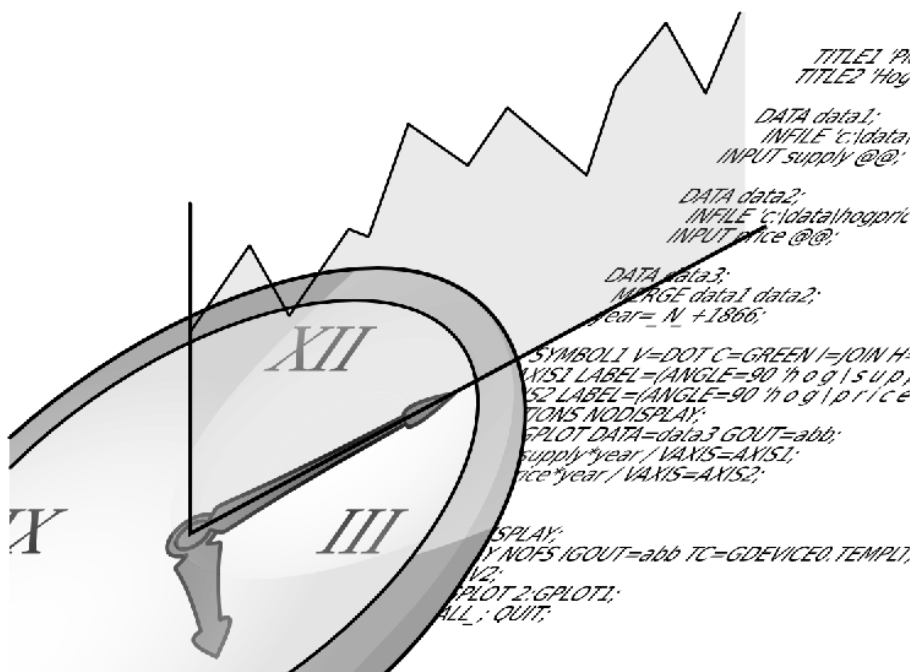


A First Course on Time Series Analysis

Examples with SAS



Chair of Statistics, University of Würzburg

March 20, 2011

**A First Course on
Time Series Analysis — Examples with SAS**
by Chair of Statistics, University of Würzburg.
Version 2011.March.01

Copyright © 2011 Michael Falk.

Editors	Michael Falk, Frank Marohn, René Michel, Daniel Hofmann, Maria Macke, Christoph Spachmann, Stefan Englert
Programs	Bernward Tewes, René Michel, Daniel Hofmann, Christoph Spachmann, Stefan Englert
Layout and Design	Peter Dinges, Stefan Englert

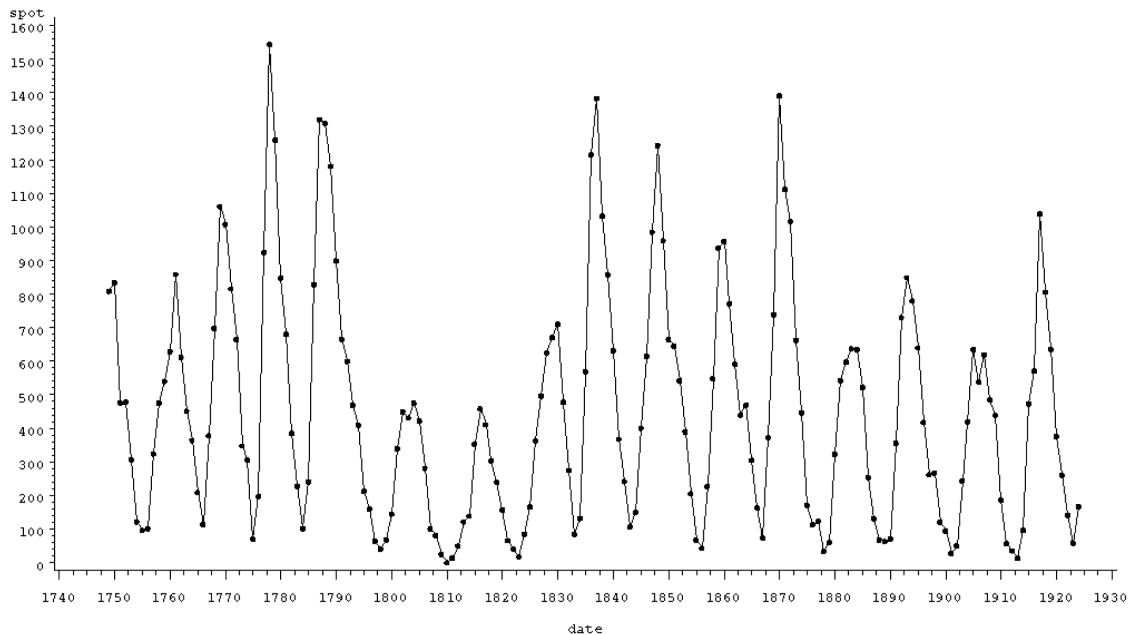
Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. Windows is a trademark, Microsoft is a registered trademark of the Microsoft Corporation. The authors accept no responsibility for errors in the programs mentioned of their consequences.

Preface

The analysis of real data by means of statistical methods with the aid of a software package common in industry and administration usually is not an integral part of mathematics studies, but it will certainly be part of a future professional work.

The practical need for an investigation of time series data is exemplified by the following plot, which displays the yearly sunspot numbers between 1749 and 1924. These data are also known as the Wolf or Wölfer (a student of Wolf) Data. For a discussion of these data and further literature we refer to Wei and Reilly (1989), Example 6.2.5.



Plot 1: Sunspot data

The present book links up elements from time series analysis with a selection of statistical procedures used in general practice including the

statistical software package SAS (Statistical Analysis System). Consequently this book addresses students of statistics as well as students of other branches such as economics, demography and engineering, where lectures on statistics belong to their academic training. But it is also intended for the practitioner who, beyond the use of statistical tools, is interested in their mathematical background. Numerous problems illustrate the applicability of the presented statistical procedures, where SAS gives the solutions. The programs used are explicitly listed and explained. No previous experience is expected neither in SAS nor in a special computer system so that a short training period is guaranteed. This book is meant for a two semester course (lecture, seminar or practical training) where the first three chapters can be dealt with in the first semester. They provide the principal components of the analysis of a time series in the time domain. Chapters 4, 5 and 6 deal with its analysis in the frequency domain and can be worked through in the second term. In order to understand the mathematical background some terms are useful such as convergence in distribution, stochastic convergence, maximum likelihood estimator as well as a basic knowledge of the test theory, so that work on the book can start after an introductory lecture on stochastics. Each chapter includes exercises. An exhaustive treatment is recommended. Chapter 7 (case study) deals with a practical case and demonstrates the presented methods. It is possible to use this chapter independent in a seminar or practical training course, if the concepts of time series analysis are already well understood.

Due to the vast field a selection of the subjects was necessary. Chapter 1 contains elements of an exploratory time series analysis, including the fit of models (logistic, Mitscherlich, Gompertz curve) to a series of data, linear filters for seasonal and trend adjustments (difference filters, Census X-11 Program) and exponential filters for monitoring a system. Autocovariances and autocorrelations as well as variance stabilizing techniques (Box-Cox transformations) are introduced. Chapter 2 provides an account of mathematical models of stationary sequences of random variables (white noise, moving averages, autoregressive processes, ARIMA models, cointegrated sequences, ARCH- and GARCH-processes) together with their mathematical background (existence of stationary processes, covariance

generating function, inverse and causal filters, stationarity condition, Yule–Walker equations, partial autocorrelation). The Box–Jenkins program for the specification of ARMA-models is discussed in detail (AIC, BIC and HQ information criterion). Gaussian processes and maximum likelihood estimation in Gaussian models are introduced as well as least squares estimators as a nonparametric alternative. The diagnostic check includes the Box–Ljung test. Many models of time series can be embedded in state-space models, which are introduced in Chapter 3. The Kalman filter as a unified prediction technique closes the analysis of a time series in the time domain. The analysis of a series of data in the frequency domain starts in Chapter 4 (harmonic waves, Fourier frequencies, periodogram, Fourier transform and its inverse). The proof of the fact that the periodogram is the Fourier transform of the empirical autocovariance function is given. This links the analysis in the time domain with the analysis in the frequency domain. Chapter 5 gives an account of the analysis of the spectrum of the stationary process (spectral distribution function, spectral density, Herglotz’s theorem). The effects of a linear filter are studied (transfer and power transfer function, low pass and high pass filters, filter design) and the spectral densities of ARMA-processes are computed. Some basic elements of a statistical analysis of a series of data in the frequency domain are provided in Chapter 6. The problem of testing for a white noise is dealt with (Fisher’s κ -statistic, Bartlett–Kolmogorov–Smirnov test) together with the estimation of the spectral density (periodogram, discrete spectral average estimator, kernel estimator, confidence intervals). Chapter 7 deals with the practical application of the Box–Jenkins Program to a real dataset consisting of 7300 discharge measurements from the Donau river at Donauwoerth. For the purpose of studying, the data have been kindly made available to the University of Würzburg. A special thank is dedicated to Rudolf Neusiedl. Additionally, the asymptotic normality of the partial and general autocorrelation estimators is proven in this chapter and some topics discussed earlier are further elaborated (order selection, diagnostic check, forecasting).

This book is consecutively subdivided in a statistical part and a SAS-specific part. For better clearness the SAS-specific part, including the diagrams generated with SAS, is between two horizontal bars,

separating it from the rest of the text.

```
1  /* This is a sample comment. */
2  /* The first comment in each program will be its name. */
3
4  Program code will be set in typewriter-font. SAS keywords like DATA or
5  PROC will be set in bold.
6
7  Also all SAS keywords are written in capital letters. This is not
8  necessary as SAS code is not case sensitive, but it makes it easier to
9  read the code.
10
11 Extra-long lines will be broken into smaller lines with continuation
12   ↪marked by an arrow and indentation.
   (Also, the line-number is missing in this case.)
```

In this area, you will find a step-by-step explanation of the above program. The keywords will be set in typewriter-font. Please note that SAS cannot be explained as a whole this way. Only the actually used commands will be mentioned.

Contents

1	Elements of Exploratory Time Series Analysis	1
1.1	The Additive Model for a Time Series	2
1.2	Linear Filtering of Time Series	16
1.3	Autocovariances and Autocorrelations	35
	Exercises	41
2	Models of Time Series	47
2.1	Linear Filters and Stochastic Processes	47
2.2	Moving Averages and Autoregressive Processes . .	61
2.3	The Box–Jenkins Program	99
	Exercises	111
3	State-Space Models	121
3.1	The State-Space Representation	121
3.2	The Kalman-Filter	125
	Exercises	132
4	The Frequency Domain Approach of a Time Series	135
4.1	Least Squares Approach with Known Frequencies .	136
4.2	The Periodogram	142
	Exercises	155

5	The Spectrum of a Stationary Process	159
5.1	Characterizations of Autocovariance Functions	160
5.2	Linear Filters and Frequencies	166
5.3	Spectral Density of an ARMA-Process	175
	Exercises	181
6	Statistical Analysis in the Frequency Domain	187
6.1	Testing for a White Noise	187
6.2	Estimating Spectral Densities	196
	Exercises	216
7	The Box–Jenkins Program: A Case Study	223
7.1	Partial Correlation and Levinson–Durbin Recursion .	224
7.2	Asymptotic Normality of Partial Autocorrelation Estimator	234
7.3	Asymptotic Normality of Autocorrelation Estimator .	259
7.4	First Examinations	272
7.5	Order Selection	284
7.6	Diagnostic Check	311
7.7	Forecasting	324
	Exercises	335
	Bibliography	337
	Index	341
	SAS-Index	348
	GNU Free Documentation Licence	351

Elements of Exploratory Time Series Analysis

A time series is a sequence of observations that are arranged according to the time of their outcome. The annual crop yield of sugar-beets and their price per ton for example is recorded in agriculture. The newspapers' business sections report daily stock prices, weekly interest rates, monthly rates of unemployment and annual turnovers. Meteorology records hourly wind speeds, daily maximum and minimum temperatures and annual rainfall. Geophysics is continuously observing the shaking or trembling of the earth in order to predict possibly impending earthquakes. An electroencephalogram traces brain waves made by an electroencephalograph in order to detect a cerebral disease, an electrocardiogram traces heart waves. The social sciences survey annual death and birth rates, the number of accidents in the home and various forms of criminal activities. Parameters in a manufacturing process are permanently monitored in order to carry out an on-line inspection in quality assurance.

There are, obviously, numerous reasons to record and to analyze the data of a time series. Among these is the wish to gain a better understanding of the data generating mechanism, the prediction of future values or the optimal control of a system. The characteristic property of a time series is the fact that the data are *not generated independently*, their dispersion varies in time, they are often governed by a trend and they have cyclic components. Statistical procedures that suppose *independent* and *identically distributed* data are, therefore, excluded from the analysis of time series. This requires proper methods that are summarized under *time series analysis*.

1.1 The Additive Model for a Time Series

The additive model for a given time series y_1, \dots, y_n is the assumption that these data are realizations of random variables Y_t that are themselves sums of four components

$$Y_t = T_t + Z_t + S_t + R_t, \quad t = 1, \dots, n. \quad (1.1)$$

where T_t is a (monotone) function of t , called *trend*, and Z_t reflects some nonrandom long term cyclic influence. Think of the famous business cycle usually consisting of recession, recovery, growth, and decline. S_t describes some nonrandom short term cyclic influence like a seasonal component whereas R_t is a random variable grasping all the deviations from the ideal non-stochastic model $y_t = T_t + Z_t + S_t$. The variables T_t and Z_t are often summarized as

$$G_t = T_t + Z_t, \quad (1.2)$$

describing the long term behavior of the time series. We suppose in the following that the expectation $E(R_t)$ of the error variable exists and equals zero, reflecting the assumption that the random deviations above or below the nonrandom model balance each other on the average. Note that $E(R_t) = 0$ can always be achieved by appropriately modifying one or more of the nonrandom components.

Example 1.1.1. (Unemployed1 Data). The following data y_t , $t = 1, \dots, 51$, are the monthly numbers of unemployed workers in the building trade in Germany from July 1975 to September 1979.

MONTH	T	UNEMPLYD
July	1	60572
August	2	52461
September	3	47357
October	4	48320
November	5	60219
December	6	84418
January	7	119916
February	8	124350
March	9	87309

April	10	57035
May	11	39903
June	12	34053
July	13	29905
August	14	28068
September	15	26634
October	16	29259
November	17	38942
December	18	65036
January	19	110728
February	20	108931
March	21	71517
April	22	54428
May	23	42911
June	24	37123
July	25	33044
August	26	30755
September	27	28742
October	28	31968
November	29	41427
December	30	63685
January	31	99189
February	32	104240
March	33	75304
April	34	43622
May	35	33990
June	36	26819
July	37	25291
August	38	24538
September	39	22685
October	40	23945
November	41	28245
December	42	47017
January	43	90920
February	44	89340
March	45	47792
April	46	28448
May	47	19139
June	48	16728
July	49	16523
August	50	16622
September	51	15499

Listing 1.1.1: Unemployed1 Data.

```

1  /* unemployed1_listing.sas */
2  TITLE1 'Listing';
3  TITLE2 'Unemployed1 Data';
4
5  /* Read in the data (Data-step) */
6  DATA data1;
7      INFILE 'c:\data\unemployed1.txt';
8      INPUT month $ t unemployd;

```

```
9
10 /* Print the data (Proc-step) */
11 PROC PRINT DATA = data1 NOOBS;
12 RUN; QUIT;
```

This program consists of two main parts, a DATA and a PROC step.

The DATA step started with the DATA statement creates a temporary dataset named data1. The purpose of INFILE is to link the DATA step to a raw dataset outside the program. The path-name of this dataset depends on the operating system; we will use the syntax of MS-DOS, which is most commonly known. INPUT tells SAS how to read the data. Three variables are defined here, where the first one contains character values. This is determined by the \$ sign behind the variable name. For each variable one value per line is read from the source into the computer's memory.

The statement PROC procedurename DATA=filename; invokes a procedure that is linked to the data from filename. Without the option DATA=filename the most recently created file is used.

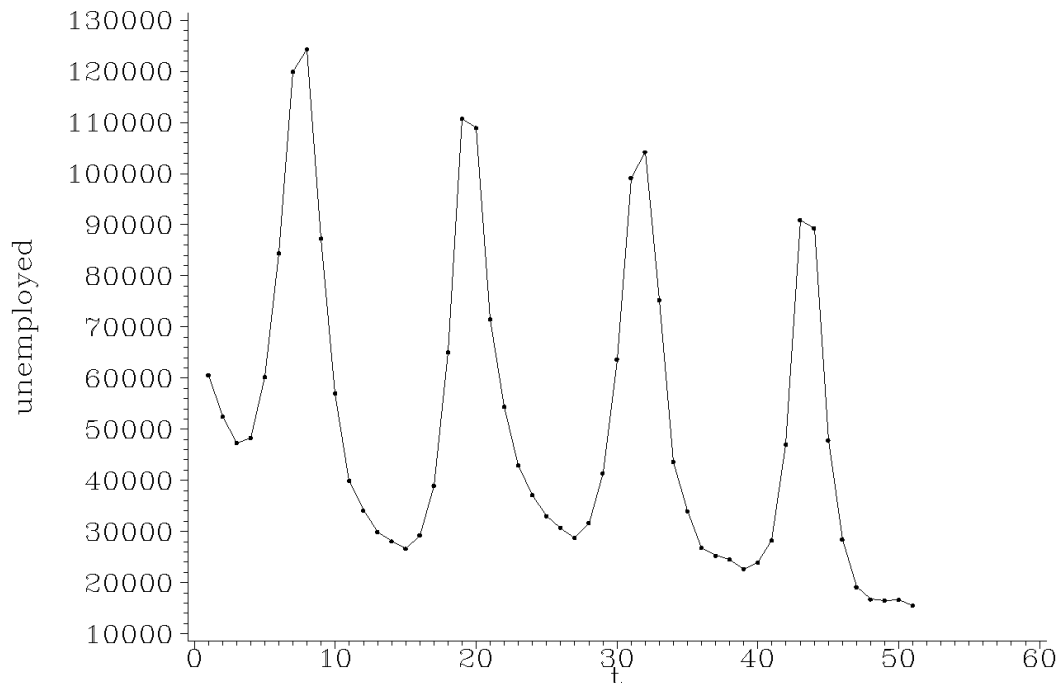
The PRINT procedure lists the data; it comes with numerous options that allow control of the

variables to be printed out, 'dress up' of the display etc. The SAS internal observation number (OBS) is printed by default, NOOBS suppresses the column of observation numbers on each line of output. An optional VAR statement determines the order (from left to right) in which variables are displayed. If not specified (like here), all variables in the data set will be printed in the order they were defined to SAS. Entering RUN; at any point of the program tells SAS that a unit of work (DATA step or PROC) ended. SAS then stops reading the program and begins to execute the unit. The QUIT; statement at the end terminates the processing of SAS.

A line starting with an asterisk * and ending with a semicolon ; is ignored. These comment statements may occur at any point of the program except within raw data or another statement.

The TITLE statement generates a title. Its printing is actually suppressed here and in the following.

The following plot of the Unemployed1 Data shows a seasonal component and a downward trend. The period from July 1975 to September 1979 might be too short to indicate a possibly underlying long term business cycle.



Plot 1.1.2: Unemployed1 Data.

```

1  /* unemployed1_plot.sas */
2  TITLE1 'Plot';
3  TITLE2 'Unemployed1 Data';
4
5  /* Read in the data */
6  DATA data1;
7  INFILE 'c:\data\unemployed1.txt';
8  INPUT month $ t unemployd;
9
10 /* Graphical Options */
11 AXIS1 LABEL=(ANGLE=90 'unemployed');
12 AXIS2 LABEL=('t');
13 SYMBOL1 V=DOT C=GREEN I=JOIN H=0.4 W=1;
14
15 /* Plot the data */
16 PROC GLOT DATA=data1;
17     PLOT unemployd*t / VAXIS=AXIS1 HAXIS=AXIS2;
18 RUN; QUIT;

```

Variables can be plotted by using the GLOT procedure, where the graphical output is controlled by numerous options.

The AXIS statements with the LABEL options control labelling of the vertical and horizontal axes. ANGLE=90 causes a rotation of the label of 90° so that it parallels the (vertical) axis in this example.

The SYMBOL statement defines the manner

in which the data are displayed. V=DOT C=GREEN I=JOIN H=0.4 W=1 tell SAS to plot green dots of height 0.4 and to join them with a line of width 1. The PLOT statement in the GLOT procedure is of the form PLOT y-variable*x-variable / options;, where the options here define the horizontal and the vertical axes.

Models with a Nonlinear Trend

In the additive model $Y_t = T_t + R_t$, where the nonstochastic component is only the trend T_t reflecting the growth of a system, and assuming $E(R_t) = 0$, we have

$$E(Y_t) = T_t =: f(t).$$

A common assumption is that the function f depends on several (unknown) *parameters* β_1, \dots, β_p , i.e.,

$$f(t) = f(t; \beta_1, \dots, \beta_p). \quad (1.3)$$

However, the *type* of the function f is known. The unknown parameters β_1, \dots, β_p are then to be estimated from the set of realizations y_t of the random variables Y_t . A common approach is a *least squares estimate* $\hat{\beta}_1, \dots, \hat{\beta}_p$ satisfying

$$\sum_t \left(y_t - f(t; \hat{\beta}_1, \dots, \hat{\beta}_p) \right)^2 = \min_{\beta_1, \dots, \beta_p} \sum_t \left(y_t - f(t; \beta_1, \dots, \beta_p) \right)^2, \quad (1.4)$$

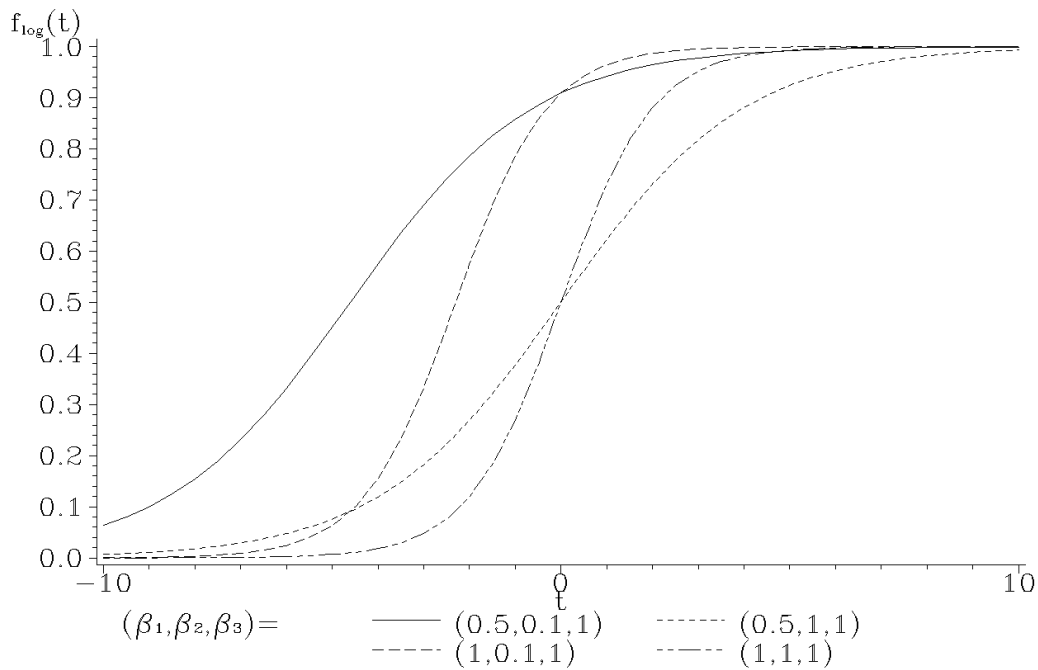
whose computation, if it exists at all, is a numerical problem. The value $\hat{y}_t := f(t; \hat{\beta}_1, \dots, \hat{\beta}_p)$ can serve as a *prediction* of a future y_t . The *observed* differences $y_t - \hat{y}_t$ are called *residuals*. They contain information about the goodness of the fit of our model to the data. In the following we list several popular examples of trend functions.

The Logistic Function

The function

$$f_{\log}(t) := f_{\log}(t; \beta_1, \beta_2, \beta_3) := \frac{\beta_3}{1 + \beta_2 \exp(-\beta_1 t)}, \quad t \in \mathbb{R}, \quad (1.5)$$

with $\beta_1, \beta_2, \beta_3 \in \mathbb{R} \setminus \{0\}$ is the widely used *logistic function*.

Plot 1.1.3: The logistic function f_{\log} with different values of $\beta_1, \beta_2, \beta_3$

```

1  /* logistic.sas */
2  TITLE1 'Plots of the Logistic Function';
3
4  /* Generate the data for different logistic functions */
5  DATA data1;
6    beta3=1;
7    DO beta1= 0.5, 1;
8      DO beta2=0.1, 1;
9        DO t=-10 TO 10 BY 0.5;
10         s=COMPRESS('(' || beta1 || ',' || beta2 || ',' || beta3 || ')');
11         f_log=beta3/(1+beta2*EXP(-beta1*t));
12         OUTPUT;
13       END;
14     END;
15   END;
16
17  /* Graphical Options */
18  SYMBOL1 C=GREEN V=NONE I=JOIN L=1;
19  SYMBOL2 C=GREEN V=NONE I=JOIN L=2;
20  SYMBOL3 C=GREEN V=NONE I=JOIN L=3;
21  SYMBOL4 C=GREEN V=NONE I=JOIN L=33;
22  AXIS1 LABEL=(H=2 'f' H=1 'log' H=2 '(t)');
23  AXIS2 LABEL=('t');
24  LEGEND1 LABEL=(F=CGREEK H=2 '(b' H=1 '1' H=2 ', b' H=1 '2' H=2 ',b' H
    ↪=1 '3' H=2 ')='));
25
26  /* Plot the functions */

```

```

27 PROC GPLOT DATA=data1;
28     PLOT f_log*t=s / VAXIS=AXIS1 HAXIS=AXIS2 LEGEND=LEGEND1;
29 RUN; QUIT;

```

A function is plotted by computing its values at numerous grid points and then joining them. The computation is done in the `DATA` step, where the data file `data1` is generated. It contains the values of `f_log`, computed at the grid $t = -10, -9.5, \dots, 10$ and indexed by the vector `s` of the different choices of parameters. This is done by nested `DO` loops. The operator `||` merges two strings and `COMPRESS` removes the empty space in the string. `OUTPUT` then stores the values of interest of `f_log`, `t` and `s` (and the other variables) in the data set `data1`.

The four functions are plotted by the `GPLOT` procedure by adding `=s` in the `PLOT` statement. This also automatically generates a legend, which is customized by the `LEGEND1` statement. Here the label is modified by using a greek font (`F=CGREEK`) and generating smaller letters of height 1 for the indices, while assuming a normal height of 2 (`H=1` and `H=2`). The last feature is also used in the axis statement. For each value of `s` SAS takes a new `SYMBOL` statement. They generate lines of different line types (`L=1, 2, 3, 33`).

We obviously have $\lim_{t \rightarrow \infty} f_{\log}(t) = \beta_3$, if $\beta_1 > 0$. The value β_3 often resembles the maximum impregnation or growth of a system. Note that

$$\begin{aligned}
 \frac{1}{f_{\log}(t)} &= \frac{1 + \beta_2 \exp(-\beta_1 t)}{\beta_3} \\
 &= \frac{1 - \exp(-\beta_1)}{\beta_3} + \exp(-\beta_1) \frac{1 + \beta_2 \exp(-\beta_1(t-1))}{\beta_3} \\
 &= \frac{1 - \exp(-\beta_1)}{\beta_3} + \exp(-\beta_1) \frac{1}{f_{\log}(t-1)} \\
 &= a + \frac{b}{f_{\log}(t-1)}.
 \end{aligned} \tag{1.6}$$

This means that there is a linear relationship among $1/f_{\log}(t)$. This can serve as a basis for estimating the parameters $\beta_1, \beta_2, \beta_3$ by an appropriate linear least squares approach, see Exercises 1.2 and 1.3. In the following example we fit the logistic trend model (1.5) to the population growth of the area of North Rhine-Westphalia (NRW), which is a federal state of Germany.

Example 1.1.2. (Population1 Data). Table 1.1.1 shows the population sizes y_t in millions of the area of North-Rhine-Westphalia in

5 years steps from 1935 to 1980 as well as their predicted values \hat{y}_t , obtained from a least squares estimation as described in (1.4) for a logistic model.

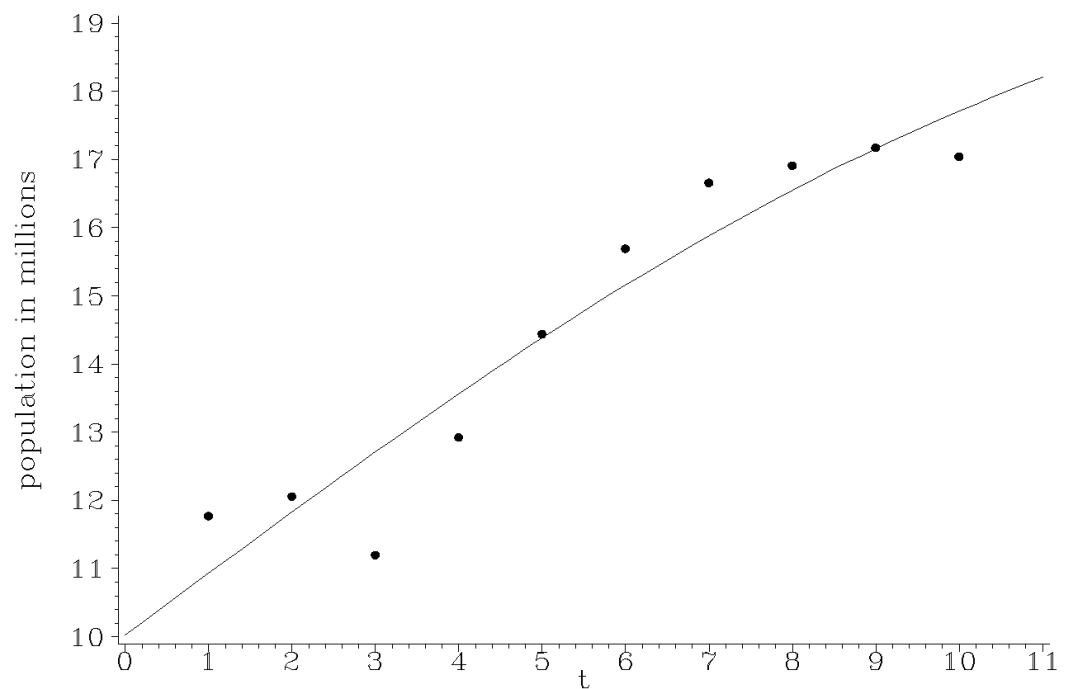
Year	t	Population sizes y_t (in millions)	Predicted values \hat{y}_t (in millions)
1935	1	11.772	10.930
1940	2	12.059	11.827
1945	3	11.200	12.709
1950	4	12.926	13.565
1955	5	14.442	14.384
1960	6	15.694	15.158
1965	7	16.661	15.881
1970	8	16.914	16.548
1975	9	17.176	17.158
1980	10	17.044	17.710

Table 1.1.1: Population1 Data

As a prediction of the population size at time t we obtain in the logistic model

$$\begin{aligned}\hat{y}_t &:= \frac{\hat{\beta}_3}{1 + \hat{\beta}_2 \exp(-\hat{\beta}_1 t)} \\ &= \frac{21.5016}{1 + 1.1436 \exp(-0.1675 t)}\end{aligned}$$

with the estimated saturation size $\hat{\beta}_3 = 21.5016$. The following plot shows the data and the fitted logistic curve.



Plot 1.1.4: NRW population sizes and fitted logistic function.

```

1  /* population1.sas */
2  TITLE1 'Population sizes and logistic fit';
3  TITLE2 'Population1 Data';
4
5  /* Read in the data */
6  DATA data1;
7  INFILE 'c:\data\population1.txt';
8  INPUT year t pop;
9
10 /* Compute parameters for fitted logistic function */
11 PROC NLIN DATA=data1 OUTEST=estimate;
12     MODEL pop=beta3/(1+beta2*EXP(-beta1*t));
13     PARAMETERS beta1=1 beta2=1 beta3=20;
14 RUN;
15
16 /* Generate fitted logistic function */
17 DATA data2;
18 SET estimate(WHERE=(_TYPE_='FINAL'));
19 DO t1=0 TO 11 BY 0.2;
20     f_log=beta3/(1+beta2*EXP(-beta1*t1));
21     OUTPUT;
22 END;
23
24 /* Merge data sets */
25 DATA data3;
26     MERGE data1 data2;
27

```

```

28 /* Graphical options */
29 AXIS1 LABEL=(ANGLE=90 'population in millions');
30 AXIS2 LABEL=('t');
31 SYMBOL1 V=DOT C=GREEN I=NONE;
32 SYMBOL2 V=NONE C=GREEN I=JOIN W=1;
33
34 /* Plot data with fitted function */
35 PROC GLOT DATA=data3;
36     PLOT pop*t=1 f_log*t1=2 / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2;
37 RUN; QUIT;

```

The procedure `NLIN` fits nonlinear regression models by least squares. The `OUTEST` option names the data set to contain the parameter estimates produced by `NLIN`. The `MODEL` statement defines the prediction equation by declaring the dependent variable and defining an expression that evaluates predicted values. A `PARAMETERS` statement must follow the `PROC NLIN` statement. Each `parameter=value` expression specifies the starting values of the pa-

rameter. Using the final estimates of `PROC NLIN` by the `SET` statement in combination with the `WHERE` data set option, the second data step generates the fitted logistic function values. The options in the `GLOT` statement cause the data points and the predicted function to be shown in one plot, after they were stored together in a new data set `data3` merging `data1` and `data2` with the `MERGE` statement.

The Mitscherlich Function

The *Mitscherlich function* is typically used for modelling the long term growth of a system:

$$f_M(t) := f_M(t; \beta_1, \beta_2, \beta_3) := \beta_1 + \beta_2 \exp(\beta_3 t), \quad t \geq 0, \quad (1.7)$$

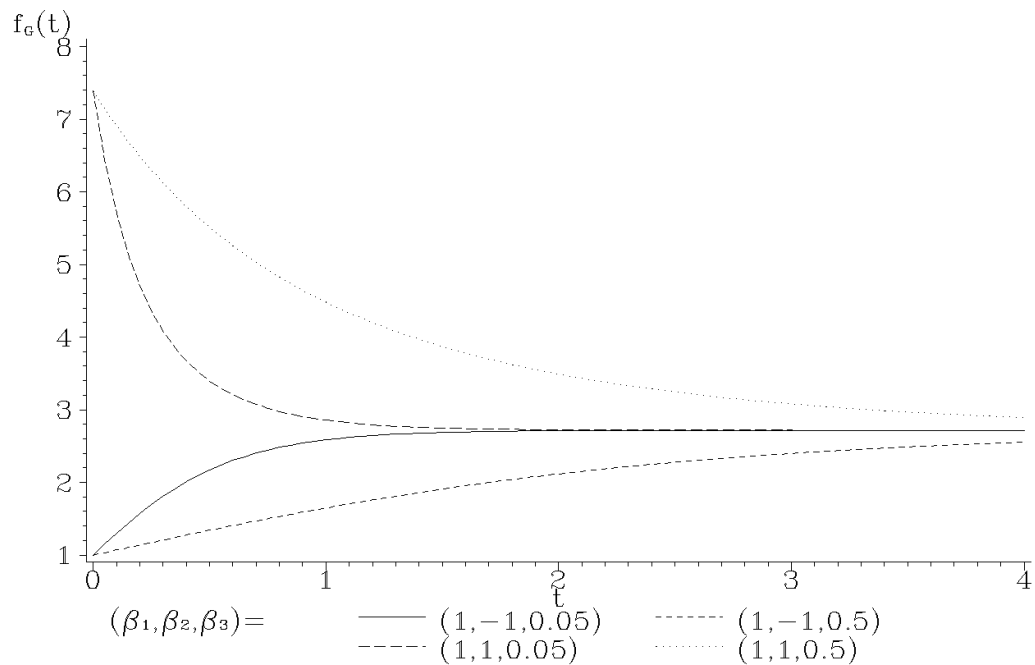
where $\beta_1, \beta_2 \in \mathbb{R}$ and $\beta_3 < 0$. Since β_3 is negative we have the asymptotic behavior $\lim_{t \rightarrow \infty} f_M(t) = \beta_1$ and thus the parameter β_1 is the saturation value of the system. The (initial) value of the system at the time $t = 0$ is $f_M(0) = \beta_1 + \beta_2$.

The Gompertz Curve

A further quite common function for modelling the increase or decrease of a system is the *Gompertz curve*

$$f_G(t) := f_G(t; \beta_1, \beta_2, \beta_3) := \exp(\beta_1 + \beta_2 \beta_3^t), \quad t \geq 0, \quad (1.8)$$

where $\beta_1, \beta_2 \in \mathbb{R}$ and $\beta_3 \in (0, 1)$.



Plot 1.1.5: Gompertz curves with different parameters.

```

1  /* gompertz.sas */
2  TITLE1 'Gompertz curves';
3
4  /* Generate the data for different Gompertz functions */
5  DATA data1;
6    beta1=1;
7    DO beta2=-1, 1;
8      DO beta3=0.05, 0.5;
9        DO t=0 TO 4 BY 0.05;
10         s=COMPRESS('(' || beta1 || ',' || beta2 || ',' || beta3 || ')');
11         f_g=EXP(beta1+beta2*beta3**t);
12         OUTPUT;
13       END;
14     END;
15   END;
16
17  /* Graphical Options */
18  SYMBOL1 C=GREEN V=NONE I=JOIN L=1;
19  SYMBOL2 C=GREEN V=NONE I=JOIN L=2;
20  SYMBOL3 C=GREEN V=NONE I=JOIN L=3;
21  SYMBOL4 C=GREEN V=NONE I=JOIN L=33;
22  AXIS1 LABEL=(H=2 'f' H=1 'G' H=2 '(t)');
23  AXIS2 LABEL=('t');

```

```

24 LEGEND1 LABEL=(F=CGREEK H=2 ' (b' H=1 '1' H=2 ',b' H=1 '2' H=2 ',b' H=1
    ↪ '3' H=2 ')=' );
25
26 /*Plot the functions */
27 PROC Gplot DATA=data1;
28     PLOT f_g*t=s / VAXIS=AXIS1 HAXIS=AXIS2 LEGEND=LEGEND1;
29 RUN; QUIT;

```

We obviously have

$$\log(f_G(t)) = \beta_1 + \beta_2 \beta_3^t = \beta_1 + \beta_2 \exp(\log(\beta_3)t),$$

and thus $\log(f_G)$ is a Mitscherlich function with parameters β_1, β_2 and $\log(\beta_3)$. The saturation size obviously is $\exp(\beta_1)$.

The Allometric Function

The *allometric function*

$$f_a(t) := f_a(t; \beta_1, \beta_2) = \beta_2 t^{\beta_1}, \quad t \geq 0, \quad (1.9)$$

with $\beta_1 \in \mathbb{R}$, $\beta_2 > 0$, is a common trend function in biometry and economics. It can be viewed as a particular *Cobb–Douglas function*, which is a popular econometric model to describe the output produced by a system depending on an input. Since

$$\log(f_a(t)) = \log(\beta_2) + \beta_1 \log(t), \quad t > 0,$$

is a linear function of $\log(t)$, with slope β_1 and intercept $\log(\beta_2)$, we can assume a linear regression model for the logarithmic data $\log(y_t)$

$$\log(y_t) = \log(\beta_2) + \beta_1 \log(t) + \varepsilon_t, \quad t \geq 1,$$

where ε_t are the error variables.

Example 1.1.3. (Income Data). Table 1.1.2 shows the (accumulated) annual average increases of gross and net incomes in thousands DM (deutsche mark) in Germany, starting in 1960.

Year	t	Gross income x_t	Net income y_t
1960	0	0	0
1961	1	0.627	0.486
1962	2	1.247	0.973
1963	3	1.702	1.323
1964	4	2.408	1.867
1965	5	3.188	2.568
1966	6	3.866	3.022
1967	7	4.201	3.259
1968	8	4.840	3.663
1969	9	5.855	4.321
1970	10	7.625	5.482

Table 1.1.2: Income Data.

We assume that the increase of the net income y_t is an allometric function of the time t and obtain

$$\log(y_t) = \log(\beta_2) + \beta_1 \log(t) + \varepsilon_t. \quad (1.10)$$

The least squares estimates of β_1 and $\log(\beta_2)$ in the above linear regression model are (see, for example Falk et al., 2002, Theorem 3.2.2)

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{10} (\log(t) - \overline{\log(t)}) (\log(y_t) - \overline{\log(y)})}{\sum_{t=1}^{10} (\log(t) - \overline{\log(t)})^2} = 1.019,$$

where $\overline{\log(t)} := \frac{1}{10} \sum_{t=1}^{10} \log(t) = 1.5104$, $\overline{\log(y)} := \frac{1}{10} \sum_{t=1}^{10} \log(y_t) = 0.7849$, and hence

$$\widehat{\log(\beta_2)} = \overline{\log(y)} - \hat{\beta}_1 \overline{\log(t)} = -0.7549$$

We estimate β_2 therefore by

$$\hat{\beta}_2 = \exp(-0.7549) = 0.4700.$$

The predicted value \hat{y}_t corresponds to the time t

$$\hat{y}_t = 0.47t^{1.019}. \quad (1.11)$$

t	$y_t - \hat{y}_t$
1	0.0159
2	0.0201
3	-0.1176
4	-0.0646
5	0.1430
6	0.1017
7	-0.1583
8	-0.2526
9	-0.0942
10	0.5662

Table 1.1.3: Residuals of Income Data.

Table 1.1.3 lists the residuals $y_t - \hat{y}_t$ by which one can judge the goodness of fit of the model (1.11).

A popular measure for assessing the fit is the *squared multiple correlation coefficient* or *R^2 -value*

$$R^2 := 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2} \quad (1.12)$$

where $\bar{y} := n^{-1} \sum_{t=1}^n y_t$ is the average of the observations y_t (cf Falk et al., 2002, Section 3.3). In the linear regression model with \hat{y}_t based on the least squares estimates of the parameters, R^2 is necessarily between zero and one with the implications $R^2 = 1$ iff¹ $\sum_{t=1}^n (y_t - \hat{y}_t)^2 = 0$ (see Exercise 1.4). A value of R^2 close to 1 is in favor of the fitted model. The model (1.10) has R^2 equal to 0.9934, whereas (1.11) has $R^2 = 0.9789$. Note, however, that the initial model (1.9) is not linear and $\hat{\beta}_2$ is not the least squares estimates, in which case R^2 is no longer necessarily between zero and one and has therefore to be viewed with care as a crude measure of fit.

The annual average gross income in 1960 was 6148 DM and the corresponding net income was 5178 DM. The actual average gross and net incomes were therefore $\tilde{x}_t := x_t + 6.148$ and $\tilde{y}_t := y_t + 5.178$ with

¹if and only if

the estimated model based on the above predicted values \hat{y}_t

$$\hat{\tilde{y}}_t = \hat{y}_t + 5.178 = 0.47t^{1.019} + 5.178.$$

Note that the residuals $\tilde{y}_t - \hat{\tilde{y}}_t = y_t - \hat{y}_t$ are not influenced by adding the constant 5.178 to y_t . The above models might help judging the average tax payer's situation between 1960 and 1970 and to predict his future one. It is apparent from the residuals in Table 1.1.3 that the net income y_t is an almost perfect multiple of t for t between 1 and 9, whereas the large increase y_{10} in 1970 seems to be an outlier. Actually, in 1969 the German government had changed and in 1970 a long strike in Germany caused an enormous increase in the income of civil servants.

1.2 Linear Filtering of Time Series

In the following we consider the additive model (1.1) and assume that there is no long term cyclic component. Nevertheless, we allow a trend, in which case the smooth nonrandom component G_t equals the trend function T_t . Our model is, therefore, the decomposition

$$Y_t = T_t + S_t + R_t, \quad t = 1, 2, \dots \quad (1.13)$$

with $E(R_t) = 0$. Given realizations y_t , $t = 1, 2, \dots, n$, of this time series, the aim of this section is the derivation of estimators \hat{T}_t, \hat{S}_t of the nonrandom functions T_t and S_t and to remove them from the time series by considering $y_t - \hat{T}_t$ or $y_t - \hat{S}_t$ instead. These series are referred to as the *trend* or *seasonally adjusted* time series. The data y_t are decomposed in smooth parts and irregular parts that fluctuate around zero.

Linear Filters

Let $a_{-r}, a_{-r+1}, \dots, a_s$ be arbitrary real numbers, where $r, s \geq 0$, $r + s + 1 \leq n$. The linear transformation

$$Y_t^* := \sum_{u=-r}^s a_u Y_{t-u}, \quad t = s + 1, \dots, n - r,$$

is referred to as a *linear filter* with *weights* a_{-r}, \dots, a_s . The Y_t are called *input* and the Y_t^* are called *output*.

Obviously, there are less output data than input data, if $(r, s) \neq (0, 0)$. A positive value $s > 0$ or $r > 0$ causes a truncation at the beginning or at the end of the time series; see Example 1.2.2 below. For convenience, we call the vector of weights $(a_u) = (a_{-r}, \dots, a_s)^T$ a (linear) filter.

A filter (a_u) , whose weights sum up to one, $\sum_{u=-r}^s a_u = 1$, is called *moving average*. The particular cases $a_u = 1/(2s+1)$, $u = -s, \dots, s$, with an odd number of equal weights, or $a_u = 1/(2s)$, $u = -s+1, \dots, s-1$, $a_{-s} = a_s = 1/(4s)$, aiming at an even number of weights, are *simple moving averages of order* $2s+1$ and $2s$, respectively.

Filtering a time series aims at smoothing the irregular part of a time series, thus detecting trends or seasonal components, which might otherwise be covered by fluctuations. While for example a *digital* speedometer in a car can provide its instantaneous velocity, thereby showing considerably large fluctuations, an *analog* instrument that comes with a hand and a built-in smoothing filter, reduces these fluctuations but takes a while to adjust. The latter instrument is much more comfortable to read and its information, reflecting a trend, is sufficient in most cases.

To compute the output of a simple moving average of order $2s+1$, the following obvious equation is useful:

$$Y_{t+1}^* = Y_t^* + \frac{1}{2s+1}(Y_{t+s+1} - Y_{t-s}).$$

This filter is a particular example of a *low-pass filter*, which preserves the slowly varying trend component of a series but removes from it the rapidly fluctuating or high frequency component. There is a trade-off between the two requirements that the irregular fluctuation should be reduced by a filter, thus leading, for example, to a large choice of s in a simple moving average, and that the long term variation in the data should not be distorted by oversmoothing, i.e., by a too large choice of s . If we assume, for example, a time series $Y_t = T_t + R_t$ without

seasonal component, a simple moving average of order $2s + 1$ leads to

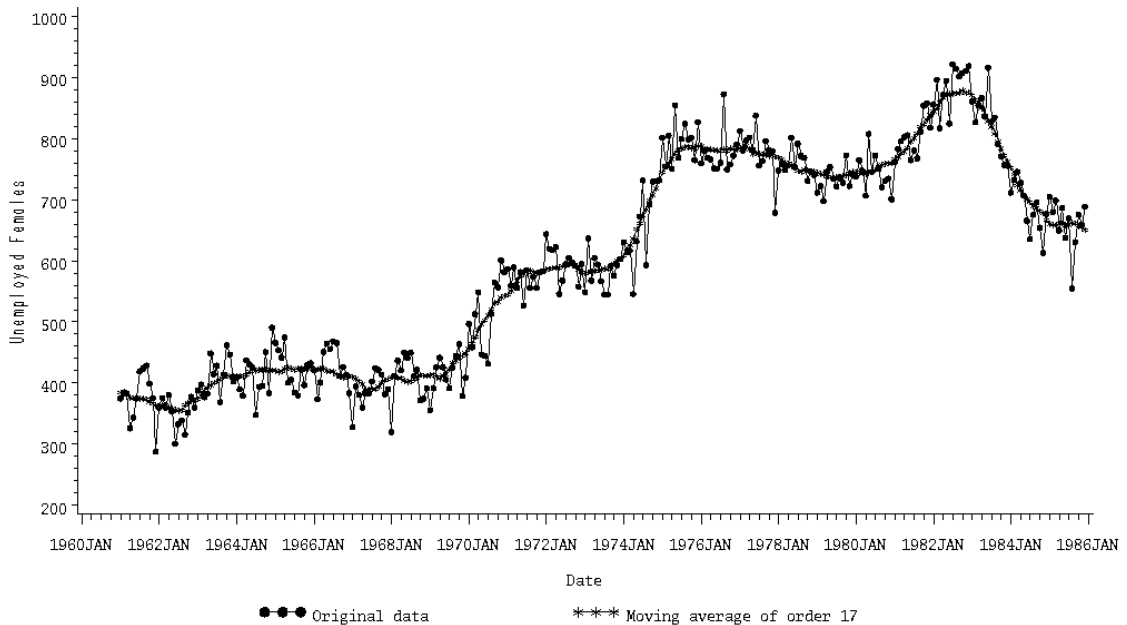
$$\begin{aligned} Y_t^* &= \frac{1}{2s+1} \sum_{u=-s}^s Y_{t-u} \\ &= \frac{1}{2s+1} \sum_{u=-s}^s T_{t-u} + \frac{1}{2s+1} \sum_{u=-s}^s R_{t-u} =: T_t^* + R_t^*, \end{aligned}$$

where by some law of large numbers argument

$$R_t^* \sim E(R_t) = 0,$$

if s is large. But T_t^* might then no longer reflect T_t . A small choice of s , however, has the effect that R_t^* is not yet close to its expectation.

Example 1.2.1. (Unemployed Females Data). The series of monthly unemployed females between ages 16 and 19 in the United States from January 1961 to December 1985 (in thousands) is smoothed by a simple moving average of order 17. The data together with their smoothed counterparts can be seen in Figure 1.2.1.



Plot 1.2.1: Unemployed young females in the US and the smoothed values (simple moving average of order 17).

```

1  /* females.sas */
2  TITLE1 'Simple Moving Average of Order 17';
3  TITLE2 'Unemployed Females Data';
4
5  /* Read in the data and generate SAS-formatted date */
6  DATA data1;
7      INFILE 'c:\data\female.txt';
8      INPUT upd @@;
9      date=INTNX('month','01jan61'd, _N_-1);
10     FORMAT date yymon.;
11
12  /* Compute the simple moving averages of order 17 */
13  PROC EXPAND DATA=data1 OUT=data2 METHOD=NONE;
14      ID date;
15      CONVERT upd=ma17 / TRANSFORM=(CMOVAVE 17);
16
17  /* Graphical options */
18  AXIS1 LABEL=(ANGLE=90 'Unemployed Females');
19  AXIS2 LABEL=('Date');
20  SYMBOL1 V=DOT C=GREEN I=JOIN H=.5 W=1;
21  SYMBOL2 V=STAR C=GREEN I=JOIN H=.5 W=1;
22  LEGEND1 LABEL=NONE VALUE=('Original data' 'Moving average of order
    ↪17');
23
24  /* Plot the data together with the simple moving average */
25  PROC GPLOT DATA=data2;
26      PLOT upd*date=1 ma17*date=2 / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2
    ↪LEGEND=LEGEND1;
27
28  RUN; QUIT;

```

In the data step the values for the variable `upd` are read from an external file. The option `@@` allows SAS to read the data over line break in the original txt-file.

By means of the function `INTNX`, a new variable in a date format is generated, containing monthly data starting from the 1st of January 1961. The temporarily created variable `_N_`, which counts the number of cases, is used to determine the distance from the starting value. The `FORMAT` statement attributes the format `yymon` to this variable, consisting of four digits for the year and three for the month.

The SAS procedure `EXPAND` computes simple moving averages and stores them in the file specified in the `OUT=` option. `EXPAND` is also able to interpolate series. For example if one has a quarterly series and wants to turn it into monthly data, this can be done by the method stated in the `METHOD=` option. Since we do not

wish to do this here, we choose `METHOD=NONE`. The `ID` variable specifies the time index, in our case the date, by which the observations are ordered. The `CONVERT` statement now computes the simple moving average. The syntax is `original=smoothed variable`. The smoothing method is given in the `TRANSFORM` option. `CMOVAVE number` specifies a simple moving average of order `number`. Remark that for the values at the boundary the arithmetic mean of the data within the moving window is computed as the simple moving average. This is an extension of our definition of a simple moving average. Also other smoothing methods can be specified in the `TRANSFORM` statement like the exponential smoother with smoothing parameter `alpha` (see page 33ff.) by `EWMA alpha`. The smoothed values are plotted together with the original series against the date in the final step.

Seasonal Adjustment

A simple moving average of a time series $Y_t = T_t + S_t + R_t$ now decomposes as

$$Y_t^* = T_t^* + S_t^* + R_t^*,$$

where S_t^* is the pertaining moving average of the seasonal components. Suppose, moreover, that S_t is a p -periodic function, i.e.,

$$S_t = S_{t+p}, \quad t = 1, \dots, n-p.$$

Take for instance monthly average temperatures Y_t measured at fixed points, in which case it is reasonable to assume a periodic seasonal component S_t with period $p = 12$ months. A simple moving average of order p then yields a *constant* value $S_t^* = S$, $t = p, p+1, \dots, n-p$. By adding this constant S to the trend function T_t and putting $T_t' := T_t + S$, we can assume in the following that $S = 0$. Thus we obtain for the differences

$$D_t := Y_t - Y_t^* \sim S_t + R_t.$$

To estimate S_t we average the differences with lag p (note that they vary around S_t) by

$$\begin{aligned} \bar{D}_t &:= \frac{1}{n_t} \sum_{j=0}^{n_t-1} D_{t+jp} \sim S_t, \quad t = 1, \dots, p, \\ \bar{D}_t &:= \bar{D}_{t-p} \quad \text{for } t > p, \end{aligned}$$

where n_t is the number of periods available for the computation of \bar{D}_t . Thus,

$$\hat{S}_t := \bar{D}_t - \frac{1}{p} \sum_{j=1}^p \bar{D}_j \sim S_t - \frac{1}{p} \sum_{j=1}^p S_j = S_t \quad (1.14)$$

is an estimator of $S_t = S_{t+p} = S_{t+2p} = \dots$ satisfying

$$\frac{1}{p} \sum_{j=0}^{p-1} \hat{S}_{t+j} = 0 = \frac{1}{p} \sum_{j=0}^{p-1} S_{t+j}.$$

The differences $Y_t - \hat{S}_t$ with a seasonal component close to zero are then the seasonally adjusted time series.

Example 1.2.2. For the 51 Unemployed1 Data in Example 1.1.1 it is obviously reasonable to assume a periodic seasonal component with $p = 12$ months. A simple moving average of order 12

$$Y_t^* = \frac{1}{12} \left(\frac{1}{2} Y_{t-6} + \sum_{u=-5}^5 Y_{t-u} + \frac{1}{2} Y_{t+6} \right), \quad t = 7, \dots, 45,$$

then has a constant seasonal component, which we assume to be zero by adding this constant to the trend function. Table 1.2.1 contains the values of D_t , \bar{D}_t and the estimates \hat{S}_t of S_t .

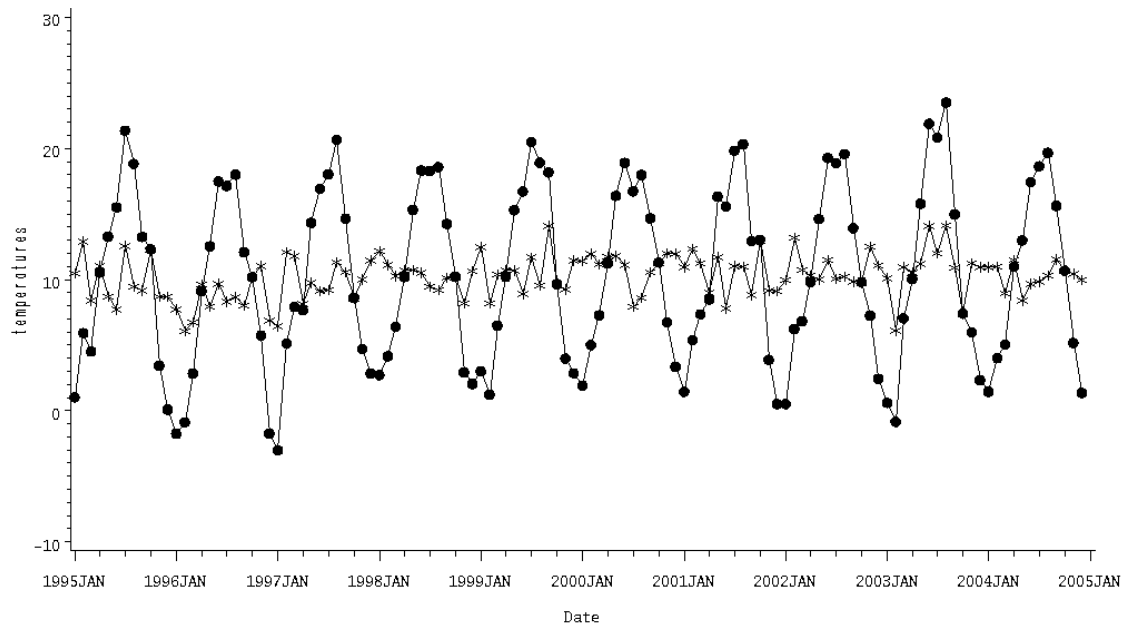
Month	d_t (rounded values)				\bar{d}_t (rounded)	\hat{s}_t (rounded)
	1976	1977	1978	1979		
January	53201	56974	48469	52611	52814	53136
February	59929	54934	54102	51727	55173	55495
March	24768	17320	25678	10808	19643	19966
April	-3848	42	-5429	—	-3079	-2756
May	-19300	-11680	-14189	—	-15056	-14734
June	-23455	-17516	-20116	—	-20362	-20040
July	-26413	-21058	-20605	—	-22692	-22370
August	-27225	-22670	-20393	—	-23429	-23107
September	-27358	-24646	-20478	—	-24161	-23839
October	-23967	-21397	-17440	—	-20935	-20612
November	-14300	-10846	-11889	—	-12345	-12023
December	11540	12213	7923	—	10559	10881

Table 1.2.1: Table of d_t , \bar{d}_t and of estimates \hat{s}_t of the seasonal component S_t in the Unemployed1 Data.

We obtain for these data

$$\hat{s}_t = \bar{d}_t - \frac{1}{12} \sum_{j=1}^{12} \bar{d}_j = \bar{d}_t + \frac{3867}{12} = \bar{d}_t + 322.25.$$

Example 1.2.3. (Temperatures Data). The monthly average temperatures near Würzburg, Germany were recorded from the 1st of January 1995 to the 31st of December 2004. The data together with their seasonally adjusted counterparts can be seen in Figure 1.2.2.



Plot 1.2.2: Monthly average temperatures near Würzburg and seasonally adjusted values.

```

1  /* temperatures.sas */
2  TITLE1 'Original and seasonally adjusted data';
3  TITLE2 'Temperatures data';
4
5  /* Read in the data and generate SAS-formatted date */
6  DATA temperatures;
7      INFILE 'c:\data\temperatures.txt';
8      INPUT temperature;
9      date=INTNX('month','01jan95'd,_N_-1);
10     FORMAT date yymon.;
11
12  /* Make seasonal adjustment */
13  PROC TIMESERIES DATA=temperatures OUT=series SEASONALITY=12 OUTDECOMP=
14     ↪deseason;
15     VAR temperature;
16     DECOMP /MODE=ADD;
17
18  /* Merge necessary data for plot */
19  DATA plotseries;
20     MERGE temperatures deseason(KEEP=SA);
21
22  /* Graphical options */
23  AXIS1 LABEL=(ANGLE=90 'temperatures');
24  AXIS2 LABEL=('Date');

```

```

24 SYMBOL1 V=DOT C=GREEN I=JOIN H=1 W=1;
25 SYMBOL2 V=STAR C=GREEN I=JOIN H=1 W=1;
26
27 /* Plot data and seasonally adjusted series */
28 PROC GPLOT data=plotseries;
29     PLOT temperature*date=1 SA*date=2 /OVERLAY VAXIS=AXIS1 HAXIS=AXIS2;
30
31 RUN; QUIT;

```

In the data step the values for the variable `temperature` are read from an external file. By means of the function `INTNX`, a date variable is generated, see Program 1.2.1 (females.sas).

The SAS procedure `TIMESERIES` together with the statement `DECOMP` computes a seasonally adjusted series, which is stored in the file after the `OUTDECOMP` option. With `MODE=ADD` an additive model of the time series is assumed. The

default is a multiplicative model. The original series together with an automated time variable (just a counter) is stored in the file specified in the `OUT` option. In the option `SEASONALITY` the underlying period is specified. Depending on the data it can be any natural number.

The seasonally adjusted values can be referenced by `SA` and are plotted together with the original series against the date in the final step.

The Census X-11 Program

In the fifties of the 20th century the U.S. Bureau of the Census has developed a program for seasonal adjustment of economic time series, called the *Census X-11 Program*. It is based on monthly observations and assumes an additive model

$$Y_t = T_t + S_t + R_t$$

as in (1.13) with a seasonal component S_t of period $p = 12$. We give a brief summary of this program following Wallis (1974), which results in a moving average with symmetric weights. The census procedure is discussed in Shiskin and Eisenpress (1957); a complete description is given by Shiskin et al. (1967). A theoretical justification based on stochastic models is provided by Cleveland and Tiao (1976)).

The X-11 Program essentially works as the seasonal adjustment described above, but it adds iterations and various moving averages. The different steps of this program are

- (i) Compute a simple moving average Y_t^* of order 12 to leave essentially a trend $Y_t^* \sim T_t$.

- (ii) The difference

$$D_t := Y_t - Y_t^* \sim S_t + R_t$$

then leaves approximately the seasonal plus irregular component.

- (iii) Apply a moving average of order 5 to *each month* separately by computing

$$\bar{D}_t^{(1)} := \frac{1}{9} \left(D_{t-24}^{(1)} + 2D_{t-12}^{(1)} + 3D_t^{(1)} + 2D_{t+12}^{(1)} + D_{t+24}^{(1)} \right) \sim S_t,$$

which gives an estimate of the seasonal component S_t . Note that the moving average with weights $(1, 2, 3, 2, 1)/9$ is a simple moving average of length 3 of simple moving averages of length 3.

- (iv) The $\bar{D}_t^{(1)}$ are adjusted to approximately sum up to 0 over any 12-months period by putting

$$\hat{S}_t^{(1)} := \bar{D}_t^{(1)} - \frac{1}{12} \left(\frac{1}{2} \bar{D}_{t-6}^{(1)} + \bar{D}_{t-5}^{(1)} + \cdots + \bar{D}_{t+5}^{(1)} + \frac{1}{2} \bar{D}_{t+6}^{(1)} \right).$$

- (v) The differences

$$Y_t^{(1)} := Y_t - \hat{S}_t^{(1)} \sim T_t + R_t$$

then are the preliminary seasonally adjusted series, quite in the manner as before.

- (vi) The adjusted data $Y_t^{(1)}$ are further smoothed by a *Henderson* moving average Y_t^{**} of order 9, 13, or 23.

- (vii) The differences

$$D_t^{(2)} := Y_t - Y_t^{**} \sim S_t + R_t$$

then leave a second estimate of the sum of the seasonal and irregular components.

(viii) A moving average of order 7 is applied to each month separately

$$\bar{D}_t^{(2)} := \sum_{u=-3}^3 a_u D_{t-12u}^{(2)},$$

where the weights a_u come from a simple moving average of order 3 applied to a simple moving average of order 5 of the original data, i.e., the vector of weights is $(1, 2, 3, 3, 3, 2, 1)/15$. This gives a second estimate of the seasonal component S_t .

(ix) Step (4) is repeated yielding approximately centered estimates $\hat{S}_t^{(2)}$ of the seasonal components.

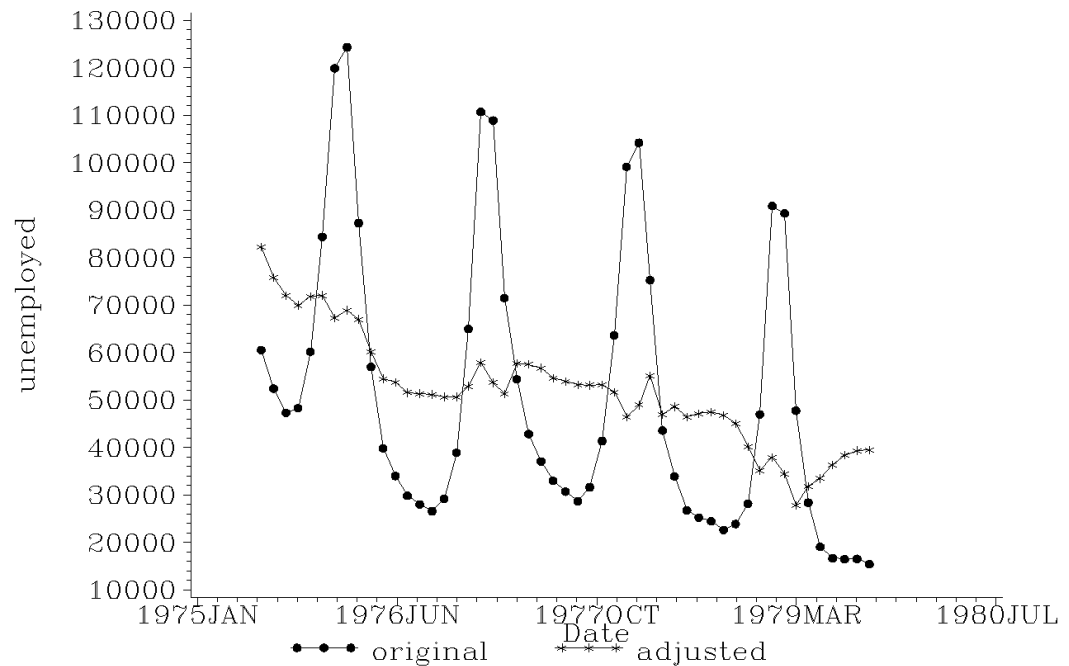
(x) The differences

$$Y_t^{(2)} := Y_t - \hat{S}_t^{(2)}$$

then finally give the seasonally adjusted series.

Depending on the length of the Henderson moving average used in step (6), $Y_t^{(2)}$ is a moving average of length 165, 169 or 179 of the original data (see Exercise 1.10). Observe that this leads to averages at time t of the past and future seven years, roughly, where seven years is a typical length of business cycles observed in economics (Juglar cycle). The U.S. Bureau of Census has recently released an extended version of the X-11 Program called *Census X-12-ARIMA*. It is implemented in SAS version 8.1 and higher as PROC X12; we refer to the SAS online documentation for details.

We will see in Example 5.2.4 on page 171 that linear filters may cause unexpected effects and so, it is not clear a priori how the seasonal adjustment filter described above behaves. Moreover, end-corrections are necessary, which cover the important problem of adjusting current observations. This can be done by some extrapolation.



Plot 1.2.3: Plot of the Unemployed1 Data y_t and of $y_t^{(2)}$, seasonally adjusted by the X-11 procedure.

```

1  /* unemployed1_x11.sas */
2  TITLE1 'Original and X-11 seasonal adjusted data';
3  TITLE2 'Unemployed1 Data';
4
5  /* Read in the data and generated SAS-formatted date */
6  DATA data1;
7  INFILE 'c:\data\unemployed1.txt';
8  INPUT month $ t upd;
9  date=INTNX('month','01jul75'd, _N_-1);
10 FORMAT date yymon.;
11
12 /* Apply X-11-Program */
13 PROC X11 DATA=data1;
14     MONTHLY DATE=date ADDITIVE;
15     VAR upd;
16     OUTPUT OUT=data2 B1=upd D11=updx11;
17
18 /* Graphical options */
19 AXIS1 LABEL=(ANGLE=90 'unemployed');
20 AXIS2 LABEL=('Date') ;
21 SYMBOL1 V=DOT C=GREEN I=JOIN H=1 W=1;
22 SYMBOL2 V=STAR C=GREEN I=JOIN H=1 W=1;
23 LEGEND1 LABEL=NONE VALUE=('original' 'adjusted');
24
25 /* Plot data and adjusted data */

```

```

26 PROC GPLOT DATA=data2;
27   PLOT upd*date=1 updx11*date=2
28     / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2 LEGEND=LEGEND1;
29 RUN; QUIT;

```

In the data step values for the variables `month`, `t` and `upd` are read from an external file, where `month` is defined as a character variable by the succeeding `$` in the `INPUT` statement. By means of the function `INTNX`, a date variable is generated, see Program 1.2.1 (females.sas). The SAS procedure `X11` applies the Census X-11 Program to the data. The `MONTHLY` statement selects an algorithm for monthly data, `DATE` defines the date variable and `ADDITIVE` selects an additive model (default: multiplicative model). The results for this analysis for the

variable `upd` (unemployed) are stored in a data set named `data2`, containing the original data in the variable `upd` and the final results of the X-11 Program in `updx11`.

The last part of this SAS program consists of statements for generating the plot. Two `AXIS` and two `SYMBOL` statements are used to customize the graphic containing two plots, the original data and the by `X11` seasonally adjusted data. A `LEGEND` statement defines the text that explains the symbols.

Best Local Polynomial Fit

A simple moving average works well for a locally almost linear time series, but it may have problems to reflect a more twisted shape. This suggests fitting higher order local polynomials. Consider $2k + 1$ consecutive data $y_{t-k}, \dots, y_t, \dots, y_{t+k}$ from a time series. A *local polynomial estimator* of order $p < 2k + 1$ is the minimizer β_0, \dots, β_p satisfying

$$\sum_{u=-k}^k (y_{t+u} - \beta_0 - \beta_1 u - \dots - \beta_p u^p)^2 = \min. \quad (1.15)$$

If we differentiate the left hand side with respect to each β_j and set the derivatives equal to zero, we see that the minimizers satisfy the $p + 1$ linear equations

$$\beta_0 \sum_{u=-k}^k u^j + \beta_1 \sum_{u=-k}^k u^{j+1} + \dots + \beta_p \sum_{u=-k}^k u^{j+p} = \sum_{u=-k}^k u^j y_{t+u}$$

for $j = 0, \dots, p$. These $p + 1$ equations, which are called *normal equations*, can be written in matrix form as

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y} \quad (1.16)$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & -k & (-k)^2 & \dots & (-k)^p \\ 1 & -k+1 & (-k+1)^2 & \dots & (-k+1)^p \\ \vdots & & & \ddots & \vdots \\ 1 & k & k^2 & \dots & k^p \end{pmatrix} \quad (1.17)$$

is the *design matrix*, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ and $\mathbf{y} = (y_{t-k}, \dots, y_{t+k})^T$. The rank of $\mathbf{X}^T \mathbf{X}$ equals that of \mathbf{X} , since their null spaces coincide (Exercise 1.12). Thus, the matrix $\mathbf{X}^T \mathbf{X}$ is invertible iff the columns of \mathbf{X} are linearly independent. But this is an immediate consequence of the fact that a polynomial of degree p has at most p different roots (Exercise 1.13). The normal equations (1.16) have, therefore, the unique solution

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (1.18)$$

The linear prediction of y_{t+u} , based on u, u^2, \dots, u^p , is

$$\hat{y}_{t+u} = (1, u, \dots, u^p) \boldsymbol{\beta} = \sum_{j=0}^p \beta_j u^j.$$

Choosing $u = 0$ we obtain in particular that $\beta_0 = \hat{y}_t$ is a predictor of the central observation y_t among y_{t-k}, \dots, y_{t+k} . The local polynomial approach consists now in replacing y_t by the intercept β_0 .

Though it seems as if this local polynomial fit requires a great deal of computational effort by calculating β_0 for each y_t , it turns out that it is actually a moving average. First observe that we can write by (1.18)

$$\beta_0 = \sum_{u=-k}^k c_u y_{t+u}$$

with some $c_u \in \mathbb{R}$ which *do not depend* on the values y_u of the time series and hence, (c_u) is a linear filter. Next we show that the c_u sum up to 1. Choose to this end $y_{t+u} = 1$ for $u = -k, \dots, k$. Then $\beta_0 = 1$, $\beta_1 = \dots = \beta_p = 0$ is an obvious solution of the minimization problem (1.15). Since this solution is unique, we obtain

$$1 = \beta_0 = \sum_{u=-k}^k c_u$$

and thus, (c_u) is a moving average. As can be seen in Exercise 1.14 it actually has symmetric weights. We summarize our considerations in the following result.

Theorem 1.2.4. *Fitting locally by least squares a polynomial of degree p to $2k+1 > p$ consecutive data points y_{t-k}, \dots, y_{t+k} and predicting y_t by the resulting intercept β_0 , leads to a moving average (c_u) of order $2k+1$, given by the first row of the matrix $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.*

Example 1.2.5. Fitting locally a polynomial of degree 2 to five consecutive data points leads to the moving average (Exercise 1.14)

$$(c_u) = \frac{1}{35}(-3, 12, 17, 12, -3)^T.$$

An extensive discussion of local polynomial fit is in Kendall and Ord (1993, Sections 3.2-3.13). For a book-length treatment of local polynomial estimation we refer to Fan and Gijbels (1996). An outline of various aspects such as the choice of the degree of the polynomial and further background material is given in Simonoff (1996, Section 5.2).

Difference Filter

We have already seen that we can remove a periodic seasonal component from a time series by utilizing an appropriate linear filter. We will next show that also a polynomial trend function can be removed by a suitable linear filter.

Lemma 1.2.6. *For a polynomial $f(t) := c_0 + c_1 t + \dots + c_p t^p$ of degree p , the difference*

$$\Delta f(t) := f(t) - f(t-1)$$

is a polynomial of degree at most $p-1$.

Proof. The assertion is an immediate consequence of the binomial expansion

$$(t-1)^p = \sum_{k=0}^p \binom{p}{k} t^k (-1)^{p-k} = t^p - p t^{p-1} + \dots + (-1)^p.$$

□

The preceding lemma shows that differencing reduces the degree of a polynomial. Hence,

$$\Delta^2 f(t) := \Delta f(t) - \Delta f(t-1) = \Delta(\Delta f(t))$$

is a polynomial of degree not greater than $p-2$, and

$$\Delta^q f(t) := \Delta(\Delta^{q-1} f(t)), \quad 1 \leq q \leq p,$$

is a polynomial of degree at most $p-q$. The function $\Delta^p f(t)$ is therefore a constant. The linear filter

$$\Delta Y_t = Y_t - Y_{t-1}$$

with weights $a_0 = 1, a_1 = -1$ is the first order *difference filter*. The recursively defined filter

$$\Delta^p Y_t = \Delta(\Delta^{p-1} Y_t), \quad t = p, \dots, n,$$

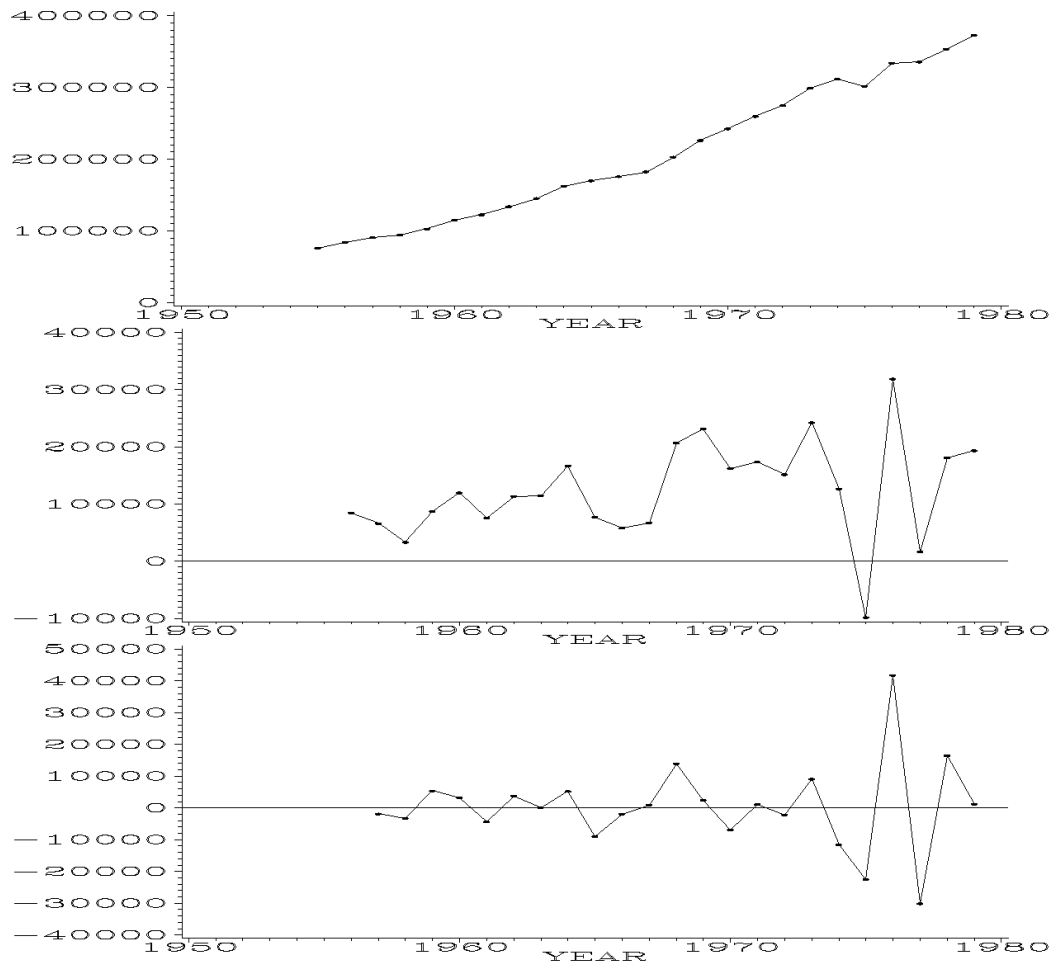
is the difference filter of order p .

The difference filter of second order has, for example, weights $a_0 = 1, a_1 = -2, a_2 = 1$

$$\begin{aligned} \Delta^2 Y_t &= \Delta Y_t - \Delta Y_{t-1} \\ &= Y_t - Y_{t-1} - Y_{t-1} + Y_{t-2} = Y_t - 2Y_{t-1} + Y_{t-2}. \end{aligned}$$

If a time series Y_t has a polynomial trend $T_t = \sum_{k=0}^p c_k t^k$ for some constants c_k , then the difference filter $\Delta^p Y_t$ of order p removes this trend up to a constant. Time series in economics often have a trend function that can be removed by a first or second order difference filter.

Example 1.2.7. (Electricity Data). The following plots show the total annual output of electricity production in Germany between 1955 and 1979 in millions of kilowatt-hours as well as their first and second order differences. While the original data show an increasing trend, the second order differences fluctuate around zero having no more trend, but there is now an increasing variability visible in the data.



Plot 1.2.4: Annual electricity output, first and second order differences.

```

1  /* electricity_differences.sas */
2  TITLE1 'First and second order differences';
3  TITLE2 'Electricity Data';
4  /* Note that this program requires the macro mkfields.sas to be
   ↪submitted before this program */
5
6  /* Read in the data, compute moving average of length as 12
7   as well as first and second order differences */
8  DATA data1(KEEP=year sum delta1 delta2);
9      INFILE 'c:\data\electric.txt';
10     INPUT year t jan feb mar apr may jun jul aug sep oct nov dec;
11     sum=jan+feb+mar+apr+may+jun+jul+aug+sep+oct+nov+dec;
12     delta1=DIF(sum);

```

```

13   delta2=DIF(delta1);
14
15   /* Graphical options */
16   AXIS1 LABEL=NONE;
17   SYMBOL1 V=DOT C=GREEN I=JOIN H=0.5 W=1;
18
19   /* Generate three plots */
20   GOPTIONS NODISPLAY;
21   PROC Gplot DATA=data1 GOUT=fig;
22       PLOT sum*year / VAXIS=AXIS1 HAXIS=AXIS2;
23       PLOT delta1*year / VAXIS=AXIS1 VREF=0;
24       PLOT delta2*year / VAXIS=AXIS1 VREF=0;
25   RUN;
26
27   /* Display them in one output */
28   GOPTIONS DISPLAY;
29   PROC GREPLAY NOFS IGOUT=fig TC=SASHELP.TEMPLT;
30       TEMPLATE=V3;
31       TREPLAY 1:Gplot 2:Gplot1 3:Gplot2;
32   RUN; DELETE _ALL_; QUIT;

```

In the first data step, the raw data are read from a file. Because the electric production is stored in different variables for each month of a year, the sum must be evaluated to get the annual output. Using the `DIF` function, the resulting variables `delta1` and `delta2` contain the first and second order differences of the original annual sums.

To display the three plots of `sum`, `delta1` and `delta2` against the variable `year` within one graphic, they are first plotted using the procedure `Gplot`. Here the option `GOUT=fig` stores the plots in a graphics catalog named `fig`, while `GOPTIONS NODISPLAY` causes no output of this procedure. After changing the `GOPTIONS` back to `DISPLAY`, the procedure `GREPLAY` is invoked. The option `NOFS` (no full-screen) suppresses the opening of a `GREPLAY`

window. The subsequent two line mode statements are read instead. The option `IGOUT` determines the input graphics catalog, while `TC=SASHELP.TEMPLT` causes SAS to take the standard template catalog. The `TEMPLATE` statement selects a template from this catalog, which puts three graphics one below the other. The `TREPLAY` statement connects the defined areas and the plots of the the graphics catalog. `Gplot`, `Gplot1` and `Gplot2` are the graphical outputs in the chronological order of the `Gplot` procedure. The `DELETE` statement after `RUN` deletes all entries in the input graphics catalog.

Note that SAS by default prints borders, in order to separate the different plots. Here these border lines are suppressed by defining `WHITE` as the border color.

For a time series $Y_t = T_t + S_t + R_t$ with a periodic seasonal component $S_t = S_{t+p} = S_{t+2p} = \dots$ the difference

$$Y_t^* := Y_t - Y_{t-p}$$

obviously removes the seasonal component. An additional differencing of proper length can moreover remove a polynomial trend, too. Note that the order of seasonal and trend adjusting makes no difference.

Exponential Smoother

Let Y_0, \dots, Y_n be a time series and let $\alpha \in [0, 1]$ be a constant. The linear filter

$$Y_t^* = \alpha Y_t + (1 - \alpha)Y_{t-1}^*, \quad t \geq 1,$$

with $Y_0^* = Y_0$ is called *exponential smoother*.

Lemma 1.2.8. *For an exponential smoother with constant $\alpha \in [0, 1]$ we have*

$$Y_t^* = \alpha \sum_{j=0}^{t-1} (1 - \alpha)^j Y_{t-j} + (1 - \alpha)^t Y_0, \quad t = 1, 2, \dots, n.$$

Proof. The assertion follows from induction. We have for $t = 1$ by definition $Y_1^* = \alpha Y_1 + (1 - \alpha)Y_0$. If the assertion holds for t , we obtain for $t + 1$

$$\begin{aligned} Y_{t+1}^* &= \alpha Y_{t+1} + (1 - \alpha)Y_t^* \\ &= \alpha Y_{t+1} + (1 - \alpha) \left(\alpha \sum_{j=0}^{t-1} (1 - \alpha)^j Y_{t-j} + (1 - \alpha)^t Y_0 \right) \\ &= \alpha \sum_{j=0}^t (1 - \alpha)^j Y_{t+1-j} + (1 - \alpha)^{t+1} Y_0. \end{aligned}$$

□

The parameter α determines the smoothness of the filtered time series. A value of α close to 1 puts most of the weight on the actual observation Y_t , resulting in a highly fluctuating series Y_t^* . On the other hand, an α close to 0 reduces the influence of Y_t and puts most of the weight to the past observations, yielding a smooth series Y_t^* . An exponential smoother is typically used for monitoring a system. Take, for example, a car having an analog speedometer with a hand. It is more convenient for the driver if the movements of this hand are smoothed, which can be achieved by α close to zero. But this, on the other hand, has the effect that an essential alteration of the speed can be read from the speedometer only with a certain delay.

Corollary 1.2.9. (i) Suppose that the random variables Y_0, \dots, Y_n have common expectation μ and common variance $\sigma^2 > 0$. Then we have for the exponentially smoothed variables with smoothing parameter $\alpha \in (0, 1)$

$$\begin{aligned} E(Y_t^*) &= \alpha \sum_{j=0}^{t-1} (1-\alpha)^j \mu + \mu(1-\alpha)^t \\ &= \mu(1 - (1-\alpha)^t) + \mu(1-\alpha)^t = \mu. \end{aligned} \quad (1.19)$$

If the Y_t are in addition uncorrelated, then

$$\begin{aligned} E((Y_t^* - \mu)^2) &= \alpha^2 \sum_{j=0}^{t-1} (1-\alpha)^{2j} \sigma^2 + (1-\alpha)^{2t} \sigma^2 \\ &= \sigma^2 \alpha^2 \frac{1 - (1-\alpha)^{2t}}{1 - (1-\alpha)^2} + (1-\alpha)^{2t} \sigma^2 \\ &\longrightarrow_{t \rightarrow \infty} \frac{\sigma^2 \alpha}{2 - \alpha} < \sigma^2. \end{aligned} \quad (1.20)$$

(ii) Suppose that the random variables Y_0, Y_1, \dots satisfy $E(Y_t) = \mu$ for $0 \leq t \leq N-1$, and $E(Y_t) = \lambda$ for $t \geq N$. Then we have for $t \geq N$

$$\begin{aligned} E(Y_t^*) &= \alpha \sum_{j=0}^{t-N} (1-\alpha)^j \lambda + \alpha \sum_{j=t-N+1}^{t-1} (1-\alpha)^j \mu + (1-\alpha)^t \mu \\ &= \lambda(1 - (1-\alpha)^{t-N+1}) + \\ &\quad \mu \left((1-\alpha)^{t-N+1} (1 - (1-\alpha)^{N-1}) + (1-\alpha)^t \right) \\ &\longrightarrow_{t \rightarrow \infty} \lambda. \end{aligned} \quad (1.21)$$

The preceding result quantifies the influence of the parameter α on the expectation and on the variance i.e., the smoothness of the filtered series Y_t^* , where we assume for the sake of a simple computation of the variance that the Y_t are uncorrelated. If the variables Y_t have common expectation μ , then this expectation carries over to Y_t^* . After a *change point* N , where the expectation of Y_t changes for $t \geq N$ from μ to $\lambda \neq \mu$, the filtered variables Y_t^* are, however, biased. This bias,

which will vanish as t increases, is due to the still inherent influence of past observations Y_t , $t < N$. The influence of these variables on the current expectation can be reduced by switching to a larger value of α . The price for the gain in correctness of the expectation is, however, a higher variability of Y_t^* (see Exercise 1.17).

An exponential smoother is often also used to make forecasts, explicitly by predicting Y_{t+1} through Y_t^* . The forecast error $Y_{t+1} - Y_t^* =: e_{t+1}$ then satisfies the equation $Y_{t+1}^* = \alpha e_{t+1} + Y_t^*$.

Also a motivation of the exponential smoother via a least squares approach is possible, see Exercise 1.18.

1.3 Autocovariances and Autocorrelations

Autocovariances and autocorrelations are measures of dependence between variables in a time series. Suppose that Y_1, \dots, Y_n are square integrable random variables with the property that the covariance $\text{Cov}(Y_{t+k}, Y_t) = E((Y_{t+k} - E(Y_{t+k}))(Y_t - E(Y_t)))$ of observations with lag k does not depend on t . Then

$$\gamma(k) := \text{Cov}(Y_{k+1}, Y_1) = \text{Cov}(Y_{k+2}, Y_2) = \dots$$

is called *autocovariance function* and

$$\rho(k) := \frac{\gamma(k)}{\gamma(0)}, \quad k = 0, 1, \dots$$

is called *autocorrelation function*.

Let y_1, \dots, y_n be realizations of a time series Y_1, \dots, Y_n . The empirical counterpart of the autocovariance function is

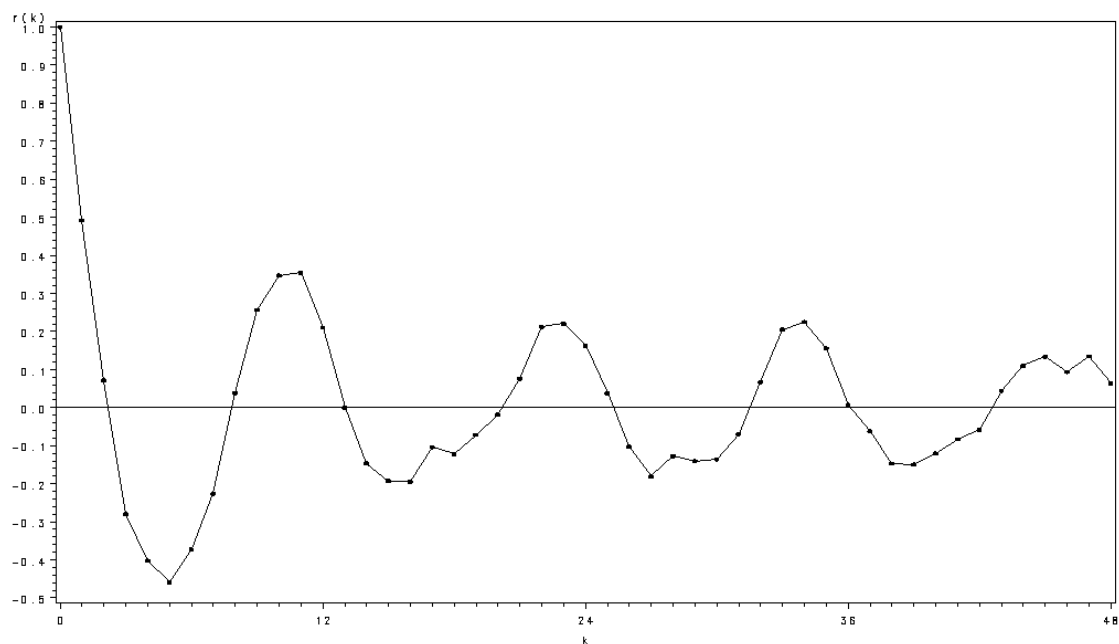
$$c(k) := \frac{1}{n} \sum_{t=1}^{n-k} (y_{t+k} - \bar{y})(y_t - \bar{y}) \quad \text{with } \bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$$

and the empirical autocorrelation is defined by

$$r(k) := \frac{c(k)}{c(0)} = \frac{\sum_{t=1}^{n-k} (y_{t+k} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}.$$

See Exercise 2.9 (ii) for the particular role of the factor $1/n$ in place of $1/(n-k)$ in the definition of $c(k)$. The graph of the function

$r(k)$, $k = 0, 1, \dots, n - 1$, is called *correlogram*. It is based on the assumption of equal expectations and should, therefore, be used for a trend adjusted series. The following plot is the correlogram of the first order differences of the Sunspot Data. The description can be found on page 207. It shows high and decreasing correlations at regular intervals.



Plot 1.3.1: Correlogram of the first order differences of the Sunspot Data.

```

1  /* sunspot_correlogram */
2  TITLE1 'Correlogram of first order differences';
3  TITLE2 'Sunspot Data';
4
5  /* Read in the data, generate year of observation and
6     compute first order differences */
7  DATA data1;
8     INFILE 'c:\data\sunspot.txt';
9     INPUT spot @@;
10    date=1748+_N_;
11    diff1=DIF(spot);
12
13  /* Compute autocorrelation function */
14  PROC ARIMA DATA=data1;
15    IDENTIFY VAR=diff1 NLAG=49 OUTCOV=corr NOPRINT;

```

```

16
17 /* Graphical options */
18 AXIS1 LABEL=(' r(k) ');
19 AXIS2 LABEL=(' k' ) ORDER=(0 12 24 36 48) MINOR=(N=11);
20 SYMBOL1 V=DOT C=GREEN I=JOIN H=0.5 W=1;
21
22 /* Plot autocorrelation function */
23 PROC GPLOT DATA=corr;
24     PLOT CORR*LAG / VAXIS=AXIS1 HAXIS=AXIS2 VREF=0;
25 RUN; QUIT;

```

In the data step, the raw data are read into the variable `spot`. The specification `@@` suppresses the automatic line feed of the `INPUT` statement after every entry in each row, see also Program 1.2.1 (females.txt). The variable `date` and the first order differences of the variable of interest `spot` are calculated.

The following procedure `ARIMA` is a crucial one in time series analysis. Here we just need the autocorrelation of `delta`, which will be calculated up to a lag of 49 (`NLAG=49`) by the `IDENTIFY` statement. The option

`OUTCOV=corr` causes SAS to create a data set `corr` containing among others the variables `LAG` and `CORR`. These two are used in the following `GPLOT` procedure to obtain a plot of the autocorrelation function. The `ORDER` option in the `AXIS2` statement specifies the values to appear on the horizontal axis as well as their order, and the `MINOR` option determines the number of minor tick marks between two major ticks. `VREF=0` generates a horizontal reference line through the value 0 on the vertical axis.

The autocovariance function γ obviously satisfies $\gamma(0) \geq 0$ and, by the Cauchy-Schwarz inequality

$$\begin{aligned}
 |\gamma(k)| &= |E((Y_{t+k} - E(Y_{t+k}))(Y_t - E(Y_t)))| \\
 &\leq E(|Y_{t+k} - E(Y_{t+k})||Y_t - E(Y_t)|) \\
 &\leq \text{Var}(Y_{t+k})^{1/2} \text{Var}(Y_t)^{1/2} \\
 &= \gamma(0) \quad \text{for } k \geq 0.
 \end{aligned}$$

Thus we obtain for the autocorrelation function the inequality

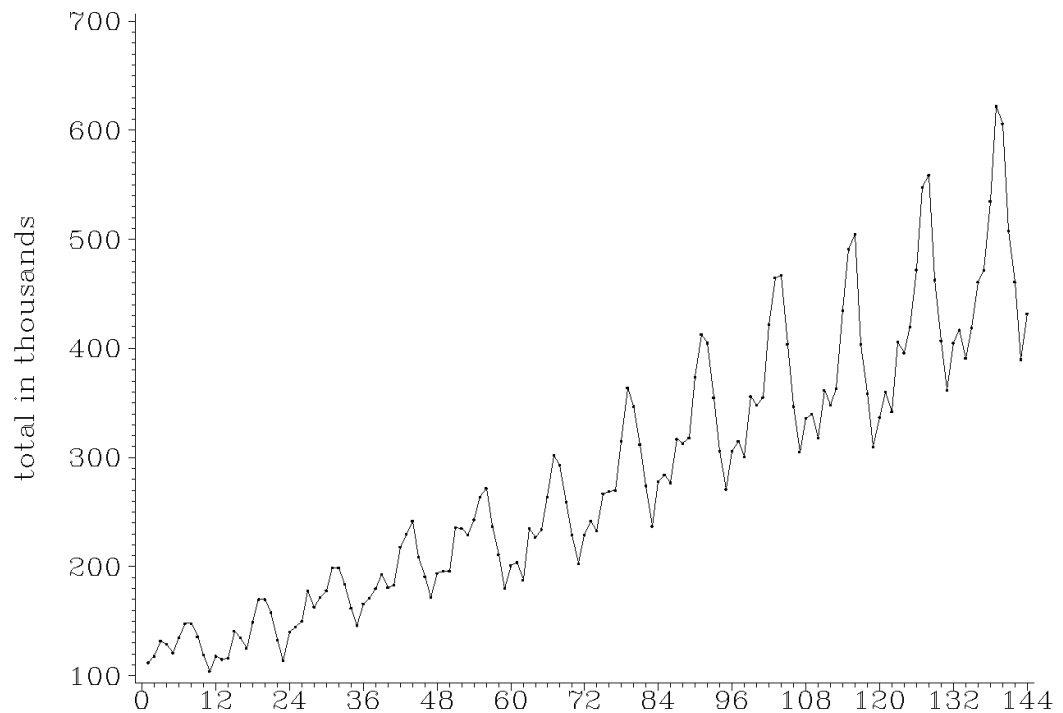
$$|\rho(k)| \leq 1 = \rho(0).$$

Variance Stabilizing Transformation

The scatterplot of the points (t, y_t) sometimes shows a variation of the data y_t depending on their height.

Example 1.3.1. (Airline Data). Plot 1.3.2, which displays monthly totals in thousands of international airline passengers from January

1949 to December 1960, exemplifies the above mentioned dependence. These Airline Data are taken from Box et al. (1994); a discussion can be found in Brockwell and Davis (1991, Section 9.2).



Plot 1.3.2: Monthly totals in thousands of international airline passengers from January 1949 to December 1960.

```

1  /* airline_plot.sas */
2  TITLE1 'Monthly totals from January 49 to December 60';
3  TITLE2 'Airline Data';
4
5  /* Read in the data */
6  DATA datal;
7      INFILE 'c:\data\airline.txt';
8      INPUT y;
9      t=_N_;
10
11  /* Graphical options */
12  AXIS1 LABEL=NONE ORDER=(0 12 24 36 48 60 72 84 96 108 120 132 144)
      ↪MINOR=(N=5);
13  AXIS2 LABEL=(ANGLE=90 'total in thousands');
14  SYMBOL1 V=DOT C=GREEN I=JOIN H=0.2;
15

```

```

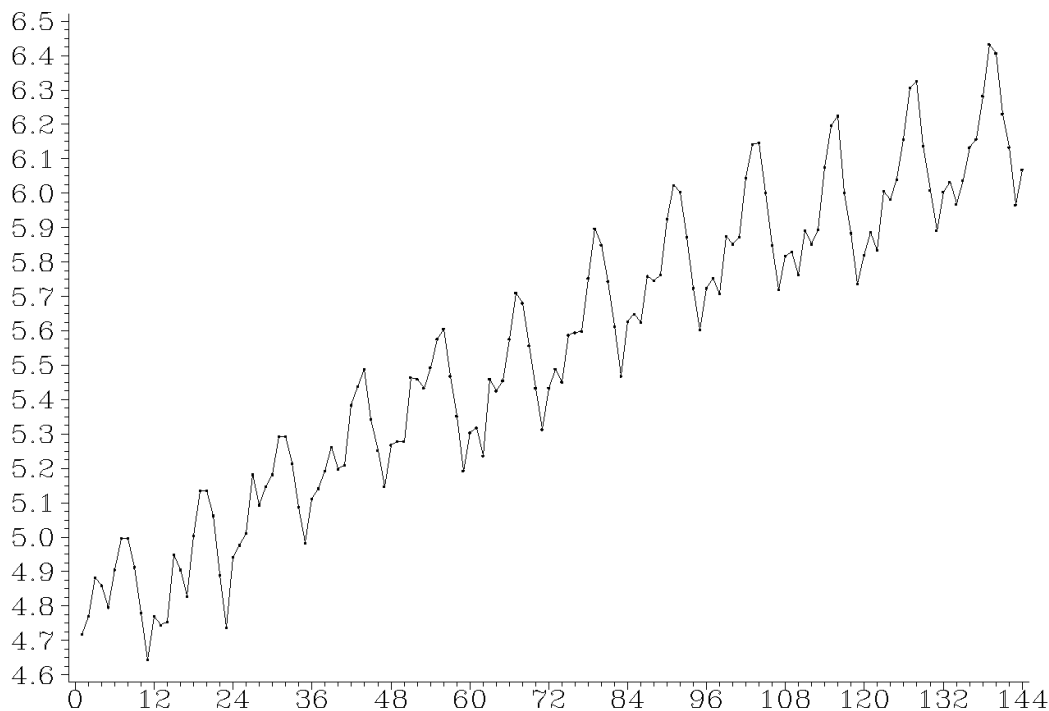
16 /* Plot the data */
17 PROC GPLOT DATA=data1;
18     PLOT y*t / HAXIS=AXIS1 VAXIS=AXIS2;
19 RUN; QUIT;

```

In the first data step, the monthly passenger totals are read into the variable y . To get a time variable t , the temporarily created SAS variable $_{N_}$ is used; it counts the observations.

The passenger totals are plotted against t with a line joining the data points, which are symbolized by small dots. On the horizontal axis a label is suppressed.

The variation of the data y_t obviously increases with their height. The logtransformed data $x_t = \log(y_t)$, displayed in the following figure, however, show no dependence of variability from height.



Plot 1.3.3: Logarithm of Airline Data $x_t = \log(y_t)$.

```

1 /* airline_log.sas */
2 TITLE1 'Logarithmic transformation';
3 TITLE2 'Airline Data';
4
5 /* Read in the data and compute log-transformed data */

```

```

6 DATA data1;
7   INFILE 'c\data\airline.txt';
8   INPUT y;
9   t=_N_;
10  x=LOG(y);
11
12  /* Graphical options */
13  AXIS1 LABEL=NONE ORDER=(0 12 24 36 48 60 72 84 96 108 120 132 144)
    ↪MINOR=(N=5);
14  AXIS2 LABEL=NONE;
15  SYMBOL1 V=DOT C=GREEN I=JOIN H=0.2;
16
17  /* Plot log-transformed data */
18  PROC GPLOT DATA=data1;
19    PLOT x*t / HAXIS=AXIS1 VAXIS=AXIS2;
20  RUN; QUIT;

```

The plot of the log-transformed data is done in the same manner as for the original data in Program 1.3.2 (airline_plot.sas). The only differences are the log-transformation by means of the `LOG` function and the suppressed label on the vertical axis.

The fact that taking the logarithm of data often reduces their variability, can be illustrated as follows. Suppose, for example, that the data were generated by random variables, which are of the form $Y_t = \sigma_t Z_t$, where $\sigma_t > 0$ is a scale factor depending on t , and Z_t , $t \in \mathbb{Z}$, are independent copies of a positive random variable Z with variance 1. The variance of Y_t is in this case σ_t^2 , whereas the variance of $\log(Y_t) = \log(\sigma_t) + \log(Z_t)$ is a constant, namely the variance of $\log(Z)$, if it exists.

A transformation of the data, which reduces the dependence of the variability on their height, is called *variance stabilizing*. The logarithm is a particular case of the general Box–Cox (1964) transformation T_λ of a time series (Y_t) , where the parameter $\lambda \geq 0$ is chosen by the statistician:

$$T_\lambda(Y_t) := \begin{cases} (Y_t^\lambda - 1)/\lambda, & Y_t \geq 0, \lambda > 0 \\ \log(Y_t), & Y_t > 0, \lambda = 0. \end{cases}$$

Note that $\lim_{\lambda \searrow 0} T_\lambda(Y_t) = T_0(Y_t) = \log(Y_t)$ if $Y_t > 0$ (Exercise 1.22). Popular choices of the parameter λ are 0 and 1/2. A variance stabilizing transformation of the data, if necessary, usually precedes any further data manipulation such as trend or seasonal adjustment.

Exercises

1.1. Plot the Mitscherlich function for different values of $\beta_1, \beta_2, \beta_3$ using PROC GPLOT.

1.2. Put in the logistic trend model (1.5) $z_t := 1/y_t \sim 1/E(Y_t) = 1/f_{\log}(t)$, $t = 1, \dots, n$. Then we have the linear regression model $z_t = a + bz_{t-1} + \varepsilon_t$, where ε_t is the error variable. Compute the least squares estimates \hat{a}, \hat{b} of a, b and motivate the estimates $\hat{\beta}_1 := -\log(\hat{b})$, $\hat{\beta}_3 := (1 - \exp(-\hat{\beta}_1))/\hat{a}$ as well as

$$\hat{\beta}_2 := \exp\left(\frac{n+1}{2}\hat{\beta}_1 + \frac{1}{n} \sum_{t=1}^n \log\left(\frac{\hat{\beta}_3}{y_t} - 1\right)\right),$$

proposed by Tintner (1958); see also Exercise 1.3.

1.3. The estimate $\hat{\beta}_2$ defined above suffers from the drawback that all observations y_t have to be strictly less than the estimate $\hat{\beta}_3$. Motivate the following substitute of $\hat{\beta}_2$

$$\tilde{\beta}_2 = \left(\sum_{t=1}^n \frac{\hat{\beta}_3 - y_t}{y_t} \exp(-\hat{\beta}_1 t) \right) / \sum_{t=1}^n \exp(-2\hat{\beta}_1 t)$$

as an estimate of the parameter β_2 in the logistic trend model (1.5).

1.4. Show that in a linear regression model $y_t = \beta_1 x_t + \beta_2$, $t = 1, \dots, n$, the squared multiple correlation coefficient R^2 based on the least squares estimates $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{y}_t := \hat{\beta}_1 x_t + \hat{\beta}_2$ is necessarily between zero and one with $R^2 = 1$ if and only if $\hat{y}_t = y_t, t = 1, \dots, n$ (see (1.12)).

1.5. (Population2 Data) Table 1.3.1 lists total population numbers of North Rhine-Westphalia between 1961 and 1979. Suppose a logistic trend for these data and compute the estimators $\hat{\beta}_1, \hat{\beta}_3$ using PROC REG. Since some observations exceed $\hat{\beta}_3$, use $\tilde{\beta}_2$ from Exercise 1.3 and do an ex post-analysis. Use PROC NLIN and do an ex post-analysis. Compare these two procedures by their residual sums of squares.

1.6. (Income Data) Suppose an allometric trend function for the income data in Example 1.1.3 and do a regression analysis. Plot the data y_t versus $\hat{\beta}_2 t^{\hat{\beta}_1}$. To this end compute the R^2 -coefficient. Estimate the parameters also with PROC NLIN and compare the results.

Year	t	Total Population in millions
1961	1	15.920
1963	2	16.280
1965	3	16.661
1967	4	16.835
1969	5	17.044
1971	6	17.091
1973	7	17.223
1975	8	17.176
1977	9	17.052
1979	10	17.002

Table 1.3.1: Population2 Data.

1.7. (Unemployed2 Data) Table 1.3.2 lists total numbers of unemployed (in thousands) in West Germany between 1950 and 1993. Compare a logistic trend function with an allometric one. Which one gives the better fit?

1.8. Give an update equation for a simple moving average of (even) order 2s.

1.9. (Public Expenditures Data) Table 1.3.3 lists West Germany's public expenditures (in billion D-Marks) between 1961 and 1990. Compute simple moving averages of order 3 and 5 to estimate a possible trend. Plot the original data as well as the filtered ones and compare the curves.

1.10. Check that the Census X-11 Program leads to a moving average of length 165, 169, or 179 of the original data, depending on the length of the Henderson moving average in step (6) of X-11.

1.11. (Unemployed Females Data) Use PROC X11 to analyze the monthly unemployed females between ages 16 and 19 in the United States from January 1961 to December 1985 (in thousands).

1.12. Show that the rank of a matrix \mathbf{A} equals the rank of $\mathbf{A}^T \mathbf{A}$.

Year	Unemployed
1950	1869
1960	271
1970	149
1975	1074
1980	889
1985	2304
1988	2242
1989	2038
1990	1883
1991	1689
1992	1808
1993	2270

Table 1.3.2: Unemployed2 Data.

Year	Public Expenditures	Year	Public Expenditures
1961	113,4	1976	546,2
1962	129,6	1977	582,7
1963	140,4	1978	620,8
1964	153,2	1979	669,8
1965	170,2	1980	722,4
1966	181,6	1981	766,2
1967	193,6	1982	796,0
1968	211,1	1983	816,4
1969	233,3	1984	849,0
1970	264,1	1985	875,5
1971	304,3	1986	912,3
1972	341,0	1987	949,6
1973	386,5	1988	991,1
1974	444,8	1989	1018,9
1975	509,1	1990	1118,1

Table 1.3.3: Public Expenditures Data.

1.13. The $p + 1$ columns of the design matrix \mathbf{X} in (1.17) are linear independent.

1.14. Let (c_u) be the moving average derived by the best local polynomial fit. Show that

- (i) fitting locally a polynomial of degree 2 to five consecutive data points leads to

$$(c_u) = \frac{1}{35}(-3, 12, 17, 12, -3)^T,$$

- (ii) the inverse matrix \mathbf{A}^{-1} of an invertible $m \times m$ -matrix $\mathbf{A} = (a_{ij})_{1 \leq i, j \leq m}$ with the property that $a_{ij} = 0$, if $i + j$ is odd, shares this property,

- (iii) (c_u) is symmetric, i.e., $c_{-u} = c_u$.

1.15. (Unemployed1 Data) Compute a seasonal and trend adjusted time series for the Unemployed1 Data in the building trade. To this end compute seasonal differences and first order differences. Compare the results with those of PROC X11.

1.16. Use the SAS function RANNOR to generate a time series $Y_t = b_0 + b_1 t + \varepsilon_t$, $t = 1, \dots, 100$, where $b_0, b_1 \neq 0$ and the ε_t are independent normal random variables with mean μ and variance σ_1^2 if $t \leq 69$ but variance $\sigma_2^2 \neq \sigma_1^2$ if $t \geq 70$. Plot the exponentially filtered variables Y_t^* for different values of the smoothing parameter $\alpha \in (0, 1)$ and compare the results.

1.17. Compute under the assumptions of Corollary 1.2.9 the variance of an exponentially filtered variable Y_t^* after a change point $t = N$ with $\sigma^2 := E(Y_t - \mu)^2$ for $t < N$ and $\tau^2 := E(Y_t - \lambda)^2$ for $t \geq N$. What is the limit for $t \rightarrow \infty$?

1.18. Show that one obtains the exponential smoother also as least squares estimator of the weighted approach

$$\sum_{j=0}^{\infty} (1 - \alpha)^j (Y_{t-j} - \mu)^2 = \min_{\mu}$$

with $Y_t = Y_0$ for $t < 0$.

1.19. (Female Unemployed Data) Compute exponentially smoothed series of the Female Unemployed Data with different smoothing parameters α . Compare the results to those obtained with simple moving averages and X-11.

1.20. (Bankruptcy Data) Table 1.3.4 lists the percentages to annual bankruptcies among all US companies between 1867 and 1932:

1.33	0.94	0.79	0.83	0.61	0.77	0.93	0.97	1.20	1.33
1.36	1.55	0.95	0.59	0.61	0.83	1.06	1.21	1.16	1.01
0.97	1.02	1.04	0.98	1.07	0.88	1.28	1.25	1.09	1.31
1.26	1.10	0.81	0.92	0.90	0.93	0.94	0.92	0.85	0.77
0.83	1.08	0.87	0.84	0.88	0.99	0.99	1.10	1.32	1.00
0.80	0.58	0.38	0.49	1.02	1.19	0.94	1.01	1.00	1.01
1.07	1.08	1.04	1.21	1.33	1.53				

Table 1.3.4: Bankruptcy Data.

Compute and plot the empirical autocovariance function and the empirical autocorrelation function using the SAS procedures PROC ARIMA and PROC GPLOT.

1.21. Verify that the empirical correlation $r(k)$ at lag k for the trend $y_t = t$, $t = 1, \dots, n$ is given by

$$r(k) = 1 - 3\frac{k}{n} + 2\frac{k(k^2 - 1)}{n(n^2 - 1)}, \quad k = 0, \dots, n.$$

Plot the correlogram for different values of n . This example shows, that the correlogram has no interpretation for non-stationary processes (see Exercise 1.20).

1.22. Show that

$$\lim_{\lambda \downarrow 0} T_\lambda(Y_t) = T_0(Y_t) = \log(Y_t), \quad Y_t > 0$$

for the Box–Cox transformation T_λ .

Models of Time Series

Each time series Y_1, \dots, Y_n can be viewed as a clipping from a sequence of random variables $\dots, Y_{-2}, Y_{-1}, Y_0, Y_1, Y_2, \dots$. In the following we will introduce several models for such a *stochastic process* Y_t with index set \mathbb{Z} .

2.1 Linear Filters and Stochastic Processes

For mathematical convenience we will consider *complex valued* random variables Y , whose range is the set of complex numbers $\mathbb{C} = \{u + iv : u, v \in \mathbb{R}\}$, where $i = \sqrt{-1}$. Therefore, we can decompose Y as $Y = Y_{(1)} + iY_{(2)}$, where $Y_{(1)} = \text{Re}(Y)$ is the *real part* of Y and $Y_{(2)} = \text{Im}(Y)$ is its *imaginary part*. The random variable Y is called *integrable* if the real valued random variables $Y_{(1)}, Y_{(2)}$ both have finite expectations, and in this case we define the expectation of Y by

$$E(Y) := E(Y_{(1)}) + i E(Y_{(2)}) \in \mathbb{C}.$$

This expectation has, up to monotonicity, the usual properties such as $E(aY + bZ) = a E(Y) + b E(Z)$ of its real counterpart (see Exercise 2.1). Here a and b are complex numbers and Z is a further integrable complex valued random variable. In addition we have $\overline{E(Y)} = E(\bar{Y})$, where $\bar{a} = u - iv$ denotes the *conjugate complex number* of $a = u + iv$. Since $|a|^2 := u^2 + v^2 = a\bar{a} = \bar{a}a$, we define the *variance* of Y by

$$\text{Var}(Y) := E((Y - E(Y))(\overline{Y - E(Y)})) \geq 0.$$

The complex random variable Y is called *square integrable* if this number is finite. To carry the equation $\text{Var}(X) = \text{Cov}(X, X)$ for a

real random variable X over to complex ones, we define the *covariance* of complex square integrable random variables Y, Z by

$$\text{Cov}(Y, Z) := E((Y - E(Y))(\overline{Z - E(Z)})).$$

Note that the covariance $\text{Cov}(Y, Z)$ is no longer symmetric with respect to Y and Z , as it is for real valued random variables, but it satisfies $\text{Cov}(Y, Z) = \overline{\text{Cov}(Z, Y)}$.

The following lemma implies that the Cauchy–Schwarz inequality carries over to complex valued random variables.

Lemma 2.1.1. *For any integrable complex valued random variable $Y = Y_{(1)} + iY_{(2)}$ we have*

$$|E(Y)| \leq E(|Y|) \leq E(|Y_{(1)}|) + E(|Y_{(2)}|).$$

Proof. We write $E(Y)$ in polar coordinates $E(Y) = re^{i\vartheta}$, where $r = |E(Y)|$ and $\vartheta \in [0, 2\pi)$. Observe that

$$\begin{aligned} \text{Re}(e^{-i\vartheta}Y) &= \text{Re}\left((\cos(\vartheta) - i\sin(\vartheta))(Y_{(1)} + iY_{(2)})\right) \\ &= \cos(\vartheta)Y_{(1)} + \sin(\vartheta)Y_{(2)} \\ &\leq (\cos^2(\vartheta) + \sin^2(\vartheta))^{1/2}(Y_{(1)}^2 + Y_{(2)}^2)^{1/2} = |Y| \end{aligned}$$

by the Cauchy–Schwarz inequality for real numbers. Thus we obtain

$$\begin{aligned} |E(Y)| = r &= E(e^{-i\vartheta}Y) \\ &= E\left(\text{Re}(e^{-i\vartheta}Y)\right) \leq E(|Y|). \end{aligned}$$

The second inequality of the lemma follows from $|Y| = (Y_{(1)}^2 + Y_{(2)}^2)^{1/2} \leq |Y_{(1)}| + |Y_{(2)}|$. \square

The next result is a consequence of the preceding lemma and the Cauchy–Schwarz inequality for real valued random variables.

Corollary 2.1.2. *For any square integrable complex valued random variable we have*

$$|E(YZ)| \leq E(|Y||Z|) \leq E(|Y|^2)^{1/2} E(|Z|^2)^{1/2}$$

and thus,

$$|\text{Cov}(Y, Z)| \leq \text{Var}(Y)^{1/2} \text{Var}(Z)^{1/2}.$$

Stationary Processes

A stochastic process $(Y_t)_{t \in \mathbb{Z}}$ of square integrable complex valued random variables is said to be (*weakly*) *stationary* if for any $t_1, t_2, k \in \mathbb{Z}$

$$E(Y_{t_1}) = E(Y_{t_1+k}) \text{ and } E(Y_{t_1} \bar{Y}_{t_2}) = E(Y_{t_1+k} \bar{Y}_{t_2+k}).$$

The random variables of a stationary process $(Y_t)_{t \in \mathbb{Z}}$ have identical means and variances. The autocovariance function satisfies moreover for $s, t \in \mathbb{Z}$

$$\begin{aligned} \gamma(t, s) &:= \text{Cov}(Y_t, Y_s) = \text{Cov}(Y_{t-s}, Y_0) =: \gamma(t - s) \\ &= \overline{\text{Cov}(Y_0, Y_{t-s})} = \overline{\text{Cov}(Y_{s-t}, Y_0)} = \overline{\gamma(s - t)}, \end{aligned}$$

and thus, the autocovariance function of a stationary process can be viewed as a function of a single argument satisfying $\gamma(t) = \overline{\gamma(-t)}$, $t \in \mathbb{Z}$.

A stationary process $(\varepsilon_t)_{t \in \mathbb{Z}}$ of square integrable and uncorrelated *real* valued random variables is called *white noise* i.e., $\text{Cov}(\varepsilon_t, \varepsilon_s) = 0$ for $t \neq s$ and there exist $\mu \in \mathbb{R}$, $\sigma \geq 0$ such that

$$E(\varepsilon_t) = \mu, \quad E((\varepsilon_t - \mu)^2) = \sigma^2, \quad t \in \mathbb{Z}.$$

In Section 1.2 we defined linear filters of a time series, which were based on a finite number of real valued weights. In the following we consider linear filters with an infinite number of complex valued weights.

Suppose that $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a white noise and let $(a_t)_{t \in \mathbb{Z}}$ be a sequence of complex numbers satisfying $\sum_{t=-\infty}^{\infty} |a_t| := \sum_{t \geq 0} |a_t| + \sum_{t \geq 1} |a_{-t}| < \infty$. Then $(a_t)_{t \in \mathbb{Z}}$ is said to be an *absolutely summable (linear) filter* and

$$Y_t := \sum_{u=-\infty}^{\infty} a_u \varepsilon_{t-u} := \sum_{u \geq 0} a_u \varepsilon_{t-u} + \sum_{u \geq 1} a_{-u} \varepsilon_{t+u}, \quad t \in \mathbb{Z},$$

is called a *general linear process*.

Existence of General Linear Processes

We will show that $\sum_{u=-\infty}^{\infty} |a_u \varepsilon_{t-u}| < \infty$ with probability one for $t \in \mathbb{Z}$ and, thus, $Y_t = \sum_{u=-\infty}^{\infty} a_u \varepsilon_{t-u}$ is well defined. Denote by

$L_2 := L_2(\Omega, \mathcal{A}, \mathcal{P})$ the set of all complex valued square integrable random variables, defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, and put $\|Y\|_2 := E(|Y|^2)^{1/2}$, which is the L_2 -pseudonorm on L_2 .

Lemma 2.1.3. *Let X_n , $n \in \mathbb{N}$, be a sequence in L_2 such that $\|X_{n+1} - X_n\|_2 \leq 2^{-n}$ for each $n \in \mathbb{N}$. Then there exists $X \in L_2$ such that $\lim_{n \rightarrow \infty} X_n = X$ with probability one.*

Proof. Write $X_n = \sum_{k \leq n} (X_k - X_{k-1})$, where $X_0 := 0$. By the monotone convergence theorem, the Cauchy–Schwarz inequality and Corollary 2.1.2 we have

$$\begin{aligned} E\left(\sum_{k \geq 1} |X_k - X_{k-1}|\right) &= \sum_{k \geq 1} E(|X_k - X_{k-1}|) \leq \sum_{k \geq 1} \|X_k - X_{k-1}\|_2 \\ &\leq \|X_1\|_2 + \sum_{k \geq 1} 2^{-k} < \infty. \end{aligned}$$

This implies that $\sum_{k \geq 1} |X_k - X_{k-1}| < \infty$ with probability one and hence, the limit $\lim_{n \rightarrow \infty} \sum_{k \leq n} (X_k - X_{k-1}) = \lim_{n \rightarrow \infty} X_n = X$ exists in \mathbb{C} almost surely. Finally, we check that $X \in L_2$:

$$\begin{aligned} E(|X|^2) &= E(\lim_{n \rightarrow \infty} |X_n|^2) \\ &\leq E\left(\lim_{n \rightarrow \infty} \left(\sum_{k \leq n} |X_k - X_{k-1}|\right)^2\right) \\ &= \lim_{n \rightarrow \infty} E\left(\left(\sum_{k \leq n} |X_k - X_{k-1}|\right)^2\right) \\ &= \lim_{n \rightarrow \infty} \sum_{k, j \leq n} E(|X_k - X_{k-1}| |X_j - X_{j-1}|) \\ &\leq \lim_{n \rightarrow \infty} \sum_{k, j \leq n} \|X_k - X_{k-1}\|_2 \|X_j - X_{j-1}\|_2 \\ &= \lim_{n \rightarrow \infty} \left(\sum_{k \leq n} \|X_k - X_{k-1}\|_2\right)^2 \\ &= \left(\sum_{k \geq 1} \|X_k - X_{k-1}\|_2\right)^2 < \infty. \end{aligned}$$

□

Theorem 2.1.4. *The space $(L_2, \|\cdot\|_2)$ is complete i.e., suppose that $X_n \in L_2$, $n \in \mathbb{N}$, has the property that for arbitrary $\varepsilon > 0$ one can find an integer $N(\varepsilon) \in \mathbb{N}$ such that $\|X_n - X_m\|_2 < \varepsilon$ if $n, m \geq N(\varepsilon)$. Then there exists a random variable $X \in L_2$ such that $\lim_{n \rightarrow \infty} \|X - X_n\|_2 = 0$.*

Proof. We can find integers $n_1 < n_2 < \dots$ such that

$$\|X_n - X_m\|_2 \leq 2^{-k} \quad \text{if } n, m \geq n_k.$$

By Lemma 2.1.3 there exists a random variable $X \in L_2$ such that $\lim_{k \rightarrow \infty} X_{n_k} = X$ with probability one. Fatou's lemma implies

$$\begin{aligned} \|X_n - X\|_2^2 &= E(|X_n - X|^2) \\ &= E\left(\liminf_{k \rightarrow \infty} |X_n - X_{n_k}|^2\right) \leq \liminf_{k \rightarrow \infty} \|X_n - X_{n_k}\|_2^2. \end{aligned}$$

The right-hand side of this inequality becomes arbitrarily small if we choose n large enough, and thus we have $\lim_{n \rightarrow \infty} \|X_n - X\|_2^2 = 0$. \square

The following result implies in particular that a general linear process is well defined.

Theorem 2.1.5. *Suppose that $(Z_t)_{t \in \mathbb{Z}}$ is a complex valued stochastic process such that $\sup_t E(|Z_t|) < \infty$ and let $(a_t)_{t \in \mathbb{Z}}$ be an absolutely summable filter. Then we have $\sum_{u \in \mathbb{Z}} |a_u Z_{t-u}| < \infty$ with probability one for $t \in \mathbb{Z}$ and, thus, $Y_t := \sum_{u \in \mathbb{Z}} a_u Z_{t-u}$ exists almost surely in \mathbb{C} . We have moreover $E(|Y_t|) < \infty$, $t \in \mathbb{Z}$, and*

$$(i) \quad E(Y_t) = \lim_{n \rightarrow \infty} \sum_{u=-n}^n a_u E(Z_{t-u}), \quad t \in \mathbb{Z},$$

$$(ii) \quad E(|Y_t - \sum_{u=-n}^n a_u Z_{t-u}|) \xrightarrow{n \rightarrow \infty} 0.$$

If, in addition, $\sup_t E(|Z_t|^2) < \infty$, then we have $E(|Y_t|^2) < \infty$, $t \in \mathbb{Z}$, and

$$(iii) \quad \|Y_t - \sum_{u=-n}^n a_u Z_{t-u}\|_2 \xrightarrow{n \rightarrow \infty} 0.$$

Proof. The monotone convergence theorem implies

$$\begin{aligned}
 \mathbb{E}\left(\sum_{u \in \mathbb{Z}} |a_u| |Z_{t-u}|\right) &= \lim_{n \rightarrow \infty} \mathbb{E}\left(\sum_{u=-n}^n |a_u| |Z_{t-u}|\right) \\
 &= \lim_{n \rightarrow \infty} \left(\sum_{u=-n}^n |a_u| \mathbb{E}(|Z_{t-u}|)\right) \\
 &\leq \lim_{n \rightarrow \infty} \left(\sum_{u=-n}^n |a_u|\right) \sup_{t \in \mathbb{Z}} \mathbb{E}(|Z_{t-u}|) < \infty
 \end{aligned}$$

and, thus, we have $\sum_{u \in \mathbb{Z}} |a_u| |Z_{t-u}| < \infty$ with probability one as well as $\mathbb{E}(|Y_t|) \leq \mathbb{E}(\sum_{u \in \mathbb{Z}} |a_u| |Z_{t-u}|) < \infty$, $t \in \mathbb{Z}$. Put $X_n(t) := \sum_{u=-n}^n a_u Z_{t-u}$. Then we have $|Y_t - X_n(t)| \xrightarrow{n \rightarrow \infty} 0$ almost surely. By the inequality $|Y_t - X_n(t)| \leq \sum_{u \in \mathbb{Z}} |a_u| |Z_{t-u}|$, $n \in \mathbb{N}$, the dominated convergence theorem implies (ii) and therefore (i):

$$\begin{aligned}
 \left| \mathbb{E}(Y_t) - \sum_{u=-n}^n a_u \mathbb{E}(Z_{t-u}) \right| &= \left| \mathbb{E}(Y_t) - \mathbb{E}(X_n(t)) \right| \\
 &\leq \mathbb{E}(|Y_t - X_n(t)|) \xrightarrow{n \rightarrow \infty} 0.
 \end{aligned}$$

Put $K := \sup_t \mathbb{E}(|Z_t|^2) < \infty$. The Cauchy-Schwarz inequality implies

for $m, n \in \mathbb{N}$ and $\varepsilon > 0$

$$\begin{aligned}
0 &\leq \mathbb{E}(|X_{n+m}(t) - X_n(t)|^2) \\
&= \mathbb{E} \left(\left| \sum_{|u|=n+1}^{n+m} a_u Z_{t-u} \right|^2 \right) \\
&= \sum_{|u|=n+1}^{n+m} \sum_{|w|=n+1}^{n+m} a_u \bar{a}_w \mathbb{E}(Z_{t-u} \bar{Z}_{t-w}) \\
&\leq \sum_{|u|=n+1}^{n+m} \sum_{|w|=n+1}^{n+m} |a_u| |a_w| \mathbb{E}(|Z_{t-u}| |Z_{t-w}|) \\
&\leq \sum_{|u|=n+1}^{n+m} \sum_{|w|=n+1}^{n+m} |a_u| |a_w| \mathbb{E}(|Z_{t-u}|^2)^{1/2} \mathbb{E}(|Z_{t-w}|^2)^{1/2} \\
&\leq K \left(\sum_{|u|=n+1}^{n+m} |a_u| \right)^2 \leq K \left(\sum_{|u| \geq n} |a_u| \right)^2 < \varepsilon
\end{aligned}$$

if n is chosen sufficiently large. Theorem 2.1.4 now implies the existence of a random variable $X(t) \in L_2$ with $\lim_{n \rightarrow \infty} \|X_n(t) - X(t)\|_2 = 0$. For the proof of (iii) it remains to show that $X(t) = Y_t$ almost surely. Markov's inequality implies

$$P\{|Y_t - X_n(t)| \geq \varepsilon\} \leq \varepsilon^{-1} \mathbb{E}(|Y_t - X_n(t)|) \longrightarrow_{n \rightarrow \infty} 0$$

by (ii), and Chebyshev's inequality yields

$$P\{|X(t) - X_n(t)| \geq \varepsilon\} \leq \varepsilon^{-2} \|X(t) - X_n(t)\|_2^2 \longrightarrow_{n \rightarrow \infty} 0$$

for arbitrary $\varepsilon > 0$. This implies

$$\begin{aligned}
&P\{|Y_t - X(t)| \geq \varepsilon\} \\
&\leq P\{|Y_t - X_n(t)| + |X_n(t) - X(t)| \geq \varepsilon\} \\
&\leq P\{|Y_t - X_n(t)| \geq \varepsilon/2\} + P\{|X(t) - X_n(t)| \geq \varepsilon/2\} \longrightarrow_{n \rightarrow \infty} 0
\end{aligned}$$

and thus $Y_t = X(t)$ almost surely, because Y_t does not depend on n . This completes the proof of Theorem 2.1.5. \square

Theorem 2.1.6. Suppose that $(Z_t)_{t \in \mathbb{Z}}$ is a stationary process with mean $\mu_Z := E(Z_0)$ and autocovariance function γ_Z and let (a_t) be an absolutely summable filter. Then $Y_t = \sum_u a_u Z_{t-u}$, $t \in \mathbb{Z}$, is also stationary with

$$\mu_Y = E(Y_0) = \left(\sum_u a_u \right) \mu_Z$$

and autocovariance function

$$\gamma_Y(t) = \sum_u \sum_w a_u \bar{a}_w \gamma_Z(t + w - u).$$

Proof. Note that

$$\begin{aligned} E(|Z_t|^2) &= E(|Z_t - \mu_Z + \mu_Z|^2) \\ &= E((Z_t - \mu_Z + \mu_Z)(\overline{Z_t - \mu_Z + \mu_Z})) \\ &= E(|Z_t - \mu_Z|^2) + |\mu_Z|^2 \\ &= \gamma_Z(0) + |\mu_Z|^2 \end{aligned}$$

and, thus,

$$\sup_{t \in \mathbb{Z}} E(|Z_t|^2) < \infty.$$

We can, therefore, now apply Theorem 2.1.5. Part (i) of Theorem 2.1.5 immediately implies $E(Y_t) = (\sum_u a_u) \mu_Z$ and part (iii) implies that the Y_t are square integrable and for $t, s \in \mathbb{Z}$ we get (see Exercise 2.16 for the second equality)

$$\begin{aligned} \gamma_Y(t - s) &= E((Y_t - \mu_Y)(\overline{Y_s - \mu_Y})) \\ &= \lim_{n \rightarrow \infty} \text{Cov} \left(\sum_{u=-n}^n a_u Z_{t-u}, \sum_{w=-n}^n a_w Z_{s-w} \right) \\ &= \lim_{n \rightarrow \infty} \sum_{u=-n}^n \sum_{w=-n}^n a_u \bar{a}_w \text{Cov}(Z_{t-u}, Z_{s-w}) \\ &= \lim_{n \rightarrow \infty} \sum_{u=-n}^n \sum_{w=-n}^n a_u \bar{a}_w \gamma_Z(t - s + w - u) \\ &= \sum_u \sum_w a_u \bar{a}_w \gamma_Z(t - s + w - u). \end{aligned}$$

The covariance of Y_t and Y_s depends, therefore, only on the difference $t-s$. Note that $|\gamma_Z(t)| \leq \gamma_Z(0) < \infty$ and thus, $\sum_u \sum_w |a_u \bar{a}_w \gamma_Z(t-s+w-u)| \leq \gamma_Z(0) (\sum_u |a_u|)^2 < \infty$, i.e., (Y_t) is a stationary process. \square

The Covariance Generating Function

The *covariance generating function* of a stationary process with autocovariance function γ is defined as the double series

$$G(z) := \sum_{t \in \mathbb{Z}} \gamma(t) z^t = \sum_{t \geq 0} \gamma(t) z^t + \sum_{t \geq 1} \gamma(-t) z^{-t},$$

known as a *Laurent series* in complex analysis. We assume that there exists a real number $r > 1$ such that $G(z)$ is defined for all $z \in \mathbb{C}$ in the annulus $1/r < |z| < r$. The covariance generating function will help us to compute the autocovariances of filtered processes.

Since the coefficients of a Laurent series are uniquely determined (see e.g. Conway, 1978, Chapter V, 1.11), the covariance generating function of a stationary process is a constant function if and only if this process is a white noise i.e., $\gamma(t) = 0$ for $t \neq 0$.

Theorem 2.1.7. *Suppose that $Y_t = \sum_u a_u \varepsilon_{t-u}$, $t \in \mathbb{Z}$, is a general linear process with $\sum_u |a_u| |z^u| < \infty$, if $r^{-1} < |z| < r$ for some $r > 1$. Put $\sigma^2 := \text{Var}(\varepsilon_0)$. The process (Y_t) then has the covariance generating function*

$$G(z) = \sigma^2 \left(\sum_u a_u z^u \right) \left(\sum_u \bar{a}_u z^{-u} \right), \quad r^{-1} < |z| < r.$$

Proof. Theorem 2.1.6 implies for $t \in \mathbb{Z}$

$$\begin{aligned} \text{Cov}(Y_t, Y_0) &= \sum_u \sum_w a_u \bar{a}_w \gamma_\varepsilon(t+w-u) \\ &= \sigma^2 \sum_u a_u \bar{a}_{u-t}. \end{aligned}$$

This implies

$$\begin{aligned}
G(z) &= \sigma^2 \sum_t \sum_u a_u \bar{a}_{u-t} z^t \\
&= \sigma^2 \left(\sum_u |a_u|^2 + \sum_{t \geq 1} \sum_u a_u \bar{a}_{u-t} z^t + \sum_{t \leq -1} \sum_u a_u \bar{a}_{u-t} z^t \right) \\
&= \sigma^2 \left(\sum_u |a_u|^2 + \sum_u \sum_{t \leq u-1} a_u \bar{a}_t z^{u-t} + \sum_u \sum_{t \geq u+1} a_u \bar{a}_t z^{u-t} \right) \\
&= \sigma^2 \sum_u \sum_t a_u \bar{a}_t z^{u-t} = \sigma^2 \left(\sum_u a_u z^u \right) \left(\sum_t \bar{a}_t z^{-t} \right).
\end{aligned}$$

□

Example 2.1.8. Let $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a white noise with $\text{Var}(\varepsilon_0) =: \sigma^2 > 0$. The covariance generating function of the simple moving average $Y_t = \sum_u a_u \varepsilon_{t-u}$ with $a_{-1} = a_0 = a_1 = 1/3$ and $a_u = 0$ elsewhere is then given by

$$\begin{aligned}
G(z) &= \frac{\sigma^2}{9} (z^{-1} + z^0 + z^1)(z^1 + z^0 + z^{-1}) \\
&= \frac{\sigma^2}{9} (z^{-2} + 2z^{-1} + 3z^0 + 2z^1 + z^2), \quad z \in \mathbb{R}.
\end{aligned}$$

Then the autocovariances are just the coefficients in the above series

$$\begin{aligned}
\gamma(0) &= \frac{\sigma^2}{3}, \\
\gamma(1) &= \gamma(-1) = \frac{2\sigma^2}{9}, \\
\gamma(2) &= \gamma(-2) = \frac{\sigma^2}{9}, \\
\gamma(k) &= 0 \quad \text{elsewhere.}
\end{aligned}$$

This explains the name covariance generating function.

The Characteristic Polynomial

Let (a_u) be an absolutely summable filter. The Laurent series

$$A(z) := \sum_{u \in \mathbb{Z}} a_u z^u$$

is called *characteristic polynomial* of (a_u) . We know from complex analysis that $A(z)$ exists either for all z in some annulus $r < |z| < R$ or almost nowhere. In the first case the coefficients a_u are uniquely determined by the function $A(z)$ (see e.g. Conway, 1978, Chapter V, 1.11).

If, for example, (a_u) is absolutely summable with $a_u = 0$ for $u \geq 1$, then $A(z)$ exists for all complex z such that $|z| \geq 1$. If $a_u = 0$ for all large $|u|$, then $A(z)$ exists for all $z \neq 0$.

Inverse Filters

Let now (a_u) and (b_u) be absolutely summable filters and denote by $Y_t := \sum_u a_u Z_{t-u}$, the filtered stationary sequence, where $(Z_u)_{u \in \mathbb{Z}}$ is a stationary process. Filtering $(Y_t)_{t \in \mathbb{Z}}$ by means of (b_u) leads to

$$\sum_w b_w Y_{t-w} = \sum_w \sum_u b_w a_u Z_{t-w-u} = \sum_v \left(\sum_{u+w=v} b_w a_u \right) Z_{t-v},$$

where $c_v := \sum_{u+w=v} b_w a_u$, $v \in \mathbb{Z}$, is an absolutely summable filter:

$$\sum_v |c_v| \leq \sum_v \sum_{u+w=v} |b_w a_u| = \left(\sum_u |a_u| \right) \left(\sum_w |b_w| \right) < \infty.$$

We call (c_v) the *product filter* of (a_u) and (b_u) .

Lemma 2.1.9. *Let (a_u) and (b_u) be absolutely summable filters with characteristic polynomials $A_1(z)$ and $A_2(z)$, which both exist on some annulus $r < |z| < R$. The product filter $(c_v) = (\sum_{u+w=v} b_w a_u)$ then has the characteristic polynomial*

$$A(z) = A_1(z)A_2(z).$$

Proof. By repeating the above arguments we obtain

$$A(z) = \sum_v \left(\sum_{u+w=v} b_w a_u \right) z^v = A_1(z)A_2(z).$$

□

Suppose now that (a_u) and (b_u) are absolutely summable filters with characteristic polynomials $A_1(z)$ and $A_2(z)$, which both exist on some annulus $r < z < R$, where they satisfy $A_1(z)A_2(z) = 1$. Since $1 = \sum_v c_v z^v$ if $c_0 = 1$ and $c_v = 0$ elsewhere, the uniquely determined coefficients of the characteristic polynomial of the product filter of (a_u) and (b_u) are given by

$$\sum_{u+w=v} b_w a_u = \begin{cases} 1 & \text{if } v = 0 \\ 0 & \text{if } v \neq 0. \end{cases}$$

In this case we obtain for a stationary process (Z_t) that almost surely

$$Y_t = \sum_u a_u Z_{t-u} \quad \text{and} \quad \sum_w b_w Y_{t-w} = Z_t, \quad t \in \mathbb{Z}. \quad (2.1)$$

The filter (b_u) is, therefore, called the *inverse filter* of (a_u) .

Causal Filters

An absolutely summable filter $(a_u)_{u \in \mathbb{Z}}$ is called *causal* if $a_u = 0$ for $u < 0$.

Lemma 2.1.10. *Let $a \in \mathbb{C}$. The filter (a_u) with $a_0 = 1$, $a_1 = -a$ and $a_u = 0$ elsewhere has an absolutely summable and causal inverse filter $(b_u)_{u \geq 0}$ if and only if $|a| < 1$. In this case we have $b_u = a^u$, $u \geq 0$.*

Proof. The characteristic polynomial of (a_u) is $A_1(z) = 1 - az$, $z \in \mathbb{C}$. Since the characteristic polynomial $A_2(z)$ of an inverse filter satisfies $A_1(z)A_2(z) = 1$ on some annulus, we have $A_2(z) = 1/(1-az)$. Observe now that

$$\frac{1}{1-az} = \sum_{u \geq 0} a^u z^u, \quad \text{if } |z| < 1/|a|.$$

As a consequence, if $|a| < 1$, then $A_2(z) = \sum_{u \geq 0} a^u z^u$ exists for all $|z| < 1$ and the inverse causal filter $(a^u)_{u \geq 0}$ is absolutely summable, i.e., $\sum_{u \geq 0} |a^u| < \infty$. If $|a| \geq 1$, then $A_2(z) = \sum_{u \geq 0} a^u z^u$ exists for all $|z| < 1/|a|$, but $\sum_{u \geq 0} |a|^u = \infty$, which completes the proof. \square

Theorem 2.1.11. *Let $a_1, a_2, \dots, a_p \in \mathbb{C}$, $a_p \neq 0$. The filter (a_u) with coefficients $a_0 = 1, a_1, \dots, a_p$ and $a_u = 0$ elsewhere has an absolutely*

summable and causal inverse filter if the p roots $z_1, \dots, z_p \in \mathbb{C}$ of $A(z) = 1 + a_1 z + a_2 z^2 + \dots + a_p z^p$ are outside of the unit circle i.e., $|z_i| > 1$ for $1 \leq i \leq p$.

Proof. We know from the Fundamental Theorem of Algebra that the equation $A(z) = 0$ has exactly p roots $z_1, \dots, z_p \in \mathbb{C}$ (see e.g. Conway, 1978, Chapter IV, 3.5), which are all different from zero, since $A(0) = 1$. Hence we can write (see e.g. Conway, 1978, Chapter IV, 3.6)

$$\begin{aligned} A(z) &= a_p(z - z_1) \cdots (z - z_p) \\ &= c \left(1 - \frac{z}{z_1}\right) \left(1 - \frac{z}{z_2}\right) \cdots \left(1 - \frac{z}{z_p}\right), \end{aligned}$$

where $c := a_p(-1)^p z_1 \cdots z_p$. In case of $|z_i| > 1$ we can write for $|z| < |z_i|$

$$\frac{1}{1 - \frac{z}{z_i}} = \sum_{u \geq 0} \left(\frac{1}{z_i}\right)^u z^u, \quad (2.2)$$

where the coefficients $(1/z_i)^u$, $u \geq 0$, are absolutely summable. In case of $|z_i| < 1$, we have for $|z| > |z_i|$

$$\frac{1}{1 - \frac{z}{z_i}} = -\frac{1}{\frac{z}{z_i}} \frac{1}{1 - \frac{z_i}{z}} = -\frac{z_i}{z} \sum_{u \geq 0} z_i^u z^{-u} = -\sum_{u \leq -1} \left(\frac{1}{z_i}\right)^u z^u,$$

where the filter with coefficients $-(1/z_i)^u$, $u \leq -1$, is not a causal one. In case of $|z_i| = 1$, we have for $|z| < 1$

$$\frac{1}{1 - \frac{z}{z_i}} = \sum_{u \geq 0} \left(\frac{1}{z_i}\right)^u z^u,$$

where the coefficients $(1/z_i)^u$, $u \geq 0$, are not absolutely summable. Since the coefficients of a Laurent series are uniquely determined, the factor $1 - z/z_i$ has an inverse $1/(1 - z/z_i) = \sum_{u \geq 0} b_u z^u$ on some annulus with $\sum_{u \geq 0} |b_u| < \infty$ if $|z_i| > 1$. A small analysis implies that this argument carries over to the product

$$\frac{1}{A(z)} = \frac{1}{c \left(1 - \frac{z}{z_1}\right) \cdots \left(1 - \frac{z}{z_p}\right)}$$

which has an expansion $1/A(z) = \sum_{u \geq 0} b_u z^u$ on some annulus with $\sum_{u \geq 0} |b_u| < \infty$ if each factor has such an expansion, and thus, the proof is complete. \square

Remark 2.1.12. Note that the roots z_1, \dots, z_p of $A(z) = 1 + a_1 z + \dots + a_p z^p$ are complex valued and thus, the coefficients b_u of the inverse causal filter will, in general, be complex valued as well. The preceding proof shows, however, that if a_p and each z_i are real numbers, then the coefficients b_u , $u \geq 0$, are real as well.

The preceding proof shows, moreover, that a filter (a_u) with complex coefficients $a_0, a_1, \dots, a_p \in \mathbb{C}$ and $a_u = 0$ elsewhere has an absolutely summable inverse filter if *no* root $z \in \mathbb{C}$ of the equation $A(z) = a_0 + a_1 z + \dots + a_p z^p = 0$ has length 1 i.e., $|z| \neq 1$ for each root. The additional condition $|z| > 1$ for each root then implies that the inverse filter is a causal one.

Example 2.1.13. The filter with coefficients $a_0 = 1, a_1 = -0.7$ and $a_2 = 0.1$ has the characteristic polynomial $A(z) = 1 - 0.7z + 0.1z^2 = 0.1(z - 2)(z - 5)$, with $z_1 = 2, z_2 = 5$ being the roots of $A(z) = 0$. Theorem 2.1.11 implies the existence of an absolutely summable inverse causal filter, whose coefficients can be obtained by expanding $1/A(z)$ as a power series of z :

$$\begin{aligned} \frac{1}{A(z)} &= \frac{1}{\left(1 - \frac{z}{2}\right)\left(1 - \frac{z}{5}\right)} = \left(\sum_{u \geq 0} \left(\frac{1}{2}\right)^u z^u\right) \left(\sum_{w \geq 0} \left(\frac{1}{5}\right)^w z^w\right) \\ &= \sum_{v \geq 0} \sum_{u+w=v} \left(\frac{1}{2}\right)^u \left(\frac{1}{5}\right)^w z^v \\ &= \sum_{v \geq 0} \sum_{w=0}^v \left(\frac{1}{2}\right)^{v-w} \left(\frac{1}{5}\right)^w z^v \\ &= \sum_{v \geq 0} \left(\frac{1}{2}\right)^v \frac{1 - \left(\frac{2}{5}\right)^{v+1}}{1 - \frac{2}{5}} z^v = \sum_{v \geq 0} \frac{10}{3} \left(\left(\frac{1}{2}\right)^{v+1} - \left(\frac{1}{5}\right)^{v+1} \right) z^v. \end{aligned}$$

The preceding expansion implies that $b_v := (10/3)(2^{-(v+1)} - 5^{-(v+1)})$, $v \geq 0$, are the coefficients of the inverse causal filter.

2.2 Moving Averages and Autoregressive Processes

Let $a_1, \dots, a_q \in \mathbb{R}$ with $a_q \neq 0$ and let $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a white noise. The process

$$Y_t := \varepsilon_t + a_1 \varepsilon_{t-1} + \dots + a_q \varepsilon_{t-q}$$

is said to be a *moving average* of order q , denoted by $\text{MA}(q)$. Put $a_0 = 1$. Theorem 2.1.6 and 2.1.7 imply that a moving average $Y_t = \sum_{u=0}^q a_u \varepsilon_{t-u}$ is a stationary process with covariance generating function

$$\begin{aligned} G(z) &= \sigma^2 \left(\sum_{u=0}^q a_u z^u \right) \left(\sum_{w=0}^q a_w z^{-w} \right) \\ &= \sigma^2 \sum_{u=0}^q \sum_{w=0}^q a_u a_w z^{u-w} \\ &= \sigma^2 \sum_{v=-q}^q \sum_{u-w=v} a_u a_w z^{u-w} \\ &= \sigma^2 \sum_{v=-q}^q \left(\sum_{w=0}^{q-v} a_{v+w} a_w \right) z^v, \quad z \in \mathbb{C}, \end{aligned}$$

where $\sigma^2 = \text{Var}(\varepsilon_0)$. The coefficients of this expansion provide the autocovariance function $\gamma(v) = \text{Cov}(Y_0, Y_v)$, $v \in \mathbb{Z}$, which cuts off after lag q .

Lemma 2.2.1. *Suppose that $Y_t = \sum_{u=0}^q a_u \varepsilon_{t-u}$, $t \in \mathbb{Z}$, is a $\text{MA}(q)$ -process. Put $\mu := \mathbb{E}(\varepsilon_0)$ and $\sigma^2 := \text{Var}(\varepsilon_0)$. Then we have*

$$(i) \quad \mathbb{E}(Y_t) = \mu \sum_{u=0}^q a_u,$$

$$(ii) \quad \gamma(v) = \text{Cov}(Y_v, Y_0) = \begin{cases} 0, & v > q, \\ \sigma^2 \sum_{w=0}^{q-v} a_{v+w} a_w, & 0 \leq v \leq q, \end{cases}$$

$$\gamma(-v) = \gamma(v),$$

$$(iii) \quad \text{Var}(Y_0) = \gamma(0) = \sigma^2 \sum_{w=0}^q a_w^2,$$

$$(iv) \quad \rho(v) = \frac{\gamma(v)}{\gamma(0)} = \begin{cases} 0, & v > q, \\ \left(\sum_{w=0}^{q-v} a_{v+w} a_w \right) / \left(\sum_{w=0}^q a_w^2 \right), & 0 < v \leq q, \\ 1, & v = 0, \end{cases}$$

$$\rho(-v) = \rho(v).$$

Example 2.2.2. The MA(1)-process $Y_t = \varepsilon_t + a\varepsilon_{t-1}$ with $a \neq 0$ has the autocorrelation function

$$\rho(v) = \begin{cases} 1, & v = 0 \\ a/(1 + a^2), & v = \pm 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Since $a/(1 + a^2) = (1/a)/(1 + (1/a)^2)$, the autocorrelation functions of the two MA(1)-processes with parameters a and $1/a$ coincide. We have, moreover, $|\rho(1)| \leq 1/2$ for an arbitrary MA(1)-process and thus, a large value of the empirical autocorrelation function $r(1)$, which exceeds $1/2$ essentially, might indicate that an MA(1)-model for a given data set is not a correct assumption.

Invertible Processes

Example 2.2.2 shows that a MA(q)-process is not uniquely determined by its autocorrelation function. In order to get a unique relationship between moving average processes and their autocorrelation function, Box and Jenkins introduced the condition of invertibility. This is useful for estimation procedures, since the coefficients of an MA(q)-process will be estimated later by the empirical autocorrelation function, see Section 2.3.

The MA(q)-process $Y_t = \sum_{u=0}^q a_u \varepsilon_{t-u}$, with $a_0 = 1$ and $a_q \neq 0$, is said to be *invertible* if all q roots $z_1, \dots, z_q \in \mathbb{C}$ of $A(z) = \sum_{u=0}^q a_u z^u$ are outside of the unit circle i.e., if $|z_i| > 1$ for $1 \leq i \leq q$. Theorem 2.1.11 and representation (2.1) imply that the white noise process (ε_t) , pertaining to an invertible MA(q)-process $Y_t = \sum_{u=0}^q a_u \varepsilon_{t-u}$, can be obtained by means of an absolutely summable and causal filter $(b_u)_{u \geq 0}$ via

$$\varepsilon_t = \sum_{u \geq 0} b_u Y_{t-u}, \quad t \in \mathbb{Z},$$

with probability one. In particular the MA(1)-process $Y_t = \varepsilon_t - a\varepsilon_{t-1}$ is invertible iff $|a| < 1$, and in this case we have by Lemma 2.1.10 with probability one

$$\varepsilon_t = \sum_{u \geq 0} a^u Y_{t-u}, \quad t \in \mathbb{Z}.$$

Autoregressive Processes

A real valued stochastic process (Y_t) is said to be an *autoregressive process of order p* , denoted by $\text{AR}(p)$ if there exist $a_1, \dots, a_p \in \mathbb{R}$ with $a_p \neq 0$, and a white noise (ε_t) such that

$$Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}. \quad (2.3)$$

The value of an $\text{AR}(p)$ -process at time t is, therefore, regressed on its own past p values plus a random shock.

The Stationarity Condition

While by Theorem 2.1.6 MA(q)-processes are automatically stationary, this is not true for $\text{AR}(p)$ -processes (see Exercise 2.28). The following result provides a sufficient condition on the constants a_1, \dots, a_p implying the existence of a uniquely determined stationary solution (Y_t) of (2.3).

Theorem 2.2.3. *The $\text{AR}(p)$ -equation (2.3) with the given constants a_1, \dots, a_p and white noise $(\varepsilon_t)_{t \in \mathbb{Z}}$ has a stationary solution $(Y_t)_{t \in \mathbb{Z}}$ if all p roots of the equation $1 - a_1 z - a_2 z^2 - \dots - a_p z^p = 0$ are outside of the unit circle. In this case, the stationary solution is almost surely uniquely determined by*

$$Y_t := \sum_{u \geq 0} b_u \varepsilon_{t-u}, \quad t \in \mathbb{Z},$$

where $(b_u)_{u \geq 0}$ is the absolutely summable inverse causal filter of $c_0 = 1$, $c_u = -a_u$, $u = 1, \dots, p$ and $c_u = 0$ elsewhere.

Proof. The existence of an absolutely summable causal filter follows from Theorem 2.1.11. The stationarity of $Y_t = \sum_{u \geq 0} b_u \varepsilon_{t-u}$ is a consequence of Theorem 2.1.6, and its uniqueness follows from

$$\varepsilon_t = Y_t - a_1 Y_{t-1} - \dots - a_p Y_{t-p}, \quad t \in \mathbb{Z},$$

and equation (2.1) on page 58. \square

The condition that all roots of the characteristic equation of an AR(p)-process $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \varepsilon_t$ are outside of the unit circle i.e.,

$$1 - a_1 z - a_2 z^2 - \dots - a_p z^p \neq 0 \text{ for } |z| \leq 1, \quad (2.4)$$

will be referred to in the following as the *stationarity condition* for an AR(p)-process.

An AR(p) process satisfying the stationarity condition can be interpreted as a MA(∞) process.

Note that a stationary solution (Y_t) of (2.1) exists in general if no root z_i of the characteristic equation lies on the unit sphere. If there are solutions in the unit circle, then the stationary solution is noncausal, i.e., Y_t is correlated with future values of ε_s , $s > t$. This is frequently regarded as unnatural.

Example 2.2.4. The AR(1)-process $Y_t = aY_{t-1} + \varepsilon_t$, $t \in \mathbb{Z}$, with $a \neq 0$ has the characteristic equation $1 - az = 0$ with the obvious solution $z_1 = 1/a$. The process (Y_t) , therefore, satisfies the stationarity condition iff $|z_1| > 1$ i.e., iff $|a| < 1$. In this case we obtain from Lemma 2.1.10 that the absolutely summable inverse causal filter of $a_0 = 1$, $a_1 = -a$ and $a_u = 0$ elsewhere is given by $b_u = a^u$, $u \geq 0$, and thus, with probability one

$$Y_t = \sum_{u \geq 0} b_u \varepsilon_{t-u} = \sum_{u \geq 0} a^u \varepsilon_{t-u}.$$

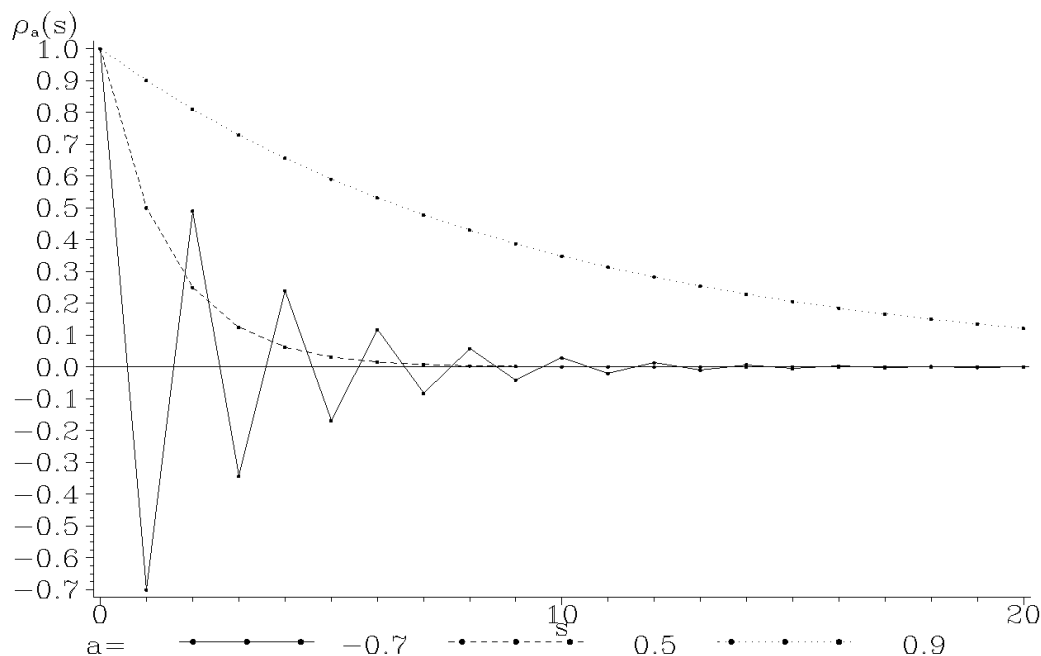
Denote by σ^2 the variance of ε_0 . From Theorem 2.1.6 we obtain the autocovariance function of (Y_t)

$$\begin{aligned} \gamma(s) &= \sum_u \sum_w b_u b_w \text{Cov}(\varepsilon_0, \varepsilon_{s+w-u}) \\ &= \sum_{u \geq 0} b_u b_{u-s} \text{Cov}(\varepsilon_0, \varepsilon_0) \\ &= \sigma^2 a^s \sum_{u \geq s} a^{2(u-s)} = \sigma^2 \frac{a^s}{1 - a^2}, \quad s = 0, 1, 2, \dots \end{aligned}$$

and $\gamma(-s) = \gamma(s)$. In particular we obtain $\gamma(0) = \sigma^2/(1 - a^2)$ and thus, the autocorrelation function of (Y_t) is given by

$$\rho(s) = a^{|s|}, \quad s \in \mathbb{Z}.$$

The autocorrelation function of an AR(1)-process $Y_t = aY_{t-1} + \varepsilon_t$ with $|a| < 1$ therefore decreases at an exponential rate. Its sign is alternating if $a \in (-1, 0)$.



Plot 2.2.1: Autocorrelation functions of AR(1)-processes $Y_t = aY_{t-1} + \varepsilon_t$ with different values of a .

```

1 /* arl_autocorrelation.sas */
2 TITLE1 'Autocorrelation functions of AR(1)-processes';
3
4 /* Generate data for different autocorrelation functions */
5 DATA data1;
6   DO a=-0.7, 0.5, 0.9;
7     DO s=0 TO 20;
8       rho=a**s;
9     OUTPUT;

```

```

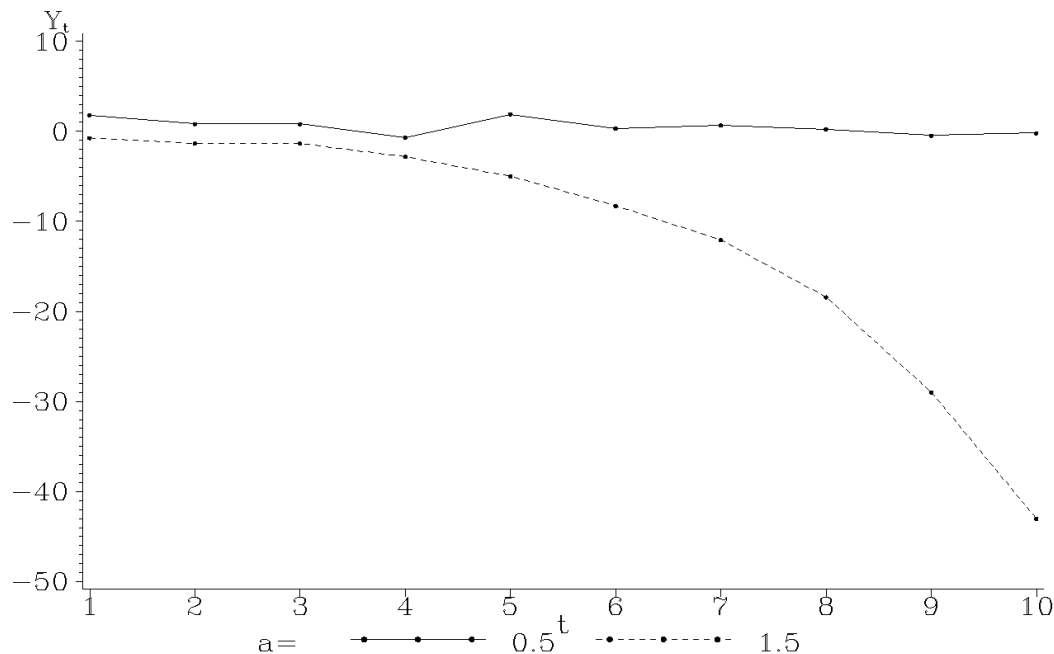
10      END;
11      END;
12
13      /* Graphical options */
14      SYMBOL1 C=GREEN V=DOT I=JOIN H=0.3 L=1;
15      SYMBOL2 C=GREEN V=DOT I=JOIN H=0.3 L=2;
16      SYMBOL3 C=GREEN V=DOT I=JOIN H=0.3 L=33;
17      AXIS1 LABEL=(' s' );
18      AXIS2 LABEL=(F=CGREEK 'r' F=COMPLEX H=1 'a' H=2 '(s)');
19      LEGEND1 LABEL=(' a=' ) SHAPE=SYMBOL(10,0.6);
20
21      /* Plot autocorrelation functions */
22      PROC GPLOT DATA=data1;
23          PLOT rho*s=a / HAXIS=AXIS1 VAXIS=AXIS2 LEGEND=LEGEND1 VREF=0;
24      RUN; QUIT;

```

The data step evaluates ρ for three different values of a and the range of s from 0 to 20 using two loops. The plot is generated by the procedure GPLOT. The LABEL option in the AXIS2 statement uses, in addition to the greek font CGREEK, the font COMPLEX

assuming this to be the default text font (GOPTION FTEXT=COMPLEX). The SHAPE option SHAPE=SYMBOL(10,0.6) in the LEGEND statement defines width and height of the symbols presented in the legend.

The following figure illustrates the significance of the stationarity condition $|a| < 1$ of an AR(1)-process. Realizations $Y_t = aY_{t-1} + \varepsilon_t$, $t = 1, \dots, 10$, are displayed for $a = 0.5$ and $a = 1.5$, where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{10}$ are independent standard normal in each case and Y_0 is assumed to be zero. While for $a = 0.5$ the sample path follows the constant zero closely, which is the expectation of each Y_t , the observations Y_t decrease rapidly in case of $a = 1.5$.



Plot 2.2.2: Realizations of the AR(1)-processes $Y_t = 0.5Y_{t-1} + \varepsilon_t$ and $Y_t = 1.5Y_{t-1} + \varepsilon_t$, $t = 1, \dots, 10$, with ε_t independent standard normal and $Y_0 = 0$.

```

1  /* arl_plot.sas */
2  TITLE1 'Realizations of AR(1)-processes';
3
4  /* Generated AR(1)-processes */
5  DATA datal;
6      DO a=0.5, 1.5;
7          t=0; y=0; OUTPUT;
8          DO t=1 TO 10;
9              y=a*y+RANNOR(1);
10             OUTPUT;
11         END;
12     END;
13
14  /* Graphical options */
15  SYMBOL1 C=GREEN V=DOT I=JOIN H=0.4 L=1;
16  SYMBOL2 C=GREEN V=DOT I=JOIN H=0.4 L=2;
17  AXIS1 LABEL=('t') MINOR=NONE;
18  AXIS2 LABEL=('Y' H=1 't');
19  LEGEND1 LABEL=('a=') SHAPE=SYMBOL(10,0.6);
20
21  /* Plot the AR(1)-processes */
22  PROC GPLOT DATA=datall(WHERE=(t>0));
23      PLOT y*t=a / HAXIS=AXIS1 VAXIS=AXIS2 LEGEND=LEGEND1;
24  RUN; QUIT;

```

The data are generated within two loops, the first one over the two values for a . The variable y is initialized with the value 0 corresponding to $t=0$. The realizations for $t=1, \dots, 10$ are created within the second loop over t and with the help of the function `RANNOR` which returns pseudo random numbers distributed as standard normal. The argument 1 is the initial seed to produce a stream of random numbers. A positive value of this seed always produces the same series of random numbers, a negative value generates a different series each time

the program is submitted. A value of y is calculated as the sum of a times the actual value of y and the random number and stored in a new observation. The resulting data set has 22 observations and 3 variables (a , t and y). In the plot created by `PROC GPLOT` the initial observations are dropped using the `WHERE` data set option. Only observations fulfilling the condition $t > 0$ are read into the data set used here. To suppress minor tick marks between the integers 0, 1, \dots , 10 the option `MINOR` in the `AXIS1` statement is set to `NONE`.

The Yule–Walker Equations

The Yule–Walker equations entail the recursive computation of the autocorrelation function ρ of an $AR(p)$ -process satisfying the stationarity condition (2.4).

Lemma 2.2.5. *Let $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \varepsilon_t$ be an $AR(p)$ -process, which satisfies the stationarity condition (2.4). Its autocorrelation function ρ then satisfies for $s = 1, 2, \dots$ the recursion*

$$\rho(s) = \sum_{u=1}^p a_u \rho(s - u), \quad (2.5)$$

known as Yule–Walker equations.

Proof. Put $\mu := E(Y_0)$ and $\nu := E(\varepsilon_0)$. Recall that $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \varepsilon_t$, $t \in \mathbb{Z}$ and taking expectations on both sides $\mu = \sum_{u=1}^p a_u \mu + \nu$. Combining both equations yields

$$Y_t - \mu = \sum_{u=1}^p a_u (Y_{t-u} - \mu) + \varepsilon_t - \nu, \quad t \in \mathbb{Z}. \quad (2.6)$$

By multiplying equation (2.6) with $Y_{t-s} - \mu$ for $s > 0$ and taking

expectations again we obtain

$$\begin{aligned}\gamma(s) &= E((Y_t - \mu)(Y_{t-s} - \mu)) \\ &= \sum_{u=1}^p a_u E((Y_{t-u} - \mu)(Y_{t-s} - \mu)) + E((\varepsilon_t - \nu)(Y_{t-s} - \mu)) \\ &= \sum_{u=1}^p a_u \gamma(s - u).\end{aligned}$$

for the autocovariance function γ of (Y_t) . The final equation follows from the fact that Y_{t-s} and ε_t are uncorrelated for $s > 0$. This is a consequence of Theorem 2.2.3, by which almost surely $Y_{t-s} = \sum_{u \geq 0} b_u \varepsilon_{t-s-u}$ with an absolutely summable causal filter (b_u) and thus, $\text{Cov}(Y_{t-s}, \varepsilon_t) = \sum_{u \geq 0} b_u \text{Cov}(\varepsilon_{t-s-u}, \varepsilon_t) = 0$, see Theorem 2.1.5 and Exercise 2.16. Dividing the above equation by $\gamma(0)$ now yields the assertion. \square

Since $\rho(-s) = \rho(s)$, equations (2.5) can be represented as

$$\begin{pmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ \vdots \\ \rho(p) \end{pmatrix} = \begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(p-1) \\ \rho(1) & 1 & \rho(1) & & \rho(p-2) \\ \rho(2) & \rho(1) & 1 & & \rho(p-3) \\ \vdots & & & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \rho(p-3) & \dots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} \quad (2.7)$$

This matrix equation offers an estimator of the coefficients a_1, \dots, a_p by replacing the autocorrelations $\rho(j)$ by their empirical counterparts $r(j)$, $1 \leq j \leq p$. Equation (2.7) then formally becomes $\mathbf{r} = \mathbf{R}\mathbf{a}$, where $\mathbf{r} = (r(1), \dots, r(p))^T$, $\mathbf{a} = (a_1, \dots, a_p)^T$ and

$$\mathbf{R} := \begin{pmatrix} 1 & r(1) & r(2) & \dots & r(p-1) \\ r(1) & 1 & r(1) & \dots & r(p-2) \\ \vdots & & & \ddots & \vdots \\ r(p-1) & r(p-2) & r(p-3) & \dots & 1 \end{pmatrix}.$$

If the $p \times p$ -matrix \mathbf{R} is invertible, we can rewrite the formal equation $\mathbf{r} = \mathbf{R}\mathbf{a}$ as $\mathbf{R}^{-1}\mathbf{r} = \mathbf{a}$, which motivates the estimator

$$\hat{\mathbf{a}} := \mathbf{R}^{-1}\mathbf{r} \quad (2.8)$$

of the vector $\mathbf{a} = (a_1, \dots, a_p)^T$ of the coefficients.

The Partial Autocorrelation Coefficients

We have seen that the autocorrelation function $\rho(k)$ of an MA(q)-process vanishes for $k > q$, see Lemma 2.2.1. This is not true for an AR(p)-process, whereas the partial autocorrelation coefficients will share this property. Note that the correlation matrix

$$\begin{aligned} \mathbf{P}_k &:= \left(\text{Corr}(Y_i, Y_j) \right)_{1 \leq i, j \leq k} \\ &= \begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-1) \\ \rho(1) & 1 & \rho(1) & & \rho(k-2) \\ \rho(2) & \rho(1) & 1 & & \rho(k-3) \\ \vdots & & & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \rho(k-3) & \dots & 1 \end{pmatrix} \end{aligned} \quad (2.9)$$

is positive semidefinite for any $k \geq 1$. If we suppose that \mathbf{P}_k is positive definite, then it is invertible, and the equation

$$\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(k) \end{pmatrix} = \mathbf{P}_k \begin{pmatrix} a_{k1} \\ \vdots \\ a_{kk} \end{pmatrix} \quad (2.10)$$

has the unique solution

$$\mathbf{a}_k := \begin{pmatrix} a_{k1} \\ \vdots \\ a_{kk} \end{pmatrix} = \mathbf{P}_k^{-1} \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(k) \end{pmatrix}.$$

The number a_{kk} is called *partial autocorrelation coefficient at lag k* , denoted by $\alpha(k)$, $k \geq 1$. Observe that for $k \geq p$ the vector $(a_1, \dots, a_p, 0, \dots, 0) \in \mathbb{R}^k$, with $k - p$ zeros added to the vector of coefficients (a_1, \dots, a_p) , is by the Yule–Walker equations (2.5) a solution of the equation (2.10). Thus we have $\alpha(p) = a_p$, $\alpha(k) = 0$ for $k > p$. Note that the coefficient $\alpha(k)$ also occurs as the coefficient of Y_{n-k} in the best linear one-step forecast $\sum_{u=0}^k c_u Y_{n-u}$ of Y_{n+1} , see equation (2.27) in Section 2.3.

If the empirical counterpart \mathbf{R}_k of \mathbf{P}_k is invertible as well, then

$$\hat{\mathbf{a}}_k := \mathbf{R}_k^{-1} \mathbf{r}_k,$$

with $\mathbf{r}_k := (r(1), \dots, r(k))^T$, is an obvious estimate of \mathbf{a}_k . The k -th component

$$\hat{\alpha}(k) := \hat{a}_{kk} \quad (2.11)$$

of $\hat{\mathbf{a}}_k = (\hat{a}_{k1}, \dots, \hat{a}_{kk})$ is the *empirical partial autocorrelation coefficient at lag k* . It can be utilized to estimate the order p of an AR(p)-process, since $\hat{\alpha}(p) \approx \alpha(p) = a_p$ is different from zero, whereas $\hat{\alpha}(k) \approx \alpha(k) = 0$ for $k > p$ should be close to zero.

Example 2.2.6. The Yule–Walker equations (2.5) for an AR(2)-process $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \varepsilon_t$ are for $s = 1, 2$

$$\rho(1) = a_1 + a_2 \rho(1), \quad \rho(2) = a_1 \rho(1) + a_2$$

with the solutions

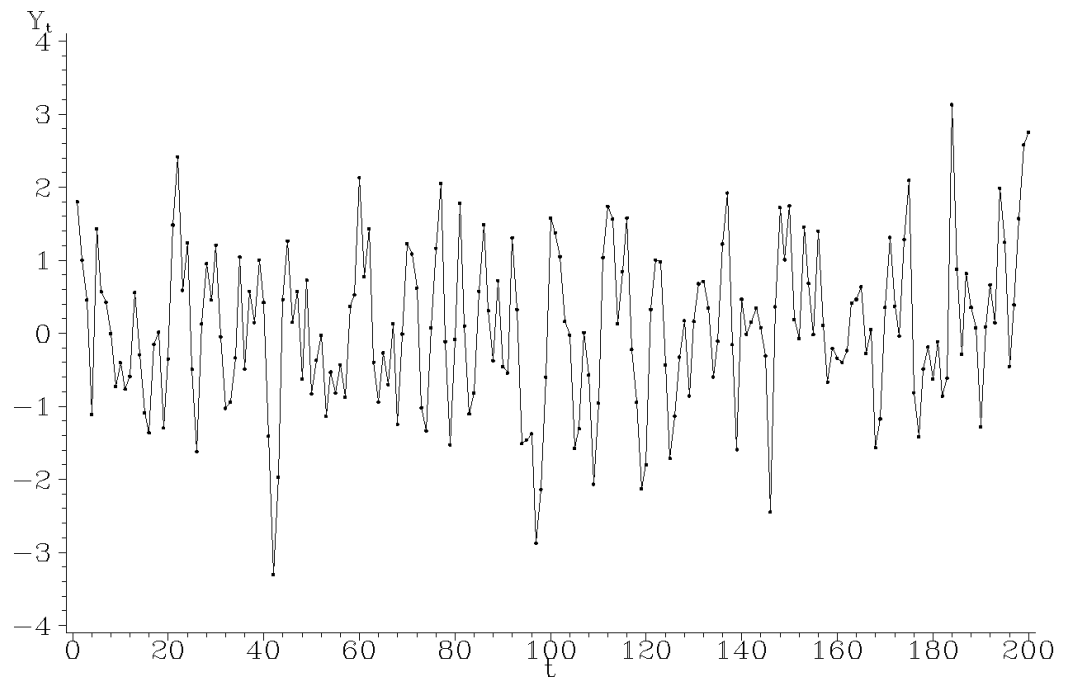
$$\rho(1) = \frac{a_1}{1 - a_2}, \quad \rho(2) = \frac{a_1^2}{1 - a_2} + a_2.$$

and thus, the partial autocorrelation coefficients are

$$\begin{aligned} \alpha(1) &= \rho(1), \\ \alpha(2) &= a_2, \\ \alpha(j) &= 0, \quad j \geq 3. \end{aligned}$$

The recursion (2.5) entails the computation of $\rho(s)$ for an arbitrary s from the two values $\rho(1)$ and $\rho(2)$.

The following figure displays realizations of the AR(2)-process $Y_t = 0.6Y_{t-1} - 0.3Y_{t-2} + \varepsilon_t$ for $1 \leq t \leq 200$, conditional on $Y_{-1} = Y_0 = 0$. The random shocks ε_t are iid standard normal. The corresponding empirical partial autocorrelation function is shown in Plot 2.2.4.



Plot 2.2.3: Realization of the AR(2)-process $Y_t = 0.6Y_{t-1} - 0.3Y_{t-2} + \varepsilon_t$, conditional on $Y_{-1} = Y_0 = 0$. The ε_t , $1 \leq t \leq 200$, are iid standard normal.

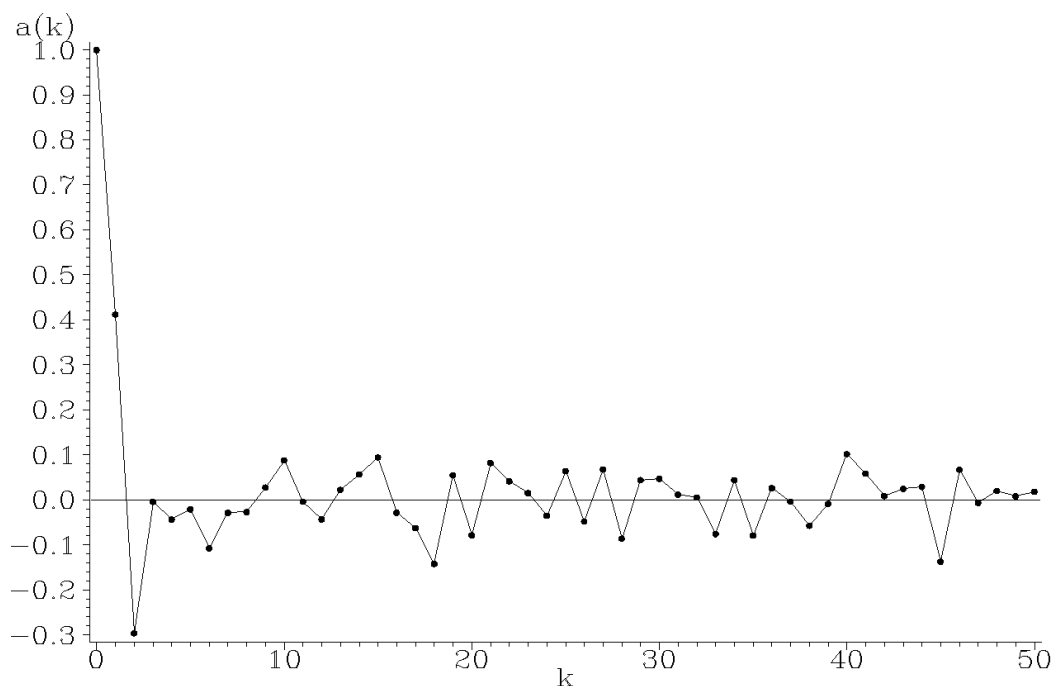
```

1  /* ar2_plot.sas */
2  TITLE1 'Realisation of an AR(2)-process';
3
4  /* Generated AR(2)-process */
5  DATA data1;
6      t=-1; y=0; OUTPUT;
7      t=0; y1=y; y=0; OUTPUT;
8      DO t=1 TO 200;
9          y2=y1;
10         y1=y;
11         y=0.6*y1-0.3*y2+RANNOR(1);
12         OUTPUT;
13     END;
14
15  /* Graphical options */
16  SYMBOL1 C=GREEN V=DOT I=JOIN H=0.3;
17  AXIS1 LABEL=('t');
18  AXIS2 LABEL=('Y' H=1 't');
19
20  /* Plot the AR(2)-processes */
21  PROC GPLOT DATA=data1 (WHERE=(t>0));
22      PLOT y*t / HAXIS=AXIS1 VAXIS=AXIS2;
23  RUN; QUIT;

```


The two initial values of y are defined and stored in an observation by the `OUTPUT` statement. The second observation contains an additional value y_1 for y_{t-1} . Within the loop the

values y_2 (for y_{t-2}), y_1 and y are updated one after the other. The data set used by `PROC GPLOT` again just contains the observations with $t > 0$.



Plot 2.2.4: Empirical partial autocorrelation function of the AR(2)-data in Plot 2.2.3.

```

1  /* ar2_epa.sas */
2  TITLE1 'Empirical partial autocorrelation function';
3  TITLE2 'of simulated AR(2)-process data';
4  /* Note that this program requires data1 generated by the previous
   ↪ program (ar2_plot.sas) */
5
6  /* Compute partial autocorrelation function */
7  PROC ARIMA DATA=data1 (WHERE=(t>0));
8      IDENTIFY VAR=y NLAG=50 OUTCOV=corr NOPRINT;
9
10 /* Graphical options */
11 SYMBOL1 C=GREEN V=DOT I=JOIN H=0.7;
12 AXIS1 LABEL=('k');

```

```

13 AXIS2 LABEL= (' a (k) ' ) ;
14
15 /* Plot autocorrelation function */
16 PROC Gplot DATA=corr;
17   PLOT PARTCORR*LAG / HAXIS=AXIS1 VAXIS=AXIS2 VREF=0;
18 RUN; QUIT;

```

This program requires to be submitted to SAS for execution within a joint session with Program 2.2.3 (ar2_plot.sas), because it uses the temporary data step `data1` generated there. Otherwise you have to add the block of statements to this program concerning the data step. Like in Program 1.3.1 (sunspot_correlogram.sas)

the procedure `ARIMA` with the `IDENTIFY` statement is used to create a data set. Here we are interested in the variable `PARTCORR` containing the values of the empirical partial autocorrelation function from the simulated `AR(2)`-process data. This variable is plotted against the lag stored in variable `LAG`.

ARMA-Processes

Moving averages $MA(q)$ and autoregressive $AR(p)$ -processes are special cases of so called autoregressive moving averages. Let $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a white noise, $p, q \geq 0$ integers and $a_0, \dots, a_p, b_0, \dots, b_q \in \mathbb{R}$. A real valued stochastic process $(Y_t)_{t \in \mathbb{Z}}$ is said to be an *autoregressive moving average process of order p, q* , denoted by $ARMA(p, q)$, if it satisfies the equation

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \dots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q}. \quad (2.12)$$

An $ARMA(p, 0)$ -process with $p \geq 1$ is obviously an $AR(p)$ -process, whereas an $ARMA(0, q)$ -process with $q \geq 1$ is a moving average $MA(q)$. The polynomials

$$A(z) := 1 - a_1 z - \dots - a_p z^p \quad (2.13)$$

and

$$B(z) := 1 + b_1 z + \dots + b_q z^q, \quad (2.14)$$

are the characteristic polynomials of the autoregressive part and of the moving average part of an $ARMA(p, q)$ -process (Y_t) , which we can represent in the form

$$Y_t - a_1 Y_{t-1} - \dots - a_p Y_{t-p} = \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q}.$$

Denote by Z_t the right-hand side of the above equation i.e., $Z_t := \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q}$. This is a MA(q)-process and, therefore, stationary by Theorem 2.1.6. If all p roots of the equation $A(z) = 1 - a_1z - \cdots - a_pz^p = 0$ are outside of the unit circle, then we deduce from Theorem 2.1.11 that the filter $c_0 = 1$, $c_u = -a_u$, $u = 1, \dots, p$, $c_u = 0$ elsewhere, has an absolutely summable causal inverse filter $(d_u)_{u \geq 0}$. Consequently we obtain from the equation $Z_t = Y_t - a_1Y_{t-1} - \cdots - a_pY_{t-p}$ and (2.1) on page 58 that with $b_0 = 1$, $b_w = 0$ if $w > q$

$$\begin{aligned} Y_t &= \sum_{u \geq 0} d_u Z_{t-u} = \sum_{u \geq 0} d_u (\varepsilon_{t-u} + b_1\varepsilon_{t-1-u} + \cdots + b_q\varepsilon_{t-q-u}) \\ &= \sum_{u \geq 0} \sum_{w \geq 0} d_u b_w \varepsilon_{t-w-u} = \sum_{v \geq 0} \left(\sum_{u+w=v} d_u b_w \right) \varepsilon_{t-v} \\ &= \sum_{v \geq 0} \left(\sum_{w=0}^{\min(v,q)} b_w d_{v-w} \right) \varepsilon_{t-v} =: \sum_{v \geq 0} \alpha_v \varepsilon_{t-v} \end{aligned}$$

is the almost surely uniquely determined stationary solution of the ARMA(p, q)-equation (2.12) for a given white noise (ε_t) .

The condition that all p roots of the characteristic equation $A(z) = 1 - a_1z - a_2z^2 - \cdots - a_pz^p = 0$ of the ARMA(p, q)-process (Y_t) are outside of the unit circle will again be referred to in the following as the *stationarity condition* (2.4).

The MA(q)-process $Z_t = \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q}$ is by definition invertible if all q roots of the polynomial $B(z) = 1 + b_1z + \cdots + b_qz^q$ are outside of the unit circle. Theorem 2.1.11 and equation (2.1) imply in this case the existence of an absolutely summable causal filter $(g_u)_{u \geq 0}$ such that with $a_0 = -1$

$$\begin{aligned} \varepsilon_t &= \sum_{u \geq 0} g_u Z_{t-u} = \sum_{u \geq 0} g_u (Y_{t-u} - a_1Y_{t-1-u} - \cdots - a_pY_{t-p-u}) \\ &= - \sum_{v \geq 0} \left(\sum_{w=0}^{\min(v,p)} a_w g_{v-w} \right) Y_{t-v}. \end{aligned}$$

In this case the ARMA(p, q)-process (Y_t) is said to be *invertible*.

The Autocovariance Function of an ARMA-Process

In order to deduce the autocovariance function of an ARMA(p, q)-process (Y_t) , which satisfies the stationarity condition (2.4), we compute at first the absolutely summable coefficients

$$\alpha_v = \sum_{w=0}^{\min(q,v)} b_w d_{v-w}, \quad v \geq 0,$$

in the above representation $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$. The characteristic polynomial $D(z)$ of the absolutely summable causal filter $(d_u)_{u \geq 0}$ coincides by Lemma 2.1.9 for $0 < |z| < 1$ with $1/A(z)$, where $A(z)$ is given in (2.13). Thus we obtain with $B(z)$ as given in (2.14) for $0 < |z| < 1$, where we set this time $a_0 := -1$ to simplify the following formulas,

$$\begin{aligned} & A(z)(B(z)D(z)) = B(z) \\ \Leftrightarrow & \left(- \sum_{u=0}^p a_u z^u \right) \left(\sum_{v \geq 0} \alpha_v z^v \right) = \sum_{w=0}^q b_w z^w \\ \Leftrightarrow & \sum_{w \geq 0} \left(- \sum_{u+v=w} a_u \alpha_v \right) z^w = \sum_{w \geq 0} b_w z^w \\ \Leftrightarrow & \sum_{w \geq 0} \left(- \sum_{u=0}^w a_u \alpha_{w-u} \right) z^w = \sum_{w \geq 0} b_w z^w \\ \Leftrightarrow & \begin{cases} \alpha_0 = 1 \\ \alpha_w - \sum_{u=1}^w a_u \alpha_{w-u} = b_w & \text{for } 1 \leq w \leq p \\ \alpha_w - \sum_{u=1}^p a_u \alpha_{w-u} = b_w & \text{for } w > p \text{ with } b_w = 0 \text{ for } w > q. \end{cases} \end{aligned} \tag{2.15}$$

Example 2.2.7. For the ARMA(1, 1)-process $Y_t - aY_{t-1} = \varepsilon_t + b\varepsilon_{t-1}$ with $|a| < 1$ we obtain from (2.15)

$$\alpha_0 = 1, \quad \alpha_1 - a = b, \quad \alpha_w - a\alpha_{w-1} = 0, \quad w \geq 2,$$

This implies $\alpha_0 = 1$, $\alpha_w = a^{w-1}(b + a)$, $w \geq 1$, and, hence,

$$Y_t = \varepsilon_t + (b + a) \sum_{w \geq 1} a^{w-1} \varepsilon_{t-w}.$$

Theorem 2.2.8. *Suppose that $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \sum_{v=0}^q b_v \varepsilon_{t-v}$, $b_0 := 1$, is an ARMA(p, q)-process, which satisfies the stationarity condition (2.4). Its autocovariance function γ then satisfies the recursion*

$$\begin{aligned} \gamma(s) - \sum_{u=1}^p a_u \gamma(s-u) &= \sigma^2 \sum_{v=s}^q b_v \alpha_{v-s}, \quad 0 \leq s \leq q, \\ \gamma(s) - \sum_{u=1}^p a_u \gamma(s-u) &= 0, \quad s \geq q+1, \end{aligned} \quad (2.16)$$

where α_v , $v \geq 0$, are the coefficients in the representation $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$, which we computed in (2.15) and σ^2 is the variance of ε_0 .

Consequently the autocorrelation function ρ of the ARMA(p, q) process (Y_t) satisfies

$$\rho(s) = \sum_{u=1}^p a_u \rho(s-u), \quad s \geq q+1,$$

which coincides with the autocorrelation function of the stationary AR(p)-process $X_t = \sum_{u=1}^p a_u X_{t-u} + \varepsilon_t$, c.f. Lemma 2.2.5.

Proof of Theorem 2.2.8. Put $\mu := E(Y_0)$ and $\nu := E(\varepsilon_0)$. Recall that $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \sum_{v=0}^q b_v \varepsilon_{t-v}$, $t \in \mathbb{Z}$ and taking expectations on both sides $\mu = \sum_{u=1}^p a_u \mu + \sum_{v=0}^q b_v \nu$. Combining both equations yields

$$Y_t - \mu = \sum_{u=1}^p a_u (Y_{t-u} - \mu) + \sum_{v=0}^q b_v (\varepsilon_{t-v} - \nu), \quad t \in \mathbb{Z}.$$

Multiplying both sides with $Y_{t-s} - \mu$, $s \geq 0$, and taking expectations, we obtain

$$\text{Cov}(Y_{t-s}, Y_t) = \sum_{u=1}^p a_u \text{Cov}(Y_{t-s}, Y_{t-u}) + \sum_{v=0}^q b_v \text{Cov}(Y_{t-s}, \varepsilon_{t-v}),$$

which implies

$$\gamma(s) - \sum_{u=1}^p a_u \gamma(s-u) = \sum_{v=0}^q b_v \operatorname{Cov}(Y_{t-s}, \varepsilon_{t-v}).$$

From the representation $Y_{t-s} = \sum_{w \geq 0} \alpha_w \varepsilon_{t-s-w}$ and Theorem 2.1.5 we obtain

$$\operatorname{Cov}(Y_{t-s}, \varepsilon_{t-v}) = \sum_{w \geq 0} \alpha_w \operatorname{Cov}(\varepsilon_{t-s-w}, \varepsilon_{t-v}) = \begin{cases} 0 & \text{if } v < s \\ \sigma^2 \alpha_{v-s} & \text{if } v \geq s. \end{cases}$$

This implies

$$\begin{aligned} \gamma(s) - \sum_{u=1}^p a_u \gamma(s-u) &= \sum_{v=s}^q b_v \operatorname{Cov}(Y_{t-s}, \varepsilon_{t-v}) \\ &= \begin{cases} \sigma^2 \sum_{v=s}^q b_v \alpha_{v-s} & \text{if } s \leq q \\ 0 & \text{if } s > q, \end{cases} \end{aligned}$$

which is the assertion. \square

Example 2.2.9. For the ARMA(1,1)-process $Y_t - aY_{t-1} = \varepsilon_t + b\varepsilon_{t-1}$ with $|a| < 1$ we obtain from Example 2.2.7 and Theorem 2.2.8 with $\sigma^2 = \operatorname{Var}(\varepsilon_0)$

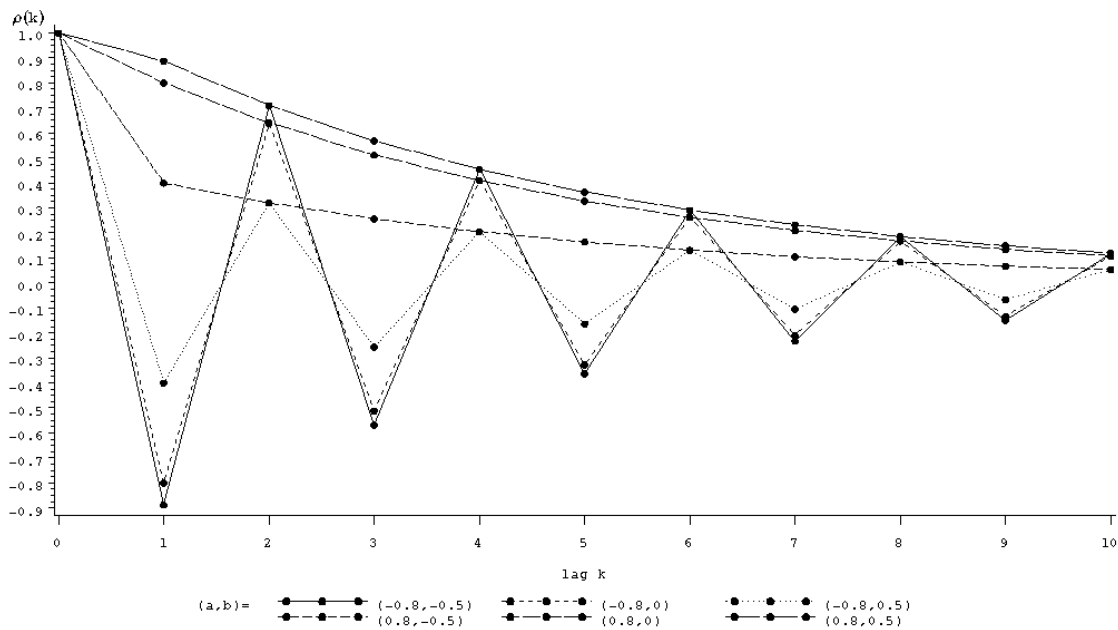
$$\gamma(0) - a\gamma(1) = \sigma^2(1 + b(b+a)), \quad \gamma(1) - a\gamma(0) = \sigma^2 b,$$

and thus

$$\gamma(0) = \sigma^2 \frac{1 + 2ab + b^2}{1 - a^2}, \quad \gamma(1) = \sigma^2 \frac{(1 + ab)(a + b)}{1 - a^2}.$$

For $s \geq 2$ we obtain from (2.16)

$$\gamma(s) = a\gamma(s-1) = \cdots = a^{s-1}\gamma(1).$$



Plot 2.2.5: Autocorrelation functions of ARMA(1,1)-processes with $a = 0.8/-0.8$, $b = 0.5/0/-0.5$ and $\sigma^2 = 1$.

```

1  /* armall_autocorrelation.sas */
2  TITLE1 'Autocorrelation functions of ARMA(1,1)-processes';
3
4  /* Compute autocorrelations functions for different ARMA(1,1)-
   ↪ processes */
5  DATA datal;
6  DO a=-0.8, 0.8;
7    DO b=-0.5, 0, 0.5;
8      s=0; rho=1;
9      q=COMPRESS('(' || a || ',' || b || ')');
10     OUTPUT;
11     s=1; rho=(1+a*b)*(a+b)/(1+2*a*b+b*b);
12     q=COMPRESS('(' || a || ',' || b || ')');
13     OUTPUT;
14     DO s=2 TO 10;
15       rho=a*rho;
16       q=COMPRESS('(' || a || ',' || b || ')');
17       OUTPUT;
18     END;
19   END;
20 END;
21
22 /* Graphical options */
23 SYMBOL1 C=RED V=DOT I=JOIN H=0.7 L=1;
24 SYMBOL2 C=YELLOW V=DOT I=JOIN H=0.7 L=2;
25 SYMBOL3 C=BLUE V=DOT I=JOIN H=0.7 L=33;
26 SYMBOL4 C=RED V=DOT I=JOIN H=0.7 L=3;
27 SYMBOL5 C=YELLOW V=DOT I=JOIN H=0.7 L=4;

```

```

28 SYMBOL6 C=BLUE V=DOT I=JOIN H=0.7 L=5;
29 AXIS1 LABEL=(F=CGREEK 'r' F=COMPLEX '(k)');
30 AXIS2 LABEL=('lag k') MINOR=NONE;
31 LEGEND1 LABEL=(' (a,b)=') SHAPE=SYMBOL(10,0.8);
32
33 /* Plot the autocorrelation functions */
34 PROC GPLOT DATA=data1;
35     PLOT rho*s=q / VAXIS=AXIS1 HAXIS=AXIS2 LEGEND=LEGEND1;
36 RUN; QUIT;

```

In the data step the values of the autocorrelation function belonging to an ARMA(1,1) process are calculated for two different values of a , the coefficient of the AR(1)-part, and three different values of b , the coefficient of the MA(1)-part. Pure AR(1)-processes result for the value $b=0$. For the arguments (lags) $s=0$ and $s=1$ the computation is done directly, for the rest up

to $s=10$ a loop is used for a recursive computation. For the COMPRESS statement see Program 1.1.3 (logistic.sas).

The second part of the program uses PROC GPLOT to plot the autocorrelation function, using known statements and options to customize the output.

ARIMA-Processes

Suppose that the time series (Y_t) has a polynomial trend of degree d . Then we can eliminate this trend by considering the process $(\Delta^d Y_t)$, obtained by d times differencing as described in Section 1.2. If the filtered process $(\Delta^d Y_d)$ is an ARMA(p, q)-process satisfying the stationarity condition (2.4), the original process (Y_t) is said to be an *autoregressive integrated moving average of order p, d, q* , denoted by ARIMA(p, d, q). In this case constants $a_1, \dots, a_p, b_0 = 1, b_1, \dots, b_q \in \mathbb{R}$ exist such that

$$\Delta^d Y_t = \sum_{u=1}^p a_u \Delta^d Y_{t-u} + \sum_{w=0}^q b_w \varepsilon_{t-w}, \quad t \in \mathbb{Z},$$

where (ε_t) is a white noise.

Example 2.2.10. An ARIMA(1, 1, 1)-process (Y_t) satisfies

$$\Delta Y_t = a \Delta Y_{t-1} + \varepsilon_t + b \varepsilon_{t-1}, \quad t \in \mathbb{Z},$$

where $|a| < 1$, $b \neq 0$ and (ε_t) is a white noise, i.e.,

$$Y_t - Y_{t-1} = a(Y_{t-1} - Y_{t-2}) + \varepsilon_t + b \varepsilon_{t-1}, \quad t \in \mathbb{Z}.$$

This implies $Y_t = (a + 1)Y_{t-1} - aY_{t-2} + \varepsilon_t + b\varepsilon_{t-1}$. Note that the characteristic polynomial of the AR-part of this ARMA(2, 1)-process has a root 1 and the process is, thus, not stationary.

A *random walk* $X_t = X_{t-1} + \varepsilon_t$ is obviously an ARIMA(0, 1, 0)-process.

Consider $Y_t = S_t + R_t$, $t \in \mathbb{Z}$, where the random component (R_t) is a stationary process and the seasonal component (S_t) is periodic of length s , i.e., $S_t = S_{t+s} = S_{t+2s} = \dots$ for $t \in \mathbb{Z}$. Then the process (Y_t) is in general not stationary, but $Y_t^* := Y_t - Y_{t-s}$ is. If this seasonally adjusted process (Y_t^*) is an ARMA(p, q)-process satisfying the stationarity condition (2.4), then the original process (Y_t) is called a *seasonal* ARMA(p, q)-process with period length s , denoted by $\text{SARMA}_s(p, q)$. One frequently encounters a time series with a trend as well as a periodic seasonal component. A stochastic process (Y_t) with the property that $(\Delta^d(Y_t - Y_{t-s}))$ is an ARMA(p, q)-process is, therefore, called a SARIMA(p, d, q)-process. This is a quite common assumption in practice.

Cointegration

In the sequel we will frequently use the notation that a time series (Y_t) is $I(d)$, $d = 0, 1, \dots$ if the sequence of differences $(\Delta^d Y_t)$ of order d is a stationary process. By the difference $\Delta^0 Y_t$ of order zero we denote the undifferenced process Y_t , $t \in \mathbb{Z}$.

Suppose that the two time series (Y_t) and (Z_t) satisfy

$$Y_t = aW_t + \varepsilon_t, \quad Z_t = W_t + \delta_t, \quad t \in \mathbb{Z},$$

for some real number $a \neq 0$, where (W_t) is $I(1)$, and (ε_t) , (δ_t) are uncorrelated white noise processes, i.e., $\text{Cov}(\varepsilon_t, \delta_s) = 0$, $t, s \in \mathbb{Z}$, and both are uncorrelated to (W_t) .

Then (Y_t) and (Z_t) are both $I(1)$, but

$$X_t := Y_t - aZ_t = \varepsilon_t - a\delta_t, \quad t \in \mathbb{Z},$$

is $I(0)$.

The fact that the combination of two nonstationary series yields a stationary process arises from a common component (W_t) , which is $I(1)$. More generally, two $I(1)$ series (Y_t) , (Z_t) are said to be *cointegrated*

(of order 1), if there exist constants μ, α_1, α_2 with α_1, α_2 different from 0, such that the process

$$X_t = \mu + \alpha_1 Y_t + \alpha_2 Z_t, \quad t \in \mathbb{Z}, \quad (2.17)$$

is $I(0)$. Without loss of generality, we can choose $\alpha_1 = 1$ in this case. Such cointegrated time series are often encountered in macroeconomics (Granger, 1981; Engle and Granger, 1987). Consider, for example, prices for the same commodity in different parts of a country. Principles of supply and demand, along with the possibility of arbitrage, mean that, while the process may fluctuate more-or-less randomly, the distance between them will, in equilibrium, be relatively constant (typically about zero).

The link between cointegration and error correction can vividly be described by the humorous tale of the drunkard and his dog, c.f. Murray (1994). In the same way a drunkard seems to follow a random walk an unleashed dog wanders aimlessly. We can, therefore, model their ways by random walks

$$\begin{aligned} Y_t &= Y_{t-1} + \varepsilon_t \quad \text{and} \\ Z_t &= Z_{t-1} + \delta_t, \end{aligned}$$

where the individual single steps $(\varepsilon_t), (\delta_t)$ of man and dog are uncorrelated white noise processes. Random walks are not stationary, since their variances increase, and so both processes (Y_t) and (Z_t) are not stationary.

And if the dog belongs to the drunkard? We assume the dog to be unleashed and thus, the distance $Y_t - Z_t$ between the drunk and his dog is a random variable. It seems reasonable to assume that these distances form a stationary process, i.e., that (Y_t) and (Z_t) are cointegrated with constants $\alpha_1 = 1$ and $\alpha_2 = -1$.

Cointegration requires that both variables in question be $I(1)$, but that a linear combination of them be $I(0)$. This means that the first step is to figure out if the series themselves are $I(1)$, typically by using *unit root tests*. If one or both are not $I(1)$, cointegration of order 1 is not an option.

Whether two processes (Y_t) and (Z_t) are cointegrated can be tested by means of a linear regression approach. This is based on the *coin-*

tegration regression

$$Y_t = \beta_0 + \beta_1 Z_t + \varepsilon_t,$$

where (ε_t) is a stationary process and $\beta_0, \beta_1 \in \mathbb{R}$ are the cointegration constants.

One can use the ordinary least squares estimates $\hat{\beta}_0, \hat{\beta}_1$ of the target parameters β_0, β_1 , which satisfy

$$\sum_{t=1}^n \left(Y_t - \hat{\beta}_0 - \hat{\beta}_1 Z_t \right)^2 = \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{t=1}^n \left(Y_t - \beta_0 - \beta_1 Z_t \right)^2,$$

and one checks, whether the estimated residuals

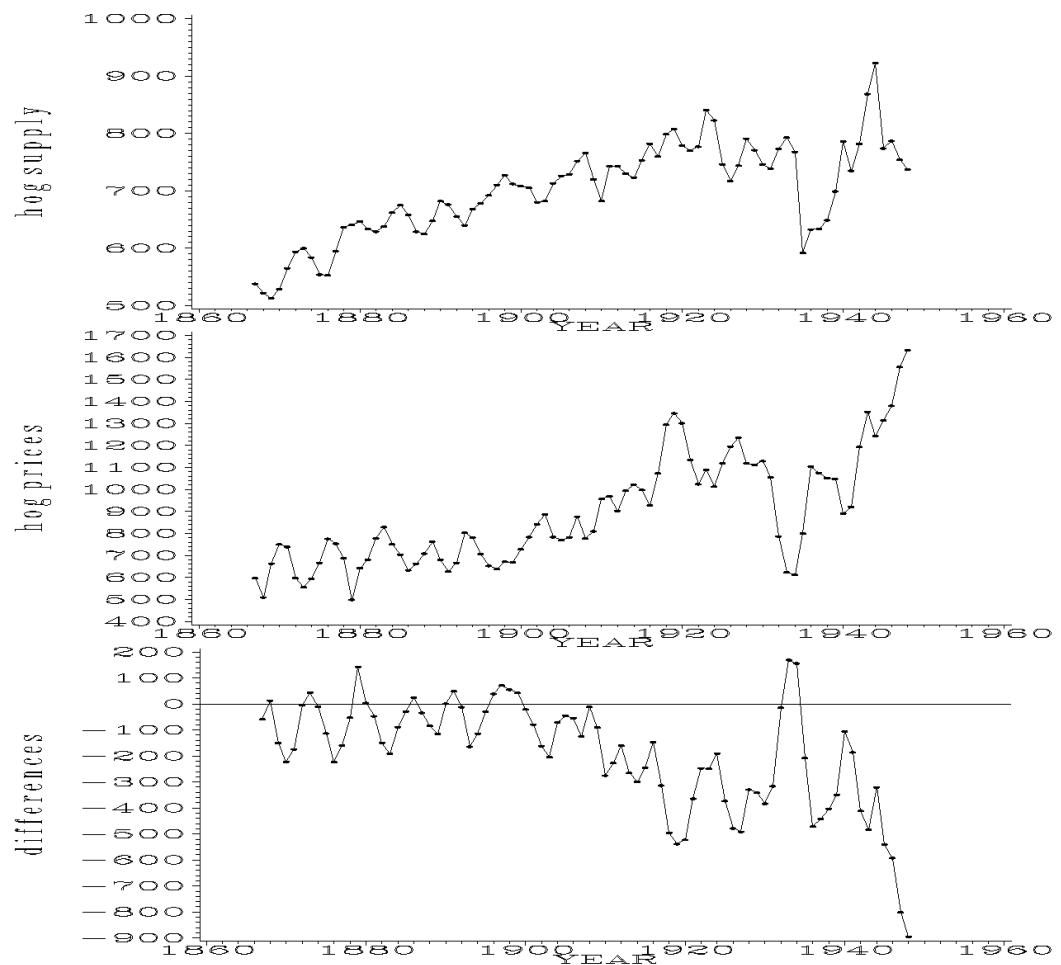
$$\hat{\varepsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 Z_t$$

are generated by a stationary process.

A general strategy for examining cointegrated series can now be summarized as follows:

- (i) Determine that the two series are $I(1)$ by standard unit root tests such as Dickey–Fuller or augmented Dickey–Fuller.
- (ii) Compute $\hat{\varepsilon}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 Z_t$ using ordinary least squares.
- (iii) Examine $\hat{\varepsilon}_t$ for stationarity, using for example the Phillips–Ouliaris test.

Example 2.2.11. (Hog Data) Quenouille (1968) Hog Data list the annual hog supply and hog prices in the U.S. between 1867 and 1948. Do they provide a typical example of cointegrated series? A discussion can be found in Box and Tiao (1977).



Plot 2.2.6: Hog Data: hog supply and hog prices.

```

1  /* hog.sas */
2  TITLE1 'Hog supply, hog prices and differences';
3  TITLE2 'Hog Data (1867-1948)';
4  /* Note that this program requires the macro mkfields.sas to be
   ↪ submitted before this program */
5
6  /* Read in the two data sets */
7  DATA data1;
8      INFILE 'c:\data\hogsuppl.txt';
9      INPUT supply @@;
10
11 DATA data2;
12     INFILE 'c:\data\hogprice.txt';
13     INPUT price @@;
14
15 /* Merge data sets, generate year and compute differences */
16 DATA data3;
17     MERGE data1 data2;

```

```

18  year=_N_+1866;
19  diff=supply-price;
20
21  /* Graphical options */
22  SYMBOL1 V=DOT C=GREEN I=JOIN H=0.5 W=1;
23  AXIS1 LABEL=(ANGLE=90 'h o g   s u p p l y');
24  AXIS2 LABEL=(ANGLE=90 'h o g   p r i c e s');
25  AXIS3 LABEL=(ANGLE=90 'd i f f e r e n c e s');
26
27  /* Generate three plots */
28  GOPTIONS NODISPLAY;
29  PROC GPLOT DATA=data3 GOUT=abb;
30      PLOT supply*year / VAXIS=AXIS1;
31      PLOT price*year / VAXIS=AXIS2;
32      PLOT diff*year / VAXIS=AXIS3 VREF=0;
33  RUN;
34
35  /* Display them in one output */
36  GOPTIONS DISPLAY;
37  PROC GREPLAY NOFS IGOUT=abb TC=SASHELP.TEMPLT;
38      TEMPLATE=V3;
39      TREPLAY 1:GPLOT 2:GPLOT1 3:GPLOT2;
40  RUN; DELETE _ALL_; QUIT;

```

The supply data and the price data read in from two external files are merged in data3. Year is an additional variable with values 1867, 1868, ..., 1932. By PROC GPLOT hog supply, hog prices and their differences diff are plotted in three different plots stored in the graphics catalog abb. The horizontal line at the zero level is plotted by the option

VREF=0. The plots are put into a common graphic using PROC GREPLAY and the template V3. Note that the labels of the vertical axes are spaced out as SAS sets their characters too close otherwise.

For the program to work properly the macro mkfields.sas has to be submitted beforehand.

Hog supply ($=: y_t$) and hog price ($=: z_t$) obviously increase in time t and do, therefore, not seem to be realizations of stationary processes; nevertheless, as they behave similarly, a linear combination of both might be stationary. In this case, hog supply and hog price would be cointegrated.

This phenomenon can easily be explained as follows. A high price z_t at time t is a good reason for farmers to breed more hogs, thus leading to a large supply y_{t+1} in the next year $t+1$. This makes the price z_{t+1} fall with the effect that farmers will reduce their supply of hogs in the following year $t+2$. However, when hogs are in short supply, their price z_{t+2} will rise etc. There is obviously some *error correction mechanism* inherent in these two processes, and the observed cointegration

helps us to detect its existence.

Before we can examine the data for cointegration however, we have to check that our two series are $I(1)$. We will do this by the Dickey-Fuller-test which can assume three different models for the series Y_t :

$$\Delta Y_t = \gamma Y_{t-1} + \varepsilon_t \quad (2.18)$$

$$\Delta Y_t = a_0 + \gamma Y_{t-1} + \varepsilon_t \quad (2.19)$$

$$\Delta Y_t = a_0 + a_2 t + \gamma Y_{t-1} + \varepsilon_t, \quad (2.20)$$

where (ε_t) is a white noise with expectation 0. Note that (2.18) is a special case of (2.19) and (2.19) is a special case of (2.20). Note also that one can bring (2.18) into an AR(1)-form by putting $a_1 = \gamma + 1$ and (2.19) into an AR(1)-form with an intercept term a_0 (so called *drift* term) by also putting $a_1 = \gamma + 1$. (2.20) can be brought into an AR(1)-form with a drift and trend term $a_0 + a_2 t$.

The null hypothesis of the Dickey-Fuller-test is now that $\gamma = 0$. The corresponding AR(1)-processes would then not be stationary, since the characteristic polynomial would then have a root on the unit circle, a so called *unit root*. Note that in the case (2.20) the series Y_t is $I(2)$ under the null hypothesis and two $I(2)$ time series are said to be cointegrated of order 2, if there is a linear combination of them which is stationary as in (2.17).

The Dickey-Fuller-test now estimates $a_1 = \gamma + 1$ by \hat{a}_1 , obtained from an ordinary regression and checks for $\gamma = 0$ by computing the test statistic

$$x := n\hat{\gamma} := n(\hat{a}_1 - 1), \quad (2.21)$$

where n is the number of observations on which the regression is based (usually one less than the initial number of observations). The test statistic follows the so called Dickey-Fuller distribution which cannot be explicitly given but has to be obtained by Monte-Carlo and bootstrap methods. P-values derived from this distribution can for example be obtained in SAS by the function `PROBDF`, see the following program. For more information on the Dickey-Fuller-test, especially the extension of the augmented Dickey-Fuller-test with more than one autoregressing variable in (2.18) to (2.20) we refer to Enders (2004, Chapter 4).

Testing for Unit Roots by Dickey-Fuller Hog Data (1867-1948)				
Beob.	xsupply	xprice	psupply	pprice
1	0.12676	.	0.70968	.
2	.	0.86027	.	0.88448

Listing 2.2.7: Dickey-Fuller test of Hog Data.

```

1  /* hog_dickey_fuller.sas */
2  TITLE1 'Testing for Unit Roots by Dickey-Fuller';
3  TITLE2 'Hog Data (1867-1948)';
4  /* Note that this program needs data3 generated by the previous
   ↪ program (hog.sas) */
5
6  /* Prepare data set for regression */
7  DATA regression;
8      SET data3;
9      supply1=LAG(supply);
10     price1=LAG(price);
11
12  /* Estimate gamma for both series by regression */
13  PROC REG DATA=regression OUTEST=est;
14      MODEL supply=supply1 / NOINT NOPRINT;
15      MODEL price=price1 / NOINT NOPRINT;
16
17  /* Compute test statistics for both series */
18  DATA dickeyfuller1;
19      SET est;
20      xsupply= 81*(supply1-1);
21      xprice= 81*(price1-1);
22
23  /* Compute p-values for the three models */
24  DATA dickeyfuller2;
25      SET dickeyfuller1;
26      psupply=PROBDF(xsupply,81,1,"RZM");
27      pprice=PROBDF(xprice,81,1,"RZM");
28
29  /* Print the results */
30  PROC PRINT DATA=dickeyfuller2(KEEP= xsupply xprice psupply pprice);
31  RUN; QUIT;

```

Unfortunately the Dickey-Fuller-test is only implemented in the High Performance Forecasting module of SAS (PROC HPFEDIAG). Since this is no standard module we compute it by hand here.

In the first DATA step the data are prepared for the regression by lagging the corresponding

variables. Assuming model (2.18), the regression is carried out for both series, suppressing an intercept by the option NOINT. The results are stored in est. If model (2.19) is to be investigated, NOINT is to be deleted, for model (2.19) the additional regression variable year has to be inserted.

In the next step the corresponding test statistics are calculated by (2.21). The factor 81 comes from the fact that the hog data contain 82 observations and the regression is carried out with 81 observations.

After that the corresponding p-values are computed. The function `PROBDF`, which completes this task, expects four arguments. First the test statistic, then the sample size of the regression, then the number of autoregressive variables in (2.18) to (2.20) (in our case 1) and

a three-letter specification which of the models (2.18) to (2.20) is to be tested. The first letter states, in which way γ is estimated (`R` for regression, `S` for a studentized test statistic which we did not explain) and the last two letters state the model (`ZM` (Zero mean) for (2.18), `SM` (single mean) for (2.19), `TR` (trend) for (2.20)).

In the final step the test statistics and corresponding p-values are given to the output window.

The p-values do not reject the hypothesis that we have two $I(1)$ series under model (2.18) at the 5%-level, since they are both larger than 0.05 and thus support that $\gamma = 0$.

Since we have checked that both hog series can be regarded as $I(1)$ we can now check for cointegration.

The AUTOREG Procedure					
Dependent Variable		supply			
Ordinary Least Squares Estimates					
SSE	338324.258		DFE	80	
MSE	4229		Root MSE	65.03117	
SBC	924.172704		AIC	919.359266	
Regress R-Square	0.3902		Total R-Square	0.3902	
Durbin-Watson	0.5839				
Phillips-Ouliaris Cointegration Test					
Lags		Rho	Tau		
1		-28.9109	-4.0142		
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	515.7978	26.6398	19.36	<.0001

price	1	0.2059	0.0288	7.15	<.0001
-------	---	--------	--------	------	--------

Listing 2.2.8: Phillips–Ouliaris test for cointegration of Hog Data.

```

1 /* hog_cointegration.sas */
2 TITLE1 'Testing for cointegration';
3 TITLE2 'Hog Data (1867–1948)';
4 /* Note that this program needs data3 generated by the previous
   ↪ program (hog.sas) */
5
6 /* Compute Phillips–Ouliaris–test for cointegration */
7 PROC AUTOREG DATA=data3;
8   MODEL supply=price / STATIONARITY=(PHILLIPS);
9 RUN; QUIT;

```

The procedure AUTOREG (for autoregressive models) uses data3 from Program 2.2.6 (hog.sas). In the MODEL statement a regression from supply on price is defined and the option STATIONARITY=(PHILLIPS) makes SAS calculate the statistics of the Phillips–Ouliaris test for cointegration.

The output of the above program contains some characteristics of the cointegration regression, the Phillips–Ouliaris test statistics and the regression coefficients with their t -ratios. The Phillips–Ouliaris test statistics need some further explanation.

The hypothesis of the Phillips–Ouliaris cointegration test is *no cointegration*. Unfortunately SAS does not provide the p -value, but only the values of the test statistics denoted by RHO and TAU. Tables of critical values of these test statistics can be found in Phillips and Ouliaris (1990). Note that in the original paper the two test statistics are denoted by \hat{Z}_α and \hat{Z}_t . The hypothesis is to be rejected if RHO or TAU are below the critical value for the desired type I level error α . For this one has to differentiate between the following cases:

- (i) If model (2.18) with $\gamma = 0$ has been validated for both series, then use the following table for critical values of RHO and TAU. This is the so-called *standard* case.

α	0.15	0.125	0.1	0.075	0.05	0.025	0.01
RHO	-10.74	-11.57	-12.54	-13.81	-15.64	-18.88	-22.83
TAU	-2.26	-2.35	-2.45	-2.58	-2.76	-3.05	-3.39

- (ii) If model (2.19) with $\gamma = 0$ has been validated for both series, then use the following table for critical values of RHO and TAU. This case is referred to as *demeaned*.

α	0.15	0.125	0.1	0.075	0.05	0.025	0.01
RHO	-14.91	-15.93	-17.04	-18.48	-20.49	-23.81	-28.32
TAU	-2.86	-2.96	-3.07	-3.20	-3.37	-3.64	-3.96

- (iii) If model (2.20) with $\gamma = 0$ has been validated for both series, then use the following table for critical values of RHO and TAU. This case is said to be *demeaned and detrended*.

α	0.15	0.125	0.1	0.075	0.05	0.025	0.01
RHO	-20.79	-21.81	-23.19	-24.75	-27.09	-30.84	-35.42
TAU	-3.33	-3.42	-3.52	-3.65	-3.80	-4.07	-4.36

In our example, the RHO-value equals -28.9109 and the TAU-value is -4.0142 . Since we have seen that model (2.18) with $\gamma = 0$ is appropriate for our series, we have to use the standard table. Both test statistics are smaller than the critical values of -15.64 and -2.76 in the above table of the standard case and, thus, lead to a rejection of the null hypothesis of no cointegration at the 5%-level.

For further information on cointegration we refer to the time series book by Hamilton (1994, Chapter 19) and the one by Enders (2004, Chapter 6).

ARCH- and GARCH-Processes

In particular the monitoring of stock prices gave rise to the idea that the *volatility* of a time series (Y_t) might not be a constant but rather a random variable, which depends on preceding realizations. The following approach to model such a change in volatility is due to Engle (1982).

We assume the multiplicative model

$$Y_t = \sigma_t Z_t, \quad t \in \mathbb{Z},$$

where the Z_t are independent and identically distributed random variables with

$$E(Z_t) = 0 \text{ and } E(Z_t^2) = 1, \quad t \in \mathbb{Z}.$$

The *scale* σ_t is supposed to be a function of the past p values of the series:

$$\sigma_t^2 = a_0 + \sum_{j=1}^p a_j Y_{t-j}^2, \quad t \in \mathbb{Z}, \quad (2.22)$$

where $p \in \{0, 1, \dots\}$ and $a_0 > 0$, $a_j \geq 0$, $1 \leq j \leq p-1$, $a_p > 0$ are constants.

The particular choice $p = 0$ yields obviously a white noise model for (Y_t) . Common choices for the distribution of Z_t are the standard normal distribution or the (standardized) t -distribution, which in the non-standardized form has the density

$$f_m(x) := \frac{\Gamma((m+1)/2)}{\Gamma(m/2)\sqrt{\pi m}} \left(1 + \frac{x^2}{m}\right)^{-(m+1)/2}, \quad x \in \mathbb{R}.$$

The number $m \geq 1$ is the *degree of freedom* of the t -distribution. The scale σ_t in the above model is determined by the past observations Y_{t-1}, \dots, Y_{t-p} , and the *innovation* on this scale is then provided by Z_t . We assume moreover that the process (Y_t) is a causal one in the sense that Z_t and Y_s , $s < t$, are independent. Some autoregressive structure is, therefore, inherent in the process (Y_t) . Conditional on $Y_{t-j} = y_{t-j}$, $1 \leq j \leq p$, the variance of Y_t is $a_0 + \sum_{j=1}^p a_j y_{t-j}^2$ and, thus, the conditional variances of the process will generally be different. The process $Y_t = \sigma_t Z_t$ is, therefore, called an *autoregressive* and *conditional heteroscedastic* process of order p , ARCH(p)-process for short.

If, in addition, the causal process (Y_t) is stationary, then we obviously have

$$E(Y_t) = E(\sigma_t) E(Z_t) = 0$$

and

$$\begin{aligned} \sigma^2 &:= E(Y_t^2) = E(\sigma_t^2) E(Z_t^2) \\ &= a_0 + \sum_{j=1}^p a_j E(Y_{t-j}^2) \\ &= a_0 + \sigma^2 \sum_{j=1}^p a_j, \end{aligned}$$

which yields

$$\sigma^2 = \frac{a_0}{1 - \sum_{j=1}^p a_j}.$$

A necessary condition for the stationarity of the process (Y_t) is, therefore, the inequality $\sum_{j=1}^p a_j < 1$. Note, moreover, that the preceding arguments immediately imply that the Y_t and Y_s are uncorrelated for different values $s < t$

$$E(Y_s Y_t) = E(\sigma_s Z_s \sigma_t Z_t) = E(\sigma_s Z_s \sigma_t) E(Z_t) = 0,$$

since Z_t is independent of σ_t , σ_s and Z_s . But they are not independent, because Y_s influences the scale σ_t of Y_t by (2.22).

The following lemma is crucial. It embeds the ARCH(p)-processes to a certain extent into the class of AR(p)-processes, so that our above tools for the analysis of autoregressive processes can be applied here as well.

Lemma 2.2.12. *Let (Y_t) be a stationary and causal ARCH(p)-process with constants a_0, a_1, \dots, a_p . If the process of squared random variables (Y_t^2) is a stationary one, then it is an AR(p)-process:*

$$Y_t^2 = a_1 Y_{t-1}^2 + \dots + a_p Y_{t-p}^2 + \varepsilon_t,$$

where (ε_t) is a white noise with $E(\varepsilon_t) = a_0$, $t \in \mathbb{Z}$.

Proof. From the assumption that (Y_t) is an ARCH(p)-process we obtain

$$\varepsilon_t := Y_t^2 - \sum_{j=1}^p a_j Y_{t-j}^2 = \sigma_t^2 Z_t^2 - \sigma_t^2 + a_0 = a_0 + \sigma_t^2 (Z_t^2 - 1), \quad t \in \mathbb{Z}.$$

This implies $E(\varepsilon_t) = a_0$ and

$$\begin{aligned} E((\varepsilon_t - a_0)^2) &= E(\sigma_t^4) E((Z_t^2 - 1)^2) \\ &= E\left(\left(a_0 + \sum_{j=1}^p a_j Y_{t-j}^2\right)^2\right) E((Z_t^2 - 1)^2) =: c, \end{aligned}$$

independent of t by the stationarity of (Y_t^2) . For $h \in \mathbb{N}$ the causality of (Y_t) finally implies

$$\begin{aligned} E((\varepsilon_t - a_0)(\varepsilon_{t+h} - a_0)) &= E(\sigma_t^2 \sigma_{t+h}^2 (Z_t^2 - 1)(Z_{t+h}^2 - 1)) \\ &= E(\sigma_t^2 \sigma_{t+h}^2 (Z_t^2 - 1)) E(Z_{t+h}^2 - 1) = 0, \end{aligned}$$

i.e., (ε_t) is a white noise with $E(\varepsilon_t) = a_0$. \square

The process (Y_t^2) satisfies, therefore, the stationarity condition (2.4) if all p roots of the equation $1 - \sum_{j=1}^p a_j z^j = 0$ are outside of the unit circle. Hence, we can estimate the order p using an estimate as in (2.11) of the partial autocorrelation function of (Y_t^2) . The Yule–Walker equations provide us, for example, with an estimate of the coefficients a_1, \dots, a_p , which then can be utilized to estimate the expectation a_0 of the error ε_t .

Note that conditional on $Y_{t-1} = y_{t-1}, \dots, Y_{t-p} = y_{t-p}$, the distribution of $Y_t = \sigma_t Z_t$ is a normal one if the Z_t are normally distributed. In this case it is possible to write down explicitly the joint density of the vector (Y_{p+1}, \dots, Y_n) , conditional on $Y_1 = y_1, \dots, Y_p = y_p$ (Exercise 2.40). A numerical maximization of this density with respect to a_0, a_1, \dots, a_p then leads to a *maximum likelihood estimate* of the vector of constants; see also Section 2.3.

A *generalized* ARCH-process, GARCH(p, q) (Bollerslev, 1986), adds an autoregressive structure to the scale σ_t by assuming the representation

$$\sigma_t^2 = a_0 + \sum_{j=1}^p a_j Y_{t-j}^2 + \sum_{k=1}^q b_k \sigma_{t-k}^2,$$

where the constants b_k are nonnegative. The set of parameters a_j, b_k can again be estimated by conditional maximum likelihood as before if a parametric model for the distribution of the innovations Z_t is assumed.

Example 2.2.13. (Hongkong Data). The daily Hang Seng closing index was recorded between July 16th, 1981 and September 30th, 1983, leading to a total amount of 552 observations p_t . The daily *log returns* are defined as

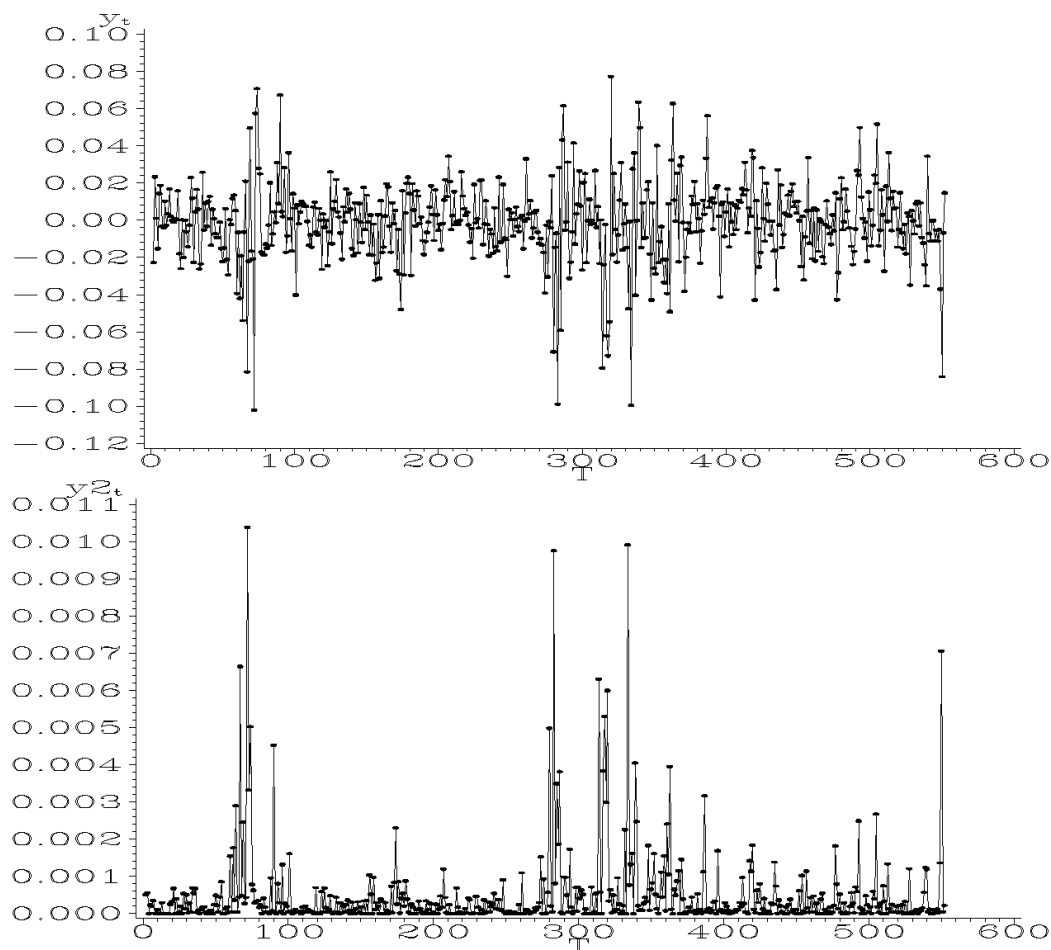
$$y_t := \log(p_t) - \log(p_{t-1}),$$

where we now have a total of $n = 551$ observations. The expansion $\log(1 + \varepsilon) \approx \varepsilon$ implies that

$$y_t = \log \left(1 + \frac{p_t - p_{t-1}}{p_{t-1}} \right) \approx \frac{p_t - p_{t-1}}{p_{t-1}},$$

provided that p_{t-1} and p_t are close to each other. In this case we can interpret the return as the difference of indices on subsequent days, relative to the initial one.

We use an ARCH(3) model for the generation of y_t , which seems to be a plausible choice by the partial autocorrelations plot. If one assumes t -distributed innovations Z_t , SAS estimates the distribution's degrees of freedom and displays the reciprocal in the TDFI-line, here $m = 1/0.1780 = 5.61$ degrees of freedom. We obtain the estimates $a_0 = 0.000214$, $a_1 = 0.147593$, $a_2 = 0.278166$ and $a_3 = 0.157807$. The SAS output also contains some general regression model information from an ordinary least squares estimation approach, some specific information for the (G)ARCH approach and as mentioned above the estimates for the ARCH model parameters in combination with t ratios and approximated p -values. The following plots show the returns of the Hang Seng index, their squares and the autocorrelation function of the log returns, indicating a possible ARCH model, since the values are close to 0. The pertaining partial autocorrelation function of the squared process and the parameter estimates are also given.



Plot 2.2.9: Log returns of Hang Seng index and their squares.

```

1  /* hongkong_plot.sas */
2  TITLE1 'Daily log returns and their squares';
3  TITLE2 'Hongkong Data ';
4
5  /* Read in the data, compute log return and their squares */
6  DATA data1;
7      INFILE 'c:\data\hongkong.txt';
8      INPUT p@@;
9      t=_N_;
10     y=DIF(LOG(p));
11     y2=y**2;
12
13  /* Graphical options */
14  SYMBOL1 C=RED V=DOT H=0.5 I=JOIN L=1;
15  AXIS1 LABEL=('y' H=1 't') ORDER=(-.12 TO .10 BY .02);
16  AXIS2 LABEL=('y2' H=1 't');
17
18  /* Generate two plots */

```

```

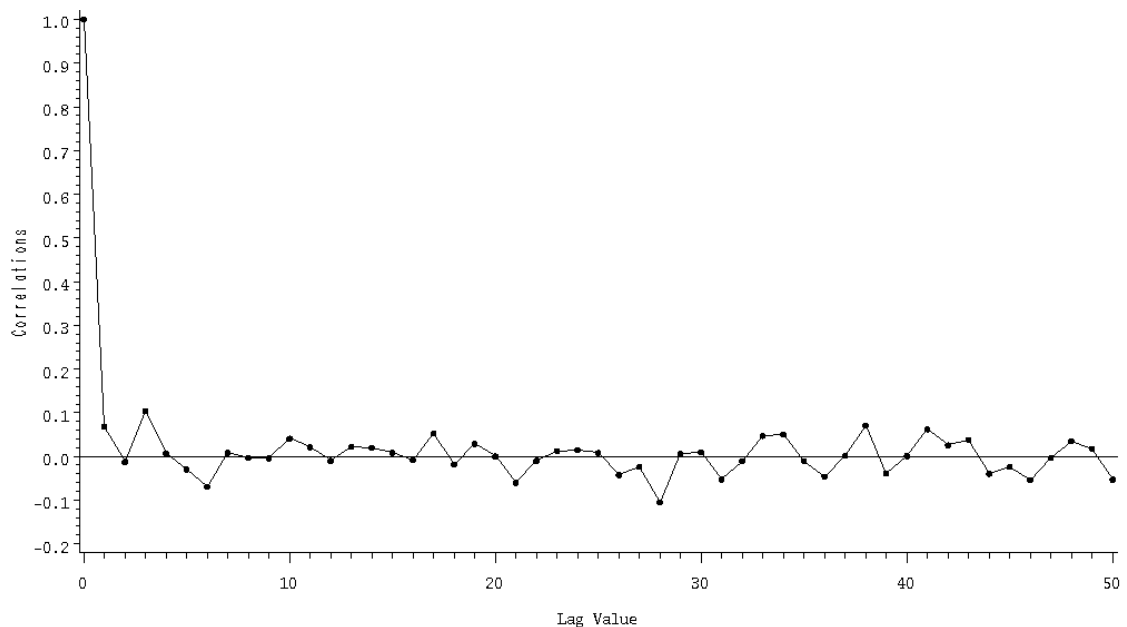
19 GOPTIONS NODISPLAY;
20 PROC GPLOT DATA=data1 GOUT=abb;
21   PLOT y*t / VAXIS=AXIS1;
22   PLOT y2*t / VAXIS=AXIS2;
23 RUN;
24
25 /* Display them in one output */
26 GOPTIONS DISPLAY;
27 PROC GREPLAY NOFS IGOUT=abb TC=SASHELP.TEMPLT;
28   TEMPLATE=V2;
29   TREPLAY 1:GPLOT 2:GPLOT1;
30 RUN; DELETE _ALL_; QUIT;

```

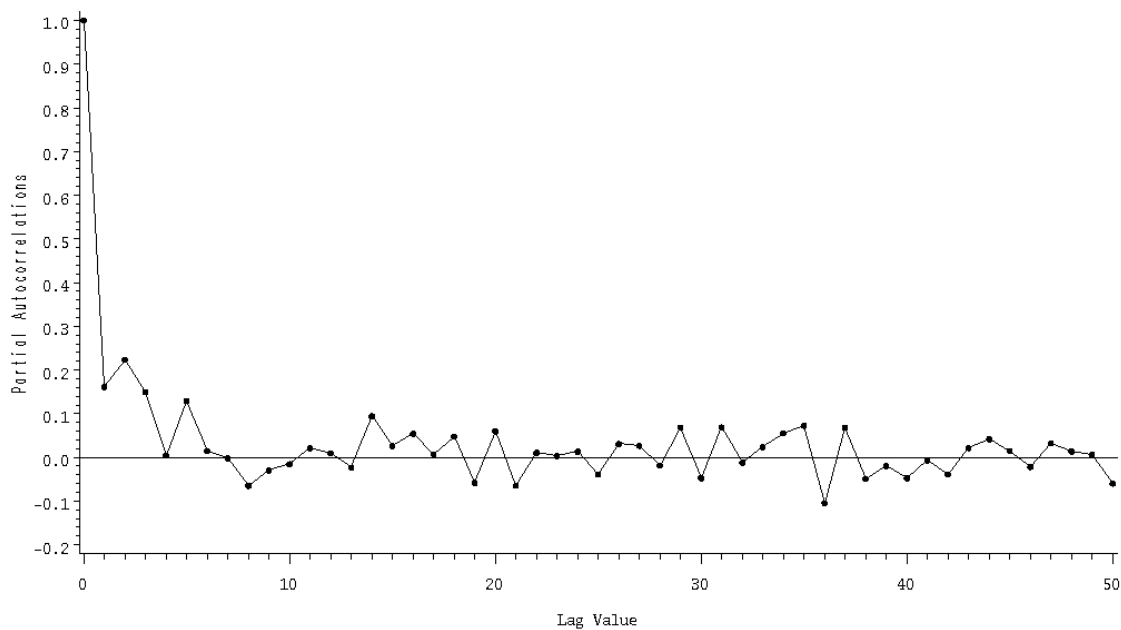
In the DATA step the observed values of the Hang Seng closing index are read into the variable *p* from an external file. The time index variable *t* uses the SAS-variable *_N_*, and the log transformed and differenced values of the index are stored in the variable *y*, their squared

values in *y2*.

After defining different axis labels, two plots are generated by two PLOT statements in PROC GPLOT, but they are not displayed. By means of PROC GREPLAY the plots are merged vertically in one graphic.



Plot 2.2.10: Autocorrelations of log returns of Hang Seng index.



Plot 2.2.10b: Partial autocorrelations of squares of log returns of Hang Seng index.

The AUTOREG Procedure

Dependent Variable = Y

Ordinary Least Squares Estimates

SSE	0.265971	DFE	551
MSE	0.000483	Root MSE	0.021971
SBC	-2643.82	AIC	-2643.82
Reg Rsq	0.0000	Total Rsq	0.0000
Durbin-Watson	1.8540		

NOTE: No intercept term is used. R-squares are redefined.

GARCH Estimates

SSE	0.265971	OBS	551
MSE	0.000483	UVAR	0.000515
Log L	1706.532	Total Rsq	0.0000
SBC	-3381.5	AIC	-3403.06
Normality Test	119.7698	Prob>Chi-Sq	0.0001

Variable	DF	B Value	Std Error	t Ratio	Approx Prob
ARCH0	1	0.000214	0.000039	5.444	0.0001
ARCH1	1	0.147593	0.0667	2.213	0.0269

ARCH2	1	0.278166	0.0846	3.287	0.0010
ARCH3	1	0.157807	0.0608	2.594	0.0095
TDFI	1	0.178074	0.0465	3.833	0.0001

Listing 2.2.10c: Parameter estimates in the ARCH(3)-model for stock returns.

```

1  /* hongkong_pa.sas */
2  TITLE1 'ARCH(3)-model';
3  TITLE2 'Hongkong Data';
4  /* Note that this program needs data generated by the previous
   ↪ program (hongkong_plot.sas) */
5
6  /* Compute (partial) autocorrelation function */
7  PROC ARIMA DATA=data1;
8      IDENTIFY VAR=y NLAG=50 OUTCOV=data3;
9      IDENTIFY VAR=y2 NLAG=50 OUTCOV=data2;
10
11  /* Graphical options */
12  SYMBOL1 C=RED V=DOT H=0.5 I=JOIN;
13  AXIS1 LABEL=(ANGLE=90);
14
15  /* Plot autocorrelation function of supposed ARCH data */
16  PROC GPLOT DATA=data3;
17      PLOT corr*lag / VREF=0 VAXIS=AXIS1;
18  RUN;
19
20
21  /* Plot partial autocorrelation function of squared data */
22  PROC GPLOT DATA=data2;
23      PLOT partcorr*lag / VREF=0 VAXIS=AXIS1;
24  RUN;
25
26  /* Estimate ARCH(3)-model */
27  PROC AUTOREG DATA=data1;
28      MODEL y = / NOINT GARCH=(q=3) DIST=T;
29  RUN; QUIT;

```

To identify the order of a possibly underlying ARCH process for the daily log returns of the Hang Seng closing index, the empirical autocorrelation of the log returns, the empirical partial autocorrelations of their squared values, which are stored in the variable `y2` of the data set `data1` in Program 2.2.9 (`hongkong_plot.sas`), are calculated by means of `PROC ARIMA` and the `IDENTIFY` statement. The subsequent procedure `GPLOT` displays these (partial) autocorrelations. A horizontal reference line helps to decide whether a value

is substantially different from 0.

`PROC AUTOREG` is used to analyze the ARCH(3) model for the daily log returns. The `MODEL` statement specifies the dependent variable `y`. The option `NOINT` suppresses an intercept parameter, `GARCH=(q=3)` selects the ARCH(3) model and `DIST=T` determines a t distribution for the innovations Z_t in the model equation. Note that, in contrast to our notation, SAS uses the letter q for the ARCH model order.

2.3 Specification of ARMA-Models: The Box–Jenkins Program

The aim of this section is to fit a time series model $(Y_t)_{t \in \mathbb{Z}}$ to a given set of data y_1, \dots, y_n collected in time t . We suppose that the data y_1, \dots, y_n are (possibly) variance-stabilized as well as trend or seasonally adjusted. We assume that they were generated by clipping Y_1, \dots, Y_n from an ARMA(p, q)-process $(Y_t)_{t \in \mathbb{Z}}$, which we will fit to the data in the following. As noted in Section 2.2, we could also fit the model $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$ to the data, where (ε_t) is a white noise. But then we would have to determine infinitely many parameters α_v , $v \geq 0$. By the *principle of parsimony* it seems, however, reasonable to fit only the finite number of parameters of an ARMA(p, q)-process. The Box–Jenkins program consists of four steps:

1. **Order selection:** Choice of the parameters p and q .
2. **Estimation of coefficients:** The coefficients of the AR part of the model a_1, \dots, a_p and the MA part b_1, \dots, b_q are estimated.
3. **Diagnostic check:** The fit of the ARMA(p, q)-model with the estimated coefficients is checked.
4. **Forecasting:** The prediction of future values of the original process.

The four steps are discussed in the following. The application of the Box–Jenkins Program is presented in a case study in Chapter 7.

Order Selection

The order q of a moving average MA(q)-process can be estimated by means of the empirical autocorrelation function $r(k)$ i.e., by the correlogram. Part (iv) of Lemma 2.2.1 shows that the autocorrelation function $\rho(k)$ vanishes for $k \geq q + 1$. This suggests to choose the order q such that $r(q)$ is clearly different from zero, whereas $r(k)$ for $k \geq q + 1$ is quite close to zero. This, however, is obviously a rather vague selection rule.

The order p of an $\text{AR}(p)$ -process can be estimated in an analogous way using the empirical partial autocorrelation function $\hat{\alpha}(k)$, $k \geq 1$, as defined in (2.11). Since $\hat{\alpha}(p)$ should be close to the p -th coefficient a_p of the $\text{AR}(p)$ -process, which is different from zero, whereas $\hat{\alpha}(k) \approx 0$ for $k > p$, the above rule can be applied again with r replaced by $\hat{\alpha}$. The choice of the orders p and q of an $\text{ARMA}(p, q)$ -process is a bit more challenging. In this case one commonly takes the pair (p, q) , minimizing some information function, which is based on an estimate $\hat{\sigma}_{p,q}^2$ of the variance of ε_0 .

Popular functions are *Akaike's Information Criterion*

$$\text{AIC}(p, q) := \log(\hat{\sigma}_{p,q}^2) + 2 \frac{p+q}{n},$$

the *Bayesian Information Criterion*

$$\text{BIC}(p, q) := \log(\hat{\sigma}_{p,q}^2) + \frac{(p+q) \log(n)}{n}$$

and the *Hannan–Quinn Criterion*

$$\text{HQ}(p, q) := \log(\hat{\sigma}_{p,q}^2) + \frac{2(p+q)c \log(\log(n))}{n} \quad \text{with } c > 1.$$

AIC and BIC are discussed in Brockwell and Davis (1991, Section 9.3) for Gaussian processes (Y_t) , where the joint distribution of an arbitrary vector $(Y_{t_1}, \dots, Y_{t_k})$ with $t_1 < \dots < t_k$ is multivariate normal, see below. For the HQ-criterion we refer to Hannan and Quinn (1979). Note that the variance estimate $\hat{\sigma}_{p,q}^2$, which uses estimated model parameters, discussed in the next section, will in general become arbitrarily small as $p+q$ increases. The additive terms in the above criteria serve, therefore, as penalties for large values, thus helping to prevent *overfitting* of the data by choosing p and q too large. It can be shown that BIC and HQ lead under certain regularity conditions to strongly consistent estimators of the model order. AIC has the tendency not to underestimate the model order. Simulations point to the fact that BIC is generally to be preferred for larger samples, see Shumway and Stoffer (2006, Section 2.2).

More sophisticated methods for selecting the orders p and q of an $\text{ARMA}(p, q)$ -process are presented in Section 7.5 within the application of the Box–Jenkins Program in a case study.

Estimation of Coefficients

Suppose we fixed the order p and q of an ARMA(p, q)-process $(Y_t)_{t \in \mathbb{Z}}$, with Y_1, \dots, Y_n now modelling the data y_1, \dots, y_n . In the next step we will derive estimators of the constants $a_1, \dots, a_p, b_1, \dots, b_q$ in the model

$$Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q}, \quad t \in \mathbb{Z}.$$

The Gaussian Model: Maximum Likelihood Estimator

We assume first that (Y_t) is a *Gaussian process* and thus, the joint distribution of (Y_1, \dots, Y_n) is a n -dimensional normal distribution

$$P\{Y_i \leq s_i, i = 1, \dots, n\} = \int_{-\infty}^{s_1} \dots \int_{-\infty}^{s_n} \varphi_{\mu, \Sigma}(x_1, \dots, x_n) dx_n \dots dx_1$$

for arbitrary $s_1, \dots, s_n \in \mathbb{R}$. Here

$$\begin{aligned} & \varphi_{\mu, \Sigma}(x_1, \dots, x_n) \\ &= \frac{1}{(2\pi)^{n/2} (\det \Sigma)^{1/2}} \cdot \\ & \quad \exp \left(-\frac{1}{2} ((x_1, \dots, x_n) - \mu^T) \Sigma^{-1} ((x_1, \dots, x_n)^T - \mu) \right) \end{aligned}$$

is for arbitrary $x_1, \dots, x_n \in \mathbb{R}$ the density of the n -dimensional normal distribution with mean vector $\mu = (\mu, \dots, \mu)^T \in \mathbb{R}^n$ and covariance matrix $\Sigma = (\gamma(i-j))_{1 \leq i, j \leq n}$, denoted by $\mathbf{N}(\mu, \Sigma)$, where $\mu = E(Y_0)$ and γ is the autocovariance function of the stationary process (Y_t) . The number $\varphi_{\mu, \Sigma}(x_1, \dots, x_n)$ reflects the probability that the random vector (Y_1, \dots, Y_n) realizes close to (x_1, \dots, x_n) . Precisely, we have for $\varepsilon \downarrow 0$

$$\begin{aligned} & P\{Y_i \in [x_i - \varepsilon, x_i + \varepsilon], i = 1, \dots, n\} \\ &= \int_{x_1 - \varepsilon}^{x_1 + \varepsilon} \dots \int_{x_n - \varepsilon}^{x_n + \varepsilon} \varphi_{\mu, \Sigma}(z_1, \dots, z_n) dz_n \dots dz_1 \\ &\approx 2^n \varepsilon^n \varphi_{\mu, \Sigma}(x_1, \dots, x_n). \end{aligned}$$

The *likelihood principle* is the fact that a random variable tends to attain its most likely value and thus, if the vector (Y_1, \dots, Y_n) actually attained the value (y_1, \dots, y_n) , the unknown underlying mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ ought to be such that $\varphi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(y_1, \dots, y_n)$ is maximized. The computation of these parameters leads to the *maximum likelihood estimator* of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

We assume that the process (Y_t) satisfies the stationarity condition (2.4), in which case $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$, $t \in \mathbb{Z}$, is invertible, where (ε_t) is a white noise and the coefficients α_v depend only on a_1, \dots, a_p and b_1, \dots, b_q . Consequently we have for $s \geq 0$

$$\gamma(s) = \text{Cov}(Y_0, Y_s) = \sum_{v \geq 0} \sum_{w \geq 0} \alpha_v \alpha_w \text{Cov}(\varepsilon_{-v}, \varepsilon_{s-w}) = \sigma^2 \sum_{v \geq 0} \alpha_v \alpha_{s+v}.$$

The matrix

$$\boldsymbol{\Sigma}' := \sigma^{-2} \boldsymbol{\Sigma},$$

therefore, depends only on a_1, \dots, a_p and b_1, \dots, b_q .

We can write now the density $\varphi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(x_1, \dots, x_n)$ as a function of $\boldsymbol{\vartheta} := (\sigma^2, \boldsymbol{\mu}, a_1, \dots, a_p, b_1, \dots, b_q) \in \mathbb{R}^{p+q+2}$ and $(x_1, \dots, x_n) \in \mathbb{R}^n$

$$\begin{aligned} p(x_1, \dots, x_n | \boldsymbol{\vartheta}) \\ &:= \varphi_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(x_1, \dots, x_n) \\ &= (2\pi\sigma^2)^{-n/2} (\det \boldsymbol{\Sigma}')^{-1/2} \exp \left(-\frac{1}{2\sigma^2} Q(\boldsymbol{\vartheta} | x_1, \dots, x_n) \right), \end{aligned}$$

where

$$Q(\boldsymbol{\vartheta} | x_1, \dots, x_n) := ((x_1, \dots, x_n) - \boldsymbol{\mu}^T) \boldsymbol{\Sigma}'^{-1} ((x_1, \dots, x_n)^T - \boldsymbol{\mu})$$

is a quadratic function. The *likelihood function* pertaining to the outcome (y_1, \dots, y_n) is

$$L(\boldsymbol{\vartheta} | y_1, \dots, y_n) := p(y_1, \dots, y_n | \boldsymbol{\vartheta}).$$

A parameter $\hat{\boldsymbol{\vartheta}}$ maximizing the likelihood function

$$L(\hat{\boldsymbol{\vartheta}} | y_1, \dots, y_n) = \sup_{\boldsymbol{\vartheta}} L(\boldsymbol{\vartheta} | y_1, \dots, y_n),$$

is then a maximum likelihood estimator of $\boldsymbol{\vartheta}$.

Due to the strict monotonicity of the logarithm, maximizing the likelihood function is in general equivalent to the maximization of the *loglikelihood function*

$$l(\boldsymbol{\vartheta}|y_1, \dots, y_n) = \log L(\boldsymbol{\vartheta}|y_1, \dots, y_n).$$

The maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}$ therefore satisfies

$$\begin{aligned} l(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n) &= \sup_{\boldsymbol{\vartheta}} l(\boldsymbol{\vartheta}|y_1, \dots, y_n) \\ &= \sup_{\boldsymbol{\vartheta}} \left(-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log(\det \boldsymbol{\Sigma}') - \frac{1}{2\sigma^2} Q(\boldsymbol{\vartheta}|y_1, \dots, y_n) \right). \end{aligned}$$

The computation of a maximizer is a numerical and usually computer intensive problem. Some further insights are given in Section 7.5.

Example 2.3.1. The AR(1)-process $Y_t = aY_{t-1} + \varepsilon_t$ with $|a| < 1$ has by Example 2.2.4 the autocovariance function

$$\gamma(s) = \sigma^2 \frac{a^s}{1 - a^2}, \quad s \geq 0,$$

and thus,

$$\boldsymbol{\Sigma}' = \frac{1}{1 - a^2} \begin{pmatrix} 1 & a & a^2 & \dots & a^{n-1} \\ a & 1 & a & & a^{n-2} \\ \vdots & & & \ddots & \vdots \\ a^{n-1} & a^{n-2} & a^{n-3} & \dots & 1 \end{pmatrix}.$$

The inverse matrix is

$$\boldsymbol{\Sigma}'^{-1} = \begin{pmatrix} 1 & -a & 0 & 0 & \dots & 0 \\ -a & 1 + a^2 & -a & 0 & & 0 \\ 0 & -a & 1 + a^2 & -a & & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \dots & & -a & 1 + a^2 & -a \\ 0 & 0 & \dots & 0 & -a & 1 \end{pmatrix}.$$

Check that the determinant of Σ'^{-1} is $\det(\Sigma'^{-1}) = 1 - a^2 = 1/\det(\Sigma')$, see Exercise 2.44. If (Y_t) is a Gaussian process, then the likelihood function of $\boldsymbol{\vartheta} = (\sigma^2, \mu, a)$ is given by

$$L(\boldsymbol{\vartheta}|y_1, \dots, y_n) = (2\pi\sigma^2)^{-n/2}(1 - a^2)^{1/2} \exp\left(-\frac{1}{2\sigma^2}Q(\boldsymbol{\vartheta}|y_1, \dots, y_n)\right),$$

where

$$\begin{aligned} Q(\boldsymbol{\vartheta}|y_1, \dots, y_n) &= ((y_1, \dots, y_n) - \boldsymbol{\mu})\Sigma'^{-1}((y_1, \dots, y_n) - \boldsymbol{\mu})^T \\ &= (y_1 - \mu)^2 + (y_n - \mu)^2 + (1 + a^2) \sum_{i=2}^{n-1} (y_i - \mu)^2 \\ &\quad - 2a \sum_{i=1}^{n-1} (y_i - \mu)(y_{i+1} - \mu). \end{aligned}$$

Nonparametric Approach: Least Squares

If $E(\varepsilon_t) = 0$, then

$$\hat{Y}_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q}$$

would obviously be a reasonable one-step forecast of the ARMA(p, q)-process

$$Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q},$$

based on Y_{t-1}, \dots, Y_{t-p} and $\varepsilon_{t-1}, \dots, \varepsilon_{t-q}$. The prediction error is given by the residual

$$Y_t - \hat{Y}_t = \varepsilon_t.$$

Suppose that $\hat{\varepsilon}_t$ is an estimator of ε_t , $t \leq n$, which depends on the choice of constants $a_1, \dots, a_p, b_1, \dots, b_q$ and satisfies the recursion

$$\hat{\varepsilon}_t = y_t - a_1 y_{t-1} - \dots - a_p y_{t-p} - b_1 \hat{\varepsilon}_{t-1} - \dots - b_q \hat{\varepsilon}_{t-q}.$$

The function

$$\begin{aligned}
 S^2(a_1, \dots, a_p, b_1, \dots, b_q) \\
 &= \sum_{t=-\infty}^n \hat{\varepsilon}_t^2 \\
 &= \sum_{t=-\infty}^n (y_t - a_1 y_{t-1} - \dots - a_p y_{t-p} - b_1 \hat{\varepsilon}_{t-1} - \dots - b_q \hat{\varepsilon}_{t-q})^2
 \end{aligned}$$

is the *residual sum of squares* and the *least squares approach* suggests to estimate the underlying set of constants by minimizers a_1, \dots, a_p and b_1, \dots, b_q of S^2 . Note that the residuals $\hat{\varepsilon}_t$ and the constants are nested.

We have no observation y_t available for $t \leq 0$. But from the assumption $E(\varepsilon_t) = 0$ and thus $E(Y_t) = 0$, it is reasonable to *backforecast* y_t by zero and to put $\hat{\varepsilon}_t := 0$ for $t \leq 0$, leading to

$$S^2(a_1, \dots, a_p, b_1, \dots, b_q) = \sum_{t=1}^n \hat{\varepsilon}_t^2.$$

The estimated residuals $\hat{\varepsilon}_t$ can then be computed from the recursion

$$\begin{aligned}
 \hat{\varepsilon}_1 &= y_1 \\
 \hat{\varepsilon}_2 &= y_2 - a_1 y_1 - b_1 \hat{\varepsilon}_1 \\
 \hat{\varepsilon}_3 &= y_3 - a_1 y_2 - a_2 y_1 - b_1 \hat{\varepsilon}_2 - b_2 \hat{\varepsilon}_1 \\
 &\vdots \\
 \hat{\varepsilon}_j &= y_j - a_1 y_{j-1} - \dots - a_p y_{j-p} - b_1 \hat{\varepsilon}_{j-1} - \dots - b_q \hat{\varepsilon}_{j-q},
 \end{aligned} \tag{2.23}$$

where j now runs from $\max\{p, q\} + 1$ to n .

For example for an ARMA(2, 3)–process we have

$$\begin{aligned}
 \hat{\varepsilon}_1 &= y_1 \\
 \hat{\varepsilon}_2 &= y_2 - a_1 y_1 - b_1 \hat{\varepsilon}_1 \\
 \hat{\varepsilon}_3 &= y_3 - a_1 y_2 - a_2 y_1 - b_1 \hat{\varepsilon}_2 - b_2 \hat{\varepsilon}_1 \\
 \hat{\varepsilon}_4 &= y_4 - a_1 y_3 - a_2 y_2 - b_1 \hat{\varepsilon}_3 - b_2 \hat{\varepsilon}_2 - b_3 \hat{\varepsilon}_1 \\
 \hat{\varepsilon}_5 &= y_5 - a_1 y_4 - a_2 y_3 - b_1 \hat{\varepsilon}_4 - b_2 \hat{\varepsilon}_3 - b_3 \hat{\varepsilon}_2 \\
 &\vdots
 \end{aligned}$$

From step 4 in this iteration procedure the order $(2, 3)$ has been attained.

The coefficients a_1, \dots, a_p of a pure $\text{AR}(p)$ -process can be estimated directly, using the Yule–Walker equations as described in (2.8).

Diagnostic Check

Suppose that the orders p and q as well as the constants a_1, \dots, a_p and b_1, \dots, b_q have been chosen in order to model an $\text{ARMA}(p, q)$ -process underlying the data. The *Portmanteau-test* of Box and Pierce (1970) checks, whether the estimated residuals $\hat{\varepsilon}_t$, $t = 1, \dots, n$, behave approximately like realizations from a white noise process. To this end one considers the pertaining empirical autocorrelation function

$$\hat{r}_{\hat{\varepsilon}}(k) := \frac{\sum_{j=1}^{n-k} (\hat{\varepsilon}_j - \bar{\varepsilon})(\hat{\varepsilon}_{j+k} - \bar{\varepsilon})}{\sum_{j=1}^n (\hat{\varepsilon}_j - \bar{\varepsilon})^2}, \quad k = 1, \dots, n-1, \quad (2.24)$$

where $\bar{\varepsilon} = n^{-1} \sum_{j=1}^n \hat{\varepsilon}_j$, and checks, whether the values $\hat{r}_{\hat{\varepsilon}}(k)$ are sufficiently close to zero. This decision is based on

$$Q(K) := n \sum_{k=1}^K \hat{r}_{\hat{\varepsilon}}(k),$$

which follows asymptotically for $n \rightarrow \infty$ and K not too large a χ^2 -distribution with $K - p - q$ degrees of freedom if (Y_t) is actually an $\text{ARMA}(p, q)$ -process (see e.g. Brockwell and Davis (1991, Section 9.4)). The parameter K must be chosen such that the sample size $n - k$ in $\hat{r}_{\hat{\varepsilon}}(k)$ is large enough to give a stable estimate of the autocorrelation function. The hypothesis H_0 that the estimated residuals result from a white noise process and therefore the $\text{ARMA}(p, q)$ -model is rejected if the *p-value* $1 - \chi_{K-p-q}^2(Q(K))$ is too small, since in this case the value $Q(K)$ is unexpectedly large. By χ_{K-p-q}^2 we denote the distribution function of the χ^2 -distribution with $K - p - q$ degrees of freedom. To accelerate the convergence to the χ_{K-p-q}^2 distribution under the null hypothesis of an $\text{ARMA}(p, q)$ -process, one often replaces the Box–Pierce statistic $Q(K)$ by the Box–Ljung statistic (Ljung and Box,

1978)

$$Q^*(K) := n \sum_{k=1}^K \left(\left(\frac{n+2}{n-k} \right)^{1/2} \hat{r}_{\hat{\varepsilon}}(k) \right)^2 = n(n+2) \sum_{k=1}^K \frac{1}{n-k} \hat{r}_{\hat{\varepsilon}}(k)$$

with weighted empirical autocorrelations.

Another method for diagnostic check is *overfitting*, which will be presented in Section 7.6 within the application of the Box–Jenkins Program in a case study.

Forecasting

We want to determine weights $c_0^*, \dots, c_{n-1}^* \in \mathbb{R}$ such that for $h \in \mathbb{N}$

$$\mathbb{E} \left(Y_{n+h} - \hat{Y}_{n+h} \right) = \min_{c_0, \dots, c_{n-1} \in \mathbb{R}} \mathbb{E} \left(\left(Y_{n+h} - \sum_{u=0}^{n-1} c_u Y_{n-u} \right)^2 \right),$$

with $\hat{Y}_{n+h} := \sum_{u=0}^{n-1} c_u^* Y_{n-u}$. Then \hat{Y}_{n+h} with minimum mean squared error is said to be a best (linear) h -step forecast of Y_{n+h} , based on Y_1, \dots, Y_n . The following result provides a sufficient condition for the optimality of weights.

Lemma 2.3.2. *Let (Y_t) be an arbitrary stochastic process with finite second moments and $h \in \mathbb{N}$. If the weights c_0^*, \dots, c_{n-1}^* have the property that*

$$\mathbb{E} \left(Y_i \left(Y_{n+h} - \sum_{u=0}^{n-1} c_u^* Y_{n-u} \right) \right) = 0, \quad i = 1, \dots, n, \quad (2.25)$$

then $\hat{Y}_{n+h} := \sum_{u=0}^{n-1} c_u^ Y_{n-u}$ is a best h -step forecast of Y_{n+h} .*

Proof. Let $\tilde{Y}_{n+h} := \sum_{u=0}^{n-1} c_u Y_{n-u}$ be an arbitrary forecast, based on

Y_1, \dots, Y_n . Then we have

$$\begin{aligned}
 & E((Y_{n+h} - \tilde{Y}_{n+h})^2) \\
 &= E((Y_{n+h} - \hat{Y}_{n+h} + \hat{Y}_{n+h} - \tilde{Y}_{n+h})^2) \\
 &= E((Y_{n+h} - \hat{Y}_{n+h})^2) + 2 \sum_{u=0}^{n-1} (c_u^* - c_u) E(Y_{n-u}(Y_{n+h} - \hat{Y}_{n+h})) \\
 &\quad + E((\hat{Y}_{n+h} - \tilde{Y}_{n+h})^2) \\
 &= E((Y_{n+h} - \hat{Y}_{n+h})^2) + E((\hat{Y}_{n+h} - \tilde{Y}_{n+h})^2) \\
 &\geq E((Y_{n+h} - \hat{Y}_{n+h})^2).
 \end{aligned}$$

□

Suppose that (Y_t) is a stationary process with mean zero and autocorrelation function ρ . The equations (2.25) are then of Yule–Walker type

$$\rho(h+s) = \sum_{u=0}^{n-1} c_u^* \rho(s-u), \quad s = 0, 1, \dots, n-1,$$

or, in matrix language

$$\begin{pmatrix} \rho(h) \\ \rho(h+1) \\ \vdots \\ \rho(h+n-1) \end{pmatrix} = \mathbf{P}_n \begin{pmatrix} c_0^* \\ c_1^* \\ \vdots \\ c_{n-1}^* \end{pmatrix} \quad (2.26)$$

with the matrix \mathbf{P}_n as defined in (2.9). If this matrix is invertible, then

$$\begin{pmatrix} c_0^* \\ \vdots \\ c_{n-1}^* \end{pmatrix} := \mathbf{P}_n^{-1} \begin{pmatrix} \rho(h) \\ \vdots \\ \rho(h+n-1) \end{pmatrix} \quad (2.27)$$

is the uniquely determined solution of (2.26).

If we put $h = 1$, then equation (2.27) implies that c_{n-1}^* equals the partial autocorrelation coefficient $\alpha(n)$. In this case, $\alpha(n)$ is the coefficient of Y_1 in the best linear one-step forecast $\hat{Y}_{n+1} = \sum_{u=0}^{n-1} c_u^* Y_{n-u}$ of Y_{n+1} .

Example 2.3.3. Consider the MA(1)-process $Y_t = \varepsilon_t + a\varepsilon_{t-1}$ with $E(\varepsilon_0) = 0$. Its autocorrelation function is by Example 2.2.2 given by $\rho(0) = 1$, $\rho(1) = a/(1 + a^2)$, $\rho(u) = 0$ for $u \geq 2$. The matrix \mathbf{P}_n equals therefore

$$\mathbf{P}_n = \begin{pmatrix} 1 & \frac{a}{1+a^2} & 0 & 0 & \cdots & 0 \\ \frac{a}{1+a^2} & 1 & \frac{a}{1+a^2} & 0 & & 0 \\ 0 & \frac{a}{1+a^2} & 1 & \frac{a}{1+a^2} & & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & & & 1 & \frac{a}{1+a^2} \\ 0 & 0 & \cdots & & \frac{a}{1+a^2} & 1 \end{pmatrix}.$$

Check that the matrix $\mathbf{P}_n = (\text{Corr}(Y_i, Y_j))_{1 \leq i, j \leq n}$ is positive definite, $\mathbf{x}^T \mathbf{P}_n \mathbf{x} > 0$ for any $\mathbf{x} \in \mathbb{R}^n$ unless $\mathbf{x} = \mathbf{0}$ (Exercise 2.44), and thus, \mathbf{P}_n is invertible. The best forecast of Y_{n+1} is by (2.27), therefore, $\sum_{u=0}^{n-1} c_u^* Y_{n-u}$ with

$$\begin{pmatrix} c_0^* \\ \vdots \\ c_{n-1}^* \end{pmatrix} = \mathbf{P}_n^{-1} \begin{pmatrix} \frac{a}{1+a^2} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

which is $a/(1 + a^2)$ times the first column of \mathbf{P}_n^{-1} . The best forecast of Y_{n+h} for $h \geq 2$ is by (2.27) the constant 0. Note that Y_{n+h} is for $h \geq 2$ uncorrelated with Y_1, \dots, Y_n and thus not really predictable by Y_1, \dots, Y_n .

Theorem 2.3.4. Suppose that $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \varepsilon_t$, $t \in \mathbb{Z}$, is an AR(p)-process, which satisfies the stationarity condition (2.4) and has zero mean $E(Y_0) = 0$. Let $n \geq p$. The best one-step forecast is

$$\hat{Y}_{n+1} = a_1 Y_n + a_2 Y_{n-1} + \cdots + a_p Y_{n+1-p}$$

and the best two-step forecast is

$$\hat{Y}_{n+2} = a_1 \hat{Y}_{n+1} + a_2 Y_n + \cdots + a_p Y_{n+2-p}.$$

The best h -step forecast for arbitrary $h \geq 2$ is recursively given by

$$\hat{Y}_{n+h} = a_1 \hat{Y}_{n+h-1} + \cdots + a_{h-1} \hat{Y}_{n+1} + a_h Y_n + \cdots + a_p Y_{n+h-p}.$$

Proof. Since (Y_t) satisfies the stationarity condition (2.4), it is invertible by Theorem 2.2.3 i.e., there exists an absolutely summable causal filter $(b_u)_{u \geq 0}$ such that $Y_t = \sum_{u \geq 0} b_u \varepsilon_{t-u}$, $t \in \mathbb{Z}$, almost surely. This implies in particular $E(Y_t \varepsilon_{t+h}) = \sum_{u \geq 0} b_u E(\varepsilon_{t-u} \varepsilon_{t+h}) = 0$ for any $h \geq 1$, cf. Theorem 2.1.5. Hence we obtain for $i = 1, \dots, n$

$$E((Y_{n+1} - \hat{Y}_{n+1})Y_i) = E(\varepsilon_{n+1}Y_i) = 0$$

from which the assertion for $h = 1$ follows by Lemma 2.3.2. The case of an arbitrary $h \geq 2$ is now a consequence of the recursion

$$\begin{aligned} & E((Y_{n+h} - \hat{Y}_{n+h})Y_i) \\ &= E \left(\left(\varepsilon_{n+h} + \sum_{u=1}^{\min(h-1,p)} a_u Y_{n+h-u} - \sum_{u=1}^{\min(h-1,p)} a_u \hat{Y}_{n+h-u} \right) Y_i \right) \\ &= \sum_{u=1}^{\min(h-1,p)} a_u E \left((Y_{n+h-u} - \hat{Y}_{n+h-u}) Y_i \right) = 0, \quad i = 1, \dots, n, \end{aligned}$$

and Lemma 2.3.2. □

A repetition of the arguments in the preceding proof implies the following result, which shows that for an ARMA(p, q)-process the forecast of Y_{n+h} for $h > q$ is controlled only by the AR-part of the process.

Theorem 2.3.5. *Suppose that $Y_t = \sum_{u=1}^p a_u Y_{t-u} + \varepsilon_t + \sum_{v=1}^q b_v \varepsilon_{t-v}$, $t \in \mathbb{Z}$, is an ARMA(p, q)-process, which satisfies the stationarity condition (2.4) and has zero mean, precisely $E(\varepsilon_0) = 0$. Suppose that $n + q - p \geq 0$. The best h -step forecast of Y_{n+h} for $h > q$ satisfies the recursion*

$$\hat{Y}_{n+h} = \sum_{u=1}^p a_u \hat{Y}_{n+h-u}.$$

Example 2.3.6. We illustrate the best forecast of the ARMA(1, 1)-process

$$Y_t = 0.4Y_{t-1} + \varepsilon_t - 0.6\varepsilon_{t-1}, \quad t \in \mathbb{Z},$$

with $E(Y_t) = E(\varepsilon_t) = 0$. First we need the optimal 1-step forecast \hat{Y}_i for $i = 1, \dots, n$. These are defined by putting unknown values of Y_t

with an index $t \leq 0$ equal to their expected value, which is zero. We, thus, obtain

$$\begin{aligned}
 \hat{Y}_1 &:= 0, & \hat{\varepsilon}_1 &:= Y_1 - \hat{Y}_1 = Y_1, \\
 \hat{Y}_2 &:= 0.4Y_1 + 0 - 0.6\hat{\varepsilon}_1 \\
 &= -0.2Y_1, & \hat{\varepsilon}_2 &:= Y_2 - \hat{Y}_2 = Y_2 + 0.2Y_1, \\
 \hat{Y}_3 &:= 0.4Y_2 + 0 - 0.6\hat{\varepsilon}_2 \\
 &= 0.4Y_2 - 0.6(Y_2 + 0.2Y_1) \\
 &= -0.2Y_2 - 0.12Y_1, & \hat{\varepsilon}_3 &:= Y_3 - \hat{Y}_3, \\
 &\vdots & &\vdots
 \end{aligned}$$

until \hat{Y}_i and $\hat{\varepsilon}_i$ are defined for $i = 1, \dots, n$. The actual forecast is then given by

$$\begin{aligned}
 \hat{Y}_{n+1} &= 0.4Y_n + 0 - 0.6\hat{\varepsilon}_n = 0.4Y_n - 0.6(Y_n - \hat{Y}_n), \\
 \hat{Y}_{n+2} &= 0.4\hat{Y}_{n+1} + 0 + 0, \\
 &\vdots \\
 \hat{Y}_{n+h} &= 0.4\hat{Y}_{n+h-1} = \dots = 0.4^{h-1}\hat{Y}_{n+1} \xrightarrow{h \rightarrow \infty} 0,
 \end{aligned}$$

where ε_t with index $t \geq n+1$ is replaced by zero, since it is uncorrelated with Y_i , $i \leq n$.

In practice one replaces the usually unknown coefficients a_u , b_v in the above forecasts by their estimated values.

Exercises

2.1. Show that the expectation of complex valued random variables is linear, i.e.,

$$E(aY + bZ) = aE(Y) + bE(Z),$$

where $a, b \in \mathbb{C}$ and Y, Z are integrable.

Show that

$$\text{Cov}(Y, Z) = E(Y\bar{Z}) - E(Y)E(\bar{Z})$$

for square integrable complex random variables Y and Z .

2.2. Suppose that the complex random variables Y and Z are square integrable. Show that

$$\text{Cov}(aY + b, Z) = a \text{Cov}(Y, Z), \quad a, b \in \mathbb{C}.$$

2.3. Give an example of a stochastic process (Y_t) such that for arbitrary $t_1, t_2 \in \mathbb{Z}$ and $k \neq 0$

$$E(Y_{t_1}) \neq E(Y_{t_1+k}) \quad \text{but} \quad \text{Cov}(Y_{t_1}, Y_{t_2}) = \text{Cov}(Y_{t_1+k}, Y_{t_2+k}).$$

2.4. Let $(X_t), (Y_t)$ be stationary processes such that $\text{Cov}(X_t, Y_s) = 0$ for $t, s \in \mathbb{Z}$. Show that for arbitrary $a, b \in \mathbb{C}$ the linear combinations $(aX_t + bY_t)$ yield a stationary process.

Suppose that the decomposition $Z_t = X_t + Y_t$, $t \in \mathbb{Z}$ holds. Show that stationarity of (Z_t) does not necessarily imply stationarity of (X_t) .

2.5. Show that the process $Y_t = X e^{iat}$, $a \in \mathbb{R}$, is stationary, where X is a complex valued random variable with mean zero and finite variance.

Show that the random variable $Y = b e^{iU}$ has mean zero, where U is a uniformly distributed random variable on $(0, 2\pi)$ and $b \in \mathbb{C}$.

2.6. Let Z_1, Z_2 be independent and normal $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, distributed random variables and choose $\lambda \in \mathbb{R}$. For which means $\mu_1, \mu_2 \in \mathbb{R}$ and variances $\sigma_1^2, \sigma_2^2 > 0$ is the *cosinoid process*

$$Y_t = Z_1 \cos(2\pi\lambda t) + Z_2 \sin(2\pi\lambda t), \quad t \in \mathbb{Z}$$

stationary?

2.7. Show that the autocovariance function $\gamma : \mathbb{Z} \rightarrow \mathbb{C}$ of a complex-valued stationary process $(Y_t)_{t \in \mathbb{Z}}$, which is defined by

$$\gamma(h) = E(Y_{t+h} \bar{Y}_t) - E(Y_{t+h}) E(\bar{Y}_t), \quad h \in \mathbb{Z},$$

has the following properties: $\gamma(0) \geq 0$, $|\gamma(h)| \leq \gamma(0)$, $\gamma(h) = \overline{\gamma(-h)}$, i.e., γ is a Hermitian function, and $\sum_{1 \leq r, s \leq n} z_r \gamma(r-s) \bar{z}_s \geq 0$ for $z_1, \dots, z_n \in \mathbb{C}$, $n \in \mathbb{N}$, i.e., γ is a positive semidefinite function.

2.8. Suppose that Y_t , $t = 1, \dots, n$, is a stationary process with mean μ . Then $\hat{\mu}_n := n^{-1} \sum_{t=1}^n Y_t$ is an unbiased estimator of μ . Express the mean square error $E(\hat{\mu}_n - \mu)^2$ in terms of the autocovariance function γ and show that $E(\hat{\mu}_n - \mu)^2 \rightarrow 0$ if $\gamma(n) \rightarrow 0$, $n \rightarrow \infty$.

2.9. Suppose that $(Y_t)_{t \in \mathbb{Z}}$ is a stationary process and denote by

$$c(k) := \begin{cases} \frac{1}{n} \sum_{t=1}^{n-|k|} (Y_t - \bar{Y})(Y_{t+|k|} - \bar{Y}), & |k| = 0, \dots, n-1, \\ 0, & |k| \geq n. \end{cases}$$

the empirical autocovariance function at lag k , $k \in \mathbb{Z}$.

(i) Show that $c(k)$ is a biased estimator of $\gamma(k)$ (even if the factor n^{-1} is replaced by $(n-k)^{-1}$) i.e., $E(c(k)) \neq \gamma(k)$.

(ii) Show that the k -dimensional empirical covariance matrix

$$\mathbf{C}_k := \begin{pmatrix} c(0) & c(1) & \dots & c(k-1) \\ c(1) & c(0) & & c(k-2) \\ \vdots & & \ddots & \vdots \\ c(k-1) & c(k-2) & \dots & c(0) \end{pmatrix}$$

is positive semidefinite. (If the factor n^{-1} in the definition of $c(j)$ is replaced by $(n-j)^{-1}$, $j = 1, \dots, k$, the resulting covariance matrix may not be positive semidefinite.) Hint: Consider $k \geq n$ and write $\mathbf{C}_k = n^{-1} \mathbf{A} \mathbf{A}^T$ with a suitable $k \times 2k$ -matrix \mathbf{A} . Show further that \mathbf{C}_m is positive semidefinite if \mathbf{C}_k is positive semidefinite for $k > m$.

(iii) If $c(0) > 0$, then \mathbf{C}_k is nonsingular, i.e., \mathbf{C}_k is positive definite.

2.10. Suppose that (Y_t) is a stationary process with autocovariance function γ_Y . Express the autocovariance function of the difference filter of first order $\Delta Y_t = Y_t - Y_{t-1}$ in terms of γ_Y . Find it when $\gamma_Y(k) = \lambda^{|k|}$.

2.11. Let $(Y_t)_{t \in \mathbb{Z}}$ be a stationary process with mean zero. If its autocovariance function satisfies $\gamma(\tau) = 0$ for some $\tau > 0$, then γ is periodic with length τ , i.e., $\gamma(t + \tau) = \gamma(t)$, $t \in \mathbb{Z}$.

2.12. Let (Y_t) be a stochastic process such that for $t \in \mathbb{Z}$

$$P\{Y_t = 1\} = p_t = 1 - P\{Y_t = -1\}, \quad 0 < p_t < 1.$$

Suppose in addition that (Y_t) is a *Markov process*, i.e., for any $t \in \mathbb{Z}$, $k \geq 1$

$$P(Y_t = y_0 | Y_{t-1} = y_1, \dots, Y_{t-k} = y_k) = P(Y_t = y_0 | Y_{t-1} = y_1).$$

- (i) Is $(Y_t)_{t \in \mathbb{N}}$ a stationary process?
- (ii) Compute the autocovariance function in case $P(Y_t = 1 | Y_{t-1} = 1) = \lambda$ and $p_t = 1/2$.

2.13. Let $(\varepsilon_t)_t$ be a white noise process with independent $\varepsilon_t \sim N(0, 1)$ and define

$$\tilde{\varepsilon}_t = \begin{cases} \varepsilon_t, & \text{if } t \text{ is even,} \\ (\varepsilon_{t-1}^2 - 1)/\sqrt{2}, & \text{if } t \text{ is odd.} \end{cases}$$

Show that $(\tilde{\varepsilon}_t)_t$ is a white noise process with $E(\tilde{\varepsilon}_t) = 0$ and $\text{Var}(\tilde{\varepsilon}_t) = 1$, where the $\tilde{\varepsilon}_t$ are neither independent nor identically distributed. Plot the path of $(\varepsilon_t)_t$ and $(\tilde{\varepsilon}_t)_t$ for $t = 1, \dots, 100$ and compare!

2.14. Let $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a white noise. The process $Y_t = \sum_{s=1}^t \varepsilon_s$ is said to be a *random walk*. Plot the path of a random walk with normal $N(\mu, \sigma^2)$ distributed ε_t for each of the cases $\mu < 0$, $\mu = 0$ and $\mu > 0$.

2.15. Let $(a_u), (b_v)$ be absolutely summable filters and let (Z_t) be a stochastic process with $\sup_{t \in \mathbb{Z}} E(Z_t^2) < \infty$. Put for $t \in \mathbb{Z}$

$$X_t = \sum_u a_u Z_{t-u}, \quad Y_t = \sum_v b_v Z_{t-v}.$$

Then we have

$$E(X_t Y_t) = \sum_u \sum_v a_u b_v E(Z_{t-u} Z_{t-v}).$$

Hint: Use the general inequality $|xy| \leq (x^2 + y^2)/2$.

2.16. Show the equality

$$E((Y_t - \mu_Y)(\overline{Y_s - \mu_Y})) = \lim_{n \rightarrow \infty} \text{Cov} \left(\sum_{u=-n}^n a_u Z_{t-u}, \sum_{w=-n}^n a_w Z_{s-w} \right)$$

in the proof of Theorem 2.1.6.

2.17. Let $Y_t = aY_{t-1} + \varepsilon_t$, $t \in \mathbb{Z}$ be an AR(1)-process with $|a| > 1$. Compute the autocorrelation function of this process.

2.18. Compute the orders p and the coefficients a_u of the process $Y_t = \sum_{u=0}^p a_u \varepsilon_{t-u}$ with $\text{Var}(\varepsilon_0) = 1$ and autocovariance function $\gamma(1) = 2$, $\gamma(2) = 1$, $\gamma(3) = -1$ and $\gamma(t) = 0$ for $t \geq 4$. Is this process invertible?

2.19. The autocorrelation function ρ of an arbitrary MA(q)-process satisfies

$$-\frac{1}{2} \leq \sum_{v=1}^q \rho(v) \leq \frac{1}{2}q.$$

Give examples of MA(q)-processes, where the lower bound and the upper bound are attained, i.e., these bounds are sharp.

2.20. Let $(Y_t)_{t \in \mathbb{Z}}$ be a stationary stochastic process with $E(Y_t) = 0$, $t \in \mathbb{Z}$, and

$$\rho(t) = \begin{cases} 1 & \text{if } t = 0 \\ \rho(1) & \text{if } t = 1 \\ 0 & \text{if } t > 1, \end{cases}$$

where $|\rho(1)| < 1/2$. Then there exists $a \in (-1, 1)$ and a white noise $(\varepsilon_t)_{t \in \mathbb{Z}}$ such that

$$Y_t = \varepsilon_t + a\varepsilon_{t-1}.$$

Hint: Example 2.2.2.

2.21. Find two MA(1)-processes with the same autocovariance functions.

2.22. Suppose that $Y_t = \varepsilon_t + a\varepsilon_{t-1}$ is a noninvertible MA(1)-process, where $|a| > 1$. Define the new process

$$\tilde{\varepsilon}_t = \sum_{j=0}^{\infty} (-a)^{-j} Y_{t-j}$$

and show that $(\tilde{\varepsilon}_t)$ is a white noise. Show that $\text{Var}(\tilde{\varepsilon}_t) = a^2 \text{Var}(\varepsilon_t)$ and (Y_t) has the *invertible* representation

$$Y_t = \tilde{\varepsilon}_t + a^{-1}\tilde{\varepsilon}_{t-1}.$$

2.23. Plot the autocorrelation functions of MA(p)-processes for different values of p .

2.24. Generate and plot AR(3)-processes (Y_t) , $t = 1, \dots, 500$ where the roots of the characteristic polynomial have the following properties:

- (i) all roots are outside the unit disk,
- (ii) all roots are inside the unit disk,
- (iii) all roots are on the unit circle,
- (iv) two roots are outside, one root inside the unit disk,
- (v) one root is outside, one root is inside the unit disk and one root is on the unit circle,
- (vi) all roots are outside the unit disk but close to the unit circle.

2.25. Show that the AR(2)-process $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \varepsilon_t$ for $a_1 = 1/3$ and $a_2 = 2/9$ has the autocorrelation function

$$\rho(k) = \frac{16}{21} \left(\frac{2}{3}\right)^{|k|} + \frac{5}{21} \left(-\frac{1}{3}\right)^{|k|}, \quad k \in \mathbb{Z}$$

and for $a_1 = a_2 = 1/12$ the autocorrelation function

$$\rho(k) = \frac{45}{77} \left(\frac{1}{3}\right)^{|k|} + \frac{32}{77} \left(-\frac{1}{4}\right)^{|k|}, \quad k \in \mathbb{Z}.$$

2.26. Let (ε_t) be a white noise with $E(\varepsilon_0) = \mu$, $\text{Var}(\varepsilon_0) = \sigma^2$ and put

$$Y_t = \varepsilon_t - Y_{t-1}, \quad t \in \mathbb{N}, Y_0 = 0.$$

Show that

$$\text{Corr}(Y_s, Y_t) = (-1)^{s+t} \min\{s, t\} / \sqrt{st}.$$

2.27. An AR(2)-process $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \varepsilon_t$ satisfies the stationarity condition (2.4), if the pair (a_1, a_2) is in the triangle

$$\Delta := \left\{ (\alpha, \beta) \in \mathbb{R}^2 : -1 < \beta < 1, \alpha + \beta < 1 \text{ and } \beta - \alpha < 1 \right\}.$$

Hint: Use the fact that necessarily $\rho(1) \in (-1, 1)$.

2.28. Let (Y_t) denote the unique stationary solution of the autoregressive equations

$$Y_t = aY_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z},$$

with $|a| > 1$. Then (Y_t) is given by the expression

$$Y_t = - \sum_{j=1}^{\infty} a^{-j} \varepsilon_{t+j}$$

(see the proof of Lemma 2.1.10). Define the new process

$$\tilde{\varepsilon}_t = Y_t - \frac{1}{a} Y_{t-1},$$

and show that $(\tilde{\varepsilon}_t)$ is a white noise with $\text{Var}(\tilde{\varepsilon}_t) = \text{Var}(\varepsilon_t)/a^2$. These calculations show that (Y_t) is the (unique stationary) solution of the causal AR-equations

$$Y_t = \frac{1}{a} Y_{t-1} + \tilde{\varepsilon}_t, \quad t \in \mathbb{Z}.$$

Thus, every AR(1)-process with $|a| > 1$ can be represented as an AR(1)-process with $|a| < 1$ and a new white noise.

Show that for $|a| = 1$ the above autoregressive equations have *no stationary* solutions. A stationary solution exists if the white noise process is degenerated, i.e., $E(\varepsilon_t^2) = 0$.

2.29. Consider the process

$$\tilde{Y}_t := \begin{cases} \varepsilon_1 & \text{for } t = 1 \\ aY_{t-1} + \varepsilon_t & \text{for } t > 1, \end{cases}$$

i.e., $\tilde{Y}_t, t \geq 1$, equals the AR(1)-process $Y_t = aY_{t-1} + \varepsilon_t$, conditional on $Y_0 = 0$. Compute $E(\tilde{Y}_t)$, $\text{Var}(\tilde{Y}_t)$ and $\text{Cov}(Y_t, Y_{t+s})$. Is there something like asymptotic stationarity for $t \rightarrow \infty$?

Choose $a \in (-1, 1)$, $a \neq 0$, and compute the correlation matrix of Y_1, \dots, Y_{10} .

2.30. Use the IML function ARMASIM to simulate the stationary AR(2)-process

$$Y_t = -0.3Y_{t-1} + 0.3Y_{t-2} + \varepsilon_t.$$

Estimate the parameters $a_1 = -0.3$ and $a_2 = 0.3$ by means of the Yule–Walker equations using the SAS procedure PROC ARIMA.

2.31. Show that the value at lag 2 of the partial autocorrelation function of the MA(1)-process

$$Y_t = \varepsilon_t + a\varepsilon_{t-1}, \quad t \in \mathbb{Z}$$

is

$$\alpha(2) = -\frac{a^2}{1 + a^2 + a^4}.$$

2.32. (Unemployed1 Data) Plot the empirical autocorrelations and partial autocorrelations of the trend and seasonally adjusted Unemployed1 Data from the building trade, introduced in Example 1.1.1. Apply the Box–Jenkins program. Is a fit of a pure MA(q)- or AR(p)-process reasonable?

2.33. Plot the autocorrelation functions of ARMA(p, q)-processes for different values of p, q using the IML function ARMACOV. Plot also their empirical counterparts.

2.34. Compute the autocovariance and autocorrelation function of an ARMA(1, 2)-process.

2.35. Derive the least squares normal equations for an $AR(p)$ -process and compare them with the Yule–Walker equations.

2.36. Let $(\varepsilon_t)_{t \in \mathbb{Z}}$ be a white noise. The process $W_t = \sum_{s=1}^t \varepsilon_s$ is then called a random walk. Generate two independent random walks $\mu_t, \nu_t, t = 1, \dots, 100$, where the ε_t are standard normal and independent. Simulate from these

$$X_t = \mu_t + \delta_t^{(1)}, \quad Y_t = \mu_t + \delta_t^{(2)}, \quad Z_t = \nu_t + \delta_t^{(3)},$$

where the $\delta_t^{(i)}$ are again independent and standard normal, $i = 1, 2, 3$. Plot the generated processes X_t, Y_t and Z_t and their first order differences. Which of these series are cointegrated? Check this by the Phillips-Ouliaris-test.

2.37. (US Interest Data) The file “us_interest_rates.txt” contains the interest rates of three-month, three-year and ten-year federal bonds of the USA, monthly collected from July 1954 to December 2002. Plot the data and the corresponding differences of first order. Test also whether the data are $I(1)$. Check next if the three series are pairwise cointegrated.

2.38. Show that the density of the t -distribution with m degrees of freedom converges to the density of the standard normal distribution as m tends to infinity. Hint: Apply the dominated convergence theorem (Lebesgue).

2.39. Let $(Y_t)_t$ be a stationary and causal ARCH(1)-process with $|a_1| < 1$.

(i) Show that $Y_t^2 = a_0 \sum_{j=0}^{\infty} a_1^j Z_t^2 Z_{t-1}^2 \cdots Z_{t-j}^2$ with probability one.

(ii) Show that $E(Y_t^2) = a_0/(1 - a_1)$.

(iii) Evaluate $E(Y_t^4)$ and deduce that $E(Z_1^4)a_1^2 < 1$ is a sufficient condition for $E(Y_t^4) < \infty$.

Hint: Theorem 2.1.5.

2.40. Determine the joint density of Y_{p+1}, \dots, Y_n for an ARCH(p)-process Y_t with normal distributed Z_t given that $Y_1 = y_1, \dots, Y_p = y_p$. Hint: Recall that the joint density $f_{X,Y}$ of a random vector (X, Y) can be written in the form $f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x)$, where $f_{Y|X}(y|x) := f_{X,Y}(x, y)/f_X(x)$ if $f_X(x) > 0$, and $f_{Y|X}(y|x) := f_Y(y)$, else, is the (conditional) density of Y given $X = x$ and f_X, f_Y is the (marginal) density of X, Y .

2.41. Generate an ARCH(1)-process $(Y_t)_t$ with $a_0 = 1$ and $a_1 = 0.5$. Plot $(Y_t)_t$ as well as $(Y_t^2)_t$ and its partial autocorrelation function. What is the value of the partial autocorrelation coefficient at lag 1 and lag 2? Use PROC ARIMA to estimate the parameter of the AR(1)-process $(Y_t^2)_t$ and apply the Box–Ljung test.

2.42. (Hong Kong Data) Fit an GARCH(p, q)-model to the daily Hang Seng closing index of Hong Kong stock prices from July 16, 1981, to September 31, 1983. Consider in particular the cases $p = q = 2$ and $p = 3, q = 2$.

2.43. (Zurich Data) The daily value of the Zurich stock index was recorded between January 1st, 1988 and December 31st, 1988. Use a difference filter of first order to remove a possible trend. Plot the (trend-adjusted) data, their squares, the pertaining partial autocorrelation function and parameter estimates. Can the squared process be considered as an AR(1)-process?

2.44. Show that the matrix Σ'^{-1} in Example 2.3.1 has the determinant $1 - a^2$.

Show that the matrix \mathbf{P}_n in Example 2.3.3 has the determinant $(1 + a^2 + a^4 + \dots + a^{2n})/(1 + a^2)^n$.

2.45. (Car Data) Apply the Box–Jenkins program to the Car Data.

State-Space Models

In state-space models we have, in general, a nonobservable target process (X_t) and an observable process (Y_t) . They are linked by the assumption that (Y_t) is a linear function of (X_t) with an added noise, where the linear function may vary in time. The aim is the derivation of best linear estimates of X_t , based on $(Y_s)_{s \leq t}$.

3.1 The State-Space Representation

Many models of time series such as ARMA(p, q)-processes can be embedded in state-space models if we allow in the following sequences of random *vectors* $\mathbf{X}_t \in \mathbb{R}^k$ and $\mathbf{Y}_t \in \mathbb{R}^m$.

A multivariate *state-space model* is now defined by the *state equation*

$$\mathbf{X}_{t+1} = \mathbf{A}_t \mathbf{X}_t + \mathbf{B}_t \boldsymbol{\varepsilon}_{t+1} \in \mathbb{R}^k, \quad (3.1)$$

describing the time-dependent behavior of the *state* $\mathbf{X}_t \in \mathbb{R}^k$, and the *observation equation*

$$\mathbf{Y}_t = \mathbf{C}_t \mathbf{X}_t + \boldsymbol{\eta}_t \in \mathbb{R}^m. \quad (3.2)$$

We assume that (\mathbf{A}_t) , (\mathbf{B}_t) and (\mathbf{C}_t) are sequences of known matrices, $(\boldsymbol{\varepsilon}_t)$ and $(\boldsymbol{\eta}_t)$ are uncorrelated sequences of white noises with mean vectors $\mathbf{0}$ and known covariance matrices $\text{Cov}(\boldsymbol{\varepsilon}_t) = \text{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^T) =: \mathbf{Q}_t$, $\text{Cov}(\boldsymbol{\eta}_t) = \text{E}(\boldsymbol{\eta}_t \boldsymbol{\eta}_t^T) =: \mathbf{R}_t$.

We suppose further that \mathbf{X}_0 and $\boldsymbol{\varepsilon}_t, \boldsymbol{\eta}_t, t \geq 1$, are uncorrelated, where two random vectors $\mathbf{W} \in \mathbb{R}^p$ and $\mathbf{V} \in \mathbb{R}^q$ are said to be uncorrelated if their components are i.e., if the matrix of their covariances vanishes

$$\text{E}((\mathbf{W} - \text{E}(\mathbf{W}))(\mathbf{V} - \text{E}(\mathbf{V}))^T) = \mathbf{0}.$$

By $E(\mathbf{W})$ we denote the vector of the componentwise expectations of \mathbf{W} . We say that a time series (\mathbf{Y}_t) has a *state-space representation* if it satisfies the representations (3.1) and (3.2).

Example 3.1.1. Let (η_t) be a white noise in \mathbb{R} and put

$$Y_t := \mu_t + \eta_t$$

with linear trend $\mu_t = a + bt$. This simple model can be represented as a state-space model as follows. Define the state vector \mathbf{X}_t as

$$\mathbf{X}_t := \begin{pmatrix} \mu_t \\ 1 \end{pmatrix},$$

and put

$$\mathbf{A} := \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$$

From the recursion $\mu_{t+1} = \mu_t + b$ we then obtain the state equation

$$\mathbf{X}_{t+1} = \begin{pmatrix} \mu_{t+1} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_t \\ 1 \end{pmatrix} = \mathbf{A}\mathbf{X}_t,$$

and with

$$\mathbf{C} := (1, 0)$$

the observation equation

$$Y_t = (1, 0) \begin{pmatrix} \mu_t \\ 1 \end{pmatrix} + \eta_t = \mathbf{C}\mathbf{X}_t + \eta_t.$$

Note that the state \mathbf{X}_t is nonstochastic i.e., $\mathbf{B}_t = \mathbf{0}$. This model is moreover time-invariant, since the matrices \mathbf{A} , $\mathbf{B} := \mathbf{B}_t$ and \mathbf{C} do not depend on t .

Example 3.1.2. An $\text{AR}(p)$ -process

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t$$

with a white noise (ε_t) has a state-space representation with state vector

$$\mathbf{X}_t = (Y_t, Y_{t-1}, \dots, Y_{t-p+1})^T.$$

If we define the $p \times p$ -matrix \mathbf{A} by

$$\mathbf{A} := \begin{pmatrix} a_1 & a_2 & \dots & a_{p-1} & a_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

and the $p \times 1$ -matrices \mathbf{B} , \mathbf{C}^T by

$$\mathbf{B} := (1, 0, \dots, 0)^T =: \mathbf{C}^T,$$

then we have the state equation

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t + \mathbf{B}\varepsilon_{t+1}$$

and the observation equation

$$Y_t = \mathbf{C}\mathbf{X}_t.$$

Example 3.1.3. For the MA(q)-process

$$Y_t = \varepsilon_t + b_1\varepsilon_{t-1} + \dots + b_q\varepsilon_{t-q}$$

we define the non observable state

$$\mathbf{X}_t := (\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q})^T \in \mathbb{R}^{q+1}.$$

With the $(q+1) \times (q+1)$ -matrix

$$\mathbf{A} := \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 \\ \vdots & & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

the $(q+1) \times 1$ -matrix

$$\mathbf{B} := (1, 0, \dots, 0)^T$$

and the $1 \times (q + 1)$ -matrix

$$\mathbf{C} := (1, b_1, \dots, b_q)$$

we obtain the state equation

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t + \mathbf{B}\varepsilon_{t+1}$$

and the observation equation

$$\mathbf{Y}_t = \mathbf{C}\mathbf{X}_t.$$

Example 3.1.4. Combining the above results for AR(p) and MA(q)-processes, we obtain a state-space representation of ARMA(p, q)-processes

$$Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q}.$$

In this case the state vector can be chosen as

$$\mathbf{X}_t := (Y_t, Y_{t-1}, \dots, Y_{t-p+1}, \varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q+1})^T \in \mathbb{R}^{p+q}.$$

We define the $(p + q) \times (p + q)$ -matrix

$$\mathbf{A} := \begin{pmatrix} a_1 & a_2 & \dots & a_{p-1} & a_p & b_1 & b_2 & \dots & b_{q-1} & b_q \\ 1 & 0 & \dots & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & \vdots & & & & \vdots \\ \vdots & & \ddots & & \vdots & \vdots & & & & \vdots \\ 0 & \dots & 0 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & & & & \vdots & 1 & 0 & \dots & \dots & 0 \\ \vdots & & & & \vdots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & \vdots & \vdots & & \ddots & & \vdots \\ 0 & \dots & \dots & \dots & 0 & 0 & \dots & 0 & 1 & 0 \end{pmatrix},$$

the $(p + q) \times 1$ -matrix

$$\mathbf{B} := (1, 0, \dots, 0, 1, 0, \dots, 0)^T$$

with the entry 1 at the first and $(p + 1)$ -th position and the $1 \times (p + q)$ -matrix

$$\mathbf{C} := (1, 0, \dots, 0).$$

Then we have the state equation

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t + \mathbf{B}\varepsilon_{t+1}$$

and the observation equation

$$Y_t = \mathbf{C}\mathbf{X}_t.$$

Remark 3.1.5. General ARIMA(p, d, q)-processes also have a state-space representation, see exercise 3.2. Consequently the same is true for random walks, which can be written as ARIMA(0, 1, 0)-processes.

3.2 The Kalman-Filter

The key problem in the state-space model (3.1), (3.2) is the estimation of the nonobservable state \mathbf{X}_t . It is possible to derive an estimate of \mathbf{X}_t recursively from an estimate of \mathbf{X}_{t-1} together with the last observation \mathbf{Y}_t , known as the *Kalman recursions* (Kalman 1960). We obtain in this way a unified prediction technique for each time series model that has a state-space representation.

We want to compute the best linear prediction

$$\hat{\mathbf{X}}_t := \mathbf{D}_1\mathbf{Y}_1 + \cdots + \mathbf{D}_t\mathbf{Y}_t \quad (3.3)$$

of \mathbf{X}_t , based on $\mathbf{Y}_1, \dots, \mathbf{Y}_t$ i.e., the $k \times m$ -matrices $\mathbf{D}_1, \dots, \mathbf{D}_t$ are such that the mean squared error is minimized

$$\begin{aligned} & \mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)^T(\mathbf{X}_t - \hat{\mathbf{X}}_t)) \\ &= \mathbb{E}((\mathbf{X}_t - \sum_{j=1}^t \mathbf{D}_j\mathbf{Y}_j)^T(\mathbf{X}_t - \sum_{j=1}^t \mathbf{D}_j\mathbf{Y}_j)) \\ &= \min_{k \times m\text{-matrices } \mathbf{D}'_1, \dots, \mathbf{D}'_t} \mathbb{E}((\mathbf{X}_t - \sum_{j=1}^t \mathbf{D}'_j\mathbf{Y}_j)^T(\mathbf{X}_t - \sum_{j=1}^t \mathbf{D}'_j\mathbf{Y}_j)). \quad (3.4) \end{aligned}$$

By repeating the arguments in the proof of Lemma 2.3.2 we will prove the following result. It states that $\hat{\mathbf{X}}_t$ is a best linear prediction of \mathbf{X}_t based on $\mathbf{Y}_1, \dots, \mathbf{Y}_t$ if each component of the vector $\mathbf{X}_t - \hat{\mathbf{X}}_t$ is orthogonal to each component of the vectors \mathbf{Y}_s , $1 \leq s \leq t$, with respect to the inner product $\mathbb{E}(XY)$ of two random variable X and Y .

Lemma 3.2.1. *If the estimate $\hat{\mathbf{X}}_t$ defined in (3.3) satisfies*

$$\mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)\mathbf{Y}_s^T) = \mathbf{0}, \quad 1 \leq s \leq t, \quad (3.5)$$

then it minimizes the mean squared error (3.4).

Note that $\mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)\mathbf{Y}_s^T)$ is a $k \times m$ -matrix, which is generated by multiplying each component of $\mathbf{X}_t - \hat{\mathbf{X}}_t \in \mathbb{R}^k$ with each component of $\mathbf{Y}_s \in \mathbb{R}^m$.

Proof. Let $\mathbf{X}'_t = \sum_{j=1}^t \mathbf{D}'_j \mathbf{Y}_j \in \mathbb{R}^k$ be an arbitrary linear combination of $\mathbf{Y}_1, \dots, \mathbf{Y}_t$. Then we have

$$\begin{aligned} & \mathbb{E}((\mathbf{X}_t - \mathbf{X}'_t)^T(\mathbf{X}_t - \mathbf{X}'_t)) \\ &= \mathbb{E} \left(\left(\mathbf{X}_t - \hat{\mathbf{X}}_t + \sum_{j=1}^t (\mathbf{D}_j - \mathbf{D}'_j) \mathbf{Y}_j \right)^T \left(\mathbf{X}_t - \hat{\mathbf{X}}_t + \sum_{j=1}^t (\mathbf{D}_j - \mathbf{D}'_j) \mathbf{Y}_j \right) \right) \\ &= \mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)^T(\mathbf{X}_t - \hat{\mathbf{X}}_t)) + 2 \sum_{j=1}^t \mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)^T(\mathbf{D}_j - \mathbf{D}'_j) \mathbf{Y}_j) \\ & \quad + \mathbb{E} \left(\left(\sum_{j=1}^t (\mathbf{D}_j - \mathbf{D}'_j) \mathbf{Y}_j \right)^T \sum_{j=1}^t (\mathbf{D}_j - \mathbf{D}'_j) \mathbf{Y}_j \right) \\ &\geq \mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)^T(\mathbf{X}_t - \hat{\mathbf{X}}_t)), \end{aligned}$$

since in the second-to-last line the final term is nonnegative and the second one vanishes by the property (3.5). \square

Let now $\hat{\mathbf{X}}_{t-1}$ be a linear prediction of \mathbf{X}_{t-1} fulfilling (3.5) based on the observations $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$. Then

$$\tilde{\mathbf{X}}_t := \mathbf{A}_{t-1} \hat{\mathbf{X}}_{t-1} \quad (3.6)$$

is the best linear prediction of \mathbf{X}_t based on $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$, which is easy to see. We simply replaced $\boldsymbol{\varepsilon}_t$ in the state equation by its expectation $\mathbf{0}$. Note that $\boldsymbol{\varepsilon}_t$ and \mathbf{Y}_s are uncorrelated if $s < t$ i.e., $\mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t)\mathbf{Y}_s^T) = \mathbf{0}$ for $1 \leq s \leq t-1$, see Exercise 3.4.

From this we obtain that

$$\tilde{\mathbf{Y}}_t := \mathbf{C}_t \tilde{\mathbf{X}}_t$$

is the best linear prediction of \mathbf{Y}_t based on $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$, since $E((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)\mathbf{Y}_s^T) = E((\mathbf{C}_t(\mathbf{X}_t - \tilde{\mathbf{X}}_t) + \boldsymbol{\eta}_t)\mathbf{Y}_s^T) = \mathbf{0}$, $1 \leq s \leq t-1$; note that $\boldsymbol{\eta}_t$ and \mathbf{Y}_s are uncorrelated if $s < t$, see also Exercise 3.4.

Define now by

$$\hat{\Delta}_t := E((\mathbf{X}_t - \hat{\mathbf{X}}_t)(\mathbf{X}_t - \hat{\mathbf{X}}_t)^T), \quad \tilde{\Delta}_t := E((\mathbf{X}_t - \tilde{\mathbf{X}}_t)(\mathbf{X}_t - \tilde{\mathbf{X}}_t)^T).$$

the covariance matrices of the approximation errors. Then we have

$$\begin{aligned} \tilde{\Delta}_t &= E((\mathbf{A}_{t-1}(\mathbf{X}_{t-1} - \hat{\mathbf{X}}_{t-1}) + \mathbf{B}_{t-1}\boldsymbol{\varepsilon}_t) \cdot \\ &\quad (\mathbf{A}_{t-1}(\mathbf{X}_{t-1} - \hat{\mathbf{X}}_{t-1}) + \mathbf{B}_{t-1}\boldsymbol{\varepsilon}_t)^T) \\ &= E(\mathbf{A}_{t-1}(\mathbf{X}_{t-1} - \hat{\mathbf{X}}_{t-1})(\mathbf{A}_{t-1}(\mathbf{X}_{t-1} - \hat{\mathbf{X}}_{t-1}))^T) \\ &\quad + E((\mathbf{B}_{t-1}\boldsymbol{\varepsilon}_t)(\mathbf{B}_{t-1}\boldsymbol{\varepsilon}_t)^T) \\ &= \mathbf{A}_{t-1}\hat{\Delta}_{t-1}\mathbf{A}_{t-1}^T + \mathbf{B}_{t-1}\mathbf{Q}_t\mathbf{B}_{t-1}^T, \end{aligned}$$

since $\boldsymbol{\varepsilon}_t$ and $\mathbf{X}_{t-1} - \hat{\mathbf{X}}_{t-1}$ are obviously uncorrelated. In complete analogy one shows that

$$E((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) = \mathbf{C}_t\tilde{\Delta}_t\mathbf{C}_t^T + \mathbf{R}_t.$$

Suppose that we have observed $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$, and that we have predicted \mathbf{X}_t by $\tilde{\mathbf{X}}_t = \mathbf{A}_{t-1}\hat{\mathbf{X}}_{t-1}$. Assume that we now also observe \mathbf{Y}_t . How can we use this additional information to improve the prediction $\tilde{\mathbf{X}}_t$ of \mathbf{X}_t ? To this end we add a matrix \mathbf{K}_t such that we obtain the best prediction $\hat{\mathbf{X}}_t$ based on $\mathbf{Y}_1, \dots, \mathbf{Y}_t$:

$$\tilde{\mathbf{X}}_t + \mathbf{K}_t(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t) = \hat{\mathbf{X}}_t \quad (3.7)$$

i.e., we have to choose the matrix \mathbf{K}_t according to Lemma 3.2.1 such that $\mathbf{X}_t - \hat{\mathbf{X}}_t$ and \mathbf{Y}_s are uncorrelated for $s = 1, \dots, t$. In this case, the matrix \mathbf{K}_t is called the *Kalman gain*.

Lemma 3.2.2. *The matrix \mathbf{K}_t in (3.7) is a solution of the equation*

$$\mathbf{K}_t(\mathbf{C}_t\tilde{\Delta}_t\mathbf{C}_t^T + \mathbf{R}_t) = \tilde{\Delta}_t\mathbf{C}_t^T. \quad (3.8)$$

Proof. The matrix \mathbf{K}_t has to be chosen such that $\mathbf{X}_t - \hat{\mathbf{X}}_t$ and \mathbf{Y}_s are uncorrelated for $s = 1, \dots, t$, i.e.,

$$\mathbf{0} = E((\mathbf{X}_t - \hat{\mathbf{X}}_t)\mathbf{Y}_s^T) = E((\mathbf{X}_t - \tilde{\mathbf{X}}_t - \mathbf{K}_t(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t))\mathbf{Y}_s^T), \quad s \leq t.$$

Note that an arbitrary $k \times m$ -matrix \mathbf{K}_t satisfies

$$\begin{aligned} & \mathbb{E} \left((\mathbf{X}_t - \tilde{\mathbf{X}}_t - \mathbf{K}_t(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)) \mathbf{Y}_s^T \right) \\ &= \mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t) \mathbf{Y}_s^T) - \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t) \mathbf{Y}_s^T) = \mathbf{0}, \quad s \leq t-1. \end{aligned}$$

In order to fulfill the above condition, the matrix \mathbf{K}_t needs to satisfy only

$$\begin{aligned} \mathbf{0} &= \mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t) \mathbf{Y}_t^T) - \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t) \mathbf{Y}_t^T) \\ &= \mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) - \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) \\ &= \mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t)(\mathbf{C}_t(\mathbf{X}_t - \tilde{\mathbf{X}}_t) + \boldsymbol{\eta}_t)^T) - \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) \\ &= \mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t)(\mathbf{X}_t - \tilde{\mathbf{X}}_t)^T) \mathbf{C}_t^T - \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) \\ &= \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T - \mathbf{K}_t(\mathbf{C}_t \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T + \mathbf{R}_t). \end{aligned}$$

But this is the assertion of Lemma 3.2.2. Note that $\tilde{\mathbf{Y}}_t$ is a linear combination of $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$ and that $\boldsymbol{\eta}_t$ and $\mathbf{X}_t - \tilde{\mathbf{X}}_t$ are uncorrelated. \square

If the matrix $\mathbf{C}_t \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T + \mathbf{R}_t$ is invertible, then

$$\mathbf{K}_t := \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T (\mathbf{C}_t \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T + \mathbf{R}_t)^{-1}$$

is the uniquely determined Kalman gain. We have, moreover, for a Kalman gain

$$\begin{aligned} \hat{\boldsymbol{\Delta}}_t &= \mathbb{E}((\mathbf{X}_t - \hat{\mathbf{X}}_t)(\mathbf{X}_t - \hat{\mathbf{X}}_t)^T) \\ &= \mathbb{E} \left((\mathbf{X}_t - \tilde{\mathbf{X}}_t - \mathbf{K}_t(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t))(\mathbf{X}_t - \tilde{\mathbf{X}}_t - \mathbf{K}_t(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t))^T \right) \\ &= \tilde{\boldsymbol{\Delta}}_t + \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) \mathbf{K}_t^T \\ &\quad - \mathbb{E}((\mathbf{X}_t - \tilde{\mathbf{X}}_t)(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)^T) \mathbf{K}_t^T - \mathbf{K}_t \mathbb{E}((\mathbf{Y}_t - \tilde{\mathbf{Y}}_t)(\mathbf{X}_t - \tilde{\mathbf{X}}_t)^T) \\ &= \tilde{\boldsymbol{\Delta}}_t + \mathbf{K}_t(\mathbf{C}_t \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T + \mathbf{R}_t) \mathbf{K}_t^T \\ &\quad - \tilde{\boldsymbol{\Delta}}_t \mathbf{C}_t^T \mathbf{K}_t^T - \mathbf{K}_t \mathbf{C}_t \tilde{\boldsymbol{\Delta}}_t \\ &= \tilde{\boldsymbol{\Delta}}_t - \mathbf{K}_t \mathbf{C}_t \tilde{\boldsymbol{\Delta}}_t \end{aligned}$$

by (3.8) and the arguments in the proof of Lemma 3.2.2.

The recursion in the discrete *Kalman filter* is done in two steps: From $\hat{\mathbf{X}}_{t-1}$ and $\hat{\mathbf{\Delta}}_{t-1}$ one computes in the *prediction step* first

$$\begin{aligned}\tilde{\mathbf{X}}_t &= \mathbf{A}_{t-1}\hat{\mathbf{X}}_{t-1}, \\ \tilde{\mathbf{Y}}_t &= \mathbf{C}_t\tilde{\mathbf{X}}_t, \\ \tilde{\mathbf{\Delta}}_t &= \mathbf{A}_{t-1}\hat{\mathbf{\Delta}}_{t-1}\mathbf{A}_{t-1}^T + \mathbf{B}_{t-1}\mathbf{Q}_t\mathbf{B}_{t-1}^T.\end{aligned}\tag{3.9}$$

In the *updating step* one then computes \mathbf{K}_t and the updated values $\hat{\mathbf{X}}_t, \hat{\mathbf{\Delta}}_t$

$$\begin{aligned}\mathbf{K}_t &= \tilde{\mathbf{\Delta}}_t\mathbf{C}_t^T(\mathbf{C}_t\tilde{\mathbf{\Delta}}_t\mathbf{C}_t^T + \mathbf{R}_t)^{-1}, \\ \hat{\mathbf{X}}_t &= \tilde{\mathbf{X}}_t + \mathbf{K}_t(\mathbf{Y}_t - \tilde{\mathbf{Y}}_t), \\ \hat{\mathbf{\Delta}}_t &= \tilde{\mathbf{\Delta}}_t - \mathbf{K}_t\mathbf{C}_t\tilde{\mathbf{\Delta}}_t.\end{aligned}\tag{3.10}$$

An obvious problem is the choice of the initial values $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{\Delta}}_1$. One frequently puts $\tilde{\mathbf{X}}_1 = \mathbf{0}$ and $\tilde{\mathbf{\Delta}}_1$ as the diagonal matrix with constant entries $\sigma^2 > 0$. The number σ^2 reflects the degree of uncertainty about the underlying model. Simulations as well as theoretical results show, however, that the estimates $\hat{\mathbf{X}}_t$ are often not affected by the initial values $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{\Delta}}_1$ if t is large, see for instance Example 3.2.3 below. If in addition we require that the state-space model (3.1), (3.2) is completely determined by some parametrization $\boldsymbol{\vartheta}$ of the distribution of (\mathbf{Y}_t) and (\mathbf{X}_t) , then we can estimate the matrices of the Kalman filter in (3.9) and (3.10) under suitable conditions by a maximum likelihood estimate of $\boldsymbol{\vartheta}$; see e.g. Brockwell and Davis (2002, Section 8.5) or Janacek and Swift (1993, Section 4.5).

By iterating the 1-step prediction $\tilde{\mathbf{X}}_t = \mathbf{A}_{t-1}\hat{\mathbf{X}}_{t-1}$ of \mathbf{X}_t in (3.6) h times, we obtain the h -step prediction of the Kalman filter

$$\tilde{\mathbf{X}}_{t+h} := \mathbf{A}_{t+h-1}\tilde{\mathbf{X}}_{t+h-1}, \quad h \geq 1,$$

with the initial value $\tilde{\mathbf{X}}_{t+0} := \hat{\mathbf{X}}_t$. The pertaining h -step prediction of \mathbf{Y}_{t+h} is then

$$\tilde{\mathbf{Y}}_{t+h} := \mathbf{C}_{t+h}\tilde{\mathbf{X}}_{t+h}, \quad h \geq 1.$$

Example 3.2.3. Let (η_t) be a white noise in \mathbb{R} with $E(\eta_t) = 0$, $E(\eta_t^2) = \sigma^2 > 0$ and put for some $\mu \in \mathbb{R}$

$$Y_t := \mu + \eta_t, \quad t \in \mathbb{Z}.$$

This process can be represented as a state-space model by putting $X_t := \mu$, with state equation $X_{t+1} = X_t$ and observation equation $Y_t = X_t + \eta_t$ i.e., $A_t = 1 = C_t$ and $B_t = 0$. The prediction step (3.9) of the Kalman filter is now given by

$$\tilde{X}_t = \hat{X}_{t-1}, \quad \tilde{Y}_t = \tilde{X}_t, \quad \tilde{\Delta}_t = \Delta_{t-1}.$$

Note that all these values are in \mathbb{R} . The h -step predictions $\tilde{X}_{t+h}, \tilde{Y}_{t+h}$ are, therefore, given by $\tilde{X}_{t+1} = \tilde{X}_t$. The update step (3.10) of the Kalman filter is

$$\begin{aligned} K_t &= \frac{\Delta_{t-1}}{\Delta_{t-1} + \sigma^2} \\ \hat{X}_t &= \tilde{X}_{t-1} + K_t(Y_t - \tilde{X}_{t-1}) \\ \hat{\Delta}_t &= \hat{\Delta}_{t-1} - K_t \hat{\Delta}_{t-1} = \hat{\Delta}_{t-1} \frac{\sigma^2}{\Delta_{t-1} + \sigma^2}. \end{aligned}$$

Note that $\hat{\Delta}_t = E((X_t - \hat{X}_t)^2) \geq 0$ and thus,

$$0 \leq \hat{\Delta}_t = \hat{\Delta}_{t-1} \frac{\sigma^2}{\Delta_{t-1} + \sigma^2} \leq \hat{\Delta}_{t-1}$$

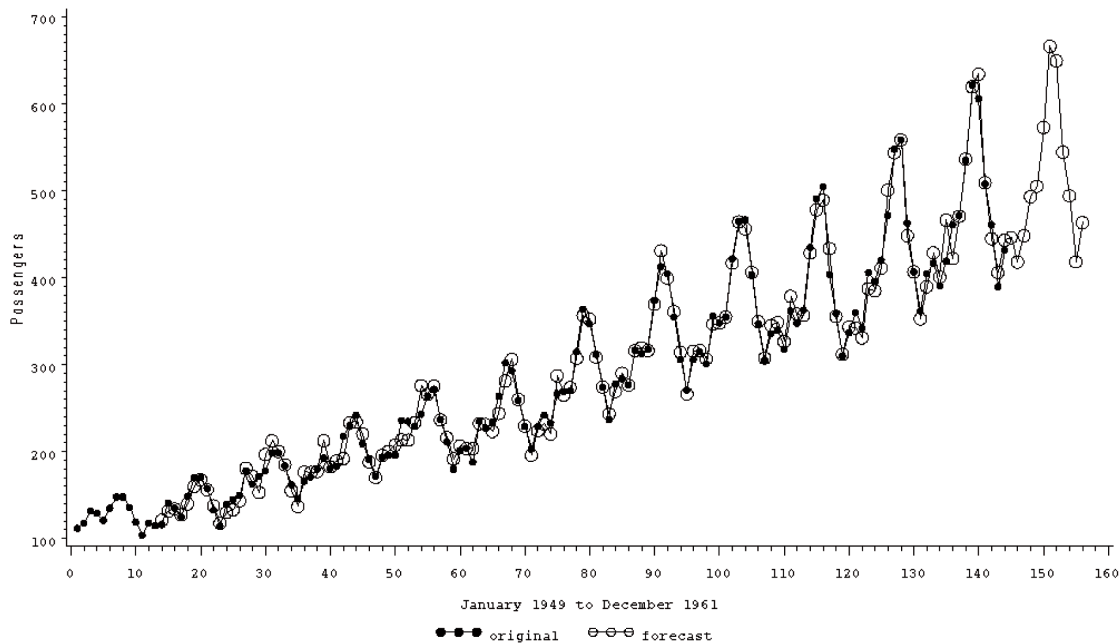
is a decreasing and bounded sequence. Its limit $\Delta := \lim_{t \rightarrow \infty} \hat{\Delta}_t$ consequently exists and satisfies

$$\Delta = \Delta \frac{\sigma^2}{\Delta + \sigma^2}$$

i.e., $\Delta = 0$. This means that the mean squared error $E((X_t - \hat{X}_t)^2) = E((\mu - \hat{X}_t)^2)$ vanishes asymptotically, no matter how the initial values \tilde{X}_1 and $\tilde{\Delta}_1$ are chosen. Further we have $\lim_{t \rightarrow \infty} K_t = 0$, which means that additional observations Y_t do not contribute to \hat{X}_t if t is large. Finally, we obtain for the mean squared error of the h -step prediction \tilde{Y}_{t+h} of Y_{t+h}

$$\begin{aligned} E((Y_{t+h} - \tilde{Y}_{t+h})^2) &= E((\mu + \eta_{t+h} - \hat{X}_t)^2) \\ &= E((\mu - \hat{X}_t)^2) + E(\eta_{t+h}^2) \longrightarrow_{t \rightarrow \infty} \sigma^2. \end{aligned}$$

Example 3.2.4. The following figure displays the Airline Data from Example 1.3.1 together with 12-step forecasts based on the Kalman filter. The original data y_t , $t = 1, \dots, 144$ were log-transformed $x_t = \log(y_t)$ to stabilize the variance; first order differences $\Delta x_t = x_t - x_{t-1}$ were used to eliminate the trend and, finally, $z_t = \Delta x_t - \Delta x_{t-12}$ were computed to remove the seasonal component of 12 months. The Kalman filter was applied to forecast z_t , $t = 145, \dots, 156$, and the results were transformed in the reverse order of the preceding steps to predict the initial values y_t , $t = 145, \dots, 156$.



Plot 3.2.1: Airline Data and predicted values using the Kalman filter.

```

1  /* airline_kalman.sas */
2  TITLE1 'Original and Forecasted Data';
3  TITLE2 'Airline Data';
4
5  /* Read in the data and compute log-transformation */
6  DATA datal;
7      INFILE 'c:\data\airline.txt';
8      INPUT y;
9      yl=LOG(y);
10     t=_N_;
11

```

```

12  /* Compute trend and seasonally adjusted data set */
13  PROC STATESPACE DATA=data1 OUT=data2 LEAD=12;
14      VAR y1(1,12); ID t;
15
16  /* Compute forecasts by inverting the log-transformation */
17  DATA data3;
18      SET data2;
19      yhat=EXP(FOR1);
20
21  /* Merge data sets */
22  DATA data4(KEEP=t y yhat);
23      MERGE data1 data3;
24      BY t;
25
26  /* Graphical options */
27  LEGEND1 LABEL=('') VALUE=('original' 'forecast');
28  SYMBOL1 C=BLACK V=DOT H=0.7 I=JOIN L=1;
29  SYMBOL2 C=BLACK V=CIRCLE H=1.5 I=JOIN L=1;
30  AXIS1 LABEL=(ANGLE=90 'Passengers');
31  AXIS2 LABEL=('January 1949 to December 1961');
32
33  /* Plot data and forecasts */
34  PROC GPLOT DATA=data4;
35      PLOT y*t=1 yhat*t=2 / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2 LEGEND=LEGEND1;
36  RUN; QUIT;

```

In the first data step the Airline Data are read into `data1`. Their logarithm is computed and stored in the variable `y1`. The variable `t` contains the observation number.

The statement `VAR y1(1,12)` of the `PROC STATESPACE` procedure tells SAS to use first order differences of the initial data to remove their trend and to adjust them to a seasonal component of 12 months. The data are iden-

tified by the time index set to `t`. The results are stored in the data set `data2` with forecasts of 12 months after the end of the input data. This is invoked by `LEAD=12`.

`data3` contains the exponentially transformed forecasts, thereby inverting the log-transformation in the first data step.

Finally, the two data sets are merged and displayed in one plot.

Exercises

3.1. Consider the two state-space models

$$\begin{aligned}
 X_{t+1} &= A_t X_t + B_t \varepsilon_{t+1} \\
 Y_t &= C_t X_t + \eta_t
 \end{aligned}$$

and

$$\begin{aligned}\tilde{\mathbf{X}}_{t+1} &= \tilde{\mathbf{A}}_t \tilde{\mathbf{X}}_t + \tilde{\mathbf{B}}_t \tilde{\boldsymbol{\varepsilon}}_{t+1} \\ \tilde{\mathbf{Y}}_t &= \tilde{\mathbf{C}}_t \tilde{\mathbf{X}}_t + \tilde{\boldsymbol{\eta}}_t,\end{aligned}$$

where $(\boldsymbol{\varepsilon}_t^T, \boldsymbol{\eta}_t^T, \tilde{\boldsymbol{\varepsilon}}_t^T, \tilde{\boldsymbol{\eta}}_t^T)^T$ is a white noise. Derive a state-space representation for $(\mathbf{Y}_t^T, \mathbf{Y}_t^T)^T$.

3.2. Find the state-space representation of an ARIMA(p, d, q)-process $(Y_t)_t$. Hint: $Y_t = \Delta^d Y_t - \sum_{j=1}^d (-1)^j \binom{d}{j} Y_{t-j}$ and consider the state vector $\mathbf{Z}_t := (\mathbf{X}_t, \mathbf{Y}_{t-1})^T$, where $\mathbf{X}_t \in \mathbb{R}^{p+q}$ is the state vector of the ARMA(p, q)-process $\Delta^d Y_t$ and $\mathbf{Y}_{t-1} := (Y_{t-d}, \dots, Y_{t-1})^T$.

3.3. Assume that the matrices \mathbf{A} and \mathbf{B} in the state-space model (3.1) are independent of t and that all eigenvalues of \mathbf{A} are in the interior of the unit circle $\{z \in \mathbb{C} : |z| \leq 1\}$. Show that the unique stationary solution of equation (3.1) is given by the infinite series $\mathbf{X}_t = \sum_{j=0}^{\infty} \mathbf{A}^j \mathbf{B} \boldsymbol{\varepsilon}_{t-j+1}$. Hint: The condition on the eigenvalues is equivalent to $\det(I_r - \mathbf{A}z) \neq 0$ for $|z| \leq 1$. Show that there exists some $\varepsilon > 0$ such that $(I_r - \mathbf{A}z)^{-1}$ has the power series representation $\sum_{j=0}^{\infty} \mathbf{A}^j z^j$ in the region $|z| < 1 + \varepsilon$.

3.4. Show that $\boldsymbol{\varepsilon}_t$ and \mathbf{Y}_s are uncorrelated and that $\boldsymbol{\eta}_t$ and \mathbf{Y}_s are uncorrelated if $s < t$.

3.5. Apply PROC STATESPACE to the simulated data of the AR(2)-process in Exercise 2.28.

3.6. (Gas Data) Apply PROC STATESPACE to the gas data. Can they be stationary? Compute the one-step predictors and plot them together with the actual data.

The Frequency Domain Approach of a Time Series

The preceding sections focussed on the analysis of a time series in the *time domain*, mainly by modelling and fitting an $\text{ARMA}(p, q)$ -process to stationary sequences of observations. Another approach towards the modelling and analysis of time series is via the *frequency domain*: A series is often the sum of a whole variety of cyclic components, from which we had already added to our model (1.2) a long term cyclic one or a short term seasonal one. In the following we show that a time series can be completely decomposed into cyclic components. Such cyclic components can be described by their *periods* and *frequencies*. The period is the interval of time required for one cycle to complete. The frequency of a cycle is its number of occurrences during a fixed time unit; in electronic media, for example, frequencies are commonly measured in *hertz*, which is the number of cycles per second, abbreviated by *Hz*. The analysis of a time series in the frequency domain aims at the detection of such cycles and the computation of their frequencies.

Note that in this chapter the results are formulated for any data y_1, \dots, y_n , which need for mathematical reasons not to be generated by a stationary process. Nevertheless it is reasonable to apply the results only to realizations of stationary processes, since the empirical autocovariance function occurring below has no interpretation for non-stationary processes, see Exercise 1.21.

4.1 Least Squares Approach with Known Frequencies

A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is said to be *periodic* with *period* $P > 0$ if $f(t + P) = f(t)$ for any $t \in \mathbb{R}$. A smallest period is called a *fundamental* one. The reciprocal value $\lambda = 1/P$ of a fundamental period is the fundamental frequency. An arbitrary (time) interval of length L consequently shows $L\lambda$ cycles of a periodic function f with fundamental frequency λ . Popular examples of periodic functions are sine and cosine, which both have the fundamental period $P = 2\pi$. Their fundamental frequency, therefore, is $\lambda = 1/(2\pi)$. The predominant family of periodic functions within time series analysis are the harmonic components

$$m(t) := A \cos(2\pi\lambda t) + B \sin(2\pi\lambda t), \quad A, B \in \mathbb{R}, \lambda > 0,$$

which have period $1/\lambda$ and frequency λ . A linear combination of harmonic components

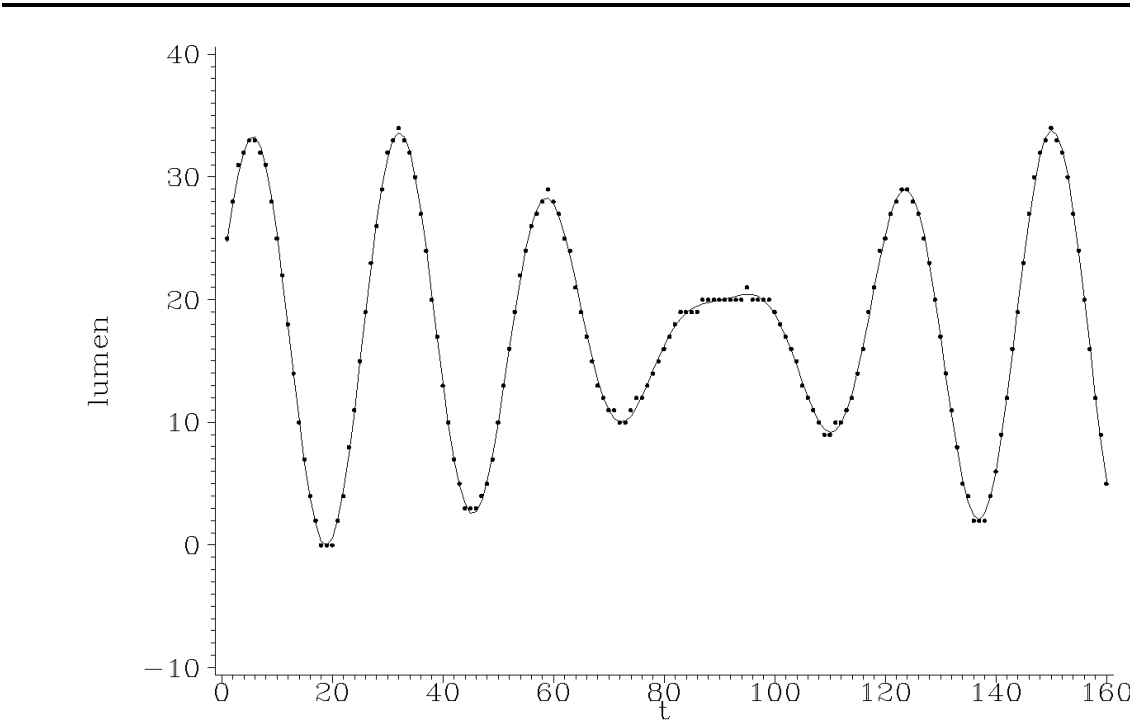
$$g(t) := \mu + \sum_{k=1}^r (A_k \cos(2\pi\lambda_k t) + B_k \sin(2\pi\lambda_k t)), \quad \mu \in \mathbb{R},$$

will be named a *harmonic wave of length r* .

Example 4.1.1. (Star Data). To analyze physical properties of a pulsating star, the intensity of light emitted by this pulsar was recorded at midnight during 600 consecutive nights. The data are taken from Newton (1988). It turns out that a harmonic wave of length two fits the data quite well. The following figure displays the first 160 data y_t and the sum of two harmonic components with period 24 and 29, respectively, plus a constant term $\mu = 17.07$ fitted to these data, i.e.,

$$\begin{aligned} \tilde{y}_t = & 17.07 - 1.86 \cos(2\pi(1/24)t) + 6.82 \sin(2\pi(1/24)t) \\ & + 6.09 \cos(2\pi(1/29)t) + 8.01 \sin(2\pi(1/29)t). \end{aligned}$$

The derivation of these particular frequencies and coefficients will be the content of this section and the following ones. For easier access we begin with the case of known frequencies but unknown constants.



Plot 4.1.1: Intensity of light emitted by a pulsating star and a fitted harmonic wave.

Model: MODEL1
Dependent Variable: LUMEN

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	48400	12100	49297.2	<.0001
Error	595	146.04384	0.24545		
C Total	599	48546			
Root MSE		0.49543	R-square	0.9970	
Dep Mean		17.09667	Adj R-sq	0.9970	
C.V.		2.89782			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Prob > T
----------	----	--------------------	----------------	---------	-----------

Intercept	1	17.06903	0.02023	843.78	<.0001
sin24	1	6.81736	0.02867	237.81	<.0001
cos24	1	-1.85779	0.02865	-64.85	<.0001
sin29	1	8.01416	0.02868	279.47	<.0001
cos29	1	6.08905	0.02865	212.57	<.0001

Listing 4.1.1b: Regression results of fitting a harmonic wave.

```

1  /* star_harmonic.sas */
2  TITLE1 'Harmonic wave';
3  TITLE2 'Star Data';
4
5  /* Read in the data and compute harmonic waves to which data are to be
   ↪ fitted */
6  DATA data1;
7      INFILE 'c:\data\star.txt';
8      INPUT lumen @@;
9      t=_N_;
10     pi=CONSTANT('PI');
11     sin24=SIN(2*pi*t/24);
12     cos24=COS(2*pi*t/24);
13     sin29=SIN(2*pi*t/29);
14     cos29=COS(2*pi*t/29);
15
16  /* Compute a regression */
17  PROC REG DATA=data1;
18      MODEL lumen=sin24 cos24 sin29 cos29;
19      OUTPUT OUT=regdat P=predi;
20
21  /* Graphical options */
22  SYMBOL1 C=GREEN V=DOT I=NONE H=.4;
23  SYMBOL2 C=RED V=NONE I=JOIN;
24  AXIS1 LABEL=(ANGLE=90 'lumen');
25  AXIS2 LABEL=('t');
26
27  /* Plot data and fitted harmonic wave */
28  PROC GPLOT DATA=regdat (OBS=160);
29      PLOT lumen*t=1 predi*t=2 / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2;
30  RUN; QUIT;

```

The number π is generated by the SAS function `CONSTANT` with the argument 'PI'. It is then stored in the variable `pi`. This is used to define the variables `cos24`, `sin24`, `cos29` and `sin29` for the harmonic components. The other variables here are `lumen` read from an external file and `t` generated by `_N_`.

The `PROC REG` statement causes SAS to make a regression from the independent variable `lumen` defined on the left side of the `MODEL` statement on the harmonic components which

are on the right side. A temporary data file named `regdat` is generated by the `OUTPUT` statement. It contains the original variables of the source data step and the values predicted by the regression for `lumen` in the variable `predi`.

The last part of the program creates a plot of the observed `lumen` values and a curve of the predicted values restricted on the first 160 observations.

The output of Program 4.1.1 (star_harmonic.sas) is the standard text output of a regression with an ANOVA table and parameter estimates. For further information on how to read the output, we refer to Falk et al. (2002, Chapter 3).

In a first step we will fit a harmonic component with fixed frequency λ to mean value adjusted data $y_t - \bar{y}$, $t = 1, \dots, n$. To this end, we put with arbitrary $A, B \in \mathbb{R}$

$$m(t) = Am_1(t) + Bm_2(t),$$

where

$$m_1(t) := \cos(2\pi\lambda t), \quad m_2(t) = \sin(2\pi\lambda t).$$

In order to get a proper fit uniformly over all t , it is reasonable to choose the constants A and B as minimizers of the *residual sum of squares*

$$R(A, B) := \sum_{t=1}^n (y_t - \bar{y} - m(t))^2.$$

Taking partial derivatives of the function R with respect to A and B and equating them to zero, we obtain that the minimizing pair A, B has to satisfy the *normal equations*

$$\begin{aligned} Ac_{11} + Bc_{12} &= \sum_{t=1}^n (y_t - \bar{y}) \cos(2\pi\lambda t) \\ Ac_{21} + Bc_{22} &= \sum_{t=1}^n (y_t - \bar{y}) \sin(2\pi\lambda t), \end{aligned}$$

where

$$c_{ij} = \sum_{t=1}^n m_i(t)m_j(t).$$

If $c_{11}c_{22} - c_{12}c_{21} \neq 0$, the uniquely determined pair of solutions A, B of these equations is

$$\begin{aligned} A &= A(\lambda) = n \frac{c_{22}C(\lambda) - c_{12}S(\lambda)}{c_{11}c_{22} - c_{12}c_{21}} \\ B &= B(\lambda) = n \frac{c_{21}C(\lambda) - c_{11}S(\lambda)}{c_{12}c_{21} - c_{11}c_{22}}, \end{aligned}$$

where

$$\begin{aligned} C(\lambda) &:= \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y}) \cos(2\pi\lambda t), \\ S(\lambda) &:= \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y}) \sin(2\pi\lambda t) \end{aligned} \quad (4.1)$$

are the *empirical (cross-)covariances* of $(y_t)_{1 \leq t \leq n}$ and $(\cos(2\pi\lambda t))_{1 \leq t \leq n}$ and of $(y_t)_{1 \leq t \leq n}$ and $(\sin(2\pi\lambda t))_{1 \leq t \leq n}$, respectively. As we will see, these cross-covariances $C(\lambda)$ and $S(\lambda)$ are fundamental to the analysis of a time series in the frequency domain.

The solutions A and B become particularly simple in the case of *Fourier frequencies* $\lambda = k/n$, $k = 0, 1, 2, \dots, [n/2]$, where $[x]$ denotes the greatest integer less than or equal to $x \in \mathbb{R}$. If $k \neq 0$ and $k \neq n/2$ in the case of an even sample size n , then we obtain from Lemma 4.1.2 below that $c_{12} = c_{21} = 0$ and $c_{11} = c_{22} = n/2$ and thus

$$A = 2C(\lambda), \quad B = 2S(\lambda).$$

Harmonic Waves with Fourier Frequencies

Next we will fit harmonic waves to data y_1, \dots, y_n , where we restrict ourselves to Fourier frequencies, which facilitates the computations. The following lemma will be crucial.

Lemma 4.1.2. *For arbitrary $0 \leq k, m \leq [n/2]$; $k, m \in \mathbb{N}$ we have*

$$\begin{aligned} \sum_{t=1}^n \cos\left(2\pi\frac{k}{n}t\right) \cos\left(2\pi\frac{m}{n}t\right) &= \begin{cases} n, & k = m = 0 \text{ or } = n/2 \\ n/2, & k = m \neq 0 \text{ and } \neq n/2 \\ 0, & k \neq m \end{cases} \\ \sum_{t=1}^n \sin\left(2\pi\frac{k}{n}t\right) \sin\left(2\pi\frac{m}{n}t\right) &= \begin{cases} 0, & k = m = 0 \text{ or } n/2 \\ n/2, & k = m \neq 0 \text{ and } \neq n/2 \\ 0, & k \neq m \end{cases} \\ \sum_{t=1}^n \cos\left(2\pi\frac{k}{n}t\right) \sin\left(2\pi\frac{m}{n}t\right) &= 0. \end{aligned}$$

Proof. Exercise 4.3. □

The above lemma implies that the $2[n/2] + 1$ vectors in \mathbb{R}^n

$$(\sin(2\pi(k/n)t))_{1 \leq t \leq n}, \quad k = 1, \dots, [n/2],$$

and

$$(\cos(2\pi(k/n)t))_{1 \leq t \leq n}, \quad k = 0, \dots, [n/2],$$

span the space \mathbb{R}^n . Note that by Lemma 4.1.2 in the case of n odd the above $2[n/2] + 1 = n$ vectors are linearly independent, precisely, they are orthogonal, whereas in the case of an even sample size n the vector $(\sin(2\pi(k/n)t))_{1 \leq t \leq n}$ with $k = n/2$ is the null vector $(0, \dots, 0)$ and the remaining n vectors are again orthogonal. As a consequence we obtain that for a given set of data y_1, \dots, y_n , there exist in any case uniquely determined coefficients A_k and B_k , $k = 0, \dots, [n/2]$, with $B_0 := 0$ and, if n is even, $B_{n/2} = 0$ as well, such that

$$y_t = \sum_{k=0}^{[n/2]} \left(A_k \cos\left(2\pi \frac{k}{n} t\right) + B_k \sin\left(2\pi \frac{k}{n} t\right) \right), \quad t = 1, \dots, n. \quad (4.2)$$

Next we determine these coefficients A_k, B_k . Let to this end $\mathbf{v}_1, \dots, \mathbf{v}_n$ be arbitrary orthogonal vectors in \mathbb{R}^n , precisely, $\mathbf{v}_i^T \mathbf{v}_j = 0$ if $i \neq j$ and $= \|\mathbf{v}_i\|^2 > 0$ if $i = j$. Then $\mathbf{v}_1, \dots, \mathbf{v}_n$ span \mathbb{R}^n and, thus, for any $\mathbf{y} \in \mathbb{R}^n$ there exists real numbers c_1, \dots, c_n such that

$$\mathbf{y} = \sum_{k=1}^n c_k \mathbf{v}_k.$$

As a consequence we obtain for $i = 1, \dots, n$

$$\mathbf{y}^T \mathbf{v}_i = \sum_{k=1}^n c_k \mathbf{v}_k^T \mathbf{v}_i = c_i \|\mathbf{v}_i\|^2$$

and, thus,

$$c_k = \frac{\mathbf{y}^T \mathbf{v}_i}{\|\mathbf{v}_i\|^2}, \quad k = 1, \dots, n.$$

Applying this to equation (4.2) we obtain from Lemma 4.1.2

$$\begin{aligned} A_k &= \begin{cases} \frac{2}{n} \sum_{t=1}^n y_t \cos \left(2\pi \frac{k}{n} t \right), & k = 1, \dots, [(n-1)/2] \\ \frac{1}{n} \sum_{t=1}^n y_t \cos \left(2\pi \frac{k}{n} t \right), & k = 0 \text{ and } k = n/2, \text{ if } n \text{ is even} \end{cases} \\ B_k &= \frac{2}{n} \sum_{t=1}^n y_t \sin \left(2\pi \frac{k}{n} t \right), \quad k = 1, \dots, [(n-1)/2]. \end{aligned} \quad (4.3)$$

A popular equivalent formulation of (4.2) is

$$y_t = \frac{\tilde{A}_0}{2} + \sum_{k=1}^{[n/2]} \left(A_k \cos \left(2\pi \frac{k}{n} t \right) + B_k \sin \left(2\pi \frac{k}{n} t \right) \right), \quad t = 1, \dots, n, \quad (4.4)$$

with A_k, B_k as in (4.3) for $k = 1, \dots, [n/2]$, $B_{n/2} = 0$ for an even n , and

$$\tilde{A}_0 = 2A_0 = \frac{2}{n} \sum_{t=1}^n y_t = 2\bar{y}.$$

Up to the factor 2, the coefficients A_k, B_k coincide with the empirical covariances $C(k/n)$ and $S(k/n)$, $k = 1, \dots, [(n-1)/2]$, defined in (4.1). This follows from the equations (Exercise 4.2)

$$\sum_{t=1}^n \cos \left(2\pi \frac{k}{n} t \right) = \sum_{t=1}^n \sin \left(2\pi \frac{k}{n} t \right) = 0, \quad k = 1, \dots, [n/2]. \quad (4.5)$$

4.2 The Periodogram

In the preceding section we exactly fitted a harmonic wave with Fourier frequencies $\lambda_k = k/n$, $k = 0, \dots, [n/2]$, to a given series y_1, \dots, y_n . Example 4.1.1 shows that a harmonic wave including only two frequencies already fits the Star Data quite well. There is a general tendency that a time series is governed by different frequencies $\lambda_1, \dots, \lambda_r$ with $r < [n/2]$, which on their part can be approximated by Fourier frequencies $k_1/n, \dots, k_r/n$ if n is sufficiently large. The question *which* frequencies actually govern a time series leads to the *intensity* of a frequency λ . This number reflects the influence of the harmonic component with frequency λ on the series. The intensity of the Fourier

frequency $\lambda = k/n$, $1 \leq k \leq [(n-1)/2]$, is defined via its residual sum of squares. We have by Lemma 4.1.2, (4.5) and (4.4)

$$\begin{aligned} \sum_{t=1}^n \left(y_t - \bar{y} - A_k \cos \left(2\pi \frac{k}{n} t \right) - B_k \sin \left(2\pi \frac{k}{n} t \right) \right)^2 \\ = \sum_{t=1}^n (y_t - \bar{y})^2 - \frac{n}{2} (A_k^2 + B_k^2), \quad k = 1, \dots, [(n-1)/2], \end{aligned}$$

and

$$\sum_{t=1}^n (y_t - \bar{y})^2 = \frac{n}{2} \sum_{k=1}^{[n/2]} (A_k^2 + B_k^2).$$

The number $(n/2)(A_k^2 + B_k^2) = 2n(C^2(k/n) + S^2(k/n))$ is therefore the contribution of the harmonic component with Fourier frequency k/n , $k = 1, \dots, [(n-1)/2]$, to the total variation $\sum_{t=1}^n (y_t - \bar{y})^2$. It is called the *intensity of the frequency k/n* . Further insight is gained from the Fourier analysis in Theorem 4.2.4. For general frequencies $\lambda \in \mathbb{R}$ we define its intensity now by

$$\begin{aligned} I(\lambda) &= n(C(\lambda)^2 + S(\lambda)^2) \\ &= \frac{1}{n} \left(\left(\sum_{t=1}^n (y_t - \bar{y}) \cos(2\pi \lambda t) \right)^2 + \left(\sum_{t=1}^n (y_t - \bar{y}) \sin(2\pi \lambda t) \right)^2 \right). \end{aligned} \tag{4.6}$$

This function is called the *periodogram*. The following Theorem implies in particular that it is sufficient to define the periodogram on the interval $[0, 1]$. For Fourier frequencies we obtain from (4.3) and (4.5)

$$I(k/n) = \frac{n}{4} (A_k^2 + B_k^2), \quad k = 1, \dots, [(n-1)/2].$$

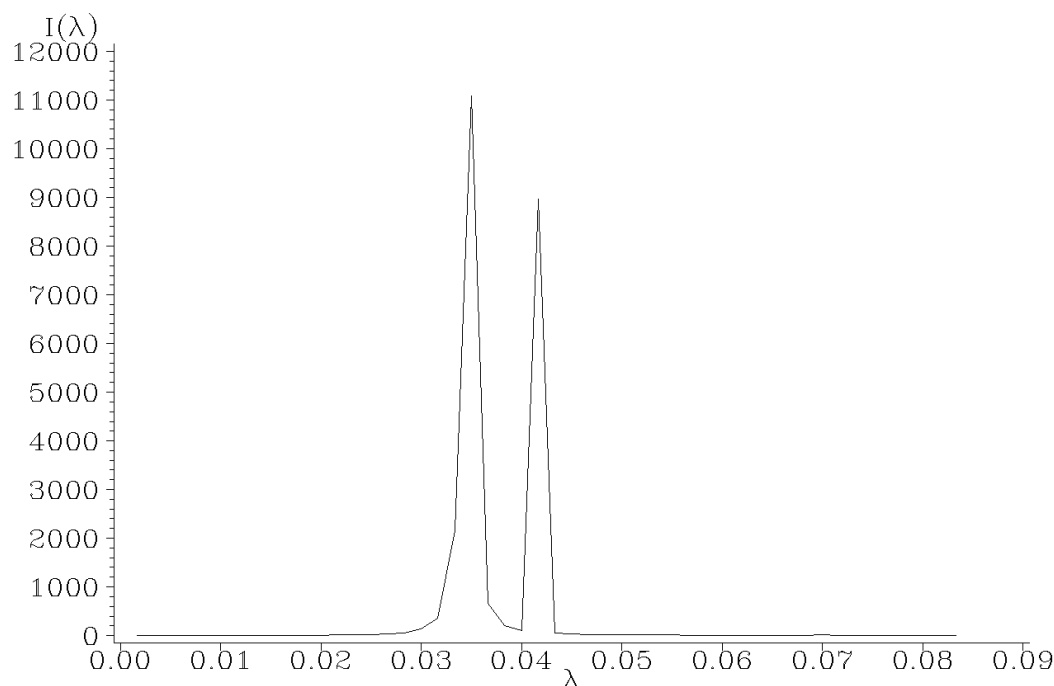
Theorem 4.2.1. *We have*

- (i) $I(0) = 0$,
- (ii) I is an even function, i.e., $I(\lambda) = I(-\lambda)$ for any $\lambda \in \mathbb{R}$,
- (iii) I has the period 1.

Proof. Part (i) follows from $\sin(0) = 0$ and $\cos(0) = 1$, while (ii) is a consequence of $\cos(-x) = \cos(x)$, $\sin(-x) = -\sin(x)$, $x \in \mathbb{R}$. Part (iii) follows from the fact that 2π is the fundamental period of \sin and \cos . \square

Theorem 4.2.1 implies that the function $I(\lambda)$ is completely determined by its values on $[0, 0.5]$. The periodogram is often defined on the scale $[0, 2\pi]$ instead of $[0, 1]$ by putting $I^*(\omega) := 2I(\omega/(2\pi))$; this version is, for example, used in SAS. In view of Theorem 4.2.4 below we prefer $I(\lambda)$, however.

The following figure displays the periodogram of the Star Data from Example 4.1.1. It has two obvious peaks at the Fourier frequencies $21/600 = 0.035 \approx 1/28.57$ and $25/600 = 1/24 \approx 0.04167$. This indicates that essentially two cycles with period 24 and 28 or 29 are inherent in the data. A least squares approach for the determination of the coefficients A_i, B_i , $i = 1, 2$ with frequencies $1/24$ and $1/29$ as done in Program 4.1.1 (star_harmonic.sas) then leads to the coefficients in Example 4.1.1.



Plot 4.2.1: Periodogram of the Star Data.

COS_01				
34.1933				

PERIOD	COS_01	SIN_01	P	LAMBDA
28.5714	-0.91071	8.54977	11089.19	0.035000
24.0000	-0.06291	7.73396	8972.71	0.041667
30.0000	0.42338	-3.76062	2148.22	0.033333
27.2727	-0.16333	2.09324	661.25	0.036667
31.5789	0.20493	-1.52404	354.71	0.031667
26.0870	-0.05822	1.18946	212.73	0.038333

Listing 4.2.1b: The constant $\tilde{A}_0 = 2A_0 = 2\bar{y}$ and the six Fourier frequencies $\lambda = k/n$ with largest $I(k/n)$ -values, their inverses and the Fourier coefficients pertaining to the Star Data.

```

1  /* star_periodogram.sas */
2  TITLE1 'Periodogram';
3  TITLE2 'Star Data';
4
5  /* Read in the data */
6  DATA datal;
7      INFILE 'c:\data\star.txt';
8      INPUT lumen @@;
9
10 /* Compute the periodogram */
11 PROC SPECTRA DATA=datall COEF P OUT=data2;
12     VAR lumen;
13
14 /* Adjusting different periodogram definitions */
15 DATA data3;
16     SET data2 (FIRSTOBS=2);
17     p=P_01/2;
18     lambda=FREQ/(2*CONSTANT('PI'));
19     DROP P_01 FREQ;
20
21 /* Graphical options */
22 SYMBOL1 V=NONE C=GREEN I=JOIN;
23 AXIS1 LABEL=(' I(' F=CGREEK 'l')');
24 AXIS2 LABEL=(F=CGREEK 'l');
25
26 /* Plot the periodogram */
27 PROC GPLOT DATA=data3 (OBS=50);
28     PLOT p*lambda=1 / VAXIS=AXIS1 HAXIS=AXIS2;
29
30 /* Sort by periodogram values */
31 PROC SORT DATA=data3 OUT=data4;
32     BY DESCENDING p;
33
34 /* Print largest periodogram values */

```

```

35 PROC PRINT DATA=data2 (OBS=1) NOOBS;
36     VAR COS_01;
37 PROC PRINT DATA=data4 (OBS=6) NOOBS;
38 RUN; QUIT;

```

The first step is to read the star data from an external file into a data set. Using the SAS procedure `SPECTRA` with the options `P` (periodogram), `COEF` (Fourier coefficients), `OUT=data2` and the `VAR` statement specifying the variable of interest, an output data set is generated. It contains periodogram data `P_01` evaluated at the Fourier frequencies, a `FREQ` variable for this frequencies, the pertaining period in the variable `PERIOD` and the variables `COS_01` and `SIN_01` with the coefficients for the harmonic waves. Because SAS uses different definitions for the frequencies and the periodogram, here in `data3` new variables `lambda` (dividing `FREQ` by 2π to eliminate an additional factor 2π) and `p` (dividing `P_01` by 2) are created

and the no more needed ones are dropped. By means of the data set option `FIRSTOBS=2` the first observation of `data2` is excluded from the resulting data set.

The following `PROC Gplot` just takes the first 50 observations of `data3` into account. This means a restriction of `lambda` up to $50/600 = 1/12$, the part with the greatest peaks in the periodogram.

The procedure `sort` generates a new data set `data4` containing the same observations as the input data set `data3`, but they are sorted in descending order of the periodogram values. The two `PROC PRINT` statements at the end make SAS print to datasets `data2` and `data4`.

The first part of the output is the coefficient \tilde{A}_0 which is equal to two times the mean of the lumen data. The results for the Fourier frequencies with the six greatest periodogram values constitute the second part of the output. Note that the meaning of `COS_01` and `SIN_01` are slightly different from the definitions of A_k and B_k in (4.3), because SAS lets the index run from 0 to $n - 1$ instead of 1 to n .

The Fourier Transform

From Euler's equation $e^{iz} = \cos(z) + i \sin(z)$, $z \in \mathbb{R}$, we obtain for $\lambda \in \mathbb{R}$

$$D(\lambda) := C(\lambda) - iS(\lambda) = \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y}) e^{-i2\pi\lambda t}.$$

The periodogram is a function of $D(\lambda)$, since $I(\lambda) = n|D(\lambda)|^2$. Unlike the periodogram, the number $D(\lambda)$ contains the complete information about $C(\lambda)$ and $S(\lambda)$, since both values can be recovered from the complex number $D(\lambda)$, being its real and negative imaginary part. In the following we view the data y_1, \dots, y_n again as a clipping from an

infinite series y_t , $t \in \mathbb{Z}$. Let $a := (a_t)_{t \in \mathbb{Z}}$ be an absolutely summable sequence of real numbers. For such a sequence a the complex valued function

$$f_a(\lambda) = \sum_{t \in \mathbb{Z}} a_t e^{-i2\pi\lambda t}, \quad \lambda \in \mathbb{R},$$

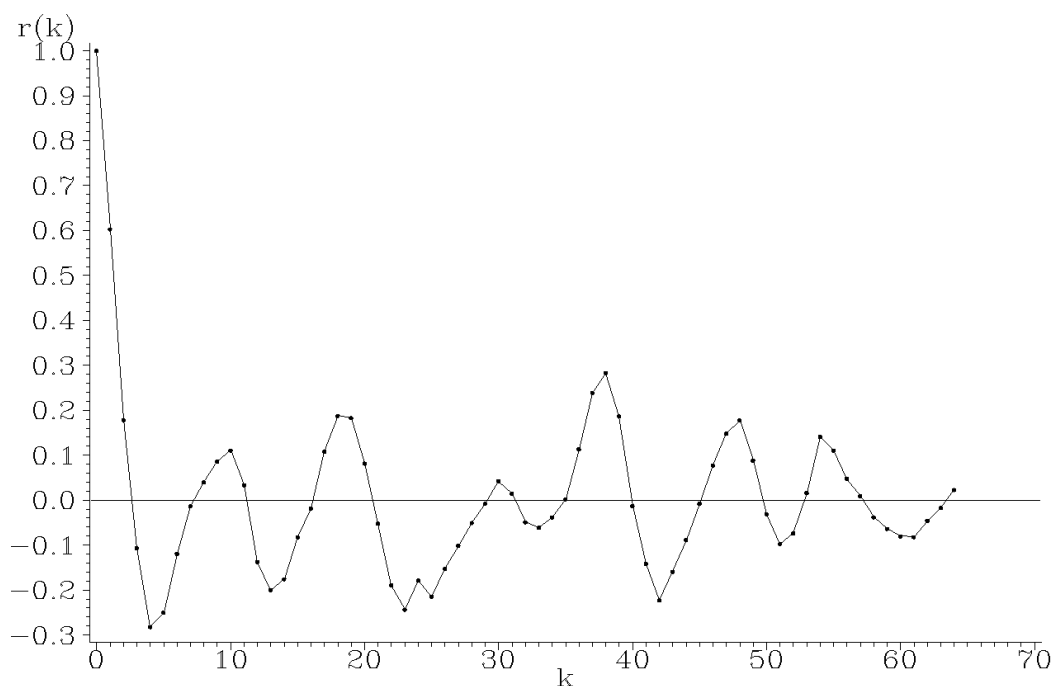
is said to be its *Fourier transform*. It links the empirical autocovariance function to the periodogram as it will turn out in Theorem 4.2.3 that the latter is the Fourier transform of the first. Note that $\sum_{t \in \mathbb{Z}} |a_t e^{-i2\pi\lambda t}| = \sum_{t \in \mathbb{Z}} |a_t| < \infty$, since $|e^{ix}| = 1$ for any $x \in \mathbb{R}$. The Fourier transform of $a_t = (y_t - \bar{y})/n$, $t = 1, \dots, n$, and $a_t = 0$ elsewhere is then given by $D(\lambda)$. The following elementary properties of the Fourier transform are immediate consequences of the arguments in the proof of Theorem 4.2.1. In particular we obtain that the Fourier transform is already determined by its values on $[0, 0.5]$.

Theorem 4.2.2. *We have*

- (i) $f_a(0) = \sum_{t \in \mathbb{Z}} a_t$,
- (ii) $f_a(-\lambda)$ and $f_a(\lambda)$ are conjugate complex numbers i.e., $f_a(-\lambda) = \overline{f_a(\lambda)}$,
- (iii) f_a has the period 1.

Autocorrelation Function and Periodogram

Information about cycles that are inherent in given data, can also be deduced from the empirical autocorrelation function. The following figure displays the autocorrelation function of the Bankruptcy Data, introduced in Exercise 1.20.



Plot 4.2.2: Autocorrelation function of the Bankruptcy Data.

```

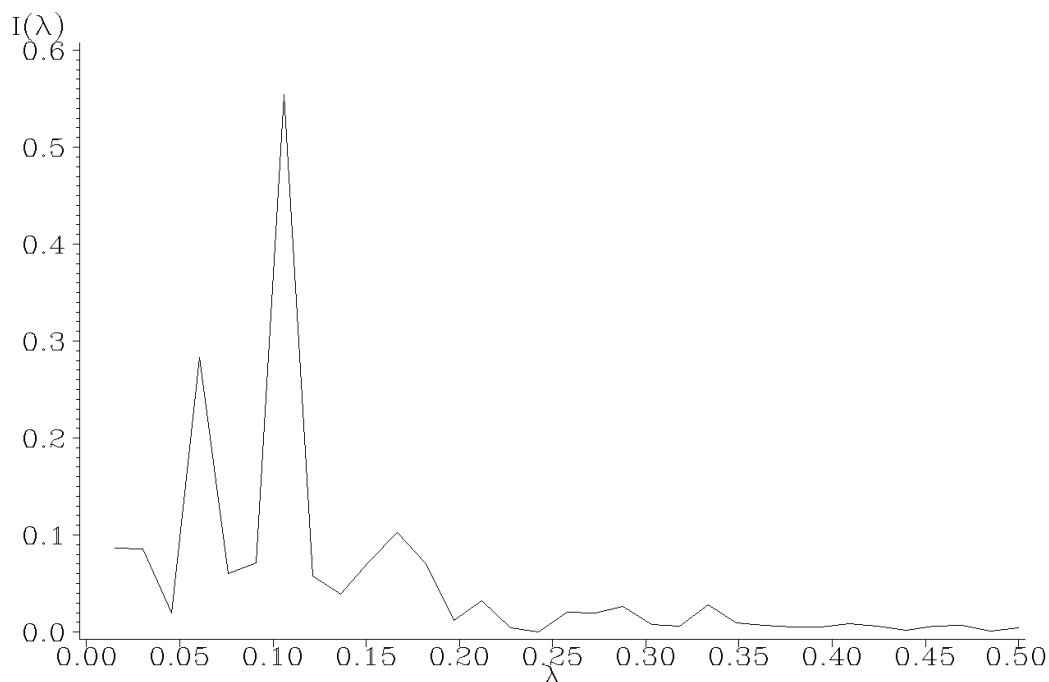
1  /* bankruptcy_correlogram.sas */
2  TITLE1 'Correlogram';
3  TITLE2 'Bankruptcy Data';
4
5  /* Read in the data */
6  DATA data1;
7      INFILE 'c:\data\bankrupt.txt';
8      INPUT year bankrupt;
9
10 /* Compute autocorrelation function */
11 PROC ARIMA DATA=data1;
12     IDENTIFY VAR=bankrupt NLAG=64 OUTCOV=corr NOPRINT;
13
14 /* Graphical options */
15 AXIS1 LABEL=('r(k)');
16 AXIS2 LABEL=('k');
17 SYMBOL1 V=DOT C=GREEN I=JOIN H=0.4 W=1;
18
19 /* Plot auto correlation function */
20 PROC GPLOT DATA=corr;
21     PLOT CORR*LAG / VAXIS=AXIS1 HAXIS=AXIS2 VREF=0;
22 RUN; QUIT;

```

After reading the data from an external file into a data step, the procedure ARIMA calculates the empirical autocorrelation function and

stores them into a new data set. The correlogram is generated using PROC GPLOT.

The next figure displays the periodogram of the Bankruptcy Data.



Plot 4.2.3: Periodogram of the Bankruptcy Data.

```

1  /* bankruptcy_periodogram.sas */
2  TITLE1 'Periodogram';
3  TITLE2 'Bankruptcy Data';
4
5  /* Read in the data */
6  DATA data1;
7      INFILE 'c:\data\bankrupt.txt';
8      INPUT year bankrupt;
9
10 /* Compute the periodogram */
11 PROC SPECTRA DATA=data1 P OUT=data2;
12     VAR bankrupt;
13
14 /* Adjusting different periodogram definitions */
15 DATA data3;
16     SET data2 (FIRSTOBS=2);
17     p=P_01/2;

```

```

18   lambda=FREQ/(2*CONSTANT('PI'));
19
20   /* Graphical options */
21   SYMBOL1 V=NONE C=GREEN I=JOIN;
22   AXIS1 LABEL=('I' F=CGREEK '(1)') ;
23   AXIS2 ORDER=(0 TO 0.5 BY 0.05) LABEL=(F=CGREEK '1');
24
25   /* Plot the periodogram */
26   PROC GPLOT DATA=data3;
27       PLOT p*lambda / VAXIS=AXIS1 HAXIS=AXIS2;
28   RUN; QUIT;

```

This program again first reads the data formations of the periodogram and the frequency values generated by PROC SPECTRA and then starts a spectral analysis by PROC SPECTRA. Due to the reasons mentioned in the comments to Program 4.2.1 (star.periodogram.sas) there are some trans-

The autocorrelation function of the Bankruptcy Data has extreme values at about multiples of 9 years, which indicates a period of length 9. This is underlined by the periodogram in Plot 4.2.3, which has a peak at $\lambda = 0.11$, corresponding to a period of $1/0.11 \sim 9$ years as well. As mentioned above, there is actually a close relationship between the *empirical* autocovariance function and the periodogram. The corresponding result for the *theoretical* autocovariances is given in Chapter 5.

Theorem 4.2.3. Denote by c the empirical autocovariance function of y_1, \dots, y_n , i.e., $c(k) = \frac{1}{n} \sum_{j=1}^{n-k} (y_j - \bar{y})(y_{j+k} - \bar{y})$, $k = 0, \dots, n-1$, where $\bar{y} := n^{-1} \sum_{j=1}^n y_j$. Then we have with $c(-k) := c(k)$

$$\begin{aligned}
 I(\lambda) &= c(0) + 2 \sum_{k=1}^{n-1} c(k) \cos(2\pi\lambda k) \\
 &= \sum_{k=-(n-1)}^{n-1} c(k) e^{-i2\pi\lambda k}.
 \end{aligned}$$

Proof. From the equation $\cos(x_1) \cos(x_2) + \sin(x_1) \sin(x_2) = \cos(x_1 -$

x_2) for $x_1, x_2 \in \mathbb{R}$ we obtain

$$\begin{aligned} I(\lambda) &= \frac{1}{n} \sum_{s=1}^n \sum_{t=1}^n (y_s - \bar{y})(y_t - \bar{y}) \\ &\quad \times (\cos(2\pi\lambda s) \cos(2\pi\lambda t) + \sin(2\pi\lambda s) \sin(2\pi\lambda t)) \\ &= \frac{1}{n} \sum_{s=1}^n \sum_{t=1}^n a_{st}, \end{aligned}$$

where $a_{st} := (y_s - \bar{y})(y_t - \bar{y}) \cos(2\pi\lambda(s - t))$. Since $a_{st} = a_{ts}$ and $\cos(0) = 1$ we have moreover

$$\begin{aligned} I(\lambda) &= \frac{1}{n} \sum_{t=1}^n a_{tt} + \frac{2}{n} \sum_{k=1}^{n-1} \sum_{j=1}^{n-k} a_{jj+k} \\ &= \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y})^2 + 2 \sum_{k=1}^{n-1} \left(\frac{1}{n} \sum_{j=1}^{n-k} (y_j - \bar{y})(y_{j+k} - \bar{y}) \right) \cos(2\pi\lambda k) \\ &= c(0) + 2 \sum_{k=1}^{n-1} c(k) \cos(2\pi\lambda k). \end{aligned}$$

The complex representation of the periodogram is then obvious:

$$\begin{aligned} \sum_{k=-(n-1)}^{n-1} c(k) e^{-i2\pi\lambda k} &= c(0) + \sum_{k=1}^{n-1} c(k) (e^{i2\pi\lambda k} + e^{-i2\pi\lambda k}) \\ &= c(0) + \sum_{k=1}^{n-1} c(k) 2 \cos(2\pi\lambda k) = I(\lambda). \end{aligned}$$

□

Inverse Fourier Transform

The empirical autocovariance function can be recovered from the periodogram, which is the content of our next result. Since the periodogram is the Fourier transform of the empirical autocovariance function, this result is a special case of the inverse Fourier transform in Theorem 4.2.5 below.

Theorem 4.2.4. *The periodogram*

$$I(\lambda) = \sum_{k=-(n-1)}^{n-1} c(k) e^{-i2\pi\lambda k}, \quad \lambda \in \mathbb{R},$$

satisfies the inverse formula

$$c(k) = \int_0^1 I(\lambda) e^{i2\pi\lambda k} d\lambda, \quad |k| \leq n-1.$$

In particular for $k = 0$ we obtain

$$c(0) = \int_0^1 I(\lambda) d\lambda.$$

The sample variance $c(0) = \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y})^2$ equals, therefore, the area under the curve $I(\lambda)$, $0 \leq \lambda \leq 1$. The integral $\int_{\lambda_1}^{\lambda_2} I(\lambda) d\lambda$ can be interpreted as that portion of the total variance $c(0)$, which is contributed by the harmonic waves with frequencies $\lambda \in [\lambda_1, \lambda_2]$, where $0 \leq \lambda_1 < \lambda_2 \leq 1$. The periodogram consequently shows the distribution of the total variance among the frequencies $\lambda \in [0, 1]$. A peak of the periodogram at a frequency λ_0 implies, therefore, that a large part of the total variation $c(0)$ of the data can be explained by the harmonic wave with that frequency λ_0 .

The following result is the inverse formula for general Fourier transforms.

Theorem 4.2.5. *For an absolutely summable sequence $a := (a_t)_{t \in \mathbb{Z}}$ with Fourier transform $f_a(\lambda) = \sum_{t \in \mathbb{Z}} a_t e^{-i2\pi\lambda t}$, $\lambda \in \mathbb{R}$, we have*

$$a_t = \int_0^1 f_a(\lambda) e^{i2\pi\lambda t} d\lambda, \quad t \in \mathbb{Z}.$$

Proof. The dominated convergence theorem implies

$$\begin{aligned} \int_0^1 f_a(\lambda) e^{i2\pi\lambda t} d\lambda &= \int_0^1 \left(\sum_{s \in \mathbb{Z}} a_s e^{-i2\pi\lambda s} \right) e^{i2\pi\lambda t} d\lambda \\ &= \sum_{s \in \mathbb{Z}} a_s \int_0^1 e^{i2\pi\lambda(t-s)} d\lambda = a_t, \end{aligned}$$

since

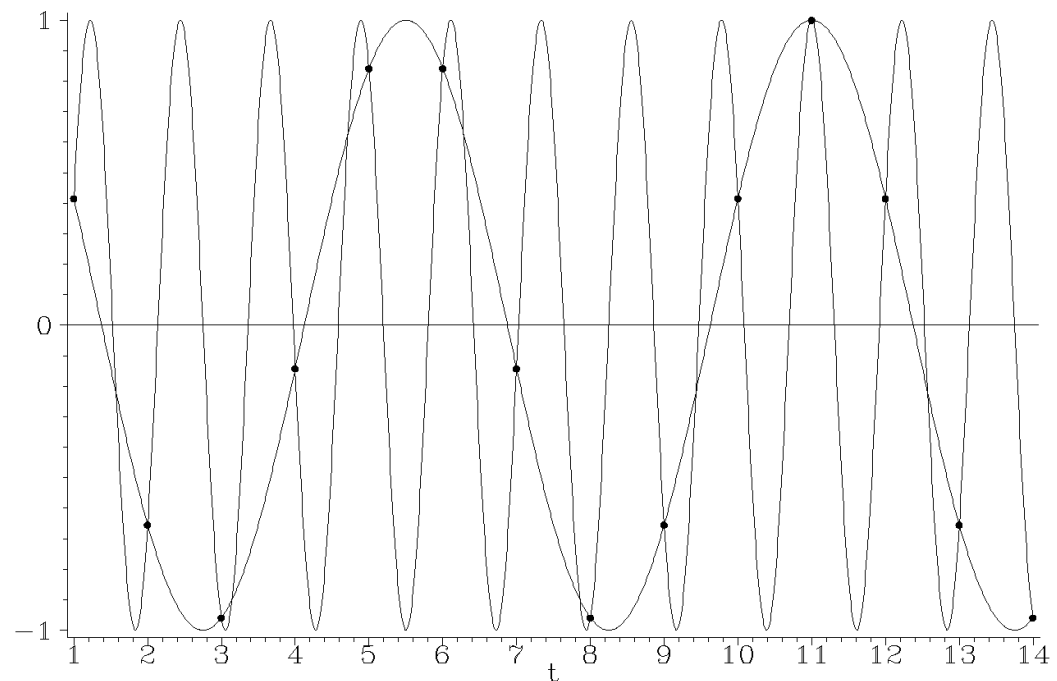
$$\int_0^1 e^{i2\pi\lambda(t-s)} d\lambda = \begin{cases} 1, & \text{if } s = t \\ 0, & \text{if } s \neq t. \end{cases}$$

□

The inverse Fourier transformation shows that the complete sequence $(a_t)_{t \in \mathbb{Z}}$ can be recovered from its Fourier transform. This implies in particular that the Fourier transforms of absolutely summable sequences are uniquely determined. The analysis of a time series in the frequency domain is, therefore, equivalent to its analysis in the time domain, which is based on an evaluation of its autocovariance function.

Aliasing

Suppose that we observe a *continuous time* process $(Z_t)_{t \in \mathbb{R}}$ only through its values at $k\Delta$, $k \in \mathbb{Z}$, where $\Delta > 0$ is the sampling interval, i.e., we actually observe $(Y_k)_{k \in \mathbb{Z}} = (Z_{k\Delta})_{k \in \mathbb{Z}}$. Take, for example, $Z_t := \cos(2\pi(9/11)t)$, $t \in \mathbb{R}$. The following figure shows that at $k \in \mathbb{Z}$, i.e., $\Delta = 1$, the observations Z_k coincide with X_k , where $X_t := \cos(2\pi(2/11)t)$, $t \in \mathbb{R}$. With the sampling interval $\Delta = 1$, the observations Z_k with high frequency 9/11 can, therefore, not be distinguished from the X_k , which have low frequency 2/11.



Plot 4.2.4: Aliasing of $\cos(2\pi(9/11)k)$ and $\cos(2\pi(2/11)k)$.

```

1  /* aliasing.sas */
2  TITLE1 'Aliasing';
3
4  /* Generate harmonic waves */
5  DATA data1;
6      DO t=1 TO 14 BY .01;
7          y1=COS(2*CONSTANT('PI')*2/11*t);
8          y2=COS(2*CONSTANT('PI')*9/11*t);
9          OUTPUT;
10     END;
11
12  /* Generate points of intersection */
13  DATA data2;
14      DO t0=1 TO 14;
15          y0=COS(2*CONSTANT('PI')*2/11*t0);
16          OUTPUT;
17     END;
18
19  /* Merge the data sets */
20  DATA data3;
21      MERGE data1 data2;
22
23  /* Graphical options */
24  SYMBOL1 V=DOT C=GREEN I=NONE H=.8;
25  SYMBOL2 V=NONE C=RED I=JOIN;
26  AXIS1 LABEL=NONE;
27  AXIS2 LABEL=('t');

```

```

28 /* Plot the curves with point of intersection */
29 PROC GPLOT DATA=data3;
30     PLOT y0*t0=1 y1*t=2 y2*t=2 / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2 VREF=0;
31 RUN; QUIT;

```

In the first data step a tight grid for the cosine waves with frequencies 2/11 and 9/11 is generated. In the second data step the values of the cosine wave with frequency 2/11 are generated just for integer values of t symbolizing the observation points.

After merging the two data sets the two waves are plotted using the `JOIN` option in the `SYMBOL` statement while the values at the observation points are displayed in the same graph by dot symbols.

This phenomenon that a high frequency component takes on the values of a lower one, is called *aliasing*. It is caused by the choice of the sampling interval Δ , which is 1 in Plot 4.2.4. If a series is not just a constant, then the shortest observable period is 2Δ . The highest observable frequency λ^* with period $1/\lambda^*$, therefore, satisfies $1/\lambda^* \geq 2\Delta$, i.e., $\lambda^* \leq 1/(2\Delta)$. This frequency $1/(2\Delta)$ is known as the *Nyquist frequency*. The sampling interval Δ should, therefore, be chosen small enough, so that $1/(2\Delta)$ is above the frequencies under study. If, for example, a process is a harmonic wave with frequencies $\lambda_1, \dots, \lambda_p$, then Δ should be chosen such that $\lambda_i \leq 1/(2\Delta)$, $1 \leq i \leq p$, in order to visualize p different frequencies. For the periodic curves in Plot 4.2.4 this means to choose $9/11 \leq 1/(2\Delta)$ or $\Delta \leq 11/18$.

Exercises

4.1. Let $y(t) = A \cos(2\pi\lambda t) + B \sin(2\pi\lambda t)$ be a harmonic component. Show that y can be written as $y(t) = \alpha \cos(2\pi\lambda t - \varphi)$, where α is the amplitude, i.e., the maximum departure of the wave from zero and φ is the phase displacement.

4.2. Show that

$$\begin{aligned}
 \sum_{t=1}^n \cos(2\pi\lambda t) &= \begin{cases} n, & \lambda \in \mathbb{Z} \\ \cos(\pi\lambda(n+1)) \frac{\sin(\pi\lambda n)}{\sin(\pi\lambda)}, & \lambda \notin \mathbb{Z} \end{cases} \\
 \sum_{t=1}^n \sin(2\pi\lambda t) &= \begin{cases} 0, & \lambda \in \mathbb{Z} \\ \sin(\pi\lambda(n+1)) \frac{\sin(\pi\lambda n)}{\sin(\pi\lambda)}, & \lambda \notin \mathbb{Z}. \end{cases}
 \end{aligned}$$

Hint: Compute $\sum_{t=1}^n e^{i2\pi\lambda t}$, where $e^{i\varphi} = \cos(\varphi) + i\sin(\varphi)$ is the complex valued exponential function.

4.3. Verify Lemma 4.1.2. Hint: Exercise 4.2.

4.4. Suppose that the time series $(y_t)_t$ satisfies the additive model with seasonal component

$$s(t) = \sum_{k=1}^s A_k \cos\left(2\pi\frac{k}{s}t\right) + \sum_{k=1}^s B_k \sin\left(2\pi\frac{k}{s}t\right).$$

Show that $s(t)$ is eliminated by the seasonal differencing $\Delta_s y_t = y_t - y_{t-s}$.

4.5. Fit a harmonic component with frequency λ to a time series y_1, \dots, y_N , where $\lambda \in \mathbb{Z}$ and $\lambda - 0.5 \in \mathbb{Z}$. Compute the least squares estimator and the pertaining residual sum of squares.

4.6. Put $y_t = t$, $t = 1, \dots, n$. Show that

$$I(k/n) = \frac{n}{4 \sin^2(\pi k/n)}, \quad k = 1, \dots, [(n-1)/2].$$

Hint: Use the equations

$$\begin{aligned} \sum_{t=1}^{n-1} t \sin(\theta t) &= \frac{\sin(n\theta)}{4 \sin^2(\theta/2)} - \frac{n \cos((n-0.5)\theta)}{2 \sin(\theta/2)} \\ \sum_{t=1}^{n-1} t \cos(\theta t) &= \frac{n \sin((n-0.5)\theta)}{2 \sin(\theta/2)} - \frac{1 - \cos(n\theta)}{4 \sin^2(\theta/2)}. \end{aligned}$$

4.7. (Unemployed1 Data) Plot the periodogram of the first order differences of the numbers of unemployed in the building trade as introduced in Example 1.1.1.

4.8. (Airline Data) Plot the periodogram of the variance stabilized and trend adjusted Airline Data, introduced in Example 1.3.1. Add a seasonal adjustment and compare the periodograms.

4.9. The contribution of the autocovariance $c(k)$, $k \geq 1$, to the periodogram can be illustrated by plotting the functions $\pm \cos(2\pi\lambda k)$, $\lambda \in [0.5]$.

- (i) Which conclusion about the intensities of large or small frequencies can be drawn from a positive value $c(1) > 0$ or a negative one $c(1) < 0$?
- (ii) Which effect has an increase of $|c(2)|$ if all other parameters remain unaltered?
- (iii) What can you say about the effect of an increase of $c(k)$ on the periodogram at the values $40, 1/k, 2/k, 3/k, \dots$ and the intermediate values $1/2k, 3/(2k), 5/(2k)$? Illustrate the effect at a time series with seasonal component $k = 12$.

4.10. Establish a version of the inverse Fourier transform in real terms.

4.11. Let $\mathbf{a} = (a_t)_{t \in \mathbb{Z}}$ and $\mathbf{b} = (b_t)_{t \in \mathbb{Z}}$ be absolute summable sequences.

- (i) Show that for $\alpha \mathbf{a} + \beta \mathbf{b} := (\alpha a_t + \beta b_t)_{t \in \mathbb{Z}}$, $\alpha, \beta \in \mathbb{R}$,

$$f_{\alpha \mathbf{a} + \beta \mathbf{b}}(\lambda) = \alpha f_{\mathbf{a}}(\lambda) + \beta f_{\mathbf{b}}(\lambda).$$

- (ii) For $\mathbf{ab} := (a_t b_t)_{t \in \mathbb{Z}}$ we have

$$f_{\mathbf{ab}}(\lambda) = f_{\mathbf{a}} * f_{\mathbf{b}}(\lambda) := \int_0^1 f_{\mathbf{a}}(\mu) f_{\mathbf{b}}(\lambda - \mu) d\mu.$$

- (iii) Show that for $\mathbf{a} * \mathbf{b} := (\sum_{s \in \mathbb{Z}} a_s b_{t-s})_{t \in \mathbb{Z}}$ (convolution)

$$f_{\mathbf{a} * \mathbf{b}}(\lambda) = f_{\mathbf{a}}(\lambda) f_{\mathbf{b}}(\lambda).$$

4.12. (Fast Fourier Transform (FFT)) The Fourier transform of a finite sequence a_0, a_1, \dots, a_{N-1} can be represented under suitable conditions as the composition of Fourier transforms. Put

$$f(s/N) = \sum_{t=0}^{N-1} a_t e^{-i2\pi st/N}, \quad s = 0, \dots, N-1,$$

which is the Fourier transform of length N . Suppose that $N = KM$ with $K, M \in \mathbb{N}$. Show that f can be represented as Fourier transform

of length K , computed for a Fourier transform of length M .

Hint: Each $t, s \in \{0, \dots, N-1\}$ can uniquely be written as

$$\begin{aligned} t &= t_0 + t_1 K, & t_0 &\in \{0, \dots, K-1\}, & t_1 &\in \{0, \dots, M-1\} \\ s &= s_0 + s_1 M, & s_0 &\in \{0, \dots, M-1\}, & s_1 &\in \{0, \dots, K-1\}. \end{aligned}$$

Sum over t_0 and t_1 .

4.13. (Star Data) Suppose that the Star Data are only observed weekly (i.e., keep only every seventh observation). Is an aliasing effect observable?

The Spectrum of a Stationary Process

In this chapter we investigate the spectrum of a real valued stationary process, which is the Fourier transform of its (theoretical) autocovariance function. Its empirical counterpart, the periodogram, was investigated in the preceding sections, cf. Theorem 4.2.3.

Let $(Y_t)_{t \in \mathbb{Z}}$ be a (real valued) stationary process with an absolutely summable autocovariance function $\gamma(t)$, $t \in \mathbb{Z}$. Its Fourier transform

$$f(\lambda) := \sum_{t \in \mathbb{Z}} \gamma(t) e^{-i2\pi\lambda t} = \gamma(0) + 2 \sum_{t \in \mathbb{N}} \gamma(t) \cos(2\pi\lambda t), \quad \lambda \in \mathbb{R},$$

is called *spectral density* or *spectrum* of the process $(Y_t)_{t \in \mathbb{Z}}$. By the inverse Fourier transform in Theorem 4.2.5 we have

$$\gamma(t) = \int_0^1 f(\lambda) e^{i2\pi\lambda t} d\lambda = \int_0^1 f(\lambda) \cos(2\pi\lambda t) d\lambda.$$

For $t = 0$ we obtain

$$\gamma(0) = \int_0^1 f(\lambda) d\lambda,$$

which shows that the spectrum is a decomposition of the variance $\gamma(0)$. In Section 5.3 we will in particular compute the spectrum of an ARMA-process. As a preparatory step we investigate properties of spectra for arbitrary absolutely summable filters.

5.1 Characterizations of Autocovariance Functions

Recall that the autocovariance function $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ of a stationary process $(Y_t)_{t \in \mathbb{Z}}$ is given by

$$\gamma(h) = E(Y_{t+h}Y_t) - E(Y_{t+h})E(Y_t), \quad h \in \mathbb{Z},$$

with the properties

$$\gamma(0) \geq 0, \quad |\gamma(h)| \leq \gamma(0), \quad \gamma(h) = \gamma(-h), \quad h \in \mathbb{Z}. \quad (5.1)$$

The following result characterizes an autocovariance function in terms of positive semidefiniteness.

Theorem 5.1.1. *A symmetric function $K : \mathbb{Z} \rightarrow \mathbb{R}$ is the autocovariance function of a stationary process $(Y_t)_{t \in \mathbb{Z}}$ iff K is a positive semidefinite function, i.e., $K(-n) = K(n)$ and*

$$\sum_{1 \leq r, s \leq n} x_r K(r-s) x_s \geq 0 \quad (5.2)$$

for arbitrary $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathbb{R}$.

Proof. It is easy to see that (5.2) is a necessary condition for K to be the autocovariance function of a stationary process, see Exercise 5.19. It remains to show that (5.2) is sufficient, i.e., we will construct a stationary process, whose autocovariance function is K .

We will define a family of finite-dimensional normal distributions, which satisfies the *consistency condition* of Kolmogorov's theorem, cf. Brockwell and Davis (1991, Theorem 1.2.1). This result implies the existence of a process $(V_t)_{t \in \mathbb{Z}}$, whose finite dimensional distributions coincide with the given family.

Define the $n \times n$ -matrix

$$K^{(n)} := (K(r-s))_{1 \leq r, s \leq n},$$

which is positive semidefinite. Consequently there exists an n -dimensional normal distributed random vector (V_1, \dots, V_n) with mean vector

zero and covariance matrix $K^{(n)}$. Define now for each $n \in \mathbb{N}$ and $t \in \mathbb{Z}$ a distribution function on \mathbb{R}^n by

$$F_{t+1,\dots,t+n}(v_1, \dots, v_n) := P\{V_1 \leq v_1, \dots, V_n \leq v_n\}.$$

This defines a family of distribution functions indexed by consecutive integers. Let now $t_1 < \dots < t_m$ be arbitrary integers. Choose $t \in \mathbb{Z}$ and $n \in \mathbb{N}$ such that $t_i = t + n_i$, where $1 \leq n_1 < \dots < n_m \leq n$. We define now

$$F_{t_1,\dots,t_m}((v_i)_{1 \leq i \leq m}) := P\{V_{n_i} \leq v_i, 1 \leq i \leq m\}.$$

Note that F_{t_1,\dots,t_m} does not depend on the special choice of t and n and thus, we have defined a family of distribution functions indexed by $t_1 < \dots < t_m$ on \mathbb{R}^m for each $m \in \mathbb{N}$, which obviously satisfies the consistency condition of *Kolmogorov's theorem*. This result implies the existence of a process $(V_t)_{t \in \mathbb{Z}}$, whose finite dimensional distribution at $t_1 < \dots < t_m$ has distribution function F_{t_1,\dots,t_m} . This process has, therefore, mean vector zero and covariances $E(V_{t+h}V_t) = K(h)$, $h \in \mathbb{Z}$. \square

Spectral Distribution Function and Spectral Density

The preceding result provides a characterization of an autocovariance function in terms of positive semidefiniteness. The following characterization of positive semidefinite functions is known as *Herglotz's theorem*. We use in the following the notation $\int_0^1 g(\lambda) dF(\lambda)$ in place of $\int_{(0,1]} g(\lambda) dF(\lambda)$.

Theorem 5.1.2. *A symmetric function $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ is positive semidefinite iff it can be represented as an integral*

$$\gamma(h) = \int_0^1 e^{i2\pi\lambda h} dF(\lambda) = \int_0^1 \cos(2\pi\lambda h) dF(\lambda), \quad h \in \mathbb{Z}, \quad (5.3)$$

where F is a real valued measure generating function on $[0, 1]$ with $F(0) = 0$. The function F is uniquely determined.

The uniquely determined function F , which is a right-continuous, increasing and bounded function, is called the *spectral distribution function* of γ . If F has a derivative f and, thus, $F(\lambda) = F(\lambda) - F(0) =$

$\int_0^\lambda f(x) dx$ for $0 \leq \lambda \leq 1$, then f is called the *spectral density* of γ . Note that the property $\sum_{h \geq 0} |\gamma(h)| < \infty$ already implies the existence of a spectral density of γ , cf. Theorem 4.2.5 and the proof of Corollary 5.1.5.

Recall that $\gamma(0) = \int_0^1 dF(\lambda) = F(1)$ and thus, the autocorrelation function $\rho(h) = \gamma(h)/\gamma(0)$ has the above integral representation, but with F replaced by the *probability distribution function* $F/\gamma(0)$.

Proof of Theorem 5.1.2. We establish first the uniqueness of F . Let G be another measure generating function with $G(\lambda) = 0$ for $\lambda \leq 0$ and constant for $\lambda \geq 1$ such that

$$\gamma(h) = \int_0^1 e^{i2\pi\lambda h} dF(\lambda) = \int_0^1 e^{i2\pi\lambda h} dG(\lambda), \quad h \in \mathbb{Z}.$$

Let now ψ be a continuous function on $[0, 1]$. From calculus we know (cf. Rudin, 1986, Section 4.24) that we can find for arbitrary $\varepsilon > 0$ a trigonometric polynomial $p_\varepsilon(\lambda) = \sum_{h=-N}^N a_h e^{i2\pi\lambda h}$, $0 \leq \lambda \leq 1$, such that

$$\sup_{0 \leq \lambda \leq 1} |\psi(\lambda) - p_\varepsilon(\lambda)| \leq \varepsilon.$$

As a consequence we obtain that

$$\begin{aligned} \int_0^1 \psi(\lambda) dF(\lambda) &= \int_0^1 p_\varepsilon(\lambda) dF(\lambda) + r_1(\varepsilon) \\ &= \int_0^1 p_\varepsilon(\lambda) dG(\lambda) + r_1(\varepsilon) \\ &= \int_0^1 \psi(\lambda) dG(\lambda) + r_2(\varepsilon), \end{aligned}$$

where $r_i(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$, $i = 1, 2$, and, thus,

$$\int_0^1 \psi(\lambda) dF(\lambda) = \int_0^1 \psi(\lambda) dG(\lambda).$$

Since ψ was an arbitrary continuous function, this in turn together with $F(0) = G(0) = 0$ implies $F = G$.

Suppose now that γ has the representation (5.3). We have for arbitrary $x_i \in \mathbb{R}$, $i = 1, \dots, n$

$$\begin{aligned} \sum_{1 \leq r, s \leq n} x_r \gamma(r-s) x_s &= \int_0^1 \sum_{1 \leq r, s \leq n} x_r x_s e^{i2\pi\lambda(r-s)} dF(\lambda) \\ &= \int_0^1 \left| \sum_{r=1}^n x_r e^{i2\pi\lambda r} \right|^2 dF(\lambda) \geq 0, \end{aligned}$$

i.e., γ is positive semidefinite.

Suppose conversely that $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ is a positive semidefinite function. This implies that for $0 \leq \lambda \leq 1$ and $N \in \mathbb{N}$ (Exercise 5.2)

$$\begin{aligned} f_N(\lambda) &:= \frac{1}{N} \sum_{1 \leq r, s \leq N} e^{-i2\pi\lambda r} \gamma(r-s) e^{i2\pi\lambda s} \\ &= \frac{1}{N} \sum_{|m| < N} (N - |m|) \gamma(m) e^{-i2\pi\lambda m} \geq 0. \end{aligned}$$

Put now

$$F_N(\lambda) := \int_0^\lambda f_N(x) dx, \quad 0 \leq \lambda \leq 1.$$

Then we have for each $h \in \mathbb{Z}$

$$\begin{aligned} \int_0^1 e^{i2\pi\lambda h} dF_N(\lambda) &= \sum_{|m| < N} \left(1 - \frac{|m|}{N}\right) \gamma(m) \int_0^1 e^{i2\pi\lambda(h-m)} d\lambda \\ &= \begin{cases} \left(1 - \frac{|h|}{N}\right) \gamma(h), & \text{if } |h| < N \\ 0, & \text{if } |h| \geq N. \end{cases} \end{aligned} \quad (5.4)$$

Since $F_N(1) = \gamma(0) < \infty$ for any $N \in \mathbb{N}$, we can apply *Helly's selection theorem* (cf. Billingsley, 1968, page 226ff) to deduce the existence of a measure generating function \tilde{F} and a subsequence $(F_{N_k})_k$ such that F_{N_k} converges weakly to \tilde{F} i.e.,

$$\int_0^1 g(\lambda) dF_{N_k}(\lambda) \xrightarrow{k \rightarrow \infty} \int_0^1 g(\lambda) d\tilde{F}(\lambda)$$

for every continuous and bounded function $g : [0, 1] \rightarrow \mathbb{R}$ (cf. Billingsley, 1968, Theorem 2.1). Put now $F(\lambda) := \tilde{F}(\lambda) - \tilde{F}(0)$. Then F is a

measure generating function with $F(0) = 0$ and

$$\int_0^1 g(\lambda) d\tilde{F}(\lambda) = \int_0^1 g(\lambda) dF(\lambda).$$

If we replace N in (5.4) by N_k and let k tend to infinity, we now obtain representation (5.3). \square

Example 5.1.3. A white noise $(\varepsilon_t)_{t \in \mathbb{Z}}$ has the autocovariance function

$$\gamma(h) = \begin{cases} \sigma^2, & h = 0 \\ 0, & h \in \mathbb{Z} \setminus \{0\}. \end{cases}$$

Since

$$\int_0^1 \sigma^2 e^{i2\pi\lambda h} d\lambda = \begin{cases} \sigma^2, & h = 0 \\ 0, & h \in \mathbb{Z} \setminus \{0\}, \end{cases}$$

the process (ε_t) has by Theorem 5.1.2 the constant spectral density $f(\lambda) = \sigma^2$, $0 \leq \lambda \leq 1$. This is the name giving property of the white noise process: As the white light is characteristically perceived to belong to objects that reflect nearly all incident energy throughout the visible spectrum, a white noise process weighs all possible frequencies equally.

Corollary 5.1.4. A symmetric function $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ is the autocovariance function of a stationary process $(Y_t)_{t \in \mathbb{Z}}$, iff it satisfies one of the following two (equivalent) conditions:

(i) $\gamma(h) = \int_0^1 e^{i2\pi\lambda h} dF(\lambda)$, $h \in \mathbb{Z}$, where F is a measure generating function on $[0, 1]$ with $F(0) = 0$.

(ii) $\sum_{1 \leq r, s \leq n} x_r \gamma(r - s) x_s \geq 0$ for each $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathbb{R}$.

Proof. Theorem 5.1.2 shows that (i) and (ii) are equivalent. The assertion is then a consequence of Theorem 5.1.1. \square

Corollary 5.1.5. A symmetric function $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ with $\sum_{t \in \mathbb{Z}} |\gamma(t)| < \infty$ is the autocovariance function of a stationary process iff

$$f(\lambda) := \sum_{t \in \mathbb{Z}} \gamma(t) e^{-i2\pi\lambda t} \geq 0, \quad \lambda \in [0, 1].$$

The function f is in this case the spectral density of γ .

Proof. Suppose first that γ is an autocovariance function. Since γ is in this case positive semidefinite by Theorem 5.1.1, and $\sum_{t \in \mathbb{Z}} |\gamma(t)| < \infty$ by assumption, we have (Exercise 5.2)

$$\begin{aligned} 0 \leq f_N(\lambda) &:= \frac{1}{N} \sum_{1 \leq r, s \leq N} e^{-i2\pi\lambda r} \gamma(r-s) e^{i2\pi\lambda s} \\ &= \sum_{|t| < N} \left(1 - \frac{|t|}{N}\right) \gamma(t) e^{-i2\pi\lambda t} \rightarrow f(\lambda) \quad \text{as } N \rightarrow \infty, \end{aligned}$$

see Exercise 5.8. The function f is consequently nonnegative. The inverse Fourier transform in Theorem 4.2.5 implies $\gamma(t) = \int_0^1 f(\lambda) e^{i2\pi\lambda t} d\lambda$, $t \in \mathbb{Z}$ i.e., f is the spectral density of γ .

Suppose on the other hand that $f(\lambda) = \sum_{t \in \mathbb{Z}} \gamma(t) e^{i2\pi\lambda t} \geq 0$, $0 \leq \lambda \leq 1$. The inverse Fourier transform implies $\gamma(t) = \int_0^1 f(\lambda) e^{i2\pi\lambda t} d\lambda = \int_0^1 e^{i2\pi\lambda t} dF(\lambda)$, where $F(\lambda) = \int_0^\lambda f(x) dx$, $0 \leq \lambda \leq 1$. Thus we have established representation (5.3), which implies that γ is positive semidefinite, and, consequently, γ is by Corollary 5.1.4 the autocovariance function of a stationary process. \square

Example 5.1.6. Choose a number $\rho \in \mathbb{R}$. The function

$$\gamma(h) = \begin{cases} 1, & \text{if } h = 0 \\ \rho, & \text{if } h \in \{-1, 1\} \\ 0, & \text{elsewhere} \end{cases}$$

is the autocovariance function of a stationary process iff $|\rho| \leq 0.5$. This follows from

$$\begin{aligned} f(\lambda) &= \sum_{t \in \mathbb{Z}} \gamma(t) e^{i2\pi\lambda t} \\ &= \gamma(0) + \gamma(1) e^{i2\pi\lambda} + \gamma(-1) e^{-i2\pi\lambda} \\ &= 1 + 2\rho \cos(2\pi\lambda) \geq 0 \end{aligned}$$

for $\lambda \in [0, 1]$ iff $|\rho| \leq 0.5$. Note that the function γ is the autocorrelation function of an MA(1)-process, cf. Example 2.2.2.

The spectral distribution function of a stationary process satisfies (Exercise 5.10)

$$F(0.5 + \lambda) - F(0.5^-) = F(0.5) - F((0.5 - \lambda)^-), \quad 0 \leq \lambda < 0.5,$$

where $F(x^-) := \lim_{\varepsilon \downarrow 0} F(x - \varepsilon)$ is the left-hand limit of F at $x \in (0, 1]$. If F has a derivative f , we obtain from the above symmetry $f(0.5 + \lambda) = f(0.5 - \lambda)$ or, equivalently, $f(1 - \lambda) = f(\lambda)$ and, hence,

$$\gamma(h) = \int_0^1 \cos(2\pi\lambda h) dF(\lambda) = 2 \int_0^{0.5} \cos(2\pi\lambda h) f(\lambda) d\lambda.$$

The autocovariance function of a stationary process is, therefore, determined by the values $f(\lambda)$, $0 \leq \lambda \leq 0.5$, if the spectral density exists. Recall, moreover, that the smallest nonconstant period P_0 visible through observations evaluated at time points $t = 1, 2, \dots$ is $P_0 = 2$ i.e., the largest observable frequency is the Nyquist frequency $\lambda_0 = 1/P_0 = 0.5$, cf. the end of Section 4.2. Hence, the spectral density $f(\lambda)$ matters only for $\lambda \in [0, 0.5]$.

Remark 5.1.7. The preceding discussion shows that a function $f : [0, 1] \rightarrow \mathbb{R}$ is the spectral density of a stationary process iff f satisfies the following three conditions

- (i) $f(\lambda) \geq 0$,
- (ii) $f(\lambda) = f(1 - \lambda)$,
- (iii) $\int_0^1 f(\lambda) d\lambda < \infty$.

5.2 Linear Filters and Frequencies

The application of a linear filter to a stationary time series has a quite complex effect on its autocovariance function, see Theorem 2.1.6. Its effect on the spectral density, if it exists, turns, however, out to be quite simple. We use in the following again the notation $\int_0^\lambda g(x) dF(x)$ in place of $\int_{(0, \lambda]} g(x) dF(x)$.

Theorem 5.2.1. *Let $(Z_t)_{t \in \mathbb{Z}}$ be a stationary process with spectral distribution function F_Z and let $(a_t)_{t \in \mathbb{Z}}$ be an absolutely summable filter with Fourier transform f_a . The linear filtered process $Y_t := \sum_{u \in \mathbb{Z}} a_u Z_{t-u}$, $t \in \mathbb{Z}$, then has the spectral distribution function*

$$F_Y(\lambda) := \int_0^\lambda |f_a(x)|^2 dF_Z(x), \quad 0 \leq \lambda \leq 1. \quad (5.5)$$

If in addition $(Z_t)_{t \in \mathbb{Z}}$ has a spectral density f_Z , then

$$f_Y(\lambda) := |f_a(\lambda)|^2 f_Z(\lambda), \quad 0 \leq \lambda \leq 1, \quad (5.6)$$

is the spectral density of $(Y_t)_{t \in \mathbb{Z}}$.

Proof. Theorem 2.1.6 yields that $(Y_t)_{t \in \mathbb{Z}}$ is stationary with autocovariance function

$$\gamma_Y(t) = \sum_{u \in \mathbb{Z}} \sum_{w \in \mathbb{Z}} a_u a_w \gamma_Z(t - u + w), \quad t \in \mathbb{Z},$$

where γ_Z is the autocovariance function of (Z_t) . Its spectral representation (5.3) implies

$$\begin{aligned} \gamma_Y(t) &= \sum_{u \in \mathbb{Z}} \sum_{w \in \mathbb{Z}} a_u a_w \int_0^1 e^{i2\pi\lambda(t-u+w)} dF_Z(\lambda) \\ &= \int_0^1 \left(\sum_{u \in \mathbb{Z}} a_u e^{-i2\pi\lambda u} \right) \left(\sum_{w \in \mathbb{Z}} a_w e^{i2\pi\lambda w} \right) e^{i2\pi\lambda t} dF_Z(\lambda) \\ &= \int_0^1 |f_a(\lambda)|^2 e^{i2\pi\lambda t} dF_Z(\lambda) \\ &= \int_0^1 e^{i2\pi\lambda t} dF_Y(\lambda). \end{aligned}$$

Theorem 5.1.2 now implies that F_Y is the uniquely determined spectral distribution function of $(Y_t)_{t \in \mathbb{Z}}$. The second to last equality yields in addition the spectral density (5.6). \square

Transfer Function and Power Transfer Function

Since the spectral density is a measure of intensity of a frequency λ inherent in a stationary process (see the discussion of the periodogram in Section 4.2), the effect (5.6) of applying a linear filter (a_t) with Fourier transform f_a can easily be interpreted. While the intensity of λ is diminished by the filter (a_t) iff $|f_a(\lambda)| < 1$, its intensity is amplified iff $|f_a(\lambda)| > 1$. The Fourier transform f_a of (a_t) is, therefore, also called *transfer function* and the function $g_a(\lambda) := |f_a(\lambda)|^2$ is referred to as the *gain* or *power transfer function* of the filter $(a_t)_{t \in \mathbb{Z}}$.

Example 5.2.2. The simple moving average of length three

$$a_u = \begin{cases} 1/3, & u \in \{-1, 0, 1\} \\ 0 & \text{elsewhere} \end{cases}$$

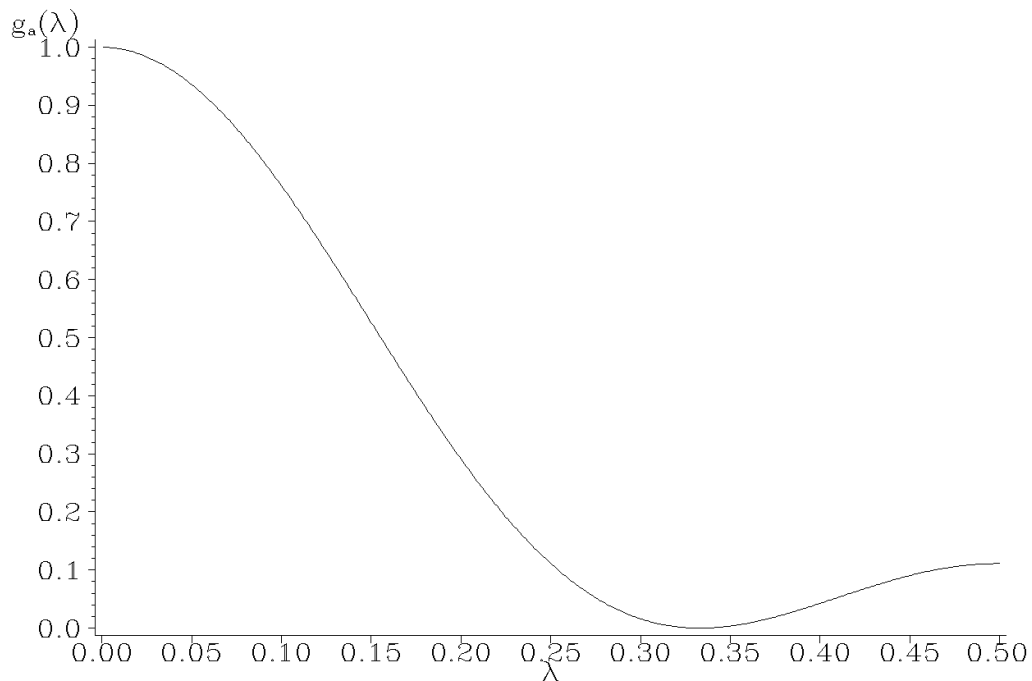
has the transfer function

$$f_a(\lambda) = \frac{1}{3} + \frac{2}{3} \cos(2\pi\lambda)$$

and the power transfer function

$$g_a(\lambda) = \begin{cases} 1, & \lambda = 0 \\ \left(\frac{\sin(3\pi\lambda)}{3\sin(\pi\lambda)} \right)^2, & \lambda \in (0, 0.5] \end{cases}$$

(see Exercise 5.13 and Theorem 4.2.2). This power transfer function is plotted in Plot 5.2.1 below. It shows that frequencies λ close to zero i.e., those corresponding to a large period, remain essentially unaltered. Frequencies λ close to 0.5, which correspond to a short period, are, however, damped by the approximate factor 0.1, when the moving average (a_u) is applied to a process. The frequency $\lambda = 1/3$ is completely eliminated, since $g_a(1/3) = 0$.



Plot 5.2.1: Power transfer function of the simple moving average of length three.

```

1  /* power_transfer_sma3.sas */
2  TITLE1 'Power transfer function';
3  TITLE2 'of the simple moving average of length 3';
4
5  /* Compute power transfer function */
6  DATA datal;
7      DO lambda=.001 TO .5 BY .001;
8          g=(SIN(3*CONSTANT('PI')*lambda)/(3*SIN(CONSTANT('PI')*lambda)))
9              ↪**2;
10         OUTPUT;
11     END;
12
13 /* Graphical options */
14 AXIS1 LABEL=('g' H=1 'a' H=2 '(' F=CGREEK 'l)');
15 AXIS2 LABEL=(F=CGREEK 'l');
16 SYMBOL1 V=NONE C=GREEN I=JOIN;
17
18 /* Plot power transfer function */
19 PROC GPLOT DATA=datal;
20     PLOT g*lambda / VAXIS=AXIS1 HAXIS=AXIS2;
21 RUN; QUIT;

```

Example 5.2.3. The first order difference filter

$$a_u = \begin{cases} 1, & u = 0 \\ -1, & u = 1 \\ 0 & \text{elsewhere} \end{cases}$$

has the transfer function

$$f_a(\lambda) = 1 - e^{-i2\pi\lambda}.$$

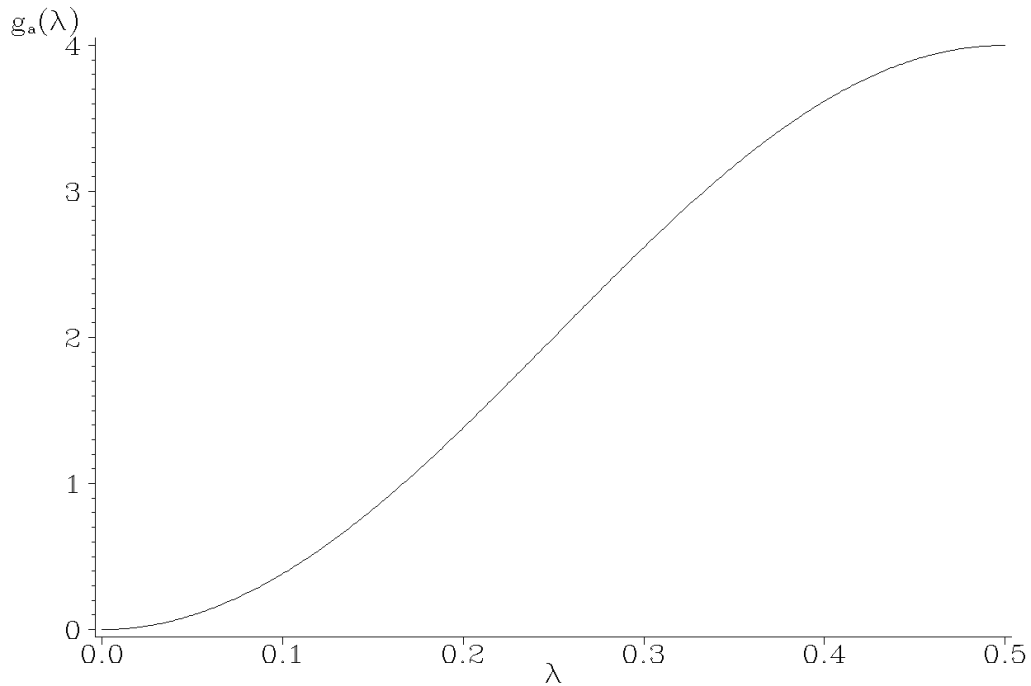
Since

$$f_a(\lambda) = e^{-i\pi\lambda}(e^{i\pi\lambda} - e^{-i\pi\lambda}) = ie^{-i\pi\lambda}2\sin(\pi\lambda),$$

its power transfer function is

$$g_a(\lambda) = 4\sin^2(\pi\lambda).$$

The first order difference filter, therefore, damps frequencies close to zero but amplifies those close to 0.5.



Plot 5.2.2: Power transfer function of the first order difference filter.

Example 5.2.4. The preceding example immediately carries over to the seasonal difference filter of arbitrary length $s \geq 0$ i.e.,

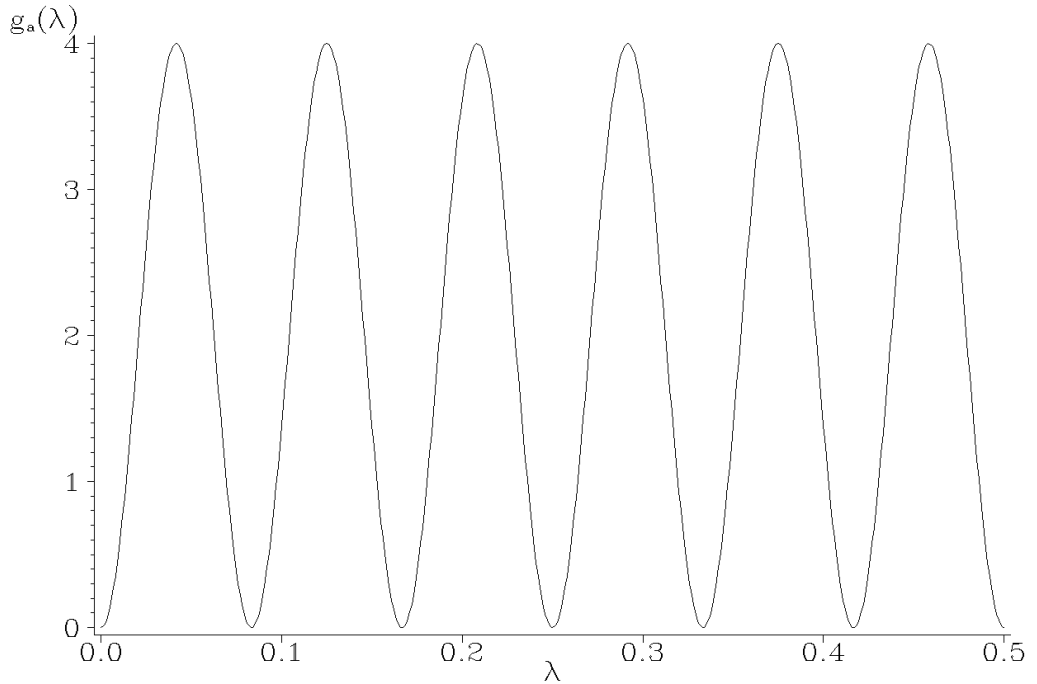
$$a_u^{(s)} = \begin{cases} 1, & u = 0 \\ -1, & u = s \\ 0 & \text{elsewhere,} \end{cases}$$

which has the transfer function

$$f_{a^{(s)}}(\lambda) = 1 - e^{-i2\pi\lambda s}$$

and the power transfer function

$$g_{a^{(s)}}(\lambda) = 4 \sin^2(\pi\lambda s).$$



Plot 5.2.3: Power transfer function of the seasonal difference filter of order 12.

Since $\sin^2(x) = 0$ iff $x = k\pi$ and $\sin^2(x) = 1$ iff $x = (2k + 1)\pi/2$, $k \in \mathbb{Z}$, the power transfer function $g_{a(s)}(\lambda)$ satisfies for $k \in \mathbb{Z}$

$$g_{a(s)}(\lambda) = \begin{cases} 0, & \text{iff } \lambda = k/s \\ 4 & \text{iff } \lambda = (2k + 1)/(2s). \end{cases}$$

This implies, for example, in the case of $s = 12$ that those frequencies, which are multiples of $1/12 = 0.0833$, are eliminated, whereas the midpoint frequencies $k/12 + 1/24$ are amplified. This means that the seasonal difference filter on the one hand does what we would like it to do, namely to eliminate the frequency $1/12$, but on the other hand it generates unwanted side effects by eliminating also multiples of $1/12$ and by amplifying midpoint frequencies. This observation gives rise to the problem, whether one can construct linear filters that have prescribed properties.

Least Squares Based Filter Design

A *low pass filter* aims at eliminating high frequencies, a *high pass filter* aims at eliminating small frequencies and a *band pass filter* allows only frequencies in a certain band $[\lambda_0 - \Delta, \lambda_0 + \Delta]$ to pass through. They consequently should have the ideal power transfer functions

$$\begin{aligned} g_{low}(\lambda) &= \begin{cases} 1, & \lambda \in [0, \lambda_0] \\ 0, & \lambda \in (\lambda_0, 0.5] \end{cases} \\ g_{high}(\lambda) &= \begin{cases} 0, & \lambda \in [0, \lambda_0] \\ 1, & \lambda \in [\lambda_0, 0.5] \end{cases} \\ g_{band}(\lambda) &= \begin{cases} 1, & \lambda \in [\lambda_0 - \Delta, \lambda_0 + \Delta] \\ 0 & \text{elsewhere,} \end{cases} \end{aligned}$$

where λ_0 is the cut off frequency in the first two cases and $[\lambda_0 - \Delta, \lambda_0 + \Delta]$ is the cut off interval with bandwidth $2\Delta > 0$ in the final one. Therefore, the question naturally arises, whether there actually exist filters, which have a prescribed power transfer function. One possible approach for fitting a linear filter with weights a_u to a given transfer function f is offered by utilizing least squares. Since only filters of finite length matter in applications, one chooses a transfer function

$$f_a(\lambda) = \sum_{u=r}^s a_u e^{-i2\pi\lambda u}$$

with fixed integers r, s and fits this function f_a to f by minimizing the integrated squared error

$$\int_0^{0.5} |f(\lambda) - f_a(\lambda)|^2 d\lambda$$

in $(a_u)_{r \leq u \leq s} \in \mathbb{R}^{s-r+1}$. This is achieved for the choice (Exercise 5.16)

$$a_u = 2 \operatorname{Re} \left(\int_0^{0.5} f(\lambda) e^{i2\pi\lambda u} d\lambda \right), \quad u = r, \dots, s,$$

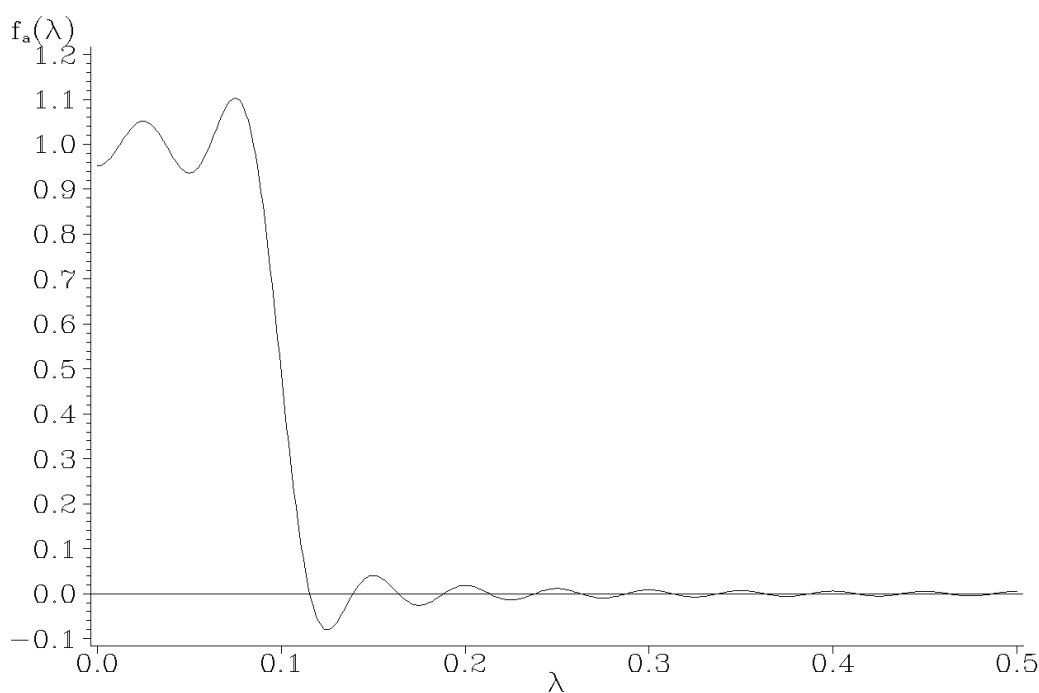
which is formally the real part of the inverse Fourier transform of f .

Example 5.2.5. For the low pass filter with cut off frequency $0 < \lambda_0 < 0.5$ and ideal transfer function

$$f(\lambda) = 1_{[0, \lambda_0]}(\lambda)$$

we obtain the weights

$$a_u = 2 \int_0^{\lambda_0} \cos(2\pi\lambda u) d\lambda = \begin{cases} 2\lambda_0, & u = 0 \\ \frac{1}{\pi u} \sin(2\pi\lambda_0 u), & u \neq 0. \end{cases}$$



Plot 5.2.4: Transfer function of least squares fitted low pass filter with cut off frequency $\lambda_0 = 1/10$ and $r = -20$, $s = 20$.

```

1  /* transfer.sas */
2  TITLE1 'Transfer function';
3  TITLE2 'of least squares fitted low pass filter';
4
5  /* Compute transfer function */
6  DATA data1;
7    DO lambda=0 TO .5 BY .001;
8      f=2*1/10;

```

```

9      DO u=1 TO 20;
10         f=f+2*1/(CONSTANT('PI')*u)*SIN(2*CONSTANT('PI')*1/10*u)*COS(2*
           ↳CONSTANT('PI')*lambda*u);
11      END;
12      OUTPUT;
13      END;
14
15      /* Graphical options */
16      AXIS1 LABEL=('f' H=1 'a' H=2 F=CGREEK '(1)');
17      AXIS2 LABEL=(F=CGREEK 'l');
18      SYMBOL1 V=NONE C=GREEN I=JOIN L=1;
19
20      /* Plot transfer function */
21      PROC GPLOT DATA=data1;
22         PLOT f*lambda / VAXIS=AXIS1 HAXIS=AXIS2 VREF=0;
23      RUN; QUIT;

```

The programs in Section 5.2 (Linear Filters and Frequencies) are just made for the purpose of generating graphics, which demonstrate the shape of power transfer functions or, in case of Program 5.2.4 (transfer.sas), of a transfer function. They all consist of two parts, a DATA step and a PROC step.

In the DATA step values of the power transfer function are calculated and stored in a variable

g by a DO loop over λ from 0 to 0.5 with a small increment. In Program 5.2.4 (transfer.sas) it is necessary to use a second DO loop within the first one to calculate the sum used in the definition of the transfer function f .

Two AXIS statements defining the axis labels and a SYMBOL statement precede the procedure PLOT, which generates the plot of g or f versus λ .

The transfer function in Plot 5.2.4, which approximates the ideal transfer function $1_{[0,0.1]}$, shows higher oscillations near the cut off point $\lambda_0 = 0.1$. This is known as *Gibbs' phenomenon* and requires further smoothing of the fitted transfer function if these oscillations are to be damped (cf. Bloomfield, 2000, Section 6.4).

5.3 Spectral Density of an ARMA-Process

Theorem 5.2.1 enables us to compute spectral densities of ARMA-processes.

Theorem 5.3.1. *Suppose that*

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}, \quad t \in \mathbb{Z},$$

is a stationary ARMA(p, q)-process, where (ε_t) is a white noise with

variance σ^2 . Put

$$\begin{aligned} A(z) &:= 1 - a_1 z - a_2 z^2 - \cdots - a_p z^p, \\ B(z) &:= 1 + b_1 z + b_2 z^2 + \cdots + b_q z^q \end{aligned}$$

and suppose that the process (Y_t) satisfies the stationarity condition (2.4), i.e., the roots of the equation $A(z) = 0$ are outside of the unit circle. The process (Y_t) then has the spectral density

$$f_Y(\lambda) = \sigma^2 \frac{|B(e^{-i2\pi\lambda})|^2}{|A(e^{-i2\pi\lambda})|^2} = \sigma^2 \frac{|1 + \sum_{v=1}^q b_v e^{-i2\pi\lambda v}|^2}{|1 - \sum_{u=1}^p a_u e^{-i2\pi\lambda u}|^2}. \quad (5.7)$$

Proof. Since the process (Y_t) is supposed to satisfy the stationarity condition (2.4) it is causal, i.e., $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$, $t \in \mathbb{Z}$, for some absolutely summable constants α_v , $v \geq 0$, see Section 2.2. The white noise process (ε_t) has by Example 5.1.3 the spectral density $f_\varepsilon(\lambda) = \sigma^2$ and, thus, (Y_t) has by Theorem 5.2.1 a spectral density f_Y . The application of Theorem 5.2.1 to the process

$$X_t := Y_t - a_1 Y_{t-1} - \cdots - a_p Y_{t-p} = \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}$$

then implies that (X_t) has the spectral density

$$f_X(\lambda) = |A(e^{-i2\pi\lambda})|^2 f_Y(\lambda) = |B(e^{-i2\pi\lambda})|^2 f_\varepsilon(\lambda).$$

Since the roots of $A(z) = 0$ are assumed to be outside of the unit circle, we have $|A(e^{-i2\pi\lambda})| \neq 0$ and, thus, the assertion of Theorem 5.3.1 follows. \square

The preceding result with $a_1 = \cdots = a_p = 0$ implies that an MA(q)-process has the spectral density

$$f_Y(\lambda) = \sigma^2 |B(e^{-i2\pi\lambda})|^2.$$

With $b_1 = \cdots = b_q = 0$, Theorem 5.3.1 implies that a stationary AR(p)-process, which satisfies the stationarity condition (2.4), has the spectral density

$$f_Y(\lambda) = \sigma^2 \frac{1}{|A(e^{-i2\pi\lambda})|^2}.$$

Example 5.3.2. The stationary ARMA(1,1)-process

$$Y_t = aY_{t-1} + \varepsilon_t + b\varepsilon_{t-1}$$

with $|a| < 1$ has the spectral density

$$f_Y(\lambda) = \sigma^2 \frac{1 + 2b \cos(2\pi\lambda) + b^2}{1 - 2a \cos(2\pi\lambda) + a^2}.$$

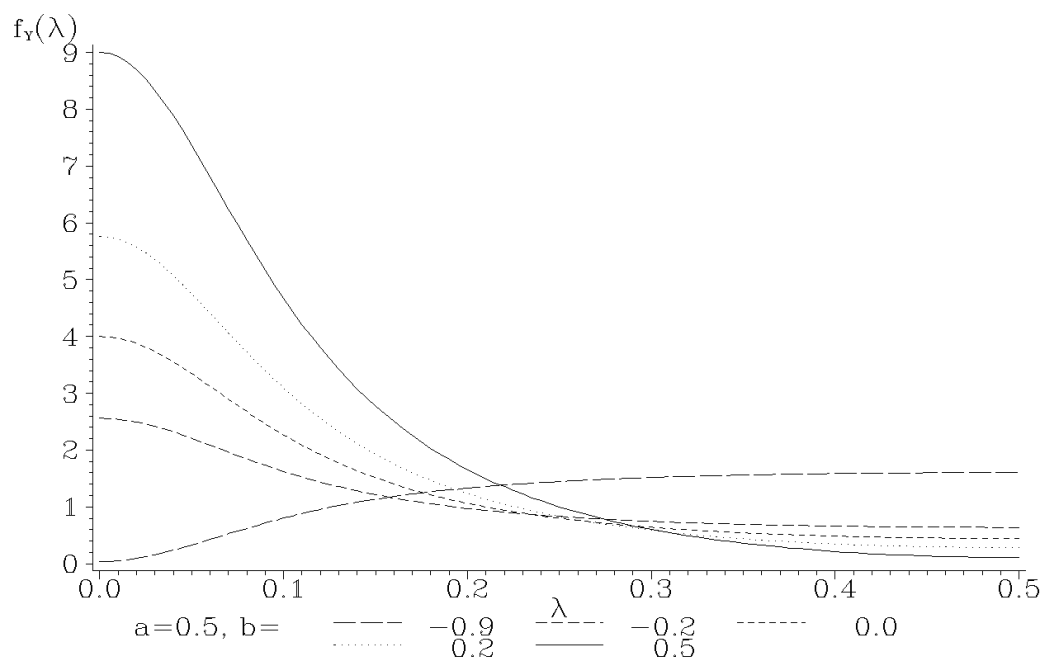
The MA(1)-process, in which case $a = 0$, has, consequently the spectral density

$$f_Y(\lambda) = \sigma^2(1 + 2b \cos(2\pi\lambda) + b^2),$$

and the stationary AR(1)-process with $|a| < 1$, for which $b = 0$, has the spectral density

$$f_Y(\lambda) = \sigma^2 \frac{1}{1 - 2a \cos(2\pi\lambda) + a^2}.$$

The following figures display spectral densities of ARMA(1,1)-processes for various a and b with $\sigma^2 = 1$.



Plot 5.3.1: Spectral densities of ARMA(1,1)-processes $Y_t = aY_{t-1} + \varepsilon_t + b\varepsilon_{t-1}$ with fixed a and various b ; $\sigma^2 = 1$.

```

1  /* armall_sd.sas */
2  TITLE1 'Spectral densities of ARMA(1,1)-processes';
3
4  /* Compute spectral densities of ARMA(1,1)-processes */
5  DATA data1;
6  a=.5;
7  DO b=-.9, -.2, 0, .2, .5;
8    DO lambda=0 TO .5 BY .005;
9      f=(1+2*b*COS(2*CONSTANT('PI')*lambda)+b*b)/(1-2*a*COS(2*CONSTANT
      ↳('PI')*lambda)+a*a);
10     OUTPUT;
11   END;
12 END;
13
14 /* Graphical options */
15 AXIS1 LABEL=('f' H=1 'Y' H=2 F=CGREEK '(l)');
16 AXIS2 LABEL=(F=CGREEK 'l');
17 SYMBOL1 V=NONE C=GREEN I=JOIN L=4;
18 SYMBOL2 V=NONE C=GREEN I=JOIN L=3;
19 SYMBOL3 V=NONE C=GREEN I=JOIN L=2;
20 SYMBOL4 V=NONE C=GREEN I=JOIN L=33;
21 SYMBOL5 V=NONE C=GREEN I=JOIN L=1;
22 LEGEND1 LABEL=('a=0.5, b=');
23
24 /* Plot spectral densities of ARMA(1,1)-processes */

```

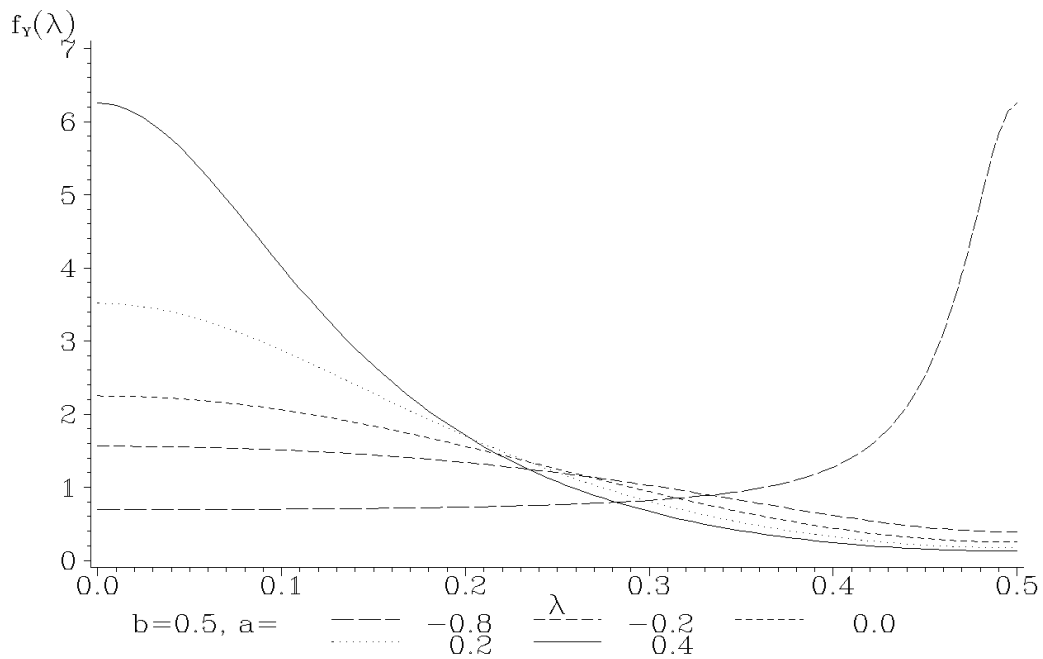
```

25 PROC Gplot DATA=data1;
26   PLOT f*lambda=b / VAXIS=AXIS1 HAXIS=AXIS2 LEGEND=LEGEND1;
27 RUN; QUIT;

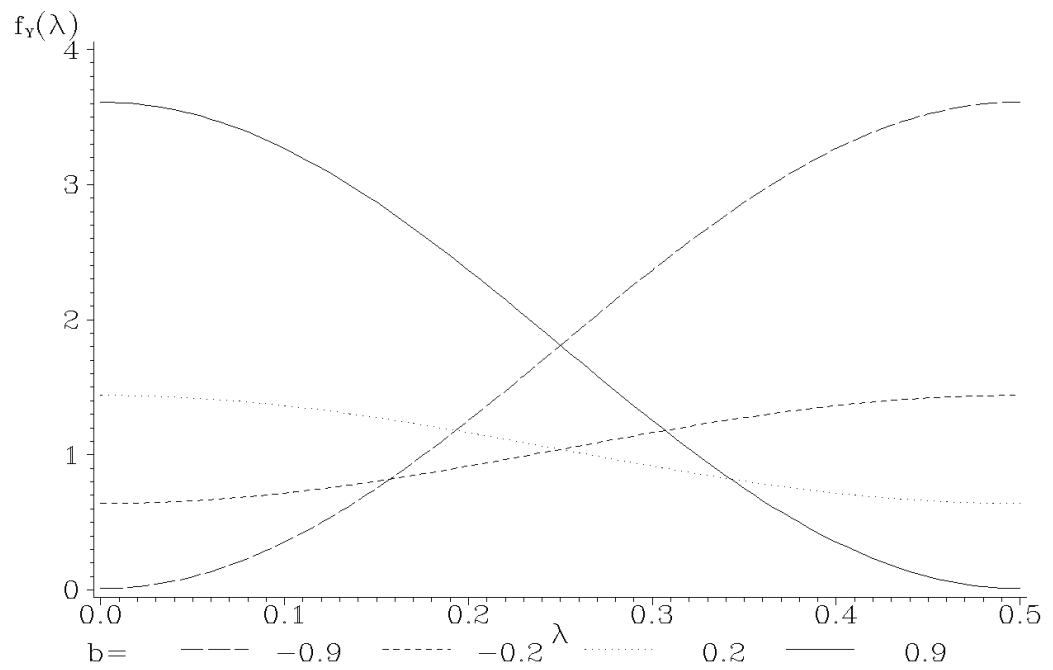
```

Like in section 5.2 (Linear Filters and Frequencies) the programs here just generate graphics. In the `DATA` step some loops over the varying parameter and over `lambda` are used to calculate the values of the spectral densities of the

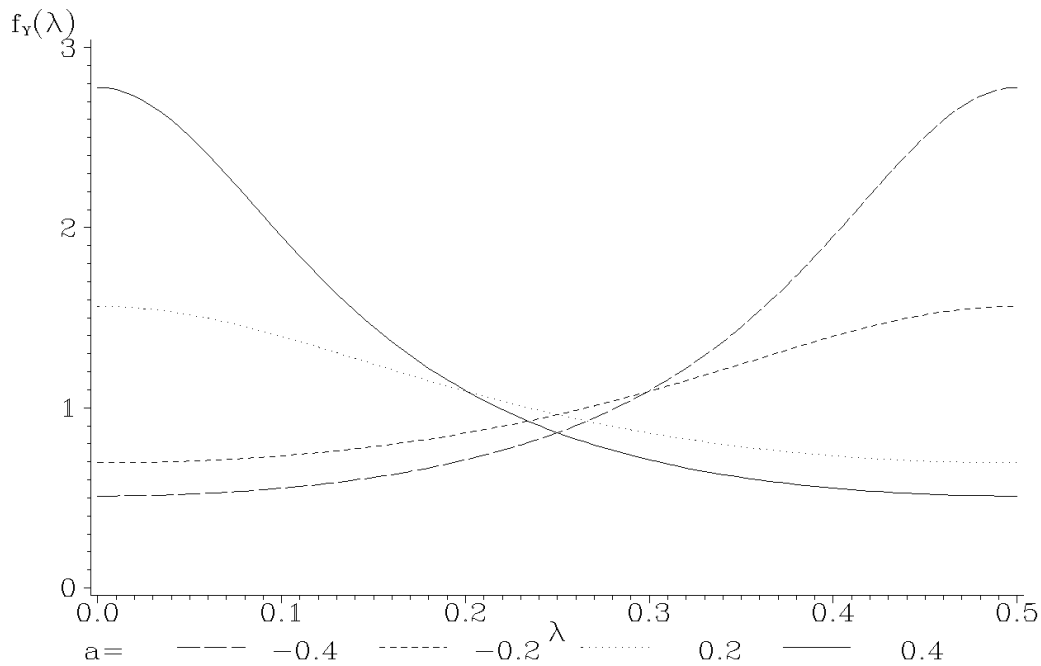
corresponding processes. Here `SYMBOL` statements are necessary and a `LABEL` statement to distinguish the different curves generated by `PROC Gplot`.



Plot 5.3.2: Spectral densities of ARMA(1, 1)-processes with parameter b fixed and various a . Corresponding file: `arma11_sd2.sas`.



Plot 5.3.2b: Spectral densities of MA(1)-processes $Y_t = \varepsilon_t + b\varepsilon_{t-1}$ for various b . Corresponding file: ma1_sd.sas.



Plot 5.3.2c: Spectral densities of AR(1)-processes $Y_t = aY_{t-1} + \varepsilon_t$ for various a . Corresponding file: ar1_sd.sas.

Exercises

5.1. Formulate and prove Theorem 5.1.1 for Hermitian functions K and complex-valued stationary processes. Hint for the sufficiency part: Let K_1 be the real part and K_2 be the imaginary part of K . Consider the real-valued $2n \times 2n$ -matrices

$$M^{(n)} = \frac{1}{2} \begin{pmatrix} K_1^{(n)} & K_2^{(n)} \\ -K_2^{(n)} & K_1^{(n)} \end{pmatrix}, \quad K_l^{(n)} = (K_l(r-s))_{1 \leq r, s \leq n}, \quad l = 1, 2.$$

Then $M^{(n)}$ is a positive semidefinite matrix (check that $z^T K \bar{z} = (x, y)^T M^{(n)}(x, y)$, $z = x + iy$, $x, y \in \mathbb{R}^n$). Proceed as in the proof of Theorem 5.1.1: Let $(V_1, \dots, V_n, W_1, \dots, W_n)$ be a $2n$ -dimensional normal distributed random vector with mean vector zero and covariance matrix $M^{(n)}$ and define for $n \in \mathbb{N}$ the family of finite dimensional

distributions

$$\begin{aligned} F_{t+1, \dots, t+n}(v_1, w_1, \dots, v_n, w_n) \\ := P\{V_1 \leq v_1, W_1 \leq w_1, \dots, V_n \leq v_n, W_n \leq w_n\}, \end{aligned}$$

$t \in \mathbb{Z}$. By Kolmogorov's theorem there exists a bivariate Gaussian process $(V_t, W_t)_{t \in \mathbb{Z}}$ with mean vector zero and covariances

$$\begin{aligned} E(V_{t+h}V_t) &= E(W_{t+h}W_t) = \frac{1}{2}K_1(h) \\ E(V_{t+h}W_t) &= -E(W_{t+h}V_t) = \frac{1}{2}K_2(h). \end{aligned}$$

Conclude by showing that the complex-valued process $Y_t := V_t - iW_t$, $t \in \mathbb{Z}$, has the autocovariance function K .

5.2. Suppose that \mathbf{A} is a real positive semidefinite $n \times n$ -matrix i.e., $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$ for $\mathbf{x} \in \mathbb{R}^n$. Show that \mathbf{A} is also positive semidefinite for complex numbers i.e., $\mathbf{z}^T \mathbf{A} \bar{\mathbf{z}} \geq 0$ for $\mathbf{z} \in \mathbb{C}^n$.

5.3. Use (5.3) to show that for $0 < a < 0.5$

$$\gamma(h) = \begin{cases} \frac{\sin(2\pi ah)}{2\pi h}, & h \in \mathbb{Z} \setminus \{0\} \\ a, & h = 0 \end{cases}$$

is the autocovariance function of a stationary process. Compute its spectral density.

5.4. Compute the autocovariance function of a stationary process with spectral density

$$f(\lambda) = \frac{0.5 - |\lambda - 0.5|}{0.5^2}, \quad 0 \leq \lambda \leq 1.$$

5.5. Suppose that F and G are measure generating functions defined on some interval $[a, b]$ with $F(a) = G(a) = 0$ and

$$\int_{[a,b]} \psi(x) F(dx) = \int_{[a,b]} \psi(x) G(dx)$$

for every continuous function $\psi : [a, b] \rightarrow \mathbb{R}$. Show that $F=G$. Hint: Approximate the indicator function $1_{[a,t]}(x)$, $x \in [a, b]$, by continuous functions.

5.6. Generate a white noise process and plot its periodogram.

5.7. A real valued stationary process $(Y_t)_{t \in \mathbb{Z}}$ is supposed to have the spectral density $f(\lambda) = a + b\lambda$, $\lambda \in [0, 0.5]$. Which conditions must be satisfied by a and b ? Compute the autocovariance function of $(Y_t)_{t \in \mathbb{Z}}$.

5.8. (Césaro convergence) Show that $\lim_{N \rightarrow \infty} \sum_{t=1}^N a_t = S$ implies $\lim_{N \rightarrow \infty} \sum_{t=1}^{N-1} (1 - t/N) a_t = S$.

Hint: $\sum_{t=1}^{N-1} (1 - t/N) a_t = (1/N) \sum_{s=1}^{N-1} \sum_{t=1}^s a_t$.

5.9. Suppose that $(Y_t)_{t \in \mathbb{Z}}$ and $(Z_t)_{t \in \mathbb{Z}}$ are stationary processes such that Y_r and Z_s are uncorrelated for arbitrary $r, s \in \mathbb{Z}$. Denote by F_Y and F_Z the pertaining spectral distribution functions and put $X_t := Y_t + Z_t$, $t \in \mathbb{Z}$. Show that the process (X_t) is also stationary and compute its spectral distribution function.

5.10. Let (Y_t) be a real valued stationary process with spectral distribution function F . Show that for any function $g : [-0.5, 0.5] \rightarrow \mathbb{C}$ with $\int_0^1 |g(\lambda - 0.5)|^2 dF(\lambda) < \infty$

$$\int_0^1 g(\lambda - 0.5) dF(\lambda) = \int_0^1 g(0.5 - \lambda) dF(\lambda).$$

In particular we have

$$F(0.5 + \lambda) - F(0.5^-) = F(0.5) - F((0.5 - \lambda)^-).$$

Hint: Verify the equality first for $g(x) = \exp(i2\pi hx)$, $h \in \mathbb{Z}$, and then use the fact that, on compact sets, the trigonometric polynomials are uniformly dense in the space of continuous functions, which in turn form a dense subset in the space of square integrable functions. Finally, consider the function $g(x) = 1_{[0, \xi]}(x)$, $0 \leq \xi \leq 0.5$ (cf. the hint in Exercise 5.5).

5.11. Let (X_t) and (Y_t) be stationary processes with mean zero and absolute summable covariance functions. If their spectral densities f_X and f_Y satisfy $f_X(\lambda) \leq f_Y(\lambda)$ for $0 \leq \lambda \leq 1$, show that

- (i) $\Gamma_{n,Y} - \Gamma_{n,X}$ is a positive semidefinite matrix, where $\Gamma_{n,X}$ and $\Gamma_{n,Y}$ are the covariance matrices of $(X_1, \dots, X_n)^T$ and $(Y_1, \dots, Y_n)^T$ respectively, and
- (ii) $\text{Var}(a^T(X_1, \dots, X_n)) \leq \text{Var}(a^T(Y_1, \dots, Y_n))$ for all vectors $a = (a_1, \dots, a_n)^T \in \mathbb{R}^n$.

5.12. Compute the gain function of the filter

$$a_u = \begin{cases} 1/4, & u \in \{-1, 1\} \\ 1/2, & u = 0 \\ 0 & \text{elsewhere.} \end{cases}$$

5.13. The simple moving average

$$a_u = \begin{cases} 1/(2q+1), & u \leq |q| \\ 0 & \text{elsewhere} \end{cases}$$

has the gain function

$$g_a(\lambda) = \begin{cases} 1, & \lambda = 0 \\ \left(\frac{\sin((2q+1)\pi\lambda)}{(2q+1)\sin(\pi\lambda)} \right)^2, & \lambda \in (0, 0.5]. \end{cases}$$

Is this filter for large q a low pass filter? Plot its power transfer functions for $q = 5/10/20$. Hint: Exercise 4.2.

5.14. Compute the gain function of the exponential smoothing filter

$$a_u = \begin{cases} \alpha(1-\alpha)^u, & u \geq 0 \\ 0, & u < 0, \end{cases}$$

where $0 < \alpha < 1$. Plot this function for various α . What is the effect of $\alpha \rightarrow 0$ or $\alpha \rightarrow 1$?

5.15. Let $(X_t)_{t \in \mathbb{Z}}$ be a stationary process, let $(a_u)_{u \in \mathbb{Z}}$ be an absolutely summable filter and put $Y_t := \sum_{u \in \mathbb{Z}} a_u X_{t-u}$, $t \in \mathbb{Z}$. If $(b_w)_{w \in \mathbb{Z}}$ is another absolutely summable filter, then the process $Z_t = \sum_{w \in \mathbb{Z}} b_w Y_{t-w}$ has the spectral distribution function

$$F_Z(\lambda) = \int_0^\lambda |F_a(\mu)|^2 |F_b(\mu)|^2 dF_X(\mu)$$

(cf. Exercise 4.11 (iii)).

5.16. Show that the function

$$D(a_r, \dots, a_s) = \int_0^{0.5} |f(\lambda) - f_a(\lambda)|^2 d\lambda$$

with $f_a(\lambda) = \sum_{u=r}^s a_u e^{-i2\pi\lambda u}$, $a_u \in \mathbb{R}$, is minimized for

$$a_u := 2 \operatorname{Re} \left(\int_0^{0.5} f(\lambda) e^{i2\pi\lambda u} d\lambda \right), \quad u = r, \dots, s.$$

Hint: Put $f(\lambda) = f_1(\lambda) + if_2(\lambda)$ and differentiate with respect to a_u .

5.17. Compute in analogy to Example 5.2.5 the transfer functions of the least squares high pass and band pass filter. Plot these functions. Is Gibbs' phenomenon observable?

5.18. An AR(2)-process

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \varepsilon_t$$

satisfying the stationarity condition (2.4) (cf. Exercise 2.25) has the spectral density

$$f_Y(\lambda) = \frac{\sigma^2}{1 + a_1^2 + a_2^2 + 2(a_1 a_2 - a_1) \cos(2\pi\lambda) - 2a_2 \cos(4\pi\lambda)}.$$

Plot this function for various choices of a_1, a_2 .

5.19. Show that (5.2) is a necessary condition for K to be the autocovariance function of a stationary process.

Statistical Analysis in the Frequency Domain

We will deal in this chapter with the problem of testing for a white noise and the estimation of the spectral density. The empirical counterpart of the spectral density, the periodogram, will be basic for both problems, though it will turn out that it is not a consistent estimate. Consistency requires extra smoothing of the periodogram via a linear filter.

6.1 Testing for a White Noise

Our initial step in a statistical analysis of a time series in the frequency domain is to test, whether the data are generated by a white noise $(\varepsilon_t)_{t \in \mathbb{Z}}$. We start with the model

$$Y_t = \mu + A \cos(2\pi\lambda t) + B \sin(2\pi\lambda t) + \varepsilon_t,$$

where we assume that the ε_t are independent and normal distributed with mean zero and variance σ^2 . We will test the nullhypothesis

$$A = B = 0$$

against the alternative

$$A \neq 0 \quad \text{or} \quad B \neq 0,$$

where the frequency λ , the variance $\sigma^2 > 0$ and the intercept $\mu \in \mathbb{R}$ are unknown. Since the periodogram is used for the detection of highly intensive frequencies inherent in the data, it seems plausible to apply it to the preceding testing problem as well. Note that $(Y_t)_{t \in \mathbb{Z}}$ is a stationary process only under the nullhypothesis $A = B = 0$.

The Distribution of the Periodogram

In the following we will derive the tests by Fisher and Bartlett–Kolmogorov–Smirnov for the above testing problem. In a preparatory step we compute the distribution of the periodogram.

Lemma 6.1.1. *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent and identically normal distributed random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Denote by $\bar{\varepsilon} := \frac{1}{n} \sum_{t=1}^n \varepsilon_t$ the sample mean of $\varepsilon_1, \dots, \varepsilon_n$ and by*

$$C_\varepsilon\left(\frac{k}{n}\right) = \frac{1}{n} \sum_{t=1}^n (\varepsilon_t - \bar{\varepsilon}) \cos\left(2\pi \frac{k}{n} t\right),$$

$$S_\varepsilon\left(\frac{k}{n}\right) = \frac{1}{n} \sum_{t=1}^n (\varepsilon_t - \bar{\varepsilon}) \sin\left(2\pi \frac{k}{n} t\right)$$

the cross covariances with Fourier frequencies k/n , $1 \leq k \leq [(n-1)/2]$, cf. (4.1). Then the $2[(n-1)/2]$ random variables

$$C_\varepsilon(k/n), S_\varepsilon(k/n), \quad 1 \leq k \leq [(n-1)/2],$$

are independent and identically $N(0, \sigma^2/(2n))$ -distributed.

Proof. Note that with $m := [(n-1)/2]$ we have

$$\begin{aligned} \mathbf{v} &:= (C_\varepsilon(1/n), S_\varepsilon(1/n), \dots, C_\varepsilon(m/n), S_\varepsilon(m/n))^T \\ &= \mathbf{A}(\varepsilon_t - \bar{\varepsilon})_{1 \leq t \leq n} \\ &= \mathbf{A}(\mathbf{I}_n - n^{-1} \mathbf{E}_n)(\varepsilon_t)_{1 \leq t \leq n}, \end{aligned}$$

where the $2m \times n$ -matrix \mathbf{A} is given by

$$\mathbf{A} := \frac{1}{n} \begin{pmatrix} \cos\left(2\pi \frac{1}{n}\right) & \cos\left(2\pi \frac{1}{n} 2\right) & \dots & \cos\left(2\pi \frac{1}{n} n\right) \\ \sin\left(2\pi \frac{1}{n}\right) & \sin\left(2\pi \frac{1}{n} 2\right) & \dots & \sin\left(2\pi \frac{1}{n} n\right) \\ \vdots & \vdots & & \vdots \\ \cos\left(2\pi \frac{m}{n}\right) & \cos\left(2\pi \frac{m}{n} 2\right) & \dots & \cos\left(2\pi \frac{m}{n} n\right) \\ \sin\left(2\pi \frac{m}{n}\right) & \sin\left(2\pi \frac{m}{n} 2\right) & \dots & \sin\left(2\pi \frac{m}{n} n\right) \end{pmatrix}.$$

\mathbf{I}_n is the $n \times n$ -unity matrix and \mathbf{E}_n is the $n \times n$ -matrix with each entry being 1. The vector \mathbf{v} is, therefore, normal distributed with mean vector zero and covariance matrix

$$\begin{aligned} & \sigma^2 \mathbf{A}(\mathbf{I}_n - n^{-1} \mathbf{E}_n)(\mathbf{I}_n - n^{-1} \mathbf{E}_n)^T \mathbf{A}^T \\ &= \sigma^2 \mathbf{A}(\mathbf{I}_n - n^{-1} \mathbf{E}_n) \mathbf{A}^T \\ &= \sigma^2 \mathbf{A} \mathbf{A}^T \\ &= \frac{\sigma^2}{2n} \mathbf{I}_{2m}, \end{aligned}$$

which is a consequence of (4.5) and the orthogonality properties from Lemma 4.1.2; see e.g. Falk et al. (2002, Definition 2.1.2). \square

Corollary 6.1.2. *Let $\varepsilon_1, \dots, \varepsilon_n$ be as in the preceding lemma and let*

$$I_\varepsilon(k/n) = n \left\{ C_\varepsilon^2(k/n) + S_\varepsilon^2(k/n) \right\}$$

be the pertaining periodogram, evaluated at the Fourier frequencies k/n , $1 \leq k \leq [(n-1)/2]$, $k \in \mathbb{N}$. The random variables $I_\varepsilon(k/n)/\sigma^2$ are independent and identically standard exponential distributed i.e.,

$$P\{I_\varepsilon(k/n)/\sigma^2 \leq x\} = \begin{cases} 1 - \exp(-x), & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Proof. Lemma 6.1.1 implies that

$$\sqrt{\frac{2n}{\sigma^2}} C_\varepsilon(k/n), \sqrt{\frac{2n}{\sigma^2}} S_\varepsilon(k/n)$$

are independent standard normal random variables and, thus,

$$\frac{2I_\varepsilon(k/n)}{\sigma^2} = \left(\sqrt{\frac{2n}{\sigma^2}} C_\varepsilon(k/n) \right)^2 + \left(\sqrt{\frac{2n}{\sigma^2}} S_\varepsilon(k/n) \right)^2$$

is χ^2 -distributed with two degrees of freedom. Since this distribution has the distribution function $1 - \exp(-x/2)$, $x \geq 0$, the assertion follows; see e.g. Falk et al. (2002, Theorem 2.1.7). \square

Denote by $U_{1:m} \leq U_{2:m} \leq \cdots \leq U_{m:m}$ the ordered values pertaining to independent and uniformly on $(0, 1)$ distributed random variables U_1, \dots, U_m . It is a well known result in the theory of order statistics that the distribution of the vector $(U_{j:m})_{1 \leq j \leq m}$ coincides with that of $((Z_1 + \cdots + Z_j)/(Z_1 + \cdots + Z_{m+1}))_{1 \leq j \leq m}$, where Z_1, \dots, Z_{m+1} are independent and identically exponential distributed random variables; see, for example, Reiss (1989, Theorem 1.6.7). The following result, which will be basic for our further considerations, is, therefore, an immediate consequence of Corollary 6.1.2; see also Exercise 6.3. By $=_D$ we denote equality in distribution.

Theorem 6.1.3. *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent $N(\mu, \sigma^2)$ -distributed random variables and denote by*

$$S_j := \frac{\sum_{k=1}^j I_\varepsilon(k/n)}{\sum_{k=1}^m I_\varepsilon(k/n)}, \quad j = 1, \dots, m := [(n-1)/2],$$

the cumulated periodogram. Note that $S_m = 1$. Then we have

$$(S_1, \dots, S_{m-1}) =_D (U_{1:m-1}, \dots, U_{m-1:m-1}).$$

The vector (S_1, \dots, S_{m-1}) has, therefore, the Lebesgue-density

$$f(s_1, \dots, s_{m-1}) = \begin{cases} (m-1)!, & \text{if } 0 < s_1 < \cdots < s_{m-1} < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

The following consequence of the preceding result is obvious.

Corollary 6.1.4. *The empirical distribution function of S_1, \dots, S_{m-1} is distributed like that of U_1, \dots, U_{m-1} , i.e.,*

$$\hat{F}_{m-1}(x) := \frac{1}{m-1} \sum_{j=1}^{m-1} 1_{(0,x]}(S_j) =_D \frac{1}{m-1} \sum_{j=1}^{m-1} 1_{(0,x]}(U_j), \quad x \in [0, 1].$$

Corollary 6.1.5. *Put $S_0 := 0$ and*

$$M_m := \max_{1 \leq j \leq m} (S_j - S_{j-1}) = \frac{\max_{1 \leq j \leq m} I_\varepsilon(j/n)}{\sum_{k=1}^m I_\varepsilon(k/n)}.$$

The maximum spacing M_m has the distribution function

$$G_m(x) := P\{M_m \leq x\} = \sum_{j=0}^m (-1)^j \binom{m}{j} (\max\{0, 1-jx\})^{m-1}, \quad x > 0.$$

Proof. Put

$$V_j := S_j - S_{j-1}, \quad j = 1, \dots, m.$$

By Theorem 6.1.3 the vector (V_1, \dots, V_m) is distributed like the length of the m consecutive intervals into which $[0, 1]$ is partitioned by the $m - 1$ random points U_1, \dots, U_{m-1} :

$$(V_1, \dots, V_m) =_D (U_{1:m-1}, U_{2:m-1} - U_{1:m-1}, \dots, 1 - U_{m-1:m-1}).$$

The probability that M_m is less than or equal to x equals the probability that all spacings V_j are less than or equal to x , and this is provided by the *covering theorem* as stated in Feller (1971, Theorem 3 in Section I.9). \square

Fisher's Test

The preceding results suggest to test the hypothesis $Y_t = \varepsilon_t$ with ε_t independent and $N(\mu, \sigma^2)$ -distributed, by testing for the uniform distribution on $[0, 1]$. Precisely, we will reject this hypothesis if *Fisher's* κ -statistic

$$\kappa_m := \frac{\max_{1 \leq j \leq m} I(j/n)}{(1/m) \sum_{k=1}^m I(k/n)} = m M_m$$

is significantly large, i.e., if one of the values $I(j/n)$ is significantly larger than the average over all. The hypothesis is, therefore, rejected at error level α if

$$\kappa_m > c_\alpha \quad \text{with} \quad 1 - G_m\left(\frac{c_\alpha}{m}\right) = \alpha.$$

This is Fisher's test for hidden periodicities. Common values are $\alpha = 0.01$ and $\alpha = 0.05$. Table 6.1.1, taken from Fuller (1995), lists several critical values c_α .

Note that these quantiles can be approximated by the corresponding quantiles of a Gumbel distribution if m is large (Exercise 6.12).

m	$c_{0.05}$	$c_{0.01}$	m	$c_{0.05}$	$c_{0.01}$
10	4.450	5.358	150	7.832	9.372
15	5.019	6.103	200	8.147	9.707
20	5.408	6.594	250	8.389	9.960
25	5.701	6.955	300	8.584	10.164
30	5.935	7.237	350	8.748	10.334
40	6.295	7.663	400	8.889	10.480
50	6.567	7.977	500	9.123	10.721
60	6.785	8.225	600	9.313	10.916
70	6.967	8.428	700	9.473	11.079
80	7.122	8.601	800	9.612	11.220
90	7.258	8.750	900	9.733	11.344
100	7.378	8.882	1000	9.842	11.454

Table 6.1.1: Critical values c_α of Fisher's test for hidden periodicities.

The Bartlett–Kolmogorov–Smirnov Test

Denote again by S_j the cumulated periodogram as in Theorem 6.1.3. If actually $Y_t = \varepsilon_t$ with ε_t independent and identically $N(\mu, \sigma^2)$ -distributed, then we know from Corollary 6.1.4 that the empirical distribution function \hat{F}_{m-1} of S_1, \dots, S_{m-1} behaves stochastically exactly like that of $m-1$ independent and uniformly on $(0, 1)$ distributed random variables. Therefore, with the Kolmogorov–Smirnov statistic

$$\Delta_{m-1} := \sup_{x \in [0,1]} |\hat{F}_{m-1}(x) - x|$$

we can measure the maximum difference between the empirical distribution function and the theoretical one $F(x) = x$, $x \in [0, 1]$. The following rule is quite common. For $m > 30$, the hypothesis $Y_t = \varepsilon_t$ with ε_t being independent and $N(\mu, \sigma^2)$ -distributed is rejected if $\Delta_{m-1} > c_\alpha / \sqrt{m-1}$, where $c_{0.05} = 1.36$ and $c_{0.01} = 1.63$ are the critical values for the levels $\alpha = 0.05$ and $\alpha = 0.01$.

This Bartlett-Kolmogorov-Smirnov test can also be carried out visually by plotting for $x \in [0, 1]$ the sample distribution function $\hat{F}_{m-1}(x)$ and the band

$$y = x \pm \frac{c_\alpha}{\sqrt{m-1}}.$$

The hypothesis $Y_t = \varepsilon_t$ is rejected if $\hat{F}_{m-1}(x)$ is for some $x \in [0, 1]$ outside of this *confidence band*.

Example 6.1.6. (Airline Data). We want to test, whether the variance stabilized, trend eliminated and seasonally adjusted Airline Data from Example 1.3.1 were generated from a white noise $(\varepsilon_t)_{t \in \mathbb{Z}}$, where ε_t are independent and identically normal distributed. The Fisher test statistic has the value $\kappa_m = 6.573$ and does, therefore, not reject the hypothesis at the levels $\alpha = 0.05$ and $\alpha = 0.01$, where $m = 65$. The Bartlett–Kolmogorov–Smirnov test, however, rejects this hypothesis at both levels, since $\Delta_{64} = 0.2319 > 1.36/\sqrt{64} = 0.17$ and also $\Delta_{64} > 1.63/\sqrt{64} = 0.20375$.

```

SPECTRA Procedure

----- Test for White Noise for variable DLOGNUM -----

Fisher's Kappa: M*MAX(P(*))/SUM(P(*))
Parameters:      M           =           65
                  MAX(P(*)) =         0.028
                  SUM(P(*)) =         0.275
Test Statistic: Kappa      =         6.5730

Bartlett's Kolmogorov-Smirnov Statistic:
    Maximum absolute difference of the standardized
    partial sums of the periodogram and the CDF of a
    uniform(0,1) random variable.

Test Statistic =         0.2319

```

Listing 6.1.1: Fisher's κ and the Bartlett-Kolmogorov-Smirnov test with $m = 65$ for testing a white noise generation of the adjusted Airline Data.

```

1  /* airline_whitenoise.sas */
2  TITLE1 'Tests for white noise';
3  TITLE2 'for the trend und seasonal';
4  TITLE3 'adjusted Airline Data';
5
6  /* Read in the data and compute log-transformation as well as seasonal
   ↪ and trend adjusted data */
7  DATA datal;
8  INFILE 'c:\data\airline.txt';

```

```

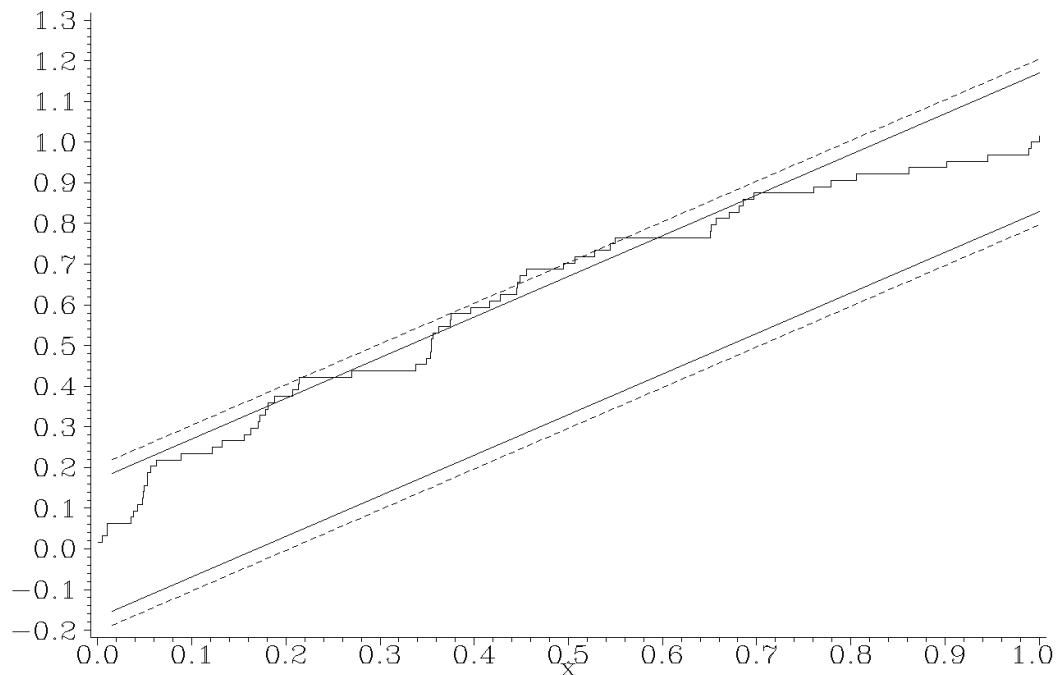
9  INPUT num @@;
10  dlognum=DIF12(DIF(LOG(num)) );
11
12  /* Compute periodogram and test for white noise */
13  PROC SPECTRA DATA=data1 P WHITETEST OUT=data2;
14      VAR dlognum;
15  RUN; QUIT;

```

In the DATA step the raw data of the airline passengers are read into the variable `num`. A log-transformation, building the first order difference for trend elimination and the 12th order difference for elimination of a seasonal component lead to the variable `dlognum`, which is supposed to be generated by a stationary process. Then PROC SPECTRA is applied to this variable, whereby the options P and OUT=data2 generate a data set containing the periodogram

data. The option WHITETEST causes SAS to carry out the two tests for a white noise, Fisher's test and the Bartlett-Kolmogorov-Smirnov test. SAS only provides the values of the test statistics but no decision. One has to compare these values with the critical values from Table 6.1.1 (Critical values for Fisher's Test in the script) and the approximative ones $c_\alpha/\sqrt{m-1}$.

The following figure visualizes the rejection at both levels by the Bartlett-Kolmogorov-Smirnov test.



Plot 6.1.2: Bartlett-Kolmogorov-Smirnov test with $m = 65$ testing for a white noise generation of the adjusted Airline Data. Solid line/broken line = confidence bands for $\hat{F}_{m-1}(x)$, $x \in [0, 1]$, at levels $\alpha = 0.05/0.01$.

```

1  /* airline_whitenoise_plot.sas */
2  TITLE1 'Visualisation of the test for white noise';
3  TITLE2 'for the trend und seasonal adjusted';
4  TITLE3 'Airline Data';
5  /* Note that this program needs data2 generated by the previous
   ↪ program (airline_whitenoise.sas) */
6
7  /* Calculate the sum of the periodogram */
8  PROC MEANS DATA=data2 (FIRSTOBS=2) NOPRINT;
9      VAR P_01;
10     OUTPUT OUT=data3 SUM=psum;
11
12  /* Compute empirical distribution function of cumulated periodogram
   ↪ and its confidence bands */
13  DATA data4;
14     SET data2 (FIRSTOBS=2);
15     IF _N_=1 THEN SET data3;
16     RETAIN s 0;
17     s=s+P_01/psum;
18     fm=_N/(_FREQ_-1);
19     yu_01=fm+1.63/SQRT(_FREQ_-1);
20     yl_01=fm-1.63/SQRT(_FREQ_-1);
21     yu_05=fm+1.36/SQRT(_FREQ_-1);

```

```

22   yl_05=fm-1.36/SQRT(_FREQ_-1);
23
24   /* Graphical options */
25   SYMBOL1 V=NONE I=STEPJ C=GREEN;
26   SYMBOL2 V=NONE I=JOIN C=RED L=2;
27   SYMBOL3 V=NONE I=JOIN C=RED L=1;
28   AXIS1 LABEL=('x') ORDER=(.0 TO 1.0 BY .1);
29   AXIS2 LABEL=NONE;
30
31   /* Plot empirical distribution function of cumulated periodogram with
      ↪ its confidence bands */
32   PROC GPLOT DATA=data4;
33     PLOT fm*s=1 yu_01*fm=2 yl_01*fm=2 yu_05*fm=3 yl_05*fm=3 / OVERLAY
      ↪ HAXIS=AXIS1 VAXIS=AXIS2;
34   RUN; QUIT;

```

This program uses the data set `data2` created by Program 6.1.1 (`airline_whitenoise.sas`), where the first observation belonging to the frequency 0 is dropped. `PROC MEANS` calculates the sum (keyword `SUM`) of the SAS periodogram variable `P_0` and stores it in the variable `psum` of the data set `data3`. The `NOPRINT` option suppresses the printing of the output.

The next `DATA` step combines every observation of `data2` with this sum by means of the `IF` statement. Furthermore a variable `s` is initialized with the value 0 by the `RETAIN` statement and then the portion of each periodogram value from the sum is cumulated. The variable

`fm` contains the values of the empirical distribution function calculated by means of the automatically generated variable `_N_` containing the number of observation and the variable `_FREQ_`, which was created by `PROC MEANS` and contains the number `m`. The values of the upper and lower band are stored in the `y` variables.

The last part of this program contains `SYMBOL` and `AXIS` statements and `PROC GPLOT` to visualize the Bartlett-Kolmogorov-Smirnov statistic. The empirical distribution of the cumulated periodogram is represented as a step function due to the `I=STEPJ` option in the `SYMBOL1` statement.

6.2 Estimating Spectral Densities

We suppose in the following that $(Y_t)_{t \in \mathbb{Z}}$ is a stationary real valued process with mean μ and absolutely summable autocovariance function γ . According to Corollary 5.1.5, the process (Y_t) has the continuous spectral density

$$f(\lambda) = \sum_{h \in \mathbb{Z}} \gamma(h) e^{-i2\pi\lambda h}.$$

In the preceding section we computed the distribution of the empirical counterpart of a spectral density, the periodogram, in the particular case, when (Y_t) is a Gaussian white noise. In this section we will inves-

tigate the limit behavior of the periodogram for arbitrary independent random variables (Y_t) .

Asymptotic Properties of the Periodogram

In order to establish asymptotic properties of the periodogram, its following modification is quite useful. For Fourier frequencies k/n , $0 \leq k \leq [n/2]$ we put

$$\begin{aligned} I_n(k/n) &= \frac{1}{n} \left| \sum_{t=1}^n Y_t e^{-i2\pi(k/n)t} \right|^2 \\ &= \frac{1}{n} \left\{ \left(\sum_{t=1}^n Y_t \cos \left(2\pi \frac{k}{n} t \right) \right)^2 + \left(\sum_{t=1}^n Y_t \sin \left(2\pi \frac{k}{n} t \right) \right)^2 \right\}. \end{aligned} \quad (6.1)$$

Up to $k = 0$, this coincides by (4.5) with the definition of the periodogram as given in (4.6) on page 143. From Theorem 4.2.3 we obtain the representation

$$I_n(k/n) = \begin{cases} n\bar{Y}_n^2, & k = 0 \\ \sum_{|h| < n} c(h) e^{-i2\pi(k/n)h}, & k = 1, \dots, [n/2] \end{cases} \quad (6.2)$$

with $\bar{Y}_n := \frac{1}{n} \sum_{t=1}^n Y_t$ and the sample autocovariance function

$$c(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (Y_t - \bar{Y}_n)(Y_{t+|h|} - \bar{Y}_n).$$

By representation (6.1), the proof of Theorem 4.2.3 and the equations (4.5), the value $I_n(k/n)$ does not change for $k \neq 0$ if we replace the sample mean \bar{Y}_n in $c(h)$ by the theoretical mean μ . This leads to the

equivalent representation of the periodogram for $k \neq 0$

$$\begin{aligned}
 I_n(k/n) &= \sum_{|h| < n} \frac{1}{n} \left(\sum_{t=1}^{n-|h|} (Y_t - \mu)(Y_{t+|h|} - \mu) \right) e^{-i2\pi(k/n)h} \\
 &= \frac{1}{n} \sum_{t=1}^n (Y_t - \mu)^2 + \\
 &\quad 2 \sum_{h=1}^{n-1} \frac{1}{n} \left(\sum_{t=1}^{n-|h|} (Y_t - \mu)(Y_{t+|h|} - \mu) \right) \cos \left(2\pi \frac{k}{n} h \right).
 \end{aligned} \tag{6.3}$$

We define now the periodogram for $\lambda \in [0, 0.5]$ as a piecewise constant function

$$I_n(\lambda) = I_n(k/n) \quad \text{if} \quad \frac{k}{n} - \frac{1}{2n} < \lambda \leq \frac{k}{n} + \frac{1}{2n}. \tag{6.4}$$

The following result shows that the periodogram $I_n(\lambda)$ is for $\lambda \neq 0$ an asymptotically unbiased estimator of the spectral density $f(\lambda)$.

Theorem 6.2.1. *Let $(Y_t)_{t \in \mathbb{Z}}$ be a stationary process with absolutely summable autocovariance function γ . Then we have with $\mu = E(Y_t)$*

$$\begin{aligned}
 E(I_n(0)) - n\mu^2 &\xrightarrow{n \rightarrow \infty} f(0), \\
 E(I_n(\lambda)) &\xrightarrow{n \rightarrow \infty} f(\lambda), \quad \lambda \neq 0.
 \end{aligned}$$

If $\mu = 0$, then the convergence $E(I_n(\lambda)) \xrightarrow{n \rightarrow \infty} f(\lambda)$ holds uniformly on $[0, 0.5]$.

Proof. By representation (6.2) and the Césaro convergence result (Exercise 5.8) we have

$$\begin{aligned}
 E(I_n(0)) - n\mu^2 &= \frac{1}{n} \left(\sum_{t=1}^n \sum_{s=1}^n E(Y_t Y_s) \right) - n\mu^2 \\
 &= \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n \text{Cov}(Y_t, Y_s) \\
 &= \sum_{|h| < n} \left(1 - \frac{|h|}{n} \right) \gamma(h) \xrightarrow{n \rightarrow \infty} \sum_{h \in \mathbb{Z}} \gamma(h) = f(0).
 \end{aligned}$$

Define now for $\lambda \in [0, 0.5]$ the auxiliary function

$$g_n(\lambda) := \frac{k}{n}, \quad \text{if } \frac{k}{n} - \frac{1}{2n} < \lambda \leq \frac{k}{n} + \frac{1}{2n}, \quad k \in \mathbb{Z}. \quad (6.5)$$

Then we obviously have

$$I_n(\lambda) = I_n(g_n(\lambda)). \quad (6.6)$$

Choose now $\lambda \in (0, 0.5]$. Since $g_n(\lambda) \rightarrow \lambda$ as $n \rightarrow \infty$, it follows that $g_n(\lambda) > 0$ for n large enough. By (6.3) and (6.6) we obtain for such n

$$\begin{aligned} E(I_n(\lambda)) &= \sum_{|h| < n} \frac{1}{n} \sum_{t=1}^{n-|h|} E\left((Y_t - \mu)(Y_{t+|h|} - \mu)\right) e^{-i2\pi g_n(\lambda)h} \\ &= \sum_{|h| < n} \left(1 - \frac{|h|}{n}\right) \gamma(h) e^{-i2\pi g_n(\lambda)h}. \end{aligned}$$

Since $\sum_{h \in \mathbb{Z}} |\gamma(h)| < \infty$, the series $\sum_{|h| < n} \gamma(h) \exp(-i2\pi \lambda h)$ converges to $f(\lambda)$ uniformly for $0 \leq \lambda \leq 0.5$. Kronecker's Lemma (Exercise 6.8) implies moreover

$$\left| \sum_{|h| < n} \frac{|h|}{n} \gamma(h) e^{-i2\pi \lambda h} \right| \leq \sum_{|h| < n} \frac{|h|}{n} |\gamma(h)| \xrightarrow{n \rightarrow \infty} 0,$$

and, thus, the series

$$f_n(\lambda) := \sum_{|h| < n} \left(1 - \frac{|h|}{n}\right) \gamma(h) e^{-i2\pi \lambda h}$$

converges to $f(\lambda)$ uniformly in λ as well. From $g_n(\lambda) \xrightarrow{n \rightarrow \infty} \lambda$ and the continuity of f we obtain for $\lambda \in (0, 0.5]$

$$\begin{aligned} |E(I_n(\lambda)) - f(\lambda)| &= |f_n(g_n(\lambda)) - f(\lambda)| \\ &\leq |f_n(g_n(\lambda)) - f(g_n(\lambda))| + |f(g_n(\lambda)) - f(\lambda)| \\ &\xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Note that $|g_n(\lambda) - \lambda| \leq 1/(2n)$. The uniform convergence in case of $\mu = 0$ then follows from the uniform convergence of $g_n(\lambda)$ to λ and the uniform continuity of f on the compact interval $[0, 0.5]$. \square

In the following result we compute the asymptotic distribution of the periodogram for independent and identically distributed random variables with zero mean, which are not necessarily Gaussian ones. The Gaussian case was already established in Corollary 6.1.2.

Theorem 6.2.2. *Let Z_1, \dots, Z_n be independent and identically distributed random variables with mean $E(Z_t) = 0$ and variance $E(Z_t^2) = \sigma^2 < \infty$. Denote by $I_n(\lambda)$ the pertaining periodogram as defined in (6.6).*

- (i) *The random vector $(I_n(\lambda_1), \dots, I_n(\lambda_r))$ with $0 < \lambda_1 < \dots < \lambda_r < 0.5$ converges in distribution for $n \rightarrow \infty$ to the distribution of r independent and identically exponential distributed random variables with mean σ^2 .*
- (ii) *If $E(Z_t^4) = \eta\sigma^4 < \infty$, then we have for $k = 0, \dots, [n/2]$, $k \in \mathbb{N}$*

$$\text{Var}(I_n(k/n)) = \begin{cases} 2\sigma^4 + n^{-1}(\eta - 3)\sigma^4, & k = 0 \text{ or } k = n/2 \\ \sigma^4 + n^{-1}(\eta - 3)\sigma^4 & \text{elsewhere} \end{cases} \quad (6.7)$$

and

$$\text{Cov}(I_n(j/n), I_n(k/n)) = n^{-1}(\eta - 3)\sigma^4, \quad j \neq k. \quad (6.8)$$

For $N(0, \sigma^2)$ -distributed random variables Z_t we have $\eta = 3$ (Exercise 6.9) and, thus, $I_n(k/n)$ and $I_n(j/n)$ are for $j \neq k$ uncorrelated. Actually, we established in Corollary 6.1.2 that they are independent in this case.

Proof. Put for $\lambda \in (0, 0.5)$

$$A_n(\lambda) := A_n(g_n(\lambda)) := \sqrt{2/n} \sum_{t=1}^n Z_t \cos(2\pi g_n(\lambda)t),$$

$$B_n(\lambda) := B_n(g_n(\lambda)) := \sqrt{2/n} \sum_{t=1}^n Z_t \sin(2\pi g_n(\lambda)t),$$

with g_n defined in (6.5). Since

$$I_n(\lambda) = \frac{1}{2} \left\{ A_n^2(\lambda) + B_n^2(\lambda) \right\},$$

it suffices, by repeating the arguments in the proof of Corollary 6.1.2, to show that

$$(A_n(\lambda_1), B_n(\lambda_1), \dots, A_n(\lambda_r), B_n(\lambda_r)) \longrightarrow_D \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{2r}),$$

where \mathbf{I}_{2r} denotes the $2r \times 2r$ -unity matrix and \longrightarrow_D convergence in distribution. Since $g_n(\lambda) \xrightarrow{n \rightarrow \infty} \lambda$, we have $g_n(\lambda) > 0$ for $\lambda \in (0, 0.5)$ if n is large enough. The independence of Z_t together with the definition of g_n and the orthogonality equations in Lemma 4.1.2 imply

$$\begin{aligned} \text{Var}(A_n(\lambda)) &= \text{Var}(A_n(g_n(\lambda))) \\ &= \sigma^2 \frac{2}{n} \sum_{t=1}^n \cos^2(2\pi g_n(\lambda)t) = \sigma^2. \end{aligned}$$

For $\varepsilon > 0$ we have

$$\begin{aligned} &\frac{1}{n} \sum_{t=1}^n \mathbb{E} \left(Z_t^2 \cos^2(2\pi g_n(\lambda)t) 1_{\{|Z_t \cos(2\pi g_n(\lambda)t)| > \varepsilon \sqrt{n\sigma^2}\}} \right) \\ &\leq \frac{1}{n} \sum_{t=1}^n \mathbb{E} \left(Z_t^2 1_{\{|Z_t| > \varepsilon \sqrt{n\sigma^2}\}} \right) = \mathbb{E} \left(Z_1^2 1_{\{|Z_1| > \varepsilon \sqrt{n\sigma^2}\}} \right) \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

i.e., the triangular array $(2/n)^{1/2} Z_t \cos(2\pi g_n(\lambda)t)$, $1 \leq t \leq n$, $n \in \mathbb{N}$, satisfies the *Lindeberg condition* implying $A_n(\lambda) \longrightarrow_D N(0, \sigma^2)$; see, for example, Billingsley (1968, Theorem 7.2).

Similarly one shows that $B_n(\lambda) \longrightarrow_D N(0, \sigma^2)$ as well. Since the random vector $(A_n(\lambda_1), B_n(\lambda_1), \dots, A_n(\lambda_r), B_n(\lambda_r))$ has by (3.2) the covariance matrix $\sigma^2 \mathbf{I}_{2r}$, its asymptotic joint normality follows easily by applying the *Cramér-Wold device*, cf. Billingsley (1968, Theorem 7.7), and proceeding as before. This proves part (i) of Theorem 6.2.2. From the definition of the periodogram in (6.1) we conclude as in the proof of Theorem 4.2.3

$$I_n(k/n) = \frac{1}{n} \sum_{s=1}^n \sum_{t=1}^n Z_s Z_t e^{-i2\pi \frac{k}{n}(s-t)}$$

and, thus, we obtain

$$\begin{aligned} & \mathbb{E}(I_n(j/n)I_n(k/n)) \\ &= \frac{1}{n^2} \sum_{s=1}^n \sum_{t=1}^n \sum_{u=1}^n \sum_{v=1}^n \mathbb{E}(Z_s Z_t Z_u Z_v) e^{-i2\pi \frac{j}{n}(s-t)} e^{-i2\pi \frac{k}{n}(u-v)}. \end{aligned}$$

We have

$$\mathbb{E}(Z_s Z_t Z_u Z_v) = \begin{cases} \eta\sigma^4, & s = t = u = v \\ \sigma^4, & s = t \neq u = v, s = u \neq t = v, s = v \neq t = u \\ 0 & \text{elsewhere} \end{cases}$$

and

$$e^{-i2\pi \frac{j}{n}(s-t)} e^{-i2\pi \frac{k}{n}(u-v)} = \begin{cases} 1, & s = t, u = v \\ e^{-i2\pi((j+k)/n)s} e^{i2\pi((j+k)/n)t}, & s = u, t = v \\ e^{-i2\pi((j-k)/n)s} e^{i2\pi((j-k)/n)t}, & s = v, t = u. \end{cases}$$

This implies

$$\begin{aligned} & \mathbb{E}(I_n(j/n)I_n(k/n)) \\ &= \frac{\eta\sigma^4}{n} + \frac{\sigma^4}{n^2} \left\{ n(n-1) + \left| \sum_{t=1}^n e^{-i2\pi((j+k)/n)t} \right|^2 + \right. \\ & \quad \left. \left| \sum_{t=1}^n e^{-i2\pi((j-k)/n)t} \right|^2 - 2n \right\} \\ &= \frac{(\eta-3)\sigma^4}{n} + \sigma^4 \left\{ 1 + \frac{1}{n^2} \left| \sum_{t=1}^n e^{i2\pi((j+k)/n)t} \right|^2 + \right. \\ & \quad \left. \frac{1}{n^2} \left| \sum_{t=1}^n e^{i2\pi((j-k)/n)t} \right|^2 \right\}. \end{aligned}$$

From $\mathbb{E}(I_n(k/n)) = n^{-1} \sum_{t=1}^n \mathbb{E}(Z_t^2) = \sigma^2$ we finally obtain

$$\begin{aligned} & \text{Cov}(I_n(j/n), I_n(k/n)) \\ &= \frac{(\eta-3)\sigma^4}{n} + \frac{\sigma^4}{n^2} \left\{ \left| \sum_{t=1}^n e^{i2\pi((j+k)/n)t} \right|^2 + \left| \sum_{t=1}^n e^{i2\pi((j-k)/n)t} \right|^2 \right\}, \end{aligned}$$

from which (6.7) and (6.8) follow by using (4.5) on page 142. \square

Remark 6.2.3. Theorem 6.2.2 can be generalized to filtered processes $Y_t = \sum_{u \in \mathbb{Z}} a_u Z_{t-u}$, with $(Z_t)_{t \in \mathbb{Z}}$ as in Theorem 6.2.2. In this case one has to replace σ^2 , which equals by Example 5.1.3 the constant spectral density $f_Z(\lambda)$, in (i) by the spectral density $f_Y(\lambda_i)$, $1 \leq i \leq r$. If in addition $\sum_{u \in \mathbb{Z}} |a_u| |u|^{1/2} < \infty$, then we have in (ii) the expansions

$$\text{Var}(I_n(k/n)) = \begin{cases} 2f_Y^2(k/n) + O(n^{-1/2}), & k = 0 \text{ or } k = n/2 \\ f_Y^2(k/n) + O(n^{-1/2}) & \text{elsewhere,} \end{cases}$$

and

$$\text{Cov}(I_n(j/n), I_n(k/n)) = O(n^{-1}), \quad j \neq k,$$

where I_n is the periodogram pertaining to Y_1, \dots, Y_n . The above terms $O(n^{-1/2})$ and $O(n^{-1})$ are uniformly bounded in k and j by a constant C .

We omit the highly technical proof and refer to Brockwell and Davis (1991, Section 10.3).

Recall that the class of processes $Y_t = \sum_{u \in \mathbb{Z}} a_u Z_{t-u}$ is a fairly rich one, which contains in particular ARMA-processes, see Section 2.2 and Remark 2.1.12.

Discrete Spectral Average Estimator

The preceding results show that the periodogram is not a consistent estimator of the spectral density. The law of large numbers together with the above remark motivates, however, that consistency can be achieved for a smoothed version of the periodogram such as a simple moving average

$$\sum_{|j| \leq m} \frac{1}{2m+1} I_n\left(\frac{k+j}{n}\right),$$

which puts equal weights $1/(2m+1)$ on adjacent values. Dropping the condition of equal weights, we define a general linear smoother by the linear filter

$$\hat{f}_n\left(\frac{k}{n}\right) := \sum_{|j| \leq m} a_{jn} I_n\left(\frac{k+j}{n}\right). \quad (6.9)$$

The sequence $m = m(n)$, defining the adjacent points of k/n , has to satisfy

$$m \xrightarrow{n \rightarrow \infty} \infty \text{ and } m/n \xrightarrow{n \rightarrow \infty} 0, \quad (6.10)$$

and the weights a_{jn} have the properties

$$\begin{aligned} & \text{(i) } a_{jn} \geq 0, \\ & \text{(ii) } a_{jn} = a_{-jn}, \\ & \text{(iii) } \sum_{|j| \leq m} a_{jn} = 1, \\ & \text{(iv) } \sum_{|j| \leq m} a_{jn}^2 \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad (6.11)$$

For the simple moving average we have, for example

$$a_{jn} = \begin{cases} 1/(2m+1), & |j| \leq m \\ 0 & \text{elsewhere} \end{cases}$$

and $\sum_{|j| \leq m} a_{jn}^2 = 1/(2m+1) \xrightarrow{n \rightarrow \infty} 0$. For $\lambda \in [0, 0.5]$ we put $I_n(0.5 + \lambda) := I_n(0.5 - \lambda)$, which defines the periodogram also on $[0.5, 1]$. If $(k+j)/n$ is outside of the interval $[0, 1]$, then $I_n((k+j)/n)$ is understood as the periodic extension of I_n with period 1. This also applies to the spectral density f . The estimator

$$\hat{f}_n(\lambda) := \hat{f}_n(g_n(\lambda)),$$

with g_n as defined in (6.5), is called the *discrete spectral average estimator*. The following result states its consistency for linear processes.

Theorem 6.2.4. *Let $Y_t = \sum_{u \in \mathbb{Z}} b_u Z_{t-u}$, $t \in \mathbb{Z}$, where Z_t are iid with $E(Z_t) = 0$, $E(Z_t^4) < \infty$ and $\sum_{u \in \mathbb{Z}} |b_u| |u|^{1/2} < \infty$. Then we have for $0 \leq \mu, \lambda \leq 0.5$*

$$\begin{aligned} & \text{(i) } \lim_{n \rightarrow \infty} E(\hat{f}_n(\lambda)) = f(\lambda), \\ & \text{(ii) } \lim_{n \rightarrow \infty} \frac{\text{Cov}(\hat{f}_n(\lambda), \hat{f}_n(\mu))}{\left(\sum_{|j| \leq m} a_{jn}^2\right)} = \begin{cases} 2f^2(\lambda), & \lambda = \mu = 0 \text{ or } 0.5 \\ f^2(\lambda), & 0 < \lambda = \mu < 0.5 \\ 0, & \lambda \neq \mu. \end{cases} \end{aligned}$$

Condition (iv) in (6.11) on the weights together with (ii) in the preceding result entails that $\text{Var}(\hat{f}_n(\lambda)) \xrightarrow{n \rightarrow \infty} 0$ for any $\lambda \in [0, 0.5]$. Together with (i) we, therefore, obtain that the mean squared error of $\hat{f}_n(\lambda)$ vanishes asymptotically:

$$\begin{aligned} \text{MSE}(\hat{f}_n(\lambda)) &= \text{E} \left\{ (\hat{f}_n(\lambda) - f(\lambda))^2 \right\} \\ &= \text{Var}(\hat{f}_n(\lambda)) + \text{Bias}^2(\hat{f}_n(\lambda)) \\ &\xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Proof. By the definition of the spectral density estimator in (6.9) we have

$$\begin{aligned} |\text{E}(\hat{f}_n(\lambda)) - f(\lambda)| &= \left| \sum_{|j| \leq m} a_{jn} \left\{ \text{E} (I_n(g_n(\lambda) + j/n)) - f(\lambda) \right\} \right| \\ &= \left| \sum_{|j| \leq m} a_{jn} \left\{ \text{E} (I_n(g_n(\lambda) + j/n)) - f(g_n(\lambda) + j/n) \right. \right. \\ &\quad \left. \left. + f(g_n(\lambda) + j/n) - f(\lambda) \right\} \right|, \end{aligned}$$

where (6.10) together with the uniform convergence of $g_n(\lambda)$ to λ implies

$$\max_{|j| \leq m} |g_n(\lambda) + j/n - \lambda| \xrightarrow{n \rightarrow \infty} 0.$$

Choose $\varepsilon > 0$. The spectral density f of the process (Y_t) is continuous (Exercise 6.16), and hence we have

$$\max_{|j| \leq m} |f(g_n(\lambda) + j/n) - f(\lambda)| < \varepsilon/2$$

if n is sufficiently large. From Theorem 6.2.1 we know that in the case $\text{E}(Y_t) = 0$

$$\max_{|j| \leq m} |\text{E}(I_n(g_n(\lambda) + j/n)) - f(g_n(\lambda) + j/n)| < \varepsilon/2$$

if n is large. Condition (iii) in (6.11) together with the triangular inequality implies $|\text{E}(\hat{f}_n(\lambda)) - f(\lambda)| < \varepsilon$ for large n . This implies part (i).

From the definition of \hat{f}_n we obtain

$$\begin{aligned} & \text{Cov}(\hat{f}_n(\lambda), \hat{f}_n(\mu)) \\ &= \sum_{|j| \leq m} \sum_{|k| \leq m} a_{jn} a_{kn} \text{Cov} \left(I_n(g_n(\lambda) + j/n), I_n(g_n(\mu) + k/n) \right). \end{aligned}$$

If $\lambda \neq \mu$ and n sufficiently large we have $g_n(\lambda) + j/n \neq g_n(\mu) + k/n$ for arbitrary $|j|, |k| \leq m$. According to Remark 6.2.3 there exists a universal constant $C_1 > 0$ such that

$$\begin{aligned} |\text{Cov}(\hat{f}_n(\lambda), \hat{f}_n(\mu))| &= \left| \sum_{|j| \leq m} \sum_{|k| \leq m} a_{jn} a_{kn} O(n^{-1}) \right| \\ &\leq C_1 n^{-1} \left(\sum_{|j| \leq m} a_{jn} \right)^2 \\ &\leq C_1 \frac{2m+1}{n} \sum_{|j| \leq m} a_{jn}^2, \end{aligned}$$

where the final inequality is an application of the Cauchy–Schwarz inequality. Since $m/n \xrightarrow{n \rightarrow \infty} 0$, we have established (ii) in the case $\lambda \neq \mu$. Suppose now $0 < \lambda = \mu < 0.5$. Utilizing again Remark 6.2.3 we have

$$\begin{aligned} \text{Var}(\hat{f}_n(\lambda)) &= \sum_{|j| \leq m} a_{jn}^2 \left\{ f^2(g_n(\lambda) + j/n) + O(n^{-1/2}) \right\} \\ &\quad + \sum_{|j| \leq m} \sum_{|k| \leq m} a_{jn} a_{kn} O(n^{-1}) + o(n^{-1}) \\ &=: S_1(\lambda) + S_2(\lambda) + o(n^{-1}). \end{aligned}$$

Repeating the arguments in the proof of part (i) one shows that

$$S_1(\lambda) = \left(\sum_{|j| \leq m} a_{jn}^2 \right) f^2(\lambda) + o \left(\sum_{|j| \leq m} a_{jn}^2 \right).$$

Furthermore, with a suitable constant $C_2 > 0$, we have

$$|S_2(\lambda)| \leq C_2 \frac{1}{n} \left(\sum_{|j| \leq m} a_{jn} \right)^2 \leq C_2 \frac{2m+1}{n} \sum_{|j| \leq m} a_{jn}^2.$$

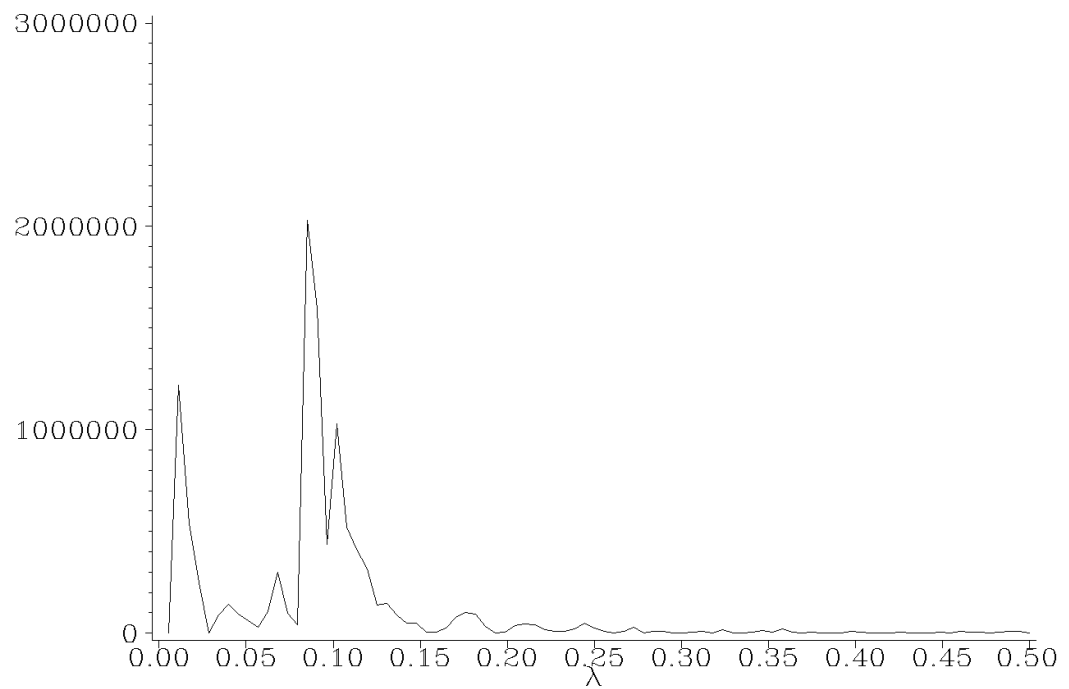
Thus we established the assertion of part (ii) also in the case $0 < \lambda = \mu < 0.5$. The remaining cases $\lambda = \mu = 0$ and $\lambda = \mu = 0.5$ are shown in a similar way (Exercise 6.17). \square

The preceding result requires zero mean variables Y_t . This might, however, be too restrictive in practice. Due to (4.5), the periodograms of $(Y_t)_{1 \leq t \leq n}$, $(Y_t - \mu)_{1 \leq t \leq n}$ and $(Y_t - \bar{Y})_{1 \leq t \leq n}$ coincide at Fourier frequencies different from zero. At frequency $\lambda = 0$, however, they will differ in general. To estimate $f(0)$ consistently also in the case $\mu = E(Y_t) \neq 0$, one puts

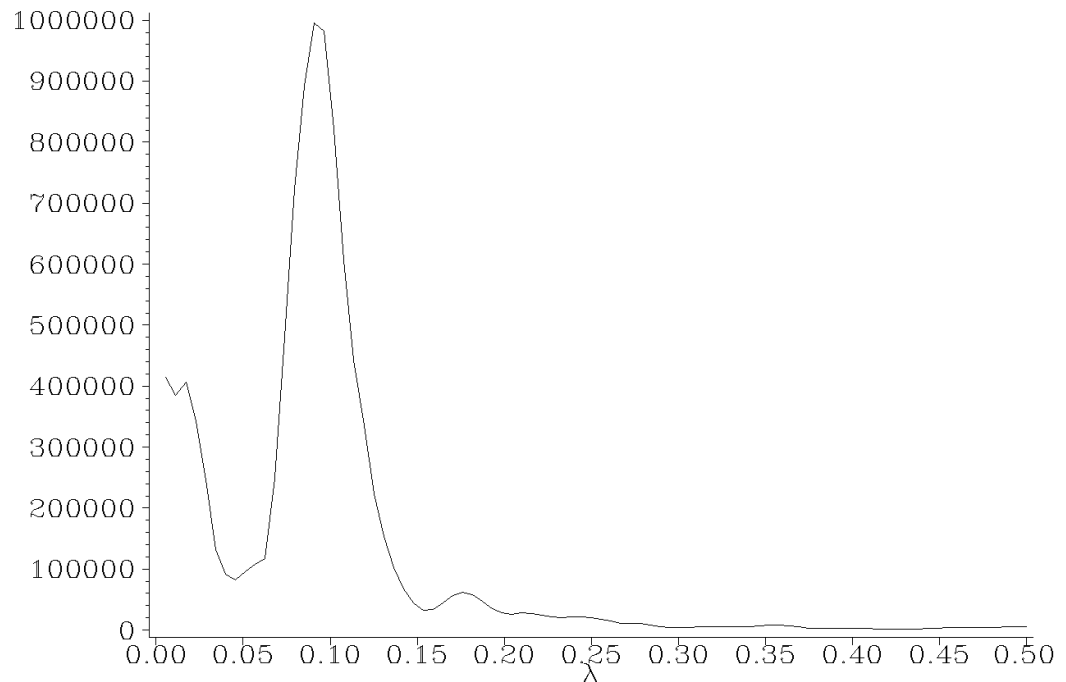
$$\hat{f}_n(0) := a_{0n}I_n(1/n) + 2 \sum_{j=1}^m a_{jn}I_n((1+j)/n). \quad (6.12)$$

Each time the value $I_n(0)$ occurs in the moving average (6.9), it is replaced by $\hat{f}_n(0)$. Since the resulting estimator of the spectral density involves only Fourier frequencies different from zero, we can assume without loss of generality that the underlying variables Y_t have zero mean.

Example 6.2.5. (Sunspot Data). We want to estimate the spectral density underlying the Sunspot Data. These data, also known as the Wolf or Wölfer (a student of Wolf) Data, are the yearly sunspot numbers between 1749 and 1924. For a discussion of these data and further literature we refer to Wei and Reilly (1989), Example 6.2. Plot 6.2.1 shows the pertaining periodogram and Plot 6.2.1b displays the discrete spectral average estimator with weights $a_{0n} = a_{1n} = a_{2n} = 3/21$, $a_{3n} = 2/21$ and $a_{4n} = 1/21$, $n = 176$. These weights pertain to a simple moving average of length 3 of a simple moving average of length 7. The smoothed version joins the two peaks close to the frequency $\lambda = 0.1$ visible in the periodogram. The observation that a periodogram has the tendency to split a peak is known as the *leakage phenomenon*.



Plot 6.2.1: Periodogram for Sunspot Data.



Plot 6.2.1b: Discrete spectral average estimate for Sunspot Data.

```
1 /* sunspot_dsae.sas */
```



```

2 TITLE1 'Periodogram and spectral density estimate';
3 TITLE2 'Woelfer Sunspot Data';
4
5 /* Read in the data */
6 DATA data1;
7     INFILE 'c:\data\sunspot.txt';
8     INPUT num @@;
9
10 /* Computation of periodogram and estimation of spectral density */
11 PROC SPECTRA DATA=data1 P S OUT=data2;
12     VAR num;
13     WEIGHTS 1 2 3 3 3 2 1;
14
15 /* Adjusting different periodogram definitions */
16 DATA data3;
17     SET data2(FIRSTOBS=2);
18     lambda=FREQ/(2*CONSTANT('PI'));
19     p=P_01/2;
20     s=S_01/2*4*CONSTANT('PI');
21
22 /* Graphical options */
23 SYMBOL1 I=JOIN C=RED V=NONE L=1;
24 AXIS1 LABEL=(F=CGREEK 'l') ORDER=(0 TO .5 BY .05);
25 AXIS2 LABEL=NONE;
26
27 /* Plot periodogram and estimated spectral density */
28 PROC GPLOT DATA=data3;
29     PLOT p*lambda / HAXIS=AXIS1 VAXIS=AXIS2;
30     PLOT s*lambda / HAXIS=AXIS1 VAXIS=AXIS2;
31 RUN; QUIT;

```

In the **DATA** step the data of the sunspots are read into the variable `num`.

Then **PROC SPECTRA** is applied to this variable, whereby the options **P** (see Program 6.1.1, `airline.whitenoise.sas`) and **S** generate a data set stored in `data2` containing the periodogram data and the estimation of the spectral density which SAS computes with

the weights given in the **WEIGHTS** statement. Note that SAS automatically normalizes these weights.

In following **DATA** step the slightly different definition of the periodogram by SAS is being adjusted to the definition used here (see Program 4.2.1, `star.periodogram.sas`). Both plots are then printed with **PROC GPLOT**.

A mathematically convenient way to generate weights a_{jn} , which satisfy the conditions (6.11), is the use of a kernel function. Let $K : [-1, 1] \rightarrow [0, \infty)$ be a symmetric function i.e., $K(-x) = K(x)$, $x \in [-1, 1]$, which satisfies $\int_{-1}^1 K^2(x) dx < \infty$. Let now $m = m(n) \xrightarrow{n \rightarrow \infty} \infty$

be an arbitrary sequence of integers with $m/n \xrightarrow{n \rightarrow \infty} 0$ and put

$$a_{jn} := \frac{K(j/m)}{\sum_{i=-m}^m K(i/m)}, \quad -m \leq j \leq m. \quad (6.13)$$

These weights satisfy the conditions (6.11) (Exercise 6.18).

Take for example $K(x) := 1 - |x|$, $-1 \leq x \leq 1$. Then we obtain

$$a_{jn} = \frac{m - |j|}{m^2}, \quad -m \leq j \leq m.$$

Example 6.2.6. (i) The *truncated kernel* is defined by

$$K_T(x) = \begin{cases} 1, & |x| \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

(ii) The *Bartlett* or *triangular kernel* is given by

$$K_B(x) := \begin{cases} 1 - |x|, & |x| \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

(iii) The *Blackman–Tukey kernel* (1949) is defined by

$$K_{BT}(x) = \begin{cases} 1 - 2a + 2a \cos(x), & |x| \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

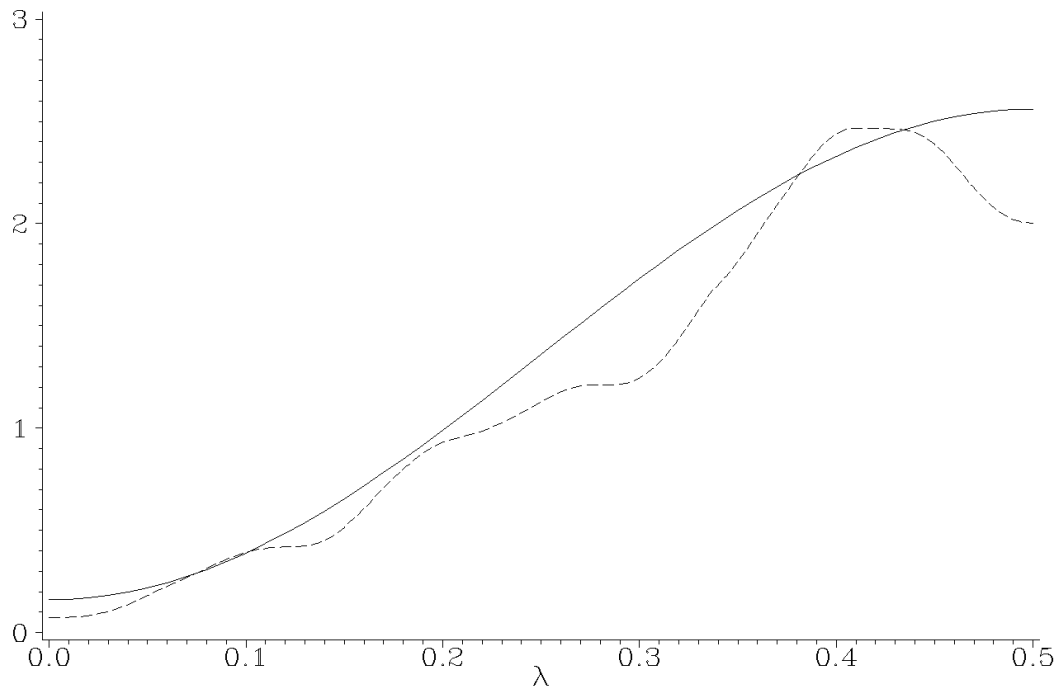
where $0 < a \leq 1/4$. The particular choice $a = 1/4$ yields the *Tukey–Hanning kernel*.

(iv) The *Parzen kernel* (1957) is given by

$$K_P(x) := \begin{cases} 1 - 6|x|^2 + 6|x|^3, & |x| < 1/2 \\ 2(1 - |x|)^3, & 1/2 \leq |x| \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

We refer to Andrews (1991) for a discussion of these kernels.

Example 6.2.7. We consider realizations of the MA(1)-process $Y_t = \varepsilon_t - 0.6\varepsilon_{t-1}$ with ε_t independent and standard normal for $t = 1, \dots, n = 160$. Example 5.3.2 implies that the process (Y_t) has the spectral density $f(\lambda) = 1 - 1.2 \cos(2\pi\lambda) + 0.36$. We estimate $f(\lambda)$ by means of the Tukey–Hanning kernel.



Plot 6.2.2: Discrete spectral average estimator (broken line) with Blackman–Tukey kernel with parameters $m = 10$, $a = 1/4$ and underlying spectral density $f(\lambda) = 1 - 1.2 \cos(2\pi\lambda) + 0.36$ (solid line).

```

1  /* mal_blackman_tukey.sas */
2  TITLE1 'Spectral density and Blackman-Tukey estimator';
3  TITLE2 'of MA(1)-process';
4
5  /* Generate MA(1)-process */
6  DATA data1;
7      DO t=0 TO 160;
8          e=RANNOR(1);
9          y=e-.6*LAG(e);
10         OUTPUT;
11     END;
12
13  /* Estimation of spectral density */
14  PROC SPECTRA DATA=data1(FIRSTOBS=2) S OUT=data2;
15      VAR y;
16      WEIGHTS TUKEY 10 0;
17  RUN;
18
19  /* Adjusting different definitions */
20  DATA data3;

```

```

21  SET data2;
22  lambda=FREQ/(2*CONSTANT('PI'));
23  s=S_01/2*4*CONSTANT('PI');
24
25  /* Compute underlying spectral density */
26  DATA data4;
27  DO l=0 TO .5 BY .01;
28      f=1-1.2*COS(2*CONSTANT('PI')*l)+.36;
29      OUTPUT;
30  END;
31
32  /* Merge the data sets */
33  DATA data5;
34      MERGE data3(KEEP=s lambda) data4;
35
36  /* Graphical options */
37  AXIS1 LABEL=NONE;
38  AXIS2 LABEL=(F=CGREEK 'l') ORDER=(0 TO .5 BY .1);
39  SYMBOL1 I=JOIN C=BLUE V=NONE L=1;
40  SYMBOL2 I=JOIN C=RED V=NONE L=3;
41
42  /* Plot underlying and estimated spectral density */
43  PROC GPLOT DATA=data5;
44      PLOT f*1=1 s*lambda=2 / OVERLAY VAXIS=AXIS1 HAXIS=AXIS2;
45  RUN; QUIT;

```

In the first DATA step the realizations of an MA(1)-process with the given parameters are created. Thereby the function RANNOR, which generates standard normally distributed data, and LAG, which accesses the value of e of the preceding loop, are used.

As in Program 6.2.1 (sunspot.dsae.sas) PROC SPECTRA computes the estimator of the spectral density (after dropping the first observation) by the option S and stores them in data2. The weights used here come from the Tukey–Hanning kernel with a specified bandwidth of

$m = 10$. The second number after the TUKEY option can be used to refine the choice of the bandwidth. Since this is not needed here it is set to 0.

The next DATA step adjusts the different definitions of the spectral density used here and by SAS (see Program 4.2.1, star.periodogram.sas). The following DATA step generates the values of the underlying spectral density. These are merged with the values of the estimated spectral density and then displayed by PROC GPLOT.

Confidence Intervals for Spectral Densities

The random variables $I_n((k+j)/n)/f((k+j)/n)$, $0 < k+j < n/2$, will by Remark 6.2.3 for large n approximately behave like independent and standard exponential distributed random variables X_j . This sug-

gests that the distribution of the discrete spectral average estimator

$$\hat{f}(k/n) = \sum_{|j| \leq m} a_{jn} I_n\left(\frac{k+j}{n}\right)$$

can be approximated by that of the weighted sum $\sum_{|j| \leq m} a_{jn} X_j f((k+j)/n)$. Tukey (1949) showed that the distribution of this weighted sum can in turn be approximated by that of cY with a suitably chosen $c > 0$, where Y follows a *gamma distribution* with parameters $p := \nu/2 > 0$ and $b = 1/2$ i.e.,

$$P\{Y \leq t\} = \frac{b^p}{\Gamma(p)} \int_0^t x^{p-1} \exp(-bx) dx, \quad t \geq 0,$$

where $\Gamma(p) := \int_0^\infty x^{p-1} \exp(-x) dx$ denotes the gamma function. The parameters ν and c are determined by the method of moments as follows: ν and c are chosen such that cY has mean $f(k/n)$ and its variance equals the leading term of the variance expansion of $\hat{f}(k/n)$ in Theorem 6.2.4 (Exercise 6.21):

$$\begin{aligned} E(cY) &= c\nu = f(k/n), \\ \text{Var}(cY) &= 2c^2\nu = f^2(k/n) \sum_{|j| \leq m} a_{jn}^2. \end{aligned}$$

The solutions are obviously

$$c = \frac{f(k/n)}{2} \sum_{|j| \leq m} a_{jn}^2$$

and

$$\nu = \frac{2}{\sum_{|j| \leq m} a_{jn}^2}.$$

Note that the gamma distribution with parameters $p = \nu/2$ and $b = 1/2$ equals the χ^2 -distribution with ν degrees of freedom if ν is an integer. The number ν is, therefore, called the *equivalent degree of freedom*. Observe that $\nu/f(k/n) = 1/c$; the random variable $\nu \hat{f}(k/n)/f(k/n) = \hat{f}(k/n)/c$ now approximately follows a $\chi^2(\nu)$ -distribution with the convention that $\chi^2(\nu)$ is the gamma distribution

with parameters $p = \nu/2$ and $b = 1/2$ if ν is not an integer. The interval

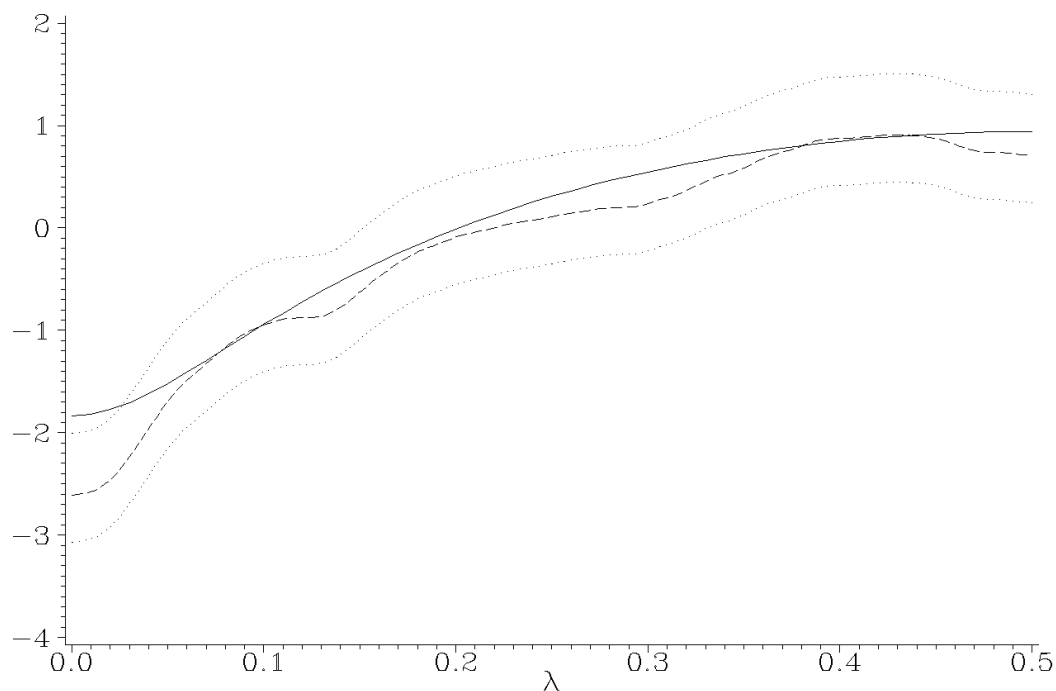
$$\left(\frac{\nu \hat{f}(k/n)}{\chi_{1-\alpha/2}^2(\nu)}, \frac{\nu \hat{f}(k/n)}{\chi_{\alpha/2}^2(\nu)} \right) \quad (6.14)$$

is a confidence interval for $f(k/n)$ of approximate level $1 - \alpha$, $\alpha \in (0, 1)$. By $\chi_q^2(\nu)$ we denote the q -quantile of the $\chi^2(\nu)$ -distribution i.e., $P\{Y \leq \chi_q^2(\nu)\} = q$, $0 < q < 1$. Taking logarithms in (6.14), we obtain the confidence interval for $\log(f(k/n))$

$$C_{\nu, \alpha}(k/n) := \left(\log(\hat{f}(k/n)) + \log(\nu) - \log(\chi_{1-\alpha/2}^2(\nu)), \right. \\ \left. \log(\hat{f}(k/n)) + \log(\nu) - \log(\chi_{\alpha/2}^2(\nu)) \right).$$

This interval has constant length $\log(\chi_{1-\alpha/2}^2(\nu)/\chi_{\alpha/2}^2(\nu))$. Note that $C_{\nu, \alpha}(k/n)$ is a level $(1 - \alpha)$ -confidence interval only for $\log(f(\lambda))$ at a fixed Fourier frequency $\lambda = k/n$, with $0 < k < [n/2]$, but not *simultaneously* for $\lambda \in (0, 0.5)$.

Example 6.2.8. In continuation of Example 6.2.7 we want to estimate the spectral density $f(\lambda) = 1 - 1.2 \cos(2\pi\lambda) + 0.36$ of the MA(1)-process $Y_t = \varepsilon_t - 0.6\varepsilon_{t-1}$ using the discrete spectral average estimator $\hat{f}_n(\lambda)$ with the weights 1, 3, 6, 9, 12, 15, 18, 20, 21, 21, 21, 20, 18, 15, 12, 9, 6, 3, 1, each divided by 231. These weights are generated by iterating simple moving averages of lengths 3, 7 and 11. Plot 6.2.3 displays the logarithms of the estimates, of the true spectral density and the pertaining confidence intervals.



Plot 6.2.3: Logarithms of discrete spectral average estimates (broken line), of spectral density $f(\lambda) = 1 - 1.2 \cos(2\pi\lambda) + 0.36$ (solid line) of MA(1)-process $Y_t = \varepsilon_t - 0.6\varepsilon_{t-1}$, $t = 1, \dots, n = 160$, and confidence intervals of level $1 - \alpha = 0.95$ for $\log(f(k/n))$.

```

1  /* mal_logdsae.sas */
2  TITLE1 'Logarithms of spectral density,';
3  TITLE2 'of their estimates and confidence intervals';
4  TITLE3 'of MA(1)-process';
5
6  /* Generate MA(1)-process */
7  DATA data1;
8      DO t=0 TO 160;
9          e=RANNOR(1);
10         y=e-.6*LAG(e);
11         OUTPUT;
12     END;
13
14  /* Estimation of spectral density */
15  PROC SPECTRA DATA=data1(FIRSTOBS=2) S OUT=data2;
16      VAR y; WEIGHTS 1 3 6 9 12 15 18 20 21 21 21 20 18 15 12 9 6 3 1;
17      RUN;
18
19  /* Adjusting different definitions and computation of confidence bands
    ↪ */
20  DATA data3; SET data2;
21      lambda=FREQ/(2*CONSTANT('PI'));
22      log_s_01=LOG(S_01/2*4*CONSTANT('PI'));

```

```

23  nu=2/(3763/53361);
24  c1=log_s_01+LOG(nu)-LOG(CINV(.975,nu));
25  c2=log_s_01+LOG(nu)-LOG(CINV(.025,nu));
26
27  /* Compute underlying spectral density */
28  DATA data4;
29      DO l=0 TO .5 BY 0.01;
30          log_f=LOG((1-1.2*COS(2*CONSTANT('PI')*l)+.36));
31          OUTPUT;
32      END;
33
34  /* Merge the data sets */
35  DATA data5;
36      MERGE data3(KEEP=log_s_01 lambda c1 c2) data4;
37
38  /* Graphical options */
39  AXIS1 LABEL=NONE;
40  AXIS2 LABEL=(F=CGREEK 'l') ORDER=(0 TO .5 BY .1);
41  SYMBOL1 I=JOIN C=BLUE V=NONE L=1;
42  SYMBOL2 I=JOIN C=RED V=NONE L=2;
43  SYMBOL3 I=JOIN C=GREEN V=NONE L=33;
44
45  /* Plot underlying and estimated spectral density */
46  PROC GPLOT DATA=data5;
47      PLOT log_f*1=1 log_s_01*lambda=2 c1*lambda=3 c2*lambda=3 / OVERLAY
         ↪ VAXIS=AXIS1 HAXIS=AXIS2;
48  RUN; QUIT;

```

This program starts identically to Program 6.2.2 (ma1_blackman_tukey.sas) with the generation of an MA(1)-process and of the computation the spectral density estimator. Only this time the weights are directly given to SAS.

In the next DATA step the usual adjustment of the frequencies is done. This is followed by the computation of ν according to its definition. The logarithm of the confidence intervals is calcu-

lated with the help of the function CINV which returns quantiles of a χ^2 -distribution with ν degrees of freedom.

The rest of the program which displays the logarithm of the estimated spectral density, of the underlying density and of the confidence intervals is analogous to Program 6.2.2 (ma1_blackman_tukey.sas).

Exercises

6.1. For independent random variables X, Y having continuous distribution functions it follows that $P\{X = Y\} = 0$. Hint: Fubini's theorem.

6.2. Let X_1, \dots, X_n be iid random variables with values in \mathbb{R} and distribution function F . Denote by $X_{1:n} \leq \dots \leq X_{n:n}$ the pertaining

order statistics. Then we have

$$P\{X_{k:n} \leq t\} = \sum_{j=k}^n \binom{n}{j} F(t)^j (1 - F(t))^{n-j}, \quad t \in \mathbb{R}.$$

The maximum $X_{n:n}$ has in particular the distribution function F^n , and the minimum $X_{1:n}$ has distribution function $1 - (1 - F)^n$. Hint: $\{X_{k:n} \leq t\} = \left\{ \sum_{j=1}^n 1_{(-\infty, t]}(X_j) \geq k \right\}$.

6.3. Suppose in addition to the conditions in Exercise 6.2 that F has a (Lebesgue) density f . The ordered vector $(X_{1:n}, \dots, X_{n:n})$ then has a density

$$f_n(x_1, \dots, x_n) = n! \prod_{j=1}^n f(x_j), \quad x_1 < \dots < x_n,$$

and zero elsewhere. Hint: Denote by S_n the group of permutations of $\{1, \dots, n\}$ i.e., $(\tau(1), \dots, \tau(n))$ with $\tau \in S_n$ is a permutation of $(1, \dots, n)$. Put for $\tau \in S_n$ the set $B_\tau := \{X_{\tau(1)} < \dots < X_{\tau(n)}\}$. These sets are disjoint and we have $P(\sum_{\tau \in S_n} B_\tau) = 1$ since $P\{X_j = X_k\} = 0$ for $i \neq j$ (cf. Exercise 6.1).

6.4. Let X and Y be independent, standard normal distributed random variables. Show that $(X, Z)^T := (X, \rho X + \sqrt{1 - \rho^2} Y)^T$, $-1 < \rho < 1$, is normal distributed with mean vector $(0, 0)$ and covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, and that X and Z are independent if and only if they are uncorrelated (i.e., $\rho = 0$).

Suppose that X and Y are normal distributed and uncorrelated. Does this imply the independence of X and Y ? Hint: Let $X \sim N(0, 1)$ -distributed and define the random variable $Y = VX$ with V independent of X and $P\{V = -1\} = 1/2 = P\{V = 1\}$.

6.5. Generate 100 independent and standard normal random variables ε_t and plot the periodogram. Is the hypothesis that the observations were generated by a white noise rejected at level $\alpha = 0.05(0.01)$? Visualize the Bartlett-Kolmogorov-Smirnov test by plotting the empirical

distribution function of the cumulated periodograms S_j , $1 \leq j \leq 48$, together with the pertaining bands for $\alpha = 0.05$ and $\alpha = 0.01$

$$y = x \pm \frac{c_\alpha}{\sqrt{m-1}}, \quad x \in (0, 1).$$

6.6. Generate the values

$$Y_t = \cos\left(2\pi \frac{1}{6}t\right) + \varepsilon_t, \quad t = 1, \dots, 300,$$

where ε_t are independent and standard normal. Plot the data and the periodogram. Is the hypothesis $Y_t = \varepsilon_t$ rejected at level $\alpha = 0.01$?

6.7. (Share Data) Test the hypothesis that the share data were generated by independent and identically normal distributed random variables and plot the periodogram. Plot also the original data.

6.8. (Kronecker's lemma) Let $(a_j)_{j \geq 0}$ be an absolute summable complex valued filter. Show that $\lim_{n \rightarrow \infty} \sum_{j=0}^n (j/n) |a_j| = 0$.

6.9. The normal distribution $N(0, \sigma^2)$ satisfies

- (i) $\int x^{2k+1} dN(0, \sigma^2)(x) = 0, \quad k \in \mathbb{N} \cup \{0\}.$
- (ii) $\int x^{2k} dN(0, \sigma^2)(x) = 1 \cdot 3 \cdot \dots \cdot (2k-1) \sigma^{2k}, \quad k \in \mathbb{N}.$
- (iii) $\int |x|^{2k+1} dN(0, \sigma^2)(x) = \frac{2^{k+1}}{\sqrt{2\pi}} k! \sigma^{2k+1}, \quad k \in \mathbb{N} \cup \{0\}.$

6.10. Show that a $\chi^2(\nu)$ -distributed random variable satisfies $E(Y) = \nu$ and $\text{Var}(Y) = 2\nu$. Hint: Exercise 6.9.

6.11. (Slutzky's lemma) Let $X, X_n, n \in \mathbb{N}$, be random variables in \mathbb{R} with distribution functions F_X and F_{X_n} , respectively. Suppose that X_n converges in distribution to X (denoted by $X_n \rightarrow_{\mathcal{D}} X$) i.e., $F_{X_n}(t) \rightarrow F_X(t)$ for every continuity point of F_X as $n \rightarrow \infty$. Let $Y_n, n \in \mathbb{N}$, be another sequence of random variables which converges stochastically to some constant $c \in \mathbb{R}$, i.e., $\lim_{n \rightarrow \infty} P\{|Y_n - c| > \varepsilon\} = 0$ for arbitrary $\varepsilon > 0$. This implies

- (i) $X_n + Y_n \rightarrow_{\mathcal{D}} X + c.$

$$(ii) \quad X_n Y_n \rightarrow_{\mathcal{D}} cX.$$

$$(iii) \quad X_n/Y_n \rightarrow_{\mathcal{D}} X/c, \quad \text{if } c \neq 0.$$

This entails in particular that stochastic convergence implies convergence in distribution. The reverse implication is not true in general. Give an example.

6.12. Show that the distribution function F_m of Fisher's test statistic κ_m satisfies under the condition of independent and identically normal observations ε_t

$$F_m(x + \ln(m)) = P\{\kappa_m \leq x + \ln(m)\} \xrightarrow{m \rightarrow \infty} \exp(-e^{-x}) =: G(x), \quad x \in \mathbb{R}.$$

The limiting distribution G is known as the *Gumbel* distribution. Hence we have $P\{\kappa_m > x\} = 1 - F_m(x) \approx 1 - \exp(-m e^{-x})$. Hint: Exercise 6.2 and 6.11.

6.13. Which effect has an outlier on the periodogram? Check this for the simple model $(Y_t)_{t, \dots, n}$ ($t_0 \in \{1, \dots, n\}$)

$$Y_t = \begin{cases} \varepsilon_t, & t \neq t_0 \\ \varepsilon_t + c, & t = t_0, \end{cases}$$

where the ε_t are independent and identically normal $N(0, \sigma^2)$ distributed and $c \neq 0$ is an arbitrary constant. Show to this end

$$E(I_Y(k/n)) = E(I_\varepsilon(k/n)) + c^2/n$$

$$\text{Var}(I_Y(k/n)) = \text{Var}(I_\varepsilon(k/n)) + 2c^2\sigma^2/n, \quad k = 1, \dots, [(n-1)/2].$$

6.14. Suppose that U_1, \dots, U_n are uniformly distributed on $(0, 1)$ and let \hat{F}_n denote the pertaining empirical distribution function. Show that

$$\sup_{0 \leq x \leq 1} |\hat{F}_n(x) - x| = \max_{1 \leq k \leq n} \left\{ \max \left\{ U_{k:n} - \frac{(k-1)}{n}, \frac{k}{n} - U_{k:n} \right\} \right\}.$$

6.15. (Monte Carlo Simulation) For m large we have under the hypothesis

$$P\{\sqrt{m-1}\Delta_{m-1} > c_\alpha\} \approx \alpha.$$

For different values of m (> 30) generate 1000 times the test statistic $\sqrt{m-1}\Delta_{m-1}$ based on independent random variables and check, how often this statistic exceeds the critical values $c_{0.05} = 1.36$ and $c_{0.01} = 1.63$. Hint: Exercise 6.14.

6.16. In the situation of Theorem 6.2.4 show that the spectral density f of $(Y_t)_t$ is continuous.

6.17. Complete the proof of Theorem 6.2.4 (ii) for the remaining cases $\lambda = \mu = 0$ and $\lambda = \mu = 0.5$.

6.18. Verify that the weights (6.13) defined via a kernel function satisfy the conditions (6.11).

6.19. Use the IML function `ARMASIM` to simulate the process

$$Y_t = 0.3Y_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1}, \quad 1 \leq t \leq 150,$$

where ε_t are independent and standard normal. Plot the periodogram and estimates of the log spectral density together with confidence intervals. Compare the estimates with the log spectral density of $(Y_t)_{t \in \mathbb{Z}}$.

6.20. Compute the distribution of the periodogram $I_\varepsilon(1/2)$ for independent and identically normal $N(0, \sigma^2)$ -distributed random variables $\varepsilon_1, \dots, \varepsilon_n$ in case of an even sample size n .

6.21. Suppose that Y follows a gamma distribution with parameters p and b . Calculate the mean and the variance of Y .

6.22. Compute the length of the confidence interval $C_{\nu, \alpha}(k/n)$ for fixed α (preferably $\alpha = 0.05$) but for various ν . For the calculation of ν use the weights generated by the kernel $K(x) = 1 - |x|$, $-1 \leq x \leq 1$ (see equation (6.13)).

6.23. Show that for $y \in \mathbb{R}$

$$\frac{1}{n} \left| \sum_{t=0}^{n-1} e^{i2\pi yt} \right|^2 = \sum_{|t| < n} \left(1 - \frac{|t|}{n}\right) e^{i2\pi yt} = K_n(y),$$

where

$$K_n(y) = \begin{cases} n, & y \in \mathbb{Z} \\ \frac{1}{n} \left(\frac{\sin(\pi y n)}{\sin(\pi y)} \right)^2, & y \notin \mathbb{Z}, \end{cases}$$

is the Fejer kernel of order n . Verify that it has the properties

- (i) $K_n(y) \geq 0$,
- (ii) the Fejer kernel is a periodic function of period length one,
- (iii) $K_n(y) = K_n(-y)$,
- (iv) $\int_{-0.5}^{0.5} K_n(y) dy = 1$,
- (v) $\int_{-\delta}^{\delta} K_n(y) dy \xrightarrow{n \rightarrow \infty} 1, \quad \delta > 0$.

6.24. (Nile Data) Between 715 and 1284 the river Nile had its lowest annual minimum levels. These data are among the longest time series in hydrology. Can the trend removed Nile Data be considered as being generated by a white noise, or are there hidden periodicities? Estimate the spectral density in this case. Use discrete spectral estimators as well as lag window spectral density estimators. Compare with the spectral density of an AR(1)-process.

The Box–Jenkins Program: A Case Study

This chapter deals with the practical application of the Box–Jenkins Program to the Donauwoerth Data, consisting of 7300 discharge measurements from the Donau river at Donauwoerth, specified in cubic centimeter per second and taken on behalf of the Bavarian State Office For Environment between January 1st, 1985 and December 31st, 2004. For the purpose of studying, the data have been kindly made available to the University of Würzburg.

As introduced in Section 2.3, the Box–Jenkins methodology can be applied to specify adequate ARMA(p, q)-model $Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}, t \in \mathbb{Z}$ for the Donauwoerth data in order to forecast future values of the time series. In short, the original time series will be adjusted to represent a possible realization of such a model. Based on the identification methods MINIC, SCAN and ESACF, appropriate pairs of orders (p, q) are chosen in Section 7.5 and the corresponding model coefficients a_1, \dots, a_p and b_1, \dots, b_q are determined. Finally, it is demonstrated by Diagnostic Checking in Section 7.6 that the resulting model is adequate and forecasts based on this model are executed in the concluding Section 7.7.

Yet, before starting the program in Section 7.4, some theoretical preparations have to be carried out. In the first Section 7.1 we introduce the general definition of the partial autocorrelation leading to the Levinson–Durbin-Algorithm, which will be needed for Diagnostic Checking. In order to verify whether pure AR(p)- or MA(q)-models might be appropriate candidates for explaining the sample, we derive in Section 7.2 and 7.3 asymptotic normal behaviors of suitable estimators of the partial and general autocorrelations.

7.1 Partial Correlation and Levinson–Durbin Recursion

In general, the correlation between two random variables is often due to the fact that both variables are correlated with other variables. Therefore, the linear influence of these variables is removed to receive the partial correlation.

Partial Correlation

The *partial correlation* of two square integrable, real valued random variables X and Y , holding the random variables $Z_1, \dots, Z_m, m \in \mathbb{N}$, fixed, is defined as

$$\begin{aligned} & \text{Corr}(X - \hat{X}_{Z_1, \dots, Z_m}, Y - \hat{Y}_{Z_1, \dots, Z_m}) \\ &= \frac{\text{Cov}(X - \hat{X}_{Z_1, \dots, Z_m}, Y - \hat{Y}_{Z_1, \dots, Z_m})}{(\text{Var}(X - \hat{X}_{Z_1, \dots, Z_m}))^{1/2} (\text{Var}(Y - \hat{Y}_{Z_1, \dots, Z_m}))^{1/2}}, \end{aligned}$$

provided that $\text{Var}(X - \hat{X}_{Z_1, \dots, Z_m}) > 0$ and $\text{Var}(Y - \hat{Y}_{Z_1, \dots, Z_m}) > 0$, where $\hat{X}_{Z_1, \dots, Z_m}$ and $\hat{Y}_{Z_1, \dots, Z_m}$ denote best linear approximations of X and Y based on Z_1, \dots, Z_m , respectively.

Let $(Y_t)_{t \in \mathbb{Z}}$ be an ARMA(p, q)-process satisfying the stationarity condition (2.4) with expectation $E(Y_t) = 0$ and variance $\gamma(0) > 0$. The *partial autocorrelation* $\alpha(t, k)$ for $k > 1$ is the partial correlation of Y_t and Y_{t-k} , where the linear influence of the intermediate variables $Y_i, t-k < i < t$, is removed, i.e., the best linear approximation of Y_t based on the $k-1$ preceding process variables, denoted by $\hat{Y}_{t, k-1} := \sum_{i=1}^{k-1} \hat{a}_i Y_{t-i}$. Since $\hat{Y}_{t, k-1}$ minimizes the mean squared error $E((Y_t - \hat{Y}_{t, \mathbf{a}, k-1})^2)$ among all linear combinations $\tilde{Y}_{t, \mathbf{a}, k-1} := \sum_{i=1}^{k-1} a_i Y_{t-i}$, $\mathbf{a} := (a_1, \dots, a_{k-1})^T \in \mathbb{R}^{k-1}$ of $Y_{t-k+1}, \dots, Y_{t-1}$, we find, due to the stationarity condition, that $\hat{Y}_{t-k, -k+1} := \sum_{i=1}^{k-1} \hat{a}_i Y_{t-k+i}$ is a best linear approximation of Y_{t-k} based on the $k-1$ subsequent process variables. Setting $\hat{Y}_{t,0} = 0$ in the case of $k = 1$, we obtain, for

$k > 0$,

$$\begin{aligned}\alpha(t, k) &:= \text{Corr}(Y_t - \hat{Y}_{t,k-1}, Y_{t-k} - \hat{Y}_{t-k,-k+1}) \\ &= \frac{\text{Cov}(Y_t - \hat{Y}_{t,k-1}, Y_{t-k} - \hat{Y}_{t-k,-k+1})}{\text{Var}(Y_t - \hat{Y}_{t,k-1})}.\end{aligned}$$

Note that $\text{Var}(Y_t - \hat{Y}_{t,k-1}) > 0$ is provided by the preliminary conditions, which will be shown later by the proof of Theorem 7.1.1. Observe moreover that $\hat{Y}_{t+h,k-1} = \sum_{i=1}^{k-1} \hat{a}_i Y_{t+h-i}$ for all $h \in \mathbb{Z}$, implying $\alpha(t, k) = \text{Corr}(Y_t - \hat{Y}_{t,k-1}, Y_{t-k} - \hat{Y}_{t-k,-k+1}) = \text{Corr}(Y_k - \hat{Y}_{k,k-1}, Y_0 - \hat{Y}_{0,-k+1}) = \alpha(k, k)$. Consequently the *partial autocorrelation function* can be more conveniently written as

$$\alpha(k) := \frac{\text{Cov}(Y_k - \hat{Y}_{k,k-1}, Y_0 - \hat{Y}_{0,-k+1})}{\text{Var}(Y_k - \hat{Y}_{k,k-1})} \quad (7.1)$$

for $k > 0$ and $\alpha(0) = 1$ for $k = 0$. For negative k , we set $\alpha(k) := \alpha(-k)$.

The determination of the partial autocorrelation coefficient $\alpha(k)$ at lag $k > 1$ entails the computation of the coefficients of the corresponding best linear approximation $\hat{Y}_{k,k-1}$, leading to an equation system similar to the normal equations coming from a regression model. Let $\tilde{Y}_{k,a,k-1} = a_1 Y_{k-1} + \cdots + a_{k-1} Y_1$ be an arbitrary linear approximation of Y_k based on Y_1, \dots, Y_{k-1} . Then, the mean squared error is given by

$$\begin{aligned}E((Y_k - \tilde{Y}_{k,a,k-1})^2) &= E(Y_k^2) - 2E(Y_k \tilde{Y}_{k,a,k-1}) + E(\tilde{Y}_{k,a,k-1}^2) \\ &= E(Y_k^2) - 2 \sum_{l=1}^{k-1} a_l E(Y_k Y_{k-l}) \\ &\quad + \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} E(a_i a_j Y_{k-i} Y_{k-j})\end{aligned}$$

Computing the partial derivatives and equating them to zero yields

$$\begin{aligned}2\hat{a}_1 E(Y_{k-1} Y_{k-1}) + \cdots + 2\hat{a}_{k-1} E(Y_{k-1} Y_1) - 2E(Y_{k-1} Y_k) &= 0, \\ 2\hat{a}_1 E(Y_{k-2} Y_{k-1}) + \cdots + 2\hat{a}_{k-1} E(Y_{k-2} Y_1) - 2E(Y_{k-2} Y_k) &= 0, \\ \vdots \\ 2\hat{a}_1 E(Y_1 Y_{k-1}) + \cdots + 2\hat{a}_{k-1} E(Y_1 Y_1) - 2E(Y_1 Y_k) &= 0.\end{aligned}$$

Or, equivalently, written in matrix notation

$$\mathbf{V}_{yy}\hat{\mathbf{a}} = \mathbf{V}_{yY_k}$$

with vector of coefficients $\hat{\mathbf{a}} := (\hat{a}_1, \dots, \hat{a}_{k-1})^T$, $\mathbf{y} := (Y_{k-1}, \dots, Y_1)^T$ and matrix representation

$$\mathbf{V}_{wz} := \left(E(W_i Z_j) \right)_{ij},$$

where $\mathbf{w} := (W_1, \dots, W_n)^T$, $\mathbf{z} := (Z_1, \dots, Z_m)^T$, $n, m \in \mathbb{N}$, denote random vectors. So, if the matrix \mathbf{V}_{yy} is regular, we attain a uniquely determined solution of the equation system

$$\hat{\mathbf{a}} = \mathbf{V}_{yy}^{-1} \mathbf{V}_{yY_k}.$$

Since the ARMA(p, q)-process $(Y_t)_{t \in \mathbb{Z}}$ has expectation zero, above representations divided by the variance $\gamma(0) > 0$ carry over to the Yule-Walker equations of order $k - 1$ as introduced in Section 2.2

$$\mathbf{P}_{k-1} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_{k-1} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k-1) \end{pmatrix}, \quad (7.2)$$

or, respectively,

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_{k-1} \end{pmatrix} = \mathbf{P}_{k-1}^{-1} \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k-1) \end{pmatrix} \quad (7.3)$$

if \mathbf{P}_{k-1}^{-1} is regular. A best linear approximation $\hat{Y}_{k,k-1}$ of Y_k based on Y_1, \dots, Y_{k-1} obviously has to share this necessary condition (7.2). Since $\hat{Y}_{k,k-1}$ equals a best linear one-step forecast of Y_k based on Y_1, \dots, Y_{k-1} , Lemma 2.3.2 shows that $\hat{Y}_{k,k-1} = \hat{\mathbf{a}}^T \mathbf{y}$ is a best linear approximation of Y_k . Thus, if \mathbf{P}_{k-1}^{-1} is regular, then $\hat{Y}_{k,k-1}$ is given by

$$\hat{Y}_{k,k-1} = \hat{\mathbf{a}}^T \mathbf{y} = \mathbf{V}_{Y_k y} \mathbf{V}_{yy}^{-1} \mathbf{y}. \quad (7.4)$$

The next Theorem will now establish the necessary regularity of \mathbf{P}_k .

Theorem 7.1.1. *If $(Y_t)_{t \in \mathbb{Z}}$ is an $\text{ARMA}(p, q)$ -process satisfying the stationarity condition with $\mathbb{E}(Y_t) = 0$, $\text{Var}(Y_t) > 0$ and $\text{Var}(\varepsilon_t) = \sigma^2$, then, the covariance matrix Σ_k of $\mathbf{Y}_{t,k} := (Y_{t+1}, \dots, Y_{t+k})^T$ is regular for every $k > 0$.*

Proof. Let $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$, $t \in \mathbb{Z}$, be the almost surely stationary solution of $(Y_t)_{t \in \mathbb{Z}}$ with absolutely summable filter $(\alpha_v)_{v \geq 0}$. The autocovariance function $\gamma(k)$ of $(Y_t)_{t \in \mathbb{Z}}$ is absolutely summable as well, since

$$\begin{aligned} \sum_{k \in \mathbb{Z}} |\gamma(k)| &< 2 \sum_{k \geq 0} |\gamma(k)| = 2 \sum_{k \geq 0} |\mathbb{E}(Y_t Y_{t+k})| \\ &= 2 \sum_{k \geq 0} \left| \sum_{u \geq 0} \sum_{v \geq 0} \alpha_u \alpha_v \mathbb{E}(\varepsilon_{t-u} \varepsilon_{t+k-v}) \right| \\ &= 2\sigma^2 \sum_{k \geq 0} \left| \sum_{u \geq 0} \alpha_u \alpha_{u+k} \right| \\ &\leq 2\sigma^2 \sum_{u \geq 0} |\alpha_u| \sum_{w \geq 0} |\alpha_w| < \infty. \end{aligned}$$

Suppose now that Σ_k is singular for a $k > 1$. Then, there exists a maximum integer $\kappa \geq 1$ such that Σ_κ is regular. Thus, $\Sigma_{\kappa+1}$ is singular. This entails in particular the existence of a non-zero vector $\boldsymbol{\lambda}_{\kappa+1} := (\lambda_1, \dots, \lambda_{\kappa+1})^T \in \mathbb{R}^{\kappa+1}$ with $\boldsymbol{\lambda}_{\kappa+1}^T \Sigma_{\kappa+1} \boldsymbol{\lambda}_{\kappa+1} = 0$. Therefore, $\text{Var}(\boldsymbol{\lambda}_{\kappa+1}^T \mathbf{Y}_{t,\kappa+1}) = 0$ since the process has expectation zero. Consequently, $\boldsymbol{\lambda}_{\kappa+1}^T \mathbf{Y}_{t,\kappa+1}$ is a constant. It has expectation zero and is therefore constant zero. Hence, $Y_{t+\kappa+1} = \sum_{i=1}^{\kappa} \frac{\lambda_i}{\lambda_{\kappa+1}} Y_{t+i}$ for each $t \in \mathbb{Z}$, which implies the existence of a non-zero vector $\boldsymbol{\lambda}_\kappa^{(t)} := (\lambda_1^{(t)}, \dots, \lambda_\kappa^{(t)})^T \in \mathbb{R}^\kappa$ with $Y_{t+\kappa+1} = \sum_{i=1}^{\kappa} \lambda_i^{(t)} Y_{t+i} = \boldsymbol{\lambda}_\kappa^{(t)T} \mathbf{Y}_{0,\kappa}$. Note that $\lambda_{\kappa+1} \neq 0$, since otherwise Σ_κ would be singular. Because of the regularity of Σ_κ , we find a diagonal matrix \mathbf{D}_κ and an orthogonal matrix \mathbf{S}_κ such that

$$\gamma(0) = \text{Var}(Y_{t+\kappa+1}) = \boldsymbol{\lambda}_\kappa^{(t)T} \Sigma_\kappa \boldsymbol{\lambda}_\kappa^{(t)} = \boldsymbol{\lambda}_\kappa^{(t)T} \mathbf{S}_\kappa \mathbf{D}_\kappa \mathbf{S}_\kappa^T \boldsymbol{\lambda}_\kappa^{(t)}.$$

The diagonal elements of \mathbf{D}_κ are the positive eigenvalues of Σ_κ , whose smallest eigenvalue will be denoted by $\mu_{\kappa,\min}$. Hence,

$$\infty > \gamma(0) \geq \mu_{\kappa,\min} \boldsymbol{\lambda}_\kappa^{(t)T} \mathbf{S}_\kappa \mathbf{S}_\kappa^T \boldsymbol{\lambda}_\kappa^{(t)} = \mu_{\kappa,\min} \sum_{i=1}^{\kappa} \lambda_i^{(t)2}.$$

This shows the boundedness of $\lambda_i^{(t)}$ for a fixed i . On the other hand,

$$\gamma(0) = \text{Cov} \left(Y_{t+\kappa+1}, \sum_{i=1}^{\kappa} \lambda_i^{(t)} Y_i \right) \leq \sum_{i=1}^{\kappa} |\lambda_i^{(t)}| |\gamma(t + \kappa + 1 - i)|.$$

Since $\gamma(0) > 0$ and $\gamma(t + \kappa + 1 - i) \rightarrow 0$ as $t \rightarrow \infty$, due to the absolutely summability of $\gamma(k)$, this inequality would produce a contradiction. Consequently, Σ_k is regular for all $k > 0$. \square

The Levinson–Durbin Recursion

The *power of an approximation* \hat{Y}_k of Y_k can be measured by the part of unexplained variance of the residuals. It is defined as follows:

$$P(k-1) := \frac{\text{Var}(Y_k - \hat{Y}_{k,k-1})}{\text{Var}(Y_k)}, \quad (7.5)$$

if $\text{Var}(Y_k) > 0$. Observe that the greater the power, the less precise the approximation \hat{Y}_k performs. Let again $(Y_t)_{t \in \mathbb{Z}}$ be an ARMA(p, q)-process satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$. Furthermore, let $\hat{Y}_{k,k-1} := \sum_{u=1}^{k-1} \hat{a}_u(k-1) Y_{k-u}$ denote the best linear approximation of Y_k based on the $k-1$ preceding random variables for $k > 1$. Then, equation (7.4) and Theorem 7.1.1 provide

$$\begin{aligned} P(k-1) &:= \frac{\text{Var}(Y_k - \hat{Y}_{k,k-1})}{\text{Var}(Y_k)} = \frac{\text{Var}(Y_k - \mathbf{V}_{Y_k \mathbf{y}} \mathbf{V}_{\mathbf{y} \mathbf{y}}^{-1} \mathbf{y})}{\text{Var}(Y_k)} \\ &= \frac{1}{\text{Var}(Y_k)} \left(\text{Var}(Y_k) + \text{Var}(\mathbf{V}_{Y_k \mathbf{y}} \mathbf{V}_{\mathbf{y} \mathbf{y}}^{-1} \mathbf{y}) - 2 E(Y_k \mathbf{V}_{Y_k \mathbf{y}} \mathbf{V}_{\mathbf{y} \mathbf{y}}^{-1} \mathbf{y}) \right) \\ &= \frac{1}{\gamma(0)} \left(\gamma(0) + \mathbf{V}_{Y_k \mathbf{y}} \mathbf{V}_{\mathbf{y} \mathbf{y}}^{-1} \mathbf{V}_{\mathbf{y} \mathbf{y}} \mathbf{V}_{\mathbf{y} \mathbf{y}}^{-1} \mathbf{V}_{\mathbf{y} Y_k} - 2 \mathbf{V}_{Y_k \mathbf{y}} \mathbf{V}_{\mathbf{y} \mathbf{y}}^{-1} \mathbf{V}_{\mathbf{y} Y_k} \right) \\ &= \frac{1}{\gamma(0)} \left(\gamma(0) - \mathbf{V}_{Y_k \mathbf{y}} \hat{\mathbf{a}}(k-1) \right) = 1 - \sum_{i=1}^{k-1} \rho(i) \hat{a}_i(k-1) \end{aligned} \quad (7.6)$$

where $\hat{\mathbf{a}}(k-1) := (\hat{a}_1(k-1), \hat{a}_2(k-1), \dots, \hat{a}_{k-1}(k-1))^T$. Note that $P(k-1) \neq 0$ is provided by the proof of the previous Theorem 7.1.1.

In order to determine the coefficients of the best linear approximation $\hat{Y}_{k,k-1}$ we have to solve (7.3). A well-known algorithm simplifies this task by evading the computation of the inverse matrix.

Theorem 7.1.2. (*The Levinson–Durbin Recursion*) Let $(Y_t)_{t \in \mathbb{Z}}$ be an ARMA(p, q)-process satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$. Let $\hat{Y}_{k,k-1} = \sum_{u=1}^{k-1} \hat{a}_u(k-1)Y_{k-u}$ denote the best linear approximation of Y_k based on the $k-1$ preceding process values Y_{k-1}, \dots, Y_1 for $k > 1$. Then,

$$(i) \quad \hat{a}_k(k) = \omega(k)/P(k-1),$$

$$(ii) \quad \hat{a}_i(k) = \hat{a}_i(k-1) - \hat{a}_k(k)\hat{a}_{k-i}(k-1) \text{ for } i = 1, \dots, k-1, \text{ as well as}$$

$$(iii) \quad P(k) = P(k-1)(1 - \hat{a}_k^2(k)),$$

where $P(k-1) = 1 - \sum_{i=1}^{k-1} \rho(i)\hat{a}_i(k-1)$ denotes the power of approximation and where

$$\omega(k) := \rho(k) - \hat{a}_1(k-1)\rho(k-1) - \dots - \hat{a}_{k-1}(k-1)\rho(1).$$

Proof. We carry out two best linear approximations of Y_t , $\hat{Y}_{t,k-1} := \sum_{u=1}^{k-1} \hat{a}_u(k-1)Y_{t-u}$ and $\hat{Y}_{t,k} := \sum_{u=1}^k \hat{a}_u(k)Y_{t-u}$, which lead to the Yule-Walker equations (7.3) of order $k-1$

$$\begin{pmatrix} 1 & \rho(1) & \dots & \rho(k-2) \\ \rho(1) & 1 & & \rho(k-3) \\ \rho(2) & \rho(1) & & \rho(k-4) \\ \vdots & & \ddots & \vdots \\ \rho(k-2) & \rho(k-3) & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{a}_1(k-1) \\ \hat{a}_2(k-1) \\ \hat{a}_3(k-1) \\ \vdots \\ \hat{a}_{k-1}(k-1) \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ \vdots \\ \rho(k-1) \end{pmatrix} \quad (7.7)$$

and order k

$$\begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-1) \\ \rho(1) & 1 & \rho(1) & & \rho(k-2) \\ \rho(2) & \rho(1) & 1 & & \rho(k-3) \\ \vdots & & & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \rho(k-3) & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{a}_1(k) \\ \hat{a}_2(k) \\ \hat{a}_3(k) \\ \vdots \\ \hat{a}_k(k) \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ \vdots \\ \rho(k) \end{pmatrix}.$$

Latter equation system can be written as

$$\begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-2) \\ \rho(1) & 1 & \rho(1) & & \rho(k-3) \\ \rho(2) & \rho(1) & 1 & & \rho(k-4) \\ \vdots & & & \ddots & \vdots \\ \rho(k-2) & \rho(k-3) & \rho(k-4) & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{a}_1(k) \\ \hat{a}_2(k) \\ \hat{a}_3(k) \\ \vdots \\ \hat{a}_{k-1}(k) \end{pmatrix} + \hat{a}_k(k) \begin{pmatrix} \rho(k-1) \\ \rho(k-2) \\ \rho(k-3) \\ \vdots \\ \rho(1) \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ \vdots \\ \rho(k-1) \end{pmatrix}$$

and

$$\rho(k-1)\hat{a}_1(k) + \rho(k-2)\hat{a}_2(k) + \dots + \rho(1)\hat{a}_{k-1}(k) + \hat{a}_k(k) = \rho(k). \quad (7.8)$$

From (7.7), we obtain

$$\begin{aligned} & \begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-2) \\ \rho(1) & 1 & \rho(1) & & \rho(k-3) \\ \rho(2) & \rho(1) & 1 & & \rho(k-4) \\ \vdots & & & \ddots & \vdots \\ \rho(k-2) & \rho(k-3) & \rho(k-4) & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{a}_1(k) \\ \hat{a}_2(k) \\ \hat{a}_3(k) \\ \vdots \\ \hat{a}_{k-1}(k) \end{pmatrix} \\ & + \hat{a}_k(k) \begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-2) \\ \rho(1) & 1 & \rho(1) & & \rho(k-3) \\ \rho(2) & \rho(1) & 1 & & \rho(k-4) \\ \vdots & & & \ddots & \vdots \\ \rho(k-2) & \rho(k-3) & \rho(k-4) & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{a}_{k-1}(k-1) \\ \hat{a}_{k-2}(k-1) \\ \hat{a}_{k-3}(k-1) \\ \vdots \\ \hat{a}_1(k-1) \end{pmatrix} \\ & = \begin{pmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(k-2) \\ \rho(1) & 1 & \rho(1) & & \rho(k-3) \\ \rho(2) & \rho(1) & 1 & & \rho(k-4) \\ \vdots & & & \ddots & \vdots \\ \rho(k-2) & \rho(k-3) & \rho(k-4) & \dots & 1 \end{pmatrix} \begin{pmatrix} \hat{a}_1(k-1) \\ \hat{a}_2(k-1) \\ \hat{a}_3(k-1) \\ \vdots \\ \hat{a}_{k-1}(k-1) \end{pmatrix} \end{aligned}$$

Multiplying \mathbf{P}_{k-1}^{-1} to the left we get

$$\begin{pmatrix} \hat{a}_1(k) \\ \hat{a}_2(k) \\ \hat{a}_3(k) \\ \vdots \\ \hat{a}_{k-1}(k) \end{pmatrix} = \begin{pmatrix} \hat{a}_1(k-1) \\ \hat{a}_2(k-1) \\ \hat{a}_3(k-1) \\ \vdots \\ \hat{a}_{k-1}(k-1) \end{pmatrix} - \hat{a}_k(k) \begin{pmatrix} \hat{a}_{k-1}(k-1) \\ \hat{a}_{k-2}(k-1) \\ \hat{a}_{k-3}(k-1) \\ \vdots \\ \hat{a}_1(k-1) \end{pmatrix}.$$

This is the central recursion equation system (ii). Applying the corresponding terms of $\hat{a}_1(k), \dots, \hat{a}_{k-1}(k)$ from the central recursion equa-

tion to the remaining equation (7.8) yields

$$\begin{aligned} \hat{a}_k(k)(1 - \hat{a}_1(k-1)\rho(1) - \hat{a}_2(k-1)\rho(2) - \dots - \hat{a}_{k-1}(k-1)\rho(k-1)) \\ = \rho(k) - (\hat{a}_1(k-1)\rho(k-1) + \dots + \hat{a}_{k-1}(k-1)\rho(1)) \end{aligned}$$

or, respectively,

$$\hat{a}_k(k)P(k-1) = \omega(k)$$

which proves (i). From (i) and (ii), it follows (iii),

$$\begin{aligned} P(k) &= 1 - \hat{a}_1(k)\rho(1) - \dots - \hat{a}_k(k)\rho(k) \\ &= 1 - \hat{a}_1(k-1)\rho(1) - \dots - \hat{a}_{k-1}(k-1)\rho(k-1) - \hat{a}_k(k)\rho(k) \\ &\quad + \hat{a}_k(k)\rho(1)\hat{a}_{k-1}(k-1) + \dots + \hat{a}_k(k)\rho(k-1)\hat{a}_1(k-1) \\ &= P(k-1) - \hat{a}_k(k)\omega(k) \\ &= P(k-1)(1 - \hat{a}_k(k)^2). \end{aligned}$$

□

We have already observed several similarities between the general definition of the partial autocorrelation (7.1) and the definition in Section 2.2. The following theorem will now establish the equivalence of both definitions in the case of zero-mean ARMA(p, q)-processes satisfying the stationarity condition and $\text{Var}(Y_t) > 0$.

Theorem 7.1.3. *The partial autocorrelation $\alpha(k)$ of an ARMA(p, q)-process $(Y_t)_{t \in \mathbb{Z}}$ satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$ equals, for $k > 1$, the coefficient $\hat{a}_k(k)$ in the best linear approximation $\hat{Y}_{k+1,k} = \hat{a}_1(k)Y_k + \dots + \hat{a}_k(k)Y_1$ of Y_{k+1} .*

Proof. Applying the definition of $\alpha(k)$, we find, for $k > 1$,

$$\begin{aligned} \alpha(k) &= \frac{\text{Cov}(Y_k - \hat{Y}_{k,k-1}, Y_0 - \hat{Y}_{0,-k+1})}{\text{Var}(Y_k - \hat{Y}_{k,k-1})} \\ &= \frac{\text{Cov}(Y_k - \hat{Y}_{k,k-1}, Y_0) + \text{Cov}(Y_k - \hat{Y}_{k,k-1}, -\hat{Y}_{0,-k+1})}{\text{Var}(Y_k - \hat{Y}_{k,k-1})} \\ &= \frac{\text{Cov}(Y_k - \hat{Y}_{k,k-1}, Y_0)}{\text{Var}(Y_k - \hat{Y}_{k,k-1})}. \end{aligned}$$

The last step follows from (7.2), which implies that $Y_k - \hat{Y}_{k,k-1}$ is uncorrelated with Y_1, \dots, Y_{k-1} . Setting $\hat{Y}_{k,k-1} = \sum_{u=1}^{k-1} \hat{a}_u(k-1)Y_{k-u}$, we attain for the numerator

$$\begin{aligned} \text{Cov}(Y_k - \hat{Y}_{k,k-1}, Y_0) &= \text{Cov}(Y_k, Y_0) - \text{Cov}(\hat{Y}_{k,k-1}, Y_0) \\ &= \gamma(k) - \sum_{u=1}^{k-1} \hat{a}_u(k-1) \text{Cov}(Y_{k-u}, Y_0) \\ &= \gamma(0)(\rho(k) - \sum_{u=1}^{k-1} \hat{a}_u(k-1)\rho(k-u)). \end{aligned}$$

Now, applying the first formula of the Levinson–Durbin Theorem 7.1.2 finally leads to

$$\begin{aligned} \alpha(k) &= \frac{\gamma(0)(\rho(k) - \sum_{u=1}^{k-1} \hat{a}_u(k-1)\rho(k-u))}{\gamma(0)P(k-1)} \\ &= \frac{\gamma(0)P(k-1)\hat{a}_k(k)}{\gamma(0)P(k-1)} \\ &= \hat{a}_k(k). \end{aligned}$$

□

Partial Correlation Matrix

We are now in position to extend the previous definitions to random vectors. Let $\mathbf{y} := (Y_1, \dots, Y_k)^T$ and $\mathbf{x} := (X_1, \dots, X_m)^T$ be in the following random vectors of square integrable, real valued random variables. The *partial covariance matrix* of \mathbf{y} with X_1, \dots, X_m being held fix is defined as

$$\mathbf{V}_{\mathbf{y}\mathbf{y},\mathbf{x}} := \left(\text{Cov}(Y_i - \hat{Y}_{i,\mathbf{x}}, Y_j - \hat{Y}_{j,\mathbf{x}}) \right)_{1 \leq i, j \leq k},$$

where $\hat{Y}_{i,\mathbf{x}}$ denotes the best linear approximation of $Y_{i,\mathbf{x}}$ based on X_1, \dots, X_m for $i = 1, \dots, k$. The *partial correlation matrix* of \mathbf{y} , where the linear influence of X_1, \dots, X_m is removed, is accordingly defined by

$$\mathbf{R}_{\mathbf{y}\mathbf{y},\mathbf{x}} := \left(\text{Corr}(Y_i - \hat{Y}_{i,\mathbf{x}}, Y_j - \hat{Y}_{j,\mathbf{x}}) \right)_{1 \leq i, j \leq k},$$

if $\text{Var}(Y_i - \hat{Y}_{i,x}) > 0$ for each $i \in \{1, \dots, k\}$.

Lemma 7.1.4. *Consider above notations with $E(\mathbf{y}) = \mathbf{0}$ as well as $E(\mathbf{x}) = \mathbf{0}$, where $\mathbf{0}$ denotes the zero vector in \mathbb{R}^k and \mathbb{R}^m , respectively. Let furthermore $\mathbf{V}_{\mathbf{wz}}$ denote the covariance matrix of two random vectors \mathbf{w} and \mathbf{z} . Then, if \mathbf{V}_{xx} is regular, the partial covariance matrix $\mathbf{V}_{yy,x}$ satisfies*

$$\mathbf{V}_{yy,x} = \mathbf{V}_{yy} - \mathbf{V}_{yx} \mathbf{V}_{xx}^{-1} \mathbf{V}_{xy}.$$

Proof. For arbitrary $i, j \in \{1, \dots, k\}$, we get

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \text{Cov}(Y_i - \hat{Y}_{i,x} + \hat{Y}_{i,x}, Y_j - \hat{Y}_{j,x} + \hat{Y}_{j,x}) \\ &= \text{Cov}(Y_i - \hat{Y}_{i,x}, Y_j - \hat{Y}_{j,x}) + \text{Cov}(\hat{Y}_{i,x}, \hat{Y}_{j,x}) \\ &\quad + E(\hat{Y}_{i,x}(Y_j - \hat{Y}_{j,x})) + E(\hat{Y}_{j,x}(Y_i - \hat{Y}_{i,x})) \\ &= \text{Cov}(Y_i - \hat{Y}_{i,x}, Y_j - \hat{Y}_{j,x}) + \text{Cov}(\hat{Y}_{i,x}, \hat{Y}_{j,x}) \end{aligned} \quad (7.9)$$

or, equivalently,

$$\mathbf{V}_{yy} = \mathbf{V}_{yy,x} + \mathbf{V}_{\hat{\mathbf{y}}_x \hat{\mathbf{y}}_x},$$

where $\hat{\mathbf{y}}_x := (\hat{Y}_{1,x}, \dots, \hat{Y}_{k,x})^T$. The last step in (7.9) follows from the circumstance that the approximation errors $(Y_i - \hat{Y}_{i,x})$ of the best linear approximation are uncorrelated with every linear combination of X_1, \dots, X_m for every $i = 1, \dots, n$, due to the partial derivatives of the mean squared error being equal to zero as described at the beginning of this section. Furthermore, (7.4) implies that

$$\hat{\mathbf{y}}_x = \mathbf{V}_{yx} \mathbf{V}_{xx}^{-1} \mathbf{x}.$$

Hence, we finally get

$$\begin{aligned} \mathbf{V}_{\hat{\mathbf{y}}_x \hat{\mathbf{y}}_x} &= \mathbf{V}_{yx} \mathbf{V}_{xx}^{-1} \mathbf{V}_{xx} \mathbf{V}_{xx}^{-1} \mathbf{V}_{xy} \\ &= \mathbf{V}_{yx} \mathbf{V}_{xx}^{-1} \mathbf{V}_{xy}, \end{aligned}$$

which completes the proof. \square

Lemma 7.1.5. *Consider the notations and conditions of Lemma 7.1.4. If $\mathbf{V}_{yy,x}$ is regular, then,*

$$\begin{pmatrix} \mathbf{V}_{xx} & \mathbf{V}_{xy} \\ \mathbf{V}_{yx} & \mathbf{V}_{yy} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{V}_{xx}^{-1} + \mathbf{V}_{xx}^{-1} \mathbf{V}_{xy} \mathbf{V}_{yy,x}^{-1} \mathbf{V}_{yx} \mathbf{V}_{xx}^{-1} & -\mathbf{V}_{xx}^{-1} \mathbf{V}_{xy} \mathbf{V}_{yy,x}^{-1} \\ -\mathbf{V}_{yy,x}^{-1} \mathbf{V}_{yx} \mathbf{V}_{xx}^{-1} & \mathbf{V}_{yy,x}^{-1} \end{pmatrix}.$$

Proof. The assertion follows directly from the previous Lemma 7.1.4. \square

7.2 Asymptotic Normality of Partial Autocorrelation Estimator

The following table comprises some behaviors of the partial and general autocorrelation function with respect to the type of process.

process type	autocorrelation function	partial autocorrelation function
MA(q)	finite	infinite
AR(p)	$\rho(k) = 0$ for $k > q$ infinite	finite $\alpha(k) = 0$ for $k > p$
ARMA(p, q)	infinite	infinite

Knowing these qualities of the true partial and general autocorrelation function we expect a similar behavior for the corresponding empirical counterparts. In general, unlike the theoretic function, the empirical partial autocorrelation computed from an outcome of an AR(p)-process won't vanish after lag p . Nevertheless the coefficients after lag p will tend to be close to zero. Thus, if a given time series was actually generated by an AR(p)-process, then the empirical partial autocorrelation coefficients will lie close to zero after lag p . The same is true for the empirical autocorrelation coefficients after lag q , if the time series was actually generated by a MA(q)-process. In order to justify AR(p)- or MA(q)-processes as appropriate model candidates for explaining a given time series, we have to verify if the values are significantly close to zero. To this end, we will derive asymptotic distributions of suitable estimators of these coefficients in the following two sections.

Cramer–Wold Device

Definition 7.2.1. A sequence of real valued random variables $(Y_n)_{n \in \mathbb{N}}$, defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, is said to be *asymptoti-*

cally normal with asymptotic mean μ_n and asymptotic variance $\sigma_n^2 > 0$ for sufficiently large n , written as $Y_n \approx N(\mu_n, \sigma_n^2)$, if $\sigma_n^{-1}(Y_n - \mu_n) \xrightarrow{D} Y$ as $n \rightarrow \infty$, where Y is $N(0,1)$ -distributed.

We want to extend asymptotic normality to random vectors, which will be motivated by the Cramer–Wold Device 7.2.4. However, in order to simplify proofs concerning convergence in distribution, we first characterize convergence in distribution in terms of the sequence of the corresponding characteristic functions.

Let \mathbf{Y} be a real valued random k -vector with distribution function $F(\mathbf{x})$. The *characteristic function* of $F(\mathbf{x})$ is then defined as the Fourier transformation

$$\phi_Y(\mathbf{t}) := E(\exp(i\mathbf{t}^T \mathbf{Y})) = \int_{\mathbb{R}^k} \exp(i\mathbf{t}^T \mathbf{x}) dF(\mathbf{x})$$

for $\mathbf{t} \in \mathbb{R}^k$. In many cases of proving weak convergence of a sequence of random k -vectors $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$, it is often easier to show the pointwise convergence of the corresponding sequence of characteristic functions $\phi_{Y_n}(\mathbf{t}) \rightarrow \phi_Y(\mathbf{t})$ for every $\mathbf{t} \in \mathbb{R}^k$. The following two Theorems will establish this equivalence.

Theorem 7.2.2. *Let $\mathbf{Y}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \dots$ be real valued random k -vectors, defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$. Then, the following conditions are equivalent*

- (i) $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$,
- (ii) $E(\vartheta(\mathbf{Y}_n)) \rightarrow E(\vartheta(\mathbf{Y}))$ as $n \rightarrow \infty$ for all bounded and continuous functions $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}$,
- (iii) $E(\vartheta(\mathbf{Y}_n)) \rightarrow E(\vartheta(\mathbf{Y}))$ as $n \rightarrow \infty$ for all bounded and uniformly continuous functions $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}$,
- (iv) $\limsup_{n \rightarrow \infty} P(\{\mathbf{Y}_n \in C\}) \leq P(\{\mathbf{Y} \in C\})$ for any closed set $C \subset \mathbb{R}^k$,
- (v) $\liminf_{n \rightarrow \infty} P(\{\mathbf{Y}_n \in O\}) \geq P(\{\mathbf{Y} \in O\})$ for any open set $O \subset \mathbb{R}^k$.

Proof. (i) \Rightarrow (ii): Suppose that $F(\mathbf{x}), F_n(\mathbf{x})$ denote the corresponding distribution functions of real valued random k -vectors \mathbf{Y}, \mathbf{Y}_n for all $n \in \mathbb{N}$, respectively, satisfying $F_n(\mathbf{x}) \rightarrow F(\mathbf{x})$ as $n \rightarrow \infty$ for every continuity point \mathbf{x} of $F(\mathbf{x})$. Let $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}$ be a bounded and continuous function, which is obviously bounded by the finite value $B := \sup_{\mathbf{x}} \{|\vartheta(\mathbf{x})|\}$. Now, given $\varepsilon > 0$, we find, due to the right-continuousness, continuity points $\pm \mathbf{C} := \pm(C_1, \dots, C_k)^T$ of $F(\mathbf{x})$, with $C_r \neq 0$ for $r = 1, \dots, k$ and a compact set $K := \{(x_1, \dots, x_k) : -C_r \leq x_r \leq C_r, r = 1, \dots, k\}$ such that $P\{\mathbf{Y} \notin K\} < \varepsilon/B$. Note that this entails $P\{\mathbf{Y}_n \notin K\} < 2\varepsilon/B$ for n sufficiently large. Now, we choose $l \geq 2$ continuity points $\mathbf{x}_j := (x_{j1}, \dots, x_{jk})^T, j \in \{1, \dots, l\}$, of $F(\mathbf{x})$ such that $-C_r = x_{1r} < \dots < x_{lr} = C_r$ for each $r \in \{1, \dots, k\}$ and such that $\sup_{\mathbf{x} \in K} |\vartheta(\mathbf{x}) - \varphi(\mathbf{x})| < \varepsilon$, where $\varphi(\mathbf{x}) := \sum_{i=1}^{l-1} \vartheta(\mathbf{x}_i) 1_{(\mathbf{x}_i, \mathbf{x}_{i+1}]}(\mathbf{x})$. Then, we attain, for n sufficiently large,

$$\begin{aligned}
& |E(\vartheta(\mathbf{Y}_n)) - E(\vartheta(\mathbf{Y}))| \\
& \leq |E(\vartheta(\mathbf{Y}_n) 1_K(\mathbf{Y}_n)) - E(\vartheta(\mathbf{Y}) 1_K(\mathbf{Y}))| \\
& \quad + E(|\vartheta(\mathbf{Y}_n|)|) 1_{K^c}(\mathbf{Y}_n) + E(|\vartheta(\mathbf{Y})|) 1_{K^c}(\mathbf{Y}) \\
& < |E(\vartheta(\mathbf{Y}_n) 1_K(\mathbf{Y}_n)) - E(\vartheta(\mathbf{Y}) 1_K(|\mathbf{Y}|))| + B \cdot 2\varepsilon/B + B \cdot \varepsilon/B \\
& \leq |E(\vartheta(\mathbf{Y}_n) 1_K(\mathbf{Y}_n)) - E(\varphi(\mathbf{Y}_n))| + |E(\varphi(\mathbf{Y}_n)) - E(\varphi(\mathbf{Y}))| \\
& \quad + |E(\varphi(\mathbf{Y})) - E(\vartheta(\mathbf{Y}) 1_K(\mathbf{Y}))| + 3\varepsilon \\
& < |E(\varphi(\mathbf{Y}_n)) - E(\varphi(\mathbf{Y}))| + 5\varepsilon,
\end{aligned}$$

where K^c denotes the complement of K and where $1_A(\cdot)$ denotes the indicator function of a set A , i.e., $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ else. As $\varepsilon > 0$ was chosen arbitrarily, it remains to show $E(\varphi(\mathbf{Y}_n)) \rightarrow E(\varphi(\mathbf{Y}))$ as $n \rightarrow \infty$, which can be seen as follows:

$$\begin{aligned}
E(\varphi(\mathbf{Y}_n)) &= \sum_{i=1}^{l-1} \vartheta(\mathbf{x}_i) (F_n(\mathbf{x}_{i+1}) - F_n(\mathbf{x}_i)) \\
&\rightarrow \sum_{i=1}^{l-1} \vartheta(\mathbf{x}_i) (F(\mathbf{x}_{i+1}) - F(\mathbf{x}_i)) = E(\varphi(\mathbf{Y}))
\end{aligned}$$

as $n \rightarrow \infty$.

(ii) \Rightarrow (iv): Let $C \subset \mathbb{R}^k$ be a closed set. We define $\psi_C(\mathbf{y}) := \inf\{\|\mathbf{y} - \mathbf{x}\| : \mathbf{x} \in C\}$, $\psi : \mathbb{R}^k \rightarrow \mathbb{R}$, which is a continuous function, as well as

$\xi_i(z) := 1_{(-\infty, 0]}(z) + (1 - iz)1_{(0, i^{-1}]}(z)$, $z \in \mathbb{R}$, for every $i \in \mathbb{N}$. Hence $\vartheta_{i,C}(\mathbf{y}) := \xi_i(\psi_C(\mathbf{y}))$ is a continuous and bounded function with $\vartheta_{i,C} : \mathbb{R}^k \rightarrow \mathbb{R}$ for each $i \in \mathbb{N}$. Observe moreover that $\vartheta_{i,C}(\mathbf{y}) \geq \vartheta_{i+1,C}(\mathbf{y})$ for each $i \in \mathbb{N}$ and each fixed $\mathbf{y} \in \mathbb{R}^k$ as well as $\vartheta_{i,C}(\mathbf{y}) \rightarrow 1_C(\mathbf{y})$ as $i \rightarrow \infty$ for each $\mathbf{y} \in \mathbb{R}^k$. From (ii), we consequently obtain

$$\limsup_{n \rightarrow \infty} P(\{\mathbf{Y}_n \in C\}) \leq \lim_{n \rightarrow \infty} E(\vartheta_{i,C}(\mathbf{Y}_n)) = E(\vartheta_{i,C}(\mathbf{Y}))$$

for each $i \in \mathbb{N}$. The dominated convergence theorem finally provides that $E(\vartheta_{i,C}(\mathbf{Y})) \rightarrow E(1_C(\mathbf{Y})) = P(\{\mathbf{Y} \in C\})$ as $i \rightarrow \infty$.

(iv) and (v) are equivalent since the complement O^c of an open set $O \subset \mathbb{R}^k$ is closed.

(iv), (v) \Rightarrow (i): (iv) and (v) imply that

$$\begin{aligned} P(\{\mathbf{Y} \in (-\infty, \mathbf{y})\}) &\leq \liminf_{n \rightarrow \infty} P(\{\mathbf{Y}_n \in (-\infty, \mathbf{y})\}) \\ &\leq \liminf_{n \rightarrow \infty} F_n(\mathbf{y}) \leq \limsup_{n \rightarrow \infty} F_n(\mathbf{y}) \\ &= \limsup_{n \rightarrow \infty} P(\{\mathbf{Y}_n \in (-\infty, \mathbf{y}]\}) \\ &\leq P(\{\mathbf{Y} \in (-\infty, \mathbf{y}]\}) = F(\mathbf{y}). \end{aligned} \quad (7.10)$$

If \mathbf{y} is a continuity point of $F(\mathbf{x})$, i.e., $P(\{\mathbf{Y} \in (-\infty, \mathbf{y})\}) = F(\mathbf{y})$, then above inequality (7.10) shows $\lim_{n \rightarrow \infty} F_n(\mathbf{y}) = F(\mathbf{y})$.

(i) \Leftrightarrow (iii): (i) \Rightarrow (iii) is provided by (i) \Rightarrow (ii). Since $\vartheta_{i,C}(\mathbf{y}) = \xi_i(\psi_C(\mathbf{y}))$ from step (ii) \Rightarrow (iv) is moreover a uniformly continuous function, the equivalence is shown. \square

Theorem 7.2.3. (*Continuity Theorem*) A sequence $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ of real valued random k -vectors, all defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$, converges in distribution to a random k -vector \mathbf{Y} as $n \rightarrow \infty$ iff the corresponding sequence of characteristic functions $(\phi_{\mathbf{Y}_n}(\mathbf{t}))_{n \in \mathbb{N}}$ converges pointwise to the characteristic function $\phi_{\mathbf{Y}}(\mathbf{t})$ of the distribution function of \mathbf{Y} for each $\mathbf{t} = (t_1, t_2, \dots, t_k) \in \mathbb{R}^k$.

Proof. \Rightarrow : Suppose that $(\mathbf{Y}_n)_{n \in \mathbb{N}} \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$. Since both the real and imaginary part of $\phi_{\mathbf{Y}_n}(\mathbf{t}) = E(\exp(i\mathbf{t}^T \mathbf{Y}_n)) = E(\cos(\mathbf{t}^T \mathbf{Y}_n)) + i E(\sin(\mathbf{t}^T \mathbf{Y}_n))$ is a bounded and continuous function, the previous Theorem 7.2.2 immediately provides that $\phi_{\mathbf{Y}_n}(\mathbf{t}) \rightarrow \phi_{\mathbf{Y}}(\mathbf{t})$ as $n \rightarrow \infty$ for every fixed $\mathbf{t} \in \mathbb{R}^k$.

\Leftarrow : Let $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}$ be a uniformly continuous and bounded function with $|\vartheta| \leq M, M \in \mathbb{R}^+$. Thus, for arbitrary $\varepsilon > 0$, we find $\delta > 0$ such that $|\mathbf{y} - \mathbf{x}| < \delta \Rightarrow |\vartheta(\mathbf{y}) - \vartheta(\mathbf{x})| < \varepsilon$ for all $\mathbf{y}, \mathbf{x} \in \mathbb{R}^k$. With Theorem 7.2.2 it suffices to show that $E(\vartheta(\mathbf{Y}_n)) \rightarrow E(\vartheta(\mathbf{Y}))$ as $n \rightarrow \infty$. Consider therefore $\mathbf{Y}_n + \sigma \mathbf{X}$ and $\mathbf{Y} + \sigma \mathbf{X}$, where $\sigma > 0$ and \mathbf{X} is a k -dimensional standard normal vector independent of \mathbf{Y} and \mathbf{Y}_n for all $n \in \mathbb{N}$. So, the triangle inequality provides

$$\begin{aligned} |E(\vartheta(\mathbf{Y}_n)) - E(\vartheta(\mathbf{Y}))| &\leq |E(\vartheta(\mathbf{Y}_n)) - E(\vartheta(\mathbf{Y}_n + \sigma \mathbf{X}))| \\ &\quad + |E(\vartheta(\mathbf{Y}_n + \sigma \mathbf{X})) - E(\vartheta(\mathbf{Y} + \sigma \mathbf{X}))| \\ &\quad + |E(\vartheta(\mathbf{Y} + \sigma \mathbf{X})) - E(\vartheta(\mathbf{Y}))|, \end{aligned} \quad (7.11)$$

where the first term on the right-hand side satisfies, for σ sufficiently small,

$$\begin{aligned} &|E(\vartheta(\mathbf{Y}_n)) - E(\vartheta(\mathbf{Y}_n + \sigma \mathbf{X}))| \\ &\leq E(|\vartheta(\mathbf{Y}_n) - \vartheta(\mathbf{Y}_n + \sigma \mathbf{X})| 1_{[-\delta, \delta]}(|\sigma \mathbf{X}|)) \\ &\quad + E(|\vartheta(\mathbf{Y}_n) - \vartheta(\mathbf{Y}_n + \sigma \mathbf{X})| 1_{(-\infty, -\delta) \cup (\delta, \infty)}(|\sigma \mathbf{X}|)) \\ &< \varepsilon + 2MP\{|\sigma \mathbf{X}| > \delta\} < 2\varepsilon. \end{aligned}$$

Analogously, the third term on the right-hand side in (7.11) is bounded by 2ε for sufficiently small σ . Hence, it remains to show that $E(\vartheta(\mathbf{Y}_n + \sigma \mathbf{X})) \rightarrow E(\vartheta(\mathbf{Y} + \sigma \mathbf{X}))$ as $n \rightarrow \infty$. By Fubini's Theorem and substituting $\mathbf{z} = \mathbf{y} + \mathbf{x}$, we attain (Pollard, 1984, p. 54)

$$\begin{aligned} &E(\vartheta(\mathbf{Y}_n + \sigma \mathbf{X})) \\ &= (2\pi\sigma^2)^{-k/2} \int_{\mathbb{R}^k} \int_{\mathbb{R}^k} \vartheta(\mathbf{y} + \mathbf{x}) \exp\left(-\frac{\mathbf{x}^T \mathbf{x}}{2\sigma^2}\right) d\mathbf{x} dF_n(\mathbf{y}) \\ &= (2\pi\sigma^2)^{-k/2} \int_{\mathbb{R}^k} \int_{\mathbb{R}^k} \vartheta(\mathbf{z}) \exp\left(-\frac{1}{2\sigma^2}(\mathbf{z} - \mathbf{y})^T(\mathbf{z} - \mathbf{y})\right) dF_n(\mathbf{y}) d\mathbf{z}, \end{aligned}$$

where F_n denotes the distribution function of \mathbf{Y}_n . Since the characteristic function of $\sigma \mathbf{X}$ is bounded and is given by

$$\begin{aligned} \phi_{\sigma \mathbf{X}}(\mathbf{t}) &= E(\exp(it^T \sigma \mathbf{X})) \\ &= \int_{\mathbb{R}^k} \exp(it^T \mathbf{x}) (2\pi\sigma^2)^{-k/2} \exp\left(-\frac{\mathbf{x}^T \mathbf{x}}{2\sigma^2}\right) d\mathbf{x} \\ &= \exp\left(-\frac{1}{2} \mathbf{t}^T \mathbf{t} \sigma^2\right), \end{aligned}$$

due to the Gaussian integral, we finally obtain the desired assertion from the dominated convergence theorem and Fubini's Theorem,

$$\begin{aligned}
& \mathbb{E}(\vartheta(\mathbf{Y}_n + \sigma \mathbf{X})) \\
&= (2\pi\sigma^2)^{-k/2} \int_{\mathbb{R}^k} \int_{\mathbb{R}^k} \vartheta(\mathbf{z}) \left(\frac{\sigma^2}{2\pi}\right)^{k/2} \\
&\quad \int_{\mathbb{R}^k} \exp\left(i\mathbf{v}^T(\mathbf{z} - \mathbf{y}) - \frac{\sigma^2}{2}\mathbf{v}^T\mathbf{v}\right) d\mathbf{v} dF_n(\mathbf{y}) d\mathbf{z} \\
&= (2\pi)^{-k} \int_{\mathbb{R}^k} \vartheta(\mathbf{z}) \int_{\mathbb{R}^k} \phi_{Y_n}(-\mathbf{v}) \exp\left(i\mathbf{v}^T\mathbf{z} - \frac{\sigma^2}{2}\mathbf{v}^T\mathbf{v}\right) d\mathbf{v} d\mathbf{z} \\
&\rightarrow (2\pi)^{-k} \int_{\mathbb{R}^k} \vartheta(\mathbf{z}) \int_{\mathbb{R}^k} \phi_Y(-\mathbf{v}) \exp\left(i\mathbf{v}^T\mathbf{z} - \frac{\sigma^2}{2}\mathbf{v}^T\mathbf{v}\right) d\mathbf{v} d\mathbf{z} \\
&= \mathbb{E}(\vartheta(\mathbf{Y} + \sigma \mathbf{X}))
\end{aligned}$$

as $n \rightarrow \infty$. □

Theorem 7.2.4. (*Cramer–Wold Device*) Let $(\mathbf{Y}_n)_{n \in \mathbb{N}}$, be a sequence of real valued random k -vectors and let \mathbf{Y} be a real valued random k -vector, all defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$. Then, $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$ iff $\boldsymbol{\lambda}^T \mathbf{Y}_n \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{Y}$ as $n \rightarrow \infty$ for every $\boldsymbol{\lambda} \in \mathbb{R}^k$.

Proof. Suppose that $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$. Hence, for arbitrary $\boldsymbol{\lambda} \in \mathbb{R}^k, t \in \mathbb{R}$,

$$\phi_{\boldsymbol{\lambda}^T \mathbf{Y}_n}(t) = \mathbb{E}(\exp(it\boldsymbol{\lambda}^T \mathbf{Y}_n)) = \phi_{Y_n}(t\boldsymbol{\lambda}) \rightarrow \phi_Y(t\boldsymbol{\lambda}) = \phi_{\boldsymbol{\lambda}^T \mathbf{Y}}(t)$$

as $n \rightarrow \infty$, showing $\boldsymbol{\lambda}^T \mathbf{Y}_n \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{Y}$ as $n \rightarrow \infty$.

If conversely $\boldsymbol{\lambda}^T \mathbf{Y}_n \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{Y}$ as $n \rightarrow \infty$ for each $\boldsymbol{\lambda} \in \mathbb{R}^k$, then we have

$$\phi_{Y_n}(\boldsymbol{\lambda}) = \mathbb{E}(\exp(i\boldsymbol{\lambda}^T \mathbf{Y}_n)) = \phi_{\boldsymbol{\lambda}^T \mathbf{Y}_n}(1) \rightarrow \phi_{\boldsymbol{\lambda}^T \mathbf{Y}}(1) = \phi_Y(\boldsymbol{\lambda})$$

as $n \rightarrow \infty$, which provides $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$. □

Definition 7.2.5. A sequence $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ of real valued random k -vectors, all defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, is said to be *asymptotically normal* with *asymptotic mean vector* $\boldsymbol{\mu}_n$ and *asymptotic covariance matrix* $\boldsymbol{\Sigma}_n$, written as $\mathbf{Y}_n \approx N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, if for all n sufficiently large $\boldsymbol{\lambda}^T \mathbf{Y}_n \approx N(\boldsymbol{\lambda}^T \boldsymbol{\mu}_n, \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_n \boldsymbol{\lambda})$ for every $\boldsymbol{\lambda} \in \mathbb{R}^k, \boldsymbol{\lambda} \neq 0$, and if $\boldsymbol{\Sigma}_n$ is symmetric and positive definite for all these sufficiently large n .

Lemma 7.2.6. Consider a sequence $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ of real valued random k -vectors $\mathbf{Y}_n := (Y_{n1}, \dots, Y_{nk})^T$, all defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, with corresponding distribution functions F_n . If $\mathbf{Y}_n \xrightarrow{D} \mathbf{c}$ as $n \rightarrow \infty$, where $\mathbf{c} := (c_1, \dots, c_k)^T$ is a constant vector in \mathbb{R}^k , then $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$ as $n \rightarrow \infty$.

Proof. In the one-dimensional case $k = 1$, we have $Y_n \xrightarrow{D} c$ as $n \rightarrow \infty$, or, respectively, $F_n(t) \rightarrow 1_{[c, \infty)}(t)$ as $n \rightarrow \infty$ for all $t \neq c$. Thereby, for any $\varepsilon > 0$,

$$\begin{aligned} \lim_{n \rightarrow \infty} P(|Y_n - c| \leq \varepsilon) &= \lim_{n \rightarrow \infty} P(c - \varepsilon \leq Y_n \leq c + \varepsilon) \\ &= 1_{[c, \infty)}(c + \varepsilon) - 1_{[c, \infty)}(c - \varepsilon) = 1, \end{aligned}$$

showing $Y_n \xrightarrow{P} c$ as $n \rightarrow \infty$. For the multidimensional case $k > 1$, we obtain $Y_{ni} \xrightarrow{D} c_i$ as $n \rightarrow \infty$ for each $i = 1, \dots, k$ by the Cramer Wold device. Thus, we have k one-dimensional cases, which leads to $Y_{ni} \xrightarrow{P} c_i$ as $n \rightarrow \infty$ for each $i = 1, \dots, k$, providing $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$ as $n \rightarrow \infty$. \square

In order to derive weak convergence for a sequence $(\mathbf{X}_n)_{n \in \mathbb{N}}$ of real valued random vectors, it is often easier to show the weak convergence for an approximative sequence $(\mathbf{Y}_n)_{n \in \mathbb{N}}$. Lemma 7.2.8 will show that both sequences share the same limit distribution if $\mathbf{Y}_n - \mathbf{X}_n \xrightarrow{P} 0$ for $n \rightarrow \infty$. The following *Basic Approximation Theorem* uses a similar strategy, where the approximative sequence is approximated by a further subsequence.

Theorem 7.2.7. Let $(\mathbf{X}_n)_{n \in \mathbb{N}}$, $(\mathbf{Y}_m)_{m \in \mathbb{N}}$ and $(\mathbf{Y}_{mn})_{n \in \mathbb{N}}$ for each $m \in \mathbb{N}$ be sequences of real valued random vectors, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$, such that

- (i) $\mathbf{Y}_{mn} \xrightarrow{D} \mathbf{Y}_m$ as $n \rightarrow \infty$ for each m ,
- (ii) $\mathbf{Y}_m \xrightarrow{D} \mathbf{Y}$ as $m \rightarrow \infty$ and
- (iii) $\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P\{|\mathbf{X}_n - \mathbf{Y}_{mn}| > \varepsilon\} = 0$ for every $\varepsilon > 0$.

Then, $\mathbf{X}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$.

Proof. By the Continuity Theorem 7.2.3, we need to show that

$$|\phi_{X_n}(\mathbf{t}) - \phi_Y(\mathbf{t})| \rightarrow 0 \text{ as } n \rightarrow \infty$$

for each $\mathbf{t} \in \mathbb{R}^k$. The triangle inequality gives

$$\begin{aligned} |\phi_{X_n}(\mathbf{t}) - \phi_Y(\mathbf{t})| &\leq |\phi_{X_n}(\mathbf{t}) - \phi_{Y_{mn}}(\mathbf{t})| + |\phi_{Y_{mn}}(\mathbf{t}) - \phi_{Y_m}(\mathbf{t})| \\ &\quad + |\phi_{Y_m}(\mathbf{t}) - \phi_Y(\mathbf{t})|, \end{aligned} \quad (7.12)$$

where the first term on the right-hand side satisfies, for $\delta > 0$,

$$\begin{aligned} &|\phi_{X_n}(\mathbf{t}) - \phi_{Y_{mn}}(\mathbf{t})| \\ &= |\mathbb{E}(\exp(i\mathbf{t}^T \mathbf{X}_n) - \exp(i\mathbf{t}^T \mathbf{Y}_{mn}))| \\ &\leq \mathbb{E}(|\exp(i\mathbf{t}^T \mathbf{X}_n)(1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n)))|) \\ &= \mathbb{E}(|1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n))|) \\ &= \mathbb{E}(|1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n))| 1_{(-\delta, \delta)}(|\mathbf{Y}_{mn} - \mathbf{X}_n|)) \\ &\quad + \mathbb{E}(|1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n))| 1_{(-\infty, -\delta] \cup [\delta, \infty)}(|\mathbf{Y}_{mn} - \mathbf{X}_n|)). \end{aligned} \quad (7.13)$$

Now, given $\mathbf{t} \in \mathbb{R}^k$ and $\varepsilon > 0$, we choose $\delta > 0$ such that

$$|\exp(i\mathbf{t}^T \mathbf{x}) - \exp(i\mathbf{t}^T \mathbf{y})| < \varepsilon \text{ if } |\mathbf{x} - \mathbf{y}| < \delta,$$

which implies

$$\mathbb{E}(|1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n))| 1_{(-\delta, \delta)}(|\mathbf{Y}_{mn} - \mathbf{X}_n|)) < \varepsilon.$$

Moreover, we find

$$|1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n))| \leq 2.$$

This shows the upper boundedness

$$\begin{aligned} &\mathbb{E}(|1 - \exp(i\mathbf{t}^T (\mathbf{Y}_{mn} - \mathbf{X}_n))| 1_{(-\infty, -\delta] \cup [\delta, \infty)}(|\mathbf{Y}_{mn} - \mathbf{X}_n|)) \\ &\leq 2P\{|\mathbf{Y}_{mn} - \mathbf{X}_n| \geq \delta\}. \end{aligned}$$

Hence, by property (iii), we have

$$\limsup_{n \rightarrow \infty} |\phi_{X_n}(\mathbf{t}) - \phi_{Y_{mn}}(\mathbf{t})| \rightarrow 0 \text{ as } m \rightarrow \infty.$$

Assumption (ii) guarantees that the last term in (7.12) vanishes as $m \rightarrow \infty$. For any $\eta > 0$, we can therefore choose m such that the upper limits of the first and last term on the right-hand side of (7.12) are both less than $\eta/2$ as $n \rightarrow \infty$. For this fixed m , assumption (i) provides $\lim_{n \rightarrow \infty} |\phi_{Y_{mn}}(\mathbf{t}) - \phi_{Y_m}(\mathbf{t})| = 0$. Altogether,

$$\limsup_{n \rightarrow \infty} |\phi_{X_n}(\mathbf{t}) - \phi_Y(\mathbf{t})| < \eta/2 + \eta/2 = \eta,$$

which completes the proof, since η was chosen arbitrarily. \square

Lemma 7.2.8. *Consider sequences of real valued random k -vectors $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ and $(\mathbf{X}_n)_{n \in \mathbb{N}}$, all defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, such that $\mathbf{Y}_n - \mathbf{X}_n \xrightarrow{P} \mathbf{0}$ as $n \rightarrow \infty$. If there exists a random k -vector \mathbf{Y} such that $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$, then $\mathbf{X}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$.*

Proof. Similar to the proof of Theorem 7.2.7, we find, for arbitrary $\mathbf{t} \in \mathbb{R}^k$ and $\varepsilon > 0$, a $\delta > 0$ satisfying

$$\begin{aligned} |\phi_{X_n}(\mathbf{t}) - \phi_{Y_n}(\mathbf{t})| &\leq E(|1 - \exp(i\mathbf{t}^T(\mathbf{Y}_n - \mathbf{X}_n))|) \\ &< \varepsilon + 2P\{|\mathbf{Y}_n - \mathbf{X}_n| \geq \delta\}. \end{aligned}$$

Consequently, $|\phi_{X_n}(\mathbf{t}) - \phi_{Y_n}(\mathbf{t})| \rightarrow 0$ as $n \rightarrow \infty$. The triangle inequality then completes the proof

$$\begin{aligned} |\phi_{X_n}(\mathbf{t}) - \phi_Y(\mathbf{t})| &\leq |\phi_{X_n}(\mathbf{t}) - \phi_{Y_n}(\mathbf{t})| + |\phi_{Y_n}(\mathbf{t}) - \phi_Y(\mathbf{t})| \\ &\rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

\square

Lemma 7.2.9. *Let $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ be a sequence of real valued random k -vectors, defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$. If $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ as $n \rightarrow \infty$, then $\vartheta(\mathbf{Y}_n) \xrightarrow{D} \vartheta(\mathbf{Y})$ as $n \rightarrow \infty$, where $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}^m$ is a continuous function.*

Proof. Since, for fixed $\mathbf{t} \in \mathbb{R}^m$, $\phi_{\vartheta(\mathbf{y})}(\mathbf{t}) = E(\exp(i\mathbf{t}^T \vartheta(\mathbf{y})))$, $\mathbf{y} \in \mathbb{R}^k$, is a bounded and continuous function of \mathbf{y} , Theorem 7.2.2 provides that $\phi_{\vartheta(\mathbf{Y}_n)}(\mathbf{t}) = E(\exp(i\mathbf{t}^T \vartheta(\mathbf{Y}_n))) \rightarrow E(\exp(i\mathbf{t}^T \vartheta(\mathbf{Y}))) = \phi_{\vartheta(\mathbf{Y})}(\mathbf{t})$ as $n \rightarrow \infty$ for each \mathbf{t} . Finally, the Continuity Theorem 7.2.3 completes the proof. \square

Lemma 7.2.10. Consider a sequence of real valued random k -vectors $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ and a sequence of real valued random m -vectors $(\mathbf{X}_n)_{n \in \mathbb{N}}$, where all random variables are defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$. If $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ and $\mathbf{X}_n \xrightarrow{P} \boldsymbol{\lambda}$ as $n \rightarrow \infty$, where $\boldsymbol{\lambda}$ is a constant m -vector, then $(\mathbf{Y}_n^T, \mathbf{X}_n^T)^T \xrightarrow{D} (\mathbf{Y}^T, \boldsymbol{\lambda}^T)^T$ as $n \rightarrow \infty$.

Proof. Defining $\mathbf{W}_n := (\mathbf{Y}_n^T, \boldsymbol{\lambda}^T)^T$, we get $\mathbf{W}_n \xrightarrow{D} (\mathbf{Y}^T, \boldsymbol{\lambda}^T)^T$ as well as $\mathbf{W}_n - (\mathbf{Y}_n^T, \mathbf{X}_n^T)^T \xrightarrow{P} \mathbf{0}$ as $n \rightarrow \infty$. The assertion follows now from Lemma 7.2.8. \square

Lemma 7.2.11. Let $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ and $(\mathbf{X}_n)_{n \in \mathbb{N}}$ be sequences of random k -vectors, where all random variables are defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$. If $\mathbf{Y}_n \xrightarrow{D} \mathbf{Y}$ and $\mathbf{X}_n \xrightarrow{D} \boldsymbol{\lambda}$ as $n \rightarrow \infty$, where $\boldsymbol{\lambda}$ is a constant k -vector, then $\mathbf{X}_n^T \mathbf{Y}_n \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{Y}$ as $n \rightarrow \infty$.

Proof. The assertion follows directly from Lemma 7.2.6, 7.2.10 and 7.2.9 by applying the continuous function $\vartheta : \mathbb{R}^{2k} \rightarrow \mathbb{R}$, $\vartheta((\mathbf{x}^T, \mathbf{y}^T)^T) := \mathbf{x}^T \mathbf{y}$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$. \square

Strictly Stationary M -Dependent Sequences

Lemma 7.2.12. Consider the process $Y_t := \sum_{u=-\infty}^{\infty} b_u Z_{t-u}$, $t \in \mathbb{Z}$, where $(b_u)_{u \in \mathbb{Z}}$ is an absolutely summable filter of real valued numbers and where $(Z_t)_{t \in \mathbb{Z}}$ is a process of square integrable, independent and identically distributed, real valued random variables with expectation $E(Z_t) = \mu$. Then $\bar{Y}_n := \frac{1}{n} \sum_{t=1}^n Y_t \xrightarrow{P} \mu \sum_{u=-\infty}^{\infty} b_u$ as $n \in \mathbb{N}$, $n \rightarrow \infty$.

Proof. We approximate \bar{Y}_n by

$$Y_{nm} := \frac{1}{n} \sum_{t=1}^n \sum_{|u| \leq m} b_u Z_{t-u}.$$

Due to the weak law of large numbers, which provides the convergence in probability $\frac{1}{n} \sum_{t=1}^n Z_{t-u} \xrightarrow{P} \mu$ as $n \rightarrow \infty$, we find $Y_{nm} \xrightarrow{P} \mu \sum_{|u| \leq m} b_u$ as $n \rightarrow \infty$. Defining now $Y_m := \mu \sum_{|u| \leq m} b_u$, entailing $Y_m \rightarrow \mu \sum_{u=-\infty}^{\infty} b_u$ as $m \rightarrow \infty$, it remains to show that

$$\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P\{|\bar{Y}_n - Y_{nm}| > \varepsilon\} = 0$$

for every $\varepsilon > 0$, since then the assertion follows at once from Theorem 7.2.7 and 7.2.6. Note that Y_m and $\mu \sum_{u=-\infty}^{\infty} b_u$ are constant numbers. By Markov's inequality, we attain the required condition

$$\begin{aligned} P\{|\bar{Y}_n - Y_{nm}| > \varepsilon\} &= P\left\{\left|\frac{1}{n} \sum_{t=1}^n \sum_{|u|>m} b_u Z_{t-u}\right| > \varepsilon\right\} \\ &\leq \frac{1}{\varepsilon} \sum_{|u|>m} |b_u| E(|Z_1|) \rightarrow 0 \end{aligned}$$

as $m \rightarrow \infty$, since $|b_u| \rightarrow 0$ as $u \rightarrow \infty$, due to the absolute summability of $(b_u)_{u \in \mathbb{Z}}$. \square

Lemma 7.2.13. *Consider the process $Y_t = \sum_{u=-\infty}^{\infty} b_u Z_{t-u}$ of the previous Lemma 7.2.12 with $E(Z_t) = \mu = 0$ and variance $E(Z_t^2) = \sigma^2 > 0$, $t \in \mathbb{Z}$. Then, $\tilde{\gamma}_n(k) := \frac{1}{n} \sum_{t=1}^n Y_t Y_{t+k} \xrightarrow{P} \gamma(k)$ for $k \in \mathbb{N}$ as $n \in \mathbb{N}, n \rightarrow \infty$, where $\gamma(k)$ denotes the autocovariance function of Y_t .*

Proof. Simple algebra gives

$$\begin{aligned} \tilde{\gamma}_n(k) &= \frac{1}{n} \sum_{t=1}^n Y_t Y_{t+k} = \frac{1}{n} \sum_{t=1}^n \sum_{u=-\infty}^{\infty} \sum_{w=-\infty}^{\infty} b_u b_w Z_{t-u} Z_{t-w+k} \\ &= \frac{1}{n} \sum_{t=1}^n \sum_{u=-\infty}^{\infty} b_u b_{u+k} Z_{t-u}^2 + \frac{1}{n} \sum_{t=1}^n \sum_{u=-\infty}^{\infty} \sum_{w \neq u+k} b_u b_w Z_{t-u} Z_{t-w+k}. \end{aligned}$$

The first term converges in probability to $\sigma^2(\sum_{u=-\infty}^{\infty} b_u b_{u+k}) = \gamma(k)$ by Lemma 7.2.12 and $\sum_{u=-\infty}^{\infty} |b_u b_{u+k}| \leq \sum_{u=-\infty}^{\infty} |b_u| \sum_{w=-\infty}^{\infty} |b_{w+k}| < \infty$. It remains to show that

$$W_n := \frac{1}{n} \sum_{t=1}^n \sum_{u=-\infty}^{\infty} \sum_{w \neq u+k} b_u b_w Z_{t-u} Z_{t-w+k} \xrightarrow{P} 0$$

as $n \rightarrow \infty$. We approximate W_n by

$$W_{nm} := \frac{1}{n} \sum_{t=1}^n \sum_{|u| \leq m} \sum_{|w| \leq m, w \neq u+k} b_u b_w Z_{t-u} Z_{t-w+k}.$$

and deduce from $E(W_{nm}) = 0$, $\text{Var}(n^{-1} \sum_{t=1}^n Z_{t-u} Z_{t-w+k}) = n^{-1} \sigma^4$, if $w \neq u + k$, and Chebychev's inequality,

$$P\{|W_{nm}| \geq \varepsilon\} \leq \varepsilon^{-2} \text{Var}(W_{nm})$$

for every $\varepsilon > 0$, that $W_{nm} \xrightarrow{P} 0$ as $n \rightarrow \infty$. By the Basic Approximation Theorem 7.2.7, it remains to show that

$$\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P\{|W_n - W_{nm}| > \varepsilon\} = 0$$

for every $\varepsilon > 0$ in order to establish $W_n \xrightarrow{P} 0$ as $n \rightarrow \infty$. Applying Markov's inequality, we attain

$$\begin{aligned} P\{|W_n - W_{nm}| > \varepsilon\} &\leq \varepsilon^{-1} E(|W_n - W_{nm}|) \\ &\leq (\varepsilon n)^{-1} \sum_{t=1}^n \sum_{|u|>m} \sum_{|w|>m, w \neq u-k} |b_u b_w| E(|Z_{t-u} Z_{t-w+k}|). \end{aligned}$$

This shows

$$\begin{aligned} &\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P\{|W_n - W_{nm}| > \varepsilon\} \\ &\leq \lim_{m \rightarrow \infty} \varepsilon^{-1} \sum_{|u|>m} \sum_{|w|>m, w \neq u-k} |b_u b_w| E(|Z_1 Z_2|) = 0, \end{aligned}$$

since $b_u \rightarrow 0$ as $u \rightarrow \infty$. □

Definition 7.2.14. A sequence $(Y_t)_{t \in \mathbb{Z}}$ of square integrable, real valued random variables is said to be *strictly stationary* if $(Y_1, \dots, Y_k)^T$ and $(Y_{1+h}, \dots, Y_{k+h})^T$ have the same joint distribution for all $k > 0$ and $h \in \mathbb{Z}$.

Observe that strict stationarity implies (weak) stationarity.

Definition 7.2.15. A strictly stationary sequence $(Y_t)_{t \in \mathbb{Z}}$ of square integrable, real valued random variables is said to be *m-dependent*, $m \geq 0$, if the two sets $\{Y_j | j \leq k\}$ and $\{Y_i | i \geq k + m + 1\}$ are independent for each $k \in \mathbb{Z}$.

In the special case of $m = 0$, *m-dependence* reduces to independence. Considering especially a $\text{MA}(q)$ -process, we find $m = q$.

Theorem 7.2.16. (*The Central Limit Theorem For Strictly Stationary M -Dependent Sequences*) Let $(Y_t)_{t \in \mathbb{Z}}$ be a strictly stationary sequence of square integrable m -dependent real valued random variables with expectation zero, $E(Y_t) = 0$. Its autocovariance function is denoted by $\gamma(k)$. Now, if $V_m := \gamma(0) + 2 \sum_{k=1}^m \gamma(k) \neq 0$, then, for $n \in \mathbb{N}$,

$$(i) \lim_{n \rightarrow \infty} \text{Var}(\sqrt{n}\bar{Y}_n) = V_m \text{ and}$$

$$(ii) \sqrt{n}\bar{Y}_n \xrightarrow{D} N(0, V_m) \text{ as } n \rightarrow \infty, \\ \text{which implies that } \bar{Y}_n \approx N(0, V_m/n) \text{ for } n \text{ sufficiently large,}$$

where $\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$.

Proof. We have

$$\begin{aligned} \text{Var}(\sqrt{n}\bar{Y}_n) &= n E \left(\left(\frac{1}{n} \sum_{i=1}^n Y_i \right) \left(\frac{1}{n} \sum_{j=1}^n Y_j \right) \right) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \gamma(i-j) \\ &= \frac{1}{n} \left(2\gamma(n-1) + 4\gamma(n-2) + \cdots + 2(n-1)\gamma(1) + n\gamma(0) \right) \\ &= \gamma(0) + 2\frac{n-1}{n}\gamma(1) + \cdots + 2\frac{1}{n}\gamma(n-1) \\ &= \sum_{|k| < n, k \in \mathbb{Z}} \left(1 - \frac{|k|}{n} \right) \gamma(k). \end{aligned}$$

The autocovariance function $\gamma(k)$ vanishes for $k > m$, due to the m -dependence of $(Y_t)_{t \in \mathbb{Z}}$. Thus, we get assertion (i),

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Var}(\sqrt{n}\bar{Y}_n) &= \lim_{n \rightarrow \infty} \sum_{|k| < n, k \in \mathbb{Z}} \left(1 - \frac{|k|}{n} \right) \gamma(k) \\ &= \sum_{|k| \leq m, k \in \mathbb{Z}} \gamma(k) = V_m. \end{aligned}$$

To prove (ii) we define random variables $\tilde{Y}_i^{(p)} := (Y_{i+1} + \cdots + Y_{i+p})$, $i \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, as the sum of p , $p > m$, sequent variables taken from the sequence $(Y_t)_{t \in \mathbb{Z}}$. Each $\tilde{Y}_i^{(p)}$ has expectation zero and variance

$$\begin{aligned} \text{Var}(\tilde{Y}_i^{(p)}) &= \sum_{l=1}^p \sum_{j=1}^p \gamma(l-j) \\ &= p\gamma(0) + 2(p-1)\gamma(1) + \cdots + 2(p-m)\gamma(m). \end{aligned}$$

Note that $\tilde{Y}_0^{(p)}, \tilde{Y}_{p+m}^{(p)}, \tilde{Y}_{2(p+m)}^{(p)}, \dots$ are independent of each other.

Defining $Y_{rp} := \tilde{Y}_0^{(p)} + \tilde{Y}_{p+m}^{(p)} + \dots + \tilde{Y}_{(r-1)(p+m)}^{(p)}$, where $r = \lfloor n/(p+m) \rfloor$ denotes the greatest integer less than or equal to $n/(p+m)$, we attain a sum of r independent, identically distributed and square integrable random variables. The central limit theorem now provides, as $r \rightarrow \infty$,

$$(r \operatorname{Var}(\tilde{Y}_0^{(p)}))^{-1/2} Y_{rp} \xrightarrow{D} N(0, 1),$$

which implies the weak convergence

$$\frac{1}{\sqrt{n}} Y_{rp} \xrightarrow{D} Y_p \quad \text{as } r \rightarrow \infty,$$

where Y_p follows the normal distribution $N(0, \operatorname{Var}(\tilde{Y}_0^{(p)})/(p+m))$. Note that $r \rightarrow \infty$ and $n \rightarrow \infty$ are nested. Since $\operatorname{Var}(\tilde{Y}_0^{(p)})/(p+m) \rightarrow V_m$ as $p \rightarrow \infty$ by the dominated convergence theorem, we attain moreover

$$Y_p \xrightarrow{D} Y \quad \text{as } p \rightarrow \infty,$$

where Y is $N(0, V_m)$ -distributed. In order to apply Theorem 7.2.7, we have to show that the condition

$$\lim_{p \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left\{ \left| \sqrt{n} \bar{Y}_n - \frac{1}{\sqrt{n}} Y_{rp} \right| > \varepsilon \right\} = 0$$

holds for every $\varepsilon > 0$. The term $\sqrt{n} \bar{Y}_n - \frac{1}{\sqrt{n}} Y_{rp}$ is a sum of r independent random variables

$$\begin{aligned} \sqrt{n} \bar{Y}_n - \frac{1}{\sqrt{n}} Y_{rp} = & \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{r-1} \left(Y_{i(p+m)-m+1} + Y_{i(p+m)-m+2} + \dots \right. \right. \\ & \left. \left. \dots + Y_{i(p+m)} \right) + (Y_{r(p+m)-m+1} + \dots + Y_n) \right) \end{aligned}$$

with variance $\operatorname{Var}(\sqrt{n} \bar{Y}_n - \frac{1}{\sqrt{n}} Y_{rp}) = \frac{1}{n} ((r-1) \operatorname{Var}(Y_1 + \dots + Y_m) + \operatorname{Var}(Y_1 + \dots + Y_{m+n-r(p+m)}))$. From Chebychev's inequality, we know

$$P \left\{ \left| \sqrt{n} \bar{Y}_n - \frac{1}{\sqrt{n}} Y_{rp} \right| \geq \varepsilon \right\} \leq \varepsilon^{-2} \operatorname{Var} \left(\sqrt{n} \bar{Y}_n - \frac{1}{\sqrt{n}} Y_{rp} \right).$$

Since the term $m+n-r(p+m)$ is bounded by $m \leq m+n-r(p+m) < 2m+p$ independent of n , we get $\limsup_{n \rightarrow \infty} \text{Var}(\sqrt{n}\bar{Y}_n - \frac{1}{\sqrt{n}}Y_{rp}) = \frac{1}{p+m} \text{Var}(Y_1 + \cdots + Y_m)$ and can finally deduce the remaining condition

$$\begin{aligned} & \lim_{p \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left\{ \left| \sqrt{n}\bar{Y}_n - \frac{1}{\sqrt{n}}Y_{rp} \right| \geq \varepsilon \right\} \\ & \leq \lim_{p \rightarrow \infty} \frac{1}{(p+m)\varepsilon^2} \text{Var}(Y_1 + \cdots + Y_m) = 0. \end{aligned}$$

Hence, $\sqrt{n}\bar{Y}_n \xrightarrow{D} N(0, V_m)$ as $n \rightarrow \infty$. \square

Example 7.2.17. Let $(Y_t)_{t \in \mathbb{Z}}$ be the MA(q)-process $Y_t = \sum_{u=0}^q b_u \varepsilon_{t-u}$ satisfying $\sum_{u=0}^q b_u \neq 0$ and $b_0 = 1$, where the ε_t are independent, identically distributed and square integrable random variables with $E(\varepsilon_t) = 0$ and $\text{Var}(\varepsilon_t) = \sigma^2 > 0$. Since the process is a q -dependent strictly stationary sequence with

$$\gamma(k) = E \left(\left(\sum_{u=0}^q b_u \varepsilon_{t-u} \right) \left(\sum_{v=0}^q b_v \varepsilon_{t+k-v} \right) \right) = \sigma^2 \sum_{u=0}^q b_u b_{u+k}$$

for $|k| \leq q$, where $b_w = 0$ for $w > q$, we find $V_q := \gamma(0) + 2 \sum_{j=1}^q \gamma(j) = \sigma^2 (\sum_{j=0}^q b_j)^2 > 0$.

Theorem 7.2.16 (ii) then implies that $\bar{Y}_n \approx N(0, \frac{\sigma^2}{n} (\sum_{j=0}^q b_j)^2)$ for sufficiently large n .

Theorem 7.2.18. Let $Y_t = \sum_{u=-\infty}^{\infty} b_u Z_{t-u}$, $t \in \mathbb{Z}$, be a stationary process with absolutely summable real valued filter $(b_u)_{u \in \mathbb{Z}}$, where $(Z_t)_{t \in \mathbb{Z}}$ is a process of independent, identically distributed, square integrable and real valued random variables with $E(Z_t) = 0$ and $\text{Var}(Z_t) = \sigma^2 > 0$. If $\sum_{u=-\infty}^{\infty} b_u \neq 0$, then, for $n \in \mathbb{N}$,

$$\begin{aligned} \sqrt{n}\bar{Y}_n & \xrightarrow{D} N \left(0, \sigma^2 \left(\sum_{u=-\infty}^{\infty} b_u \right)^2 \right) \quad \text{as } n \rightarrow \infty, \text{ as well as} \\ \bar{Y}_n & \approx N \left(0, \frac{\sigma^2}{n} \left(\sum_{u=-\infty}^{\infty} b_u \right)^2 \right) \quad \text{for } n \text{ sufficiently large,} \end{aligned}$$

where $\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$.

Proof. We approximate Y_t by $Y_t^{(m)} := \sum_{u=-m}^m b_u Z_{t-u}$ and let $\bar{Y}_n^{(m)} := \frac{1}{n} \sum_{t=1}^n Y_t^{(m)}$. With Theorem 7.2.16 and Example 7.2.17 above, we attain, as $n \rightarrow \infty$,

$$\sqrt{n} \bar{Y}_n^{(m)} \xrightarrow{D} Y^{(m)},$$

where $Y^{(m)}$ is $N(0, \sigma^2(\sum_{u=-m}^m b_u)^2)$ -distributed. Furthermore, since the filter $(b_u)_{u \in \mathbb{Z}}$ is absolutely summable $\sum_{u=-\infty}^{\infty} |b_u| < \infty$, we conclude by the dominated convergence theorem that, as $m \rightarrow \infty$,

$$Y^{(m)} \xrightarrow{D} Y, \quad \text{where } Y \text{ is } N\left(0, \sigma^2\left(\sum_{u=-\infty}^{\infty} b_u\right)^2\right)\text{-distributed.}$$

In order to show $\bar{Y}_n \xrightarrow{D} Y$ as $n \rightarrow \infty$ by Theorem 7.2.7 and Chebyshev's inequality, we have to proof

$$\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} \text{Var}(\sqrt{n}(\bar{Y}_n - \bar{Y}_n^{(m)})) = 0.$$

The dominated convergence theorem gives

$$\begin{aligned} & \text{Var}(\sqrt{n}(\bar{Y}_n - \bar{Y}_n^{(m)})) \\ &= n \text{Var}\left(\frac{1}{n} \sum_{t=1}^n \sum_{|u|>m} b_u Z_{t-u}\right) \\ &\rightarrow \sigma^2\left(\sum_{|u|>m} b_u\right)^2 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Hence, for every $\varepsilon > 0$, we get

$$\begin{aligned} & \lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P\{|\sqrt{n}(\bar{Y}_n - \bar{Y}_n^{(m)})| \geq \varepsilon\} \\ & \leq \lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} \frac{1}{\varepsilon^2} \text{Var}(\sqrt{n}(\bar{Y}_n - \bar{Y}_n^{(m)})) = 0, \end{aligned}$$

since $b_u \rightarrow 0$ as $u \rightarrow \infty$, showing that

$$\sqrt{n} \bar{Y}_n \xrightarrow{D} N\left(0, \sigma^2\left(\sum_{u=-\infty}^{\infty} b_u\right)^2\right)$$

as $n \rightarrow \infty$. □

Asymptotic Normality of Partial Autocorrelation Estimator

Recall the empirical autocovariance function $c(k) = \frac{1}{n} \sum_{t=1}^{n-k} (y_{t+k} - \bar{y})(y_t - \bar{y})$ for a given realization y_1, \dots, y_n of a process $(Y_t)_{t \in \mathbb{Z}}$, where $\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$. Hence, $\hat{c}_n(k) := \frac{1}{n} \sum_{t=1}^{n-k} (Y_{t+k} - \bar{Y})(Y_t - \bar{Y})$, where $\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$, is an obvious estimator of the autocovariance function at lag k . Consider now a stationary process $Y_t = \sum_{u=-\infty}^{\infty} b_u Z_{t-u}$, $t \in \mathbb{Z}$, with absolutely summable real valued filter $(b_u)_{u \in \mathbb{Z}}$, where $(Z_t)_{t \in \mathbb{Z}}$ is a process of independent, identically distributed and square integrable, real valued random variables with $E(Z_t) = 0$ and $\text{Var}(Z_t) = \sigma^2 > 0$. Then, we know from Theorem 7.2.18 that $n^{1/4} \bar{Y}_n \xrightarrow{D} 0$ as $n \rightarrow \infty$, due to the vanishing variance. Lemma 7.2.6 implies that $n^{1/4} \bar{Y}_n \xrightarrow{P} 0$ as $n \rightarrow \infty$ and, consequently, we obtain together with Lemma 7.2.13

$$\hat{c}_n(k) \xrightarrow{P} \gamma(k) \quad (7.14)$$

as $n \rightarrow \infty$. Thus, the Yule-Walker equations (7.3) of an $\text{AR}(p)$ -process satisfying above process conditions motivate the following *Yule-Walker estimator*

$$\hat{\mathbf{a}}_{k,n} := \begin{pmatrix} \hat{a}_{k1,n} \\ \vdots \\ \hat{a}_{kk,n} \end{pmatrix} = \hat{\mathbf{C}}_{k,n}^{-1} \begin{pmatrix} \hat{c}_n(1) \\ \vdots \\ \hat{c}_n(k) \end{pmatrix} \quad (7.15)$$

for $k \in \mathbb{N}$, where $\hat{\mathbf{C}}_{k,n} := (\hat{c}_n(|i-j|))_{1 \leq i,j \leq k}$ denotes an estimation of the autocovariance matrix $\mathbf{\Sigma}_k$ of k sequent process variables. $\hat{c}_n(k) \xrightarrow{P} \gamma(k)$ as $n \rightarrow \infty$ implies the convergence $\hat{\mathbf{C}}_{k,n} \xrightarrow{P} \mathbf{\Sigma}_k$ as $n \rightarrow \infty$. Since $\mathbf{\Sigma}_k$ is regular by Theorem 7.1.1, above estimator (7.15) is well defined for n sufficiently large. The k -th component of the Yule-Walker estimator $\hat{\alpha}_n(k) := \hat{a}_{kk,n}$ is the *partial autocorrelation estimator* of the partial autocorrelation coefficient $\alpha(k)$. Our aim is to derive an asymptotic distribution of this partial autocorrelation estimator or, respectively, of the corresponding Yule-Walker estimator. Since the direct derivation from (7.15) is very tedious, we will choose an approximative estimator.

Consider the $\text{AR}(p)$ -process $Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t$ satisfying the stationarity condition, where the errors ε_t are independent and

identically distributed with expectation $E(\varepsilon_t) = 0$. Since the process, written in matrix notation

$$\mathbf{Y}_n = \mathbf{X}_n \mathbf{a}_p + \mathbf{E}_n \quad (7.16)$$

with $\mathbf{Y}_n := (Y_1, Y_2, \dots, Y_n)^T$, $\mathbf{E}_n := (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T \in \mathbb{R}^n$, $\mathbf{a}_p := (a_1, a_2, \dots, a_p)^T \in \mathbb{R}^p$ and $n \times p$ -design matrix

$$\mathbf{X}_n = \begin{pmatrix} Y_0 & Y_{-1} & Y_{-2} & \dots & Y_{1-p} \\ Y_1 & Y_0 & Y_{-1} & \dots & Y_{2-p} \\ Y_2 & Y_1 & Y_0 & \dots & Y_{3-p} \\ \vdots & \vdots & \vdots & & \vdots \\ Y_{n-1} & Y_{n-2} & Y_{n-3} & \dots & Y_{n-p} \end{pmatrix},$$

has the pattern of a regression model, we define

$$\mathbf{a}_{p,n}^* := (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n \quad (7.17)$$

as a possibly appropriate approximation of $\hat{\mathbf{a}}_{p,n}$. Since, by Lemma 7.2.13,

$$\begin{aligned} \frac{1}{n} (\mathbf{X}_n^T \mathbf{X}_n) &\xrightarrow{P} \Sigma_p \text{ as well as} \\ \frac{1}{n} \mathbf{X}_n^T \mathbf{Y}_n &\xrightarrow{P} (\gamma(1), \dots, \gamma(p))^T \end{aligned}$$

as $n \rightarrow \infty$, where the covariance matrix Σ_p of p sequent process variables is regular by Theorem 7.1.1, above estimator (7.17) is well defined for n sufficiently large. Note that the first convergence in probability is again a conveniently written form of the entrywise convergence

$$\frac{1}{n} \sum_{k=1-i}^{n-i} Y_k Y_{k+i-j} \xrightarrow{P} \gamma(i-j)$$

as $n \rightarrow \infty$ of all the (i, j) -elements of $\frac{1}{n} (\mathbf{X}_n^T \mathbf{X}_n)$, $i = 1, \dots, n$ and $j = 1, \dots, p$. The next two theorems will now show that there exists a relationship in the asymptotic behavior of $\mathbf{a}_{p,n}^*$ and $\hat{\mathbf{a}}_{p,n}$.

Theorem 7.2.19. *Suppose $(Y_t)_{t \in \mathbb{Z}}$ is an $\text{AR}(p)$ -process $Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t, t \in \mathbb{Z}$, satisfying the stationarity condition, where the errors ε_t are independent and identically distributed with expectation $E(\varepsilon_t) = 0$ and variance $\text{Var}(\varepsilon_t) = \sigma^2 > 0$. Then,*

$$\sqrt{n}(\mathbf{a}_{p,n}^* - \mathbf{a}_p) \xrightarrow{D} N(\mathbf{0}, \sigma^2 \Sigma_p^{-1})$$

as $n \in \mathbb{N}, n \rightarrow \infty$, with $\mathbf{a}_{p,n}^* := (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n$ and the vector of coefficients $\mathbf{a}_p = (a_1, a_2, \dots, a_p)^T \in \mathbb{R}^p$.

Accordingly, $\mathbf{a}_{p,n}^*$ is asymptotically normal distributed with asymptotic expectation $E(\mathbf{a}_{p,n}^*) = \mathbf{a}_p$ and asymptotic covariance matrix $\frac{1}{n} \sigma^2 \Sigma_p^{-1}$ for sufficiently large n .

Proof. By (7.16) and (7.17), we have

$$\begin{aligned} \sqrt{n}(\mathbf{a}_{p,n}^* - \mathbf{a}_p) &= \sqrt{n} \left((\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T (\mathbf{X}_n \mathbf{a}_p + \mathbf{E}_n) - \mathbf{a}_p \right) \\ &= n(\mathbf{X}_n^T \mathbf{X}_n)^{-1} (n^{-1/2} \mathbf{X}_n^T \mathbf{E}_n). \end{aligned}$$

Lemma 7.2.13 already implies $\frac{1}{n} \mathbf{X}_n^T \mathbf{X}_n \xrightarrow{P} \Sigma_p$ as $n \rightarrow \infty$. As Σ_p is a matrix of constant elements, it remains to show that

$$n^{-1/2} \mathbf{X}_n^T \mathbf{E}_n \xrightarrow{D} N(\mathbf{0}, \sigma^2 \Sigma_p)$$

as $n \rightarrow \infty$, since then, the assertion follows directly from Lemma 7.2.11. Defining $\mathbf{W}_t := (Y_{t-1} \varepsilon_t, \dots, Y_{t-p} \varepsilon_t)^T$, we can rewrite the term as

$$n^{-1/2} \mathbf{X}_n^T \mathbf{E}_n = n^{-1/2} \sum_{t=1}^n \mathbf{W}_t.$$

The considered process $(Y_t)_{t \in \mathbb{Z}}$ satisfies the stationarity condition. So, Theorem 2.2.3 provides the almost surely stationary solution $Y_t = \sum_{u \geq 0} b_u \varepsilon_{t-u}, t \in \mathbb{Z}$. Consequently, we are able to approximate Y_t by $Y_t^{(m)} := \sum_{u=0}^m b_u \varepsilon_{t-u}$ and furthermore \mathbf{W}_t by the term $\mathbf{W}_t^{(m)} := (Y_{t-1}^{(m)} \varepsilon_t, \dots, Y_{t-p}^{(m)} \varepsilon_t)^T$. Taking an arbitrary vector $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_p)^T \in \mathbb{R}^p, \boldsymbol{\lambda} \neq \mathbf{0}$, we gain a strictly stationary $(m+p)$ -dependent sequence

$(R_t^{(m)})_{t \in \mathbb{Z}}$ defined by

$$\begin{aligned} R_t^{(m)} &:= \boldsymbol{\lambda}^T \mathbf{W}_t^{(m)} = \lambda_1 Y_{t-1}^{(m)} \varepsilon_t + \cdots + \lambda_p Y_{t-p}^{(m)} \varepsilon_t \\ &= \lambda_1 \left(\sum_{u=0}^m b_u \varepsilon_{t-u-1} \right) \varepsilon_t + \lambda_2 \left(\sum_{u=0}^m b_u \varepsilon_{t-u-2} \right) \varepsilon_t + \cdots \\ &\quad + \lambda_p \left(\sum_{u=0}^m b_u \varepsilon_{t-u-p} \right) \varepsilon_t \end{aligned}$$

with expectation $E(R_t^{(m)}) = 0$ and variance

$$\text{Var}(R_t^{(m)}) = \boldsymbol{\lambda}^T \text{Var}(\mathbf{W}_t^{(m)}) \boldsymbol{\lambda} = \sigma^2 \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_p^{(m)} \boldsymbol{\lambda} > 0,$$

where $\boldsymbol{\Sigma}_p^{(m)}$ is the regular covariance matrix of $(Y_{t-1}^{(m)}, \dots, Y_{t-p}^{(m)})^T$, due to Theorem 7.1.1 and $b_0 = 1$. In order to apply Theorem 7.2.16 to the sequence $(R_t^{(m)})_{t \in \mathbb{Z}}$, we have to check if the autocovariance function of $R_t^{(m)}$ satisfies the required condition $V_m := \gamma(0) + 2 \sum_{k=1}^m \gamma(k) \neq 0$. In view of the independence of the errors ε_t , we obtain, for $k \neq 0$, $\gamma(k) = E(R_t^{(m)} R_{t+k}^{(m)}) = 0$. Consequently, $V_m = \gamma(0) = \text{Var}(R_t^{(m)}) > 0$ and it follows

$$n^{-1/2} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t^{(m)} \xrightarrow{D} N(0, \sigma^2 \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_p^{(m)} \boldsymbol{\lambda})$$

as $n \rightarrow \infty$. Thus, applying the Cramer–Wold device 7.2.4 leads to

$$n^{-1/2} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t^{(m)} \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{U}^{(m)}$$

as $n \rightarrow \infty$, where $\mathbf{U}^{(m)}$ is $N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_p^{(m)})$ -distributed. Since, entrywise, $\sigma^2 \boldsymbol{\Sigma}_p^{(m)} \rightarrow \sigma^2 \boldsymbol{\Sigma}_p$ as $m \rightarrow \infty$, we attain $\boldsymbol{\lambda}^T \mathbf{U}^{(m)} \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{U}$ as $m \rightarrow \infty$ by the dominated convergence theorem, where \mathbf{U} follows the normal distribution $N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_p)$. With Theorem 7.2.7, it only remains to show that, for every $\varepsilon > 0$,

$$\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left\{ \left| \frac{1}{\sqrt{n}} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t - \frac{1}{\sqrt{n}} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t^{(m)} \right| > \varepsilon \right\} = 0$$

to establish

$$\frac{1}{\sqrt{n}} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t \xrightarrow{D} \boldsymbol{\lambda}^T \mathbf{U}$$

as $n \rightarrow \infty$, which, by the Cramer–Wold device 7.2.4, finally leads to the desired result

$$\frac{1}{\sqrt{n}} \mathbf{X}_n^T \mathbf{E}_n \xrightarrow{D} N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_p)$$

as $n \rightarrow \infty$. Because of the identically distributed and independent $\mathbf{W}_t - \mathbf{W}_t^{(m)}$, $1 \leq t \leq n$, we find the variance

$$\begin{aligned} & \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t - \frac{1}{\sqrt{n}} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{W}_t^{(m)} \right) \\ &= \frac{1}{n} \text{Var} \left(\boldsymbol{\lambda}^T \sum_{t=1}^n (\mathbf{W}_t - \mathbf{W}_t^{(m)}) \right) = \boldsymbol{\lambda}^T \text{Var}(\mathbf{W}_t - \mathbf{W}_t^{(m)}) \boldsymbol{\lambda} \end{aligned}$$

being independent of n . Since almost surely $\mathbf{W}_t^{(m)} \rightarrow \mathbf{W}_t$ as $m \rightarrow \infty$, Chebychev's inequality finally gives

$$\begin{aligned} & \lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left\{ \left| \frac{1}{\sqrt{n}} \boldsymbol{\lambda}^T \sum_{t=1}^n (\mathbf{W}_t^{(m)} - \mathbf{W}_t) \right| \geq \varepsilon \right\} \\ & \leq \lim_{m \rightarrow \infty} \frac{1}{\varepsilon^2} \boldsymbol{\lambda}^T \text{Var}(\mathbf{W}_t - \mathbf{W}_t^{(m)}) \boldsymbol{\lambda} = 0. \end{aligned}$$

□

Theorem 7.2.20. *Consider the AR(p)-process from Theorem 7.2.19. Then,*

$$\sqrt{n}(\hat{\mathbf{a}}_{p,n} - \mathbf{a}_p) \xrightarrow{D} N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_p^{-1})$$

as $n \in \mathbb{N}$, $n \rightarrow \infty$, where $\hat{\mathbf{a}}_{p,n}$ is the Yule–Walker estimator and $\mathbf{a}_p = (a_1, a_2, \dots, a_p)^T \in \mathbb{R}^p$ is the vector of coefficients.

Proof. In view of the previous Theorem 7.2.19, it remains to show that the Yule-Walker estimator $\hat{\mathbf{a}}_{p,n}$ and $\mathbf{a}_{p,n}^*$ follow the same limit law. Therefore, together with Lemma 7.2.8, it suffices to prove that

$$\sqrt{n}(\hat{\mathbf{a}}_{p,n} - \mathbf{a}_{p,n}^*) \xrightarrow{P} 0$$

as $n \rightarrow \infty$. Applying the definitions, we get

$$\begin{aligned} \sqrt{n}(\hat{\mathbf{a}}_{p,n} - \mathbf{a}_{p,n}^*) &= \sqrt{n}(\hat{\mathbf{C}}_{p,n}^{-1} \hat{\mathbf{c}}_{p,n} - (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n) \\ &= \sqrt{n} \hat{\mathbf{C}}_{p,n}^{-1} (\hat{\mathbf{c}}_{p,n} - \frac{1}{n} \mathbf{X}_n^T \mathbf{Y}_n) + \sqrt{n} (\hat{\mathbf{C}}_{p,n}^{-1} - n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}) \frac{1}{n} \mathbf{X}_n^T \mathbf{Y}_n, \end{aligned} \quad (7.18)$$

where $\hat{\mathbf{c}}_{p,n} := (\hat{c}_n(1), \dots, \hat{c}_n(p))^T$. The k -th component of $\sqrt{n}(\hat{\mathbf{c}}_{p,n} - \frac{1}{n} \mathbf{X}_n^T \mathbf{Y}_n)$ is given by

$$\begin{aligned} &\frac{1}{\sqrt{n}} \left(\sum_{t=1}^{n-k} (Y_{t+k} - \bar{Y}_n)(Y_t - \bar{Y}_n) - \sum_{s=1-k}^{n-k} Y_s Y_{s+k} \right) \\ &= -\frac{1}{\sqrt{n}} \sum_{t=1-k}^0 Y_t Y_{t+k} - \frac{1}{\sqrt{n}} \bar{Y}_n \left(\sum_{s=1}^{n-k} Y_{s+k} + Y_s \right) + \frac{n-k}{\sqrt{n}} \bar{Y}_n^2, \end{aligned} \quad (7.19)$$

where the latter terms can be written as

$$\begin{aligned} &-\frac{1}{\sqrt{n}} \bar{Y}_n \left(\sum_{s=1}^{n-k} Y_{s+k} + Y_s \right) + \frac{n-k}{\sqrt{n}} \bar{Y}_n^2 \\ &= -2\sqrt{n} \bar{Y}_n^2 + \frac{n-k}{\sqrt{n}} \bar{Y}_n^2 + \frac{1}{\sqrt{n}} \bar{Y}_n \left(\sum_{t=n-k+1}^n Y_t + \sum_{s=1}^k Y_s \right) \\ &= -n^{1/4} \bar{Y}_n n^{1/4} \bar{Y}_n - \frac{k}{\sqrt{n}} \bar{Y}_n^2 + \frac{1}{\sqrt{n}} \bar{Y}_n \left(\sum_{t=n-k+1}^n Y_t + \sum_{s=1}^k Y_s \right). \end{aligned} \quad (7.20)$$

Because of the vanishing variance as $n \rightarrow \infty$, Theorem 7.2.18 gives $n^{1/4} \bar{Y}_n \xrightarrow{D} 0$ as $n \rightarrow \infty$, which leads together with Lemma 7.2.6 to $n^{1/4} \bar{Y}_n \xrightarrow{P} 0$ as $n \rightarrow \infty$, showing that (7.20) converges in probability to zero as $n \rightarrow \infty$. Consequently, (7.19) converges in probability to zero.

Focusing now on the second term in (7.18), Lemma 7.2.13 implies that

$$\frac{1}{n} \mathbf{X}_n^T \mathbf{Y}_n \xrightarrow{P} \gamma(p)$$

as $n \rightarrow \infty$. Hence, we need to show the convergence in probability

$$\sqrt{n} \|\hat{\mathbf{C}}_{p,n}^{-1} - n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}\| \xrightarrow{P} 0 \text{ as } n \rightarrow \infty,$$

where $\|\hat{\mathbf{C}}_{p,n}^{-1} - n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}\|$ denotes the Euclidean norm of the p^2 dimensional vector consisting of all entries of the $p \times p$ -matrix $\hat{\mathbf{C}}_{p,n}^{-1} - n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}$. Thus, we attain

$$\sqrt{n} \|\hat{\mathbf{C}}_{p,n}^{-1} - n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}\| \tag{7.21}$$

$$\begin{aligned} &= \sqrt{n} \|\hat{\mathbf{C}}_{p,n}^{-1}(n^{-1} \mathbf{X}_n^T \mathbf{X}_n - \hat{\mathbf{C}}_{p,n})n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}\| \\ &\leq \sqrt{n} \|\hat{\mathbf{C}}_{p,n}^{-1}\| \|n^{-1} \mathbf{X}_n^T \mathbf{X}_n - \hat{\mathbf{C}}_{p,n}\| \|n(\mathbf{X}_n^T \mathbf{X}_n)^{-1}\|, \end{aligned} \tag{7.22}$$

where

$$\begin{aligned} &\sqrt{n} \|n^{-1} \mathbf{X}_n^T \mathbf{X}_n - \hat{\mathbf{C}}_{p,n}\|^2 \\ &= n^{-1/2} \sum_{i=1}^p \sum_{j=1}^p \left(n^{-1/2} \sum_{s=1}^n Y_{s-i} Y_{s-j} \right. \\ &\quad \left. - n^{-1/2} \left(\sum_{t=1}^{n-|i-j|} (Y_{t+|i-j|} - \bar{Y}_n)(Y_t - \bar{Y}_n) \right) \right)^2. \end{aligned}$$

Regarding (7.19) with $k = |i - j|$, it only remains to show that

$$n^{-1/2} \sum_{s=1}^n Y_{s-i} Y_{s-j} - n^{-1/2} \sum_{t=1-k}^{n-k} Y_t Y_{t+k} \xrightarrow{P} 0 \text{ as } n \rightarrow \infty$$

or, equivalently,

$$n^{-1/2} \sum_{s=1-i}^{n-i} Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=1-k}^{n-k} Y_t Y_{t+k} \xrightarrow{P} 0 \text{ as } n \rightarrow \infty.$$

In the case of $k = i - j$, we find

$$\begin{aligned} & n^{-1/2} \sum_{s=1-i}^{n-i} Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=1-i+j}^{n-i+j} Y_t Y_{t+i-j} \\ &= n^{-1/2} \sum_{s=1-i}^{-i+j} Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=n-i+1}^{n-i+j} Y_t Y_{t-j+i}. \end{aligned}$$

Applied to Markov's inequality yields

$$\begin{aligned} & P \left\{ \left| n^{-1/2} \sum_{s=1-i}^{-i+j} Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=n-i+1}^{n-i+j} Y_t Y_{t-j+i} \right| \geq \varepsilon \right\} \\ & \leq P \left\{ \left| n^{-1/2} \sum_{s=1-i}^{-i+j} Y_s Y_{s-j+i} \right| \geq \varepsilon/2 \right\} \\ & \quad + P \left\{ \left| n^{-1/2} \sum_{t=n-i+1}^{n-i+j} Y_t Y_{t-j+i} \right| \geq \varepsilon/2 \right\} \\ & \leq 4(\sqrt{n}\varepsilon)^{-1} j \gamma(0) \xrightarrow{P} 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

In the case of $k = j - i$, on the other hand,

$$\begin{aligned} & n^{-1/2} \sum_{s=1-i}^{n-i} Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=1-j+i}^{n-j+i} Y_t Y_{t+j-i} \\ &= n^{-1/2} \sum_{s=1-i}^{n-i} Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=1}^n Y_{t+i-j} Y_t \end{aligned}$$

entails that

$$P \left\{ \left| n^{-1/2} \sum_{s=1-i}^0 Y_s Y_{s-j+i} - n^{-1/2} \sum_{t=n-i+1}^n Y_{t+i-j} Y_t \right| \geq \varepsilon \right\} \xrightarrow{P} 0 \text{ as } n \rightarrow \infty.$$

This completes the proof, since $\hat{\mathbf{C}}_{p,n}^{-1} \xrightarrow{P} \mathbf{\Sigma}_p^{-1}$ as well as $n(\mathbf{X}_n^T \mathbf{X}_n)^{-1} \xrightarrow{P} \mathbf{\Sigma}_p^{-1}$ as $n \rightarrow \infty$ by Lemma 7.2.13. \square

Theorem 7.2.21. : *Let $(Y_t)_{t \in \mathbb{Z}}$ be an $\text{AR}(p)$ -process $Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t, t \in \mathbb{Z}$, satisfying the stationarity condition, where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a white noise process of independent and identically distributed random variables with expectation $E(\varepsilon_t) = 0$ and variance $E(\varepsilon_t^2) = \sigma^2 > 0$. Then the partial autocorrelation estimator $\hat{\alpha}_n(k)$ of order $k > p$ based on Y_1, \dots, Y_n , is asymptotically normal distributed with expectation $E(\hat{\alpha}_n(k)) = 0$ and variance $E(\hat{\alpha}_n(k)^2) = 1/n$ for sufficiently large $n \in \mathbb{N}$.*

Proof. The $\text{AR}(p)$ -process can be regarded as an $\text{AR}(k)$ -process $\tilde{Y}_t := a_1 Y_{t-1} + \cdots + a_k Y_{t-k} + \varepsilon_t$ with $k > p$ and $a_i = 0$ for $p < i \leq k$. The partial autocorrelation estimator $\hat{\alpha}_n(k)$ is the k -th component $\hat{a}_{k,n}$ of the Yule-Walker estimator $\hat{\mathbf{a}}_{k,n}$. Defining

$$(\sigma_{ij})_{1 \leq i, j \leq k} := \Sigma_k^{-1}$$

as the inverse matrix of the covariance matrix Σ_k of k sequent process variables, Theorem 7.2.20 provides the asymptotic behavior

$$\hat{\alpha}_n(k) \approx N\left(a_k, \frac{\sigma^2}{n} \sigma_{kk}\right)$$

for n sufficiently large. We obtain from Lemma 7.1.5 with $\mathbf{y} := Y_k$ and $\mathbf{x} := (Y_1, \dots, Y_{k-1})$ that

$$\sigma_{kk} = \mathbf{V}_{\mathbf{y}\mathbf{y}, \mathbf{x}}^{-1} = \frac{1}{\text{Var}(Y_k - \hat{Y}_{k,k-1})},$$

since \mathbf{y} and so $\mathbf{V}_{\mathbf{y}\mathbf{y}, \mathbf{x}}$ have dimension one. As $(Y_t)_{t \in \mathbb{Z}}$ is an $\text{AR}(p)$ -process satisfying the Yule-Walker equations, the best linear approximation is given by $\hat{Y}_{k,k-1} = \sum_{u=1}^p a_u Y_{k-u}$ and, therefore, $Y_k - \hat{Y}_{k,k-1} = \varepsilon_k$. Thus, for $k > p$

$$\hat{\alpha}_n(k) \approx N\left(0, \frac{1}{n}\right).$$

□

7.3 Asymptotic Normality of Autocorrelation Estimator

The Landau-Symbols

A sequence $(Y_n)_{n \in \mathbb{N}}$ of real valued random variables, defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, is said to be *bounded in probability (or tight)*, if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$P\{|Y_n| > \delta\} < \varepsilon$$

for all n , conveniently notated as $Y_n = \mathcal{O}_P(1)$. Given a sequence $(h_n)_{n \in \mathbb{N}}$ of positive real valued numbers we define moreover

$$Y_n = \mathcal{O}_P(h_n)$$

if and only if Y_n/h_n is bounded in probability and

$$Y_n = o_P(h_n)$$

if and only if Y_n/h_n converges in probability to zero. A sequence $(\mathbf{Y}_n)_{n \in \mathbb{N}}$, $\mathbf{Y}_n := (Y_{n1}, \dots, Y_{nk})$, of real valued random vectors, all defined on some probability space $(\Omega, \mathcal{A}, \mathcal{P})$, is said to be bounded in probability

$$\mathbf{Y}_n = \mathcal{O}_P(h_n)$$

if and only if Y_{ni}/h_n is bounded in probability for every $i = 1, \dots, k$. Furthermore,

$$\mathbf{Y}_n = \mathbf{o}_P(h_n)$$

if and only if Y_{ni}/h_n converges in probability to zero for every $i = 1, \dots, k$.

The next four Lemmas will show basic properties concerning these Landau-symbols.

Lemma 7.3.1. *Consider two sequences $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ and $(\mathbf{X}_n)_{n \in \mathbb{N}}$ of real valued random k -vectors, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$, such that $\mathbf{Y}_n = \mathcal{O}_P(1)$ and $\mathbf{X}_n = \mathbf{o}_P(1)$. Then, $\mathbf{Y}_n \mathbf{X}_n = \mathbf{o}_P(1)$.*

Proof. Let $\mathbf{Y}_n := (Y_{n1}, \dots, Y_{nk})^T$ and $\mathbf{X}_n := (X_{n1}, \dots, X_{nk})^T$. For any fixed $i \in \{1, \dots, k\}$, arbitrary $\varepsilon > 0$ and $\delta > 0$, we have

$$\begin{aligned} P\{|Y_{ni}X_{ni}| > \varepsilon\} &= P\{|Y_{ni}X_{ni}| > \varepsilon \cap |Y_{ni}| > \delta\} \\ &\quad + P\{|Y_{ni}X_{ni}| > \varepsilon \cap |Y_{ni}| \leq \delta\} \\ &\leq P\{|Y_{ni}| > \delta\} + P\{|X_{ni}| > \varepsilon/\delta\}. \end{aligned}$$

As $Y_{ni} = \mathcal{O}_P(1)$, for arbitrary $\eta > 0$, there exists $\kappa > 0$ such that $P\{|Y_{ni}| > \kappa\} < \eta$ for all $n \in \mathbb{N}$. Choosing $\delta := \kappa$, we finally attain

$$\lim_{n \rightarrow \infty} P\{|Y_{ni}X_{ni}| > \varepsilon\} = 0,$$

since η was chosen arbitrarily. \square

Lemma 7.3.2. *Consider a sequence of real valued random k -vectors $(\mathbf{Y}_n)_{n \in \mathbb{N}}$, where all $\mathbf{Y}_n := (Y_{n1}, \dots, Y_{nk})^T$ are defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$. If $\mathbf{Y}_n = \mathcal{O}_P(h_n)$, where $h_n \rightarrow 0$ as $n \rightarrow \infty$, then $\mathbf{Y}_n = \mathbf{o}_P(1)$.*

Proof. For any fixed $i \in \{1, \dots, k\}$, we have $Y_{ni}/h_n = \mathcal{O}_P(1)$. Thus, for every $\varepsilon > 0$, there exists $\delta > 0$ with $P\{|Y_{ni}/h_n| > \delta\} = P\{|Y_{ni}| > \delta|h_n|\} < \varepsilon$ for all n . Now, for arbitrary $\eta > 0$, there exists $N \in \mathbb{N}$ such that $\delta|h_n| \leq \eta$ for all $n \geq N$. Hence, we have $P\{|Y_{ni}| > \eta\} \rightarrow 0$ as $n \rightarrow \infty$ for all $\eta > 0$. \square

Lemma 7.3.3. *Let Y be a random variable on $(\Omega, \mathcal{A}, \mathcal{P})$ with corresponding distribution function F and $(Y_n)_{n \in \mathbb{N}}$ is supposed to be a sequence of random variables on $(\Omega, \mathcal{A}, \mathcal{P})$ with corresponding distribution functions F_n such that $Y_n \xrightarrow{D} Y$ as $n \rightarrow \infty$. Then, $Y_n = \mathcal{O}_P(1)$.*

Proof. Let $\varepsilon > 0$. Due to the right-continuousness of distribution functions, we always find continuity points t_1 and $-t_1$ of F satisfying $F(t_1) > 1 - \varepsilon/4$ and $F(-t_1) < \varepsilon/4$. Moreover, there exists N_1 such that $|F_n(t_1) - F(t_1)| < \varepsilon/4$ for $n \geq N_1$, entailing $F_n(t_1) > 1 - \varepsilon/2$ for $n \geq N_1$. Analogously, we find N_2 such that $F_n(-t_1) < \varepsilon/2$ for $n \geq N_2$. Consequently, $P\{|Y_n| > t_1\} < \varepsilon$ for $n \geq \max\{N_1, N_2\}$. Since we always find a continuity point t_2 of F with $\max_{n < \max\{N_1, N_2\}} \{P\{|Y_n| > t_2\}\} < \varepsilon$, the boundedness follows: $P(|Y_n| > t) < \varepsilon$ for all n , where $t := \max\{t_1, t_2\}$. \square

Lemma 7.3.4. *Consider two sequences $(\mathbf{Y}_n)_{n \in \mathbb{N}}$ and $(\mathbf{X}_n)_{n \in \mathbb{N}}$ of real valued random k -vectors, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$, such that $\mathbf{Y}_n - \mathbf{X}_n = o_P(1)$. If furthermore $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$ as $n \rightarrow \infty$, then $\mathbf{X}_n \xrightarrow{P} \mathbf{Y}$ as $n \rightarrow \infty$.*

Proof. The triangle inequality immediately provides

$$|\mathbf{X}_n - \mathbf{Y}| \leq |\mathbf{X}_n - \mathbf{Y}_n| + |\mathbf{Y}_n - \mathbf{Y}| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Note that $\mathbf{Y}_n - \mathbf{Y} = o_P(1) \Leftrightarrow |\mathbf{Y}_n - \mathbf{Y}| = o_P(1)$, due to the Euclidean distance. \square

Taylor Series Expansion in Probability

Lemma 7.3.5. *Let $(\mathbf{Y}_n)_{n \in \mathbb{N}}$, $\mathbf{Y}_n := (Y_{n1}, \dots, Y_{nk})$, be a sequence of real valued random k -vectors, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$, and let \mathbf{c} be a constant vector in \mathbb{R}^k such that $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$ as $n \rightarrow \infty$. If the function $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}^m$ is continuous at \mathbf{c} , then $\vartheta(\mathbf{Y}_n) \xrightarrow{P} \vartheta(\mathbf{c})$ as $n \rightarrow \infty$.*

Proof. Let $\varepsilon > 0$. Since ϑ is continuous at \mathbf{c} , there exists $\delta > 0$ such that, for all n ,

$$\{|\mathbf{Y}_n - \mathbf{c}| < \delta\} \subseteq \{|\vartheta(\mathbf{Y}_n) - \vartheta(\mathbf{c})| < \varepsilon\},$$

which implies

$$P\{|\vartheta(\mathbf{Y}_n) - \vartheta(\mathbf{c})| > \varepsilon\} \leq P\{|\mathbf{Y}_n - \mathbf{c}| > \delta/2\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

\square

If we assume in addition the existence of the partial derivatives of ϑ in a neighborhood of \mathbf{c} , we attain a stochastic analogue of the Taylor series expansion.

Theorem 7.3.6. *Let $(\mathbf{Y}_n)_{n \in \mathbb{N}}$, $\mathbf{Y}_n := (Y_{n1}, \dots, Y_{nk})$, be a sequence of real valued random k -vectors, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$, and let $\mathbf{c} := (c_1, \dots, c_k)^T$ be an arbitrary constant vector in \mathbb{R}^k such that $\mathbf{Y}_n - \mathbf{c} = \mathcal{O}_P(h_n)$, where $h_n \rightarrow 0$ as $n \rightarrow \infty$. If the*

function $\vartheta : \mathbb{R}^k \rightarrow \mathbb{R}$, $\mathbf{y} \rightarrow \vartheta(\mathbf{y})$, has continuous partial derivatives $\frac{\partial \vartheta}{\partial y_i}$, $i = 1, \dots, k$, in a neighborhood of \mathbf{c} , then

$$\vartheta(\mathbf{Y}_n) = \vartheta(\mathbf{c}) + \sum_{i=1}^k \frac{\partial \vartheta}{\partial y_i}(\mathbf{c})(Y_{ni} - c_i) + o_P(h_n).$$

Proof. The Taylor series expansion (e.g. Seeley, 1970, Section 5.3) gives, as $\mathbf{y} \rightarrow \mathbf{c}$,

$$\vartheta(\mathbf{y}) = \vartheta(\mathbf{c}) + \sum_{i=1}^k \frac{\partial \vartheta}{\partial y_i}(\mathbf{c})(y_i - c_i) + o(|\mathbf{y} - \mathbf{c}|),$$

where $\mathbf{y} := (y_1, \dots, y_k)^T$. Defining

$$\varphi(\mathbf{y}) := \frac{1}{|\mathbf{y} - \mathbf{c}|} \left(\vartheta(\mathbf{y}) - \vartheta(\mathbf{c}) - \sum_{i=1}^k \frac{\partial \vartheta}{\partial y_i}(\mathbf{c})(y_i - c_i) \right) = \frac{o(|\mathbf{y} - \mathbf{c}|)}{|\mathbf{y} - \mathbf{c}|}$$

for $\mathbf{y} \neq \mathbf{c}$ and $\varphi(\mathbf{c}) = 0$, we attain a function $\varphi : \mathbb{R}^k \rightarrow \mathbb{R}$ that is continuous at \mathbf{c} . Since $\mathbf{Y}_n - \mathbf{c} = \mathcal{O}_P(h_n)$, where $h_n \rightarrow 0$ as $n \rightarrow \infty$, Lemma 7.3.2 and the definition of stochastically boundedness directly imply $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$. Together with Lemma 7.3.5, we receive therefore $\varphi(\mathbf{Y}_n) \xrightarrow{P} \varphi(\mathbf{c}) = 0$ as $n \rightarrow \infty$. Finally, from Lemma 7.3.1 and Exercise 7.1, the assertion follows: $\varphi(\mathbf{Y}_n)|\mathbf{Y}_n - \mathbf{c}| = o_P(h_n)$. \square

The Delta-Method

Theorem 7.3.7. (*The Multivariate Delta-Method*) Consider a sequence of real valued random k -vectors $(\mathbf{Y}_n)_{n \in \mathbb{N}}$, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$, such that $h_n^{-1}(\mathbf{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \rightarrow \infty$ with $\boldsymbol{\mu} := (\mu_1, \dots, \mu_k)^T \in \mathbb{R}^k$, $h_n \rightarrow 0$ as $n \rightarrow \infty$ and $\boldsymbol{\Sigma} := (\sigma_{rs})_{1 \leq r, s \leq k}$ being a symmetric and positive definite $k \times k$ -matrix. Moreover, let $\vartheta = (\vartheta_1, \dots, \vartheta_m)^T : \mathbf{y} \rightarrow \vartheta(\mathbf{y})$ be a function from \mathbb{R}^k into \mathbb{R}^m , $m \leq k$, where each ϑ_j , $1 \leq j \leq m$, is continuously differentiable in a neighborhood of $\boldsymbol{\mu}$. If

$$\Delta := \left(\frac{\partial \vartheta_j}{\partial y_i}(\mathbf{y}) \Big|_{\boldsymbol{\mu}} \right)_{1 \leq j \leq m, 1 \leq i \leq k}$$

is a $m \times k$ -matrix with $\text{rank}(\Delta) = m$, then

$$h_n^{-1}(\vartheta(\mathbf{Y}_n) - \vartheta(\boldsymbol{\mu})) \xrightarrow{D} N(\mathbf{0}, \Delta \Sigma \Delta^T)$$

as $n \rightarrow \infty$.

Proof. By Lemma 7.3.3, $\mathbf{Y}_n - \boldsymbol{\mu} = \mathcal{O}_P(h_n)$. Hence, the conditions of Theorem 7.3.6 are satisfied and we receive

$$\vartheta_j(\mathbf{Y}_n) = \vartheta_j(\boldsymbol{\mu}) + \sum_{i=1}^k \frac{\partial \vartheta_j}{\partial y_i}(\boldsymbol{\mu})(Y_{ni} - \mu_i) + o_P(h_n)$$

for $j = 1, \dots, m$. Conveniently written in matrix notation yields

$$\vartheta(\mathbf{Y}_n) - \vartheta(\boldsymbol{\mu}) = \Delta(\mathbf{Y}_n - \boldsymbol{\mu}) + o_P(h_n)$$

or, respectively,

$$h_n^{-1}(\vartheta(\mathbf{Y}_n) - \vartheta(\boldsymbol{\mu})) = h_n^{-1} \Delta(\mathbf{Y}_n - \boldsymbol{\mu}) + o_P(1).$$

We know $h_n^{-1} \Delta(\mathbf{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} N(\mathbf{0}, \Delta \Sigma \Delta^T)$ as $n \rightarrow \infty$, thus, we conclude from Lemma 7.2.8 that $h_n^{-1}(\vartheta(\mathbf{Y}_n) - \vartheta(\boldsymbol{\mu})) \xrightarrow{D} N(\mathbf{0}, \Delta \Sigma \Delta^T)$ as $n \rightarrow \infty$ as well. \square

Asymptotic Normality of Autocorrelation Estimator

Since $\hat{c}_n(k) = \frac{1}{n} \sum_{t=1}^{n-k} (Y_{t+k} - \bar{Y})(Y_t - \bar{Y})$ is an estimator of the autocovariance $\gamma(k)$ at lag $k \in \mathbb{N}$, $\hat{r}_n(k) := \hat{c}_n(k)/\hat{c}_n(0)$, $\hat{c}_n(0) \neq 0$, is an obvious estimator of the autocorrelation $\rho(k)$. As the direct derivation of an asymptotic distribution of $\hat{c}_n(k)$ or $\hat{r}_n(k)$ is a complex problem, we consider another estimator of $\gamma(k)$. Lemma 7.2.13 motivates $\tilde{\gamma}_n(k) := \frac{1}{n} \sum_{t=1}^n Y_t Y_{t+k}$ as an appropriate candidate.

Lemma 7.3.8. *Consider the stationary process $Y_t = \sum_{u=-\infty}^{\infty} b_u \varepsilon_{t-u}$, where the filter $(b_u)_{u \in \mathbb{Z}}$ is absolutely summable and $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a white noise process of independent and identically distributed random variables with expectation $E(\varepsilon_t) = 0$ and variance $E(\varepsilon_t^2) = \sigma^2 > 0$. If*

$$\begin{aligned}
E(\varepsilon_t^4) &:= \alpha\sigma^4 < \infty, \alpha > 0, \text{ then, for } l \geq 0, k \geq 0 \text{ and } n \in \mathbb{N}, \\
&\lim_{n \rightarrow \infty} n \operatorname{Cov}(\tilde{\gamma}_n(k), \tilde{\gamma}_n(l)) \\
&= (\alpha - 3)\gamma(k)\gamma(l) \\
&\quad + \sum_{m=-\infty}^{\infty} \left(\gamma(m+l)\gamma(m-k) + \gamma(m+l-k)\gamma(m) \right),
\end{aligned}$$

where $\tilde{\gamma}_n(k) = \frac{1}{n} \sum_{t=1}^n Y_t Y_{t+k}$.

Proof. The autocovariance function of $(Y_t)_{t \in \mathbb{Z}}$ is given by

$$\begin{aligned}
\gamma(k) &= E(Y_t Y_{t+k}) = \sum_{u=-\infty}^{\infty} \sum_{w=-\infty}^{\infty} b_u b_w E(\varepsilon_{t-u} \varepsilon_{t+k-w}) \\
&= \sigma^2 \sum_{u=-\infty}^{\infty} b_u b_{u+k}.
\end{aligned}$$

Observe that

$$E(\varepsilon_g \varepsilon_h \varepsilon_i \varepsilon_j) = \begin{cases} \alpha\sigma^4 & \text{if } g = h = i = j, \\ \sigma^4 & \text{if } g = h \neq i = j, g = i \neq h = j, g = j \neq h = i, \\ 0 & \text{elsewhere.} \end{cases}$$

We therefore find

$$\begin{aligned}
&E(Y_t Y_{t+s} Y_{t+s+r} Y_{t+s+r+v}) \\
&= \sum_{g=-\infty}^{\infty} \sum_{h=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} b_g b_h b_i b_j E(\varepsilon_{t-g} \varepsilon_{t-h} \varepsilon_{t-i} \varepsilon_{t-j}) \\
&= \sigma^4 \sum_{g=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} (b_g b_{g+s} b_{i+s+r} b_{i+s+r+v} + b_g b_{g+s+r} b_{i+s} b_{i+s+r+v} \\
&\quad + b_g b_{g+s+r+v} b_{i+s} b_{i+s+r}) \\
&\quad + (\alpha - 3)\sigma^4 \sum_{j=-\infty}^{\infty} b_j b_{j+s} b_{j+s+r} b_{j+s+r+v} \\
&= (\alpha - 3)\sigma^4 \sum_{g=-\infty}^{\infty} b_g b_{g+s} b_{g+s+r} b_{g+s+r+v} \\
&\quad + \gamma(s)\gamma(v) + \gamma(r+s)\gamma(r+v) + \gamma(r+s+v)\gamma(r).
\end{aligned}$$

Applying the result to the covariance of $\tilde{\gamma}_n(k)$ and $\tilde{\gamma}_n(l)$ provides

$$\begin{aligned}
& \text{Cov}(\tilde{\gamma}_n(k), \tilde{\gamma}_n(l)) \\
&= \mathbb{E}(\tilde{\gamma}_n(k)\tilde{\gamma}_n(l)) - \mathbb{E}(\tilde{\gamma}_n(k))\mathbb{E}(\tilde{\gamma}_n(l)) \\
&= \frac{1}{n^2} \mathbb{E} \left(\sum_{s=1}^n Y_s Y_{s+k} \sum_{t=1}^n Y_t Y_{t+l} \right) - \gamma(l)\gamma(k) \\
&= \frac{1}{n^2} \sum_{s=1}^n \sum_{t=1}^n \left(\gamma(t-s)\gamma(t-s-k+l) \right. \\
&\quad \left. + \gamma(t-s+l)\gamma(t-s-k) + \gamma(k)\gamma(l) \right. \\
&\quad \left. + (\alpha-3)\sigma^4 \sum_{g=-\infty}^{\infty} b_g b_{g+k} b_{g+t-s} b_{g+t-s+l} \right) - \gamma(l)\gamma(k). \tag{7.23}
\end{aligned}$$

Since the two indices t and s occur as linear combination $t-s$, we can apply the following useful form

$$\begin{aligned}
\sum_{s=1}^n \sum_{t=1}^n C_{t-s} &= nC_0 + (n-1)C_1 + (n-2)C_2 \cdots + C_{n-1} \\
&\quad + (n-1)C_{-1} + (n-2)C_{-2} \cdots + C_{1-n} \\
&= \sum_{|m| \leq n} (n-|m|)C_m.
\end{aligned}$$

Defining

$$\begin{aligned}
C_{t-s} &:= \gamma(t-s)\gamma(t-s-k+l) + \gamma(t-s+l)\gamma(t-s-k) \\
&\quad + (\alpha-3)\sigma^4 \sum_{g=-\infty}^{\infty} b_g b_{g+k} b_{g+t-s} b_{g+t-s+l},
\end{aligned}$$

we can conveniently rewrite the above covariance (7.23) as

$$\text{Cov}(\tilde{\gamma}_n(k), \tilde{\gamma}_n(l)) = \frac{1}{n^2} \sum_{s=1}^n \sum_{t=1}^n C_{t-s} = \frac{1}{n^2} \sum_{|m| \leq n} (n-|m|)C_m.$$

The absolutely summable filter $(b_u)_{u \in \mathbb{Z}}$ entails the absolute summability of the sequence $(C_m)_{m \in \mathbb{Z}}$. Hence, by the dominated convergence

theorem, it finally follows

$$\begin{aligned}
\lim_{n \rightarrow \infty} n \operatorname{Cov}(\tilde{\gamma}_n(k), \tilde{\gamma}_n(l)) &= \sum_{m=-\infty}^{\infty} C_m \\
&= \sum_{m=-\infty}^{\infty} \left(\gamma(m)\gamma(m-k+l) + \gamma(m+l)\gamma(m-k) \right. \\
&\quad \left. + (\alpha-3)\sigma^4 \sum_{g=-\infty}^{\infty} b_g b_{g+k} b_{g+m} b_{g+m+l} \right) \\
&= (\alpha-3)\gamma(k)\gamma(l) + \sum_{m=-\infty}^{\infty} \left(\gamma(m)\gamma(m-k+l) + \gamma(m+l)\gamma(m-k) \right).
\end{aligned}$$

□

Lemma 7.3.9. *Consider the stationary process $(Y_t)_{t \in \mathbb{Z}}$ from the previous Lemma 7.3.8 satisfying $E(\varepsilon_t^4) := \alpha\sigma^4 < \infty, \alpha > 0, b_0 = 1$ and $b_u = 0$ for $u < 0$. Let $\tilde{\gamma}_{p,n} := (\tilde{\gamma}_n(0), \dots, \tilde{\gamma}_n(p))^T, \gamma_p := (\gamma(0), \dots, \gamma(p))^T$ for $p \geq 0, n \in \mathbb{N}$, and let the $p \times p$ -matrix $\Sigma_c := (c_{kl})_{0 \leq k, l \leq p}$ be given by*

$$\begin{aligned}
c_{kl} &:= (\alpha-3)\gamma(k)\gamma(l) \\
&\quad + \sum_{m=-\infty}^{\infty} \left(\gamma(m)\gamma(m-k+l) + \gamma(m+l)\gamma(m-k) \right).
\end{aligned}$$

Furthermore, consider the MA(q)-process $Y_t^{(q)} := \sum_{u=0}^q b_u \varepsilon_{t-u}, q \in \mathbb{N}, t \in \mathbb{Z}$, with corresponding autocovariance function $\gamma(k)^{(q)}$ and the $p \times p$ -matrix $\Sigma_c^{(q)} := (c_{kl}^{(q)})_{0 \leq k, l \leq p}$ with elements

$$\begin{aligned}
c_{kl}^{(q)} &:= (\alpha-3)\gamma(k)^{(q)}\gamma(l)^{(q)} \\
&\quad + \sum_{m=-\infty}^{\infty} \left(\gamma(m)^{(q)}\gamma(m-k+l)^{(q)} + \gamma(m+l)^{(q)}\gamma(m-k)^{(q)} \right).
\end{aligned}$$

Then, if Σ_c and $\Sigma_c^{(q)}$ are regular,

$$\sqrt{n}(\tilde{\gamma}_{p,n} - \gamma_p) \xrightarrow{D} N(\mathbf{0}, \Sigma_c),$$

as $n \rightarrow \infty$.

Proof. Consider the MA(q)-process $Y_t^{(q)} := \sum_{u=0}^q b_u \varepsilon_{t-u}$, $q \in \mathbb{N}$, $t \in \mathbb{Z}$, with corresponding autocovariance function $\gamma(k)^{(q)}$. We define $\tilde{\gamma}_n(k)^{(q)} := \frac{1}{n} \sum_{t=1}^n Y_t^{(q)} Y_{t+k}^{(q)}$ as well as $\tilde{\gamma}_{p,n}^{(q)} := (\tilde{\gamma}_n(0)^{(q)}, \dots, \tilde{\gamma}_n(p)^{(q)})^T$. Defining moreover

$$\mathbf{Y}_t^{(q)} := (Y_t^{(q)} Y_t^{(q)}, Y_t^{(q)} Y_{t+1}^{(q)}, \dots, Y_t^{(q)} Y_{t+p}^{(q)})^T, \quad t \in \mathbb{Z},$$

we attain a strictly stationary $(q+p)$ -dependence sequence $(\boldsymbol{\lambda}^T \mathbf{Y}_t^{(q)})_{t \in \mathbb{Z}}$ for any $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_{p+1})^T \in \mathbb{R}^{p+1}$, $\boldsymbol{\lambda} \neq \mathbf{0}$. Since

$$\frac{1}{n} \sum_{t=1}^n \mathbf{Y}_t^{(q)} = \tilde{\gamma}_{p,n}^{(q)},$$

the previous Lemma 7.3.8 gives

$$\lim_{n \rightarrow \infty} n \operatorname{Var} \left(\frac{1}{n} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{Y}_t^{(q)} \right) = \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_c^{(q)} \boldsymbol{\lambda} > 0.$$

Now, we can apply Theorem 7.2.16 and receive

$$n^{-1/2} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{Y}_t^{(q)} - n^{1/2} \boldsymbol{\lambda}^T \boldsymbol{\gamma}_p^{(q)} \xrightarrow{D} N(0, \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_c^{(q)} \boldsymbol{\lambda})$$

as $n \rightarrow \infty$, where $\boldsymbol{\gamma}_p^{(q)} := (\gamma(0)^{(q)}, \dots, \gamma(p)^{(q)})^T$. Since, entrywise, $\boldsymbol{\Sigma}_c^{(q)} \rightarrow \boldsymbol{\Sigma}_c$ as $q \rightarrow \infty$, the dominated convergence theorem gives

$$n^{-1/2} \sum_{t=1}^n \boldsymbol{\lambda}^T \mathbf{Y}_t - n^{1/2} \boldsymbol{\lambda}^T \boldsymbol{\gamma}_p \xrightarrow{D} N(0, \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_c \boldsymbol{\lambda})$$

as $n \rightarrow \infty$. The Cramer–Wold device then provides

$$\sqrt{n}(\tilde{\gamma}_{p,n} - \boldsymbol{\gamma}_p) \xrightarrow{D} \mathbf{Y}$$

as $n \rightarrow \infty$, where \mathbf{Y} is $N(\mathbf{0}, \boldsymbol{\Sigma}_c)$ -distributed. Recalling once more Theorem 7.2.7, it remains to show

$$\lim_{q \rightarrow \infty} \limsup_{n \rightarrow \infty} P\{\sqrt{n}|\tilde{\gamma}_n(k)^{(q)} - \gamma(k)^{(q)} - \tilde{\gamma}_n(k) + \gamma(k)| > \varepsilon\} = 0 \quad (7.24)$$

for every $\varepsilon > 0$ and $k = 0, \dots, p$. We attain, by Chebychev's inequality,

$$\begin{aligned} & P\{\sqrt{n}|\tilde{\gamma}_n(k)^{(q)} - \gamma(k)^{(q)} - \tilde{\gamma}_n(k) + \gamma(k)| \geq \varepsilon\} \\ & \leq \frac{n}{\varepsilon^2} \text{Var}(\tilde{\gamma}_n(k)^{(q)} - \tilde{\gamma}_n(k)) \\ & = \frac{1}{\varepsilon^2} \left(n \text{Var}(\tilde{\gamma}_n(k))^{(q)} + n \text{Var}(\tilde{\gamma}_n(k)) - 2n \text{Cov}(\tilde{\gamma}_n(k)^{(q)} \tilde{\gamma}_n(k)) \right). \end{aligned}$$

By the dominated convergence theorem and Lemma 7.3.8, the first term satisfies

$$\lim_{q \rightarrow \infty} \lim_{n \rightarrow \infty} n \text{Var}(\tilde{\gamma}_n(k)^{(q)}) = \lim_{n \rightarrow \infty} n \text{Var}(\tilde{\gamma}_n(k)) = c_{kk},$$

the last one

$$\lim_{q \rightarrow \infty} \lim_{n \rightarrow \infty} 2n \text{Cov}(\tilde{\gamma}_n(k)^{(q)} \tilde{\gamma}_n(k)) = 2c_{kk},$$

showing altogether (7.24). \square

Lemma 7.3.10. *Consider the stationary process and the notations from Lemma 7.3.9, where the matrices Σ_c and $\Sigma_c^{(q)}$ are assumed to be positive definite. Let $\hat{c}_n(s) = \frac{1}{n} \sum_{t=1}^{n-k} (Y_{t+s} - \bar{Y})(Y_t - \bar{Y})$, with $\bar{Y} = \frac{1}{n} \sum_{t=1}^n Y_t$, denote the autocovariance estimator at lag $s \in \mathbb{N}$. Then, for $p \geq 0$,*

$$\sqrt{n}(\hat{\mathbf{c}}_{p,n} - \boldsymbol{\gamma}_p) \xrightarrow{D} N(\mathbf{0}, \Sigma_c)$$

as $n \rightarrow \infty$, where $\hat{\mathbf{c}}_{p,n} := (\hat{c}_n(0), \dots, \hat{c}_n(p))^T$.

Proof. In view of Lemma 7.3.9, it suffices to show that $\sqrt{n}(\hat{c}_n(k) - \tilde{\gamma}_n(k)) \xrightarrow{P} 0$ as $n \rightarrow \infty$ for $k = 0, \dots, p$, since then Lemma 7.2.8 provides the assertion. We have

$$\begin{aligned} \sqrt{n}(\hat{c}_n(k) - \tilde{\gamma}_n(k)) &= \sqrt{n} \left(\frac{1}{n} \sum_{t=1}^{n-k} (Y_{t+k} - \bar{Y})(Y_t - \bar{Y}) - \frac{1}{n} \sum_{t=1}^n Y_t Y_{t+k} \right) \\ &= \sqrt{n} \bar{Y} \left(\frac{n-k}{n} \bar{Y} - \frac{1}{n} \sum_{t=1}^{n-k} (Y_{t+k} + Y_t) \right) - \frac{1}{\sqrt{n}} \sum_{t=n-k+1}^n Y_{t+k} Y_t. \end{aligned}$$

We know from Theorem 7.2.18 that either $\sqrt{n}\bar{Y} \xrightarrow{P} 0$, if $\sum_{u=0}^{\infty} b_u = 0$ or

$$\sqrt{n}\bar{Y} \xrightarrow{D} N\left(0, \sigma^2 \left(\sum_{u=0}^{\infty} b_u\right)^2\right)$$

as $n \rightarrow \infty$, if $\sum_{u=0}^{\infty} b_u \neq 0$. This entails the boundedness in probability $\sqrt{n}\bar{Y} = \mathcal{O}_P(1)$ by Lemma 7.3.3. By Markov's inequality, it follows, for every $\varepsilon > 0$,

$$\begin{aligned} P\left\{\left|\frac{1}{\sqrt{n}} \sum_{t=n-k+1}^n Y_{t+k} Y_t\right| \geq \varepsilon\right\} &\leq \frac{1}{\varepsilon} E\left(\left|\frac{1}{\sqrt{n}} \sum_{t=n-k+1}^n Y_{t+k} Y_t\right|\right) \\ &\leq \frac{k}{\varepsilon \sqrt{n}} \gamma(0) \rightarrow 0 \text{ as } n \rightarrow \infty, \end{aligned}$$

which shows that $n^{-1/2} \sum_{t=n-k+1}^n Y_{t+k} Y_t \xrightarrow{P} 0$ as $n \rightarrow \infty$. Applying Lemma 7.2.12 leads to

$$\frac{n-k}{n} \bar{Y} - \frac{1}{n} \sum_{t=1}^{n-k} (Y_{t+k} + Y_t) \xrightarrow{P} 0 \text{ as } n \rightarrow \infty.$$

The required condition

$$\sqrt{n}(\hat{c}_n(k) - \tilde{\gamma}_n(k)) \xrightarrow{P} 0 \text{ as } n \rightarrow \infty$$

is then attained by Lemma 7.3.1. \square

Theorem 7.3.11. *Consider the stationary process and the notations from Lemma 7.3.10, where the matrices Σ_c and $\Sigma_c^{(q)}$ are assumed to be positive definite. Let $\rho(k)$ be the autocorrelation function of $(Y_t)_{t \in \mathbb{Z}}$ and $\boldsymbol{\rho}_p := (\rho(1), \dots, \rho(p))^T$ the autocorrelation vector in \mathbb{R}^p . Let furthermore $\hat{r}_n(k) := \hat{c}_n(k)/\hat{c}_n(0)$ be the autocorrelation estimator of the process for sufficiently large n with corresponding estimator vector $\hat{\mathbf{r}}_{p,n} := (\hat{r}_n(1), \dots, \hat{r}_n(p))^T$, $p > 0$. Then, for sufficiently large n ,*

$$\sqrt{n}(\hat{\mathbf{r}}_{p,n} - \boldsymbol{\rho}_p) \xrightarrow{D} N(\mathbf{0}, \Sigma_r)$$

as $n \rightarrow \infty$, where Σ_r is the covariance matrix $(r_{ij})_{1 \leq i, j \leq p}$ with entries

$$r_{ij} = \sum_{m=-\infty}^{\infty} \left(2\rho(m) \left(\rho(i)\rho(j)\rho(m) - \rho(i)\rho(m+j) - \rho(j)\rho(m+i) \right) + \rho(m+j) \left(\rho(m+i) + \rho(m-i) \right) \right). \quad (7.25)$$

Proof. Note that $\hat{r}_n(k)$ is well defined for sufficiently large n since $\hat{c}_n(0) \xrightarrow{P} \gamma(0) = \sigma^2 \sum_{u=-\infty}^{\infty} b_u^2 > 0$ as $n \rightarrow \infty$ by (7.14). Let ϑ be the function defined by $\vartheta((x_0, x_1, \dots, x_p)^T) = (\frac{x_1}{x_0}, \frac{x_2}{x_0}, \dots, \frac{x_p}{x_0})^T$, where $x_s \in \mathbb{R}$ for $s \in \{0, \dots, p\}$ and $x_0 \neq 0$. The multivariate delta-method 7.3.7 and Lemma 7.3.10 show that

$$\sqrt{n} \left(\vartheta \begin{pmatrix} \hat{c}_n(0) \\ \vdots \\ \hat{c}_n(p) \end{pmatrix} - \vartheta \begin{pmatrix} \gamma(0) \\ \vdots \\ \gamma(p) \end{pmatrix} \right) = \sqrt{n}(\hat{\mathbf{r}}_{p,n} - \boldsymbol{\rho}_p) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Delta} \boldsymbol{\Sigma}_c \boldsymbol{\Delta}^T)$$

as $n \rightarrow \infty$, where the $p \times p+1$ -matrix $\boldsymbol{\Delta}$ is given by the block matrix

$$(\delta_{ij})_{1 \leq i \leq p, 0 \leq j \leq p} := \boldsymbol{\Delta} = \frac{1}{\gamma(0)} \begin{pmatrix} -\boldsymbol{\rho}_p & \mathbf{I}_p \end{pmatrix}$$

with $p \times p$ -identity matrix \mathbf{I}_p .

The (i, j) -element r_{ij} of $\Sigma_r := \boldsymbol{\Delta} \boldsymbol{\Sigma}_c \boldsymbol{\Delta}^T$ satisfies

$$\begin{aligned}
r_{ij} &= \sum_{k=0}^p \delta_{ik} \sum_{l=0}^p c_{kl} \delta_{jl} \\
&= \frac{1}{\gamma(0)^2} \left(-\rho(i) \sum_{l=0}^p (c_{0l} \delta_{jl} + c_{il} \delta_{jl}) \right) \\
&= \frac{1}{\gamma(0)^2} \left(\rho(i) \rho(j) c_{00} - \rho(i) c_{0j} - \rho(j) c_{10} + c_{11} \right) \\
&= \rho(i) \rho(j) \left((\alpha - 3) + \sum_{m=-\infty}^{\infty} 2\rho(m)^2 \right) \\
&\quad - \rho(i) \left((\alpha - 3) \rho(j) + \sum_{m'=-\infty}^{\infty} 2\rho(m') \rho(m' + j) \right) \\
&\quad - \rho(j) \left((\alpha - 3) \rho(i) + \sum_{m^*=-\infty}^{\infty} 2\rho(m^*) \rho(m^* - i) \right) + (\alpha - 3) \rho(i) \rho(j) \\
&\quad + \sum_{m^\circ=-\infty}^{\infty} \left(\rho(m^\circ) \rho(m^\circ - i + j) + \rho(m^\circ + j) \rho(m^\circ - i) \right) \\
&= \sum_{m=-\infty}^{\infty} \left(2\rho(i) \rho(j) \rho(m)^2 - 2\rho(i) \rho(m) \rho(m + j) \right. \\
&\quad \left. - 2\rho(j) \rho(m) \rho(m - i) + \rho(m) \rho(m - i + j) + \rho(m + j) \rho(m - i) \right). \tag{7.26}
\end{aligned}$$

We may write

$$\sum_{m=-\infty}^{\infty} \rho(j) \rho(m) \rho(m - i) = \sum_{m=-\infty}^{\infty} \rho(j) \rho(m + i) \rho(m)$$

as well as

$$\sum_{m=-\infty}^{\infty} \rho(m) \rho(m - i + j) = \sum_{m=-\infty}^{\infty} \rho(m + i) \rho(m + j). \tag{7.27}$$

Now, applied to (7.26) yields the representation (7.25). \square

The representation (7.25) is the so-called *Bartlett's formula*, which can be more conveniently written as

$$r_{ij} = \sum_{m=1}^{\infty} \left(\rho(m+i) + \rho(m-i) - 2\rho(i)\rho(m) \right) \cdot \left(\rho(m+j) + \rho(m-j) - 2\rho(j)\rho(m) \right) \quad (7.28)$$

using (7.27).

Remark 7.3.12. The derived distributions in the previous Lemmata and Theorems remain valid even without the assumption of regular matrices $\Sigma_c^{(q)}$ and Σ_c (Brockwell and Davis, 2002, Section 7.2 and 7.3). Hence, we may use above formula for ARMA(p, q)-processes that satisfy the process conditions of Lemma 7.3.9.

7.4 First Examinations

The Box–Jenkins program was introduced in Section 2.3 and deals with the problem of selecting an invertible and zero-mean ARMA(p, q)-model that satisfies the stationary condition and $\text{Var}(Y_t) > 0$ for the purpose of appropriately explaining, estimating and forecasting univariate time series. In general, the original time series has to be prepared in order to represent a possible realization of such an ARMA(p, q)-model satisfying above conditions. Given an adequate ARMA(p, q)-model, forecasts of the original time series can be attained by reversing the applied modifications.

In the following sections, we will apply the program to the Donauwoerth time series y_1, \dots, y_n .

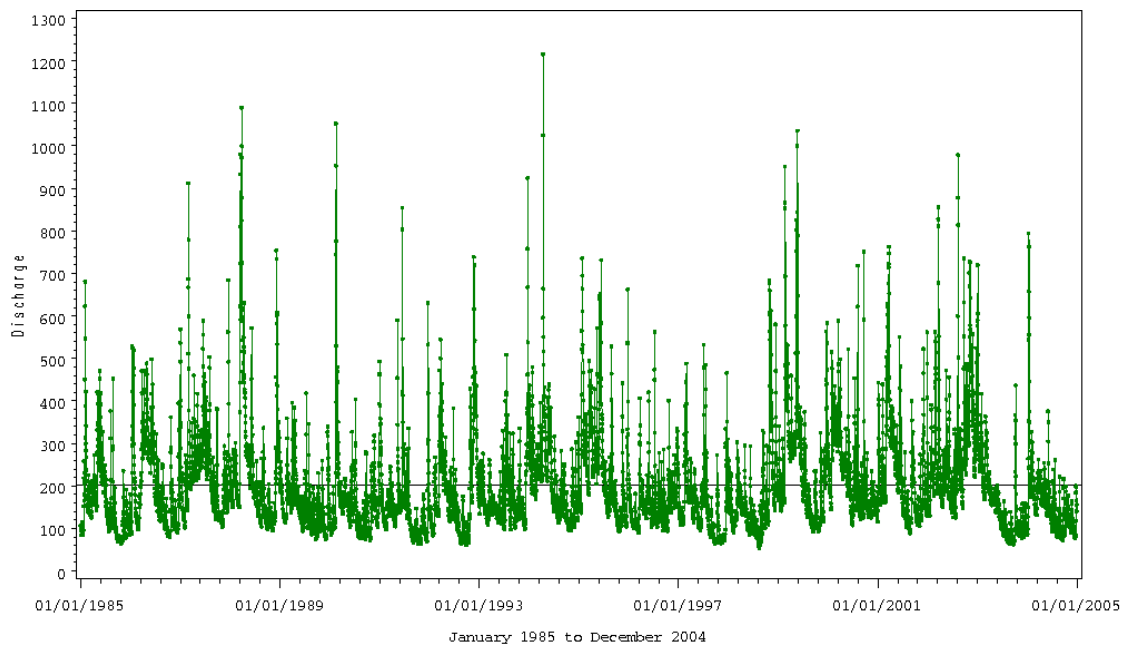
First, we have to check, whether there is need to eliminate an occurring trend, seasonal influence or spread variation over time. Also, in general, a mean correction is necessary in order to attain a zero-mean time series. The following plots of the time series and empirical autocorrelation as well as the periodogram of the time series provide first indications.

The MEANS Procedure

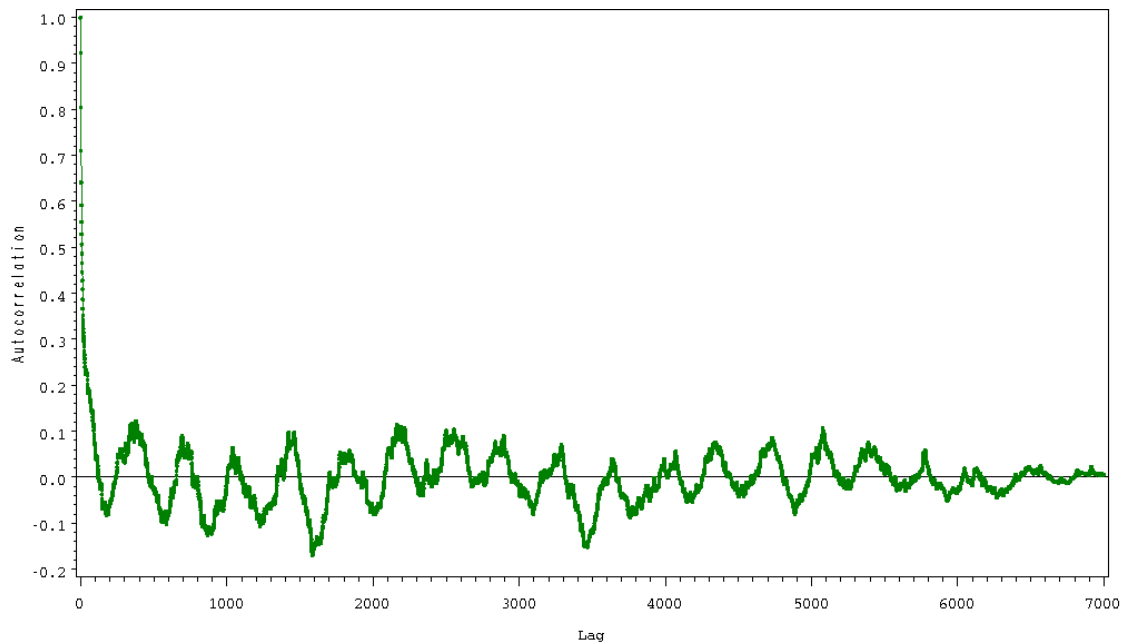
Analysis Variable : discharge

N	Mean	Std Dev	Minimum	Maximum
7300	201.5951932	117.6864736	54.0590000	1216.09

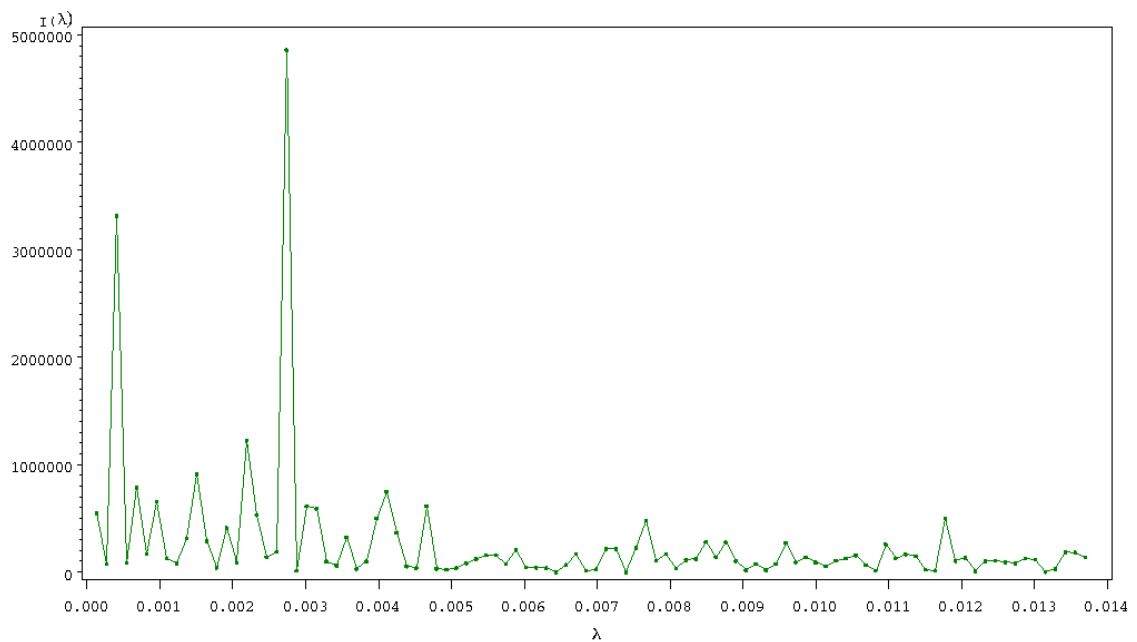
Listing 7.4.1: Summary statistics of the original Donauwoerth Data.



Plot 7.4.1b: Plot of the original Donauwoerth Data.



Plot 7.4.1c: Empirical autocorrelation function of original Donauwoerth Data.



Plot 7.4.1d: Periodogram of the original Donauwoerth Data.

PERIOD	COS_01	SIN_01	p	lambda
365.00	7.1096	51.1112	4859796.14	.002739726

2433.33	-37.4101	20.4289	3315765.43	.000410959
456.25	-1.0611	25.8759	1224005.98	.002191781

Listing 7.4.1e: Greatest periods inherent in the original Donauwoerth Data.

```

1  /* donauwoerth_firstanalysis.sas */
2  TITLE1 'First Analysis';
3  TITLE2 'Donauwoerth Data';
4
5  /* Read in data set */
6  DATA donau;
7      INFILE '/scratch/perm/stat/daten/donauwoerth.txt';
8      INPUT day month year discharge;
9      date=MDY(month, day, year);
10     FORMAT date mmddyy10.;
11
12  /* Compute mean */
13  PROC MEANS DATA=donau;
14      VAR discharge;
15  RUN;
16
17  /* Graphical options */
18  SYMBOL1 V=DOT I=JOIN C=GREEN H=0.3 W=1;
19  AXIS1 LABEL=(ANGLE=90 'Discharge');
20  AXIS2 LABEL=('January 1985 to December 2004') ORDER=('01JAN85'd '01
    ↳JAN89'd '01JAN93'd '01JAN97'd '01JAN01'd '01JAN05'd);
21  AXIS3 LABEL=(ANGLE=90 'Autocorrelation');
22  AXIS4 LABEL=('Lag') ORDER = (0 1000 2000 3000 4000 5000 6000 7000);
23  AXIS5 LABEL=('I(' F=CGREEK 'l)');
24  AXIS6 LABEL=(F=CGREEK 'l');
25
26  /* Generate data plot */
27  PROC GPLOT DATA=donau;
28      PLOT discharge*date=1 / VREF=201.6 VAXIS=AXIS1 HAXIS=AXIS2;
29  RUN;
30
31  /* Compute and plot empirical autocorrelation */
32  PROC ARIMA DATA=donau;
33      IDENTIFY VAR=discharge NLAG=7000 OUTCOV=autocorr NOPRINT;
34  PROC GPLOT DATA=autocorr;
35      PLOT corr*lag=1 /VAXIS=AXIS3 HAXIS=AXIS4 VREF=0;
36  RUN;
37
38  /* Compute periodogram */
39  PROC SPECTRA DATA=donau COEF P OUT=datall;
40      VAR discharge;
41
42  /* Adjusting different periodogram definitions */
43  DATA data2;
44      SET datall(FIRSTOBS=2);
45      p=P_01/2;
46      lambda=FREQ/(2*CONSTANT('PI'));

```

```

47 DROP P_01 FREQ;
48
49 /* Plot periodogram */
50 PROC GPLOT DATA=data2(OBS=100);
51 PLOT p*lambda=1 / VAXIS=AXIS5 HAXIS=AXIS6;
52 RUN;
53
54 /* Print three largest periodogram values */
55 PROC SORT DATA=data2 OUT=data3; BY DESCENDING p;
56 PROC PRINT DATA=data3(OBS=3) NOOBS;
57 RUN; QUIT;

```

In the DATA step the observed measurements of discharge as well as the dates of observations are read from the external file 'donauwoerth.txt' and written into corresponding variables *day*, *month*, *year* and *discharge*. The MDJ function together with the FORMAT statement then creates a variable *date* consisting of month, day and year. The raw discharge values are plotted by the procedure PROC GPLOT with respect to *date* in order to obtain a first impression of the data. The option VREF creates a horizontal line at 201.6,

indicating the arithmetic mean of the variable discharge, which was computed by PROC MEANS. Next, PROC ARIMA computes the empirical autocorrelations up to a lag of 7000 and PROC SPECTRA computes the periodogram of the variable discharge as described in the previous chapters. The corresponding plots are again created with PROC GPLOT. Finally, PROC SORT sorts the values of the periodogram in decreasing order and the three largest ones are printed by PROC PRINT, due to option OBS=3.

The plot of the original Donauwoerth time series indicate no trend, but the plot of the empirical autocorrelation function indicate seasonal variations. The periodogram shows a strong influence of the Fourier frequencies $0.00274 \approx 1/365$ and $0.000411 \approx 1/2433$, indicating cycles with period 365 and 2433.

Carrying out a seasonal and mean adjustment as presented in Chapter 1 leads to an apparently stationary shape, as we can see in the following plots of the adjusted time series, empirical autocorrelation function and periodogram. Furthermore, the variation of the data seems to be independent of time. Hence, there is no need to apply variance stabilizing methods.

Next, we execute Dickey–Fuller’s test for stationarity as introduced in Section 2.2. Since we assume an invertible ARMA(p, q)-process, we approximated it by high-ordered autoregressive models with orders of 7, 8 and 9. Under the three model assumptions 2.18, 2.19 and 2.20, the test of Dickey-Fuller finally provides significantly small p -values for each considered case, cf. Listing 7.4.2d. Accordingly, we have no reason to doubt that the adjusted zero-mean time series, which will

henceforth be denoted by $\tilde{y}_1, \dots, \tilde{y}_n$, can be interpreted as a realization of a stationary ARMA(p, q)-process.

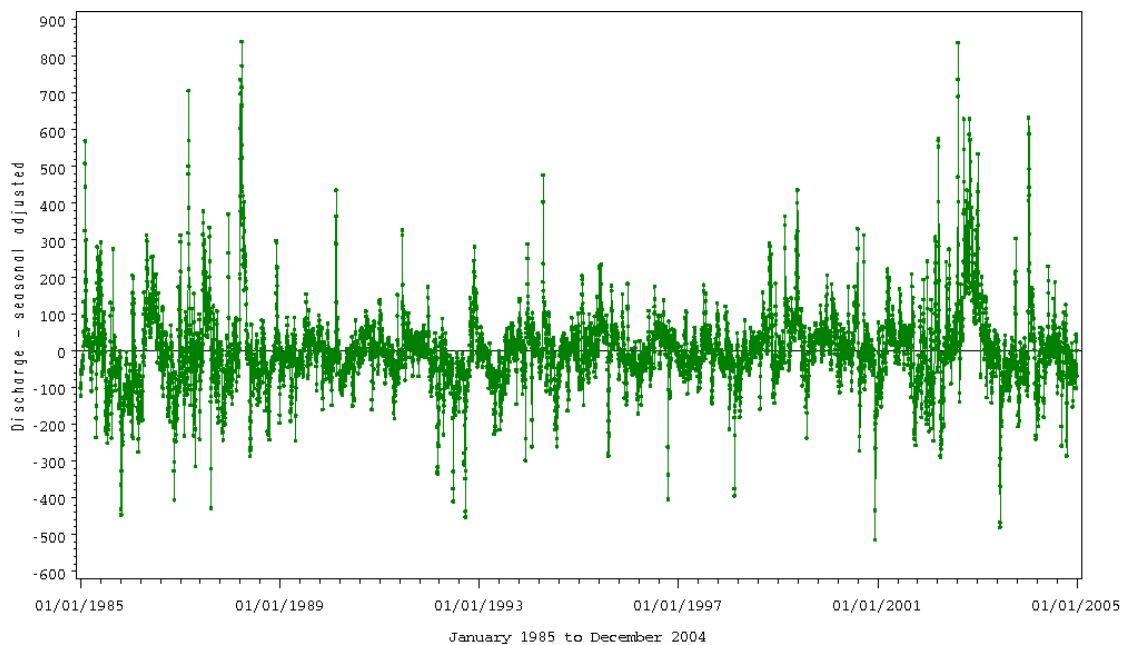
Simultaneously, we tested, if the adjusted time series can be regarded as an outcome of a white noise process. Then there would be no need to fit a model at all. The plot of the empirical autocorrelation function 7.4.2b clearly rejects this assumption. A more formal test for white noise can be executed, using for example the Portmanteau-test of Box–Ljung, cf. Section 7.6.

The MEANS Procedure

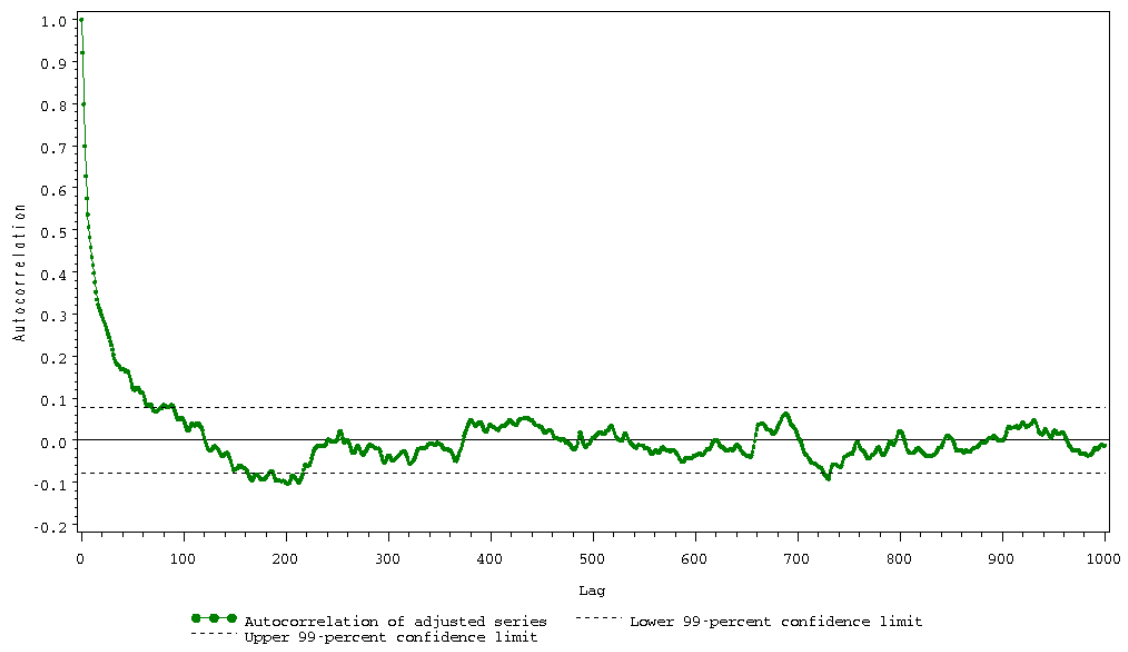
Analysis Variable : discharge

N	Mean	Std Dev	Minimum	Maximum
7300	201.5951932	117.6864736	54.0590000	1216.09

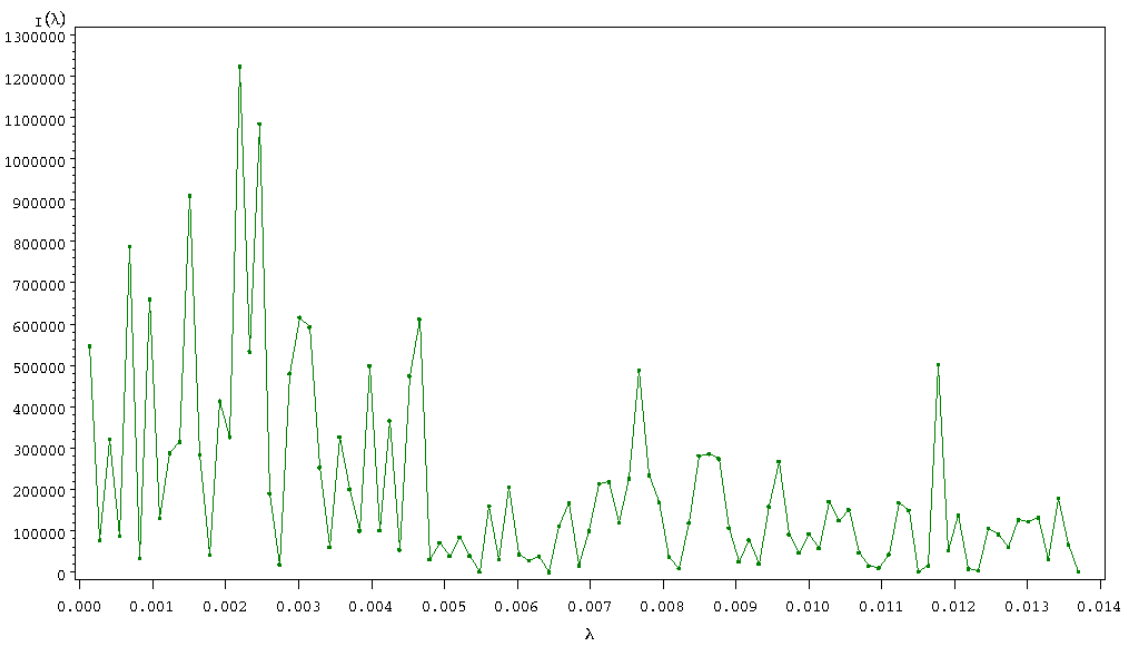
Listing 7.4.2: Summary statistics of the adjusted Donauwoerth Data.



Plot 7.4.2b: Plot of the adjusted Donauwoerth Data.



Plot 7.4.2c: Empirical autocorrelation function of adjusted Donauwoerth Data.



Plot 7.4.2d: Periodogram of adjusted Donauwoerth Data.

Augmented Dickey–Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F

Zero Mean	7	-69.1526	<.0001	-5.79	<.0001		
	8	-63.5820	<.0001	-5.55	<.0001		
	9	-58.4508	<.0001	-5.32	<.0001		
Single Mean	7	-519.331	0.0001	-14.62	<.0001	106.92	0.0010
	8	-497.216	0.0001	-14.16	<.0001	100.25	0.0010
	9	-474.254	0.0001	-13.71	<.0001	93.96	0.0010
Trend	7	-527.202	0.0001	-14.71	<.0001	108.16	0.0010
	8	-505.162	0.0001	-14.25	<.0001	101.46	0.0010
	9	-482.181	0.0001	-13.79	<.0001	95.13	0.0010

Listing 7.4.2e: Augmented Dickey-Fuller Tests for stationarity.

```

1  /* donauwoerth_adjustedment.sas */
2  TITLE1 'Analysis of Adjusted Data';
3  TITLE2 'Donauwoerth Data';
4
5  /* Remove period of 2433 and 365 days */
6  PROC TIMESERIES DATA=donau SEASONALITY=2433 OUTDECOMP=seasad1;
7      VAR discharge;
8      DECOMP / MODE=ADD;
9
10 PROC TIMESERIES DATA=seasad1 SEASONALITY=365 OUTDECOMP=seasad;
11     VAR sa;
12     DECOMP / MODE=ADD;
13 RUN;
14
15 /* Create zero-mean data set */
16 PROC MEANS DATA=seasad;
17 VAR sa;
18
19 DATA seasad;
20 MERGE seasad donau(KEEP=date);
21 sa=sa-201.6;
22
23 /* Graphical options */
24 SYMBOL1 V=DOT I=JOIN C=GREEN H=0.3;
25 SYMBOL2 V=NONE I=JOIN C=BLACK L=2;
26 AXIS1 LABEL=(ANGLE=90 'Adjusted Data');
27 AXIS2 LABEL=('January 1985 to December 2004')
28     ORDER=('01JAN85'd '01JAN89'd '01JAN93'd '01JAN97'd '01JAN01'd
29         ↪ '01JAN05'd);
30 AXIS3 LABEL=('I' F=CGREEK '(1)');
31 AXIS4 LABEL=(F=CGREEK '1');
32 AXIS5 LABEL=(ANGLE=90 'Autocorrelation') ORDER=(-0.2 TO 1 BY 0.1);
33 AXIS6 LABEL=('Lag') ORDER=(0 TO 1000 BY 100);
34 LEGEND1 LABEL=NONE VALUE=('Autocorrelation of adjusted series' 'Lower
35     ↪ 99-percent confidence limit' 'Upper 99-percent confidence limit')
36     ↪;
37
38 /* Plot data */
39 PROC GPLOT DATA=seasad;
40     PLOT sa*date=1 / VREF=0 VAXIS=AXIS1 HAXIS=AXIS2;

```

```

38 RUN;
39
40 /* Compute and plot autocorrelation of adjusted data */
41 PROC ARIMA DATA=seasad;
42     IDENTIFY VAR=sa NLAG=1000 OUTCOV=corrseasad NOPRINT;
43
44 /* Add confidence intervals */
45 DATA corrseasad;
46 SET corrseasad;
47 u99=0.079;
48 l99=-0.079;
49
50 PROC GPLOT DATA=corrseasad;
51     PLOT corr*lag=1 u99*lag=2 l99*lag=2 / OVERLAY VAXIS=AXIS5 HAXIS=
        ↪AXIS6 VREF=0 LEGEND=LEGEND1;
52 RUN;
53
54 /* Compute periodogram of adjusted data */
55 PROC SPECTRA DATA=seasad COEF P OUT=data1;
56     VAR sa;
57
58 /* Adjust different periodogram definitions */
59 DATA data2;
60     SET data1(FIRSTOBS=2);
61     p=P_01/2;
62     lambda=FREQ/(2*CONSTANT('PI'));
63     DROP P_01 FREQ;
64
65 /* Plot periodogram of adjusted data */
66 PROC GPLOT DATA=data2(OBS=100);
67     PLOT p*lambda=1 / VAXIS=AXIS3 HAXIS=AXIS4;
68 RUN;
69
70 /* Test for stationarity */
71 PROC ARIMA DATA=seasad;
72     IDENTIFY VAR=sa NLAG=100 STATIONARITY=(ADF=(7,8,9));
73 RUN; QUIT;

```

The seasonal influences of the periods of both 365 and 2433 days are removed by the `SEASONALITY` option in the procedure `PROC TIMESERIES`. By `MODE=ADD`, an additive model for the time series is assumed. The mean correction of -201.6 finally completes the adjustment. The dates of the original Donauwoerth dataset need to be restored by means of `MERGE` within the `DATA` statement. Then, the same steps as in the previous Program 7.4.1 (donauwoerth.firstanalysis.sas) are executed, i.e., the general plot as well as the plots of the empirical autocorrelation function and pe-

riodogram are created from the adjusted data. The confidence interval $I := [-0.079, 0.079]$ pertaining to the 3σ -rule is displayed by the variables `l99` and `u99`.

The final part of the program deals with the test for stationarity. The augmented Dickey-Fuller test is initiated by `ADF` in the `STATIONARITY` option in `PROC ARIMA`. Since we deal with true $ARMA(p, q)$ -processes, we have to choose a high-ordered autoregressive model, whose order selection is specified by the subsequent numbers 7, 8 and 9.

In a next step we analyze the empirical partial autocorrelation function $\hat{\alpha}(k)$ and the empirical autocorrelation function $r(k)$ as estimations of the theoretical partial autocorrelation function $\alpha(k)$ and autocorrelation function $\rho(k)$ of the underlying stationary process. As shown before, both estimators become less reliable as $k \rightarrow n$, $k < n = 7300$.

Given a MA(q)-process satisfying the conditions of Lemma 7.3.9, we attain, using Bartlett's formula (7.28), for $i > q$ and n sufficiently large,

$$\frac{1}{n}r_{ii} = \frac{1}{n} \sum_{m=1}^{\infty} \rho(m-i)^2 = \frac{1}{n} \left(1 + 2\rho(1)^2 + 2\rho(2)^2 \cdots + 2\rho(q)^2 \right) \quad (7.29)$$

as asymptotic variance of the autocorrelation estimators $\hat{r}_n(i)$. Thus, if the zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$ is an outcome of the MA(q)-process with a sufficiently large sample size n , then we can approximate above variance by

$$\sigma_q^2 := \frac{1}{n} \left(1 + 2r(1)^2 + 2r(2)^2 \cdots + 2r(q)^2 \right). \quad (7.30)$$

Since $\hat{r}_n(i)$ is asymptotically normal distributed (see Section 7.3), we will reject the hypothesis H_0 that actually a MA(q)-process is underlying the time series $\tilde{y}_1, \dots, \tilde{y}_n$, if significantly more than $1 - \alpha$ percent of the empirical autocorrelation coefficients with lag greater than q lie outside of the confidence interval $I := [-\sigma_q \cdot q_{1-\alpha/2}, \sigma_q \cdot q_{1-\alpha/2}]$ of level α , where $q_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution. Setting $q = 10$, we attain in our case $\sigma_q \approx \sqrt{5/7300} \approx 0.026$. By the 3σ rule, we would expect that almost all empirical autocorrelation coefficients with lag greater than 10 will be elements of the confidence interval $J := [-0.079, 0.079]$. Hence, Plot 7.4.2b makes us doubt that we can express the time series by a mere MA(q)-process with a small $q \leq 10$.

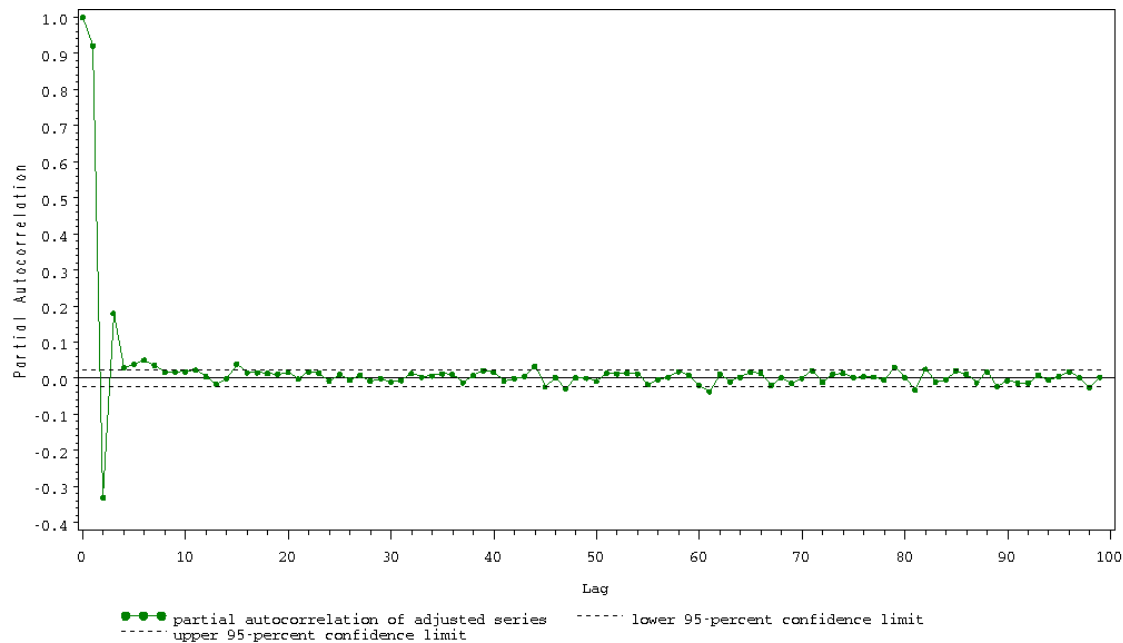
The following figure of the empirical partial autocorrelation function can mislead to the deceptive conclusion that an AR(3)-process might

explain the time series quite well. Actually, if an $\text{AR}(p)$ -process underlies the adjusted time series $\tilde{y}_1, \dots, \tilde{y}_n$ with a sample size n of 7300, then the asymptotic variances of the partial autocorrelation estimators $\hat{\alpha}_n(k)$ with lag greater than p are equal to $1/7300$, cf Theorem 7.2.21. Since $\hat{\alpha}_n(k)$ is asymptotically normal distributed, we attain the 95-percent confidence interval $I := [-0.0234, 0.0234]$, due to the 2σ rule. In view of the numerous quantity of empirical partial autocorrelation coefficients lying outside of the interval I , summarized in Table 7.4.1, we have no reason to choose mere $\text{AR}(p)$ -processes as adequate models.

Nevertheless, focusing on true $\text{ARMA}(p, q)$ -processes, the shape of the empirical partial autocorrelation might indicate that the autoregressive order p will be equal to 2 or 3.

lag	partial autocorrelation	lag	partial autocorrelation
1	0.92142	44	0.03368
2	-0.33057	45	-0.02412
3	0.17989	47	-0.02892
4	0.02893	61	-0.03667
5	0.03871	79	0.03072
6	0.05010	81	-0.03248
7	0.03633	82	0.02521
15	0.03937	98	-0.02534

Table 7.4.1: Greatest coefficients in absolute value of empirical partial autocorrelation of adjusted time series.



Plot 7.4.3: Empirical partial autocorrelation function of adjusted Donauwoerth Data.

```

1  /* donauwoerth_pacf.sas */
2  TITLE1 'Partial Autocorrelation';
3  TITLE2 'Donauwoerth Data';
4
5  /* Note that this program requires the file 'seasad' generated by the
   ↪ previous program (donauwoerth_adjustment.sas) */
6
7  /* Graphical options */
8  SYMBOL1 V=DOT I=JOIN C=GREEN H=0.5;
9  SYMBOL2 V=NONE I=JOIN C=BLACK L=2;
10 AXIS1 LABEL=(ANGLE=90 'Partial Autocorrelation') ORDER=(-0.4 TO 1 BY
   ↪ 0.1);
11 AXIS2 LABEL=('Lag');
12 LEGEND2 LABEL=NONE VALUE=('Partial autocorrelation of adjusted series'
13   'Lower 95-percent confidence limit' 'Upper 95-percent confidence
   ↪ limit');
14
15 /* Compute partial autocorrelation of the seasonal adjusted data */
16 PROC ARIMA DATA=seasad;
17   IDENTIFY VAR=sa NLAG=100 OUTCOV=partcorr NOPRINT;
18 RUN;
19
20 /* Add confidence intervals */
21 DATA partcorr;
22 SET partcorr;
23 u95=0.0234;
24 l95=-0.0234;
25

```

```

26 /* Plot partial autocorrelation */
27 PROC GPLOT DATA=partcorr(OBS=100);
28     PLOT partcorr*lag=1 u95*lag=2 195*lag=2 / OVERLAY VAXIS=AXIS1 HAXIS=
    ↪AXIS2 VREF=0 LEGEND=LEGEND2;
29 RUN;
30
31 /* Indicate greatest values of partial autocorrelation */
32 DATA gp(KEEP=lag partcorr);
33     SET partcorr;
34     IF ABS(partcorr) le 0.0234 THEN delete;
35 RUN; QUIT;

```

The empirical partial autocorrelation coefficients of the adjusted data are computed by PROC ARIMA up to lag 100 and are plotted afterwards by PROC GPLOT. The confidence interval $[-0.0234, 0.0234]$ pertaining to the 2σ -rule

is again visualized by two variables u95 and 195. The last DATA step finally generates a file consisting of the empirical partial autocorrelation coefficients outside of the confidence interval.

7.5 Order Selection

As seen in Section 2, every invertible ARMA(p, q)-process satisfying the stationarity condition can be both rewritten as $Y_t = \sum_{v \geq 0} \alpha_v \varepsilon_{t-v}$ and as $\varepsilon_t = \sum_{w \geq 0} \beta_w Y_{t-w}$, almost surely. Thus, they can be approximated by a high-order MA(q)- or AR(p)-process, respectively. Accordingly, we could entirely omit the consideration of true ARMA(p, q)-processes. However, if the time series actually results from a true ARMA(p, q)-process, we would only attain an appropriate approximation by admitting a lot of parameters.

Box et al. (1994) suggested the choose of the model orders by means of the *principle of parsimony*, preferring the model with least amount of parameters. Suppose that a MA(q')-, an AR(p')- and an ARMA(p, q)-model have been adjusted to a sample with almost equal goodness of fit. Then, studies of Box and Jenkins have demonstrated that, in general, $p + q \leq q'$ as well as $p + q \leq p'$. Thus, the ARMA(p, q)-model requires the smallest quantity of parameters.

The principle of parsimony entails not only a comparatively small number of parameters, which can reduce the efforts of estimating the unknown model coefficients. In general, it also provides a more reliable forecast, since the forecast is based on fewer estimated model

coefficients.

To specify the order of an invertible ARMA(p, q)-model

$$Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}, \quad t \in \mathbb{Z}, \quad (7.31)$$

satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$ from a given zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$, which is assumed to be a realization of the model, SAS offers several methods within the framework of the ARIMA procedure. The minimum information criterion, short MINIC method, the extended sample autocorrelation function method (ESACF-method) and the smallest canonical correlation method (SCAN-method) are provided in the following sections. Since the methods depend on estimations, it is always advisable to compare all three approaches.

Note that SAS only considers ARMA(p, q)-model satisfying the preliminary conditions.

Information Criteria

To choose the orders p and q of an ARMA(p, q)-process one commonly takes the pair (p, q) minimizing some information function, which is based on the loglikelihood function.

For Gaussian ARMA(p, q)-processes we have derived the loglikelihood function in Section 2.3

$$\begin{aligned} l(\boldsymbol{\vartheta} | y_1, \dots, y_n) \\ = \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log(\det \boldsymbol{\Sigma}') - \frac{1}{2\sigma^2} Q(\boldsymbol{\vartheta} | y_1, \dots, y_n). \end{aligned} \quad (7.32)$$

The maximum likelihood estimator $\hat{\boldsymbol{\vartheta}} := (\hat{\sigma}^2, \hat{\mu}, \hat{a}_1, \dots, \hat{a}_p, \hat{b}_1, \dots, \hat{b}_q)$, maximizing $l(\boldsymbol{\vartheta} | y_1, \dots, y_n)$, can often be computed by deriving the ordinary derivative and the partial derivatives of $l(\boldsymbol{\vartheta} | y_1, \dots, y_n)$ and equating them to zero. These are the so-called *maximum likelihood equations*, which $\hat{\boldsymbol{\vartheta}}$ necessarily has to satisfy. Holding $\sigma^2, a_1, \dots, a_p, b_1,$

\dots, b_q , fix we obtain

$$\begin{aligned}
 \frac{\partial l(\boldsymbol{\vartheta}|y_1, \dots, y_n)}{\partial \mu} &= -\frac{1}{2\sigma^2} \frac{\partial Q(\boldsymbol{\vartheta}|y_1, \dots, y_n)}{\partial \mu} \\
 &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \mu} \left((\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}'^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right) \\
 &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \mu} \left(\mathbf{y}^T \boldsymbol{\Sigma}'^{-1} \mathbf{y} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}'^{-1} \boldsymbol{\mu} \right. \\
 &\quad \left. - \boldsymbol{\mu} \cdot \mathbf{y}^T \boldsymbol{\Sigma}'^{-1} \mathbf{1} - \boldsymbol{\mu} \cdot \mathbf{1}^T \boldsymbol{\Sigma}'^{-1} \mathbf{y} \right) \\
 &= -\frac{1}{2\sigma^2} (2\boldsymbol{\mu} \cdot \mathbf{1}^T \boldsymbol{\Sigma}'^{-1} \mathbf{1} - 2 \cdot \mathbf{1}^T \boldsymbol{\Sigma}'^{-1} \mathbf{y}),
 \end{aligned}$$

where $\mathbf{1} := (1, 1, \dots, 1)^T$, $\mathbf{y} := (y_1, \dots, y_n)^T \in \mathbb{R}^n$. Equating the partial derivative to zero yields finally the maximum likelihood estimator $\hat{\mu}$ of μ

$$\hat{\mu} = \frac{\mathbf{1}^T \hat{\boldsymbol{\Sigma}}'^{-1} \mathbf{y}}{\mathbf{1}^T \hat{\boldsymbol{\Sigma}}'^{-1} \mathbf{1}}, \quad (7.33)$$

where $\hat{\boldsymbol{\Sigma}}'$ equals $\boldsymbol{\Sigma}'$, where the unknown parameters a_1, \dots, a_p and b_1, \dots, b_q are replaced by maximum likelihood estimators $\hat{a}_1, \dots, \hat{a}_p$ and $\hat{b}_1, \dots, \hat{b}_q$.

The maximum likelihood estimator $\hat{\sigma}^2$ of σ^2 can be achieved in an analogous way with $\mu, a_1, \dots, a_p, b_1, \dots, b_q$ being held fix. From

$$\frac{\partial l(\boldsymbol{\vartheta}|y_1, \dots, y_n)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{Q(\boldsymbol{\vartheta}|y_1, \dots, y_n)}{2\sigma^4},$$

we attain

$$\hat{\sigma}^2 = \frac{Q(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n)}{n}. \quad (7.34)$$

The computation of the remaining maximum likelihood estimators $\hat{a}_1, \dots, \hat{a}_p, \hat{b}_1, \dots, \hat{b}_q$ is usually a computer intensive problem.

Since the maximized loglikelihood function only depends on the model assumption $M_{p,q}$, $l_{p,q}(y_1, \dots, y_n) := l(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n)$ is suitable to create a measure for comparing the goodness of fit of different models. The greater $l_{p,q}(y_1, \dots, y_n)$, the higher the likelihood that y_1, \dots, y_n actually results from a Gaussian ARMA(p, q)-process.

Comparing a Gaussian ARMA(p^*, q^*)- and ARMA(p', q')-process, with $p^* \geq p', q^* \geq q'$, immediately yields $l_{p^*, q^*}(y_1, \dots, y_n) \geq l_{p', q'}(y_1, \dots, y_n)$. Thus, without modification, a decision rule based on the pure maximized loglikelihood function would always suggest comprehensive models.

Akaike (1977) therefore investigated the goodness of fit of $\hat{\boldsymbol{\vartheta}}$ in view of a new realization (z_1, \dots, z_n) from the same ARMA(p, q)-process, which is independent from (y_1, \dots, y_n) . The loglikelihood function (7.32) then provides

$$\begin{aligned} l(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n) \\ = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \log(\det \hat{\boldsymbol{\Sigma}}') - \frac{1}{2\hat{\sigma}^2} Q(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n) \end{aligned}$$

as well as

$$\begin{aligned} l(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n) \\ = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \log(\det \hat{\boldsymbol{\Sigma}}') - \frac{1}{2\hat{\sigma}^2} Q(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n). \end{aligned}$$

This leads to the representation

$$\begin{aligned} l(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n) \\ = l(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n) - \frac{1}{2\hat{\sigma}^2} (Q(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n) - Q(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n)). \quad (7.35) \end{aligned}$$

The asymptotic distribution of $Q(\hat{\boldsymbol{\vartheta}}|Z_1, \dots, Z_n) - Q(\hat{\boldsymbol{\vartheta}}|Y_1, \dots, Y_n)$, where Z_1, \dots, Z_n and Y_1, \dots, Y_n are independent samples resulting from the ARMA(p, q)-process, can be computed for n sufficiently large, which reveals an asymptotic expectation of $2\sigma^2(p + q)$ (see for example Brockwell and Davis, 1991, Section 8.11). If we now replace the term $Q(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n) - Q(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n)$ in (7.35) by its approximately expected value $2\hat{\sigma}^2(p + q)$, we attain with (7.34)

$$\begin{aligned} l(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n) &= l(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n) - (p + q) \\ &= -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \log(\det \hat{\boldsymbol{\Sigma}}') - \frac{1}{2\hat{\sigma}^2} Q(\hat{\boldsymbol{\vartheta}}|y_1, \dots, y_n) - (p + q) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2} \log(\det \hat{\boldsymbol{\Sigma}}') - \frac{n}{2} - (p + q). \end{aligned}$$

The greater the value of $l(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n)$, the more precisely the estimated parameters $\hat{\sigma}^2, \hat{\mu}, \hat{a}_1, \dots, \hat{a}_p, \hat{b}_1, \dots, \hat{b}_q$ reproduce the true underlying ARMA(p, q)-process. Due to this result, Akaike's Information Criterion (AIC) is defined as

$$\begin{aligned} \text{AIC} &:= -\frac{2}{n}l(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n) - \log(2\pi) - \frac{1}{n}\log(\det \hat{\boldsymbol{\Sigma}}') - 1 \\ &= \log(\hat{\sigma}^2) + \frac{2(p+q)}{n} \end{aligned}$$

for sufficiently large n , since then $n^{-1}\log(\det \hat{\boldsymbol{\Sigma}}')$ becomes negligible. Thus, the smaller the value of the AIC, the greater the loglikelihood function $l(\hat{\boldsymbol{\vartheta}}|z_1, \dots, z_n)$ and, accordingly, the more precisely $\hat{\boldsymbol{\vartheta}}$ estimates $\boldsymbol{\vartheta}$.

The AIC adheres to the principle of parsimony as the model orders are included in an additional term, the so-called *penalty function*.

However, comprehensive studies came to the conclusion that the AIC has the tendency to overestimate the order p . Therefore, modified criteria based on the AIC approach were suggested, like the Bayesian Information Criterion (BIC)

$$\text{BIC}(p, q) := \log(\hat{\sigma}_{p,q}^2) + \frac{(p+q)\log(n)}{n}$$

and the Hannan-Quinn Criterion

$$\text{HQ}(p, q) := \log(\hat{\sigma}_{p,q}^2) + \frac{2(p+q)c\log(\log(n))}{n} \quad \text{with } c > 1,$$

for which we refer to Akaike (1977) and Hannan and Quinn (1979).

Remark 7.5.1. Although these criteria are derived under the assumption of an underlying Gaussian process, they nevertheless can tentatively identify the orders p and q , even if the original process is not Gaussian. This is due to the fact that the criteria can be regarded as a measure of the goodness of fit of the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}'$ to the time series (Brockwell and Davis, 2002).

MINIC Method

The *MINIC method* is based on the previously introduced AIC and BIC. The approach seeks the order pair (p, q) minimizing the BIC-Criterion

$$\text{BIC}(p, q) = \log(\hat{\sigma}_{p,q}^2) + \frac{(p + q) \log(n)}{n}$$

in a chosen order range of $p_{\min} \leq p \leq p_{\max}$ and $q_{\min} \leq q \leq q_{\max}$, where $\hat{\sigma}_{p,q}^2$ is an estimate of the variance $\text{Var}(\varepsilon_t) = \sigma^2$ of the errors ε_t in the model (7.31). In order to estimate σ^2 , we first have to estimate the model parameters from the zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$. Since their computation by means of the maximum likelihood method is generally very extensive, the parameters are estimated from the regression model

$$\tilde{y}_t = a_1 \tilde{y}_{t-1} + \dots + a_p \tilde{y}_{t-p} + b_1 \hat{\varepsilon}_{t-1} + \dots + b_q \hat{\varepsilon}_{t-q} + \hat{\varepsilon}_t \quad (7.36)$$

for $\max\{p, q\} + 1 \leq t \leq n$, where $\hat{\varepsilon}_t$ is an estimator of ε_t . The least squares approach, as introduced in Section 2.3, minimizes the residual sum of squares

$$\begin{aligned} S^2(a_1, \dots, a_p, b_1, \dots, b_q) \\ &= \sum_{t=\max\{p,q\}+1}^n (\tilde{y}_t - a_1 \tilde{y}_{t-1} - \dots - a_p \tilde{y}_{t-p} - b_1 \hat{\varepsilon}_{t-1} - \dots - b_q \hat{\varepsilon}_{t-q})^2 \\ &= \sum_{t=\max\{p,q\}+1}^n \hat{\varepsilon}_t^2 \end{aligned}$$

with respect to the parameters $a_1, \dots, a_p, b_1, \dots, b_q$. It provides minimizers $\hat{a}_1, \dots, \hat{a}_p, \hat{b}_1, \dots, \hat{b}_q$. Note that the residuals $\hat{\varepsilon}_t$ and the constants are nested. Then,

$$\hat{\sigma}_{p,q}^2 := \frac{1}{n} \sum_{t=\max\{p,q\}+1}^n \hat{\varepsilon}_t^2, \quad (7.37)$$

where

$$\hat{\varepsilon}_t := \tilde{y}_t - \hat{a}_1 \tilde{y}_{t-1} - \dots - \hat{a}_p \tilde{y}_{t-p} - \hat{b}_1 \hat{\varepsilon}_{t-1} - \dots - \hat{b}_q \hat{\varepsilon}_{t-q},$$

is an obvious estimate of σ^2 for n sufficiently large. Now, we want to estimate the error sequence $(\varepsilon_t)_{t \in \mathbb{Z}}$. Due to the invertibility of the ARMA(p, q)-model (7.31), we find the almost surely autoregressive representation of the errors $\varepsilon_t = \sum_{u \geq 0} c_u Y_{t-u}$, which we can approximate by a finite high-order autoregressive process $\varepsilon_{t,p_\varepsilon} := \sum_{u=0}^{p_\varepsilon} c_u Y_{t-u}$ of order p_ε . Consequently, $\hat{\varepsilon}_t := \sum_{u=0}^{p_\varepsilon} \hat{c}_u \tilde{y}_{t-u}$ is an obvious estimate of the error ε_t for $1 + p_\varepsilon \leq t \leq n$, where the coefficients \hat{c}_i are obtained from the Yule-Walker-estimate (2.8). The length p_ε of the autoregressive model is determined by the order that minimizes the AIC

$$\text{AIC}(p_\varepsilon, 0) := \log(\hat{\sigma}_{p_\varepsilon, 0}^2) + 2 \frac{p_\varepsilon + 0}{n}$$

in a chosen range of $0 < p_\varepsilon \leq p_{\varepsilon, \max}$, where $\hat{\sigma}_{p_\varepsilon, 0}^2 := \frac{1}{n} \sum_{t=1}^n \hat{\varepsilon}_t^2$ with $\hat{\varepsilon}_t = 0$ for $t \in \{1, \dots, p_\varepsilon\}$ is again an obvious estimate of the error variance $\text{Var}(\varepsilon_{t,p_\varepsilon})$ for n sufficiently large.

Hence, we can apply above regression (7.36) for $p_\varepsilon + 1 + \max\{p, q\} \leq t \leq n$ and receive (7.37) as an obvious estimate of σ^2 , where $\tilde{\varepsilon}_t = 0$ for $t = \max\{p, q\} + 1, \dots, \max\{p, q\} + p_\varepsilon$.

ESACF Method

The second method provided by SAS is the *ESACF method*, developed by Tsay and Tiao (1984). The essential idea of this approach is based on the estimation of the autoregressive part of the underlying process in order to receive a pure moving average representation, whose empirical autocorrelation function, the so-called extended sample autocorrelation function, will be close to 0 after the lag of order. For this purpose, we utilize a recursion initiated by the best linear approximation with real valued coefficients $\hat{Y}_t^{(k)} := \lambda_{0,1}^{(k)} Y_{t-1} + \dots + \lambda_{0,k}^{(k)} Y_{t-k}$ of Y_t in the model (7.31) for a chosen $k \in \mathbb{N}$. Recall that the coefficients satisfy

$$\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(k) \end{pmatrix} = \begin{pmatrix} \rho(0) & \dots & \rho(k-1) \\ \vdots & \ddots & \vdots \\ \rho(-k+1) & \dots & \rho(0) \end{pmatrix} \begin{pmatrix} \lambda_{0,1}^{(k)} \\ \vdots \\ \lambda_{0,k}^{(k)} \end{pmatrix} \quad (7.38)$$

and are uniquely determined by Theorem 7.1.1. The lagged residual $R_{t-1,0}^{(k)}, R_{s,0}^{(k)} := Y_s - \lambda_{0,1}^{(k)} Y_{s-1} - \dots - \lambda_{0,k}^{(k)} Y_{s-k}, s \in \mathbb{Z}$, is a linear combi-

nation of $Y_{t-1}, \dots, Y_{t-k-1}$. Hence, the best linear approximation of Y_t based on $Y_{t-1}, \dots, Y_{t-k-1}$ equals the best linear approximation of Y_t based on $Y_{t-1}, \dots, Y_{t-k}, R_{t-1,0}^{(k)}$. Thus, we have

$$\begin{aligned} & \lambda_{0,1}^{(k+1)} Y_{t-1} + \dots + \lambda_{0,k+1}^{(k+1)} Y_{t-k-1} \\ &= \lambda_{1,1}^{(k)} Y_{t-1} + \dots + \lambda_{1,k}^{(k)} Y_{t-k} + \lambda_{1,k+1}^{(k)} R_{t-1,0}^{(k)} \\ &= \lambda_{1,1}^{(k)} Y_{t-1} + \dots + \lambda_{1,k}^{(k)} Y_{t-k} + \lambda_{1,k+1}^{(k)} (Y_{t-1} - \lambda_{0,1}^{(k)} Y_{t-2} - \dots - \lambda_{0,k}^{(k)} Y_{t-k-1}), \end{aligned}$$

where the real valued coefficients consequently satisfy

$$\begin{pmatrix} \lambda_{0,1}^{(k+1)} \\ \lambda_{0,2}^{(k+1)} \\ \vdots \\ \lambda_{0,k-1}^{(k+1)} \\ \lambda_{0,k}^{(k+1)} \\ \lambda_{0,k+1}^{(k+1)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & & \vdots & -\lambda_{0,1}^{(k)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & 0 - \lambda_{0,k-2}^{(k)} \\ \vdots & & & & 1 - \lambda_{0,k-1}^{(k)} \\ 0 & \dots & & 0 & -\lambda_{0,k}^{(k)} \end{pmatrix} \begin{pmatrix} \lambda_{1,1}^{(k)} \\ \lambda_{1,2}^{(k)} \\ \vdots \\ \lambda_{1,k-1}^{(k)} \\ \lambda_{1,k}^{(k)} \\ \lambda_{1,k+1}^{(k)} \end{pmatrix}.$$

Together with (7.38), we attain

$$\begin{pmatrix} \rho(0) & \dots & \rho(k-1) & \omega_0(k) \\ \rho(-1) & \dots & \rho(k-2) & \omega_{-1}(k) \\ \vdots & & \vdots & \vdots \\ \rho(-k) & \dots & \rho(-1) & \omega_{-k}(k) \end{pmatrix} \begin{pmatrix} \lambda_{1,1}^{(k)} \\ \lambda_{1,2}^{(k)} \\ \vdots \\ \lambda_{1,k+1}^{(k)} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k+1) \end{pmatrix}, \quad (7.39)$$

where $\omega_s(i) := \rho(s) - \lambda_{0,1}^{(i)} \rho(s+1) - \dots - \lambda_{0,i}^{(i)} \rho(s+i)$, $s \in \mathbb{Z}$, $i \in \mathbb{N}$. Note that $\omega_s(k) = 0$ for $s = -1, \dots, -k$ by (7.38).

The recursion now proceeds with the best linear approximation of Y_t based on $Y_{t-1}, \dots, Y_{t-k}, R_{t-1,1}^{(k)}, R_{t-2,0}^{(k)}$ with

$$R_{t-1,1}^{(k)} := Y_{t-1} - \lambda_{1,1}^{(k)} Y_{t-2} - \dots - \lambda_{1,k}^{(k)} Y_{t-k-1} - \lambda_{1,k+1}^{(k)} R_{t-2,0}^{(k)}$$

being the lagged residual of the previously executed regression. After $l > 0$ such iterations, we get

$$\begin{aligned} & \lambda_{0,1}^{(k+l)} Y_{t-1} + \dots + \lambda_{0,k+l}^{(k+l)} Y_{t-k-l} \\ &= \lambda_{l,1}^{(k)} Y_{t-1} + \dots + \lambda_{l,k}^{(k)} Y_{t-k} \\ & \quad + \lambda_{l,k+1}^{(k)} R_{t-1,l-1}^{(k)} + \lambda_{l,k+2}^{(k)} R_{t-2,l-2}^{(k)} + \dots + \lambda_{l,k+l}^{(k)} R_{t-l,0}^{(k)}, \end{aligned}$$

where

$$R_{t',l'}^{(k)} := Y_{t'} - \lambda_{l',1}^{(k)} Y_{t'-1} - \cdots - \lambda_{l',k}^{(k)} Y_{t'-k} - \sum_{i=1}^{l'} \lambda_{l',k+i}^{(k)} R_{t'-i,l'-i}^{(k)}$$

for $t' \in \mathbb{Z}$, $l' \in \mathbb{N}_0$, $\sum_{i=1}^0 \lambda_{l',k+i}^{(k)} R_{t'-i,l'-i}^{(k)} := 0$ and where all occurring coefficients are real valued. Since, on the other hand, $R_{t,l}^{(k)} = Y_t - \lambda_{0,1}^{(k+l)} Y_{t-1} - \cdots - \lambda_{0,k+l}^{(k+l)} Y_{t-k-l}$, the coefficients satisfy

$$\begin{pmatrix} \mathbf{I}_k & \begin{matrix} 1 & 0 & \cdots & 0 \\ -\lambda_{0,1}^{(k+l-1)} & 1 & \ddots & \vdots \\ -\lambda_{0,2}^{(k+l-1)} & -\lambda_{0,1}^{(k+l-2)} & \ddots & \\ \vdots & \vdots & & 0 \\ \vdots & \vdots & & 1 \\ \vdots & \vdots & & -\lambda_{0,1}^{(k)} \end{matrix} \\ \mathbf{0}_{lk} & \begin{matrix} \vdots & \vdots & \ddots & \vdots \\ -\lambda_{0,k+l-2}^{(k+l-1)} & -\lambda_{0,k+l-3}^{(k+l-2)} & \cdots & -\lambda_{0,k-1}^{(k)} \\ -\lambda_{0,k+l-1}^{(k+l-1)} & -\lambda_{0,k+l-2}^{(k+l-2)} & \cdots & -\lambda_{0,k}^{(k)} \end{matrix} \end{pmatrix} \begin{pmatrix} \lambda_{l,1}^{(k)} \\ \lambda_{l,2}^{(k)} \\ \lambda_{l,3}^{(k)} \\ \vdots \\ \lambda_{l,k+l-1}^{(k)} \\ \lambda_{l,k+l}^{(k)} \end{pmatrix} = \begin{pmatrix} \lambda_{0,1}^{(k+l)} \\ \vdots \\ \lambda_{0,k+l}^{(k+l)} \end{pmatrix},$$

where \mathbf{I}_k denotes the $k \times k$ -identity matrix and $\mathbf{0}_{lk}$ denotes the $l \times k$ -matrix consisting of zeros. Applying again (7.38) yields

$$\begin{pmatrix} \rho(0) & \cdots & \rho(k-1) & \omega_0(k+l-1) & \omega_1(k+l-2) & \cdots & \omega_{l-1}(k) \\ \rho(-1) & \cdots & \rho(k-2) & 0 & \omega_0(k+l-1) & \cdots & \omega_{l-2}(k) \\ \vdots & & \vdots & 0 & 0 & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \omega_0(k) \\ \vdots & & \vdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \rho(-k-l+1) & \cdots & \rho(-l) & 0 & 0 & \cdots & 0 \end{pmatrix} \cdot \begin{pmatrix} \lambda_{l,1}^{(k)} \\ \lambda_{l,2}^{(k)} \\ \lambda_{l,3}^{(k)} \\ \vdots \\ \lambda_{l,k+l}^{(k)} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \rho(3) \\ \vdots \\ \rho(k+l) \end{pmatrix}. \quad (7.40)$$

If we assume that the coefficients $\lambda_{l,1}^{(k)}, \dots, \lambda_{l,k+l}^{(k)}$ are uniquely deter-

mined, then, for $k = p$ and $l \geq q$,

$$\lambda_{l,s}^{(p)} := \begin{cases} a_s & \text{if } 1 \leq s \leq p \\ \left(\sigma^2 \sum_{v=s-p}^q b_q \alpha_{v-s+p} - \sum_{i=1}^{q+p-s} \omega_i (2p+l-s-i) \lambda_{l,s+i}^{(p)} \right) / \\ \quad \omega_0 (2p+l-s) & \text{if } p+1 \leq s \leq p+q \\ 0 & \text{if } s > q+p \end{cases} \quad (7.41)$$

turns out to be the uniquely determined solution of the equation system (7.40) by means of Theorem 2.2.8. Thus, we have the representation of a MA(q)-process

$$Y_t - \sum_{s=1}^p \lambda_{l,s}^{(p)} Y_{t-s} = Y_t - \sum_{s=1}^p a_s Y_{t-s} = \sum_{v=0}^q b_v \varepsilon_{t-v},$$

whose autocorrelation function will cut off after lag q .

Now, in practice, we initiate the iteration by fitting a pure AR(k)-regression model

$$\hat{y}_{t,0}^{(k)} := \hat{\lambda}_{0,1}^{(k)} \tilde{y}_{t-1} + \cdots + \hat{\lambda}_{0,k}^{(k)} \tilde{y}_{t-k}$$

to the zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$ for $t = 1+k, \dots, n$ by means of the least squares method. That is, we need to minimize the empirical residual sum of squares

$$\begin{aligned} RSS(\hat{\lambda}_{0,1}^{(k)}, \dots, \hat{\lambda}_{0,k}^{(k)}) &= \sum_{t=1}^{n-k} (\tilde{y}_{t+k} - \hat{y}_{t+k,0}^{(k)})^2 \\ &= \sum_{t=1}^{n-k} (\tilde{y}_{t+k} - \hat{\lambda}_{0,1}^{(k)} \tilde{y}_{t+k-1} + \cdots + \hat{\lambda}_{0,k}^{(k)} \tilde{y}_t)^2. \end{aligned}$$

Using partial derivatives and equating them to zero yields the follow-

ing equation system

$$\begin{pmatrix} \sum_{t=1}^{n-k} \tilde{y}_{t+k} \tilde{y}_{t+k-1} \\ \vdots \\ \sum_{t=1}^{n-k} \tilde{y}_{t+k} \tilde{y}_t \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n-k} \tilde{y}_{t+k-1} \tilde{y}_{t+k-1} & \cdots & \sum_{t=1}^{n-k} \tilde{y}_{t+k-1} \tilde{y}_t \\ \vdots & \ddots & \vdots \\ \sum_{t=1}^{n-k} \tilde{y}_t \tilde{y}_{t+k-1} & \cdots & \sum_{t=1}^{n-k} \tilde{y}_t \tilde{y}_t \end{pmatrix} \begin{pmatrix} \hat{\lambda}_{0,1}^{(k)} \\ \vdots \\ \hat{\lambda}_{0,k}^{(k)} \end{pmatrix}, \quad (7.42)$$

which can be approximated by

$$\begin{pmatrix} c(1) \\ \vdots \\ c(k) \end{pmatrix} = \begin{pmatrix} c(0) & \cdots & c(k-1) \\ \vdots & \ddots & \vdots \\ c(-k+1) & \cdots & c(0) \end{pmatrix} \begin{pmatrix} \hat{\lambda}_{0,1}^{(k)} \\ \vdots \\ \hat{\lambda}_{0,k}^{(k)} \end{pmatrix}$$

for n sufficiently large and k sufficiently small, where the terms $c(i) = \frac{1}{n} \sum_{t=1}^{n-i} \tilde{y}_{t+i} \tilde{y}_t$, $i = 1, \dots, k$, denote the empirical autocorrelation coefficients at lag i . We define until the end of this section for a given real valued series $(c_i)_{i \in \mathbb{N}}$

$$\sum_{i=j}^h c_i := 0 \text{ if } h < j \text{ as well as } \prod_{i=j}^h c_i := 1 \text{ if } h < j.$$

Thus, proceeding the iteration, we recursively obtain after $l > 0$ steps

$$\hat{y}_{t,l}^{(k)} := \hat{\lambda}_{l,1}^{(k)} \tilde{y}_{t-1} + \cdots + \hat{\lambda}_{l,k}^{(k)} \tilde{y}_{t-k} + \hat{\lambda}_{l,k+1}^{(k)} r_{t-1,l-1}^{(k)} + \cdots + \hat{\lambda}_{l,k+l}^{(k)} r_{t-l,0}^{(k)}$$

for $t = 1 + k + l, \dots, n$, where

$$r_{t',l'}^{(k)} := \tilde{y}_{t'} - \sum_{i=1}^k \tilde{y}_{t'-i} \hat{\lambda}_{l',i}^{(k)} - \sum_{j=1}^{l'} \hat{\lambda}_{l',k+j}^{(k)} r_{t'-j,l'-j}^{(k)} \quad (7.43)$$

for $t' \in \mathbb{Z}$, $l' \in \mathbb{N}_0$.

Furthermore, we receive the equation system

$$\begin{pmatrix} c(0) & \dots & c(k-1) & \tilde{\omega}_0(k+l-1) & \tilde{\omega}_1(k+l-2) & \dots & \tilde{\omega}_{l-1}(k) \\ c(-1) & \dots & c(k-2) & 0 & \tilde{\omega}_0(k+l-1) & \dots & \tilde{\omega}_{l-2}(k) \\ \vdots & & \vdots & 0 & 0 & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \tilde{\omega}_0(k) \\ \vdots & & \vdots & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ c(-k-l) & \dots & c(-l-1) & 0 & 0 & \dots & 0 \end{pmatrix} \cdot \begin{pmatrix} \hat{\lambda}_{l1}^{(k)} \\ \hat{\lambda}_{l2}^{(k)} \\ \hat{\lambda}_{l3}^{(k)} \\ \vdots \\ \hat{\lambda}_{lk+l}^{(k)} \end{pmatrix} = \begin{pmatrix} c(1) \\ c(2) \\ c(3) \\ \vdots \\ c(k+l) \end{pmatrix},$$

where $\tilde{\omega}_s(i) := c(s) - \lambda_{0,1}^{(i)}c(s+1) - \dots - \lambda_{0,i}^{(i)}c(s+i)$, $s \in \mathbb{Z}$, $i \in \mathbb{N}$. Since $\frac{1}{n} \sum_{t=1}^{n-k} Y_{t-k} Y_t \xrightarrow{P} \gamma(k)$ as $n \rightarrow \infty$ by Theorem 7.2.13, we attain, for $k = p$, that $(\hat{\lambda}_{l,1}^{(p)}, \dots, \hat{\lambda}_{l,p}^{(p)})^T \approx (\lambda_{l,1}^{(p)}, \dots, \lambda_{l,p}^{(p)})^T = (a_1, \dots, a_p)^T$ for n sufficiently large, if, actually, the considered model (7.31) with $p = k$ and $q \leq l$ is underlying the zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$. Thereby,

$$\begin{aligned} Z_{t,l}^{(p)} &:= Y_t - \hat{\lambda}_{l,1}^{(p)} Y_{t-1} - \hat{\lambda}_{l,2}^{(p)} Y_{t-2} - \dots - \hat{\lambda}_{l,p}^{(p)} Y_{t-p} \\ &\approx Y_t - a_1 Y_{t-1} - a_2 Y_{t-2} - \dots - a_p Y_{t-p} \\ &= \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q} \end{aligned}$$

for $t \in \mathbb{Z}$ behaves like a MA(q)-process, if n is large enough. Defining

$$z_{t,l}^{(p)} := \tilde{y}_t - \hat{\lambda}_{l,1}^{(p)} \tilde{y}_{t-1} - \hat{\lambda}_{l,2}^{(p)} \tilde{y}_{t-2} - \dots - \hat{\lambda}_{l,p}^{(p)} \tilde{y}_{t-p}, \quad t = p+1, \dots, n,$$

we attain a realization $z_{p+1,l}^{(p)}, \dots, z_{n,l}^{(p)}$ of the process. Now, the empirical autocorrelation coefficient at lag $l+1$ can be computed, which will be close to zero, if the zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$ actually results from an invertible ARMA(p, q)-process satisfying the stationarity condition with expectation $E(Y_t) = 0$, variance $\text{Var}(Y_t) > 0$ and $q \leq l$.

Now, if $k > p, l \geq q$, then, in general, $\hat{\lambda}_{lk}^{(k)} \neq 0$. Hence, we can't assume that $Z_{t,l}^{(k)}$ will behave like a MA(q)-process. We can handle this by considering the following property.

An ARMA(p, q)-process $Y_t = a_1 Y_{t-1} + \cdots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}$ can be more conveniently rewritten as

$$\alpha(B)[Y_t] = \beta(B)[\varepsilon_t], \quad (7.44)$$

where

$$\begin{aligned} \alpha(B) &:= 1 - a_1 B - \cdots - a_p B^p, \\ \beta(B) &:= 1 + b_1 B + \cdots + b_q B^q \end{aligned}$$

are the characteristic polynomials in B of the autoregressive and moving average part of the ARMA(p, q)-process, respectively, and where B denotes the *backward shift operator* defined by $B^s Y_t = Y_{t-s}$ and $B^s \varepsilon_t = \varepsilon_{t-s}$, $s \in \mathbb{Z}$. If the process satisfies the stationarity condition, then, by Theorem 2.1.11, there exists an absolutely summable causal inverse filter $(d_u)_{u \geq 0}$ of the filter $c_0 = 1, c_u = a_u, u = 1, \dots, p, c_u = 0$ elsewhere. Hence, we may define $\alpha(B)^{-1} := \sum_{u \geq 0} d_u B^u$ and attain

$$Y_t = \alpha(B)^{-1} \beta(B)[\varepsilon_t]$$

as the almost surely stationary solution of $(Y_t)_{t \in \mathbb{Z}}$. Note that $B^i B^j = B^{i+j}$ for $i, j \in \mathbb{Z}$. Theorem 2.1.7 provides the covariance generating function

$$\begin{aligned} G(B) &= \sigma^2 \left(\alpha(B)^{-1} \beta(B) \right) \left(\alpha(B^{-1})^{-1} \beta(B^{-1}) \right) \\ &= \sigma^2 \left(\alpha(B)^{-1} \beta(B) \right) \left(\alpha(F)^{-1} \beta(F) \right), \end{aligned}$$

where $F = B^{-1}$ denotes the *forward shift operator*. Consequently, if we multiply both sides in (7.44) by an arbitrary factor $(1 - \lambda B)$, $\lambda \in \mathbb{R}$, we receive the identical covariance generating function $G(B)$. Hence, we find infinitely many ARMA($p + l, q + l$)-processes, $l \in \mathbb{N}$, having the same autocovariance function as the ARMA(p, q)-process.

Thus, $Z_{t,l}^{(k)}$ behaves approximately like a MA($q + k - p$)-process and the empirical autocorrelation coefficient at lag $l + 1 + k - p$, $k > p, l \geq q$, of the sample $z_{k+1,l}^{(k)}, \dots, z_{n,l}^{(k)}$ is expected to be close to zero, if the zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$ results from the considered model (7.31). Altogether, we expect, for an invertible ARMA(p, q)-process satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance

$\text{Var}(Y_t) > 0$, the following theoretical ESACF-table, where the mentioned autocorrelation coefficients are summarized with respect to k and l

$k \backslash l$...	$q-1$	q	$q+1$	$q+2$	$q+3$...
\vdots		\vdots	\vdots	\vdots	\vdots	\vdots	
$p-1$...	X	X	X	X	X	...
p	...	X	0	0	0	0	...
$p+1$...	X	X	0	0	0	...
$p+2$...	X	X	X	0	0	...
\vdots		\vdots	\vdots	\vdots	\ddots	\ddots	\ddots

Table 7.5.1: Theoretical ESACF-table for an ARMA(p, q)-process, where the entries 'X' represent arbitrary numbers.

Furthermore, we can formulate a test $H_0 : \rho_k(l+1) = 0$ against $H_1 : \rho_k(l+1) \neq 0$ based on the approximation of Bartlett's formula (7.30) for MA(l)-processes, where $\rho_k(l+1)$ denotes the autocorrelation function of $Z_{t,l}^{(k)}$ at lag $l+1$. The hypothesis $H_0 : \rho_k(l+1) = 0$ that the autocorrelation function at lag $l+1$ is equal to zero is rejected at level α and so is the presumed ARMA(k, l)-model, if its empirical counterpart $r_k(l+1)$ lies outside of the confidence interval $I := [-\sigma_l q_{1-\alpha/2}, \sigma_l q_{1-\alpha/2}]$, where $q_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution. Equivalent, the hypothesis $H_0 : \rho_k(l+1) = 0$ is rejected, if the pertaining approximate p -value $2(1 - \Phi_{0, \sigma_l^2}(|\hat{r}_k(l+1)|))$ is significantly small, where Φ_{0, σ_l^2} denotes the normal distribution with expectation 0 and variance σ_l^2 and where $\hat{r}_k(l+1)$ denotes the autocorrelation estimator at lag $l+1$. The corresponding ESACF-table and p -values of the present case study are summarized in the concluding analysis of the order selection methods at the end of this section, cf. Listings 7.5.1e and 7.5.1f.

ESACF Algorithm

The computation of the relevant autoregressive coefficients in the ESACF-approach follows a simple algorithm

$$\hat{\lambda}_{l,r}^{(k)} = \hat{\lambda}_{l-1,r}^{(k+1)} - \hat{\lambda}_{l-1,r-1}^{(k)} \frac{\hat{\lambda}_{l-1,k+1}^{(k+1)}}{\hat{\lambda}_{l-1,k}^{(k)}} \quad (7.45)$$

for $k, l \in \mathbb{N}$ not too large and $r \in \{1, \dots, k\}$, where $\hat{\lambda}_{\cdot,0}^{(\cdot)} = -1$. Observe that the estimations in the previously discussed iteration become less reliable, if k or l are too large, due to the finite sample size n . Note that the formula remains valid for $r < 1$, if we define furthermore $\hat{\lambda}_{\cdot,i}^{(\cdot)} = 0$ for negative i . Notice moreover that the algorithm as well as the ESACF approach can only be executed, if $\hat{\lambda}_{l^*,k^*}^{(k^*)}$ turns out to be unequal to zero for $l^* \in \{0, \dots, l-1\}$ and $k^* \in \{k, \dots, k+l-l^*-1\}$, which, in general, is satisfied. Under latter condition above algorithm can be shown inductively.

Proof. For $k \in \mathbb{N}$ not too large and $t = k+2, \dots, n$,

$$\hat{y}_{t,0}^{(k+1)} = \hat{\lambda}_{0,1}^{(k+1)} \tilde{y}_{t-1} + \dots + \hat{\lambda}_{0,k+1}^{(k+1)} \tilde{y}_{t-k-1}$$

as well as

$$\begin{aligned} \hat{y}_{t,1}^{(k)} &= \hat{\lambda}_{1,1}^{(k)} \tilde{y}_{t-1} + \dots + \hat{\lambda}_{1,k}^{(k)} \tilde{y}_{t-k} + \\ &\quad + \hat{\lambda}_{1,k+1}^{(k)} \left(\tilde{y}_{t-1} - \hat{\lambda}_{0,1}^{(k)} \tilde{y}_{t-2} - \dots - \hat{\lambda}_{0,k}^{(k)} \tilde{y}_{t-k-1} \right) \end{aligned}$$

is provided. Due to the equality of the coefficients, we find

$$\begin{aligned} \hat{\lambda}_{0,k+1}^{(k+1)} &= -\hat{\lambda}_{1,k+1}^{(k)} \hat{\lambda}_{0,k}^{(k)} \quad \text{as well as} \\ \hat{\lambda}_{0,i}^{(k+1)} &= \hat{\lambda}_{1,i}^{(k)} - \hat{\lambda}_{1,k+1}^{(k)} \hat{\lambda}_{0,i-1}^{(k)} = \hat{\lambda}_{1,i}^{(k)} + \hat{\lambda}_{0,i-1}^{(k)} \frac{\hat{\lambda}_{0,k+1}^{(k+1)}}{\hat{\lambda}_{0,k}^{(k)}} \end{aligned} \quad (7.46)$$

for $i = 1, \dots, k$. So the assertion (7.45) is proved for $l = 1$, k not too large and $r \in \{1, \dots, k\}$.

Now, let $\hat{\lambda}_{l^*,r}^{(k)} = \hat{\lambda}_{l^*-1,r}^{(k+1)} - \hat{\lambda}_{l^*-1,r-1}^{(k)} (\hat{\lambda}_{l^*-1,k+1}^{(k+1)} / (\hat{\lambda}_{l^*-1,k}^{(k)}))$ be shown for every $l^* \in \mathbb{N}$ with $1 \leq l^* < l$, $l, k \in \mathbb{N}$ not too large, $l \neq 1$, and

$r \in \{1, \dots, k\}$. Applying successively (7.43), we show, for $t = k + l' + 1, \dots, n$ and inductively on $l' \in \mathbb{N}_0$, that

$$\begin{aligned}
\hat{y}_{t,l'}^{(k)} &= \hat{\lambda}_{l',1}^{(k)} \tilde{y}_{t-1} + \dots + \hat{\lambda}_{l',k}^{(k)} \tilde{y}_{t-k} + \hat{\lambda}_{l',k+1}^{(k)} r_{t-1,l'-1}^{(k)} + \dots + \hat{\lambda}_{l',k+l'}^{(k)} r_{t-l',0}^{(k)} \\
&= \sum_{i=1}^k \hat{\lambda}_{l',i}^{(k)} \tilde{y}_{t-i} - \hat{\lambda}_{l',k+1}^{(k)} \sum_{\iota=0}^k \tilde{y}_{t-\iota-1} \hat{\lambda}_{l'-1,\iota}^{(k)} \\
&\quad - \hat{\lambda}_{l',k+1}^{(k)} \sum_{j=1}^{l'-1} \hat{\lambda}_{l'-1,k+j}^{(k)} r_{t-j-1,l'-j-1}^{(k)} + \sum_{h=2}^{l'} \hat{\lambda}_{l',k+h}^{(k)} r_{t-h,l'-h}^{(k)} \\
&= \sum_{s=0}^{l'} \sum_{i=\max\{0,1-s\}}^k \tilde{y}_{t-i-s} \hat{\lambda}_{l'-s,i}^{(k)} \left[\sum_{s_j=0 \text{ if } s_{j-1}=0}^{s_1+s_2+\dots=s} (-\hat{\lambda}_{l',k+s_1}^{(k)}) \right. \\
&\quad \left. (-\hat{\lambda}_{l'-(s-s_2),k+s_2}^{*(k)}) \dots (-\hat{\lambda}_{l'-(s-s_2-\dots-s_l'),k+s_l'}^{*(k)}) \right] \\
&= \sum_{s=0}^{l'} \sum_{i=\max\{0,1-s\}}^k \tilde{y}_{t-i-s} \hat{\lambda}_{l'-s,i}^{(k)} \Phi_{s,l'}^{(k)} \tag{7.47}
\end{aligned}$$

with

$$\Phi_{s,l'}^{(k)} := \left[\sum_{s_j=0 \text{ if } s_{j-1}=0}^{s_1+s_2+\dots=s} (-\hat{\lambda}_{l',k+s_1}^{(k)}) \prod_{j=2}^s \left(-\hat{\lambda}_{l'-(s-s_2-\dots-s_j),k+s_j}^{*(k)} \right) \right],$$

where

$$(-\hat{\lambda}_{l'-(s-s_2-\dots-s_j),k+s_j}^{*(k)}) = \begin{cases} 1 & \text{if } s_j = 0, \\ -\hat{\lambda}_{l'-(s-s_2-\dots-s_j),k+s_j}^{(k)} & \text{else.} \end{cases}$$

In the case of $l' = 0$, we receive

$$\begin{aligned}
\hat{y}_{t,0}^{(k)} &= \hat{\lambda}_{0,1}^{(k)} \tilde{y}_{t-1} + \dots + \hat{\lambda}_{0,k}^{(k)} \tilde{y}_{t-k} \\
&= \sum_{i=1}^k \tilde{y}_{t-i} \hat{\lambda}_{0,i}^{(k)}.
\end{aligned}$$

So, let the assertion (7.47) be shown for one $l' \in \mathbb{N}_0$. The induction

is then accomplished by considering

$$\begin{aligned}
 \hat{y}_{t,l'+1}^{(k)} &= \hat{\lambda}_{l'+1,1}^{(k)} \tilde{y}_{t-1} + \cdots + \hat{\lambda}_{l'+1,k}^{(k)} \tilde{y}_{t-k} + \sum_{h=1}^{l'} \hat{\lambda}_{l'+1,k+h}^{(k)} r_{t-h,l'+1-h}^{(k)*} \\
 &\quad + \sum_{i=0}^k \tilde{y}_{t-i-l'-1} \hat{\lambda}_{0,i}^{(k)} \Phi_{l'+1,l'+1}^{(k)} \\
 &= \sum_{s=0}^{l'} \sum_{i=\max\{0,1-s\}}^k \tilde{y}_{t-i-s} \hat{\lambda}_{l'+1-s,i}^{(k)} \Phi_{s,l'+1}^{(k)} + \sum_{i=0}^k \tilde{y}_{t-i-l'-1} \hat{\lambda}_{0,i}^{(k)} \Phi_{l'+1,l'+1}^{(k)} \\
 &= \sum_{s=0}^{l'+1} \sum_{i=\max\{0,1-s\}}^k \tilde{y}_{t-i-s} \hat{\lambda}_{l'+1-s,i}^{(k)} \Phi_{s,l'+1}^{(k)}.
 \end{aligned}$$

Next, we show, inductively on m , that

$$\hat{\lambda}_{m,k}^{(k)} \Phi_{l-m,l}^{(k)} = \hat{\lambda}_{m,k+1}^{(k+1)} \Phi_{l-m-1,l-1}^{(k+1)} + \hat{\lambda}_{m,k}^{(k)} \Phi_{l-m,l-1}^{(k+1)} \quad (7.48)$$

for $m = 1, \dots, l-1$. Since $\hat{y}_{t,l'}^{(k)} = \hat{y}_{t,l'-1}^{(k+1)}$, we find, using (7.47), the equal factors of $\tilde{y}_{t-l'-k}$, which satisfy in conjunction with (7.46)

$$\hat{\lambda}_{0,k}^{(k)} \Phi_{l',l'}^{(k)} = \hat{\lambda}_{0,k+1}^{(k+1)} \Phi_{l'-1,l'-1}^{(k+1)} \Leftrightarrow \Phi_{l',l'}^{(k)} = -\hat{\lambda}_{1,k+1}^{(k)} \Phi_{l'-1,l'-1}^{(k+1)}. \quad (7.49)$$

Equating the factors of $\tilde{y}_{t-k-l'+1}$ yields

$$\hat{\lambda}_{0,k-1}^{(k)} \Phi_{l',l'}^{(k)} + \hat{\lambda}_{1,k}^{(k)} \Phi_{l'-1,l'}^{(k)} = \hat{\lambda}_{0,k}^{(k+1)} \Phi_{l'-1,l'-1}^{(k+1)} + \hat{\lambda}_{1,k+1}^{(k+1)} \Phi_{l'-2,l'-1}^{(k+1)}. \quad (7.50)$$

Multiplying both sides of the first equation in (7.49) with $\hat{\lambda}_{0,k-1}^{(k)}/\hat{\lambda}_{0,k}^{(k)}$, where l' is set equal to 1, and utilizing afterwards formula (7.45) with $r = k$ and $l = 1$

$$\hat{\lambda}_{0,k-1}^{(k)} \frac{\hat{\lambda}_{0,k+1}^{(k+1)}}{\hat{\lambda}_{0,k}^{(k)}} = \hat{\lambda}_{0,k}^{(k+1)} - \hat{\lambda}_{1,k}^{(k)},$$

we get

$$\hat{\lambda}_{0,k-1}^{(k)} \Phi_{l',l'}^{(k)} = \left(\hat{\lambda}_{0,k}^{(k+1)} - \hat{\lambda}_{1,k}^{(k)} \right) \Phi_{l'-1,l'-1}^{(k+1)},$$

which leads together with (7.50) to

$$\hat{\lambda}_{1,k}^{(k)} \Phi_{l'-1,l'}^{(k)} = \hat{\lambda}_{1,k+1}^{(k+1)} \Phi_{l'-2,l'-1}^{(k+1)} + \hat{\lambda}_{1,k}^{(k)} \Phi_{l'-1,l'-1}^{(k+1)}.$$

So, the assertion (7.48) is shown for $m = 1$, since we may choose $l' = l$.

Let $\hat{\lambda}_{m',k}^{(k)} \Phi_{l-m',l}^{(k)} = \hat{\lambda}_{m',k+1}^{(k+1)} \Phi_{l-m'-1,l-1}^{(k+1)} + \hat{\lambda}_{m',k}^{(k)} \Phi_{l-m',l-1}^{(k+1)}$ now be shown for every $m' \in \mathbb{N}$ with $1 \leq m' < m \leq l-1$. We define

$$\Psi_{l,m'}^{(k)} := -\hat{\lambda}_{m',k+1}^{(k+1)} \Phi_{l-1-m',l-1}^{(k+1)} + \hat{\lambda}_{m',k}^{(k)} \Phi_{l-m',l}^{(k)} - \hat{\lambda}_{m',k}^{(k)} \Phi_{l-m',l-1}^{(k+1)}$$

as well as

$$\Psi_{l,0}^{(k)} := \hat{\lambda}_{0,k}^{(k)} \Phi_{l,l}^{(k)} - \hat{\lambda}_{0,k+1}^{(k+1)} \Phi_{l-1,l-1}^{(k+1)},$$

which are equal to zero by (7.48) and (7.49). Together with (7.45)

$$\hat{\lambda}_{i,k-m+i}^{(k)} \frac{\hat{\lambda}_{i,k+1}^{(k+1)}}{\hat{\lambda}_{i,k}^{(k)}} = \hat{\lambda}_{i,k-m+i+1}^{(k+1)} - \hat{\lambda}_{i+1,k-m+i+1}^{(k)},$$

which is true for every i satisfying $0 \leq i < m$, the required condition can be derived

$$\begin{aligned} 0 &= \sum_{i=0}^{m-1} \frac{\hat{\lambda}_{i,k-m+i}^{(k)}}{\hat{\lambda}_{i,k}^{(k)}} \Psi_{l,i}^{(k)} \\ &= \sum_{s=0}^l \sum_{i=\max\{0,1-s\}, i+s=k+l-m}^{k-1} \hat{\lambda}_{l-s,i}^{(k)} \Phi_{s,l}^{(k)} \\ &\quad - \sum_{v=0}^{l-1} \sum_{j=\max\{0,1-v\}, v+j=k+l-m}^k \hat{\lambda}_{l-1-v,j}^{(k+1)} \Phi_{v,l-1}^{(k+1)} \\ &\quad + \hat{\lambda}_{m,k}^{(k)} \Phi_{l-m,l-1}^{(k+1)}. \end{aligned}$$

Noting the factors of $\tilde{y}_{t-k-l+m}$

$$\begin{aligned} &\sum_{s=0}^l \sum_{i=\max\{0,1-s\}, i+s=k+l-m}^k \hat{\lambda}_{l-s,i}^{(k)} \Phi_{s,l}^{(k)} \\ &= \sum_{v=0}^{l-1} \sum_{j=\max\{0,1-v\}, v+j=k+l-m}^{k+1} \hat{\lambda}_{l-1-v,j}^{(k+1)} \Phi_{v,l-1}^{(k+1)}, \end{aligned}$$

we finally attain

$$\hat{\lambda}_{m,k}^{(k)} \Phi_{l-m,l}^{(k)} = \hat{\lambda}_{m,k+1}^{(k+1)} \Phi_{l-m-1,l-1}^{(k+1)} + \hat{\lambda}_{m,k}^{(k)} \Phi_{l-m,l-1}^{(k+1)},$$

which completes the inductive proof of formula (7.48). Note that (7.48) becomes in the case of $m = l - 1$

$$-\hat{\lambda}_{l-1,k}^{(k)} \hat{\lambda}_{l,k+1}^{(k)} = \hat{\lambda}_{l-1,k+1}^{(k+1)} - \hat{\lambda}_{l-1,k}^{(k)} \hat{\lambda}_{l-1,k+2}^{(k+1)}. \quad (7.51)$$

Applied to the equal factors of \tilde{y}_{t-1}

$$\hat{\lambda}_{l,1}^{(k)} + \hat{\lambda}_{l,k+1}^{(k)} = \hat{\lambda}_{l-1,1}^{(k+1)} + \hat{\lambda}_{l-1,k+2}^{(k+1)}$$

yields assertion (7.45) for l, k not too large and $r = 1$,

$$\hat{\lambda}_{l,1}^{(k)} = \hat{\lambda}_{l-1,1}^{(k+1)} + \frac{\hat{\lambda}_{l-1,k+1}^{(k+1)}}{\hat{\lambda}_{l-1,k}^{(k)}}.$$

For the other cases of $r \in \{2, \dots, k\}$, we consider

$$\begin{aligned} 0 &= \sum_{i=0}^{l-2} \frac{\hat{\lambda}_{i,r-l+i}^{(k)}}{\hat{\lambda}_{i,k}^{(k)}} \Psi_{l,i}^{(k)} \\ &= \sum_{s=2}^l \sum_{i=\max\{0,1-s\}, s+i=r}^k \hat{\lambda}_{l-s,i}^{(k)} \Phi_{s,l}^{(k)} \\ &\quad - \sum_{v=1}^{l-1} \sum_{j=\max\{0,1-v\}, v+j=r}^{k+1} \hat{\lambda}_{l-1-v,j}^{(k+1)} \Phi_{v,l-1}^{(k+1)} \\ &\quad - \hat{\lambda}_{l-1,r-1}^{(k)} \hat{\lambda}_{l-1,k+2}^{(k+1)}. \end{aligned}$$

Applied to the equation of the equal factors of \tilde{y}_{t-r} leads by (7.51) to

$$\begin{aligned}
& \sum_{s=0}^l \sum_{i=\max\{0,1-s\}, s+i=r}^k \hat{\lambda}_{l-s,i}^{(k)} \Phi_{s,l}^{(k)} \\
&= \sum_{v=0}^{l-1} \sum_{j=\max\{0,1-v\}, v+j=r}^{k+1} \hat{\lambda}_{l-1-v,j}^{(k+1)} \Phi_{v,l-1}^{(k+1)} \\
&\Leftrightarrow \\
& \hat{\lambda}_{l,r}^{(k)} - \hat{\lambda}_{l-1,r-1}^{(k)} \hat{\lambda}_{l,k+1}^{(k)} + \sum_{s=2}^l \sum_{i=\max\{0,1-s\}, s+i=r}^k \hat{\lambda}_{l-s,i}^{(k)} \Phi_{s,l}^{(k)} \\
&= \hat{\lambda}_{l-1,r}^{(k+1)} + \sum_{v=1}^{l-1} \sum_{j=\max\{0,1-v\}, v+j=r}^{k+1} \hat{\lambda}_{l-1-v,j}^{(k+1)} \Phi_{v,l-1}^{(k+1)} \\
&\Leftrightarrow \\
& \hat{\lambda}_{l,r}^{(k)} = \hat{\lambda}_{l-1,r}^{(k+1)} + \hat{\lambda}_{l-1,r-1}^{(k)} \hat{\lambda}_{l,k+1}^{(k)} - \hat{\lambda}_{l-1,r-1}^{(k)} \hat{\lambda}_{l-1,k+2}^{(k+1)} \\
&\Leftrightarrow \\
& \hat{\lambda}_{l,r}^{(k)} = \hat{\lambda}_{l-1,r}^{(k+1)} - \hat{\lambda}_{l-1,r-1}^{(k)} \frac{\hat{\lambda}_{l-1,k+1}^{(k+1)}}{\hat{\lambda}_{l-1,k}^{(k)}}
\end{aligned}$$

□

Remark 7.5.2. The autoregressive coefficients of the subsequent iterations are uniquely determined by (7.45), once the initial approximation coefficients $\hat{\lambda}_{0,1}^{(k)}, \dots, \hat{\lambda}_{0,k}^{(k)}$ have been provided for all k not too large. Since we may carry over the algorithm to the theoretical case, the solution (7.41) is in deed uniquely determined, if we provide $\lambda_{l^*,k^*}^{(k^*)} \neq 0$ in the iteration for $l^* \in \{0, \dots, l-1\}$ and $k^* \in \{p, \dots, p+l-l^*-1\}$.

SCAN Method

The smallest canonical correlation method, short *SCAN-method*, suggested by Tsay and Tiao (1985), seeks nonzero vectors \mathbf{a} and $\mathbf{b} \in \mathbb{R}^{k+1}$

such that the correlation

$$\text{Corr}(\mathbf{a}^T \mathbf{y}_{t,k}, \mathbf{b}^T \mathbf{y}_{s,k}) = \frac{\mathbf{a}^T \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{s,k}} \mathbf{b}}{(\mathbf{a}^T \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}} \mathbf{a})^{1/2} (\mathbf{b}^T \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{s,k}} \mathbf{b})^{1/2}} \quad (7.52)$$

is minimized, where $\mathbf{y}_{t,k} := (Y_t, \dots, Y_{t-k})^T$ and $\mathbf{y}_{s,k} := (Y_s, \dots, Y_{s-k})^T$, $s, t \in \mathbb{Z}, t > s$, are random vectors of $k+1$ process variables, and $\Sigma_{VW} := \text{Cov}(V, W)$. Setting $\boldsymbol{\alpha}_{t,k} := \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1/2} \mathbf{a}$ and $\boldsymbol{\beta}_{s,k} := \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{s,k}}^{-1/2} \mathbf{b}$, we attain

$$\begin{aligned} \left| \text{Corr}(\mathbf{a}^T \mathbf{y}_{t,k}, \mathbf{b}^T \mathbf{y}_{s,k}) \right| &= \frac{\left| \boldsymbol{\alpha}_{t,k}^T \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1/2} \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{s,k}} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{s,k}}^{-1/2} \boldsymbol{\beta}_{s,k} \right|}{(\boldsymbol{\alpha}_{t,k}^T \boldsymbol{\alpha}_{t,k})^{1/2} (\boldsymbol{\beta}_{s,k}^T \boldsymbol{\beta}_{s,k})^{1/2}} \\ &\leq \frac{(\boldsymbol{\alpha}_{t,k}^T \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1/2} \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{s,k}} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{s,k}}^{-1} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{t,k}} \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1/2} \boldsymbol{\alpha}_{t,k})^{1/2}}{(\boldsymbol{\alpha}_{t,k}^T \boldsymbol{\alpha}_{t,k})^{1/2}} \\ &=: R_{s,t,k}(\boldsymbol{\alpha}_{t,k}) \end{aligned} \quad (7.53)$$

by the Cauchy-Schwarz inequality. Since $R_{s,t,k}(\boldsymbol{\alpha}_{t,k})^2$ is a Rayleigh quotient, it maps the $k+1$ eigenvectors of the matrix

$$\mathbf{M}_{t-s,k} := \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1/2} \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{s,k}} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{s,k}}^{-1} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{t,k}} \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1/2}, \quad (7.54)$$

which only depends on the lag $t-s$ and k , on the corresponding eigenvalues, which are called *squared canonical correlations* (see for example Horn and Johnson, 1985). Hence, there exist nonzero vectors \mathbf{a} and $\mathbf{b} \in \mathbb{R}^{k+1}$ such that the squared value of above correlation (7.52) is lower or equal than the smallest eigenvalue of $\mathbf{M}_{t-s,k}$. As $\mathbf{M}_{t-s,k}$ and

$$\Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{-1} \Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{s,k}} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{s,k}}^{-1} \Sigma_{\mathbf{y}_{s,k} \mathbf{y}_{t,k}}, \quad (7.55)$$

have the same $k+1$ eigenvalues $\lambda_{1,t-s,k} \leq \dots \leq \lambda_{k+1,t-s,k}$ with corresponding eigenvectors $\Sigma_{\mathbf{y}_{t,k} \mathbf{y}_{t,k}}^{1/2} \boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_i, i = 1, \dots, k+1$, respectively, we henceforth focus on matrix (7.55).

Now, given an invertible ARMA(p, q)-process satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$. For $k+1 > p$ and $l+1 := t-s > q$, we find, by Theorem 2.2.8, the normalized eigenvector $\boldsymbol{\mu}_1 := \eta \cdot (1, -a_1, \dots, -a_p, 0, \dots, 0)^T \in \mathbb{R}^{k+1}, \eta \in$

\mathbb{R} , of $\Sigma_{\mathbf{y}_{t,k}\mathbf{y}_{t,k}}^{-1} \Sigma_{\mathbf{y}_{t,k}\mathbf{y}_{s,k}} \Sigma_{\mathbf{y}_{s,k}\mathbf{y}_{s,k}}^{-1} \Sigma_{\mathbf{y}_{s,k}\mathbf{y}_{t,k}}$ with corresponding eigenvalue zero, since $\Sigma_{\mathbf{y}_{s,k}\mathbf{y}_{t,k}} \boldsymbol{\mu}_1 = 0 \cdot \boldsymbol{\mu}_1$. Accordingly, the two linear combinations $Y_t - a_1 Y_{t-1} - \dots - a_p Y_{t-p}$ and $Y_s - a_1 Y_{s-1} - \dots - a_p Y_{s-p}$ are uncorrelated for $t - s > q$ by (7.53), which isn't surprising, since they are representations of two MA(q)-processes, which are in deed uncorrelated for $t - s > q$.

In practice, given a zero-mean time series $\tilde{y}_1, \dots, \tilde{y}_n$ with sufficiently large n , the SCAN method computes the smallest eigenvalues $\hat{\lambda}_{1,t-s,k}$ of the empirical counterpart of (7.55)

$$\hat{\Sigma}_{\mathbf{y}_{t,k}\mathbf{y}_{t,k}}^{-1} \hat{\Sigma}_{\mathbf{y}_{t,k}\mathbf{y}_{s,k}} \hat{\Sigma}_{\mathbf{y}_{s,k}\mathbf{y}_{s,k}}^{-1} \hat{\Sigma}_{\mathbf{y}_{s,k}\mathbf{y}_{t,k}},$$

where the autocovariance coefficients $\gamma(m)$ are replaced by their empirical counterparts $c(m)$. Since $\hat{c}(m) \xrightarrow{P} \gamma(m)$ as $n \rightarrow \infty$ by Theorem 7.2.13, we expect that $\hat{\lambda}_{1,t-s,k} \approx \lambda_{1,t-s,k}$, if $t - s$ and k are not too large. Thus, $\hat{\lambda}_{1,t-s,k} \approx 0$ for $k + 1 > p$ and $l + 1 = t - s > q$, if the zero-mean time series is actually a realization of an invertible ARMA(p, q)-process satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$.

Applying Bartlett's formula (7.28) from page 272 and the asymptotic distribution of canonical correlations (e.g. Anderson, 1984, Section 13.5), the following asymptotic test statistic can be derived for testing if the smallest eigenvalue equals zero,

$$c(k, l) = -(n - l - k) \ln(1 - \hat{\lambda}_{1,l,k}^* / d(k, l)),$$

where $d(k, l) / (n - l - k)$ is an approximation of the variance of the root of an appropriate estimator $\hat{\lambda}_{1,l,k}^*$ of $\lambda_{1,l,k}$. Note that $\hat{\lambda}_{1,l,k}^* \approx 0$ entails $c(k, l) \approx 0$. The test statistics $c(k, l)$ are computed with respect to k and l in a chosen order range of $p_{\min} \leq k \leq p_{\max}$ and $q_{\min} \leq l \leq q_{\max}$. In the theoretical case, if actually an invertible ARMA(p, q)-model satisfying the stationarity condition with expectation $E(Y_t) = 0$ and variance $\text{Var}(Y_t) > 0$ underlies the sample, we would attain the following SCAN-table, consisting of the corresponding test statistics $c(k, l)$

Under the hypothesis H_0 that the ARMA(p, q)-model (7.31) is actually underlying the time series $\tilde{y}_1, \dots, \tilde{y}_n$, it can be shown that $c(k, l)$ asymptotically follows a χ_1^2 -distribution, if $p \leq k \leq p_{\max}$ and

$k \backslash l$...	$q-1$	q	$q+1$	$q+2$	$q+3$...
\vdots		\vdots	\vdots	\vdots	\vdots	\vdots	
$p-1$...	X	X	X	X	X	...
p	...	X	0	0	0	0	...
$p+1$...	X	0	0	0	0	...
$p+2$...	X	0	0	0	0	...
\vdots		\vdots	\vdots	\vdots	\vdots	\vdots	

Table 7.5.2: Theoretical SCAN-table for an ARMA(p, q)-process, where the entries 'X' represent arbitrary numbers.

$q \leq l \leq q_{max}$. We would therefore reject the ARMA(p, q)-model (7.31), if one of the test statistics $c(k, l)$ is unexpectedly large or, respectively, if one of the corresponding p -values $1 - \chi_1^2(c(k, l))$ is too small for $k \in \{p, \dots, p_{max}\}$ and $l \in \{q, \dots, q_{max}\}$.

Note that the SCAN method is also valid in the case of nonstationary ARMA(p, q)-processes (Tsay and Tiao, 1985).

Minimum Information Criterion									
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 0	9.275	9.042	8.830	8.674	8.555	8.461	8.386	8.320	8.256
AR 1	7.385	7.257	7.257	7.246	7.240	7.238	7.237	7.237	7.230
AR 2	7.270	7.257	7.241	7.234	7.235	7.236	7.237	7.238	7.239
AR 3	7.238	7.230	7.234	7.235	7.236	7.237	7.238	7.239	7.240
AR 4	7.239	7.234	7.235	7.236	7.237	7.238	7.239	7.240	7.241
AR 5	7.238	7.235	7.237	7.230	7.238	7.239	7.240	7.241	7.242
AR 6	7.237	7.236	7.237	7.238	7.239	7.240	7.241	7.242	7.243
AR 7	7.237	7.237	7.238	7.239	7.240	7.241	7.242	7.243	7.244
AR 8	7.238	7.238	7.239	7.240	7.241	7.242	7.243	7.244	7.246
Error series model: AR(15)									
Minimum Table Value: BIC(2,3) = 7.234557									

Listing 7.5.1: MINIC statistics (BIC values).

Squared Canonical Correlation Estimates								
Lags	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 1	0.0017	0.0110	0.0057	0.0037	0.0022	0.0003	<.0001	<.0001
AR 2	0.0110	0.0052	<.0001	<.0001	0.0003	<.0001	<.0001	<.0001
AR 3	0.0011	0.0006	<.0001	<.0001	0.0002	<.0001	<.0001	0.0001
AR 4	<.0001	0.0002	0.0003	0.0002	<.0001	<.0001	<.0001	<.0001
AR 5	0.0002	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
AR 6	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
AR 7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
AR 8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

Listing 7.5.1b: SCAN table.

SCAN Chi-Square[1] Probability Values								
Lags	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 1	0.0019	<.0001	<.0001	<.0001	0.0004	0.1913	0.9996	0.9196
AR 2	<.0001	<.0001	0.5599	0.9704	0.1981	0.5792	0.9209	0.9869
AR 3	0.0068	0.0696	0.9662	0.5624	0.3517	0.6279	0.5379	0.5200
AR 4	0.9829	0.3445	0.1976	0.3302	0.5856	0.7427	0.5776	0.8706
AR 5	0.3654	0.6321	0.7352	0.9621	0.7129	0.5468	0.6914	0.9008
AR 6	0.6658	0.6065	0.8951	0.8164	0.6402	0.6789	0.6236	0.9755
AR 7	0.6065	0.6318	0.7999	0.9358	0.7356	0.7688	0.9253	0.7836
AR 8	0.9712	0.9404	0.6906	0.7235	0.7644	0.8052	0.8068	0.9596

Listing 7.5.1c: P-Values of SCAN-approach.

Extended Sample Autocorrelation Function								
Lags	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 1	-0.0471	-0.1214	-0.0846	-0.0662	-0.0502	-0.0184	-0.0000	-0.0014
AR 2	-0.1627	-0.0826	0.0111	-0.0007	-0.0275	-0.0184	-0.0002	-0.0002
AR 3	-0.1219	-0.1148	0.0054	-0.0005	-0.0252	-0.0200	0.0020	-0.0005
AR 4	-0.0016	-0.1120	-0.0035	-0.0377	-0.0194	0.0086	-0.0119	-0.0029
AR 5	0.0152	-0.0607	-0.0494	0.0062	-0.0567	-0.0098	-0.0087	0.0034
AR 6	0.1803	-0.1915	-0.0708	-0.0067	-0.0260	0.0264	-0.0081	0.0013
AR 7	-0.4009	-0.2661	-0.1695	0.0091	-0.0240	0.0326	-0.0066	0.0008
AR 8	0.0266	0.0471	-0.2037	-0.0243	-0.0643	0.0226	0.0140	-0.0008

Listing 7.5.1d: ESACF table.

ESACF Probability Values								
Lags	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 1	0.0003	<.0001	<.0001	<.0001	0.0002	0.1677	0.9992	0.9151
AR 2	<.0001	<.0001	0.3874	0.9565	0.0261	0.1675	0.9856	0.9845
AR 3	<.0001	<.0001	0.6730	0.9716	0.0496	0.1593	0.8740	0.9679
AR 4	0.9103	<.0001	0.7851	0.0058	0.1695	0.6034	0.3700	0.8187

AR 5	0.2879	0.0004	0.0020	0.6925	<.0001	0.4564	0.5431	0.8464
AR 6	<.0001	<.0001	<.0001	0.6504	0.0497	0.0917	0.5944	0.9342
AR 7	<.0001	<.0001	<.0001	0.5307	0.0961	0.0231	0.6584	0.9595
AR 8	0.0681	0.0008	<.0001	0.0802	<.0001	0.0893	0.2590	0.9524

Listing 7.5.1e: P-Values of ESACF-approach.

ARMA (p+d,q) Tentative Order Selection Tests					
-----SCAN-----			-----ESACF-----		
p+d	q	BIC	p+d	q	BIC
3	2	7.234889	4	5	7.238272
2	3	7.234557	1	6	7.237108
4	1	7.234974	2	6	7.237619
1	6	7.237108	3	6	7.238428
			6	6	7.241525
			8	6	7.243719
(5% Significance Level)					

Listing 7.5.1f: Order suggestions of SAS, based on the methods SCAN and ESACF.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	-3.77827	7.19241	-0.53	0.5994	0
MA1,1	0.34553	0.03635	9.51	<.0001	1
MA1,2	0.34821	0.01508	23.10	<.0001	2
MA1,3	0.10327	0.01501	6.88	<.0001	3
AR1,1	1.61943	0.03526	45.93	<.0001	1
AR1,2	-0.63143	0.03272	-19.30	<.0001	2

Listing 7.5.1g: Parameter estimates of ARMA(2,3)-model.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	-3.58738	7.01935	-0.51	0.6093	0
MA1,1	0.58116	0.03196	18.19	<.0001	1
MA1,2	0.23329	0.03051	7.65	<.0001	2
AR1,1	1.85383	0.03216	57.64	<.0001	1
AR1,2	-1.04408	0.05319	-19.63	<.0001	2
AR1,3	0.17899	0.02792	6.41	<.0001	3

Listing 7.5.1h: Parameter estimates of ARMA(3,2)-model.

```

1  /* donauwoerth_orderselection.sas */
2  TITLE1 'Order Selection';
3  TITLE2 'Donauwoerth Data';
4
5  /* Note that this program requires the file 'seasad' generated by
   ↪ program donauwoerth_adjustment.sas */
6
7  /* Order selection by means of MINIC, SCAN and ESACF approach */
8  PROC ARIMA DATA=seasad;
9      IDENTIFY VAR=sa MINIC PERROR=(8:20) SCAN ESACF p=(1:8) q=(1:8);
10 RUN;
11
12 /* Estimate model coefficients for chosen orders p and q */
13 PROC ARIMA DATA=seasad;
14     IDENTIFY VAR=sa NLAG=100 NOPRINT;
15     ESTIMATE METHOD=CLS p=2 q=3;
16 RUN;
17
18 PROC ARIMA DATA=seasad;
19     IDENTIFY VAR=sa NLAG=100 NOPRINT;
20     ESTIMATE METHOD=CLS p=3 q=2;
21 RUN; QUIT;

```

The discussed order identification methods are carried out in the framework of PROC ARIMA, invoked by the options MINIC, SCAN and ESACF in the IDENTIFY statement. $p = (1 : 8)$ and $q = (1 : 8)$ restrict the order ranges, admitting integer numbers in the range of 1 to 8 for the autoregressive and moving average order, respectively. By PERROR, the order

of the autoregressive model is specified that is used for estimating the error sequence in the MINIC approach. The second part of the program estimates the coefficients of the preferred ARMA(2,3)- and ARMA(3,2)-models, using conditional least squares, specified by METHOD=CLS in the ESTIMATE statement.

p	q	BIC
2	3	7.234557
3	2	7.234889
4	1	7.234974
3	3	7.235731
2	4	7.235736
5	1	7.235882
4	2	7.235901

Table 7.5.3: Smallest BIC-values with respect to orders p and q .

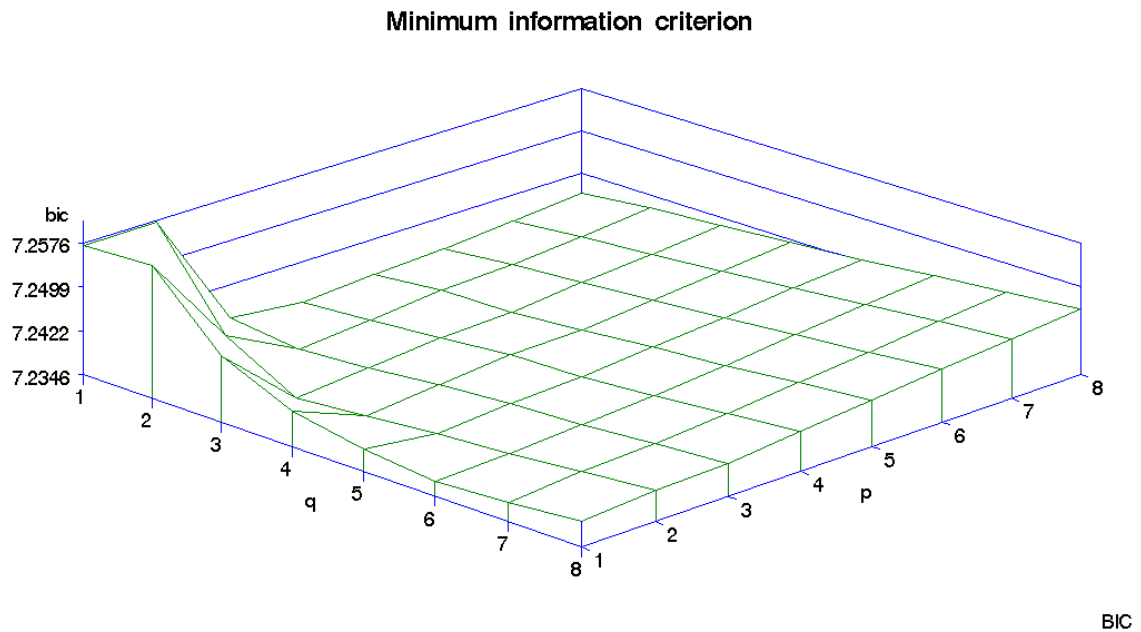


Figure 7.5.2: Graphical illustration of BIC-values.

SAS advocates an ARMA(2,3)-model as appropriate model choose, referring to the minimal BIC-value 7.234557 in Listing 7.5.1. The seven smallest BIC-values are summarized in Table 7.5.3 and a graphical illustration of the BIC-values is given in Figure 7.5.2.

The next smallest BIC-value 7.234889 pertains to the adjustment of an ARMA(3,2)-model. The SCAN method also prefers these two models, whereas the ESACF technique suggests models with a numerous quantity of parameters and great BIC-values. Hence, in the following section, we will investigate the goodness of fit of both the ARMA(2,3)- and the ARMA(3,2)-model.

The coefficients of the models are estimated by the least squares method, as introduced in Section 2.3, conditional on the past estimated residuals $\hat{\varepsilon}_t$ and unobserved values \tilde{y}_t being equal to 0 for $t \leq 0$. Accordingly, this method is also called the *conditional least squares method*. SAS provides

$$\begin{aligned} Y_t - 1.85383Y_{t-1} + 1.04408Y_{t-2} - 0.17899Y_{t-3} \\ = \varepsilon_t - 0.58116\varepsilon_{t-1} - 0.23329\varepsilon_{t-2}, \end{aligned} \quad (7.56)$$

for the ARMA(3, 2)-model and

$$\begin{aligned} Y_t - 1.61943Y_{t-1} + 0.63143Y_{t-2} \\ = \varepsilon_t - 0.34553\varepsilon_{t-1} - 0.34821\varepsilon_{t-2} - 0.10327\varepsilon_{t-3} \end{aligned} \quad (7.57)$$

for the ARMA(2, 3)-model. The negative means of -3.77827 and -3.58738 are somewhat irritating in view of the used data with an approximate zero mean. Yet, this fact presents no particular difficulty, since it almost doesn't affect the estimation of the model coefficients. Both models satisfy the stationarity and invertibility condition, since the roots 1.04, 1.79, 3.01 and 1.04, 1.53 pertaining to the characteristic polynomials of the autoregressive part as well as the roots -3.66, 1.17 and 1.14, $-2.26 + 1.84i$, $-2.26 - 1.84i$ pertaining to those of the moving average part, respectively, lie outside of the unit circle. Note that SAS estimates the coefficients in the ARMA(p, q)-model $Y_t - a_1Y_{t-1} - a_2Y_{t-2} - \dots - a_pY_{t-p} = \varepsilon_t - b_1\varepsilon_{t-1} - \dots - b_q\varepsilon_{t-q}$.

7.6 Diagnostic Check

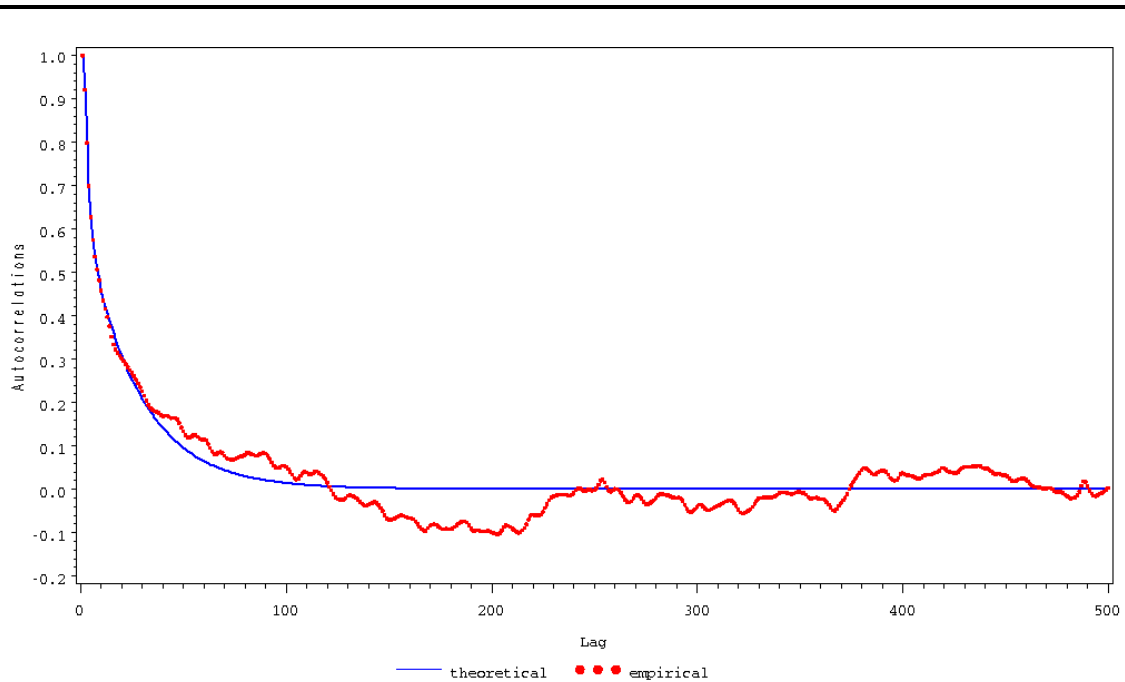
In the previous section we have adjusted an ARMA(3, 2)- as well as an ARMA(2, 3)-model to the prepared Donauwoerth time series $\tilde{y}_1, \dots, \tilde{y}_n$. The next step is to check, if the time series can be interpreted as a typical realization of these models. This *diagnostic check* can be divided in several classes:

1. Comparing theoretical structures with the corresponding empirical counterparts
2. Test of the residuals for white noise
3. Overfitting

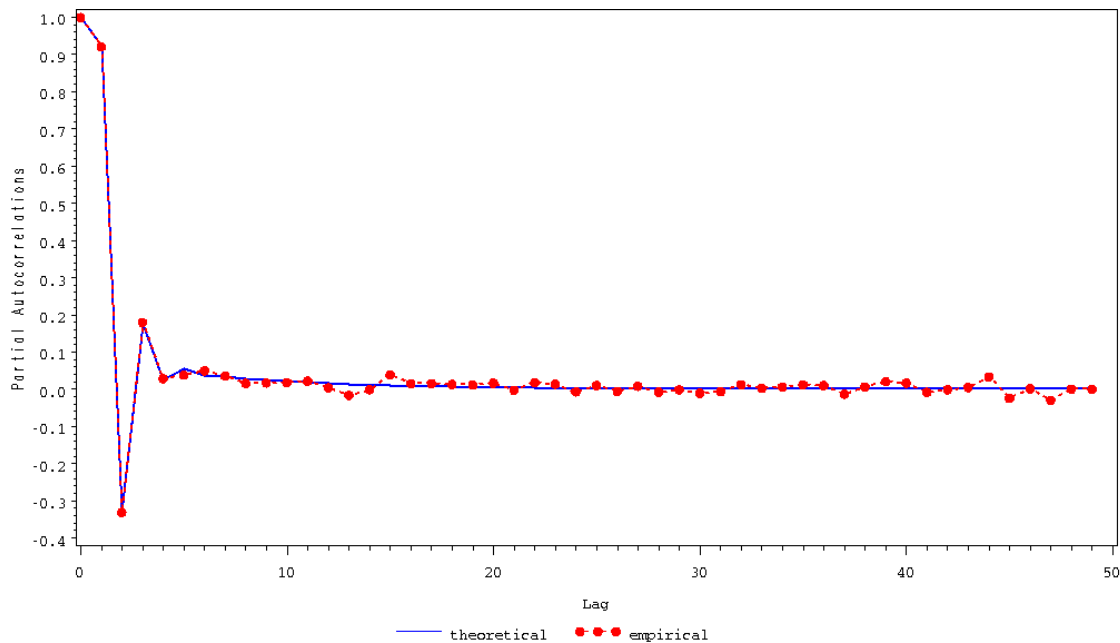
Comparing Theoretical and Empirical Structures

An obvious method of the first class is the comparison of the empirical partial and empirical general autocorrelation function, $\hat{\alpha}(k)$ and $r(k)$, with the theoretical partial and theoretical general autocorrelation function, $\alpha_{p,q}(k)$ and $\rho_{p,q}(k)$, of the assumed ARMA(p, q)-model. The

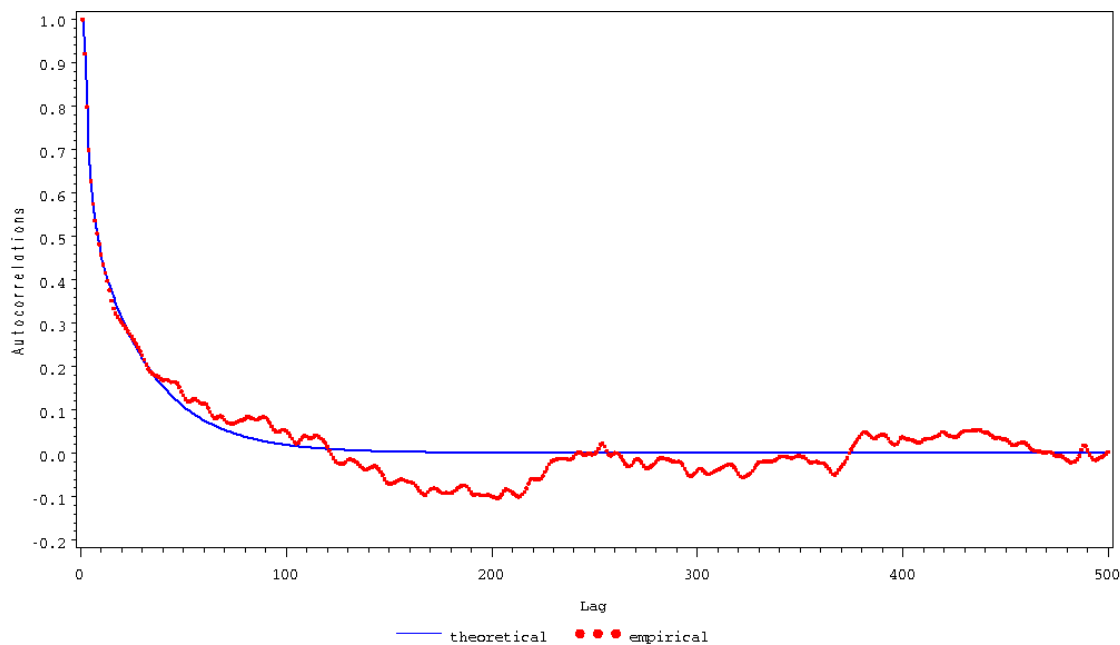
following plots show these comparisons for the identified ARMA(3, 2)- and ARMA(2, 3)-model.



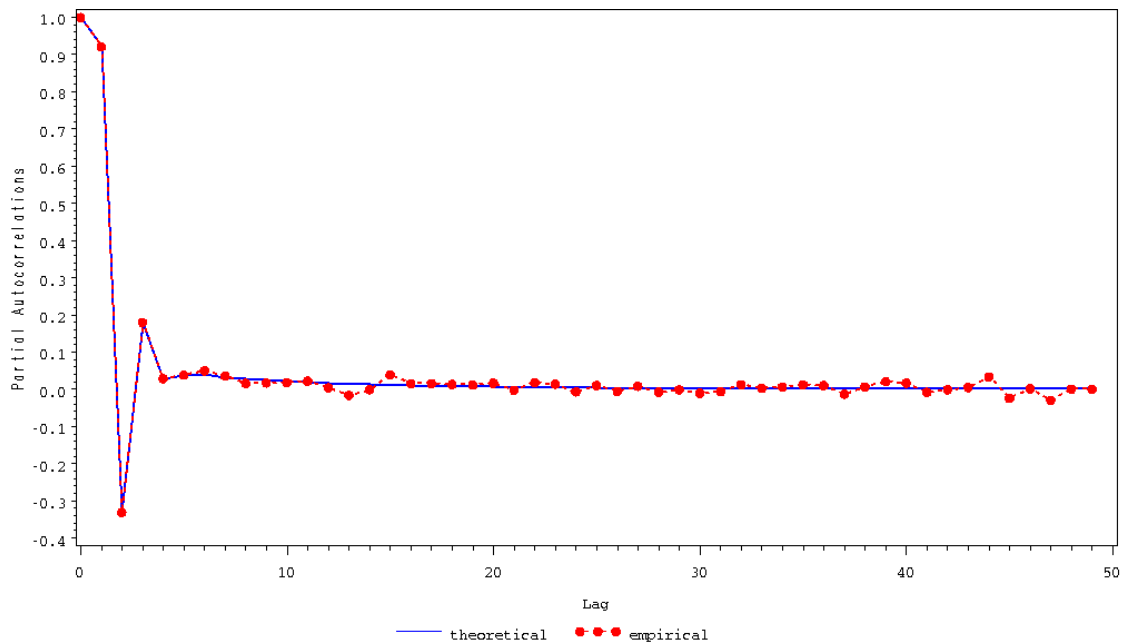
Plot 7.6.1: Theoretical autocorrelation function of the ARMA(3, 2)-model with the empirical autocorrelation function of the adjusted time series.



Plot 7.6.1b: Theoretical partial autocorrelation function of the ARMA(3, 2)-model with the empirical partial autocorrelation function of the adjusted time series.



Plot 7.6.1c: Theoretical autocorrelation function of the ARMA(2, 3)-model with the empirical autocorrelation function of the adjusted time series.



Plot 7.6.1d: Theoretical partial autocorrelation function of the ARMA(2,3)-model with the empirical partial autocorrelation function of the adjusted time series.

```

1  /* donauwoerth_dcheck1.sas */
2  TITLE1 'Theoretical process analysis';
3  TITLE2 'Donauwoerth Data';
4  /* Note that this program requires the file 'corrseasad' and 'seasad'
   ↪ generated by program donauwoerth_adjustment.sas */
5
6  /* Computation of theoretical autocorrelation */
7  PROC IML;
8    PHI={1 -1.61943 0.63143};
9    THETA={1 -0.34553 -0.34821 -0.10327};
10   LAG=1000;
11   CALL ARMACOV(COV, CROSS, CONVOL, PHI, THETA, LAG);
12   N=1:1000;
13   theocorr=COV/7.7685992485;
14   CREATE autocorr23;
15   APPEND;
16  QUIT;
17
18  PROC IML;
19   PHI={1 -1.85383 1.04408 -0.17899};
20   THETA={1 -0.58116 -0.23329};
21   LAG=1000;
22   CALL ARMACOV(COV, CROSS, CONVOL, PHI, THETA, LAG);
23   N=1:1000;
24   theocorr=COV/7.7608247812;
25   CREATE autocorr32;

```

```

26  APPEND;
27  QUIT;
28
29  %MACRO Simproc(p,q);
30
31  /* Graphical options */
32  SYMBOL1 V=NONE C=BLUE I=JOIN W=2;
33  SYMBOL2 V=DOT C=RED I=NONE H=0.3 W=1;
34  SYMBOL3 V=DOT C=RED I=JOIN L=2 H=1 W=2;
35  AXIS1 LABEL=(ANGLE=90 'Autocorrelations');
36  AXIS2 LABEL=('Lag') ORDER = (0 TO 500 BY 100);
37  AXIS3 LABEL=(ANGLE=90 'Partial Autocorrelations');
38  AXIS4 LABEL=('Lag') ORDER = (0 TO 50 BY 10);
39  LEGEND1 LABEL=NONE VALUE=('theoretical' 'empirical');
40
41  /* Comparison of theoretical and empirical autocorrelation */
42  DATA compare&p&q;
43  MERGE autocorr&p&q(KEEP=N theocorr) corrseasad(KEEP=corr);
44  PROC GLOT DATA=compare&p&q(OBS=500);
45  PLOT theocorr*N=1 corr*N=2 / OVERLAY LEGEND=LEGEND1 VAXIS=AXIS1
    ↪HAXIS=AXIS2;
46  RUN;
47
48  /* Computing of theoretical partial autocorrelation */
49  PROC TRANSPOSE DATA=autocorr&p&q OUT=transposed&p&q PREFIX=CORR;
50  DATA pacf&p&q(KEEP=PCORR0-PCORR100);
51  PCORR0=1;
52  SET transposed&p&q(WHERE=( _NAME_='THEOCORR' ));
53  ARRAY CORRS(100) CORR2-CORR101;
54  ARRAY PCORRS(100) PCORR1-PCORR100;
55  ARRAY P(100) P1-P100;
56  ARRAY w(100) w1-w100;
57  ARRAY A(100) A1-A100;
58  ARRAY B(100) B1-B100;
59  ARRAY u(100) u1-u100;
60  PCORRS(1)=CORRS(1);
61  P(1)=1-(CORRS(1)**2);
62  DO i=1 TO 100; B(i)=0; u(i)=0; END;
63  DO j=2 TO 100;
64  IF j > 2 THEN DO n=1 TO j-2; A(n)=B(n); END;
65  A(j-1)=PCORRS(j-1);
66  DO k=1 TO j-1; u(j)=u(j)-A(k)*CORRS(j-k); END;
67  w(j)=u(j)+CORRS(j);
68  PCORRS(j)=w(j)/P(j-1);
69  P(j)=P(j-1)*(1-PCORRS(j)**2);
70  DO m=1 TO j-1; B(m)=A(m)-PCORRS(j)*A(j-m); END;
71  END;
72  PROC TRANSPOSE DATA=pacf&p&q OUT=plotdata&p&q(KEEP=PACF1) PREFIX=PACF;
73
74  /* Comparison of theoretical and empirical partial autocorrelation */
75  DATA compareplot&p&q;
76  MERGE plotdata&p&q corrseasad(KEEP=partcorr LAG);

```

```

77 PROC GPLOT DATA=compareplot&p&q(OBS=50);
78   PLOT PACF1*LAG=1 partcorr*LAG=3 / OVERLAY VAXIS=AXIS3 HAXIS=AXIS4
    ↪LEGEND=LEGEND1;
79 RUN;
80
81 %MEND;
82 %Simproc(2,3);
83 %Simproc(3,2);
84 QUIT;

```

The procedure PROC IML generates an ARMA(p, q)-process, where the p autoregressive and q moving-average coefficients are specified by the entries in brackets after the PHI and THETA option, respectively. Note that the coefficients are taken from the corresponding characteristic polynomials with negative signs. The routine ARMACOV simulates the theoretical autocovariance function up to LAG 1000. Afterwards, the theoretical autocorrelation coefficients are attained by dividing the theoretical autocovariance coefficients by the variance. The results are written into the files autocorr23 and autocorr32, respectively. The next part of the program is a macro-step. The macro is named Simproc and has two parameters p and q . Macros enable the user to repeat a specific sequent of steps and procedures with respect to some dependent parameters, in our case the process orders p and q . A macro step is initiated by %MACRO and terminated by %MEND. Finally, by the concluding statements %Simproc(2,3) and %Simproc(3,2), the macro step is ex-

ecuted for the chosen pairs of process orders (2,3) and (3,2).

The macro step: First, the theoretical autocorrelation and the empirical autocorrelation function computed in Program 7.4.2, are summarized in one file by the option MERGE in order to derive a mutual, visual comparison. Based on the theoretical autocorrelation function, the theoretical partial autocorrelation is computed by means of the Levinson–Durbin recursion, introduced in Section 7.1. For this reason, the autocorrelation data are transposed by the procedure PROC TRANSPOSE, such that the recursion can be carried out by means of arrays, as executed in the sequent DATA step. The PREFIX option specifies the prefix of the transposed and consecutively numbered variables. Transposing back, the theoretical partial autocorrelation coefficients are merged with the coefficients of the empirical partial autocorrelations. PROC GPLOT together with the OVERLAY option produces a single plot of both autocorrelation functions.

In all cases there exists a great agreement between the theoretical and empirical parts. Hence, there is no reason to doubt the validity of any of the underlying models.

Examination of Residuals

We suppose that the adjusted time series $\tilde{y}_1, \dots, \tilde{y}_n$ was generated by an invertible ARMA(p, q)-model

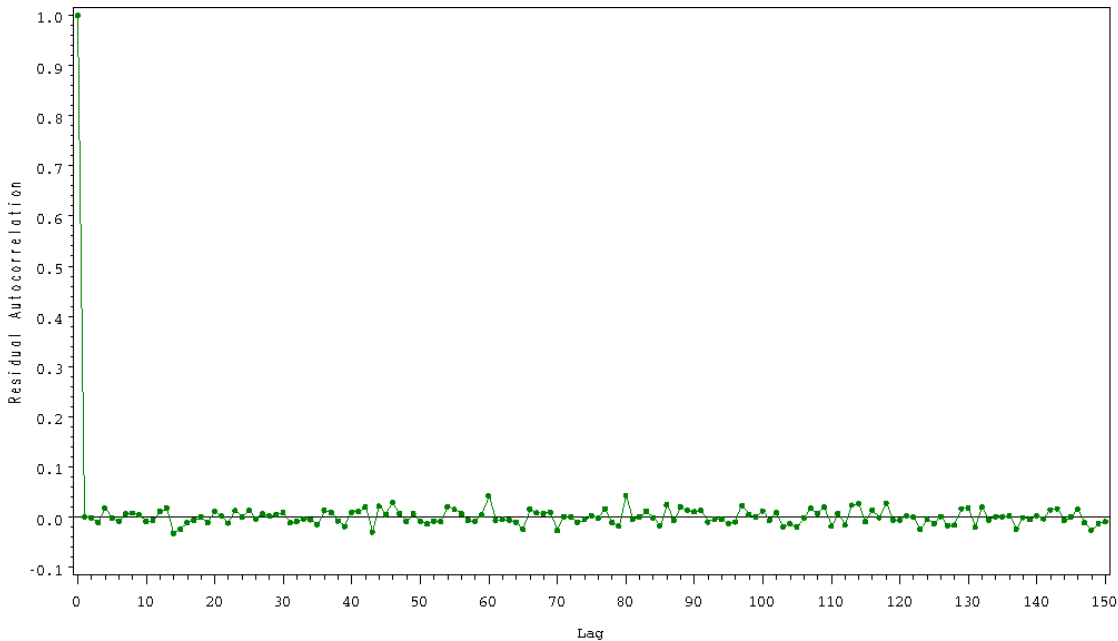
$$Y_t = a_1 Y_{t-1} + \dots + a_p Y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q} \quad (7.58)$$

satisfying the stationarity condition, $E(Y_t) = 0$ and $\text{Var}(Y_t) > 0$, entailing that $E(\varepsilon_t) = 0$. As seen in recursion (2.23) on page 105 the residuals can be estimated recursively by

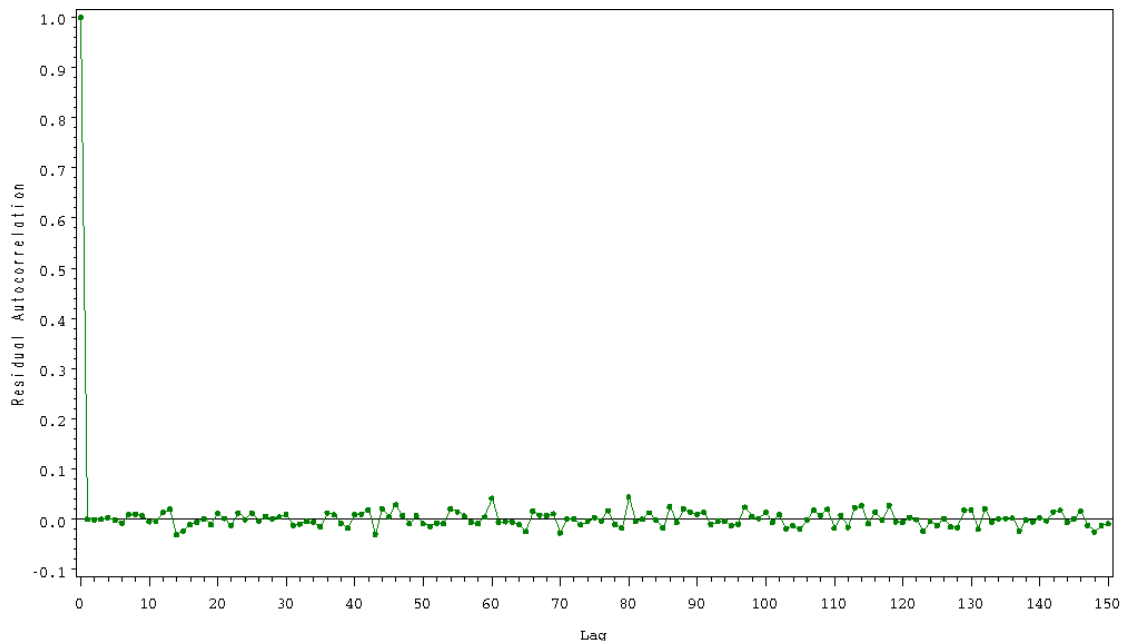
$$\hat{\varepsilon}_t = \tilde{y}_t - \hat{y}_t = \tilde{y}_t - a_1\tilde{y}_{t-1} - \cdots - a_p\tilde{y}_{t-p} - b_1\hat{\varepsilon}_{t-1} - \cdots - b_q\hat{\varepsilon}_{t-q} \approx \varepsilon_t$$

for $t = 1, \dots, n$, where $\hat{\varepsilon}_t$ and \tilde{y}_t are set equal to zero for $t \leq 0$. Under the model assumptions they show an approximate white noise behavior and therefore, their empirical autocorrelation function $\hat{r}_{\varepsilon}(k)$ (2.24) will be close to zero for $k > 0$ not too large. Note that above estimation of the empirical autocorrelation coefficients becomes less reliable as $k \rightarrow n$.

The following two figures show the empirical autocorrelation functions of the estimated residuals up to lag 100 based on the ARMA(3, 2)-model (7.56) and ARMA(2, 3)-model (7.57), respectively. Their pattern indicate uncorrelated residuals in both cases, confirming the adequacy of the chosen models.



Plot 7.6.2: Empirical autocorrelation function of estimated residuals based on ARMA(3, 2)-model.



Plot 7.6.2b: Empirical autocorrelation function of estimated residuals based on ARMA(2, 3)-model.

```

1  /* donauwoerth_dcheck2.sas */
2  TITLE1 'Residual Analysis';
3  TITLE2 'Donauwoerth Data';
4
5  /* Note that this program requires 'seasad' and 'donau' generated by
   ↪ the program donauwoerth_adjustment.sas */
6
7  /* Preparations */
8  DATA donau2;
9      MERGE donau seasad(KEEP=SA);
10
11  /* Test for white noise */
12  PROC ARIMA DATA=donau2;
13      IDENTIFY VAR=sa NLAG=60 NOPRINT;
14      ESTIMATE METHOD=CLS p=3 q=2 NOPRINT;
15      FORECAST LEAD=0 OUT=forecast1 NOPRINT;
16      ESTIMATE METHOD=CLS p=2 q=3 NOPRINT;
17      FORECAST LEAD=0 OUT=forecast2 NOPRINT;
18
19  /* Compute empirical autocorrelations of estimated residuals */
20  PROC ARIMA DATA=forecast1;
21      IDENTIFY VAR=residual NLAG=150 OUTCOV=cov1 NOPRINT;
22  RUN;
23
24  PROC ARIMA DATA=forecast2;
25      IDENTIFY VAR=residual NLAG=150 OUTCOV=cov2 NOPRINT;
26  RUN;

```

```

27
28 /* Graphical options */
29 SYMBOL1 V=DOT C=GREEN I=JOIN H=0.5;
30 AXIS1 LABEL=(ANGLE=90 'Residual Autocorrelation');
31 AXIS2 LABEL=('Lag') ORDER=(0 TO 150 BY 10);
32
33 /* Plot empirical autocorrelation function of estimated residuals */
34 PROC GPLOT DATA=cov1;
35   PLOT corr*LAG=1 / VREF=0 VAXIS=AXIS1 HAXIS=AXIS2;
36 RUN;
37
38 PROC GPLOT DATA=cov2;
39   PLOT corr*LAG=1 / VREF=0 VAXIS=AXIS1 HAXIS=AXIS2;
40 RUN; QUIT;

```

One-step forecasts of the adjusted series $\tilde{y}_1, \dots, \tilde{y}_n$ are computed by means of the FORECAST statement in the ARIMA procedure. LEAD=0 prevents future forecasts beyond the sample size 7300. The corresponding forecasts

are written together with the estimated residuals into the data file notated after the option OUT. Finally, the empirical autocorrelations of the estimated residuals are computed by PROC ARIMA and plotted by PROC GPLOT.

A more formal test is the *Portmanteau-test* of Box and Pierce (1970) presented in Section 2.3. SAS uses the Box–Ljung statistic with weighted empirical autocorrelations

$$\hat{r}_{\hat{\varepsilon}}(k)^* := \left(\frac{n+2}{n-k} \right)^{1/2} \hat{r}_{\hat{\varepsilon}}(k)$$

and computes the test statistic for different lags k .

Autocorrelation Check of Residuals									
To Lag	Chi- Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	3.61	1	0.0573	0.001	-0.001	-0.010	0.018	-0.001	-0.008
12	6.44	7	0.4898	0.007	0.008	0.005	-0.008	-0.006	0.012
18	21.77	13	0.0590	0.018	-0.032	-0.024	-0.011	-0.006	0.001
24	25.98	19	0.1307	-0.011	0.012	0.002	-0.012	0.013	0.000
30	28.66	25	0.2785	0.014	-0.003	0.006	0.002	0.005	0.010
36	33.22	31	0.3597	-0.011	-0.008	-0.004	-0.005	-0.015	0.013
42	41.36	37	0.2861	0.009	-0.007	-0.018	0.010	0.011	0.020
48	58.93	43	0.0535	-0.030	0.022	0.006	0.030	0.007	-0.008
54	65.37	49	0.0589	0.007	-0.008	-0.014	-0.007	-0.009	0.021

60	81.64	55	0.0114	0.015	0.007	-0.007	-0.008	0.005	0.042
----	-------	----	--------	-------	-------	--------	--------	-------	-------

Listing 7.6.3: Box–Ljung test of ARMA(3,2)-model.

Autocorrelation Check of Residuals									
To Lag	Chi- Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	0.58	1	0.4478	-0.000	-0.001	-0.000	0.003	-0.002	-0.008
12	3.91	7	0.7901	0.009	0.010	0.007	-0.005	-0.004	0.013
18	18.68	13	0.1333	0.020	-0.031	-0.023	-0.011	-0.006	0.001
24	22.91	19	0.2411	-0.011	0.012	0.002	-0.012	0.013	-0.001
30	25.12	25	0.4559	0.012	-0.004	0.005	0.001	0.004	0.009
36	29.99	31	0.5177	-0.012	-0.009	-0.004	-0.006	-0.015	0.012
42	37.38	37	0.4517	0.009	-0.008	-0.018	0.009	0.010	0.019
48	54.56	43	0.1112	-0.030	0.021	0.005	0.029	0.008	-0.009
54	61.23	49	0.1128	0.007	-0.009	-0.014	-0.007	-0.009	0.021
60	76.95	55	0.0270	0.015	0.006	-0.006	-0.009	0.004	0.042

Listing 7.6.3b: Box–Ljung test of ARMA(2,3)-model.

```

1  /* donauwoerth_dcheck3.sas */
2  TITLE1 'Residual Analysis';
3  TITLE2 'Donauwoerth Data';
4
5  /* Note that this program requires 'seasad' and 'donau' generated by
   ↪ the program donauwoerth_adjustment.sas */
6
7  /* Test preparation */
8  DATA donau2;
9      MERGE donau seasad(KEEP=SA);
10
11 /* Test for white noise */
12 PROC ARIMA DATA=donau2;
13     IDENTIFY VAR=sa NLAG=60 NOPRINT;
14     ESTIMATE METHOD=CLS p=3 q=2 WHITENOISE=IGNOREMISS;
15     FORECAST LEAD=0 OUT=out32(KEEP=residual RENAME=residual=res32);
16     ESTIMATE METHOD=CLS p=2 q=3 WHITENOISE=IGNOREMISS;
17     FORECAST LEAD=0 OUT=out23(KEEP=residual RENAME=residual=res23);
18 RUN; QUIT;

```

The Portmanteau-test of Box–Ljung is initiated by the WHITENOISE=IGNOREMISS option in the ESTIMATE statement of PROC ARIMA. The residuals are also written in the files out32 and out23 for further use.

For both models, the Listings 7.6.3 and 7.6.3 indicate sufficiently great p -values for the Box–Ljung test statistic up to lag 54, letting us presume that the residuals are representations of a white noise process.

Furthermore, we might deduce that the ARMA(2,3)-model is the more appropriate model, since the p -values are remarkably greater than those of the ARMA(3,2)-model.

Overfitting

The third diagnostic class is the method of overfitting. To verify the fitting of an ARMA(p, q)-model, slightly more comprehensive models with additional parameters are estimated, usually an ARMA($p, q + 1$)- and ARMA($p + 1, q$)-model. It is then expected that these new additional parameters will be close to zero, if the initial ARMA(p, q)-model is adequate.

The original ARMA(p, q)-model should be considered hazardous or critical, if one of the following aspects occur in the more comprehensive models:

1. The 'old' parameters, which already have been estimated in the ARMA(p, q)-model, are instable, i.e., the parameters differ extremely.
2. The 'new' parameters differ significantly from zero.
3. The variance of the residuals is remarkably smaller than the corresponding one in the ARMA(p, q)-model.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	-3.82315	7.23684	-0.53	0.5973	0
MA1,1	0.29901	0.14414	2.07	0.0381	1
MA1,2	0.37478	0.07795	4.81	<.0001	2
MA1,3	0.12273	0.05845	2.10	0.0358	3
AR1,1	1.57265	0.14479	10.86	<.0001	1
AR1,2	-0.54576	0.25535	-2.14	0.0326	2
AR1,3	-0.03884	0.11312	-0.34	0.7313	3

Listing 7.6.4: Estimated coefficients of ARMA(3,3)-model.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	-3.93060	7.36255	-0.53	0.5935	0
MA1,1	0.79694	0.11780	6.77	<.0001	1
MA1,2	0.08912	0.09289	0.96	0.3374	2
AR1,1	2.07048	0.11761	17.60	<.0001	1
AR1,2	-1.46595	0.23778	-6.17	<.0001	2
AR1,3	0.47350	0.16658	2.84	0.0045	3
AR1,4	-0.08461	0.04472	-1.89	0.0585	4

Listing 7.6.4b: Estimated coefficients of ARMA(4,2)-model.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
MU	-3.83016	7.24197	-0.53	0.5969	0
MA1,1	0.36178	0.05170	7.00	<.0001	1
MA1,2	0.35398	0.02026	17.47	<.0001	2
MA1,3	0.10115	0.01585	6.38	<.0001	3
MA1,4	-0.0069054	0.01850	-0.37	0.7089	4
AR1,1	1.63540	0.05031	32.51	<.0001	1
AR1,2	-0.64655	0.04735	-13.65	<.0001	2

Listing 7.6.4c: Estimated coefficients of ARMA(2,4)-model.

The MEANS Procedure					
Variable	N	Mean	Std Dev	Minimum	Maximum
res33	7300	0.2153979	37.1219190	-254.0443970	293.3993079
res42	7300	0.2172515	37.1232437	-253.8610208	294.0049942
res24	7300	0.2155696	37.1218896	-254.0365135	293.4197058
res32	7300	0.2090862	37.1311356	-254.2828124	292.9240859
res23	7300	0.2143327	37.1222160	-254.0171144	293.3822232

Listing 7.6.4d: Statistics of estimated residuals of considered overfitted models.

```

1  /* donauwoerth_overfitting.sas */
2  TITLE1 'Overfitting';
3  TITLE2 'Donauwoerth Data';
4  /* Note that this program requires the file 'seasad' generated by
   ↪ program donauwoerth_adjustment.sas as well as 'out32' and 'out23'
   ↪ generated by donauwoerth_dcheck3.sas */
5
6  /* Estimate coefficients for chosen orders p and q */
7  PROC ARIMA DATA=seasad;

```

```

8  IDENTIFY VAR=sa NLAG=100 NOPRINT;
9  ESTIMATE METHOD=CLS p=3 q=3;
10 FORECAST LEAD=0 OUT=out33(KEEP=residual RENAME=residual=res33);
11 RUN;
12
13 PROC ARIMA DATA=seasad;
14   IDENTIFY VAR=sa NLAG=100 NOPRINT;
15   ESTIMATE METHOD=CLS p=4 q=2;
16   FORECAST LEAD=0 OUT=out42(KEEP=residual RENAME=residual=res42);
17 RUN;
18
19 PROC ARIMA DATA=seasad;
20   IDENTIFY VAR=sa NLAG=100 NOPRINT;
21   ESTIMATE METHOD=CLS p=2 q=4;
22   FORECAST LEAD=0 OUT=out24(KEEP=residual RENAME=residual=res24);
23 RUN;
24
25 /* Merge estimated residuals */
26 DATA residuals;
27   MERGE out33 out42 out24 out32 out23;
28
29 /* Compute standard deviations of estimated residuals */
30 PROC MEANS DATA=residuals;
31   VAR res33 res42 res24 res32 res23;
32 RUN; QUIT;

```

The coefficients of the overfitted models are again computed by the conditional least squares approach (CLS) in the framework of the procedure PROC ARIMA. Finally, the estimated residuals of all considered models are merged into one data file and the procedure PROC MEANS computes their standard deviations.

SAS computes the model coefficients for the chosen orders (3, 3), (2, 4) and (4, 2) by means of the conditional least squares method and provides the ARMA(3, 3)-model

$$\begin{aligned}
 Y_t - 1.57265Y_{t-1} + 0.54576Y_{t-2} + 0.03884Y_{t-3} \\
 = \varepsilon_t - 0.29901\varepsilon_{t-1} - 0.37478\varepsilon_{t-2} - 0.12273\varepsilon_{t-3},
 \end{aligned}$$

the ARMA(4, 2)-model

$$\begin{aligned}
 Y_t - 2.07048Y_{t-1} + 1.46595Y_{t-2} - 0.47350Y_{t-3} + 0.08461Y_{t-4} \\
 = \varepsilon_t - 0.79694\varepsilon_{t-1} - 0.08912\varepsilon_{t-2}
 \end{aligned}$$

as well as the ARMA(2, 4)-model

$$\begin{aligned}
 Y_t - 1.63540Y_{t-1} + 0.64655Y_{t-2} \\
 = \varepsilon_t - 0.36178\varepsilon_{t-1} - 0.35398\varepsilon_{t-2} - 0.10115\varepsilon_{t-3} + 0.0069054\varepsilon_{t-4}.
 \end{aligned}$$

This overfitting method indicates that, once more, we have reason to doubt the suitability of the adjusted ARMA(3, 2)-model

$$\begin{aligned} Y_t - 1.85383Y_{t-1} + 1.04408Y_{t-2} - 0.17899Y_{t-3} \\ = \varepsilon_t - 0.58116\varepsilon_{t-1} - 0.23329\varepsilon_{t-2}, \end{aligned}$$

since the 'old' parameters in the ARMA(4, 2)- and ARMA(3, 3)-model differ extremely from those inherent in the above ARMA(3, 2)-model and the new added parameter -0.12273 in the ARMA(3, 3)-model is not close enough to zero, as can be seen by the corresponding p -value. Whereas the adjusted ARMA(2, 3)-model

$$\begin{aligned} Y_t &= -1.61943Y_{t-1} + 0.63143Y_{t-2} \\ &= \varepsilon_t - 0.34553\varepsilon_{t-1} - 0.34821\varepsilon_{t-2} - 0.10327\varepsilon_{t-3} \end{aligned} \quad (7.59)$$

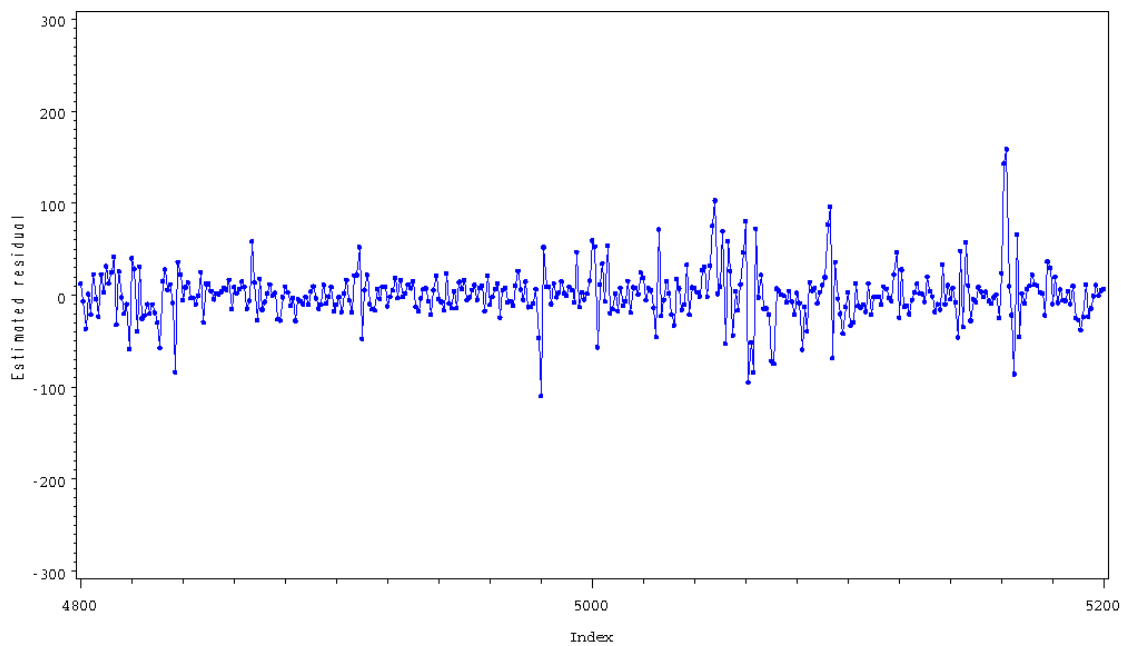
satisfies these conditions. Finally, Listing 7.6.4d indicates the almost equal standard deviations of 37.1 of all estimated residuals. Accordingly, we assume almost equal residual variances. Therefore none of the considered models has remarkably smaller variance in the residuals.

7.7 Forecasting

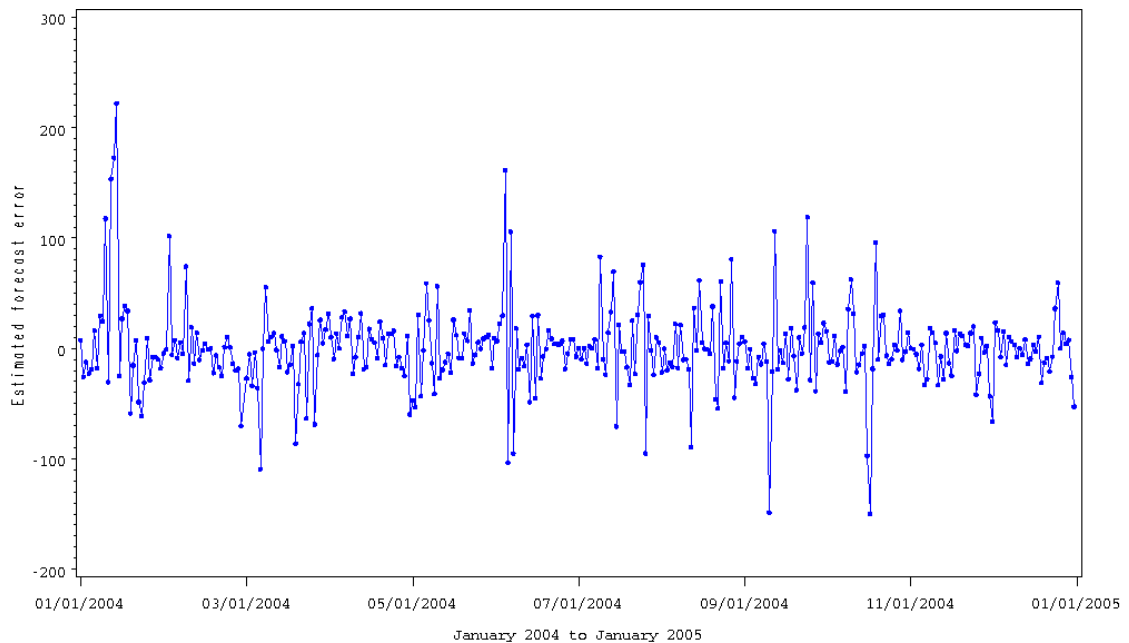
To justify forecasts based on the apparently adequate ARMA(2, 3)-model (7.59), we have to verify its forecasting ability. It can be achieved by either *ex-ante* or *ex-post* best one-step forecasts of the sample Y_1, \dots, Y_n . Latter method already has been executed in the residual examination step. We have shown that the estimated forecast errors $\hat{\varepsilon}_t := \tilde{y}_t - \hat{y}_t, t = 1, \dots, n$ can be regarded as a realization of a white noise process. The ex-ante forecasts are based on the first $n - m$ observations $\tilde{y}_1, \dots, \tilde{y}_{n-m}$, i.e., we adjust an ARMA(2, 3)-model to the reduced time series $\tilde{y}_1, \dots, \tilde{y}_{n-m}$. Afterwards, best one-step forecasts \hat{y}_t of Y_t are estimated for $t = 1, \dots, n$, based on the parameters of this new ARMA(2, 3)-model.

Now, if the ARMA(2, 3)-model (7.59) is actually adequate, we will again expect that the estimated forecast errors $\hat{\varepsilon}_t = \tilde{y}_t - \hat{y}_t, t = n - m + 1, \dots, n$ behave like realizations of a white noise process, where

the standard deviations of the estimated residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{n-m}$ and estimated forecast errors $\hat{\varepsilon}_{n-m+1}, \dots, \hat{\varepsilon}_n$ should be close to each other.



Plot 7.7.1: Extract of estimated residuals.



Plot 7.7.1b: Estimated forecast errors of ARMA(2,3)-model based on the reduced sample.

The SPECTRA Procedure

Test for White Noise for Variable RESIDUAL

M = 3467
 Max(P(*)) 44294.13
 Sum(P(*)) 9531509

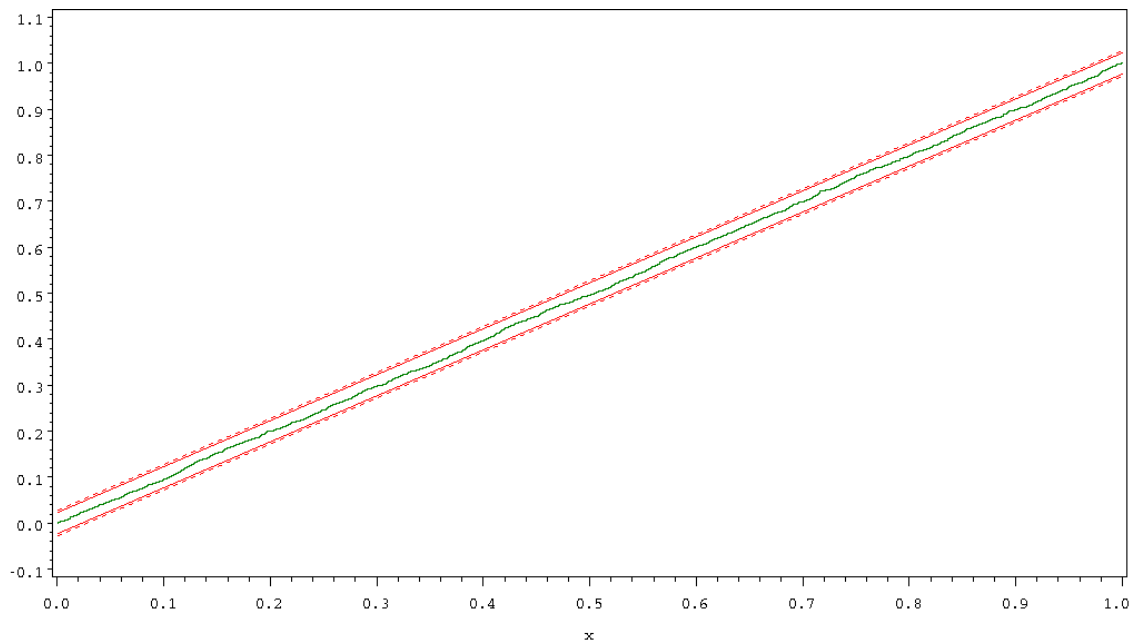
Fisher's Kappa: $M \cdot \text{MAX}(P(*)) / \text{SUM}(P(*))$

Kappa 16.11159

Bartlett's Kolmogorov-Smirnov Statistic:
 Maximum absolute difference of the standardized
 partial sums of the periodogram and the CDF of a
 uniform(0,1) random variable.

Test Statistic 0.006689
 Approximate P-Value 0.9978

Listing 7.7.1c: Test for white noise of estimated residuals.



Plot 7.7.1d: Visual representation of Bartlett-Kolmogorov-Smirnov test for white noise of estimated residuals.

The SPECTRA Procedure

Test for White Noise for Variable RESIDUAL

M = 182
 Max(P(*)) 18698.19
 Sum(P(*)) 528566

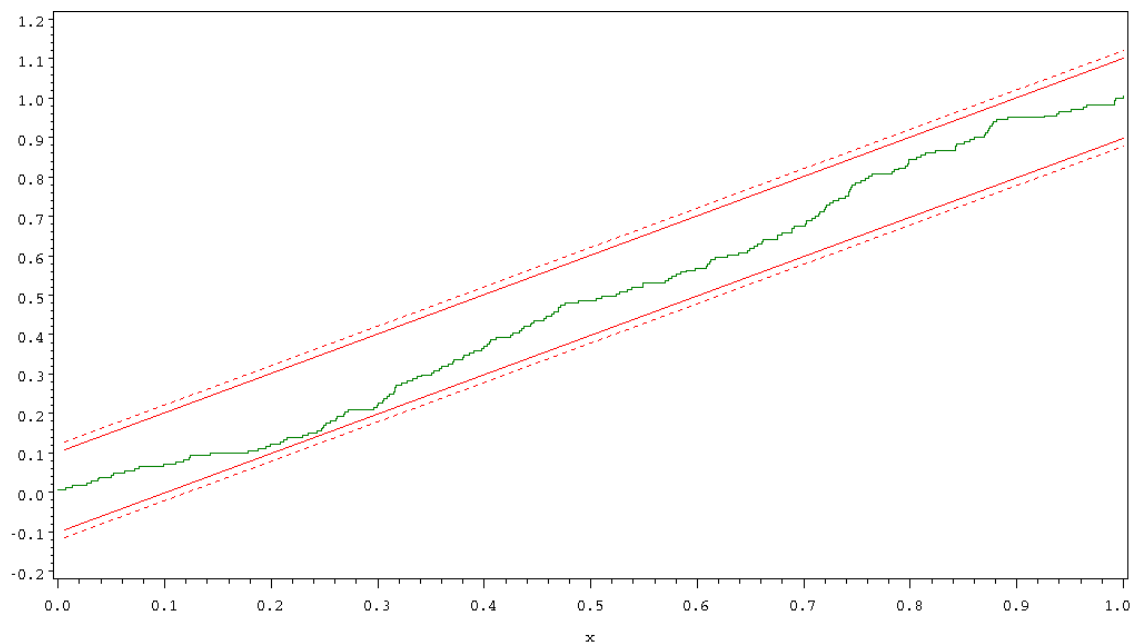
Fisher's Kappa: $M \cdot \text{MAX}(P(*)) / \text{SUM}(P(*))$

Kappa 6.438309

Bartlett's Kolmogorov-Smirnov Statistic:
 Maximum absolute difference of the standardized
 partial sums of the periodogram and the CDF of a
 uniform(0,1) random variable.

Test Statistic 0.09159
 Approximate P-Value 0.0944

Listing 7.7.1e: Test for white noise of estimated forecast errors.



Plot 7.7.1f: Visual representation of Bartlett-Kolmogorov-Smirnov test for white noise of estimated forecast errors.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
RESIDUAL	6935	-0.0038082	37.0756613	-254.3711804	293.3417594
forecasterror	365	-0.0658013	38.1064859	-149.4454536	222.4778315

Listing 7.7.1g: Statistics of estimated residuals and estimated forecast errors.

The SPECTRA Procedure

Test for White Noise for Variable RESIDUAL

M-1 2584
 Max(P(*)) 24742.39
 Sum(P(*)) 6268989

Fisher's Kappa: (M-1)*Max(P(*))/Sum(P(*))

Kappa 10.19851

Bartlett's Kolmogorov-Smirnov Statistic:
 Maximum absolute difference of the standardized

partial sums of the periodogram and the CDF of a
uniform(0,1) random variable.

Test Statistic	0.021987
Approximate P-Value	0.1644

Listing 7.7.1h: Test for white noise of the first 5170 estimated residuals.

The SPECTRA Procedure

Test for White Noise for Variable RESIDUAL

M-1	2599
Max(P(*))	39669.63
Sum(P(*))	6277826

Fisher's Kappa: $(M-1) * \text{Max}(P(*)) / \text{Sum}(P(*))$

Kappa	16.4231
-------	---------

Bartlett's Kolmogorov-Smirnov Statistic:
Maximum absolute difference of the standardized
partial sums of the periodogram and the CDF of a
uniform(0,1) random variable.

Test Statistic	0.022286
Approximate P-Value	0.1512

Listing 7.7.1i: Test for white noise of the first 5200 estimated residuals.

```

1  /* donauwoerth_forecasting.sas */
2  TITLE1 'Forecasting';
3  TITLE2 'Donauwoerth Data';
4
5  /* Note that this program requires the file 'seasad' generated by
   ↪ program donauwoerth_adjustment.sas */
6
7  /* Cut off the last 365 values */
8  DATA reduced;
9      SET seasad(OBS=6935);
10     N=_N_;
11
12  /* Estimate model coefficients from reduced sample */
13  PROC ARIMA DATA=reduced;
14      IDENTIFY VAR=sa NLAG=300 NOPRINT;
15      ESTIMATE METHOD=CLS p=2 q=3 OUTMODEL=model NOPRINT;
16  RUN;
17
18  /* One-step-forecasts of previous estimated model */
19  PROC ARIMA DATA=seasad;
```

```

20 IDENTIFY VAR=sa NLAG=300 NOPRINT;
21 ESTIMATE METHOD=CLS p=2 q=3 MU=-0.04928 AR=1.60880 -0.62077 MA
    ↪=0.33764 0.35157 0.10282 NOEST NOPRINT;
22 FORECAST LEAD=0 OUT=forecast NOPRINT;
23 RUN;
24
25 DATA forecast;
26 SET forecast;
27 lag=_N_;
28
29 DATA residual1(KEEP=residual lag);
30 SET forecast(OBS=6935);
31
32 DATA residual2(KEEP=residual date);
33 MERGE forecast(FIRSTOBS=6936) donau(FIRSTOBS=6936 KEEP=date);
34
35 DATA residual3(KEEP=residual lag);
36 SET forecast(OBS=5170);
37
38 DATA residual4(KEEP=residual lag);
39 SET forecast(OBS=5200);
40
41 /* Graphical options */
42 SYMBOL1 V=DOT C=GREEN I=JOIN H=0.5;
43 SYMBOL2 V=DOT I=JOIN H=0.4 C=BLUE W=1 L=1;
44 AXIS1 LABEL=('Index')
45 ORDER=(4800 5000 5200);
46 AXIS2 LABEL=(ANGLE=90 'Estimated residual');
47 AXIS3 LABEL=('January 2004 to January 2005')
48 ORDER=('01JAN04'd '01MAR04'd '01MAY04'd '01JUL04'd '01SEP04'd
    ↪'01NOV04'd '01JAN05'd);
49 AXIS4 LABEL=(ANGLE=90 'Estimated forecast error');
50
51
52 /* Plot estimated residuals */
53 PROC GPLOT DATA=residual1(FIRSTOBS=4800);
54 PLOT residual*lag=2 /HAXIS=AXIS1 VAXIS=AXIS2;
55 RUN;
56
57 /* Plot estimated forecast errors of best one-step forecasts */
58 PROC GPLOT DATA=residual2;
59 PLOT residual*date=2 /HAXIS=AXIS3 VAXIS=AXIS4;
60 RUN;
61
62 /* Testing for white noise */
63 %MACRO wn(k);
64 PROC SPECTRA DATA=residual&k P WHITETEST OUT=periodo&k;
65 VAR residual;
66 RUN;
67
68 /* Calculate the sum of the periodogram */
69 PROC MEANS DATA=periodo&k(FIRSTOBS=2) NOPRINT;

```

```

70  VAR P_01;
71  OUTPUT OUT=psum&k SUM=psum;
72  RUN;
73
74  /* Compute empirical distribution function of cumulated periodogram
    ↪and its confidence bands */
75  DATA conf&k;
76  SET periodo&k (FIRSTOBS=2);
77  IF _N_=1 THEN SET psum&k;
78  RETAIN s 0;
79  s=s+P_01/psum;
80  fm=_N/_FREQ_-1;
81  yu_01=fm+1.63/SQRT(_FREQ_-1);
82  yl_01=fm-1.63/SQRT(_FREQ_-1);
83  yu_05=fm+1.36/SQRT(_FREQ_-1);
84  yl_05=fm-1.36/SQRT(_FREQ_-1);
85
86  /* Graphical options */
87  SYMBOL3 V=NONE I=STEPJ C=GREEN;
88  SYMBOL4 V=NONE I=JOIN C=RED L=2;
89  SYMBOL5 V=NONE I=JOIN C=RED L=1;
90  AXIS5 LABEL=('x') ORDER=(.0 TO 1.0 BY .1);
91  AXIS6 LABEL=NONE;
92
93  /* Plot empirical distribution function of cumulated periodogram with
    ↪its confidence bands */
94  PROC GPLOT DATA=conf&k;
95  PLOT fm*s=3 yu_01*fm=4 yl_01*fm=4 yu_05*fm=5 yl_05*fm=5 / OVERLAY
    ↪HAXIS=AXIS5 VAXIS=AXIS6;
96  RUN;
97  %MEND;
98
99  %wn(1);
100 %wn(2);
101
102 /* Merge estimated residuals and estimated forecast errors */
103 DATA residuals;
104 MERGE residual1(KEEP=residual) residual2(KEEP=residual RENAME=(
    ↪residual=forecasterror));
105
106 /* Compute standard deviations */
107 PROC MEANS DATA=residuals;
108 VAR residual forecasterror;
109 RUN;
110
111 %wn(3);
112 %wn(4);
113
114 QUIT;

```

In the first `DATA` step, the adjusted discharge values pertaining to the discharge values measured during the last year are removed from the sample. In the framework of the procedure `PROC ARIMA`, an $\text{ARMA}(2,3)$ -model is adjusted to the reduced sample. The corresponding estimated model coefficients and further model details are written into the file 'model', by the option `OUTMODEL`. Best one-step forecasts of the $\text{ARMA}(2,3)$ -model are then computed by `PROC ARIMA` again, where the model orders, model coefficients and model mean are specified in the `ESTIMATE` statement, where the option `NOEST` suppresses a new model estimation. The following steps deal with the examination

of the residuals and forecast errors. The residuals coincide with the one-step forecasts errors of the first 6935 observations, whereas the forecast errors are given by one-step forecasts errors of the last 365 observations. Visual representation is created by `PROC GPLOT`.

The macro-step then executes Fisher's test for hidden periodicities as well as the Bartlett–Kolmogorov–Smirnov test for both residuals and forecast errors within the procedure `PROC SPECTRA`. After renaming, they are written into a single file. `PROC MEANS` computes their standard deviations for a direct comparison.

Afterwards tests for white noise are computed for other numbers of estimated residuals by the macro.

The ex-ante forecasts are computed from the dataset without the last 365 observations. Plot 7.7.1b of the estimated forecast errors $\hat{\varepsilon}_t = \tilde{y}_t - \hat{y}_t, t = n - m + 1, \dots, n$ indicates an accidental pattern. Applying Fisher's test as well as Bartlett–Kolmogorov–Smirnov's test for white noise, as introduced in Section 6.1, the p -value of 0.0944 and Fisher's κ -statistic of 6.438309 do not reject the hypothesis that the estimated forecast errors are generated from a white noise process $(\varepsilon_t)_{t \in \mathbb{Z}}$, where the ε_t are independent and identically normal distributed (Listing 7.7.1e). Whereas, testing the estimated residuals for white noise behavior yields a conflicting result. Bartlett–Kolmogorov–Smirnov's test clearly doesn't reject the hypothesis that the estimated residuals are generated from a white noise process $(\varepsilon_t)_{t \in \mathbb{Z}}$, where the ε_t are independent and identically normal distributed, due to the great p -value of 0.9978 (Listing 7.7.1c). On the other hand, the conservative test of Fisher produces an irritating, high κ -statistic of 16.11159, which would lead to a rejection of the hypothesis. Further examinations with Fisher's test lead to the κ -statistic of 10.19851, when testing the first 5170 estimated residuals, and it leads to the κ -statistic of 16.4231, when testing the first 5200 estimated residuals (Listing 7.7.1h and 7.7.1i). Applying the result of Exercise 6.12 $P\{\kappa_m > x\} \approx 1 - \exp(-m e^{-x})$, we obtain the approximate p -values of 0.09 and < 0.001 , respectively. That is, Fisher's test doesn't reject the hypothesis for the first 5170 estimated residuals but it clearly re-

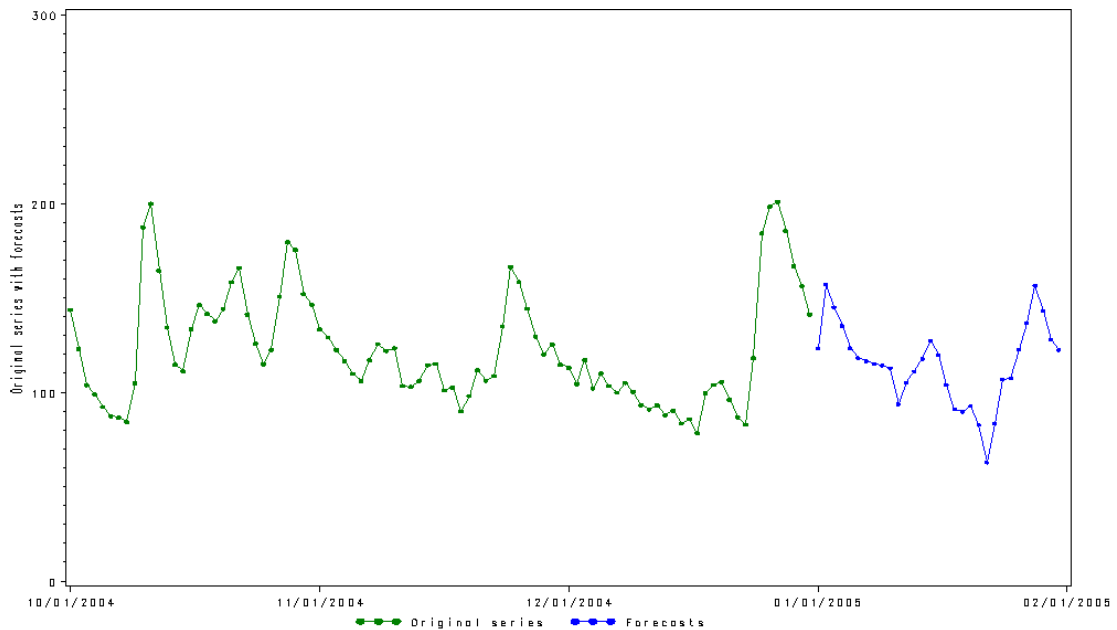
jects the hypothesis for the first 5200 ones. This would imply that the additional 30 values cause such a great periodical influence such that a white noise behavior is strictly rejected. In view of the small number of 30, relatively to 5170, Plot 7.7.1 and the great p -value 0.9978 of Bartlett-Kolmogorov-Smirnov's test we can't trust Fisher's κ -statistic. Recall, as well, that the Portmanteau-test of Box-Ljung has not rejected the hypothesis that the 7300 estimated residuals result from a white noise, carried out in Section 7.6. Finally, Listing 7.7.1g displays that the standard deviations 37.1 and 38.1 of the estimated residuals and estimated forecast errors, respectively, are close to each other.

Hence, the validity of the ARMA(2, 3)-model (7.59) has been established, justifying us to predict future values using best h -step forecasts \hat{Y}_{n+h} based on this model. Since the model is invertible, it can be almost surely rewritten as AR(∞)-process $Y_t = \sum_{u>0} c_u Y_{t-u} + \varepsilon_t$. Consequently, the best h -step forecast \hat{Y}_{n+h} of Y_{n+h} can be estimated by $\hat{y}_{n+h} = \sum_{u=1}^{h-1} c_u \hat{y}_{n+h-u} + \sum_{v=h}^{t+h-1} c_v \tilde{y}_{t+h-v}$, cf. Theorem 2.3.4, where the estimated forecasts \hat{y}_{n+i} , $i = 1, \dots, h$ are computed recursively.

An estimated best h -step forecast $\hat{y}_{n+h}^{(o)}$ of the original Donauwoerth time series y_1, \dots, y_n is then obviously given by

$$\hat{y}_{n+h}^{(o)} := \hat{y}_{n+h} + \mu + \hat{S}_{n+h}^{(365)} + \hat{S}_{n+h}^{(2433)}, \quad (7.60)$$

where μ denotes the arithmetic mean of y_1, \dots, y_n and where $\hat{S}_{n+h}^{(365)}$ and $\hat{S}_{n+h}^{(2433)}$ are estimations of the seasonal nonrandom components at lag $n+h$ pertaining to the cycles with periods of 365 and 2433 days, respectively. These estimations already have been computed by Program 7.4.2. Note that we can only make reasonable forecasts for the next few days, as \hat{y}_{n+h} approaches its expectation zero for increasing h .



Plot 7.7.2: Estimated best h -step forecasts of Donauwoerth Data for the next month by means of identified ARMA(2, 3)-model.

```

1  /* donauwoerth_final.sas */
2  TITLE1 ;
3  TITLE2 ;
4
5  /* Note that this program requires the file 'seasad' and 'seasad1'
   ↪ generated by program donauwoerth_adjustment.sas */
6
7  /* Computations of forecasts for next 31 days */
8  PROC ARIMA DATA=seasad;
9      IDENTIFY VAR=sa NLAG=300 NOPRINT;
10     ESTIMATE METHOD=CLS p=2 q=3 OUTMODEL=model NOPRINT;
11     FORECAST LEAD=31 OUT=forecast;
12 RUN;
13
14 /* Plot preparations */
15 DATA forecast(DROP=residual sa);
16     SET forecast(FIRSTOBS=7301);
17     N=_N_;
18     date=MDY(1,N,2005);
19     FORMAT date ddmmyy10.;
20
21 DATA seascomp(KEEP=SC RENAME=(SC=SC1));
22     SET seasad(FIRSTOBS=1 OBS=31);
23
24 /* Adding the seasonal components and mean */
25 DATA seascomp;
26     MERGE forecast seascomp seasad1(OBS=31 KEEP=SC RENAME=(SC=SC2));
27     forecast=forecast+SC1+SC2+201.6;

```

```

28
29
30 DATA plotforecast;
31     SET donau seascomp;
32
33 /* Graphical display */
34 SYMBOL1 V=DOT C=GREEN I=JOIN H=0.5;
35 SYMBOL2 V=DOT I=JOIN H=0.4 C=BLUE W=1 L=1;
36 AXIS1 LABEL=NONE ORDER=('01OCT04'd '01NOV04'd '01DEC04'd '01JAN05'd
    ↪ '01FEB05'd);
37 AXIS2 LABEL=(ANGLE=90 'Original series with forecasts') ORDER=(0 to
    ↪ 300 by 100);
38 LEGEND1 LABEL=NONE VALUE=('Original series' 'Forecasts' 'Lower 95-
    ↪ percent confidence limit' 'Upper 95-percent confidence limit');
39
40 PROC GPLOT DATA=plotforecast (FIRSTOBS=7209);
41     PLOT discharge*date=1 forecast*date=2 / OVERLAY HAXIS=AXIS1 VAXIS=
    ↪ AXIS2 LEGEND=LEGEND1;
42 RUN; QUIT;

```

The best h -step forecasts for the next month are computed within PROC ARIMA by the FORECAST statement and the option LEAD. Adding the corresponding seasonal components and the arithmetic mean (see (7.60)), forecasts of the original time series are attained. The procedure PROC GPLOT then creates a graphical visualization of the best h -step forecasts.

Result and Conclusion

The case study indicates that an ARMA(2, 3)-model can be adjusted to the prepared data of discharge values taken by the gaging station nearby the Donau at Donauwoerth for appropriately explaining the data. For this reason, we assume that a new set of data containing further or entirely future discharge values taken during the years after 2004 can be modeled by means of an ARMA(2, 3)-process as well, entailing reasonable forecasts for the next few days.

Exercises

7.1. Show the following generalization of Lemma 7.3.1: Consider two sequences $(Y_n)_{n \in \mathbb{N}}$ and $(X_n)_{n \in \mathbb{N}}$ of real valued random k -vectors, all defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and a sequence

$(h_n)_{n \in \mathbb{N}}$ of positive real valued numbers, such that $\mathbf{Y}_n = \mathcal{O}_P(h_n)$ and $\mathbf{X}_n = \mathcal{O}_P(1)$. Then, $\mathbf{Y}_n \mathbf{X}_n = \mathcal{O}_P(h_n)$.

7.2. Prove the derivation of Equation 7.42.

7.3. (IQ Data) Apply the Box–Jenkins program to the IQ Data. This dataset contains a monthly index of quality ranging from 0 to 100.

Bibliography

- Akaike, H. (1977). On entropy maximization principle. *Applications of Statistics, Amsterdam*, pages 27–41.
- Akaike, H. (1978). On the likelihood of a time series model. *The Statistician*, 27:271–235.
- Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 2nd edition.
- Andrews, D. W. (1991). Heteroscedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.
- Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley, New York, 2nd edition.
- Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*. John Wiley, New York, 2nd edition.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327.
- Box, G. E. and Cox, D. (1964). An analysis of transformations (with discussion). *J.R. Stat. Soc. B.*, 26:211–252.
- Box, G. E., Jenkins, G. M., and Reinsel, G. (1994). *Times Series Analysis: Forecasting and Control*. Prentice Hall, 3rd edition.
- Box, G. E. and Pierce, D. A. (1970). Distribution of residual autocorrelation in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526.

- Box, G. E. and Tiao, G. (1977). A canonical analysis of multiple time series. *Biometrika*, 64:355–365.
- Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, New York, 2nd edition.
- Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer, New York.
- Choi, B. (1992). *ARMA Model Identification*. Springer, New York.
- Cleveland, W. and Tiao, G. (1976). Decomposition of seasonal time series: a model for the census x-11 program. *Journal of the American Statistical Association*, 71(355):581–587.
- Conway, J. B. (1978). *Functions of One Complex Variable*. Springer, New York, 2nd edition.
- Enders, W. (2004). *Applied econometric time series*. Wiley, New York.
- Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of uk inflation. *Econometrica*, 50:987–1007.
- Engle, R. and Granger, C. (1987). Co-integration and error correction: representation, estimation and testing. *Econometrica*, 55:251–276.
- Falk, M., Marohn, F., and Tewes, B. (2002). *Foundations of Statistical Analyses and Applications with SAS*. Birkhäuser.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Application*. Chapman and Hall, New York.
- Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*. John Wiley, New York, 2nd edition.
- Ferguson, T. (1996). *A course in large sample theory*. Texts in Statistical Science, Chapman and Hall.

- Fuller, W. A. (1995). *Introduction to Statistical Time Series*. Wiley-Interscience, 2nd edition.
- Granger, C. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16:121–130.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.
- Hannan, E. and Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, 41(2):190–195.
- Horn, R. and Johnson, C. (1985). Matrix analysis. *Cambridge University Press*, pages 176—180.
- Janacek, G. and Swift, L. (1993). *Time Series. Forecasting, Simulation, Applications*. Ellis Horwood, New York.
- Johnson, R. and Wichern, D. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall, 5th edition.
- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, 82:35–45.
- Karr, A. (1993). *Probability*. Springer, New York.
- Kendall, M. and Ord, J. (1993). *Time Series*. Arnold, Sevenoaks, 3rd edition.
- Ljung, G. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65:297–303.
- Murray, M. (1994). A drunk and her dog: an illustration of cointegration and error correction. *The American Statistician*, 48:37–39.
- Newton, J. H. (1988). *Timeslab: A Time Series Analysis Laboratory*. Brooks/Cole, Pacific Grove.

- Parzen, E. (1957). On consistent estimates of the spectrum of the stationary time series. *Annals of Mathematical Statistics*, 28(2):329–348.
- Phillips, P. and Ouliaris, S. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica*, 58:165–193.
- Phillips, P. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75:335–346.
- Pollard, D. (1984). *Convergence of stochastic processes*. Springer, New York.
- Quenouille, M. (1968). *The Analysis of Multiple Time Series*. Griffin, London.
- Reiss, R. (1989). *Approximate Distributions of Order Statistics. With Applications to Nonparametric Statistics*. Springer, New York.
- Rudin, W. (1986). *Real and Complex Analysis*. McGraw-Hill, New York, 3rd edition.
- SAS Institute Inc. (1992). *SAS Procedures Guide, Version 6*. SAS Institute Inc., 3rd edition.
- SAS Institute Inc. (2010a). Sas 9.2 documentation. http://support.sas.com/documentation/cdl_main/index.html.
- SAS Institute Inc. (2010b). Sas onlinedoc 9.1.3. <http://support.sas.com/onlinedoc/913/docMainpage.jsp>.
- Schlittgen, J. and Streitberg, B. (2001). *Zeitreihenanalyse*. Oldenbourg, Munich, 9th edition.
- Seeley, R. (1970). *Calculus of Several Variables*. Scott Foresman, Glenview, Illinois.
- Shao, J. (2003). *Mathematical Statistics*. Springer, New York, 2nd edition.

- Shiskin, J. and Eisenpress, H. (1957). Seasonal adjustment by electronic computer methods. *Journal of the American Statistical Association*, 52:415–449.
- Shiskin, J., Young, A., and Musgrave, J. (1967). The x-11 variant of census method ii seasonal adjustment program. Technical paper 15, Bureau of the Census, U.S. Dept. of Commerce.
- Shumway, R. and Stoffer, D. (2006). *Time Series Analysis and Its Applications — With R Examples*. Springer, New York, 2nd edition.
- Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Series in Statistics, Springer, New York.
- Tintner, G. (1958). Eine neue methode für die schätzung der logistischen funktion. *Metrika*, 1:154–157.
- Tsay, R. and Tiao, G. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary arma models. *Journal of the American Statistical Association*, 79:84–96.
- Tsay, R. and Tiao, G. (1985). Use of canonical analysis in time series model identification. *Biometrika*, 72:229–315.
- Tukey, J. (1949). The sampling theory of power spectrum estimates. proc. symp. on applications of autocorrelation analysis to physical problems. Technical Report NAVEXOS-P-735, Office of Naval Research, Washington.
- Wallis, K. (1974). Seasonal adjustment and relations between variables. *Journal of the American Statistical Association*, 69(345):18–31.
- Wei, W. and Reilly, D. P. (1989). *Time Series Analysis. Univariate and Multivariate Methods*. Addison–Wesley, New York.

Index

- χ^2 -distribution, *see* Distribution
- t -distribution, *see* Distribution
- Absolutely summable filter, *see* Filter
- AIC, *see* Information criterion
- Akaike's information criterion, *see* Information criterion
- Aliasing, 155
- Allometric function, *see* Function
- AR-process, *see* Process
- ARCH-process, *see* Process
- ARIMA-process, *see* Process
- ARMA-process, *see* Process
- Asymptotic normality
 - of random variables, 235
 - of random vectors, 239
- Augmented Dickey–Fuller test,
see Test
- Autocorrelation
 - function, 35
- Autocovariance
 - function, 35
- Backforecasting, 105
- Backward shift operator, 296
- Band pass filter, *see* Filter
- Bandwidth, 173
- Bartlett kernel, *see* Kernel
- Bartlett's formula, 272
- Bayesian information criterion,
see Information criterion
- BIC, *see* Information criterion
- Blackman–Tukey kernel, *see* Kernel
- Boundedness in probability, 259
- Box–Cox Transformation, 40
- Box–Jenkins program, 99, 223,
272
- Box–Ljung test, *see* Test
- Cauchy–Schwarz inequality, 37
- Causal, 58
- Census
 - U.S. Bureau of the, 23
 - X–11 Program, 23
 - X–12 Program, 25
- Change point, 34
- Characteristic function, 235
- Cointegration, 82, 89
 - regression, 83
- Complex number, 47
- Conditional least squares method,
see Least squares method
- Confidence band, 193
- Conjugate complex number, 47
- Continuity theorem, 237
- Correlogram, 36
- Covariance, 48

- Covariance generating function, 55, 56
- Covering theorem, 191
- Cramér–Wold device, 201, 234, 239
- Critical value, 191
- Crosscovariances
 empirical, 140
- Cycle
 Juglar, 25
- Césaro convergence, 183
- Data
 Airline, 37, 131, 193
 Bankruptcy, 45, 147
 Car, 120
 Donauwoerth, 223
 Electricity, 30
 Gas, 133
 Hog, 83, 87, 89
 Hogprice, 83
 Hogsuppl, 83
 Hongkong, 93
 Income, 13
 IQ, 336
 Nile, 221
 Population1, 8
 Population2, 41
 Public Expenditures, 42
 Share, 218
 Star, 136, 144
 Sunspot, iii, 36, 207
 Temperatures, 21
 Unemployed Females, 18, 42
 Unemployed1, 2, 21, 26, 44
 Unemployed2, 42
 Wolf, *see* Sunspot Data
 Wölfer, *see* Sunspot Data
- Degree of Freedom, 91
- Delta-Method, 262
- Demeaned and detrended case, 90
- Demeaned case, 90
- Design matrix, 28
- Diagnostic check, 106, 311
- Dickey–Fuller test, *see* Test
- Difference
 seasonal, 32
- Difference filter, *see* Filter
- Discrete spectral average estimator, 203, 204
- Distribution
 χ^2 -, 189, 213
 t -, 91
 Dickey–Fuller, 86
 exponential, 189, 190
 gamma, 213
 Gumbel, 191, 219
 uniform, 190
- Distribution function
 empirical, 190
- Drift, 86
- Drunkard’s walk, 82
- Equivalent degree of freedom, 213
- Error correction, 85
- ESACF method, 290
- Euler’s equation, 146
- Exponential distribution, *see* Distribution
- Exponential smoother, *see* Filter
- Fatou’s lemma, 51
- Filter
 absolutely summable, 49

- band pass, 173
- difference, 30
- exponential smoother, 33
- high pass, 173
- inverse, 58
- linear, 16, 17, 203
- low pass, 17, 173
- product, 57
- Fisher test for hidden periodicities, *see* Test
- Fisher's kappa statistic, 191
- Forecast
 - by an exponential smoother, 35
 - ex-ante, 324
 - ex-post, 324
- Forecasting, 107, 324
- Forward shift operator, 296
- Fourier frequency, *see* Frequency
- Fourier transform, 147
 - inverse, 152
- Frequency, 135
 - domain, 135
 - Fourier, 140
 - Nyquist, 155, 166
- Function
 - quadratic, 102
 - allometric, 13
 - Cobb–Douglas, 13
 - gamma, 91, 213
 - logistic, 6
 - Mitscherlich, 11
 - positive semidefinite, 160
 - transfer, 167
- Fundamental period, *see* Period
- Gain, *see* Power transfer function
- Gamma distribution, *see* Distribution
- Gamma function, *see* Function
- GARCH-process, *see* Process
- Gaussian process, *see* Process
- General linear process, *see* Process
- Gibb's phenomenon, 175
- Gompertz curve, 11
- Gumbel distribution, *see* Distribution
- Hang Seng
 - closing index, 93
- Hannan–Quinn information criterion, *see* Information criterion
- Harmonic wave, 136
- Helly's selection theorem, 163
- Henderson moving average, 24
- Herglotz's theorem, 161
- Hertz, 135
- High pass filter, *see* Filter
- Imaginary part
 - of a complex number, 47
- Information criterion
 - Akaike's, 100
 - Bayesian, 100
 - Hannan–Quinn, 100
- Innovation, 91
- Input, 17
- Intensity, 142
- Intercept, 13
- Inverse filter, *see* Filter
- Kalman
 - filter, 125, 129
 - h -step prediction, 129

- prediction step of, 129
- updating step of, 129
- gain, 127, 128
- recursions, 125
- Kernel
 - Bartlett, 210
 - Blackman–Tukey, 210, 211
 - Fejer, 221
 - function, 209
 - Parzen, 210
 - triangular, 210
 - truncated, 210
 - Tukey–Hanning, 210, 212
- Kolmogorov’s theorem, 160, 161
- Kolmogorov–Smirnov statistic, 192
- Kronecker’s lemma, 218
- Landau-symbol, 259
- Laurent series, 55
- Leakage phenomenon, 207
- Least squares, 104
 - estimate, 6, 83
 - approach, 105
 - filter design, 173
- Least squares method
 - conditional, 310
- Levinson–Durbin recursion, 229, 316
- Likelihood function, 102
- Lindeberg condition, 201
- Linear process, 204
- Local polynomial estimator, 27
- Log returns, 93
- Logistic function, *see* Function
- Loglikelihood function, 103
- Low pass filter, *see* Filter
- MA-process, *see* Process
- Maximum impregnation, 8
- Maximum likelihood
 - equations, 285
 - estimator, 93, 102, 285
 - principle, 102
- MINIC method, 289
- Mitscherlich function, *see* Function
- Moving average, 17, 203
 - simple, 17
- Normal equations, 27, 139
- North Rhine-Westphalia, 8
- Nyquist frequency, *see* Frequency
- Observation equation, 121
- Order selection, 99, 284
- Order statistics, 190
- Output, 17
- Overfitting, 100, 311, 321
- Partial autocorrelation, 224, 231
 - coefficient, 70
 - empirical, 71
 - estimator, 250
 - function, 225
- Partial correlation, 224
 - matrix, 232
- Partial covariance
 - matrix, 232
- Parzen kernel, *see* Kernel
- Penalty function, 288
- Period, 135
 - fundamental, 136
- Periodic, 136
- Periodogram, 143, 189
 - cumulated, 190
- Phillips–Ouliaris test, *see* Test
- Polynomial

- characteristic, 57
- Portmanteau test, *see* Test
- Power of approximation, 228
- Power transfer function, 167
- Prediction, 6
- Principle of parsimony, 99, 284
- Process
 - autoregressive , 63
 - AR, 63
 - ARCH, 91
 - ARIMA, 80
 - ARMA, 74
 - autoregressive moving average, 74
 - cosinoid, 112
 - GARCH, 93
 - Gaussian, 101
 - general linear, 49
 - MA, 61
 - Markov, 114
 - SARIMA, 81
 - SARMA, 81
 - stationary, 49
 - stochastic, 47
- Product filter, *see* Filter
- Portmanteau test, 320
- R^2 , 15
- Random walk, 81, 114
- Real part
 - of a complex number, 47
- Residual, 6, 316
- Residual sum of squares, 105, 139
- SARIMA-process, *see* Process
- SARMA-process, *see* Process
- Scale, 91
- SCAN-method, 303
- Seasonally adjusted, 16
- Simple moving average, *see* Moving average
- Slope, 13
- Slutsky's lemma, 218
- Spectral
 - density, 159, 162
 - of AR-processes, 176
 - of ARMA-processes, 175
 - of MA-processes, 176
 - distribution function, 161
- Spectrum, 159
- Square integrable, 47
- Squared canonical correlations, 304
- Standard case, 89
- State, 121
 - equation, 121
- State-space
 - model, 121
 - representation, 122
- Stationarity condition, 64, 75
- Stationary process, *see* Process
- Stochastic process, *see* Process
- Strict stationarity, 245
- Taylor series expansion in probability, 261
- Test
 - augmented Dickey–Fuller, 83
 - Bartlett–Kolmogorov–Smirnov, 192, 332
 - Box–Ljung, 107
 - Box–Pierce, 106
 - Dickey–Fuller, 83, 86, 87, 276
 - Fisher for hidden periodicities, 191, 332
 - Phillips–Ouliaris, 83

- Phillips–Ouliaris for cointegration, 89
- Portmanteau, 106, 319
- Tight, 259
- Time domain, 135
- Time series
 - seasonally adjusted, 21
- Time Series Analysis, 1
- Transfer function, *see* Function
- Trend, 2, 16
- Tukey–Hanning kernel, *see* Kernel
- Uniform distribution, *see* Distribution
- Unit root, 86
 - test, 82
- Variance, 47
- Variance stabilizing transformation, 40
- Volatility, 90
- Weights, 16, 17
- White noise, 49
 - spectral density of a, 164
- X–11 Program, *see* Census
- X–12 Program, *see* Census
- Yule–Walker
 - equations, 68, 71, 226
 - estimator, 250

SAS-Index

||, 8
*, 4
;, 4
@@, 37
\$, 4
%MACRO, 316, 332
%MEND, 316
FREQ, 196
N, 19, 39, 96, 138, 196

ADD, 23
ADDITIVE, 27
ADF, 280
ANGLE, 5
ARIMA, *see* PROC
ARMACOV, 118, 316
ARMASIM, 118, 220
ARRAY, 316
AUTOREG, *see* PROC
AXIS, 5, 27, 37, 66, 68, 175, 196

C=*color*, 5
CGREEK, 66
CINV, 216
CLS, 309, 323
CMOVAVE, 19
COEF, 146
COMPLEX, 66
COMPRESS, 8, 80
CONSTANT, 138
CONVERT, 19

CORR, 37

DATA, 4, 8, 96, 175, 276, 280, 284, 316, 332
DATE, 27
DECOMP, 23
DELETE, 32
DESCENDING, 276
DIF, 32
DISPLAY, 32
DIST, 98
DO, 8, 175
DOT, 5

ESACF, 309
ESTIMATE, 309, 320, 332
EWMA, 19
EXPAND, *see* PROC

F=*font*, 8
FIRSTOBS, 146
FORECAST, 319
FORMAT, 19, 276
FREQ, 146
FTEXT, 66

GARCH, 98
GOPTION, 66
GOPTIONS, 32
GOUT, 32
GPLOT, *see* PROC
GREEK, 8

- GREEN, 5
GREPLAY, *see* PROC
H=*height*, 5, 8
HPFDIAG, *see* PROC
I=*display style*, 5, 196
ID, 19
IDENTIFY, 37, 74, 98, 309
IF, 196
IGOUT, 32
IML, 118, 220, *see* PROC
INFILE, 4
INPUT, 4, 27, 37
INTNX, 19, 23, 27

JOIN, 5, 155

L=*line type*, 8
LABEL, 5, 66, 179
LAG, 37, 74, 212
LEAD, 132, 335
LEGEND, 8, 27, 66
LOG, 40

MACRO, *see* %MACRO
MDJ, 276
MEANS, *see* PROC
MEND, *see* %MEND
MERGE, 11, 280, 316, 323
METHOD, 19, 309
MINIC, 309
MINOR, 37, 68
MODE, 23, 280
MODEL, 11, 89, 98, 138
MONTHLY, 27

NDISPLAY, 32
NLAG, 37
NLIN, *see* PROC

NOEST, 332
NOFS, 32
NOINT, 87, 98
NONE, 68
NOOBS, 4
NOPRINT, 196

OBS, 4, 276
ORDER, 37
OUT, 19, 23, 146, 194, 319
OUTCOV, 37
OUTDECOMP, 23
OUTEST, 11
OUTMODEL, 332
OUTPUT, 8, 73, 138

P (*periodogram*), 146, 194, 209
PARAMETERS, 11
PARTCORR, 74
PERIOD, 146
PERROR, 309
PHILLIPS, 89
PI, 138
PLOT, 5, 8, 96, 175
PREFIX, 316
PRINT, *see* PROC
PROBDF, 86
PROC, 4, 175
 ARIMA, 37, 45, 74, 98, 118,
 120, 148, 276, 280, 284,
 309, 319, 320, 323, 332,
 335
 AUTOREG, 89, 98
 EXPAND, 19
 GPLOT, 5, 8, 11, 32, 37, 41,
 45, 66, 68, 73, 80, 85, 96,
 98, 146, 148, 150, 179, 196,
 209, 212, 276, 284, 316,

- 319, 332, 335
 - GREPLAY, 32, 85, 96
 - HPFDIAG, 87
 - IML, 316
 - MEANS, 196, 276, 323
 - NLIN, 11, 41
 - PRINT, 4, 146, 276
 - REG, 41, 138
 - SORT, 146, 276
 - SPECTRA, 146, 150, 194, 209, 212, 276, 332
 - STATESPACE, 132, 133
 - TIMESERIES, 23, 280
 - TRANSPOSE, 316
 - X11, 27, 42, 44
- QUIT, 4
- R (*regression*), 88
- RANNOR, 44, 68, 212
- REG, *see* PROC
- RETAIN, 196
- RHO, 89
- RUN, 4, 32
- S (*spectral density*), 209, 212
- S (*studentized statistic*), 88
- SA, 23
- SCAN, 309
- SEASONALITY, 23, 280
- SET, 11
- SHAPE, 66
- SM, 88
- SORT, *see* PROC
- SPECTRA,, *see* PROC
- STATESPACE, *see* PROC
- STATIONARITY, 89, 280
- STEPJ (*display style*), 196
- SUM, 196
- SYMBOL, 5, 8, 27, 66, 155, 175, 179, 196
- T, 98
- TAU, 89
- TC=*template catalog*, 32
- TDFI, 94
- TEMPLATE, 32
- TIMESERIES, *see* PROC
- TITLE, 4
- TR, 88
- TRANSFORM, 19
- TRANSPOSE, *see* PROC
- TREPLAY, 32
- TUKEY, 212
- V= *display style*, 5
- VAR, 4, 132, 146
- VREF, 37, 85, 276
- W=*width*, 5
- WEIGHTS, 209
- WHERE, 11, 68
- WHITE, 32
- WHITETEST, 194
- X11, *see* PROC
- ZM, 88

GNU Free Documentation License

Version 1.3, 3 November 2008

Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.

<http://fsf.org/>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document “free” in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others. This License is a kind of “copyleft”, which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “**Document**”, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “**you**”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A “**Modified Version**” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A “**Secondary Section**” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics,

a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “**Invariant Sections**” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “**Cover Texts**” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “**Transparent**” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “**Opaque**”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the

machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “**Title Page**” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

The “**publisher**” means any person or entity that distributes copies of the Document to the public.

A section “**Entitled XYZ**” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “**Acknowledgements**”, “**Dedications**”, “**Endorsements**”, or “**History**”.) To “**Preserve the Title**” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on

its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the “History” section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled “Acknowledgements” or “Dedications”, Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled “Endorsements”. Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled “Endorsements” or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version’s license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version.

Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection,

provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a

draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copy-left versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on

the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.