $a(z)y_t$ instead of $a(z)(y_t | t \in \mathbf{Z})$. (The backward shift z operates on processes not on random variables.) The matrices $a(z)$ and $b(z)$ are polynomial matrices in the shift operator.

The following theorem characterizes the set of all solutions of (4.1).

Theorem 4.1 *The set of all solutions of (4.1) is of the form: one* particular solution *plus the set of all* homogenous solutions *(i.e the set of all processes* (y_t) *satisfisfying* $a(z)y_t = 0$*).*

Proof: See exercises. □

We will mainly consider one particular solution the socalled *steady state solution* which is stationary for stationary inputs.

A method obtaining such a solution of (4.1) is the socalled "z-transform" method which may be derived as follows. The basic idea is to multiply the system (4.2) by the inverse of $a(z)$ if this inverse exists. Clearly in this case we obtain

$$y_t = a^{-1}(z)b(z)u_t \tag{4.3}$$

We have to consider two problems in this context. The first is under which conditions does the inverse transformation of $a(z)$ exist and the second is the representation of this inverse in terms of z.

From section 3.2 we see that the inverse of the linear transformation $a(z)$ exists if for its transferfunction $a(z)$, where z now is a complex variable, the condition

$$\det(a(z)) \neq 0 \quad \forall |z| = 1 \tag{4.4}$$

holds. Note that we use z both for the shift operator and for a complex variable. Basically because Laurent series

$$k(z) = \sum_{j=-\infty}^{\infty} k_j z^j$$

in the shift operator z and Laurent series in the complex variable z are isomorphic with respect to the rules of multiplication. This is a rather complicated way to express the simple fact that for successive application of linear transformations as well as for the multiplication of (matrix-) Laurent series in the complex variable z the relation

$$h(z) = k(z)l(z) = (\sum_{i=-\infty}^{\infty} k_i z^i)(\sum_{j=-\infty}^{\infty} l_j z^j) = \sum_{i=-\infty}^{\infty} (\sum_{j=-\infty}^{\infty} k_j l_{i-j}) z^i = \sum_{i=-\infty}^{\infty} h_i z^i$$

holds.

The inverse of the linear transformation $a(z)$ (here z denotes the backward shift operator) exists if the polynomial matrix $a(z)$ (here z denotes a complex variable $z \in \mathbb{C}$) is nonsingular for all $|z| = 1$.

The inverse of $a(z)$ is given by

$$a^{-1}(z) = \frac{1}{\det(a(z))} \operatorname{adj}(a(z)) \quad ; \quad z \in \mathbb{C} \tag{4.5}$$

where adj$(a(z))$ denotes the *adjoint* of $a(z)$. Note that adj$(a(z))$ is by definition a polynomial matrix but $a^{-1}(z)$ in general is no polynomial matrix since $\det(a(z))^{-1}$ is not polynomial in general.

From (4.5) we get the Laurent series expansion of $a^{-1}(z)$ as follows: The assumption $\det(a(z)) \neq 0$ for all z with $|z| = 1$ implies that there exist two constants $0 < r_1 < 1 < r_2$ such that $\det(a(z)) \neq 0$ holds for all z's contained in the annulus $r_1 < |z| < r_2$. (Note that $\det(a(z))$ is a polynomial and thus has only a finite number of zeros.)

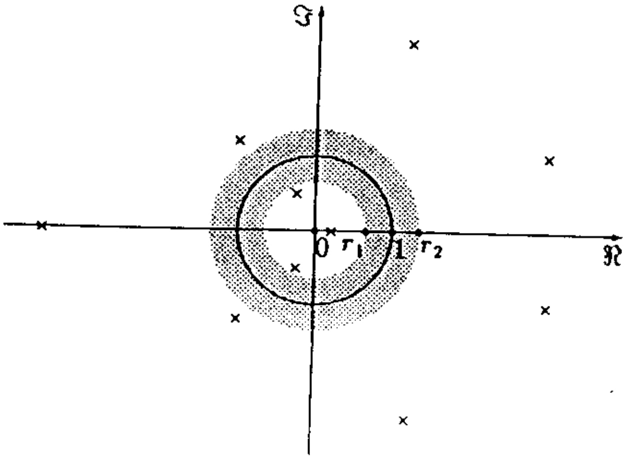


Figure 4.1: Roots of $\det(a(z))$.

Therefore we can expand $\det(a(z))^{-1}$ in a Laurentseries which is convergent on this annulus, i.e.

$$\frac{1}{\det(a(z))} = \sum_{j=-\infty}^{\infty} h_j z^j \text{ for all } r_1 < |z| < r_2.$$

This expansion can be obtained by factorizing $\det(a(z))$ as $\det(a(z)) = c(z - z_1)(z - z_2)\cdots(z - z_d)$, where $z_1, \ldots, z_d$ are the roots of $\det(a(z))$, and inverting each of this factors which gives a geometric series of the form

$$\frac{1}{z - z_i} = \begin{cases} -\frac{1}{z_i}\sum_{j=0}^{\infty}(z_i^{-j})z^j & \text{for all } |z| < |z_i| \text{ in the case } |z_i| > 1 \\ \frac{1}{z_i}\sum_{j=-\infty}^{-1}(z_i^{-j})z^j & \text{for all } |z| > |z_i| \text{ in the case } |z_i| < 1 \end{cases} \tag{4.6}$$

By multiplying these Laurentseries for $(z - z_i)^{-1}$ we get the Laurent series for $\det(a(z))^{-1}$. Note that the coefficients h_j will geometrically converge to zero for $j \to +\infty$ and $j \to -\infty$. From (4.5) we get a Laurent series expansion for $a(z)^{-1} = l(z) = \sum_{j=-\infty}^{\infty} l_j z^j$

for $z \in \mathbf{C}$. Thus using the isomorphism mentioned above $\sum_j l_j z^j$ is the inverse of the linear transformation $a(z)$ where z now denotes the backward shift operator. Due to the geometrically decreasing norms $\|l_j\|$ for $j \to \pm\infty$ the output of the inverse linear transformation $l(z)$ exists for every stationary input. Here $\|l_j\|$ denotes some matrix norm, for instance the norm defined by $\|l_j\| = \sup_{\|x\|=1} \|l_j x\|$. (Let $\|u_t\| = \sqrt{\mathbf{E}\, u_t' u_t}$ denote the norm of u_t in the linear space $\mathbf{L}_2^n$, where $\mathbf{L}_2$ is the Hilbert space of section 2.4. If (u_t) is stationary then $\|u_t\|$ does not depend on the time t and we have $\|\sum_j l_j u_{t-j}\| \leq \sum_j \|l_j u_{t-j}\| \leq \sum_j \|l_j\| \|u_{t-j}\| = \|u_t\| \sum_j \|l_j\| < \infty$.)

It is clear from the formula (4.6) that under the condition

$$\det(a(z)) \neq 0 \text{ for all } |z| \leq 1 \tag{4.7}$$

$l_j = 0$ for all $j < 0$ holds and thus in this case the solution will be causal.

In this way we have proved the following theorem:

Theorem 4.2

(i) If $\det(a(z)) \neq 0$ *for all* $|z| = 1$ *then the steady state solution exists and is obtained by expanding* $\det(a(z))^{-1}$ *as a Laurent series*

$$y_t = \sum_{j=-\infty}^{\infty} k_j u_{t-j} = k(z) u_t \text{ where } k(z) = \underbrace{\frac{1}{\det(a(z))}}_{\sum_{j=-\infty}^{\infty} h_j z^j} \operatorname{adj}(a(z)) b(z)$$

(ii) If $\det(a(z)) \neq 0$ *for all* $|z| \leq 1$ *then the steady state solution exists and is* causal.

$$y_t = \sum_{j=0}^{\infty} k_j u_{t-j} = k(z) u_t \text{ where } k(z) = \underbrace{\frac{1}{\det(a(z))}}_{\sum_{j=0}^{\infty} h_j z^j} \operatorname{adj}(a(z)) b(z)$$

Let us make some remarks to this theorem:

(i) Let us consider the homogenous solutions, i.e. the solutions of $a(z)y_t = 0$. Let z_i be a root of $\det(a(z))$ and let v denote any vector $v \in \mathbb{C}^n$, $v \neq 0$ which satisfies $a(z_1)v = 0$. Note that $\det(a(z_1)) = 0$ implies that $a(z_i)$ is a singular matrix. By inserting $y_t = z_i^{-t} v$ in $a(z)y_t$ we get $a_0 z_i^{-t} v + a_1 z_i^{-t+1} v + \cdots + a_p z_i^{-t+p} v = z_i^{-t}(a(z_i)v) = 0$ and thus $y_t = z_i^{-t} v$ is a homogenous solution.

For the case of simple roots every solution of the homogenous equation can be represented as a linear combination of solutions of the above type.

Note that the assumption $\det(a(z)) \neq 0$ for all $|z| = 1$ implies that all homogenous solutions except for the zero are non stationary. Therefore it is straightforward to show that the steady state solution defined above is unique.

For $|z_i| > 1$ we have $y_t = z_i^{-t}v \to 0$ for $t \to \infty$, which implies that under the assumption (4.7) each solution of (4.1) converges to the steady state solution if t goes to infinity.

(ii) Consider the scalar first order VDE

$$y_t = ay_{t-1} + u_t \tag{4.8}$$

Note that the condition $\det(a(z)) \neq 0$ for all $|z| \leq 1$ in this case corresponds to $|a| < 1$ and then the steady state solution is given by $y_t = \sum_{j=0}^{\infty} a^j u_{t-j}$. For $|a| > 1$ e.g. for inputs (u_t) satisfying $u_t = 0$ for $t < 0$ we have a causal nonstationary (non-stable) solution of the form $y_t = \sum_{j=0}^{t} a^j u_{t-j}$. The steady state solution described above however is different and of the form $y_t = -\sum_{j=1}^{\infty} a^{-j} u_{t+j}$ and this solution is noncausal and stable (and stationary for stationary inputs). This solution is the steady state solution for backward substitution in (4.8).

The condition (4.7) is called the *stability* assumption, since if we consider only causal solutions then (4.7) ensures that this causal solution is stable; i.e. bounded inputs generate bounded outputs. On the contrary if we a priori impose the condition that the solution is stable, then (4.7) implies that the solution is causal.

(iii) If $a(z)$ has rank less than n for all $z \in \mathbb{C}$ (i.e. if $\det(a(z)) \equiv 0$) then the system is *incomplete* or *inconsistent*, e.g.

$$\begin{array}{rcrcl} y_{t,1} & + & 0.5y_{t,2} & = & u_{t,1} \\ 2y_{t,1} & + & y_{t,2} & = & u_{t,2} \end{array} \quad ; \quad a(z) = a_0 = \begin{pmatrix} 1 & 0.5 \\ 2 & 1 \end{pmatrix}, \quad b(z) = b_0 = I$$

In this case the steady state solution (y_t) is not unique (if $u_{t,2} = 2u_{t,1}$ a.e.) or there exists no solution of this VDE.

(iv) If $\det(a(z)) \neq 0$ for all but a finite number of z's but if $\det(a(z)) = 0$ for some $|z| = 1$ then the steady state solution may not exist! E.g. for the difference equation

$$y_t - y_{t-1} = \epsilon_t \quad ; \quad a(z) = 1 - z \quad , \quad a(1) = 0$$

there exists no stationary solution for white noise input $\epsilon_t \neq 0$.

The stability condition (4.7) implies $\det(a(0)) = \det(a_0) \neq 0$ and thus a_0 must be a nonsingular matrix. This enables us to compute the power series expansion of $a^{-1}(z)$

and thus the coefficients of the transferfunction $k(z) = a^{-1}(z)b(z)$ by the following block recursive equations which we get from a comparison of coefficients on both sides of the equation $a(z)k(z) = b(z)$:

$$\begin{array}{lll} z^0: & k_0a_0 = b_0 & \Rightarrow k_0 = a_0^{-1}b_0 \\ z^1: & k_0a_1 + k_1a_0 = b_1 & \Rightarrow k_1 = a_0^{-1}(b_1 - a_1a_0^{-1}b_0) \\ & \vdots & \vdots \\ z^j: & k_0a_j + \cdots + k_ja_0 = b_j & \Rightarrow k_i = \ldots \\ & \vdots & \vdots \end{array}$$

From the results of section 3.2 we can easily compute the second moments of the steady state solutions. If the spectral density f_u of the inputs (u_t) exists then we have

$$\begin{aligned} f_y(\lambda) &= k(e^{-i\lambda})\, f_u(\lambda) k(e^{-i\lambda})^* \\ f_{yu}(\lambda) &= k(e^{-i\lambda})\, f_u(\lambda) \end{aligned}$$

where $k = a^{-1}b$.

4.2 Representations of VDE's

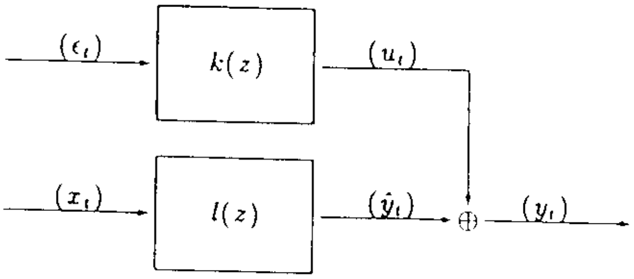
In this section we distinguish between two kinds of inputs

(i) observed inputs (x_t) and

(ii) unobserved white noise (ϵ_t).

We will always assume that the noise process (ϵ_t) is orthogonal to the input process (x_t) i.e.

$$\mathrm{E}\, x_t\epsilon_s' = 0 \quad \forall t, s. \tag{4.9}$$



The above block diagram represents how the output (y_t) is generated from the inputs (x_t) and (ϵ_t). The output (y_t) may be written as

$$y_t = \underbrace{l(z)x_t}_{\hat{y}_t} + \underbrace{k(z)\epsilon_t}_{u_t} = \hat{y}_t + u_t. \tag{4.10}$$

Thus (y_t) is the sum of the "true outputs" $\hat{y}_t$ and the "noise" (u_t) which e.g. can be interpreted as a measurement error. There are two key assumptions in this representation of the output:

(i) the inputs (x_t) are observed without noise and

(ii) the errors (u_t) (respectively (ϵ_t)) are uncorrelated with the inputs (x_t).

The *ARMAX representation* (Autoregressive Moving Average with exogenous variables) is defined as

$$a(z)y_t = d(z)x_t + b(z)\epsilon_t \tag{4.11}$$

where (x_t) are the observed inputs (ϵ_t) is white noise and a, b and d are polynomial matrices. We always will impose the stability assumption (4.7) and thus the steady state solution (y_t) is given by

$$y_t = l(z)x_t + k(z)\epsilon_t \text{ , where } l(z) = a^{-1}(z)d(z),\ k(z) = a^{-1}(z)b(z).$$

Note that under our assumptions $k(z)$, $l(z)$ are two causal, stable and rational transfer-functions.

The most important special cases are defined as follows:

$$\left.\begin{array}{llcl} \text{AR system} & a(z)y_t & = & \epsilon_t \\ \text{MA system} & y_t & = & b(z)\epsilon_t \\ \text{ARX system} & a(z)y_t & = & d(z)x_t + \epsilon_t \\ \text{ARMA system} & a(z)y_t & = & b(z)\epsilon_t \end{array}\right\} \begin{array}{l}\text{the corresponding} \\ \text{steady state} \\ \text{solution is called}\end{array} \left\{\begin{array}{l} \text{AR process} \\ \text{MA process} \\ \\ \text{ARMA process} \end{array}\right.$$

Note that an AR system has a representation as an MA system iff $\det(a(z))$ is constant. On the opposite an MA system has a representation as an AR system if and only if $\det(b(z))$ is constant.

The spectra for AR, MA and ARMA processes are given by

$$\begin{array}{ll} \text{AR:} & f_y = \frac{1}{2\pi}a^{-1}(e^{-i\lambda})\Sigma(a^{-1}(e^{-i\lambda}))^* \\ \text{MA:} & f_y = \frac{1}{2\pi}b(e^{-i\lambda})\Sigma b(e^{-i\lambda})^* \\ \text{ARMA:} & f_y = \frac{1}{2\pi}a^{-1}(e^{-i\lambda})b(e^{-i\lambda})\Sigma b(e^{-i\lambda})^*(a^{-1}(e^{-i\lambda}))^* \end{array}$$

where $\Sigma = E\epsilon_t\epsilon_t'$.

One advantage of AR-systems over MA-systems is that they are good for the modelling of *peaks* in the spectrum. We may interpret the spectral density $\mathfrak{f}_y$ as function defined on the unit circle $\{z \in \mathbf{C} \mid |z| = 1\}$ of the complex plane by identifying the frequency $\lambda \in [-\pi, \pi]$ with the point $z = e^{i\lambda}$ on the unit circle. This spectral density defined on the unit circle can be extended in an unique way to a rational function defined on the complex plane $\mathbf{C}$ by substituting z for $e^{-i\lambda}$ and more generally z^j for $e^{-i\lambda j}$. (Note that we substitute $1/z$ for $e^{i\lambda} = \overline{e^{-i\lambda}} = 1/e^{-i\lambda}$.) Using the same symbolf $\mathfrak{f}_y$ for this spectral density defined on $\mathbf{C}$ we have e.g. for a scalar AR process

$$\mathfrak{f}_y(z) = \frac{\sigma^2}{2\pi} \frac{1}{a(z)a(1/z)}$$

A pole of $\mathfrak{f}_y(z)$, $z_j = |z_j|e^{-i\lambda_j}$ say, which is close to the unit circle (i.e. $|z_j| \approx 1$) will generate a "peak" of the spectral density approximately at the frequency λ_j which corresponds to the phase of z_j. This peak will be the higher the closer this pole is to the unit circle. In this way the shape and the location of the peaks of $\mathfrak{f}_y(\lambda)$ are closely related to the location of the poles of $\mathfrak{f}_y(z)$.

The poles of $\mathfrak{f}_y(z)$ of the above scalar AR process correspond to the zeros of $a(z)$ and $a(1/z)$. Since $a(z)$ has real coefficients all zeros of $a(z)$ must occur in conjugate pairs, i.e. if z_j is a zero of $a(z)$ then also $\overline{z_j}$ must be a zero of $a(z)$. It is also immediate to see that if z_j is a zero of $a(z)$ then $1/z_j$ is a zero of $a(1/z)$ and vice versa. Note that the stability assumption excludes the case that $z = 0$ is a root of $a(z)$. Thus we see that the complex poles of $\mathfrak{f}_y(z)$ occur in "quadrupels" $(z_j, \overline{z_j}, 1/z_j, 1/\overline{z_j})$ and the real roots in pairs $(z_j, 1/z_j)$. See also figure 4.2. In other words the poles of $\mathfrak{f}_y(z)$ are symmetric to the real axis and they are reflected at the unit circle where we interpret the transformation $z \rightarrow 1/z$ as a reflection at the unit circle.

For an AR(1) process, $a(z) = 1 - a_1 z$, we can have only one real root for $a(z)$ and thus $\mathfrak{f}_y(\lambda)$ can have only one peak at 0 or at π.

If for an AR(2)-process $a(z)$ has two complex roots, $z_1 = re^{i\lambda}$ and $z_2 = \overline{z_1} = re^{-i\lambda}$ say, then the spectral density (as a function on $[0, \pi]$ will have a single peak at the frequency λ and the peak will be the higher the closer these two roots are to the unit circle. Thus by choosing the parameters of the AR(2) model we can choose the frequency of the peak and the shape of the peak. If $a(z)$ has two real roots then $\mathfrak{f}_y$ may have a single peak at $\lambda = 0$ or a single peak at $\lambda = \pi$ or two peaks at $\lambda = 0$ and $\lambda = \pi$ respectively.
The *State space representation* is defined as:

$$\begin{aligned} s_{t+1} &= As_t + Dx_t + B\epsilon_t \\ y_t &= Cs_t + \epsilon_t \end{aligned} \qquad (4.12)$$

Here $s_t \in \mathbb{R}^l$ denotes the *state vector*. $A \in \mathbb{R}^{l \times l}$, $D \in \mathbb{R}^{l \times m}$, $B \in \mathbb{R}^{l \times n}$ and $C \in \mathbb{R}^{n \times l}$ are constant matrices. An integer parameter is the dimension l of the state vector and the real parameters are the entries of the matrices A, D, B, C.

The state contains all information about the past that is necessary to generate the present and future outputs. In other words s_t is a *sufficient statistic* for the "past".

The steady state solution (y_t) may be computed as follows:

$$(z^{-1}I - A)s_t = Dx_t + B\epsilon_t$$

and thus

$$s_t = (z^{-1} - A)^{-1}(D, B)\begin{pmatrix} x_t \\ \epsilon_t \end{pmatrix}$$

$$y_t = C(z^{-1} - A)^{-1}(D, B)\begin{pmatrix} x_t \\ \epsilon_t \end{pmatrix} + \epsilon_t$$

The stability condition (4.7) is equivalent to the condition that all eigenvalues of the matrix A have modulus less than one, i.e.

$$|\lambda_{\max}(A)| < 1 \tag{4.13}$$

where $\lambda_{\max}(A)$ denotes an eigenvalue of A of maximal modulus. Using this stability condition (4.13) we have $(z^{-1}I - A)^{-1} = z(I - Az)^{-1} = z\sum_{j=0}^{j=\infty} A^j z^j$ and thus the steady solution is given by

$$y_t = \sum_{j=0}^{\infty} CA^j \begin{pmatrix} D & B \end{pmatrix} \begin{pmatrix} x_{t-j-1} \\ \epsilon_{t-j-1} \end{pmatrix} + \epsilon_t$$

We now have three different representations of VDE's namely

(i) transferfunction: $y_t = l(z)x_t + k(z)\epsilon_t$ where $l(z) = \sum_{j=0}^{\infty} l_j z^j$ and $k(z) = \sum_{j=0}^{\infty} k_j z^j$ are causal, stable and rational transferfunctions. This transferfunction representation is described by a pair of transferfunctions (l, k).

(ii) ARMAX: $a(z)y_t = d(z)x_t + b(z)\epsilon_t$ which may be described by a triple of polynomial matrices (a, b, d).

(iii) SP: $\begin{pmatrix} s_{t+1} & = & As_t + B\epsilon_t \\ y_t & = & Cs_t + \epsilon_t \end{pmatrix}$ which may be described by a quadruple of real matrices (A, B, C, D).

We now want to proof that these three representations are equivalent in a certain sense. For simplicity of notation we will do this for the case without exogenous variables, e.g. we restrict us to the case of ARMA systems.

We have already shown that every stable ARMA and SP system can be represented by a causal rational transferfunction.

We next give a state space representation for an ARMA system. (Note that the same idea also applies for ARMAX systems.) From

$$a_0 y_t + \ldots + a_p y_{t-p} = b_0 \epsilon_t + \cdots + b_q \epsilon_{t-q}$$

and $\det(a(0)) = \det(a_0) \neq 0$ we have

$$y_t = \bar{a}_1 y_{t-1} + \cdots + \bar{a}_p y_{t-p} + \bar{b}_0 \epsilon_t + \cdots + \bar{b}_q \epsilon_{t-q},$$

where $\bar{a}_j = -a_0^{-1} a_j$, $j = 1, \ldots, p$ and $\bar{b}_j = a_0^{-1} b_j$, $j = 0, \ldots, q$. Thus we can rewrite this ARMA system as a state space system of the form

$$\begin{aligned} s_{t+1} &= A s_t + B \epsilon_t \\ y_t &= C s_t + D \epsilon_t \end{aligned}$$

where

$$s_t = \left(\begin{array}{c} y_{t-1} \\ \vdots \\ \vdots \\ \vdots \\ y_{t-p+1} \\ \hline \epsilon_{t-1} \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_{t-q+1} \end{array} \right) \quad A = \left(\begin{array}{ccccc|ccccc} \bar{a}_1 & \cdots & \cdots & \cdots & \bar{a}_p & \bar{b}_1 & \cdots & \cdots & \cdots & \bar{b}_q \\ I & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots & \vdots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & & & & \vdots \\ 0 & \cdots & 0 & I & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \hline 0 & \cdots & \cdots & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & \vdots & I & \ddots & & & \vdots \\ \vdots & & & & \vdots & 0 & \ddots & \ddots & & \vdots \\ \vdots & & & & \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 0 & I & 0 \end{array} \right) \quad B = \left(\begin{array}{c} \bar{b}_0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \hline I \\ 0 \\ \vdots \\ \vdots \\ 0 \end{array} \right)$$

$$C = \left(\begin{array}{ccc|cccc} \bar{a}_1 & \cdots & \cdots & \bar{a}_p \; \bar{b}_1 & \cdots & \cdots & \cdots \; \bar{b}_q \end{array} \right), \quad D = \bar{b}_0.$$

Note that this in general will not be an "optimal" state space representation for the ARMA system, since in general the dimension of the state vector will be too high.

We now consider the *inverse* problem: Given the transfer function $k(z)$, how can we determine the ARMA representation (a, b)? We assume that $k(z)$ is a rational function which implies that the entries $k_{ij}(z)$ of $k(z)$ have a representation as $k_{ij}(z) = \frac{d_{ij}(z)}{f_{ij}(z)}$, where $d_{ij}(z)$ and $f_{ij}(z)$ are polynomials which without of restriction of generality have no common zeros. If $c(z)$ is the least common multiple of all f_{ij}'s then we have $k_{ij}(z) = \frac{1}{c(z)} n_{ij}(z)$ with polynomials $n_{ij}(z)$. If we write $k(z) = \frac{1}{c(z)} n(z)$ we have immediately an ARMA representation of the form $a = c(z) I$ and $b(z) = n(z)$.

If the transferfunction has a convergent powerseries expansion within the unit circle $\{z \mid |z| \leq 1\}$ then the ARMA system constructed above is stable, i.e. it satisfies the stability conditon (4.7): If the transferfunction $k(z)$ has no poles for $|z| \leq 1$ then none of the f_{ij}'s may have a zero within the unit circle. Therefore also $c(z)$ has no zero within the unit circle and thus $\det(a(z)) = \det(c(z)I) = c(z)^n \neq 0$ for $|z| \leq 1$ holds.

In general this representation will not be "optimal" as the degrees of $a(z)$ and $b(z)$ will be too high in general. Note also,that without further restrictions the ARMA system (a, b) is not uniquely determined from the transferfunction $k(z) = a^{-1}(z)b(z)$.

Let us discuss this point for the scalar case $n = 1$: Clearly the two ARMA systems

$$\begin{aligned} y_t &= \epsilon_t \\ y_t + ay_{t-1} &= \epsilon_t + a\epsilon_{t-1} \end{aligned}$$

have the same steady state solutions, since for both systems the transferfunction $k(z) = \frac{a(z)}{b(z)} = 1$. But the first one is *nonredundant* whereas the second one is *redundant* since it has too many parameters. Note that the *transients* of both systems in general will be different since the homogenous solutions are different.

Thus (in the scalar case) we always will assume that a and b have no common zeros, which gives us a nonredundant ARMA representation. Note in addition that with this assumption a, b are uniquely determined from the transferfunction $k = \frac{b}{a}$ up to a constant. If we for example impose the normalization $a_0 = 1$ then a and b are uniquely determined from the transferfunction.

We have shown

- how to construct the transferfunction (l, k) from a ARMAX or state space system
- how to construct an ARMAX system from given transferfunctions
- how to construct a SP system from an ARMAX system.

Putting these results together we have the following theorem:

Theorem 4.3

(i) Every (stable) ARMA system and every (stable) state space system has a rational transferfunction which has a convergent powerseries expansion within the unit circle $\{z \in \mathbb{C} \mid |z| \leq 1\}$.

(ii) Conversely for every rational transferfunction which has a convergent power series expansion in $\{z \in \mathbb{C} \mid |z| \leq 1\}$ *there is a (stable) ARMA and a (stable) state space representation.*

4.3 Identifiability

In this section we will pose the question how the underlying ARMAX system can be determined from the second moments of the observations (inputs and outputs).

For a stable ARMAX-system where the inputs are orthogonal to the errors, see (4.9), and the spectral density of the inputs $\mathfrak{f}_x$ exists, the autospectrum of the outputs and the cross-spectrum between the outputs and the inputs are given by

$$\begin{aligned} \mathfrak{f}_{yx}(e^{-i\lambda}) &= l(e^{-i\lambda})\,\mathfrak{f}_x(e^{-i\lambda}) \\ \mathfrak{f}_y(e^{-i\lambda}) &= l(e^{-i\lambda})\,\mathfrak{f}_x(e^{-i\lambda})l(e^{-i\lambda})^* + \tfrac{1}{2\pi}k(e^{-i\lambda})\Sigma k(e^{-i\lambda})^* \\ &\quad \text{where } l = a^{-1}d, k = a^{-1}b, \Sigma = \mathrm{E}\,\epsilon_t\epsilon_t'. \end{aligned} \tag{4.14}$$

This is an easy consequence of the fact that because of (4.9)

$$\begin{aligned} \mathrm{E}\,y_t x_s' &= \mathrm{E}\,\hat{y}_t x_s' + \underbrace{\mathrm{E}\,u_t x_s'}_{0} = \mathrm{E}\,\hat{y}_t x_s' \\ \mathrm{E}\,y_t y_s' &= \mathrm{E}\,\hat{y}_t \hat{y}_s' + \underbrace{\mathrm{E}\,\hat{y}_t u_s'}_{0} + \underbrace{\mathrm{E}\,u_t \hat{y}_s'}_{0} + \mathrm{E}\,u_t u_s' = \mathrm{E}\,\hat{y}_t\hat{y}_s' + \mathrm{E}\,u_t u_s' \end{aligned}$$

holds.

These equations show the relations between the *external characteristics* $\mathfrak{f}_y$, $\mathfrak{f}_x$ and $\mathfrak{f}_{yx}$ and the *internal characteristics* i.e. the degrees of the polynomial matrices a, b and d and the entries of the coefficientmatrices a_j, b_j and d_j and Σ.

Definition 4.1 *Two ARMAX systems* (a, b, d) *and* $(\bar{a}, \bar{b}, \bar{d})$ *are called* observationally equivalent *if for given* $\mathfrak{f}_x$, *for given* Σ *and a suitably chosen* $\bar{\Sigma}$ *they generate the same* $\mathfrak{f}_{yx}$ *and* $\mathfrak{f}_y$. *This must hold for all* $\Sigma \geq 0$ *and all* $\mathfrak{f}_x \geq 0$. *Thereby* Σ *and* $\bar{\Sigma}$ *are the variance matrices of the respective white noise processes.*

Note that we here defined observationally equivalence in terms of the stationary solutions and their second moments only. We don't use the information coming from the homogenous solutions or from higher order moments.

We also want to stress the fact that we here assume perfect knowledge of the second moments of the obeservations, i.e. of the spectra $\mathfrak{f}_x$, $\mathfrak{f}_{yx}$ and $\mathfrak{f}_y$. We here don't deal with the problem of estimation of the parameters of the underlying ARMAX system or SP system from a given (finite) time series. But of course we first have to answer the question whether the ARMAX system can be identified in such an idealized situation. If this is not the case then of course estimation based an a "limited information" makes not too much sense!

In the following we will always assume that

$$\begin{aligned} \mathfrak{f}_x(e^{-i\lambda}) &> 0 \quad \textit{(persistent excitation condition)} \\ \Sigma &> 0 \end{aligned} \tag{4.15}$$

hold.

Definition 4.2 *A class of ARMAX systems is called identifiable if it does not contain two different observationally equivalent systems.*

$$[(a,b,d),\Sigma,f_x] \underset{\Longleftarrow}{\overset{\Longrightarrow}{?}} [f_y, f_{yx}, f_x]$$

We are looking for restrictions on (a,b,d) such that the "$\Leftarrow$" of the above relation is defined.

In general this "inverse problem" is split into two steps:

(i) Given (f_y, f_{yx}, f_x) try to determine the transferfunctions $l = a^{-1}d$ and $k = a^{-1}b$.

(ii) Given the transferfunctions k and l try to determine the polynomial matrices (a,b,d).

From equations (4.11) and the persistent excitation condition (4.15) we immediatly get

$$\begin{aligned} l = a^{-1}b &= f_{yx} f_x^{-1} \\ f_y - l f_x l^* = f_y - f_{yx} f_x^{-1} f_x f_x^{-*} f_{yx}^* &= \tfrac{1}{2\pi} k \Sigma k^* \end{aligned} \tag{4.16}$$

Here f_x^{-*} is a short notation for $(f_x^{-1})^* = (f_x^*)^{-1}$. Note that the left-hand side of the last equation is known from the spectral densities of the observations, so that we know $k\Sigma k^*$. The problem now is to determine k from $k\Sigma k^*$, which is known as the *spectral factorization problem.*

We first consider some (simple) special cases:

Scalar MA systems: $y_t = b(z)\epsilon_t$

The spectral density $f_y(e^{-i\lambda}) = \frac{\sigma^2}{2\pi} b(e^{-i\lambda})\overline{b(e^{-i\lambda})}$ is a rational function defined on the unit circle. The unique rational extension of this function to $\mathbb{C}$ is given by: $f_y(z) = \frac{\sigma^2}{2\pi} b(z) b(z^{-1})$.

There are two trivial normalization problems associated with this spectral factorization problem:

It is trivial to see that $b(z)$ and $\tilde{b}(z) = z^j b(z)$ both give the same f_y. This is a consequence of the fact that (ϵ_t) and (ϵ_{t-j}) are both white noise processes with the same variance. Since we can't observe the noise there is no way to distinguish between these two noise processes from given f_y. Therefore w.l.o.g. we can assume that $b_0 \neq 0$ holds.

It is easy to to see that $b(z)$ together with the variance σ^2 and $\tilde{b}(z) = cb(z)$ together with a noise variance $\tilde{\sigma}^2 = \sigma^2/c^2$ generate both the same output spectrum f_y. This transformation of b corresponds to a scaling of the white noise process ϵ_t. We therefore assume $b_0 = 1$.

From the representation of the output spectrum f_y it is easy to see that if z_j is a zero of f_y then also $\overline{z_j}$, $\frac{1}{z_j}$ and $\frac{1}{\overline{z_j}}$ must be zeros of f_y. (Note that by the condition $b_0 = 1$ the point $z = 0$ can't be a zero of f_y.) This is a consequence of the fact that $b(z)$ has real coefficients (and thus the roots must occur in conjugate pairs) and that f_y is the product of $b(z)$ and $b(1/z)$.

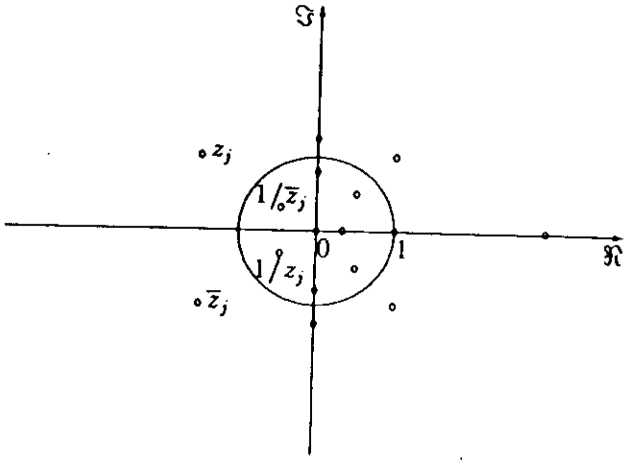


Figure 4.2: Roots of the spectral density $f_y(z)$ of an MA(7) process.

Thus if we want to (re)construct $b(z)$ from given f_y we have to make a decision which of the (pairs of) roots of f_y correspond to roots of $b(z)$ and which to roots of $b(1/z)$. To make this decision unique we impose the condition

$$b(z) \neq 0 \text{ for all } |z| \leq 1 \quad \textit{(miniphase assumption)} \tag{4.17}$$

With this condition we have $b(z) = c(z - z_1)(z - z_2)\cdots(z - z_q)$, where $z_1, \ldots, z_q$ are the zeros of f_y which have modulus greater than one. The scaling factor c is determined from $b_0 = 1$ and σ^2 is determined from $f_y(z) = \frac{\sigma^2}{2\pi} b(z)b(1/z)$.

Note that by this miniphase assumption we have

$$\epsilon_t = \frac{1}{b(z)} x_t = \sum_{j=0}^{\infty} h_j y_{t-j}$$

and thus the white noise process (ϵ_t) is a causal linear transformation of the outputs.

In this way we have shown: The class of scalar MA systems with $b_0 = 1$ and $b(z) \neq 0$ for all $|z| \leq 1$ is identifiable.

Scalar AR systems: $a(z)y_t = \epsilon_t$.

It is easy to see that the normalization $a_0 = 1$ together with the stability assumption $a(z) \neq 0$ for $|z| \leq 1$ guarantees identifiability:

Instead of looking at the zeros of $f_y(z)$ we here deal with the poles of $f(z) = \frac{\sigma^2}{2\pi}\frac{1}{a(z)a(1/z)}$ which are the zeros of $a(z)$ and $a(1/z)$. Here the stability assumption gives us immediately the possibility to decide whether a pole of $f_y(z)$ corresponds to zero a of $a(z)$ or of $a(1/z)$.

Thus the class of stable AR systems with $a_0 = 1$ is identifiable.

Scalar ARMA systems: $a(z)y_t = b(z)\epsilon_t$.

Theorem 4.4 *The class of (scalar) ARMA systems satisfying*

(i) $a(z) \neq 0$ *for all* $|z| \leq 1$ *(stability assumption)*

(ii) $b(z) \neq 0$ *for all* $|z| \leq 1$ *(miniphase assumption)*

(iii) $a_0 = b_0 = 1$ *and*

(iv) $a(z)$ *and* $b(z)$ *have no common zeros*

is identifiable.

Proof: The spectral density of (y_t) is given by

$$f_y(z) = \frac{\sigma^2}{2\pi}\frac{b(z)b(1/z)}{a(z)a(1/z)}.$$

By assumptions (iv), (i) and (ii) none of the zeros of the denominator of $f_y(z)$ can cancel with a zero of the nominator. Thus collecting all zeros of $f_y(z)$ outside the unit circle gives us the MA part $b(z)$ and the poles of $f_y(z)$ outside the unit circle define $a(z)$. The scaling factor of a and b and σ^2 are determined by the normalization condition (iii). □

Note that the assumptions of the above theorem are slightly to restrictive since not every spectral density may be factorized under these conditions (i)–(iv). By (ii) we exclude spectra which have zeros on the unit circle. This is called *overidentifiablity* since not all ARMA spectra allow for a factorization satisfying these identifiability restrictions.

Vector ARX systems: $a(z)y_t = b(z)x_t + \epsilon_t$:

From equation (4.16) we have

$$f_y - l f_x l^* = \frac{1}{2\pi}a^{-1}(z)\Sigma a^{-*}(z)$$

where the left hand side of this equation is known. Suppose that we have two observationally equivalent ARX systems (a, d) and $(\tilde{a}, \tilde{d})$. Then for each Σ there must exist a $\tilde{\Sigma}$ such that

$$a^{-1}(z)\Sigma a^{-*}(z) = \tilde{a}^{-1}(z)\tilde{\Sigma}\tilde{a}^{-*}(z)$$

holds, which implies

$$\tilde{a}a^{-1}\Sigma = \tilde{\Sigma}\tilde{a}^{-*}a^{*}.$$

By the stability assumption we can expand $r(z) = \tilde{a}(z)a^{-1}(z)$ in convergent Taylor series $r(z) = \sum_{j=0}^{\infty} r_j z^j$ for $|z| < r_1$ where $r_1 > 1$. Since $\tilde{a}^{-*}a^{*} = (a\tilde{a}^{-1})^{*}$ and $\tilde{a}(z)$ is stable we have $\tilde{a}^{-*}a^{*} = \sum_{j=-\infty}^{0} t_j z^j$ for $|z| > r_2$ where $r_2 < 1$. Putting this together we have for $r_2 < |z| < r_1$

$$\sum_{j=0}^{\infty} r_j \Sigma z^j = \sum_{j=-\infty}^{0} \tilde{\Sigma} t_j z^j.$$

By a comparison of the coefficients of these two Laurent series we can conclude that $r_j = 0$ for all $j > 0$ and $t_j = 0$ for all $j < 0$. (Note that $\Sigma > 0$ holds.) Thus we have

$$\tilde{a}(z)a^{-1}(z) = r(z) = r_0 \Longrightarrow \tilde{a}(z) = r_0 a(z).$$

Now since $t_0 = \tilde{a}^{-*}a^{*} = (\tilde{a}a^{-1})^{-*} = r_0^{-*}$ we have

$$\tilde{\Sigma} = r_0 \Sigma t_0^{-1} = r_0 \Sigma r_0^{*}.$$

Of course r_0 must be real and nonsingular. This together with $f_{yx} f_x^{-1} = l = a^{-1}d = \tilde{a}^{-1}\tilde{d}$ implies $\tilde{d} = r_0 d$. Thus we have shown:

Theorem 4.5 *Let $f_x(\lambda) > 0$ for all $\lambda \in [-\pi, \pi]$, $\Sigma > 0$ and $\det(a) \neq 0$ for all $|z| \leq 1$. Then two ARX systems (a, d) and $(\tilde{a}, \tilde{d})$ are observationally equivalent iff there is a constant nonsingular (real) matrix r such that*

$$\tilde{a} = ra \quad \tilde{d} = rd \text{ and } \tilde{\Sigma} = r\Sigma r'$$

holds.

This theorem of course implies that under the normalization condition $a_0 = I$ we have identifiability for ARX systems.

4.4 Structural Identifiability

In many cases (economic) theory (or some other a priori knowledge) gives us relations of the form

$$a_0 y_t + \cdots + a_p y_{t-p} = b_0 x_t + \cdots + b_q x_{t-q} + \epsilon_t$$

where $a_0 \neq I$ but where there are some other restrictions on the coefficients of the matrices a_j and b_j. Especially we often know that some of these entries are zero, e.g. because some variable x_t^i does not influence y_t^j.

The question here is, are the *structural parameters* (*deep parameters* which have an economic meaning) identifiable from the second moments of the data and from the a priori restrictions.

The simplest restrictions on the parameters in this *structural form* (i.e. the form which directly comes from economic theory) are given in the form of "zero restrictions". If we transform this structural form to a *final form* by multiplying by a_0^{-1} and thus obtaining $a_0 = I$ these restrictions are transformed to nonlinear restrictions on the final form parameters in general. Thus it is often more convinient to estimate the parameters in the structural form.

But although the parameter restrictions in the structural form are given in a very simple form there are other problems associated with the estimation of structural parameters. Consider the following simple Keynesian model

$$\begin{aligned} C_t &= \alpha + \beta Y_t + \epsilon_t \\ Y_t &= C_t + I_t \end{aligned}$$

where C_t is consumption Y_t is income and I_t denotes the investments. I_t is considered as an exogenous variable whereas C_t and Y_t are endogenous variables. By substituting C_t from the first equation into the second one we get

$$Y_t = \frac{1}{1-\beta}(\alpha + \beta I_t + \epsilon_t).$$

Thus using OLS to estimate β from the first equation will give a biased estimate since the regressor Y_t is correlated to the noise ϵ_t! (*Haavelmo Bias*) From a system theoretic point of view this bias is caused from the instantaneous *feed back* between C_t and Y_t.

This little example shows that in general OLS is not a suitable method to estimate *simultaneous equations* systems in structural form.

We now want to consider the problem of *structural identifiability*; i.e. the problem whether there is enough a priori information in order to uniquely determine the underlying system from the second moments of the observations and thus from the data? As we have seen in the case of ARX systems two systems, (a,d) and $(\bar{a},\bar{d})$ say, are observationally equivalent if and only if there exists a nonsingular constant matrix such that $(\tilde{a}, \tilde{d}) = r(a, d)$ holds. This equation is equivalent to

$$(\tilde{a}_0, \ldots, \tilde{a}_p, \tilde{d}_0, \ldots, \tilde{d}_s) = r(a_0, \ldots, a_p, d_0, \ldots, d_s).$$

We now assume that there are at least $(n-1)$ zero restrictions in the first equation. Let $\tilde{C}$ and C be the matrices which contain all columns of $(\tilde{a}_0, \ldots)$ and $(a_0, \ldots)$ respectively corresponding to these zero restrictions then we have

$$\tilde{C} = rC$$

(Of course by construction the first row of C and $\tilde{C}$ just contain zeros.) From the first row of this matrix equation we have

$$0 = \tilde{c}_1 = r_1 C = \begin{pmatrix} r_{11} & r_{12} \end{pmatrix} \begin{pmatrix} c_1 \\ C_2 \end{pmatrix} = r_{12} C_2$$

where $\tilde{c}_1$ denotes the first row of $\tilde{C}$, c_1 denotes the first row of C, C_2 the matrix consisting of the last $n-1$ rows of C and $r_1 = (r_{11} r_{12})$ is the first row of r. Note that C_2 has $(n-1)$ rows and by assumption at least $(n-1)$ columns. If we assume that C_2 has full rank $(n-1)$ then $r_{12} = 0$ must hold. If the analogous conditions hold for every equation, then r must be a diagonal matrix. If we in addition impose the simple normalization condition that the diagonal elements of a_0 are equal to one then $r = I$ must hold. (This normalization means that we express in the i-th equation the i-th endogenous variable in terms of its own lags and in terms of the other endogenous variables and exogenous variables.) Thus under these conditions we have identifiability.

Note that the normalization $a_0 = I$ of course satisfies all the conditions for structural identifiability.

In practice often only the condition that the i-th equation contains at least $(n-1)$ a priori zeros is checked, since the condition that C_2 has full rank will be satisfied in "general". (If it is no contradiction to the a priori zeros!) This condition for structural identifiability is often called the *counting condition*.

4.5 ARMAX and state space systems

The problem of finding the internal parameters (a, d, b) from the external characteristics $\mathfrak{f}_y$, $\mathfrak{f}_x$ and $\mathfrak{f}_{yx}$ is split into two parts.

In a first step the transferfunctions l and k are determined from the spectra. Under the persistent excitation condition we have $l = \mathfrak{f}_{yx} \mathfrak{f}_x^{-1}$. Thus we know spectrum of the "error term" $a^{-1}(z)b(z)\epsilon_t$ which is given by $\frac{1}{2\pi} k \Sigma k^*$. The next theorem states under which condition the transferfunction k is uniquely determined from $k \Sigma k^*$.

Theorem 4.6 *Under the following conditions*

$$\begin{array}{l} \det k(z) \neq 0 \text{ for all } |z| \leq 1 \\ k(z) \text{ has a Taylor series expansion in } |z| \leq 1 \\ k(0) = I \\ f_x(\lambda) > 0 \text{ for all } \lambda \in [-\pi, \pi] \\ \Sigma > 0 \end{array} \tag{4.18}$$

the transferfunction l and k are uniquely determined from f_x, f_y and f_{yx}.

Proof: For the transferfunction l the statement is immediate from (4.14).

Suppose the two transferfunction k and $\tilde{k}$ are observationally equivalent, i.e. for every $\Sigma > 0$ there exists a $\tilde{\Sigma}$ such that

$$k\Sigma k^* = \tilde{k}\tilde{\Sigma}\tilde{k}^*$$

holds. From this equation we get

$$\tilde{k}^{-1}k\Sigma = \tilde{\Sigma}\tilde{k}^* k^{-*}.$$

From our assumptions we can conclude that $\tilde{k}^{-1}k$ has a Taylor series expansion for $|z| < r_1$ for some $r_1 > 1$, since $\det \tilde{k}(z) \neq 0$ holds for all $|z| \leq 1$ and since k has a Taylor series expansion within the unit circle. Let $\tilde{k}^{-1}k = \sum_{j=0}^{\infty} r_j z^j$. By analogous consideratios we can conclude that $\tilde{k}^* k^{-*}$ has a convergent Laurent series expansion for $|z| > r_2$ for some $r_2 < 1$ and thus this expansion may not contain any positive powers of z, i.e. $\tilde{k}^* k^{-*} = \sum_{j=-\infty}^{0} t_j z^j$. By a comparison of coefficients of both Laurent series expansions we see that $r_j\Sigma = 0$ and thus $r_j = 0$ must hold for all $j > 0$ since $\Sigma > 0$ holds. We now have shown that $\tilde{k}^{-1}k = r = r_0$ and thus $k = \tilde{k}r_0$ holds. Since $k(0) = \tilde{k}(0) = I$ we have shown that $r_0 = I$ and thus $k = \tilde{k}$ holds. □

If we start from an ARMAX system satisfying the conditions from the above theorem then the transferfunctions are uniquely determined from the spectral densities of the observations. Of course every ARMAX systems generates rational spectral densities. The next theorem now states that also the converse is true. This is one reason why ARMAX systems are that important.

Theorem 4.7 *Any rational and (a.c) nonsingular spectral density may be uniquely factorized as*

$$f = \frac{1}{2\pi}k\Sigma k^*$$

where $k(z)$ is rational, has a Taylor series expansion for $|z| \leq 1$, $\det(k(z)) \neq 0$ for all $|z| \leq 1$, $k(0) = I$ and $\Sigma > 0$ hold.

For a proof of this theorem see e.g. Hannan and Deistler [4].

4.6 Exercises

(4.1) Given a linear VDE

$$a(z)y_t = b(z)x_t$$

and let (y_t^1) be an arbitrary but fixed solution (*particular solution*). Prove that (y_t) is a solution if and only if $y_t = y_t^1 + y_t^0$ holds, where (y_t^0) is a solution of the homogenous difference equation; i.e. $a(z)y_t^0 = 0$.

(4.2) Consider a scalar AR(p) model defined by

$$y_t + a_1 y_{t-1} + \cdots + a_p y_{t-p} = \epsilon_t$$

and a corresponding state space model

$$\begin{aligned} s_{t+1} &= Fs_t + B\epsilon_t \\ y_t &= Cs_t + \epsilon_t \end{aligned}$$

where

$$s_t = \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ \vdots \\ y_{t-p} \end{pmatrix}, A = \begin{pmatrix} -a_1 & -a_2 & \cdots & \cdots & -a_p \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}, B = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, C = \begin{pmatrix} -a_1 \\ -a_2 \\ \vdots \\ \vdots \\ -a_p \end{pmatrix}'.$$

Prove that the stability condition for the AR model ($\det(a(z)) \neq 0$ for all $|z| \leq 1$) is equivalent to the stability condition for the state space model ($|\lambda_{\max}(A)| < 1$).

(4.3) Prove that the normalization condition $a_0 = I$ satisfies all of the conditions of structural identifiability for ARX systems.

(4.4) Compute the steady state solution of the bivariate AR(1) system

$$x_t = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix} x_{t-1} + \epsilon_t$$

where (ϵ_t) is white noise.

5 Prediction and Filtering

In this chapter we are concerned with prediction i.e. with approximation of future values of the observations by past values and with filtering i.e. with approximation of one process by an other.

To be more precise in our context the problem of *prediction* can be described as follows. We commence from a stationary process (x_t) and we want to approximate the future value x_{t+h} $(h > 0)$ by a function of some present and past values $x_s, s \leq t$, where t denotes the present time. In order to make the problem well posed we have to specify the class of feasible approximation functions and the approximation criterion. Here we consider *linear* (or more generally *affine*) approximation functions and the approximation criterion is the *least squares criterion*.

Thus we have to solve the following minimization problem:

$$\min_{b \in \mathbb{R}^n, a_j \in \mathbb{R}^{n \times n}} \mathrm{E}(x_{t+h} - (b + \sum_{j \in \mathcal{J}} a_j x_{t-j}))^*(x_{t+h} - (b + \sum_{j \in \mathcal{J}} a_j x_{t-j}))$$

where $\mathcal{J}$ is either of the form $\mathcal{J} = \{0, 1, \ldots, r\}$ or $\mathcal{J} = \mathbf{Z}^+ = \{0, 1, 2, \ldots\}$. Accordingly we speak about prediction from a finite past or from the infinite past respectively.

In *filtering* in our context we commence from two jointly stationary processes (x_t) and (y_t) (of dimension n and m respectively). We want to approximate y_t by a linear (affine) function of (x_t). Again we are looking for the best approximation in the least squares sense. Thus the problem is as follows:

$$\min_{b \in \mathbb{R}^m, a_j \in \mathbb{R}^{m \times n}} \mathrm{E}(y_t - (b + \sum_{j=-\infty}^{\infty} a_j x_{t-j}))^*(y_t - (b + \sum_{j=-\infty}^{\infty} a_j x_{t-j}))$$

The standing assumption in this chapter is that we know the population (first and) second moments of the processes under consideration. Clearly in applications in most cases these moments have to be estimated. However the analysis presented in this chapter is an important step for prediction and filtering commencing from data. It will turn out that the knowledge of the first and second moments is sufficient to solve the linear least squares problems described above.

Clearly in general the restriction to linear approximation functions is a proper restriction. Note that the (general) least squares approximation (i.e. when the class of approximation functions is the class of measurable functions) is the conditional expectation. E.g. for predicting from the finite past we have

$$\hat{x}_{t,h} = \mathrm{E}(x_{t+h} | x_t, \ldots, x_{t-r}).$$

This approximation problem is significantly more complicated compared to the linear problem. Note that for Gaussian processes the conditional expectation is a linear function.

Thus for such processes, approximation by linear functions is no restriction of generality. Vaguely speaking the best least squares approximation will differ from the linear least squares approximation the more the distributions differ from the Gaussian distribution.

5.1 Prediction from a Finite Past

Here we want to approximate x_{t+h} by an affine function of a finite number of past and present values $x_t, x_{t-1}, \ldots, x_{t-r}$. Thus we have to solve the following minimization problem:

$$\min_{b\in\mathbb{R}, a_j\in\mathbb{R}^{n\times n}} \mathrm{E}(x_{t+h} - \tilde{x}_{t,h})^*(x_{t+h} - \tilde{x}_{t,h}) \quad \text{where } \tilde{x}_{t,h} = b + \sum_{j=0}^{r} a_j x_{t-j}. \tag{5.1}$$

The minimizing $\tilde{x}_{t,h}$ in (5.1) is called the *predictor* of x_{t+h} from $x_t, \ldots, x_{t-r}$. $(x_{t+h} - \hat{x}_{t,h})$ is called the *prediction error* and $\Sigma_h = \mathrm{E}(x_{t+h} - \hat{x}_{t,h})(x_{t+h} - \hat{x}_{t,h})^*$ is called the prediction error variance matrix.

It is clear that this minimization problem may be decomposed into n independent problems (corresponding to each component x_t^i of x_t) of the form

$$\min_{b_i\in\mathbb{R}, a_{i,j}\in\mathbb{R}^{1\times n}} \mathrm{E}(x_{t+h}^i - \tilde{x}_{t,h}^i)^2 \quad \text{where } \tilde{x}_{t,h}^i = b_i + \sum_{j=0}^{r} a_{i,j} x_{t-j}. \tag{5.2}$$

Here b_i corresponds to the i-th component of b in (5.1) and $a_{i,j}$ to the i-th row of a_j in (5.1).

We could solve this problem either by setting the first derivative of the criterion function equal to zero or by using the projection theorem (2.3) in chapter 2. We use the second approach. Let $\mathbf{H}(x_t, x_{t-1}, \ldots, x_{t-r}, 1)$ be the Hilbert space spanned by the components x_{t-j}^i, $i = 1, \ldots, n$ and $j = 0, \ldots, r$ and by the constant 1. If (x_t) is a stochastic process defined on $(\Omega, \mathcal{A}, P)$ then of course $\mathbf{H}(x_t, \ldots, 1)$ is a subspace of the Hilbert space $\mathbf{L}_2(\Omega, \mathcal{A}, P)$. The minimizing $\tilde{x}_{t,h}^i$ in (5.2) by the projection theorem is the projection of x_{t+h}^i on $\mathbf{H}(x_t, \ldots, x_{t-r}, 1)$ and of course putting these $\hat{x}_{t,h}^i$ together into a vector $\hat{x}_{t,h}$ gives us the solution of (5.1). We will often use the notation

$$\hat{x}_{t,h} = \mathrm{P}_{\mathbf{H}(x_t, \ldots, x_{t-r}, 1)} x_{t+h}$$

to indicate that $\hat{x}_{t,h}$ is the vector which is obtained by projecting each component of x_{t+h} onto the space $\mathbf{H}(x_t, \ldots, x_{t-r}, 1)$.

We now want to show how to actually compute the coefficients b and a_j.

Without loss of generality we may assume that $\mathrm{E}\,x_t = 0$ holds. This can be seen as follows: Let (y_t) be some stationary process with $\mathrm{E}\,y_t = \mu \neq 0$. The best affine predictor of y_{t+h} given the values of $y_t, \ldots, y_{t-r}$ is the projection (in the component wise sense as described above) of y_{t+h} on the space spanned by the y^i_{t-j}, $i = 1, \dots, n$, $j = 0, \dots, r$ and the constant 1. If we define a new (centered) process (x_t) by $x_t = y_t - \mu$, then it is clear that the x^i_{t-j}, $i = 1, \dots, n$, $j = 0, \dots, r$ and 1 span the same space.

Since $< x^i_{t-j}, 1 >= \mathrm{E}\, x^i_{t-j} \cdot 1 = 0$ we also see, that $\mathbf{H}(x_t, \dots, x_{t-r}, 1)$ is the orthogonal sum of the spaces $\mathbf{H}(x_t, \dots, x_{t-r})$ and $\mathbf{H}(1)$. Thus the projection on $\mathbf{H}(x_t, \dots, x_{t-r}, 1)$ is the sum of the two corresponding projections on $\mathbf{H}(x_t, \dots, x_{t-r})$ and $\mathbf{H}(1)$. (See exercises.) Therefore we have

$$\hat{x}_{t,h} = \mathrm{P}_{\mathbf{H}(x_t,\dots,x_{t-r},1)}\, x_{t+h} = \mathrm{P}_{\mathbf{H}(x_t,\dots,x_{t-r})}\, x_{t+h} + \underbrace{\mathrm{P}_{\mathbf{H}(1)}\, x_{t+h}}_{0} = \sum_{j=0}^{r} a_j x_{t-j}.$$

Since $\mathrm{E}\, x_t = 0$ the best affine predictor for x_{t+h} is a *linear* function of the x_{t-j}'s. Given a representation of $\hat{x}_{t,h}$ in terms of the x_{t-j}'s, it is easy to give the representation of $\hat{y}_{t,h}$ in terms of the y_{t-j}'s:

$$\begin{aligned}\hat{y}_{t,h} &= \mathrm{P}_{\mathbf{H}(x_t,\dots,x_{t-r},1)}\, y_{t+h} = \underbrace{\mathrm{P}_{\mathbf{H}(x_t,\dots,x_{t-r},1)}\, x_{t+h}}_{\hat{x}_{t,h}} + \underbrace{\mathrm{P}_{\mathbf{H}(x_t,\dots,x_{t-r},1)}\, \mu}_{\mu} \\ &= \textstyle\sum_{j=0}^{r} a_j x_{t-j} + \mu = \sum_{j=0}^{r} a_j y_{t-j} + (\mu - \sum_{j=0}^{r} a_j \mu).\end{aligned}$$

Analogous considerations also hold for the prediction from the infinite past and for the filtering problem. We therefore will from now on assume that the mean of all processes considered is equal to zero.

From the projection theorem we immediately see that $\hat{x}_{t,h} = \sum_{j=0}^{h} a_j x_{t-j}$ is the best linear predictor if and only if the components of $(x_{t+h} - \hat{x}_{t,h})$ are uncorrelated with the components of all x_{t-j}, $j = 0, \dots, s$. Thus we have the following equations:

$$\mathrm{E}(x^i_{t+h} - \hat{x}^i_{t,h})x^j_{t-s} = 0 \quad \text{for } i, j = 1, \dots, n \text{ and } s = 0, \dots, r \quad \Longleftrightarrow$$

$$\mathrm{E}(x_{t+h} - \hat{x}_{t,h})x'_{t-s} = 0 \quad \text{for } s = 0, \dots, r \quad \Longleftrightarrow$$

$$\textstyle\sum_{j=0}^{r} a_j\, \gamma(s-j) = \gamma(h+s) \quad \text{for } s = 0, \dots, r \quad \Longleftrightarrow$$

$$\begin{pmatrix} a_0 & \cdots & a_r \end{pmatrix} \underbrace{\begin{pmatrix} \gamma(0) & \cdots & \gamma(r) \\ \vdots & \ddots & \vdots \\ \gamma(-r) & \cdots & \gamma(0) \end{pmatrix}}_{\Gamma_r} = \begin{pmatrix} \gamma(h) & \cdots & \gamma(h+r) \end{pmatrix} \tag{5.3}$$

This gives us an easy way to compute the coefficients a_j by solving the above matrix equation. Note that the second moments are sufficient to determine this best linear least squares predictor.

If Γ_r is nonsingular then

$$\begin{pmatrix} a_0 & \cdots & a_r \end{pmatrix} = \begin{pmatrix} \gamma(h) & \cdots & \gamma(h+r) \end{pmatrix} \Gamma_r^{-1}$$

We now want to discuss the case where Γ_r is singular. From the projection theorem we know that $\hat{x}_{t,h}$ exists and is unique in any case. If Γ_r is singular, then there are in principle two possibilities:

(i) Equation (5.3) has no solution. But by the projection theorem this cannot occur.

(ii) There are infinite many solutions for $(a_0, \ldots, a_r)$. But although the "coefficients" are not unique the projection $\hat{x}_{t,h}$ is unique.

Lemma 5.1 *Let x be a random vector and $\Sigma = \mathrm{E}\,xx'$. The matrix Σ is singular if and only if there exists a vector $a \in \mathbf{R}^n$, $a \neq 0$ such that $a'x = 0$ holds a.e.*

Proof: If $a'x = 0$ a.e. then $a'(xx') = 0$ a.e. and thus $0 = \mathrm{E}\,a'(xx') = a'\,\mathrm{E}\,xx' = a'\Sigma$.
If Σ is singular then there exists a vector $a \in \mathbf{R}^n$, $a \neq 0$ such that $\Sigma a = 0$ holds. Therefore we have $0 = a'\Sigma a = a'(\mathrm{E}\,xx')a = \mathrm{E}(a'x)^2$. Since $(a'x)^2 \geq 0$ we have $a'x = 0$ a.e.
□

This lemma shows that Γ_r is singular if and only if there exists a vector a such that

$$a' \begin{pmatrix} x_t \\ \vdots \\ x_{t-r} \end{pmatrix} = 0 \quad a.e.$$

This means that the random variables $x_t, \ldots, x_{t-r}$ are linearly dependent and thus form no basis. Thus the representation of $\hat{x}_{t+h}$ as a linear combination of these random variables is not unique; i.e. the coefficients $(a_0, \ldots, a_r)$ are not uniquely determined.

In general we will have

$$x_{t+h} - \hat{x}_{t,h} \neq 0 \text{ a.e.}$$

i.e. we will have no perfect prediction. The variance

$$\Sigma_h = \mathrm{E}(x_{t+h} - \hat{x}_{t,h})(x_{t+h} - \hat{x}_{t,h})'$$

of the prediction errors is a measure for the quality of the prediction.

Note that for this derivation the stationarity of (x_t) is not needed. Stationarity however implies that the matrix Γ_r is a (block) Toeplitz matrix and that the coefficients a_j and Σ_h do not depend on the time t.

Example: Consider an AR(p) process

$$a(z)x_t = \epsilon_t \quad ; \quad a_0 = I \text{ and } \det(a(z)) \neq 0 \quad \forall |z| \leq 1$$

We have

$$x_{t+1} = -a_1 x_t \cdots - a_p x_{t-p} + \epsilon_{t+1}$$

The one-step ahead predictor for $r \geq p$ is given by

$$\hat{x}_{t,1} = -a_1 x_t \cdots - a_p x_{t-p+1}$$

since $\hat{x}_{t,1}$ is a linear combination of the x_{t-j}'s, $j = 0, \ldots, r$ and since the prediction error $x_{t+1} - \hat{x}_{t,1} = \epsilon_{t+1}$ is orthogonal to all x_s, $s = t, \ldots, t-r$. The last statement follows from the stability assumption which implies $x_t = \sum_{j=0}^{\infty} k_j \epsilon_{t-j}$.

The two step ahead predictor is given by

$$\hat{x}_{t,2} = -a_1 \hat{x}_{t,1} - a_2 x_t \cdots - a_p x_{t-p+2}$$

because $\hat{x}_{t,2}$ is a linear combination of the x_{t-j}'s, $j = 0, \ldots, r$ and $x_{t+2} - \hat{x}_{t,2} = \epsilon_{t+2} - a_1 \epsilon_{t+1}$ which is by the same argument as above orthogonal to all x_s, $s = t, \ldots, t-r$.

In a completely analogous way we can also compute the "general" h-step predictor.

5.2 Prediction from the Infinite Past: the ARMA Case

Consider an ARMA process

$$a(z)x_t = b(z)\epsilon_t \quad \text{where} \begin{cases} \det(a(z)) \neq 0 & \forall |z| \leq 1 \\ \det(b(z)) \neq 0 & \forall |z| \leq 1 \\ a_0 = b_0 = I \end{cases}$$

The stability and miniphase assumptions imply that

$$\begin{aligned} x_t &= a^{-1}(z)b(z)\epsilon_t = k(z)\epsilon_t = \textstyle\sum_{j=0}^{\infty} k_j \epsilon_{t-j} \text{ and} \\ \epsilon_t &= b^{-1}(z)a(z)x_t = r(z)x_t = \textstyle\sum_{j=0}^{\infty} r_j x_{t-j}. \end{aligned}$$

Let $\mathbf{H}_x(t)$ be the Hilbert space spanned by all x_s^i, $i = 1, \ldots, n$, $s \leq t$, i.e. the set of all linear combinations of x_s^i, $i = 1, \ldots, n$, $s \leq t$ and their limiting elements. In an analogous way let $\mathbf{H}_\epsilon(t)$ be the Hilbert space spanned by all ϵ_s^i, $i = 1, \ldots, n$, $s \leq t$. Then the above equations imply

$$\mathbf{H}_x(t) = \mathbf{H}_\epsilon(t)$$

Let us write x_{t+h} as

$$x_{t+h} = \sum_{j=0}^{\infty} k_j \epsilon_{t+h-j} = \underbrace{\sum_{j=0}^{h-1} k_j \epsilon_{t+h-j}}_{A} + \underbrace{\sum_{j=h}^{\infty} k_j \epsilon_{t+h-j}}_{B}$$

Then the last term (B) of the above equation is contained in $\mathbf{H}_\epsilon(t) = \mathbf{H}_x(t)$ and thus is a linear combination of past and present values x_s^i, $i = 1, \ldots, n$, $s \leq t$. The last but one term (A) is orthogonal to all elements in $\mathbf{H}_\epsilon(t) = \mathbf{H}_x(t)$ since (ϵ_t) is white noise. Thus using the projecion theorem, we have proved that the h-step predictor $\hat{x}_{t,h}$ is given by the term (B). But we still have to express $\hat{x}_{t,h}$ in terms of the x_{t-j}'s, $j \geq 0$. We have $\epsilon_t = b^{-1}(z)a(z)x_t$ and thus

$$\hat{x}_{t,h} = (k(z) - \sum_{j=0}^{h-1} k_j z^j) b^{-1}(z)a(z)x_{t+h}$$

where $k(z) = a^{-1}(z)b(z)$.

5.3 Wold decomposition

In this secion we deal with a very important result concerning the structure of stationary processes.

Definition 5.1 *A stationary process* (x_t) *is called* singular *if*

$$\hat{x}_{t,h} = x_{t,h} \tag{5.4}$$

holds for some t and $h > 0$ (and thus for all t and $h > 0$).

It is easy to see, that if this equality (5.4) holds for one t and h then it is fullfilled for every t and h. The above condition that x_t is exactly predictable from its own past might also be represented in the form $x_{t+h} \in \mathbf{H}_x(t)$. Singular processes are sometimes also called *deterministic* processes, not because they are nonstochastic but because prediction is deterministic.

An important class of singular processes are the harmonic processes introduced in chapter 2. For simplicity we consider only scalar harmonic processes. The equation

$$x_t = \sum_{j=1}^{h} e^{i\lambda_j t} z_j$$

can be written in the form

$$\begin{pmatrix} x_t \\ x_{t-1} \\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} e^{i\lambda_1 t} & \dots & e^{i\lambda_h t} \\ e^{i\lambda_1 (t-1)} & \dots & e^{i\lambda_h (t-1)} \\ \vdots & & \vdots \end{pmatrix}}_{A} \begin{pmatrix} z_1 \\ \vdots \\ z_h \end{pmatrix}$$

From the proof of the linear independency of the functions $e^{i\lambda\cdot} : \mathbf{Z} \rightarrow \mathbf{C}$ in chapter 2, we see that the matrix A has full rank h. Thus the z_j's may be expressed as a linear combination of the x_s, $s \leq t$. This implies that all z_j, $j = 1, \dots, h$ are contained in $\mathbf{H}_x(t)$ and since $x_{t+1} = \sum_{j=1}^{h} e^{i\lambda_j (t+1)} z_j$ is a linear combination of the z_j's also x_{t+1} is contained in $\mathbf{H}_x(t)$.

Definition 5.2 *A stationary process (x_t) is called* regular, *if*

$$\operatorname*{l.i.m}_{h \rightarrow \infty} \hat{x}_{t,h} = 0$$

holds for one t (and thus for all t).

ARMA processes are examples for regular processes. This is easily seen from:

$$x_{t+h} = \sum_{j=0}^{\infty} k_j \epsilon_{t+h-j} \text{ and } \hat{x}_{t,h} = \sum_{j=h}^{\infty} \epsilon_{t+h-j} \longrightarrow 0 \text{ for } h \rightarrow \infty.$$

The predictor is just given by the "tail" of the infinite sum which represents x_{t+h}. Since this infinite sum converges, the tail must converge to zero if h converges to infinity.

Theorem 5.2 *(Wold decomposition)*

(i) Every stationary process (x_t) can be represented in an unique way as

$$x_t = y_t + z_t \tag{5.5}$$

where (y_t) is regular, (z_t) is singular, (y_t) and (z_t) are orthogonal (i.e. $\mathrm{E}\, y_t z_s' = 0$ for all t, s) and $y_t, z_t \in \mathbf{H}_x(t)$ holds.

(ii) Every regular process (y_t) can be represented as

$$y_t = \sum_{j=0}^{\infty} k_j \epsilon_{t-j} \quad ; \quad \sum_{j=0}^{\infty} \|k_j\|^2 < \infty \tag{5.6}$$

where (ϵ_t) is white noise and $\mathbf{H}_y(t) = \mathbf{H}_\epsilon(t)$ holds.

Proof: We here give just a short sketch of the essential idea of the proof.

Define the one step ahead prediction errors by

$$\epsilon_t = x_t - \hat{x}_{t-1,1} = x_t - \mathrm{P}_{\mathbf{H}_x(t-1)} x_t \tag{5.7}$$

where $\mathbf{H}_x(t-1)$ is the Hilbert space spanned by the past values x^i_s, $i = 1, \ldots, n$, $s \leq t$ and $\mathrm{P}_{\mathbf{H}_x(t-1)} x_t$ denotes the component wise projection of x_t onto this space $\mathbf{H}_x(t-1)$. Clearly we have $\epsilon_t \in \mathbf{H}_x(t)$ and ϵ_t is orthogonal to all elements of $\mathbf{H}_x(s)$ for all $s < t$. Thus $\mathrm{E}\,\epsilon_t \epsilon'_s = 0$ for all $t \neq s$ and since the above construction of ϵ_t is time invariant we can conclude that the process (ϵ_t) is white noise.

Let $y_t = \mathrm{P}_{\mathbf{H}_\epsilon(t)} x_t$ be the projection of x_t on the space $\mathbf{H}_\epsilon(t)$ spanned by all ϵ^i_s, $i = 1, \ldots, n$, $s \leq t$. Since (ϵ_t) is white noise y_t has a representation as a causal infinite MA process

$$y_t = \sum_{j=0}^{\infty} k_j \epsilon_{t-j}$$

and thus is a regular process. Note that $k_0 = I$ holds, since $x_t = \epsilon_t + \hat{x}_{t-1,1}$ and $\hat{x}_{t-1,1}$ is orthogonal to ϵ_t.

In general y_t will not be equal to x_t and we therefore define

$$z_t = x_t - y_t = x_t - \mathrm{P}_{\mathbf{H}_\epsilon(t)} x_t \tag{5.8}$$

Since $z_t \in \mathbf{H}_x(t)$ we have $\mathrm{E}\, z_t \epsilon'_s = 0$ for all $s > t$ and because of the definition (5.8) of z_t also $\mathrm{E}\, z_t \epsilon'_s = 0$ for all $s \leq t$ holds.

To show that z_t is a singular process we use the fact that

$$\begin{aligned} \mathbf{H}_x(t) &= \mathbf{H}(\epsilon_t) \oplus \mathbf{H}_x(t-1) \\ &= \mathbf{H}(\epsilon_t) \oplus \mathbf{H}(\epsilon_{t-1}) \oplus \mathbf{H}_x(t-2) = \ldots \\ &= \mathbf{H}(\epsilon_t) \oplus \cdots \oplus \mathbf{H}(\epsilon_{t-h+1}) \oplus \mathbf{H}_x(t-h) \end{aligned}$$

which implies that $z_t \in \mathbf{H}_x(t-h)$ holds since $z_t \in \mathbf{H}_x(t)$ and z_t is orthogonal to all ϵ_{t-j}, $j = 1, \ldots, h-1$. Thus we have $z_t \in \mathbf{H}_x(s)$, $s \leq t$ and

$$z_t \in \bigcap_{s \leq t} \mathbf{H}_x(t).$$

Because of $x_t = y_t + z_t$ and the orthogonality of (ϵ_t) and (z_t) we have $\mathbf{H}_x(t) = \mathbf{H}_\epsilon(t) \oplus \mathbf{H}_z(t)$, where $\mathbf{H}_z(t)$ denotes the Hilbert space spanned by all z^i_s, $i = 1, \ldots, n$, $s \leq t$. Thus

$$z_{t+1} = \mathrm{P}_{\mathbf{H}_x(t)} z_{t+1} = \underbrace{\mathrm{P}_{\mathbf{H}_\epsilon(t)} z_{t+1}}_{\shortparallel} + \mathrm{P}_{\mathbf{H}_z(t)} z_{t+1}$$

which completes our proof. □

The prediction error ϵ_t is by definition (5.7) the part of x_t which cannot be explained by the past and thus the ϵ_t's are called the *innovations* of the process (x_t).

Let us make some remarks to this theorem:

(i) By (5.6) regular processes have a spectral density, which is given by

$$\mathrm{f}_y = \frac{1}{2\pi} k(\lambda)\Sigma k(\lambda)^*$$

where $k(\lambda) = \sum_j k_j e^{-i\lambda j}$ and $\Sigma = \mathrm{E}\epsilon_t\epsilon_t'$. Note that $\sum_j \|k_j\|^2 < \infty$ implies that the transferfunction $k(\lambda)$ (defined on the interval $[-\pi,\pi]$) exists in the sense of a limit in the mean squares sense. (Note that the space of all square integrable functions defined on $[-\pi,\pi]$ form a Hilbert space if we define an inner product $< f, g >= \int_{-\pi}^{\pi} f(\lambda)\overline{g(\lambda)}d\lambda$. Similar to the Hilbert space L_2 of square integrable random variables we have to consider equivalence classes of functions which are identical on a set of Lebesques measure one, i.e. $f \equiv g \;\Leftrightarrow\; f(\lambda) = g(\lambda)$ λ.a.e.) In general $\sum_j k_j e^{-i\lambda j}$ will not converge pointwise.

Note that the opposite of the above statement is not true in general. Not every stationary process with a spectral density is regular.

(ii) If the representation (5.6) is known, then the prediction problem is straightforward. Let $\hat{y}_{t,1} = \sum_{j=1}^{\infty} k_j\epsilon_{t+1-j}$. Then $\hat{y}_{t,1}$ is an element of $\mathbf{H}_\epsilon(t) = \mathbf{H}_y(t)$ and $y_{t+1} - \hat{y}_{t,1} = k_0\epsilon_{t+1}$ is orthogonal to $\mathbf{H}_\epsilon(t) = \mathbf{H}_y(t)$. Thus $\hat{y}_{t,1}$ is the one step ahead predictor.

The problem is to find the representation (5.6). For the ARMA-case $k(z)$ is given by $a^{-1}b$, but in general computing k from the spectral density f_y may be quite complicated. The problem of finding k given f_y again is the socalled spectral factorization problem.

(iii) Every regular process can be approximated with arbitrary accuracy by an (AR)MA process since:

$$\underset{N\to\infty}{\mathrm{l.i.m}} \sum_{j=0}^{N} k_j\epsilon_{t-j} = \sum_{j=0}^{\infty} k_j\epsilon_{t-j} = y_t$$

Note that this approximation is uniform in t, wheras the approximation of a stationary process by harmonic processes is in general not uniform in t.

For a regular process the influence of a shock in ϵ_0 will decrease over time. The influence of a shock at time $t = 0$ on y_t is given by k_t and since the sum (5.6) converges we must have $||k_t|| \to 0$ for $t \to \infty$. Thus the influence of shocks for regular processes is ***not persistent.***

In economics it is often desirable to have processes where the effect of shock does not vanish with time. One simple example of such a process is a random walk process

$$x_t = \sum_{j=0}^{t} \epsilon_j$$

Here of course the effect of a shock, in e.g. ϵ_0, will never vanish. Thus here the shocks are *persistent.*

5.4 Filtering

The filtering problem may be stated as follows: Let $(x_t', y_t')'$ be a stationary process. We are looking for the best linear approximation of y_t by (x_t) in the least squares sense, i.e we have to solve the minimization problem

$$\min_{L \text{ is linear}} \mathrm{E}(y_t - L(x_t))^*(y_t - L(x(t))$$

If $\mathbf{H}_x$ denotes the Hilbert space spanned by all x_s^i, $i = 1, \ldots, n$, $s \in \mathbf{Z}$ then by the projection theorem we know that the best linear approximation $\hat{y}_t$ is given by

$$\hat{y}_t = \mathrm{P}_{\mathbf{H}_x} y_t$$

The problem now is to express the projection as a linear function of the x_t's. In practically all cases the linear function L may be represented as an infinite sum

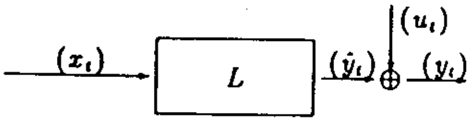
$$L(x_t) = \sum_{j=-\infty}^{\infty} l_j x_{t-j}$$

We thus can write y_t as

$$y_t = \hat{y}_t + \underbrace{(y_t - \hat{y}_t)}_{u_t} = \sum_{j=-\infty}^{\infty} l_j x_{t-j} + u_t.$$

By the projection theorem we know that the approximation error u_t is orthogonal to $\mathbf{H}_x$ and thus $\mathrm{E}\, x_t u_s' = 0$ holds for all t, s.

Thus we can consider the filtering problem as an "infinite" linear least squares regression problem. But we can also think of the filtering problem as finding the "best" linear system relating the inputs (x_t) to the outputs (y_t).



Theorem 5.3 *Let $(x_t', y_t')'$ be stationary with spectral density*

$$\mathfrak{f} = \begin{pmatrix} \mathfrak{f}_x & \mathfrak{f}_{xy} \\ \mathfrak{f}_{yx} & \mathfrak{f}_y \end{pmatrix}$$

with $\mathfrak{f}_x > 0$ for all $\lambda \in [-\pi, \pi]$. Then the transferfunction for the best linear least squares filter is given by

$$l = \mathfrak{f}_{yx}\,\mathfrak{f}_x^{-1}. \tag{5.9}$$

Proof: It is easy to see that the optimal filter L does not depend on time since the two processes (x_t) and (y_t) are jointly stationary. Thus we can define a process $(\hat{y}_t) = L(x_t)$ by $\hat{y}_t = \sum_{j=-\infty}^{\infty} l_j x_{t-j}$. For $\hat{y}_t$ to be the best linear approximation we must have $\mathrm{E}(y_t - \hat{y}_t)x_s' = 0$ for all s. This gives

$$0 = \mathrm{E}(y_t - \hat{y}_t)x_s' = \gamma_{yx}(t-s) - \gamma_{\hat{y}x}(t-s) \text{ for all } s \in \mathbb{Z},$$

where γ_{yx} and $\gamma_{\hat{y}x}$ denotes the cross covariance function between (y_t) and (x_t) and $(\hat{y}_t)$ and (x_t) respectively. This implies

$$\mathfrak{f}_{yx} = \mathfrak{f}_{\hat{y}x} = l\,\mathfrak{f}_x$$

for the corresponding cross spectral densities. Here we have used the formula (3.17) in chapter 3. □

Of course there is a one-to-one relation between the transferfunction $l(z) = \sum_j l_j z^j$ and the corresponding filterweights l_j so that we can actually construct the filterweights l_j from the formula (5.9).

In general the best linear approximation will not be causal. If however $l_j = 0$, $j < 0$, i.e. if $\hat{y}_t = \sum_{j=0}^{\infty} l_j x_{t-j}$, then we say (x_t) is *causing* (y_t). Of course this not *causality* in a strict sense because of the problem of the socalled post hoc ergo propter hoc fellacy (i.e. the fellacy of concludig from precedence to causal influence).

Note that the filter formula (5.9) is a natural generalization of the OLS formula to the case of an infinite number of regressors. By stationarity of course the linear filter L (and thus the filterweights l_j) do not depend on time t.

In chapter 3 we defined the coherence $C^2(\lambda)$ as a frequency dependent squared correlation coefficient. Using the filter formula we can give an additional interpretation. Consider two scalar processes (x_t) and (y_t). Then it is easy to see that

$$C^2(\lambda) = \frac{|\mathfrak{f}_{yx}|^2}{\mathfrak{f}_y \mathfrak{f}_x} = \frac{l \mathfrak{f}_x l^*}{\mathfrak{f}_y} = \frac{\mathfrak{f}_{\hat{y}}}{\mathfrak{f}_y},$$

where $l = \mathfrak{f}_{yx} / \mathfrak{f}_x$ and $\mathfrak{f}_{\hat{y}}$ is the spectrum of the optimal approximation $\hat{y}_t = l(z)x_t$ of y_t. Let us define $u_t = y_t - \hat{y}_t$ as the approximation error, then by the projection theorem (u_t) is orthogonal to (x_t) and thus the spectrum $\mathfrak{f}_y$ may be decomposed as $\mathfrak{f}_y = \mathfrak{f}_{\hat{y}} + \mathfrak{f}_u$. From this decomposition it is clear that $0 \leq C^2(\lambda) \leq 1$ holds and that we may interpret $C^2(\lambda)$ as a frequency dependent coefficient of determination (R^2). If $C^2(\lambda)$ is close to one then for this frequency (-band) $\hat{y}_t$ is a good approximation fo y_t. If $C^2(\lambda)$ is close to zero then at this frequency (-band) y_t cannot be well approximated by a linear function of (x_t).

5.5 Exercises

(5.1) Consider the prediction problem for an n-dimensional stationary process (x_t) and let h, r be two positive integers. Prove that

$$\mathrm{E}(x_{t+h} - \sum_{j=0}^{r} a_j^0 x_{t-j})^*(x_{t+h} - \sum_{j=0}^{r} a_j^0 x_{t-j}) \leq \mathrm{E}(x_{t+h} - \sum_{j=0}^{r} a_j x_{t-j})^*(x_{t+h} - \sum_{j=0}^{r} a_j x_{t-j})$$

for all $a_j \in \mathbf{R}^{n\times n}$, $j = 0, \ldots, r$ holds if and only if

$$\mathrm{E}(x_{t+h} - \sum_{j=0}^{r} a_j^0 x_{t-j})(x_{t+h} - \sum_{j=0}^{r} a_j^0 x_{t-j})^* \leq \mathrm{E}(x_{t+h} - \sum_{j=0}^{r} a_j x_{t-j})(x_{t+h} - \sum_{j=0}^{r} a_j x_{t-j})^*$$

for all $a_j \in \mathbf{R}^{n\times n}$, $j = 0, \ldots, r$ holds.

(5.2) Let $\mathbf{H}_1$ and $\mathbf{H}_2$ be two orthogonal subspaces of some Hilbert space $\mathbf{H}$, i.e. $< x, y >= 0$ for all $x \in \mathbf{H}_1$ and $y \in \mathbf{H}_2$ where $< \cdot, \cdot >$ denotes the inner product in $\mathbf{H}$. Prove that the projection on the sum of $\mathbf{H}_1$ and $\mathbf{H}_2$ is equal to the sum of the two projections on these spaces:

$$\mathrm{P}_{\mathbf{H}_1 \oplus \mathbf{H}_2} y = \mathrm{P}_{\mathbf{H}_1} y + \mathrm{P}_{\mathbf{H}_2} y \quad \forall y \in \mathbf{H}$$

(5.3) Let (y_t) be some stationary process with $\mathrm{E} y_t = \mu$. Define the "centered" process (x_t) by $x_t = y_t - \mu_t$ and let $\hat{x}_{t,h} = \sum_{j=0}^{r} a_j x_{t-j}$ be the best affine predictor of x_{t+h} given $x_t, \ldots, x_{t-r}$. Prove that best affine predictor for y_{t+h} given $y_t, y_{t-1}, \ldots, y_{t-r}$ is given by

$$\hat{y}_{t,h} = \sum_{j=0}^{r} a_j y_{t-j} + \mu - \sum_{j=0}^{r} a_j \mu$$

(5.4) Prove that the normal equations for the best linear predictor always have a solution.

(5.5) Consider the steady state solution of the AR(1) system $x_t = ax_{t-1} + \epsilon_t$ where (ϵ_t) is white noise and $|a| > 1$. Compute the best linear predictor for x_{t+1} given x_t and the best linear predictor for x_{t+1} given x_t, x_{t-1}.

(5.6) Consider the MA(1) process defined by $x_t = \epsilon_t + b\epsilon_{t-1}$ where (ϵ_t) is white noise. Compute the best linear predictor for x_{t+1} given x_t and the best linear predictor for x_{t+1} given x_t, x_{t-1}.

6 Estimation of the Mean and of the Covariance Function

Estimation and inference in the context of time series analysis is an essential generalization of classical i.i.d analysis. In time series analysis in most cases finite sample properties are difficult to obtain analytically. Therefore in the next chapters the emphasis will be on asymptotic theory in order to obtain some indications of the properties of estimation and test procedures for large samples. It should be noted that besides the analytical results also simulation results are important for the evaluation of estimation and test procedures.

In this chapter we consider estimation of the mean and the covariance function. The estimation of the covariance function $\gamma(\cdot)$ is a nonparametric problem since there are infinite many values to be estimated.

6.1 Convergence Concepts

In this preliminary section we will repeat concepts for the convergence of random variables and some related properties.

Definition 6.1 *A sequence x_t of scalar random variables is said to converge in* probability *to x_0 if for every $\epsilon > 0$*

$$\lim_{t\to\infty} P\{|x_t - x_0| > \epsilon\} = 0 \textit{ holds.}$$

We will use the notations $x_t \xrightarrow{p} x_0$ and $\operatorname{plim}_{t\to\infty} x_t = x_0$ for convergence in probability.

In the vector case we define the convergence in probability component wise, i.e.

$$x_t = \begin{pmatrix} x_t^1 \\ \vdots \\ x_t^n \end{pmatrix} \xrightarrow{p} x_0 = \begin{pmatrix} x_0^1 \\ \vdots \\ x_0^n \end{pmatrix} \iff x_t^i \xrightarrow{p} x_0^i \textit{ for } i = 1,\ldots,n.$$

Convergence in probability is also called stochastic convergence.

It is easy to see that the limit x_0 is unique a.e.

Lemma 6.1 *If $x_0 = \text{l.i.m}_{t\to\infty} x_t$ then $x_0 = \operatorname{plim}_{t\to\infty} x_t$.*

Proof: We can proof this lemma component wise. By the Chebyshev inequality we have

$$P\{|x_t^i - x_0^i| > \epsilon\} \leq \epsilon^{2}\, \mathrm{E}(x_t^i - x_0^i)^2 \to 0. \; \square$$

Definition 6.2 *A sequence of random variables x_t with distribution functions F_t is said to* converge in distribution *to a random variable x_0 with distribution function F_0 if*

$$\lim_{t\to\infty} F_t(x) = F_0(x)$$

holds for all continuity points x of F_0. We will use the notation $x_t \xrightarrow{d} x_0$ for convergence in distribution.

Of course this definition of convergence in distribution is related to the convergence of measures rather than to the convergence of random variables. The limit x_0 is not unique which can be seen from the following example: Let (x_t) be a sequence of i.i.d. (independent and identically distributed) random variables then $x_t \xrightarrow{d} x_s$ holds for every x_s.

Lemma 6.2 *If* $\operatorname{plim}_{t\to\infty} x_t = x_0$ *holds then* $x_t \xrightarrow{d} x_0$.

Proof: We only prove the scalar case. For each $\epsilon > 0$ we have

$$\begin{aligned} F_t(x) &= P(x_t \leq x) \\ &= \underbrace{P(x_0 \leq x + (x_0 - x_t) \text{ and } |x_t - x_0| \leq \epsilon)}_{A} + \\ &\quad \underbrace{P(x_0 \leq x + (x_0 - x_t) \text{ and } |x_t - x_0| > \epsilon)}_{\leq P(|x_t - x_0| > \epsilon) \to 0} \end{aligned}$$

For the term A we have

$$\begin{aligned} A \leq\ & P(x_0 \leq x + \epsilon \text{ and } |x_t - x_0| \leq \epsilon) = \\ & P(x_0 \leq x + \epsilon) + \underbrace{P(|x_t - x_0| \leq \epsilon)}_{\to 1} - \underbrace{P(x_0 \leq x + \epsilon \text{ or } |x_t - x_0| \leq \epsilon)}_{\to 1} \end{aligned}$$

and

$$\begin{aligned} A \geq\ & P(x_0 \leq x - \epsilon \text{ and } |x_t - x_0| \leq \epsilon) = \\ & P(x_0 \leq x - \epsilon) + \underbrace{P(|x_t - x_0| \leq \epsilon)}_{\to 1} - \underbrace{P(x_0 \leq x - \epsilon \text{ or } |x_t - x_0| \leq \epsilon)}_{\to 1} \end{aligned}$$

Thus we have

$$F_0(x - \epsilon) \leq \lim_{t\to\infty} F_t(x) \leq F_0(x + \epsilon)$$

which proves the lemma. □

Definition 6.3 *A sequence x_t of random variables is said to be* asymptotically normal *with "expectation" μ_t and "variance" Σ_t, if Σ_t is nonsingular from a certain t_0 onwards and if*

$$\Sigma_t^{-1/2}(x_t - \mu_t) \xrightarrow{d} z$$

where $z \sim \mathrm{N}(0, I)$. We will use the notation $x_t \sim \mathrm{AN}(\mu_t, \Sigma_t)$.

Note that in general we do not impose the condition that $\mathrm{E}\, x_t = \mu_t$ and $\mathrm{Var}\, x_t = \Sigma_t$ must hold.

Each nonnegative definite matrix $\Sigma \geq 0$ may be factorized as $\Sigma = AA'$. Such a factor A is often denoted by $A = \Sigma^{1/2}$. Thus $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$ satisfies $\Sigma^{-1/2}\Sigma(\Sigma^{-1/2})' = I$. Furthermore if $\mu_t = \mathrm{E}\, x_t$ and $\mathrm{Var}\, x_t = \Sigma_t$ holds then $\tilde{x}_t = \Sigma^{-1/2}(x_t - \mu_t)$ has mean zero and variance I.

Lemma 6.3 *(Slutzky) Let x_t be a sequence of random vectors with $\operatorname{plim}_{t \to \infty} x_t = x_0 \in \mathbf{R}^n$ and $\mathrm{g} : \mathbf{R}^n \to \mathbf{R}$ be a function which is continuous in x_0. Then*

$$\operatorname*{plim}_{t \to \infty} \mathrm{g}(x_t) = \mathrm{g}(\operatorname*{plim}_{t \to \infty} x_t) = \mathrm{g}(x_0)$$

holds.

Proof: Since g is continuous in x_0 there exists for every $\epsilon > 0$ a $\delta > 0$ such that $|\mathrm{g}(x) - \mathrm{g}(x_0)| \leq \epsilon$ holds for all $x \in \mathbf{R}^n$ with $|x^i - x_0^i| \leq \delta$, $i = 1, \ldots, n$. Thus

$$P(|\mathrm{g}(x_t) - \mathrm{g}(x_0)| > \epsilon) \leq \sum_{i=1}^{n} P(|x_t^i - x_t^i| > \delta) \longrightarrow 0. \ \Box$$

6.2 Estimation of the Mean

A natural estimate of the population mean of the process is the sample mean

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

Since the estimation of the mean is done component wise we will restrict ourselves to the scalar case in this section.

It is trivial to see that

$$\mathrm{E}\, \bar{x}_t = \mathrm{E}\, x_t = \mu$$

holds for every stationary process and thus the sample mean is an unbiased estimate of the population mean.

Theorem 6.4 *Let (x_t) be a scalar stationary process with* $\mathrm{E}\,x_t = \mu$ *and* $\gamma(s) = \mathrm{E}\,x_s x_0$. *Then*

(i) $\lim_{T\to\infty} \operatorname{Var} \bar{x}_T = \lim_{T\to\infty} \mathrm{E}(\bar{x}_T - \mu)^2 = 0$ *if* $\lim_{s\to\infty} \gamma(s) = 0$ *holds and*

(ii) $\lim_{T\to\infty} T \operatorname{Var} \bar{x}_T = \sum_{s=-\infty}^{\infty} \gamma(s)$ *if* $\sum_{s=-\infty}^{\infty} |\gamma(s)| < \infty$ *holds.*

Proof:

$$\begin{aligned} T \operatorname{Var} \bar{x}_T &= T\,\mathrm{E}(\bar{x}_T - \mu)^2 = \tfrac{1}{T}\,\mathrm{E}\left(\textstyle\sum_{t=1}^{T}(x_t - \mu)\sum_{s=1}^{T}(x_s - \mu)\right) \\ &= \tfrac{1}{T}\textstyle\sum_{t,s=1}^{T} \gamma(t-s) = \tfrac{1}{T}\sum_{s=-T+1}^{T-1} \gamma(s)(T - |s|) \\ &= \textstyle\sum_{|s|<T} \gamma(s)(1 - \frac{|s|}{T}) \end{aligned}$$

If $\gamma(s) \to 0$ then also $\gamma(s)(1 - \frac{|s|}{T}) \to 0$ and thus

$$\lim_{T\to\infty} \operatorname{Var} \bar{x}_T = \lim_{T\to\infty} \frac{1}{T} \sum_{|s|<T} \gamma(s)(1 - \frac{|s|}{T}) = 0$$

which proves (i). (See exercises.)

If $\sum_s |\gamma(s)| < \infty$ holds then for every $\epsilon > 0$ there exist $0 < H < T < \infty$ such that $H/T \sum_s |\gamma(s)| \leq \epsilon/2$ and $\sum_{|s|>H} |\gamma(s)| \leq \epsilon/2$ holds. Thus we have

$$\begin{aligned} \left|T \operatorname{Var} \bar{x}_T - \textstyle\sum_{s=-\infty}^{\infty} \gamma(s)\right| &= \left|-\textstyle\sum_{|s|<T} \frac{|s|}{T}\gamma(s) - \sum_{|s|\geq T} \gamma(s)\right| \\ &\leq \textstyle\sum_{|s|\leq H} \frac{|s|}{T} |\gamma(s)| + \sum_{H<|s|<T} \frac{|s|}{T} |\gamma(s)| + \sum_{|s|\geq T} |\gamma(s)| \\ &\leq \tfrac{H}{T} \textstyle\sum_{s=-\infty}^{\infty} |\gamma(s)| + \sum_{H<|s|} |\gamma(s)| \\ &\leq \epsilon \end{aligned}$$

which implies $T \operatorname{Var} \bar{x}_T \longrightarrow \sum_s \gamma(s)$ for $T \longrightarrow \infty$. □

Let us make some remarks to the theorem above:

(i) The essential condition in this theorem is that the memory of the process is fading i.e. $\gamma(s) \to 0$ for $s \to \infty$.

(ii) The first part (i) of the theorem states that $\text{l.i.m}\,\bar{x}_T = \mu$ and thus also $\operatorname{plim} \bar{x}_T = \mu$, which means that the sample mean is a *weakly consistent* estimate for the population mean (if $\lim_{s\to\infty} \gamma(s) = 0$).

(iii) The second part (ii) of the theorem states that the speed of convergence is $\sqrt{T}$ if $\sum_s \gamma(s) \neq 0$ holds.

(iv) The assumption $\sum_s |\gamma(s)| < \infty$ implies that the spectral density of the process exists and is given by $f_x(\lambda) = (2\pi)^{-1} \sum_s \gamma(s) e^{-i\lambda s}$ and thus $T \operatorname{Var} \bar{x}_T \to \sum_s \gamma(s) = 2\pi f_x(0)$. Since $T \operatorname{Var} \bar{x}_T$ is a measure of the quality of the estimate $\bar{x}_T$ we see that the estimate is the better the smaller $f_x(0)$ is. E.g. if $f_x(0)$ is high then the process contains big "slow variations" and therefore it is difficult to estimate the mean of the process.

(v) If $f_x(0) = 0$ then we may have a higher rate of convergence as the following simple example shows. Consider the MA(1) process $x_t = \epsilon_t - \epsilon_{t-1}$. The spectral density is given by $f_x(\lambda) = \frac{\sigma^2}{2\pi}|1 - e^{-i\lambda}|^2 = \frac{\sigma^2}{2\pi}(2 - 2\cos\lambda)$ which is zero for $\lambda = 0$. The sample mean is given by

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T}(\epsilon_t - \epsilon_{t-1}) = \frac{1}{T}(\epsilon_T - \epsilon_0)$$

which of course converges to zero at rate which is faster than $\sqrt{T}$.

(vi) For an infinite MA process $x_t = \sum_{j=-\infty}^{\infty} k_j \epsilon_{t-j}$ where $\sum_j |k_j| < \infty$ holds the spectral density is given by $f_x = |k|^2 \frac{\sigma^2}{2\pi}$ and $2\pi f_x(0) = (\sum_j k_j)^2 \sigma^2$. This formula again shows that $f_x(0)$ is a measure for the memory of the process.

The following theorem gives us a more general result, stating that the sample mean always converges in the mean squares sense to some random variable and gives us a nice condition under which the sample mean converges to the population mean.

Theorem 6.5 *For every stationary process (x_t) we have*

$$\operatorname*{l.i.m}_{T\to\infty} \bar{x}_T = z(0) - z(0-)$$

where $z(\lambda)$ is the process with orthogonal increments associated with (x_t) and $z(0-)$ denotes the left limit of $z(\lambda)$ at the frequency $\lambda = 0$, i.e. $z(0-) = \operatorname{l.i.m}_{\epsilon\downarrow 0} z(0-\epsilon)$.

Proof: From the spectral representation of the process (x_t) we immediately get

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T}\int_{-\pi}^{\pi} e^{i\lambda t} dz(\lambda) = \int_{-\pi}^{\pi} \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T} e^{i\lambda t}\right)}_{J_T(\lambda)} dz(\lambda)$$

Since $|e^{i\lambda t}| = 1$, we have $|J_T(\lambda)| \leq 1$. Furthermore by using the formula for (finite) geometric sums we get

$$J_T(\lambda) = \begin{cases} 1 & \text{for } \lambda = 0 \\ \frac{1}{T} e^{i\lambda} \frac{1 - e^{i\lambda T}}{1 - e^{i\lambda}} & \text{for } \lambda \neq 0 \end{cases}$$

From this formula it is easy to see that $J_T(\lambda)$ converges to the indicator function $\kappa_{\{0\}}(\lambda)$ which is defined by

$$\kappa_{\{0\}}(\lambda) = \begin{cases} 1 & \text{for } \lambda = 0 \\ 0 & \text{else.} \end{cases}$$

Finaly we have

$$\underset{T\to\infty}{\text{l.i.m}}\, \bar{x}_T = \underset{T\to\infty}{\text{l.i.m}} \int_{-\pi}^{\pi} J_T(\lambda) dz(\lambda) = \int_{-\pi}^{\pi} \underbrace{\lim_{T\to\infty} J_T(\lambda)}_{\kappa_{\{0\}}(\lambda)} dz(\lambda) = z(0) - z(0-)$$

□

This theorem has a very nice interpretation. It states that the sample mean converges to the jump of the spectral process $z(\lambda)$ at freqency 0. For a stationary process we also have

$$\mu = \mathrm{E}\, x_t = \mathrm{E}(z(0) - z(0-)$$

and thus the mean is just the expectation of this jump of $z(\lambda)$ at frequency zero. Therefore the sample mean will converge to μ if and only if the variance of this jump is zero, i.e. if $z(0) - z(0-) = \mu$ holds. If this jump is not a degenerated random variable, i.e. if it has a nonzero variance, then the sample mean will not be a consistent estimate of μ.

Let us decompose the process (x_t) into two parts of the form

$$x_t = \int_{-\pi}^{\pi} e^{i\lambda t} dz(\lambda) = \underbrace{\int_{-\pi}^{\pi} e^{i\lambda t} d\tilde{z}(\lambda)}_{\tilde{x}_t} + \underbrace{z(0) - z(0-)}_{m_t}$$

where $\tilde{z}(\lambda)$ is defined by

$$\tilde{z}(\lambda) = \begin{cases} z(\lambda) & \text{for } \lambda < 0 \\ z(\lambda) - (z(0) - z(0-)) & \text{for } \lambda \geq 0 \end{cases}$$

By construction $\tilde{z}(\lambda)$ has no jump at $\lambda = 0$ (i.e. $\tilde{z}(0) - \tilde{z}(0-) = 0$) and thus $(\tilde{x}_t)$ has mean zero and the sample mean of $\tilde{x}_t$ converges to zero. The second part of x_t namely $m_t = (z(0) - z(0-))$ is a stochastic process whose trajectories are constant. Thus given only one trajectory we never can consistently estimate the mean of m_t unless $(z(0) - z(0-)$ is a constant random variable which in other words means that its variance is zero. (See also figure 2.4.)

We can check this condition for consistency of the sample mean also by looking at the spectral distribution function. Since $z(\lambda)$ is process with orthogonal increments and by Steiner's lemma we have

$$F(0) - F(0-) = \mathrm{E}\,|z(0) - z(0-)|^2 = \mathrm{Var}(z(0) - z(0-)) + (\underbrace{\mathrm{E}(z(0) - z(0-))}_{\mu})^2$$

Thus the sample mean is consistent iff the jump of the spectral distribution function equals the square of the mean μ. If the jump of $F(\lambda)$ at frequency 0 is larger than μ^2 then the variance of $z(0) - z(0-)$ is not equal to zero and therefore $\bar{x}_T$ is not consistent.

From this formula above it is also clear that if the spectral distribution function has no jump at $\lambda = 0$ (e.g. if the spectral density exists) then $z(0) - z(0-)$ must be zero and therefore the mean μ of the process must be zero and the sample mean will converge to zero too.

The next theorem gives a set conditions under which the sample mean is asymptotically normal.

Theorem 6.6 *Let (x_t) be a stationary process with $x_t = \mu + \sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j}$, $\sum_j |b_j| < \infty$, $(\epsilon_t) \sim \mathrm{IID}(0, \sigma^2)$ and $\sum_j b_j \neq 0$. Then $\bar{x}_T \sim \mathrm{AN}(\mu, T^{-1}v)$ where $v = \sum_{s=-\infty}^{\infty} \gamma(s) = \sigma^2 (\sum_{j=-\infty}^{\infty} b_j)^2$.*

For a proof of this theorem see e.g. Brockwell and Davis [3].

Here we used the notation $(\epsilon_t) \sim \mathrm{IID}(0, \sigma^2)$ to indicate that (ϵ_t) is a sequence of indipendent and identically distributed random variables with mean zero and variance σ^2. Of course this assumption implies that (ϵ_t) is a white noise process.

Results of asymptotic normality and thus also the result of the theorem above can be used for the construction of tests and confidence intervals.

6.3 Estimation of the Covariances and Correlations

In chapter 1 we have introduced the sample covariances

$$\begin{array}{rcll} \hat{\gamma}_T(s) & = & \frac{1}{T}\sum_{t=1}^{T-s}(x_{t+s} - \bar{x}_T)(x_t - \bar{x}_T)' & \text{for } s \geq 0 \\ \hat{\gamma}_T(s) & = & \hat{\gamma}_T(-s)' & \text{for } s < 0 \end{array} \tag{6.1}$$

which are of course estimates for the (population) covariances. By some simple algebra we obtain the following expression for the sample autocovariance for $s \geq 0$:

$$\hat{\gamma}_T(s) = \frac{1}{T}\sum_{t=1}^{T-s} x_{t+s}x_t' - \bar{x}_T\bar{x}_T' + \frac{1}{T}(\sum_{t=1}^{s} x_t\bar{x}_T' + \sum_{t=T-s+1}^{T} \bar{x}_T x_t').$$

If $\mu = \mathrm{E}\, x_t$ denotes the mean and $\Sigma_{\bar{x}_T} = \mathrm{E}(\bar{x}_T - \mu)(\bar{x}_T - \mu)'$ denotes the variance of the sample mean then using Steiner's lemma we obtain for the expectation of the sample autocovariance $\hat{\gamma}_T(s)$:

$$\begin{array}{rcl} \mathrm{E}\,\hat{\gamma}_T(s) & = & \frac{T-s}{T}(\gamma(s) + \mu\mu') - (\Sigma_{\bar{x}_T} + \mu\mu') + \\ & & + \frac{1}{T^2}\left(\sum_{r=1}^{T}\sum_{t=1}^{s}(\gamma(t-r) + \mu\mu') + \sum_{r=1}^{T}\sum_{t=T-s+1}^{T}(\gamma(r-t) + \mu\mu')\right) \end{array}$$

This expression for $\mathrm{E}\,\hat{\gamma}_T(s)$ is quite complicated and indicates that the sample autocovariance in general is a biased estimate for the autocovariance function. But if $\bar{x}_T$ is a consistent estimate for the mean in the mean squares sense, i.e. if $\mathrm{l.i.m}_{T\to\infty}\,\bar{x}_T = \mu$ holds, then $\hat{\gamma}_T(s)$ is asymptotically unbiased, i.e. $\lim_{T\to\infty}\mathrm{E}\,\hat{\gamma}_T(s) = \gamma(s)$. This can easily be seen, since all summands in the last term of the above equation are bounded, i.e. $\|\gamma(k)+\mu\mu'\| \leq M$ for all k and since $\mathrm{l.i.m}_{T\to\infty}\,\bar{x}_T = \mu$ implies $\Sigma_{\bar{x}_T} \to 0$.

The matrix

$$\hat{\Gamma}_N = \begin{pmatrix} \hat{\gamma}_T(0) & \cdots & \hat{\gamma}_T(-N+1) \\ \vdots & \ddots & \vdots \\ \hat{\gamma}_T(N-1) & \cdots & \hat{\gamma}_T(0) \end{pmatrix}$$

is an estimate of the corresponding matrix Γ_N (where the block entries are the $\gamma(s)$'s). Γ_N is non negative definite since $\gamma(s)$ is a covariance function of a stationary process. It is easy to prove that also $\hat{\Gamma}_N$ is non negative definite. (See exercises in chapter 1). Thus $\hat{\Gamma}_N$ ($\hat{\gamma}_T(s)$) is a *proper* estimate of Γ_N ($\gamma(s)$ respectively) since it satisfies the same restrictions as Γ_N. Note that this is not true for the estimate $\hat{\hat{\gamma}}(s)$ defined by

$$\begin{array}{rcll} \hat{\hat{\gamma}}(s) & = & \frac{1}{T-s}\sum_{t=1}^{T-s}(x_{t+s}-\bar{x}_T)(x_t-\bar{x}_T)' & \text{for } s \geq 0 \\ \hat{\hat{\gamma}}(s) & = & \hat{\hat{\gamma}}(-s)' & \text{for } s < 0 \end{array}$$

since the corresponding matrix $\hat{\hat{\Gamma}}_N$ is not non negative definite in general.

The estimate $\hat{\gamma}_T(s)$ is equal to zero for $s \geq T$, which in many cases might not be a good estimate for $\gamma(s)$.

The quality of the estimate $\hat{\gamma}_T(s)$ depends on the lag s. For small s there will be more summands in the sum (6.1) than for large s. Thus the estimate will be more reliable for small s than for large s (for given T).

For simplicity of notation and because of the importance of the scalar case we will from now on consider the scalar case only.

We now want to give sufficient conditions under which the sample covariances are consistent. The essential idea here is to define an artificial process $(y_t|t \in \mathbb{Z})$ by

$$y_t = x_{t+s}x_t \quad ; \quad t \in \mathbb{Z}$$

The mean of this process is equal to the noncentral autocovariance $\mathrm{E}\,x_{t+s}x_t$ at lag s and the sample mean of this process (y_t) is essentially equal to the noncentral sample covariance of the original process (x_t). Thus if (y_t) satisfies the conditions for consistency of the sample mean then the sample autocovariances for (x_t) will be consistent too. We state this result in the following theorem.

Theorem 6.7 *Let (x_t) and $(y_t = x_{t+s}x_t)$ be stationary. If*

$$\begin{aligned} \textstyle\sum_{r=-\infty}^{\infty} |\gamma_x(r)| &< \infty \quad \textit{and} \\ \textstyle\sum_{r=-\infty}^{\infty} |\mathrm{E}(y_r - \mathrm{E}\, y_r)(y_0 - \mathrm{E}\, y_0)| &< \infty \end{aligned}$$

then $\operatorname{plim}_{T\to\infty} \hat{\gamma}_{x,T}(s) = \gamma_x(s)$ *holds where $\gamma_x(s)$, $\hat{\gamma}_{x,T}(s)$ denote the autocovariance and the sample autocovariance function of the process (x_t) respectively.*

Proof: For simplicity we consider only the case where $\mathrm{E}\, x_t = 0$ holds.

The sample mean $\bar{y}_{T-s}$ for the artificial process (y_t) based on a sample $1, \ldots, T-s$ is given by

$$\bar{y}_{T-s} = \frac{1}{T-s} \sum_{t=1}^{T-s} x_{t+s}x_t = \frac{T}{T-s} \hat{\gamma}_{x,T}(s)$$

Since the assumptions imply that the spectral density of the process (y_t) exists we have from the results of the last section that the sample mean $\bar{y}_{T-s}$ and thus also $\hat{\gamma}_{x,T}(s)$ is a consistent estimate of $\gamma_x(s)$. □

The condition that $(y_t) = (x_{t+s}x_t)$ is a stationary process is mainly a condition on the fourth moments of the process (x_t). They must exist and have a "stationary" structure.

Lemma (6.3) implies for $\gamma(0) \neq 0$ that the estimate

$$\hat{\rho}_T(s) = \frac{\hat{\gamma}_T(s)}{\hat{\gamma}_T(0)}$$

is a (weakly) consistent estimate of the autocorrelation $\rho(s)$ at lag s if $\hat{\gamma}_T$ is a (weakly) consistent estimate for the autocovariance function.

The next theorem states that the sample autocovariances are asymptotically normal under fairly general assumptions.

Theorem 6.8 *If (x_t) is stationary and of the form $x_t = \mu + \sum_{j=-\infty}^{\infty} b_j\epsilon_{t-j}$ where $\sum_j |b_j| < \infty$, $(\epsilon_t) \sim \mathrm{IID}(0, \sigma^2)$ and $\mathrm{E}\,\epsilon_t^4 < \infty$ then*

$$\hat{\rho}_{H,T} = \begin{pmatrix} \hat{\rho}_T(1) \\ \vdots \\ \hat{\rho}_T(H) \end{pmatrix} \sim \mathrm{AN}(\rho_H, T^{-1}W) \textit{ where } \boldsymbol{\rho}_H = \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(H) \end{pmatrix}$$

and where the i, j-th entry of W is given by

$$\begin{aligned} w_{ij} = \textstyle\sum_{k=-\infty}^{\infty}(\quad & \rho(k+i)\,\rho(k+j) + \rho(k-i)\,\rho(k+j) + \\ & 2\,\rho(i)\,\rho(j)\,\rho^2(k) - 2\,\rho(i)\,\rho(k)\,\rho(k+j) - 2\,\rho(j)\,\rho(k)\,\rho(k+i)). \end{aligned}$$

For a proof of this theorem see e.g. Brockwell and Davis [3].

Example: Consider white noise (ϵ_t) which in addition satisfies $(\epsilon_t) \sim \text{IID}(0, \sigma^2)$. Thus the ϵ_t's are not only uncorrelated but also independent. In this case we have

$$w_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{else.} \end{cases}$$

Thus $\hat{\rho}_{H,T} \sim \text{AN}(0, T^{-1}I)$.

Now it is easy to construct a test for white noise. The test statistic

$$Q = T \sum_{s=1}^{H} \hat{\rho}_T(s)^2$$

should be close to zero for white noise. Under the Null hypothesis that (ϵ_t) is white noise, the test statistic Q is asymptotically Chi squared with H degrees of freedom, i.e. $Q \xrightarrow{d} X$ where $X \sim \mathcal{X}^2_H$. This test is very often used and is called a *Portmanteau* test.

In a similar way we could also test wether a process is an MA process of order less than or equal to q by using a test statistic of the form

$$Q = T \sum_{s=q+1}^{H} \hat{\rho}_T(s)^2$$

But in this case the asymptotic variance W of $\hat{\rho}_{H,T}$ is much more complicated to compute.

6.4 Exercises

(6.1) Given a covariance function $\gamma(s)$ where $\lim_{s\to\infty} \gamma(s) = 0$. Prove that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{|s|<T} \gamma(s) = 0$$

holds.

(6.2) Consider a stationary process (x_t) defined by $x_t = \mu + \epsilon_t$ where (ϵ_t) is white noise and $\text{E}\,\epsilon_t\epsilon_t' = \Sigma_\epsilon$. Prove that

$$\widehat{\text{Var}}\, x_t = \frac{1}{T-1} \sum_{t=1}^{T} (x_t - \bar{x}_T)(x_t - \bar{x}_T)'$$

is an unbiased estimate for $\text{Var}\, x_t = \text{Var}\, \epsilon_t = \Sigma_\epsilon$.

(6.3) Show that for every time series $x_1, \ldots, x_T$, the sample autocovariances satisfy $\sum_{|s|<T} \hat{\gamma}_T(s) = 0$.

(6.4) Compute the asymptotic variance of $\hat{\rho}_T(1)$ for an AR(1) process $x_t = ax_{t-1} + \epsilon_t$.

(6.5) For an AR(1) process $x_t = ax_{t-1} + \epsilon_t$ the sample autocorrelation $\hat{\rho}_T(1)$ is AN$(a, (1 - a^2)\frac{1}{T})$. Show that $\sqrt{T}(\hat{\rho}_T(1) - a)/\sqrt{(1 - a^2)}$ is AN$(0, 1)$. If a sample size of $T = 100$ from an AR(1) process gives $\hat{\rho}_T(1) = 0.638$, construct an 95% confidence interval for a. Is the data consistent with $a = 0.7$?

7 Estimation of the spectrum

7.1 Properties of the Periodogram

In chapter 1 we have defined the periodogram by

$$\mathrm{I}_T(\lambda) = \frac{1}{T}\left|\sum_{t=1}^{T} x_t e^{-i\lambda t}\right|^2$$

We mainly considered the socalled Fourier frequencies

$$\lambda_j = \frac{2\pi j}{T} \in (0, \pi] \quad ; \quad j = -\lfloor (T-1)/2 \rfloor, \ldots, 0, \ldots, \lfloor (T/2) \rfloor$$

because the vectors

$$e_j = \frac{1}{\sqrt{T}}\begin{pmatrix} e^{i\lambda_j 1} \\ e^{i\lambda_j 2} \\ \vdots \\ e^{i\lambda_j T} \end{pmatrix} \quad ; \quad j = -\lfloor (T-1)/2 \rfloor, \ldots, 0, \ldots, \lfloor (T/2) \rfloor$$

form an orthonormal basis for $\mathbf{C}^T$. This may be seen by

$$< e_j, e_k >= \frac{1}{T}\sum_{t=1}^{T} e^{-i\lambda_j t} e^{i\lambda_k t} = \begin{cases} 1 & \text{for } j = k \quad (\lambda_j = \lambda_k) \\ \frac{e^{i(\lambda_k - \lambda_j)}}{T}\frac{1-e^{i(\lambda_k-\lambda_j)T}}{1-e^{i(\lambda_k-\lambda_j)}} & = 0 \text{ else} \end{cases}$$

since $e^{i(\lambda_k - \lambda_j)T} = e^{i(2\pi(k-j))} = 1$.

If we define $x(T) = (x_1, \ldots, x_T)'$ then $\mathrm{I}_T(\lambda_j) = | < e_j, x(T) > |^2$. Since $< e_j, e_0 >= 0$ for all $j \neq 0$ we also have for an arbitrary $a \in \mathbb{C}$ that $< e_j, x(T) - a\,1 >=< e_j, x(T) >$ where $1 = (1, \ldots, 1)' = \sqrt{T} e_0 \in \mathbb{C}^T$. This gives for a Fourier frequency $\lambda_j \neq 0$

$$\begin{aligned} \mathrm{I}_T(\lambda_j) &= | < e_j, x(T) > |^2 = | < e_j, x(T) - \bar{x}_T\,1 > |^2 \\ &= \tfrac{1}{T}|\textstyle\sum_{t=1}^{T} e^{-i\lambda_j t}(x_t - \bar{x}_T)|^2 \\ &= \tfrac{1}{T}(\textstyle\sum_{t=1}^{T} e^{i\lambda_j t}(x_t - \bar{x}_T))(\textstyle\sum_{s=1}^{T} e^{-i\lambda_j s}(x_s - \bar{x}_T)) \\ &= \tfrac{1}{T}\textstyle\sum_{s,t=1}^{T} e^{i\lambda_j(t-s)}(x_t - \bar{x}_T)(x_s - \bar{x}_T) \\ &= \textstyle\sum_{|s|<T} \hat{\gamma}(s) e^{-i\lambda_j s}. \end{aligned}$$

The periodogram $\mathrm{I}_T(\lambda_j)$ for $\lambda_j \neq 0$ is the Fourier transform of the sample autocovariance function. A comparison of this representation of $\mathrm{I}_T(\lambda_j)$ with the representation of the spectral density $\mathrm{f}(\lambda) = \frac{1}{2\pi}\sum_{s=-\infty}^{\infty}\gamma(s)e^{-i\lambda s}$ suggests that the periodogram is an estimate for $2\pi\,\mathrm{f}(\lambda)$. For the frequency zero we have

$$\mathrm{I}_T(0) = | < e_0, x(T) > |^2 = T\bar{x}_T^2.$$

In our analysis we restrict ourselves to the Fourier frequencies. This has the following reasons: Because of the orthogonality of the e_j's the analysis for the Fourier frequencies is simple. For increasing T the grid of the Fourier frequencies becomes finer and thus we may approximate every frequency λ arbitrary close by a Fourier frequency.

Note that the property for a frequency to be a Fourier frequency depends on T. So for T going to infinity a fixed frequency λ may be a Fourier frequency for some T and not for others. However if $\lambda = \lambda_j$ is a Fourier frequency for a given T_0 then clearly it will also be a Fourier frequency all $T = kT_0$, $k \in \mathrm{N}$. In the following we will use the symbol $T \to \infty$ in the sense that we are considering indices $T = kT_0$ with $k \to \infty$. Since for a fixed Fourier frequency the index j is going to infinity with $T \to \infty$ we in the following use λ to indicate a Fourier frequency.

The next theorem investigates the asymptotic mean of the peridogram:

Theorem 7.1 *Let (x_t) be a stationary process with $\mathrm{E}\,x_t = \mu$ and let $\sum_{s=-\infty}^{\infty} |\gamma(s| < \infty$ hold then*

(i) $\mathrm{E}(\mathrm{I}_T(0)) - T\mu^2 \longrightarrow 2\pi\,\mathrm{f}(0)$ *and*

(ii) $\mathrm{E}(\mathrm{I}_T(\lambda)) \longrightarrow 2\pi\,\mathrm{f}(\lambda)$ *for* $T \to \infty$.

Proof: We have

$$\mathrm{E}(\mathrm{I}_T(0)) - T\mu^2 = T\,\mathrm{E}(\bar{x}_T)^2 - T\mu^2 = T\,\mathrm{Var}\,\bar{x}_T + T\mu^2 - T\mu^2 \longrightarrow 2\pi\,\mathrm{f}(0)$$

due to our assumptions using theorem (6.4).

For the second part we use $< e_j, x(T) > = < e_j, x(T) - \mu\,1 >$ which gives

$$\mathrm{I}_T(\lambda_j) = \sum_{|s|<T} \left(\frac{1}{T} \sum_{t=1}^{N-|s|} (x_{t+|s|} - \mu)(x_t - \mu) \right) e^{-i\lambda_j s}$$

and thus

$$\mathrm{E}(\mathrm{I}_T(\lambda_j)) = \sum_{|s|<T} \left(\frac{1}{T} \sum_{t=1}^{N-|s|} \gamma(s) \right) e^{-i\lambda_j s} = \sum_{|s|<T} (1 - \frac{|s|}{T})\,\gamma(s) e^{-i\lambda_j s}. \qquad (7.1)$$

Due to the assumption $\sum_s |\gamma(s)| < \infty$ using a similar argument as in the proof of the second part of theorem (6.4) we have $\mathrm{E}(\mathrm{I}_T(\lambda)) \to \sum_{s=-\infty}^{\infty} \gamma(s) e^{-i\lambda s} = 2\pi\,\mathrm{f}(\lambda)$. □

This theorem shows that the periodogram is an asymptotic unbiased estimate for the spectral density (for $\lambda \neq 0$). But as the following simple example shows the variance of the periodogram does not converge to zero. Thus the periodogram is not a consistent estimate of the spectral density.

Example: Let (x_t) be Gaussian white noise, i.e. white noise where each finite vector $(x_{t-H}, \ldots, x_t)'$ has a normal distribution $N(0, \sigma^2 I)$.

We define the vectors

$$c_j = \frac{2}{\sqrt{T}} \begin{pmatrix} \cos \lambda_j 1 \\ \vdots \\ \cos \lambda_j T \end{pmatrix}, s_j = \frac{2}{\sqrt{T}} \begin{pmatrix} \sin \lambda_j 1 \\ \vdots \\ \sin \lambda_j T \end{pmatrix}, \text{ for } 1 \leq j < T/2$$

and

$$c_0 = \frac{1}{\sqrt{T}} \begin{pmatrix} \cos \lambda_0 1 \\ \vdots \\ \cos \lambda_0 T \end{pmatrix}; \quad \text{and for even } T \;\; c_{T/2} = \frac{1}{\sqrt{T}} \begin{pmatrix} \cos \lambda_{T/2} 1 \\ \vdots \\ \cos \lambda_{T/2} T \end{pmatrix}.$$

It is easy to prove that these T real vectors form an orthonormal basis for $\mathbf{R}^T$ (and $\mathbf{C}^T$). Thus the inner products $\alpha_j = < c_j, x(T) >$ and $\beta_j = < s_j, x(T) >$ are uncorrelated random variables which are $N(0, \sigma^2)$ distributed and thus also independent. We can easily express the periodogram by these random variables α_j and β_j:

$$\mathrm{I}_T(\lambda_j) = \begin{cases} \alpha_j^2 & \text{for } \lambda_j = 0, \pi \\ \frac{1}{2}(\alpha_j^2 + \beta_j^2) & \text{for } 0 < \lambda_j < \pi \end{cases}$$

Note that because of the symmetry of the spectral density we here consider only the frequencies $0 \leq \lambda \leq \pi$. From this result we get the distribution of the peridogram as follows

$$\begin{array}{ll} \frac{1}{\sigma^2} \mathrm{I}_T(\lambda_j) \sim \mathcal{X}_1^2 & \text{for } \lambda_j = 0, \pi \\ \frac{2}{\sigma^2} \mathrm{I}_T(\lambda_j) \sim \mathcal{X}_2^2 & \text{for } 0 < \lambda_j < \pi \end{array}$$

where $\mathrm{I}_T(\lambda_j)$ and $\mathrm{I}_T(\lambda_k)$ are independent for $j \neq k$. Note that the Chi square distribution with two degrees of freedom is an exponential distribution.

The expectation and the variance of the peridogram are given by

$$\mathrm{E}(\mathrm{I}_T(\lambda_j)) = \sigma^2 = 2\pi \, \mathrm{f}(\lambda_j)$$

and

$$\mathrm{Var}(\mathrm{I}_T(\lambda_j)) = \begin{cases} 2\sigma^4 = 2(2\pi \, \mathrm{f}(\lambda_j))^2 & \text{for } \lambda_j = 0, \pi \\ \sigma^4 = (2\pi \, \mathrm{f}(\lambda_j))^2 & \text{for } 0 < \lambda_j < \pi \end{cases}$$

In addition we have in this case

$$\mathrm{Cov}(\mathrm{I}_T(\lambda_j), \mathrm{I}_T(\lambda_k)) = 0 \text{ for } \lambda_j \neq \lambda_k$$

This shows that the standard deviation of the peridogram is (essentially) equal to the spectral density and does not depend on the sample size T and thus does not converge to zero for $T \to \infty$. Therefore the periodogram is not a consistent estimate of the spectral density. The essential reason for this is that in $\mathrm{I}_T(\lambda_j) = \sum_{|s|<T} \hat{\gamma}(s) e^{-i\lambda_j s}$ we always have summands $\hat{\gamma}(s)$, $|s| \approx T$ which are only poor estimates of $\gamma(s)$.

For a very general class of stationary processes these results hold asymptotically.

Theorem 7.2 *Let (x_t) be a scalar stationary process of the form*

$$x_t = \sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j} \quad ; \quad (\epsilon_t) \sim \mathrm{IID}(0,\sigma^2) \quad ; \quad \sum_{j=-\infty}^{\infty} |b_j| < \infty$$

(i) If $\mathrm{f}(\lambda) > 0$ for all $\lambda \in [-\pi, \pi]$ and if $0 < \lambda_1 < \lambda_2 \cdots < \lambda_m < \pi$ then $(\mathrm{I}_T(\lambda_1), \mathrm{I}_T(\lambda_2), \ldots, \mathrm{I}_T(\lambda_m))'$ converges in distribution to a vector of independent and exponentially distributed random variables whose i-th component has mean $2\pi\, \mathrm{f}(\lambda_i)$ (and variance $(2\pi\, \mathrm{f}(\lambda_i))^2$).

(ii) If $\sum_{j=-\infty}^{\infty} |b_j| \sqrt{|j|} < \infty$ and $\mathrm{E}\,\epsilon_t^4 = \eta\sigma^4 < \infty$ holds then we have for $\lambda_j, \lambda_k \geq 0$

$$\mathrm{Cov}(\mathrm{I}_T(\lambda_i), \mathrm{I}_T(\lambda_j)) = \begin{cases} 2(2\pi)^2 \mathrm{f}(\lambda_i)^2 + O(T^{-1/2}) & \text{for } \lambda_j = \lambda_k = 0, \pi \\ (2\pi)^2 \mathrm{f}(\lambda_i)^2 + O(T^{-1/2}) & \text{for } 0 < \lambda_j = \lambda_k < \pi \\ O(T^{-1}) & \text{for } \lambda_j \neq \lambda_k \end{cases} \qquad (7.2)$$

For a sequence (x_k) the statement $x_k = O(k^{-\alpha})$ means that $|x_k k^{\alpha}|$ is bounded. Thus e.g. for $\lambda_j \neq \lambda_k$ the covariance $\mathrm{Cov}(\mathrm{I}_T(\lambda_j), \mathrm{I}_T(\lambda_k))$ converges to zero with the same rate as T^{-1}.

For a proof of this theorem see e.g. [3].

Note that the condition $\sum_{j=-\infty}^{\infty} |b_j| \sqrt{|j|} < \infty$ is e.g. fulfilled for ARMA processes.

Using the result of this theorem we can write the periodogram in the form

$$\mathrm{I}_T(\lambda_j) = 2\pi\, \mathrm{f}(\lambda_j) + u_j$$

where the u_j's are asymptotically uncorrelated and have asymptotically mean zero. Thus we may interpret the estimation problem of the spectral density as a regression problem in the frequency domain, since we cannot observe the "true values" $\mathrm{f}(\lambda_j)$ but only the true values corrupted with some noise u_j which gives the periodogram.

This representation also suggests that we may diminish the variance of periodogram by averaging over frequencies and thus instead of using the peridogram as an estimate we use a smoothed periodogram as an estimate for $f(\lambda)$.

7.2 Smoothed Spectral Estimates

We define a *direct spectral estimate* by

$$\hat{f}(\lambda_j) = \frac{1}{2\pi} \sum_{u \leq m_T} w_{u,T}\, I_T(\underbrace{\lambda_j - \frac{2\pi u}{T}}_{\lambda_{j-u}}). \tag{7.3}$$

We will always assume that the *filter weights* $w_{u,T}$ are symmetric and nonnegative, i.e. $w_{u,T} = w_{-u,T}$ and $w_{u,T} \geq 0$. By this assumption $\hat{f}(\lambda_j) \geq 0$ holds. In the formula (7.3) we used the periodic extension of the periodogram from the interval $(-\pi, \pi]$ to $\mathbf{R}$. Thus for example $I_T(\frac{2\pi(T+3)}{T}) = I_T(\frac{2\pi 3}{T})$.

Often we use the notation m and w_u rather than m_T and $w_{u,T}$.

A simple example for such a direct estimate is the Daniell estimate where the weights w_u are defined by

$$w_u = \begin{cases} \frac{1}{2m+1} & \text{for } |u| \leq m \\ 0 & \text{else.} \end{cases}$$

Thus the Daniel estimate is a simple moving average of the periodogram.

From the equations (7.2) it is easy to obtain the following results for the asymptotic behavior of the direct estimate $\hat{f}(\lambda)$. We use the symbol $a_T \cong a$ if $\lim_{T \to \infty} a_T = a$.

$$E\,\hat{f}(\lambda) = \sum_{|u| \leq m_T} w_u \underbrace{(2\pi)^{-1} E\, I_T(\lambda_{j-u})}_{\cong f(\lambda_{j-u}) \cong f(\lambda)} \cong f(\lambda) \sum_{|u| \leq m_T} w_u$$

Here we have used the assumption that the spectral density $f(\lambda)$ is continuous and that m_T/T is converging to zero for $T \to \infty$. For the variance we have under the same assumptions

$$\begin{aligned} \mathrm{Var}(\hat{f}(\lambda)) &= \textstyle\sum_{|u| \leq m_T} w_u^2 \underbrace{\mathrm{Var}((2\pi)^{-1} I_T(\lambda_{j-u}))}_{\cong f^2(\lambda_{j-u}) \cong f^2(\lambda)} + \\ &\quad \textstyle\sum_{|u|,|v| \leq m_T, u \neq v} w_u w_v \underbrace{\mathrm{Cov}((2\pi)^{-1} I_T(\lambda_{j-u}), (2\pi)^{-1} I_T(\lambda_{j-v}))}_{\cong 0} \\ &\cong f^2(\lambda) \textstyle\sum_{|u| \leq m_T} w_u^2 \end{aligned}$$

From these considerations we can conclude that $\hat{f}(\lambda)$ is a consistent estimate if

(i) $\sum_{|u| \leq m_T} w_u = 1$,

(ii) $\frac{m_T}{T} \longrightarrow 0$ and

(iii) $\sum_{|u| \leq m_T} w_u^2 \longrightarrow 0$ holds.

E.g. for the Daniell estimate this conditions are equivalent to $m_T \to \infty$ and $\frac{m_T}{T} \to 0$ for $T \to \infty$.

Let us state this result in a formal theorem

Theorem 7.3 *Let (x_t) be a stationary process of the form $x_t = \sum_{j=-\infty}^{\infty} b_j \epsilon_{t-j}$ where $(\epsilon_t) \sim \mathrm{IID}(0,\sigma^2)$, $\sum_{j=-\infty}^{\infty} |b_j|\sqrt{|j|} < \infty$ and $\mathrm{E}\,\epsilon_t^4 < \infty$ holds. If $\hat{\mathrm{f}}_T(\lambda_j)$ is direct estimator with $m_T/T \to 0$, $w_{u,T} \geq 0$, $w_{u,T} = w_{-u,T}$, $\sum_{|u|\leq m_T} w_{u,T} = 1$ and $\sum_{|u|\leq m_T} w_{u,T}^2 \to 0$ then*

(i) $\lim_{T\to\infty} \mathrm{E}\,\hat{\mathrm{f}}_T(\lambda) = \mathrm{f}(\lambda)$

(ii)

$$\lim_{T\to\infty} \Big(\sum_{|u|\leq m_T} w_{u,T}^2 \Big)^{-1} \mathrm{Cov}(\hat{\mathrm{f}}_T(\lambda), \hat{\mathrm{f}}_T(\omega)) = \begin{cases} 2\,\mathrm{f}(\lambda)^2 & \textit{for } \lambda = \omega = 0, \pi \\ \mathrm{f}(\lambda)^2 & \textit{for } 0 < \lambda = \omega < \pi \\ 0 & \textit{else.} \end{cases}$$

For a proof of this theorem see e.g. Brockwell and Davis [3].

Spectral windows and the finite sample bias:

In addition to the asymptotic properties finite sample considerations are important for the design of the filter weights w_u. First let us look at the expected value of the periodogram (for Fourier frequencies unequal to zero).

Theorem 7.4

$$\mathrm{E}\Big(\frac{1}{2\pi}\,\mathrm{I}_T(\lambda_j)\Big) = \int_{-\pi}^{\pi} \mathrm{f}(\omega)\, g_T^F(\lambda_j - \omega) d\omega \qquad \textit{for } \lambda_j \neq 0$$

where

$$g_T^F(\omega) = \begin{cases} \frac{T}{2\pi} & \textit{for } \omega = 0 \\ \frac{1}{2\pi}\frac{\sin^2(T\omega/2)}{\sin^2(\omega/2)} & \textit{else} \end{cases}$$

is the socalled Fejer kernel.

Proof: From equation (7.1) and the spectral representation of the autocovariance function we get

$$\begin{aligned} \mathrm{E}\big(\tfrac{1}{2\pi}\,\mathrm{I}_T(\lambda_j)\big) &= \tfrac{1}{2\pi} \textstyle\sum_{|s|<T}\big(1 - \frac{|s|}{T}\big)\, \gamma(s) e^{-i\lambda_j s} \\ &= \tfrac{1}{2\pi} \textstyle\sum_{|s|<T}\big(1 - \frac{|s|}{T}\big)\big(\int_{-\pi}^{\pi} \mathrm{f}(\omega) e^{i\omega s} d\omega\big) e^{-i\lambda_j s} \\ &= \textstyle\int_{-\pi}^{\pi} \mathrm{f}(\omega) \underbrace{\left(\frac{1}{2\pi} \sum_{|s|<T} \big(1 - \frac{|s|}{T}\big) e^{-i(\lambda_j - \omega)s} \right)}_{g_T^F(\lambda_j - \omega)} d\omega \end{aligned}$$

For the Fejer kernel $g_T^F(\omega)$ we further get

$$g_T^F(\omega) = \frac{1}{2\pi}\sum_{|s|<T}(1 - \frac{|s|}{T})e^{-i\omega s} = \frac{1}{2\pi T}\left|\sum_{s=0}^{T-1} e^{i\omega s}\right|^2 = \begin{cases} \frac{T}{2\pi} & \text{for } \omega = 0 \\ \frac{1}{2\pi}\left|\frac{1-e^{i\omega T}}{1-e^{i\omega}}\right|^2 & \text{else} \end{cases}$$

which gives us the desired result since $1 - e^{i\alpha} = -2e^{i\alpha/2}\sin\alpha/2$. □

Let us list some of the most important properties of the Fejer kernel (see also figure (7.1)):

(i) g_T^F is symmetric and non negative, i.e. $g_T^F(\omega) = g_T^F(-\omega)$ and $g_T^F(\omega) \geq 0$.

(ii) $g_T^F(\omega)$ is zero for the Fourier frequencies $\omega = \lambda_j$, $j \neq 0$.

(iii) $g_T^F(\omega)$ converges to zero for $T \rightarrow \infty$ and $\omega \neq 0$. (This together with (iv) again shows that the periodogram is asymptotically unbiased.)

(iv) $\int_{-\pi}^{\pi} g_T^F(\omega)d\omega = 1$.

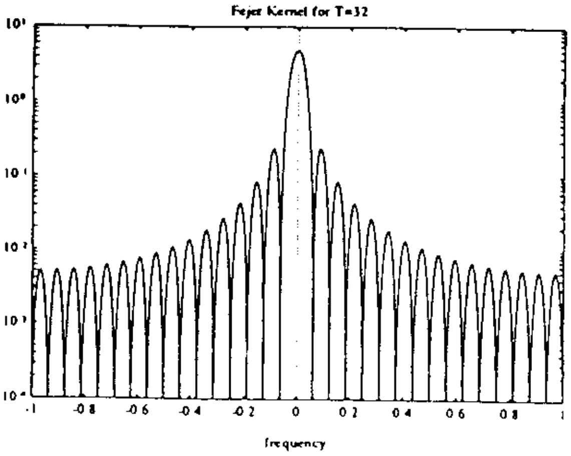


Figure 7.1: Fejer kernel $T = 32$.

This result can easily be generalized to direct spectral estimators

Theorem 7.5 *For the direct spectral estimator* $\hat{f}(\lambda)$ *we have*

$$E(\hat{f}(\lambda_j)) = \int_{-\pi}^{\pi} f(\omega)\, g_T(\lambda_j - \omega)d\omega \qquad (7.4)$$

where

$$g_T(\omega) = \sum_{|u| \leq m_T} w_u \, g_T^F(\omega - \frac{2\pi u}{T})$$

Proof: The proof of this is immediate.

These functions g_T are called *spectral windows.* Of course also the Fejer kernel is a spectral window since the periodogram is one special type of direct spectral estimators (it is the Daniell estimator for $m_T = 0$).

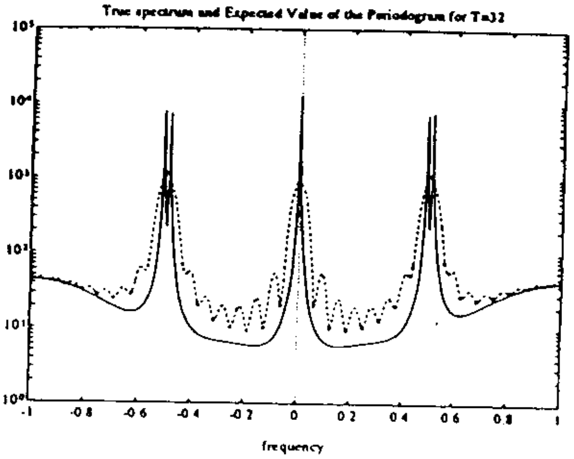


Figure 7.2: True spectral density (solid) and the expectation of the periodogram (dashed) for $T = 32$. The Fourier frequencies unequal to zero are marked by stars.

From the formula (7.4) we can see that there are two sources for the finite sample bias of a direct spectral estimator. See also figure (7.2).

(i) *Resolution problem:* First let us introduce a measure for the width of the main lobe of the spectral window. The *bandwidth* for the Daniell window is defined by $B = 2\pi(2m+1)/T$. For general windows the bandwidth is defined by $B = 2\pi(T\sum_u w_u^2)^{-1}$. Now it is easy to see from (7.4) that, if the spectrum has two peaks with a distance less than the bandwidth, the estimate cannot distinguish these two peaks. In other words the estimate cannot resolve these two peaks. In addition a single sharp peak will be made smaller and broader due to the nonzero bandwith. The bias caused by the nonzero bandwidth is sometimes called the *narrow band bias.*

(ii) *Leakage:* If side maxima of the spectral window $g(\lambda_j - \omega)$ concide with a peak of the spectral density at ω this may have a big effect on the bias at the frequency λ_j. This sometimes is called the ***broad band bias.***

It is easy to see that problems of (non)resolution and leakage are less severe in case of flat and smooth spectral densities. This is the reason why one often applies a linear transformation to the data in order to make the underlying process similar to white noise. Such a procedure is called *prewithening.*

In order to reduce the bias, the filter weights w_u should be designed such that the side maxima of the spectral window are small and the bandwidth is small. The bandwidth can be decreased by decreasing the filter length m_T for given T. However this is done at the cost of increasing the variance of the estimator. So high resolution and low variance are two conflicting aims in spectral estimation. This is reflected in the socalled *uncertainty relation* in spectral estimation:

$$B(\hat{f}(\lambda))\,\mathrm{Var}(\hat{f}(\lambda)) \approx \frac{1}{T\sum_u w_u^2} \sum_u w_u^2 f(\lambda)^2 = \frac{1}{T} f^2(\lambda)$$

The problem is now how to choose in actual computations the filter length m_T. This is often done by a procedure called *window closing* where one compares by visual inspection the direct estimators for different m_T's and one chooses that one that is on the one hand not to smooth (the bias is not too large) and on the other hand not to erratic (the variance is not too large).

The main use of spectral estimates in the case of economic time series are:

(i) "second look" on data in order to learn the main features of the time series before parametric models are used.

(ii) Analysis of underlying cycles such as business cycles and seasonals.

(iii) Controlling the action of (nonlinear) transformations of the data.

Final remark:

In chapters 6 and 7 we have treated the problem of the estimation of the second moments i.e. the estimation of the autocovariance function and the estimation of the spectral density. Note that e.g. under the assumption $\sum_s |\gamma(s)| < \infty$ the autocovariance and the spectral density are in a one-to-one relation. In the next chapter we will consider estimation for parametric models.

These approaches to estimation of second moments may be classified as in the following table:

Time Domain	Frequency Domain	
$\hat{\gamma}(s) = \frac{1}{T}\sum x_{t+s}x_t$ $\hat{\rho}(s) = \frac{\hat{\gamma}(s)}{\hat{\gamma}(0)}$	$\mathrm{I}(\lambda_j) = \frac{1}{T}\lvert\sum x_t e^{-i\lambda_j t}\rvert^2$ $\hat{\mathrm{f}}(\lambda_j) = \frac{1}{2\pi}\sum w_u\,\mathrm{I}(\lambda_{j-u})$	**non parametric approach** (We do not assume that the covariance function (spectral density) may be described by a finite number of parameters.)
$\hat{\gamma}(s) = \int_{-\pi}^{\pi}\hat{\mathrm{f}}(\lambda)e^{i\lambda s}d\lambda$	$\hat{\mathrm{f}}(\lambda) = \frac{\hat{\sigma}^2}{2\pi}\frac{1}{\lvert\hat{a}(e^{-i\lambda})\rvert^2}$	**parametric approach** (E.g. AR(p) model. The coefficients a_j may be consistently estimated by OLS. These estimates $\hat{a}_j$ then give estimates for $\gamma(s)$ and $\mathrm{f}(\lambda)$.)

In the parametric approach more a priori information is needed compared to the non parametric approach. E.g. we have to assume that the underlying process is an AR(p) process. But as we know every regular process can be approximated with arbitrary accuracy by an AR process and we can estimate the order p from the data. In the latter case we could speak of a semi parametric approach.

The non parametric approach needs less information and can handle a wider class of processes. But on the other hand we loose efficiency. Thus parametric approaches are of particular use for economic time series where typically the sample size is small.

7.3 Exercises

(7.1) Computer example: (Spectral estimation – continuation of Example 9.9) In example 9.9 you had to compute the periodogram of a differenced time series. In this example you start from this periodogram to compute an estimate of the spectral density (of the differenced process). Compute the Daniell estimate of the spectral density for different window widths and try to find an "optimal" window width.

There are several possible variations for this task:

- Take the log of the data and then difference it. These transformation of the data give a measure for the growth rates.
- Use seasonal differences instead of first differences.

- Plot the log of the spectral estimates. This has the advantage that the very high peaks in the spectrum caused e.g. by seasonals do not hide the rest of the spectrum.

8 Identification of Linear Systems

In this chapter we are concerned with the problem of determining an AR, ARX, ARMA, ARMAX system from data. This problem may be decomposed into four main steps.

(i) Preliminary data transformations to make the underlying processes stationary, e.g. detrending, nonlinear transformations e.g. taking the logarithm of the data, ...

(ii) Specification of the model class, e.g. for an ARMA model we have to determine p and q from the data. (Estimation of integer parameters.)

(iii) Estimation of the real valued parameters of the model, e.g. for an ARMA model we have to estimate the a_j's, b_j's and σ^2.

(iv) Evaluation of the estimated model.

8.1 ARIMA processes

Definition 8.1 *A process (x_t) is called an* ARIMA(p,d,q) *process if $(1-z)^d x_t$ is an ARMA(p,q) process, i.e.*

$$a(z)(1-z)^d x_t = b(z)\epsilon_t \text{ for } t \geq d \text{ and } a(z) \neq 0 \text{ for all } |z| \leq 1$$

and $(1-z)^{d-1}x_t$ is not stationary. Such processes are also called integrated processes *of order d.*

Consider for example an integrated process (x_t) of order $d = 1$:

$$a(z)(1-z)x_t = b(z)\epsilon_t$$

This gives

$$(1-z)x_t = a^{-1}(z)b(z)\epsilon_t = \sum_{j=0}^{\infty} k_j\epsilon_{t-j} = y_t$$

If we for example assume that $x_0 = c = \text{const}$ then we have

$$x_t = \sum_{j=1}^{t} y_j + (t-1)c$$

which gives a linear trend in mean and a trend in variance.

ARIMA models in some sense are very special models, since we allow only roots of $a(z)$ which lie outside of the unit circle and roots which are equal to 1. But nevertheless these models can explain many features which are very common for economic data.

The sample autocorrelation function of such an ARIMA process (with $d > 0$) shows typically a very slow decay. Thus as a rule of thumb one can determine the *integration order d* by iteratively differencing the data until the sample autocorrelation function shows a "nice" geometrical decay which is typical for ARMA pocesses.

8.2 Identification of Scalar AR Systems

From now on we assume stationarity, i.e. we assume that the data have been transformed in a way that makes the underlying process stationary.

There are two problems related to the identification of scalar AR systems. We have to determine the order p and for given p we have to estimate the real valued parameters $a_1, \dots, a_p, \sigma^2$ of the model.

Let us first consider the estimation of the real valued parameters given the model specification, i.e. given the order p. We suppose that (x_t) is an AR(p) process defined by

$$x_t = a_1 x_{t-1} + \cdots + a_p x_{t-p} + \epsilon_t \tag{8.1}$$

where (ϵ_t) is white noise with $\mathrm{E}\,\epsilon_t^2 = \sigma^2$ and the stability condition

$$a(z) = 1 - a_1 z - \ldots - a_p z^p \neq 0 \quad \forall |z| \leq 1$$

holds. Thus x_t has a representation as a causal infinite MA process

$$x_t = \sum_{j=0}^{\infty} k_j \epsilon_{t-j} = k(z)\epsilon_t \text{ where } k(z) = a^{-1}(z).$$

If we multiply both sides of equation (8.1) with x_s and take the expectation on both sides we get the following relations:

$$\gamma(t-s) = a_1\,\gamma(t-1-s) + \cdots + a_p\,\gamma(t-p-s) + \mathrm{E}\,x_s\epsilon_t$$

Since (ϵ_t) is white noise we have

$$\mathrm{E}\,x_s\epsilon_t = \sum_{j=0}^{\infty} k_j\,\mathrm{E}\,\epsilon_{s-j}\epsilon_t = \begin{cases} k_0\,\mathrm{E}\,\epsilon_t^2 = \sigma^2 & \text{for } s = t \\ 0 & \text{for } s < t \end{cases}$$

Note that $a(0) = a_0 = 1$ implies that $k(0) = k_0 = 1$. These relations for $s = t-1, \dots, t-p$ give the following linear equations for $a_1, a_2, \dots, a_p$

$$\underbrace{\begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma(1) \\ \gamma(p-1) & \cdots & \gamma(1) & \gamma(0) \end{pmatrix}}_{\Gamma_p} \underbrace{\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}}_{a} = \underbrace{\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}}_{\gamma_p} \tag{8.2}$$

and for $s = t$ we obtain for σ^2

$$\sigma^2 = \gamma(0) - a' \gamma_p \tag{8.3}$$

Here we used the symmetry of the autocovariance function $\gamma(\cdot)$. The equations (8.2) and (8.3) are called the Yule–Walker equations.

Note that the equations (8.2) are the same as the normal equations for the linear prediction problem from a finite past $(x_{t-1}, \ldots, x_{t-p})$ and the normal equations for the theoretical regression problem.

The Yule–Walker equations enable us to determine the parameters a_j and σ^2 given the population second moments. If we assume $\Gamma_p > 0$ then we have

$$a = \Gamma_p^{-1} \gamma_p \quad ; \quad \sigma^2 = \gamma(0) - \gamma_p' \Gamma_p^{-1} \gamma_p$$

If we substitute the sample moments for the population moments we get *moment estimators* for the parameters. Let us define

$$\hat{\Gamma}_p = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(p-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{\gamma}(1) \\ \hat{\gamma}(p-1) & \cdots & \hat{\gamma}(1) & \hat{\gamma}(0) \end{pmatrix}, \quad \hat{\gamma}_p = \begin{pmatrix} \hat{\gamma}(1) \\ \hat{\gamma}(2) \\ \vdots \\ \hat{\gamma}(p) \end{pmatrix} \quad \text{and } \hat{a} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix}$$

then we have (for $\hat{\Gamma}_p > 0$)

$$\hat{a} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p \quad ; \quad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\gamma}_p' \hat{\Gamma}_p^{-1} \hat{\gamma}_p \tag{8.4}$$

$\hat{a}$, $\hat{\sigma}^2$ defined by the equations (8.4) are called the Yule–Walker estimates. It easy to express the Yule–Walker estimates by the sample autocorrelation function as

$$\hat{a} = \hat{\Gamma}_p^{-1} \gamma_p = \hat{R}_p^{-1} \hat{\rho}_p \qquad \hat{\sigma}^2 = \hat{\gamma}(0)(1 - \hat{\rho}_p' \hat{R}_p^{-1} \hat{\rho}_p)$$

where $\hat{R}_p = \hat{\gamma}(0)^{-1} \hat{\Gamma}_p$ and $\hat{\rho}_p = \hat{\gamma}(0)^{-1} \hat{\gamma}_p = (\hat{\rho}(1), \ldots, \hat{\rho}(p))'$.

Lemma 8.1 *If (x_t) is a (scalar) stable $AR(p)$ process with $\gamma(0) > 0$ then Γ_r is non singular and for $\hat{\gamma}(0) > 0$ also $\hat{\Gamma}_r$ is nonsingular for every r.*

We will first prove a more general result:

Lemma 8.2 *For a scalar stationary process (x_t) with $\gamma(0) > 0$ and $\lim_{s \to \infty} \gamma(s) = 0$ the covariance matrix Γ_p is positive definite.*

Proof: Suppose that Γ_p is singular for some p. Then there exist an integer $r > 0$ such that $\Gamma_r > 0$ holds and Γ_{r+1} is singular. Thus by lemma (5.1) there exists a vector $a \in \mathbf{R}^{1\times r}$ such that $x_{r+1} = a\mathrm{x}(r)$ a.e. where $\mathrm{x}(r) = (x_1, \ldots, x_r)'$. Consequently by stationarity of (x_t) for all $t > r$ there exist real vectors $a^{(t)}$ such that $x_t = a^{(t)}\mathrm{x}(r)$ a.e.

Thus

$$\gamma(0) = \mathrm{E}\, x_t x_t = a^{(t)}\, \mathrm{E}\, \mathrm{x}(r)\mathrm{x}(r)' a^{(t)\prime} = a^{(t)}\, \Gamma_r\, a^{(t)\prime} \geq \|a^{(t)}\|^2 \lambda_1$$

where λ_1 is the smallest eigenvalue of Γ_r. (Note that $\Gamma_r > 0$ implies that $\lambda_1 > 0$ holds.) Thus the vectors $a^{(t)}$ are bounded.

We also have

$$\gamma(0) = \mathrm{E}\, x_t x_t = \mathrm{E}(x_t \sum_{j=1}^{r} a_j^{(t)} x_j) = \sum_{j=1}^{r} a_j^{(t)}\, \gamma(t-j) \leq \sum_{j=1}^{r} |a_j^{(t)}|\, |\gamma(t-j)|$$

But this inequality is a contradiction to the boundedness of the $a^{(t)}$'s and the assumption $\gamma(t) \to 0$. □

Proof of lemma (8.1): For a stable AR(p) process (with $\mathrm{E}\epsilon_t^2 = \sigma^2 > 0$) the autocovariances converge to zero with a geometrical rate which gives $\Gamma_r > 0$ by the lemma above.

Since $\hat{\gamma}(s) = 0$ for all $|s| \geq T$ we also have $\hat{\gamma}(s) \to 0$ for $s \to \infty$. Since the sample autocovarance function is nonnegative definite, there exist a stationary process (y_t) such that the autocovariance function of y_t is equal to the sample autocovariance function $\hat{\gamma}(s)$. This implies $\hat{\Gamma}_r > 0$ by lemma (8.2). □

Lemma 8.3 *If* $\mathrm{plim}_{T\to\infty}\, \hat{\gamma}(s) = \gamma(s)$ *holds for* $s = 0, \ldots, p$ *then* $\mathrm{plim}_{T\to\infty}\, \hat{a} = a$ *holds.*

Proof: This result holds by the Slutzky lemma (6.3) since the inversion of a nonsingular matrix is a continuous function and thus $\hat{a} = \hat{\Gamma}_p^{-1}\, \hat{\gamma}_p \xrightarrow{p} \Gamma_p^{-1}\, \gamma_p = a$.

Theorem 8.4 *If* (x_t) *is a (stable) AR(p) process with* $(\epsilon_t) \sim \mathrm{IID}(0, \sigma^2)$ *then*

$$\sqrt{T}(\hat{a} - a) \xrightarrow{d} \mathrm{N}(0, \sigma^2\, \Gamma_p^{-1}).$$

For a proof of this theorem see e.g. Brockwell and Davis [3].

In addition the Yule–Walker estimators $\hat{a}$ can be shown to be be asymptotically efficient, i.e. $\sigma^2\, \Gamma_p^{-1}$ is the smallest covariance matrix of the limiting distribution of $\sqrt{T}(\tilde{a} - a)$, for any estimator $\tilde{a}$.

Lemma 8.5 *If* $\hat{\Gamma}_p$ *is non singular then the Yule-Walker estimate* $\hat{a}$ *gives a stable AR-polynomial, i.e. the transferfunction* $\hat{a}(z) = 1 - \hat{a}_1 z - \cdots - \hat{a}_p z^p$ *satisfies the stability condition* $\hat{a}(z) \neq 0$ *for all* $|z| \leq 1$.

Proof: Let us define $\tilde{x}_t = x_t - \bar{x}_T$ and

$$y = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_T \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} \text{ and } X = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \tilde{x}_1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{x}_T & & \ddots & 0 \\ 0 & \ddots & & \tilde{x}_1 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \tilde{x}_T \end{pmatrix}.$$

It is easy to see that $\hat{a} = \hat{\Gamma}_p^{-1}\hat{\gamma}_p = (X'X)^{-1}(X'y)$ holds. Therefore $\hat{a}$ is the OLS estimate for these "data matrices" y and X and minimizes the residual sum of squares $u'u$ for $u = y - Xa$. If $(1/c)$ is a root of $\hat{a}(z) = 1 - \hat{a}_1 z - \cdots - \hat{a}_p z^p$ then we have $\hat{a}(z) = b(z)(1 - cz) = (1 - b_1 z - \cdots - b_{p-1} z^{p-1})(1 - cz)$. If we define $b_{(1)} = (1, -b_1, \ldots, -b_{p-1}, 0)' \in \mathbf{R}^{p+1}$ and $b_{(2)} = (0, 1, -b_1, \ldots, -b_{p-1})' \in \mathbf{R}^{p+1}$ we may write $\hat{u} = y - X\hat{a} = (y, X)b_{(1)} - c(y, X)b_{(2)} = \hat{u}_{(1)} - c\hat{u}_{(2)}$ and

$$\hat{u}'\hat{u} = \hat{u}'_{(1)}\hat{u}_{(1)} - 2c\hat{u}'_{(1)}\hat{u}_{(2)} + c^2\hat{u}'_{(2)}\hat{u}_{(2)}.$$

Since $\hat{u}'\hat{u}$ is minimal the derivative of this expression with respect to c must be equal to zero which gives

$$c\hat{u}'_{(2)}\hat{u}_{(2)} = \hat{u}'_{(1)}\hat{u}_{(2)}$$

By the assumption $\hat{\Gamma}_p = \frac{1}{T}(X'X) > 0$ we have $\hat{u}'_{(2)}\hat{u}_{(2)} > 0$. It is immediate to see that $\hat{u}_{(1)}$ is the vector $\hat{u}_{(2)}$ shifted by one entry. Thus we have by the Cauchy-Schwarz inequality

$$|c| = \frac{|\hat{u}'_{(1)}\hat{u}_{(2)}|}{\sqrt{(\hat{u}'_{(1)}\hat{u}_{(1)})(\hat{u}'_{(2)}\hat{u}_{(2)})}} \leq 1$$

We also know that the case $|c| = 1$ can only occur in the case when $\hat{u}_{(1)}$ and $\hat{u}_{(2)}$ are linearly dependent, which by the special structure of these two vectors would imply that they are equal to the zero vector which is a contradiction to $\hat{u}'_{(1)}\hat{u}_{(1)} > 0$. □

Thus the Yule-Walker estimates are proper estimates for the AR-parameters in the sense that they always give a stable AR-polynomial.

The OLS estimate $\hat{b}$ for the AR(p) parameters a is given by

$$\hat{b} = (\frac{1}{T}X'X)^{-1}(\frac{1}{T}X'y)$$

where

$$X = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ x_1 & 0 & & 0 \\ x_2 & x_1 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \vdots & & & x_1 \\ \vdots & & & \vdots \\ x_{T-1} & x_{T-2} & \cdots & x_{T-p} \end{pmatrix} \qquad y = \begin{pmatrix} x_1 \\ x_1 \\ \vdots \\ \\ \vdots \\ x_T \end{pmatrix}$$

Note that here the matrix X corresponds to a special choice of starting values x_s, $s \leq 0$, namely where these starting values simply are set equal to zero. It is easy to see that the matrices $\frac{1}{T}X'X$ and $\frac{1}{T}X'y$ converge to the matrices $\hat{\Gamma}_p$ and $\hat{\gamma}_p$ and thus the OLS estimates $\hat{b}$ are consistent. Since plim $\sqrt{T}(\frac{1}{T}X'X - \hat{\Gamma}_p) = 0$ and plim $\sqrt{T}(\frac{1}{T}X'y - \hat{\gamma}_p) = 0$ it can be shown that the OLS estimates $\hat{b}$ is is asymptotically equivalent to the Yule–Walker estimates. (In the sense that they have the same asymptotical distribution.)

Problems of missspecification

We assume that the data generating process is an AR process of "true order" p_0 and thus we have

$$a_0(z)x_t = \epsilon_t$$

where $a_0(z)$ is a stable polynomial of order p_0 and (ϵ_t) is white noise.

If we specify a order p there are two possible cases for missspecification

- Underfitting $p < p_0$ or
- Overfitting $p > p_0$.

Let us first discuss the problem of underfitting. If (x_t) is a general stationary process (not necessarily an AR(p) process) then the Yule–Walker equations define a parameter vector $\tilde{a}$ by

$$\Gamma_p \tilde{a} = \gamma_p \quad \tilde{a} = \Gamma_p^{-1}\gamma_p$$

if we in addition assume that $\Gamma_p > 0$ holds. This parameter vector also defines the best linear least squares predictor for x_t given the past $x_{t-1}, \ldots, x_{t-p}$.

If the sample estimates for the autocovariances are consistent then the Yule–Walker estimate will converge to $\tilde{a}$. Thus in the case of underfitting the Yule–Walker estimate will converge to the AR-model which gives the best prediction within the class of AR(h), $h \leq p$ systems. Thus even in the case of underfitting the Yule–Walker estimates have some optimality properties.

Consider e.g. a regular process (x_t). By the Wold decomposition (5.6) we know that x_t has a representation as a causal infinite MA process of the form

$$x_t = \sum_{j=0}^{\infty} b_j \epsilon_{t-j} = b(z)\epsilon_t$$

where ϵ_t are the innovations of the process x_t. If $\tilde{a}(z)$ denotes the optimal AR(p) model in the sense defined above then the residuals

$$\tilde{\epsilon}_t = \underbrace{\tilde{a}(z)b(z)}_{l(z)}\epsilon_t = l(z)\epsilon_t$$

in general (i.e. for the case of underfitting) will not be white noise and the variance

$$\mathrm{E}\tilde{\epsilon}_t^2 = \sigma^2 \sum_{j=0}^{\infty} l_j^2 > \sigma^2$$

will be greater than the variance of the innovations ϵ_t. (Note that $\tilde{a}(0) = 1$ and $b(0) = 1$ imply that $l(0) = l_0 = 1$.) Thus due to underfitting the residuals $\tilde{\epsilon}_t$ will not be uncorrelated and have a higher variance than the predicion errors from the infinite past. These facts can be used for a test of not having underfitted, e.g. by testing $\tilde{\epsilon}_t$ for white noise.

In the case of overfitting the specified order p is larger than the true order p_0. But as the following theorem shows overfitting is not that troublesome as underfitting.

Theorem 8.6 *Let (x_t) be a stable AR(p_0) process with $(\epsilon_t) \sim \mathrm{IID}(0, \sigma^2)$ and let $p > p_0$,*

$$\hat{a}^{(p)} = \begin{pmatrix} \hat{a}_{p,1} \\ \vdots \\ \hat{a}_{p,p_0} \\ \hat{a}_{p,p_0+1} \\ \vdots \\ \hat{a}_{p,p} \end{pmatrix} = \hat{\Gamma}_p^{-1}\hat{\gamma}_p \quad \textit{and} \quad a^{(p)} = \begin{pmatrix} a_1 \\ \vdots \\ a_{p_0} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \Gamma_p^{-1}\gamma_p$$

then $\sqrt{T}(\hat{a}^{(p)} - a^{(p)}) \xrightarrow{d} \mathrm{N}(0, \sigma^2\Gamma_p^{-1})$.

For a proof of this theorem see e.g. Brockwell and Davis [3].

Note that the last entries of the true parameter vector $a^{(p)}$ are zero since we have assumed that the data generating process is an AR(p_0) process and that $p_0 < p$ holds.

This theorem can be used to construct a test wether $a_{p_0+1} = \cdots = a_p = 0$ holds.

Order Estimation

There are three commonly used procedures to determine the true order p_0:

(i) Looking at the partial autocorrelation function

(ii) Sequence of tests for AR coefficients

(iii) Information criteria

(ii) Partial autocorrelation function

Definition 8.2 *Let (x_t) be a scalar stationary process then the* partial autocorrelation function (pacf) $\alpha : \mathbf{Z} \to \mathbf{R}$ *is defined by*

$$\begin{array}{rcll}
\alpha(0) & = & 1 = \rho(0) & \\
\alpha(1) & = & \mathrm{Corr}(x_0, x_1) = \rho(1) & \\
\alpha(s) & = & \mathrm{Corr}(x_0 - P_{\mathbf{H}(1,x_1,\ldots,x_{s-1})}\, x_0, x_s - P_{\mathbf{H}(1,x_1,\ldots,x_{s-1})}\, x_s) & \textit{for } s > 1 \\
\alpha(s) & = & \alpha(-s) & \textit{for } s < 0
\end{array}$$

where $P_{\mathbf{H}(1,x_1,\ldots,x_{s-1})}$ denotes the projection on the Hilbert space spanned by $1, x_1, \ldots, x_{s-1}$.

Lemma 8.7 *Let (x_t) be a (scalar) mean zero stationary process with autocovariance function $\gamma(s)$ such that $\gamma(s) \to 0$ as $s \to \infty$. If $P_{\mathbf{H}(x_0,\ldots,x_s)}\, x_s = \sum_{j=1}^{n} b_j x_{s-j}$, then $\alpha(s) = b_s$.*

Proof: Let $\hat{x}_0$, $\hat{x}_s$ denote the projections of x_0 and x_s onto the Hilbertspace spanned by $x_1, \ldots, x_{s-1}$. If $b = (b_1, \ldots, b_s)'$ and $\mu_{s-1} = (\gamma(s-1), \ldots, \gamma(1))'$ then we have from equation (5.3)

$$\begin{pmatrix} \Gamma_{s-1} & \mu_{s-1} \\ \mu'_{s-1} & \gamma(0) \end{pmatrix} a = \begin{pmatrix} \gamma_{s-1} \\ \gamma(s) \end{pmatrix}$$

Note that by our assumptions $\Gamma_{s-1} > 0$ holds. By Cramer's rule we get

$$b_s = \frac{\det \begin{pmatrix} \Gamma_{s-1} & \gamma_{s-1} \\ \mu'_{s-1} & \gamma(s) \end{pmatrix}}{\det \begin{pmatrix} \Gamma_{s-1} & \mu_{s-1} \\ \mu'_{s-1} & \gamma(0) \end{pmatrix}} = \frac{(\det \Gamma_{s-1})(\gamma(s) - \mu'_{s-1}\, \Gamma_{s-1}^{-1}\, \gamma_{s-1})}{(\det \Gamma_{s-1})(\gamma(0) - \mu'_{s-1}\, \Gamma_{s-1}^{-1}\, \mu_{s-1})} = \frac{\gamma(s) - \mu'_{s-1}\, \Gamma_{s-1}^{-1}\, \gamma_{s-1}}{\gamma(0) - \mu'_{s-1}\, \Gamma_{s-1}^{-1}\, \mu_{s-1}}$$

By equation (5.3) we also get

$$\hat{x}_{s-1} = (x_1, \ldots, x_{s-1})\, \Gamma_{s-1}^{-1}\, \gamma_{s-1}$$

and by analogous computations

$$\hat{x}_0 = (x_1, \ldots, x_{s-1})\, \Gamma_{s-1}^{-1}\, \mu_{s-1}$$

Thus

$$\mathrm{Var}(x_{s-1} - \hat{x}_s) = \gamma(0) - \gamma'_{s-1}\,\Gamma_{s-1}^{-1}\,\gamma_{s-1} = \gamma(0) - \mu'_{s-1}\,\Gamma_{s-1}^{-1}\,\mu_{s-1} = \mathrm{Var}(x_0 - \hat{x}_0)$$

and

$$\mathrm{Cov}(x_0 - \hat{x}_0, x_s - \hat{x}_s) = \mathrm{E}\,x_0 x_s - \mathrm{E}\,x_0 \hat{x}_s = \gamma(s) - \mu'_{s-1}\,\Gamma_{s-1}^{-1}\,\gamma_{s-1}\,.$$

□

Consider for example the pacf of an AR(p) process

$$x_t = a_1 x_{t-1} + a_p x_{t-p} + \epsilon_t \quad (1 - a_1\,z - \ldots - a_p\,z^p) \neq 0\,\forall |\,z\,| \leq 1$$

: For $s > p$ we have

$$\alpha(s) = \mathrm{Corr}(\underbrace{x_0 - \mathrm{P}_{\mathbf{H}(1,x_1,\ldots,x_{s-1})}\,x_0}_{\in \mathbf{H}_\epsilon(s-1)}, \underbrace{x_s - \mathrm{P}_{\mathbf{H}(1,x_1,\ldots,x_{s-1})}\,x_s}_{=\epsilon_s}) = 0$$

Here $\mathbf{H}_\epsilon(s-1)$ denotes the Hilbert space spanned by all ϵ_j, $j \leq s-1$. (See the section 5.1 on the prediction from a finite past.) For $s = p$ we have by the above lemma (and the Yule Walker equations)

$$\alpha(p) = a_p \neq 0.$$

Thus the pacf of an AR(p) process is equal to zero for $|s| > p$ and not equal to zero for $s = p$. (Note the duality between AR(p) and MA(q) processes. For MA(q) processes the autocovariance function $\gamma(s)$ is equal to zero for $|s| > q$.)

This fact can be used to construct a test wether the true order is smaller or less than p_0 by testing $\alpha(p_0 + 1) = \cdots = \alpha(H) = 0$.

(ii) Test procedures for AR coefficients

We can use a sequence of likelihood ratio tests to determine the order p_0:
Bottom up test sequence:

$$\begin{array}{ll} H_{0,1} & a_1 = a_2 = \cdots = a_p = 0 \\ H_{0,2} & a_1 \neq 0, a_2 = \cdots = a_p = 0 \\ \vdots & \vdots \end{array}$$

Top down test sequence:

$$\begin{array}{ll} H_{0,1} & a_p = 0 \\ H_{0,2} & a_{p-1} = a_p = 0 \\ \vdots & \vdots \end{array}$$

For a detailed discussion see Anderson [1]

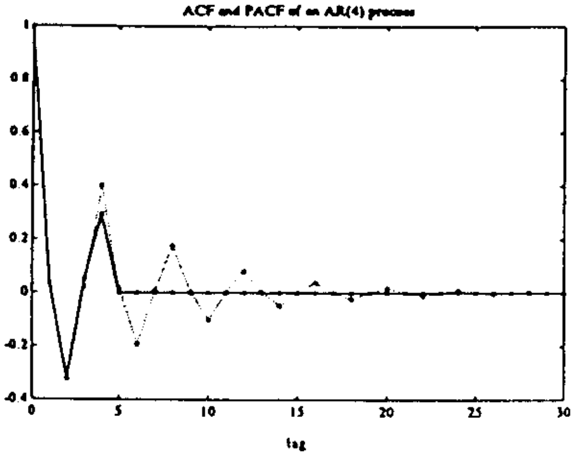


Figure 8.1: Autocorrelation function (dotted) and Partial autocorrelation function (solid) of the AR(4) process $y_t = -0.1y_{t-1} + 0.66y_{t-2} - 0.14y_{t-3} + 3.024y_{t-4} + \epsilon_t$.

(iii) Information criteria

A class of information criteria is defined by

$$A(p) = \hat{\sigma}^2(p) + \frac{c(T)}{T}p \tag{8.5}$$

Here $\hat{\sigma}^2(p)$ is the Yule Walker estimate of the variance, when the specified order is p. In more generality the first term on the r.h.s. of (8.5) is a measure for goodnes of (miss)fit, wheras p is a measure of the complexity of the model class (i.e. its dimension). $\frac{c(T)}{T}$ defines a certain trade off between these two measures. The order p_0 is estimated by

$$\hat{p}_T = \operatorname{argmin}_p A(p)$$

For $c(T) = 2$ we get Akaikes information criterion AIC and for $c(T) = \ln T$ we get the BIC criterion.

Theorem 8.8 *(Hannan) Under a certain set of assumptions the following holds:*

(i) BIC gives a strong consistent estimator of p_0, i.e. $\hat{p}_T \to p_0$ a.c.

(ii) AIC gives no consistent estimator of p_0, it tends to overestimate the true order.

For a complete set of assumptions and the proof see e.g. Hannan and Deistler [4].

8.3 Identification of Vector ARX systems

Vector ARX systems include (linear) simultaneous equation systems of the form

$$a_0 y_t = a_1 y_{t-1} + \cdots a_p y_{t-p} + d_0 z_t + \cdots + d_r z_{t-r} + \epsilon_t \tag{8.6}$$

This form which often comes from some economic theory is called the *structural form*. In many cases (especially for macro economic models) there are many a priori restrictions on the entries of the matrices a_j and d_j in the form of simple zero restrictions.

We will always assume

$$\mathrm{E}\,\epsilon_t\epsilon_t' = \Sigma > 0, \quad \mathrm{E}\,z_t\epsilon_s' = 0 \quad \text{and } \det a(z) \neq 0 \quad \forall |z| \leq 1$$

The stability assumption implies $\det a_0 \neq 0$ and thus we can rearrange this equation to give

$$y_t = \tilde{a}_1 y_{t-1} + \cdots + \tilde{a}_p y_{t-p} + \tilde{d}_0 z_t + \cdots + \tilde{d}_r z_{t-r} + \tilde{\epsilon}_t \tag{8.7}$$

which is called the *reduced form*. Here $\tilde{a}_j = a_0^{-1} a_j$, $\tilde{d}_j = a_0^{-1} d_j$ and $\tilde{\epsilon}_t = a_0^{-1}\epsilon_t$. If we define the vector $x_t = (y_{t-1}', \ldots, y_{t-p}', z_t', \ldots, z_{t-r}')'$ and the stacked parameter matrix $\tilde{\beta} = (\tilde{a}_1, \ldots, \tilde{a}_p, \tilde{d}_0, \ldots, \tilde{d}_r)$ then we have

$$y_t = \tilde{\beta} x_t + \tilde{\epsilon}_t$$

The vector x_t is often called the vector of *predetermined variables*. Since the exogeneous variables z_j are orthogonal to the noise ϵ_t and since $a(z)$ is stable we see that x_t is orthogonal to $\tilde{\epsilon}_t$, i.e. $\mathrm{E}\,x_t\tilde{\epsilon}_t{}' = 0$.

We may split y_t into two (orthogonal) parts:

$$y_t = \underbrace{a^{-1}(z)d(z)}_{l(z)} z_t + \underbrace{a^{-1}(z)}_{k(z)} \epsilon_t = y_{z,t} + y_{\epsilon,t}$$

which gives a corresponding split of the vector x_t:

$$x_t = \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \\ z_t \\ \vdots \\ z_{t-r} \end{pmatrix} = \underbrace{\begin{pmatrix} y_{z,t-1} \\ \vdots \\ y_{z,t-p} \\ z_t \\ \vdots \\ z_{t-r} \end{pmatrix}}_{x_{z,t}} + \underbrace{\begin{pmatrix} y_{\epsilon,t-1} \\ \vdots \\ y_{\epsilon,t-p} \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{x_{\epsilon,t}}$$

Lemma 8.9 *If* $\Sigma = \mathrm{E}\epsilon_t\epsilon_t' > 0$ *and* $M_{zz} = \mathrm{E}(z_t, \dots, z_{t-r})'(z_t, \dots, z_{t-r}) > 0$ *holds then* $M_{xx} = \mathrm{E}\, x_t x_t'$ *is non singular.*

Proof: We first prove for a stable AR process with $\Sigma > 0$ that $\Gamma_r > 0$ holds. We can compute the autocovariance function $\gamma(s)$ from the transferfunction $k(z) = a^{-1}(z)$ by the following infinite matrix equation

$$\begin{pmatrix} \gamma(0) & \gamma(1)' & \cdots & \cdots \\ \gamma(1) & \gamma(0) & \ddots & \\ \vdots & \ddots & \ddots & \end{pmatrix} = \underbrace{\begin{pmatrix} k_0 & k_1 & \cdots & \cdots \\ 0 & k_0 & \ddots & \\ \vdots & \ddots & \ddots & \end{pmatrix}}_{K} \begin{pmatrix} \Sigma & 0 & \cdots & \cdots \\ 0 & \Sigma & 0 & \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} K'$$

Since we can assume that $k_0 = I$ holds and by the assumption $\Sigma > 0$ all three matrices on the right hand side of this equation are nonsingular. Thus also matrix on the left hand side is nonsingular. Thus the matrix Γ_r is nonsingular since it is a finite section of this infinite matrix.

Now we have

$$M_{xx} = M^z + M^\epsilon = \begin{pmatrix} M_{11}^z & M_{12}^z \\ M_{21}^z & M_{22}^z \end{pmatrix} + \begin{pmatrix} M_{11}^\epsilon & 0 \\ 0 & 0 \end{pmatrix}$$

where M^z and M^ϵ are the covariances of the vectors $x_{z,t}$ and $x_{\epsilon,t}$ which we have partitioned conformingly. We have $M_{11}^\epsilon > 0$ by the above considerations and $M_{22}^z = M_{zz} > 0$ by assumption.

Let $c = (c_1', c_2')'$ be a vector which is partitioned conformingly. Then $0 = c' M_{xx} c$ implies $c_1' M_{11}^2 c_1 = 0$ and thus $c_1 = 0$. Thus we further get $c_2' M_{22}^1 c_2 = 0$ and $c_2 = 0$. □

From equation (8.7) we get

$$M_{yx} = \mathrm{E}\, y_t x_t' = \bar{\beta}\, \mathrm{E}\, x_t x_t' + \mathrm{E}\, \bar{\epsilon}_t x_t' = \bar{\beta} M_{xx} + 0$$

and thus by the above considerations

$$\bar{\beta} = M_{yx} M_{xx}^{-1}.$$

We now introduce a more compact notation, which is similar to usual notation for simple regression models. We define:

$$Y = \begin{pmatrix} y_1' \\ y_2' \\ \cdot \\ y_T' \end{pmatrix} = \begin{pmatrix} y_1^{(1)} & \cdots & y_1^{(n)} \\ y_2^{(1)} & \cdots & y_2^{(n)} \\ \vdots & & \vdots \\ y_T^{(1)} & \cdot & y_T^{(n)} \end{pmatrix}, \quad U = \begin{pmatrix} \epsilon_1' \\ \epsilon_2' \\ \vdots \\ \epsilon_T' \end{pmatrix} = \begin{pmatrix} \epsilon_1^{(1)} & \cdots & \epsilon_1^{(n)} \\ \epsilon_2^{(1)} & \cdots & \epsilon_2^{(n)} \\ \vdots & & \vdots \\ \epsilon_T^{(1)} & & \epsilon_T^{(n)} \end{pmatrix}$$

$$\Gamma = a_0', \quad B' = \begin{pmatrix} a_1 & a_2 & \cdots & a_p & d_0 & d_1 & \cdots & d_r \end{pmatrix},$$

$$X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{pmatrix} =$$

$$\begin{pmatrix} 0 & \cdots & 0 & \cdots\cdots & 0 & \cdots & 0 & z_1^{(1)} & \cdots & z_1^{(m)} & \cdots\cdots & 0 & \cdots & 0 \\ y_1^{(1)} & \cdots & y_1^{(n)} & \cdots\cdots & \vdots & & \vdots & z_2^{(1)} & \cdots & z_2^{(m)} & \cdots\cdots & \vdots & & \vdots \\ \vdots & & \vdots & & y_1^{(1)} & \cdots & y_1^{(n)} & \vdots & & \vdots & & z_1^{(1)} & \cdots & z_1^{(n)} \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ y_{T-1}^{(1)} & \cdots & y_{T-1}^{(n)} & \cdots\cdots & y_{T-p}^{(1)} & \cdots & y_{T-p}^{(n)} & z_T^{(1)} & \cdots & z_T^{(m)} & \cdots\cdots & z_{T-r}^{(1)} & \cdots & z_{T-r}^{(n)} \end{pmatrix}$$

Here we have replaced the unknown "starting values" $y_s^{(i)}$, $z_s^{(i)}$, $s \leq 0$ by zeros. Using this notation we may write the structural form as

$$Y\Gamma = XB + U \tag{8.8}$$

and the reduced form as

$$Y = X\underbrace{B\Gamma^{-1}}_{\Pi} + \underbrace{U\Gamma^{-1}}_{V} \tag{8.9}$$

Note that $\Pi = B\Gamma^{-1} = (a_1, \cdots, c_r)'(a_0^{-1})' = (a_0^{-1}(a_1, \cdots, c_r))' = \tilde{\beta}'$ and that the t-th row of V contains $\epsilon_t'(a_0^{-1})' = (a_0^{-1}\epsilon_t)' = \tilde{\epsilon}_t'$. The OLS estimate for Π is given by

$$\hat{\Pi} = (X'X)^{-1}(X'Y)$$

If the sample moments $T^{-1}(X'X)$, $T^{-1}(X'Y)$ are weakly consistent estimates of their population counterparts M_{xx} and M_{xy} respectively then the OLS estimate $\hat{\Pi}$ is a consistent estimate of Π:

$$\hat{\Pi} = \left(\frac{1}{T}X'X\right)^{-1}\left(\frac{1}{T}X'Y\right) \xrightarrow{p} M_{xx}^{-1}M_{xy} = (M_{yx}M_{xx}^{-1})' = \tilde{\beta}_t' = \Pi$$

If we use the normalization condition $a_{0,ii} = 1$ and rearrange the i-th equation of (8.8) such that all variables except $y_t^{(i)}$ are on the right hand side, using the a priori zero restriction in the structrural form, then we may write the i-th equation in the form

$$\mathbf{y}_i = Y_i\gamma_i + X_i\beta_i + \mathbf{u}_i = W_i\alpha_i + \mathbf{u}_i$$

Here $\mathbf{y}_i$ denotes the i-th column of Y, and $\mathbf{u}_i$ the i-th column of U. Thus e.g. $\mathbf{y}_1$ contains all observations of the first endogenous variable from time $t = 1$ to time $t = T$. γ_i and

β_i are the i-th columns of $(-\Gamma)$ and B respectively, where all elements which are a priori zero are ommited. Y_i and X_i are the corresponding columns of Y and X, $W_i = (Y_i, X_i)$ and $\alpha_i = (\gamma_i', \beta_i')'$.

The OLS estimate of α_i then is given by

$$\hat{\alpha}_i = (W_i'W_i)^{-1}(W_i'y_i) = (W_i'W_i)^{-1}(W_i'(W_i\alpha_i + \mathrm{u}_i)) = \alpha_i + (W_i'W_i)^{-1}(W_i'\mathrm{u}_i)$$

The (probability) limit of the estimation error $(\hat{\alpha}_i - \alpha_i)$ is given by

$$\begin{gathered}\mathrm{plim}(\hat{\alpha}_i - \alpha_i) = \mathrm{plim}(\tfrac{1}{T}W_i'W_i)^{-1}\,\mathrm{plim}(\tfrac{1}{T}W_i'\mathrm{u}_i) = \\ \mathrm{plim}(\tfrac{1}{T}W_i'W_i)^{-1}\,\mathrm{plim}(\tfrac{1}{T}\begin{pmatrix} Y_i'\mathrm{u}_i \\ X_i'\mathrm{u}_i \end{pmatrix}) = \mathrm{plim}(\tfrac{1}{T}W_i'W_i)^{-1}\begin{pmatrix} \mathrm{plim}(\tfrac{1}{T}Y_i'\mathrm{u}_i) \\ 0 \end{pmatrix}\end{gathered} \tag{8.10}$$

Here we have assumed that the sample moments are (weakly) consistent estimates of their population counterparts. Thus $T^{-1}W_i'W_i$ will converge to its population counterpart and thus to some (constant) nonsingular matrix. Since the predetermined variables are orthogonal to the errors the sample covariance $T^{-1}X_i'\mathrm{u}_i$ will converge to zero. We may write $Y_i = YS^{(i)}$ and $U_i = Ue_i$, where $S^{(i)}$ is a "selection matrix", which picks out of Y the columns corresponding to Y_i and e_i is the i-th unit vector. Using this notation we get

$$\begin{aligned}\mathrm{plim}\,\tfrac{1}{T}Y_i'\mathrm{u}_i &= \mathrm{plim}\,\tfrac{1}{T}(S^{(i)})'(\Pi'X' + V')Ue_i \\ &= (S^{(i)})'\Pi'\underbrace{\mathrm{plim}(T^{-1}X'U)}_{=0}e_i + (S^{(i)})'(\Gamma^{-1})'\underbrace{\mathrm{plim}(T^{-1}U'U)}_{=\Sigma}e_i \\ &= (S^{(i)})'(\Gamma^{-1})'\Sigma e_i\end{aligned}$$

This probability limit is not equal to zero in general. Therefore the OLS-estimate $\hat{\alpha}_i$ in general will be asymptotically biased. The essential reason for this asymptotic bias is that the "regressor" Y_i contains present vaues of y_t which are in general correlated with the noise $\epsilon_t^{(i)}$.

We now want to give two examples of linear simultaneous equation systems:

Example 1: Consider the simple Keynesian model

$$\begin{aligned}C_t &= \alpha Y_t + u_t \\ Y_t &= C_t + G_t\end{aligned}$$

where

C_t ··· Consumption (endogenous)
Y_t ··· (disposable) income (endogenous)
G_t ··· private investments and goverment expenditures (exogeneous)

By substituting $C_t = Y_t - G_t$ into the first equation we get $Y_t = \frac{1}{1-\alpha}(G_t + u_t)$ and

$$\mathrm{E}\,Y_t u_t = \mathrm{E}(\frac{1}{1-\alpha}(G_t + u_t)u_t) = \frac{1}{1-\alpha}\,\mathrm{E}\,u_t^2 \neq 0$$

using the orthogonality of the noise and the exogenous variable ($\mathrm{E}G_t u_t = 0$). Therefore OLS will give a biased and non consistent estimate for α in the first equation, since the noise u_t is not orthogonal to the regressor Y_t! We have

$$\operatorname{plim}(\hat{\alpha}-\alpha) = \operatorname{plim}(\frac{1}{T}\sum Y_t^2)^{-1}\underbrace{\operatorname{plim}\frac{1}{T}\sum Y_t u_t}_{\mathrm{E}Y_t u_t} \neq 0$$

Example 2: The Klein model I:

$$\begin{aligned}
C_t &= c_0 + c_1 Q_t + c_2 Q_{t-1} + c_3 W_t + u_t^{(1)} \\
I_t &= i_0 + i_1 Q_t + i_2 Q_{t-1} + i_3 K_{t-1} + u_t^{(2)} \\
W_t^{\mathrm{pr}} &= w_0 + w_1 E_t + w_2 E_{t-1} + w_3(t-t_0) + u_t^{(3)} \\
Y_t &= C_t + I_t + G_t - T_t^{\mathrm{ind}} \\
Q_t &= Y_t - W_t \\
K_t &= K_{t-1} + I_t \\
W_t &= W_t^{\mathrm{pr}} + W_t^{\mathrm{G}} \\
E_t &= Y_t + T_t^{\mathrm{ind}} - W_t^{\mathrm{G}}
\end{aligned}$$

The endogenous variables are

C_t	··· consumption
I_t	··· investments
W_t^{pr}	··· wages in the private sector
Y_t	··· income
Q_t	··· profits
K_t	··· capital stock
W_t	··· wages
E_t	··· net social product of the private sector

and the exogenous variables are

W_t^{G}	··· wages in the govermental sector
T_t^{ind}	··· indirect taxes
G_t	··· goverment expenditures
$t-t_0$	··· time
1	··· constant

The vector y_t of endogenous variables and the vector z_t of exogenous variables are

$$y_t = \begin{pmatrix} C_t \\ I_t \\ W_t^{\mathrm{pr}} \\ Y_t \\ Q_t \\ K_t \\ W_t \\ E_t \end{pmatrix} \quad \text{and } z_t = \begin{pmatrix} W_t^{\mathrm{G}} \\ T_t^{\mathrm{ind}} \\ G_t \\ t - t_0 \\ 1 \end{pmatrix}$$

The orders are $p = 1$ and $r = 0$ and the coefficient matrices a_j and d_j of equation (8.6) are given by:

$$a_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & -c_1 & 0 & -c_3 & 0 \\ 0 & 1 & 0 & 0 & -i_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -w_1 \\ -1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}, a_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & c_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & i_2 & i_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$d_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & c_0 \\ 0 & 0 & 0 & 0 & i_0 \\ 0 & 0 & 0 & w_3 & w_0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

For the first equation (consumption) we have

$$\mathbf{y}_1 = \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_T \end{pmatrix}, \quad \mathbf{u}_1 = \begin{pmatrix} u_1^{(1)} \\ u_2^{(1)} \\ \vdots \\ u_T^{(1)} \end{pmatrix}, \quad Y_1 = \begin{pmatrix} Q_1 & W_1 \\ Q_2 & W_2 \\ \vdots & \vdots \\ Q_T & W_T \end{pmatrix} \quad \text{and } X_1 = \begin{pmatrix} 0 & 1 \\ Q_1 & 1 \\ \vdots & \vdots \\ Q_{T-1} & 1 \end{pmatrix}$$

and the corresponding parameter matrices are

$$\gamma_1 = \begin{pmatrix} c_1 \\ c_3 \end{pmatrix}, \quad \beta_1 = \begin{pmatrix} c_2 \\ c_0 \end{pmatrix} \quad \text{and } \alpha_1 = \begin{pmatrix} c_1 \\ c_3 \\ c_2 \\ c_0 \end{pmatrix}.$$

The Klein model I satisfies the counting conditions for structural identifiability: The diagonal elements of a_0 are equal to one and each row of (a_0, a_1, d_0) contains at least $(n-1) = 7$ a priori zeros.

Estimation of the Structural Form:

Note that Π is uniquely determined from the second moments of the observations. Thus in the identifiable case we may uniquely determine B, Γ from the equation $B = \Pi\Gamma$ or $\Pi = B\Gamma^{-1}$ using the a priori restriction on B and Γ. Since Π may be consistently estimated by OLS, this is a possible way to estimate B and Γ.

The problem is that usually we have much more a priori restricions than are needed for identifiability. (See e.g. the coefficient matrices a_j, d_j of the Klein model.) Thus the set $\mathcal{P}_r$ of all $\Pi = B\Gamma^{-1}$ where B, Γ satisfy the restrictions is a very "thin" subset of the set $\mathcal{P}$ of all Π's. Thus the OLS estimate $\hat{\Pi}$ will not be contained in the set $\mathcal{P}_r$ allthough the true Π is contained in $\mathcal{P}_r$ and $\hat{\Pi}$ is a consistent estimate of Π. Thus in general there will not exist $\hat{B}$ and $\hat{\Gamma}$ satisfying our restrictions and $\hat{\Pi} = \hat{B}\hat{\Gamma}^{-1}$. (Overidentifiability)

Recursive systems:

If $\Sigma = \mathrm{E}\epsilon_t\epsilon_t'$ is diagonal and Γ (after possible reordering of endogenous variables) is lower triangular then (equation wise) OLS will give consistent estimates:

$$\begin{array}{rrcl} n - \text{th equation:} & \mathbf{y}_n & = & X_n\beta_n + \mathbf{u}_n \\ (n-1) - \text{th equation:} & \mathbf{y}_{n-1} & = & -\gamma_{n,n-1}\mathbf{y}_n + X_{n-1}\beta_{n-1} + \mathbf{u}_{n-1} \\ (n-2) - \text{th equation:} & \mathbf{y}_{n-2} & = & -\gamma_{n-1,n-2}\mathbf{y}_{n-1} + -\gamma_{n,n-2}\mathbf{y}_n + X_{n-2}\beta_{n-2} + \mathbf{u}_{n-2} \\ \cdots\cdots & & & \cdots\cdots\cdots\cdots\cdots\cdots \end{array}$$

Note that by assumptions the entries γ_{ij} of the matrix Γ satisfy $\gamma_{ii} = 1$ and $\gamma_{ij} = 0$ for $j > i$. OLS gives consistent estimate in the n-th equation since the predetermined variables are orthogonal to the noise ϵ_t. The regressor $\mathbf{y}_n = X_n\beta_n + \mathbf{u}_n$ in the $(n-1)$-th equation is orthogonal to the noise $\mathbf{u}_{n-1}$ since X_n is orthogonal to $\mathbf{u}_{n-1}$ and because by assumption the n-component of ϵ_t is uncorrelated to the $(n-1)$-th component of ϵ_t. Proceeding in a similar way we can see that also the other equations may be consistently estimated.

2-Stage Least Squares Estimator (Theil, Basemann)

As the above considerations show the bias is caused by the regressors contained in Y_i. Thus the idea here is to use some instrumental variables P_i for Y_i. These instruments must satisfy the following conditions:

$$\text{plim}\,\frac{1}{T}(P_i, X_i)'(Y_i, X_i) = \text{plim}\,\frac{1}{T}\begin{pmatrix} P_i'Y_i & P_i'X_i \\ X_i'Y_i & X_i'X_i \end{pmatrix} \text{ exists and is non singular}$$

and

$$\text{plim}\,\frac{1}{T}(P_i'\mathbf{u}_i) = 0.$$

Under these conditions the instrumental variables estimate

$$\hat{\alpha}_i = [(P_i, X_i)'(Y_i, X_i)]^{-1}\,[(P_i, X_i)'\mathbf{y}_i]$$

is (weakly) consistent as can easily be seen.

It is easy to see that $\tilde{P}_i = X\Pi S^{(i)}$ satisfies these two conditions for consistency. But Π is not known and thus $\tilde{P}_i$ is not a feasible instrument. It can be shown that $P_i = X\hat{\Pi}S^{(i)}$ is a feasible instrument for Y_i satisfying the above conditions. Here $\hat{\Pi}$ denotes the OLS estimate of Π (which, as we know, is consistent for Π).

Therefore the whole estimation procedure is as follows

1. step: Compute the OLS estimate $\hat{\Pi}$ for the reduced form parameters Π.

2. step: Compute the instrumental variable estimate using $P_i = X\hat{\Pi}S^{(i)}$ as instruments for Y_i.

8.4 Identification of ARMA systems

Under a Gaussian assumption $-2T^{-1}$ times the logarithm of the likelihood is (up to constants) given by

$$l_T(a, b, \Sigma) = \frac{1}{T}\log\det\Gamma_T + \frac{1}{T}\mathbf{x}(T)\,\Gamma_T\,\mathbf{x}(T)'$$

where $\mathbf{x}(T)$ is the stacked vectors of observations

$$\mathbf{x}(T) = \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}$$

and Γ_T is the covariance matrix $\mathrm{E}\,\mathbf{x}(T)\mathbf{x}(T)'$. Γ_T and thus the likelihood depends on the ARMA parameters a, b and Σ. $l_T(a, b, \Sigma)$ is often called the *pseudo likelihood*, because we use a Gaussian distribution for the noise process, although we do not "believe" on normality.

By maximizing the likelihood (or by minimizing $l_T(a, b, \Sigma)$) as a function of a, b and Σ the parameters a, b and Σ may be estimated.

List of Figures

0.1 Examples of time series 1

1.1 Interpretation of the covariance 4

1.2 Differencing of time series 8

1.3 Periodogram 11

1.4 Information criteria 13

2.1 White noise 20

2.2 Random walk process 24

2.3 AR(1) processes 25

2.4 A simple harmonic process 27

2.5 Autorrelation of a harmonic process 29

2.6 Spectral distribution function of a harmonic process 30

3.1 Spectral densities of AR(1) processes 43

3.2 Spectral density of an AR(4) process 44

3.3 Transferfunction of first differences 49

3.4 Transferfunction of fourth differences 49

3.5 Optimal seasonal filter 51

4.1 Roots of det($a(z)$) 60

4.2 Roots of $f_y(z)$ 71

7.1 Fejer kernel 108

7.2 Expectation of the periodogram 109

8.1 PACF of an AR(4) process 122

References

[1] T.W. Anderson. *The Statistical Analysis of Time Series.* John Wiley, New York, 1971.

[2] G.E.P. Box and G.M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, 1970.

[3] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer Verlag, New York, 2nd edition, 1989.

[4] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems.* John Wiley & Sons, New York, 1988.

[5] Yu.A. Rozanov. *Stationary Random Processes.* Holden-Day, San Francisco, 1967.