

# Time Series Analysis in Economics

Klaus Neusser

May 26, 2015

# Contents

<b>I</b>	<b>Univariate Time Series Analysis</b>	<b>3</b>
<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Some examples . . . . .	2
1.2	Formal definitions . . . . .	7
1.3	Stationarity . . . . .	10
1.4	Construction of stochastic processes . . . . .	14
1.4.1	White noise . . . . .	14
1.4.2	Construction of stochastic processes: some examples . . . . .	14
1.4.3	Moving average process of order one . . . . .	15
1.4.4	Random walk . . . . .	18
1.4.5	Changing mean . . . . .	19
1.5	Properties of the autocovariance function . . . . .	19
1.5.1	Autocovariance function of a MA(1) process . . . . .	20
1.6	Exercises . . . . .	23
<b>2</b>	<b>ARMA Models</b>	<b>25</b>
2.1	The lag operator . . . . .	26
2.2	Some important special cases . . . . .	27
2.2.1	Moving-average process of order $q$ . . . . .	27
2.2.2	First order autoregressive process . . . . .	28
2.3	Causality and invertibility . . . . .	32
2.4	Computation of Autocovariance Function . . . . .	39
2.4.1	First procedure . . . . .	40
2.4.2	Second procedure . . . . .	43
2.4.3	Third procedure . . . . .	44
2.5	Exercises . . . . .	45
<b>3</b>	<b>Forecasting Stationary Processes</b>	<b>47</b>
3.1	Linear Least-Squares Forecasts . . . . .	47
3.1.1	Forecasting with an AR(p) process . . . . .	51
3.1.2	Forecasting with MA(q) processes . . . . .	52

3.1.3	Forecasting from the Infinite Past . . . . .	55
3.2	The Wold Decomposition . . . . .	57
3.3	Exponential smoothing . . . . .	60
3.4	Exercises . . . . .	64
3.5	Partial Autocorrelation . . . . .	65
3.5.1	Definition . . . . .	65
3.5.2	Interpretation of ACF and PACF . . . . .	68
3.6	Exercises . . . . .	69
<b>4</b>	<b>Estimation of Mean and ACF</b>	<b>71</b>
4.1	Estimation of the mean . . . . .	71
4.2	Estimation of ACF . . . . .	77
4.3	Estimation of PACF . . . . .	82
4.4	Estimation of the long-run variance . . . . .	83
4.4.1	An example . . . . .	88
4.5	Exercises . . . . .	91
<b>5</b>	<b>Estimation of ARMA Models</b>	<b>93</b>
5.1	The Yule-Walker estimator . . . . .	94
5.2	OLS Estimation of an AR( $p$ ) model . . . . .	97
5.3	Estimation of an ARMA( $p,q$ ) model . . . . .	101
5.4	Estimation of the orders $p$ and $q$ . . . . .	105
5.5	Modeling a Stochastic Process . . . . .	109
5.6	Modeling real GDP of Switzerland . . . . .	110
<b>6</b>	<b>Spectral Analysis and Linear Filters</b>	<b>117</b>
6.1	Spectral Density . . . . .	118
6.2	Spectral Decomposition . . . . .	122
6.3	Estimation of Spectral Densities . . . . .	125
6.3.1	Non-parametric Estimation . . . . .	125
6.3.2	Parametric Estimation . . . . .	128
6.4	Linear Time-invariant Filters . . . . .	129
6.5	Some Important Filters . . . . .	134
6.5.1	Construction of Low- and High-Pass Filters . . . . .	134
6.5.2	The Hodrick-Prescott Filter . . . . .	135
6.5.3	Seasonal Filters . . . . .	136
6.5.4	Using Filtered Data . . . . .	138
<b>7</b>	<b>Integrated Processes</b>	<b>141</b>
7.1	Definition, Properties and Interpretation . . . . .	141
7.1.1	Long-run Forecast . . . . .	143

7.1.2	Variance of Forecast Error . . . . .	144
7.1.3	Impulse Response Function . . . . .	145
7.1.4	The Beveridge-Nelson Decomposition . . . . .	146
7.2	OLS estimator with integrated variables . . . . .	149
7.3	Unit-Root Tests . . . . .	153
7.3.1	Dickey-Fuller Test . . . . .	155
7.3.2	Phillips-Perron Test . . . . .	158
7.3.3	Unit-Root Test: Testing Strategy . . . . .	159
7.3.4	Examples for unit root tests . . . . .	161
7.4	Generalizations of Unit-Root Tests . . . . .	164
7.4.1	Structural Breaks in the Trend Function . . . . .	164
7.4.2	KPSS-Test . . . . .	168
7.5	Regression with Integrated Variables . . . . .	169
7.5.1	The Spurious Regression Problem . . . . .	169
7.5.2	Cointegration . . . . .	172
7.5.3	Some rules to deal with integrated times series in re- gressions . . . . .	173
<b>8</b>	<b>Models of Volatility</b>	<b>179</b>
8.1	Specification and Interpretation . . . . .	180
8.1.1	Summary of the Forecasting Properties of the AR(1)- Model . . . . .	180
8.1.2	The ARCH(1) model . . . . .	181
8.1.3	General Models of Volatility . . . . .	185
8.1.4	The GARCH(1,1) Model . . . . .	190
8.2	Tests for Heteroskedasticity . . . . .	196
8.2.1	Autocorrelation of quadratic residuals . . . . .	196
8.2.2	Engle's Lagrange-multiplier . . . . .	197
8.3	Estimation of GARCH(p,q) Models . . . . .	197
8.3.1	Maximum-Likelihood Estimation . . . . .	197
8.3.2	Method of Moment Estimation . . . . .	200
8.4	Example: Swiss Market Index (SMI) . . . . .	201
<b>II</b>	<b>Multivariate Time Series Analysis</b>	<b>209</b>
<b>9</b>	<b>Introduction</b>	<b>211</b>
<b>10</b>	<b>Definitions and Stationarity</b>	<b>215</b>

<b>11 Estimation of Covariance Function</b>	<b>221</b>
11.1 Estimators and their Asymptotic Distributions . . . . .	221
11.2 Testing Cross-Correlations of Time Series . . . . .	224
11.3 Some Examples for Independence Tests . . . . .	225
<b>12 VARMA Processes</b>	<b>231</b>
12.1 The VAR(1) Process . . . . .	232
12.2 Representation in Companion Form . . . . .	234
12.3 Causal Representation . . . . .	235
12.4 Computation of Covariance Function . . . . .	237
<b>13 Estimation of VAR Models</b>	<b>241</b>
13.1 Introduction . . . . .	241
13.2 The Least-Squares Estimator . . . . .	242
13.3 Proofs of Asymptotic Normality . . . . .	247
13.4 The Yule-Walker Estimator . . . . .	254
<b>14 Forecasting with VAR Models</b>	<b>257</b>
14.1 Forecasting with Known Parameters . . . . .	257
14.1.1 Wold Decomposition Theorem . . . . .	261
14.2 Forecasting with Estimated Parameters . . . . .	261
14.3 Modeling of VAR models . . . . .	263
14.4 Example: VAR Model . . . . .	264
<b>15 Interpretation of VAR Models</b>	<b>273</b>
15.1 Wiener-Granger Causality . . . . .	273
15.1.1 VAR Approach . . . . .	274
15.1.2 Wiener-Granger Causality and Causal Representation .	277
15.1.3 Cross-correlation Approach . . . . .	277
15.2 Structural and Reduced Form . . . . .	278
15.2.1 A Prototypical Example . . . . .	278
15.2.2 Identification: the General Case . . . . .	281
15.2.3 Identification: the Case $n = 2$ . . . . .	285
15.3 Identification via Short-run Restrictions . . . . .	286
15.4 Interpretation of VAR Models . . . . .	288
15.4.1 Impulse Response Functions . . . . .	288
15.4.2 Variance Decomposition . . . . .	289
15.4.3 Confidence Intervals . . . . .	290
15.4.4 Example 1: Advertisement and Sales . . . . .	292
15.4.5 Example 2: IS-LM Model with Phillips Curve . . . . .	294
15.5 Identification via Long-Run Restrictions . . . . .	301

15.5.1	A Prototypical Example . . . . .	301
15.5.2	The General Approach . . . . .	305
<b>16</b>	<b>Cointegration</b>	<b>311</b>
16.1	A Theoretical Example . . . . .	312
16.2	Definition and Representation . . . . .	318
16.2.1	Definition . . . . .	318
16.2.2	VAR and VEC Models . . . . .	322
16.2.3	Beveridge-Nelson Decomposition . . . . .	325
16.2.4	Common Trend Representation . . . . .	327
16.3	Johansen Test . . . . .	328
16.4	An Example . . . . .	335
<b>17</b>	<b>Kalman Filter</b>	<b>339</b>
17.1	The State Space Representation . . . . .	340
17.1.1	Examples . . . . .	342
17.2	Filtering and Smoothing . . . . .	349
17.2.1	The Kalman Filter . . . . .	352
17.2.2	The Kalman Smoother . . . . .	354
17.3	Estimation of State Space Models . . . . .	357
17.3.1	The Likelihood Function . . . . .	357
17.3.2	Identification . . . . .	359
17.4	Examples . . . . .	360
17.4.1	Estimation of Quarterly GDP . . . . .	360
17.4.2	Structural Time Series Analysis . . . . .	363
17.5	Exercises . . . . .	364
<b>A</b>	<b>Complex Numbers</b>	<b>367</b>
<b>B</b>	<b>Linear Difference Equations</b>	<b>371</b>
<b>C</b>	<b>Stochastic Convergence</b>	<b>375</b>
<b>D</b>	<b>BN-Decomposition</b>	<b>381</b>
<b>E</b>	<b>The Delta Method</b>	<b>385</b>



# List of Figures

1.1	Real gross domestic product (GDP) . . . . .	3
1.2	Growth rate of real gross domestic product (GDP) . . . . .	4
1.3	Swiss real gross domestic product . . . . .	4
1.4	Short- and long-term Swiss interest rates . . . . .	5
1.5	Swiss Market Index (SMI): . . . . .	6
1.6	Unemployment rate in Switzerland . . . . .	7
1.7	Realization of a Random Walk . . . . .	10
1.8	Realization of a Branching process . . . . .	11
1.9	Processes constructed from a given white noise process . . . . .	16
1.10	Relation between the autocorrelation coefficient of order one, $\rho(1)$ , and the parameter $\theta$ of a MA(1) process . . . . .	22
2.1	Realization and estimated ACF of MA(1) process . . . . .	29
2.2	Realization and estimated ACF of an AR(1) process . . . . .	31
2.3	Autocorrelation function of an ARMA(2,1) process . . . . .	42
3.1	Autocorrelation and partial autocorrelation functions . . . . .	70
4.1	Estimated autocorrelation function of a WN(0,1) process . . . . .	79
4.2	Estimated autocorrelation function of MA(1) process . . . . .	81
4.3	Estimated autocorrelation function of an AR(1) process . . . . .	82
4.4	Estimated PACF of an AR(1) process . . . . .	83
4.5	Estimated PACF for a MA(1) process . . . . .	84
4.6	Common kernel functions . . . . .	86
4.7	Estimated autocorrelation function for the growth rate of GDP . . . . .	89
5.1	Parameter space of a causal and invertible ARMA(1,1) process . . . . .	107
5.2	Real GDP growth rates of Switzerland . . . . .	111
5.3	ACF and PACF of GDP growth rate . . . . .	112
5.4	Inverted roots of the ARMA(1,3) model . . . . .	114
5.5	ACF of the residuals from the AR(2) and the ARMA(1,3) model . . . . .	115
5.6	Impulse responses of the AR(2) and the ARMA(1,3) model . . . . .	116

5.7	Forecasts of real GDP growth rates . . . . .	116
6.1	Examples of spectral densities with $Z_t \sim \text{WN}(0,1)$ . . . . .	122
6.2	Non-parametric direct estimates of a spectral density . . . . .	128
6.3	Nonparametric and parametric estimates of spectral density . . . . .	130
6.4	Transfer function of the Kuznets filters . . . . .	134
6.5	Transfer function of HP-filter . . . . .	137
6.6	HP-filtered US GDP . . . . .	137
6.7	Transfer function of growth rate of investment in the construction sector with and without seasonal adjustment . . . . .	139
7.1	Distribution of the OLS estimator . . . . .	152
7.2	Distribution of t-statistic and standard normal distribution . . . . .	153
7.3	ACF of a random walk with 100 observations . . . . .	154
7.4	Three Structural Breaks at $T_B$ . . . . .	165
7.5	Distribution of $\hat{\beta}$ and t-statistic $t_{\hat{\beta}}$ . . . . .	171
7.6	Cointegration of inflation and three-month LIBOR . . . . .	174
8.1	Simulation of two ARCH(1) processes . . . . .	186
8.2	Parameter region for which a strictly stationary solution to the GARCH(1,1) process exists assuming $\nu_t \sim \text{IIDN}(0,1)$ . . . . .	193
8.3	Daily return of the SMI (Swiss Market Index) . . . . .	202
8.4	Normal-Quantile Plot of SMI returns . . . . .	203
8.5	Histogram of SMI returns . . . . .	204
8.6	ACF of the returns and the squared returns of the SMI . . . . .	205
11.1	Cross-correlations between two independent AR(1) processes . . . . .	226
11.2	Cross-correlations between consumption and advertisement . . . . .	228
11.3	Cross-correlations between GDP and consumer sentiment . . . . .	229
14.1	Forecast Comparison of Alternative Models . . . . .	268
14.2	Forecast of VAR(8) model and 80 percent confidence intervals (red dotted lines) . . . . .	270
15.1	Identification in a two-dimensional structural VAR . . . . .	286
15.2	Impulse response functions for advertisement and sales . . . . .	295
15.3	Impulse response functions of IS-LM model . . . . .	300
15.4	Impulse response functions of the Blanchard-Quah Model . . . . .	310
16.1	Impulse responses of present discounted value model . . . . .	318
16.2	Stochastic simulation of present discounted value model . . . . .	319
17.1	Spectral density of cyclical component . . . . .	347

17.2	Estimates of quarterly GDP growth rates . . . . .	363
17.3	Components of the Basic Structural Model (BSM) for real GDP of Switzerland . . . . .	364
A.1	Representation of a complex number . . . . .	368



# List of Tables

1.1	Construction of stochastic processes . . . . .	16
3.1	Forecast function for a MA(1) process with $\theta = -0.9$ and $\sigma^2 = 1$	55
3.2	Properties of the ACF and the PACF . . . . .	68
4.1	Some popular kernel functions . . . . .	85
5.1	AIC for alternative ARMA(p,q) models . . . . .	113
5.2	BIC for alternative ARMA(p,q) models . . . . .	114
7.1	The four most important cases for the unit-root test . . . . .	156
7.2	Examples of unit root tests . . . . .	163
7.3	Dickey-Fuller Regression allowing for Structural Breaks . . . . .	166
7.4	Critical Values of the KPSS-Test . . . . .	169
7.5	Rules of thumb in regressions with integrated processes (summary) . . . . .	176
8.1	AIC criterion for the variance equation in the GARCH(p,q) model . . . . .	204
8.2	BIC criterion for the variance equation in the GARCH(p,q) model . . . . .	205
8.3	One Percent VaR for the Next Day of the Return to the SMI .	208
8.4	One Percent VaR for the Next 10 Days of the Return to the SMI . . . . .	208
14.1	Information Criteria for the VAR Models of Different Orders .	265
14.2	Forecast evaluation of alternative VAR models . . . . .	269
15.1	Forecast error variance decomposition (FEVD) in terms of demand, supply, price, wage, and money shocks (percentages) . .	302
16.1	Evaluation of the results of Johansen's cointegration test . . .	337



# List of Definitions

1.3	Definition (Model)	9
1.4	Definition (Autocovariance function)	11
1.5	Definition (Stationarity)	11
1.6	Definition (Strict stationarity)	13
1.7	Definition (Strict stationarity)	13
1.8	Definition (Gaussian process)	13
1.9	Definition (White noise)	14
2.1	Definition (ARMA Models)	25
2.2	Definition (Causality)	33
2.3	Definition (Invertibility)	38
6.2	Definition (Periodogram)	126
7.2	Definition (Cointegration, bivariate)	172
8.1	Definition (ARCH(1) model)	181
10.2	Definition (Stationarity)	216
10.3	Definition (Strict Stationarity)	217
12.1	Definition (VARMA process)	231
16.3	Definition (Cointegration)	321
C.1	Definition (Almost Sure Convergence)	376
C.2	Definition (Convergence in Probability)	376
C.3	Definition (Convergence in r-th mean)	376
C.4	Definition (Convergence in Distribution)	377
C.5	Definition (Characteristic function)	378
C.6	Definition (Asymptotic Normality)	379
C.7	Definition (m-Dependence)	379



# List of Theorems

3.1	Theorem (Wold Decomposition)	57
4.1	Theorem (Convergence of arithmetic average)	72
4.2	Theorem (Asymptotic distribution of sample mean)	73
5.1	Theorem (Asymptotic property of Yule-Walker estimator)	95
5.2	Theorem (Asymptotic property of the Least-Squares estimator)	98
5.3	Theorem (Asymptotic Distribution of ML Estimator)	104
6.1	Theorem (Properties of a spectral density)	120
6.2	Theorem (Spectral representation)	123
6.4	Theorem (Autocovariance function of filtered process)	130
13.1	Theorem (Asymptotic distribution of OLS estimator)	245
16.1	Theorem (Beveridge-Nelson decomposition)	320
C.1	Theorem (Cauchy-Bunyakovskii-Schwarz inequality)	375
C.2	Theorem (Minkowski's inequality)	375
C.3	Theorem (Chebyshev's inequality)	375
C.4	Theorem (Borel-Cantelli lemma)	376
C.5	Theorem (Kolmogorov's Strong Law of Large Numbers (SLLN))	376
C.6	Theorem (Riesz-Fisher)	377
C.9	Theorem (Continuous Mapping Theorem)	378
C.10	Theorem (Slutzky's Lemma)	378
C.11	Theorem (Convergence of characteristic functions, Lévy)	379
C.12	Theorem (Central Limit Theorem)	379
C.13	Theorem (CLT for m-dependent processes)	380
C.14	Theorem (Basis Approximation Theorem)	380



**Part I**

**Univariate Time Series  
Analysis**



# Chapter 1

## Introduction and Basic Theoretical Concepts

Time series analysis is an integral part of every empirical investigation which aims at describing and modeling the evolution over time of a variable or a set of variables in a statistically coherent way. The economics of time series analysis is thus very much intermingled with macroeconomics and finance which are concerned with the construction of dynamic models. In principle, one can approach the subject from two complementary perspectives. The first one focuses on *descriptive statistics*. It characterizes the empirical properties and regularities using basic statistical concepts like mean, variance, and covariance. These properties can be directly measured and estimated from the data using standard statistical tools. Thus, they summarize the *external* (observable) or outside characteristics of the time series. The second perspective tries to capture the *internal data generating mechanism*. This mechanism is usually unknown in economics as the models developed in economic theory are mostly of a qualitative nature and are usually not specific enough to single out a particular mechanism.<sup>1</sup> Thus, one has to consider some larger class of models. By far most widely used is the class of autoregressive moving-average (ARMA) models which rely on linear stochastic difference equations with constant coefficients. Of course, one wants to know how the two perspectives are related which leads to the important problem of *identifying* a model from the data.

---

<sup>1</sup> One prominent exception is the random-walk hypothesis of real private consumption first derived and analyzed by Hall (1978). This hypothesis states that the current level of private consumption should just depend on private consumption one period ago and on no other variable, in particular not on disposable income. The random-walk property of asset prices is another very much discussed hypothesis. See Campbell et al. (1997) for a general exposition and Samuelson (1965) for a first rigorous derivation from market efficiency.

The observed regularities summarized in the form of descriptive statistics or as a specific model are, of course, of principal interest to economics. They can be used to test particular theories or to uncover new features. One of the main assumptions underlying time series analysis is that the regularities observed in the sample period are not specific to that period, but can be extrapolated into the future. This leads to the issue of forecasting which is another major application of time series analysis.

Although time series analysis has its roots in the natural sciences and in engineering, several techniques are specific to economics. I just want to mention the analysis of univariate and multivariate integrated, respectively cointegrated time series (see Chapters 7 and 16), the identification of vector autoregressive (VAR) models (see Chapter 15), and the analysis of volatility of financial market data in Chapter 8. Each of these topics alone would justify the treatment of time series analysis in economics as a separate subfield.

## 1.1 Some examples

Before going into more formal analysis, it is useful to examine some prototypical economic time series by plotting them against time. This simple graphical inspection already reveals some of the issues encountered in this book. One of the most popular time series is the real gross domestic product. Figure 1.1 plots the data for the U.S. from 1947 first quarter to 2011 last quarter on logarithmic scale. Several observations are in order. *First*, the data at hand cover just a part of the time series. There are data available before 1947 and there will be data available after 2011. As there is no natural starting nor end point, we think of a time series as extending back into the infinite past and into the infinite future. *Second*, the observations are treated as the realizations of a random mechanism. This implies that we observe only one realization. If we could turn back time and let run history again, we would obtain a second realization. This is, of course, impossible, at least in the macroeconomics context. Thus, typically, we are faced with just one realization on which to base our analysis. However, sound statistical analysis needs many realizations. This implies that we have to make some assumption on the constancy of the random mechanism over time. This leads to the concept of stationarity which will be introduced more rigorously in the next section. *Third*, even a cursory look at the plot reveals that the mean of real GDP is not constant, but is upward *trending*. As we will see, this feature is typical of many economic time series.<sup>2</sup> The investigation into the nature of the trend and the statistical consequences thereof have been the

---

<sup>2</sup>See footnote 1 for some theories predicting non-stationary behavior.

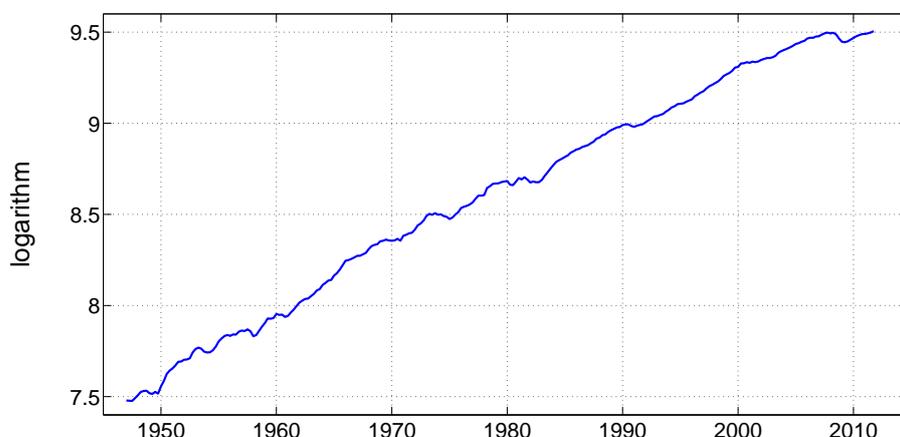


Figure 1.1: Real gross domestic product (GDP) of the U.S. (chained 2005 dollars; seasonally adjusted annual rate)

subject of intense research over the last couple of decades. *Fourth*, a simple way to overcome this problem is to take *first differences*. As the data have been logged, this amounts to taking growth rates.<sup>3</sup> The corresponding plot is given in Figure 1.2 which shows no trend anymore.

Another feature often encountered in economic time series is *seasonality*. This issue arises, for example in the case of real GDP, because of a particular regularity within a year: the first quarter being the quarter with the lowest values, the second and fourth quarter those with the highest values, and the third quarter being in between. These movements are due to climatical and holiday seasonal variations within the year and are viewed to be of minor economic importance. Moreover, these seasonal variations, because of their size, hide the more important business cycle movements. It is therefore customary to work with time series which have been adjusted for seasonality before hand. Figure 1.3 shows the unadjusted and the adjusted real gross domestic product for Switzerland. The adjustment has been achieved by taking a moving-average. This makes the time series much smoother and evens out the seasonal movements.

Other typical economic time series are interest rates plotted in Figure 1.4. Over the period considered these two variables also seem to trend. However, the nature of this trend must be different because of the theoretically binding zero lower bound. Although the relative level of the two series changes over

---

<sup>3</sup>This is obtained by using the approximation  $\ln(1 + \varepsilon) \approx \varepsilon$  for small  $\varepsilon$  where  $\varepsilon$  equals the growth rate of GDP.

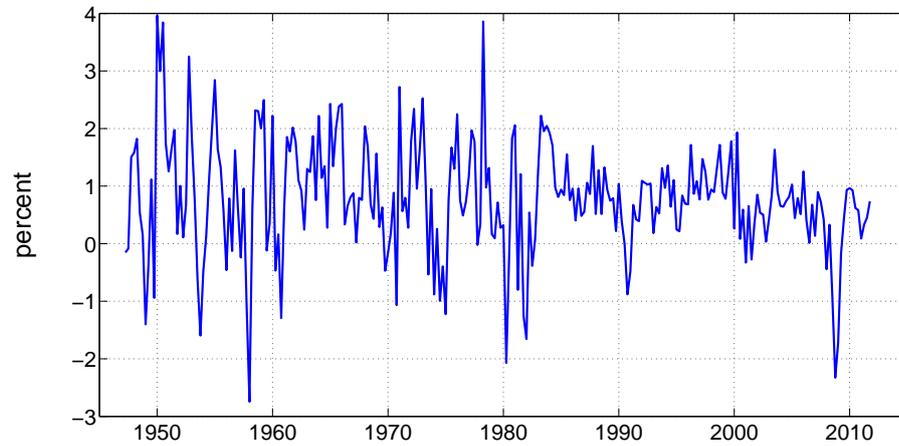


Figure 1.2: Quarterly growth rate of U.S. real gross domestic product (GDP) (chained 2005 dollars)

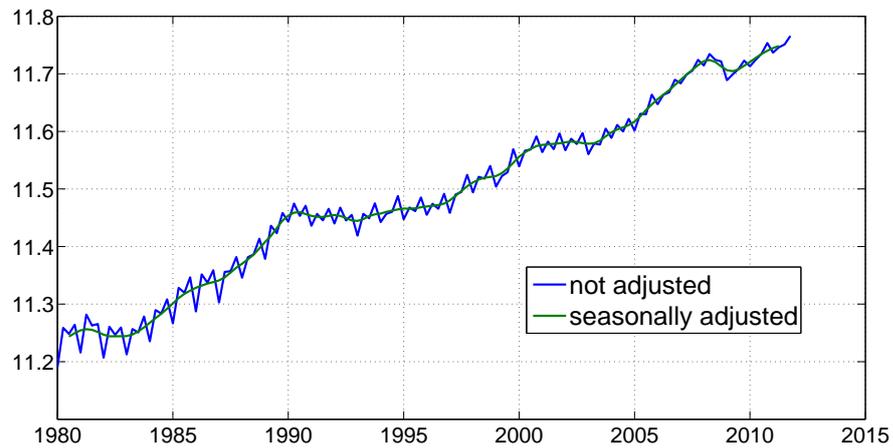


Figure 1.3: Comparison of unadjusted and seasonally adjusted Swiss real gross domestic product (GDP)

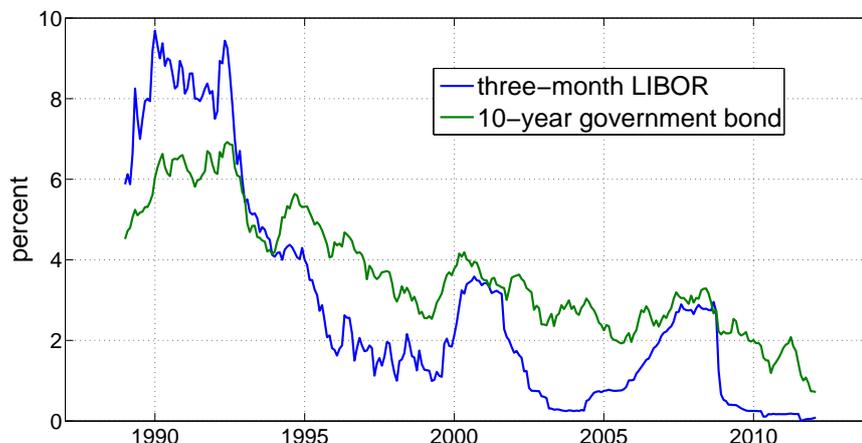


Figure 1.4: Short- and long-term Swiss interest rates (three-month LIBOR and 10 year government bond)

time – at the beginning of the sample, short-term rates are higher than long-term ones – they move more or less together. This *comovement* is true in particular true with respect to the medium- and long-term.

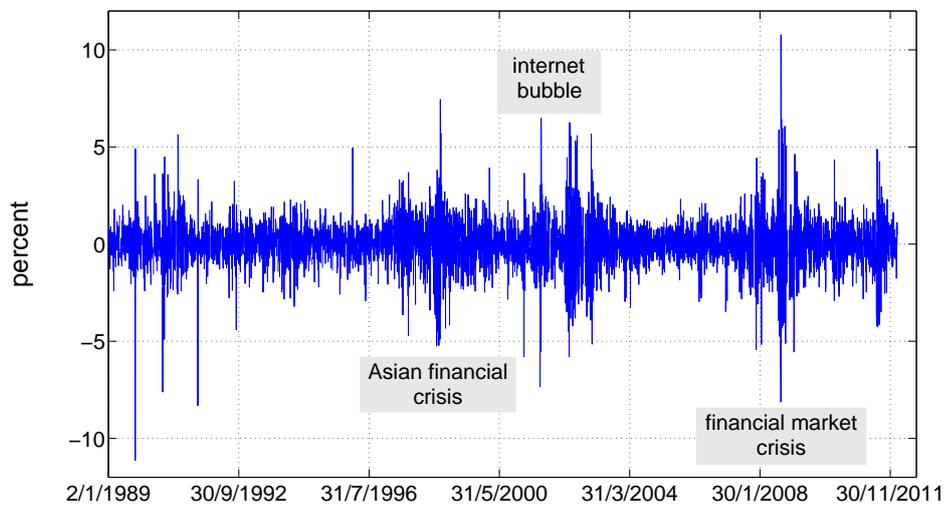
Other prominent time series are stock market indices. In Figure 1.5 the Swiss Market Index (SMI) is plotted as an example. The first panel displays the raw data on a logarithmic scale. One can clearly discern the different crises: the internet bubble in 2001 and the most recent financial market crisis in 2008. More interesting than the index itself is the return on the index plotted in the second panel. Whereas the mean seems to stay relatively constant over time, the volatility is not: in the periods of crisis volatility is much higher. This *clustering of volatility* is a typical feature of financial market data and will be analyzed in detail in Chapter 8.

Finally, Figure 1.6 plots the unemployment rate for Switzerland. This is another widely discussed time series. However, the Swiss data have a particular feature in that the behavior of the series changes over time. Whereas unemployment was practically nonexistent in Switzerland up to the end of 1990's, several policy changes (introduction of unemployment insurance, liberalization of immigration laws) led to drastic shifts. Although such dramatic *structural breaks* are rare, one has to be always aware of such a possibility. Reasons for breaks are policy changes and simply structural changes in the economy at large.<sup>4</sup>

<sup>4</sup>Burren and Neusser (2012) investigate, for example, how systematic sectoral shifts affect volatility of real GDP growth.



(a) Index



(b) daily return

Figure 1.5: Swiss Market Index (SMI):

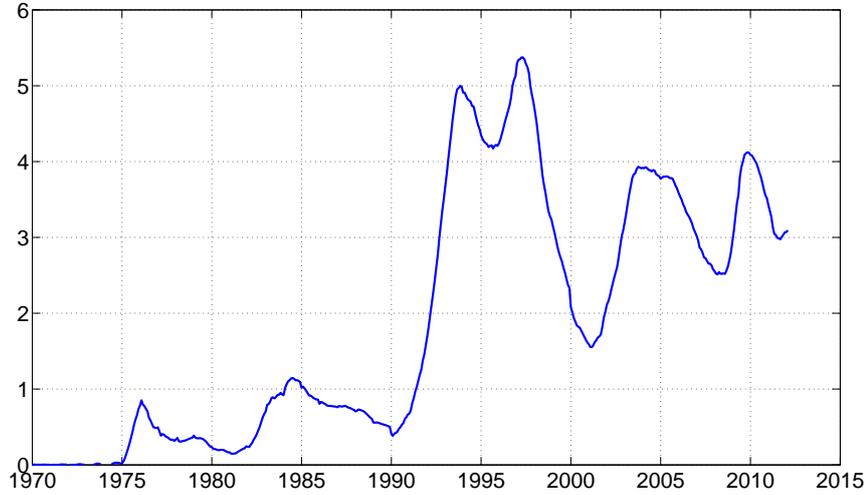


Figure 1.6: Unemployment rate in Switzerland

## 1.2 Formal definitions

The previous section attempted to give an intuitive approach of the subject. The analysis to follow necessitates, however, more precise definitions and concepts. At the heart of the exposition stands the concept of a stochastic process. For this purpose we view the observation at some time  $t$  as the realization of random variable  $X_t$ . In time series analysis we are, however, in general not interested in a particular point in time, but rather in a whole sequence. This leads to the following definition.

**Definition 1.1.** *A stochastic process  $\{X_t\}$  is a family of random variables indexed by  $t \in \mathcal{T}$  and defined on some given probability space.*

Thereby  $\mathcal{T}$  denotes an ordered index set which is typically identified with time. In the literature one can encounter the following index sets:

$$\begin{aligned} \text{discrete time: } \mathcal{T} &= \{1, 2, \dots\} = \mathbb{N} \\ \text{discrete time: } \mathcal{T} &= \{\dots, -2, -1, 0, 1, 2, \dots\} = \mathbb{Z} \\ \text{continuous time: } \mathcal{T} &= [0, \infty) = \mathbb{R}^+ \text{ or } \mathcal{T} = (-\infty, \infty) = \mathbb{R} \end{aligned}$$

**Remark 1.1.** *Given that  $\mathcal{T}$  is identified with time and thus has a direction, a characteristic of time series analysis is the distinction between past, present, and future.*

For technical reasons which will become clear later, we will work with  $\mathcal{T} = \mathbb{Z}$ , the set of integers. This choice is consistent with use of time indices in economics as there is, usually, no natural starting point nor a foreseeable endpoint. Although models in continuous time are well established in the theoretical finance literature, we will disregard them because observations are always of a discrete nature and because models in continuous time would need substantially higher mathematical requirements.

**Remark 1.2.** *The random variables  $\{X_t\}$  take values in a so-called state space. In the first part of this treatise, we take as the state space the space of real numbers  $\mathbb{R}$  and thus consider only univariate time series. In part II we extend the state space to  $\mathbb{R}^n$  and study multivariate times series. Theoretically, it is possible to consider other state spaces (for example,  $\{0, 1\}$ , the integers, or the complex numbers), but this will not be pursued here.*

**Definition 1.2.** *The function  $t \rightarrow x_t$  which assigns to each point in time  $t$  the realization of the random variable  $X_t$ ,  $x_t$ , is called a realization or a trajectory of the stochastic process. We denote such a realization by  $\{x_t\}$ .*

We denominate by a *time series* the realization or trajectory (observations or data), or the underlying stochastic process. Usually, there is no room for misunderstandings. A trajectory therefore represents one observation of the stochastic process. Whereas in standard statistics a sample consists of several, typically, independent draws from the same distribution, a sample in time series analysis is just one trajectory. Thus we are confronted with a situation where there is in principle just one observation. We cannot turn back the clock and get additional trajectories. The situation is even worse as we typically observe only the realizations in a particular time window. For example, we might have data on US GDP from the first quarter 1960 up to the last quarter in 2011. But it is clear, the United States existed before 1960 and will continue to exist after 2011, so that there are in principle observations before 1960 and after 2011. In order to make a meaningful statistical analysis, it is therefore necessary to assume that the observed part of the trajectory is typical for the time series as a whole. This idea is related to the concept of *stationarity* which we will introduce more formally below. In addition, we want to require that the observations cover in principle all possible events. This leads to the concept of *ergodicity*. We avoid a formal definition of ergodicity as this would require a sizeable amount of theoretical probabilistic background material which goes beyond the scope this treatise.<sup>5</sup>

---

<sup>5</sup>In the theoretical probability theory ergodicity is an important concept which asks the question under which conditions the time average of a property is equal to the cor-

An important goal of time series analysis is to build a model given the realization (data) at hand. This amounts to specify the *joint distribution* of some set of  $X_t$ 's with corresponding realization  $\{x_t\}$ .

**Definition 1.3** (Model). *A time series model or a model for the observations (data)  $\{x_t\}$  is a specification of the joint distribution of  $\{X_t\}$  for which  $\{x_t\}$  is a realization.*

The Kolmogorov existence theorem ensures that the specification of all *finite dimensional* distributions is sufficient to characterize the whole stochastic process (see Billingsley (1986), Brockwell and Davis (1991), or Kallenberg (2002)).

Most of the time it is too involved to specify the complete distribution so that one relies on only the first two moments. These moments are then given by the means  $\mathbb{E}X_t$ , the variances  $\mathbb{V}X_t$ ,  $t \in \mathbb{Z}$ , and the covariances  $\text{cov}(X_t, X_s) = \mathbb{E}(X_t - \mathbb{E}X_t)(X_s - \mathbb{E}X_s) = \mathbb{E}(X_t X_s) - \mathbb{E}X_t \mathbb{E}X_s$ , respectively the correlations  $\text{corr}(X_t, X_s) = \text{cov}(X_t, X_s) / (\sqrt{\mathbb{V}X_t} \sqrt{\mathbb{V}X_s})$ ,  $t, s \in \mathbb{Z}$ . If the random variables are jointly normally distributed then the specification of the first two moments is sufficient to characterize the whole distribution.

### Examples of stochastic processes

- $\{X_t\}$  is a sequence of independently distributed random variables with values in  $\{-1, 1\}$  such that  $\mathbf{P}[X_t = 1] = \mathbf{P}[X_t = -1] = 1/2$ .  $X_t$  represents, for example, the payoff after tossing a coin: if head occurs one gets a Euro whereas if tail occurs one has to pay a Euro.
- The simple *random walk*  $\{S_t\}$  is defined by

$$S_t = S_{t-1} + X_t = \sum_{i=1}^t X_i \quad \text{with } t \geq 0 \text{ and } S_0 = 0,$$

where  $\{X_t\}$  is the process from the example just above. In this case  $S_t$  is the proceeds after  $t$  rounds of coin tossing. More generally,  $\{X_t\}$  could be any sequence of identically and independently distributed random variables. Figure 1.7 shows a realization of  $\{X_t\}$  for  $t = 1, 2, \dots, 100$  and the corresponding random walk  $\{S_t\}$ . For more on random walks see Section 1.4.4 and, in particular, Chapter 7.

---

responding ensemble average, i.e. the average over the entire state space. In particular, ergodicity ensures that the arithmetic averages over time converge to their theoretical counterparts. In chapter 4 we allude to this principle in the estimation of the mean and the autocovariance function of a time series.

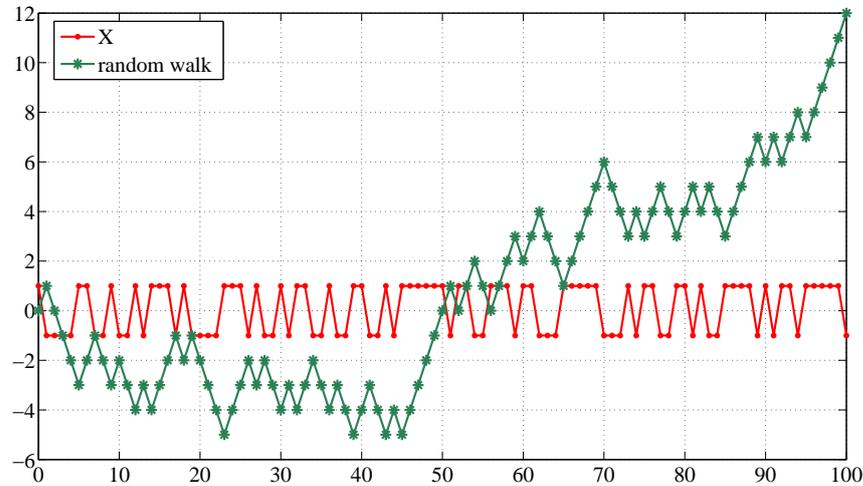


Figure 1.7: Realization of a Random Walk

- The simple branching process is defined through the recursion

$$X_{t+1} = \sum_{j=1}^{X_t} Z_{t,j} \quad \text{with starting value: } X_0 = x_0.$$

In this example  $X_t$  represents the size of a population where each member lives just one period and reproduces itself with some probability.  $Z_{t,j}$  thereby denotes the number of offsprings of the  $j$ -th member of the population in period  $t$ . In the simplest case  $\{Z_{t,j}\}$  is nonnegative integer valued and identically and independently distributed. A realization with  $X_0 = 100$  and with probabilities of one third each that the member has no, one, or two offsprings is shown as an example in Figure 1.8.

### 1.3 Stationarity

An important insight in time series analysis is that the realizations in different periods are related with each other. The value of GDP in some year obviously depends on the values from previous years. This temporal dependence can be represented either by an explicit model or, in a descriptive way, by covariances, respectively correlations. Because the realization of  $X_t$  in some year  $t$  may depend, in principle, on all past realizations  $X_{t-1}, X_{t-2}, \dots$ ,

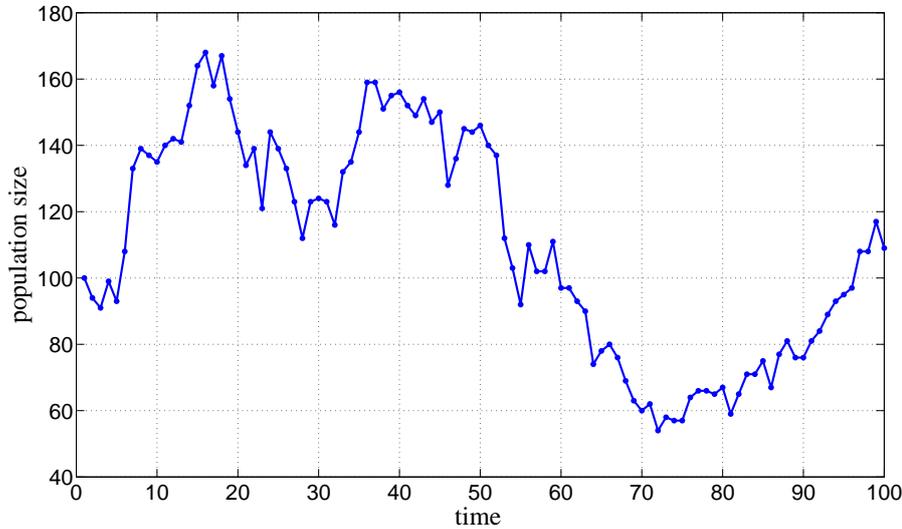


Figure 1.8: Realization of a Branching process

we do not have to specify just a finite number of covariances, but infinitely many covariances. This leads to the concept of the *covariance function*. The covariance function is not only a tool for summarizing the statistical properties of a time series, but is also instrumental in the derivation of forecasts (Chapter 3), in the estimation of ARMA models, the most important class of models (Chapter 5), and in the Wold representation (Section 3.2 in Chapter 3). It is therefore of utmost importance to get a thorough understanding of the meaning and properties of the covariance function.

**Definition 1.4** (Autocovariance function). *Let  $\{X_t\}$  be a stochastic process with  $\forall X_t < \infty$  for all  $t \in \mathbb{Z}$  then the function which assigns to any two time periods  $t$  and  $s$ ,  $t, s \in \mathbb{Z}$ , the covariance between  $X_t$  and  $X_s$  is called the autocovariance function of  $\{X_t\}$ . The autocovariance function is denoted by  $\gamma_X(t, s)$ . Formally this function is given by*

$$\gamma_X(t, s) = \text{cov}(X_t, X_s) = \mathbb{E}[(X_t - \mathbb{E}X_t)(X_s - \mathbb{E}X_s)] = \mathbb{E}X_t X_s - \mathbb{E}X_t \mathbb{E}X_s.$$

**Remark 1.3.** *The acronym auto emphasizes that the covariance is computed with respect to the same variable taken at different points in time. Alternatively, one may use the term covariance function for short.*

**Definition 1.5** (Stationarity). *A stochastic process  $\{X_t\}$  is called stationary if and only if for all integers  $r$ ,  $s$  and  $t$  the following properties hold:*

- (i)  $\mathbb{E}X_t = \mu$  constant;

- (ii)  $\mathbb{V}X_t < \infty$ ;
- (iii)  $\gamma_X(t, s) = \gamma_X(t + r, s + r)$ .

**Remark 1.4.** Processes with these properties are often called *weakly stationary*, *wide-sense stationary*, *covariance stationary*, or *second order stationary*. As we will not deal with other forms of stationarity, we just speak of *stationary processes*, for short.

**Remark 1.5.** For  $t = s$ , we have  $\gamma_X(t, s) = \gamma_X(t, t) = \mathbb{V}X_t$ . Thus, if  $\{X_t\}$  is stationary  $\gamma_X(t, t) = \mathbb{V}X_t = \text{constant}$  which is nothing but the unconditional variance of  $X_t$ .

**Remark 1.6.** If  $\{X_t\}$  is stationary, by setting  $r = -s$  the autocovariance function becomes:

$$\gamma_X(t, s) = \gamma_X(t - s, 0).$$

Thus the covariance  $\gamma_X(t, s)$  does not depend on the points in time  $t$  and  $s$ , but only on the number of periods  $t$  and  $s$  are apart from each other, i.e. from  $t - s$ . For stationary processes it is therefore possible to view the autocovariance function as a function of just one argument. We denote the autocovariance function in this case by  $\gamma_X(h)$ ,  $h \in \mathbb{Z}$ . Because the covariance is symmetric in  $t$  and  $s$ , i.e.  $\gamma_X(t, s) = \gamma_X(s, t)$ , we have

$$\gamma_X(h) = \gamma_X(-h) \quad \text{for all integers } h.$$

It is thus sufficient to look at the autocovariance function for positive integers only, i.e. for  $h = 0, 1, 2, \dots$ . In this case we refer to  $h$  as the *order of the autocovariance*. For  $h = 0$ , we get the unconditional variance of  $X_t$ , i.e.  $\gamma_X(0) = \mathbb{V}X_t$ .

In practice it is more convenient to look at the autocorrelation coefficients instead of the autocovariances. The *autocorrelation function* (ACF) for stationary processes is defined as:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \text{corr}(X_{t+h}, X_t) \quad \text{for all integers } h$$

where  $h$  is referred to as the order. Note that this definition is equivalent to the ordinary correlation coefficients  $\rho(h) = \frac{\text{cov}(X_t, X_{t-h})}{\sqrt{\mathbb{V}X_t}\sqrt{\mathbb{V}X_{t-h}}}$  because stationarity implies that  $\mathbb{V}X_t = \mathbb{V}X_{t-h}$  so that  $\sqrt{\mathbb{V}X_t}\sqrt{\mathbb{V}X_{t-h}} = \mathbb{V}X_t = \gamma_X(0)$ .

Most of the time it is sufficient to concentrate on the first two moments. However, there are situations where it is necessary to look at the whole distribution. This leads to the concept of *strict stationarity*.

**Definition 1.6** (Strict stationarity). *A stochastic process is called strictly stationary if the joint distributions of  $(X_{t_1}, \dots, X_{t_n})$  and  $(X_{t_1+h}, \dots, X_{t_n+h})$  are the same for all  $h \in \mathbb{Z}$  and all  $(t_1, \dots, t_n) \in \mathcal{T}^n$ ,  $n = 1, 2, \dots$*

**Definition 1.7** (Strict stationarity). *A stochastic process is called strictly stationary if for all integers  $h$  and  $n \geq 1$   $(X_1, \dots, X_n)$  and  $(X_{1+h}, \dots, X_{n+h})$  have the same distribution.*

**Remark 1.7.** *Both definitions are equivalent.*

**Remark 1.8.** *If  $\{X_t\}$  is strictly stationary then  $X_t$  has the same distribution for all  $t$  ( $n=1$ ). For  $n = 2$  we have that  $X_{t+h}$  and  $X_t$  have a joint distribution which is independent of  $t$ . This implies that the covariance depends only on  $h$ . Thus, every strictly stationary process with  $\forall X_t < \infty$  is also stationary.<sup>6</sup> The converse is, however, not true as shown by the following example::*

$$X_t \sim \begin{cases} \text{exponentially distributed with mean 1 (i.e. } f(x) = e^{-x}), & t \text{ uneven;} \\ N(1, 1), & t \text{ even;} \end{cases}$$

whereby the  $X_t$ 's are independently distributed. In this example we have:

- $\mathbb{E}X_t = 1$
- $\gamma_X(0) = 1$  and  $\gamma_X(h) = 0$  for  $h \neq 0$

Thus  $\{X_t\}$  is stationary, but not strictly stationary, because the distribution changes depending on whether  $t$  is even or uneven.

**Definition 1.8** (Gaussian process). *A stochastic process  $\{X_t\}$  is called a Gaussian process if all finite dimensional distributions of  $\{X_t\}$  are multivariate normally distributed.*

**Remark 1.9.** *A Gaussian process is strictly stationary. For all  $n, h, t_1, \dots, t_n$ ,  $(X_{t_1}, \dots, X_{t_n})$  and  $(X_{t_1+h}, \dots, X_{t_n+h})$  have the same mean and the same covariance matrix.*

At this point we will not delve into the relation between stationarity, strict stationarity and Gaussian processes, rather some of these issues will be further discussed in Chapter 8.

---

<sup>6</sup>An example of a process which is strictly stationary, but not stationary, is given by the IGARCH process (see Section 8.1.4). This process is strictly stationary with infinite variance.

## 1.4 Construction of stochastic processes

One important idea in time series analysis is to build up more complicated process from simple ones. The simplest building block is a process with no autocorrelation called a *white noise* process which is introduced below. Taking moving-averages from this process or using it in a recursion gives rise to more sophisticated process with more elaborated autocovariance functions.

### 1.4.1 White noise

The simplest building block is a process with no autocorrelation called a *white noise* process.

**Definition 1.9** (White noise).  $\{Z_t\}$  is called a white noise process if  $\{Z_t\}$  satisfies:

- $\mathbb{E}Z_t = 0$
- $\gamma_Z(h) = \begin{cases} \sigma^2 & h = 0; \\ 0 & h \neq 0. \end{cases}$

We denote this by  $Z_t \sim \text{WN}(0, \sigma^2)$ .

The white noise process is therefore stationary and temporally uncorrelated, i.e. the ACF is always equal to zero, except for  $h = 0$  where it is equal to one. As the ACF possesses no structure it is impossible to draw inferences from past observations to its future development, at least in a least square setting with linear forecasting functions (see Chapter 3). Therefore one can say that a white noise process has no memory.

If  $\{Z_t\}$  is not only temporally uncorrelated, but also independently and identically distributed we write  $Z_t \sim \text{IID}(0, \sigma^2)$ . If in addition  $Z_t$  is normally distributed, we write  $Z_t \sim \text{IIN}(0, \sigma^2)$ . An  $\text{IID}(0, \sigma^2)$  process is always a white noise process. The converse is, however, not true as will be shown in Chapter 8.

### 1.4.2 Construction of stochastic processes: some examples

We will now illustrate how complex stationary processes can be constructed by manipulating of a white noise process. In Table 1.1 we report in column 2 the first 6 realizations of a white noise process  $\{Z_t\}$ . Figure 1.4.2 plots the first 100 observations. We can now construct a new process  $\{X_t^{(\text{MA})}\}$

by taking moving-averages over adjacent periods. More specifically, we take  $X_t = Z_t + 0.9Z_{t-1}$ ,  $t = 2, 3, \dots$ . Thus, the realization of  $\{X_t^{(MA)}\}$  in period 2 is  $\{x_2^{(MA)}\} = -0.6387 = -0.8718 + 0.9 \times 0.2590$ .<sup>7</sup> The realization in period 3 is  $\{x_3^{(MA)}\} = -1.5726 = -0.7879 + 0.9 \times -0.6387$ , and so on. The resulting realizations of  $\{X_t^{(MA)}\}$  for  $t = 2, \dots, 6$  are reported in the third column of Table 1.1 and the plot is shown in Figure 1.4.2. One can see that the averaging makes the series more smooth. In section 1.4.3 we will provide a more detailed analysis of this moving-average process.

Another construction device is a recursion  $X_t^{(AR)} = \phi X_{t-1}^{(AR)} + Z_t$ ,  $t = 2, 3, \dots$ , with starting value  $X_1^{(AR)} = Z_1$ . Such a process is called autoregressive because it refers to its own past. Taking  $\phi = 0.9$ , the realization of  $\{X_t^{(AR)}\}$  in period 2 is  $\{x_2^{(AR)}\} = -0.6387 = 0.9 \times 0.2590 - 0.8718$ , in period 3  $\{x_3^{(AR)}\} = -1.3627 = 0.9 \times -0.6387 - 0.7879$ , and so on. Again the resulting realizations of  $\{X_t^{(AR)}\}$  for  $t = 2, \dots, 6$  are reported in the fourth column of Table 1.1 and the plot is shown in Figure 1.4.2. One can see how the series becomes more persistent. In section 2.2.2 we will provide a more detailed analysis of this autoregressive process.

Finally, we construct a new process by taking cumulative sums:  $X_t^{(RW)} = \sum_{\tau=1}^t Z_\tau$ . This process can also be obtained from the recursion above by taking  $\phi = 1$  so that  $X_t^{(RW)} = X_{t-1}^{(RW)} + Z_t$ . It is called a random walk. Thus, the realization of  $\{X_t^{(RW)}\}$  for period 2 is  $\{x_2^{(RW)}\} = -0.6128 = 0.2590 - 0.8718$ , for period 3  $\{x_3^{(RW)}\} = -1.4007 = -0.6128 - 0.7879$ , and so on. Again the resulting realizations of  $\{X_t^{(RW)}\}$  for  $t = 2, \dots, 6$  are reported in the last column of Table 1.1 and the plot is shown in Figure 1.4.2. One can see how the series moves away from its mean of zero more persistently than all the other three processes considered. In section 1.4.4 we will provide a more detailed analysis of this so-called random walk process and show that it is not stationary.

### 1.4.3 Moving average process of order one

The white noise process can be used as a building block to construct more complex processes with a more involved autocorrelation structure. The simplest procedure is to take moving averages over consecutive periods.<sup>8</sup> This leads to the moving-average processes. The moving-average process of order

<sup>7</sup>The following calculations are subject to rounding to four digits.

<sup>8</sup>This procedure is an example of a filter. Section 6.4 provides a general introduction to filters.

time	white noise	moving-average	auto-regressive	random walk
1	0.2590	0.2590	0.2590	0.2590
2	-0.8718	-0.6387	-0.6387	-0.6128
3	-0.7879	-1.5726	-1.3627	-1.4007
4	-0.3443	-1.0535	-1.5708	-1.7451
5	0.6476	0.3377	-0.7661	-1.0974
6	2.0541	2.6370	1.3646	0.9567
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1.1: Construction of stochastic processes

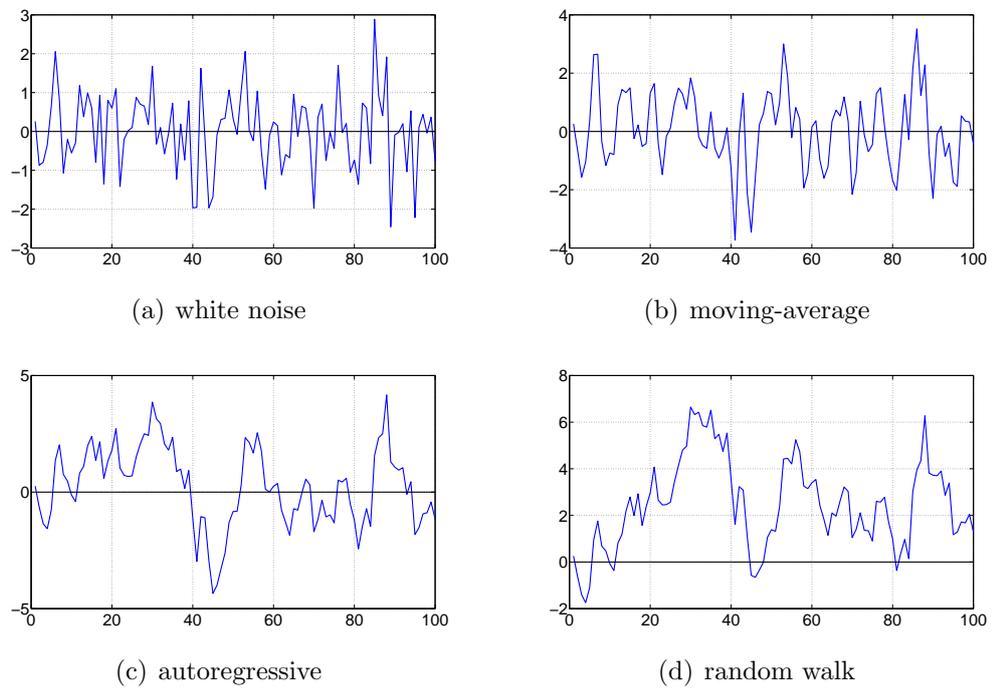


Figure 1.9: Processes constructed from a given white noise process

one, MA(1) process, is defined as

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

Clearly,  $\mathbb{E}X_t = \mathbb{E}Z_t + \theta\mathbb{E}Z_{t-1} = 0$ . The mean is therefore constant and equal to zero.

The autocovariance function can be generated as follows:

$$\begin{aligned} \gamma_X(t+h, t) &= \text{cov}(X_{t+h}, X_t) \\ &= \text{cov}(Z_{t+h} + \theta Z_{t+h-1}, Z_t + \theta Z_{t-1}) \\ &= \mathbb{E}Z_{t+h}Z_t + \theta\mathbb{E}Z_{t+h}Z_{t-1} + \theta\mathbb{E}Z_{t+h-1}Z_t + \theta^2\mathbb{E}Z_{t+h-1}Z_{t-1}. \end{aligned}$$

Recalling that  $\{Z_t\}$  is white noise so that  $\mathbb{E}Z_t^2 = \sigma^2$  and  $\mathbb{E}Z_tZ_{t+h} = 0$  for  $h \neq 0$ . We therefore get the following autocovariance function of  $\{X_t\}$ :

$$\gamma_X(h) = \begin{cases} (1 + \theta^2)\sigma^2 & h = 0; \\ \theta\sigma^2 & h = \pm 1; \\ 0 & \text{otherwise.} \end{cases} \quad (1.1)$$

Thus  $\{X_t\}$  is stationary irrespective of the value of  $\theta$ . The autocorrelation function is:

$$\rho_X(h) = \begin{cases} 1 & h = 0; \\ \frac{\theta}{1+\theta^2} & h = \pm 1; \\ 0 & \text{otherwise.} \end{cases}$$

Note that  $0 \leq |\rho_X(1)| \leq \frac{1}{2}$ . As the correlation between  $X_t$  and  $X_s$  is zero when  $t$  and  $s$  are more than one period apart, we call a moving-average process a process with finite memory or a process with finite-range dependence.

**Remark 1.10.** *To motivate the name moving-average, we can define the MA(1) process more generally as*

$$X_t = \theta_0 Z_t + \theta_1 Z_{t-1} \quad \text{with} \quad Z_t \sim \text{WN}(0, \sigma^2) \quad \text{and} \quad \theta_0 \neq 0.$$

*Thus  $X_t$  is a weighted average of  $Z_t$  and  $Z_{t-1}$ . If  $\theta_0 = \theta_1 = 1/2$ ,  $X_t$  would be arithmetic mean of  $Z_t$  and  $Z_{t-1}$ . This process is, however, equivalent to the process*

$$X_t = \tilde{Z}_t + \tilde{\theta}\tilde{Z}_{t-1} \quad \text{with} \quad \tilde{Z}_t \sim \text{WN}(0, \tilde{\sigma}^2)$$

*where  $\tilde{\theta} = \theta_1/\theta_0$  and  $\tilde{\sigma}^2 = \theta_0^2\sigma^2$ . Both processes would generate the same first two moments and are therefore observationally indistinguishable from each other. Thus, we can set  $\theta_0 = 1$  without loss of generality.*

### 1.4.4 Random walk

Let  $Z_t \sim \text{WN}(0, \sigma^2)$  be a white noise process then the new process

$$X_t = Z_1 + Z_2 + \dots + Z_t = \sum_{j=1}^t Z_j, \quad t > 0, \quad (1.2)$$

is called a *random walk*. Note that, in contrast to  $\{Z_t\}$ ,  $\{X_t\}$  is only defined for  $t > 0$ . The random walk may alternatively be defined through the recursion

$$X_t = X_{t-1} + Z_t, \quad t > 0 \text{ and } X_0 = 0.$$

If in each time period a constant  $\delta$  is added such that

$$X_t = \delta + X_{t-1} + Z_t,$$

the process  $\{X_t\}$  is called a *random walk with drift*.

Although the random walk has a constant mean of zero, it is a nonstationary process.

**Proposition 1.1.** *The random walk  $\{X_t\}$  as defined in equation (1.2) is nonstationary.*

*Proof.* The variance of  $X_{t+1} - X_1$  equals  $\mathbb{V}(X_{t+1} - X_1) = \mathbb{V}\left(\sum_{j=2}^{t+1} Z_j\right) = \sum_{j=2}^{t+1} \mathbb{V}Z_j = t\sigma^2$ .

Assume for the moment that  $\{X_t\}$  is stationary then the triangular inequality implies for  $t > 0$ :

$$0 < \sqrt{t\sigma^2} = \text{std}(X_{t+1} - X_1) \leq \text{std}(X_{t+1}) + \text{std}(X_1) = 2\text{std}(X_1)$$

where “std” denotes the standard deviation. As the left hand side of the inequality converges to infinity for  $t$  going to infinity, also the right hand side must go to infinity. This means that the variance of  $X_1$  must be infinite. This, however, contradicts the assumption of stationarity. Thus  $\{X_t\}$  cannot be stationary.  $\square$

The random walk represents by far the most widely used nonstationary process. It has proven to be an important ingredient in many economic time series. Typical nonstationary time series which are or are driven by random walks are stock market prices, exchange rates, or the gross domestic product (GDP). Usually it is necessary to apply some transformation (filter) first to achieve stationarity. In the example above, one has to replace  $\{X_t\}$  by its first difference  $\{\Delta X_t\} = \{X_t - X_{t-1}\} = \{Z_t\}$  which is stationary by construction. Time series which become stationary after differencing are called

integrated processes and are the subject of a more in depth analysis in Chapter 7. Besides ordinary differencing other transformations are encountered: seasonal differencing, inclusion of a time trend, seasonal dummies, moving averages, etc.

### 1.4.5 Changing mean

Finally, here is another simple example of a nonstationary process.

$$X_t = \begin{cases} Y_t, & t < t_c; \\ Y_t + c, & t \geq t_c \text{ und } c \neq 0 \end{cases}$$

where  $t_c$  is some specific point in time.  $\{X_t\}$  is clearly not stationary because the mean is not constant. In econometrics we refer to such a situation as a *structural change* which can be accommodated by introducing a so-called dummy variable.

## 1.5 Properties of the autocovariance function

The autocovariance function represents the directly accessible external properties of the time series. It is therefore important to understand its properties and how it is related to its inner structure. We will deepen the connection between the autocovariance function and a particular class of models in Chapter 2. The estimation of the autocovariance function will be treated in Chapter 4. For the moment we will just give its properties and analyze the case of the MA(1) model as a prototypical example.

**Theorem 1.1.** *The autocovariance function of a stationary process  $\{X_t\}$  is characterized by the following properties:*

- (i)  $\gamma_X(0) \geq 0$ ;
- (ii)  $0 \leq |\gamma_X(h)| \leq \gamma_X(0)$ ;
- (iii)  $\gamma_X(h) = \gamma_X(-h)$ ;
- (iv)  $\sum_{i,j=1}^n a_i \gamma_X(t_i - t_j) a_j \geq 0$  for all  $n$  and all vectors  $(a_1, \dots, a_n)'$  and  $(t_1, \dots, t_n)$ . This property is called non-negative definiteness.

*Proof.* The first property is obvious as the variance is always nonnegative. The second property follows from the Cauchy-Bunyakovskii-Schwarz inequality (see Theorem C.1) applied to  $X_t$  and  $X_{t+h}$  which yields  $0 \leq |\gamma_X(h)| \leq \gamma_X(0)$ . The third property follows immediately from the definition of the

covariance. Define  $a = (a_1, \dots, a_n)'$  and  $X = (X_{t_1}, \dots, X_{t_n})'$  then the last property follows from the fact that the variance is always nonnegative:  $0 \leq \mathbb{V}(a'X) = a'\mathbb{V}(X)a = \sum_{i,j=1}^n a_i \gamma_X(t_i - t_j) a_j$ .  $\square$

Similar properties hold for the correlation function  $\rho_X$ , except that we have  $\rho_X(0) = 1$ .

**Theorem 1.2.** *The autocorrelation function of a stationary stochastic process  $\{X_t\}$  is characterized by the following properties:*

- (i)  $\rho_X(0) = 1$ ;
- (ii)  $0 \leq |\rho_X(h)| \leq 1$ ;
- (iii)  $\rho_X(h) = \rho_X(-h)$ ;
- (iv)  $\sum_{i,j=1}^n a_i \rho_X(t_i - t_j) a_j \geq 0$  for all  $n$  and all vectors  $(a_1, \dots, a_n)'$  and  $(t_1, \dots, t_n)$ .

*Proof.* The proof follows immediately from the properties of the autocovariance function.  $\square$

It can be shown that for any given function with the above properties there exists a stationary process (Gaussian process) which has this function as its autocovariance function, respectively autocorrelation function.

### 1.5.1 Autocovariance function of a MA(1) process

The autocovariance function describes the external observable characteristics of a time series which can be estimated from the data. Usually, we want to understand the internal mechanism which generates the data at hand. For this we need a model. Hence it is important to understand the relation between the autocovariance function and a certain class of models. In this section, by analyzing the MA(1) model, we will show that this relationship is not one-to-one. Thus we are confronted with a fundamental *identification problem*.

In order to make the point, consider the following given autocovariance function:

$$\gamma(h) = \begin{cases} \gamma_0, & h = 0; \\ \gamma_1, & h = \pm 1; \\ 0, & |h| > 1. \end{cases}$$

The problem now consists of determining the parameters of the MA(1) model,  $\theta$  and  $\sigma^2$  from the values of the autocovariance function. For this purpose we

equate  $\gamma_0 = (1 + \theta^2)\sigma^2$  and  $\gamma_1 = \theta\sigma^2$  (see equation (1.1)). This leads to an equation system in the two unknowns  $\theta$  and  $\sigma^2$ . This system can be simplified by dividing the second equation by the first one to obtain:  $\gamma_1/\gamma_0 = \theta/(1 + \theta^2)$ . Because  $\gamma_1/\gamma_0 = \rho(1) = \rho_1$  one gets a quadratic equation in  $\theta$ :

$$\rho_1\theta^2 - \theta + \rho_1 = 0.$$

The two solutions of this equation are

$$\theta = \frac{1}{2\rho_1} \left( 1 \pm \sqrt{1 - 4\rho_1^2} \right).$$

The solutions are real if and only if the discriminant  $1 - 4\rho_1^2$  is positive. This is the case if and only if  $\rho_1^2 \leq 1/4$ , respectively  $|\rho_1| \leq 1/2$ . Note that one root is the inverse of the other. The identification problem thus takes the following form:

$|\rho_1| < 1/2$ : there exists two observationally equivalent MA(1) processes.

$\rho_1 = \pm 1/2$ : there exists exactly one MA(1) process with  $\theta = \pm 1$ .

$|\rho_1| > 1/2$ : there exists no MA(1) process with this autocovariance function.

The relation between the first order autocorrelation coefficient,  $\rho_1 = \rho(1)$ , and the parameter  $\theta$  of the MA(1) process is represented in Figure 1.10. As can be seen, there exists for each  $\rho(1)$  with  $|\rho(1)| < \frac{1}{2}$  two solutions. The two solutions are inverses of each other. Hence one solution is absolutely smaller than one whereas the other is bigger than one. In Section 2.3 we will argue in favor of the solution smaller than one. For  $\rho(1) = \pm 1/2$  there exists exactly one solution, namely  $\theta = \pm 1$ . For  $|\rho(1)| > 1/2$  there is no solution. For  $|\rho_1| > 1/2$ ,  $\rho(h)$  actually does not represent a genuine autocorrelation function as the fourth condition in Theorem 1.1, respectively Theorem 1.2 is violated. Taking  $\rho_1 = \frac{1}{2}$  and setting  $a = (1, -1, 1, -1, \dots, 1, -1)'$  one gets:

$$\sum_{i,j=1}^n a_i \rho(i-j) a_j = n - 2(n-1)\rho_1 < 0, \quad \text{if } n > \frac{2\rho_1}{2\rho_1 - 1}.$$

For  $\rho_1 = -\frac{1}{2}$  one sets  $a = (1, 1, \dots, 1)'$ . Hence the fourth property is violated.

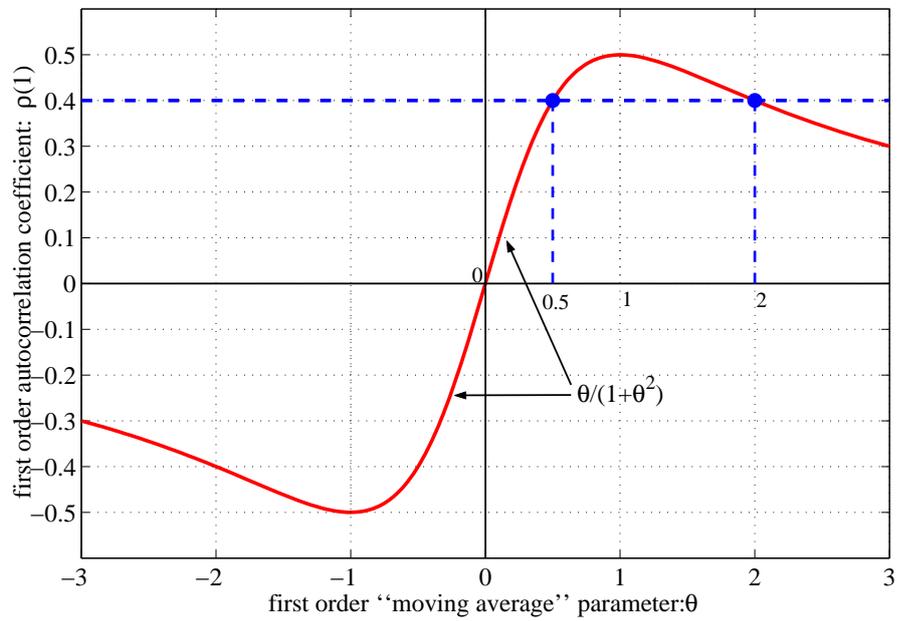


Figure 1.10: Relation between the autocorrelation coefficient of order one,  $\rho(1)$ , and the parameter  $\theta$  of a MA(1) process

## 1.6 Exercises

**Exercise 1.6.1.** Let the process  $\{X_t\}$  be generated by a two-sided moving-average process

$$X_t = 0.5Z_{t+1} + 0.5Z_{t-1} \quad \text{with } Z_t \sim \text{WN}(0, \sigma^2).$$

Determine the autocovariance and the autocorrelation function of  $\{X_t\}$ .

**Exercise 1.6.2.** Let  $\{X_t\}$  be the MA(1) process

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

- (i) Determine the autocovariance and the autocorrelation function of  $\{X_t\}$  for  $\theta = 0.9$ .
- (ii) Determine the variance of the mean  $(X_1 + X_2 + X_3 + X_4)/4$ .
- (iii) How do the previous results change if  $\theta = -0.9$ ?

**Exercise 1.6.3.** Consider the autocovariance function

$$\gamma(h) = \begin{cases} 4, & h = 0; \\ -2, & h = \pm 1; \\ 0, & \text{otherwise.} \end{cases}$$

Determine the parameters  $\theta$  and  $\sigma^2$ , if they exist, of the first order moving-average process  $X_t = Z_t + \theta Z_{t-1}$  with  $Z_t \sim \text{WN}(0, \sigma^2)$  such that autocovariance function above is the autocovariance function corresponding to  $\{X_t\}$ .

**Exercise 1.6.4.** Let the stochastic process  $\{X_t\}$  be defined as

$$\begin{cases} Z_t, & \text{if } t \text{ is even;} \\ (Z_{t-1}^2 - 1)/\sqrt{2}, & \text{if } t \text{ is uneven,} \end{cases}$$

where  $\{Z_t\}$  is identically and independently distributed as  $Z_t \sim \text{N}(0, 1)$ . Show that  $\{X_t\} \sim \text{WN}(0, 1)$ , but not IID(0, 1).

**Exercise 1.6.5.** Which of the following processes is stationary?

- (i)  $X_t = Z_t + \theta Z_{t-1}$
- (ii)  $X_t = Z_t Z_{t-1}$
- (iii)  $X_t = a + \theta Z_0$
- (iv)  $X_t = Z_0 \sin(at)$

In all cases we assume that  $\{Z_t\}$  is identically and independently distributed with  $Z_t \sim \text{N}(0, \sigma^2)$ .  $\theta$  and  $a$  are arbitrary parameters.



# Chapter 2

## Autoregressive Moving-Average Models

A basic idea in time series analysis is to construct more complex processes from simple ones. In the previous chapter we showed how the averaging of a white noise process leads to a process with first order autocorrelation. In this chapter we generalize this idea and consider processes which are solutions of linear stochastic difference equations. These so-called *ARMA processes* constitute the most widely used class of models for stationary processes.

**Definition 2.1** (ARMA Models). *A stochastic process  $\{X_t\}$  with  $t \in \mathbb{Z}$  is called an autoregressive moving-average process (ARMA process) of order  $(p, q)$ , denoted by  $ARMA(p, q)$  process if the process is stationary and satisfies a linear stochastic difference equation of the form*

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q} \quad (2.1)$$

with  $Z_t \sim \text{WN}(0, \sigma^2)$  and  $\phi_p \theta_q \neq 0$ .  $\{X_t\}$  is called an  $ARMA(p, q)$  process with mean  $\mu$  if  $\{X_t - \mu\}$  is an  $ARMA(p, q)$  process.

The importance of ARMA processes is due to the fact that every stationary process can be approximated arbitrarily well by an ARMA process. In particular, it can be shown that for any given autocovariance function  $\gamma$  with the property  $\lim_{h \rightarrow \infty} \gamma(h) = 0$  and any positive integer  $k$  there exists an autoregressive moving-average process (ARMA process)  $\{X_t\}$  such that  $\gamma_X(h) = \gamma(h)$ ,  $h = 0, 1, \dots, k$ .

For an ARMA process with mean  $\mu$  one often adds a constant  $c$  to the right hand side of the difference equation:

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = c + Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}.$$

The mean of  $X_t$  is then:  $\mu = \frac{c}{1-\phi_1-\dots-\phi_p}$ . The mean is therefore only well-defined if  $\phi_1 + \dots + \phi_p \neq 1$ . The case  $\phi_1 + \dots + \phi_p = 1$  can, however, be excluded because there exists no stationary solution in this case (see Remark 2.2) and thus no ARMA process.

## 2.1 The lag operator

In times series analysis it customary to rewrite the above difference equation more compactly in terms of the *lag operator*  $L$ . This is, however, not only a compact notation, but will open the way to analyze the inner structure of ARMA processes. The lag or back-shift operator  $L$  moves the time index one period back:

$$L\{X_t\} = \{X_{t-1}\}.$$

For ease of notation we write:  $LX_t = X_{t-1}$ . The lag operator is a linear operator with the following calculation rules:

- (i)  $L$  applied to the process  $\{X_t = c\}$  where  $c$  is an arbitrary constant gives:

$$Lc = c.$$

- (ii) Applying  $L$   $n$  times:

$$\underbrace{L \dots L}_{n \text{ times}} X_t = L^n X_t = X_{t-n}.$$

- (iii) The inverse of the lag operator is the lead or forward operator. This operator shifts the time index one period into the future.<sup>1</sup> We can write  $L^{-1}$ :

$$L^{-1}X_t = X_{t+1}.$$

- (iv) As  $L^{-1}LX_t = X_t$  we have that

$$L^0 = \mathbf{1}.$$

- (v) For any integers  $m$  and  $n$  we have:

$$L^m L^n X_t = L^{m+n} X_t = X_{t-m-n}.$$

---

<sup>1</sup>One technical advantage of using the double-infinite index set  $\mathbb{Z}$  is that the lag operators form a group.

- (vi) For any real numbers  $a$  and  $b$ , any integers  $m$  and  $n$ , and arbitrary stochastic processes  $\{X_t\}$  and  $\{Y_t\}$  we have:

$$(aL^m + bL^n)(X_t + Y_t) = aX_{t-m} + bX_{t-n} + aY_{t-m} + bY_{t-n}.$$

In this way it is possible to define *lag polynomials*:  $A(L) = a_0 + a_1L + a_2L^2 + \dots + a_pL^p$  where  $a_0, a_1, \dots, a_p$  are any real numbers. For these polynomials the usual calculation rules apply. Let, for example,  $A(L) = 1 - 0.5L$  and  $B(L) = 1 + 4L^2$  then  $C(L) = A(L)B(L) = 1 - 0.5L + 4L^2 - 2L^3$ .

The case of the stochastic difference equation we get an autoregressive and a moving-average polynomial as follows:

$$\begin{aligned}\Phi(L) &= 1 - \phi_1L - \dots - \phi_pL^p, \\ \Theta(L) &= 1 + \theta_1L + \dots + \theta_qL^q.\end{aligned}$$

The stochastic difference equation defining the ARMA process can then be written compactly as

$$\Phi(L)X_t = \Theta(L)Z_t.$$

Thus, the use of lag polynomials provides a compact notation for ARMA processes. Moreover and most importantly,  $\Phi(z)$  and  $\Theta(z)$ , viewed as polynomials of the complex number  $z$ , also reveal much of their inherent structural properties as will become clear in Section 2.3.

## 2.2 Some important special cases

Before we deal with the general theory of ARMA processes, we will analyze some important special cases first:

$q = 0$ : autoregressive process of order  $p$ , AR( $p$ ) process

$p = 0$ : moving-average process of order  $q$ , MA( $q$ ) process

### 2.2.1 The Moving-average process of order $q$ (MA( $q$ ) process)

The MA( $q$ ) process is defined by the following stochastic difference equation:

$$X_t = \Theta(L)Z_t = \theta_0Z_t + \theta_1Z_{t-1} + \dots + \theta_qZ_{t-q} \quad \text{with } \theta_0 = 1 \text{ and } \theta_q \neq 0$$

and  $Z_t \sim \text{WN}(0, \sigma^2)$ . Obviously,

$$\mathbb{E}X_t = \mathbb{E}Z_t + \theta_1\mathbb{E}Z_{t-1} + \dots + \theta_q\mathbb{E}Z_{t-q} = 0,$$

because  $Z_t \sim \text{WN}(0, \sigma^2)$ . As can be easily verified using the properties of  $\{Z_t\}$ , the autocovariance function of the MA( $q$ ) processes are:

$$\gamma_X(h) = \text{cov}(X_{t+h}, X_t) = \begin{cases} \sigma^2 \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|}, & |h| \leq q; \\ 0, & |h| > q. \end{cases}$$

This implies the following autocorrelation function:

$$\rho_X(h) = \text{corr}(X_{t+h}, X_t) = \begin{cases} \frac{1}{\sum_{i=0}^q \theta_i^2} \sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|}, & |h| \leq q; \\ 0, & |h| > q. \end{cases}$$

Every MA( $q$ ) process is therefore stationary irrespective of its parameters  $\theta_0, \theta_1, \dots, \theta_q$ . Because the correlation between  $X_t$  and  $X_s$  is equal to zero if the two time points  $t$  and  $s$  are more than  $q$  periods apart, such processes are sometimes called processes with *short memory* or processes with *short range dependence*.

Figure 2.1 displays an MA(1) process and its autocorrelation function.

### 2.2.2 The first order autoregressive process (AR(1) process)

The AR( $p$ ) process requires a more thorough analysis as will already become clear from the AR(1) process. This process is defined by the following stochastic difference equation:

$$X_t = \phi X_{t-1} + Z_t, \quad Z_t \sim \text{WN}(0, \sigma^2) \text{ and } \phi \neq 0$$

The above stochastic difference equation has in general several solutions. Given a sequence  $\{Z_t\}$  and an arbitrary distribution for  $X_0$ , it determines all random variables  $X_t$ ,  $t \in \mathbb{Z} \setminus \{0\}$ , by applying the above recursion. The solutions are, however, not necessarily stationary. But, according to the Definition 2.1, only stationary processes qualify for ARMA processes. As we will demonstrate, depending on the value of  $\phi$ , there may exist no or just one solution.

Consider first the case of  $|\phi| < 1$ . Inserting into the difference equation several times leads to:

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t = \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\ &= \dots \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \dots + \phi^k Z_{t-k} + \phi^{k+1} X_{t-k-1}. \end{aligned}$$

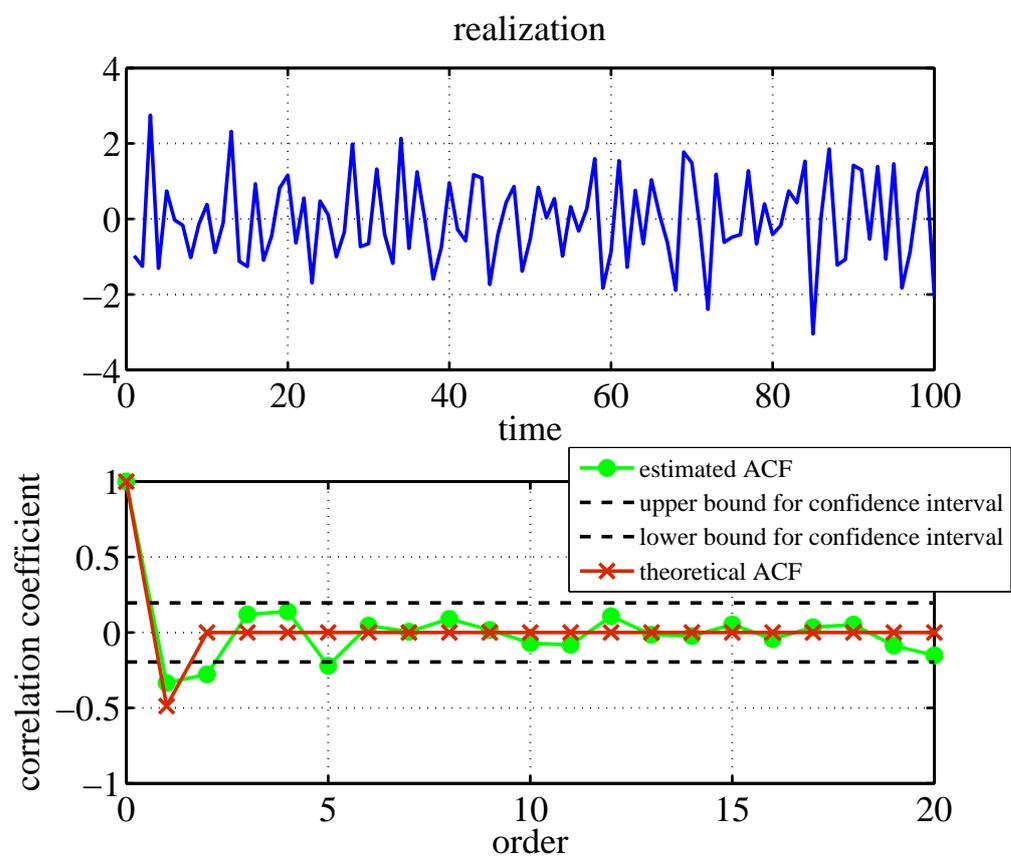


Figure 2.1: Realization and estimated ACF of a MA(1) process:  $X_t = Z_t - 0.8Z_{t-1}$  with  $Z_t \sim \text{IID } N(0, 1)$

If  $\{X_t\}$  is a stationary solution,  $\mathbb{V}X_{t-k-1}$  remains constant independently of  $k$ . Thus

$$\mathbb{V}\left(X_t - \sum_{j=0}^k \phi^j Z_{t-j}\right) = \phi^{2k+2} \mathbb{V}X_{t-k-1} \rightarrow 0 \text{ for } k \rightarrow \infty.$$

This shows that  $\sum_{j=0}^k \phi^j Z_{t-j}$  converges in the mean square sense, and thus also in probability, to  $X_t$  for  $k \rightarrow \infty$  (see Theorem C.8 in Appendix C). This suggests to take

$$X_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \dots = \sum_{j=0}^{\infty} \phi^j Z_{t-j} \quad (2.2)$$

as the solution to the stochastic difference equation. As  $\sum_{j=0}^{\infty} |\phi^j| = \frac{1}{1-\phi} < \infty$  this solution is well-defined according to Theorem 6.4 and has the following properties:

$$\begin{aligned} \mathbb{E}X_t &= \sum_{j=0}^{\infty} \phi^j \mathbb{E}Z_{t-j} = 0, \\ \gamma_X(h) &= \text{cov}(X_{t+h}, X_t) = \lim_{k \rightarrow \infty} \mathbb{E}\left(\sum_{j=0}^k \phi^j Z_{t+h-j}\right) \left(\sum_{j=0}^k \phi^j Z_{t-j}\right) \\ &= \sigma^2 \phi^{|h|} \sum_{j=0}^{\infty} \phi^{2j} = \frac{\phi^{|h|}}{1-\phi^2} \sigma^2, \quad h \in \mathbb{Z}, \\ \rho_X(h) &= \phi^{|h|}. \end{aligned}$$

Thus the solution  $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$  is stationary and fulfills the difference equation as can be easily verified. It is also the only stationary solution which is compatible with the difference equation. Assume that there is second solution  $\{\tilde{X}_t\}$  with these properties. Inserting into the difference equation yields again

$$\mathbb{V}\left(\tilde{X}_t - \sum_{j=0}^k \phi^j Z_{t-j}\right) = \phi^{2k+2} \mathbb{V}\tilde{X}_{t-k-1}.$$

This variance converges to zero for  $k$  going to infinity because  $|\phi| < 1$  and because  $\{\tilde{X}_t\}$  is stationary. The two processes  $\{\tilde{X}_t\}$  and  $\{X_t\}$  with  $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$  are therefore identical in the mean square sense and thus with probability one. Figure 2.2 shows a realization of such a process and its estimated autocorrelation function.

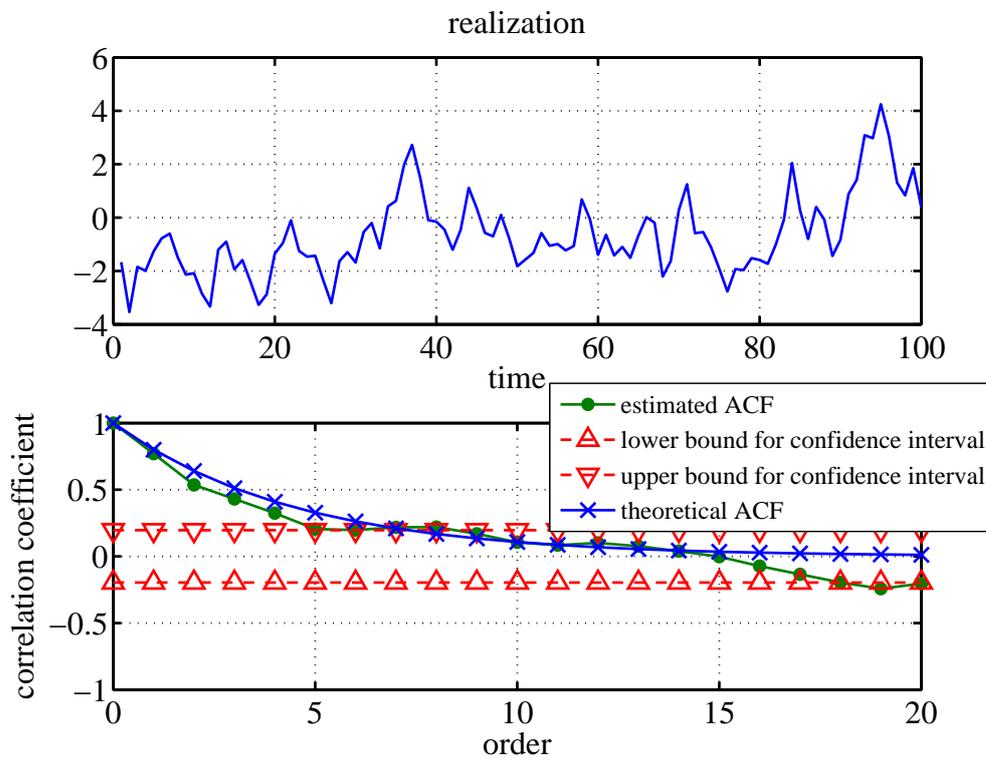


Figure 2.2: Realization and estimated ACF of an AR(1) process:  $X_t = 0.8X_{t-1} + Z_t$  with  $Z_t \sim \text{IIN}(0, 1)$

In the case  $|\phi| > 1$  the solution (2.2) does not converge. It is, however, possible to iterate the difference equation forward to obtain:

$$\begin{aligned} X_t &= \phi^{-1}X_{t+1} - \phi^{-1}Z_{t+1} \\ &= \phi^{-k-1}X_{t+k+1} - \phi^{-1}Z_{t+1} - \phi^{-2}Z_{t+2} - \dots - \phi^{-k-1}Z_{t+k+1}. \end{aligned}$$

This suggests to take

$$X_t = - \sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$$

as the solution. Going through similar arguments as before it is possible to show that this is indeed the only stationary solution. This solution is, however, viewed to be inadequate because  $X_t$  depends on contemporaneous and future shocks  $Z_{t+j}, j = 0, 1, \dots$ . Note, however, that there exists an AR(1) process with  $|\phi| < 1$  which is observationally equivalent, in the sense that it generates the same autocorrelation function, but with new shock or forcing  $\{\tilde{Z}_t\}$  (see next section).

In the case  $|\phi| = 1$  there exists no stationary solution (see Section 1.4.4) and therefore, according to our definition, no ARMA process. Processes with this property are called random walk, unit root processes or integrated processes. They play an important role in economics and are treated separately in Chapter 7.

## 2.3 Causality and invertibility

If we view  $\{X_t\}$  as the state variable and  $\{Z_t\}$  as an impulse or shock, we can ask whether it is possible to represent today's state  $X_t$  as the outcome of current and past shocks  $Z_t, Z_{t-1}, Z_{t-2}, \dots$ . In this case we can view  $X_t$  as being *caused* by past shocks and call this a *causal representation*. Thus, shocks to current  $Z_t$  will not only influence current  $X_t$ , but will propagate to affect also future  $X_t$ 's. This notion of causality rest on the assumption that the past can cause the future but that the future cannot cause the past. See section 15.1 for an elaboration of the concept of causality and its generalization to the multivariate context.

In the case that  $\{X_t\}$  is a moving-average process of order  $q$ ,  $X_t$  is given as a weighted sum of current and past shocks  $Z_t, Z_{t-1}, \dots, Z_{t-q}$ . Thus, the moving-average representation is already the causal representation. In the case of an AR(1) process, we have seen that this is not always feasible. For  $|\phi| < 1$ , the solution (2.2) represents  $X_t$  as a weighted sum of current and past shocks and is thus the corresponding causal representation. For  $|\phi| > 1$ , no such representation is possible. The following Definition 2.2 makes the

notion of a causal representation precise and Theorem 2.1 gives a general condition for its existence.

**Definition 2.2** (Causality). *An ARMA( $p, q$ ) process  $\{X_t\}$  with  $\Phi(L)X_t = \Theta(L)Z_t$  is called causal with respect to  $\{Z_t\}$  if there exists a sequence  $\{\psi_j\}$  with the property  $\sum_{j=0}^{\infty} |\psi_j| < \infty$  such that*

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \dots = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \quad \text{with } \psi_0 = 1.$$

The above equation is referred to as the causal representation of  $\{X_t\}$  with respect to  $\{Z_t\}$ .

The coefficients  $\{\psi_j\}$  are of great importance because they determine how an impulse or a shock in period  $t$  propagates to affect current and future  $X_{t+j}$ ,  $j = 0, 1, 2, \dots$ . In particular, consider an impulse  $e_{t_0}$  at time  $t_0$ , i.e. a time series which is equal to zero except for the time  $t_0$  where it takes on the value  $e_{t_0}$ . Then,  $\{\psi_{t-t_0} e_{t_0}\}$  traces out the time history of this impulse. For this reason, the function  $\psi_j$  with  $j = t - t_0$ ,  $t = t_0, t_0 + 1, t_0 + 2, \dots$ , is called the *impulse response function*. If  $e_{t_0} = 1$ , it is called a unit impulse. Alternatively,  $e_{t_0}$  is sometimes taken to be equal to  $\sigma$ , the standard deviation of  $Z_t$ . It is customary to plot  $\psi_j$  as a function of  $j$ ,  $j = 0, 1, 2, \dots$

Note that the notion of causality is not an attribute of  $\{X_t\}$ , but is defined relative to another process  $\{Z_t\}$ . It is therefore possible that a stationary process is causal with respect to one process, but not with respect to another process. In order to make this point more concrete, consider again the AR(1) process defined by the equation  $X_t = \phi X_{t-1} + Z_t$  with  $|\phi| > 1$ . As we have seen, the only stationary solution is given by  $X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$  which is clearly not causal with respect  $\{Z_t\}$ . Consider now the process

$$\tilde{Z}_t = X_t - \frac{1}{\phi} X_{t-1} = \phi^{-2} Z_t + (\phi^{-2} - 1) \sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}.$$

This new process is white noise with variance  $\tilde{\sigma}^2 = (1 - \phi^{-2} + \phi^{-4})\sigma^2$ .<sup>2</sup> Because  $\{X_t\}$  fulfills the difference equation

$$X_t = \frac{1}{\phi} X_{t-1} + \tilde{Z}_t,$$

$\{X_t\}$  is causal with respect to  $\{\tilde{Z}_t\}$ . This remark shows that there is no loss of generality involved if we concentrate on causal ARMA processes.

<sup>2</sup>The reader is invited to verify this.

**Theorem 2.1.** *Let  $\{X_t\}$  be an ARMA( $p, q$ ) process with  $\Phi(L)X_t = \Theta(L)Z_t$  and assume that the polynomials  $\Phi(z)$  and  $\Theta(z)$  have no common root.  $\{X_t\}$  is causal with respect to  $\{Z_t\}$  if and only if  $\Phi(z) \neq 0$  for  $|z| \leq 1$ , i.e. all roots of the equation  $\Phi(z) = 0$  are outside the unit circle. The coefficients  $\{\psi_j\}$  are then uniquely defined by identity :*

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)}.$$

*Proof.* Given that  $\Phi(z)$  is a finite order polynomial with  $\Phi(z) \neq 0$  for  $|z| \leq 1$ , there exists  $\epsilon > 0$  such that  $\Phi(z) \neq 0$  for  $|z| \leq 1 + \epsilon$ . This implies that  $1/\Phi(z)$  is an analytic function on the circle with radius  $1 + \epsilon$  and therefore possesses a power series expansion:

$$\frac{1}{\Phi(z)} = \sum_{j=0}^{\infty} \xi_j z^j = \Xi(z), \quad \text{for } |z| < 1 + \epsilon.$$

This implies that  $\xi_j(1 + \epsilon/2)^j$  goes to zero for  $j$  to infinity. Thus there exists a positive and finite constant  $C$  such that

$$|\xi_j| < C(1 + \epsilon/2)^{-j}, \quad \text{for all } j = 0, 1, 2, \dots$$

This in turn implies that  $\sum_{j=0}^{\infty} |\xi_j| < \infty$  and that  $\Xi(z)\Phi(z) = 1$  for  $|z| \leq 1$ . Applying  $\Xi(L)$  on both sides of  $\Phi(L)X_t = \Theta(L)Z_t$ , gives:

$$X_t = \Xi(L)\Phi(L)X_t = \Xi(L)\Theta(L)Z_t.$$

Theorem 6.4 implies that the right hand side is well-defined. Thus  $\Psi(L) = \Xi(L)\Theta(L)$  is the sought polynomial. Its coefficients are determined by the relation  $\Psi(z) = \Theta(z)/\Phi(z)$ .

Assume now that there exists a causal representation  $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$  with  $\sum_{j=0}^{\infty} |\psi_j| < \infty$ . Therefore

$$\Theta(L)Z_t = \Phi(L)X_t = \Phi(L)\Psi(L)Z_t.$$

Take  $\eta(z) = \Phi(z)\Psi(z) = \sum_{j=0}^{\infty} \eta_j z^j$ ,  $|z| \leq 1$ . Multiplying the above equation by  $Z_{t-k}$  and taking expectations shows that  $\eta_k = \theta_k$ ,  $k = 0, 1, 2, \dots, q$ , and that  $\eta_k = 0$  for  $k > q$ . Thus we get  $\Theta(z) = \eta(z) = \Phi(z)\Psi(z)$  for  $|z| \leq 1$ . As  $\Theta(z)$  and  $\Phi(z)$  have no common roots and because  $|\Psi(z)| < \infty$  for  $|z| \leq 1$ ,  $\Phi(z)$  cannot be equal to zero for  $|z| \leq 1$ .  $\square$

**Remark 2.1.** *If the AR and the MA polynomial have common roots, there are two possibilities:*

- *No common roots lies on the unit circle. In this situation there exists a unique stationary solution which can be obtained by canceling the common factors of the polynomials.*
- *If at least one common root lies on the unit circle then more than one stationary solution may exist (see the last example below).*

### Some examples

We concretize the above Theorem and Remark by investigating some examples starting from the ARMA model  $\Phi(L)X_t = \Theta(L)Z_t$  with  $Z_t \sim \text{WN}(0, \sigma^2)$ .

$\Phi(L) = 1 - 0.05L - 0.6L^2$  **and**  $\Theta(L) = 1$ : The roots of the polynomial  $\Phi(z)$  are  $z_1 = -4/3$  and  $z_2 = 5/4$ . Because both roots are absolutely greater than one, there exists a causal representation with respect to  $\{Z_t\}$ .

$\Phi(L) = 1 + 2L + 5/4L^2$  **and**  $\Theta(L) = 1$ : The roots are conjugate complex and equal to  $z_1 = -4/5 + 2/5i$  and  $z_2 = -4/5 - 2/5i$ . The modulus or absolute value of  $z_1$  and  $z_2$  equals  $|z_1| = |z_2| = \sqrt{20/25}$ . This number is smaller than one. Therefore there exists a stationary solution, but this solution is not causal with respect to  $\{Z_t\}$ .

$\Phi(L) = 1 - 0.05L - 0.6L^2$  **and**  $\Theta(L) = 1 + 0.75L$ :  $\Phi(z)$  and  $\Theta(z)$  have the common root  $z = -4/3 \neq 1$ . Thus one can cancel both  $\Phi(L)$  and  $\Theta(L)$  by  $1 + \frac{3}{4}L$  to obtain the polynomials  $\tilde{\Phi}(L) = 1 - 0.8L$  and  $\tilde{\Theta}(L) = 1$ . Because the root of  $\tilde{\Phi}(z)$  equals  $5/4$  which is greater than one, there exists a unique stationary and causal representation with respect to  $\{Z_t\}$ .

$\Phi(L) = 1 + 1.2L - 1.6L^2$  **and**  $\Theta(L) = 1 + 2L$ : The roots of  $\Phi(z)$  are  $z_1 = 5/4$  and  $z_2 = -0.5$ . Thus one root is outside the unit circle whereas one is inside. This would suggest that there is no causal solution. However, the root  $-0.5 \neq 1$  is shared by  $\Phi(z)$  and  $\Theta(z)$  and can therefore be canceled to obtain  $\tilde{\Phi}(L) = 1 - 0.8L$  and  $\tilde{\Theta}(L) = 1$ . Because the root of  $\tilde{\Phi}(z)$  equals  $5/4 > 1$ , there exists a unique stationary and causal solution with respect to  $\{Z_t\}$ .

$\Phi(L) = 1 + L$  **and**  $\Theta(L) = 1 + L$ :  $\Phi(z)$  and  $\Theta(z)$  have the common root  $-1$  which lies on the unit circle. As before one might cancel both polynomials by  $1 + L$  to obtain the trivial stationary and causal solution  $\{X_t\} = \{Z_t\}$ . This is, however, not the only solution. Additional solutions are given by  $\{Y_t\} = \{Z_t + A(-1)^t\}$  where  $A$  is an arbitrary

random variable with mean zero and finite variance  $\sigma_A^2$  which is independent from both  $\{X_t\}$  and  $\{Z_t\}$ . The process  $\{Y_t\}$  has a mean of zero and an autocovariance function  $\gamma_Y(h)$  which is equal to

$$\gamma_Y(h) = \begin{cases} \sigma^2 + \sigma_A^2, & h = 0; \\ (-1)^h \sigma_A^2, & h = \pm 1, \pm 2, \dots \end{cases}$$

Thus this new process is therefore stationary and fulfills the difference equation.

**Remark 2.2.** *If the AR and the MA polynomial in the stochastic difference equation  $\Phi(L)X_t = \Theta(L)Z_t$  have no common root, but  $\Phi(z) = 0$  for some  $z$  on the unit circle, there exists no stationary solution. In this sense the stochastic difference equation does no longer define an ARMA model. Models with this property are said to have a unit root and are treated in Chapter 7. If  $\Phi(z)$  has no root on the unit circle, there exists a unique stationary solution.*

As explained in the previous Theorem, the coefficients  $\{\psi_j\}$  of the causal representation are uniquely determined by the relation  $\Psi(z)\Phi(z) = \Theta(z)$ . If  $\{X_t\}$  is a MA process,  $\Phi(z) = 1$  and the coefficients  $\{\psi_j\}$  just correspond to the coefficients of the MA polynomial, i.e.  $\psi_j = \theta_j$  for  $0 \leq j \leq q$  and  $\psi_j = 0$  for  $j > q$ . Thus in this case no additional computations are necessary. In general this is not the case. In principle there are two ways to find the coefficients  $\{\psi_j\}$ . The first one uses polynomial division or partial fractions, the second one uses the method of undetermined coefficients. This book relies on the second method because it is more intuitive and presents some additional insides. For this purpose let us write out the defining relation  $\Psi(z)\Phi(z) = \Theta(z)$ :

$$(\psi_0 + \psi_1 z + \psi_2 z^2 + \dots) (1 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p) = 1 + \theta_1 z + \theta_2 z^2 + \dots + \theta_q z^q$$

Multiplying out the left hand side one gets:

$$\begin{aligned} & \psi_0 - \psi_0 \phi_1 z - \psi_0 \phi_2 z^2 - \psi_0 \phi_3 z^3 - \dots - \psi_0 \phi_p z^p \\ & \quad \psi_1 z - \psi_1 \phi_1 z^2 - \psi_1 \phi_2 z^3 - \dots - \psi_1 \phi_p z^{p+1} \\ & \quad + \psi_2 z^2 - \psi_2 \phi_1 z^3 - \dots - \psi_2 \phi_p z^{p+2} \\ & \quad \dots \\ & = 1 + \theta_1 z + \theta_2 z^2 + \theta_3 z^3 + \dots + \theta_q z^q \end{aligned}$$

Equating the coefficients of the powers of  $z$ ,  $z^j$ ,  $j = 0, 1, 2, \dots$ , one obtains the following equations:

$$\begin{aligned} z^0 : \quad & \psi_0 = 1, \\ z^1 : \quad & \psi_1 = \theta_1 + \phi_1\psi_0 = \theta_1 + \phi_1, \\ z^2 : \quad & \psi_2 = \theta_2 + \phi_2\psi_0 + \phi_1\psi_1 = \theta_2 + \phi_2 + \phi_1\theta_1 + \phi_1^2, \\ & \dots \end{aligned}$$

As can be seen, it is possible to solve recursively for the unknown coefficients  $\{\psi_j\}$ . This is convenient when it comes to numerical computations, but in some cases one wants an analytical solution. Such a solution can be obtained by observing that, for  $j \geq \max\{p, q + 1\}$ , the recursion leads to the following difference equation of order  $p$ :

$$\psi_j = \sum_{k=1}^p \phi_k \psi_{j-k} = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2} + \dots + \phi_p \psi_{j-p}, \quad j \geq \max\{p, q + 1\}.$$

This is a linear homogeneous difference equation with constant coefficients. The solution of such an equation is of the form (see equation (B.1) in Appendix B):

$$\psi_j = c_1 z_1^{-j} + \dots + c_p z_p^{-j}, \quad j \geq \max\{p, q + 1\} - p, \quad (2.3)$$

where  $z_1, \dots, z_p$  denote the roots of  $\Phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p = 0$ .<sup>3</sup> Note that the roots are exactly those which have been computed to assess the existence of a causal representation. The coefficients  $c_1, \dots, c_p$  can be obtained using the  $p$  boundary conditions obtained from  $\psi_j = \sum_{0 < k \leq j} \phi_k \psi_{j-k} = \theta_j$ ,  $\max\{p, q + 1\} - p \leq j < \max\{p, q + 1\}$ . Finally, the values for  $\psi_j$ ,  $0 \leq j < \max\{p, q + 1\} - p$ , must be computed from the first  $\max\{p, q + 1\} - p$  iterations (see the example in Section 2.4).

As mentioned previously, the coefficients  $\{\psi_j\}$  are of great importance as they quantify the effect of a shock to  $Z_{t-j}$  on  $X_t$ , respectively of  $Z_t$  on  $X_{t+j}$ . In macroeconomics they are sometimes called *dynamic multipliers* of a transitory or temporary shock. Because the underlying ARMA process is stationary and causal, the infinite sum  $\sum_{j=0}^{\infty} |\psi_j|$  converges. This implies that the effect  $\psi_j$  converges to zero as  $j \rightarrow \infty$ . Thus the effect of a shock dies out eventually:<sup>4</sup>

$$\frac{\partial X_{t+j}}{\partial Z_t} = \psi_j \rightarrow 0 \text{ for } j \rightarrow \infty.$$

<sup>3</sup>In the case of multiple roots one has to modify the formula according to equation (B.2).

<sup>4</sup>The use of the partial derivative sign actually represents an abuse of notation. It is inspired by an alternative definition of the impulse responses:  $\psi_j = \frac{\partial \bar{\mathbb{P}}_t X_{t+j}}{\partial x_t}$  where

As can be seen from equation (2.3), the coefficients  $\{\psi_j\}$  even converge to zero exponentially fast to zero because each term  $c_i z_i^{-j}$ ,  $i = 1, \dots, p$ , goes to zero exponentially fast as the roots  $z_i$  are greater than one in absolute value. Viewing  $\{\psi_j\}$  as a function of  $j$  one gets the so-called *impulse response function* which is usually displayed graphically.

The effect of a *permanent shock* in period  $t$  on  $X_{t+j}$  is defined as the cumulative effect of all the transitory shocks. Thus the effect of a permanent shock to  $X_{t+j}$  is given by  $\sum_{i=0}^j \psi_i$ . Because  $\sum_{i=0}^j \psi_i \leq \sum_{i=0}^j |\psi_i| \leq \sum_{i=0}^{\infty} |\psi_i| < \infty$ , the cumulative effect remains finite.

In time series analysis we view the observations as realizations of  $\{X_t\}$  and treat the realizations of  $\{Z_t\}$  as unobserved. It is therefore of interest to know whether it is possible to recover the unobserved shocks from the observations on  $\{X_t\}$ . This idea leads to the concept of invertibility.

**Definition 2.3** (Invertibility). *An ARMA( $p, q$ ) process for  $\{X_t\}$  with  $\Phi(L)X_t = \Theta(L)Z_t$  is called invertible with respect to  $\{Z_t\}$  if and only if there exists a sequence  $\{\pi_j\}$  with the property  $\sum_{j=0}^{\infty} |\pi_j| < \infty$  such that*

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

Note that like causality, invertibility is not an attribute of  $\{X_t\}$ , but is defined only relative to another process  $\{Z_t\}$ . In the literature, one often refers to invertibility as the strict miniphase property.<sup>5</sup>

**Theorem 2.2.** *Let  $\{X_t\}$  be an ARMA( $p, q$ ) process with  $\Phi(L)X_t = \Theta(L)Z_t$  such that polynomials  $\Phi(z)$  and  $\Theta(z)$  have no common roots. Then  $\{X_t\}$  is invertible with respect to  $\{Z_t\}$  if and only if  $\Theta(z) \neq 0$  for  $|z| \leq 1$ . The coefficients  $\{\pi_j\}$  are then uniquely determined through the relation:*

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\Phi(z)}{\Theta(z)}.$$

*Proof.* The proof follows from Theorem 2.1 with  $X_t$  and  $Z_t$  interchanged.  $\square$

---

$\tilde{\mathbb{P}}_t$  denotes the optimal (in the mean squared error sense) linear predictor of  $X_{t+j}$  given a realization back to infinite remote past  $\{x_t, x_{t-1}, x_{t-2}, \dots\}$  (see Section 3.1.3). Thus,  $\psi_j$  represents the sensitivity of the forecast of  $X_{t+j}$  with respect to the observation  $x_t$ . The equivalence of alternative definitions in the linear and especially nonlinear context is discussed in Potter (2000).

<sup>5</sup>Without the qualification strict, the miniphase property allows for roots of  $\Theta(z)$  on the unit circle. The terminology is, however, not uniform in the literature.

The discussion in Section 1.3 showed that there are in general two MA(1) processes compatible with the same autocorrelation function  $\rho(h)$  given by  $\rho(0) = 1$ ,  $\rho(1) = \rho$  with  $|\rho| \leq \frac{1}{2}$ , and  $\rho(h) = 0$  for  $h \geq 2$ . However, only one of these solutions is invertible because the two solutions for  $\theta$  are inverses of each other. As it is important to be able to recover  $Z_t$  from current and past  $X_t$ , one prefers the invertible solution.

**Remark 2.3.** *If  $\{X_t\}$  is a stationary solution to the stochastic difference equation  $\Phi(L)X_t = \Theta(L)Z_t$  with  $Z_t \sim \text{WN}(0, \sigma^2)$  and if in addition  $\Phi(z)\Theta(z) \neq 0$  for  $|z| \leq 1$  then*

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

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

where the coefficients  $\{\psi_j\}$  and  $\{\pi_j\}$  are determined for  $|z| \leq 1$  by  $\Psi(z) = \frac{\Theta(z)}{\Phi(z)}$  and  $\Pi(z) = \frac{\Phi(z)}{\Theta(z)}$ , respectively. In this case  $\{X_t\}$  is causal and invertible with respect to  $\{Z_t\}$ .

**Remark 2.4.** *If  $\{X_t\}$  is an ARMA process with  $\Phi(L)X_t = \Theta(L)Z_t$  such that  $\Phi(z) \neq 0$  for  $|z| = 1$  then there exists polynomials  $\tilde{\Phi}(z)$  and  $\tilde{\Theta}(z)$  and a white noise process  $\{\tilde{Z}_t\}$  such that  $\{X_t\}$  fulfills the stochastic difference equation  $\tilde{\Phi}(L)X_t = \tilde{\Theta}(L)\tilde{Z}_t$  and is causal with respect to  $\{\tilde{Z}_t\}$ . If in addition  $\Theta(z) \neq 0$  for  $|z| = 1$  then  $\tilde{\Theta}(L)$  can be chosen such that  $\{X_t\}$  is also invertible with respect to  $\{\tilde{Z}_t\}$  (see the discussion of the AR(1) process after the definition of causality and Brockwell and Davis (1991, p. 88)).*

## 2.4 Computation of the autocovariance function of an ARMA process

Whereas the autocovariance function summarizes the external and directly observable properties of a time series, the coefficients of the ARMA process give information of its internal structure. Although there exists for each ARMA model a corresponding autocovariance function, the converse is not true as we have seen in Section 1.3 where we showed that two MA(1) processes are compatible with the same autocovariance function. This brings up a fundamental identification problem. In order to shed some light on the relation between autocovariance function and ARMA models it is necessary

to be able to compute the autocovariance function for a given ARMA model. In the following, we will discuss three such procedures. Each procedure relies on the assumption that the ARMA process  $\Phi(L)X_t = \Theta(L)Z_t$  with  $Z_t \sim \text{WN}(0, \sigma^2)$  is causal with respect to  $\{Z_t\}$ . Thus there exists a representation of  $X_t$  as a weighted sum of current and past  $Z_t$ 's:  $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$  with  $\sum_{j=0}^{\infty} |\psi_j| < \infty$ .

### 2.4.1 First procedure

Starting from the causal representation of  $\{X_t\}$ , it is easy to calculate its autocovariance function given that  $\{Z_t\}$  is white noise. The exact formula is proved in Theorem (6.4).

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|},$$

where

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)} \quad \text{for } |z| \leq 1.$$

The first step consists in determining the coefficients  $\psi_j$  by the method of undetermined coefficients. This leads to the following system of equations:

$$\begin{aligned} \psi_j - \sum_{0 < k \leq j} \phi_k \psi_{j-k} &= \theta_j, & 0 \leq j < \max\{p, q+1\}, \\ \psi_j - \sum_{0 < k \leq p} \phi_k \psi_{j-k} &= 0, & j \geq \max\{p, q+1\}. \end{aligned}$$

This equation system can be solved recursively (see Section 2.3):

$$\begin{aligned} \psi_0 &= \theta_0 = 1, \\ \psi_1 &= \theta_1 + \psi_0 \phi_1 = \theta_1 + \phi_1, \\ \psi_2 &= \theta_2 + \psi_0 \phi_2 + \psi_1 \phi_1 = \theta_2 + \phi_2 + \phi_1 \theta_1 + \phi_1^2, \\ &\dots \end{aligned}$$

Alternatively one may view the second part of the equation system as a linear homogeneous difference equation with constant coefficients (see Section 2.3). Its solution is given by equation (2.3). The first part of the equation system delivers the necessary initial conditions to determine the coefficients  $c_1, c_2, \dots, c_p$ . Finally one can insert the  $\psi$ 's in the above formula for the autocovariance function.

**A numerical example**

Consider the ARMA(2,1) process with  $\Phi(L) = 1 - 1.3L + 0.4L^2$  and  $\Theta(L) = 1 + 0.4L$ . Writing out the defining equation for  $\Psi(z)$ ,  $\Psi(z)\Phi(z) = \Theta(z)$ , gives:

$$\begin{aligned} 1 + \psi_1 z + \psi_2 z^2 + \psi_3 z^3 + \dots \\ - 1.3z - 1.3\psi_1 z^2 - 1.3\psi_2 z^3 - \dots \\ + 0.4z^2 + 0.4\psi_1 z^3 + \dots \\ \dots = 1 + 0.4z. \end{aligned}$$

Equating the coefficients of the powers of  $z$  leads to the following equation system:

$$\begin{aligned} z^0 : \quad & \psi_0 = 1, \\ z : \quad & \psi_1 - 1.3 = 0.4, \\ z^2 : \quad & \psi_2 - 1.3\psi_1 + 0.4 = 0, \\ z^3 : \quad & \psi_3 - 1.3\psi_2 + 0.4\psi_1 = 0, \\ & \dots \\ & \psi_j - 1.3\psi_{j-1} + 0.4\psi_{j-2} = 0, \quad \text{for } j \geq 2. \end{aligned}$$

The last equation represents a linear difference equation of order two. Its solution is given by

$$\psi_j = c_1 z_1^{-j} + c_2 z_2^{-j}, \quad j \geq \max\{p, q + 1\} - p = 0,$$

whereby  $z_1$  and  $z_2$  are the two distinct roots of the characteristic polynomial  $\Phi(z) = 1 - 1.3z + 0.4z^2 = 0$  (see equation (2.3)) and where the coefficients  $c_1$  and  $c_2$  are determined from the initial conditions. The two roots are  $\frac{1.3 \pm \sqrt{1.69 - 4 \times 0.4}}{2 \times 0.4} = 5/4 = 1.25$  and 2. The general solution to the homogeneous equation therefore is  $\psi_j = c_1 0.8^j + c_2 0.5^j$ . The constants  $c_1$  and  $c_2$  are determined by the equations:

$$\begin{aligned} j = 0 : \quad \psi_0 = 1 &= c_1 0.8^0 + c_2 0.5^0 = c_1 + c_2 \\ j = 1 : \quad \psi_1 = 1.7 &= c_1 0.8^1 + c_2 0.5^1 = 0.8c_1 + 0.5c_2. \end{aligned}$$

Solving this equation system in the two unknowns  $c_1$  and  $c_2$  gives:  $c_1 = 4$  and  $c_2 = -3$ . Thus the solution to the difference equation is given by:

$$\psi_j = 4(0.8)^j - 3(0.5)^j.$$

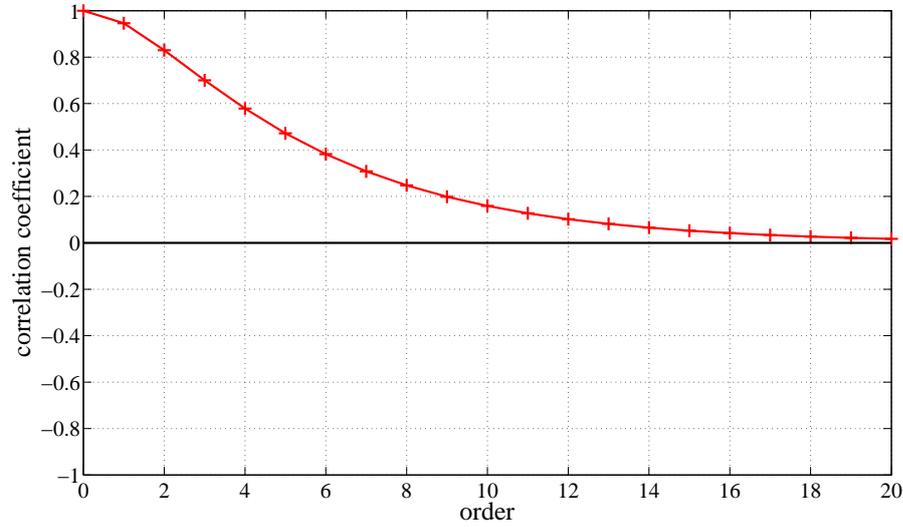


Figure 2.3: Autocorrelation function of the ARMA(2,1) process:  $(1 - 1.3L + 0.4L^2)X_t = (1 + 0.4L)Z_t$

Inserting this solution for  $\psi_j$  into the above formula for  $\gamma(h)$  one obtains after using the formula for the geometric sum:

$$\begin{aligned}
 \gamma(h) &= \sigma^2 \sum_{j=0}^{\infty} (4 \times 0.8^j - 3 \times 0.5^j) (4 \times 0.8^{j+h} - 3 \times 0.5^{j+h}) \\
 &= \sigma^2 \sum_{j=0}^{\infty} (16 \times 0.8^{2j+h} - 12 \times 0.5^j \times 0.8^{j+h} \\
 &\quad - 12 \times 0.8^j \times 0.5^{j+h} + 9 \times 0.5^{2j+h}) \\
 &= 16\sigma^2 \frac{0.8^h}{1 - 0.64} - 12\sigma^2 \frac{0.8^h}{1 - 0.4} - 12\sigma^2 \frac{0.5^h}{1 - 0.4} + 9\sigma^2 \frac{0.5^h}{1 - 0.25} \\
 &= \frac{220}{9}\sigma^2(0.8)^h - 8\sigma^2(0.5)^h.
 \end{aligned}$$

Dividing  $\gamma(h)$  by  $\gamma(0)$ , one gets the autocorrelation function:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{55}{37} \times 0.8^h - \frac{18}{37} \times 0.5^h$$

which is represented in Figure 2.3.

### 2.4.2 Second procedure

Instead of determining the  $\psi_j$  coefficients first, it is possible to compute the autocovariance function directly from the ARMA model. To see this multiply the ARMA equation successively by  $X_{t-h}$ ,  $h = 0, 1, \dots$  and apply the expectations operator:

$$\begin{aligned} \mathbb{E}X_t X_{t-h} - \phi_1 \mathbb{E}X_{t-1} X_{t-h} - \dots - \phi_p \mathbb{E}X_{t-p} X_{t-h} \\ = \mathbb{E}Z_t X_{t-h} + \theta_1 \mathbb{E}Z_{t-1} X_{t-h} + \dots + \theta_q \mathbb{E}Z_{t-q} X_{t-h}. \end{aligned}$$

This leads to an equation system for the autocovariances  $\gamma(h)$ ,  $h = 0, 1, 2, \dots$ :

$$\begin{aligned} \gamma(h) - \phi_1 \gamma(h-1) - \dots - \phi_p \gamma(h-p) &= \sigma^2 \sum_{h \leq j \leq q} \theta_j \psi_{j-h}, & h < \max\{p, q+1\} \\ \gamma(h) - \phi_1 \gamma(h-1) - \dots - \phi_p \gamma(h-p) &= 0, & h \geq \max\{p, q+1\}. \end{aligned}$$

The second part of the equation system consists again of a linear homogeneous difference equation in  $\gamma(h)$  whereas the first part can be used to determine the initial conditions. Note that the initial conditions depend  $\psi_1, \dots, \psi_q$  which have to be determined before hand. The general solution of the difference equation is:

$$\gamma(h) = c_1 z_1^{-h} + \dots + c_p z_p^{-h} \quad (2.4)$$

where  $z_1, \dots, z_p$  are the distinct roots of the polynomial  $\Phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p = 0$ .<sup>6</sup> The constants  $c_1, \dots, c_p$  can be computed from the first  $p$  initial conditions after the  $\psi_1, \dots, \psi_q$  have been calculated like in the first procedure. The form of the solution shows that the autocovariance and hence the autocorrelation function converges to zero exponentially fast.

#### A numerical example

We consider the same example as before. The second part of the above equation system delivers a difference equation for  $\gamma(h)$ :  $\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2) = 1.3\gamma(h-1) - 0.4\gamma(h-2)$ ,  $h \geq 2$ . The general solution of this difference equation is (see Appendix B):

$$\gamma(h) = c_1 (0.8)^h + c_2 (0.5)^h, \quad h \geq 2$$

where 0.8 and 0.5 are the inverses of the roots computed from the same polynomial  $\Phi(z) = 1 - 1.3z - 0.4z^2 = 0$ .

---

<sup>6</sup>In case of multiple roots the formula has to be adapted accordingly. See equation (B.2) in the Appendix.

The first part of the system delivers the initial conditions which determine the constants  $c_1$  and  $c_2$ :

$$\begin{aligned}\gamma(0) - 1.3\gamma(-1) + 0.4\gamma(-2) &= \sigma^2(1 + 0.4 \times 1.7) \\ \gamma(1) - 1.3\gamma(0) + 0.4\gamma(-1) &= \sigma^2 0.4\end{aligned}$$

where the numbers on the right hand side are taken from the first procedure. Inserting the general solution in this equation system and bearing in mind that  $\gamma(h) = \gamma(-h)$  leads to:

$$\begin{aligned}0.216c_1 + 0.450c_2 &= 1.68\sigma^2 \\ -0.180c_1 - 0.600c_2 &= 0.40\sigma^2\end{aligned}$$

Solving this equation system in the unknowns  $c_1$  and  $c_2$  one gets finally gets:  $c_1 = (220/9)\sigma^2$  and  $c_2 = -8\sigma^2$ .

### 2.4.3 Third procedure

Whereas the first two procedures produce an analytical solution which relies on the solution of a linear difference equation, the third procedure is more suited for numerical computation using a computer. It rests on the same equation system as in the second procedure. The first step determines the values  $\gamma(0), \gamma(1), \dots, \gamma(p)$  from the first part of the equation system. The following  $\gamma(h)$ ,  $h > p$  are then computed recursively using the second part of the equation system.

#### A numerical example

Using again the same example as before, the first of the equation delivers  $\gamma(2), \gamma(1)$  and  $\gamma(0)$  from the equation system:

$$\begin{aligned}\gamma(0) - 1.3\gamma(-1) + 0.4\gamma(-2) &= \sigma^2(1 + 0.4 \times 1.7) \\ \gamma(1) - 1.3\gamma(0) + 0.4\gamma(-1) &= \sigma^2 0.4 \\ \gamma(2) - 1.3\gamma(1) + 0.4\gamma(0) &= 0\end{aligned}$$

Bearing in mind that  $\gamma(h) = \gamma(-h)$ , this system has three equations in three unknowns  $\gamma(0), \gamma(1)$  and  $\gamma(2)$ . The solution is:  $\gamma(0) = (148/9)\sigma^2$ ,  $\gamma(1) = (140/9)\sigma^2$ ,  $\gamma(2) = (614/45)\sigma^2$ . This corresponds, of course, to the same numerical values as before. The subsequent values for  $\gamma(h)$ ,  $h > 2$  are then determined recursively from the difference equation  $\gamma(h) = 1.3\gamma(h-1) - 0.4\gamma(h-2)$ .

## 2.5 Exercises

**Exercise 2.5.1.** Consider the AR(1) process  $X_t = 0.8X_{t-1} + Z_t$  with  $Z_t \sim \text{WN}(0, \sigma^2)$ . Compute the variance of  $(X_1 + X_2 + X_3 + X_4)/4$ .

**Exercise 2.5.2.** Check whether the following stochastic difference equations possess a stationary solution. If yes, is the solution causal and/or invertible with respect to  $Z_t \sim \text{WN}(0, \sigma^2)$ ?

(i)  $X_t = Z_t + 2Z_{t-1}$

(ii)  $X_t = 1.3X_{t-1} + Z_t$

(iii)  $X_t = 1.3X_{t-1} - 0.4X_{t-2} + Z_t$

(iv)  $X_t = 1.3X_{t-1} - 0.4X_{t-2} + Z_t - 0.3Z_{t-1}$

(v)  $X_t = 0.2X_{t-1} + 0.8X_{t-2} + Z_t$

(vi)  $X_t = 0.2X_{t-1} + 0.8X_{t-2} + Z_t - 1.5Z_{t-1} + 0.5Z_{t-2}$

**Exercise 2.5.3.** Compute the causal representation with respect  $Z_t \sim \text{WN}(0, \sigma^2)$  of the following ARMA processes:

(i)  $X_t = 1.3X_{t-1} - 0.4X_{t-2} + Z_t$

(ii)  $X_t = 1.3X_{t-1} - 0.4X_{t-2} + Z_t - 0.2Z_{t-1}$

(iii)  $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$  with  $|\phi| < 1$

**Exercise 2.5.4.** Compute the autocovariance function of the following ARMA processes:

(i)  $X_t = 0.5X_{t-1} + 0.36X_{t-2} + Z_t$

(ii)  $X_t = 0.5X_{t-1} + 0.36X_{t-2} + Z_t + 0.5Z_{t-1}$

Thereby  $Z_t \sim \text{WN}(0, \sigma^2)$ .



# Chapter 3

## Forecasting Stationary Processes

An important goal of time series analysis is forecasting. In the following we will consider the problem of forecasting  $X_{T+h}$ ,  $h > 0$ , given  $\{X_T, \dots, X_1\}$  where  $\{X_t\}$  is a stationary stochastic process with known mean  $\mu$  and known autocovariance function  $\gamma(h)$ . In practical applications  $\mu$  and  $\gamma$  are unknown so that we must replace these entities by their estimates. These estimates can be obtained directly from the data as explained in Section 4.2 or indirectly by first estimating an appropriate ARMA model (see Chapter 5) and then inferring the corresponding autocovariance function using one of the methods explained in Section 2.4. Thus the forecasting problem is inherently linked to the problem of identifying an appropriate ARMA model from the data (see Deistler and Neusser, 2012).

### 3.1 The theory of linear least-squares forecasts

We restrict our discussion to linear *forecast functions*, also called *linear predictors*,  $\mathbb{P}_T X_{T+h}$ . Given observation from period 1 up to period  $T$ , these predictors have the form:

$$\mathbb{P}_T X_{T+h} = a_0 + a_1 X_T + \dots + a_T X_1 = a_0 + \sum_{i=1}^T a_i X_{T+1-i}$$

with unknown coefficients  $a_0, a_1, a_2, \dots, a_T$ . In principle, we should index these coefficients by  $T$  because they may change with every new observations.

See the example of the MA(1) process in Section 3.1.2. In order not to overload the notation, we will omit this additional index.

In the Hilbert space of random variables with finite second moments the optimal forecast in the mean squared error sense is given by the conditional expectation  $\mathbb{E}(X_{T+h}|c, X_1, X_2, \dots, X_T)$ . However, having practical applications in mind, we restrict ourself to linear predictors for the following reasons:<sup>1</sup>

- (i) Linear predictors are easy to compute.
- (ii) The coefficients of the optimal (in the sense of means squared errors) linear forecasting function depend only on the first two moments of the time series, i.e. on  $\mathbb{E}X_t$  and  $\gamma(j)$ ,  $j = 0, 1, \dots, h + T - 1$ .
- (iii) In the case of Gaussian processes the conditional expectation coincides with the linear predictor.
- (iv) The optimal predictor is linear when the process is a causal and invertible ARMA process even when  $Z_t$  follows an arbitrary distribution with finite variance (see Rosenblatt, 2000, chapter 5)).
- (v) Practical experience has shown that even non-linear processes can be predicted accurately by linear predictors.

The coefficients  $a_0, \dots, a_T$  of the forecasting function are determined such that the mean squared errors are minimized. The use of mean squared errors as a criterion leads to a compact representation and solution of the forecasting problem. It implies that over- and underestimation are treated equally. Thus we have to solve the following minimization problem:

$$\begin{aligned} S &= S(a_0, \dots, a_T) = \mathbb{E}(X_{T+h} - \mathbb{P}_T X_{T+h})^2 \\ &= \mathbb{E}(X_{T+h} - a_0 - a_1 X_T - \dots - a_T X_1)^2 \longrightarrow \min_{a_0, a_1, \dots, a_T} \end{aligned}$$

As  $S$  is a quadratic equation the coefficients,  $a_j$ ,  $j = 0, 1, \dots, T$ , are uniquely determined by the so-called normal equation. These are obtained from the first order conditions of the minimization problem, i.e. from  $\frac{\partial S}{\partial a_j} = 0$ ,  $j =$

---

<sup>1</sup>Elliott and Timmermann (2008) provide a general overview of forecasting procedures and their evaluations.

$0, 1, \dots, T$ :

$$\frac{\partial S}{\partial a_0} = \mathbb{E} \left( X_{T+h} - a_0 - \sum_{i=1}^T a_i X_{T+1-i} \right) = 0, \quad (3.1)$$

$$\frac{\partial S}{\partial a_j} = \mathbb{E} \left[ \left( X_{T+h} - a_0 - \sum_{i=1}^T a_i X_{T+1-i} \right) X_{T+1-j} \right] = 0, \quad j = 1, \dots, T. \quad (3.2)$$

The first equation can be rewritten as  $a_0 = \mu - \sum_{i=1}^T a_i \mu$  so that the forecasting function becomes:

$$\mathbb{P}_T X_{T+h} = \mu + \sum_{i=1}^T a_i (X_{T+1-i} - \mu).$$

The unconditional mean of the forecast error,  $\mathbb{E}(X_{T+h} - \mathbb{P}_T X_{T+h})$ , is therefore equal to zero. This means that there is no bias, neither upward nor downward, in the forecasts. The forecasts correspond on average to the “true” value.

Inserting in the second normal equation the expression for  $\mathbb{P}_T X_{T+h}$  from above, we get:

$$\mathbb{E} [(X_{T+h} - \mathbb{P}_T X_{T+h}) X_{T+1-j}] = 0, \quad j = 1, 2, \dots, T.$$

The forecast error is therefore uncorrelated with the available information represented by past observations. Thus the forecast errors  $X_{T+h} - \mathbb{P}_T X_{T+h}$  are orthogonal to  $X_T, X_{T-1}, \dots, X_1$ . Geometrically speaking, the best linear forecast is obtained by finding the point in the linear subspace spanned by  $\{X_T, X_{T-1}, \dots, X_1\}$  which is closest to  $X_{T+h}$ . This point is found by projecting  $X_{T+h}$  on this linear subspace.<sup>2</sup>

The normal equations (3.1) can be rewritten in matrix notation as follows:

$$a_0 = \mu \left( 1 - \sum_{i=1}^T a_i \right) \quad (3.3)$$

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(T-1) & \gamma(T-2) & \dots & \gamma(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+T-1) \end{pmatrix}. \quad (3.4)$$

---

<sup>2</sup>Note the similarity of the forecast errors with the least-square residuals of a linear regression.

Denoting by  $\iota$ ,  $\alpha_T$  and  $\gamma_T(h)$  the vectors  $(1, 1, \dots, 1)'$ ,  $(a_1, \dots, a_T)'$  and  $(\gamma(h), \dots, \gamma(h + T - 1))'$  and by  $\Gamma_T = [\gamma(i - j)]_{i,j=1,\dots,T}$  the symmetric  $T \times T$  covariance matrix of  $(X_1, \dots, X_T)'$  the normal equations can be written compactly as:

$$a_0 = \mu(1 - \iota' \alpha_T) \quad (3.5)$$

$$\Gamma_T \alpha_T = \gamma_T(h). \quad (3.6)$$

Dividing the second equation by  $\gamma(0)$ , one obtains an equation in terms autocorrelations instead of autocovariances:

$$R_T \alpha_T = \rho_T(h), \quad (3.7)$$

where  $R_T = \Gamma_T / \gamma(0)$  and  $\rho_T(h) = (\rho(h), \dots, \rho(h + T - 1))'$ . The coefficients of the forecasting function  $\alpha_T$  are then obtained by inverting  $\Gamma_T$ , respectively  $R_T$ :

$$\alpha_T = \begin{pmatrix} a_1 \\ \vdots \\ a_T \end{pmatrix} = \Gamma_T^{-1} \gamma_T(h) = R_T^{-1} \rho_T(h).$$

A sufficient condition which ensures the invertibility of  $\Gamma_T$ , respectively  $R_T$ , is given by assuming  $\gamma(0) > 0$  and  $\lim_{h \rightarrow \infty} \gamma(h) = 0$ .<sup>3</sup> The last condition is automatically satisfied for ARMA processes because  $\gamma(h)$  converges even exponentially fast to zero (see Section 2.4).

The mean squared error or *variance of the forecast error* for the forecasting horizon  $h$ ,  $v_T(h)$ , is given by:

$$\begin{aligned} v_T(h) &= \mathbb{E}(X_{T+h} - \mathbb{P}_T X_{T+h})^2 \\ &= \gamma(0) - 2 \sum_{i=1}^T a_i \gamma(h + i - 1) + \sum_{i=1}^T \sum_{j=1}^T a_i \gamma(i - j) a_j \\ &= \gamma(0) - 2 \alpha_T' \gamma_T(h) + \alpha_T' \Gamma_T \alpha_T \\ &= \gamma(0) - \alpha_T' \gamma_T(h), \end{aligned}$$

because  $\Gamma_T \alpha_T = \gamma_T(h)$ . Bracketing out  $\gamma(0)$ , one can write the mean squared forecast error as:

$$v_T(h) = \gamma(0) (1 - \alpha_T' \rho_T(h)). \quad (3.8)$$

Because the coefficients of the forecast function have to be recomputed with the arrival of each new observation, it is necessary to have a fast and reliable algorithm at hand. These numerical problems have been solved by the development of appropriate computer algorithms, like the Durbin-Levinson algorithm or the innovation algorithm (see Brockwell and Davis, 1991, Chapter 5).

<sup>3</sup>See Brockwell and Davis (1991, p. 167)

### 3.1.1 Forecasting with an AR(p) process

Consider first the case of an AR(1) process:

$$X_t = \phi X_{t-1} + Z_t \quad \text{with } |\phi| < 1 \text{ and } Z_t \sim \text{WN}(0, \sigma^2).$$

The equation system (3.7) becomes:

$$\begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \phi^h \\ \phi^{h+1} \\ \phi^{h+2} \\ \vdots \\ \phi^{h+T-1} \end{pmatrix}.$$

The guess-and-verify method immediately leads to the solution:

$$\alpha_T = (a_1, a_2, a_3, \dots, a_T)' = (\phi^h, 0, 0, \dots, 0)'.$$

We therefore get the following predictor:

$$\mathbb{P}_T X_{T+h} = \phi^h X_T$$

The forecast therefore just depends on the last observation. All previous observations can be disregarded, they cannot improve the forecast further. To put it otherwise, all the useful information about  $X_{t+1}$  in the entire realization previous to  $X_{t+1}$  is contained in  $X_t$ .

The variance of the prediction error is given by

$$v_T(h) = \frac{1 - \phi^{2h}}{1 - \phi^2} \sigma^2$$

For  $h = 1$ , the formula simplifies to  $\sigma^2$  and for  $h \rightarrow \infty$ ,  $v_T(h) \rightarrow \frac{1}{1 - \phi^2} \sigma^2$ . Note also that the variance of the forecast error  $v_T(h)$  increases with  $h$ .

The general case of an AR(p) process,  $p > 1$ , can be treated in the same way. The autocovariances follow a  $p$ -th order difference equation (see Section 2.4):

$$\gamma(j) = \phi_1 \gamma(j-1) + \phi_2 \gamma(j-2) + \dots + \phi_p \gamma(j-p).$$

Applying again the guess-and-verify method for the case  $h = 1$  and assuming that  $T > p$ , the solution is given by  $\alpha_T = (\phi_1, \phi_2, \dots, \phi_p, 0, \dots, 0)'$ . Thus the one-step ahead predictor is

$$\mathbb{P}_T X_{T+1} = \phi_1 X_T + \phi_2 X_{T-1} + \dots + \phi_p X_{T+1-p}, \quad T > p. \quad (3.9)$$

The one-step ahead forecast of an AR( $p$ ) process therefore depends only on the last  $p$  observations.

The above predictor can also be obtained in a different way. View for this purpose  $\mathbb{P}_T$  as an *operator* with the following meaning: Take the linear least-squares forecast with respect to the information  $\{X_T, \dots, X_1\}$ . Apply this operator to the defining stochastic difference equation of the AR( $p$ ) process forwarded one period:

$$\mathbb{P}_T X_{T+1} = \mathbb{P}_T (\phi_1 X_T) + \mathbb{P}_T (\phi_2 X_{T-1}) + \dots + \mathbb{P}_T (\phi_p X_{T+1-p}) + \mathbb{P}_T (Z_{T+1}).$$

In period  $T$  the observations of  $X_T, X_{T-1}, \dots, X_1$  are known so that  $\mathbb{P}_T X_{T-j} = X_{T-j}$ ,  $j = 0, 1, \dots, T-1$ . Because  $\{Z_t\}$  is a white noise process and because  $\{X_t\}$  is causal function with respect to  $\{Z_t\}$ ,  $Z_{T+1}$  is uncorrelated with  $X_T, \dots, X_1$ . This reasoning leads to the same predictor as in equation (3.9).

The forecasting functions for  $h > 1$  can be obtained recursively by successively applying the forecast operator. Take, for example, the case  $h = 2$ :

$$\begin{aligned} \mathbb{P}_T X_{T+2} &= \mathbb{P}_T (\phi_1 X_{T+1}) + \mathbb{P}_T (\phi_2 X_T) + \dots + \mathbb{P}_T (\phi_p X_{T+2-p}) + \mathbb{P}_T (Z_{T+2}) \\ &= \phi_1 (\phi_1 X_T + \phi_2 X_{T-1} + \dots + \phi_p X_{T+1-p}) \\ &\quad + \phi_2 X_T + \dots + \phi_p X_{T+2-p} \\ &= (\phi_1^2 + \phi_2) X_T + (\phi_1 \phi_2 + \phi_3) X_{T-1} + \dots + (\phi_1 \phi_{p-1} + \phi_p) X_{T+2-p} \\ &\quad + \phi_1 \phi_p X_{T+1-p}. \end{aligned}$$

In this way forecasting functions for  $h > 2$  can be obtained.

Note that in the case of AR( $p$ ) processes the coefficient of the forecast function remain constant as long as  $T > p$ . Thus with each new observation it is not necessary to recompute the equation system and solve it again. This will be different in the case of MA processes. In practice, the parameters of the AR model are usually unknown and have therefore be replaced by some estimate. Section 14.2 investigates how this substitution affects the results.

### 3.1.2 Forecasting with MA( $q$ ) processes

The forecasting problem becomes more complicated in the case of MA( $q$ ) processes. In order to get a better understanding we analyze the case of a MA(1) process:

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

Taking a forecast horizon of one period, i.e.  $h = 1$ , the equation system (3.7) in the case of a MA(1) process becomes:

$$\begin{pmatrix} 1 & \frac{\theta}{1+\theta^2} & 0 & \dots & 0 \\ \frac{\theta}{1+\theta^2} & 1 & \frac{\theta}{1+\theta^2} & \dots & 0 \\ 0 & \frac{\theta}{1+\theta^2} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_T \end{pmatrix} = \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (3.10)$$

Despite the fact that the equation system has a simple structure, the forecasting function will depend in general on all past observations of  $X_{T-j}$ ,  $j \geq 0$ . We illustrate this point by a numerical example which will allow us to get a deeper understanding.

Suppose that we know the parameters of the MA(1) process to be  $\theta = -0.9$  and  $\sigma^2 = 1$ . We start the forecasting exercise in period  $T = 0$  and assume that, at this point in time, we have no observation at hand. The best forecast is therefore just the unconditional mean which in this example is zero. Thus,  $\mathbb{P}_0 X_1 = 0$ . The variance of the forecast error then is  $\mathbb{V}(X_1 - \mathbb{P}_0 X_1) = v_0(1) = \sigma^2 + \theta^2 \sigma^2 = 1.81$ . This result is summarized in the first row of table 3.1. In period 1, the realization of  $X_1$  is observed. This information can be used and the forecasting function becomes  $\mathbb{P}_1 X_2 = a_1 X_1$ . The coefficient  $a_1$  is found by solving the equation system (3.10) for  $T = 1$ . This gives  $a_1 = \theta/(1 + \theta^2) = -0.4972$ . The corresponding variance of the forecast error according to equation (3.8) is  $\mathbb{V}(X_2 - \mathbb{P}_1 X_2) = v_1(1) = \gamma(0)(1 - \alpha_1' \rho_1(1)) = 1.81(1 - 0.4972 \times 0.4972) = 1.3625$ . This value is lower compared to the previous forecast because additional information, the observation of the realization of  $X_1$ , is taken into account. Row 2 in table 3.1 summarizes these results.

In period 2, not only  $X_1$ , but also  $X_2$  is observed which allows us to base our forecast on both observations:  $\mathbb{P}_2 X_3 = a_1 X_2 + a_2 X_1$ . The coefficients can be found by solving the equation system (3.10) for  $T = 2$ . This amounts to solving the simultaneous equation system

$$\begin{pmatrix} 1 & \frac{\theta}{1+\theta^2} \\ \frac{\theta}{1+\theta^2} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \end{pmatrix}.$$

Inserting  $\theta = -0.9$ , the solution is  $\alpha_2 = (a_1, a_2)' = (-0.6606, -0.3285)'$ . The

corresponding variance of the forecast error becomes

$$\begin{aligned}\mathbb{V}(X_3 - \mathbb{P}_2 X_3) &= v_2(1) = \gamma(0)(1 - \alpha'_2 \rho_2(1)) \\ &= \gamma(0) \left( 1 - (a_1 \ a_2) \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \end{pmatrix} \right) \\ &= 1.81 \left( 1 - (-0.6606 \ -0.3285) \begin{pmatrix} -0.4972 \\ 0 \end{pmatrix} \right) = 1.2155.\end{aligned}$$

These results are summarized in row 3 of table 3.1.

In period 3, the realizations of  $X_1$ ,  $X_2$  and  $X_3$  are so that the forecast function becomes  $\mathbb{P}_3 X_4 = a_1 X_3 + a_2 X_2 + a_3 X_1$ . The coefficients can again be found by solving the equation system (3.10) for  $T = 3$ :

$$\begin{pmatrix} 1 & \frac{\theta}{1+\theta^2} & 0 \\ \frac{\theta}{1+\theta^2} & 1 & \frac{\theta}{1+\theta^2} \\ 0 & \frac{\theta}{1+\theta^2} & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \\ 0 \end{pmatrix}.$$

Inserting  $\theta = -0.9$ , the solution is  $\alpha_2 = (a_1, a_2, a_3)' = (-0.7404, -0.4891, -0.2432)'$ . The corresponding variance of the forecast error becomes

$$\begin{aligned}\mathbb{V}(X_4 - \mathbb{P}_3 X_4) &= v_3(1) = \gamma(0)(1 - \alpha'_3 \rho_3(1)) \\ &= \gamma(0) \left( 1 - (a_1 \ a_2 \ a_3) \begin{pmatrix} \frac{\theta}{1+\theta^2} \\ 0 \\ 0 \end{pmatrix} \right) \\ &= 1.81 \left( 1 - (-0.7404 \ -0.4891 \ -0.2432) \begin{pmatrix} -0.4972 \\ 0 \\ 0 \end{pmatrix} \right) \\ &= 1.1436.\end{aligned}$$

These results are summarized in row 4 of table 3.1. We can, of course, continue in this way and derive successively the forecast functions for  $T = 4, 5, \dots$

From this exercise we can make several observations.

- In contrast to the AR process, every new information is used. The forecast  $\mathbb{P}_T X_{T+1}$  depends on all available information, in particular on  $X_T, X_{T-1}, \dots, X_1$ .
- The coefficients of the forecast function are not constant. They change as more and more information comes in.

Table 3.1: Forecast function for a MA(1) process with  $\theta = -0.9$  and  $\sigma^2 = 1$ 

time	forecasting function $\alpha_T = (a_1, a_2, \dots, a_T)'$	forecast error variance
$T = 0 :$		$v_0(1) = 1.8100$
$T = 1 :$	$\alpha_1 = (-0.4972)'$	$v_1(1) = 1.3625$
$T = 2 :$	$\alpha_2 = (-0.6606, -0.3285)'$	$v_2(1) = 1.2155$
$T = 3 :$	$\alpha_3 = (-0.7404, -0.4891, -0.2432)'$	$v_3(1) = 1.1436$
$T = 4 :$	$\alpha_4 = (-0.7870, -0.5827, -0.3849, -0.1914)'$	$v_4(1) = 1.1017$
$\dots$	$\dots$	$\dots$
$T = \infty :$	$\alpha_\infty = (-0.9000, -0.8100, -0.7290, \dots)'$	$v_\infty(1) = 1$

- The importance of the new information can be “measured” by the last coefficient of  $\alpha_T$ . These coefficients are termed *partial autocorrelations* (see Definition 3.2) and are of particular relevance as will be explained in Chapter 3.5. In our example they are -0.4972, -0.3285, -0.2432, and -0.1914.
- As more information becomes available, the variance of the forecast error (mean squared error) declines monotonically. It will converge to  $\sigma^2 = 1$ . The reason for this result can be explained as follows. Applying the forecasting operator to the defining MA(1) stochastic difference equation forwarded by one period gives:  $\mathbb{P}_T X_{T+1} = \mathbb{P}_T Z_{T+1} + \theta \mathbb{P}_T Z_T = \theta \mathbb{P}_T Z_T$  with forecast error  $X_{T+1} - \mathbb{P}_T X_{T+1} = Z_{T+1}$ . As more and more observation become available, it becomes better and better possible to recover the “true” value of the unobserved  $Z_T$  from the observations  $X_T, X_{T-1}, \dots, X_1$ . As the process is invertible, in the limit it is possible to recover the value of  $Z_T$  exactly (almost surely). The only uncertainty remaining is with respect to  $Z_{T+1}$  which has a mean of zero and a variance of  $\sigma^2 = 1$ .

### 3.1.3 Forecasting from the Infinite Past

The forecasting function based on the *infinitely remote past* is of particular theoretical interest. Thereby we look at the problem of finding the optimal linear forecast of  $X_{T+h}$  given  $X_T, X_{T-1}, \dots, X_1, X_0, X_{-1}, \dots$  taking the mean squared error again as the criterion function. The corresponding forecasting function (predictor) will be denoted by  $\tilde{\mathbb{P}}_T X_{T+h}$ ,  $h > 0$ .

Noting that the MA(1) process with  $|\theta| < 1$  is invertible, we have

$$Z_t = X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \dots$$

We can therefore write  $X_{t+1}$  as

$$X_{t+1} = Z_{t+1} + \theta \underbrace{(X_t - \theta X_{t-1} + \theta^2 X_{t-2} - \dots)}_{Z_t}$$

The predictor of  $X_{T+1}$  from the infinite past,  $\tilde{\mathbb{P}}_T$ , is then given by:

$$\tilde{\mathbb{P}}_T X_{T+1} = \theta (X_T - \theta X_{T-1} + \theta^2 X_{T-2} - \dots)$$

where the mean squared forecasting error is

$$v_\infty(1) = \mathbb{E} \left( X_{T+1} - \tilde{\mathbb{P}}_T X_{T+1} \right)^2 = \sigma^2.$$

Applying this result to our example gives:

$$\tilde{\mathbb{P}}_T X_{T+1} = -0.9X_T - 0.81X_{T-1} - 0.729\theta^2 X_{T-2} - \dots$$

with  $v_\infty(1) = 1$ . See last row in table 3.1.

### Example of an ARMA(1,1) process

Consider now the case of a causal and invertible ARMA(1,1) process  $\{X_t\}$ :

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1},$$

where  $|\phi| < 1$ ,  $|\theta| < 1$  and  $Z_t \sim \text{WN}(0, \sigma^2)$ . Because  $\{X_t\}$  is causal and invertible with respect to  $\{Z_t\}$ ,

$$\begin{aligned} X_{T+1} &= Z_{T+1} + (\phi + \theta) \sum_{j=0}^{\infty} \phi^j Z_{T-j}, \\ Z_{T+1} &= X_{T+1} - (\phi + \theta) \sum_{j=0}^{\infty} (-\theta)^j X_{T-j}. \end{aligned}$$

Applying the forecast operator  $\tilde{\mathbb{P}}_T$  to the second equation and noting that  $\tilde{\mathbb{P}}_T Z_{T+1} = 0$ , one obtains the following one-step ahead predictor

$$\tilde{\mathbb{P}}_T X_{T+1} = (\phi + \theta) \sum_{j=0}^{\infty} (-\theta)^j X_{T-j}.$$

Applying the forecast operator to the first equation, we obtain

$$\tilde{\mathbb{P}}_T X_{T+1} = (\phi + \theta) \sum_{j=0}^{\infty} (\phi)^j Z_{T-j}.$$

This implies that one-step ahead prediction error is equal to  $X_{T+1} - \tilde{\mathbb{P}}_T X_{T+1} = Z_{T+1}$  and that the mean squared forecasting error of the one-step ahead predictor given the infinite past is equal to  $\mathbb{E} Z_{T+1}^2 = \sigma^2$ .

## 3.2 The Wold Decomposition

The *Wold Decomposition theorem* is essential for the theoretical understanding of stationary stochastic processes. It shows that *any* stationary process can essentially be represented as a linear combination of current and past forecast errors. Before we can state the theorem precisely, we have to introduce the following definition.

**Definition 3.1.** *A stochastic process  $\{X_t\}$  is called deterministic if and only if it can be forecasted exactly from the infinite past. More precisely, if and only if*

$$\sigma^2 = \mathbb{E} \left( X_{t+1} - \tilde{\mathbb{P}}_t X_{t+1} \right)^2 = 0 \quad \text{for all } t \in \mathbb{Z}$$

where  $\tilde{\mathbb{P}}_t X_{t+1}$  denotes the best linear forecast of  $X_{t+1}$  given its infinite past, i.e. given  $\{X_t, X_{t-1}, \dots\}$ .

The most important class of deterministic processes are the *harmonic processes*. These processes are characterized by finite or infinite sums of sine and cosine functions with stochastic amplitude.<sup>4</sup> A simple example of a harmonic process is given by

$$X_t = A \cos(\omega t) + B \sin(\omega t) \quad \text{with } \omega \in (0, \pi).$$

Thereby,  $A$  and  $B$  denote two uncorrelated random variables with mean zero and finite variance. One can check that  $X_t$  satisfies the deterministic difference equation

$$X_t = (2 \cos \omega) X_{t-1} - X_{t-2}.$$

Thus,  $X_t$  can be forecasted exactly from its past. In this example the last two observations are sufficient. We are now in a position to state the Wold Decomposition Theorem.

**Theorem 3.1** (Wold Decomposition). *Every stationary stochastic process  $\{X_t\}$  with mean zero and finite positive variance can be represented as*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t = \Psi(L)Z_t + V_t, \quad (3.11)$$

where

$$(i) \quad Z_t = X_t - \tilde{\mathbb{P}}_{t-1} X_t = \tilde{\mathbb{P}}_t Z_t;$$

---

<sup>4</sup>More about harmonic processes can be found in Section 6.2.

- (ii)  $Z_t \sim \text{WN}(0, \sigma^2)$  where  $\sigma^2 = \mathbb{E} \left( X_{t+1} - \tilde{\mathbb{P}}_t X_{t+1} \right)^2 > 0$ ;
- (iii)  $\psi_0 = 1$  and  $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ ;
- (iv)  $\{V_t\}$  is deterministic;
- (v)  $\mathbb{E}(Z_t V_s) = 0$  for all  $t, s \in \mathbb{Z}$ .

The sequences  $\{\psi_j\}$ ,  $\{Z_t\}$ , and  $\{V_t\}$  are uniquely determined by (3.11).

*Proof.* The proof, although insightful, requires some knowledge about Hilbert spaces which is beyond the scope of this book. A rigorous proof can be found in Brockwell and Davis (1991, Section 5.7).

It is nevertheless instructive to give an intuition of the proof. Following the MA(1) example from the previous section, we start in period 0 and assume that no information is available. Thus, the best forecast  $\mathbb{P}_0 X_1$  is zero so that trivially

$$X_1 = X_1 - \mathbb{P}_0 X_1 = Z_1.$$

Starting with  $X_1 = Z_1$ ,  $X_2, X_3, \dots$  can then be constructed recursively:

$$\begin{aligned} X_2 &= X_2 - \mathbb{P}_1 X_2 + \mathbb{P}_1 X_2 = Z_2 + a_1^{(1)} X_1 = Z_2 + a_1^{(1)} Z_1 \\ X_3 &= X_3 - \mathbb{P}_2 X_3 + \mathbb{P}_2 X_3 = Z_3 + a_1^{(2)} X_2 + a_2^{(2)} X_1 \\ &= Z_3 + a_1^{(2)} Z_2 + \left( a_1^{(2)} a_1^{(1)} + a_2^{(2)} \right) Z_1 \\ X_4 &= X_4 - \mathbb{P}_3 X_4 + \mathbb{P}_3 X_4 = Z_4 + a_1^{(3)} X_3 + a_2^{(3)} X_2 + a_3^{(3)} X_1 \\ &= Z_4 + a_1^{(3)} Z_3 + \left( a_1^{(3)} a_1^{(2)} + a_2^{(3)} \right) Z_2 \\ &\quad + \left( a_1^{(3)} a_1^{(2)} a_1^{(1)} + a_1^{(3)} a_2^{(2)} + a_2^{(3)} a_1^{(1)} + a_3^{(3)} \right) Z_1 \\ &\dots \end{aligned}$$

where  $a_j^{(t-1)}$ ,  $j = 1, \dots, t-1$ , denote the coefficients of the forecast function for  $X_t$  based on  $X_{t-1}, \dots, X_1$ . This shows how  $X_t$  unfolds into the sum of forecast errors. The stationarity of  $\{X_t\}$  ensures that the coefficients of  $Z_j$  converge, as  $t$  goes to infinity, to  $\psi_j$  which are independent of  $t$ .  $\square$

Every stationary stochastic process is thus representable as the sum of a moving-average of infinite order and a purely deterministic process.<sup>5</sup> The

<sup>5</sup>The Wold Decomposition corresponds to the decomposition of the spectral distribution function of  $F$  into the sum of  $F_Z$  and  $F_V$  (see Section 6.2). Thereby the spectral distribution function  $F_Z$  has spectral density  $f_Z(\lambda) = \frac{\sigma^2}{2\pi} |\Psi(e^{-i\lambda})|^2$ .

weights of the infinite moving average are thereby normalized such that  $\psi_0 = 1$ . In addition, the coefficients  $\psi_i$  are square summable. This property is less strong than absolute summability which is required for a causal representation (see Definition 2.2). The process  $\{Z_t\}$  is a white noise process with positive variance  $\sigma^2 > 0$ . The  $Z_t$ 's are called *innovations* as they represent the one-period ahead forecast errors based on the infinite past, i.e.  $Z_t = X_t - \tilde{\mathbb{P}}_{t-1}X_t$ .  $Z_t$  is the additional information revealed from the  $t$ -th observation. Thus, the Wold Decomposition Theorem serves as a justification for the use of causal ARMA models.

The second part of Property (i) further means that the innovation process  $\{Z_t\}$  is *fundamental* with respect to  $\{X_t\}$ , i.e. that  $Z_t$  lies in the linear space spanned by  $\{X_t, X_{t-1}, X_{t-2}, \dots\}$  or that  $Z_t = \tilde{\mathbb{P}}_t Z_t$ . This implies that  $\Psi(L)$  must be invertible and that  $Z_t$  can be perfectly (almost surely) recovered from the observations of  $X_t, X_{t-1}, \dots$ . Finally, property (v) says that the two components  $\{Z_t\}$  and  $\{V_t\}$  are uncorrelated with each other at all leads and lags. Thus, in essence, the Wold Decomposition Theorem states that every stationary stochastic process can be uniquely decomposed into a weighted sum of current and past forecast errors plus a deterministic process.

Although the Wold Decomposition is very appealing from a theoretical perspective, it is not directly implementable in practice because it requires the estimation of infinitely many parameters ( $\psi_1, \psi_2, \dots$ ). This is impossible with only a finite amount of observations. It is therefore necessary to place some assumptions on ( $\psi_1, \psi_2, \dots$ ). One possibility is to assume that  $\{X_t\}$  is a causal ARMA process and to recover the  $\psi_j$ 's from the causal representation. This amounts to say that  $\Psi(L)$  is a rational polynomial which means that

$$\Psi(L) = \frac{\Theta(L)}{\Phi(L)} = \frac{1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p}.$$

This necessitates the estimation of only  $p+q$  parameters. Another possibility is to place restrictions on the smoothness of the spectrum (see Chapter 6).

The Wold Decomposition Theorem has several implications which are presented in the following remarks.

**Remark 3.1.** *In the case of ARMA processes, the purely deterministic part can be disregarded so that the process is represented only as a weighted sum of current and past innovations. Processes with this property are called purely non-deterministic or regular.*

**Remark 3.2.** *The process  $\{Z_t\}$  is white noise, but not necessarily Gaussian. In particular,  $\{Z_t\}$  need not be independently and identically distributed (IID). Thus,  $\mathbb{E}(Z_{t+1}|X_t, X_{t-1}, \dots)$  need not be equal to zero although*

$\tilde{P}_t Z_{t+1} = 0$ . The reason is that  $\tilde{P}_t Z_{t+1}$  is only the best linear forecast function, whereas  $\mathbb{E}(Z_{t+1}|X_t, X_{t-1}, \dots)$  is the best forecast function among all linear and non-linear functions. Examples of processes which are white noise, but not IID, are GARCH processes discussed in Chapter 8.

**Remark 3.3.** The innovations  $\{Z_t\}$  may not correspond to the “true” shocks of the underlying economic system. In this case, the shocks to the economic system cannot be recovered from the Wold Decomposition. Thus, they are not fundamental with respect to  $\{X_t\}$ . Suppose, as a simple example, that  $\{X_t\}$  is generated by a noninvertible MA(1):

$$X_t = U_t + \theta U_{t-1}, \quad U_t \sim \text{WN}(0, \sigma^2) \quad \text{and } |\theta| > 1.$$

This generates an impulse response function with respect to the true shocks of the system equal to  $(1, \theta, 0, \dots)$ . The above mechanism can, however, not be the Wold Decomposition because the noninvertibility implies that  $U_t$  cannot be recovered from the observation of  $\{X_t\}$ . As shown in the introduction, there is an observationally equivalent MA(1) process, i.e. a process which generates the same ACF. Based on the computation in section 1.5, this MA(1) process is

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

with  $\tilde{\theta} = \theta^{-1}$  and  $\tilde{\sigma}^2 = \frac{1+\theta^2}{1+\theta^{-2}}\sigma^2$ . This is already the Wold Decomposition. The impulse response function for this process is  $(1, \theta^{-1}, 0, \dots)$  which is different from the original system. As  $|\tilde{\theta}| = |\theta^{-1}| < 1$ , the innovations  $\{Z_t\}$  can be recovered from the observations as  $Z_t = \sum_{j=0}^{\infty} (-\tilde{\theta})^j X_{t-j}$ , but they do not correspond to the shocks of the system  $\{U_t\}$ . Hansen and Sargent (1991), Quah (1990), and Lippi and Reichlin (1993) among others provide a deeper discussion and present additional more interesting economic examples.

### 3.3 Exponential smoothing

Besides the method of least-squares forecasting *exponential smoothing* can often be seen as a valid alternative. This method views  $X_t$  as a function of time:

$$X_t = f(t; \beta) + \varepsilon_t,$$

whereby  $f(t; \beta)$  typically represents a polynomial in  $t$  with coefficients  $\beta$ . The above equation is similar to a regression model with error term  $\varepsilon_t$ . This error term is usually specified as a white noise process  $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$ .

Consider first the simplest case where  $X_t$  just moves randomly around a fixed mean  $\beta$ . This corresponds to the case where  $f(t; \beta)$  is a polynomial

of degree zero:

$$X_t = \beta + \varepsilon_t.$$

If  $\beta$  is known then  $\mathbb{P}_T X_{T+h}$ , the forecast of  $X_{T+h}$  given the observations  $X_T, \dots, X_1$ , clearly is  $\beta$ . If, however,  $\beta$  is unknown, we can substitute  $\beta$  by  $\bar{X}_T$ , the average of the observations:

$$\hat{\mathbb{P}}_T X_{T+h} = \hat{\beta} = \bar{X}_T = \frac{1}{T} \sum_{t=1}^T X_t,$$

where “ $\hat{\cdot}$ ” means that the model parameter  $\beta$  has been replaced by its estimate. The one-period ahead forecast function can then be rewritten as follows:

$$\begin{aligned} \hat{\mathbb{P}}_T X_{T+1} &= \frac{T-1}{T} \hat{\mathbb{P}}_{T-1} X_T + \frac{1}{T} X_T \\ &= \hat{\mathbb{P}}_{T-1} X_T + \frac{1}{T} \left( X_T - \hat{\mathbb{P}}_{T-1} X_T \right). \end{aligned}$$

The first equation represents the forecast for  $T+1$  as a linear combination of the forecast for  $T$  and of the last additional information, i.e. the last observation. The weight given to the last observation is equal to  $1/T$  because we assumed that the mean remains constant and because the contribution of one observation to the mean is  $1/T$ . The second equation represents the forecast for  $T+1$  as the forecast for  $T$  plus a correction term which is proportional to the last forecast error. One advantage of this second representation is that the computation of the new forecast, i.e. the forecast for  $T+1$ , only depends on the forecast for  $T$  and additional observation. In this way the storage requirements are minimized.

In many applications, the mean does not remain constant, but is a slowly moving function of time. In this case it is no longer meaningful to give each observation the same weight. Instead, it seems plausible to weigh the more recent observation higher than the older ones. A simple idea is to let the weights decline exponentially which leads to the following forecast function:

$$\mathbb{P}_T X_{T+1} = \frac{1-\omega}{1-\omega^T} \sum_{t=0}^{T-1} \omega^t X_{T-t} \quad \text{with } |\omega| < 1.$$

$\omega$  thereby acts like a discount factor which controls the rate at which agents forget information.  $1-\omega$  is often called the smoothing parameter. The value of  $\omega$  should depend on the speed at which the mean changes. In case when the mean changes only slowly,  $\omega$  should be large so that all observations are almost equally weighted; in case when the mean changes quickly,  $\omega$  should

be small so that only the most recent observations are taken into account. The normalizing constant  $\frac{1-\omega}{1-\omega^T}$  ensures that the weights sum up to one. For large  $T$  the term  $\omega^T$  can be disregarded so that one obtains the following forecasting function based on *simple exponential smoothing*:

$$\begin{aligned}\mathbb{P}_T X_{T+1} &= (1 - \omega) [X_T + \omega X_{T-1} + \omega^2 X_{T-2} + \dots] \\ &= (1 - \omega) X_T + \omega \mathbb{P}_{T-1} X_T \\ &= \mathbb{P}_{T-1} X_T + (1 - \omega) (X_T - \mathbb{P}_{T-1} X_T).\end{aligned}$$

In the economics literature this forecasting method is called *adaptive expectation*. Similar to the model with constant mean, the new forecast is a weighted average between the old forecast and the last (newest) observation, respectively between the previous forecast and a term proportional to the last forecast error.

One important advantage of adaptive forecasting methods is that they can be computed recursively. Starting with value  $S_0$ , the following values can be computed as follows:

$$\begin{aligned}\mathbb{P}_0 X_1 &= S_0 \\ \mathbb{P}_1 X_2 &= \omega \mathbb{P}_0 X_1 + (1 - \omega) X_1 \\ \mathbb{P}_2 X_3 &= \omega \mathbb{P}_1 X_2 + (1 - \omega) X_2 \\ &\dots \\ \mathbb{P}_T X_{T+1} &= \omega \mathbb{P}_{T-1} X_T + (1 - \omega) X_T.\end{aligned}$$

Thereby  $S_0$  has to be determined. Because

$$\mathbb{P}_T X_{T+1} = (1 - \omega) [X_T + \omega X_{T-1} + \dots + \omega^{T-1} X_1] + \omega^T S_0,$$

the effect of the starting value declines exponentially with time. In practice, we can take  $S_0 = X_1$  or  $S_0 = \bar{X}_T$ . The discount factor  $\omega$  is usually set a priori to be a number between 0.7 and 0.95. It is, however, possible to determine  $\omega$  optimally by choosing a value which minimizes the mean squared one-period forecast error:

$$\sum_{t=1}^T (X_t - \mathbb{P}_{t-1} X_t)^2 \longrightarrow \min_{|\omega| < 1}.$$

From a theoretical perspective one can ask the question for which class of models exponential smoothing represents the optimal procedure. Muth (1960) showed that this class of models is given by

$$\Delta X_t = X_t - X_{t-1} = Z_t - \omega Z_{t-1}.$$

Note that the process generated by the above equation is no longer stationary. This has to be expected as the exponential smoothing assumes a non-constant mean. Despite the fact that this class seems rather restrictive at first, practice has shown that it delivers reasonable forecasts, especially in situations when it becomes costly to specify a particular model.<sup>6</sup> Additional results and more general exponential smoothing methods can be found in Abraham and Ledolter (1983) and Mertens and Rässler (2005).

---

<sup>6</sup>This happens, for example, when many, perhaps thousands of time series have to be forecasted in a real time situation.

### 3.4 Exercises

**Exercise 3.4.1.** Compute the linear least-squares forecasting function  $\mathbb{P}_T X_{T+h}$ ,  $T > 2$ , and the mean squared error  $v_T(h)$ ,  $h = 1, 2, 3$ , if  $\{X_t\}$  is given by the AR(2) process

$$X_t = 1.3X_{t-1} - 0.4X_{t-2} + Z_t \quad \text{with} \quad Z_t \sim \text{WN}(0, 2).$$

To which values do  $\mathbb{P}_T X_{T+h}$  and  $v_T(h)$  converge for  $h$  going to infinity?

**Exercise 3.4.2.** Compute the linear least-squares forecasting function  $\mathbb{P}_T(X_{T+1})$  and the mean squared error  $v_T(1)$ ,  $T = 0, 1, 2, 3$ , if  $\{X_t\}$  is given by the MA(1) process

$$X_t = Z_t + 0.8Z_{t-1} \quad \text{with} \quad Z_t \sim \text{WN}(0, 2).$$

To which values do  $\mathbb{P}_T X_{T+h}$  and  $v_T(h)$  converge for  $h$  going to infinity?

**Exercise 3.4.3.** Suppose that you observe  $\{X_t\}$  for the two periods  $t = 1$  and  $t = 3$ , but not for  $t = 2$ .

- (i) Compute the linear least-squares forecast for  $X_2$  if

$$X_t = \phi X_{t-1} + Z_t \quad \text{with} \quad |\phi| < 1 \quad \text{and} \quad Z_t \sim \text{WN}(0, 4)$$

Compute the mean squared error for this forecast.

- (ii) Assume now that  $\{X_t\}$  is the MA(1) process

$$X_t = Z_t + \theta Z_{t-1} \quad \text{with} \quad Z_t \sim \text{WN}(0, 4).$$

Compute the mean squared error for the forecast of  $X_2$ .

## 3.5 The Partial Autocorrelation Function

Consider again the problem of forecasting  $X_{T+1}$  from observations  $X_T, X_{T-1}, \dots, X_2, X_1$ . Denoting, as before, the best linear predictor by  $\mathbb{P}_T X_{T+1} = a_1 X_T + a_2 X_{T-1} + a_{T-1} X_2 + a_T X_1$ , we can express  $X_{T+1}$  as

$$X_{T+1} = \mathbb{P}_T X_{T+1} + Z_{T+1} = a_1 X_T + a_2 X_{T-1} + a_{T-1} X_2 + a_T X_1 + Z_{T+1}$$

where  $Z_{T+1}$  denotes the forecast error which is uncorrelated with  $X_T, \dots, X_1$ . We can now ask the question whether  $X_1$  contributes to the forecast of  $X_{T+1}$  *controlling* for  $X_T, X_{T-2}, \dots, X_2$  or, equivalently, whether  $a_T$  is equal to zero. Thus,  $a_T$  can be viewed as a measure of the importance of the additional information provided by  $X_1$ . It is referred to as the *partial autocorrelation*. In the case of an AR(p) process, the whole information useful for forecasting  $X_{T+1}$ ,  $T > p$ , is incorporated in the last  $p$  observations so that  $a_T = 0$ . In the case of the MA process, the observations on  $X_T, \dots, X_1$  can be used to retrieve the unobserved  $Z_T, Z_{T-1}, \dots, Z_{t-q+1}$ . As  $Z_t$  is an infinite weighted sum of past  $X_t$ 's, every new observation contributes to the recovering of the  $Z_t$ 's. Thus, the partial autocorrelation  $a_T$  is not zero. Taking  $T$  successively equal to 0, 1, 2, etc. we get the partial autocorrelation function (PACF).

We can, however, interpret the above equation as a regression equation. From the Frisch-Lovell-Waugh Theorem (See Davidson and MacKinnon, 1993), we can obtain  $a_T$  by a two-stage procedure. Project (regress) in a first stage  $X_{T+1}$  on  $X_T, \dots, X_2$  and take the residual. Similarly, project (regress)  $X_1$  on  $X_T, \dots, X_2$  and take the residual. The coefficient  $a_T$  is then obtained by projecting (regressing) the first residual on the second. Stationarity implies that this is nothing but the correlation coefficient between the two residuals.

### 3.5.1 Definition

The above intuition suggests two equivalent definitions of the partial autocorrelation function (PACF).

**Definition 3.2.** *The partial autocorrelation function (PACF)  $\alpha(h)$ ,  $h = 0, 1, 2, \dots$ , of a stationary process is defined as follows:*

$$\begin{aligned}\alpha(0) &= 1 \\ \alpha(h) &= a_h,\end{aligned}$$

where  $a_h$  denotes the last element of the vector  $\alpha_h = \Gamma_h^{-1} \gamma_h(1) = R_h^{-1} \rho_h(1)$  (see Section 3.1 and equation (3.7)).

**Definition 3.3.** The partial autocorrelation function (PACF)  $\alpha(h)$ ,  $h = 0, 1, 2, \dots$ , of a stationary process is defined as follows:

$$\begin{aligned}\alpha(0) &= 1 \\ \alpha(1) &= \text{corr}(X_2, X_1) = \rho(1) \\ \alpha(h) &= \text{corr}[X_{h+1} - \mathbb{P}(X_{h+1}|1, X_2, \dots, X_h), X_1 - \mathbb{P}(X_1|1, X_2, \dots, X_h)],\end{aligned}$$

where  $\mathbb{P}(X_{h+1}|1, X_2, \dots, X_h)$  and  $\mathbb{P}(X_1|1, X_2, \dots, X_h)$  denote the best, in the sense mean squared forecast errors, linear forecasts of  $X_{h+1}$ , respectively  $X_1$  given  $\{1, X_2, \dots, X_h\}$ .

**Remark 3.4.** If  $\{X_t\}$  has a mean of zero, then the constant in the projection operator can be omitted.

The first definition implies that the partial autocorrelations are determined from the coefficients of the forecasting function which are themselves functions of the autocorrelation coefficients. It is therefore possible to express the partial autocorrelations as a function of the autocorrelations. More specifically, the partial autocorrelation functions can be computed recursively from the autocorrelation function according to Durbin-Levinson algorithm (Durbin, 1960):

$$\begin{aligned}\alpha(0) &= 1 \\ \alpha(1) &= a_{11} = \rho(1) \\ \alpha(2) &= a_{22} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} \\ &\dots \\ \alpha(h) &= a_{hh} = \frac{\rho(h) - \sum_{j=1}^{h-1} a_{h-1,j} \rho_{h-j}}{1 - \sum_{j=1}^{h-1} a_{h-1,j} \rho_j},\end{aligned}$$

where  $a_{h,j} = a_{h-1,j} - a_{hh} a_{h-1,h-j}$  for  $j = 1, 2, \dots, h-1$ .

### Autoregressive processes

The idea of the PACF can be well illustrated in the case of an AR(1) process

$$X_t = \phi X_{t-1} + Z_t \quad \text{with } 0 < |\phi| < 1 \text{ and } Z_t \sim \text{WN}(0, \sigma^2).$$

As shown in Chapter 2,  $X_t$  and  $X_{t-2}$  are correlated with each other despite the fact that there is no direct relationship between the two. The correlation is obtained “indirectly” because  $X_t$  is correlated with  $X_{t-1}$  which is itself

correlated with  $X_{t-2}$ . Because both correlation are equal to  $\phi$ , the correlation between  $X_t$  and  $X_{t-2}$  is equal to  $\rho(2) = \phi^2$ . The ACF therefore accounts for all correlation, including the indirect ones. The partial autocorrelation on the other hand only accounts for the direct relationships. In the case of the AR(1) process, there is only an indirect relation between  $X_t$  and  $X_{t-h}$  for  $h \geq 2$ , thus the PACF is zero.

Based on the results in Section 3.1 for the AR(1) process, the definition 3.2 of the PACF implies:

$$\begin{aligned}\alpha_1 &= \phi && \Rightarrow \alpha(1) = \rho(1) = \phi, \\ \alpha_2 &= (\phi, 0)' && \Rightarrow \alpha(2) = 0, \\ \alpha_3 &= (\phi, 0, 0)' && \Rightarrow \alpha(3) = 0.\end{aligned}$$

The partial autocorrelation function of an AR(1) process is therefore equal to zero for  $h \geq 2$ .

This logic can be easily generalized. The PACF of a causal AR(p) process is equal to zero for  $h > p$ , i.e.  $\alpha(h) = 0$  for  $h > p$ . This property characterizes an AR(p) process as shown in the next section.

### Moving average processes

Consider now the case of an invertible MA process. For this process we have:

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \quad \Rightarrow \quad X_t = - \sum_{j=1}^{\infty} \pi_j X_{t-j} + Z_t.$$

$X_t$  is therefore “directly” correlated with each  $X_{t-h}$ ,  $h = 1, 2, \dots$ . Consequently, the PACF is never exactly equal to zero, but converges exponentially to zero. This convergence can be monotonic or oscillating.

Take the MA(1) process as an illustration:

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

The computations in Section 3.1 showed that

$$\begin{aligned}\alpha_1 &= \frac{\theta}{1 + \theta^2} && \Rightarrow \alpha(1) = \rho(1) = \frac{\theta}{1 + \theta^2}, \\ \alpha_2 &= \left( \frac{\theta(1 + \theta^2)}{1 + \theta^2 + \theta^4}, \frac{-\theta^2}{1 + \theta^2 + \theta^4} \right)' && \Rightarrow \alpha(2) = \frac{-\theta^2}{1 + \theta^2 + \theta^4}.\end{aligned}$$

Thus we get for the MA(1) process:

$$\alpha(h) = - \frac{(-\theta)^h}{1 + \theta^2 + \dots + \theta^{2h}} = - \frac{(-\theta)^h(1 - \theta^2)}{1 - \theta^{2(h+1)}}$$

Table 3.2: Properties of the ACF and the PACF

processes	ACF	PACF
AR(p)	declines exponentially (monotonically or oscillating) to zero	$\alpha(h) = 0$ for $h > p$
MA(q)	$\rho(h) = 0$ for $h > q$	declines exponentially (monotonically or oscillating) to zero

### 3.5.2 Interpretation of ACF and PACF

The ACF and the PACF are two important tools in determining the nature of the underlying mechanism of a stochastic process. In particular, they can be used to determine the orders of the underlying AR, respectively MA processes. The analysis of ACF and PACF to identify appropriate models is known as the Box-Jenkins methodology (Box and Jenkins, 1976). Table 3.2 summarizes the properties of both tools for the case of a causal AR and an invertible MA process.

If  $\{X_t\}$  is a causal and invertible ARMA(p,q) process, we have the following properties. As shown in Section 2.4, the ACF is characterized for  $h > \max\{p, q + 1\}$  by the homogeneous difference equation  $\rho(h) = \phi_1\rho(h - 1) + \dots + \phi_p\rho(h - p)$ . Causality implies that the roots of the characteristic equation are all inside the unit circle. The autocorrelation coefficients therefore decline exponentially to zero. Whether this convergence is monotonic or oscillating depends on the signs of the roots. The PACF starts to decline to zero for  $h > p$ . Thereby the coefficients of the PACF exhibit the same behavior as the autocorrelation coefficients of  $\theta^{-1}(L)X_t$ .

## 3.6 Exercises

**Exercise 3.6.1.** *Assign the ACF and the PACF from Figure 3.1 to the following processes:*

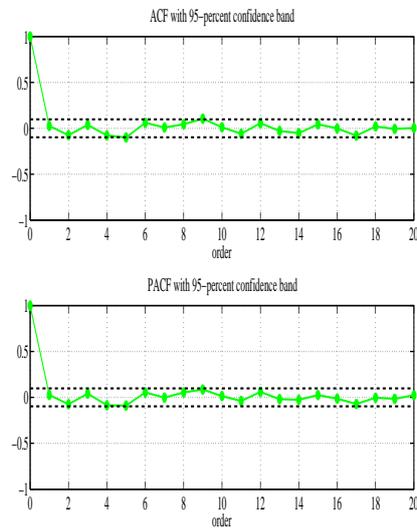
$$X_t = Z_t,$$

$$X_t = 0.9X_{t-1} + Z_t,$$

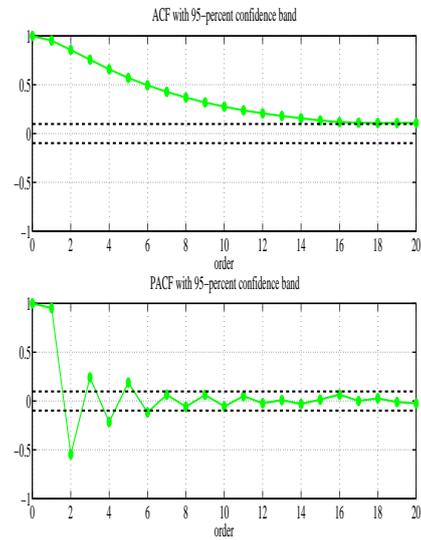
$$X_t = Z_t + 0.8Z_{t-1},$$

$$X_t = 0.9X_{t-1} + Z_t + 0.8Z_{t-1}$$

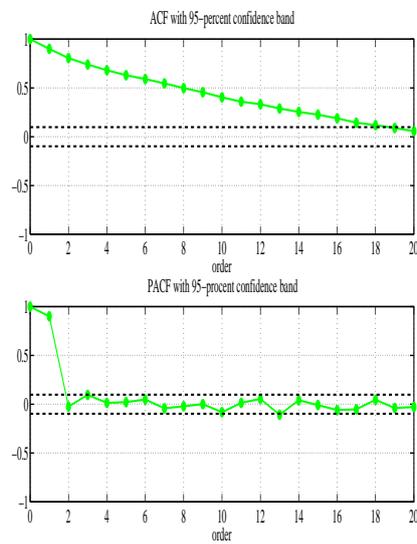
with  $Z_t \sim \text{WN}(0, \sigma^2)$ .



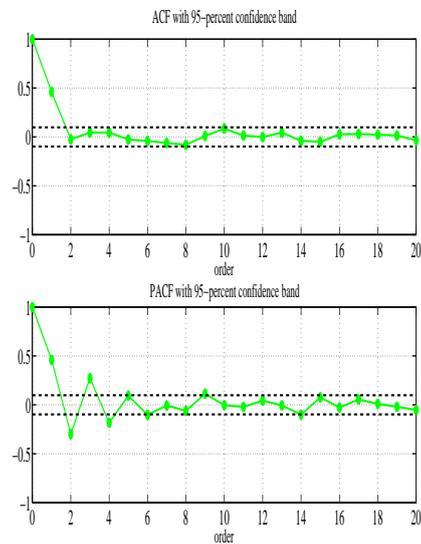
(a) Process 1



(b) Process 2



(c) Process 3



(d) Process 4

Figure 3.1: Autocorrelation and partial autocorrelation functions

# Chapter 4

## Estimation of the Mean and the Autocorrelation Function

In the previous chapters we have seen in which way the mean  $\mu$ , and, more importantly, the autocovariance function,  $\gamma(h)$ ,  $h = 0, \pm 1, \pm 2, \dots$ , of a stationary stochastic process  $\{X_t\}$  characterize its dynamic properties, at least if we restrict ourselves to the first two moments. In particular, we have investigated how the autocovariance function is related to the coefficients of the corresponding ARMA process. Thus the estimation of the ACF is not only interesting for its own sake, but also for the specification and identification of appropriate ARMA models. It is therefore of outmost importance to have reliable (consistent) estimators for these entities. Moreover, we want to test specific features for a given time series. This means that we have to develop corresponding testing theory. As the small sample distributions are hard to get, we rely for this purpose on asymptotic theory.<sup>1</sup>

In this section we will assume that the process is stationary and observed for the time periods  $t = 1, 2, \dots, T$ . We will refer to  $T$  as the sample size. As mentioned previously, the standard sampling theory is not appropriate in the times series context because the  $X_t$ 's are not independent draws from some underlying distribution, but are systematically related to each other.

### 4.1 Estimation of the mean

The arithmetic average constitutes a “natural” estimator of the mean  $\mu$  of the stochastic process. The arithmetic mean  $\bar{X}_T$  is defined as usual by

$$\bar{X}_T = \frac{1}{T} (X_1 + X_2 + \dots + X_T).$$

---

<sup>1</sup>Recently, bootstrap methods have also been introduced in the time series context.

It is immediately clear that arithmetic average is an unbiased estimator of the mean:

$$\mathbb{E}\bar{X}_T = \frac{1}{T} (\mathbb{E}X_1 + \mathbb{E}X_2 + \dots + \mathbb{E}X_T) = \mu.$$

Of greater interest are the asymptotic properties of the variance of the arithmetic mean  $\mathbb{V}\bar{X}_T$  which are summarized in the following theorem:

**Theorem 4.1** (Convergence of arithmetic average). *If  $\{X_t\}$  is a stationary stochastic process with mean  $\mu$  and ACF  $\gamma(h)$  then the variance of the arithmetic mean  $\mathbb{V}\bar{X}_T$  has the following asymptotic properties:*

$$\begin{aligned} \mathbb{V}\bar{X}_T &= \mathbb{E}(\bar{X}_T - \mu)^2 \rightarrow 0, & \text{if } \gamma(T) \rightarrow 0; \\ T\mathbb{V}\bar{X}_T &= T\mathbb{E}(\bar{X}_T - \mu)^2 \rightarrow \sum_{h=-\infty}^{\infty} \gamma(h), & \text{if } \sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \end{aligned}$$

for  $T$  going to infinity.

*Proof.* Immediate algebra establishes:

$$\begin{aligned} 0 \leq T\mathbb{V}\bar{X}_T &= \frac{1}{T} \sum_{i,j=1}^T \text{cov}(X_i, X_j) = \sum_{|h|<T} \left(1 - \frac{|h|}{T}\right) \gamma(h) \\ &\leq \sum_{|h|<T} |\gamma(h)| = 2 \sum_{h=1}^T |\gamma(h)| + \gamma(0). \end{aligned}$$

The assumption  $\gamma(h) \rightarrow 0$  for  $h \rightarrow \infty$  implies that for any given  $\varepsilon > 0$ , we can find  $T_0$  such that  $|\gamma(h)| < \varepsilon/2$  for  $h \geq T_0$ . If  $T > T_0$  and  $T > 2T_0 \gamma(0)/\varepsilon$  then

$$\begin{aligned} 0 \leq \frac{1}{T} \sum_{h=1}^T |\gamma(h)| &= \frac{1}{T} \sum_{h=1}^{T_0-1} |\gamma(h)| + \frac{1}{T} \sum_{h=T_0}^T |\gamma(h)| \\ &\leq \frac{T_0 \gamma(0)}{T} + \frac{1}{T} (T - T_0) \varepsilon/2 \leq \frac{T_0 \gamma(0)}{T} + \varepsilon/2 \leq \frac{T_0 \gamma(0) \varepsilon}{2T_0 \gamma(0)} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Therefore  $\mathbb{V}\bar{X}_T$  converges to zero for  $T \rightarrow \infty$  which establishes the first property. Moreover, we have

$$\lim_{T \rightarrow \infty} T\mathbb{V}\bar{X}_T = \lim_{T \rightarrow \infty} \sum_{|h|<T} \left(1 - \frac{|h|}{T}\right) \gamma(h) = \sum_{h=-\infty}^{\infty} \gamma(h) < \infty.$$

The infinite sum  $\sum_{h=-\infty}^{\infty} \gamma(h)$  converges because it converges absolutely by assumption.  $\square$

This Theorem establishes that the arithmetic average is not only an unbiased estimator of the mean, but also a consistent one. In particular, the arithmetic average converges in the mean-square sense, and therefore also in probability, to the true mean (see appendix C). This result can be interpreted as a reflection of the concept of ergodicity (see Section 1.2). The assumptions are relatively mild and are fulfilled for the ARMA processes because for these processes  $\gamma(h)$  converges exponentially fast to zero (see Section 2.4.2, in particular equation (2.4)). Under little more restrictive assumptions it is even possible to show that the arithmetic mean is asymptotically normally distributed.

**Theorem 4.2** (Asymptotic distribution of sample mean). *For any stationary process  $\{X_t\}$  given by*

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

*such that  $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$  and  $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$ , the arithmetic average  $\bar{X}_T$  is asymptotically normal:*

$$\begin{aligned} \sqrt{T}(\bar{X}_T - \mu) &\xrightarrow{d} N\left(0, \sum_{h=-\infty}^{\infty} \gamma(h)\right) \\ &= N\left(0, \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2\right) = N(0, \sigma^2 \Psi(1)^2) \end{aligned}$$

where  $\gamma$  is the autocovariance function of  $\{X_t\}$ .

*Proof.* The standard proof uses the Basic Approximation Theorem C.14 and the Central Limit Theorem for m-dependent processes C.13. To this end we define the approximate process

$$X_t^{(m)} = \mu + \sum_{j=-m}^m \psi_j Z_{t-j}.$$

Clearly,  $\{X_t^{(m)}\}$  is a  $2m$ -dependent process with  $V_m = \sum_{h=-m}^m \gamma(h) = \sigma^2 (\sum_{j=-m}^m \psi_j)^2$ . This last assertion can be verified by noting that

$$\begin{aligned} V &= \sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \\ &= \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \sum_{h=-\infty}^{\infty} \psi_{j+h} = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2. \end{aligned}$$

Note that the assumption  $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$  guarantees the convergence of the infinite sums. Applying this result to the special case  $\psi_j = 0$  for  $|j| > m$ , we obtain  $V_m$ .

The arithmetic average of the approximating process is

$$\bar{X}_T^{(m)} = \frac{1}{T} \sum_{t=1}^T X_t^{(m)}.$$

The CLT for  $m$ -dependent processes C.13 then implies that for  $T \rightarrow \infty$

$$\sqrt{T} \left( \bar{X}_T^{(m)} - \mu \right) \xrightarrow{d} X^{(m)} = N(0, V_m).$$

As  $m \rightarrow \infty$ ,  $\sigma^2(\sum_{j=-m}^m \psi_j)^2$  converges to  $\sigma^2(\sum_{j=-\infty}^{\infty} \psi_j)^2$  and thus

$$X^{(m)} \xrightarrow{d} X = N(0, V) = N \left( 0, \sigma^2 \left( \sum_{j=-\infty}^{\infty} \psi_j \right)^2 \right).$$

This assertion can be established by noting that the characteristic functions of  $X^{(m)}$  approaches the characteristic function of  $X$  so that by Theorem C.11  $X^{(m)} \xrightarrow{d} X$ .

Finally, we show that the approximation error becomes negligible as  $T$  goes to infinity:

$$\sqrt{T} (\bar{X}_T - \mu) - \sqrt{T} (\bar{X}_T^{(m)} - \mu) = T^{-1/2} \sum_{t=1}^T (X_t - X_t^{(m)}) = T^{-1/2} \sum_{t=1}^T e_t^{(m)}$$

where the error  $e_t^{(m)}$  is

$$e_t^{(m)} = \sum_{|j|>m} \psi_j Z_{t-j}.$$

Clearly,  $\{e_t^{(m)}\}$  is a stationary process with autocovariance function  $\gamma_e$  such that  $\sum_{h=-\infty}^{\infty} \gamma_e(h) = \sigma^2 \left( \sum_{|j|>m} \psi_j \right)^2 < \infty$ . We can therefore invoke Theorem 4.1 to show that

$$\mathbb{V} \left( \sqrt{T} (\bar{X}_T - \mu) - \sqrt{T} (\bar{X}_T^{(m)} - \mu) \right) = T \mathbb{V} \left( \frac{1}{T} \sum_{t=1}^T e_t^{(m)} \right)$$

converges to  $\sigma^2 \left( \sum_{|j|>m} \psi_j \right)^2$  as  $T \rightarrow \infty$ . This term converges to zero as  $m \rightarrow \infty$ . The approximation error  $\sqrt{T} (\bar{X}_T - \mu) - \sqrt{T} (\bar{X}_T^{(m)} - \mu)$  therefore

converges in mean square to zero and thus, using Chebyshev's inequality (see Theorem C.3 or C.7), also in probability. We have therefore established the third condition of Theorem C.14 as well. Thus, we can conclude that  $\sqrt{T}(\bar{X}_T - \mu) \xrightarrow{d} X$ .  $\square$

Under a more restrictive summability condition which holds, however, within the context of causal ARMA processes, we can provide a less technical proof. This proof follows an idea of Phillips and Solo (1992) and is based on the Beveridge-Nelson decomposition (see Appendix D).<sup>2</sup>

**Theorem 4.3.** *For any stationary process*

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

with the properties  $Z_t \sim \text{IID}(0, \sigma^2)$  and  $\sum_{j=0}^{\infty} j^2 |\psi_j|^2 < \infty$ , the arithmetic average  $\bar{X}_T$  is asymptotically normal:

$$\begin{aligned} \sqrt{T}(\bar{X}_T - \mu) &\xrightarrow{d} N\left(0, \sum_{h=-\infty}^{\infty} \gamma(h)\right) \\ &= N\left(0, \sigma^2 \left(\sum_{j=0}^{\infty} \psi_j\right)^2\right) = N(0, \sigma^2 \Psi(1)^2). \end{aligned}$$

*Proof.* The application of the Beveridge-Nelson decomposition (see Theorem D.1 in Appendix D) leads to

$$\begin{aligned} \bar{X}_T - \mu &= \frac{1}{T} \sum_{t=1}^T \Psi(L) Z_t = \frac{1}{T} \sum_{t=1}^T (\Psi(1) - (L-1)) \tilde{\Psi}(L) Z_t \\ &= \Psi(1) \left( \frac{1}{T} \sum_{t=1}^T Z_t \right) + \frac{1}{T} \tilde{\Psi}(L) (Z_0 - Z_T) \\ \sqrt{T}(\bar{X}_T - \mu) &= \Psi(1) \left( \sqrt{T} \frac{\sum_{t=1}^T Z_t}{T} \right) + \frac{1}{\sqrt{T}} \tilde{\Psi}(L) Z_0 - \frac{1}{\sqrt{T}} \tilde{\Psi}(L) Z_T. \end{aligned}$$

The assumption  $Z_t \sim \text{IID}(0, \sigma^2)$  allows to apply the Central Limit Theorem C.12 of Appendix C to the first term. Thus,  $\sqrt{T} \frac{\sum_{t=1}^T Z_t}{T}$  is asymptotical normal with mean zero and variance  $\sigma^2$ . Theorem D.1 also implies

---

<sup>2</sup>The Beveridge-Nelson decomposition is an indispensable tool for the understanding of integrated and cointegrated processes analyzed in Chapters 7 and 16.

$|\Psi(1)| < \infty$ . Therefore, the term  $\Psi(1)\sqrt{T} \frac{\sum_{t=1}^T Z_t}{T}$  is asymptotically normal with mean zero and variance  $\sigma^2\Psi(1)^2$ .

The variances of the second and third term are equal to  $\frac{\sigma^2}{T} \sum_{j=0}^T \tilde{\psi}_j^2$ . The summability condition then implies according to Theorem D.1 that  $\sum_{j=0}^T \tilde{\psi}_j^2$  converges for  $T \rightarrow \infty$ . Thus, the variances of the last two terms converge to zero implying that these terms converge also to zero in probability (see Theorem C.7) and thus also in distribution. We can then invoke Theorem C.10 to establish the Theorem. Finally, the equality of  $\sum_{h=-\infty}^{\infty} \gamma(h)$  and  $\sigma^2\Psi(1)^2$  can be obtained from direct computations or by the application of Theorem 6.4.  $\square$

**Remark 4.1.** *Theorem 4.2 holds for any causal ARMA process as the  $\psi_j$ 's converge exponentially fast to zero (see the discussion following equation (2.3)).*

**Remark 4.2.** *If  $\{X_t\}$  is a Gaussian process, then for any given fixed  $T$ ,  $\bar{X}_T$  is distributed as*

$$\sqrt{T}(\bar{X}_T - \mu) \sim N\left(0, \sum_{|h| < T} \left(1 - \frac{|h|}{T}\right) \gamma(h)\right).$$

According to Theorem 4.2, the asymptotic variance of the average depends on the sum of all covariances  $\gamma(h)$ . This entity, denoted by  $J$ , is called the *long-run variance* of  $\{X_t\}$ :

$$J = \sum_{h=-\infty}^{\infty} \gamma(h) = \gamma(0) \left(1 + 2 \sum_{h=1}^{\infty} \rho(h)\right). \quad (4.1)$$

Note that the the long-run variance equals  $2\pi$  times the spectral density  $f(\lambda)$  evaluated at  $\lambda = 0$  (see the Definition 6.1 of the spectral density in Section 6.1).

As the long-run variance takes into account the serial properties of the time series, it is also called heteroskedastic and autocorrelation consistent variance (HAC variance). If  $\{X_t\}$  has some nontrivial autocorrelation (i.e.  $\rho(h) \neq 0$  for  $h \neq 0$ ), the long-run variance  $J$  is different from  $\gamma(0)$ . This implies among other things that the construction of the t-statistic for testing the simple hypothesis  $H_0: \mu = \mu_0$  should use  $J$  instead of  $\gamma(0)$ .

In case that  $\{X_t\}$  is a causal ARMA process with  $\Phi(L)X_t = \Theta(L)Z_t$ ,  $Z_t \sim \text{WN}(0, \sigma^2)$ , the long-run variance is given by

$$J = \left(\frac{\Theta(1)}{\Phi(1)}\right)^2 \sigma^2 = \Psi(1)^2 \sigma^2.$$

If  $\{X_t\}$  is a AR(1) process with  $X_t = \phi X_{t-1} + Z_t$ ,  $Z_t \sim \text{WN}(0, \sigma^2)$  and  $|\phi| < 1$ ,  $\gamma(0) = \frac{\sigma^2}{1-\phi^2}$  and  $\rho(h) = \phi^{|h|}$ . Thus the long-run variance is given by  $J = \frac{\sigma^2}{(1-\phi)^2} = \gamma(0) \times \frac{1+\phi}{1-\phi}$ . From this example it is clear that the long-run variance can be smaller or larger than  $\gamma(0)$ , depending on the sign of  $\phi$ : for negative values of  $\phi$ ,  $\gamma(0)$  overestimates the long-run variance; for positive values, it underestimates  $J$ . The estimation of the long-run variance is dealt with in Section 4.4.

## 4.2 Estimation of the autocovariance and the autocorrelation function

With some slight, asymptotically unimportant modifications, we can use the standard estimators for the autocovariances,  $\gamma(h)$ , and the autocorrelations,  $\rho(h)$ , of a stationary stochastic process:

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (X_t - \bar{X}_T) (X_{t+h} - \bar{X}_T), \quad (4.2)$$

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \quad (4.3)$$

These estimators are biased because the sums are normalized (divided) by  $T$  rather than  $T-h$ . The normalization with  $T-h$  delivers an unbiased estimate only if  $\bar{X}_T$  is replaced by  $\mu$  which, however is typically unknown in practice. The second modification concerns the use of the complete sample for the estimation of  $\mu$ .<sup>3</sup> The main advantage of using the above estimators is that the implied estimator for the covariance matrix,  $\hat{\Gamma}_T$ , respectively the autocorrelation matrix,  $\hat{R}_T$ , of  $(X_1, \dots, X_T)'$ ,

$$\hat{\Gamma}_T = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \dots & \hat{\gamma}(T-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \dots & \hat{\gamma}(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(T-1) & \hat{\gamma}(T-2) & \dots & \hat{\gamma}(0) \end{pmatrix}$$

$$\hat{R}_T = \frac{\hat{\Gamma}_T}{\hat{\gamma}(0)}$$

---

<sup>3</sup>The standard statistical formulas would suggest estimate the mean appearing in first multiplicand from  $X_1, \dots, X_{t-h}$ , and the mean appearing in the second multiplicand from  $X_{h+1}, \dots, X_T$ .

always delivers, independently of the realized observations, non-negative definite and for  $\widehat{\gamma}(0) > 0$  non-singular matrices. The resulting estimated autocovariance function will then satisfy the characterization given in Theorem 1.1, in particular property (iv).

According to Box and Jenkins (1976, p. 33), one can expect reasonable estimates for  $\gamma(h)$  and  $\rho(h)$  if the sample size is larger than 50 and if the order of the autocorrelation coefficient is smaller than  $T/4$ .

The theorem below establishes that these estimators lead under rather mild conditions to consistent and asymptotically normally distributed estimators.

**Theorem 4.4.** *Let  $\{X_t\}$  be the stationary process*

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

with  $Z_t \sim \text{IID}(0, \sigma^2)$ ,  $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$  and  $\sum_{j=-\infty}^{\infty} j|\psi_j|^2 < \infty$ . Then we have for  $h = 1, 2, \dots$

$$\begin{pmatrix} \widehat{\rho}(1) \\ \vdots \\ \widehat{\rho}(h) \end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} \rho(1) \\ \vdots \\ \rho(h) \end{pmatrix}, \frac{W}{T} \right)$$

where the elements of  $W = (w_{ij})_{i,j \in \{1, \dots, h\}}$  are given Bartlett's formula

$$w_{ij} = \sum_{k=1}^{\infty} [\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)][\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)].$$

*Proof.* Brockwell and Davis (1991, section 7.3) □

Brockwell and Davis (1991) offer a second version of the above theorem where  $\sum_{j=-\infty}^{\infty} j|\psi_j|^2 < \infty$  is replaced by the assumption of finite fourth moments, i.e. by assuming  $\mathbb{E}Z_t^4 < \infty$ . As we rely mainly on ARMA processes, we do not pursue this distinction further because this class of process automatically fulfills the above assumptions as soon as  $\{Z_t\}$  is identically and independently distributed (IID). A proof which relies on the Beveridge-Nelson polynomial decomposition (see Theorem D.1 in Appendix D) can be gathered from Phillips and Solo (1992).

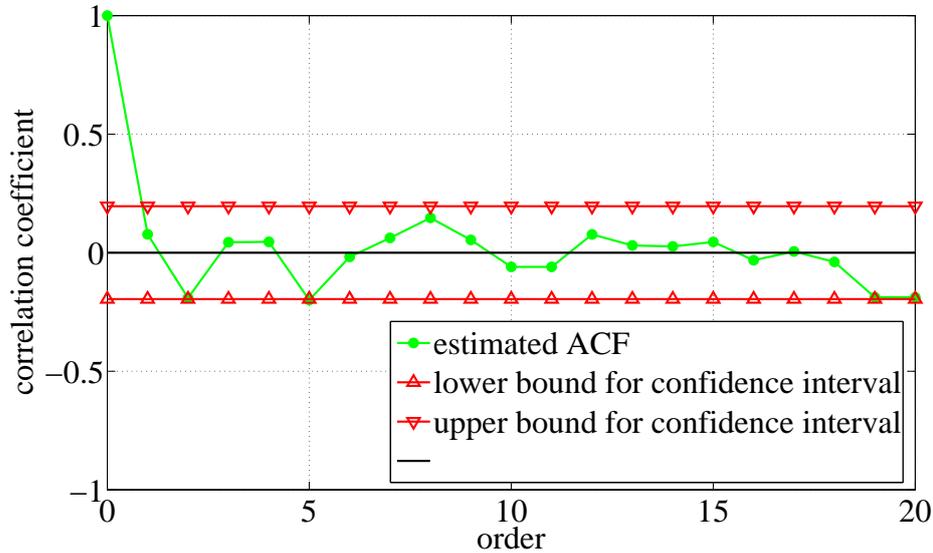


Figure 4.1: Estimated autocorrelation function of a  $WN(0,1)$  process with 95 percent confidence interval for sample size  $T = 100$

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

The most important application of the above theorem, is related to the case of a white noise process for which  $\rho(h)$  is equal to zero for  $|h| > 0$ . The above theorem then implies that

$$w_{ij} = \begin{cases} 1, & \text{for } i = j; \\ 0, & \text{otherwise.} \end{cases}$$

The estimated autocorrelation coefficients converge to the true autocorrelation coefficient, in this case zero. The asymptotic distribution of  $\sqrt{T}\hat{\rho}(h)$  converges to the standard normal distribution. This implies that for large  $T$  we can approximate the distribution of  $\hat{\rho}(h)$  by a normal distribution with mean zero and variance  $1/T$ . This allows the construction of a 95 percent confidence interval for the case that the true process is white noise. This confidence interval is therefore given by  $\pm 1.96T^{-\frac{1}{2}}$ . It can be used to verify if the observed process is indeed white noise.

Figure 4.1 plots the empirical autocorrelation function of a  $WN(0,1)$  process with a sample size of  $T = 100$ . The implied 95 percent confidence is therefore equal to  $\pm 0.196$ . As each estimated autocorrelation coefficient falls within the confidence interval so that we can conclude that the observed times series may indeed represent a white noise process.

Instead of examining each correlation coefficient separately, we can test the joint hypothesis that all correlation coefficients up to order  $N$  are equal to zero, i.e.  $\rho(1) = \rho(2) = \dots = \rho(N) = 0$ ,  $N = 1, 2, \dots$ . As each  $\sqrt{T}\hat{\rho}(h)$  has an asymptotic standard normal distribution and uncorrelated with  $\sqrt{T}\hat{\rho}(k)$ ,  $h \neq k$ , the sum of the squared estimated autocorrelation coefficients is  $\chi^2$  distributed with  $N$  degrees of freedom. This test statistic is called *Box-Pierce statistic*:

$$Q = T \sum_{h=1}^N \hat{\rho}^2(h) \sim \chi_N^2.$$

A refinement of this test statistic is given by the *Ljung-Box statistic*:

$$Q' = T(T+2) \sum_{h=1}^N \frac{\hat{\rho}^2(h)}{T-h} \sim \chi_N^2. \quad (4.4)$$

This test statistic is also asymptotically  $\chi^2$  distributed with the same degree of freedom  $N$ . This statistic accounts for the fact that the estimates for high orders  $h$  are based on a smaller number of observations and are thus less precise and more noisy. The two test statistics are used in the usual way. The null hypothesis that all correlation coefficients are jointly equal to zero is rejected if  $Q$ , respectively  $Q'$  is larger than the critical value corresponding the  $\chi_N^2$  distribution. The number of summands  $N$  is usually taken to be rather large, for a sample size of 150 in the range between 15 and 20. The two test are also referred to as *Portemanteau tests*.

**Example: MA(q) process:**  $X_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$  with  $Z_t \sim \text{IID}(0, \sigma^2)$

In this case the covariance matrix is determined as

$$w_{ii} = 1 + 2\rho(1)^2 + \dots + 2\rho(q)^2 \quad \text{for } i > q.$$

For  $i, j > q$ ,  $w_{ij}$  is equal to zero. The 95 percent confidence interval for the MA(1) process  $X_t = Z_t - 0.8Z_{t-1}$  is therefore given for a sample size of  $T = 200$  by  $\pm 1.96T^{-\frac{1}{2}}[1 + 2\rho(1)^2]^{\frac{1}{2}} = \pm 0.1684$ .

Figure 4.2 shows the estimated autocorrelation function of the above MA(1) process together with 95 percent confidence interval based on a white noise process and a MA(1) process with  $\theta = -0.8$ . As the first order autocorrelation coefficient is clearly outside the confidence interval whereas all other autocorrelation coefficients are inside it, the figure demonstrate that the observations are evidently the realization of MA(1) process.

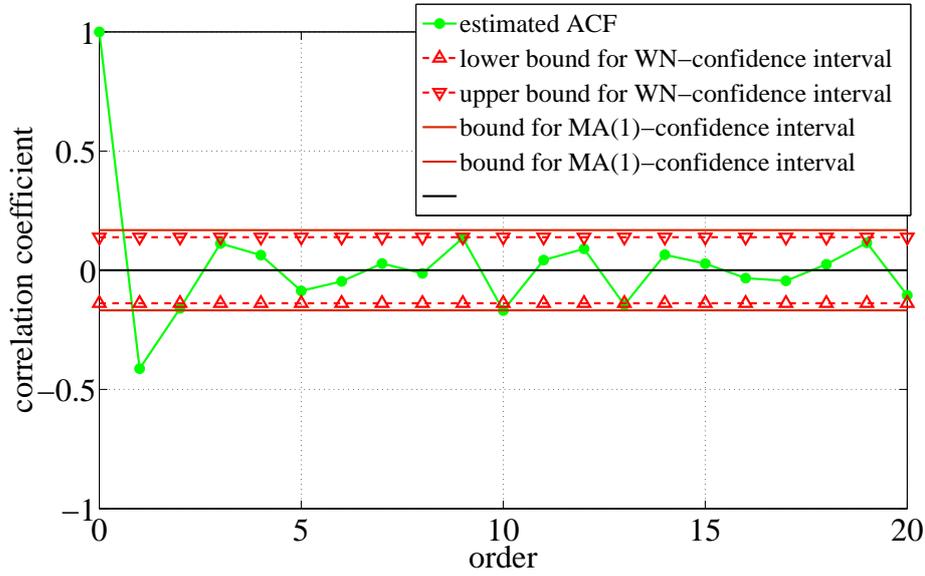


Figure 4.2: Estimated autocorrelation function of a MA(1) process with  $\theta = -0.8$  with corresponding 95 percent confidence interval for  $T = 200$

**Example: AR(1) process**  $X_t - \phi X_{t-1} = Z_t$  with  $Z_t \sim \text{IID}(0, \sigma^2)$

In this case the covariance matrix is determined as

$$\begin{aligned} w_{ii} &= \sum_{k=1}^i \phi^{2i} (\phi^k - \phi^{-k})^2 + \sum_{k=i+1}^{\infty} \phi^{2k} (\phi^i - \phi^{-i})^2 \\ &= \frac{(1 - \phi^{2i})(1 + \phi^2)}{1 - \phi^2} - 2i\phi^{2i} \\ &\approx \frac{1 + \phi^2}{1 - \phi^2} \quad \text{for large } i. \end{aligned}$$

The formula for  $w_{ij}$  with  $i \neq j$  are not shown. In any case, this formula is of relatively little importance because the partial autocorrelations are better suited for the identification of AR processes (see Section 3.5 and 4.3).

Figure 4.3 shows an estimated autocorrelation function of an AR(1) process. The autocorrelation coefficients decline exponentially which is a characteristic for an AR(1) process.<sup>4</sup> Furthermore the coefficients are outside the confidence interval up to order 8 for white noise processes.

<sup>4</sup>As a reminder: the theoretical autocorrelation coefficients are  $\rho(h) = \phi^{|h|}$ .

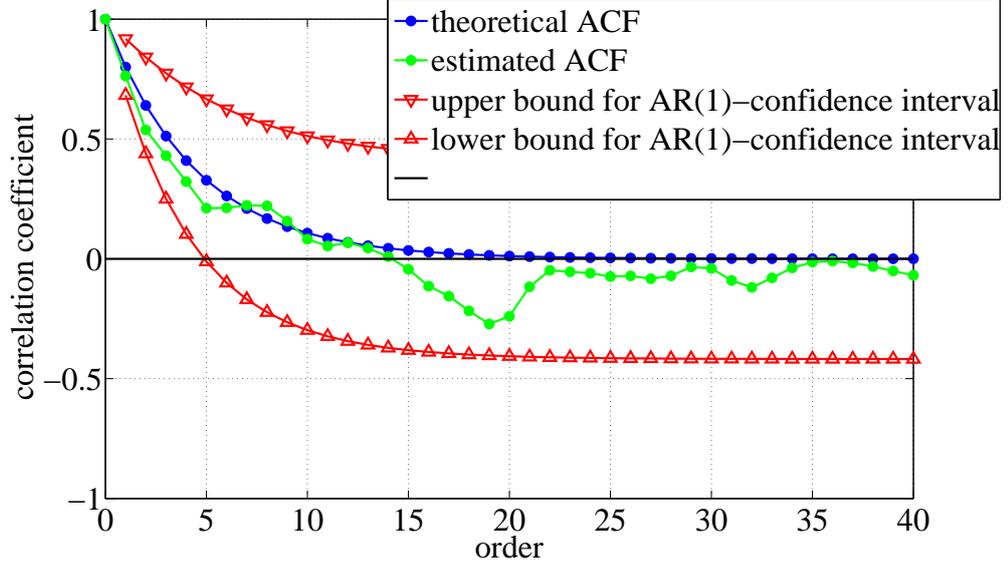


Figure 4.3: Estimated autocorrelation function of an AR(1) process with  $\phi = 0.8$  and corresponding 95 percent confidence interval for  $T = 100$

### 4.3 Estimation of the partial autocorrelation function

According to its definition (see definition 3.2), the partial autocorrelation of order  $h$ ,  $\alpha(h)$ , is equal to  $a_h$ , the last element of the vector  $\alpha_h = \Gamma_h^{-1}\gamma_h(1) = R_h^{-1}\rho_h(1)$ . Thus  $\alpha_h$  and consequently  $a_h$  can be estimated by  $\hat{\alpha}_h = \hat{\Gamma}_h^{-1}\hat{\gamma}_h(1) = \hat{R}_h^{-1}\hat{\rho}_h(1)$ . As  $\rho(h)$  can be consistently estimated and is asymptotically normally distributed (see Section 4.2), the *continuous mapping theorem* (see Appendix C) ensures that the above estimator for  $\alpha(h)$  is also consistent and asymptotically normal. In particular we have for an AR( $p$ ) process (Brockwell and Davis, 1991)

$$\sqrt{T}\hat{\alpha}(h) \xrightarrow{d} N(0, 1) \quad \text{for } T \rightarrow \infty \text{ and } h > p.$$

This result allows to construct, as in the case of the autocorrelation coefficients, confidence intervals for the partial autocorrelations coefficients. The 95 percent confidence interval is given by  $\pm \frac{1.96}{\sqrt{T}}$ . The AR( $p$ ) process is characterized by the fact that the partial autocorrelation coefficients are zero for  $h > p$ .  $\hat{\alpha}(h)$  should therefore be inside the confidence interval for  $h > p$  and outside for  $h \leq p$ . Figure 4.4 confirms this for an AR(1) process with  $\phi = 0.8$ .

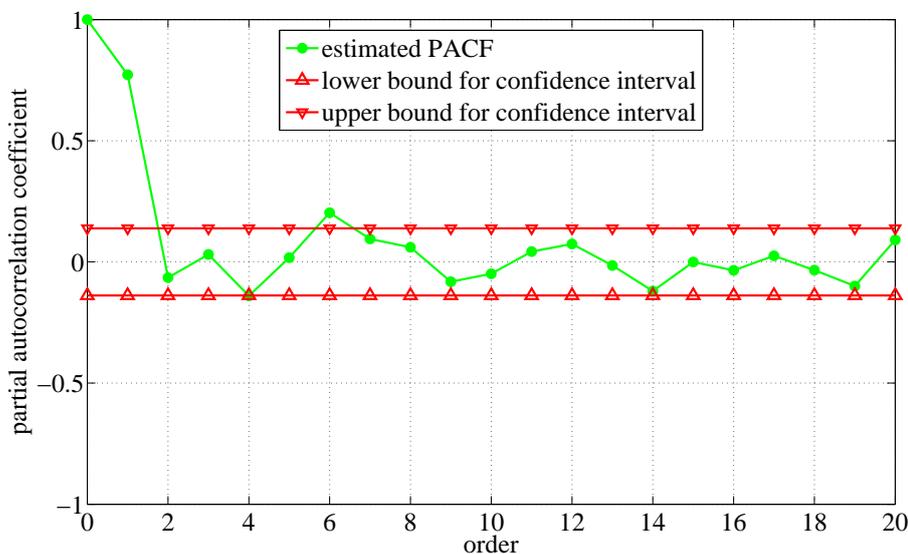


Figure 4.4: Estimated PACF for an AR(1) process with  $\phi = 0.8$  and corresponding 95 percent confidence interval for  $T = 200$

Figure 4.5 shows the estimated PACF for an MA(1) process with  $\theta = 0.8$ . In conformity with the theory, the partial autocorrelation coefficients converge to zero. They do so in an oscillating manner because  $\theta$  is positive (see formula in Section 3.5)

## 4.4 Estimation of the long-run variance

For many applications<sup>5</sup> it is necessary to estimate the long-run variance  $J$  which is defined according to equation (4.1) as follows<sup>6</sup>

$$J = \sum_{h=-\infty}^{\infty} \gamma(h) = \gamma(0) + 2 \sum_{h=1}^{\infty} \gamma(h) = \gamma(0) \left( 1 + 2 \sum_{h=1}^{\infty} \rho(h) \right). \quad (4.5)$$

This can, in principle, be done in two different ways. The first one consists in the estimation of an ARMA model which is then used to derive the implied covariances as explained in Section 2.4 which are then inserted into

<sup>5</sup>For example, when testing the null hypothesis  $H_0: \mu = \mu_0$  in the case of serially correlated observations (see Section 4.1); for the Phillips-Perron unit-root test explained in Section 7.3.2.

<sup>6</sup>See Theorem 4.2 and the comments following it.

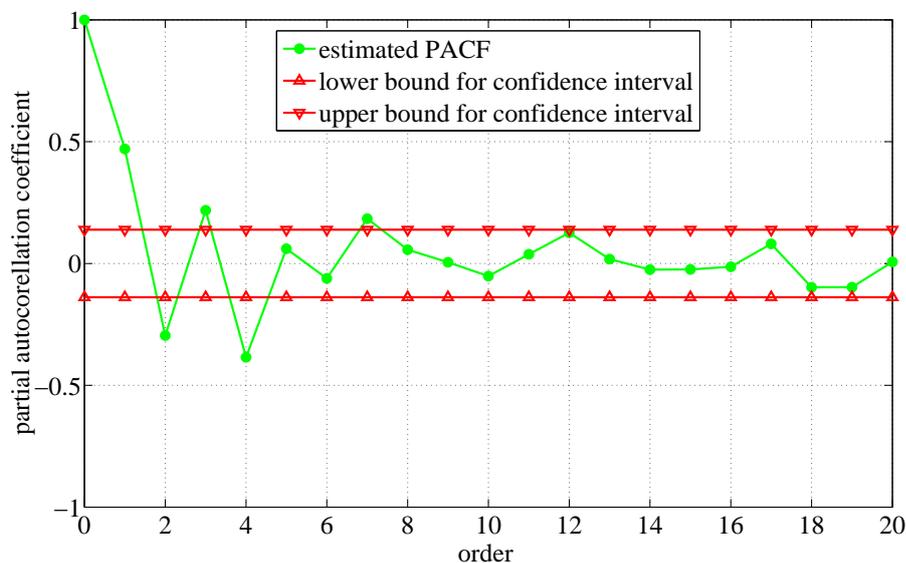


Figure 4.5: Estimated PACF for a MA(1) process with  $\theta = 0.8$  and corresponding 95 percent confidence interval for  $T = 200$

equation (4.5). The second method is a nonparametric one. It has the advantage that it is not necessary to identify and estimate an appropriate ARMA model, a step which can be cumbersome in practice. Additional and more advanced material on this topic can be found in Andrews (1991), Andrews and Monahan (1992), or among others in Haan and Levin (1997).<sup>7</sup>

If the sample size is  $T$ , only  $T - 1$  covariances can be estimated. Thus, a first naive estimator of  $J$  is given by  $\hat{J}_T$  defined as

$$\hat{J}_T = \sum_{h=-T+1}^{T-1} \hat{\gamma}(h) = \hat{\gamma}(0) + 2 \sum_{h=1}^{T-1} \hat{\gamma}(h) = \hat{\gamma}(0) \left( 1 + 2 \sum_{h=1}^{T-1} \hat{\rho}(h) \right),$$

where  $\hat{\gamma}(h)$  and  $\hat{\rho}(h)$  are the estimators for  $\gamma(h)$  and  $\rho(h)$ , respectively, given in Section 4.2. As the estimators of the higher order autocovariances are based on smaller samples, their estimates can be quite erratic. A remedy for this problem is to use only a certain number  $\ell_T$  of autocovariances and/or to use a weighted sum instead of an unweighted one. This idea leads to the

<sup>7</sup>Note the connection between the long-run variance and the spectral density at frequency zero:  $J = 2\pi f(0)$  where  $f$  is the spectral density function (see Section 6.3).

Table 4.1: Some popular kernel functions

name	$k(x) =$
<b>Boxcar</b> (“truncated”)	1
<b>Bartlett</b>	$1 -  x $
<b>Daniell</b>	$\frac{\sin(\pi x)}{\pi x}$
<b>Tukey-Hanning</b>	$(1 + \cos(\pi x))/2$
<b>Quadratic Spectral</b>	$\frac{25}{12\pi^2 x^2} \left( \frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5) \right)$

The function are, with the exception of the quadratic spectral function, only defined for  $|x| \leq 1$ . Outside this interval they are set to zero.

following class estimators:

$$\hat{J}_T = \hat{J}_T(\ell_T) = \frac{T}{T-r} \sum_{h=-T+1}^{T-1} k\left(\frac{h}{\ell_T}\right) \hat{\gamma}(h),$$

where  $k$  is a *weighting* or *kernel* function.<sup>8</sup> The kernel functions are required to have the following properties:

- (i)  $k : \mathbb{R} \rightarrow [-1, 1]$  is, with the exception of a finite number of points a continuous function. In particular,  $k$  is continuous at  $x = 0$ .
- (ii)  $k$  is quadratically integrable, i.e.  $\int_{\mathbb{R}} k(x)^2 dx < \infty$ ;
- (iii)  $k(0) = 1$ ;
- (iv)  $k(x) = k(-x)$  for all  $x \in \mathbb{R}$ .

The basic idea of the kernel function is to give relatively little weight to the higher order autocovariances and relatively more weight to the smaller order ones. As  $k(0)$  equals one, the variance  $\hat{\gamma}(0)$  receives weight one by construction. The continuity assumption implies that also the covariances of smaller order, i.e. for  $h$  small, receive a weight close to one. The Table 4.1 lists some of the most popular kernel functions used in practice:

<sup>8</sup>Kernel functions are also relevant for spectral estimators. See in particular Section 6.3.

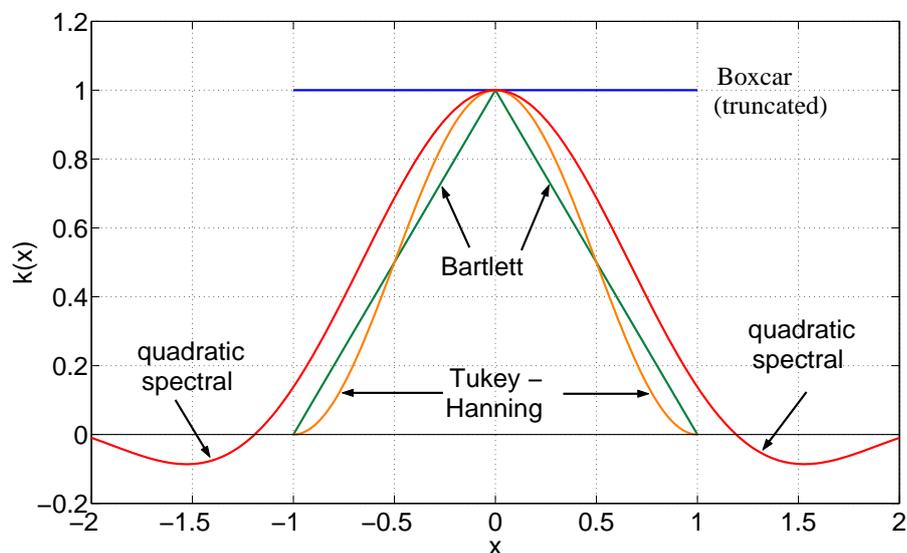


Figure 4.6: Common kernel functions

Figure 4.6 shows a plot of these functions. The first three functions are nonzero only for  $|x| < 1$ . This implies that only the orders  $h$  for which  $|h| \leq \ell_T$  are taken into account.  $\ell_T$  is called the *lag truncation parameter* or the *bandwidth*. The quadratic spectral kernel function is an example of a kernel function which takes all covariances into account. Note that some weights are negative in this case as shown in Figure 4.6.<sup>9</sup>

The estimator for the long-run variance is subject to the correction term  $\frac{T}{T-r}$ . This factor depends on the number of parameters estimated in a first step and is only relevant when the sample size is relatively small. In the case of the estimation of the mean  $r$  would be equal to one and the correction term is negligible. If on the other hand  $X_t, t = 1, \dots, T$ , are the residuals from multivariate regression,  $r$  designates the number of regressors. In many applications the correction term is omitted.

The lag truncation parameter or bandwidth,  $\ell_T$ , depends on the number of observations. It is intuitive that the number of autocovariances accounted for in the computation of the long-run variance should increase with the sample size, i.e. we should have  $\ell_T \rightarrow \infty$  for  $T \rightarrow \infty$ .<sup>10</sup> The relevant issue is, at

<sup>9</sup>Phillips (2004) has proposed a nonparametric regression-based method which does not require a kernel function.

<sup>10</sup>This is true even when the underlying process is known to be a MA( $q$ ) process. Even in this case it is advantageous to include also the autocovariances for  $h > q$ . The reason is twofold. First, only when  $\ell_T \rightarrow \infty$  for  $T \rightarrow \infty$ , do we get a consistent estimator, i.e.

which rate the lag truncation parameter should go to infinity. The literature made several suggestions.<sup>11</sup> In the following we concentrate on the Bartlett and the quadratic spectral kernel because these function always deliver a positive long-run variance in small samples. Andrews (1991) proposes the following formula to determine the optimal bandwidth:

$$\begin{aligned} \text{Bartlett :} & \quad \ell_T = 1.1447 [\alpha_{\text{Bartlett}} T]^{1/3} \\ \text{Quadratic Spectral :} & \quad \ell_T = 1.3221 [\alpha_{\text{QuadraticSpectral}} T]^{1/5}, \end{aligned}$$

whereby  $\alpha_{\text{Bartlett}}$  and  $\alpha_{\text{QuadraticSpectral}}$  are data dependent constants which have to be determined in a first step from the data (see (Andrews, 1991, 832–839), (Andrews and Monahan, 1992, 958) and (Haan and Levin, 1997)). If the underlying process is approximated by an AR(1) model, we get:

$$\begin{aligned} \alpha_{\text{Bartlett}} &= \frac{4\hat{\rho}^2}{(1 - \hat{\rho}^2)(1 + \hat{\rho}^2)} \\ \alpha_{\text{QuadraticSpectral}} &= \frac{4\hat{\rho}^2}{(1 - \hat{\rho})^4}, \end{aligned}$$

where  $\hat{\rho}$  is the first order empirical autocorrelation coefficient.

In order to avoid the cumbersome determination of the  $\alpha$ 's Newey and West (1994) suggest the following rules of thumb:

$$\begin{aligned} \text{Bartlett :} & \quad \ell_T = \beta_{\text{Bartlett}} \left[ \frac{T}{100} \right]^{2/9} \\ \text{Quadratic Spectral :} & \quad \ell_T = \beta_{\text{QuadraticSpectral}} \left[ \frac{T}{100} \right]^{2/25}. \end{aligned}$$

It has been shown that values of 4 for  $\beta_{\text{Bartlett}}$  as well as for  $\beta_{\text{QuadraticSpectral}}$  lead to acceptable results. A comparison of these formulas with the ones provided by Andrews shows that the latter imply larger values for  $\ell_T$  when the sample sizes gets larger. Both approaches lead to consistent estimates, i.e.  $\hat{J}_T(\ell_T) - J_T \xrightarrow{p} 0$  for  $T \rightarrow \infty$ .

In practice, a combination of both parametric and nonparametric methods proved to deliver the best results. This combined method consists of five steps:

---

$\hat{J}_T \rightarrow J_T$ , respectively  $J$ . Second, the restriction to  $\hat{\gamma}(h)$ ,  $|h| \leq q$ , does not necessarily lead to positive value for the estimated long-run variance  $\hat{J}_T$ , even when the Bartlett kernel is used. See Ogaki (1992) for details.

<sup>11</sup>See Haan and Levin (1997) for an overview.

- (i) The first step is called *prewhitening* and consists in the estimation of a simple ARMA model for the process  $\{X_t\}$  to remove the most obvious serial correlations. The idea, which goes back to Press and Tukey (1956) (see also Priestley (1981)), is to get a process for the residuals  $\hat{Z}_t$  which is close to a white noise process. Usually, an AR(1) model is sufficient.<sup>12</sup>
- (ii) Choose a kernel function and, if the method of Andrews has been chosen, the corresponding data dependent constants, i.e.  $\alpha_{\text{Bartlett}}$  or  $\alpha_{\text{QuadraticSpectral}}$  for the Bartlett, respectively the quadratic spectral kernel function.
- (iii) Compute the lag truncation parameter for the residuals using the the above formulas.
- (iv) Estimate the long-run variance for the residuals  $\hat{Z}_t$ .
- (v) Compute the long-run variance for the original time series  $\{X_t\}$ .

If in the first step an AR(1) model,  $X_t = \phi X_{t-1} + Z_t$ , was used, the last step is given by:

$$\hat{J}_T^X(\ell_T) = \frac{\hat{J}_T^Z(\ell_T)}{(1 - \hat{\phi})^2},$$

where  $\hat{J}_T^Z(\ell_T)$  and  $\hat{J}_T^X(\ell_T)$  denote the estimated long-run variances of  $\{X_t\}$  and  $\{\hat{Z}_t\}$ . In the general case, of an arbitrary ARMA model,  $\Phi(L)X_t = \Theta(L)Z_t$ , we get:

$$\hat{J}_T^X(\ell_T) = \left( \frac{\Theta(1)}{\Phi(1)} \right)^2 \hat{J}_T^Z(\ell_T).$$

#### 4.4.1 An example

Suppose we want to test whether the yearly growth rate of Switzerland's real GDP in the last 25 years was higher than one percent. For this purpose we compute the percentage change against the corresponding quarter of the last year over the period 1982:1 to 2006:1 (97 observations in total), i.e. we compute  $X_t = (1 - L^4) \log(\text{BIP}_t)$ . The arithmetic average of these growth rates is 1.4960 with a variance of 3.0608. We test the null hypothesis that the

---

<sup>12</sup>If in this step an AR(1) model is used and if a first order correlation  $\hat{\phi}$  larger in absolute terms than 0.97 is obtained, Andrews and Monahan (1992, 457) suggest to replace  $\hat{\phi}$  by -0.97, respectively 0.97. Instead of using a arbitrary fixed value, it turns out that a data driven value is superior. Sul et al. (2005) suggest to replace 0.97 by  $1 - 1/\sqrt{T}$  and -0.97 by  $-1 + 1/\sqrt{T}$ .

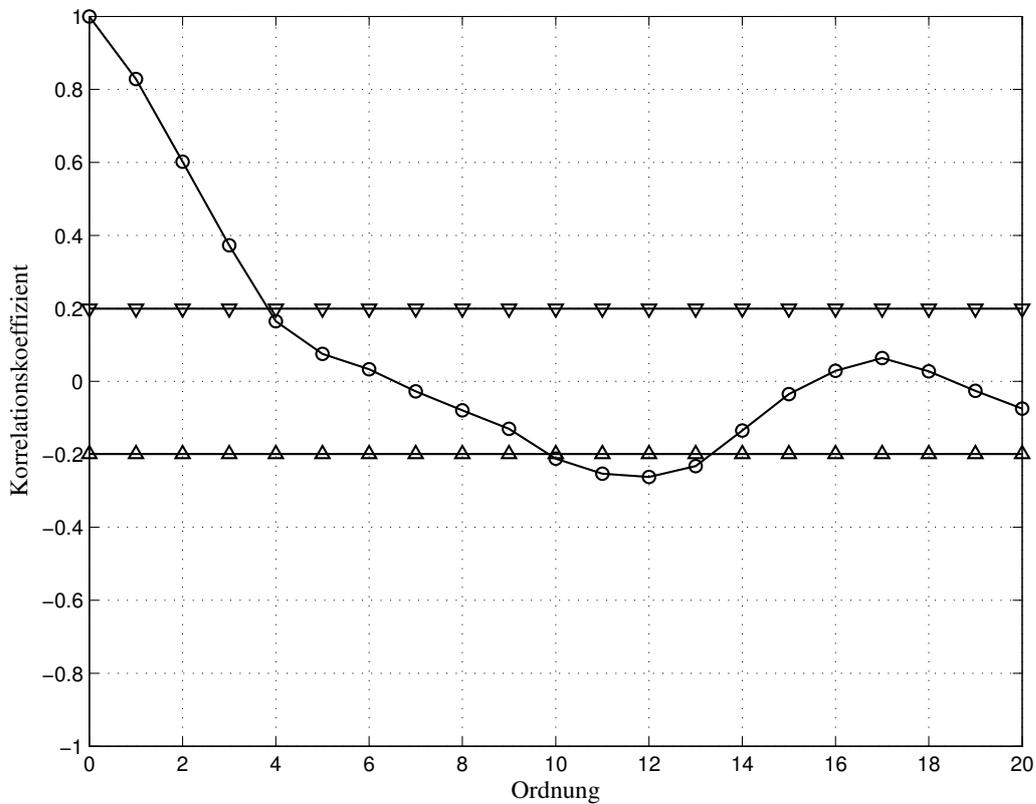


Figure 4.7: Estimated autocorrelation function for Switzerland's real GDP growth (percentage change against corresponding last year's quarter)

growth rate is smaller than one against the alternative that it is greater than one. The corresponding value of the t-statistic is  $(1.4960 - 1) / \sqrt{3.0608/97} = 2.7922$ . Taking a 5 percent significance level, the critical value for this one-sided test is 1.661. Thus the null hypothesis is clearly rejected.

The above computation is, however, not valid because the serial correlation of the time series was not taken into account. Indeed the estimated autocorrelation function shown in Figure 4.7 clearly shows that the growth rate is indeed subject to high and statistically significant autocorrelations.

Taking the Bartlett function as the kernel function, the rule of thumb formula for the lag truncation parameter suggest  $\ell_T = 4$ . The weights in the

computation of the long-run variance are therefore

$$k(h/\ell_T) = \begin{cases} 1, & h = 0; \\ 3/4, & h = \pm 1; \\ 2/4, & h = \pm 2; \\ 1/4, & h = \pm 3; \\ 0, & |h| \geq 4. \end{cases}$$

The corresponding estimate for the long-run variance is therefore given by:

$$\hat{J}_T = 3.0608 \left( 1 + 2\frac{3}{4}0.8287 + 2\frac{2}{4}0.6019 + 2\frac{1}{4}0.3727 \right) = 9.2783.$$

Using the long-run variance instead of the simple variance leads to a quite different value of the t-statistic:  $(1.4960 - 1)/\sqrt{9.2783/97} = 1.6037$ . The null hypothesis is thus not rejected at the five percent significance level when the serial correlation of the process is taken into account.

## 4.5 Exercises

**Exercise 4.5.1.** *You regress 100 realizations of a stationary stochastic process  $\{X_t\}$  against a constant  $c$ . The least-squares estimate of  $c$  equals  $\hat{c} = 0.04$  with an estimated standard deviation of  $\hat{\sigma}_c = 0.15$ . In addition, you have estimated the autocorrelation function up to order  $h = 5$  and obtained the following values:*

$$\hat{\rho}(1) = -0.43, \hat{\rho}(2) = 0.13, \hat{\rho}(3) = -0.12, \hat{\rho}(4) = 0.18, \hat{\rho}(5) = -0.23.$$

- (i) *How do you interpret the estimated parameter value of 0.4?*
- (ii) *Examine the autocorrelation function. Do you think that  $\{X_t\}$  is white noise?*
- (iii) *Why is the estimated standard deviation  $\hat{\sigma}_c = 0.15$  incorrect?*
- (iv) *Estimate the long-run variance using the Bartlett kernel.*
- (v) *Test the null hypothesis that  $\{X_t\}$  is a mean-zero process.*



# Chapter 5

## Estimation of ARMA Models

The specification and estimation of an ARMA( $p,q$ ) model for a given realization involves several intermingled steps. First one must determine the orders  $p$  and  $q$ . Given the orders one can then estimate the parameters  $\phi_j$ ,  $\theta_j$  and  $\sigma^2$ . Finally, the model has to pass several robustness checks in order to be accepted as a valid model. These checks may involve tests of parameter constancy, forecasting performance or tests for the inclusion of additional exogenous variables. This is usually an iterative process in which several models are examined. It is rarely the case that one model imposes itself. All too often, one is confronted in the modeling process with several trade-offs, like simple versus complex models or data fit versus forecasting performance. Finding the right balance among the different dimensions therefore requires some judgement based on experience.

We start the discussion by assuming that the orders of the ARMA process is known and the problem just consists in the estimation of the corresponding parameters from a realization of length  $T$ . For simplicity, we assume that the data are mean adjusted. We will introduce three estimation methods. The first one is a method of moments procedure where the theoretical moments are equated to the empirical ones. This procedure is known under the name of Yule-Walker estimator. The second procedure interprets the stochastic difference as a regression model and estimates the parameters by ordinary least-squares (OLS). These two methods work well if the underlying model is just an AR model and thus involves no MA terms. If the model comprises MA terms, a maximum likelihood (ML) approach must be pursued.

## 5.1 The Yule-Walker estimator

We assume that the stochastic process is governed by a causal purely autoregressive model of order  $p$  and mean zero:

$$\Phi(L)X_t = Z_t \quad \text{with } Z_t \sim \text{WN}(0, \sigma^2).$$

Causality with respect to  $\{Z_t\}$  implies that there exists a sequence  $\{\psi_j\}$  with  $\sum_{j=0}^{\infty} |\psi_j| < \infty$  such that  $X_t = \Psi(L)Z_t$ . Multiplying the above difference equation by  $X_{t-j}$ ,  $j = 0, 1, \dots, p$  and taking expectations leads to the following equation system for the parameters  $\Phi = (\phi_1, \dots, \phi_p)'$  and  $\sigma^2$ :

$$\begin{aligned} \gamma(0) - \phi_1\gamma(1) - \dots - \phi_p\gamma(p) &= \sigma^2 \\ \gamma(1) - \phi_1\gamma(0) - \dots - \phi_p\gamma(p-1) &= 0 \\ &\dots \\ \gamma(p) - \phi_1\gamma(p-1) - \dots - \phi_p\gamma(0) &= 0 \end{aligned}$$

This equation system is known as the *Yule-Walker equations*. It can be written compactly in matrix algebra as:

$$\begin{aligned} \gamma(0) - \Phi' \gamma_p(1) &= \sigma^2, \\ \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \dots & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} &= \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}, \end{aligned}$$

respectively

$$\begin{aligned} \gamma(0) - \Phi' \gamma_p(1) &= \sigma^2, \\ \Gamma_p \Phi &= \gamma_p(1). \end{aligned}$$

The *Yule-Walker estimator* is obtained by replacing the theoretical moments by the empirical ones and solving the resulting equation system for the unknown parameters:

$$\begin{aligned} \hat{\Phi} &= \hat{\Gamma}_p^{-1} \hat{\gamma}_p(1) = \hat{R}_p^{-1} \hat{\rho}_p(1) \\ \hat{\sigma}^2 &= \hat{\gamma}(0) - \hat{\Phi}' \hat{\gamma}_p(1) = \hat{\gamma}(0) \left( 1 - \hat{\rho}_p(1)' \hat{R}_p^{-1} \hat{\rho}_p(1) \right) \end{aligned}$$

Note the recursiveness of the equation system: the estimate  $\hat{\Phi}$  is obtained without knowledge of  $\hat{\sigma}^2$  as the estimator  $\hat{R}_p^{-1} \hat{\rho}_p(1)$  involves only autocor-

relations. The estimates  $\widehat{\Gamma}_p$ ,  $\widehat{R}_p$ ,  $\widehat{\gamma}_p(1)$ ,  $\widehat{\rho}_p(1)$ , and  $\widehat{\gamma}(0)$  are obtained in the usual way as explained in Chapter 4.<sup>1</sup>

The construction of the Yule-Walker estimator implies that the first  $p$  values of the autocovariance, respectively the autocorrelation function, implied by the estimated model exactly correspond to their estimated counterparts. It can be shown that this moment estimator always delivers coefficients  $\widehat{\Phi}$  which imply that  $\{X_t\}$  is causal with respect to  $\{Z_t\}$ . In addition, the following Theorem establishes that the estimated coefficients are asymptotically normal.

**Theorem 5.1** (Asymptotic property of Yule-Walker estimator). *Let  $\{X_t\}$  be an AR( $p$ ) process which is causal with respect to  $\{Z_t\}$  whereby  $\{Z_t\} \sim \text{IID}(0, \sigma^2)$ . Then the Yule-Walker estimator is consistent and  $\widehat{\Phi}$  is asymptotically normal with distribution given by:*

$$\sqrt{T} \left( \widehat{\Phi} - \Phi \right) \xrightarrow{d} \text{N} \left( 0, \sigma^2 \Gamma_p^{-1} \right).$$

In addition we have that

$$\widehat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

*Proof.* See Brockwell and Davis (1991, 233-234). □

Noteworthy, the asymptotic covariance matrix of the Yule-Walker estimate is independent of  $\sigma^2$ .

### Example: AR(1) process

In the case of an AR(1) process, the Yule-Walker equation is  $\widehat{\Gamma}_1 = \widehat{\gamma}(0)$ . The Yule-Walker estimator thus becomes:

$$\widehat{\Phi} = \widehat{\phi} = \frac{\widehat{\gamma}(1)}{\widehat{\gamma}(0)} = \widehat{\rho}(1).$$

The asymptotic distribution then is

$$\sqrt{T} \left( \widehat{\phi} - \phi \right) \xrightarrow{d} \text{N} \left( 0, \frac{\sigma^2}{\gamma(0)} \right) = \text{N} \left( 0, 1 - \phi^2 \right).$$

This shows that the assumption of causality, i.e.  $|\phi| < 1$ , is crucial. Otherwise no positive value for the variance would exist.

---

<sup>1</sup>Note that the application of the estimator introduced in Section 4.2 guarantees that  $\widehat{\Gamma}_p$  is always invertible.

In practice the order of the model is usually unknown. However, one can expect when estimating an AR( $m$ ) model whereby the true order  $p$  is strictly smaller than  $m$  that the estimated coefficients  $\hat{\phi}_{p+1}, \dots, \hat{\phi}_m$  should be close to zero. This is indeed the case as shown in Brockwell and Davis (1991, 241). In particular, under the assumptions of Theorem 5.1 it holds that

$$\sqrt{T} \hat{\phi}_m \xrightarrow{d} N(0, 1) \quad \text{for } m > p. \quad (5.1)$$

This result justifies the following strategy to identify the order of an AR-model. Estimate in a first step a highly parameterized model (overfitted model), i.e. a model with a large value of  $m$ , and test via a t-test whether  $\hat{\phi}_m$  is zero. If the hypothesis cannot be rejected, reduce the order of the model from  $m$  to  $m - 1$  and repeat the same procedure now with respect to  $\hat{\phi}_{m-1}$ . This is done until the hypothesis can no longer be rejected.

If the order of the initial model is too low (underfitted model) so that the true order is higher than  $m$ , one incurs an “omitted variable bias”. The corresponding estimates are no longer consistent. In Section 5.4, we take closer look at the problem of determining the order of a model.

### Example: MA( $q$ ) process

The Yule-Walker estimator can, in principle, also be applied to MA( $q$ ) or ARMA( $p, q$ ) processes with  $q > 0$ . However, the analysis of the simple MA(1) process in Section 1.5.1 showed that the relation between the autocorrelations and the model parameters is nonlinear and may have two, one, or no solution. Consider again the MA(1) process as an example. It is given by the stochastic difference equation  $X_t = Z_t + \theta Z_{t-1}$  with  $Z_t \sim \text{IID}(0, \sigma^2)$ . The Yule-Walker equations are then as follows:

$$\begin{aligned} \hat{\gamma}(0) &= \hat{\sigma}^2(1 + \hat{\theta}^2) \\ \hat{\gamma}(1) &= \hat{\sigma}^2\hat{\theta} \end{aligned}$$

As shown in Section 1.5.1, this system of equations has for  $|\hat{\rho}(1)| = |\hat{\gamma}(1)/\hat{\gamma}(0)| < 1/2$  two solutions; for  $|\hat{\rho}(1)| = |\hat{\gamma}(1)/\hat{\gamma}(0)| = 1/2$  one solution; and for  $|\hat{\rho}(1)| = |\hat{\gamma}(1)/\hat{\gamma}(0)| > 1/2$  no real solution. In the case of several solutions, we usually take the invertible one which leads to  $|\theta| < 1$ . Invertibility is, however, a restriction which is hard to implement in the case of higher order MA processes. Moreover, it can be shown that Yule-Walker estimator is no longer consistent in general (see Brockwell and Davis (1991, 246) for details). For these reasons, it is not advisable to use the Yule-Walker estimator in the case of MA processes, especially when there exist consistent and efficient alternatives.

## 5.2 Ordinary least-squares (OLS) estimation of an AR(p) model

An alternative approach is to view the AR model as a regression model for  $X_t$  with regressors  $X_{t-1}, \dots, X_{t-p}$  and error term  $Z_t$ :

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

Given observation for  $X_1, \dots, X_T$ , the regression model can be compactly written in matrix algebra as follows:

$$\begin{pmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_T \end{pmatrix} = \begin{pmatrix} X_p & X_{p-1} & \dots & X_1 \\ X_{p+1} & X_p & \dots & X_2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{T-1} & X_{T-2} & \dots & X_{T-p} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} + \begin{pmatrix} Z_{p+1} \\ Z_{p+2} \\ \vdots \\ Z_T \end{pmatrix},$$

$$Y = \mathbf{X}\Phi + Z. \quad (5.2)$$

Note that the first  $p$  observations are lost and that the effective sample size is thus reduced to  $T - p$ . The least-squares estimator (OLS estimator) is obtained as the minimizer of the sum of squares  $S(\Phi)$ :

$$\begin{aligned} S(\Phi) &= Z'Z = (Y - \mathbf{X}\Phi)'(Y - \mathbf{X}\Phi) \\ &= \sum_{t=p+1}^T (X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p})^2 \\ &= \sum_{t=p+1}^T (X_t - \mathbb{P}_{t-1} X_t)^2 \longrightarrow \min_{\Phi}. \end{aligned} \quad (5.3)$$

Note that the optimization problem involves no constraints, in particular causality is not imposed as a restriction. The solution of this minimization problem is given by usual formula:

$$\hat{\Phi} = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'Y).$$

Although equation (5.2) resembles very much an ordinary regression model, there are some important differences. First, the usual orthogonality assumption between regressors and error is violated. The regressors  $X_{t-j}$ ,  $j = 1, \dots, p$ , are correlated with the error terms  $Z_{t-j}$ ,  $j = 1, 2, \dots$ . In addition, there is a dependency on the starting values  $X_p, \dots, X_1$ . The assumption of causality, however, insures that these features do not play a role asymptotically. It can be shown that  $(\mathbf{X}'\mathbf{X})/T$  converges in probability to  $\hat{\Gamma}_p$  and

$(\mathbf{X}'Y)/T$  to  $\hat{\gamma}_p$ . In addition, under quite general conditions,  $T^{-1/2}\mathbf{X}'Z$  is asymptotically normally distributed with mean 0 and variance  $\sigma^2\Gamma_p$ . Then by Slutsky's Lemma C.10,  $\sqrt{T}(\hat{\Phi} - \Phi) = \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1} \left(\frac{\mathbf{X}'Z}{\sqrt{T}}\right)$  converges in distribution to  $N(0, \sigma^2\Gamma_p^{-1})$ . Thus, the OLS estimator is asymptotically equivalent to the Yule-Walker estimator.

**Theorem 5.2** (Asymptotic property of the Least-Squares estimator). *Under the same conditions as in Theorem 5.1 the OLS estimator  $\hat{\Phi} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'Y)$ :*

$$\sqrt{T}(\hat{\Phi} - \Phi) \xrightarrow{d} N(0, \sigma^2\Gamma_p^{-1}),$$

$$\text{plim } s_T^2 = \sigma^2$$

where  $s_T^2 = \hat{Z}'\hat{Z}/T$  and  $\hat{Z}_t$  are the OLS residuals.

*Proof.* See Chapter 13 and in particular section 13.3 for a proof in the multivariate case. Additional details may be gathered from Brockwell and Davis (1991, chapter 8).  $\square$

**Remark 5.1.** *In practice  $\sigma^2\Gamma_p^{-1}$  is approximated by  $s_T^2(\mathbf{X}'\mathbf{X}/T)^{-1}$ . Thus, for large  $T$ ,  $\hat{\Phi}$  can be viewed as being normally distributed as  $N(\Phi, s_T^2(\mathbf{X}'\mathbf{X})^{-1})$ . This result allows the application of the usual  $t$ - and  $F$ -tests.*

Because the regressors  $X_{t-j}$ ,  $j = 1, \dots, p$  are correlated with the errors terms  $Z_{t-j}$ ,  $j = 1, 2, \dots$ , the Gauss-Markov theorem cannot be applied. This implies that the least-squares estimator is no longer unbiased in finite samples. It can be shown that the estimates of an AR(1) model are downward biased when the true value of  $\phi$  is between zero and one. MacKinnon and Smith (1998, figure 1) plots the bias as a function of the sample size and the true parameter (see also Figure 7.1). As the bias function is almost linear in the range  $-0.85 < \phi < 0.85$ , an approximately unbiased estimator for the AR(1) model has been proposed by Marriott and Pope (1954), Kendall (1954), and Orcutt and Winokur (1969) MacKinnon and Smith (1998, for further details see):

$$\hat{\phi}_{\text{corrected}} = \frac{1}{T-3}(T\hat{\phi}_{\text{OLS}} + 1).$$

**Remark 5.2.** *The OLS estimator does in general not deliver coefficients  $\hat{\Phi}$  for which  $\{X_t\}$  is causal with respect  $\{Z_t\}$ . In particular, in the case of an AR(1) model, it can happen that, in contrast to the Yule-Walker estimator,  $|\hat{\phi}|$  is larger than one despite the fact that the true parameter is absolutely*

smaller than one. Nevertheless, the least-squares estimator is to be preferred in practice because it delivers small-sample biases of the coefficients which are smaller than those of Yule-Walker estimator, especially for roots of  $\Phi(z)$  close to the unit circle (Tjøstheim and Paulsen, 1983; Shaman and Stine, 1988; Reinsel, 1993).

### Appendix: Proof of the asymptotic normality of the OLS estimator

The proofs of Theorems 5.1 and 5.2 are rather involved and will therefore not be pursued here. A proof for the more general multivariate case will be given in Chapter 13. It is, however, instructive to look at a simple case, namely the AR(1) model with  $|\phi| < 1$ ,  $Z_t \sim \text{IIN}(0, \sigma^2)$  and  $X_0 = 0$ . Denoting by  $\hat{\phi}_T$  the OLS estimator of  $\phi$ , we have:

$$\sqrt{T} \left( \hat{\phi}_T - \phi \right) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} Z_t}{\frac{1}{T} \sum_{t=1}^T X_{t-1}^2}. \quad (5.4)$$

Moreover,  $X_t$  can be written as follows:

$$X_t = Z_t + \phi Z_{t-1} + \dots + \phi^{t-1} Z_1.$$

By assumption each  $Z_j$ ,  $j = 1, \dots, t$ , is normally distributed so that  $X_t$  as a sum normally distributed random variables is also normally distributed. Because the  $Z_j$ 's are independent we have:  $X_t \sim N\left(0, \sigma^2 \frac{1-\phi^{2t}}{1-\phi^2}\right)$ .

The expected value of  $\frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} Z_t$  is zero because  $Z_t \sim \text{IIN}(0, \sigma^2)$ . The variance of this expression is given by

$$\begin{aligned} \mathbb{V} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1} Z_t \right) &= \frac{1}{T} \sum_{t=1}^T \mathbb{E} X_{t-1}^2 Z_t^2 + \frac{2}{T} \sum_{t=1}^T \sum_{j=1}^{t-1} \mathbb{E} Z_t \mathbb{E} X_{t-1} X_j Z_j \\ &= \frac{\sigma^2}{T} \sum_{t=1}^T \mathbb{E} X_{t-1}^2. \end{aligned}$$

Moreover,  $\sum_{t=1}^T X_t^2 = \sum_{t=1}^T X_{t-1}^2 - (X_0^2 - X_T^2) = \phi^2 \sum_{t=1}^T X_{t-1}^2 + \sum_{t=1}^T Z_t^2 + 2\phi \sum_{t=1}^T X_{t-1} Z_t$  so that

$$\sum_{t=1}^T X_{t-1}^2 = \frac{1}{1-\phi^2} (X_0^2 - X_T^2) + \frac{1}{1-\phi^2} \sum_{t=1}^T Z_t^2 + \frac{2\phi}{1-\phi^2} \sum_{t=1}^T X_{t-1} Z_t.$$

The expected value multiplied by  $\sigma^2/T$  thus is equal to

$$\begin{aligned} \frac{\sigma^2}{T} \sum_{t=1}^T \mathbb{E}X_{t-1}^2 &= \frac{\sigma^2}{1-\phi^2} \frac{\mathbb{E}X_0^2 - \mathbb{E}X_T^2}{T} + \frac{\sigma^2}{1-\phi^2} \frac{\sum_{t=1}^T \mathbb{E}Z_t^2}{T} + \frac{2\phi}{1-\phi^2} \frac{\sum_{t=1}^T X_{t-1}Z_t}{T} \\ &= -\frac{\sigma^4(1-\phi^{2T})}{T(1-\phi^2)^2} + \frac{\sigma^4}{1-\phi^2}. \end{aligned}$$

For  $T$  going to infinity, we finally get:

$$\lim_{T \rightarrow \infty} \mathbb{V} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T X_{t-1}Z_t \right) = \frac{\sigma^4}{1-\phi^2}.$$

The numerator in equation (5.4) therefore converges to a normal random variable with mean zero and variance  $\frac{\sigma^4}{1-\phi^2}$ .

The denominator in equation (5.4) can be rewritten as

$$\frac{1}{T} \sum_{t=1}^T X_{t-1}^2 = \frac{X_0^2 - X_T^2}{(1-\phi^2)T} + \frac{1}{(1-\phi^2)T} \sum_{t=1}^T Z_t^2 + \frac{2\phi}{(1-\phi^2)T} \sum_{t=1}^T X_{t-1}Z_t.$$

The expected value and the variance of  $X_T^2/T$  converge to zero. Chebyshev's inequality (see Theorem C.3 in Appendix C) then implies that the first term converges also in probability to zero.  $X_0$  is equal to zero by assumption. The second term has a constant mean equal to  $\sigma^2/(1-\phi^2)$  and a variance which converges to zero. Theorem C.8 in Appendix C then implies that the second term converges in probability to  $\sigma^2/(1-\phi^2)$ . The third term has a mean of zero and a variance which converges to zero. Thus the third term converges to zero in probability. This implies:

$$\frac{1}{T} \sum_{t=1}^T X_{t-1}^2 \xrightarrow{p} \frac{\sigma^2}{1-\phi^2}.$$

Putting the results for the numerator and the denominator together and applying Theorem C.10 and the continuous mapping theorem for the convergence in distribution one finally obtains:

$$\sqrt{T} \left( \hat{\phi}_T - \phi \right) \xrightarrow{d} N(0, 1-\phi^2). \quad (5.5)$$

Thereby the value for the variance is derived from

$$\frac{\sigma^4}{1-\phi^2} \times \frac{1}{\left( \frac{\sigma^2}{1-\phi^2} \right)^2} = 1-\phi^2.$$

### 5.3 Estimation of an ARMA(p,q) model

While the estimation of AR models by OLS is rather straightforward and leads to consistent and asymptotically efficient estimates, the estimation of ARMA models is more complex. The reason is that, in contrast to past  $X_t$ 's  $Z_t, Z_{t-1}, \dots, Z_{t-q}$  are not directly observable from the data. They must be inferred from the observations of  $X_t$ . The standard method for the estimation of ARMA models is the method of maximum likelihood which will be explained in this section.

We assume that the  $X_t$ 's are generated by a causal and invertible ARMA(p,q) process:

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

with  $Z_t \sim \text{IID}(0, \sigma^2)$ . We also assume that  $\Phi(z)$  and  $\Theta(z)$  have no roots in common. We then stack the parameters of the model into a vector  $\beta$  and a scalar  $\sigma^2$ :

$$\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)' \quad \text{and} \quad \sigma^2.$$

Given the assumption above the admissible parameter space for  $\beta$ ,  $\mathcal{C}$ , is described by the following set:

$$\mathcal{C} = \{ \beta \in \mathbb{R}^{p+q} : \Phi(z)\Theta(z) \neq 0 \text{ for } |z| \leq 1, \phi_p \theta_q \neq 0, \\ \Phi(z) \text{ and } \Theta(z) \text{ have no roots in common.} \}$$

The estimation by the *method of maximum likelihood* (ML method) is based on some assumption about the joint distribution of  $\mathbf{X}_T = (X_1, \dots, X_T)'$  given the parameters  $\beta$  and  $\sigma^2$ . This joint distribution function is called the likelihood function. The method of maximum likelihood then determines the parameters such that the probability of observing a given sample  $\mathbf{x}_T = (x_1, \dots, x_T)$  is maximized. This is achieved by maximizing the likelihood function with respect to the parameters.

By far the most important case is given by assuming that  $\{X_t\}$  is a Gaussian process with mean zero and autocovariance function  $\gamma$ . This implies that  $\mathbf{X}_T = (X_1, \dots, X_T)'$  is distributed as a multivariate normal with mean zero and variance  $\Gamma_T$ .<sup>2</sup> The *Gaussian likelihood function* given the observations  $\mathbf{x}_T$ ,  $L_T(\beta, \sigma^2 | \mathbf{x}_T)$ , is then given by

$$L_T(\beta, \sigma^2 | \mathbf{x}_T) = (2\pi)^{-T/2} (\det \Gamma_T)^{-1/2} \exp \left( -\frac{1}{2} \mathbf{x}_T' \Gamma_T^{-1} \mathbf{x}_T \right) \\ = (2\pi\sigma^2)^{-T/2} (\det G_T)^{-1/2} \exp \left( -\frac{1}{2\sigma^2} \mathbf{x}_T' G_T^{-1} \mathbf{x}_T \right)$$

<sup>2</sup>If the process does not have a mean of zero, we can demean the data in a preliminary step.

where  $G_T = \sigma^{-2}\Gamma_T$ . Note that, in contrast to  $\Gamma_T$ ,  $G_T$  does only depend on  $\beta$  and not on  $\sigma^2$ .<sup>3</sup> If one wants to point out the dependence of  $G_T$  from  $\beta$  we write  $G_T(\beta)$ . The method of maximum likelihood then consists in the maximization of the likelihood function with respect to  $\beta$  and  $\sigma^2$  taking the data  $\mathbf{x}_T$  as given.

The first order condition of this maximization problem with respect to  $\sigma^2$  is obtained by taking the logarithm of the likelihood function  $L_T(\beta, \sigma^2|\mathbf{x}_T)$  and differentiating with respect  $\sigma^2$  and setting the resulting equation equal to zero:

$$\frac{\partial \ln L_T(\beta, \sigma^2|\mathbf{x}_T)}{\partial \sigma^2} = -\frac{T}{2} \frac{1}{\sigma^2} + \frac{\mathbf{X}'_T G_T^{-1} \mathbf{X}_T}{2\sigma^4} = 0.$$

Solving this equation with respect to  $\sigma^2$  we get as the solution:  $\sigma^2 = T^{-1} \mathbf{x}'_T G_T^{-1} \mathbf{x}_T$ . Inserting this value into the original likelihood function and taking the logarithm, one gets the concentrated log-likelihood function:

$$\ln L_T(\beta|\mathbf{x}_T) = -\ln(2\pi) - \frac{T}{2} \ln (T^{-1} \mathbf{x}'_T G_T(\beta)^{-1} \mathbf{x}_T) - \frac{1}{2} \ln \det G_T(\beta) - \frac{T}{2}.$$

This function is then maximized with respect to  $\beta \in \mathcal{C}$ . This is, however, equivalent to minimizing the function

$$\ell_T(\beta|\mathbf{x}_T) = \ln (T^{-1} \mathbf{x}'_T G_T(\beta)^{-1} \mathbf{x}_T) + T^{-1} \ln \det G_T(\beta) \longrightarrow \min_{\beta \in \mathcal{C}}.$$

The value of  $\beta$  which minimizes the above function is called *maximum-likelihood estimator* of  $\beta$ . It will be denoted by  $\hat{\beta}_{\text{ML}}$ . The maximum-likelihood estimator for  $\sigma^2$ ,  $\hat{\sigma}_{\text{ML}}^2$ , is then given by

$$\hat{\sigma}_{\text{ML}}^2 = T^{-1} \mathbf{x}'_T G_T(\hat{\beta}_{\text{ML}})^{-1} \mathbf{x}_T.$$

The actual computation of  $\det G_T(\beta)$  and  $G_T(\beta)^{-1}$  is numerically involved, especially when  $T$  is large, and should therefore be avoided. It is therefore convenient to rewrite the likelihood function in a different, but equivalent form:

$$L_T(\beta, \sigma^2|\mathbf{x}_T) = (2\pi\sigma^2)^{-T/2} (r_0 r_1 \dots r_{T-1})^{-1/2} \exp \left( -\frac{1}{2\sigma^2} \sum_{t=1}^T \frac{(X_t - \mathbb{P}_{t-1} X_t)^2}{r_{t-1}} \right).$$

<sup>3</sup>In Section 2.4 we showed how the autocovariance function  $\gamma$  and as a consequence  $\Gamma_T$ , respectively  $G_T$  can be inferred from a given ARMA model, i.e from a given  $\beta$ .

Thereby  $\mathbb{P}_{t-1}X_t$  denotes least-squares predictor of  $X_t$  given  $X_{t-1}, \dots, X_1$  and  $r_t = v_t/\sigma^2$  where  $v_t$  is the mean squared forecast error as defined in Section 3.1. Several numerical algorithms have been developed to compute these forecast in a numerically efficient and stable way.<sup>4</sup>

$\mathbb{P}_{t-1}X_t$  and  $r_t$  do not depend on  $\sigma^2$  so that the partial differentiation of the log-likelihood function  $\ln L(\beta, \sigma^2|\mathbf{x}_T)$  with respect to the parameters leads to the maximum likelihood estimator. This estimator fulfills the following equations:

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{T}S(\hat{\beta}_{\text{ML}}),$$

where

$$S(\hat{\beta}_{\text{ML}}) = \sum_{t=1}^T \frac{(X_t - \mathbb{P}_{t-1}X_t)^2}{r_{t-1}}$$

and where  $\hat{\beta}_{\text{ML}}$  denote the value of  $\beta$  which minimizes the function

$$\ell_T(\beta|\mathbf{x}_T) = \ln \left( \frac{1}{T}S(\beta) \right) + \frac{1}{T} \sum_{t=1}^T \ln r_{t-1}$$

subject to  $\beta \in \mathcal{C}$ . This optimization problem must be solved numerically. In practice, one chooses as a starting value  $\beta_0$  for the iteration an initial estimate such that  $\beta_0 \in \mathcal{C}$ . In the following iterations this restriction is no longer imposed to enhance speed and reduce the complexity of the optimization problem. This implies that one must check whether the so obtained final estimates are indeed in  $\mathcal{C}$ .

If instead of  $\ell_T(\beta|\mathbf{x}_T)$ , the function

$$S(\beta) = \sum_{t=1}^T \frac{(X_t - \mathbb{P}_{t-1}X_t)^2}{r_{t-1}}$$

is minimized subject to constraint  $\beta \in \mathcal{C}$ , we obtain the *least-squares estimator* of  $\beta$  denoted by  $\hat{\beta}_{\text{LS}}$ . The least-squares estimator of  $\sigma^2$ ,  $\hat{\sigma}_{\text{LS}}^2$ , is then

$$\hat{\sigma}_{\text{LS}}^2 = \frac{S(\hat{\beta}_{\text{LS}})}{T - p - q}.$$

---

<sup>4</sup>One such algorithm is the innovation algorithm. See Brockwell and Davis (1991, section 5.) for details.

The term  $\frac{1}{T} \sum_{t=1}^T \ln r_{t-1}$  disappears asymptotically because, given the restriction  $\beta \in \mathcal{C}$ , the mean-squared forecast error  $v_T$  converges to  $\sigma^2$  and thus  $r_T$  goes to one as  $T$  goes to infinity. This implies that for  $T$  going to infinity the maximization of the likelihood function becomes equivalent to the minimization of the least-squares criterion. Thus the maximum-likelihood estimator and the least-squares estimator share the same asymptotic normal distribution.

Note also that in the case of autoregressive models  $r_t$  is constant and equal to one. In this case, the least-squares criterion  $S(\beta)$  reduces to the criterion (5.3) discussed in the previous Section 5.2.

**Theorem 5.3** (Asymptotic Distribution of ML Estimator). *If  $\{X_t\}$  is an ARMA process with true parameters  $\beta \in \mathcal{C}$  and  $Z_t \sim \text{IID}(0, \sigma^2)$  with  $\sigma^2 > 0$  then the maximum-likelihood estimator and the least-squares estimator have asymptotically the same normal distribution. Thus we have*

$$\begin{aligned}\sqrt{T} \left( \hat{\beta}_{\text{ML}} - \beta \right) &\xrightarrow{d} \text{N} \left( 0, V(\beta) \right), \\ \sqrt{T} \left( \hat{\beta}_{\text{LS}} - \beta \right) &\xrightarrow{d} \text{N} \left( 0, V(\beta) \right).\end{aligned}$$

The asymptotic covariance matrix  $V(\beta)$  is thereby given by

$$\begin{aligned}V(\beta) &= \begin{pmatrix} \mathbb{E}U_t U_t' & \mathbb{E}U_t V_t' \\ \mathbb{E}V_t U_t' & \mathbb{E}V_t V_t' \end{pmatrix}^{-1} \\ U_t &= (u_t, u_{t-1}, \dots, u_{t-p+1})' \\ V_t &= (v_t, v_{t-1}, \dots, v_{t-q+1})'\end{aligned}$$

where  $\{u_t\}$  and  $\{v_t\}$  denote autoregressive processes defined as  $\Phi(L)u_t = w_t$  and  $\Theta(L)v_t = w_t$  with  $w_t \sim \text{WN}(0, 1)$ .

*Proof.* See Brockwell and Davis (1991, Section 8.8). □

It can be shown that both estimators are asymptotically efficient.<sup>5</sup> Note that the asymptotic covariance matrix  $V(\beta)$  is independent of  $\sigma^2$ .

The use of the Gaussian likelihood function makes sense even when the process is not Gaussian. First, the Gaussian likelihood can still be interpreted as a measure of fit of the ARMA model to the data. Second, the asymptotic distribution is still Gaussian even when the process is not Gaussian as long as  $Z_t \sim \text{IID}(0, \sigma^2)$ . The Gaussian likelihood is then called the quasi Gaussian likelihood. The use of the Gaussian likelihood under this circumstance is, however, in general no longer efficient.

<sup>5</sup>See Brockwell and Davis (1991) and Fan and Yao (2003) for details.

**Example: AR( $p$ ) process**

In this case  $\beta = (\phi_1, \dots, \phi_p)$  and  $V(\beta) = (\mathbb{E}U_t U_t')^{-1} = \sigma^2 \Gamma_p^{-1}$ . This is, however, the same asymptotic distribution as the Yule-Walker estimator. The Yule-Walker, the least-squares, and the maximum likelihood estimator are therefore asymptotically equivalent in the case of an AR( $p$ ) process. The main difference lies in the treatment of the first  $p$  observations.

In particular, we have:

$$\begin{aligned} \text{AR}(1) : \quad & \hat{\phi} \sim N(\phi, (1 - \phi^2)/T), \\ \text{AR}(2) : \quad & \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \frac{1}{T} \begin{pmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & 1 - \phi_2^2 \end{pmatrix} \right). \end{aligned}$$

**Example: MA( $q$ ) process**

Similarly, one can compute the asymptotic distribution for an MA( $q$ ) process. In particular, we have:

$$\begin{aligned} \text{MA}(1) : \quad & \hat{\theta} \sim N(\theta, (1 - \theta^2)/T), \\ \text{MA}(2) : \quad & \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim N \left( \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \frac{1}{T} \begin{pmatrix} 1 - \theta_2^2 & \theta_1(1 - \theta_2) \\ \theta_1(1 - \theta_2) & 1 - \theta_2^2 \end{pmatrix} \right). \end{aligned}$$

**Example: ARMA(1,1) process**

For an ARMA(1,1) process the asymptotic covariance matrix is given by

$$V(\phi, \theta) = \begin{pmatrix} (1 - \phi^2)^{-1} & (1 + \phi\theta)^{-1} \\ (1 + \phi\theta)^{-1} & (1 - \theta^2)^{-1} \end{pmatrix}^{-1}.$$

Therefore we have:

$$\begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix} \sim N \left( \begin{pmatrix} \phi \\ \theta \end{pmatrix}, \frac{1}{T} \frac{1 + \phi\theta}{(\phi + \theta)^2} \begin{pmatrix} (1 - \phi^2)(1 + \phi\theta) & -(1 - \theta^2)(1 - \phi^2) \\ -(1 - \theta^2)(1 - \phi^2) & (1 - \theta^2)(1 + \phi\theta) \end{pmatrix} \right).$$

## 5.4 Estimation of the orders $p$ and $q$

Up to now we have always assumed that the true orders of the ARMA model  $p$  and  $q$  are known. This is, however, seldom the case in practice. As economic theory does usually not provide an indication, it is all too often the case that the orders of the ARMA model must be identified from the data. In this process one can make two types of errors:  $p$  and  $q$  are too large in which

case we speak of overfitting;  $p$  and  $q$  are too low in which case we speak of underfitting.

In the case of overfitting, the maximum likelihood estimator is no longer consistent for the true parameter, but still consistent for the coefficients of the causal representation  $\psi_j$ ,  $j = 0, 1, 2, \dots$ , where  $\psi(z) = \frac{\theta(z)}{\phi(z)}$ . This can be illustrated by the following example. Suppose that  $\{X_t\}$  is a white noise process, i.e.  $X_t = Z_t \sim \text{WN}(0, \sigma^2)$ , but we fit an ARMA(1,1) model given by  $X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}$ . Then, the maximum likelihood estimator does not converge to  $\phi = \theta = 0$ , but only to the line-segment  $\phi = -\theta$  with  $|\phi| < 1$  and  $|\theta| < 1$ . For values of  $\phi$  and  $\theta$  on this line-segment we have  $\psi(z) = \theta(z)/\phi(z) = 1$ . The maximum likelihood estimator converges to the true values of  $\psi_j$ , i.e. to the values  $\psi_0 = 1$  and  $\psi_j = 0$  for  $j > 0$ . The situation is depicted in Figure 5.1. There it is shown that the estimator has a tendency to converge to the points  $(-1, 1)$  and  $(1, -1)$ , depending on the starting values. This indeterminacy of the estimator manifest itself as a numerical problem in the optimization of the likelihood function. Thus models with similar roots for the AR and MA polynomials which are close in absolute value to the unit circle are probably overparametrized. The problem can be overcome by reducing the orders of the AR and MA polynomial by one.

This problem does appear in a purely autoregressive models. As explained in Section 5.1, the estimator for the redundant coefficients converges to zero with asymptotic distribution  $N(0, 1/T)$  (see the result in equation (5.1)). This is one reason why purely autoregressive models are often preferred. In addition the estimator is easily implemented and every stationary stochastic process can be arbitrarily well approximated by an AR process. This approximation may, however, necessitate high order models when the true process encompasses a MA component.

In the case of underfitting the maximum likelihood estimator converges to those values which are closest to the true parameters given the restricted parameter space. The estimates are, however, inconsistent due to the “omitted variable bias”.

For these reasons the identification of the orders is an important step. One method which goes back to Box and Jenkins (1976) consists in the analysis of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) (see Section 3.5). Although this method requires some experience, especially when the process is not a purely AR or MA process, the analysis of the ACF und PACF remains an important first step in every practical investigation of a time series.

An alternative procedure relies on the automatic order selection. The

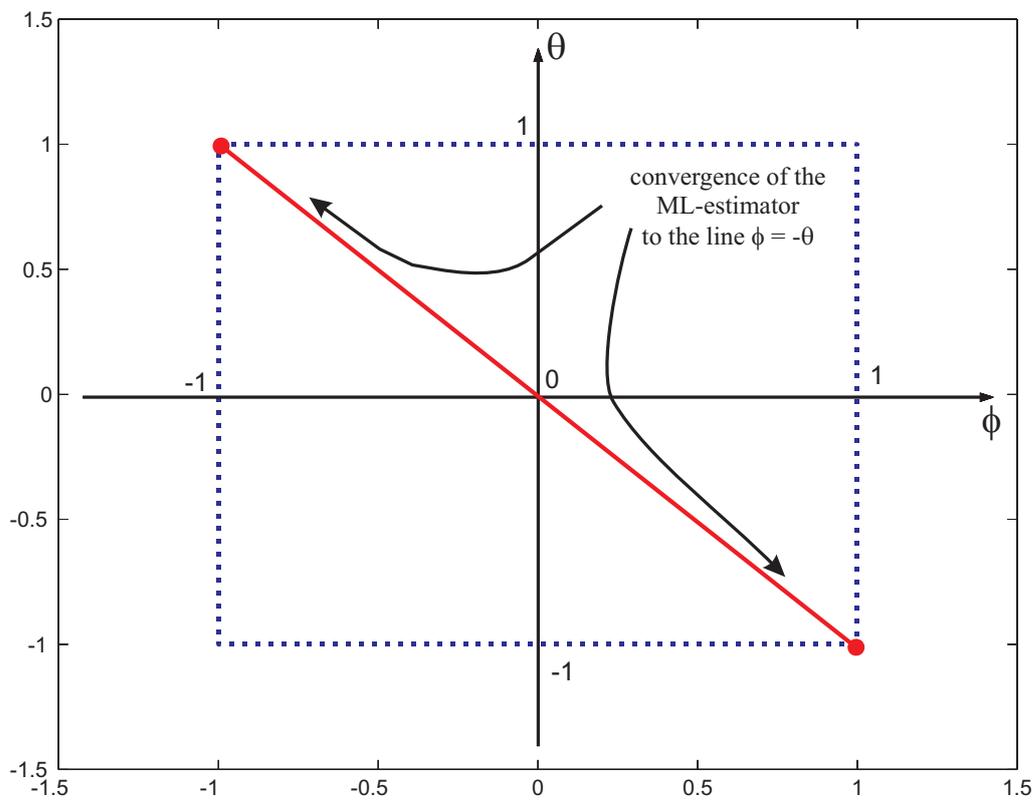


Figure 5.1: Parameter space of a causal and invertible ARMA(1,1) process

objective is to minimize a so-called *information criterion* over different values of  $p$  and  $q$ . These criteria are based on the following consideration. Given a fixed number of observations, the successive increase of the orders  $p$  and  $q$  increases the fit of the model so that variance of the residuals  $\hat{\sigma}_{p,q}^2$  steadily decreases. In order to compensate for this tendency to overfitting a penalty is introduced. This penalty term depends on the number of free parameters and on the number of observations at hand.<sup>6</sup> The most important information criteria have the following additive form:

$$\ln \hat{\sigma}_{p,q}^2 + (\# \text{ free parameters}) \frac{C(T)}{T} = \ln \hat{\sigma}_{p,q}^2 + (p + q) \frac{C(T)}{T} \longrightarrow \min_{p,q},$$

where  $\ln \hat{\sigma}_{p,q}^2$  measures the goodness of fit of the ARMA( $p,q$ ) model and  $(p + q) \frac{C(T)}{T}$  denotes the penalty term. Thereby  $C(T)$  represents a nondecreasing function of  $T$  which governs the trade-off between goodness of fit and complexity (dimension) of the model. Thus, the information criteria chooses higher order models for larger sample sizes  $T$ . If the model includes a constant term or other exogenous variables, the criterion must be adjusted accordingly. However, this will introduce, for a given sample size, just a constant term in the objective function and will therefore not influence the choice of  $p$  and  $q$ .

The most common criteria are the Akaike information criterion (AIC), the Schwarz or Bayesian information criterion (BIC), and the Hannan-Quinn information criterion (HQ criterion):

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

Because  $\text{AIC} < \text{HQC} < \text{BIC}$  for a given sample size  $T \geq 16$ , Akaike's criterion delivers the largest models, i.e. the highest order  $p+q$ ; the Bayesian criterion is more restrictive and delivers therefore the smallest models, i.e. the lowest  $p+q$ . Although Akaike's criterion is not consistent with respect to  $p$  and  $q$  and has a tendency to deliver overfitted models, it is still widely used in practice. This feature is sometimes desired as overfitting is seen as less damaging than underfitting.<sup>7</sup> Only the BIC and HQC lead to consistent estimates of the orders  $p$  and  $q$ .

<sup>6</sup>See Brockwell and Davis (1991) for details and a deeper appreciation.

<sup>7</sup>See, for example, the section on the unit root tests 7.3

## 5.5 Modeling a Stochastic Process

The identification of a satisfactory ARMA model typically involves in practice several steps.

### Step 1: Transformations to achieve stationary time series

Economic time series are often of a non-stationary nature. It is therefore necessary to transform the time series in a first step to achieve stationarity. Time series which exhibit a pronounced trend (GDP, stock market indices, etc.) should not be modeled in levels, but in differences. If the variable under consideration is already in logs, as is often case, then taking first differences effectively amounts to working with growth rates. Sometimes first differences are not enough and further differences have to be taken. Price indices or monetary aggregates are typical examples where first differences may not be sufficient to achieve stationarity. Thus instead of  $X_t$  one works with the series  $Y_t = (1 - L)^d X_t$  with  $d = 1, 2, \dots$ . A non-stationary process  $\{X_t\}$  which needs to be differentiated  $d$ -times to arrive at a stationary time series is called *integrated of order  $d$* ,  $X_t \sim I(d)$ .<sup>8</sup> If  $Y_t = (1 - L)^d X_t$  is generated by an ARMA(p,q) process,  $\{X_t\}$  is said to be an ARIMA(p,d,q) process.

An alternative method to eliminate the trend is to regress the time series against a polynomial  $t$  of degree  $s$ , i.e. against  $(1, t, \dots, t^s)$ , and to proceed with the residuals. These residuals can then be modeled as an ARMA(p,q) process. Chapter 7 discusses in detail which of the two detrending methods is to be preferred under which circumstances.

Often the data are subject to seasonal fluctuations. As with the trend there are several alternative available. The first possibility is to pass the time series through a seasonal filter and work with the seasonally adjusted data. The construction of seasonal filters is discussed in Chapter 6. A second alternative is to include seasonal dummies in the ARMA model. A third alternative is to take seasonal differences. In the case of quarterly observations, this amounts to work with  $Y_t = (1 - L^4)X_t$ . As  $1 - L^4 = (1 - L)(1 + L + L^2 + L^3)$ , this transformation involves a first difference and will therefore also account for the trend.

### Step 2: Finding the orders $p$ and $q$

Having achieved stationarity, one has to find the appropriate orders  $p$  and  $q$  of the ARMA model. Thereby one can rely either on the analysis of the

---

<sup>8</sup>An exact definition will be provided in Chapter 7. In this chapter we will analyze the consequences of non-stationarity and discuss tests for specific forms of non-stationarity.

ACF and the PACF, or on the information criteria outlined in the previous Section 5.4.

### Schritt 3: Checking the plausibility

After having identified a particular model or a set of models, one has to inspect its adequacy. There are several dimensions along which the model(s) can be checked.

- (i) Are the residuals white noise? This can be checked by investigating at the ACF and by applying the Ljung-Box test (4.4). If they are not this means that the model failed to capture all the dynamics inherent in the data.
- (ii) Are the parameters plausible?
- (iii) Are the parameters constant over time? Are there structural breaks? This can be done by looking at the residuals or by comparing parameter estimates across subsamples. More systematic approaches are discussed in Perron (2006). These involve the revolving estimation of parameters by allowing the break point to vary over the sample. Thereby different type of structural breaks can be distinguished.
- (iv) Does the model deliver sensible forecasts? It is particularly useful to investigate the out-of-sample forecasting performance. If one has several candidate models, one can perform a horse-race among them.

In case the model turns out to be unsatisfactory, one has to go back to steps 1 and 2.

## 5.6 An example: modeling real GDP of Switzerland

This section illustrates the concepts and ideas just presented by working out a specific example. We take the seasonally unadjusted Swiss real GDP as an example. The data are plotted in Figure 1.3. To take the seasonality into account we transform the logged time series by taking first seasonal differences, i.e.  $X_t = (1 - L^4) \ln \text{GDP}_t$ . Thus, the variable corresponds to the growth rate with respect to quarter of the previous year. The data are plotted in Figure 5.2. A cursory inspection of the plot reveals that this transformation eliminated the trend as well as the seasonality.

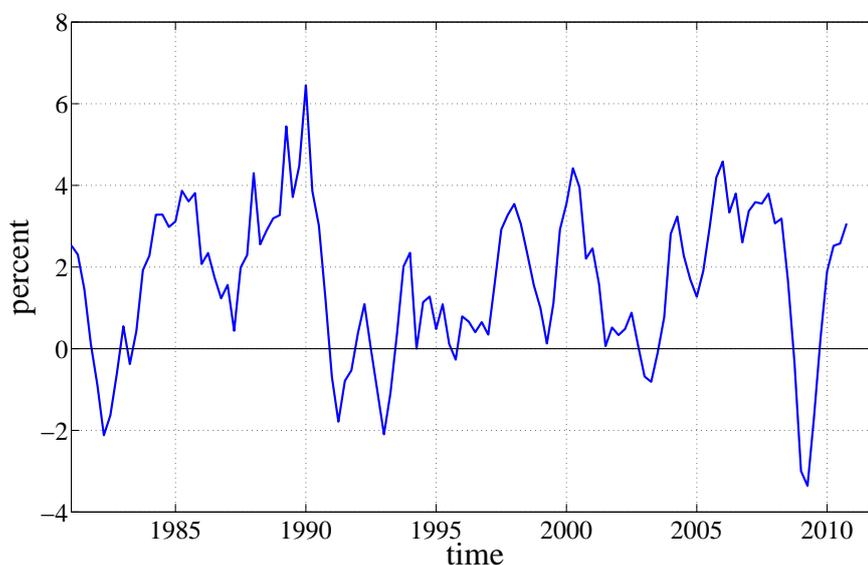


Figure 5.2: Real GDP growth rates of Switzerland

First we analyze the ACF and the PACF. They are plotted together with corresponding confidence intervals in Figure 5.3. The slowly monotonically declining ACF suggests an AR process. As only the first two orders of the PACF are significantly different from zero, it seems that an AR(2) model is appropriate. The least-squares estimate of this model are:

$$X_t - \underset{(0,103)}{1.134} X_{t-1} + \underset{(0,104)}{0.310} X_{t-2} = 0.218 + Z_t \quad \text{with } \hat{\sigma}^2 = 0.728$$

The numbers in parenthesis are the estimated standard errors of the corresponding parameter above. The roots of the AR-polynomial are 1.484 and 2.174. They are clearly outside the unit circle so that there exists a stationary and causal representation.

Next, we investigate the information criteria AIC and BIC to identify the orders of the ARMA(p,q) model. We examine all models with  $0 \leq p, q \leq 4$ . The AIC and the BIC values, are reported in Table 5.1 and 5.2. Both criteria reach a minimum at  $(p, q) = (1, 3)$  (bold numbers) so that both criteria prefer an ARMA(1,3) model. The parameters of this models are as follows:

$$X_t - \underset{(0,134)}{0.527} X_{t-1} = 0.6354 + Z_t + \underset{(0,1395)}{0.5106} Z_{t-1} \\ + \underset{(0,1233)}{0.5611} Z_{t-2} + \underset{(0,1238)}{0.4635} Z_{t-3} \quad \text{with } \hat{\sigma}^2 = 0.648.$$

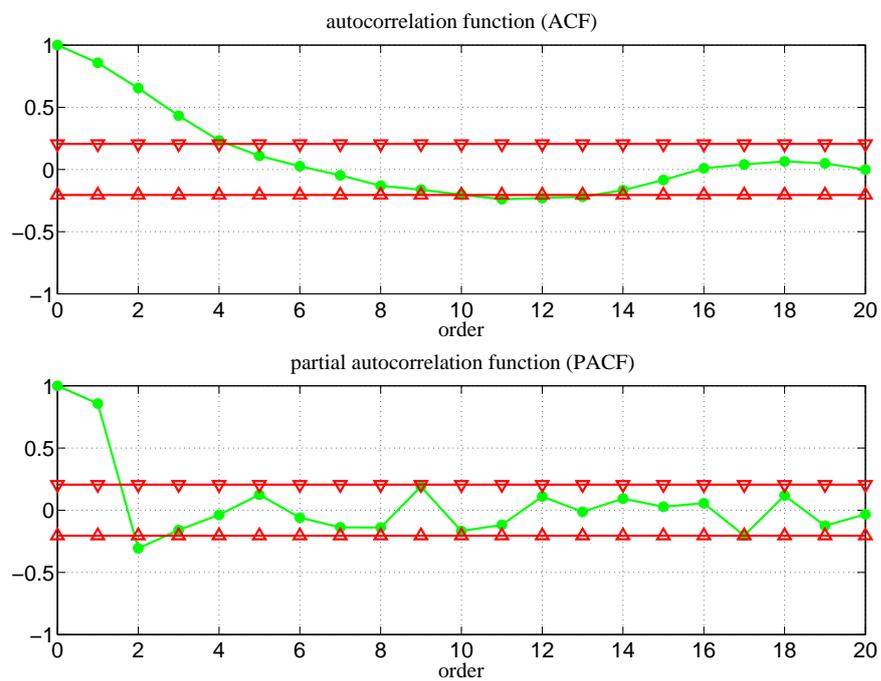


Figure 5.3: Autocorrelation (ACF) and partial autocorrelation (PACF) function (PACF) of real GDP growth rates of Switzerland with 95 percent confidence interval

Table 5.1: Values of Akaike's information criterium (AIC) for alternative ARMA(p,q) models

p	q				
	0	1	2	3	4
0		0.3021	0.0188	-0.2788	-0.3067
1	-0.2174	-0.2425	-0.2433	<b>-0.3446</b>	-0.2991
2	-0.2721	-0.2639	-0.2613	-0.3144	-0.2832
3	-0.2616	-0.2276	-0.2780	-0.2663	-0.2469
4	-0.2186	-0.1990	-0.2291	-0.2574	-0.2099

minimum in bold

The estimated standard errors of the estimated parameters are again reported in parenthesis below. The AR(2) model is not considerably worse than the ARMA(1,3) model, according to the BIC criterion it is even the second best model.

The inverted roots of the AR- and the MA-polynomial are plotted together with their corresponding 95 percent confidence regions in Figure 5.4.<sup>9</sup> As the confidence regions are all inside the unit circle, also the ARMA(1,3) has a stationary and causal representation. Moreover, the estimated process is also invertible. In addition, the roots of the AR- and the MA-polynomial are distinct.

The autocorrelation functions of the AR(2) and the ARMA(1,3) model are plotted in Figure 5.5. They show no sign of significant autocorrelations so that both residual series are practically white noise. We can examine this hypothesis formally by the Ljung-Box test (see Section 4.2 equation (4.4)). Taking  $N = 20$  the values of the test statistics are  $Q'_{AR(2)} = 33.80$  and  $Q'_{ARMA(1,3)} = 21.70$ , respectively. The 5 percent critical value according to the  $\chi^2_{20}$  distribution is 31.41. Thus the null hypothesis  $\rho(1) = \dots = \rho(20) = 0$  is rejected for the AR(2) model, but not for the ARMA(1,3) model. This implies that the AR(2) model does not capture the full dynamics of the data.

Although the AR(2) and the ARMA(1,3) model seem to be quite different at first glance, they deliver similar impulse response functions as can be

<sup>9</sup>The confidence regions are determined by the delta-method (see Appendix E).

Table 5.2: Values of Bayes' information criterium (BIC) for alternative ARMA(p,q) models

<b>p</b>	<b>q</b>				
	0	1	2	3	4
0		0.3297	0.0740	-0.1961	-0.1963
1	-0.1896	-0.1869	-0.1600	<b>-0.2335</b>	-0.1603
2	-0.2162	-0.1801	-0.1495	-0.1746	-0.1154
3	-0.1772	-0.1150	-0.1373	-0.0974	-0.0499
4	-0.1052	-0.0573	-0.0591	-0.0590	0.0169

minimum in bold

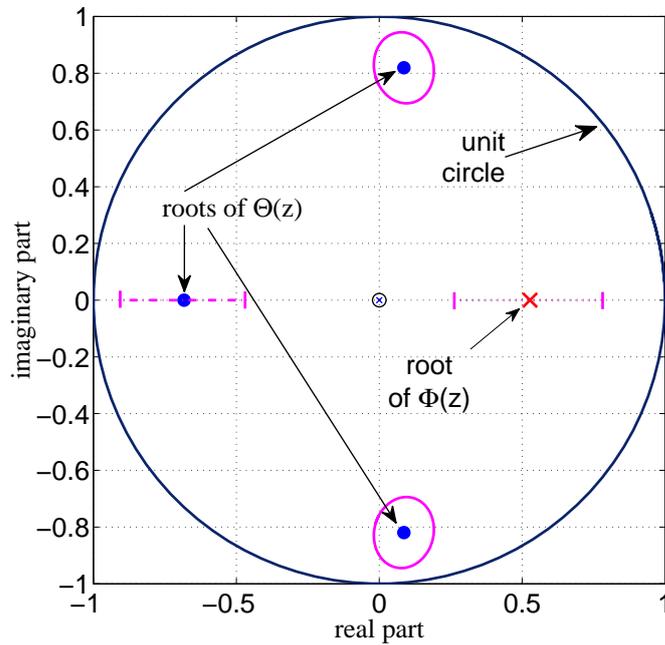


Figure 5.4: Inverted roots of the AR- and the MA-polynomial of the ARMA(1,3) model together with the corresponding 95 percent confidence regions

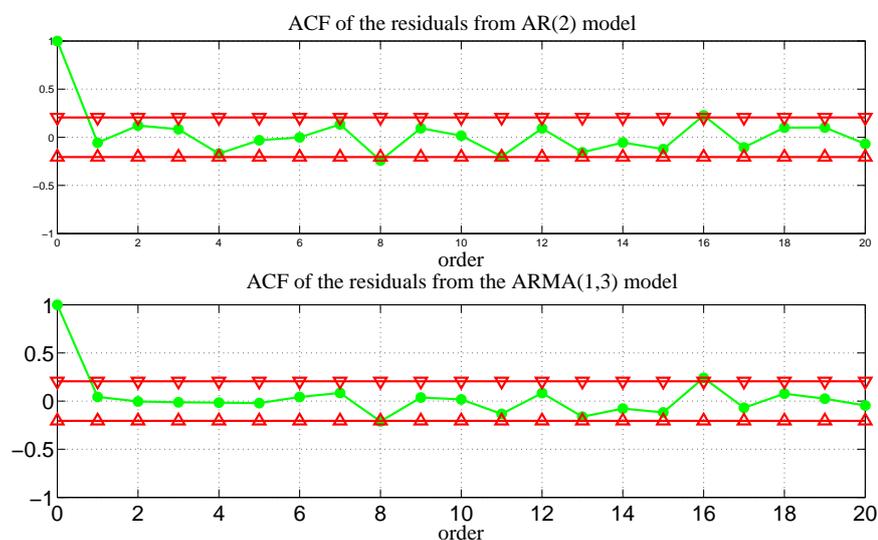


Figure 5.5: Autocorrelation function (ACF) of the residuals from the AR(2) and the ARMA(1,3) model

gathered from Figure 5.6. In both models, the impact of the initial shock is first built up to values higher than 1.1 in quarters one and two, respectively. Then the effect monotonically declines to zero. After 10 to 12 quarters the effect of the shock has practically dissipated.

As a final exercise, we use both models to forecast real GDP growth over the next nine quarters, i.e. for the period fourth quarter 2003 to fourth quarter 2005. As can be seen from Figure 5.7, both models predict that the Swiss economy should move out of recession in the coming quarters. However, the ARMA(1,3) model indicates that the recovery is taking place more quickly and the growth overshooting its long-run mean of 1.3 percent in about a year. The forecast of the AR(2) predicts a more steady approach to the long-run mean.

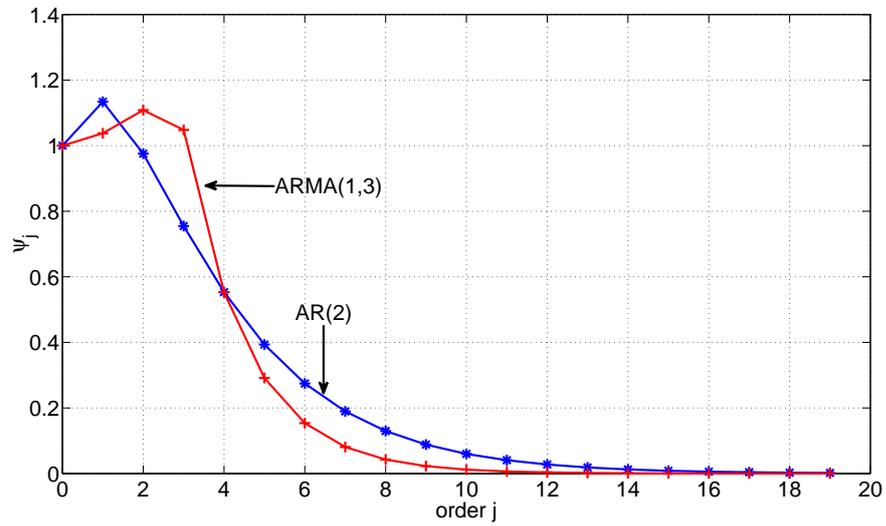


Figure 5.6: Impulse responses of the AR(2) and the ARMA(1,3) model

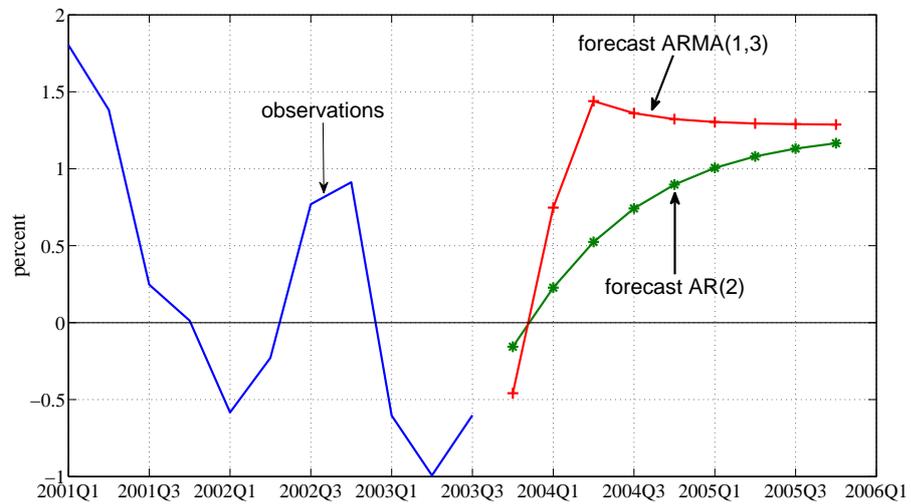


Figure 5.7: Forecasts of real GDP growth rates for Switzerland

## Chapter 6

# Spectral Analysis and Linear Filters

Up to now we have viewed a time series as a time indexed sequence of random variables. The class of ARMA process was seen as an adequate class of models for the analysis of stationary time series. This approach is usually termed as time series analysis in the *time domain*. There is, however, an equivalent perspective which views a time series as overlaid waves of different frequencies. This view point is termed time series analysis in the *frequency domain*. The decomposition of a time series into sinusoids of different frequencies is called the *spectral representation*. The estimation of the importance of the waves at particular frequencies is referred to as *spetral or spectrum estimation*. Priestley (1981) provides an excellent account of these methods. The use of frequency domain methods, in particular *spectrum estimation*, which originated in the natural sciences was introduced to economics by Granger (1964).<sup>1</sup> Notably, he showed that most of the fluctuations in economic time series can be attributed low frequencies cycles (Granger, 1966).

Although both approaches are equivalent, the analysis in the frequency domain is more convenient when it comes to the analysis and construction of *linear filters*. The application of a filter to a time series amounts to take some moving-average of the time series. These moving-average may extend, at least in theory, into the infinite past, but also into the infinite future. A causal ARMA process  $\{X_t\}$  may be regarded as filtered white-noise process with filter weights given by  $\psi_j, j = 1, 2, \dots$ . In economics, filters are usually applied to remove cycles of a particular frequency, like seasonal cycles (for example Christmas sales in a store), or to highlight particular cycles, like

---

<sup>1</sup>The use of spectral methods in the natural sciences can be traced many centuries back. The modern statistical approach builds on to the work of N. Wiener, G. U. Yule, J. W. Tukey, and many others. See the interesting survey by Robinson (1982).

business cycles.

From a mathematical point of view, the equivalence between time and frequency domain analysis rest on the theory of *Fourier series*. An adequate representation of this theory is beyond the scope of this book. The interested reader may consult Brockwell and Davis (1991, chapter 4). An introduction to the underlying mathematical theory can be found in standard textbooks like Rudin (1987).

## 6.1 The Spectral Density

In the following, we assume that  $\{X_t\}$  is a mean-zero (centered) stationary stochastic process with autocovariance function  $\gamma(h)$ ,  $h = 0, \pm 1, \pm 2, \dots$ . Mathematically,  $\gamma(h)$  represents an double-infinite sequence which can be mapped into a real valued function  $f(\lambda)$ ,  $\lambda \in \mathbb{R}$ , by the Fourier transform. This function is called the *spectral density function* or *spectral density*. Conversely, we retrieve from the spectral density each covariance. Thus, we have a one-to-one relation between autocovariance functions and spectral densities: both objects summarize the same properties of the time series, but represent them differently.

**Definition 6.1.** *Let  $\{X_t\}$  be a mean-zero stationary stochastic process absolutely summable autocovariance function  $\gamma$  then the function*

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h)e^{-ih\lambda}, \quad -\infty < \lambda < \infty, \quad (6.1)$$

*is called the spectral density function or spectral density of  $\{X_t\}$ . Thereby  $i$  denotes the imaginary unit (see Appendix A).*

The sine is an odd function whereas the cosine and the autocovariance function are even functions.<sup>2</sup> This implies that the spectral density can be rewritten as:

$$\begin{aligned} f(\lambda) &= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h)(\cos(h\lambda) - i \sin(h\lambda)) \\ &= \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \cos(h\lambda) + 0 = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) \cos(-h\lambda) \\ &= \frac{\gamma(0)}{2\pi} + \frac{1}{\pi} \sum_{h=1}^{\infty} \gamma(h) \cos(h\lambda). \end{aligned} \quad (6.2)$$

---

<sup>2</sup>A function  $f$  is called even if  $f(-x) = f(x)$ ; the function is called odd if  $f(-x) = -f(x)$ . Thus, we have  $\sin(-\theta) = -\sin(\theta)$  and  $\cos(-\theta) = \cos(\theta)$ .

Because of the periodicity of the cosine function, i.e. because

$$f(\lambda + 2k\pi) = f(\lambda), \quad \text{for all } k \in \mathbb{Z},$$

it is sufficient to consider the spectral density only in the interval  $(-\pi, \pi]$ . As the cosine is an even function so is  $f$ . Thus, we restrict the analysis of the spectral density  $f(\lambda)$  further to the domain  $\lambda \in [0, \pi]$ .

In practice, we often use the *period* or *oscillation length* instead of the radiant  $\lambda$ . They are related by the formula:

$$\text{period length} = \frac{2\pi}{\lambda}. \quad (6.3)$$

If, for example, the data are quarterly observations, a value of 0.3 for  $\lambda$  corresponds to a period length of approximately 21 quarters.

**Remark 6.1.** *We gather some properties of the spectral density function  $f$ :*

- *Because  $f(0) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h)$ , the long-run variance of  $\{X_t\}$   $J$  (see Section 4.4) equals  $2\pi f(0)$ , i.e.  $2\pi f(0) = J$ .*
- *$f$  is an even function so that  $f(\lambda) = f(-\lambda)$ .*
- *$f(\lambda) \geq 0$  for all  $\lambda \in (-\pi, \pi]$ . The proof of this proposition can be found in Brockwell and Davis (1996, chapter 4). This property corresponds to the non-negative definiteness of the autocovariance function (see property 4 in Theorem 1.1 of Section 1.3).*
- *The single autocovariances are the Fourier-coefficients of the spectral density  $f$ :*

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

*For  $h = 0$ , we therefore get  $\gamma(0) = \int_{-\pi}^{\pi} f(\lambda) d\lambda$ .*

The last property allows us to compute the autocovariances from a given spectral density. It shows how time and frequency domain analysis are related to each other and how a property in one domain is reflected as a property in the other.

These properties of a non-negative definite function can be used to characterize the spectral density of a stationary process  $\{X_t\}$  with autocovariance function  $\gamma$ .

**Theorem 6.1** (Properties of a spectral density). *A function  $f$  defined on  $(-\pi, \pi]$  is the spectral density of a stationary process if and only if the following properties hold:*

- $f(\lambda) = f(-\lambda)$ ;
- $f(\lambda) \geq 0$ ;
- $\int_{-\pi}^{\pi} f(\lambda) d\lambda < \infty$ .

**Corollary 6.1.** *An absolutely summable function  $\gamma$  is the autocovariance function of a stationary process if and only if*

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-ih\lambda} \geq 0, \quad \text{for all } \lambda \in (-\pi, \pi].$$

In this case  $f$  is called the spectral density of  $\gamma$ .

The function  $f(\lambda)/\gamma(0)$  can be considered as a density function of some probability distribution defined on  $[-\pi, \pi]$  because  $\frac{f(\lambda)}{\gamma(0)} \geq 0$  and

$$\int_{-\pi}^{\pi} \frac{f(\lambda)}{\gamma(0)} d\lambda = 1.$$

The corresponding cumulative distribution function  $G$  is then defined as:

$$G(\lambda) = \int_{-\pi}^{\lambda} \frac{f(\omega)}{\gamma(0)} d\omega, \quad -\pi \leq \lambda \leq \pi.$$

It satisfies:  $G(-\pi) = 0$ ,  $G(\pi) = 1$ ,  $1 - G(\lambda) = G(\lambda)$ , and  $G(0) = 1/2$ . The autocorrelation function  $\rho$  is then given by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \int_{-\pi}^{\pi} e^{ih\lambda} dG(\lambda).$$

### Some examples

Some relevant examples illustrating the above are:

**white noise:** Let  $\{X_t\}$  be a white noise process with  $X_t \sim \text{WN}(0, \sigma^2)$ . For this process all autocovariances, except  $\gamma(0)$ , are equal to zero. The spectral density therefore is equal to

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

Thus, the spectral density is equal to a constant which is proportional to the variance. This means that no particular frequency dominates the spectral density. This is the reason why such a process is called white noise.

**MA(1):** Let  $\{X_t\}$  be a MA(1) process with autocovariance function

$$\gamma(h) = \begin{cases} 1, & h = 0; \\ \rho, & h = \pm 1; \\ 0, & \text{otherwise.} \end{cases}$$

The spectral density therefore is:

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h)e^{-ih\lambda} = \frac{\rho e^{i\lambda} + 1 + \rho e^{-i\lambda}}{2\pi} = \frac{1 + 2\rho \cos \lambda}{2\pi}.$$

Thus,  $f(\lambda) \geq 0$  if and only if  $|\rho| \leq 1/2$ . According to Corollary 6.1 above,  $\gamma$  is the autocovariance function of a stationary stochastic process if and only if  $|\rho| \leq 1/2$ . This condition corresponds exactly to the condition derived in the time domain (see Section 1.3). The spectral density for  $\rho = 0.4$  or equivalently  $\theta = 0.5$ , respectively for  $\rho = -0.4$  or equivalently  $\theta = -0.5$ , and  $\sigma^2 = 1$  is plotted in Figure 6.1(a). As the process is rather smooth when the first order autocorrelation is positive, the spectral density is large in the neighborhood of zero and small in the neighborhood of  $\pi$ . For a negative autocorrelation the picture is just reversed.

**AR(1):** The spectral density of an AR(1) process  $X_t = \phi X_{t-1} + Z_t$  with  $Z_t \sim \text{WN}(0, \sigma^2)$  is:

$$\begin{aligned} f(\lambda) &= \frac{\gamma(0)}{2\pi} \left( 1 + \sum_{h=1}^{\infty} \phi^h (e^{-ih\lambda} + e^{ih\lambda}) \right) \\ &= \frac{\sigma^2}{2\pi(1-\phi^2)} \left( 1 + \frac{\phi e^{i\lambda}}{1-\phi e^{i\lambda}} + \frac{\phi e^{-i\lambda}}{1-\phi e^{-i\lambda}} \right) = \frac{\sigma^2}{2\pi} \frac{1}{1-2\phi \cos \lambda + \phi^2} \end{aligned}$$

The spectral density for  $\phi = 0.6$  and  $\phi = -0.6$  and  $\sigma^2 = 1$  are plotted in Figure 6.1(b). As the process with  $\phi = 0.6$  exhibits a relatively large positive autocorrelation so that it is rather smooth, the spectral density takes large values for low frequencies. In contrast, the process with  $\phi = -0.6$  is rather volatile due to the negative first order autocorrelation. Thus, high frequencies are more important than low frequencies as reflected in the corresponding figure.

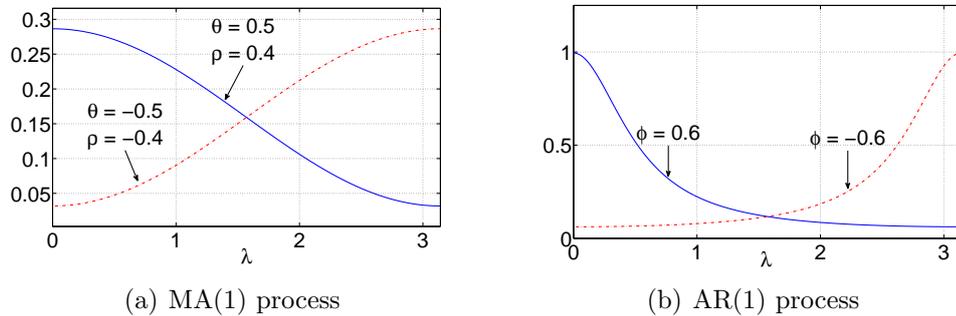


Figure 6.1: Examples of spectral densities with  $Z_t \sim \text{WN}(0, 1)$

Note that, as  $\phi$  approaches one, the spectral density evaluated at zero tends to infinity, i.e.  $\lim_{\lambda \downarrow 0} f(\lambda) = \infty$ . This can be interpreted in the following way. As the process gets closer to a random walk more and more weight is given to long-run fluctuations (cycles with very low frequency or very high periodicity) (Granger, 1966).

## 6.2 Spectral Decomposition of a Time Series

Consider the simple harmonic process  $\{X_t\}$  which just consists of a cosine and a sine wave:

$$X_t = A \cos(\omega t) + B \sin(\omega t). \quad (6.4)$$

Thereby  $A$  and  $B$  are two uncorrelated random variables with  $\mathbb{E}A = \mathbb{E}B = 0$  and  $\mathbb{V}A = \mathbb{V}B = 1$ . The autocovariance function of this process is  $\gamma(h) = \cos(\omega h)$ . This autocovariance function cannot be represented as  $\int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda$ . However, it can be regarded as the Fourier transform of a discrete distribution function  $F$ :

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

where

$$F(\lambda) = \begin{cases} 0, & \text{for } \lambda < -\omega; \\ 1/2, & \text{for } -\omega \leq \lambda < \omega; \\ 1, & \text{for } \lambda \geq \omega. \end{cases} \quad (6.5)$$

The integral with respect to the discrete distribution function is a so-called Riemann-Stieltjes integral.<sup>3</sup>  $F$  is a step function with jumps at  $-\omega$  and  $\omega$  and step size of  $1/2$  so that the above integral equals  $\frac{1}{2}e^{-i h \omega} + \frac{1}{2}e^{i h \omega} = \cos(h \omega)$ .

These considerations lead to a representation, called the *spectral representation*, of the autocovariance function as the Fourier transform a distribution function over  $[-\pi, \pi]$ .

**Theorem 6.2** (Spectral representation).  $\gamma$  is the autocovariance function of a stationary process  $\{X_t\}$  if and only if there exists a right-continuous, nondecreasing, bounded function  $F$  on  $(-\pi, \pi]$  with the properties  $F(-\pi) = 0$  and

$$\gamma(h) = \int_{(-\pi, \pi]} e^{i h \lambda} dF(\lambda). \quad (6.6)$$

$F$  is called the spectral distribution function of  $\gamma$ .

**Remark 6.2.** If the spectral distribution function  $F$  has a density  $f$  such that  $F(\lambda) = \int_{-\pi}^{\lambda} f(\omega) d\omega$  then  $f$  is called the spectral density and the time series is said to have a continuous spectrum.

**Remark 6.3.** According to the Lebesgue-Radon-Nikodym Theorem (see, for example, Rudin (1987)), the spectral distribution function  $F$  can be represented uniquely as the sum of a distribution function  $F_Z$  which is absolutely continuous with respect to the Lebesgue measure and a discrete distribution function  $F_V$ . The distribution function  $F_Z$  corresponds to the regular part of the Wold Decomposition (see Theorem 3.1 in Section 3.2) and has spectral density

$$f_Z(\lambda) = \frac{\sigma^2}{2\pi} |\Psi(e^{-i\lambda})|^2 = \frac{\sigma^2}{2\pi} \left| \sum_{j=0}^{\infty} \psi_j e^{-ij\lambda} \right|^2.$$

The discrete distribution  $F_V$  corresponds to the deterministic part  $\{V_t\}$ .

The process (6.4) considers just a single frequency  $\omega$ . We may, however, generalize this process by superposing several sinusoids. This leads to the class of *harmonic processes*:

$$X_t = \sum_{j=1}^k A_j \cos(\omega_j t) + B_j \sin(\omega_j t), \quad 0 < \omega_1 < \dots < \omega_k < \pi \quad (6.7)$$

---

<sup>3</sup>The Riemann-Stieltjes integral is a generalization of the Riemann integral. Let  $f$  and  $g$  be two bounded functions defined on the interval  $[a, b]$  then the Riemann-Stieltjes integral  $\int_a^b f(x) dg(x)$  is defined as  $\lim_{n \rightarrow \infty} \sum_{i=1}^n f(\xi_i)[g(x_i) - g(x_{i-1})]$  where  $a = x_1 < x_2 < \dots < x_{n-1} < x_n = b$ . For  $g(x) = x$  we obtain the standard Riemann integral. If  $g$  is a step function with a countable number of steps  $x_i$  of height  $h_i$  then  $\int_a^b f(x) dg(x) = \sum_i f(x_i) h_i$ .

where  $A_1, B_1, \dots, A_k, B_k$  are random variables which are uncorrelated with each other and which have means  $\mathbb{E}A_j = \mathbb{E}B_j = 0$  and variances  $\mathbb{V}A_j = \mathbb{V}B_j = \sigma_j^2$ ,  $j = 1, \dots, k$ . The autocovariance function of such a process is given by  $\gamma(h) = \sum_{j=1}^k \sigma_j^2 \cos(\omega_j h)$ . According to the spectral representation theorem the corresponding distribution function  $F$  can be represented as a weighted sum of distribution functions like those in equation (6.5):

$$F(\lambda) = \sum_{j=1}^k \sigma_j^2 F_j(\lambda)$$

with

$$F_j(\lambda) = \begin{cases} 0, & \text{for } \lambda < -\omega_j; \\ 1/2, & \text{for } -\omega_j \leq \lambda < \omega_j; \\ 1, & \text{for } \lambda \geq \omega_j. \end{cases}$$

This generalization points to the the following properties:

- Each of the  $k$  components  $A_j \cos(\omega_j t) + B_j \sin(\omega_j t)$ ,  $j = 1, \dots, k$ , is completely associated to a specific frequency  $\omega_j$ .
- The  $k$  components are uncorrelated with each other.
- The variance of each component is  $\sigma_j^2$ . The contribution of each component to the variance of  $X_t$  given by  $\sum_{j=1}^k \sigma_j^2$  therefore is  $\sigma_j^2$ .
- $F$  is a nondecreasing step-function with jumps at frequencies  $\omega = \pm\omega_j$  and step sizes  $\frac{1}{2}\sigma_j^2$ .
- The corresponding probability distribution is discrete with values  $\frac{1}{2}\sigma_j^2$  at the frequencies  $\omega = \pm\omega_j$  and zero otherwise.

The interesting feature of harmonic processes as represented in equation (6.7) is that *every* stationary process can be represented as the superposition of uncorrelated sinusoids. However, in general infinitely many (even uncountably many) of these processes have to be superimposed. The generalization of (6.7) then leads to the spectral representation of a stationary stochastic process:

$$X_t = \int_{(-\pi, \pi]} e^{it\lambda} dZ(\lambda). \quad (6.8)$$

Thereby  $\{Z(\lambda)\}$  is a complex-valued stochastic process with uncorrelated increments defined on the interval  $(-\pi, \pi]$ . The above representation is known

as the *spectral representation* of the process  $\{X_t\}$ .<sup>4</sup> Note the analogy to the spectral representation of the autocovariance function in equation (6.6).

For the harmonic processes in equation (6.7), we have:

$$dZ(\lambda) = \begin{cases} \frac{A_j + iB_j}{2}, & \text{for } \lambda = -\omega_j \text{ and } j = 1, 2, \dots, k; \\ \frac{A_j - iB_j}{2}, & \text{for } \lambda = \omega_j \text{ and } j = 1, 2, \dots, k; \\ 0, & \text{otherwise.} \end{cases}$$

In this case the variance of  $dZ$  is given by:

$$\mathbb{E}dZ(\lambda)\overline{dZ(\lambda)} = \begin{cases} \frac{\sigma_j^2}{2}, & \text{if } \lambda = \pm\omega_j; \\ 0, & \text{otherwise.} \end{cases}$$

In general, we have:

$$= \begin{cases} F(\lambda) - F(\lambda^-), & \text{discrete spectrum;} \\ f(\lambda)d\lambda, & \text{continuous spectrum.} \end{cases}$$

Thus, a large jump of the spectrum at frequency  $\lambda$  is associated with a large sinusoidal component with frequency  $\lambda$ .<sup>5</sup>

## 6.3 The Periodogram and the Estimation of Spectral Densities

Although the spectral distribution function is uniquely determined, its estimation from a finite sample with realizations  $\{x_1, x_2, \dots, x_T\}$  is not easy. This has to do with the problem of estimating a function from a finite number of points. We will present two-approaches: a non-parametric and a parametric one.

### 6.3.1 Non-parametric Estimation

A simple estimator of the spectral density,  $\hat{f}_T(\lambda)$ , can be obtained by replacing in the defining equation (6.1) the theoretical autocovariances  $\gamma$  by their estimates  $\hat{\gamma}$ . However, instead of a simple sum, we consider a weighted sum:

$$\hat{f}_T(\lambda) = \frac{1}{2\pi} \sum_{|h| \leq \ell_T} k \left( \frac{h}{\ell_T} \right) \hat{\gamma}(h) e^{-ih\lambda}. \quad (6.9)$$

<sup>4</sup>A mathematically precise statement is given in Brockwell and Davis (1991, chapter 4) where also the notion of stochastic integration is explained

<sup>5</sup>Thereby  $F(\lambda^-)$  denotes the left-sided limit, i.e.  $F(\lambda^-) = \lim_{\omega \uparrow \lambda} F(\omega)$ .

The weighting function  $k$ , also known as the *lag window*, is assumed to have exactly the same properties as the kernel function introduced in Section 4.4. This correspondence is not accidental, indeed the long-run variance defined in equation (4.1) is just  $2\pi$  times the spectral density evaluated at  $\lambda = 0$ . Thus, one might choose a weighting, kernel or lag window from Table 4.1, like the Bartlett-window, and use it to estimate the spectral density. The lag truncation parameter is chosen in such a way that  $\ell_T \rightarrow \infty$  as  $T \rightarrow \infty$ . The rate of divergence should, however, be smaller than  $T$  so that  $\frac{\ell_T}{T}$  approaches zero as  $T$  goes to infinity. As an estimator of the autocovariances one uses the estimator given in equation (4.2) of Section 4.2.

The above estimator is called an *indirect spectral estimator* because it requires the estimation of the autocovariances in the first step. The *periodogram* provides an alternative *direct spectral estimator*. For this purpose, we represent the observations as linear combinations of sinusoids of specific frequencies. These so-called *Fourier frequencies* are defined as  $\omega_k = \frac{2\pi k}{T}$ ,  $k = -\lfloor \frac{T-1}{2} \rfloor, \dots, \lfloor \frac{T}{2} \rfloor$ . Thereby  $\lfloor x \rfloor$  denotes the largest integer smaller or equal to  $x$ . With this notation, the observations  $x_t$ ,  $t = 1, \dots, T$ , can be represented as a sum of sinusoids:

$$x_t = \sum_{k=-\lfloor \frac{T-1}{2} \rfloor}^{\lfloor \frac{T}{2} \rfloor} a_k e^{i\omega_k t} = \sum_{k=-\lfloor \frac{T-1}{2} \rfloor}^{\lfloor \frac{T}{2} \rfloor} a_k (\cos(\omega_k t) + i \sin(\omega_k t)).$$

The coefficients  $\{a_k\}$  are the *discrete Fourier-transform* of the observations  $\{x_1, x_2, \dots, x_T\}$ . The periodogram  $I_T$  is then defined as follows.

**Definition 6.2** (Periodogram). *The periodogram of the observations  $\{x_1, x_2, \dots, x_T\}$  is the function*

$$I_T(\lambda) = \frac{1}{T} \left| \sum_{t=1}^T x_t e^{-it\lambda} \right|^2.$$

For each Fourier-frequency  $\omega_k$ , the periodogram  $I_T(\omega_k)$  equals  $|a_k|^2$ . This implies that

$$\sum_{t=1}^T |x_t|^2 = \sum_{k=-\lfloor \frac{T-1}{2} \rfloor}^{\lfloor \frac{T}{2} \rfloor} |a_k|^2 = \sum_{k=-\lfloor \frac{T-1}{2} \rfloor}^{\lfloor \frac{T}{2} \rfloor} I_T(\omega_k).$$

The value of the periodogram evaluated at the Fourier-frequency  $\omega_k$  is therefore nothing but the contribution of the sinusoid with frequency  $\omega_k$  to the variation of  $\{x_t\}$  as measured by sum of squares. In particular, we get for

any Fourier-frequency different from zero that

$$I_T(\omega_k) = \sum_{h=-T+1}^{T-1} \hat{\gamma}(h) e^{-ih\omega_k}.$$

Thus the periodogram represents, disregarding the proportionality factor  $2\pi$ , the sample analogue of the spectral density.

Unfortunately, it turns out that the periodogram is not a consistent estimator of the spectral density. In particular, the covariance between  $I_T(\lambda_1)$  and  $I_T(\lambda_2)$ ,  $\lambda_1 \neq \lambda_2$ , goes to zero for  $T$  going to infinity. The periodogram thus has tendency to get very jagged for large  $T$  leading to the detection of spurious sinusoids. A way out of this problem is to average the periodogram between neighboring frequencies. This makes sense because the variance is relatively constant within a small frequency band. The averaging (smoothing) of the periodogram over neighboring frequencies leads to the class of *discrete spectral average estimators* which turn out to be consistent:

$$\hat{f}_T(\lambda) = \frac{1}{2\pi} \sum_{|h| \leq \ell_T} K_T(h) I_T \left( \tilde{\omega}_{T,\lambda} + \frac{2\pi h}{T} \right) \quad (6.10)$$

where  $\tilde{\omega}_{T,\lambda}$  denotes the multiple of  $\frac{2\pi}{T}$  which is closest to  $\lambda$ .  $\ell_T$  is the bandwidth of the estimator, i.e. the number of ordinates over which the average is taken.  $\ell_T$  satisfies the same properties as in the case of the indirect spectral estimator (6.9):  $\ell_T \rightarrow \infty$  and  $\ell_T/T \rightarrow 0$  for  $T \rightarrow \infty$ . Thus, as  $T$  goes to infinity, on the one hand the average is taken over more and more values, but on the other hand the frequency band over which the average is taken is getting smaller and smaller. The *spectral weighting function*  $K_T$  is a positive even function satisfying  $\sum_{|h| \leq \ell_T} K_T(h) = 1$  and  $\sum_{|h| \leq \ell_T} K_T^2(h) \rightarrow 0$  for  $T \rightarrow \infty$ . It can be shown that under these conditions the discrete spectral average estimator is consistent. For details and the asymptotic distribution the interested reader is referred to Brockwell and Davis (1991, Chapter 10).

A simple weighting function is given by  $K_T(h) = (2\ell_T + 1)^{-1}$  for  $|h| \leq \ell_T$  where  $\ell_T = \sqrt{T}$ . This function corresponds to the Daniell kernel function (see Section 4.4). It averages over  $2\ell_T + 1$  values within a frequency band of approximate width  $\frac{4\pi}{\sqrt{T}}$ . In practice, the sample size is fixed and the researcher faced with a trade-off between variance and bias. On the one hand, a weighting function which averages over a wide frequency band produces a smooth spectral density, but has probably a large bias because the estimate of  $f(\lambda)$  depends on frequencies which are rather far away from  $\lambda$ . On the other hand, a weighting function which averages only over a small frequency band produces a small bias, but probably a large variance. It is thus advisable in

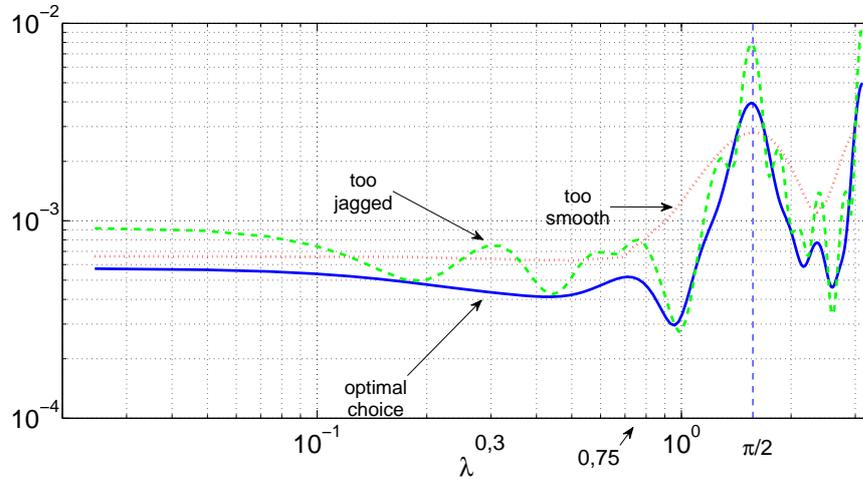


Figure 6.2: Non-parametric direct estimates of a spectral density with alternative weighting functions

practice to work with alternative weighting functions and to choose the one which delivers a satisfying balance between bias and variance.

Figure 6.2 illustrates these considerations by estimating the spectral density of quarterly growth rates of real investment in constructions for the Swiss economy using alternative weighting functions. To obtain a better graphical resolution we have plotted the estimates on a logarithmic scale. All three estimates show a peak (local maximum) at the frequency  $\lambda = \frac{\pi}{2}$ . This corresponds to a wave with a period of one year. The estimator with a comparably wide frequency band (dotted line) smoothes the minimum  $\lambda = 1$  away. The estimator with a comparable small frequency band (dashed line), on the contrary, reveals additional waves with frequencies  $\lambda = 0.75$  and  $0.3$  which correspond to periods of approximately two, respectively five years. Whether these waves are just artifacts of the weighting function or whether there really exist cycles of that periodicity remains open.

### 6.3.2 Parametric Estimation

An alternative to the nonparametric approaches just outlined consists in the estimation of an ARMA model and followed by deducing the spectral density from it. This approach was essentially first proposed by Yule (1927).

**Theorem 6.3.** *Let  $\{X_t\}$  be a causal ARMA( $p, q$ ) process given by  $\Phi(L)X_t =$*

$\Theta(L)Z_t$  and  $Z_t \sim WN(0, \sigma^2)$ . Then the spectral density  $f_X$  is given by

$$f_X(\lambda) = \frac{\sigma^2 |\Theta(e^{-i\lambda})|^2}{2\pi |\Phi(e^{-i\lambda})|^2}, \quad -\pi \leq \lambda \leq \pi. \quad (6.11)$$

*Proof.*  $\{X_t\}$  is generated by applying the linear filter  $\Psi(L)$  with transfer function  $\Psi(e^{-i\lambda}) = \frac{\Theta(e^{-i\lambda})}{\Phi(e^{-i\lambda})}$  to  $\{Z_t\}$  (see Section 6.4). Formula (6.11) is then an immediate consequence of theorem 6.5 because the spectral density of  $\{Z_t\}$  is equal to  $\frac{\sigma^2}{2\pi}$ .  $\square$

**Remark 6.4.** As the spectral density of an ARMA process  $\{X_t\}$  is given by a quotient of trigonometric functions, the process is said to have a rational spectral density.

The spectral density of the AR(2) process  $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$  with  $Z_t \sim WN(0, \sigma^2)$ , for example, is then given by

$$f_X(\lambda) = \frac{\sigma^2}{2\pi(1 + \phi_1^2 + 2\phi_2 + \phi_2^2 + 2(\phi_1\phi_2 - \phi_1)\cos\lambda - 4\phi_2\cos^2\lambda)}.$$

The spectral density of an ARMA(1,1) process  $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$  with  $Z_t \sim WN(0, \sigma^2)$  is

$$f_X(\lambda) = \frac{\sigma^2(1 + \theta^2 + 2\theta\cos\lambda)}{2\pi(1 + \phi^2 + 2\phi\cos\lambda)}.$$

An estimate of the spectral density is then obtained by replacing the unknown coefficients of the ARMA model by their corresponding estimates and by applying Formula 6.11 to the estimated model. Figure 6.3 compares the nonparametric to the parametric method based on an AR(4) model using the same data as in Figure 6.2. Both methods produce similar estimates. They clearly show waves of periodicity of half a year and a year, corresponding to frequencies  $\frac{\pi}{2}$  and  $\pi$ . The nonparametric estimate is, however, more volatile in the frequency band  $[0.6, 1]$  and around 2.5.

## 6.4 Linear Time-invariant Filters

Time-invariant linear filters are an indispensable tool in time series analysis. Their objective is to eliminate or amplify waves of a particular periodicity. For example, they may be used to purge a series from seasonal movements to get a seasonally adjusted time series which should reflect the cyclical movements more strongly. The spectral analysis provides just the right tools to construct and analyze such filters.

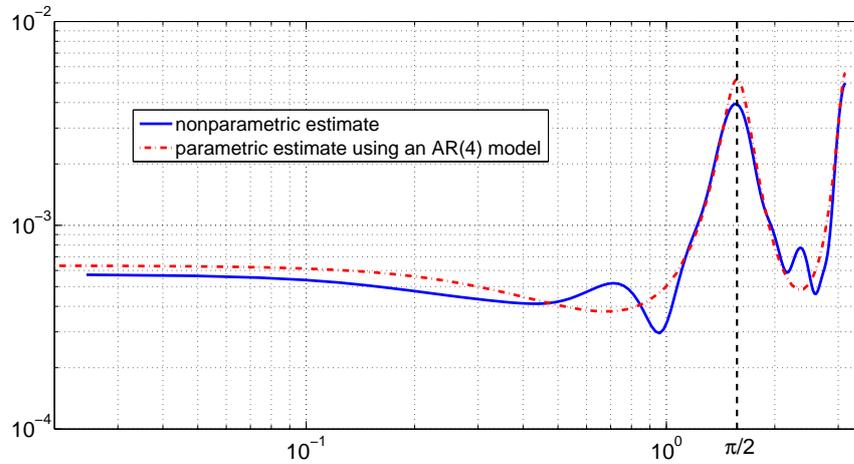


Figure 6.3: Comparison of nonparametric and parametric estimates of the spectral density of the growth rate of investment in the construction sector

**Definition 6.3.**  $\{Y_t\}$  is the output of the linear time-invariant filter (LTF)  $\Psi = \{\psi_j, j = 0, \pm 1, \pm 2, \dots\}$  applied to the input  $\{X_t\}$  if

$$Y_t = \Psi(L)X_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} \quad \text{with} \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

The filter is called causal or one-sided if  $\psi_j = 0$  for  $j < 0$ ; otherwise it is called two-sided.

**Remark 6.5.** Time-invariance in this context means that the lagged process  $\{Y_{t-s}\}$  is obtained for all  $s \in \mathbb{Z}$  from  $\{X_{t-s}\}$  by applying the same filter  $\Psi$ .

**Remark 6.6.** MA processes, causal AR processes and causal ARMA processes can be viewed as filtered white noise processes.

It is important to recognize that the application of a filter systematically changes the autocorrelation properties of the original time series. This may be warranted in some cases, but may lead to the “discovery” of spurious regularities which just reflect the properties of the filter. See the example of the Kuznets filter below.

**Theorem 6.4** (Autocovariance function of filtered process). Let  $\{X_t\}$  be a mean-zero stationary process with autocovariance function  $\gamma_X$ . Then the

filtered process  $\{Y_t\}$  defined as

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} = \Psi(L)X_t$$

with  $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$  is also a mean-zero stationary process with autocovariance function  $\gamma_Y$ . Thereby the two autocovariance functions are related as follows:

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h+k-j), \quad h = 0, \pm 1, \pm 2, \dots$$

*Proof.* We first show the existence of the output process  $\{Y_t\}$ . For this end, consider the sequence of random variables  $\{Y_t^{(m)}\}_{m=1,2,\dots}$  defined as

$$Y_t^{(m)} = \sum_{j=-m}^m \psi_j X_{t-j}.$$

To show that the limit for  $m \rightarrow \infty$  exists in the mean square sense, it is, according to Theorem C.6, enough to verify the Cauchy criterion

$$\mathbb{E} \left| Y_t^{(m)} - Y_t^{(n)} \right|^2 \longrightarrow 0, \quad \text{for } m, n \rightarrow \infty.$$

Taking without loss of generality  $m > n$ , Minkowski's inequality (see Theorem C.2 or triangular inequality) leads to

$$\begin{aligned} & \left( \mathbb{E} \left| \sum_{j=-m}^m \psi_j X_{t-j} - \sum_{j=-n}^n \psi_j X_{t-j} \right|^2 \right)^{1/2} \\ & \leq \left( \mathbb{E} \left| \sum_{j=n+1}^m \psi_j X_{t-j} \right|^2 \right)^{1/2} + \left( \mathbb{E} \left| \sum_{j=-n-1}^{-m} \psi_j X_{t-j} \right|^2 \right)^{1/2}. \end{aligned}$$

Using the Cauchy-Bunyakovskii-Schwarz inequality and the stationarity of  $\{X_t\}$ , the first term on the right hand side is bounded by

$$\begin{aligned} & \left( \mathbb{E} \sum_{j,k=n+1}^m |\psi_j X_{t-j} \psi_k X_{t-k}| \right)^{1/2} \leq \left( \sum_{j,k=n+1}^m |\psi_j| |\psi_k| \mathbb{E}(|X_{t-j}| |X_{t-k}|) \right)^{1/2} \\ & \leq \left( \sum_{j,k=n+1}^m |\psi_j| |\psi_k| (\mathbb{E}|X_{t-j}|^2)^{1/2} (\mathbb{E}|X_{t-k}|^2)^{1/2} \right)^{1/2} \\ & = \gamma_X(0)^{1/2} \sum_{j=n+1}^m |\psi_j|. \end{aligned}$$

As  $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$  by assumption, the last term converges to zero. Thus, the limit of  $\{Y_t^{(m)}\}$ ,  $m \rightarrow \infty$ , denoted by  $S_t$ , exists in the mean square sense with  $\mathbb{E}|S_t|^2 < \infty$ .

It remains to show that  $S_t$  and  $\sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$  are actually equal with probability one. This is established by noting that

$$\begin{aligned} \mathbb{E}|S_t - \Psi(L)X_t|^2 &= \mathbb{E} \liminf_{m \rightarrow \infty} \left| S_t - \sum_{j=-m}^m X_{t-j} \right|^2 \\ &\leq \liminf_{m \rightarrow \infty} \mathbb{E} \left| S_t - \sum_{j=-m}^m X_{t-j} \right|^2 = 0 \end{aligned}$$

where use has been of Fatou's lemma.

The stationarity of  $\{Y_t\}$  can be checked as follows:

$$\begin{aligned} \mathbb{E}Y_t &= \lim_{m \rightarrow \infty} \sum_{j=-m}^m \psi_j \mathbb{E}X_{t-j} = 0, \\ \mathbb{E}Y_t Y_{t-h} &= \lim_{m \rightarrow \infty} \mathbb{E} \left[ \left( \sum_{j=-m}^m \psi_j X_{t-j} \right) \left( \sum_{k=-m}^m \psi_k X_{t-h-k} \right) \right] \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h+k-j). \end{aligned}$$

Thus,  $\mathbb{E}Y_t$  and  $\mathbb{E}Y_t Y_{t-h}$  are finite and independent of  $t$ .  $\{Y_t\}$  is therefore stationary.  $\square$

**Corollary 6.2.** *If  $X_t \sim \text{WN}(0, \sigma^2)$  and  $Y_t = \sum_{j=0}^{\infty} \psi_j X_{t-j}$  with  $\sum_{j=0}^{\infty} |\psi_j| < \infty$  then the above expression for  $\gamma_Y(h)$  simplifies to*

$$\gamma_Y(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

**Remark 6.1.** *In the proof of the existence of  $\{Y_t\}$ , the assumption of the stationarity of  $\{X_t\}$  can be weakened by assuming only  $\sup_t \mathbb{E}|X_t|^2 < \infty$ .*

**Theorem 6.5.** *Under the conditions of Theorem 6.4, the spectral densities of  $\{X_t\}$  and  $\{Y_t\}$  are related as*

$$f_Y(\lambda) = |\Psi(e^{-i\lambda})|^2 f_X(\lambda) = \Psi(e^{i\lambda}) \Psi(e^{-i\lambda}) f_X(\lambda)$$

where  $\Psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$ .  $\Psi(e^{-i\lambda})$  is called the transfer function of the filter.

To understand the effect of the filter  $\Psi$ , consider the simple harmonic process  $X_t = 2 \cos(\lambda t) = e^{i\lambda t} + e^{-i\lambda t}$ . Passing  $\{X_t\}$  through the filter  $\Psi$  leads to a transformed time series  $\{Y_t\}$  defined as

$$Y_t = 2 |\Psi(e^{-i\lambda})| \cos \left( \lambda \left( t - \frac{\theta(\lambda)}{\lambda} \right) \right)$$

where  $\theta(\lambda) = \arg \Psi(e^{-i\lambda})$ . The filter therefore amplifies some frequencies by the factor  $g(\lambda) = |\Psi(e^{-i\lambda})|$  and delays  $X_t$  by  $\frac{\theta(\lambda)}{\lambda}$  periods. Thus, we have a change in amplitude given by the *amplitude gain* function  $g(\lambda)$  and a phase shift given by the *phase gain* function  $\theta(\lambda)$ . If the gain function is bigger than one the corresponding frequency is amplified. On the other hand, if the value is smaller than one the corresponding frequency is dampened.

#### Examples of filters:

- First differences (changes with respect to previous period):

$$\Psi(L) = \Delta = 1 - L.$$

The transfer function of this filter is  $(1 - e^{-i\lambda})$  and the gain function is  $2(1 - \cos \lambda)$ . These functions take the value zero for  $\lambda = 0$ . Thus, the filter eliminates the trend which can be considered as a wave with an infinite periodicity.

- Change with respect to same quarter last year, assuming that the data are quarterly observations:

$$\Psi(L) = 1 - L^4.$$

The transfer function and the gain function are  $1 - e^{-4i\lambda}$  and  $2(1 - \cos(4\lambda))$ , respectively. Thus, the filter eliminates all frequencies which are multiples of  $\frac{\pi}{2}$  including the zero frequency. In particular, it eliminates the trend and waves with periodicity of four quarters.

- A famous example of a filter which led to wrong conclusions is the Kuznets filter (see Sargent, 1987, 273-276). This filter is obtained from two transformations carried out in a row. The first transformation which should eliminate cyclical movements takes centered five year moving averages. The second one take centered non-overlapping first differences. Thus, the filter can be written as:

$$\Psi(L) = \frac{1}{5} \underbrace{(L^{-2} + L^{-1} + 1 + L + L^2)}_{\text{first transformation}} \underbrace{(L^{-5} - L^5)}_{\text{second transformation}} .$$

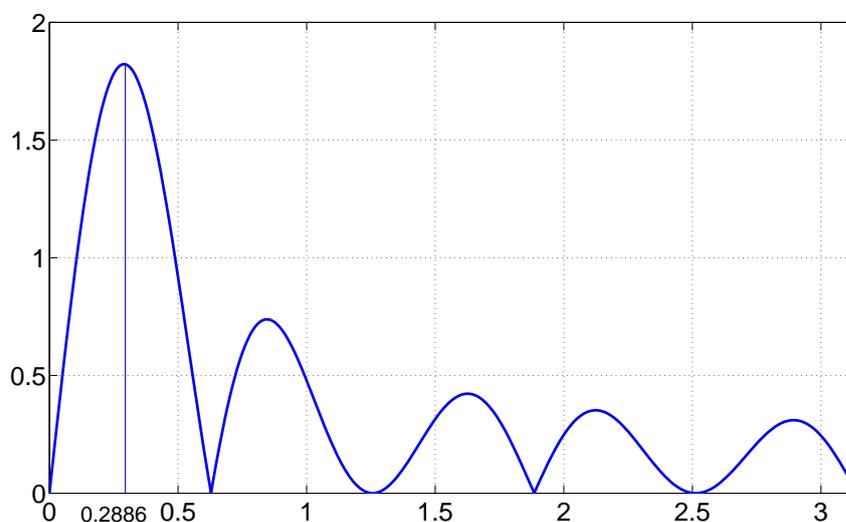


Figure 6.4: Transfer function of the Kuznets filters

Figure 6.4 gives a plot of the transfer function of the Kuznets filter. Thereby it can be seen that all frequencies are dampened, except those around  $\lambda = 0.2886$ . The value  $\lambda = 0.2886$  corresponds to a wave with periodicity of approximately 21 years. Thus, as first claimed by Howrey (1968), even a filtered white noise time series would exhibit a 21 year cycle. This demonstrates that cycles of this length, related by Kuznets (1930) to demographic processes and infrastructure investment swings, may just be an artefact produced by the filter and are therefore not endorsed by the data.

## 6.5 Some Important Filters

### 6.5.1 Construction of Low- and High-Pass Filters

For some purposes it is desirable to eliminate specific frequencies. Suppose, we want to purge a time series from all movements with frequencies above  $\lambda_c$ , but leave those below this value unchanged. The transfer function of such an ideal low-pass filter would be:

$$\Psi(e^{-i\lambda}) = \begin{cases} 1, & \text{for } \lambda \leq \lambda_c; \\ 0, & \text{for } \lambda > \lambda_c. \end{cases}$$

By expanding  $\Psi(e^{-i\lambda})$  into a Fourier-series  $\Psi(e^{-i\lambda}) = \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda}$ , it is possible to determine the appropriate filter coefficients  $\{\psi_j\}$ . In the case of a low-pass filter they are given by:

$$\psi_j = \frac{1}{2\pi} \int_{-\lambda_c}^{\lambda_c} e^{-ij\omega} d\omega = \begin{cases} \frac{\lambda_c}{\pi}, & j = 0; \\ \frac{\sin(j\lambda_c)}{j\pi}, & j \neq 0. \end{cases}$$

The implementation of the filter in practice is not straightforward because only a finite number of coefficients can be used. Depending on the number of observations, the filter must be truncated such that only those  $\psi_j$  with  $|j| \leq q$  are actually employed. The problem becomes more severe as one gets to the more recent observations because less future observations are available. For the most recent period even no future observation is available. This problem is usually overcome by replacing the missing future values by their corresponding forecast. Despite this remedy, the filter works best in the middle of the sample and is more and more distorted as one approaches the beginning or the end of the sample.

Analogously to low-pass filters, it is possible to construct high-pass filters. Figure 6.5 compares the transfer function of an ideal high-pass filter with two filters truncated at  $q = 8$  and  $q = 32$ , respectively. Obviously, the transfer function with the higher  $q$  approximates the ideal filter better. In the neighborhood of the critical frequency, in our case  $\pi/16$ , however, the approximation remains inaccurate. This is known as the Gibbs phenomenon.

### 6.5.2 The Hodrick-Prescott Filter

The Hodrick-Prescott filter (HP-Filter) has gained great popularity in the macroeconomic literature, particularly in the context of the real business cycles theory. This high-pass filter is designed to eliminate the trend and cycles of high periodicity and to emphasize movements at business cycles frequencies (see Hodrick and Prescott, 1980; King and Rebelo, 1993; Brandner and Neusser, 1992).

One way to introduce the HP-filter is to examine the problem of decomposing a time series  $\{X_t\}$  additively into a growth component  $\{G_t\}$  and a cyclical component  $\{C_t\}$ :

$$X_t = G_t + C_t.$$

This decomposition is, without further information, not unique. Following the suggestion of Whittaker (1923), the growth component should be approximated by a smooth curve. Based on this recommendation Hodrick and Prescott suggest to solve the following *restricted least-squares problem* given

a sample  $\{X_t\}_{t=1,\dots,T}$ :

$$\sum_{t=1}^T (X_t - G_t)^2 + \lambda \sum_{t=2}^{T-1} [(G_{t+1} - G_t) - (G_t - G_{t-1})]^2 \longrightarrow \min_{\{G_t\}}.$$

The above objective function has two terms. The first one measures the fit of  $\{G_t\}$  to the data. The closer  $\{G_t\}$  is to  $\{X_t\}$  the smaller this term becomes. In the limit when  $G_t = X_t$  for all  $t$ , the term is minimized and equal to zero. The second term measures the smoothness of the growth component by looking at the discrete analogue to the second derivative. This term is minimized if the changes of the growth component from one period to the next are constant. This, however, implies that  $G_t$  is a linear function. Thus the above objective function represents a trade-off between fitting the data and smoothness of the approximating function. This trade-off is governed by the meta-parameter  $\lambda$  which must be fixed a priori.

The value of  $\lambda$  depends on the critical frequency and on the periodicity of the data (see Uhlig and Ravn, 2002, for the latter). Following the proposal by Hodrick and Prescott (1980) the following values for  $\lambda$  are common in the literature:

$$\lambda = \begin{cases} 6.25, & \text{yearly observations;} \\ 1600, & \text{quarterly observations;} \\ 14400, & \text{monthly observations.} \end{cases}$$

It can be shown that these choices for  $\lambda$  practically eliminate waves of periodicity longer than eight years. The cyclical or business cycle component is therefore composed of waves with periodicity less than eight years. Thus, the choice of  $\lambda$  implicitly defines the business cycle. Figure 6.5 compares the transfer function of the HP-filter to the ideal high-pass filter and two approximate high-pass filters.<sup>6</sup>

As an example, figure 6.6 displays the HP-filtered US logged GDP together with the original series in the upper panel and the implied business cycle component in the lower panel.

### 6.5.3 Seasonal Filters

Besides the elimination of trends, the removal of seasonal movements represents another important application in practice. Seasonal movements typically arise when a time series is observed several times within a year giving rise to the possibility of waves with periodicity less than twelve month in

---

<sup>6</sup>As all filters, the HP-filter systematically distorts the properties of the time series. Harvey and Jaeger (1993) show how the blind application of the HP-filter can lead to the detection of spurious cyclical behavior.

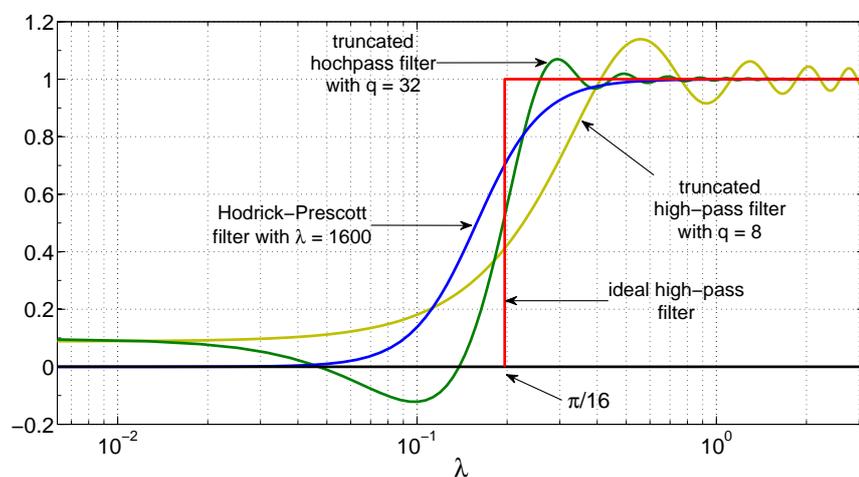


Figure 6.5: Transfer function of HP-filter in comparison to high-pass filters

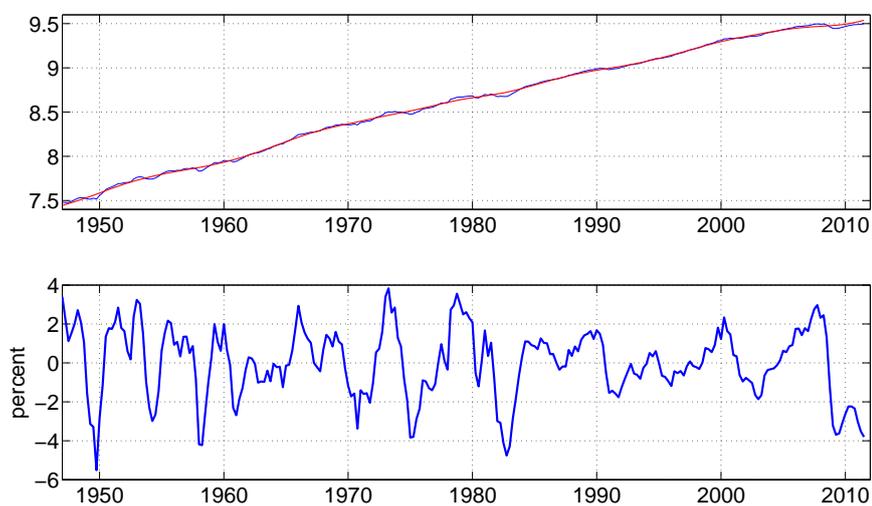


Figure 6.6: HP-filtered US logged GDP (upper panel) and cyclical component (lower panel)

the case of monthly observations, respectively of four quarters in the case of quarterly observations. These cycles are usually considered to be of minor economic interest because they are due to systematic seasonal variations in weather conditions or holidays (Easter, for example).<sup>7</sup> Such variations can, for example, reduce construction activity during the winter season or production in general during holiday time. These cycles have usually quite large amplitude so that they obstruct the view to the economically and politically more important business cycles. In practice, one may therefore prefer to work with seasonally adjusted series. This means that one must remove the seasonal components from the time series in a preliminary stage of the analysis. In section, we will confine ourself to only few remarks. Comprehensive treatment of seasonality can be found in Hylleberg (1986) and Ghysels and Osborn (2001).

Two simple filters for the elimination of seasonal cycles in the case of quarterly data are given by the one-sided filter

$$\Psi(L) = (1 + L + L^2 + L^3)/4$$

or the two-sided filter

$$\Psi(L) = \frac{1}{8}L^2 + \frac{1}{4}L + \frac{1}{4} + \frac{1}{4}L^{-1} + \frac{1}{8}L^{-2}.$$

In practice, the so-called X-11-Filter or its enhanced versions X-12 and X-13 filter developed by the United States Census Bureau is often applied. This filter is a two-sided filter which makes, in contrast to two examples above, use of all sample observations. As this filter not only adjusts for seasonality, but also corrects for outliers, a blind mechanical use is not recommended. Gómez and Maravall (1996) developed an alternative method known under the name TRAMO-SEATS. More details on the implementation of both methods can be found in Eurostat (2009).

Figure 6.7 shows the effect of seasonal adjustment using TRAMO-SEATS by looking at the corresponding transfer functions of the growth rate of construction investment. One can clearly discern how the yearly and the half-yearly waves corresponding to the frequencies  $\pi/2$  und  $\pi$  are dampened. On the other hand, the seasonal filter weakly amplifies a cycle of frequency 0.72 corresponding to a cycle of periodicity of two years.

#### 6.5.4 Using Filtered Data

Whether or not to use filtered, especially seasonally adjusted, data is still an ongoing debate. Although the use of unadjusted data together with a

---

<sup>7</sup>An exception to this view is provided by Miron (1996).

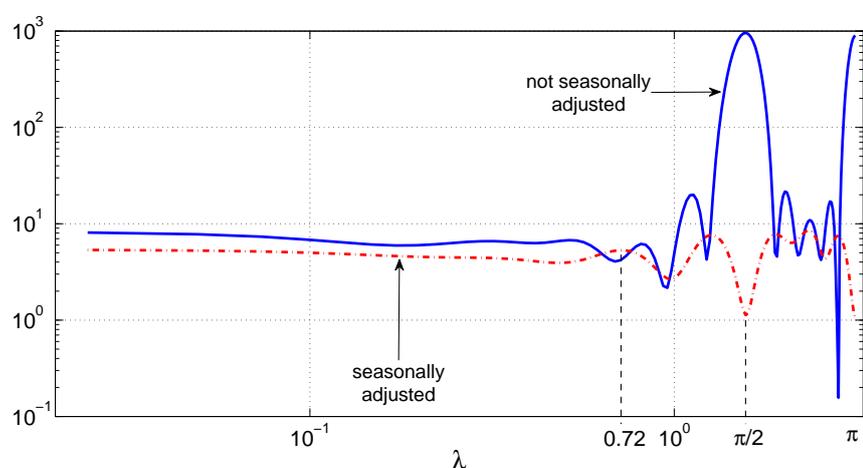


Figure 6.7: Transfer function of growth rate of investment in the construction sector with and without seasonal adjustment

correctly specified model is clearly the best choice, there is a nonnegligible uncertainty in modeling economic time series. Thus, in practice one faces several trade-offs which must be taken into account and which may depend on the particular context (Sims, 1974, 1993; Hansen and Sargent, 1993). On the one hand, the use of adjusted data may disregard important information on the dynamics of the time series and introduce some biases. On the other hand, the use of unadjusted data encounters the risk of misspecification, especially because usual measures of fit may put too large emphasis on fitting the seasonal frequencies thereby neglecting other frequencies.



# Chapter 7

## Integrated Processes

### 7.1 Definition, Properties and Interpretation

Up to now the discussion concentrated on stationary processes and in particular ARMA processes. According to the Wold decomposition theorem (see Theorem 3.1) every purely non-deterministic processes possesses the following representation:

$$X_t = \mu + \Psi(L)Z_t,$$

where  $\{Z_t\} \sim \text{WN}(0, \sigma^2)$  and  $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ . Typically, we model  $X_t$  as an ARMA process so that  $\Psi(L) = \frac{\Theta(L)}{\Phi(L)}$ . This representation implies:

- $\mathbb{E}X_t = \mu,$
- $\lim_{h \rightarrow \infty} \mathbb{P}_t X_{t+h} = \mu.$

The above property is often referred to as mean reverting because the process moves around a constant mean. Deviations from this mean are only temporary or transitory. Thus the best long-run forecast is just the mean of the process.

This property is often violated by economic time series which typically show a tendency to growth. Classic examples are time series for GDP (see Figure 1.3) or some stock market index (see Figure 1.5). This trending property is not compatible with stationarity as the mean is no longer constant. In order to cope with this characteristic of economic time series, two very different alternatives have been proposed. The first one consists in letting the mean  $\mu$  be a function of time  $\mu(t)$ . The most popular specification for  $\mu(t)$  is a linear function, i.e.  $\mu(t) = \alpha + \delta t$ . In this case we get:

$$X_t = \underbrace{\alpha + \delta t}_{\text{linear trend}} + \Psi(L)Z_t$$

The process  $\{X_t\}$  is then referred to as a *trend-stationary process*. In practice one also encounters quadratic polynomials of  $t$  or piecewise linear functions. For example,  $\mu(t) = \alpha_1 + \delta_1 t$  for  $t \leq t_0$  and  $\mu(t) = \alpha_2 + \delta_2 t$  for  $t > t_0$ . In the following, we restrict ourselves to linear trend functions.

The second alternative assumes that the time series becomes stationary after differentiation. The number of times one has to differentiate the process to achieve stationarity is called the order of integration. If  $d$  times differentiation is necessary, the process is called integrated of order  $d$  and is denoted by  $X_t \sim I(d)$ . If the resulting time series,  $\Delta^d X_t = (1 - L)^d X_t$ , is an ARMA(p,q) process, the original process is called an ARIMA(p,d,q) process. Usually it is sufficient to differentiate the time series only once, i.e.  $d = 1$ . For expositional purposes we will stick to this case.

The formal definition of an  $I(1)$  process is given as follows.

**Definition 7.1.** *The stochastic process  $\{X_t\}$  is called integrated of order one or difference-stationary, denoted as  $X_t \sim I(1)$ , if and only if  $\Delta X_t = X_t - X_{t-1}$  can be represented as*

$$\Delta X_t = (1 - L)X_t = \delta + \Psi(L)Z_t, \quad \Psi(1) \neq 0,$$

with  $\{Z_t\} \sim \text{WN}(0, \sigma^2)$  and  $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ .

The qualification  $\Psi(1) \neq 0$  is necessary to avoid trivial and uninteresting cases. Suppose for the moment that  $\Psi(1) = 0$  then it would be possible to write  $\Psi(L)$  as  $(1 - L)\tilde{\Psi}(L)$  for some lag polynomial  $\tilde{\Psi}(L)$ . This would, however, imply that the factor  $1 - L$  could be canceled in the above definition so that  $\{X_t\}$  is already stationary and that the differentiation would be unnecessary. The assumption  $\Psi(1) \neq 0$  thus excludes the case where a trend-stationary process could be regarded as an integrated process. For each trend-stationary process  $X_t = \alpha + \delta t + \tilde{\Psi}(L)Z_t$  we have  $\Delta X_t = \delta + \Psi(L)Z_t$  with  $\Psi(L) = (1 - L)\tilde{\Psi}(L)$ . This would violate the condition  $\Psi(1) \neq 0$ . Thus a trend-stationary process cannot be a difference-stationary process.

The condition  $\sum_{j=0}^{\infty} j|\psi_j| < \infty$  implies  $\sum_{j=0}^{\infty} \psi_j^2 < \infty$  and is therefore stronger than necessary for the Wold representation to hold. In particular, it implies the Beveridge-Nelson decomposition of integrated processes into a linear trend, a random walk, and a stationary component (see Section 7.1.4). The condition is automatically fulfilled for all ARMA processes because  $\{\psi_j\}$  decays exponentially to zero.

Integrated processes with  $d > 0$  are also called unit-root processes. This designation results from the fact that ARIMA processes with  $d > 0$  can be viewed as ARMA processes, whereby the AR polynomial has a  $d$ -fold root of

one.<sup>1</sup> An important prototype of an integrated process is the random walk with drift  $\delta$ :

$$X_t = \delta + X_{t-1} + Z_t, \quad Z_t \sim \text{WN}(0, \sigma^2).$$

Trend-stationary and difference-stationary processes have quite different characteristics. In particular, they imply different behavior with respect to the long-run forecast, the variance of the forecast error, and the impulse response function. In the next section, we will explore these properties in detail.

### 7.1.1 Long-run Forecast

The optimal forecast in the least-squares sense given the infinite past of a trend-stationary process is given by

$$\tilde{\mathbb{P}}_t X_{t+h} = \alpha + \delta(t+h) + \psi_h Z_t + \psi_{h+1} Z_{t-1} + \dots$$

Thus we have

$$\lim_{h \rightarrow \infty} \mathbb{E} \left( \tilde{\mathbb{P}}_t X_{t+h} - \alpha - \delta(t+h) \right)^2 = \sigma^2 \lim_{h \rightarrow \infty} \sum_{j=0}^{\infty} \psi_{h+j}^2 = 0$$

because  $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ . Thus the long-run forecast is given by the linear trend. Even if  $X_t$  deviates temporarily from the trend line, it is assumed to return to it. A trend-stationary process therefore behaves in the long-run like  $\mu(t) = \alpha + \delta t$ .

The forecast of the differentiated series is

$$\tilde{\mathbb{P}}_t \Delta X_{t+h} = \delta + \psi_h Z_t + \psi_{h+1} Z_{t-1} + \psi_{h+2} Z_{t-2} + \dots$$

The level of  $X_{t+h}$  is by definition

$$X_{t+h} = (X_{t+h} - X_{t+h-1}) + (X_{t+h-1} - X_{t+h-2}) + \dots + (X_{t+1} - X_t) + X_t$$

---

<sup>1</sup>Strictly speaking this does not conform to the definitions used in this book because our definition of ARMA processes assumes stationarity.

so that

$$\begin{aligned}
\tilde{\mathbb{P}}_t X_{t+h} &= \tilde{\mathbb{P}}_t \Delta X_{t+h} + \tilde{\mathbb{P}}_t \Delta X_{t+h-1} + \dots + \tilde{\mathbb{P}}_t \Delta X_{t+1} + X_t \\
&= \delta + \psi_h Z_t + \psi_{h+1} Z_{t-1} + \psi_{h+2} Z_{t-2} + \dots \\
&\quad + \delta + \psi_{h-1} Z_t + \psi_h Z_{t-1} + \psi_{h+1} Z_{t-2} + \dots \\
&\quad + \delta + \psi_{h-2} Z_t + \psi_{h-1} Z_{t-1} + \psi_h Z_{t-2} + \dots \\
&\quad + \dots + X_t \\
&= X_t + \delta h \\
&\quad + (\psi_h + \psi_{h-1} + \dots + \psi_1) Z_t \\
&\quad + (\psi_{h+1} + \psi_h + \dots + \psi_2) Z_{t-1} \\
&\quad \dots
\end{aligned}$$

This shows that also for the integrated process the long-run forecast depends on a linear trend with slope  $\delta$ . However, the intercept is no longer a fixed number, but given by  $X_t$  which is stochastic. With each new realization of  $X_t$  the intercept changes so that the trend line moves in parallel up and down. This issue can be well illustrated by the following two examples.

**Example 1:** Let  $\{X_t\}$  be a random walk with drift  $\delta$ . Then best forecast of  $X_{t+h}$ ,  $\mathbb{P}_t X_{t+h}$ , is

$$\mathbb{P}_t X_{t+h} = \delta h + X_t.$$

The forecast thus increases at rate  $\delta$  starting from the initial value of  $X_t$ .  $\delta$  is therefore the slope of a linear trend. The intercept of this trend is stochastic and equal to  $X_t$ . Thus the trend line moves in parallel up or down depending on the realization of  $X_t$ .

**Example 2:** Let  $\{X_t\}$  be an ARIMA(0,1,1) process given by  $\Delta X_t = \delta + Z_t + \theta Z_{t-1}$  with  $|\theta| < 1$ . The best forecast of  $X_{t+h}$  is then given by

$$\mathbb{P}_t X_{t+h} = \delta h + X_t + \theta Z_t.$$

As before the intercept changes in a stochastic way, but in contrary to the previous example it is now given by  $X_t + \theta Z_t$ . If we consider the forecast given the infinite past, the invertibility of the process implies that  $Z_t$  can be expressed as a weighted sum of current and past realizations of  $\Delta X_t$  (see Sections 2.3 and 3.1).

### 7.1.2 Variance of Forecast Error

In the case of a trend-stationary process the forecast error is

$$X_{t+h} - \tilde{\mathbb{P}}_t X_{t+h} = Z_{t+h} + \psi_1 Z_{t+h-1} + \dots + \psi_{h-1} Z_{t+1}.$$

As the mean of the forecast error is zero, the variance is

$$\mathbb{E} \left( X_{t+h} - \tilde{\mathbb{P}}_t X_{t+h} \right)^2 = (1 + \psi_1^2 + \psi_2^2 + \dots + \psi_{h-1}^2) \sigma^2.$$

For  $h$  going to infinity this expression converges to  $\sigma^2 \sum_{j=0}^{\infty} \psi_j^2 < \infty$ . This is nothing but the unconditional variance of  $X_t$ . Thus the variance of the forecast error increases with the length of the forecasting horizon, but remains bounded.

For the integrated process the forecast error can be written as

$$\begin{aligned} X_{t+h} - \tilde{\mathbb{P}}_t X_{t+h} &= Z_{t+h} + (1 + \psi_1) Z_{t+h-1} + \\ &\dots + (1 + \psi_1 + \psi_2 + \dots + \psi_{h-1}) Z_{t+1}. \end{aligned}$$

The forecast error variance therefore is

$$\mathbb{E} \left( X_{t+h} - \tilde{\mathbb{P}}_t X_{t+h} \right)^2 = [1 + (1 + \psi_1)^2 + \dots + (1 + \psi_1 + \dots + \psi_{h-1})^2] \sigma^2.$$

This expression increases with the length of the forecast horizon  $h$ , but is no longer bounded. It increases linearly in  $h$  to infinity.<sup>2</sup> The precision of the forecast therefore not only decreases with the forecasting horizon  $h$  as in the case of the trend-stationary model, but converges to zero. In the example above of the ARIMA(0,1,1) process the forecasting error variance is

$$\mathbb{E} (X_{t+h} - \mathbb{P}_t X_{t+h})^2 = [1 + (h-1)(1+\theta)^2] \sigma^2.$$

This expression clearly increases linearly with  $h$ .

### 7.1.3 Impulse Response Function

The impulse response function (dynamic multiplier) is an important analytical tool as it gives the response of the variable  $X_t$  to the underlying shocks. In the case of the trend-stationary process the impulse response function is

$$\frac{\partial \tilde{\mathbb{P}}_t X_{t+h}}{\partial Z_t} = \psi_h \longrightarrow 0 \quad \text{for } h \rightarrow \infty.$$

---

<sup>2</sup>*Proof:* By assumption  $\{\psi_j\}$  is absolutely summable so that  $\Psi(1)$  converges. Moreover, as  $\Psi(1) \neq 0$ , there exists  $\varepsilon > 0$  and an integer  $m$  such that  $\left| \sum_{j=0}^h \psi_j \right| > \varepsilon$  for all  $h > m$ . The squares are therefore bounded from below by  $\varepsilon^2 > 0$  so that their infinite sum diverges to infinity.

The effect of a shock thus declines with time and dies out. Shocks have therefore only transitory or temporary effects. In the case of an ARMA process the effect even declines exponentially (see the considerations in Section 2.3).<sup>3</sup>

In the case of integrated processes the impulse response function for  $\Delta X_t$  implies :

$$\frac{\partial \tilde{\mathbb{P}}_t X_{t+h}}{\partial Z_t} = 1 + \psi_1 + \psi_2 + \dots + \psi_h.$$

For  $h$  going to infinity, this expression converges  $\sum_{j=0}^{\infty} \psi_j = \Psi(1) \neq 0$ . This implies that a shock experienced in period  $t$  will have a long-run or permanent effect. This long-run effect is called *persistence*. If  $\{\Delta X_t\}$  is an ARMA process then the persistence is given by the expression

$$\Psi(1) = \frac{\Theta(1)}{\Phi(1)}.$$

Thus, for an ARIMA(0,1,1) the persistence is  $\Psi(1) = \Theta(1) = 1 + \theta$ . In the next section we will discuss some examples.

#### 7.1.4 The Beveridge-Nelson Decomposition

The Beveridge-Nelson decomposition represents an important tool for the understanding of integrated processes.<sup>4</sup> It shows how an integrated time series of order one can be represented as the sum of a linear trend, a random walk, and a stationary series. It may therefore be used to extract the cyclical component (business cycle component) of a time series and can thus be viewed as an alternative to the HP-filter (see section 6.5.2) or to more elaborated so-called structural time series models (see sections 17.1 and 17.4.2).

Assuming that  $\{X_t\}$  is an integrated process of order one, there exists, according to Definition 7.1, a causal representation for  $\{\Delta X_t\}$ :

$$\Delta X_t = \delta + \Psi(L)Z_t \quad \text{with } Z_t \sim \text{WN}(0, \sigma^2)$$

with the property  $\Psi(1) \neq 0$  and  $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ . Before proceeding to the main theorem, we notice the following simple, but extremely useful poly-

<sup>3</sup>The use of the partial derivative is just for convenience. It does not mean that  $X_{t+h}$  is differentiated in the literal sense.

<sup>4</sup>Neusser (2000) shows how a Beveridge-Nelson decomposition can also be derived for higher order integrated processes.

mial decomposition of  $\Psi(L)$ :

$$\begin{aligned}
\Psi(L) - \Psi(1) &= 1 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \psi_4 L^4 + \dots \\
&\quad - 1 - \psi_1 - \psi_2 - \psi_3 - \psi_4 - \dots \\
&= \psi_1(L - 1) + \psi_2(L^2 - 1) + \psi_3(L^3 - 1) + \psi_4(L^4 - 1) + \dots \\
&= (L - 1) [\psi_1 + \psi_2(L + 1) + \psi_3(L^2 + L + 1) + \dots] \\
&= (L - 1) [(\psi_1 + \psi_2 + \psi_3 + \dots) + (\psi_2 + \psi_3 + \psi_4 + \dots)L \\
&\quad + (\psi_3 + \psi_4 + \psi_5 + \dots)L^2 + \dots].
\end{aligned}$$

We state this results in the following Lemma:

**Lemma 7.1.** *Let  $\Psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$ , then*

$$\Psi(L) = \Psi(1) + (L - 1)\tilde{\Psi}(L)$$

where  $\tilde{\Psi}(L) = \sum_{j=0}^{\infty} \tilde{\psi}_j L^j$  with  $\tilde{\psi}_j = \sum_{i=j+1}^{\infty} \psi_i$ .

As  $\{X_t\}$  is integrated and because  $\Psi(1) \neq 0$  we can express  $X_t$  as follows:

$$\begin{aligned}
X_t &= X_0 + \sum_{j=1}^t \Delta X_j \\
&= X_0 + \sum_{j=1}^t \left\{ \delta + [\Psi(1) + (L - 1)\tilde{\Psi}(L)] Z_j \right\} \\
&= X_0 + \delta t + \Psi(1) \sum_{j=1}^t Z_j + \sum_{j=1}^t (L - 1)\tilde{\Psi}(L) Z_j \\
&= \underbrace{X_0 + \delta t}_{\text{linear trend}} + \underbrace{\Psi(1) \sum_{j=1}^t Z_j}_{\text{random walk}} + \underbrace{\tilde{\Psi}(L)Z_0 - \tilde{\Psi}(L)Z_t}_{\text{stationary component}}.
\end{aligned}$$

This leads to the following theorem.

**Theorem 7.1.** *[Beveridge-Nelson decomposition] Every integrated process  $\{X_t\}$  has a decomposition of the following form:*

$$X_t = \underbrace{X_0 + \delta t}_{\text{linear trend}} + \underbrace{\Psi(1) \sum_{j=1}^t Z_j}_{\text{random walk}} + \underbrace{\tilde{\Psi}(L)Z_0 - \tilde{\Psi}(L)Z_t}_{\text{stationary component}}.$$

The above representation is referred to as the Beveridge-Nelson decomposition.

*Proof.* The only substantial issue is to show that  $\tilde{\Psi}(L)Z_0 - \tilde{\Psi}(L)Z_t$  defines a stationary process. According to Theorem 6.4 it is sufficient to show that the coefficients of  $\tilde{\Psi}(L)$  are absolutely summable. We have that:

$$\sum_{j=0}^{\infty} |\tilde{\psi}_j| = \sum_{j=0}^{\infty} \left| \sum_{i=j+1}^{\infty} \psi_i \right| \leq \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} |\psi_i| = \sum_{j=1}^{\infty} j|\psi_j| < \infty,$$

where the first inequality is a consequence of the triangular inequality and the second inequality follows from the Definition 7.1 of an integrated process.  $\square$

Shocks of a random walk component have a permanent effect. This effect is measured by the persistence  $\Psi(1)$ , the coefficient of the random walk component. In macroeconomics aggregate supply shocks are ascribed to have a long-run effect as they affect productivity. In contrast monetary or demand shocks are viewed to have temporary effects only. Thus the persistence  $\Psi(1)$  can be interpreted as a measure for the importance of supply shocks (see Campbell and Mankiw (1987), Cochrane (1988) or Christiano and Eichenbaum (1990)). For a critical view from an econometric standpoint see Hauser et al. (1999). A more sophisticated multivariate approach to identify supply and demand shocks and to disentangle their relative importance is provided in Section 15.5.

In business cycle analysis it is often useful to decompose  $\{X_t\}$  into a sum of a trend component  $\mu_t$  and a cyclical component  $\varepsilon_t$ :

$$X_t = \mu_t + \varepsilon_t.$$

In the case of a difference-stationary series, the cyclical component can be identified with the stationary component in the Beveridge-Nelson decomposition and the trend component with the random walk plus the linear trend. Suppose that  $\{\Delta X_t\}$  follows an ARMA process  $\Phi(L)\Delta X_t = c + \Theta(L)Z_t$  then  $\Delta\mu_t = \delta + \Psi(1)Z_t$  can be identified as the trend component. This means that the trend component can be recursively determined from the observations by applying the formula

$$\mu_t = \frac{\Phi(L)}{\Theta(L)}\Psi(1)X_t.$$

The cyclical component is then simply the residual:  $\varepsilon_t = X_t - \mu_t$ .

In the above decomposition both the permanent (trend) component as well as the stationary (cyclical) component are driven by the same shock  $Z_t$ . A more sophisticated model would, however, allow that the two components are driven by different shocks. This idea is exploited in the so-called structural time series analysis where the different components (trend, cycle,

season, and irregular) are modeled as being driven by separated shocks. As only the series  $\{X_t\}$  is observed, not its components, this approach leads to serious identification problems. See the discussion in Harvey (1989), Hannan and Deistler (1988), or Mills (2003). In Sections 17.1 and 17.4.2 we will provide an overall framework to deal with these issues.

### Examples

Let  $\{\Delta X_t\}$  be a MA(q) process with  $\Delta X_t = \delta + Z_t + \dots + \theta_q Z_{t-q}$  then the persistence is given simply by the sum of the MA-coefficients:  $\Psi(1) = 1 + \theta_1 + \dots + \theta_q$ . Depending on the value of these coefficients. The persistence can be smaller or greater than one.

If  $\{\Delta X_t\}$  is an AR(1) process with  $\Delta X_t = \delta + \phi \Delta X_{t-1} + Z_t$  and assuming  $|\phi| < 1$  then we get:  $\Delta X_t = \frac{\delta}{1-\phi} + \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ . The persistence is then given as  $\Psi(1) = \sum_{j=0}^{\infty} \phi^j = \frac{1}{1-\phi}$ . For positive values of  $\phi$ , the persistence is greater than one. Thus, a shock of one is amplified to have an effect larger than one in the long-run.

If  $\{\Delta X_t\}$  is assumed to be an ARMA(1,1) process with  $\Delta X_t = \delta + \phi \Delta X_{t-1} + Z_t + \theta Z_{t-1}$  and  $|\phi| < 1$  then  $\Delta X_t = \frac{\delta}{1-\phi} + Z_t + (\phi + \theta) \sum_{j=0}^{\infty} \phi^j Z_{t-j-1}$ . The persistence is therefore given by  $\Psi(1) = 1 + (\phi + \theta) \sum_{j=0}^{\infty} \phi^j = \frac{1+\theta}{1-\phi}$ .

The computation of the persistence for the model estimated for Swiss GDP in Section 5.6 is more complicated because a fourth order difference  $1 - L^4$  has been used instead of a first order one. As  $1 - L^4 = (1 - L)(1 + L + L^2 + L^3)$ , it is possible to extend the above computations also to this case. For this purpose we compute the persistence for  $(1 + L + L^2 + L^3) \ln \text{BIP}_t$  in the usual way. The long-run effect on  $\ln \text{BIP}_t$  is therefore given by  $\Psi(1)/4$  because  $(1 + L + L^2 + L^3) \ln \text{BIP}_t$  is nothing but four times the moving-average of the last four values. For the AR(2) model we get a persistence of 1.42 whereas for the ARMA(1,3) model the persistence is 1.34. Both values are definitely above one so that the permanent effect of a one-percent shock to Swiss GDP is amplified to be larger than one in the long-run. Campbell and Mankiw (1987) and Cochrane (1988) report similar values for the US.

## 7.2 Properties of the OLS estimator in the case of integrated variables

The estimation and testing of coefficients of models involving integrated variables is not without complications and traps because the usual asymptotic theory may become invalid. The reason being that the asymptotic distributions are in general no longer normal so that the usual critical values for the

test statistics are no longer valid. A general treatment of these issues is beyond this text, but can be found in Banerjee et al. (1993) and Stock (1994). We may, however, illustrate the kind of problems encountered by looking at the Gaussian AR(1) case:<sup>5</sup>

$$X_t = \phi X_{t-1} + Z_t, \quad t = 1, 2, \dots,$$

where  $Z_t \sim \text{IID } N(0, \sigma^2)$  and  $X_0 = 0$ . For observations on  $X_1, X_2, \dots, X_T$  the OLS-estimator of  $\phi$  is given by the usual expression:

$$\hat{\phi}_T = \frac{\sum_{t=1}^T X_{t-1} X_t}{\sum_{t=1}^T X_{t-1}^2} = \phi + \frac{\sum_{t=1}^T X_{t-1} Z_t}{\sum_{t=1}^T X_{t-1}^2}.$$

For  $|\phi| < 1$ , the OLS estimator of  $\phi$ ,  $\hat{\phi}_T$ , converges in distribution to a normal random variable (see Chapter 5 and in particular Section 5.2):

$$\sqrt{T} \left( \hat{\phi}_T - \phi \right) \xrightarrow{d} N(0, 1 - \phi^2).$$

The estimated density of the OLS estimator of  $\phi$  for different values of  $\phi$  is represented in Figure 7.1. This figure was constructed using a Monte-Carlo simulation of the above model for a sample size of  $T = 100$  using 10,000 replications for each value of  $\phi$ .<sup>6</sup> The figure shows that the distribution of  $\hat{\phi}_T$  becomes more and more concentrated if the true value of  $\phi$  gets closer and closer to one. Moreover the distribution gets also more and more skewed to the left. This implies that the OLS estimator is downward biased and that this bias gets relatively more and more pronounced in small samples as  $\phi$  approaches one.

The asymptotic distribution would be degenerated for  $\phi = 1$  because the variance approaches zero as  $\phi$  goes to one. Thus the asymptotic distribution becomes useless for statistical inferences under this circumstance. In order to obtain a non-degenerate distribution the estimator must be scaled by  $T$  instead by  $\sqrt{T}$ . It can be shown that

$$T \left( \hat{\phi}_T - \phi \right) \xrightarrow{d} \nu.$$

This result was first established by Dickey and Fuller (1976) and Dickey and Fuller (1981). However, the asymptotic distribution  $\nu$  need no longer be normal. It was first tabulated in Fuller (1976). The scaling with  $T$  instead

<sup>5</sup>We will treat more general cases in Section 7.5 and Chapter 16.

<sup>6</sup>The densities were estimated using an adaptive kernel density estimator with Epanechnikov window (see Silverman (1986)).

of  $\sqrt{T}$  means that the OLS-estimator converges, if the true value of  $\phi$  equals one, at a higher rate to  $\phi = 1$ . This property is known as *superconsistency*.

In order to understand this result better, in particular in the light of the derivation in the Appendix of Section 5.2, we take a closer look at the asymptotic distribution of  $T(\hat{\phi}_T - \phi)$ :

$$T(\hat{\phi}_T - \phi) = \frac{\frac{1}{\sigma^2 T} \sum_{t=1}^T X_{t-1} Z_t}{\frac{1}{\sigma^2 T^2} \sum_{t=1}^T X_{t-1}^2}.$$

Under the assumption  $\phi = 1$ ,  $X_t$  becomes a random walk so that  $X_t$  can be written as  $X_t = Z_t + \dots + Z_1$ . Moreover, as a sum of normally distributed random variables  $X_t$  becomes itself normally distributed as  $X_t \sim N(0, \sigma^2 t)$ . In addition, we get

$$\begin{aligned} X_t^2 &= (X_{t-1} + Z_t)^2 = X_{t-1}^2 + 2X_{t-1}Z_t + Z_t^2 \\ &\Rightarrow X_{t-1}Z_t = (X_t^2 - X_{t-1}^2 - Z_t^2) / 2 \\ &\Rightarrow \sum_{t=1}^T X_{t-1}Z_t = \frac{X_T^2 - X_0^2}{2} - \frac{\sum_{t=1}^T Z_t^2}{2} \\ &\Rightarrow \frac{1}{T} \sum_{t=1}^T X_{t-1}Z_t = \frac{1}{2} \left[ \frac{X_T^2}{T} - \frac{\sum_{t=1}^T Z_t^2}{T} \right] \\ &\Rightarrow \frac{1}{\sigma^2 T} \sum_{t=1}^T X_{t-1}Z_t = \frac{1}{2} \left( \frac{X_T}{\sigma\sqrt{T}} \right)^2 - \frac{1}{2\sigma^2} \frac{\sum_{t=1}^T Z_t^2}{T} \xrightarrow{d} \frac{1}{2} (\chi_1^2 - 1). \end{aligned}$$

The numerator therefore converges to a  $\chi_1^2$  distribution. The distribution of the denominator is more involved, but its expected value is given by:

$$\mathbb{E} \sum_{t=1}^T X_{t-1}^2 = \sigma^2 \sum_{t=1}^T (t-1) = \frac{\sigma^2 T(T-1)}{2},$$

because  $X_{t-1} \sim N(0, \sigma^2(t-1))$ . To obtain a nondegenerate random variable one must scale by  $T^2$ . Thus, intuitively,  $T(\hat{\phi}_T - \phi)$  will no longer converge to a degenerate distribution.

Using similar arguments it can be shown that the t-statistic

$$t_T = \frac{\hat{\phi}_T - 1}{\hat{\sigma}_{\hat{\phi}}} = \frac{\hat{\phi}_T - 1}{\sqrt{\frac{s_T^2}{\sum_{t=1}^T X_{t-1}^2}}}$$

with  $s_T^2 = \frac{1}{T-2} \sum_{t=2}^T (X_t - \hat{\phi}_T X_{t-1})^2$  is not asymptotically normal. Its distribution was also first tabulated by Fuller (1976). Figure 7.2 compares its

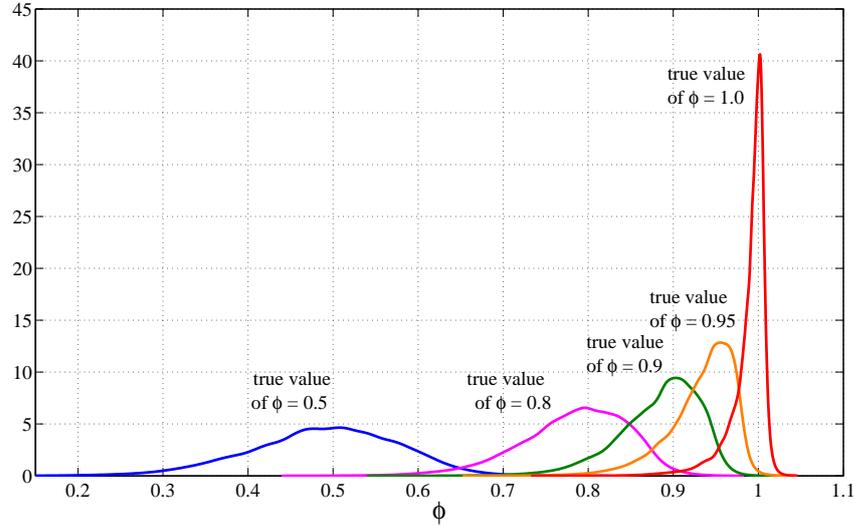


Figure 7.1: Distribution of the OLS estimator of  $\phi$  for  $T = 100$  and 10000 replications

density with the standard normal distribution in a Monte-Carlo experiment using again a sample of  $T = 100$  and 10,000 replications. It is obvious that the t-distribution is shifted to the left. This implies that the critical values will be absolutely higher than for the standard case. In addition, one may observe a slight skewness.

Finally, we also want to investigate the autocovariance function of a random walk. Using similar arguments as in Section 1.3 we get:

$$\begin{aligned} \gamma(h) &= \mathbb{E}(X_T X_{T-h}) \\ &= \mathbb{E}[(Z_T + Z_{T-1} + \dots + Z_1)(Z_{T-h} + Z_{T-h-1} + \dots + Z_1)] \\ &= \mathbb{E}(Z_{T-h}^2 + Z_{T-h-1}^2 + \dots + Z_1^2) = (T-h)\sigma^2. \end{aligned}$$

Thus the correlation coefficient between  $X_T$  and  $X_{T-h}$  is:

$$\rho(h) = \frac{\gamma(h)}{\sqrt{\mathbb{V}X_T} \sqrt{\mathbb{V}X_{T-h}}} = \frac{T-h}{\sqrt{T(T-h)}} = \sqrt{\frac{T-h}{T}}, \quad h \leq T.$$

The autocorrelation coefficient  $\rho(h)$  therefore monotonically decreases with  $h$ , holding the sample size  $T$  constant. The rate at which  $\rho(h)$  falls is, however, smaller than for ARMA processes for which  $\rho(h)$  declines exponentially fast to zero. Given  $h$ , the autocorrelation coefficient converges to one for  $T \rightarrow \infty$ . Figure 7.3 compares the theoretical and the estimated ACF of a

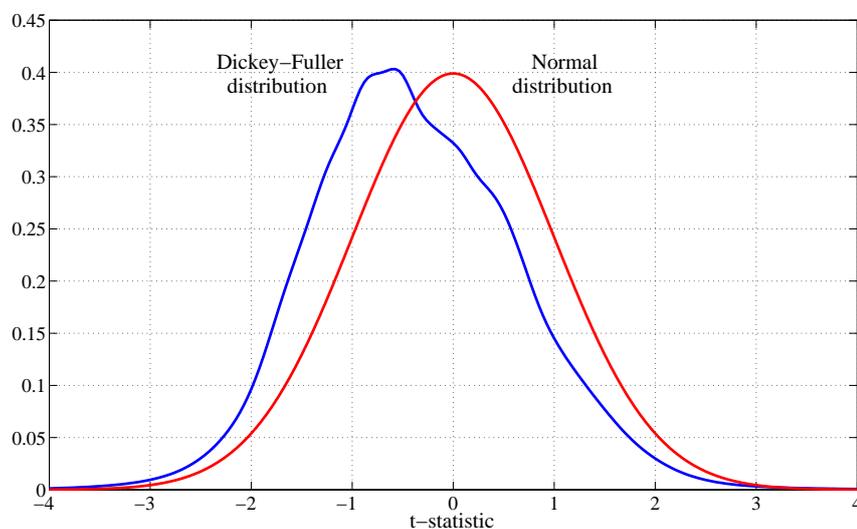


Figure 7.2: Distribution of t-statistic for  $T = 100$  and 10'000 replications and standard normal distribution

simulated random walk with  $T = 100$ . Typically, the estimated coefficients lie below the theoretical ones. In addition, we show the estimated ACF for an AR(1) process with  $\phi = 0.9$  and using the same realizations of the white noise process as in the construction of the random walk. Despite the large differences between the ACF of an AR(1) process and a random walk, the ACF is only of limited use to discriminate between an (stationary) ARMA process and a random walk.

The above calculation also shows that  $\rho(1) < 1$  so that the expected value of the OLS estimator is downward biased in finite samples:  $\mathbb{E}\hat{\phi}_T < 1$ .

### 7.3 Unit-Root Tests

The previous Sections 7.1 and 7.2 have shown that, depending on the nature of the non-stationarity (trend versus difference stationarity), the stochastic process has quite different algebraic (forecast, forecast error variance, persistence) and statistical (asymptotic distribution of OLS-estimator) properties. It is therefore important to be able to discriminate among these two different types of processes. This also pertains to standard regression models for which the presence of integrated variables can lead to non-normal asymptotic distributions.

The ability to differentiate between trend- and difference-stationary pro-

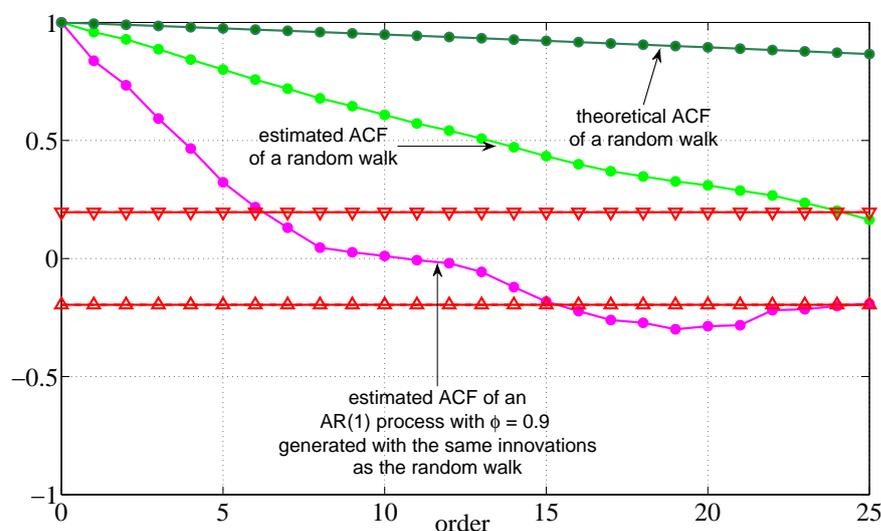


Figure 7.3: ACF of a random walk with 100 observations

cesses is not only important from a statistical point of view, but can be given an economic interpretation. In macroeconomic theory, monetary and demand disturbances are alleged to have only temporary effects whereas supply disturbances, in particular technology shocks, are supposed to have permanent effects. To put it in the language of time series analysis: monetary and demand shocks have a persistence of zero whereas supply shocks have nonzero (positive) persistence. Nelson and Plosser (1982) were the first to investigate the trend properties of economic time series from this angle. In their influential study they reached the conclusion that, with the important exception of the unemployment rate, most economic time series in the US are better characterized as being difference stationary. Although this conclusion came under severe scrutiny (see Cochrane (1988) and Campbell and Perron (1991)), this issue resurfaces in many economic debates. The latest discussion relates to the nature and effect of technology shocks (see Galí (1999) or Christiano et al. (2003)).

The following exposition focuses on the Dickey-Fuller test (DF-test) and the Phillips-Perron test (PP-test). Although other test procedures and variants thereof have been developed in the meantime, these two remain the most widely applied in practice. These types of tests are also called unit-root tests.

Both the DF- as well as the PP-test rely on a regression of  $X_t$  on  $X_{t-1}$  which may include further deterministic regressors like a constant or a linear

time trend. We call this regression the *Dickey-Fuller regression*:

$$X_t = \begin{array}{c} \text{deterministic} \\ \text{variables} \end{array} + \phi X_{t-1} + Z_t. \quad (7.1)$$

Alternatively and numerically equivalent, one may run the Dickey-Fuller regression in difference form:

$$\Delta X_t = \begin{array}{c} \text{deterministic} \\ \text{variables} \end{array} + \beta X_{t-1} + Z_t$$

with  $\beta = \phi - 1$ . For both tests, the null hypothesis is that the process is integrated of order one, difference stationary, or has a unit-root. Thus we have

$$\mathbf{H}_0 : \phi = 1 \quad \text{or} \quad \beta = 0.$$

The alternative hypothesis  $\mathbf{H}_1$  is that the process is trend-stationary or stationary with constant mean and is given by:

$$\mathbf{H}_1 : -1 < \phi < 1 \quad \text{or} \quad -2 < \beta = \phi - 1 < 0.$$

Thus the unit root test is a *one-sided test*. The advantage of the second formulation of the Dickey-Fuller regression is that the corresponding t-statistic can be readily read off from standard outputs of many computer packages which makes additional computations unnecessary.

### 7.3.1 The Dickey-Fuller Test (DF-Test)

The Dickey-Fuller test comes in two forms. The first one, sometimes called the  $\rho$ -test, takes  $T(\hat{\phi} - 1)$  as the test statistic. As shown previously, this statistic is no longer asymptotically normally distributed. However, it was first tabulated by Fuller and can be found in textbooks like Fuller (1976) or Hamilton (1994a). The second and much more common one relies on the usual t-statistic for the hypothesis  $\phi = 1$ :

$$t_{\hat{\phi}} = (\hat{\phi}_T - 1) / \hat{\sigma}_{\hat{\phi}}.$$

This test-statistic is also not asymptotically normally distributed. It was for the first time tabulated by Fuller (1976) and can be found, for example, in Hamilton (1994a). Later MacKinnon (1991) presented much more detailed tables where the critical values can be approximated for any sample size  $T$  by using interpolation formulas (see also Banerjee et al. (1993)).<sup>7</sup>

---

<sup>7</sup>These interpolation formula are now implemented in many software packages, like EViews, to compute the appropriate critical values.

Table 7.1: The four most important cases for the unit-root test

data generating process (null hypothesis)	estimated regression (Dickey-Fuller regression)	$\rho$ -test: $T(\hat{\phi} - 1)$	t-test
$X_t = X_{t-1} + Z_t$	$X_t = \phi X_{t-1} + Z_t$	case 1	case 1
$X_t = X_{t-1} + Z_t$	$X_t = \alpha + \phi X_{t-1} + Z_t$	case 2	case 2
$X_t = \alpha + X_{t-1} + Z_t, \alpha \neq 0$	$X_t = \alpha + \phi X_{t-1} + Z_t$		N(0,1)
$X_t = \alpha + X_{t-1} + Z_t$	$X_t = \alpha + \delta t + \phi X_{t-1} + Z_t$	case 4	case 4

The application of the Dickey-Fuller test as well as the Phillips-Perron test is obfuscated by the fact that the asymptotic distribution of the test statistic ( $\rho$ - or t-test) depends on the specification of the deterministic components and on the true data generating process. This implies that depending on whether the Dickey-Fuller regression includes, for example, a constant and/or a time trend and on the nature of the true data generating process one has to use different tables and thus critical values. In the following we will focus on the most common cases listed in table 7.1.

In case 1 the Dickey-Fuller regression includes no deterministic component. Thus, a rejection of the null hypothesis implies that  $\{X_t\}$  has to be a mean zero stationary process. This specification is, therefore, only warranted if one can make sure that the data have indeed mean zero. As this is rarely the case, except, for example, when the data are the residuals from a previous regression,<sup>8</sup> case 1 is very uncommon in practice. Thus, if the data do not display a trend, which can be checked by a simple time plot, the Dickey-Fuller regression should include a constant. A rejection of the null hypothesis then implies that  $\{X_t\}$  is a stationary process with mean  $\mu = \frac{c}{1-\phi}$ . If the data display a time trend, the Dickey-Fuller regression should also include a linear time trend as in case 4. A rejection of the null hypothesis then implies that the process is trend-stationary. In the case that the Dickey-Fuller regression contains no time trend and there is no time trend under the alternative hypothesis, asymptotic normality holds. This case is only of theoretical interest as it should a priori be clear whether the data are trending or not. In the instance where one is not confident about the trending nature of the time series see the procedure outlined in Section 7.3.3.

<sup>8</sup>This fact may pose a problem by itself.

In the cases 2 and 4 it is of interest to investigate the joint hypothesis  $\mathbf{H}_0 : \alpha = 0$  and  $\phi = 1$ , and  $\mathbf{H}_0 : \delta = 0$  and  $\phi = 1$  respectively. Again the corresponding F-statistic is no longer F-distributed, but has been tabulated (see Hamilton (1994a, Table B7)). The trade-off between t- and F-test is discussed in Section 7.3.3.

Most economic time series display a significant amount of autocorrelation. To take this feature into account it is necessary to include lagged differences  $\Delta X_{t-1}, \dots, \Delta X_{t-p+1}$  as additional regressors. The so modified Dickey-Fuller regression then becomes:

$$X_t = \begin{array}{l} \text{deterministic} \\ \text{variables} \end{array} + \phi X_{t-1} + \gamma_1 \Delta X_{t-1} + \dots + \gamma_{p-1} \Delta X_{t-p+1} + Z_t.$$

This modified test is called the augmented Dickey-Fuller test (ADF-test). This autoregressive correction does not change the asymptotic distribution of the test statistics. Thus the same tables can be used as before. For the coefficients of the autoregressive terms asymptotic normality holds. This implies that the standard testing procedures (t-test, F-test) can be applied in the usual way. This is true if instead of autoregressive correction terms moving-average terms are used instead (see Said and Dickey (1984)).

For the ADF-test the order  $p$  of the model should be chosen such that the residuals are close to being white noise. This can be checked, for example, by looking at the ACF of the residuals or by carrying out a Ljung-Box test (see Section 4.2). In case of doubt, it is better to choose a higher order. A consistent procedure to find the right order is to use the Akaike's criterion (AIC). Another alternative strategy advocated by Ng and Perron (1995) is an iterative testing procedure which makes use of the asymptotic normality of the autoregressive correction terms. Starting from a maximal order  $p-1 = p_{max}$ , the method amounts to the test whether the coefficient corresponding to the highest order is significantly different from zero. If the null hypothesis that the coefficient is zero is not rejected, the order of the model is reduced by one and the test is repeated. This is done as long as the null hypothesis is not rejected. If the null hypothesis is finally rejected, one sticks with the model and performs the ADF-test. The successive test are standard t-tests. It is advisable to use a rather high significance level, for example a 10 percent level. The simulation results by Ng and Perron (1995) show that this procedure leads to a smaller bias compared to using the AIC criterion and that the reduction in power remains negligible.

### 7.3.2 The Phillips-Perron Test (PP-test)

The Phillips-Perron test represents a valid alternative to the ADF-test. It is based on the simple Dickey-Fuller regression (without autoregressive correction terms) and corrects for autocorrelation by modifying the OLS-estimate or the corresponding value of the t-statistic. The simple Dickey-Fuller regression with either constant and/or trend is:

$$X_t = \begin{array}{c} \text{deterministic} \\ \text{variables} \end{array} + \phi X_{t-1} + Z_t,$$

where  $\{Z_t\}$  need no longer be a white noise process, but can be any mean zero stationary process.  $\{Z_t\}$  may, for example, be an ARMA process. In principle, the approach also allows for heteroskedasticity.<sup>9</sup>

The first step in the Phillips-Perron unit-root test estimates the above appropriately specified Dickey-Fuller regression. The second step consists in the estimation of the unconditional variance  $\gamma_Z(0)$  and the long-run variance  $J$  of the residuals  $\hat{Z}_t$ . This can be done using one of the methods prescribed in Section 4.4. These two estimates are then used in a third step to correct the  $\rho$ - and the t-test statistics. This correction would then take care of the autocorrelation present in the data. Finally, one can use the so modified test statistics to carry out the unit-root test applying the same tables for the critical values as before.

In case 1 where no deterministic components are taken into account (see case 1 in Table 7.1) the modified test statistics according to Phillips (1987) are:

$$\begin{aligned} \rho\text{-Test :} & \quad T \left( \hat{\phi} - 1 \right) - \frac{1}{2} \left( \hat{J}_T - \hat{\gamma}_Z(0) \right) \left( \frac{1}{T^2} \sum_{t=1}^T X_{t-1}^2 \right)^{-1} \\ \text{t-Test :} & \quad \sqrt{\frac{\hat{\gamma}_Z(0)}{\hat{J}_T}} t_{\hat{\phi}} - \frac{1}{2} \left( \hat{J}_T - \hat{\gamma}_Z(0) \right) \left( \frac{\hat{J}_T}{T^2} \sum_{t=1}^T X_{t-1}^2 \right)^{-1/2}. \end{aligned}$$

If  $\{Z_t\}$  would be white noise so that  $J = \gamma(0)$ , respectively  $\hat{J}_T \approx \hat{\gamma}_Z(0)$  one gets the ordinary Dickey-Fuller test statistic. Similar formulas can be derived for the cases 2 and 4. As already mentioned these modifications will not alter the asymptotic distributions so the same critical values as for the ADF-test can be used.

The main advantage of the Phillips-Perron test is that the non-parametric correction allows for very general  $\{Z_t\}$  processes. The PP-test is particularly

<sup>9</sup>The exact assumptions can be read in Phillips (1987) and Phillips and Perron (1988).

appropriate if  $\{Z_t\}$  has some MA-components which can be only poorly approximated by low order autoregressive terms. Another advantage is that one can avoid the exact modeling of the process. It has been shown by Monte-Carlo studies that the PP-test has more power compared to the DF-test, i.e. the PP-test rejects the null hypothesis more often when it is false, but that, on the other hand, it has also a higher size distortion, i.e. that it rejects the null hypothesis too often.

### 7.3.3 Unit-Root Test: Testing Strategy

Independently whether the Dickey-Fuller or the Phillips-Perron test is used, the specification of the deterministic component is important and can pose a problem in practice. On the one hand, if the deterministic part is underrepresented, for example when only a constant, but no time trend is used, the test results are biased in favor of the null hypothesis, if the data do indeed have a trend. On the other hand, if too many deterministic components are used, the power of the test is reduced. It is therefore advisable to examine a plot of the series in order to check whether a long run trend is visible or not. In some circumstances economic reasoning may help in this regard.

Sometimes, however, it is difficult to make an appropriate choice a priori. We therefore propose the following testing strategy based on Elder and Kennedy (2001).

**$X_t$  has a long-run trend:** As  $X_t$  grows in the long-run, the Dickey-Fuller regression

$$X_t = \alpha + \delta t + \phi X_{t-1} + Z_t$$

should contain a linear trend.<sup>10</sup> In this case either  $\phi = 1$ ,  $\delta = 0$  and  $\alpha \neq 0$  (unit root case) or  $\phi < 1$  with  $\delta \neq 0$  (trend stationary case). We can then test the joint null hypothesis

$$H_0 : \quad \phi = 1 \text{ and } \delta = 0$$

by a corresponding F-test. Note that the F-statistic, like the t-test, is not distributed according to the F-distribution. If the test does not reject the null, we conclude that  $\{X_t\}$  is a unit root process with drift or equivalently a difference-stationary (integrated) process. If the F-test rejects the null hypothesis, there are three possible situations:

- (i) The possibility  $\phi < 1$  and  $\delta = 0$  contradicts the primary observation that  $\{X_t\}$  has a trend and can therefore be eliminated.

---

<sup>10</sup>In case of the ADF-test additional regressors,  $\Delta X_{t-j}$ ,  $j > 0$ , might be necessary.

- (ii) The possibility  $\phi = 1$  and  $\delta \neq 0$  can also be excluded because this would imply that  $\{X_t\}$  has a quadratic trend, which is unrealistic.
- (iii) The possibility  $\phi < 1$  and  $\delta \neq 0$  represents the only valid alternative. It implies that  $\{X_t\}$  is stationary around a linear trend, i.e. that  $\{X_t\}$  is trend-stationary.

Similar conclusions can be reached if, instead of the F-test, a t-test is used to test the null hypothesis  $H_0 : \phi = 1$  against the alternative  $H_1 : \phi < 1$ . Thereby a non-rejection of  $H_0$  is interpreted that  $\delta = 0$ . If, however, the null hypothesis  $H_0$  is rejected, this implies that  $\delta \neq 0$ , because  $\{X_t\}$  exhibits a long-run trend. The F-test is more powerful than the t-test. The t-test, however, is a one-sided test, which has the advantage that it actually corresponds to the primary objective of the test. In Monte-Carlo simulations the t-test has proven to be marginally superior to the F-test.

**$X_t$  has no long-run trend:** In this case  $\delta = 0$  and the Dickey-Fuller regression should be run without a trend:<sup>11</sup>

$$X_t = \alpha + \phi X_{t-1} + Z_t.$$

Thus we have either  $\phi = 1$  and  $\alpha = 0$  or  $\phi < 1$  and  $\alpha \neq 0$ . The null hypothesis in this case therefore is

$$H_0 : \phi = 1 \text{ and } \alpha = 0.$$

A rejection of the null hypothesis can be interpreted in three alternative ways:

- (i) The case  $\phi < 1$  and  $\alpha = 0$  can be eliminated because it implies that  $\{X_t\}$  would have a mean of zero which is unrealistic for most economic time series.
- (ii) The case  $\phi = 1$  and  $\alpha \neq 0$  can equally be eliminated because it implies that  $\{X_t\}$  has a long-run trend which contradicts our primary assumption.
- (iii) The case  $\phi < 1$  and  $\alpha \neq 0$  is the only realistic alternative. It implies that the time series is stationary around a constant mean given by  $\frac{\alpha}{1-\phi}$ .

---

<sup>11</sup>In case of the ADF-test additional regressors,  $\Delta X_{t-j}, j > 0$ , might be necessary.

As before one can use, instead of a F-test, a t-test of the null hypothesis  $H_0 : \phi = 1$  against the alternative hypothesis  $H_1 : \phi < 1$ . If the null hypothesis is not rejected, we interpret this to imply that  $\alpha = 0$ . If, however, the null hypothesis  $H_0$  is rejected, we conclude that  $\alpha \neq 0$ . Similarly, Monte-Carlo simulations have proven that the t-test is superior to the F-test.

**The trend behavior of  $X_t$  is uncertain:** This situation poses the following problem. Should the data exhibit a trend, but the Dickey-Fuller regression contains no trend, then the test is biased in favor of the null hypothesis. If the data have no trend, but the Dickey-Fuller regression contains a trend, the power of the test is reduced. In such a situation one can adapt a two-stage strategy. Estimate the Dickey-Fuller regression with a linear trend:

$$X_t = \alpha + \delta t + \phi X_{t-1} + Z_t.$$

Use the t-test to test the null hypothesis  $H_0 : \phi = 1$  against the alternative hypothesis  $H_1 : \phi < 1$ . If  $H_0$  is not rejected, we conclude the process has a unit root with or without drift. The presence of a drift can then be investigated by a simple regression of  $\Delta X_t$  against a constant followed by a simple t-test of the null hypothesis that the constant is zero against the alternative hypothesis that the constant is nonzero. As  $\Delta X_t$  is stationary, the usual critical values can be used.<sup>12</sup> If the t-test rejects the null hypothesis  $H_0$ , we conclude that there is no unit root. The trend behavior can then be investigated by a simple t-test of the hypothesis  $H_0 : \delta = 0$ . In this test the usual critical values can be used as  $\{X_t\}$  is already viewed as being stationary.

### 7.3.4 Examples for unit root tests

As our first example, we examine the logged real GDP for Switzerland,  $\ln(\text{BIP}_t)$ , where we have adjusted the series for seasonality by taking a moving-average. The corresponding data are plotted in Figure 1.3. As is evident from this plot, this variable exhibits a clear trend so that the Dickey-Fuller regression should include a constant and a linear time trend. Moreover,  $\{\Delta \ln(\text{BIP}_t)\}$  is typically highly autocorrelated which makes an autoregressive correction necessary. One way to make this correction is by augmenting

---

<sup>12</sup>Eventually, one must correct the corresponding standard deviation by taking the autocorrelation in the residual into account. This can be done by using the long-run variance. In the literature this correction is known as the Newey-West correction.

the Dickey-Fuller regression by lagged  $\{\Delta \ln(\text{BIP}_t)\}$  as additional regressors. Thereby the number of lags is determined by AIC. The corresponding result is reported in the first column of table 7.2. It shows that AIC chooses only one autoregressive correction term. The value of t-test statistic is  $-3.110$  which is just above the 5-percent critical value. Thus, the null hypothesis is not rejected. If the autoregressive correction is chosen according to the method proposed by Ng and Perron five autoregressive lags have to be included. With this specification, the value of the t-test statistic is clearly above the critical value, implying that the null hypothesis of the presence of a unit root cannot be rejected (see second column in table 7.2).<sup>13</sup> The results of the ADF-tests is confirmed by the PP-test (column 3 in table 7.2) with quadratic spectral kernel function and band width 20.3 chosen according to Andrews' formula (see Section 4.4).

The second example, examines the three-month LIBOR,  $\{\text{R3M}_t\}$ . The series is plotted in Figure 1.4. The issue whether this series has a linear trend or not is not easy to decide. On the one hand, the series clearly has a negative trend over the sample period considered. On the other hand, a negative time trend does not make sense from an economic point of view because interest rates are bounded from below by zero. Because of this uncertainty, it is advisable to include in the Dickey-Fuller regression both a constant and a trend to be on the safe side. Column 5 in table 7.2 reports the corresponding results. The value of the t-statistic of the PP-test with Bartlett kernel function and band width of 5 according to the Newey-West rule of thumb is  $-2.142$  and thus higher than the corresponding 5-percent critical of  $-3.435$ . Thus, we cannot reject the null hypothesis of the presence of a unit root. We therefore conclude that the process  $\{\text{R3M}_t\}$  is integrated of order one, respectively difference-stationary. Based on this conclusion, the issue of the trend can now be decided by running a simple regression of  $\Delta \text{R3M}_t$  against a constant. This leads to the following results:

$$\Delta \text{R3M}_t = -0.0315 + e_t.$$

(0.0281)

where  $e_t$  denotes the least-squares residual. The mean of  $\Delta \text{R3M}_t$  is therefore  $-0.0315$ . This value is, however, statistically not significantly different from zero as indicated by the estimated standard error in parenthesis. Note that this estimate of the standard error has been corrected for autocorrelation (Newey-West correction). Thus,  $\{\text{R3M}_t\}$  is not subject to a linear trend. One could have therefore run the Dickey-Fuller regression without the trend

---

<sup>13</sup>The critical value changes slightly because the inclusion of additional autoregressive terms changes the sample size.

Table 7.2: Examples of unit root tests

	$\ln(\text{BIP}_t)$	$\ln(\text{BIP}_t)$	$\ln(\text{BIP}_t)$	$\text{R3M}_t$	$\text{R3M}_t$
test	ADF	ADF	PP	PP	PP
autoregressive correction	AIC	Ng and Perron	quadratic spectral	Bartlett	Bartlett
band width			20.3	5	5
$\alpha$	0.337	0.275	0.121	0.595	-0.014
$\delta$	0.0001	0.0001	0.0002	-0.0021	
$\phi$	0.970	0.975	0.989	0.963	-0.996
$\gamma_1$	0.885	1.047			
$\gamma_2$		-0.060			
$\gamma_3$		-0.085			
$\gamma_4$		-0.254			
$\gamma_5$		0.231			
$t_{\hat{\phi}}$	-3.110	-2.243	-1.543	-2.142	-0.568
critical value (5 percent)	-3.460	-3.463	-3.460	-3.435	-2.878

critical values from MacKinnon (1996)

term. The result of corresponding to this specification is reported in the last column of table 7.2. It confirms the presence of a unit root.

## 7.4 Generalizations of Unit-Root Tests

### 7.4.1 Structural Breaks in the Trend Function

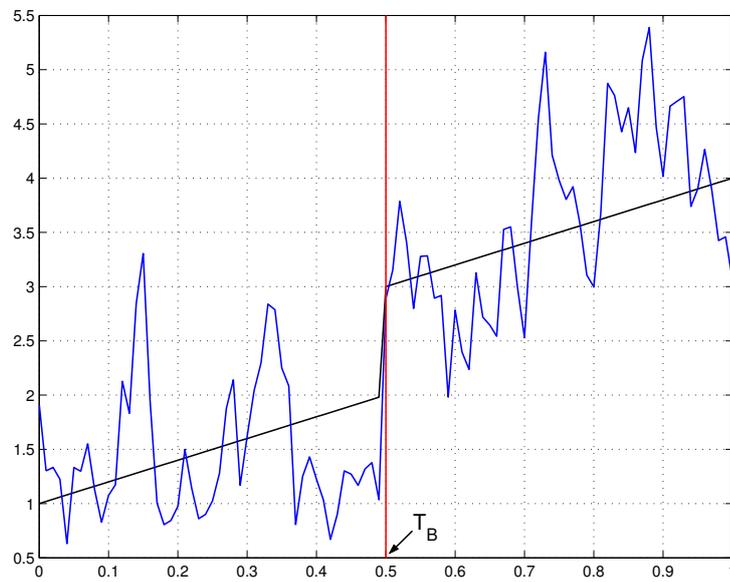
As we have seen, the unit-root test depends heavily on the correct specification of the deterministic part. Most of the time this amounts to decide whether a linear trend is present in the data or not. In the previous section we presented a rule how to proceed in case of uncertainty about the trend. Sometimes, however, the data exhibit a structural break in their deterministic component. If this structural break is ignored, the unit-root test is biased in favor of the null hypothesis (i. e. in favor of a unit root) as demonstrated by Perron (1989). Unfortunately, the distribution of the test statistic under the null hypothesis, in our case the t-statistic, depends on the exact nature of the structural break and on its date of occurrence in the data. Following Perron (1989) we concentrate on three exemplary cases: a level shift, a change in the slope (change in the growth rate), and a combination of both possibilities. Figure 7.4 shows the three possibilities assuming that a break occurred in period  $T_B$ . Thereby an AR(1) process with  $\phi = 0.8$  was superimposed on the deterministic part.

The unit-root test with the possibility of a structural break in period  $T_B$  is carried out using the Dickey-Fuller test. Thereby the date of the structural break is assumed to be known. This assumption, although restrictive, is justifiable in many applications. The first oil price shock in 1973 or the German reunification in 1989 are examples of structural breaks which can be dated exactly. Other examples would include changes in the way the data are constructed. These changes are usually documented by the data collecting agencies. Table 7.3 summarizes the three variants of Dickey-Fuller regression allowing for structural breaks.<sup>14</sup>

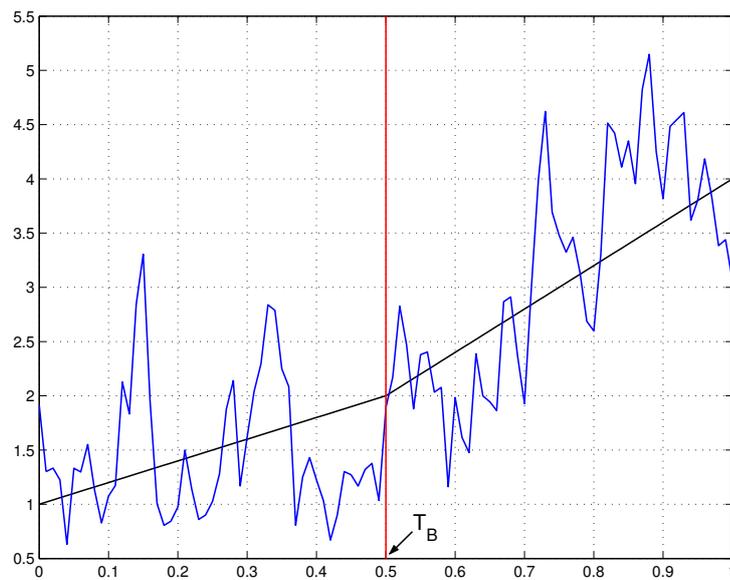
Model A allows only for a level shift. Under the null hypothesis the series undergoes a one-time shift at time  $T_B$ . This level shift is maintained under the null hypothesis which posits a random walk. Under the alternative, the process is viewed as being trend-stationary whereby the trend line shifts parallel by  $\alpha_B - \alpha$  at time  $T_B$ . Model B considers a change in the mean growth rate from  $\alpha$  to  $\alpha_B$  at time  $T_B$ . Under the alternative, the slope of time trend changes from  $\delta$  to  $\delta_B$ . Model C allows for both types of break to occur at the same time.

---

<sup>14</sup>See equation (7.1) and Table 7.1 for comparison.



(a) Level Shift



(b) Change in Slope

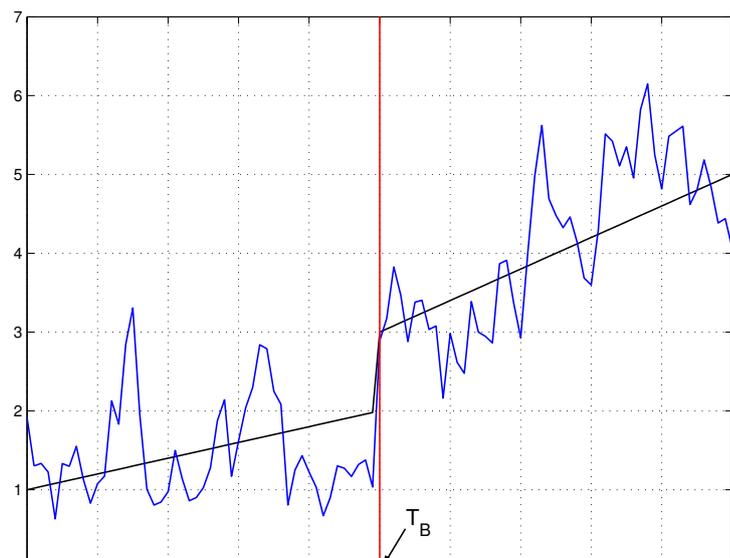


Table 7.3: Dickey-Fuller Regression allowing for Structural Breaks

---

<b>Model A: Level Shift</b>	
$\mathbf{H}_0 :$	$X_t = \alpha + \mathbf{1}_{\{t=T_B+1\}}\delta_B + X_{t-1} + Z_t$
$\mathbf{H}_1 :$	$X_t = \alpha + \delta t + \mathbf{1}_{\{t>T_B\}}(\alpha_B - \alpha) + \phi X_{t-1} + Z_t, \quad \phi < 1$

---

<b>Model B: Change in Slope (Change in Growth Rate)</b>	
$\mathbf{H}_0 :$	$X_t = \alpha + \mathbf{1}_{\{t>T_B\}}(\alpha_B - \alpha) + X_{t-1} + Z_t$
$\mathbf{H}_1 :$	$X_t = \alpha + \delta t + \mathbf{1}_{\{t>T_B\}}(\delta_B - \delta)(t - T_B) + \phi X_{t-1} + Z_t, \quad \phi < 1$

---

<b>Model C: Level Shift and Change in Slope</b>	
$\mathbf{H}_0 :$	$X_t = \alpha + \mathbf{1}_{\{t=T_B+1\}}\delta_B + \mathbf{1}_{\{t>T_B\}}(\alpha_B - \alpha) + X_{t-1} + Z_t$
$\mathbf{H}_1 :$	$X_t = \alpha + \delta t + \mathbf{1}_{\{t>T_B\}}(\alpha_B - \alpha) + \mathbf{1}_{\{t>T_B\}}(\delta_B - \delta)(t - T_B) + \phi X_{t-1} + Z_t, \quad \phi < 1$

---

$\mathbf{1}_{\{t=T_B+1\}}$  und  $\mathbf{1}_{\{t>T_B\}}$  denotes the indicator function which takes the value one if the condition is satisfied and the value zero otherwise.

The unit-root test with possible structural break for a time series  $X_t$ ,  $t = 0, 1, \dots, T$ , is implemented in two stages as follows. In the first stage, we regress  $X_t$  on the corresponding deterministic component using OLS. The residuals  $\tilde{X}_0, \tilde{X}_1, \dots, \tilde{X}_T$  from this regression are then used to carry out a Dickey-Fuller test:

$$\tilde{X}_t = \phi \tilde{X}_{t-1} + Z_t, \quad t = 1, \dots, T.$$

The distribution of the corresponding t-statistic under the null hypothesis depends not only on the type of the structural break, but also on the relative date of the break in the sample. Let this relative date be parameterized by  $\lambda = T_B/T$ . The asymptotic distribution of the t-statistic has been tabulated by Perron (1989). This table can be used to determine the critical values for the test. These critical values are smaller than those from the normal Dickey-Fuller table. Using a five percent significance level, the critical values range between  $-3.80$  and  $-3.68$  for model A, between  $-3.96$  and  $-3.65$  for model B, and between  $-4.24$  and  $-3.75$  for model C, depending on the value of  $\lambda$ . These values also show that the dependence on  $\lambda$  is only weak.

In the practical application of the test one has to control for the autocorrelation in the data. This can be done by using the Augmented Dickey-Fuller

(ADF) test. This amounts to the introduction of  $\Delta\tilde{X}_{t-j}$ ,  $t = 1, 2, \dots, p - 1$ , as additional regressors in the above Dickey-Fuller regression. Thereby the order  $p$  can be determined by Akaike's information criterion (AIC) or by the iterative testing procedure of Ng and Perron (1995). Alternatively, one may use, instead of the ADF test, the Phillips-Perron test. In this case one computes the usual t-statistic for the null hypothesis  $\phi = 1$  and corrects it using the formulas in Phillips and Perron (1988) as explained in Section 7.3.2. Which of the two methods is used, is irrelevant for the determination of the critical values which can be extracted from Perron (1989).

Although it may be legitimate in some cases to assume that the time of the structural break is known, we cannot take this for granted. It is therefore important to generalize the test allowing for an unknown date for the occurrence of a structural break. The work of Zivot and Andrews (1992) has shown that the procedure proposed by Perron can be easily expanded in this direction. We keep the three alternative models presented in Table 7.3, but change the null hypothesis to a random walk with drift with no exogenous structural break. Under the null hypothesis,  $\{X_t\}$  is therefore assumed to be generated by

$$X_t = \alpha + X_{t-1} + Z_t, \quad Z_t \sim \text{WN}(0, \sigma^2).$$

The time of the structural  $T_B$ , respectively  $\lambda = T_B/T$ , is estimated in such a way that  $\{X_t\}$  comes as close as possible to a trend-stationary process. Under the alternative hypothesis  $\{X_t\}$  is viewed as a trend-stationary process with unknown break point. The goal of the estimation strategy is to choose  $T_B$ , respectively  $\lambda$ , in such a way that the trend-stationary alternative receives the highest weight. Zivot and Andrews (1992) propose to estimate  $\lambda$  by minimizing the value of the t-statistic  $t_{\hat{\phi}}(\lambda)$  under the hypothesis  $\phi = 1$ :

$$t_{\hat{\phi}}(\hat{\lambda}_{\text{inf}}) = \inf_{\lambda \in \Lambda} t_{\hat{\phi}}(\lambda) \quad (7.2)$$

where  $\Lambda$  is a closed subinterval of  $(0, 1)$ .<sup>15</sup> The distribution of the test statistic under the null hypothesis for the three cases is tabulated in Zivot and Andrews (1992). This table then allows to determine the appropriate critical values for the test. In practice, one has to take the autocorrelation of the time series into account by one of the methods discussed previously.

This testing strategy can be adapted to determine the time of a structural break in the linear trend irrespective of whether the process is trend-

---

<sup>15</sup>Taking the infimum over  $\Lambda$  instead over  $(0, 1)$  is for theoretical reasons only. In practice, the choice of  $\Lambda$  plays no important role. For example, one may take  $\Lambda = [0.01, 0.99]$ .

stationary or integrated of order one. The distributions of the corresponding test statistics have been tabulated by Vogelsang (1997).<sup>16</sup>

### 7.4.2 Testing for Stationarity (KPSS-Test)

The unit-root tests we discussed so far tested the null hypothesis that the process is integrated of order one against the alternative hypothesis that the process is integrated of order zero (i. e. is stationary). However, one may be interested in reversing the null and the alternative hypothesis and test the hypothesis of stationarity against the alternative that the process is integrated of order one. Such a test has been proposed by Kwiatkowski et al. (1992), called the KPSS-Test. This test rests on the idea that according to the Beveridge-Nelson decomposition (see Section 7.1.4) each integrated process of order one can be seen as the sum of a linear time trend, a random walk and a stationary process:

$$X_t = \alpha + \delta t + d \sum_{j=1}^t Z_j + U_t,$$

where  $\{U_t\}$  denotes a stationary process. If  $d = 0$  then the process becomes trend-stationary, otherwise it is integrated of order one.<sup>17</sup> Thus one can state the null and the alternative hypothesis as follows:

$$H_0 : d = 0 \quad \text{against} \quad H_1 : d \neq 0.$$

Denote by  $\{S_t\}$  the process of partial sums obtained from the residuals  $\{e_t\}$  of a regression of  $X_t$  against a constant and a linear time trend, i.e.  $S_t = \sum_{j=1}^t e_j$ . Under the null hypothesis  $d = 0$ ,  $\{S_t\}$  is integrated of order one whereas under the alternative  $\{S_t\}$  is intergrated of order two. Based on this consideration Kwiatkowski et al. propose the following test statistic for a time series consisting of  $T$  observations:

$$\text{KPSS test statistic:} \quad W_T = \frac{\sum_{t=1}^T S_t^2}{T^2 \widehat{J}_T} \quad (7.3)$$

where  $\widehat{J}_T$  is an estimate of the long-run variance of  $\{U_t\}$  (see Section 4.4). As  $\{S_t\}$  is an integrated process under the null hypothesis, the variance of  $\{S_t\}$  grows linearly in  $t$  (see Section 1.4.4 or 7.2) so that the sum of squared  $S_t$  diverges at rate  $T^2$ . Therefore the test statistic is independent from further nuisance parameters. Under the alternative hypothesis, however,  $\{S_t\}$

<sup>16</sup>See also the survey by Perron (2006)

<sup>17</sup>If the data exhibit no trend, one can set  $\delta$  equal to zero.

Table 7.4: Critical Values of the KPSS-Test

	regression without time trend		
significance level:	0.1	0.05	0.01
critical value:	0.347	0.463	0.739
	regression with time trend		
significance level:	0.1	0.05	0.01
critical value:	0.119	0.146	0.216

See Kwiatkowski et al. (1992)

is integrated of order two. Thus, the null hypothesis will be rejected for large values of  $W_T$ . The corresponding asymptotic critical values of the test statistic are reported in Table 7.4.

## 7.5 Regression with Integrated Variables

### 7.5.1 The Spurious Regression Problem

The discussion on the Dickey-Fuller and Phillips-Perron tests showed that in a regression of the integrated variables  $X_t$  on its past  $X_{t-1}$  the standard  $\sqrt{T}$ -asymptotics no longer apply. A similar conclusion also holds if we regress an integrated variable  $X_t$  against another integrated variable  $Y_t$ . Suppose that both processes  $\{X_t\}$  and  $\{Y_t\}$  are generated as a random walk:

$$\begin{aligned} X_t &= X_{t-1} + U_t, & U_t &\sim \text{IID}(0, \sigma_U^2) \\ Y_t &= Y_{t-1} + V_t, & V_t &\sim \text{IID}(0, \sigma_V^2) \end{aligned}$$

where the processes  $\{U_t\}$  and  $\{V_t\}$  are uncorrelated with each other at all leads and lags. Thus,

$$\mathbb{E}(U_t V_s) = 0, \quad \text{for all } t, s \in \mathbb{Z}.$$

Consider now the regression of  $Y_t$  on  $X_t$  and a constant:

$$Y_t = \alpha + \beta X_t + \varepsilon_t.$$

As  $\{X_t\}$  and  $\{Y_t\}$  are two random walks which are uncorrelated with each other by construction, one would expect that the OLS-estimate of the coefficient of  $X_t$ ,  $\hat{\beta}$ , should tend to zero as the sample size  $T$  goes to infinity. The

same is expected for the coefficient of determination  $R^2$ . This is, however, not true as has already been remarked by Yule (1926) and, more recently, by Granger and Newbold (1974). The above regression will have a tendency to “discover” a relationship between  $Y_t$  and  $X_t$  despite the fact that there is none. This phenomenon is called *spurious correlation* or *spurious regression*. Similarly, unreliable results would be obtained by using a simple t-test for the null hypothesis  $\beta = 0$  against the alternative hypothesis  $\beta \neq 0$ . The reason for these treacherous findings is that the model is incorrect under the null as well as under the alternative hypothesis. Under the null hypothesis  $\{\varepsilon_t\}$  is an integrated process which violates the standard assumption for OLS. The alternative hypothesis is not true by construction. Thus, OLS-estimates should be interpreted with caution when a highly autocorrelated process  $\{Y_t\}$  is regressed on another highly correlated process  $\{X_t\}$ . A detailed analysis of the spurious regression problem is provided by Phillips (1986).

The spurious regression problem can be illustrated by a simple Monte Carlo study. Specifying  $U_t \sim \text{IIDN}(0, 1)$  and  $V_t \sim \text{IIDN}(0, 1)$ , we constructed  $N = 1000$  samples for  $\{Y_t\}$  and  $\{X_t\}$  of size  $T = 1000$  according to the specification above. The sample size was chosen especially large to demonstrate that this is not a small sample issue. As a contrast, we constructed two independent AR(1) processes with AR-coefficients  $\phi_X = 0.8$  and  $\phi_Y = -0.5$ .

Figures 7.5(a) and 7.5(b) show the drastic difference between a regression with stationary variables and integrated variables. Whereas the distribution of the OLS-estimates of  $\beta$  is highly concentrated around the true value  $\beta = 0$  in the stationary case, the distribution is very flat in the case of integrated variables. A similar conclusion holds for the corresponding t-value. The probability of obtaining a t-value greater than 1.96 is bigger than 0.9. This means that in more than 90 percent of the time the t-statistic leads to a rejection of the null hypothesis and therefore suggests a relationship between  $Y_t$  and  $X_t$  despite their independence. In the stationary case, this probability turns out to be smaller than 0.05. These results are also reflected in the coefficient of determination  $R^2$ . The median  $R^2$  is approximately 0.17 in the case of the random walks, but only 0.0002 in the case of AR(1) processes.

The problem remains the same if  $\{X_t\}$  and  $\{Y_t\}$  are specified as random walks with drift:

$$\begin{aligned} X_t &= \delta_X + X_{t-1} + U_t, & U_t &\sim \text{IID}(0, \sigma_U^2) \\ Y_t &= \delta_Y + Y_{t-1} + V_t, & V_t &\sim \text{IID}(0, \sigma_V^2) \end{aligned}$$

where  $\{U_t\}$  and  $\{V_t\}$  are again independent from each other at all leads and lags. The regression would be same as above:

$$Y_t = \alpha + \beta X_t + \varepsilon_t.$$

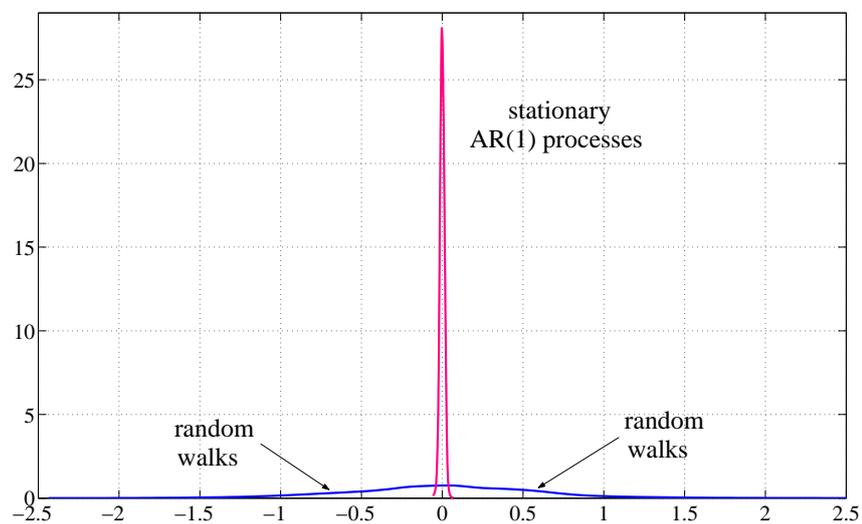
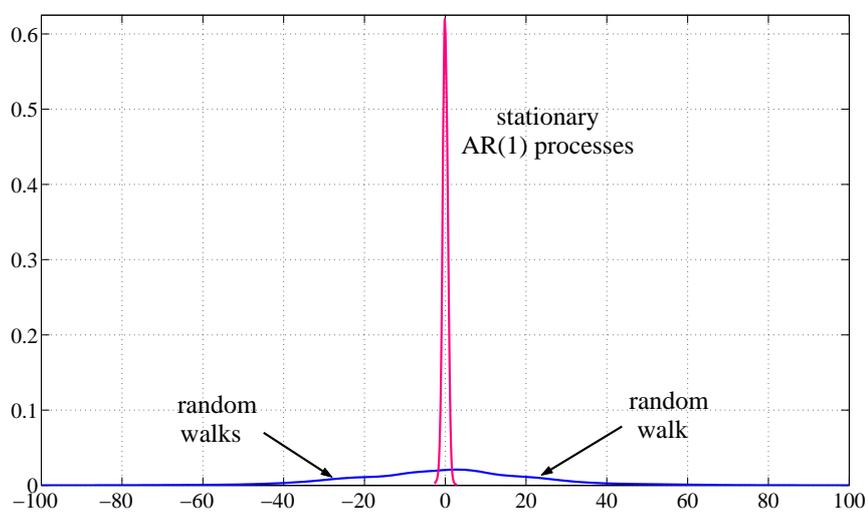
(a) distribution of  $\hat{\beta}$ (b) Distribution of  $t_{\hat{\beta}}$ 

Figure 7.5: Distribution of OLS-estimate  $\hat{\beta}$  and t-statistic  $t_{\hat{\beta}}$  for two independent random walks and two independent AR(1) processes

### 7.5.2 Cointegration

The spurious regression problem cannot be circumvented by first testing for a unit root in  $Y_t$  and  $X_t$  and then running the regression in first differences in case of no rejection of the null hypothesis. The reason being that a regression in the levels of  $Y_t$  and  $X_t$  may be sensible even when both variables are integrated. This is the case when both variables are cointegrated. The concept of *cointegration* goes back to Engle and Granger (1987) and initiated a literal boom in research. We will give a more general definition in Chapter 16 when we deal with multivariate time series. Here we stick to the case of two variables and present the following definition.

**Definition 7.2** (Cointegration, bivariate). *Two stochastic processes  $\{X_t\}$  and  $\{Y_t\}$  are called cointegrated if the following two conditions are fulfilled:*

- (i)  $\{X_t\}$  and  $\{Y_t\}$  are both integrated processes of order one, i.e.  $X_t \sim I(1)$  and  $Y_t \sim I(1)$ ;
- (ii) there exists a constant  $\beta \neq 0$  such that  $\{Y_t - \beta X_t\}$  is a stationary process, i.e.  $\{Y_t - \beta X_t\} \sim I(0)$ .

The issue whether two integrated processes are cointegrated can be decided on the basis of a unit root test. Two cases can be distinguished. In the first one,  $\beta$  is assumed to be known. Thus, one can immediately apply the augmented Dickey-Fuller (ADF) or the Phillips-Perron (PP) test to the process  $\{Y_t - \beta X_t\}$ . Thereby the same issue regarding the specification of the deterministic part arises. The critical values can be retrieved from the usual tables (for example from MacKinnon, 1991). In the second case,  $\beta$  is not known and must be estimated from the data. This can be done running, as a first step, a simple (cointegrating) regression of  $Y_t$  on  $X_t$  including a constant and/or a time trend.<sup>18</sup> Thereby the specification of the deterministic part follows the same rules as before. The unit root test is then applied, in the second step, to the residuals from this regression. As the residuals have been obtained from a preceding regression, we are faced with the so-called “generated regressor problem”.<sup>19</sup> This implies that the usual Dickey-Fuller tables can no longer be used, instead the tables provided by Phillips and Ouliaris (1990) become the relevant ones. As before, the relevant asymptotic distribution depends on the specification of the deterministic part in the cointegrating regression. If this regression included a constant, the residuals

<sup>18</sup>Thereby, in contrast to ordinary OLS regressions, it is irrelevant which variable is treated as the left hand, respectively right hand variable.

<sup>19</sup>This problem was first analyzed by Nicholls and Pagan (1984) and Pagan (1984) in a stationary context.

have necessary a mean of zero so that the Dickey-Fuller regression should include no constant (case 1 in Table 7.1):

$$e_t = \phi e_{t-1} + \xi_t$$

where  $e_t$  and  $\xi_t$  denote the residuals from the cointegrating and the residuals of the Dickey-Fuller regression, respectively. In most applications it is necessary to correct for autocorrelation which can be done by including additional lagged differences  $\Delta \hat{\varepsilon}_{t-1}, \dots, \Delta \hat{\varepsilon}_{t-p+1}$  as in the ADF-test or by adjusting the t-statistic as in the PP-test. The test where  $\beta$  is estimated from a regression is called the *regression test for cointegration*.

### An example for bivariate cointegration

As an example, we consider the relation between the short-term interest rate,  $\{R3M_t\}$ , and inflation,  $\{INFL_t\}$ , in Switzerland over the period January 1989 to February 2012. As the short-term interest rate we take the three month LIBOR. Both time series are plotted in Figure 7.6(a). As they are integrated according to the unit root tests (not shown here), we can look for cointegration. The cointegrating regression delivers:

$$INFL_t = -0.0088 + 0.5535 R3M_t + e_t, \quad R^2 = 0.7798. \quad (7.4)$$

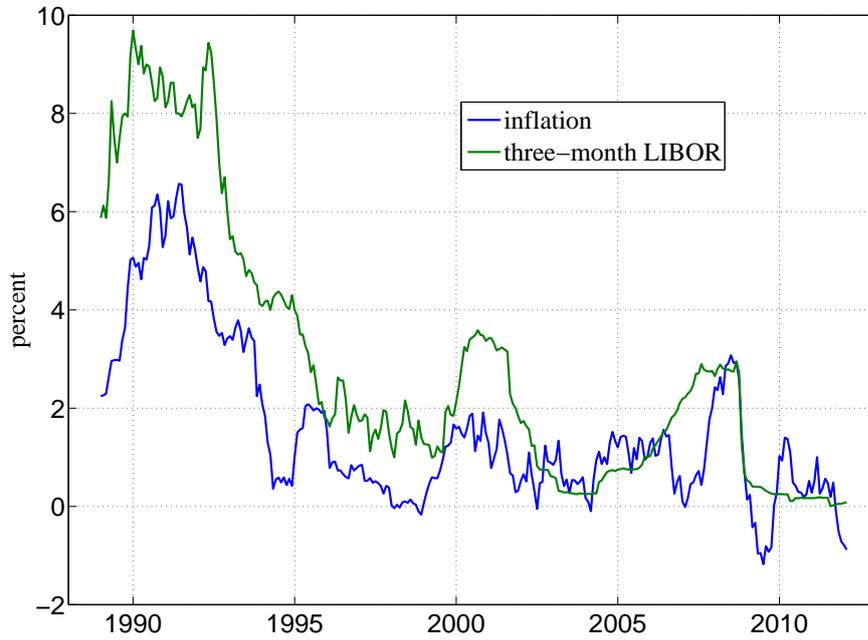
The residuals from this regression, denoted by  $e_t$ , are represented in Figure 7.6(b). The ADF unit root test of these residuals leads to a value of  $-3.617$  for the t-statistic. Thereby an autoregressive correction of 13 lags was necessary according to the AIC criterion. The corresponding value of the t-statistic resulting from the PP unit root test using a Bartlett window with band width 7 is  $-4.294$ . Taking a significance level of five percent, the critical value according to Phillips and Ouliaris (1990, Table IIb) is  $-3.365$ .<sup>20</sup> Thus, the ADF as well as the PP test reject the null hypothesis of a unit root in the residuals. This implies that inflation and the short-term interest rate are cointegrated.

### 7.5.3 Some rules to deal with integrated times series in regressions

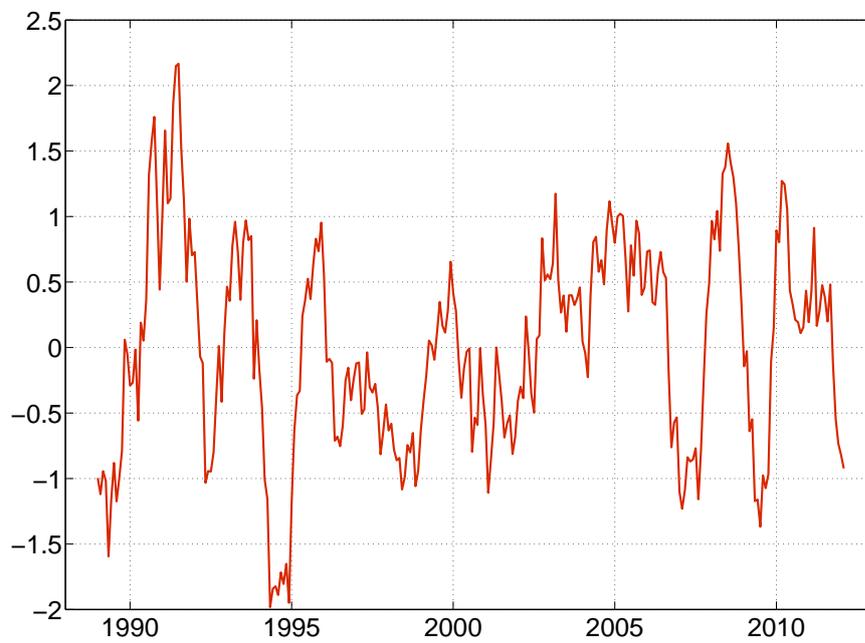
The previous sections demonstrated that the handling of integrated variables has to be done with care. We will therefore in this section examine some rules

---

<sup>20</sup>For comparison, the corresponding critical value according to MacKinnon (1991) is  $-2.872$ .



(a) Inflation and three-month LIBOR



(b) Residuals from cointegrating regression

Figure 7.6: Cointegration of inflation and three-month LIBOR

of thumb which should serve as a guideline in practical empirical work. These rules are summarized in Table 7.5. In that this section follows very closely the paper by Stock and Watson (1988a) (see also Campbell and Perron, 1991).<sup>21</sup> Consider the linear regression model:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \dots + \beta_K X_{K,t} + \varepsilon_t. \quad (7.5)$$

This model is usually based on two assumptions:

- (1) The disturbance term  $\varepsilon_t$  is white noise and is uncorrelated with any regressor. This is, for example, the case if the regressors are deterministic or exogenous.
- (2) All regressors are either deterministic or stationary processes.

If equation (7.5) represents the true data generating process,  $\{Y_t\}$  must be a stationary process. Under the above assumptions, the OLS-estimator is consistent and the OLS-estimates are asymptotically normally distributed so that the corresponding t- and F-statistics will be approximately distributed as t- and F- distributions.

Consider now the case that assumption 2 is violated and that some or all regressors are integrated, but that instead one of the two following assumptions holds:

- (2.a) The relevant coefficients are coefficients of mean-zero stationary variables.
- (2.b) Although the relevant coefficients are those of integrated variables, the regression can be rearranged in such a way that the relevant coefficients become coefficients of mean-zero stationary variables.

Under assumptions 1 and 2.a or 2.b the OLS-estimator remains consistent. Also the corresponding t- and F-statistics remain valid so that the appropriate critical values can be retrieved from the t-, respectively F-distribution. If neither assumption 2.a nor 2.b holds, but the following assumption:

- (2.c) The relevant coefficients are coefficients of integrated variables and the regression cannot be rewritten in a way that they become coefficients of stationary variables.

If assumption 1 remains valid, but assumption 2.c holds instead of 2.a and 2.b, the OLS-estimator is still consistent. However, the standard asymptotic

---

<sup>21</sup>For a thorough analysis the interested reader is referred to Sims et al. (1990).

Table 7.5: Rules of thumb in regressions with integrated processes (summary)

assumptions		OLS-estimator		remarks
		consistency	standard asymptotics	
(1)	(2)	yes	yes	classic results for OLS
(1)	(2.a)	yes	yes	
(1)	(2.b)	yes	yes	
(1)	(2.c)	yes	no	
(1.a)	(2.a)	no	no	omitted variable bias
(1.a)	(2.b)	no	no	omitted variable bias
(1.a)	(2.c)	yes	no	
neither (1) nor (1.a)	(2.c)	no	no	spurious regression

Source: Stock and Watson (1988a); results for the coefficients of interest

theory for the  $t$ - and the  $F$ -statistic fails so that they become useless for normal statistical inferences.

If we simply regress one variable on another in levels, the error term  $\varepsilon_t$  is likely not to follow a white noise process. In addition, it may even be correlated with some regressors. Suppose that we replace assumption 1 by:

- (1.a) The integrated dependent variable is cointegrated with at least one integrated regressor such that the error term is stationary, but may remain autocorrelated or correlated with the regressors.

Under assumptions 1.a and 2.a, respectively 2.b, the regressors are stationary, but correlated with the disturbance term, in this case the OLS-estimator becomes inconsistent. This situation is known as the classic omitted variable bias, simultaneous equation bias or errors-in-variable bias. However, under assumptions 1.a and 2.c, the OLS-estimator is consistent for the coefficients of interest. However, the standard asymptotic theory fails. Finally, if both the dependent variable and the regressors are integrated without being cointegrated, then the disturbance term is integrated and the OLS-estimator becomes inconsistent. This is the spurious regression problem treated in Section 7.5.1.

### Example: term structure of interest

We illustrate the above rules of thumb by investigating again the relation between inflation ( $\{\text{INFL}_t\}$ ) and the short-term interest rate ( $\{\text{R3M}_t\}$ ). In

Section 7.5.2 we found that the two variables are cointegrated with coefficient  $\hat{\beta} = 0.5535$  (see equation (7.4)). In a further step we want to investigate a dynamic relation between the two variables and estimate the following equation:

$$\text{R3M}_t = c + \phi_1 \text{R3M}_{t-1} + \phi_2 \text{R3M}_{t-2} + \phi_3 \text{R3M}_{t-3} + \delta_1 \text{INFL}_{t-1} + \delta_2 \text{INFL}_{t-2} + \varepsilon_t$$

where  $\varepsilon_t \sim WN(0, \sigma^2)$ . In this regression we want to test the hypotheses  $\phi_3 = 0$  against  $\phi_3 \neq 0$  and  $\delta_1 = 0$  against  $\delta_1 \neq 0$  by examining the corresponding simple t-statistics. Note that we are in the context of integrated variables so that the rules of thumb summarized in table 7.5 apply. We can rearrange the above equation as

$$\begin{aligned} \Delta \text{R3M}_t &= c \\ &+ (\phi_1 + \phi_2 + \phi_3 - 1) \text{R3M}_{t-1} - (\phi_2 + \phi_3) \Delta \text{R3M}_{t-1} - \phi_3 \Delta \text{R3M}_{t-2} \\ &\quad + \delta_1 \text{INFL}_{t-1} + \delta_2 \text{INFL}_{t-2} + \varepsilon_t. \end{aligned}$$

$\phi_3$  is now a coefficient of a stationary variable in a regression with a stationary dependent variables. In addition  $\varepsilon_t \sim WN(0, \sigma^2)$  so that assumptions (1) und (2.b) are satisfied. We can therefore use the ordinary t-statistic to test the hypothesis  $\phi_3 = 0$  against  $\phi_3 \neq 0$ . Note that it is not necessary to actually carry out the rearrangement of the equation. All relevant item can be retrieved from the original equation.

To test the hypothesis  $\delta_1 = 0$ , we rearrange the equation to yield:

$$\begin{aligned} \Delta \text{R3M}_t &= c \\ &+ (\phi_1 + \delta_1 \hat{\beta} - 1) \text{R3M}_{t-1} + \phi_2 \text{R3M}_{t-2} + \phi_3 \text{R3M}_{t-3} \\ &\quad + \delta_1 (\text{INFL}_{t-1} - \hat{\beta} \text{R3M}_{t-1}) + \delta_2 \text{INFL}_{t-2} + \varepsilon_t. \end{aligned}$$

As  $\{\text{R3M}_t\}$  and  $\{\text{INFL}_t\}$  are cointegrated,  $\text{INFL}_{t-1} - \hat{\beta} \text{R3M}_{t-1}$  is stationary. Thus assumptions (1) and (2.b) hold again and we use once more the simple t-test. As before it is not necessary to actually carry out the transformation.



# Chapter 8

## Models of Volatility

The prices of financial market securities are often shaken by large and time-varying shocks. The amplitudes of these price movements are not constant. There are periods of high volatility and periods of low volatility. Within these periods volatility seems to be positively autocorrelated: high amplitudes are likely to be followed by high amplitudes and low amplitudes by low amplitudes. This observation which is particularly relevant for high frequency data such as, for example, daily stock market returns implies that the conditional variance of the one-period forecast error is no longer constant (homoskedastic), but time-varying (heteroskedastic). This insight motivated Engle (1982) and Bollerslev (1986) to model the time-varying variance thereby triggering a huge and still growing literature.<sup>1</sup> The importance of volatility models stems from the fact that the price of an option crucially depends on the variance of the underlying security price. Thus with the surge of derivative markets in the last decades the application of such models has seen a tremendous rise. Another use of volatility models is to assess the risk of an investment. In the computation of the so-called value at risk (VaR), these models have become an indispensable tool. In the banking industry, due to the regulations of the Basel accords, such assessments are in particular relevant for the computation of the required equity capital backing-up assets of different risk categories.

The following exposition focuses on the class of autoregressive conditional heteroskedasticity models (ARCH models) and their generalization the generalized autoregressive conditional heteroskedasticity models (GARCH models). These models form the basis for even more generalized models (see Bollerslev et al. (1994) or Gouriéroux (1997)). Campbell et al. (1997) pro-

---

<sup>1</sup>Robert F. Engle III was awarded the Nobel prize in 2003 for his work on time-varying volatility. His Nobel lecture (Engle, 2004) is a nice and readable introduction to this literature.

vide a broader economically motivated approach to the econometric analysis of financial market data.

## 8.1 Specification and Interpretation

### 8.1.1 Summary of the Forecasting Properties of the AR(1)-Model

Models of volatility play an important role in explaining the behavior of financial market data. They start from the observation that periods of high (low) volatility are clustered in specific time intervals. In these intervals high (low) volatility periods are typically followed by high (low) volatility periods. Thus volatility is usually positively autocorrelated as can be observed in Figure 8.3. In order to understand this phenomenon we recapitulate the forecasting properties of the AR(1) model.<sup>2</sup> Starting from the model

$$X_t = c + \phi X_{t-1} + Z_t, \quad Z_t \sim \text{IID}(0, \sigma^2) \text{ and } |\phi| < 1,$$

the best linear forecast in the mean-squared-error sense of  $X_{t+1}$  conditional on  $\{X_t, X_{t-1}, \dots\}$ , denoted by  $\mathbb{P}_t X_{t+1}$ , is given by (see Chapter 3)

$$\mathbb{P}_t X_{t+1} = c + \phi X_t.$$

In practice the parameters  $c$  and  $\phi$  are replaced by an estimate.

The *conditional* variance of the forecast error then becomes:

$$\mathbb{E}_t (X_{t+1} - \mathbb{P}_t X_{t+1})^2 = \mathbb{E}_t Z_{t+1}^2 = \sigma^2,$$

where  $\mathbb{E}_t$  denotes the conditional expectation operator based on information  $X_t, X_{t-1}, \dots$ . The conditional variance of the forecast error is therefore constant, irrespective of the current state.

The *unconditional* forecast is simply the expected value of  $\mathbb{E}X_{t+1} = \mu = \frac{c}{1-\phi}$  with forecast error variance:

$$\mathbb{E} \left( X_{t+1} - \frac{c}{1-\phi} \right)^2 = \mathbb{E} (Z_{t+1} + \phi Z_t + \phi^2 Z_{t-1} + \dots)^2 = \frac{\sigma^2}{1-\phi^2} > \sigma^2.$$

Thus the conditional as well as the unconditional variance of the forecast error are constant. In addition, the conditional variance is smaller and thus more precise because it uses more information. Similar arguments can be made for ARMA models in general.

<sup>2</sup>Instead of assuming  $Z_t \sim \text{WN}(0, \sigma^2)$ , we make for convenience the stronger assumption that  $Z_t \sim \text{IID}(0, \sigma^2)$ .

### 8.1.2 The ARCH(1) model

The volatility of financial market prices exhibit a systematic behavior so that the conditional forecast error variance is no longer constant. This observation led Engle (1982) to consider the following simple model for heteroskedasticity (non-constant variance).

**Definition 8.1** (ARCH(1) model). *A stochastic process  $\{Z_t\}$ ,  $t \in \mathbb{Z}$ , is called an autoregressive conditional heteroskedastic process of order one, ARCH(1) process, if it is the solution of the following stochastic difference equation:*

$$Z_t = \nu_t \sqrt{\alpha_0 + \alpha_1 Z_{t-1}^2} \quad \text{with } \alpha_0 > 0 \text{ and } 0 < \alpha_1 < 1, \quad (8.1)$$

where  $\nu_t \sim \text{IID } N(0,1)$  and where  $\nu_t$  and  $Z_{t-1}$  are independent from each other for all  $t \in \mathbb{Z}$ .

We will discuss the implications of this simple model below and consider generalizations in the next sections. First we prove the following theorem.

**Theorem 8.1.** *Under conditions stated in the definition of the ARCH(1) process, the difference equation (8.1) possesses a unique and strictly stationary solution with  $\mathbb{E}Z_t^2 < \infty$ . This solution is given by*

$$Z_t = \nu_t \sqrt{\alpha_0 \left( 1 + \sum_{j=1}^{\infty} \alpha_1^j \nu_{t-1}^2 \nu_{t-2}^2 \cdots \nu_{t-j}^2 \right)}. \quad (8.2)$$

*Proof.* Define the process

$$Y_t = Z_t^2 = \nu_t^2 (\alpha_0 + \alpha_1 Y_{t-1}) \quad (8.3)$$

Iterating backwards  $k$  times we get:

$$\begin{aligned} Y_t &= \alpha_0 \nu_t^2 + \alpha_1 \nu_t^2 Y_{t-1} = \alpha_0 \nu_t^2 + \alpha_1 \nu_t^2 \nu_{t-1}^2 (\alpha_0 + \alpha_1 Y_{t-2}) \\ &= \alpha_0 \nu_t^2 + \alpha_0 \alpha_1 \nu_t^2 \nu_{t-1}^2 + \alpha_1^2 \nu_t^2 \nu_{t-1}^2 Y_{t-2} \\ &\quad \dots \\ &= \alpha_0 \nu_t^2 + \alpha_0 \alpha_1 \nu_t^2 \nu_{t-1}^2 + \dots + \alpha_0 \alpha_1^k \nu_t^2 \nu_{t-1}^2 \cdots \nu_{t-k}^2 \\ &\quad + \alpha_1^{k+1} \nu_t^2 \nu_{t-1}^2 \cdots \nu_{t-k}^2 Y_{t-k-1}. \end{aligned}$$

Define the process  $\{Y'_t\}$  as

$$Y'_t = \alpha_0 \nu_t^2 + \alpha_0 \sum_{j=1}^{\infty} \alpha_1^j \nu_t^2 \nu_{t-1}^2 \cdots \nu_{t-j}^2.$$

The right-hand side of the above expression just contains nonnegative terms. Moreover, making use of the IID  $N(0, 1)$  assumption of  $\{\nu_t\}$ ,

$$\begin{aligned}\mathbb{E}Y'_t &= \mathbb{E}(\alpha_0\nu_t^2) + \alpha_0\mathbb{E}\left(\sum_{j=1}^{\infty}\alpha_1^j\nu_t^2\nu_{t-1}^2\cdots\nu_{t-j}^2\right) \\ &= \alpha_0\sum_{j=0}^{\infty}\alpha_1^j = \frac{\alpha_0}{1-\alpha_1}.\end{aligned}$$

Thus,  $0 \leq Y'_t < \infty$  a.s. Therefore,  $\{Y'_t\}$  is strictly stationary and satisfies the difference equation (8.3). This implies that  $Z_t = \sqrt{Y'_t}$  is also strictly stationary and satisfies the difference equation (8.1).

To prove uniqueness, we follow Giraitis et al. (2000). For any fixed  $t$ , it follows from the definitions of  $Y_t$  and  $Y'_t$  that for any  $k \geq 1$

$$|Y_t - Y'_t| \leq \alpha_1^{k+1}\nu_t^2\nu_{t-1}^2\cdots\nu_{t-k}^2|Y_{t-k-1}| + \alpha_0\sum_{j=k+1}^{\infty}\alpha_1^j\nu_t^2\nu_{t-1}^2\cdots\nu_{t-j}^2.$$

The expectation of the right-hand side is bounded by

$$\left(\mathbb{E}|Y_1| + \frac{\alpha_0}{1-\alpha_1}\right)\alpha_1^{k+1}.$$

Define the event  $A_k$  by  $A_k = \{|Y_t - Y'_t| > 1/k\}$ . Then,

$$\mathbf{P}(A_k) \leq k\mathbb{E}|Y_t - Y'_t| \leq k\left(\mathbb{E}|Y_1| + \frac{\alpha_0}{1-\alpha_1}\right)\alpha_1^{k+1}$$

where the first inequality follows from Chebyshev's inequality setting  $r = 1$  (see Theorem C.3). Thus,  $\sum_{k=1}^{\infty}\mathbf{P}(A_k) < \infty$ . The Borel-Cantelli lemma (see Theorem C.4) then implies that  $\mathbf{P}\{A_k \text{ i.o.}\} = 0$ . However, as  $A_k \subset A_{k+1}$ ,  $\mathbf{P}(A_k) = 0$  for any  $k$ . Thus,  $Y_t = Y'_t$  a.s.  $\square$

**Remark 8.1.** *Note that the normality assumption is not necessary for the proof. The assumption  $\nu_t \sim \text{IID}(0, 1)$  would be sufficient. Indeed, in practice it has been proven useful to adopt distributions with fatter tail than the normal, like the  $t$ -distribution (see the discussion in Section 8.1.3).*

Given the assumptions made above  $\{Z_t\}$  has the following properties:

- (i) The expected value of  $Z_t$  is:

$$\mathbb{E}Z_t = \mathbb{E}\nu_t\mathbb{E}\sqrt{\alpha_0 + \alpha_1 Z_{t-1}^2} = 0.$$

This follows from the assumption that  $\nu_t$  and  $Z_{t-1}$  are independent.

(ii) The covariances between  $Z_t$  and  $Z_{t-h}$ ,  $\mathbb{E}Z_t Z_{t-h}$ , for  $h \neq 0$  are given by:

$$\begin{aligned}\mathbb{E}Z_t Z_{t-h} &= \mathbb{E} \left( \nu_t \sqrt{\alpha_0 + \alpha_1 Z_{t-1}^2} \nu_{t-h} \sqrt{\alpha_0 + \alpha_1 Z_{t-h-1}^2} \right) \\ &= \mathbb{E} \nu_t \nu_{t-h} \mathbb{E} \sqrt{\alpha_0 + \alpha_1 Z_{t-1}^2} \sqrt{\alpha_0 + \alpha_1 Z_{t-h-1}^2} = 0.\end{aligned}$$

This is also a consequence of the independence assumption between  $\nu_t$  and  $Z_{t-1}$ , respectively between  $\nu_{t-h}$  and  $Z_{t-h-1}$ .

(iii) The variance of  $Z_t$  is:

$$\begin{aligned}\mathbb{V}Z_t &= \mathbb{E}Z_t^2 = \mathbb{E}\nu_t^2 (\alpha_0 + \alpha_1 Z_{t-1}^2) \\ &= \mathbb{E}\nu_t^2 \mathbb{E} (\alpha_0 + \alpha_1 Z_{t-1}^2) = \frac{\alpha_0}{1 - \alpha_1} < \infty.\end{aligned}$$

This follows from the independence assumption between  $\nu_t$  and  $Z_{t-1}$  and from the stationarity of  $\{Z_t\}$ . Because  $\alpha_0 > 0$  and  $0 < \alpha_1 < 1$ , the variance is always strictly positive and finite.

(iv) As  $\nu_t$  is normally distributed, its skewness,  $\mathbb{E}\nu_t^3$ , equals zero. The independence assumption between  $\nu_t$  and  $Z_{t-1}^2$  then implies that the skewness of  $Z_t$  is also zero, i.e.

$$\mathbb{E}Z_t^3 = 0.$$

$Z_t$  therefore has a symmetric distribution.

The properties (i), (ii) and (iii) show that  $\{Z_t\}$  is a white noise process. According to Theorem 8.1 it is not only stationary but even strictly stationary. Thus  $\{Z_t\}$  is *uncorrelated* with  $Z_{t-1}, Z_{t-2}, \dots$ , but not *independent* from its past! In particular we have:

$$\begin{aligned}\mathbb{E}(Z_t | Z_{t-1}, Z_{t-2}, \dots) &= \mathbb{E}_t \nu_t \sqrt{\alpha_0 + \alpha_1 Z_{t-1}^2} = 0 \\ \mathbb{V}(Z_t | Z_{t-1}, Z_{t-2}, \dots) &= \mathbb{E}(Z_t^2 | Z_{t-1}, Z_{t-2}, \dots) \\ &= \mathbb{E}_t \nu_t^2 (\alpha_0 + \alpha_1 Z_{t-1}^2) = \alpha_0 + \alpha_1 Z_{t-1}^2.\end{aligned}$$

The conditional variance of  $Z_t$  therefore depends on  $Z_{t-1}$ . Note that this dependence is positive because  $\alpha_1 > 0$ .

In order to guarantee that this conditional variance is always positive, we must postulate that  $\alpha_0 > 0$  and  $\alpha_1 > 0$ . The stability of the difference equation requires in addition that  $\alpha_1 < 1$ .<sup>3</sup> Thus high volatility in the

<sup>3</sup>The case  $\alpha_1 = 1$  is treated in Section 8.1.4.

past, a large realization of  $Z_{t-1}$ , is followed by high volatility in the future. The precision of the forecast, measured by the conditional variance of the forecast error, thus depends on the history of the process. This feature is not compatible with linear models and thus underlines the *non-linear* character of the ARCH model and its generalizations.

Despite the fact that  $\nu_t$  was assumed to be normally distributed,  $Z_t$  is not normally distributed. Its distribution deviates from the normal distribution in that extreme realizations are more probable. This property is called the *heavy-tail* property. In particular we have:<sup>4</sup>

$$\begin{aligned}\mathbb{E}Z_t^4 &= \mathbb{E}\nu_t^4 (\alpha_0 + \alpha_1 Z_{t-1}^2)^2 = \mathbb{E}\nu_t^4 (\alpha_0^2 + 2\alpha_0\alpha_1 Z_{t-1}^2 + \alpha_1^2 Z_{t-1}^4) \\ &= 3\alpha_0^2 + \frac{6\alpha_0^2\alpha_1}{1-\alpha_1} + 3\alpha_1^2\mathbb{E}Z_{t-1}^4.\end{aligned}$$

The strict stationarity of  $\{Z_t\}$  implies  $\mathbb{E}Z_t^4 = \mathbb{E}Z_{t-1}^4$  so that

$$\begin{aligned}(1 - 3\alpha_1^2)\mathbb{E}Z_t^4 &= \frac{3\alpha_0^2(1 + \alpha_1)}{1 - \alpha_1} \implies \\ \mathbb{E}Z_t^4 &= \frac{1}{1 - 3\alpha_1^2} \times \frac{3\alpha_0^2(1 + \alpha_1)}{1 - \alpha_1}.\end{aligned}$$

$\mathbb{E}Z_t^4$  is therefore positive and finite if and only if  $3\alpha_1^2 < 1$ , respectively if  $0 < \alpha_1 < 1/\sqrt{3} = 0.5774$ . For high correlation of the conditional variance, i.e. high  $\alpha_1 > 1/\sqrt{3}$ , the fourth moment and therefore also all higher even moments will no longer exist. The kurtosis  $\kappa$  is

$$\kappa = \frac{\mathbb{E}Z_t^4}{[\mathbb{E}Z_t^2]^2} = 3 \times \frac{1 - \alpha_1^2}{1 - 3\alpha_1^2} > 3,$$

if  $\mathbb{E}Z_t^4$  exists. The heavy-tail property manifests itself by a kurtosis greater than 3 which is the kurtosis of the normal distribution. The distribution of  $Z_t$  is therefore leptokurtic and thus more vaulted than the normal distribution.

Finally, we want to examine the autocorrelation function of  $Z_t^2$ . This will lead to a test for ARCH effects, i.e. for time varying volatility (see Section 8.2 below).

**Theorem 8.2.** *Assuming that  $\mathbb{E}Z_t^4$  exists,  $Y_t = \frac{Z_t^2}{\alpha_0}$  has the same autocorrelation function as the AR(1) process  $W_t = \alpha_1 W_{t-1} + U_t$  with  $U_t \sim \text{WN}(0, 1)$ . In addition, under the assumption  $0 < \alpha_1 < 1$ , the process  $\{W_t\}$  is also causal with respect to  $\{U_t\}$ .*

<sup>4</sup>As  $\nu_t \sim \text{N}(0, 1)$  its even moments,  $m_{2k} = \mathbb{E}\nu_t^{2k}$ ,  $k = 1, 2, \dots$ , are given by  $m_{2k} = \prod_{j=1}^k (2j-1)$ . Thus we get  $m_4 = 3$ ,  $m_6 = 15$ , etc. As the normal distribution is symmetric, all odd moments are equal to zero.

*Proof.* From  $Y_t = \nu_t^2(1 + \alpha_1 Y_{t-1})$  we get:

$$\begin{aligned}
\gamma_Y(h) &= \mathbb{E}Y_t Y_{t-h} - \mathbb{E}Y_t \mathbb{E}Y_{t-h} = \mathbb{E}Y_t Y_{t-h} - \frac{1}{(1 - \alpha_1)^2} \\
&= \mathbb{E}\nu_t^2 (1 + \alpha_1 Y_{t-1}) Y_{t-h} - \frac{1}{(1 - \alpha_1)^2} \\
&= \mathbb{E}Y_{t-h} + \alpha_1 \mathbb{E}Y_{t-1} Y_{t-h} - \frac{1}{(1 - \alpha_1)^2} \\
&= \frac{1}{1 - \alpha_1} + \alpha_1 \left( \gamma_Y(h-1) + \frac{1}{(1 - \alpha_1)^2} \right) - \frac{1}{(1 - \alpha_1)^2} \\
&= \alpha_1 \gamma_Y(h-1) + \frac{1 - \alpha_1 + \alpha_1 - 1}{(1 - \alpha_1)^2} = \alpha_1 \gamma_Y(h-1).
\end{aligned}$$

Therefore,  $\gamma_Y(h) = \alpha_1^h \gamma_Y(0) \Rightarrow \rho(h) = \alpha_1^h$ .  $\square$

The unconditional variance of  $X_t$  is:

$$\mathbb{V}X_t = \mathbb{V} \left( \frac{c}{1 - \phi} + \sum_{j=0}^{\infty} \phi^j Z_{t-j} \right) = \frac{1}{1 - \phi^2} \mathbb{V}Z_t = \frac{\alpha_0}{1 - \alpha_1} \times \frac{1}{1 - \phi^2}.$$

The unconditional variance of  $X_t$  involves all parameters of the model. Thus modeling the variance of  $X_t$  induces a trade-off between  $\phi$ ,  $\alpha_0$  and  $\alpha_1$ .

Figure 8.1 plots the realizations of two AR(1)-ARCH(1) processes. Both processes have been generated with the same realization of  $\{\nu_t\}$  and the same parameters  $\phi = 0.9$  and  $\alpha_0 = 1$ . Whereas the first process (shown on the left panel of the figure) was generated with a value of  $\alpha_1 = 0.9$ , the second one had a value of  $\alpha_1 = 0.5$ . In both cases the stability condition,  $\alpha_1 < 1$ , is fulfilled, but for the first process  $3\alpha_1^2 > 1$ , so that the fourth moment does not exist. One can clearly discern the large fluctuations, in particular for the first process.

### 8.1.3 General Models of Volatility

The simple ARCH(1) model can be and has been generalized in several directions. A straightforward generalization proposed by Engle (1982) consists by allowing further lags to enter the ARCH equation (8.1). This leads to the ARCH(p) model:

$$\text{ARCH}(p) : \quad Z_t = \nu_t \sigma_t \quad \text{with } \sigma_t^2 = \alpha_0 + \sum_{j=1}^p \alpha_j Z_{t-j}^2 \quad (8.4)$$

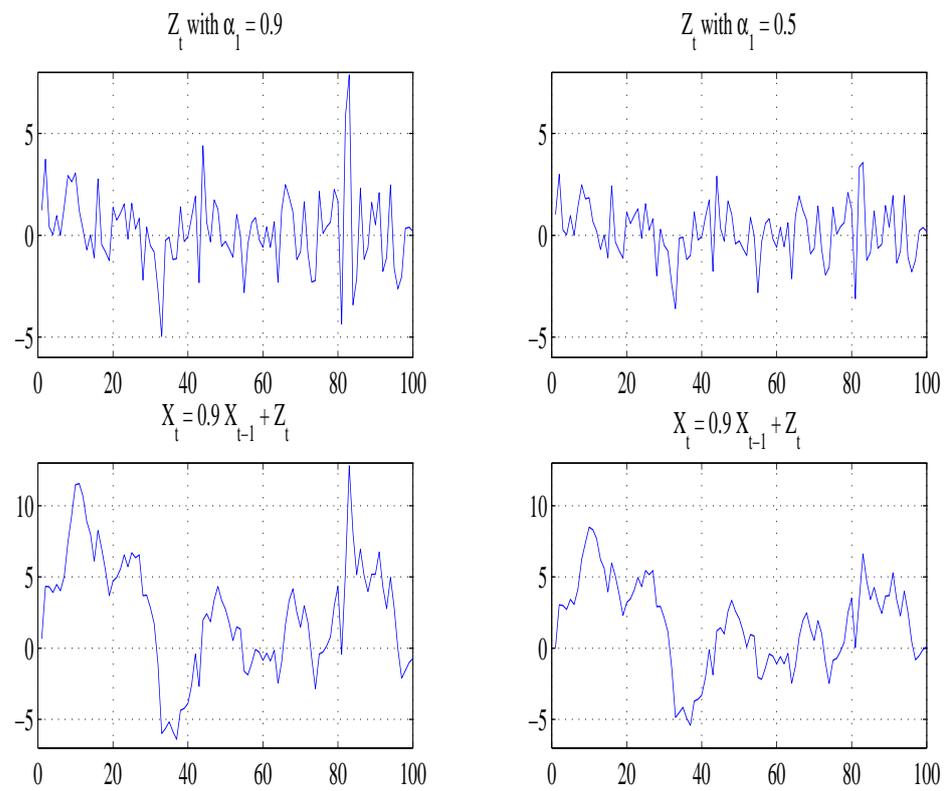


Figure 8.1: Simulation of two ARCH(1) processes ( $\alpha_1 = 0.9$  and  $\alpha_1 = 0.5$ )

where  $\alpha_0 \geq 0$ ,  $\alpha_j \geq 0$  and  $\nu_t \sim \text{IID } N(0, 1)$  with  $\nu_t$  independent from  $Z_{t-j}$ ,  $j \geq 1$ . A further popular generalization was proposed by Bollerslev (1986):

$$\text{GARCH}(p, q) : \quad Z_t = \nu_t \sigma_t \quad \text{with } \sigma_t^2 = \alpha_0 + \sum_{j=1}^p \alpha_j Z_{t-j}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \quad (8.5)$$

where we assume  $\alpha_0 \geq 0$ ,  $\alpha_j \geq 0$ ,  $\beta_j \geq 0$  and  $\nu_t \sim \text{IID } N(0, 1)$  with  $\nu_t$  independent from  $Z_{t-j}$ ,  $j \geq 1$ , as before. This model is analogous the ordinary ARMA model and allows for a parsimonious specification of the volatility process. All coefficients should be positive to guarantee that the variance is always positive. In addition it can be shown (see for example Fan and Yao (2003, 150) and the literature cited therein) that  $\{Z_t\}$  is (strictly) stationary with finite variance if and only if  $\sum_{j=1}^p \alpha_j + \sum_{j=1}^q \beta_j < 1$ .<sup>5</sup> Under this condition  $\{Z_t\} \sim \text{WN}(0, \sigma_Z^2)$  with

$$\sigma_Z^2 = \mathbb{V}(Z_t) = \frac{\alpha_0}{1 - \sum_{j=1}^p \alpha_j - \sum_{j=1}^q \beta_j}.$$

As  $\nu_t$  is still normally distributed the uneven moments of the distribution of  $Z_t$  are zero and the distribution is thus symmetric. The fourth moment of  $Z_t$ ,  $\mathbb{E}Z_t^4$ , exists if

$$\sqrt{3} \frac{\sum_{j=1}^p \alpha_j}{1 - \sum_{j=1}^q \beta_j} < 1.$$

This condition is sufficient, but not necessary.<sup>6</sup> Furthermore,  $\{Z_t\}$  is a white noise process with heavy-tail property if  $\{Z_t\}$  is strictly stationary with finite fourth moment.

In addition,  $\{Z_t^2\}$  is a causal and invertible ARMA( $\max\{p, q\}, q$ ) process satisfying the following difference equation:

$$\begin{aligned} Z_t^2 &= \alpha_0 + \sum_{j=1}^p \alpha_j Z_{t-j}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + e_t \\ &= \alpha_0 + \sum_{j=1}^{\max\{p, q\}} (\alpha_j + \beta_j) Z_{t-j}^2 + e_t - \sum_{j=1}^q \beta_j e_{t-j}, \end{aligned}$$

<sup>5</sup>A detailed exposition of the GARCH(1,1) model is given in Section 8.1.4.

<sup>6</sup>Zadrozny (2005) derives a necessary and sufficient condition for the existence of the fourth moment.

where  $\alpha_{p+j} = \beta_{q+j} = 0$  for  $j \geq 1$  and “error term”

$$e_t = Z_t^2 - \sigma_t^2 = (\nu_t^2 - 1) \left( \alpha_0 + \sum_{j=1}^p \alpha_j Z_{t-j}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \right).$$

Note, however, there is a circularity here because the noise process  $\{e_t\}$  is defined in terms of  $Z_t^2$  and is therefore not an exogenous process driving  $Z_t^2$ . Thus, one has to be precautious in the interpretation of  $\{Z_t^2\}$  as an ARMA process.

Further generalizations of the GARCH(p,q) model can be obtained by allowing deviations from the normal distribution for  $\nu_t$ . In particular, distributions such as the t-distribution which put more weight on extreme values have become popular. This seems warranted as prices on financial markets exhibit large and sudden fluctuations.<sup>7</sup>

### Threshold GARCH Model

Assuming a symmetric distribution for  $\nu_t$  and specifying a linear relationship between  $\sigma_t^2$  and  $Z_{t-j}^2$  bzw.  $\sigma_{t-j}^2$ ,  $j > 0$ , leads to a symmetric distribution for  $Z_t$ . It has, however, been observed that downward movements seem to be different from upward movements. This asymmetric behavior is accounted for by the asymmetric GARCH(1,1) model or threshold GARCH(1,1) model (TGARCH(1,1) model). This model was proposed by Glosten et al. (1993) and Zakoian (1994):

$$\begin{aligned} \text{asymmetric GARCH}(1, 1) : Z_t = \nu_t \sigma_t \quad \text{with} \\ \sigma_t^2 = \alpha_0 + \alpha_1 Z_{t-1}^2 + \beta \sigma_{t-1}^2 \\ + \gamma \mathbf{1}_{\{Z_{t-1} < 0\}} Z_{t-1}^2. \end{aligned}$$

$\mathbf{1}_{\{Z_{t-1} < 0\}}$  denotes the indicator function which takes on the value one if  $Z_{t-1}$  is negative and the value zero otherwise. Assuming, as before, that all parameters  $\alpha_0$ ,  $\alpha_1$ ,  $\beta$  and  $\gamma$  are greater than zero, this specification postulates a leverage effect because negative realizations have a greater impact than positive ones. In order to obtain a stationary process the condition  $\alpha_1 + \beta + \gamma/2 < 1$  must hold. This model can be generalized in an obvious way by allowing additional lags  $Z_{t-j}^2$  and  $\sigma_{t-j}^2$ ,  $j > 1$  to enter the above specification.

---

<sup>7</sup>A thorough treatment of the probabilistic properties of GARCH processes can be found in Nelson (1990), Bougerol and Picard (1992), Giraitis et al. (2000), and Klüppelberg et al. (2004, theorem 2.1).

### The Exponential GARCH model

Another interesting and popular class of volatility models was introduced by Nelson (1991). The so-called exponential GARCH models or EGARCH models are defined as follows:

$$\begin{aligned}\log \sigma_t^2 &= \alpha_0 + \beta \log \sigma_{t-1}^2 + \gamma \left| \frac{Z_{t-1}}{\sigma_{t-1}} \right| + \delta \frac{Z_{t-1}}{\sigma_{t-1}} \\ &= \alpha_0 + \beta \log \sigma_{t-1}^2 + \gamma |\nu_{t-1}| + \delta \nu_{t-1}.\end{aligned}$$

Note that, in contrast to the previous specifications, the dependent variable is the logarithm of  $\sigma_t^2$  and not  $\sigma_t^2$  itself. This has the advantage that the variance is always positive irrespective of the values of the coefficients. Furthermore, the leverage effect is exponential rather than quadratic because  $Z_t = \nu_t \exp(\sigma_t/2)$ . The EGARCH model is also less recursive than the GARCH model as the volatility is specified directly in terms of the noise process  $\{\nu_t\}$ . Thus, the above EGARCH model can be treated as an AR(1) model of  $\log \sigma_t^2$  with noise process  $\gamma |\nu_{t-1}| + \delta \nu_{t-1}$ . It is obvious that the model can be generalized to allow for additional lags both in  $\sigma_t^2$  and  $\nu_t$ . This results in an ARMA process for  $\{\log \sigma_t^2\}$  for which the usual conditions for the existence of a causal and invertible solution can be applied (see Section 2.3). A detailed analysis and further properties of this model class can be found in Bollerslev et al. (1994), Gouriéroux (1997) and Fan and Yao (2003, 143-180).

### The ARCH-in-Mean Model

The ARCH-in-mean model or ARCH-M model was introduced by Engle et al. (1987) to allow for a feedback of volatility into the mean equation. More specifically, assume for the sake of simplicity that the variance equation is just represented the ARCH(1) model

$$Z_t = \nu_t \sigma_t \quad \text{with } \sigma_t^2 = \alpha_0 + \alpha_1 Z_{t-1}^2.$$

Then, the ARCH-M model is given

$$X_t = M_t \beta + g(\sigma_t^2) + Z_t \tag{8.6}$$

where  $g$  is a function of the volatility  $\sigma_t^2$  and where  $M_t'$  consists of a vector of regressors, including lagged values of  $X_t$ . If  $M_t = (1, X_{t-1})$  then we get the AR(1)-ARCH-M model. The most commonly used specification for  $g$  is a linear function:  $g(\sigma_t^2) = \delta_0 + \delta_1 \sigma_t^2$ . In the asset pricing literature, higher volatility would require a higher return to compensate the investor for the additional risk. Thus, if  $X_t$  denotes the return on some asset, we expect  $\delta_1$

to be positive. Note that any time variation in  $\sigma_t^2$  translates into a serial correlation of  $\{X_t\}$  (see Hong, 1991, for details). Of course, one could easily generalize the model to allow for more sophisticated mean and variance equations.

### 8.1.4 The GARCH(1,1) Model

The Generalized Autoregressive Conditional Heteroskedasticity model of order (1,1), GARCH(1,1) model for short, is considered as a benchmark for more general specifications and often serves as a starting point for further empirical investigations. We therefore want to explore its properties in more detail. Many of its properties generalize in a straightforward way to the GARCH(p,q) process. According to equation (8.5) the GARCH(1,1) model is defined as:

$$\text{GARCH}(1,1) : Z_t = \nu_t \sigma_t \quad \text{with } \sigma_t^2 = \alpha_0 + \alpha_1 Z_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (8.7)$$

where  $\alpha_0, \alpha_1$ , and  $\beta \geq 0$ . We assume  $\alpha_1 + \beta > 0$  to avoid the degenerate case  $\alpha_1 = \beta = 0$  which implies that  $\{Z_t\}$  is just a sequence of IID random variables. Moreover,  $\nu_t \sim \text{IID}(0,1)$  with  $\nu_t$  being independent of  $Z_{t-j}$ ,  $j \geq 1$ . Note that we do not make further distributional assumption. In particular,  $\nu_t$  need not required to be normally distributed. For this model, we can formulate a similar theorem as for the ARCH(1) model (see Theorem: 8.1):

**Theorem 8.3.** *Let  $\{Z_t\}$  be a GARCH(1,1) process as defined above. Under the assumption*

$$\mathbb{E} \log(\alpha_1 \nu_t^2 + \beta) < 0,$$

*the difference equation (8.7) possess strictly stationary solution:*

$$Z_t = \nu_t \sqrt{\alpha_0 \sum_{j=0}^{\infty} \prod_{i=1}^j (\alpha_1 \nu_{t-i}^2 + \beta)}$$

*where  $\prod_i^j = 1$  whenever  $i > j$ . The solution is also unique given the sequence  $\{\nu_t\}$ . The solution is unique and (weakly) stationary with variance  $\mathbb{E} Z_t^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta} < \infty$  if  $\alpha_1 + \beta < 1$ .*

*Proof.* The proof proceeds similarly to Theorem 8.1. For this purpose, we define  $Y_t = \sigma_t^2$  and rewrite the GARCH(1,1) model as

$$Y_t = \alpha_0 + \alpha_1 \nu_{t-1}^2 Y_{t-1} + \beta Y_{t-1} = \alpha_0 + (\alpha_1 \nu_{t-1}^2 + \beta) Y_{t-1}.$$

This defines an AR(1) process with time-varying coefficients  $\xi_t = \alpha_1 \nu_t^2 + \beta \geq 0$ . Iterate this equation backwards  $k$  times to obtain:

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_0 \xi_{t-1} + \dots + \alpha_0 \xi_{t-1} \dots \xi_{t-k} + \xi_{t-1} \dots \xi_{t-k} \xi_{t-k-1} Y_{t-k-1} \\ &= \alpha_0 \sum_{j=0}^k \prod_{i=1}^j \xi_{t-i} + \prod_{i=1}^{k+1} \xi_{t-i} Y_{t-k-1}. \end{aligned}$$

Taking the limit  $k \rightarrow \infty$ , we define the process  $\{Y'_t\}$

$$Y'_t = \alpha_0 \sum_{j=0}^{\infty} \prod_{i=1}^j \xi_{t-i}. \quad (8.8)$$

The right-hand side of this expression converges almost surely as can be seen from the following argument. Given that  $\nu_t \sim \text{IID}$  and given the assumption  $\mathbb{E} \log(\alpha_1 \nu_t^2 + \beta) < 0$ , the strong law of large numbers (Theorem C.5) implies that

$$\limsup_{j \rightarrow \infty} \frac{1}{j} \left( \sum_{i=1}^j \log(\xi_{t-i}) \right) < 0 \quad \text{a.s.},$$

or equivalently,

$$\limsup_{j \rightarrow \infty} \log \left( \prod_{i=1}^j \xi_{t-i} \right)^{1/j} < 0 \quad \text{a.s.}$$

Thus,

$$\limsup_{j \rightarrow \infty} \left( \prod_{i=1}^j \xi_{t-i} \right)^{1/j} < 1 \quad \text{a.s.}$$

The application of the root test then shows that the infinite series (8.8) converges almost surely. Thus,  $\{Y'_t\}$  is well-defined. It is easy to see that  $\{Y'_t\}$  is strictly stationary and satisfies the difference equation. Moreover, if  $\alpha_1 + \beta < 1$ , we get

$$\mathbb{E} Y'_t = \alpha_0 \sum_{j=0}^{\infty} \mathbb{E} \prod_{i=1}^j \xi_{t-i} = \alpha_0 \sum_{j=0}^{\infty} (\alpha_1 + \beta)^j = \frac{\alpha_0}{1 - \alpha_1 - \beta}.$$

Thus,  $\mathbb{E} Z_t^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta} < \infty$ .

To show uniqueness, we assume that there exists another strictly stationary process  $\{Y_t\}$  which also satisfies the difference equation. This implies that

$$\begin{aligned} |Y_t - Y'_t| &= |\xi_{t-1}| |Y_{t-1} - Y'_{t-1}| = \left( \prod_{i=1}^k \xi_{t-i} \right) |Y_{t-k} - Y'_{t-k}| \\ &\leq \left( \prod_{i=1}^k \xi_{t-i} \right) |Y_{t-k}| + \left( \prod_{i=1}^k \xi_{t-i} \right) |Y'_{t-k}| \end{aligned}$$

The assumption  $\mathbb{E} \log \xi_t = \mathbb{E} \log(\alpha_1 \nu_t^2 + \beta) < 0$  together with the strong law of large numbers (Theorem C.5) imply

$$\prod_{i=1}^k \xi_{t-i} = \left( \exp \left( \frac{1}{k} \sum_{i=1}^k \log \xi_{t-i} \right) \right)^k \longrightarrow 0 \quad \text{a.s.}$$

As both solutions are strictly stationary so that the distribution of  $|Y_{t-k}|$  and  $|Y'_{t-k}|$  do not depend on  $t$ , this implies that both  $\left( \prod_{i=1}^k \xi_{t-i} \right) |Y_{t-k}|$  and  $\left( \prod_{i=1}^k \xi_{t-i} \right) |Y'_{t-k}|$  converge in probability to zero. Thus,  $Y_t = Y'_t$  a.s. once the sequence  $\xi_t$ , respectively  $\nu_t$ , is given. Because  $Z_t = \nu_t \sqrt{Y'_t}$  this completes the proof.  $\square$

**Remark 8.2.** *Using Jensen's inequality, we see that*

$$\mathbb{E} \log(\alpha_1 \nu_t^2 + \beta) \leq \log \mathbb{E}(\alpha_1 \nu_t^2 + \beta) = \log(\alpha_1 + \beta).$$

*Thus, the condition  $\alpha_1 + \beta < 1$  is sufficient, but not necessary, to ensure the existence of a strictly stationary solution. Thus even when  $\alpha_1 + \beta = 1$ , a strictly stationary solution exists, albeit one with infinite variance. This case is known as the IGARCH model and is discussed below. In the case  $\alpha_1 + \beta < 1$ , the Borel-Cantelli lemma can be used as Theorem 8.1 to establish the uniqueness of the solution. Further details can be found in the references listed in footnote 7.*

Assume that  $\alpha_1 + \beta < 1$ , then a unique strictly stationary process  $\{Z_t\}$  with finite variance which satisfies the above difference equation exists. In particular  $Z_t \sim \text{WN}(0, \sigma_Z^2)$  such that

$$\mathbb{V}(Z_t) = \frac{\alpha_0}{1 - \alpha_1 - \beta}.$$

The assumption  $1 - \alpha_1 - \beta > 0$  guarantees that the variance exists. The third moment of  $Z_t$  is zero due to the assumption of a symmetric distribution for

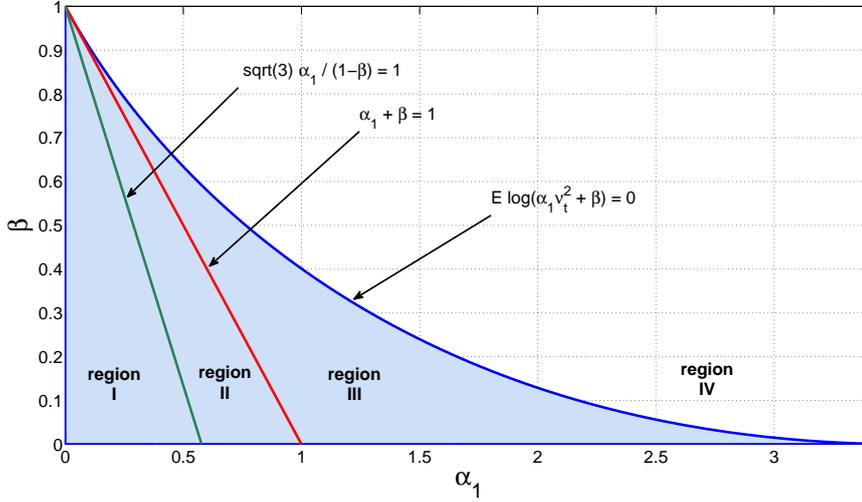


Figure 8.2: Parameter region for which a strictly stationary solution to the GARCH(1,1) process exists assuming  $\nu_t \sim \text{IID } N(0, 1)$

$\nu_t$ . The condition for the existence of the fourth moment is:  $\sqrt{3} \frac{\alpha_1}{1-\beta} < 1$ .<sup>8</sup> The kurtosis is then

$$\kappa = \frac{\mathbb{E}Z_t^4}{[\mathbb{E}Z_t^2]^2} = 3 \times \frac{1 - (\alpha_1 + \beta)^2}{1 - (\alpha_1 + \beta)^2 - 2\alpha_1^2} > 3$$

if  $\mathbb{E}Z_t^4$  exists.<sup>9</sup> Therefore the GARCH(1,1) model also possesses the heavy-tail property because  $Z_t$  is more peaked than the normal distribution.

Figure 8.2 shows how the different assumptions and conditions divide up the parameter space. In region I, all conditions are fulfilled. The process has a strictly stationary solution with finite variance and kurtosis. In region II, the kurtosis does no longer exist, but the variance does as  $\alpha_1 + \beta < 1$  still holds. In region III, the process has infinite variance, but a strictly stationary solution yet exists. In region IV, no such solution exists.

Viewing the equation for  $\sigma_t^2$  as a stochastic difference equation, its solution is given by

$$\sigma_t^2 = \frac{\alpha_0}{1 - \beta} + \alpha_1 \sum_{j=0}^{\infty} \beta^j Z_{t-1-j}^2. \tag{8.9}$$

<sup>8</sup>A necessary and sufficient condition is  $(\alpha_1 + \beta)^2 + 2\alpha_1^2 < 1$  (see Zdrozny (2005)).

<sup>9</sup>The condition for the existence of the fourth moment implies  $3\alpha_1^2 < (1 - \beta)^2$  so that the denominator  $1 - \beta^2 - 2\alpha_1\beta - 3\alpha_1^2 > 1 - \beta^2 - 2\alpha_1\beta - 1 - \beta^2 + 2\beta = 2\beta(1 - \alpha_1 - \beta) > 0$ .

This expression is well-defined because  $0 < \beta < 1$  so that the infinite sum converges. The conditional variance given the infinite past is therefore equal to

$$\mathbb{V}(Z_t | Z_{t-1}, Z_{t-2}, \dots) = \mathbb{E}(Z_t^2 | Z_{t-1}, Z_{t-2}, \dots) = \frac{\alpha_0}{1 - \beta} + \alpha_1 \sum_{j=0}^{\infty} \beta^j Z_{t-1-j}^2.$$

Thus, the conditional variance depends on the entire history of the time series and not just on  $Z_{t-1}$  as in the case of the ARCH(1) model. As all coefficients are assumed to be positive, the clustering of volatility is more persistent than for the ARCH(1) model.

Defining a new time series  $\{e_t\}$  by  $e_t = Z_t^2 - \sigma_t^2 = (\nu_t^2 - 1)(\alpha_0 + \alpha_1 Z_{t-1}^2 + \beta \sigma_{t-1}^2)$ , one can verify that  $Z_t^2$  obeys the stochastic difference equation

$$\begin{aligned} Z_t^2 &= \alpha_0 + \alpha_1 Z_{t-1}^2 + \beta \sigma_{t-1}^2 + e_t = \alpha_0 + \alpha_1 Z_{t-1}^2 + \beta(Z_{t-1}^2 - e_{t-1}) + e_t \\ &= \alpha_0 + (\alpha_1 + \beta)Z_{t-1}^2 + e_t - \beta e_{t-1}. \end{aligned} \quad (8.10)$$

This difference equation defines an ARMA(1,1) process if  $e_t$  has finite variance which is the case if the fourth moment of  $Z_t$  exists. In this case, it is easy to verify that  $\{e_t\}$  is white noise. The so-defined ARMA(1,1) process is causal and invertible with respect to  $\{e_t\}$  because  $0 < \alpha_1 + \beta < 1$  and  $0 < \beta < 1$ . The autocorrelation function (ACF),  $\rho_{Z^2}(h)$ , can be computed using the methods laid out Section 2.4. This gives

$$\begin{aligned} \rho_{Z^2}(1) &= \frac{(1 - \beta^2 - \alpha_1 \beta)\alpha_1}{1 - \beta^2 - 2\alpha_1 \beta} = \frac{(1 - \beta\varphi)(\varphi - \beta)}{1 + \beta^2 - 2\varphi\beta}, \\ \rho_{Z^2}(h) &= (\alpha_1 + \beta)\rho_{Z^2}(h-1) = \varphi\rho_{Z^2}(h-1), \quad h = 2, 3, \dots \end{aligned} \quad (8.11)$$

with  $\varphi = \alpha_1 + \beta$  (see also Bollerslev (1988)).

### The IGARCH Model

Practice has shown that the sum  $\alpha_1 + \beta$  is often close to one. Thus, it seems interesting to examine the limiting case where  $\alpha_1 + \beta = 1$ . This model was proposed by Engle and Bollerslev (1986) and was termed the *integrated GARCH* (IGARCH) model in analogy to the notion of integrated processes (see Chapter 7). From equation (8.10) we get

$$Z_t^2 = \alpha_0 + Z_{t-1}^2 + e_t - \beta e_{t-1}$$

with  $e_t = Z_t^2 - \sigma_t^2 = (\nu_t^2 - 1)(\alpha_0 + (1 - \beta)Z_{t-1}^2 + \beta\sigma_{t-1}^2)$ . As  $\{e_t\}$  is white noise, the squared innovations  $Z_t^2$  behave like a random walk with a MA(1) error

term. Although the variance of  $Z_t$  becomes infinite, the difference equation still allows for a strictly stationary solution provided that  $\mathbb{E} \log(\alpha_1 \nu_t^2 + \beta) < 0$  (see Theorem 8.3 and the citations in footnote 7 for further details).<sup>10</sup> It has been shown by Lumsdaine (1986) and Lee and Hansen (1994) that standard inferences can still be applied although  $\alpha_1 + \beta = 1$ . The model may easily be generalized to higher lag orders.

### Forecasting

On many occasions it is necessary to obtain forecasts of the conditional variance  $\sigma_t^2$ . An example is given in Section 8.4 where the value at risk (VaR) of a portfolio several periods ahead must be evaluated. Denote by  $\mathbb{P}_t \sigma_{t+h}^2$  the  $h$  period ahead forecast based on information available in period  $t$ . We assume that predictions are based on the infinite past. Then the one-period ahead forecast based on  $Z_t$  and  $\sigma_t^2$ , respectively  $\nu_t$  and  $\sigma_t^2$ , is:

$$\mathbb{P}_t \sigma_{t+1}^2 = \alpha_0 + \alpha_1 Z_t^2 + \beta \sigma_t^2 = \alpha_0 + (\alpha_1 \nu_t^2 + \beta) \sigma_t^2.$$

As  $\nu_t \sim \text{IID}(0, 1)$  and independent of  $Z_{t-j}$ ,  $j \geq 1$ ,

$$\mathbb{P}_t \sigma_{t+2}^2 = \alpha_0 + (\alpha_1 + \beta) \mathbb{P}_t \sigma_{t+1}^2.$$

Thus, forecast for  $h \geq 2$  can be obtained recursively as follows:

$$\begin{aligned} \mathbb{P}_t \sigma_{t+h}^2 &= \alpha_0 + (\alpha_1 + \beta) \mathbb{P}_t \sigma_{t+h-1}^2 \\ &= \alpha_0 \sum_{j=0}^{h-2} (\alpha_1 + \beta)^j + (\alpha_1 + \beta)^{h-1} \mathbb{P}_t \sigma_{t+1}^2. \end{aligned}$$

Assuming  $\alpha_1 + \beta < 1$ , the second term in the above expression vanishes as  $h$  goes to infinity. Thus, the contribution of the current conditional variance vanishes when we look further and further into the future. The forecast of the conditional variance then approaches the unconditional one:  $\lim_{h \rightarrow \infty} \mathbb{P}_t \sigma_{t+h}^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta}$ . If  $\alpha_1 + \beta = 1$  as in the IGARCH model, the contribution of the current conditional variance is constant, but diminishes to zero relative to the first term. Finally, if  $\alpha_1 + \beta > 1$ , the two terms are of the same order and we have a particularly persistent situation.

In practice, the parameters of the model are unknown and have therefore to be replaced by an estimate. The method can be easily adapted for higher order models. Instead of using the recursive approach outlined above, it is

---

<sup>10</sup>As the variance becomes infinite, the IGARCH process is an example of a stochastic process which is strictly stationary, but not stationary.

possible to use simulation methods by drawing repeatedly from the actual empirical distribution of the  $\hat{\nu}_t = \hat{Z}_t/\hat{\sigma}_t$ . This has the advantage to capture deviations from the underlying distributional assumptions (see Section 8.4 for a comparison of both methods). Such methods must be applied if nonlinear models for the conditional variance, like the TARARCH model, are employed.

## 8.2 Tests for Heteroskedasticity

Before modeling the volatility of a time series it is advisable to test whether heteroskedasticity is actually present in the data. For this purpose the literature proposed several tests of which we are going to examine two. For both tests the null hypothesis is that there is no heteroskedasticity i.e. that there are no ARCH effects. These tests can also be useful in a conventional regression setting.

### 8.2.1 Autocorrelation of quadratic residuals

The first test is based on the autocorrelation function of squared residuals from a preliminary regression. This preliminary regression or mean regression produces a series  $\hat{Z}_t$  which should be approximately white noise if the equation is well specified. Then we can look at the ACF of the squared residuals  $\{\hat{Z}_t^2\}$  and apply the Ljung-Box test (see equation (4.4)). Thus the test can be broken down into three steps.

- (i) Estimate an ARMA model for  $\{X_t\}$  and retrieve the residuals  $\hat{Z}_t$  from this model. Compute  $\hat{Z}_t^2$ . These data can be used to estimate  $\sigma^2$  as

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{Z}_t^2$$

Note that the ARMA model should be specified such that the residuals are approximately white noise.

- (ii) Estimate the ACF for the squared residuals in the usual way:

$$\hat{\rho}_{Z^2}(h) = \frac{\sum_{t=h+1}^T (\hat{Z}_t^2 - \hat{\sigma}^2) (\hat{Z}_{t-h}^2 - \hat{\sigma}^2)}{\sum_{t=1}^T (\hat{Z}_t^2 - \hat{\sigma}^2)^2}$$

- (iii) It is now possible to use one of the methods laid out in Chapter 4 to test the null hypothesis that  $\{\hat{Z}_t^2\}$  is white noise. It can be shown that

under the null hypothesis  $\sqrt{T}\widehat{\rho}_{Z^2}(h) \xrightarrow{d} N(0, 1)$ . One can therefore construct confidence intervals for the ACF in the usual way. Alternatively, one may use the Ljung-Box test statistic (see equation (4.4)) to test the hypothesis that all correlation coefficients up to order  $N$  are simultaneously equal to zero.

$$Q' = T(T + 2) \sum_{h=1}^N \frac{\widehat{\rho}_{Z^2}^2(h)}{T - h}.$$

Under the null hypothesis this statistic is distributed as  $\chi_N^2$ . To carry out the test,  $N$  should be chosen rather high, for example equal to  $T/4$ .

### 8.2.2 Engle's Lagrange-multiplier

Engle (1982) proposed a Lagrange-Multiplier test. This test rests on an ancillary regression of the squared residuals against a constant and lagged values of  $\widehat{Z}_{t-1}^2, \widehat{Z}_{t-2}^2, \dots, \widehat{Z}_{t-p}^2$  where the  $\{\widehat{Z}_t\}$  is again obtained from a preliminary regression. The auxiliary regression thus is

$$\widehat{Z}_t^2 = \alpha_0 + \alpha_1 \widehat{Z}_{t-1}^2 + \alpha_2 \widehat{Z}_{t-2}^2 + \dots + \alpha_p \widehat{Z}_{t-p}^2 + \varepsilon_t,$$

where  $\varepsilon_t$  denotes the error term. Then the null hypothesis  $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$  is tested against the alternative hypothesis  $H_1 : \alpha_j \neq 0$  for at least one  $j$ . As a test statistic one can use the coefficient of determination times  $T$ , i.e.  $TR^2$ . This test statistic is distributed as a  $\chi^2$  with  $p$  degrees of freedom. Alternatively, one may use the conventional F-test.

## 8.3 Estimation of GARCH(p,q) Models

### 8.3.1 Maximum-Likelihood Estimation

The literature has proposed several approaches to estimate models of volatility (see Fan and Yao (2003, 156-162)). The most popular one, however, rest on the method of maximum-likelihood. We will describe this method using the GARCH(p,q) model. Related and more detailed accounts can be found in Weiss (1986), Bollerslev et al. (1994) and Hall and Yao (2003).

In particular we consider the following model:

$$\text{mean equation:} \quad X_t = c + \phi_1 X_{t-1} + \dots + \phi_r X_{t-r} + Z_t,$$

where

$$Z_t = \nu_t \sigma_t \quad \text{with } \nu_t \sim \text{IIDN}(0, 1) \text{ and}$$

$$\text{variance equation:} \quad \sigma_t^2 = \alpha_0 + \sum_{j=1}^p \alpha_j Z_{t-j}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2.$$

The mean equation represents a simple AR( $r$ ) process for which we assume that it is causal with respect to  $\{Z_t\}$ , i.e. that all roots of  $\Phi(z)$  are outside the unit circle. The method demonstrated here can be easily generalized to ARMA processes or even ARMA process with additional exogenous variables (so-called ARMAX processes) as noted by Weiss (1986). The method also incorporates the ARCH-in-mean model (see equation(8.6)) which allows for an effect of the conditional variance  $\sigma_t$  on  $X_t$ . In addition, we assume that the coefficients of the variance equation are all positive, that  $\sum_{j=1}^p \alpha_j + \sum_{j=1}^q \beta_j < 1$  and that  $\mathbb{E}Z_t^4 < \infty$  exists.<sup>11</sup>

As  $\nu_t$  is identically and independently standard normally distributed, the distribution of  $X_t$  conditional on  $\mathcal{X}_{t-1} = \{X_{t-1}, X_{t-2}, \dots\}$  is normal with mean  $c + \phi_1 X_{t-1} + \dots + \phi_r X_{t-r}$  and variance  $\sigma_t^2$ . The conditional density,  $f(X_t | \mathcal{X}_{t-1})$ , therefore is:

$$f(X_t | \mathcal{X}_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{Z_t^2}{2\sigma_t^2}\right)$$

where  $Z_t$  equals  $X_t - c - \phi_1 X_{t-1} - \dots - \phi_r X_{t-r}$  and  $\sigma_t^2$  is given by the variance equation.<sup>12</sup> The joint density  $f(X_1, X_2, \dots, X_T)$  of a random sample  $(X_1, X_2, \dots, X_T)$  can therefore be factorized as

$$f(X_1, X_2, \dots, X_T) = f(X_1, X_2, \dots, X_{s-1}) \prod_{t=s}^T f(X_t | \mathcal{X}_{t-1})$$

where  $s$  is an integer greater than  $p$ . The necessity, not to factorize the first  $s - 1$  observations, relates to the fact that  $\sigma_t^2$  can only be evaluated for  $s > p$  in the ARCH( $p$ ) model. For the ARCH( $p$ ) model  $s$  can be set to  $p + 1$ . In the case of a GARCH model  $\sigma_t^2$  is given by weighted infinite sum of the  $Z_{t-1}^2, Z_{t-2}^2, \dots$  (see the expression (8.9) for  $\sigma_t^2$  in the GARCH(1,1) model). For finite samples, this infinite sum must be approximated by a finite sum

<sup>11</sup>The existence of the fourth moment is necessary for the asymptotic normality of the maximum-likelihood estimator, but not for the consistence. It is possible to relax this assumption somewhat (see Hall and Yao (2003)).

<sup>12</sup>If  $\nu_t$  is assumed to follow another distribution than the normal, one may use this distribution instead.

of  $s$  summands such that the numbers of summands  $s$  is increasing with the sample size.(see Hall and Yao (2003)).

We then merge all parameters of the model as follows:  $\phi = (c, \phi_1, \dots, \phi_r)'$ ,  $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)'$  and  $\beta = (\beta_1, \dots, \beta_q)'$ . For a given realization  $x = (x_1, x_2, \dots, x_T)$  the *likelihood function* conditional on  $x$ ,  $L(\phi, \alpha, \beta|x)$ , is defined as

$$L(\phi, \alpha, \beta|x) = f(x_1, x_2, \dots, x_{s-1}) \prod_{t=s}^T f(x_t|\mathcal{X}_{t-1})$$

where in  $\mathcal{X}_{t-1}$  the random variables are replaced by their realizations. The likelihood function can be seen as the probability of observing the data at hand given the values for the parameters. The method of maximum likelihood then consist in choosing the parameters  $(\phi, \alpha, \beta)$  such that the likelihood function is maximized. Thus we chose the parameter so that the probability of observing the data is maximized. In this way we obtain the maximum likelihood estimator. Taking the first  $s$  realizations as given deterministic starting values, we then get the *conditional likelihood function*.

In practice we do not maximize the likelihood function but the logarithm of it where we take  $f(x_1, \dots, x_{s-1})$  as a fixed constant which can be neglected in the optimization:

$$\begin{aligned} \log L(\phi, \alpha, \beta|x) &= \sum_{t=s}^T \log f(x_t|\mathcal{X}_t) \\ &= -\frac{T}{2} \log 2\pi - \frac{1}{2} \sum_{t=s}^T \log \sigma_t^2 - \frac{1}{2} \sum_{t=s}^T \frac{z_t^2}{\sigma_t^2} \end{aligned}$$

where  $z_t = x_t - c - \phi_1 x_{t-1} - \dots - \phi_r x_{t-r}$  denotes the realization of  $Z_t$ . The maximum likelihood estimator is obtained by maximizing the likelihood function over the admissible parameter space. Usually, the implementation of the stationarity condition and the condition for the existence of the fourth moment turns out to be difficult and cumbersome so that often these conditions are neglected and only checked in retrospect or some ad hoc solutions are envisaged. It can be shown that the (conditional) maximum likelihood estimator leads to asymptotically normally distributed estimates.<sup>13</sup> The maximum likelihood estimator remains meaningful even when  $\{\nu_t\}$  is not normally distributed. In this case the quasi maximum likelihood estimator is obtained (see Hall and Yao (2003) and Fan and Yao (2003)).

For numerical reasons it is often convenient to treat the mean equation and the variance equation separately. As the mean equation is a simple

<sup>13</sup>Jensen and Rahbek (2004) showed that, at least for the GARCH(1,1) case, the stationarity condition is not necessary.

AR( $r$ ) model, it can be estimated by ordinary least-squares (OLS) in a first step. This leads to consistent parameter estimates. However, due to the heteroskedasticity, this is no longer true for the covariance matrix of the coefficients so that the usual  $t$ - and  $F$ -tests are not reliable. This problem can be circumvented by the use of the White correction (see White (1980)). In this way it is possible to find an appropriate specification for the mean equation without having to estimate the complete model. In the second step, one can then work with the residuals to find an appropriate ARMA model for the squared residuals. This leads to consistent estimates of the parameters of the variance equation. These estimates are under additional weakly assumptions asymptotically normally distributed (see Weiss (1986)). It should, however, be noted that this way of proceeding is, in contrast to the maximum likelihood estimator, not efficient because it neglects the nonlinear character of the GARCH model. The parameters found in this way can, however, serve as meaningful starting values for the numerical maximization procedure which underlies the maximum likelihood estimation.

A final remark concerns the choice of the parameter  $r$ ,  $p$  and  $q$ . Similarly to the ordinary ARMA models, one can use information criteria such as the Akaike or the Bayes criterion, to determine the order of the model (see Section 5.4).

### 8.3.2 Method of Moment Estimation

The maximization of the likelihood function requires the use of numerical optimization routines. Depending on the routine actually used and on the starting value, different results may be obtained if the likelihood function is not well-behaved. It is therefore of interest to have alternative estimation methods at hand. The method of moments is such an alternative. It is similar to the Yule-Walker estimator (see Section 5.1) applied to the autocorrelation function of  $\{Z_t^2\}$ . This method not only leads to an analytic solution, but can also be easily implemented. Following Kristensen and Linton (2006), we will illustrate the method for the GARCH(1,1) model.

Equation (8.11) applied to  $\rho_{Z^2}(1)$  and  $\rho_{Z^2}(2)$  constitutes a nonlinear equation system in the unknown parameters  $\beta$  and  $\alpha_1$ . This system can be reparameterized to yield an equation system in  $\varphi = \alpha_1 + \beta$  and  $\beta$  which can be reduced to a single quadratic equation in  $\beta$ :

$$\beta^2 - b\beta - 1 = 0 \quad \text{where } b = \frac{\varphi^2 + 1 - 2\rho_{Z^2}(1)\varphi}{\varphi - \rho_{Z^2}(1)}.$$

The parameter  $b$  is well-defined because  $\varphi = \alpha_1 + \beta \geq \rho_{Z^2}(1)$  with equality only if  $\beta = 0$ . In the following we will assume that  $\beta > 0$ . Under this

assumption  $b > 2$  so that the only solution with the property  $0 < \beta < 1$  is given by

$$\beta = \frac{b - \sqrt{b^2 - 4}}{2}.$$

The moment estimator can therefore be constructed as follows:

- (i) Estimate the correlations  $\rho_{Z^2}(1)$  and  $\rho_{Z^2}(2)$  and  $\sigma^2$  based on the formulas (8.11) in Section 8.2.
- (ii) An estimate for  $\varphi = \alpha_1 + \beta$  is then given by

$$\hat{\varphi} = \widehat{(\alpha_1 + \beta)} = \frac{\hat{\rho}_{Z^2}(2)}{\hat{\rho}_{Z^2}(1)}.$$

- (iii) use the estimate  $\hat{\varphi}$  to compute an estimate for  $b$ :

$$\hat{b} = \frac{\hat{\varphi}^2 + 1 - 2\hat{\rho}_{Z^2}(1)\hat{\varphi}}{\hat{\varphi} - \hat{\rho}_{Z^2}(1)}.$$

The estimate  $\hat{\beta}$  for  $\beta$  is then

$$\hat{\beta} = \frac{\hat{b} - \sqrt{\hat{b}^2 - 4}}{2}.$$

- (iv) The estimate for  $\alpha_1$  is  $\hat{\alpha}_1 = \hat{\varphi} - \hat{\beta}$ . Because  $\alpha_0 = \sigma^2(1 - (\alpha_1 + \beta))$ , the estimate for  $\alpha_0$  is equal to  $\hat{\alpha}_0 = \hat{\sigma}^2(1 - \hat{\varphi})$ .

Kristensen and Linton (2006) show that, given the existence of the fourth moment of  $Z_t$ , this method of moment leads to consistent and asymptotically normal distributed estimates. These estimates may then serve as starting values for the maximization of the likelihood function to improve efficiency.

## 8.4 Example: Swiss Market Index (SMI)

In this section, we will illustrate the methods discussed previously to analyze the volatility of the Swiss Market Index (SMI). The SMI is the most important stock market index for Swiss blue chip companies. It is constructed solely from stock market prices, dividends are not accounted for. The data are the daily values of the index between the 3rd of January 1989 and the 13th of February 2004. Figure 1.5 shows a plot of the data. Instead of analyzing the level of the SMI, we will investigate the daily return computed as

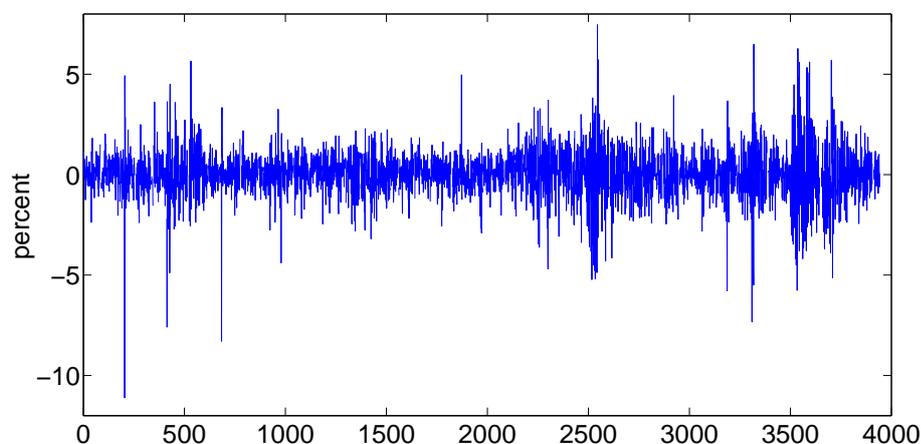


Figure 8.3: Daily return of the SMI (Swiss Market Index) computed as  $\Delta \log(\text{SMI}_t)$  between January 3rd 1989 and February 13th 2004

the logged difference. This time series is denoted by  $X_t$  and plotted in Figure 8.3. One can clearly discern phases of high (observations around  $t = 2500$  and  $t = 3500$ ) and low ( $t = 1000$  and  $t = 2000$ ) volatility. This represents a first sign of heteroskedasticity and positively correlated volatility.

Figure 8.4 shows a normal-quantile plot to compare the empirical distribution of the returns with those from a normal distribution. This plot clearly demonstrates that the probability of observing large returns is bigger than warranted from a normal distribution. Thus the distribution of returns exhibits the heavy-tail property. A similar argument can be made by comparing the histogram of the returns and the density of a normal distribution with the same mean and the same variance, shown in Figure 8.5. Again one can see that absolutely large returns are more probable than expected from a normal distribution. Moreover, the histogram shows no obvious sign for an asymmetric distribution, but a higher peakedness.

After the examination of some preliminary graphical devices, we are going to analyze the autocorrelation functions of  $\{X_t\}$  and  $\{X_t^2\}$ . Figure 8.6 shows the estimated ACFs. The estimated ACF of  $\{X_t\}$  shows practically no significant autocorrelation so that we can consider  $\{X_t\}$  be approximately white noise. The corresponding Ljung-Box statistic with  $L = 100$ , however, has a value of 129.62 which is just above the 5 percent critical value of 124.34. Thus there is some sign of weak autocorrelation. This feature is not in line with efficiency of the Swiss stock market (see Campbell et al. (1997)). The estimated ACF of  $X_t^2$  is clearly outside the 95 percent confidence interval for

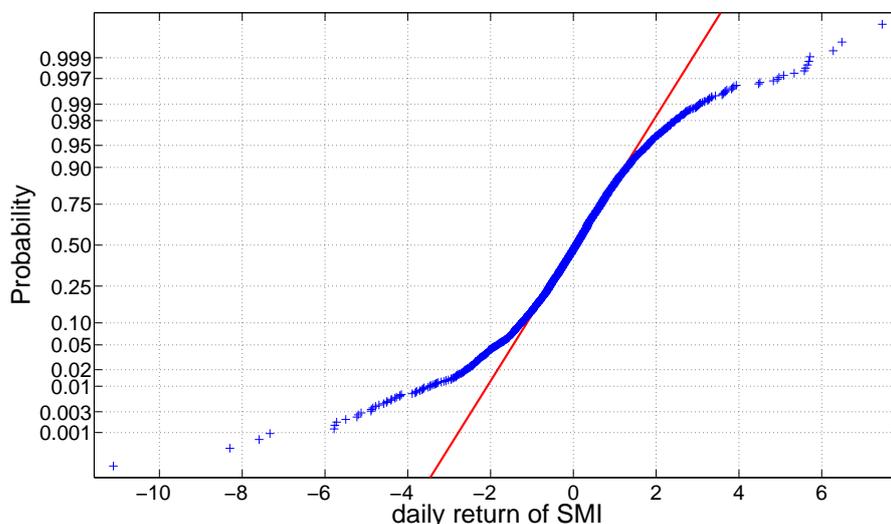


Figure 8.4: Normal-Quantile Plot of the daily returns of the SMI (Swiss Market Index)

at least up to order 20. Thus we can reject the hypothesis of homoskedasticity in favor of heteroskedasticity. This is confirmed by the Ljung-Box statistic with  $L = 100$  with a value of 2000.93 which is much higher than the critical value of 124.34.

After these first investigations, we want to find an appropriate model for the mean equation. We will use OLS with the White-correction. It turns out that a MA(1) model fits the data best although an AR(1) model leads to almost the same results. In the next step we will estimate a GARCH( $p,q$ ) model with the method of maximum likelihood where  $p$  is varied between 1 and 3 and  $q$  between 0 and 3. The values of the AIC, respectively BIC criterion corresponding to the variance equation are listed in tables 8.1 and 8.2.

The results in these tables show that the AIC criterion favors a GARCH(3,3) model corresponding to the bold number in table 8.1 whereas the BIC criterion opts for a GARCH(1,1) model corresponding to the bold number in table 8.2. It also turns out that high dimensional models, in particular those for which  $q > 0$ , the maximization algorithm has problems to find an optimum. Furthermore, the roots of the implicit AR and the MA polynomial corresponding to the variance equation of the GARCH(3,3) model are very similar. These two arguments lead us to prefer the GARCH(1,1) over the GARCH(3,3) model. This model was estimated to have the following mean

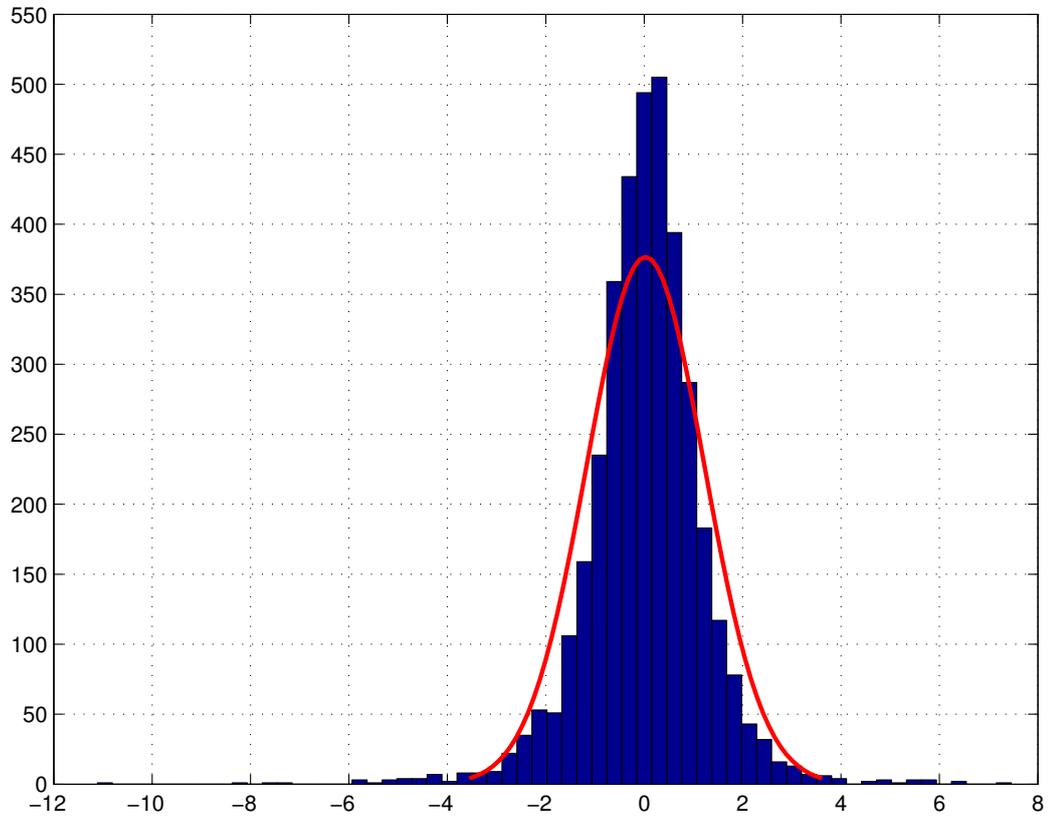


Figure 8.5: Histogram of the daily returns of the SMI (Swiss Market Index) and the density of a fitted normal distribution (red line)

Table 8.1: AIC criterion for the variance equation in the GARCH( $p,q$ ) model

$p$	$q$			
	0	1	2	3
1	3.0886	2.9491	2.9491	2.9482
2	3.0349	2.9496	2.9491	2.9486
3	2.9842	2.9477	2.9472	<b>2.9460</b>

minimum value in bold

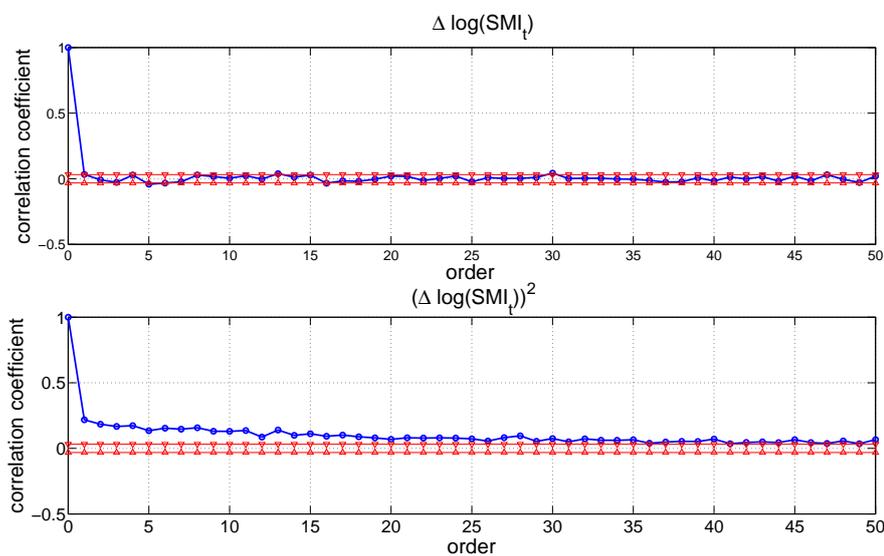


Figure 8.6: ACF of the daily returns and the squared daily returns of the SMI

Table 8.2: BIC criterion for the variance equation in the GARCH(p,q) model

<b>p</b>	<b>q</b>			
	0	1	2	3
1	3.0952	<b>2.9573</b>	2.9590	2.9597
2	3.0431	2.9595	2.9606	2.9617
3	2.9941	2.9592	2.9604	2.9607

minimum value in bold

equation:

$$X_t = 0.0755 + Z_t + 0.0484 Z_{t-1}$$

$$(0.0174) \quad (0.0184)$$

with the corresponding variance equation

$$\sigma_t^2 = 0.0765 + 0.1388 Z_{t-1}^2 + 0.8081 \sigma_{t-1}^2,$$

$$(0.0046) \quad (0.0095) \quad (0.0099)$$

where the estimated standard deviations are reported below the corresponding coefficient estimate. The small, but significant value of 0.0484 for the MA(1) coefficient shows that there is a small but systematic correlation of the returns from one day to the next. The coefficients of the GARCH model are all positive and their sum  $\alpha_1 + \beta = 0.1388 + 0.8081 = 0.9469$  is statistically below one so that all conditions for a stationary process are fulfilled.<sup>14</sup> Because  $\sqrt{3} \frac{\alpha_1}{1-\beta} = \sqrt{3} \frac{0.1388}{1-0.8081} = 1.2528 > 1$ , the condition for the existence of the fourth moment of  $Z_t$  is violated.

As a comparison we also estimate the GARCH(1,1) model using the methods of moments. First we estimate a MA(1) model for  $\Delta \log SMI$ . This results in an estimate  $\hat{\theta} = 0.034$  (compare this with the ML estimate). The squared residuals have correlation coefficients

$$\hat{\rho}_{Z^2}(1) = 0.228 \quad \text{and} \quad \hat{\rho}_{Z^2}(2) = 0.181.$$

The estimate of  $b$  therefore is  $\hat{b} = 2.241$  which leads to an estimate of  $\beta$  equal to  $\hat{\beta} = 0.615$ . This finally results in the estimates of  $\alpha_1$  and  $\alpha_0$  equal to  $\hat{\alpha}_1 = 0.179$  and  $\hat{\alpha}_0 = 0.287$  with an estimate for  $\sigma^2$  equal to  $\hat{\sigma}^2 = 1.391$ . Thus these estimates are quite different from those obtained by the ML method.

### Value at Risk

We are now in a position to use our ML estimates to compute the *Value-at-risk* (VaR). The VaR is a very popular measure to estimate the risk of an investment. In our case we consider the market portfolio represented by the stocks in the SMI. The VaR is defined as the maximal loss (in absolute value) of an investment which occurs with probability  $\alpha$  over a time horizon  $h$ . Thus a one percent VaR for the return on the SMI for the next day is the threshold value of the return such that one can be confident with 99 percent that the loss will not exceed this value. Thus the  $\alpha$  VaR at time  $t$  for  $h$

<sup>14</sup>The corresponding Wald test clearly rejects the null hypothesis  $\alpha_1 + \beta = 1$  at a significance level of one percent.

periods,  $\text{VaR}_{t,t+h}^\alpha$ , is nothing but the  $\alpha$ -quantile of the distribution of the forecast of the return in  $h$  periods given information  $X_{t-k}$ ,  $k = 0, 1, 2, \dots$ . Formally, we have:

$$\text{VaR}_{t,t+h}^\alpha = \inf \left\{ x : \mathbf{P} \left[ \tilde{X}_{t+h} \leq x | X_t, X_{t-1}, \dots \right] \geq \alpha \right\},$$

where  $\tilde{X}_{t+h}$  is the return of the portfolio over an investment horizon of  $h$  periods. This return is approximately equal to the sum of the daily returns:  $\tilde{X}_{t+h} = \sum_{j=1}^h X_{t+j}$ .

The one period forecast error is given by  $X_{t+1} - \tilde{\mathbb{P}}_t X_{t+1}$  which is equal to  $Z_{t+1} = \sigma_{t+1} \nu_{t+1}$ . Thus the VaR for the next day is

$$\text{VaR}_{t,t+1}^\alpha = \inf \left\{ x : \mathbf{P} \left[ \nu_{t+1} \leq \frac{x - \tilde{\mathbb{P}}_t X_{t+1}}{\sigma_{t+1}} \right] \geq \alpha \right\}.$$

This entity can be computed by replacing the forecast given the infinite past,  $\tilde{\mathbb{P}}_t X_{t+1}$ , by a forecast given the finite sample information  $X_{t-k}$ ,  $k = 0, 1, 2, \dots, t-1$ ,  $\mathbb{P}_t X_{t+1}$ , and by substituting  $\sigma_{t+1}$  by the corresponding forecast from variance equation,  $\hat{\sigma}_{t,t+1}$ . Thus we get:

$$\widehat{\text{VaR}}_{t,t+1}^\alpha = \inf \left\{ x : \mathbf{P} \left[ \nu_{t+1} \leq \frac{x - \mathbb{P}_t X_{t+1}}{\hat{\sigma}_{t,t+1}} \right] \geq \alpha \right\}.$$

The computation of  $\widehat{\text{VaR}}_{t,t+1}^\alpha$  requires to determine the  $\alpha$ -quantile of the distribution of  $\nu_t$ . This can be done in two ways. The first one uses the assumption about the distribution of  $\nu_t$  explicitly. In the simplest case,  $\nu_t$  is distributed as a standard normal so that the appropriate quantile can be easily retrieved. The one percent quantile for the standard normal distribution is -2.33. The second approach is a non-parametric one and uses the empirical distribution function of  $\hat{\nu}_t = \hat{Z}_t / \hat{\sigma}_t$  to determine the required quantile. This approach has the advantage that deviations from the standard normal distribution are accounted for. In our case, the one percent quantile is -2.56 and thus considerably lower than the -2.33 obtained from the normal distribution. Thus the VaR is under estimated by using the assumption of the normal distribution.

The corresponding computations for the SMI based on the estimated ARMA(0,1)-GARCH(1,1) model are reported in table 8.3. A value of 5.71 for 31st of December 2001 means that one can be 99 percent sure that the return of an investment in the stocks of the SMI will not be lower than -5.71 percent. The values for the non-parametric approach are typically higher.

Table 8.3: One Percent VaR for the Next Day of the Return to the SMI according to the ARMA(0,1)-GARCH(1,1) model

Date	$\mathbb{P}_t X_{t+1}$	$\hat{\sigma}_{t,t+1}^2$	$\widehat{\text{VaR}}(\widehat{\text{VaR}}_{t,t+1}^{0,01})$	
			parametric	non-parametric
31. 12. 2001	0.28	6.61	5.71	6.30
5. 2. 2002	-0.109	6.80	6.19	6.79
24. 7. 2003	0.0754	0.625	1.77	1.95

Table 8.4: One Percent VaR for the Next 10 Day of the Return to the SMI according to the ARMA(0,1)-GARCH(1,1) model

Date	$\mathbb{P}_t \tilde{X}_{t+1}$	$\widehat{\text{VaR}}(\widehat{\text{VaR}}_{t,t+10}^{0,01})$	
		parametric	non-parametric
31. 12. 2001	0.84	18.39	22.28
5. 2. 2002	0.65	19.41	21.53
24. 7. 2003	0.78	6.53	7.70

The comparison of the VaR for different dates clearly shows how the risk evolves over time.

Due to the nonlinear character of the model, the VaR for more than one day can only be gathered from simulating the one period returns over the corresponding horizon. Starting from a given date 10'000 realizations of the returns over the next 10 days have been simulated whereby the corresponding values for  $\nu_t$  are either drawn from a standard normal distribution (parametric case) or from the empirical distribution function of  $\hat{\nu}_t$  (non-parametric case). The results from this exercise are reported in table 8.4. Obviously, the risk is much higher for a 10 day than for a one day investment.

## Part II

# Multivariate Time Series Analysis



# Chapter 9

## Introduction

The Keynesian macroeconomic theory developed in the 1930's and 1940's, in particular its representation in terms of IS- and LM-curve, opened a new area in the application of statistical methods to economics. Based on the path breaking work by Tinbergen (1939) and Klein (1950) this research gave rise to simultaneous equation systems which should capture all relevant aspect of an economy. The goal was to establish an empirically grounded tool which would enable the politicians to analyze the consequences of their policies and thereby fine tune the economy to overcome or at least mitigate major business cycle fluctuations. This development was enhanced by the systematic compilation of national accounting data and by the advances in computer sciences.<sup>1</sup> These systems were typically built by putting together single equations such as consumption, investment, money demand, export- and import, Phillips-curve equations to an overall model. In the end, the model could well account for several hundreds or even thousands of equations. The climax of this development was the project LINK which linked the different national models to a world model by accounting for their interrelation through trade flows. Although this research program brought many insights and spurred the development of econometrics as a separate field, by the mid 1970's one had to admit that the idea to use large and very detailed models for forecasting and policy analysis was overly optimistic. In particular, the inability to forecast and cope with the oil crisis of the beginning 1970's raised doubts about the viability of this research strategy. In addition, more and more economist had concerns about the theoretical foundations of these models.

The critique had several facets. First, it was argued that the bottom-up strategy of building a system from single equations is not compatible with

---

<sup>1</sup>See Epstein (1987) for an historical overview.

general equilibrium theory which stresses the interdependence of economic activities. This insight was even reinforced by the advent of the theory of rational expectations. This theory postulated that expectations should be formed on the basis of all available information and not just by mechanically extrapolating from the past. This implies that developments in every part of the economy, in particular in the realm of economic policy making, should in principle be taken into account and shape the expectation formation. As expectations are omnipresent in almost every economic decision, all aspects of economic activities (consumption capital accumulation, investment, etc.) are inherently linked. Thus the strategy of using zero restrictions – which meant that certain variables were omitted from a particular equation – to identify the parameters in a simultaneous equation system was considered to be flawed. Second, the theory of rational expectations implied that the typical behavioral equations underlying these models are not invariant to changes in policies because economic agents would take into account systematic changes in the economic environment in their decision making. This so-called Lucas-critique (Lucas, 1976) undermined the basis for the existence of large simultaneous equation models. Third, simple univariate ARMA models proved to be as good in forecasting as the sophisticated large simultaneous models. Thus it was argued that the effort or at least part of the effort devoted to these models was wasted.

In 1980 Sims (1980b) and proposed an alternative modeling strategy. This strategy concentrates the modeling activity to only a few core variables, but places no restrictions what so ever on the dynamic interrelation among them. Thus every variable is considered to be endogenous and, in principle, dependent on all other variables of the model. In the linear context, the class of vector autoregressive (VAR) models has proven to be most convenient to capture this modeling strategy. They are easy to implement and to analyze. In contrast to the simultaneous equation approach, however, it is no longer possible to perform comparative static exercises and to analyze the effect of one variable on another one because every variable is endogenous a priori. Instead, one tries to identify and quantify the effect of shocks over time. These shocks are usually given some economic content, like demand or supply disturbances. However, these shocks are not directly observed, but are disguised behind the residuals from the VAR. Thus, the VAR approach also faces a fundamental identification problem. Since the seminal contribution by Sims, the literature has proposed several alternative identification schemes which will be discussed in Chapter 15 under the header of structural vector autoregressive (SVAR) models. The effects of these shocks are then

further analyzed by computing impulse responses.<sup>2</sup>

The reliance on shocks can be seen as a substitute for the lack of experiments in macroeconomics. The approach can be interpreted as a statistical analogue to the identification of specific episodes where some unforeseen event (shock) impinges on and propagates throughout the economy. Singling-out these episodes of quasi “natural experiments” as in Friedman and Schwartz (1963) convinced many economist of the role and effects of monetary policy.

Many concepts which have been introduced in the univariate context carry over in a straightforward manner to the multivariate context. However there are some new aspects. First, we will analyze the interaction among several variables. This can be done in a nonparametric way by examining the cross-correlations between time series or by building an explicit model. We will restrict ourself to the class of VAR models as they are easy to handle and are overwhelmingly used in practice. Second, we will discuss alternative approaches to identify the structural shocks from VAR models. Third, we will discuss the modeling of integrated variables in a more systematic way. In particular, we will extend the concept of cointegration to more than two variables. Finally, we will provide an introduction to the state space models as a general modeling approach. State space models are becoming increasingly popular in economics as they can be more directly linked to theoretical economic models.

---

<sup>2</sup>Watson (1994) provides a general introduction to this topic.



# Chapter 10

## Definitions and Stationarity

Similarly to the univariate case, we start our exposition with the concept of stationarity which is also crucial in the multivariate setting. Before doing so let us define the multivariate stochastic process.

**Definition 10.1.** *A multivariate stochastic Process,  $\{X_t\}$ , is a family of random variables indexed by  $t$ ,  $t \in \mathbb{Z}$ , which take values in  $\mathbb{R}^n$ ,  $n \geq 1$ .  $n$  is called the dimension of the process.*

Setting  $n = 1$ , the above definition includes as a special case univariate stochastic processes. This implies that the statements for multivariate processes carry over analogously to the univariate case. We view  $X_t$  as a column vector:

$$X_t = \begin{pmatrix} X_{1t} \\ \vdots \\ X_{nt} \end{pmatrix}.$$

Each element  $\{X_{it}\}$  thereby represents a particular variable which may be treated as a univariate process. As in the example of Section 15.4.5,  $\{X_t\}$  represents the multivariate process consisting of the growth rate of GDP  $Y_t$ , the unemployment rate  $U_t$ , the inflation rate  $P_t$ , the wage inflation rate  $W_t$ , and the growth rate of money  $M_t$ . Thus,  $X_t = (Y_t, U_t, P_t, W_t, M_t)'$ .

As in the univariate case, we characterize the joint distribution of the elements  $X_{it}$  and  $X_{jt}$  by the first two moments (if they exist), i.e. by the mean and the variance, respectively covariance:

$$\begin{aligned} \mu_{it} &= \mathbb{E}X_{it}, & i &= 1, \dots, n \\ \gamma_{ij}(t, s) &= \mathbb{E}(X_{it} - \mu_{it})(X_{js} - \mu_{js}), & i, j &= 1, \dots, n; t, s \in \mathbb{Z} \end{aligned} \quad (10.1)$$

It is convenient to write these entities compactly as vectors and matrices:

$$\mu_t = \begin{pmatrix} \mu_{1t} \\ \vdots \\ \mu_{nt} \end{pmatrix} = \mathbb{E}X_t = \begin{pmatrix} \mathbb{E}X_{1t} \\ \vdots \\ \mathbb{E}X_{nt} \end{pmatrix}$$

$$\Gamma(t, s) = \begin{pmatrix} \gamma_{11}(t, s) & \dots & \gamma_{1n}(t, s) \\ \vdots & \ddots & \vdots \\ \gamma_{n1}(t, s) & \dots & \gamma_{nn}(t, s) \end{pmatrix} = \mathbb{E}(X_t - \mu_t)(X_s - \mu_s)'$$

Thus, we apply the expectations operator element-wise to vectors and matrices. The matrix-valued function  $\Gamma(t, s)$  is called the *covariance function* of  $\{X_t\}$ .

In analogy to the univariate case, we define stationarity as the invariance of the first two moments to time shifts:

**Definition 10.2** (Stationarity). *A multivariate stochastic process  $\{X_t\}$  is stationary if and only if for all integers  $r, s$  and  $t$  we have:*

- (i)  $\mu = \mu_t = \mathbb{E}X_t$  is constant (independent of  $t$ );
- (ii)  $\mathbb{E}X_t'X_t < \infty$ ;
- (iii)  $\Gamma(t, s) = \Gamma(t + r, s + r)$ .

In the literature these properties are often called weak stationarity, covariance stationarity, or stationarity of second order. If  $\{X_t\}$  is stationary, the covariance function only depends on the number of periods between  $t$  and  $s$  (i.e. on  $t - s$ ) and not on  $t$  or  $s$  themselves. This implies that by setting  $r = -s$  and  $h = t - s$  the covariance function simplifies to

$$\begin{aligned} \Gamma(h) &= \Gamma(t - s) = \Gamma(t + r, s + r) = \Gamma(t + h, t) \\ &= \mathbb{E}(X_{t+h} - \mu)(X_t - \mu)' = \mathbb{E}(X_t - \mu)(X_{t-h} - \mu)'. \end{aligned}$$

For  $h = 0$ ,  $\Gamma(0)$  is the unconditional covariance matrix of  $X_t$ . Using the definition of the covariances in equation (10.1) we get:

$$\Gamma(h) = \Gamma(-h)'$$

Note that  $\Gamma(h)$  is in general not symmetric for  $h \neq 0$  because  $\gamma_{ij}(h) \neq \gamma_{ji}(h)$  for  $h \neq 0$ .

Based on the covariance function of a stationary process, we can define the *correlation function*  $R(h)$  where  $R(h) = (\rho_{ij}(h))_{i,j}$  with

$$\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}.$$

In the case  $i \neq j$  we refer to the *cross-correlations* between two variables  $\{X_{it}\}$  and  $\{X_{jt}\}$ . The correlation function can be written in matrix notation as

$$R(h) = V^{-1/2}\Gamma(h)V^{-1/2}$$

where  $V$  represents a diagonal matrix with diagonal elements equal to  $\gamma_{ii}(0)$ . Clearly,  $\rho_{ii}(0) = 1$ . As for the covariance matrix we have that in general  $\rho_{ij}(h) \neq \rho_{ji}(h)$  for  $h \neq 0$ . It is possible that  $\rho_{ij}(h) > \rho_{ij}(0)$ . We can summarize the properties of the covariance function by the following theorem.<sup>1</sup>

**Theorem 10.1.** *The covariance function of a stationary process  $\{X_t\}$  has the following properties:*

- (i) For all  $h \in \mathbb{Z}$ ,  $\Gamma(h) = \Gamma(-h)'$ ;
- (ii) for all  $h \in \mathbb{Z}$ ,  $|\gamma_{ij}(h)| \leq \sqrt{\gamma_{ii}(0) \times \gamma_{jj}(0)}$ ;
- (iii) for each  $i = 1, \dots, n$ ,  $\gamma_{ii}(h)$  is a univariate autocovariance function;
- (iv)  $\sum_{r,k=1}^m a'_r \Gamma(r-k) a_k \geq 0$  for all  $m \in \mathbb{N}$  and all  $a_1, \dots, a_m \in \mathbb{R}^n$ . This property is called *non-negative definiteness* (see Property 4 in Theorem 1.1 of Section 1.3 in the univariate case).

*Proof.* Property (i) follows immediately from the definition. Property (ii) follows from the fact that the correlation coefficient is always smaller or equal to one.  $\gamma_{ii}(h)$  is the autocovariance function of  $\{X_{it}\}$  which delivers property (iii). Property (iv) follows from  $\mathbb{E}(\sum_{k=1}^m a'_k (X_{t-k} - \mu))^2 \geq 0$ .  $\square$

If not only the first two moments are invariant to time shifts, but the distribution as a whole we arrive at the concept of strict stationarity.

**Definition 10.3** (Strict Stationarity). *A process multivariate  $\{X_t\}$  is called strictly stationary if and only if, for all  $n \in \mathbb{N}$ ,  $t_1, \dots, t_n$ ,  $h \in \mathbb{Z}$ , the joint distributions of  $(X_{t_1}, \dots, X_{t_n})$  and  $(X_{t_1+h}, \dots, X_{t_n+h})$  are the same.*

### An Example

Consider the following example for  $n = 2$ :

$$\begin{aligned} X_{1t} &= Z_t \\ X_{2t} &= Z_t + 0.75Z_{t-2} \end{aligned}$$

---

<sup>1</sup>We leave it to the reader to derive an analogous theorem for the correlation function.

where  $Z_t \sim \text{WN}(0, 1)$ . We then have  $\mu = \mathbb{E}X_t = 0$ . The covariance function is given by

$$\Gamma(h) = \begin{cases} \begin{pmatrix} 1 & 1 \\ 1 & 1.5625 \end{pmatrix}, & h = 0; \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & h = 1; \\ \begin{pmatrix} 0 & 0 \\ 0.75 & 0.75 \end{pmatrix}, & h = 2. \end{cases}$$

The covariance function is zero for  $h > 2$ . The values for  $h < 0$  are obtained from property (i) in Theorem 10.1. The correlation function is:

$$R(h) = \begin{cases} \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, & h = 0; \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & h = 1; \\ \begin{pmatrix} 0 & 0 \\ 0.60 & 0.48 \end{pmatrix}, & h = 2. \end{cases}$$

The correlation function is zero for  $h > 2$ . The values for  $h < 0$  are obtained from property (i) in Theorem 10.1.

One idea in time series analysis is to construct more complicated process from simple ones, for example by taking moving-averages. The simplest process is the white noise process which is uncorrelated with its own past. In the multivariate context the white noise process is defined as follows.

**Definition 10.4.** A stochastic process  $\{Z_t\}$  is called (multivariate) white noise with mean zero and covariance matrix  $\Sigma > 0$ , denoted by  $Z_t \sim \text{WN}(0, \Sigma)$ , if it is stationary  $\{Z_t\}$  and

$$\mathbb{E}Z_t = 0, \\ \Gamma(h) = \begin{cases} \Sigma, & h = 0; \\ 0, & h \neq 0. \end{cases}$$

If  $\{Z_t\}$  is not only white noise, but independently and identically distributed we write  $Z_t \sim \text{IID}(0, \Sigma)$ .

**Remark 10.1.** Even if each component of  $\{Z_{it}\}$  is univariate white noise, this does not imply that  $\{Z_t\} = \{(Z_{1t}, \dots, Z_{nt})'\}$  is multivariate white noise. Take, for example the process  $Z_t = (u_t, u_{t-1})'$  where  $u_t \sim \text{WN}(0, \sigma_u^2)$ . Then  $\Gamma(1) = \begin{pmatrix} 0 & 0 \\ \sigma_u^2 & 0 \end{pmatrix} \neq 0$ .

Taking moving averages of a white noise process it is possible to generate new stationary processes. This leads to the definition of a linear process.

**Definition 10.5.** A stochastic process  $\{X_t\}$  is called linear if there exists a representation

$$X_t = \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

where  $Z_t \sim \text{IID}(0, \Sigma)$  and where the sequence  $\{\Psi_j\}$  of the  $n \times n$  matrices is absolutely summable, i.e.  $\sum_{j=-\infty}^{\infty} \|\Psi_j\| < \infty$ . If for all  $j < 0$   $\Psi_j = 0$ , the linear process is also called an MA( $\infty$ ) process.

**Theorem 10.2.** A linear process is stationary with a mean of zero and with covariance function

$$\Gamma(h) = \sum_{j=-\infty}^{\infty} \Psi_{j+h} \Sigma \Psi_j' = \sum_{j=-\infty}^{\infty} \Psi_j \Sigma \Psi_{j-h}', \quad h = 0, \pm 1, \pm 2, \dots$$

*Proof.* The required result is obtained by applying the properties of  $\{Z_t\}$  to  $\Gamma(h) = \mathbb{E}X_{t+h}X_t' = \lim_{m \rightarrow \infty} \mathbb{E} \left( \sum_{j=-m}^m \Psi_j Z_{t+h-j} \right) \left( \sum_{k=-m}^m \Psi_k Z_{t-k} \right)'$ .  $\square$

**Remark 10.1.** The same conclusion is reached if  $\{Z_t\}$  is a white noise process and not an IID process.

## Appendix: Norm and Summability of Matrices

As in the definition of a linear process it is often necessary to analyze the convergence of a sequence of matrices  $\{\Psi_j\}$ ,  $j = 0, 1, 2, \dots$ . For this we need to define a norm for matrices. The literature considers different alternative approaches. For our purposes, the choice is not relevant as all norms are equivalent in the finite dimensional vector space. We therefore choose the Frobenius, Hilbert-Schmidt or Schur norm which is easy to compute.<sup>2</sup> This norm treats the elements of a  $m \times n$  matrix  $A = (a_{ij})$  as an element of the  $\mathbb{R}^{m \times n}$  Euclidean space and defines the length of  $A$ , denoted by  $\|A\|$ , as  $\|A\| = \sqrt{\sum_{i,j} a_{ij}^2}$ . This leads to the formal definition below.

**Definition 10.6.** The Frobenius, Hilbert-Schmidt or Schur norm of a  $m \times n$  matrix  $A$ , denoted by  $\|A\|$ , is defined as:

$$\|A\|^2 = \sum_{i,j} a_{ij}^2 = \text{tr}(A'A) = \sum_{i=1}^n \lambda_i$$

<sup>2</sup>For details see Meyer (2000, 279ff).

where  $\text{tr}(A'A)$  denotes the trace of  $A'A$ , i.e. the sum of the diagonal elements of  $A'A$ , and where  $\lambda_i$  are the  $n$  eigenvalues of  $A'A$ .

The matrix norm has the following properties:

$$\begin{aligned} \|A\| &\geq 0 \quad \text{and} \quad \|A\| = 0 \text{ is equivalent to } A = 0, \\ \|\alpha A\| &= |\alpha| \|A\| \text{ for all } \alpha \in \mathbb{R}, \\ \|A\| &= \|A'\|, \\ \|A + B\| &\leq \|A\| + \|B\| \text{ for all matrices } A \text{ and } B \text{ of the same dimension,} \\ \|AB\| &\leq \|A\| \|B\| \text{ for all conformable matrices } A \text{ and } B. \end{aligned}$$

The last property is called *submultiplicativity*.

A sequence of matrices  $\{\Psi_j\}$ ,  $j = 0, 1, \dots$ , is called *absolutely summable* if and only if  $\sum_{j=0}^{\infty} \|\Psi_j\| < \infty$  and *quadratic summable* if and only if  $\sum_{j=0}^{\infty} \|\Psi_j\|^2 < \infty$ . The absolute summability implies the quadratic summability, but not vice versa.

**Corollary 10.1.** *Absolute summability of  $\{\Psi_j\}$  is equivalent to the absolute summability of each sequence  $\{[\Psi_j]_{kl}\}$ ,  $k, l = 1, \dots, n$ , i.e. to  $\lim_{j \rightarrow \infty} |[\Psi_j]_{kl}|$  exists and is finite.*

*Proof.* In particular, he have:

$$|[\Psi_j]_{kl}| \leq \|\Psi_j\| = \sqrt{\sum_{k=1}^n \sum_{l=1}^n [\Psi_j]_{kl}^2} \leq \sum_{k=1}^n \sum_{l=1}^n |[\Psi_j]_{kl}|.$$

Summation over  $j$  gives

$$\sum_{j=0}^{\infty} |[\Psi_j]_{kl}| \leq \sum_{j=0}^{\infty} \|\Psi_j\| \leq \sum_{j=0}^{\infty} \sum_{k=1}^n \sum_{l=1}^n |[\Psi_j]_{kl}| = \sum_{k=1}^n \sum_{l=1}^n \sum_{j=0}^{\infty} |[\Psi_j]_{kl}|$$

The absolute convergence of each sequence  $\{[\Psi_j]_{kl}\}$ ,  $k, l = 1, \dots, n$ , follows from the absolute convergence of  $\{\Psi_j\}$  by the first inequality. Conversely, the absolute convergence of  $\{\Psi_j\}$  is implied by the absolute convergence of each sequence  $\{[\Psi_j]_{kl}\}$ ,  $k, l = 1, \dots, n$ , from the second inequality.  $\square$

# Chapter 11

## Estimation of Mean and Covariance Function

### 11.1 Estimators and their Asymptotic Distributions

We characterize the stationary process  $\{X_t\}$  by its mean and its (matrix) covariance function. In the Gaussian case, this already characterizes the whole distribution. The estimation of these entities becomes crucial in the empirical analysis. As it turns out, the results from the univariate process carry over analogously to the multivariate case. If the process is observed over the periods  $t = 1, 2, \dots, T$ , then a natural estimator for the mean  $\mu$  is the arithmetic mean or sample average:

$$\hat{\mu} = \bar{X}_T = \frac{1}{T} (X_1 + \dots + X_T) = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_n \end{pmatrix}.$$

We get a theorem analogously to Theorem 4.1 in Section 4.1.

**Theorem 11.1.** *Let  $\{X_t\}$  be stationary process with mean  $\mu$  and covariance function  $\Gamma(h)$  then asymptotically, for  $T \rightarrow \infty$ , we get*

$$\begin{aligned} \mathbb{E} (\bar{X}_T - \mu)' (\bar{X}_T - \mu) &\rightarrow 0, \text{ if } \gamma_{ii}(T) \rightarrow 0 \text{ for all } 1 \leq i \leq n; \\ T\mathbb{E} (\bar{X}_T - \mu)' (\bar{X}_T - \mu) &\rightarrow \sum_{i=1}^n \sum_{h=-\infty}^{\infty} \gamma_{ii}(h), \\ &\text{if } \sum_{h=-\infty}^{\infty} |\gamma_{ii}(h)| < \infty \text{ for all } 1 \leq i \leq n. \end{aligned}$$

*Proof.* The Theorem can be established by applying Theorem 4.1 individually to each time series  $\{X_{it}\}$ ,  $i = 1, 2, \dots, n$ .  $\square$

Thus, the sample average converges in mean square and therefore also in probability to the true mean. Thereby the second condition is more restrictive than the first one. They are, in particular, fulfilled for all VARMA processes (see Chapter 12). As in the univariate case analyzed in Section 4.1, it can be shown with some mild additional assumptions that  $\bar{X}_T$  is also asymptotically normally distributed.

**Theorem 11.2.** *For any stationary process  $\{X_t\}$*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

*with  $Z_t \sim \text{IID}(0, \Sigma)$  and  $\sum_{j=-\infty}^{\infty} \|\Psi_j\| < \infty$ , the arithmetic average  $\bar{X}_T$  is asymptotically normal:*

$$\begin{aligned} \sqrt{T} (\bar{X}_T - \mu) &\xrightarrow{d} \text{N} \left( 0, \sum_{h=-\infty}^{\infty} \Gamma(h) \right) \\ &= \text{N} \left( 0, \left( \sum_{j=-\infty}^{\infty} \Psi_j \right) \Sigma \left( \sum_{j=-\infty}^{\infty} \Psi_j' \right) \right) \\ &= \text{N} (0, \Psi(1) \Sigma \Psi(1)'). \end{aligned}$$

*Proof.* The proof is a straightforward extension to the multivariate case of the one given for Theorem 4.2 of Section 4.1.  $\square$

The summability condition is quite general. It is, in particular, fulfilled by causal VARMA processes (see Chapter 12) as their coefficients matrices  $\Psi_j$  go exponentially fast to zero. Remarks similar to those following Theorem 4.2 apply also in the multivariate case.

The above formula can be used to construct confidence regions for  $\mu$ . This turns out, however, to be relatively complicated in practice so that often univariate approximations are used instead (Brockwell and Davis, 1996, 228-229).

As in the univariate case, a natural estimator for the covariance matrix function  $\Gamma(h)$  is given by the corresponding empirical moments  $\hat{\Gamma}(h)$ :

$$\hat{\Gamma}(h) = \begin{cases} \frac{1}{T} \sum_{t=1}^{T-h} (X_{t+h} - \bar{X}_T) (X_t - \bar{X}_T)', & 0 \leq h \leq T-1; \\ \hat{\Gamma}'(-h), & -T+1 \leq h < 0. \end{cases}$$

The estimator of the covariance function can then be applied to derive an estimator for the correlation function:

$$\widehat{R}(h) = \widehat{V}^{-1/2} \widehat{\Gamma}(h) \widehat{V}^{-1/2}$$

where  $\widehat{V}^{1/2} = \text{diag} \left( \sqrt{\widehat{\gamma}_{11}(0)}, \dots, \sqrt{\widehat{\gamma}_{nn}(0)} \right)$ . Under the conditions given in Theorem 11.2 the estimator of the covariance matrix of order  $h$ ,  $\widehat{\Gamma}(h)$ , converges to the true covariance matrix  $\Gamma(h)$ . Moreover,  $\sqrt{T} \left( \widehat{\Gamma}(h) - \Gamma(h) \right)$  is asymptotically normally distributed. In particular, we can state the following Theorem:

**Theorem 11.3.** *Let  $\{X_t\}$  be a stationary process with*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \Psi_j Z_{t-j}$$

where  $Z_t \sim \text{IID}(0, \Sigma)$ ,  $\sum_{j=-\infty}^{\infty} \|\Psi_j\| < \infty$ , and  $\sum_{j=-\infty}^{\infty} \Psi_j \neq 0$ . Then, for each fixed  $h$ ,  $\widehat{\Gamma}(h)$  converges in probability as  $T \rightarrow \infty$  to  $\Gamma(h)$ :

$$\widehat{\Gamma}(h) \xrightarrow{p} \Gamma(h)$$

*Proof.* A proof can be given along the lines given in Proposition 13.1. □

As for the univariate case, we can define the long-run covariance matrix  $J$  as

$$J = \sum_{h=-\infty}^{\infty} \Gamma(h). \tag{11.1}$$

As a non-parametric estimator we can again consider the following class of estimators:

$$\widehat{J}_T = \sum_{h=-T+1}^{T-1} k \left( \frac{h}{\ell_T} \right) \widehat{\Gamma}(h)$$

where  $k(x)$  is a kernel function and where  $\widehat{\Gamma}(h)$  is the corresponding estimate of the covariance matrix at lag  $h$ . For the choice of the kernel function and the lag truncation parameter the same principles apply as in the univariate case (see Section 4.4 and Haan and Levin (1997)).

## 11.2 Testing the Cross-Correlation of two Time Series

The determination of the asymptotic distribution of  $\widehat{\Gamma}(h)$  is complicated. We therefore restrict ourselves to the case of two time series.

**Theorem 11.4.** *Let  $\{X_t\}$  be a bivariate stochastic process whose components can be described by*

$$X_{1t} = \sum_{j=-\infty}^{\infty} \alpha_j Z_{1,t-j} \quad \text{with } Z_{1t} \sim \text{IID}(0, \sigma_1^2)$$

$$X_{2t} = \sum_{j=-\infty}^{\infty} \beta_j Z_{2,t-j} \quad \text{with } Z_{2t} \sim \text{IID}(0, \sigma_2^2)$$

where  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$  are independent from each other at all leads and lags and where  $\sum_j |\alpha_j| < \infty$  and  $\sum_j |\beta_j| < \infty$ . Under these conditions the asymptotic distribution of the estimator of the cross-correlation function  $\rho_{12}(h)$  between  $\{X_{1t}\}$  and  $\{X_{2t}\}$  is

$$\sqrt{T}\widehat{\rho}_{12}(h) \xrightarrow{d} N\left(0, \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)\right), \quad h \geq 0. \quad (11.2)$$

For all  $h$  and  $k$  with  $h \neq k$ ,  $(\sqrt{T}\widehat{\rho}_{12}(h), \sqrt{T}\widehat{\rho}_{12}(k))'$  converges in distribution to a bivariate normal distribution with mean zero, variances  $\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)$  and covariances  $\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j+k-h)$ .

This result can be used to construct a test of independence, respectively uncorrelatedness, between two time series. The above theorem, however, shows that the asymptotic distribution of  $\sqrt{T}\widehat{\rho}_{12}(h)$  depends on  $\rho_{11}(h)$  and  $\rho_{22}(h)$  and is therefore unknown. Thus, the test cannot be based on the cross-correlation alone.<sup>1</sup>

This problem can, however, be overcome by implementing the following two-step procedure suggested by Haugh (1976).

**First step:** Estimate for each times series separately a univariate invertible ARMA model and compute the resulting residuals  $\widehat{Z}_{it} = \sum_{j=0}^{\infty} \widehat{\pi}_j^{(i)} X_{i,t-j}$ ,  $i = 1, 2$ . If the ARMA models correspond to the true ones, these residuals should approximately be white noise. This first step is called pre-whitening.

<sup>1</sup>The theorem may also be used to conduct a causality test between two times series (see Section 15.1).

**Second step:** Under the null hypothesis the two time series  $\{X_{1t}\}$  and  $\{X_{2t}\}$  are uncorrelated with each other. This implies that the residuals  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$  should also be uncorrelated with each other. The variance of the cross-correlations between  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$  are therefore asymptotically equal to  $1/T$  under the null hypothesis. Thus, one can apply the result of Theorem 11.4 to construct confidence intervals based on formula (11.2). A 95-percent confidence interval is therefore given by  $\pm 1.96T^{-1/2}$ . The Theorem may also be used to construct a test whether the two series are uncorrelated.

If one is not interested in modeling the two time series explicitly, the simplest way is to estimate a high order AR model in the first step. Thereby, the order should be chosen high enough to obtain white noise residuals in the first step. Instead of looking at each cross-correlation separately, one may also test the joint null hypothesis that all cross-correlations are simultaneously equal to zero. Such a test can be based on  $T$  times the sum of the squared cross-correlation coefficients. This statistic is distributed as a  $\chi^2$  with  $L$  degrees of freedom where  $L$  is the number of summands (see the Haugh-Pierce statistic (15.1) in Section 15.1).

## 11.3 Some Examples for Independence Tests

### Two independent AR processes

Consider two AR(1) process  $\{X_{1t}\}$  and  $\{X_{2t}\}$  governed by the following stochastic difference equation  $X_{it} = 0.8X_{i,t-1} + Z_{it}$ ,  $i = 1, 2$ . The two white noise processes  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$  are such that they are independent from each other.  $\{X_{1t}\}$  and  $\{X_{2t}\}$  are therefore independent from each other too. We simulate realizations of these two processes over 400 periods. The estimated cross-correlation function of these so-generated processes are plotted in the upper panel of Figure 11.1. From there one can see that many values are outside the 95 percent confidence interval given by  $\pm 1.96T^{-1/2} = 0.098$ , despite the fact that by construction both series are independent of each other. The reason is that the so computed confidence interval is not correct because it does not take the autocorrelation of each series into account. The application

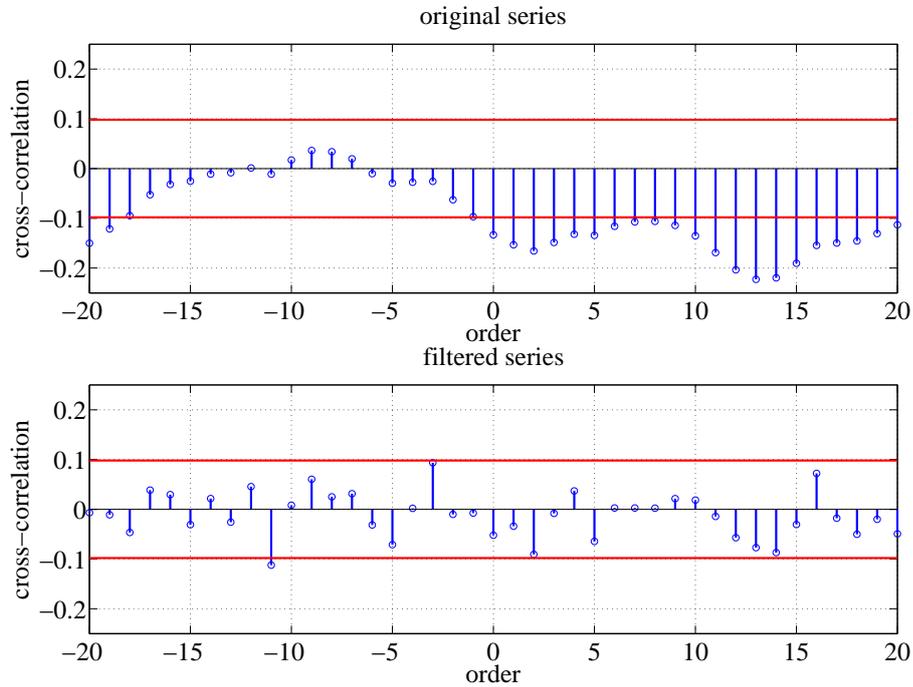


Figure 11.1: Cross-correlations between two independent AR(1) processes with  $\phi = 0.8$

of Theorem 11.4 leads to the much larger 95-percent confidence interval of

$$\begin{aligned} \frac{\pm 1.96}{\sqrt{T}} \sqrt{\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)} &= \frac{\pm 1.96}{20} \sqrt{\sum_{j=-\infty}^{\infty} 0.8^j 0.8^j} \\ &= \frac{\pm 1.96}{20} \sqrt{1 + \frac{2 \times 0.64}{1 - 0.64}} = \pm 0.209 \end{aligned}$$

which is more than twice as large. This confidence interval then encompasses most the cross-correlations computed with respect to the original series.

If one follows the testing procedure outline above instead and fits an AR(10) model for each process and then estimates the cross-correlation function of the corresponding residual series (filtered or pre-whitened time series), the plot in the lower panel of Figure 11.1 is obtained.<sup>2</sup> This figure shows no significant cross-correlation anymore so that one cannot reject the null hypothesis that both time series are independent from each other.

<sup>2</sup>The order of the AR processes are set arbitrarily equal to 10 which is more than enough to obtain white noise residuals.

### Consumption Expenditure and Advertisement

This application focuses on the interaction between nominal aggregate private consumption expenditure and nominal aggregate advertisement expenditures. Such an investigation was first conducted by Ashley et al. (1980) for the United States.<sup>3</sup> The upper panel of Figure 11.2 shows the raw cross-correlations between the two time series where the order  $h$  runs from  $-20$  to  $+20$ . Although almost all cross-correlations are positive and outside the conventional confidence interval, it would be misleading to infer a statistically significant positive cross-correlation. In order to test for independence, we filter both time series by an AR(10) model and estimate the cross-correlations for the residuals.<sup>4</sup> These are displayed in the lower panel of Figure 11.2. In this figure, only the correlations of order 0 and 16 fall outside the confidence interval and can thus be considered as statistically significant. Thus, we can reject the null hypothesis of independence between the two series. However, most of the interdependence seems to come from the correlation within the same quarter. This is confirmed by a more detailed investigation in Berndt (1991) where no significant lead and/or lag relations are found.

### Real Gross Domestic Product and Consumer Sentiment

The procedure outlined above can be used to examine whether one of the two time series is systematically leading the other one. This is, for example, important in the judgment of the current state of the economy because first provisional national accounting data are usually published with a lag of at least one quarter. However, in the conduct of monetary policy more up-to-date knowledge is necessary. Such a knowledge can be retrieved from leading indicator variables. These variables should be available more quickly and should be highly correlated with the variable of interest at a lead.

We investigate whether the Consumer Sentiment Index is a leading indicator for the percentage changes in real Gross Domestic Product (GDP).<sup>5</sup> The raw cross-correlations are plotted in the upper panel of Figure 11.3. It shows several correlations outside the conventional confidence interval. The use of this confidence interval is, however, misleading as the distribution of the raw cross-correlations depends on the autocorrelations of each series. Thus,

---

<sup>3</sup>The quarterly data are taken from Berndt (1991). They cover the period from the first quarter 1956 to the fourth quarter 1975. In order to achieve stationarity, we work with first differences.

<sup>4</sup>The order of the AR processes are set arbitrarily equal to 10 which is more than enough to obtain white noise residuals.

<sup>5</sup>We use data for Switzerland as published by the State Secretariat for Economic Affairs SECO

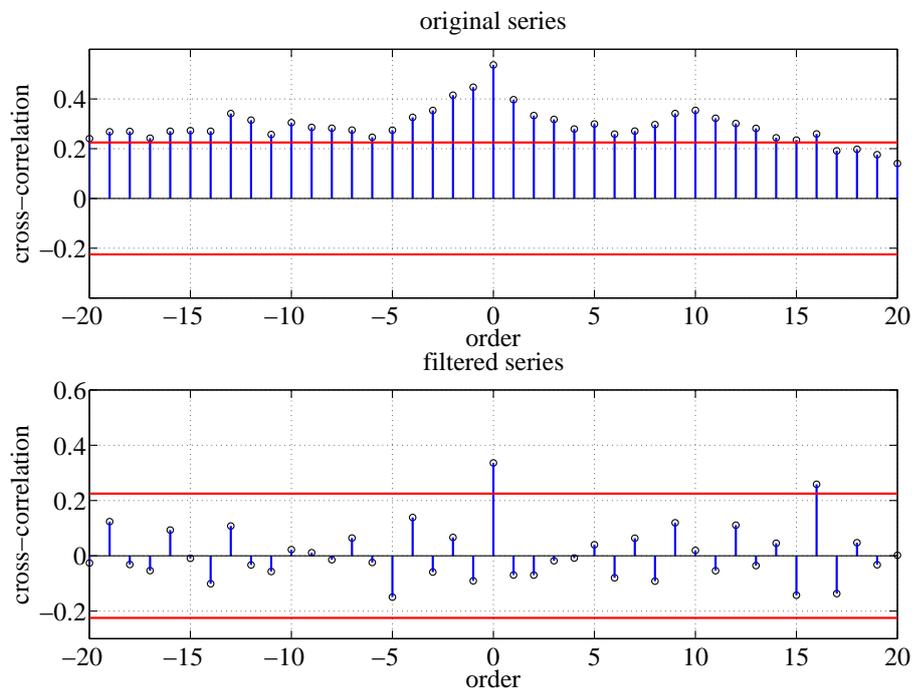


Figure 11.2: Cross-correlations between aggregate nominal private consumption expenditures and aggregate nominal advertisement expenditures

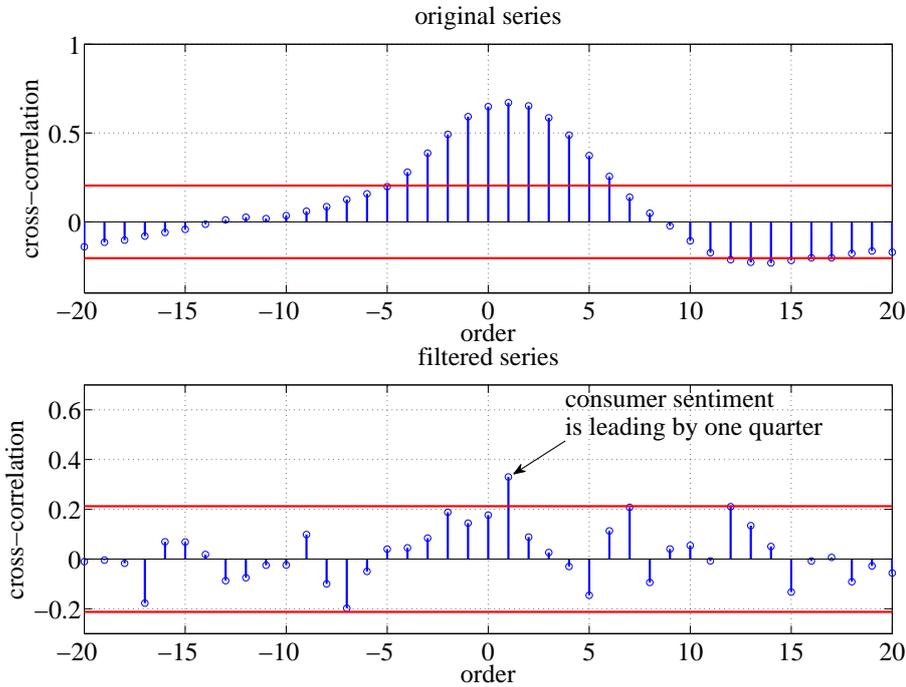


Figure 11.3: Cross-correlations between real growth of GDP and the Consumer Sentiment Index

instead we filter both time series by an AR(8) model and investigate the cross-correlations of the residuals.<sup>6</sup> The order of the AR model was chosen deliberately high to account for all autocorrelations. The cross-correlations of the filtered data are displayed in the lower panel of Figure 11.3. As it turns out, only the cross-correlation which is significantly different from zero is for  $h = 1$ . Thus the Consumer Sentiment Index is leading the growth rate in GDP. In other words, an unexpected higher consumer sentiment is reflected in a positive change in the GDP growth rate of next quarter.<sup>7</sup>

<sup>6</sup>With quarterly data it is wise to set to order as a multiple of four to account for possible seasonal movements. As it turns out  $p = 8$  is more than enough to obtain white noise residuals.

<sup>7</sup>During the interpretation of the cross-correlations be aware of the ordering of the variables because  $\rho_{12}(1) = \rho_{21}(-1) \neq \rho_{21}(1)$ .



# Chapter 12

## Stationary Time Series Models: Vector Autoregressive Moving-average Processes (VARMA Processes)

The most important class of models is obtained by requiring  $\{X_t\}$  to be the solution of a linear stochastic difference equation with constant coefficients. In analogy to the univariate case, this leads to the theory of vector autoregressive moving-average processes (VARMA processes or just ARMA processes).

**Definition 12.1** (VARMA process). *A multivariate stochastic process  $\{X_t\}$  is a vector autoregressive moving-average process of order  $(p, q)$ , VARMA $(p, q)$  process, if it is stationary and fulfills the stochastic difference equation*

$$X_t - \Phi_1 X_{t-1} - \dots - \Phi_p X_{t-p} = Z_t + \Theta_1 Z_{t-1} + \dots + \Theta_q Z_{t-q} \quad (12.1)$$

where  $\Phi_p \neq 0$ ,  $\Theta_q \neq 0$  and  $Z_t \sim \text{WN}(0, \Sigma)$ .  $\{X_t\}$  is called a VARMA $(p, q)$  process with mean  $\mu$  if  $\{X_t - \mu\}$  is a VARMA $(p, q)$  process.

With the aid of the lag operator we can write the difference equation more compactly as

$$\Phi(L)X_t = \Theta(L)Z_t$$

where  $\Phi(L) = I_n - \Phi_1 L - \dots - \Phi_p L^p$  and  $\Theta(L) = I_n + \Theta_1 L + \dots + \Theta_q L^q$ .  $\Phi(L)$  and  $\Theta(L)$  are  $n \times n$  matrices whose elements are lag polynomials of order smaller or equal to  $p$ , respectively  $q$ . If  $q = 0$ ,  $\Theta(L) = I_n$  so that there is no moving-average part. The process is then a purely autoregressive one which is simply called a VAR $(p)$  process. Similarly if  $p = 0$ ,  $\Phi(L) = I_n$  and there

is no autoregressive part. The process is then a purely moving-average one and simply called a VMA(q) process. The importance of VARMA processes stems from the fact that every stationary process can be arbitrarily well approximated by a VARMA process, VAR process, or VMA process.

## 12.1 The VAR(1) Process

we start our discussion by analyzing the properties of the VAR(1) process which is defined as the solution the following stochastic difference equation:

$$X_t = \Phi X_{t-1} + Z_t \quad \text{with } Z_t \sim \text{WN}(0, \Sigma).$$

We assume that all eigenvalues of  $\Phi$  are absolutely strictly smaller than one. As the eigenvalues correspond to the inverses of the roots of the matrix polynomial  $\det(\Phi(z)) = \det(I_n - \Phi z)$ , this assumption implies that all roots must lie outside the unit circle:

$$\det(I_n - \Phi z) \neq 0 \text{ for all } z \in \mathbb{C} \text{ with } |z| \leq 1.$$

For the sake of exposition, we will further assume that  $\Phi$  is diagonalizable, i.e. there exists an invertible matrix  $P$  such that  $J = P^{-1}\Phi P$  is a diagonal matrix with the eigenvalues of  $\Phi$  on the diagonal.<sup>1</sup>

Consider now the stochastic process

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

We will show that this process is stationary and fulfills the first order difference equation above. For  $\{X_t\}$  to be well-defined, we must show that  $\sum_{j=0}^{\infty} \|\Phi^j\| < \infty$ . Using the properties of the matrix norm we get:

$$\begin{aligned} \sum_{j=0}^{\infty} \|\Phi^j\| &= \sum_{j=0}^{\infty} \|P J^j P^{-1}\| \leq \sum_{j=0}^{\infty} \|P\| \|J^j\| \|P^{-1}\| \\ &\leq \sum_{j=0}^{\infty} \|P\| \|P^{-1}\| \sqrt{\sum_{i=1}^n |\lambda_i|^{2j}} \\ &\leq \|P\| \|P^{-1}\| \sqrt{n} \sum_{j=0}^{\infty} |\lambda_{\max}|^{2j} < \infty, \end{aligned}$$

---

<sup>1</sup>The following exposition remains valid even if  $\Phi$  is not diagonalizable. In this case one has to rely on the Jordan form which complicates the computations (Meyer, 2000).

where  $\lambda_{\max}$  denotes the maximal eigenvalue of  $\Phi$  in absolute terms. As all eigenvalues are required to be strictly smaller than one, this clearly also holds for  $\lambda_{\max}$  so that infinite matrix sum converges. This implies that the process  $\{X_t\}$  is stationary. In addition, we have that

$$X_t = \sum_{j=0}^{\infty} \Phi^j Z_{t-j} = Z_t + \Phi \sum_{j=0}^{\infty} \Phi^j Z_{t-1-j} = \Phi X_{t-1} + Z_t.$$

Thus, the process  $\{X_t\}$  also fulfills the difference equation.

Next we demonstrate that this process is also the unique stationary solution to the difference equation. Suppose that there exists another stationary process  $\{Y_t\}$  which also fulfills the difference equation. By successively iterating the difference equation one obtains:

$$\begin{aligned} Y_t &= Z_t + \Phi Z_{t-1} + \Phi^2 Y_{t-2} \\ &\dots \\ &= Z_t + \Phi Z_{t-1} + \Phi^2 Z_{t-2} + \dots + \Phi^k Z_{t-k} + \Phi^{k+1} Y_{t-k-1}. \end{aligned}$$

Because  $\{Y_t\}$  is assumed to be stationary,  $\mathbb{V}Y_t = \mathbb{V}Y_{t-k-1} = \Gamma(0)$  so that

$$\mathbb{V} \left( Y_t - \sum_{j=0}^k \Phi^j Z_{t-j} \right) = \Phi^{k+1} \mathbb{V}(Y_{t-k-1}) \Phi'^{k+1} = \Phi^{k+1} \Gamma(0) \Phi'^{k+1}.$$

The submultiplicativity of the norm then implies:

$$\|\Phi^{k+1} \Gamma(0) \Phi'^{k+1}\| \leq \|\Phi^{k+1}\|^2 \|\Gamma(0)\| = \|P\|^2 \|P^{-1}\|^2 \|\Gamma(0)\| \left( \sum_{i=1}^n |\lambda_i|^{2(k+1)} \right).$$

As all eigenvalues of  $\Phi$  are absolutely strictly smaller than one, the right hand side of the above expression converges to zero for  $k$  going to infinity. This implies that  $Y_t$  and  $X_t = \sum_{j=0}^{\infty} \Phi^j Z_{t-j}$  are equal in the mean square sense and thus also in probability.

Based on Theorem 10.2, the mean and the covariance function of the VAR(1) process is:

$$\begin{aligned} \mathbb{E}X_t &= \sum_{j=0}^{\infty} \Phi^j \mathbb{E}Z_{t-j} = 0, \\ \Gamma(h) &= \sum_{j=0}^{\infty} \Phi^{j+h} \Sigma \Phi'^j = \Phi^h \sum_{j=0}^{\infty} \Phi^j \Sigma \Phi'^j = \Phi^h \Gamma(0). \end{aligned}$$

Analogously to the univariate case, it can be shown that there still exists a unique stationary solution if all eigenvalues are absolutely strictly greater than one. This solution is, however, no longer causal with respect to  $\{Z_t\}$ . If some of the eigenvalues of  $\Phi$  are on the unit circle, there exists no stationary solution.

## 12.2 Representation in Companion Form

A VAR(p) process of dimension  $n$  can be represented as a VAR(1) process of dimension  $p \times n$ . For this purpose we define the  $pn$  vector  $Y_t = (X_t', X_{t-1}', \dots, X_{t-p+1}')'$ . This new process  $\{Y_t\}$  is characterized by the following first order stochastic difference equation:

$$\begin{aligned} Y_t &= \begin{pmatrix} X_t \\ X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-p+1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \dots & 0 & 0 \\ 0 & I_n & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_n & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ X_{t-3} \\ \vdots \\ X_{t-p} \end{pmatrix} + \begin{pmatrix} Z_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \Phi Y_{t-1} + U_t \end{aligned}$$

where  $U_t = (Z_t, 0, 0, \dots, 0)'$  with  $U_t \sim \text{WN} \left( 0, \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \right)$ . This representation is also known as the *companion form* or *state space representation* (see also Chapter 17). In this representation the last  $p(n-1)$  equations are simply identities so that there is no error term attached. The latter name stems from the fact that  $Y_t$  encompasses all the information necessary to describe the state of the system. The matrix  $\Phi$  is called the companion matrix of the VAR(p) process.<sup>2</sup>

The main advantage of the companion form is that by studying the properties of the VAR(1) model, one implicitly encompasses VAR models of higher order and also univariate AR(p) models which can be considered as special cases. The relation between the eigenvalues of the companion matrix and the roots of the polynomial matrix  $\Phi(z)$  is given by the formula (Gohberg et al., 1982):

$$\det(I_{np} - \Phi z) = \det(I_n - \Phi_1 z - \dots - \Phi_p z^p). \quad (12.2)$$

<sup>2</sup>The representation of a VAR(p) process in companion form is not uniquely defined. Permutations of the elements in  $Y_t$  will lead to changes in the companion matrix.

In the case of the AR(p) process the eigenvalues of  $\Phi$  are just the inverses of the roots of the polynomial  $\Phi(z)$ . Further elaboration of state space models is given in Chapter 17.

## 12.3 Causal Representation

As will become clear in Chapter 15 and particularly in Section 15.2, the issue of the existence of a *causal representation* is even more important than in the univariate case. Before stating the main theorem let us generalize the definition of a causal representation from the univariate case (see Definition 2.2 in Section 2.3) to the multivariate one.

**Definition 12.2.** A process  $\{X_t\}$  with  $\Phi(L)X_t = \Theta(L)Z_t$  is called causal with respect to  $\{Z_t\}$  if and only if there exists a sequence of absolutely summable matrices  $\{\Psi_j\}$ ,  $j = 0, 1, 2, \dots$ , i.e.  $\sum_{j=0}^{\infty} \|\Psi_j\| < \infty$ , such that

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j}.$$

**Theorem 12.1.** Let  $\{X_t\}$  be a VARMA(p,q) process with  $\Phi(L)X_t = \Theta(L)Z_t$  and assume that

$$\det \Phi(z) \neq 0 \quad \text{for all } z \in \mathbb{C} \text{ with } |z| \leq 1,$$

then the stochastic difference equation  $\Phi(L)X_t = \Theta(L)Z_t$  has exactly one stationary solution with causal representation

$$X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j},$$

whereby the sequence of matrices  $\{\Psi_j\}$  is absolutely summable and where the matrices are uniquely determined by the identity

$$\Phi(z)\Psi(z) = \Theta(z).$$

*Proof.* The proof is a straightforward extension of the univariate case.  $\square$

As in the univariate case, the coefficient matrices which make up the causal representation can be found by the method of undetermined coefficients, i.e. by equating  $\Phi(z)\Psi(z) = \Theta(z)$ . In the case of the VAR(1) process,

the  $\{\Psi_j\}$  have to obey the following recursion:

$$\begin{aligned} 0 & : \quad \Psi_0 = I_n \\ z & : \quad \Psi_1 = \Phi\Psi_0 = \Phi \\ z^2 & : \quad \Psi_2 = \Phi\Psi_1 = \Phi^2 \\ & \quad \dots \\ z^j & : \quad \Psi_j = \Phi\Psi_{j-1} = \Phi^j \end{aligned}$$

The recursion in the VAR(2) case is:

$$\begin{aligned} 0 & : \quad \Psi_0 = I_n \\ z & : \quad -\Phi_1 + \Psi_1 = 0 & \Rightarrow \quad \Psi_1 = \Phi_1 \\ z^2 & : \quad -\Phi_2 - \Phi_1\Psi_1 + \Psi_2 = 0 & \Rightarrow \quad \Psi_2 = \Phi_2 + \Phi_1^2 \\ z^3 & : \quad -\Phi_1\Psi_2 - \Phi_2\Psi_1 + \Psi_3 = 0 & \Rightarrow \quad \Psi_3 = \Phi_1^3 + \Phi_1\Phi_2 + \Phi_2\Phi_1 \\ & \quad \dots \end{aligned}$$

**Remark 12.1.** Consider a VAR(1) process with  $\Phi = \begin{pmatrix} 0 & \phi \\ 0 & 0 \end{pmatrix}$  with  $\phi \neq 0$  then the matrices in the causal representation are  $\Psi_j = \Phi^j = 0$  for  $j > 1$ . This means that  $\{X_t\}$  has an alternative representation as a VMA(1) process because  $X_t = Z_t + \Phi Z_{t-1}$ . This simple example demonstrates that the representation of  $\{X_t\}$  as a VARMA process is not unique. It is therefore impossible to always distinguish between VAR and VMA process of higher orders without imposing additional assumptions. These additional assumptions are much more complex in the multivariate case and are known as identifying assumptions. Thus, a general treatment of this identification problem is outside the scope of this book. See Hannan and Deistler (1988) for a general treatment of this issue. For this reason we will concentrate exclusively on VAR processes where these identification issues do not arise.

### Example

We illustrate the above concept by the following VAR(2) model:

$$\begin{aligned} X_t &= \begin{pmatrix} 0.8 & -0.5 \\ 0.1 & -0.5 \end{pmatrix} X_{t-1} + \begin{pmatrix} -0.3 & -0.3 \\ -0.2 & 0.3 \end{pmatrix} X_{t-2} + Z_t \\ & \text{with } Z_t \sim \text{WN} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 2.0 \end{pmatrix} \right). \end{aligned}$$

In a first step, we check whether the VAR model admits a causal representation with respect to  $\{Z_t\}$ . For this purpose we have to compute the roots of the equation  $\det(I_2 - \Phi_1 z - \Phi_2 z^2) = 0$ :

$$\det \begin{pmatrix} 1 - 0.8z + 0.3z^2 & 0.5z + 0.3z^2 \\ -0.1z + 0.2z^2 & 1 + 0.5z - 0.3z^2 \end{pmatrix} \\ = 1 - 0.3z - 0.35z^2 + 0.32z^3 - 0.15z^4 = 0.$$

The four roots are:  $-1.1973, 0.8828 \pm 1.6669i, 1.5650$ . As they are all outside the unit circle, there exists a causal representation which can be found from the equation  $\Phi(z)\Psi(z) = I_2$  by the method of undetermined coefficients. Multiplying the equation system out, we get:

$$\begin{aligned} I_2 - \Phi_1 z - \Phi_2 z^2 \\ + \Psi_1 z - \Phi_1 \Psi_1 z^2 - \Phi_2 \Psi_1 z^3 \\ + \Psi_2 z^2 - \Phi_1 \Psi_2 z^3 - \Phi_2 \Psi_2 z^4 \\ \dots \\ = I_2. \end{aligned}$$

Equating the coefficients corresponding to  $z^j$ ,  $j = 1, 2, \dots$ :

$$\begin{aligned} z : \quad \Psi_1 &= \Phi_1 \\ z^2 : \quad \Psi_2 &= \Phi_1 \Psi_1 + \Phi_2 \\ z^3 : \quad \Psi_3 &= \Phi_1 \Psi_2 + \Phi_2 \Psi_1 \\ \dots & \quad \dots \\ z^j : \quad \Psi_j &= \Phi_1 \Psi_{j-1} + \Phi_2 \Psi_{j-2}. \end{aligned}$$

The last equation shows how to compute the sequence  $\{\Psi_j\}$  recursively:

$$\Psi_1 = \begin{pmatrix} 0.8 & -0.5 \\ 0.1 & -0.5 \end{pmatrix} \quad \Psi_2 = \begin{pmatrix} 0.29 & -0.45 \\ -0.17 & 0.50 \end{pmatrix} \quad \Psi_3 = \begin{pmatrix} 0.047 & -0.310 \\ -0.016 & -0.345 \end{pmatrix} \quad \dots$$

## 12.4 Computation of the Covariance Function of a Causal VAR Process

As in the univariate case, it is important to be able to compute the covariance and the correlation function of VARMA process (see Section 2.4). As explained in Remark 12.1 we will concentrate on VAR processes. Consider first the case of a causal VAR(1) process:

$$X_t = \Phi X_{t-1} + Z_t \quad Z_t \sim \text{WN}(0, \Sigma).$$

Multiplying the above equation first by  $X_t'$  and then successively by  $X_{t-h}'$  from the left,  $h = 1, 2, \dots$ , and taking expectations, we obtain the Yule-Walker equations:

$$\begin{aligned}\mathbb{E}(X_t X_t') &= \Gamma(0) = \Phi \mathbb{E}(X_{t-1} X_t') + \mathbb{E}(Z_t X_t') = \Phi \Gamma(-1) + \Sigma, \\ \mathbb{E}(X_t X_{t-h}') &= \Gamma(h) = \Phi \mathbb{E}(X_{t-1} X_{t-h}') + \mathbb{E}(Z_t X_{t-h}') = \Phi \Gamma(h-1).\end{aligned}$$

Knowing  $\Gamma(0)$  and  $\Phi$ ,  $\Gamma(h)$ ,  $h > 0$ , can be computed recursively from the second equation as

$$\Gamma(h) = \Phi^h \Gamma(0), \quad h = 1, 2, \dots \quad (12.3)$$

Given  $\Phi$  and  $\Sigma$ , we can compute  $\Gamma(0)$ . For  $h = 1$ , the second equation above implies  $\Gamma(1) = \Phi \Gamma(0)$ . Inserting this expression in the first equation and using the fact that  $\Gamma(-1) = \Gamma(1)'$ , we get an equation in  $\Gamma(0)$ :

$$\Gamma(0) = \Phi \Gamma(0) \Phi' + \Sigma.$$

This equation can be solved for  $\Gamma(0)$ :

$$\begin{aligned}\text{vec} \Gamma(0) &= \text{vec}(\Phi \Gamma(0) \Phi') + \text{vec} \Sigma \\ &= (\Phi \otimes \Phi) \text{vec} \Gamma(0) + \text{vec} \Sigma,\end{aligned}$$

where  $\otimes$  and “vec” denote the Kronecker-product and the vec-operator, respectively.<sup>3</sup> Thus,

$$\text{vec} \Gamma(0) = (I_{n^2} - \Phi \otimes \Phi)^{-1} \text{vec} \Sigma. \quad (12.4)$$

The assumption that  $\{X_t\}$  is causal with respect to  $\{Z_t\}$  guarantees that the eigenvalues of  $\Phi \otimes \Phi$  are strictly smaller than one in absolute value, implying that  $I_{n^2} - \Phi \otimes \Phi$  is invertible.<sup>4</sup>

If the process is a causal VAR(p) process the covariance function can be found in two ways. The first one rewrites the process in companion form as a VAR(1) process and applies the procedure just outlined. The second way relies on the Yule-Walker equation. This equation is obtained by multiplying the stochastic difference equation from the left by  $X_t'$  and then successively by  $X_{t-h}'$ ,  $h > 0$ , and taking expectations:

$$\begin{aligned}\Gamma(0) &= \Phi_1 \Gamma(-1) + \dots + \Phi_p \Gamma(-p) + \Sigma, \\ &= \Phi_1 \Gamma(1)' + \dots + \Phi_p \Gamma(p)' + \Sigma, \\ \Gamma(h) &= \Phi_1 \Gamma(h-1) + \dots + \Phi_p \Gamma(h-p).\end{aligned} \quad (12.5)$$

<sup>3</sup>The vec-operator stacks the column of a  $n \times m$  matrix to a column vector of dimension  $nm$ . The properties of  $\otimes$  and vec can be found, e.g. in Magnus and Neudecker (1988).

<sup>4</sup>If the eigenvalues of  $\Phi$  are  $\lambda_i$ ,  $i = 1, \dots, n$ , then the eigenvalues of  $\Phi \otimes \Phi$  are  $\lambda_i \lambda_j$ ,  $i, j = 1, \dots, n$  (see Magnus and Neudecker (1988)).

The second equation can be used to compute  $\Gamma(h)$ ,  $h \geq p$ , recursively taking  $\Phi_1, \dots, \Phi_p$  and the starting values  $\Gamma(p-1), \dots, \Gamma(0)$  as given. The starting value can be retrieved by transforming the VAR(p) model into the companion form and proceeding as explained above.

### Example

We illustrate the computation of the covariance function using the same example as in Section 12.3. First, we transform the model into the companion form:

$$Y_t = \begin{pmatrix} X_{1,t} \\ X_{2,t} \\ X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} = \begin{pmatrix} 0.8 & -0.5 & -0.3 & -0.3 \\ 0.1 & -0.5 & -0.2 & 0.3 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{1,t-2} \\ X_{2,t-2} \end{pmatrix} + \begin{pmatrix} Z_{1,t} \\ Z_{2,t} \\ 0 \\ 0 \end{pmatrix}.$$

Equation (12.4) implies that  $\Gamma_Y(0)$  is given by:

$$\text{vec}\Gamma_Y(0) = \text{vec} \begin{pmatrix} \Gamma_X(0) & \Gamma_X(1) \\ \Gamma_X(1)' & \Gamma_X(0) \end{pmatrix} = (I_{16} - \Phi \otimes \Phi)^{-1} \text{vec} \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$$

so that  $\Gamma_X(0)$  and  $\Gamma_X(1)$  become:

$$\Gamma_X(0) = \begin{pmatrix} 2.4201 & 0.5759 \\ 0.5759 & 3.8978 \end{pmatrix} \quad \Gamma_X(1) = \begin{pmatrix} 1.3996 & -0.5711 \\ -0.4972 & -2.5599 \end{pmatrix}.$$

The other covariance matrices can then be computed recursively according to equation (12.5):

$$\begin{aligned} \Gamma_X(2) &= \Phi_1 \Gamma_X(1) + \Phi_2 \Gamma_X(0) = \begin{pmatrix} 0.4695 & -0.5191 \\ 0.0773 & 2.2770 \end{pmatrix}, \\ \Gamma_X(3) &= \Phi_1 \Gamma_X(2) + \Phi_2 \Gamma_X(1) = \begin{pmatrix} 0.0662 & -0.6145 \\ -0.4208 & -1.8441 \end{pmatrix}. \end{aligned}$$

### Appendix: Autoregressive Final Form

Definition 12.1 defined the VARMA process  $\{X_t\}$  as a solution to the corresponding multivariate stochastic difference equation (12.1). However, as pointed out by Zellner and Palm (1974) there is an equivalent representation in the form of  $n$  univariate ARMA processes, one for each  $X_{it}$ . Formally, these representations, also called *autoregressive final form* or *transfer function form* (Box and Jenkins, 1976), can be written as

$$\det \Phi(L) X_{it} = [\Phi^*(L)\Theta(L)]_{i\bullet} Z_t$$

where the index  $i\bullet$  indicates the  $i$ -th row of  $\Phi^*(L)\Theta(L)$ . Thereby  $\Phi^*(L)$  denotes the adjugate matrix of  $\Phi(L)$ .<sup>5</sup> Thus each variable in  $X_t$  may be investigated separately as an univariate ARMA process. Thereby the autoregressive part will be the same for each variable. Note, however, that the moving-average processes will be correlated across variables.

The disadvantage of this approach is that it involves rather long AR and MA lags as will become clear from the following example.<sup>6</sup> Take a simple two-dimensional VAR of order one, i.e.  $X_t = \Phi X_{t-1} + Z_t$ ,  $Z_t \sim \text{WN}(0, \Sigma)$ . Then the implied univariate processes will be ARMA(2,1) processes. After some straightforward manipulations we obtain:

$$\begin{aligned} (1 - (\phi_{11} + \phi_{22})L + (\phi_{11}\phi_{22} - \phi_{12}\phi_{21})L^2)X_{1t} &= Z_{1t} - \phi_{22}Z_{1,t-1} + \phi_{12}Z_{2,t-1}, \\ (1 - (\phi_{11} + \phi_{22})L + (\phi_{11}\phi_{22} - \phi_{12}\phi_{21})L^2)X_{2t} &= \phi_{21}Z_{1,t-1} + Z_{2t} - \phi_{11}Z_{2,t-1}. \end{aligned}$$

It can be shown by the means given in Sections 1.4.3 and 1.5.1 that the right hand sides are observationally equivalent to MA(1) processes.

---

<sup>5</sup>The elements of the adjugate matrix  $A^*$  of some matrix  $A$  are given by  $[A^*]_{ij} = (-1)^{i+j}M_{ij}$  where  $M_{ij}$  is the minor (minor determinant) obtained by deleting the  $i$ -th column and the  $j$ -th row of  $A$  (Meyer, 2000, p. 477).

<sup>6</sup>The degrees of the AR and the MA polynomial can be as large as  $np$  and  $(n-1)p+q$ , respectively.

# Chapter 13

## Estimation of Vector Autoregressive Models

### 13.1 Introduction

In this chapter we derive the Least-Squares (LS) estimator for vector autoregressive (VAR) models and its asymptotic distribution. For this end, we have to make several assumption which we maintain throughout this chapter.

**Assumption 13.1.** *The VAR process  $\{X_t\}$  is generated by*

$$\begin{aligned} \Phi(L)X_t &= Z_t \\ X_t - \Phi_1 X_{t-1} - \dots - \Phi_p X_{t-p} &= Z_t \quad \text{with } Z_t \sim \text{WN}(0, \Sigma), \end{aligned}$$

$\Sigma$  nonsingular, and admits a stationary and causal representation with respect to  $\{Z_t\}$ :

$$X_t = Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \dots = \sum_{j=0}^{\infty} \Psi_j Z_t = \Psi(L)Z_t$$

with  $\sum_{j=0}^{\infty} \|\Psi_j\| < \infty$ .

**Assumption 13.2.** *The residual process  $\{Z_t\}$  is not only white noise, but also independently and identically distributed:*

$$Z_t \sim \text{IID}(0, \Sigma).$$

**Assumption 13.3.** *All fourth moments of  $Z_t$  exist. In particular, there exists a finite constant  $c > 0$  such that*

$$\mathbb{E}(Z_{it}Z_{jt}Z_{kt}Z_{lt}) \leq c \quad \text{for all } i, j, k, l = 1, 2, \dots, n, \text{ and for all } t.$$

Note that the moment condition is automatically fulfilled by Gaussian processes. For the ease of exposition, we omit a constant in the VAR. Thus, we consider the demeaned process.

## 13.2 The Least-Squares Estimator

Let us denote by  $\phi_{ij}^{(k)}$  the  $(i, j)$ -th element of the matrix  $\Phi_k$ ,  $k = 1, 2, \dots, p$ , then the  $i$ -th equation,  $i = 1, \dots, n$ , can be written as

$$X_{it} = \phi_{i1}^{(1)} X_{1,t-1} + \dots + \phi_{in}^{(1)} X_{n,t-1} + \dots + \phi_{i1}^{(p)} X_{1,t-p} + \dots + \phi_{in}^{(p)} X_{n,t-p} + Z_{it}.$$

We can view this equation as a regression equation of  $X_{it}$  on all lagged variables  $X_{1,t-1}, \dots, X_{n,t-1}, \dots, X_{1,t-p}, \dots, X_{n,t-p}$  with error term  $Z_{it}$ . Note that the regressors are the same for each equation. The  $np$  regressors have coefficient vector  $(\phi_{i1}^{(1)}, \dots, \phi_{in}^{(1)}, \dots, \phi_{i1}^{(p)}, \dots, \phi_{in}^{(p)})'$ . Thus, the complete VAR(p) model has  $n^2 p$  coefficients in total to be estimated. In addition, there are  $n(n+1)/2$  independent elements of the covariance matrix  $\Sigma$  that have to be estimated too.

It is clear that the  $n$  different equations are linked through the regressors and the errors terms which in general have non-zero covariances  $\sigma_{ij} = \mathbb{E}Z_{it}Z_{jt}$ . Hence, it seems warranted to take a systems approach and to estimate all equations of the VAR jointly. Below, we will see that an equation-by-equation approach is, however, still appropriate.

Suppose that we have  $T+p$  observations with  $t = -p+1, \dots, 0, 1, \dots, T$ , then we can write the regressor matrix for each equation compactly as a  $T \times np$  matrix  $\mathbf{X}$ :

$$\mathbf{X} = \begin{pmatrix} X_{1,0} & \dots & X_{n,0} & \dots & X_{1,-p+1} & \dots & X_{n,-p+1} \\ X_{1,1} & \dots & X_{n,1} & \dots & X_{1,-p+2} & \dots & X_{n,-p+2} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{1,T-1} & \dots & X_{n,T-1} & \dots & X_{1,T-p} & \dots & X_{n,T-p} \end{pmatrix}.$$

Using this notation, we can write the VAR for observations  $t = 1, 2, \dots, T$  as

$$\underbrace{(X_1, X_2, \dots, X_T)}_{=Y} = \underbrace{(\Phi_1, \Phi_2, \dots, \Phi_p)}_{=\Phi} \underbrace{\begin{pmatrix} X_0 & X_1 & \dots & X_{T-1} \\ X_{-1} & X_0 & \dots & X_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{-p+1} & X_{-p+2} & \dots & X_{T-p} \end{pmatrix}}_{=\mathbf{X}'} + \underbrace{(Z_1, Z_2, \dots, Z_T)}_{=Z}$$

or more compactly

$$Y = \Phi \mathbf{X}' + Z.$$

There are two ways to bring this equation system in the usual multivariate regression framework. One can either arrange the data according to observations or according to equations. Ordered in terms of observations yields:

$$\text{vec } Y = \text{vec}(\Phi \mathbf{X}') + \text{vec } Z = (\mathbf{X} \otimes I_n) \text{vec } \Phi + \text{vec } Z \quad (13.1)$$

where  $\text{vec } Y = (X_{11}, X_{21}, \dots, X_{n1}, X_{12}, X_{22}, \dots, X_{n2}, \dots, X_{1T}, X_{2T}, \dots, X_{nT})'$ . If the data are arranged equation by equation, the dependent variable becomes  $\text{vec } Y' = (X_{11}, X_{12}, \dots, X_{1T}, X_{21}, X_{22}, \dots, X_{2T}, \dots, X_{n1}, X_{n2}, \dots, X_{nT})'$ . As both representations, obviously, contain the same information, there exists a  $nT \times nT$  permutation or commutation matrix  $K_{nT}$  such that  $\text{vec } Y' = K_{nT} \text{vec } Y$ . Using the computation rules for the Kronecker product, the vec operator, and the permutation matrix (see Magnus and Neudecker, 1988), we get for the ordering in terms of equations

$$\begin{aligned} \text{vec } Y' &= K_{nT} \text{vec } Y = K_{nT} (\text{vec}(\Phi \mathbf{X}') + \text{vec } Z) \\ &= K_{nT} (\mathbf{X} \otimes I_n) \text{vec } \Phi + K_{nT} \text{vec } Z \\ &= (I_n \otimes \mathbf{X}) K_{n^2p} \text{vec } \Phi + K_{nT} \text{vec } Z \\ &= (I_n \otimes \mathbf{X}) \text{vec } \Phi' + \text{vec } Z' \end{aligned} \quad (13.2)$$

where  $K_{n^2p}$  is the corresponding  $n^2 \times p$  permutation matrix relating  $\text{vec } \Phi$  and  $\text{vec } \Phi'$ .

The error terms of the different equations are correlated because, in general, the covariances  $\sigma_{ij} = \mathbb{E}Z_{it}Z_{jt}$  are nonzero. In the case of an arrangement by observation the covariance matrix of the error term  $\text{vec } Z$  is

$$\begin{aligned} \mathbb{V} \text{vec } Z &= \mathbb{E}(\text{vec } Z)(\text{vec } Z)' \\ &= \begin{pmatrix} \sigma_1^2 & \dots & \sigma_{1n} & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_n^2 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & 0 & 0 & \sigma_1^2 & \dots & \sigma_{1n} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_{n1} & \dots & \sigma_n^2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \sigma_1^2 & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \sigma_{n1} & \dots & \sigma_n^2 \end{pmatrix} = I_T \otimes \Sigma. \end{aligned}$$

In the second case, the arrangement by equation, the covariance matrix of the error term  $\text{vec } Z'$  is

$$\begin{aligned} \mathbb{V} \text{vec } Z' &= \mathbb{E}(\text{vec } Z')(\text{vec } Z')' \\ &= \begin{pmatrix} \sigma_1^2 & \dots & 0 & \sigma_{12} & \dots & 0 & \dots & \sigma_{1n} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_1^2 & 0 & \dots & \sigma_{12} & \dots & 0 & \dots & \sigma_{1n} \\ \sigma_{21} & \dots & 0 & \sigma_2^2 & \dots & 0 & \dots & \sigma_{2n} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{21} & 0 & \dots & \sigma_2^2 & \dots & 0 & \dots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \dots & 0 & \sigma_{n2} & \dots & 0 & \dots & \sigma_n^2 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{n1} & 0 & \dots & \sigma_{n2} & \dots & 0 & \dots & \sigma_n^2 \end{pmatrix} = \Sigma \otimes I_T. \end{aligned}$$

Given that the covariance matrix is not a multiple of the identity matrix, efficient estimation requires the use of generalized least squares (GLS). The GLS estimator minimizes the weighted sum of squared errors

$$S(\text{vec } \Phi) = (\text{vec } Z)'(I_T \otimes \Sigma)^{-1}(\text{vec } Z) \longrightarrow \min_{\Phi}.$$

The solution of this minimization problem can be found in standard econometric textbooks like (Dhrymes, 1978; Greene, 2008; Hamilton, 1994a) and is given by

$$\begin{aligned} (\text{vec } \hat{\Phi})_{\text{GLS}} &= ((\mathbf{X} \otimes I_n)'(I_T \otimes \Sigma)^{-1}(\mathbf{X} \otimes I_n))^{-1}(\mathbf{X} \otimes I_n)'(I_T \otimes \Sigma)^{-1} \text{vec } Y \\ &= ((\mathbf{X}' \otimes I_n)(I_T \otimes \Sigma^{-1})(\mathbf{X} \otimes I_n))^{-1}(\mathbf{X}' \otimes I_n)(I_T \otimes \Sigma^{-1}) \text{vec } Y \\ &= ((\mathbf{X}' \otimes \Sigma^{-1})(\mathbf{X} \otimes I_n))^{-1}(\mathbf{X}' \otimes \Sigma^{-1}) \text{vec } Y \\ &= ((\mathbf{X}'\mathbf{X}) \otimes \Sigma^{-1})^{-1}(\mathbf{X}' \otimes \Sigma^{-1}) \text{vec } Y \\ &= ((\mathbf{X}'\mathbf{X})^{-1} \otimes \Sigma)(\mathbf{X}' \otimes \Sigma^{-1}) \text{vec } Y = (((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n) \text{vec } Y \\ &= (\text{vec } \hat{\Phi})_{\text{OLS}} \end{aligned}$$

As the covariance matrix  $\Sigma$  cancels, the GLS and the OLS-estimator deliver numerically exactly the same solution. The reason for this result is that the regressors are the same in each equation. If this does not hold, for example when some coefficients are set a priori to zero, efficient estimation would require the use of GLS.

Further insights can be gained by rewriting the estimation problem in terms of the arrangement by equation (see equation (13.2)). For this purpose, multiply the above estimator from the left by the commutation matrix  $K_{n^2p}$ :<sup>1</sup>

$$\begin{aligned} (\text{vec } \widehat{\Phi}')_{\text{OLS}} &= K_{n^2p}(\text{vec } \widehat{\Phi})_{\text{OLS}} = K_{n^2p} \left( ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n \right) \text{vec } Y \\ &= (I_n \otimes ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) K_{nT} \text{vec } Y = (I_n \otimes ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) \text{vec } Y'. \end{aligned}$$

This can be written in a more explicit form as

$$\begin{aligned} (\text{vec } \widehat{\Phi}')_{\text{OLS}} &= \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' & 0 & \dots & 0 \\ 0 & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{pmatrix} \text{vec } Y' \\ &= \begin{pmatrix} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y_1 & 0 & \dots & 0 \\ 0 & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y_n \end{pmatrix} \end{aligned}$$

where  $Y_i$ ,  $i = 1, \dots, n$ , stacks the observations of the  $i$ -th variable such that  $Y_i = (X_{i1}, X_{i2}, \dots, X_{iT})'$ . Thus, the estimation of VAR as a system can be broken down into the estimation of  $n$  regression equations with dependent variable  $X_{it}$ . Each of these equations can then be estimated by OLS.

Thus, we have proven that

$$\text{vec } \widehat{\Phi} = (\text{vec } \widehat{\Phi})_{\text{GLS}} = (\text{vec } \widehat{\Phi})_{\text{OLS}} = \left( ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n \right) \text{vec } Y, \quad (13.3)$$

$$\text{vec } \widehat{\Phi}' = (\text{vec } \widehat{\Phi}')_{\text{GLS}} = (\text{vec } \widehat{\Phi}')_{\text{OLS}} = (I_n \otimes ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) \text{vec } Y'. \quad (13.4)$$

The least squares estimator can also be rewritten without the use of the *vec*-operator:

$$\widehat{\Phi} = YX(X'X)^{-1}.$$

Under the assumptions stated in the Introduction 13.1, these estimators are consistent and asymptotically normal.

**Theorem 13.1** (Asymptotic distribution of OLS estimator). *Under the assumption stated in the Introduction 13.1, it holds that*

$$\text{plim } \widehat{\Phi} = \Phi$$

<sup>1</sup>Alternatively, one could start from scratch and investigate the minimization problem  $S(\text{vec } \Phi') = (\text{vec } Z')'(\Sigma^{-1} \otimes I_T)(\text{vec } Z') \rightarrow \min_{\Phi}$ .

and that

$$\text{by observation:} \quad \sqrt{T} \left( \text{vec } \hat{\Phi} - \text{vec } \Phi \right) \xrightarrow{d} N \left( 0, \Gamma_p^{-1} \otimes \Sigma \right),$$

respectively,

$$\text{by equation:} \quad \sqrt{T} \left( \text{vec } \hat{\Phi}' - \text{vec } \Phi' \right) \xrightarrow{d} N \left( 0, \Sigma \otimes \Gamma_p^{-1} \right)$$

where  $\Gamma_p = \text{plim } \frac{1}{T} (\mathbf{X}'\mathbf{X})$ .

*Proof.* See Section 13.3. □

In order to make use of this result in practice, we have to replace the matrices  $\Sigma$  and  $\Gamma_p$  by some estimate. A natural consistent estimate of  $\Gamma_p$  is given according to Proposition 13.1 by

$$\hat{\Gamma}_p = \frac{\mathbf{X}'\mathbf{X}}{T}.$$

In analogy to the multivariate regression model, a natural estimator for  $\Sigma$  can be obtained from the Least-Squares residuals  $\hat{Z}$ :

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T \hat{Z}_t \hat{Z}_t' = \frac{\hat{Z} \hat{Z}'}{T} = \frac{(Y - \hat{\Phi} \mathbf{X}') (Y - \hat{\Phi} \mathbf{X}')'}{T}.$$

The property of this estimator is summarized in the proposition below.

**Theorem 13.2.** *Under the condition of Theorem 13.1*

$$\text{plim } \sqrt{T} \left( \hat{\Sigma} - \frac{ZZ'}{T} \right) = 0$$

*Proof.* See Section 13.3 □

An alternative, but asymptotically equivalent estimator  $\tilde{\Sigma}$  is obtained by adjusting  $\hat{\Sigma}$  for the degrees of freedom:

$$\tilde{\Sigma} = \frac{T}{T - np} \hat{\Sigma}. \tag{13.5}$$

If the VAR contains a constant, as is normally the case in practice, the degrees of freedom correction should be  $T - np - 1$ .

Small sample inference with respect to the parameters  $\Phi$  can therefore be carried out using the approximate distribution

$$\text{vec } \hat{\Phi} \sim N \left( \text{vec } \Phi, \hat{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1} \right). \tag{13.6}$$

This implies that hypothesis testing can be carried out using the conventional t- and F-statistics. From a system perspective, the appropriate degree of freedom for the t-ratio would be  $nT - n^2p - n$ , taking a constant in each equation into account. However, as that the system can be estimated on an equation by equation basis, it seems reasonable to use  $T - np - 1$  instead. This corresponds to a multivariate regression setting with  $T$  observation and  $np + 1$  regressors, including a constant.

However, as in the univariate case the Gauss Markov theorem does not apply because the lagged regressors are correlated with past error terms. This results in biased estimates in small samples. The amount of the bias can be assessed and corrected either by analytical or bootstrap methods. For an overview, a comparison of the different corrections proposed in the literature, and further references see Engsteg and Pedersen (2014).

### 13.3 Proofs of the Asymptotic Properties of the Least-Squares Estimator

**Lemma 13.1.** *Given the assumptions made in Section 13.1, the process  $\{\text{vec } Z_{t-j}Z'_{t-i}\}$ ,  $i, j \in \mathbb{Z}$  and  $i \neq j$ , is white noise.*

*Proof.* Using the independence assumption of  $\{Z_t\}$ , we immediately get

$$\begin{aligned} \mathbb{E} \text{vec } Z_{t-j}Z'_{t-i} &= \mathbb{E}(Z_{t-i} \otimes Z_{t-j}) = 0, \\ \mathbb{V}(\text{vec } Z_{t-j}Z'_{t-i}) &= \mathbb{E}((Z_{t-i} \otimes Z_{t-j})(Z_{t-i} \otimes Z_{t-j})') \\ &= \mathbb{E}((Z_{t-i}Z'_{t-i}) \otimes (Z_{t-j}Z'_{t-j})) = \Sigma \otimes \Sigma, \\ \Gamma_{\text{vec } Z_{t-j}Z'_{t-i}}(h) &= \mathbb{E}((Z_{t-i} \otimes Z_{t-j})(Z_{t-i-h} \otimes Z_{t-j-h})') \\ &= \mathbb{E}((Z_{t-i}Z'_{t-i-h}) \otimes (Z_{t-j}Z'_{t-j-h})) = 0, \quad h \neq 0. \end{aligned}$$

□

Under the assumption put forward in the Introduction,  $\frac{1}{T}(\mathbf{X}'\mathbf{X})$  converges in probability for  $T \rightarrow \infty$  to a  $np \times np$  matrix  $\mathbf{\Gamma}_p$ . This matrix consists of  $p^2$  blocks where each  $(i, j)$ -th block corresponds to the covariance matrix  $\Gamma(i - j)$ . Thus we have the following proposition:

**Proposition 13.1.** *Under the assumption stated in the Introduction 13.1*

$$\frac{\mathbf{X}'\mathbf{X}}{T} \xrightarrow{p} \mathbf{\Gamma}_p = \begin{pmatrix} \Gamma(0) & \Gamma(1) & \dots & \Gamma(p-1) \\ \Gamma'(1) & \Gamma(0) & \dots & \Gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma'(p-1) & \Gamma'(p-2) & \dots & \Gamma(0) \end{pmatrix}.$$

with  $\Gamma_p$  being nonsingular.

*Proof.* Write  $\frac{1}{T}(\mathbf{X}'\mathbf{X})$  as

$$\frac{\mathbf{X}'\mathbf{X}}{T} = \begin{pmatrix} \hat{\Gamma}(0) & \hat{\Gamma}(1) & \dots & \hat{\Gamma}(p-1) \\ \hat{\Gamma}'(1) & \hat{\Gamma}(0) & \dots & \hat{\Gamma}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Gamma}'(p-1) & \hat{\Gamma}'(p-2) & \dots & \hat{\Gamma}(0) \end{pmatrix}$$

where

$$\hat{\Gamma}(h) = \frac{1}{T} \sum_{t=0}^{T-1} X_t X'_{t-h}, \quad h = 0, 1, \dots, p-1.$$

We will show that each component  $\hat{\Gamma}(h)$  of  $\frac{1}{T}\mathbf{X}'\mathbf{X}$  converges in probability to  $\Gamma(h)$ . Taking the causal representation of  $\{X_t\}$  into account

$$\begin{aligned} \hat{\Gamma}(h) &= \frac{1}{T} \sum_{t=0}^{T-1} X_t X'_{t-h} = \frac{1}{T} \sum_{t=0}^{T-1} \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \Psi_j Z_{t-j} Z'_{t-h-i} \Psi'_i \\ &= \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \Psi_j \left( \frac{1}{T} \sum_{t=0}^{T-1} Z_{t-j} Z'_{t-h-i} \right) \Psi'_i \\ &= \sum_{j=0}^{\infty} \sum_{i=h}^{\infty} \Psi_j \left( \frac{1}{T} \sum_{t=0}^{T-1} Z_{t-j} Z'_{t-i} \right) \Psi'_{i-h}. \end{aligned}$$

According to Lemma 13.1 above  $\{Z_{t-j} Z'_{t-i}\}$ ,  $i \neq j$ , is white noise. Thus,

$$\frac{1}{T} \sum_{t=0}^{T-1} Z_{t-j} Z'_{t-i} \xrightarrow{P} 0, \quad i \neq j,$$

according to Theorem 11.1. Hence, for  $m$  fixed,

$$G_m(h) = \sum_{j=0}^m \sum_{\substack{i=h \\ i \neq j}}^{m+h} \Psi_j \left( \frac{1}{T} \sum_{t=0}^{T-1} Z_{t-j} Z'_{t-i} \right) \Psi'_{i-h} \xrightarrow{P} 0.$$

Taking absolute values and expectations element-wise,

$$\begin{aligned}
 \mathbb{E} |G_\infty(h) - G_m(h)| &= \mathbb{E} \left| \sum_{\substack{j>m \text{ or } i>m+h \\ i \neq j}} \Psi_j \left( \frac{1}{T} \sum_{t=0}^{T-1} Z_{t-j} Z'_{t-i} \right) \Psi'_{i-h} \right| \\
 &\leq \sum_{\substack{j>m \text{ or } i>m+h \\ i \neq j}} |\Psi_j| \left( \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} |Z_{t-j} Z'_{t-i}| \right) |\Psi'_{i-h}| \\
 &\leq \sum_{\substack{j>m \text{ or } i>m+h \\ i \neq j}} |\Psi_j| (\mathbb{E} |Z_1 Z'_2|) |\Psi'_{i-h}| \\
 &\leq \sum_{\substack{j>m \text{ or } i>m \\ i \neq j}} |\Psi_j| (\mathbb{E} |Z_1 Z'_2|) |\Psi'_i| \\
 &\leq \sum_{j>m} |\Psi_j| \left( \mathbb{E} |Z_1 Z'_2| \sum_i |\Psi'_i| \right) \\
 &\quad + \left( \sum_j |\Psi_j| \mathbb{E} |Z_1 Z'_2| \right) \sum_{i>m} |\Psi'_i|
 \end{aligned}$$

As the bound is independent of  $T$  and converges to 0 as  $m \rightarrow \infty$ , we have

$$\lim_{m \rightarrow \infty} \limsup_{T \rightarrow \infty} \mathbb{E} |G_\infty(h) - G_m(h)| = 0.$$

The Basic Approximation Theorem C.14 then establishes that

$$G_\infty(h) \xrightarrow{P} 0.$$

Henceforth

$$\begin{aligned}
 \hat{\Gamma}(h) &= G_\infty(h) + \sum_{j=h}^{\infty} \Psi_j \left( \frac{1}{T} \sum_{t=0}^{T-1} Z_{t-j} Z'_{t-j} \right) \Psi'_{j-h} \\
 &= G_\infty(h) + \sum_{j=h}^{\infty} \Psi_j \left( \frac{1}{T} \sum_{t=0}^{T-1} Z_t Z'_t \right) \Psi'_{j-h} + \text{remainder}
 \end{aligned}$$

where the remainder only depends on initial conditions<sup>2</sup> and is therefore negligible as  $T \rightarrow \infty$ . As

$$\frac{1}{T} \sum_{t=0}^{T-1} Z_t Z'_t \xrightarrow{P} \Sigma,$$

---

<sup>2</sup>See the proof of Theorem 11.2.2 in Brockwell and Davis (1991) for details.

we finally get

$$\widehat{\Gamma}(h) \xrightarrow{P} \sum_{j=h}^{\infty} \Psi_j \Sigma \Psi'_{j-h} = \Gamma(h).$$

The last equality follows from Theorem 10.2.  $\square$

**Proposition 13.2.** *Under the assumption stated in the Introduction 13.1*

$$\begin{aligned} \frac{1}{\sqrt{T}} \sum_{t=1}^T \text{vec}(Z_t X'_{t-1}, Z_t X'_{t-2}, \dots, Z_t X'_{t-p}) \\ = \frac{1}{\sqrt{T}} \text{vec}(Z\mathbf{X}) = \frac{1}{\sqrt{T}} (\mathbf{X}' \otimes I_n) \text{vec } Z \\ \xrightarrow{d} N(0, \mathbf{\Gamma}_p \otimes \Sigma) \end{aligned}$$

*Proof.* The idea of the proof is to approximate  $\{X_t\}$  by some simpler process  $\{X_t^{(m)}\}$  which allows the application of the CLT for dependent processes (Theorem C.13). This leads to an asymptotic distribution which by the virtue of the Basic Approximation Theorem C.14 converges to the asymptotic distribution of the original process. Define  $X_t^{(m)}$  as the truncated process from the causal presentation of  $X_t$ :

$$X_t^{(m)} = Z_t + \Psi_1 Z_{t-1} + \dots + \Psi_m Z_{t-m}, \quad m = p, p+1, p+2, \dots$$

Using this approximation, we can then define the process  $\{Y_t^{(m)}\}$  as

$$Y_t^{(m)} = \text{vec} \left( Z_t X_{t-1}^{(m)'} , Z_t X_{t-2}^{(m)'} , \dots , Z_t X_{t-p}^{(m)'} \right) = \begin{pmatrix} X_{t-1}^{(m)} \\ X_{t-2}^{(m)} \\ \vdots \\ X_{t-p}^{(m)} \end{pmatrix} \otimes Z_t.$$

Due to the independence of  $\{Z_t\}$  this process is a mean zero white noise process, but is clearly not independent. It is easy to see that the process is

actually  $(m + p)$ -dependent with variance  $\mathbf{V}_m$  given by

$$\begin{aligned}
\mathbf{V}_m &= \mathbb{E}Y_t^{(m)}Y_t^{(m)'} = \mathbb{E} \left( \begin{pmatrix} X_{t-1}^{(m)} \\ X_{t-2}^{(m)} \\ \vdots \\ X_{t-p}^{(m)} \end{pmatrix} \otimes Z_t \begin{pmatrix} X_{t-1}^{(m)} \\ X_{t-2}^{(m)} \\ \vdots \\ X_{t-p}^{(m)} \end{pmatrix} \right)' \\
&= \mathbb{E} \left( \begin{pmatrix} X_{t-1}^{(m)} \\ X_{t-2}^{(m)} \\ \dots \\ X_{t-p}^{(m)} \end{pmatrix} \begin{pmatrix} X_{t-1}^{(m)} \\ X_{t-2}^{(m)} \\ \dots \\ X_{t-p}^{(m)} \end{pmatrix}' \right) \otimes \mathbb{E}Z_tZ_t' \\
&= \begin{pmatrix} \Gamma^{(m)}(0) & \Gamma^{(m)}(1) & \dots & \Gamma^{(m)}(p-1) \\ \Gamma^{(m)}(1)' & \Gamma^{(m)}(0) & \dots & \Gamma^{(m)}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma^{(m)}(p-1)' & \Gamma^{(m)}(p-2)' & \dots & \Gamma^{(m)}(0) \end{pmatrix} \otimes \Sigma \\
&= \mathbf{\Gamma}_p^{(m)} \otimes \Sigma
\end{aligned}$$

where  $\mathbf{\Gamma}_p^{(m)}$  is composed of

$$\begin{aligned}
\Gamma^{(m)}(h) &= \mathbb{E}X_{t-1}^{(m)}X_{t-1-h}^{(m)'} \\
&= \mathbb{E}(Z_{t-1} + \Psi_1Z_{t-2} + \dots + \Psi_mZ_{t-1-m}) \\
&\quad (Z_{t-1-h} + \Psi_1Z_{t-2-h} + \dots + \Psi_mZ_{t-1-m-h})' \\
&= \sum_{j=h}^m \Psi_j\Sigma\Psi_j', \quad h = 0, 1, \dots, p-1.
\end{aligned}$$

Thus, we can invoke the CLT for  $(m + p)$ -dependent process (see Theorem C.13) to establish that

$$\sqrt{T} \left( \frac{1}{T} \sum_{t=1}^T Y_t^{(m)} \right) \xrightarrow{d} N(0, \mathbf{V}_m).$$

For  $m \rightarrow \infty$ ,  $\Gamma^{(m)}(h)$  converges to  $\Gamma(h)$  and thus  $\mathbf{\Gamma}_p^{(m)}$  to  $\mathbf{\Gamma}_p$ . Therefore,  $\mathbf{V}_m \rightarrow \mathbf{\Gamma}_p \otimes \Sigma$ .

The variance of the approximation error is equal to

$$\begin{aligned}
& \mathbb{V} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T \left( \text{vec}(Z_t X'_{t-1}, \dots, Z_t X'_{t-p}) - Y_t^{(m)} \right) \right) \\
&= \frac{1}{T} \mathbb{V} \left( \sum_{t=1}^T \text{vec} \left( Z_t (X_{t-1} - X_{t-1}^{(m)})', \dots, Z_t (X_{t-p} - X_{t-p}^{(m)})' \right) \right) \\
&= \frac{1}{T} \mathbb{V} \left( \sum_{t=1}^T \text{vec} \left( Z_t \left( \sum_{j=m+1}^{\infty} \Psi_j Z_{t-1-j} \right)', \dots, Z_t \left( \sum_{j=m+1}^{\infty} \Psi_j Z_{t-p-j} \right)' \right) \right) \\
&= \mathbb{V} \left( \text{vec} \left( Z_t \left( \sum_{j=m+1}^{\infty} \Psi_j Z_{t-1-j} \right)', \dots, Z_t \left( \sum_{j=m+1}^{\infty} \Psi_j Z_{t-p-j} \right)' \right) \right) \\
&= \mathbb{E} \left( \left( \begin{pmatrix} \sum_{j=m+1}^{\infty} \Psi_j Z_{t-1-j} \\ \vdots \\ \sum_{j=m+1}^{\infty} \Psi_j Z_{t-p-j} \end{pmatrix} \otimes Z_t \right) \left( \begin{pmatrix} \sum_{j=m+1}^{\infty} \Psi_j Z_{t-1-j} \\ \vdots \\ \sum_{j=m+1}^{\infty} \Psi_j Z_{t-p-j} \end{pmatrix} \otimes Z_t \right)' \right) \\
&= \begin{pmatrix} \sum_{j=m+1}^{\infty} \Psi_j \Sigma \Psi_j' & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \sum_{j=m+1}^{\infty} \Psi_j \Sigma \Psi_j' \end{pmatrix} \otimes \Sigma.
\end{aligned}$$

The absolute summability of  $\Psi_j$  then implies that the infinite sums converge to zero as  $m \rightarrow \infty$ . As  $X_t^{(m)} \xrightarrow{\text{m.s.}} X_t$  for  $m \rightarrow \infty$ , we can apply the Basic Approximation Theorem C.14 to reach the required conclusion

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \text{vec}(Z_t X'_{t-1}, Z_t X'_{t-2}, \dots, Z_t X'_{t-p}) \xrightarrow{\text{d}} \text{N}(0, \mathbf{\Gamma}_p \otimes \Sigma).$$

□

**Proof of Theorem 13.1**

*Proof.* We prove the Theorem for the arrangement by observation. The prove for the arrangement by equation can be proven in a completely analogous way. Inserting the regression formula (13.1) into the least-squares formula (13.3) leads to:

$$\begin{aligned} \text{vec } \widehat{\Phi} &= (((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n)(\mathbf{X} \otimes I_n) \text{vec } \Phi + (((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n) \text{vec } Z \\ &= \text{vec } \Phi + (((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n) \text{vec } Z. \end{aligned} \quad (13.7)$$

Bringing  $\text{vec } \Phi$  to the left hand side and taking the probability limit, we get using Slutsky's Lemma C.10 for the product of probability limits

$$\begin{aligned} \text{plim}(\text{vec } \widehat{\Phi} - \text{vec } \Phi) &= \text{plim } \text{vec } (Z\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= \text{vec} \left( \text{plim } \frac{Z\mathbf{X}}{T} \text{plim} \left( \frac{\mathbf{X}'\mathbf{X}}{T} \right)^{-1} \right) = 0. \end{aligned}$$

The last equality follows from the observation that proposition 13.1 implies  $\text{plim } \frac{\mathbf{X}'\mathbf{X}}{T} = \mathbf{\Gamma}_p$  nonsingular and that proposition 13.2 implies  $\text{plim } \frac{Z\mathbf{X}}{T} = 0$ . Thus, we have established that the Least-Squares estimator is consistent.

Equation (13.7) further implies:

$$\begin{aligned} \sqrt{T}(\text{vec } \widehat{\Phi} - \text{vec } \Phi) &= \sqrt{T} (((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \otimes I_n) \text{vec } Z \\ &= \left( \left( \frac{\mathbf{X}'\mathbf{X}^{-1}}{T} \right) \otimes I_n \right) \frac{1}{\sqrt{T}} (\mathbf{X}' \otimes I_n) \text{vec } Z \end{aligned}$$

As  $\text{plim } \frac{\mathbf{X}'\mathbf{X}}{T} = \mathbf{\Gamma}_p$  nonsingular, the above expression converges in distribution according to Theorem C.10 and Proposition 13.2 to a normally distributed random variable with mean zero and covariance matrix

$$(\mathbf{\Gamma}_p^{-1} \otimes I_n)(\mathbf{\Gamma}_p \otimes \Sigma)(\mathbf{\Gamma}_p^{-1} \otimes I_n) = \mathbf{\Gamma}_p^{-1} \otimes \Sigma$$

□

**Proof of Theorem 13.2***Proof.*

$$\begin{aligned}
\widehat{\Sigma} &= \frac{(Y - \widehat{\Phi}\mathbf{X}')(Y - \widehat{\Phi}\mathbf{X}')'}{T} \\
&= \frac{(Y - \Phi\mathbf{X}' + (\Phi - \widehat{\Phi})\mathbf{X}')(Y - \Phi\mathbf{X}' + (\Phi - \widehat{\Phi})\mathbf{X}')'}{T} \\
&= \frac{1}{T}(Z + (\Phi - \widehat{\Phi})\mathbf{X}')(Z + (\Phi - \widehat{\Phi})\mathbf{X}')' \\
&= \frac{ZZ'}{T} + \frac{Z\mathbf{X}}{T}(\Phi - \widehat{\Phi})' + (\Phi - \widehat{\Phi})\frac{\mathbf{X}'Z'}{T} + (\Phi - \widehat{\Phi})\frac{\mathbf{X}'\mathbf{X}}{T}(\Phi - \widehat{\Phi})'
\end{aligned}$$

Applying Theorem C.7 and the results of propositions 13.1 and 13.2 shows that

$$\frac{Z\mathbf{X}(\Phi - \widehat{\Phi})'}{\sqrt{T}} \xrightarrow{p} 0$$

and

$$(\Phi - \widehat{\Phi})\frac{\mathbf{X}'\mathbf{X}}{T}\sqrt{T}(\Phi - \widehat{\Phi})' \xrightarrow{p} 0.$$

Hence,

$$\sqrt{T} \left( \frac{(Y - \widehat{\Phi}\mathbf{X}')(Y - \widehat{\Phi}\mathbf{X}')'}{T} - \frac{ZZ'}{T} \right) = \sqrt{T} \left( \widehat{\Sigma} - \frac{ZZ'}{T} \right) \xrightarrow{p} 0$$

□

**13.4 The Yule-Walker Estimator**

An alternative estimation method can be derived from the Yule-Walker equations. Consider first a VAR(1) model. The Yule-Walker equation in this case simply is:

$$\begin{aligned}
\Gamma(0) &= \Phi\Gamma(-1) + \Sigma \\
\Gamma(1) &= \Phi\Gamma(0)
\end{aligned}$$

or

$$\begin{aligned}
\Gamma(0) &= \Phi\Gamma(0)\Phi' + \Sigma \\
\Gamma(1) &= \Phi\Gamma(0).
\end{aligned}$$

The solution of this system of equations is:

$$\begin{aligned}\Phi &= \Gamma(1)\Gamma(0)^{-1} \\ \Sigma &= \Gamma(0) - \Phi\Gamma(0)\Phi' = \Gamma(0) - \Gamma(1)\Gamma(0)^{-1}\Gamma(0)\Gamma(0)^{-1}\Gamma(1)' \\ &= \Gamma(0) - \Gamma(1)\Gamma(0)^{-1}\Gamma(1)'\end{aligned}$$

Replacing the theoretical moments by their empirical counterparts, we get the *Yule-Walker estimator* for  $\Phi$  and  $\Sigma$ :

$$\begin{aligned}\hat{\Phi} &= \hat{\Gamma}(1)\hat{\Gamma}(0)^{-1}, \\ \hat{\Sigma} &= \hat{\Gamma}(0) - \hat{\Phi}\hat{\Gamma}(0)\hat{\Phi}'.\end{aligned}$$

In the general case of a VAR(p) model the Yule-Walker estimator is given as the solution of the equation system

$$\begin{aligned}\hat{\Gamma}(h) &= \sum_{j=1}^p \hat{\Phi}_j \hat{\Gamma}(h-j), \quad k = 1, \dots, p, \\ \hat{\Sigma} &= \hat{\Gamma}(0) - \hat{\Phi}_1 \hat{\Gamma}(-1) - \dots - \hat{\Phi}_p \hat{\Gamma}(-p)\end{aligned}$$

As the the least-squares and the Yule-Walker estimator differ only in the treatment of the starting values, they are asymptotically equivalent. In fact, they yield very similar estimates even for finite samples (see e.g. Reinsel (1993)). However, as in the univariate case, the Yule-Walker estimator always delivers, in contrast to the least-square estimator, coefficient estimates with the property  $\det(I_n - \hat{\Phi}_1 z - \dots - \hat{\Phi}_p z^p) \neq 0$  for all  $z \in \mathbb{C}$  with  $|z| \leq 1$ . Thus, the Yule-Walker estimator guarantees that the estimated VAR possesses a causal representation. This, however, comes at the price that the Yule-Walker estimator has a larger small-sample bias than the least-squares estimator, especially when the roots of  $\Phi(z)$  get close to the unit circle (Tjøstheim and Paulsen, 1983; Shaman and Stine, 1988; Reinsel, 1993). Thus, it is generally preferable to use the least-squares estimator in practice.



# Chapter 14

## Forecasting with VAR Models

### 14.1 Forecasting with Known Parameters

The discussion of forecasting with VAR models proceeds in two steps. First, we assume that the parameters of the model are known. Although this assumption is unrealistic, it will nevertheless allow us to introduce and analyze important concepts and ideas. In a second step, we then investigate how the results established in the first step have to be amended if the parameters are estimated. The analysis will focus on stationary and causal VAR(1) processes. Processes of higher order can be accommodated by rewriting them in companion form. Thus we have:

$$\begin{aligned} X_t &= \Phi X_{t-1} + Z_t, & Z_t &\sim \text{WN}(0, \Sigma), \\ X_t &= Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \dots = \sum_{j=0}^{\infty} \Psi_j Z_{t-j}, \end{aligned}$$

where  $\Psi_j = \Phi^j$ . Consider then the following forecasting problem: Given observations  $\{X_T, X_{T-1}, \dots, X_1\}$ , find a linear function, called predictor or forecast function,  $\mathbb{P}_T X_{T+h}$ ,  $h \geq 1$ , which minimizes the expected quadratic forecast error

$$\begin{aligned} \mathbb{E} (X_{T+h} - \mathbb{P}_T X_{T+h})' (X_{T+h} - \mathbb{P}_T X_{T+h}) \\ = \mathbb{E} \text{tr} (X_{T+h} - \mathbb{P}_T X_{T+h})(X_{T+h} - \mathbb{P}_T X_{T+h})'. \end{aligned}$$

Thereby “tr” denotes the trace operator which takes the sum of the diagonal elements of a matrix. As we rely on linear forecasting functions,  $\mathbb{P}_T X_{T+h}$  can be expressed as

$$\mathbb{P}_T X_{T+h} = A_1 X_T + A_2 X_{T-1} + \dots + A_T X_1 \quad (14.1)$$

with matrices  $A_1, A_2, \dots, A_T$  still to be determined. In order to simplify the exposition, we already accounted for the fact that the mean of  $\{X_t\}$  is zero.<sup>1</sup> A justification for focusing on linear least-squares forecasts is given in Chapter 3. The first order conditions for the least-squares minimization problem are given by the normal equations:

$$\begin{aligned}\mathbb{E}(X_{T+h} - \mathbb{P}_T X_{T+h}) X'_s &= \mathbb{E}(X_{T+h} - A_1 X_T - \dots - A_T X_1) X'_s \\ &= \mathbb{E}X_{T+h} X'_s - A_1 \mathbb{E}X_T X'_s - \dots - A_T \mathbb{E}X_1 X'_s = 0, \quad 1 \leq s \leq T.\end{aligned}$$

These equations state that the forecast error  $(X_{T+h} - \mathbb{P}_T X_{T+h})$  must be uncorrelated with the available information  $X_s$ ,  $s = 1, 2, \dots, T$ . The normal equations can be written as

$$\begin{aligned}(A_1, A_2, \dots, A_T) \begin{pmatrix} \Gamma(0) & \Gamma(1) & \dots & \Gamma(T-1) \\ \Gamma'(1) & \Gamma(0) & \dots & \Gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma'(T-1) & \Gamma'(T-2) & \dots & \Gamma(0) \end{pmatrix} \\ = (\Gamma(h) \quad \Gamma(h+1) \quad \dots \quad \Gamma(T+h-1)).\end{aligned}$$

Denoting by  $\mathbf{\Gamma}_T$  the matrix

$$\mathbf{\Gamma}_T = \begin{pmatrix} \Gamma(0) & \Gamma(1) & \dots & \Gamma(T-1) \\ \Gamma'(1) & \Gamma(0) & \dots & \Gamma(T-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma'(T-1) & \Gamma'(T-2) & \dots & \Gamma(0) \end{pmatrix},$$

the normal equations can be written more compactly as

$$(A_1, A_2, \dots, A_T) \mathbf{\Gamma}_T = (\Gamma(h) \quad \Gamma(h+1) \quad \dots \quad \Gamma(T+h-1)).$$

Using the assumption that  $\{X_t\}$  is a VAR(1),  $\Gamma(h)$  can be expressed as  $\Gamma(h) = \Phi^h \Gamma(0)$  (see equation(12.3)) so that the normal equations become

$$\begin{aligned}(A_1, A_2, \dots, A_T) \begin{pmatrix} \Gamma(0) & \Phi \Gamma(0) & \dots & \Phi^{T-1} \Gamma(0) \\ \Gamma(0) \Phi' & \Gamma(0) & \dots & \Phi^{T-2} \Gamma(0) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(0) \Phi'^{T-1} & \Gamma(0) \Phi'^{T-2} & \dots & \Gamma(0) \end{pmatrix} \\ = (\Phi^h \Gamma(0) \quad \Phi^{h+1} \Gamma(0) \quad \dots \quad \Phi^{T+h-1} \Gamma(0)).\end{aligned}$$

<sup>1</sup>If the mean is non-zero, a constant  $A_0$  must be added to the forecast function.

The easily guessed solution is given by  $A_1 = \Phi^h$  and  $A_2 = \dots = A_T = 0$ . Thus, the sought-after forecasting function for the VAR(1) process is

$$\mathbb{P}_T X_{T+h} = \Phi^h X_T. \quad (14.2)$$

The forecast error  $X_{T+h} - \mathbb{P}_T X_{T+h}$  has expectation zero. Thus, the linear least-squares predictor delivers unbiased forecasts. As

$$X_{T+h} = Z_{T+h} + \Phi Z_{T+h-1} + \dots + \Phi^{h-1} Z_{T+1} + \Phi^h X_T,$$

the expected squared forecast error (mean squared error)  $\text{MSE}(h)$  is

$$\begin{aligned} \text{MSE}(h) &= \mathbb{E} (X_{T+h} - \Phi^h X_T) (X_{T+h} - \Phi^h X_T)' \\ &= \Sigma + \Phi \Sigma \Phi' + \dots + \Phi^{h-1} \Sigma \Phi^{h-1} = \sum_{j=0}^{h-1} \Phi^j \Sigma \Phi^{j'}. \end{aligned} \quad (14.3)$$

In order to analyze the case of a causal VAR(p) process with  $T > p$ , we transform the model into the companion form. For  $h = 1$ , we can apply the result above to get:

$$\mathbb{P}_T Y_{T+1} = \Phi Y_T = \begin{pmatrix} \mathbb{P}_T X_{T+1} \\ X_T \\ X_{T-1} \\ \vdots \\ X_{T-p+2} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \dots & 0 & 0 \\ 0 & I_n & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_n & 0 \end{pmatrix} \begin{pmatrix} X_T \\ X_{T-1} \\ X_{T-2} \\ \vdots \\ X_{T-p+1} \end{pmatrix}.$$

This implies that

$$\mathbb{P}_T X_{T+1} = \Phi_1 X_T + \Phi_2 X_{T-1} + \dots + \Phi_p X_{T-p+1}. \quad (14.4)$$

The forecast error is  $X_{T+1} - \mathbb{P}_T X_{T+1} = Z_t$  which has mean zero and covariance variance matrix  $\Sigma$ . In general we have that  $\mathbb{P}_T Y_{T+h} = \Phi^h Y_T$  so that  $\mathbb{P}_T X_{T+h}$  is equal to

$$\mathbb{P}_T X_{T+h} = \Phi_1^{(h)} X_T + \Phi_2^{(h)} X_{T-1} + \dots + \Phi_p^{(h)} X_{T-p+1}$$

where  $\Phi_i^{(h)}$ ,  $i = 1, \dots, p$ , denote the blocks in the first row of  $\Phi^h$ . Alternatively, the forecast for  $h > 1$  can be computed recursively. For  $h = 2$  this leads to:

$$\begin{aligned} \mathbb{P}_T X_{T+2} &= \mathbb{P}_T (\Phi_1 X_{T+1}) + \mathbb{P}_T (\Phi_2 X_T) + \dots + \mathbb{P}_T (\Phi_p X_{T+2-p}) + \mathbb{P}_T (Z_{T+2}) \\ &= \Phi_1 (\Phi_1 X_T + \Phi_2 X_{T-1} + \dots + \Phi_p X_{T+1-p}) \\ &\quad + \Phi_2 X_T + \dots + \Phi_p X_{T+2-p} \\ &= (\Phi_1^2 + \Phi_2) X_T + (\Phi_1 \Phi_2 + \Phi_3) X_{T-1} + \dots + (\Phi_1 \Phi_{p-1} + \Phi_p) X_{T+2-p} \\ &\quad + \Phi_1 \Phi_p X_{T+1-p}. \end{aligned}$$

For  $h > 2$  we proceed analogously. This way of producing forecasts is sometimes called *iterated* forecasts.

In general, the forecast error of a causal VAR(p) process can be expressed as

$$\begin{aligned} X_{T+h} - \mathbb{P}_T X_{T+h} &= Z_{T+h} + \Psi_1 Z_{T+h-1} + \dots + \Psi_{h-1} Z_{T+1} \\ &= \sum_{j=0}^{h-1} \Phi_j Z_{T+h-j}. \end{aligned}$$

The MSE( $h$ ) then is:

$$\text{MSE}(h) = \Sigma + \Psi_1 \Sigma \Psi_1' + \dots + \Psi_{h-1} \Sigma \Psi_{h-1}' = \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j'. \quad (14.5)$$

### Example

Consider again the VAR(2) model of Section 12.3. The forecast function in this case is then:

$$\begin{aligned} \mathbb{P}_T X_{T+1} &= \Phi_1 X_t + \Phi_2 X_{t-1} \\ &= \begin{pmatrix} 0.8 & -0.5 \\ 0.1 & -0.5 \end{pmatrix} X_t + \begin{pmatrix} -0.3 & -0.3 \\ -0.2 & 0.3 \end{pmatrix} X_{t-1}, \\ \mathbb{P}_T X_{T+2} &= (\Phi_1^2 + \Phi_2) X_t + \Phi_1 \Phi_2 X_{t-1} \\ &= \begin{pmatrix} 0.29 & -0.45 \\ -0.17 & 0.50 \end{pmatrix} X_t + \begin{pmatrix} -0.14 & -0.39 \\ 0.07 & -0.18 \end{pmatrix} X_{t-1}, \\ \mathbb{P}_T X_{T+3} &= (\Phi_1^3 + \Phi_1 \Phi_2 + \Phi_2 \Phi_1) X_t + (\Phi_1^2 \Phi_2 + \Phi_2^2) X_{t-1} \\ &= \begin{pmatrix} 0.047 & -0.310 \\ -0.016 & -0.345 \end{pmatrix} X_t + \begin{pmatrix} 0.003 & -0.222 \\ -0.049 & 0.201 \end{pmatrix} X_{t-1}. \end{aligned}$$

Based on the results computed in Section 12.3, we can calculate the corresponding mean squared errors (MSE):

$$\begin{aligned} \text{MSE}(1) &= \Sigma = \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 2.0 \end{pmatrix}, \\ \text{MSE}(2) &= \Sigma + \Psi_1 \Sigma \Psi_1' = \begin{pmatrix} 1.82 & 0.80 \\ 0.80 & 2.47 \end{pmatrix}, \\ \text{MSE}(3) &= \Sigma + \Psi_1 \Sigma \Psi_1' + \Psi_2 \Sigma \Psi_2' = \begin{pmatrix} 2.2047 & 0.3893 \\ 0.3893 & 2.9309 \end{pmatrix}. \end{aligned}$$

A practical forecasting exercise with additional material is presented in section 14.4.

### 14.1.1 Wold Decomposition Theorem

At this stage we note that *Wold's theorem* or *Wold's Decomposition* carries over to the multivariate case (see Section 3.2 for the univariate case). This Theorem asserts that there exists for each purely non-deterministic stationary process<sup>2</sup> a decomposition, respectively representation, of the form:

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

where  $\Psi_0 = I_n$ ,  $Z_t \sim \text{WN}(0, \Sigma)$  with  $\Sigma > 0$  and  $\sum_{j=0}^{\infty} \|\Psi_j\|^2 < \infty$ . The innovations  $\{Z_t\}$  have the property  $Z_t = X_t - \tilde{\mathbb{P}}_{t-1} X_t$  and consequently  $Z_t = \tilde{\mathbb{P}}_t Z_t$ . Thereby  $\tilde{\mathbb{P}}_t$  denotes the linear least-squares predictor based on the infinite past  $\{X_t, X_{t-1}, \dots\}$ . The interpretation of the multivariate case is analogous to the univariate one.

## 14.2 Forecasting with Estimated Parameters

In practice the parameters of the VAR model are usually unknown and have therefore to be estimated. In the previous Section we have demonstrated that

$$\mathbb{P}_T X_{T+h} = \Phi_1 \mathbb{P}_T X_{T+h-1} + \dots + \Phi_p \mathbb{P}_T X_{T+h-p}$$

where  $\mathbb{P}_T X_{T+h-j} = Y_{T+h-j}$  if  $j \geq h$ . Replacing the true parameters by their estimates, we get the forecast function

$$\hat{\mathbb{P}}_T X_{T+h} = \hat{\Phi}_1 \hat{\mathbb{P}}_T X_{T+h-1} + \dots + \hat{\Phi}_p \hat{\mathbb{P}}_T X_{T+h-p}.$$

where a hat indicates the use of estimates. The forecast error can then be decomposed into two components:

$$\begin{aligned} X_{T+h} - \hat{\mathbb{P}}_T X_{T+h} &= (X_{T+h} - \mathbb{P}_T X_{T+h}) + (\mathbb{P}_T X_{T+h} - \hat{\mathbb{P}}_T X_{T+h}) \\ &= \sum_{j=0}^{h-1} \Phi_j Z_{T+h-j} + (\mathbb{P}_T X_{T+h} - \hat{\mathbb{P}}_T X_{T+h}). \end{aligned} \quad (14.6)$$

Dufour (1985) has shown that, under the assumption of symmetrically distributed  $Z_t$ 's (i.e. if  $Z_t$  and  $-Z_t$  have the same distribution) the expectation

---

<sup>2</sup>A stationary stochastic process is called deterministic if it can be perfectly forecasted from its infinite past. It is called purely non-deterministic if there is no deterministic component (see Section 3.2).

of the forecast error is zero even when the parameters are replaced by their least-squares estimates. This result holds despite the fact that these estimates are biased in small samples. Moreover, the results do not assume that the model is correctly specified in terms of the order  $p$ . Thus, under quite general conditions the forecast with estimated coefficients remains unbiased so that  $\mathbb{E}\left(X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right) = 0$ .

If the estimation is based on a different sample than the one used for forecasting, the two terms in the above expression are uncorrelated so that its mean squared error is by the sum of the two mean squared errors:

$$\begin{aligned} \widehat{\text{MSE}}(h) &= \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j' \\ &\quad + \mathbb{E}\left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right) \left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right)'. \end{aligned} \quad (14.7)$$

The last term can be evaluated by using the asymptotic distribution of the coefficients as an approximation. The corresponding formula turns out to be cumbersome. The technical details can be found in Lütkepohl (2006) and Reinsel (1993). The formula can, however, be simplified considerably if we consider a forecast horizon of only one period. We deduce the formula for a VAR of order one, i.e. taking  $X_t = \Phi X_{t-1} + Z_t$ ,  $Z_t \sim \text{WN}(0, \Sigma)$ .

$$\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h} = (\Phi - \widehat{\Phi}) X_T = \text{vec}\left((\Phi - \widehat{\Phi}) X_T\right) = (X_T' \otimes I_n) \text{vec}(\Phi - \widehat{\Phi}).$$

This implies that

$$\begin{aligned} &\mathbb{E}\left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right) \left(\mathbb{P}_T X_{T+h} - \widehat{\mathbb{P}}_T X_{T+h}\right)' \\ &= \mathbb{E}(X_T' \otimes I_n) \text{vec}(\Phi - \widehat{\Phi}) (\text{vec}(\Phi - \widehat{\Phi}))' (X_T \otimes I_n) \\ &= \mathbb{E}(X_T' \otimes I_n) \frac{\mathbf{\Gamma}_1^{-1} \otimes \Sigma}{T} (X_T \otimes I_n) = \frac{1}{T} \mathbb{E}(X_T' \mathbf{\Gamma}_1^{-1} X_T) \otimes \Sigma \\ &= \frac{1}{T} \mathbb{E}(\text{tr} X_T' \mathbf{\Gamma}_1^{-1} X_T) \otimes \Sigma = \frac{1}{T} \text{tr}(\mathbf{\Gamma}_1^{-1} \mathbb{E}(X_T X_T')) \otimes \Sigma \\ &= \frac{1}{T} (\text{tr}(I_n) \otimes \Sigma) = \frac{n}{T} \Sigma. \end{aligned}$$

Thereby, we have used the asymptotic normality of the least-squares estimator (see Theorem 13.1) and the assumption that forecasting and estimation uses different realizations of the stochastic process. Thus, for  $h = 1$  and  $p = 1$ , we get

$$\widehat{\text{MSE}}(1) = \Sigma + \frac{n}{T} \Sigma = \frac{T+n}{T} \Sigma.$$

Higher order models can be treated similarly using the companion form of VAR(p). In this case:

$$\widehat{\text{MSE}}(1) = \Sigma + \frac{np}{T}\Sigma = \frac{T + np}{T}\Sigma. \quad (14.8)$$

This is only an approximation as we applied asymptotic results to small sample entities. The expression shows that the effect of the substitution of the coefficients by their least-squares estimates vanishes as the sample becomes large. However, in small sample the factor  $\frac{T+np}{T}$  can be sizeable. In the example treated in Section 14.4, the covariance matrix  $\Sigma$ , taking the use of a constant into account and assuming 8 lags, has to be inflated by  $\frac{T+np+1}{T} = \frac{196+4 \times 8+1}{196} = 1.168$ . Note also that the precision of the forecast, given  $\Sigma$ , diminishes with the number of parameters.

### 14.3 Modeling of VAR models

The previous section treated the estimation of VAR models under the assumption that the order of the VAR,  $p$ , is known. In most cases, this assumption is unrealistic as the order  $p$  is unknown and must be retrieved from the data. We can proceed analogously as in the univariate case (see Section 5.1) and iteratively test the hypothesis that coefficients corresponding to the highest lag, i.e.  $\Phi_p = 0$ , are simultaneously equal to zero. Starting from a maximal order  $p_{max}$ , we test the null hypothesis that  $\Phi_{p_{max}} = 0$  in the corresponding VAR( $p_{max}$ ) model. If the hypothesis is not rejected, we reduce the order by one to  $p_{max} - 1$  and test anew the null hypothesis  $\Phi_{p_{max}-1} = 0$  using the smaller VAR( $p_{max} - 1$ ) model. One continues in this way until the null hypothesis is rejected. This gives, then, the appropriate order of the VAR. The different tests can be carried out either as Wald-tests (F-tests) or as likelihood-ratio tests ( $\chi^2$ -tests) with  $n^2$  degrees of freedom.

An alternative procedure to determine the order of the VAR relies on some information criteria. As in the univariate case, the most popular ones are the Akaike (AIC), the Schwarz or Bayesian (BIC) and the Hannan-Quinn criterion (HQC). The corresponding formula are:

$$\begin{aligned} \text{AIC}(p): & \quad \ln \det \tilde{\Sigma}_p + \frac{2pn^2}{T}, \\ \text{BIC}(p): & \quad \ln \det \tilde{\Sigma}_p + \frac{pn^2}{T} \ln T, \\ \text{HQC}(p): & \quad \ln \det \tilde{\Sigma}_p + \frac{2pn^2}{T} \ln(\ln T), \end{aligned}$$

where  $\tilde{\Sigma}_p$  denotes the degree of freedom adjusted estimate of the covariance matrix  $\Sigma$  for a model of order  $p$  (see equation (13.5)).  $n^2p$  is the number of estimated coefficients. The estimated order is then given as the minimizer of one of these criteria. In practice the Akaike's criterion is the most popular one although it has a tendency to deliver orders which are too high. The BIC and the HQ-criterion on the other hand deliver the correct order on average, but can lead to models which suffer from the omitted variable bias when the estimated order is too low. Examples are discussed in Sections 14.4 and 15.4.5.

Following Lütkepohl (2006), Akaike's information criterion can be rationalized as follows. Take as a measure of fit the determinant of the one period approximate mean-squared errors  $\widehat{\text{MSE}}(1)$  from equation (14.8) and take as an estimate of  $\Sigma$  the degrees of freedom corrected version in equation (13.5). The resulting criterion is called according to Akaike (1969) the *final prediction error* (FPE):

$$\text{FPE}(p) = \det \left( \frac{T+np}{T} \times \frac{T}{T-np} \tilde{\Sigma} \right) = \left( \frac{T+np}{T-np} \right)^n \det \tilde{\Sigma}. \quad (14.9)$$

Taking logs and using the approximations  $\frac{T+np}{T-np} \approx 1 + \frac{2np}{T}$  and  $\log(1 + \frac{2np}{T}) \approx \frac{2np}{T}$ , we arrive at

$$\text{AIC}(p) \approx \log \text{FPE}(p).$$

## 14.4 Example: A Simple VAR Macroeconomic Model for the U.S. Economy

In this section, we illustrate how to build and use VAR models for forecasting key macroeconomic variables. For this purpose, we consider the following four variables: GDP per capita ( $\{Y_t\}$ ), price level in terms of the consumer price index (CPI) ( $\{P_t\}$ ), real money stock M1 ( $\{M_t\}$ ), and the three month treasury bill rate ( $\{R_t\}$ ). All variables are for the U.S. and are, with the exception of the interest rate, in logged differences.<sup>3</sup> The components of  $X_t$  are with the exception of the interest rate stationary.<sup>4</sup> Thus, we aim at modeling  $X_t = (\Delta \log Y_t, \Delta \log P_t, \Delta \log M_t, R_t)'$ . The sample runs from the first quarter 1959 to the first quarter 2012. We estimate our models, however,

<sup>3</sup>Thus,  $\Delta \log P_t$  equals the inflation rate.

<sup>4</sup>Although the unit root test indicate that  $R_t$  is integrated of order one, we do not difference this variable. This specification will not affect the consistency of the estimates nor the choice of the lag-length (Sims et al., 1990), but has the advantage that each component of  $X_t$  is expressed in percentage points which facilitates the interpretation.

Table 14.1: Information Criteria for the VAR Models of Different Orders

order	AIC	BIC	HQ
0	-14.498	-14.429	-14.470
1	-17.956	-17.611	-17.817
2	-18.638	<b>-18.016</b>	-18.386
3	-18.741	-17.843	-18.377
4	-18.943	-17.768	-18.467
5	-19.081	-17.630	<b>-18.493</b>
6	-19.077	-17.349	-18.377
7	-19.076	-17.072	-18.264
8	<b>-19.120</b>	-16.839	-18.195
9	-18.988	-16.431	-17.952
10	-18.995	-16.162	-17.847
11	-18.900	-15.789	-17.639
12	-18.884	-15.497	-17.512

minimum in bold

only up to the fourth quarter 2008 and reserve the last thirteen quarters, i.e. the period from the first quarter 2009 to first quarter of 2012, for an out-of-sample evaluation of the forecast performance. This forecast assessment has the advantage to account explicitly of the sampling variability in estimated parameter models.

The first step in the modeling process is the determination of the lag-length. Allowing for a maximum of twelve lags, the different information criteria produce the values reported in Table 14.1. Unfortunately, the three criteria deliver different orders: AIC suggests 8 lags, HQ 5 lags, and BIC 2 lags. In such a situation it is wise to keep all three models and to perform additional diagnostic tests.<sup>5</sup> One such test is to run a horse-race between the three models in terms of their forecasting performance.

Forecasts are evaluated according to the root-mean-squared-error (RMSE)

<sup>5</sup>Such tests would include an analysis of the autocorrelation properties of the residuals and tests of structural breaks.

and the mean-absolute-error (MAE)<sup>6</sup>:

$$\text{RMSE} : \sqrt{\frac{1}{h} \sum_{T+1}^{T+h} (\hat{X}_{it} - X_{it})^2} \quad (14.10)$$

$$\text{MAE} : \frac{1}{h} \sum_{T+1}^{T+h} |\hat{X}_{it} - X_{it}| \quad (14.11)$$

where  $\hat{X}_{it}$  and  $X_{it}$  denote the forecast and the actual value of variable  $i$  in period  $t$ . Forecasts are computed for a horizon  $h$  starting in period  $T$ . We can gain further insights by decomposing the mean-squared-error additively into three components:

$$\begin{aligned} \frac{1}{h} \sum_{T+1}^{T+h} (\hat{X}_{it} - X_{it})^2 &= \left( \left( \frac{1}{h} \sum_{T+1}^{T+h} \hat{X}_{it} \right) - \bar{X}_i \right)^2 \\ &\quad + (\sigma_{\hat{X}_i} - \sigma_{X_i})^2 + 2(1 - \rho)\sigma_{\hat{X}_i}\sigma_{X_i}. \end{aligned}$$

The first component measures how far the mean of the forecasts  $\frac{1}{h} \sum_{T+1}^{T+h} \hat{X}_{it}$  is away from the actual mean of the data  $\bar{X}_i$ . It therefore measures the *bias* of the forecasts. The second one compares the standard deviation of the forecast  $\sigma_{\hat{X}_i}$  to those of the data  $\sigma_{X_i}$ . Finally, the last component is a measure of the unsystematic forecast errors where  $\rho$  denotes the correlation between the forecast and the data. Ideally, each of the three components should be close to zero: there should be no bias, the variation of the forecasts should correspond to those of the data, and the forecasts and the data should be highly positively correlated. In order to avoid scaling problems, all three components are usually expressed as a proportion of  $\frac{1}{h} \sum_{T+1}^{T+h} (\hat{X}_{it} - X_{it})^2$ :

$$\text{bias proportion:} \quad \frac{\left( \left( \frac{1}{h} \sum_{T+1}^{T+h} \hat{X}_{it} \right) - \bar{X}_i \right)^2}{\frac{1}{h} \sum_{T+1}^{T+h} (\hat{X}_{it} - X_{it})^2} \quad (14.12)$$

$$\text{variance proportion:} \quad \frac{(\sigma_{\hat{X}_i} - \sigma_{X_i})^2}{\frac{1}{h} \sum_{T+1}^{T+h} (\hat{X}_{it} - X_{it})^2} \quad (14.13)$$

$$\text{covariance proportion:} \quad \frac{2(1 - \rho)\sigma_{\hat{X}_i}\sigma_{X_i}}{\frac{1}{h} \sum_{T+1}^{T+h} (\hat{X}_{it} - X_{it})^2} \quad (14.14)$$

<sup>6</sup>Alternatively one could use the mean-absolute-percentage-error (MAPE). However, as all variables are already in percentages, the MAE is to be preferred.

We use the three models to produce dynamic or iterated forecasts  $\mathbb{P}_T X_{T+1}$ ,  $\mathbb{P}_T X_{T+2}, \dots, \mathbb{P}_T X_{T+h}$ . Thus, forecasts for  $h \geq 2$  are computed iteratively by inserting for the lagged variables the forecasts obtained in the previous steps. For details see Chapter 14. Alternatively, one may consider a recursive or rolling out-of-sample strategy where the model is reestimated each time a new observation becomes available. Thus, we would evaluate the one-period-ahead forecasts  $\mathbb{P}_T X_{T+1}, \mathbb{P}_{T+1} X_{T+2}, \dots, \mathbb{P}_{T+h-1} X_{T+h}$ , the two-period-ahead forecasts  $\mathbb{P}_T X_{T+2}, \mathbb{P}_{T+1} X_{T+3}, \dots, \mathbb{P}_{T+h-2} X_{T+h}$ , and so on. The difference between the recursive and the rolling strategy is that in the first case all observations are used for estimation whereas in the second case the sample is rolled over so that its size is kept fixed at  $T$ .

Figure 14.1 displays dynamic or iterated forecasts for the four variables expressed in log-levels, respectively in levels for the interest rate. Forecasts are evaluated according to the performance measures explained above. The corresponding values are reported in Table 14.2. All models see a quick recovery after the recession in 2008 and are thus much too optimistic. The lowest RMSE for  $\log Y_t$  is 5.678 for the VAR(8) model. Thus, GDP per capita is predicted to be on average almost 6 percent too high over the forecast period. This overly optimistic forecast is reflected in a large bias proportion which amounts to more than 95 percent. The situation looks much better for the price level. All models see an increase in inflation starting in 2009. Especially, the two higher order models fare much better. Their RMSE is just over 1 percent. The bias proportion is practically zero for the VAR(8) model. The forecast results of the real money stock are mixed. All models predict a quick recovery. This took indeed place, but first at a more moderate pace. Starting in mid-2010 the unconventional monetary policy of quantitative easing, however, led to an unforeseen acceleration so that the forecasts turned out to be systematically too low for the later period. Interestingly, the smallest model fared significantly better than the other two. Finally, the results for the interest rates are very diverse. Whereas the VAR(2) model predicts a rise in the interest rate, the other models foresee a decline. The VAR(8) model even predicts a very drastic fall. However, all models miss the continuation of the low interest rate regime and forecasts an increase starting already in 2009. This error can again be attributed to the unforeseen low interest rate monetary policy which was implemented in conjunction with the quantitative easing. This misjudgement resulted in a relatively large bias proportion.

Up to now, we have just been concerned with *point forecasts*. Point forecasts, however, describe only one possible outcome and do not reflect the inherent uncertainty surrounding the prediction problem. It is, thus, a question of scientific integrity to present in addition to the point forecasts also

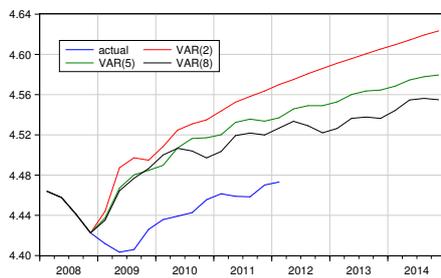
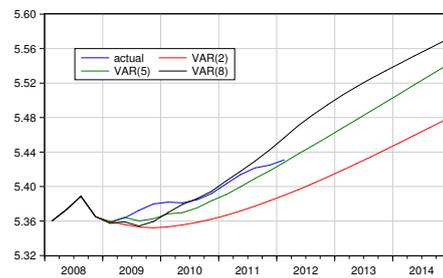
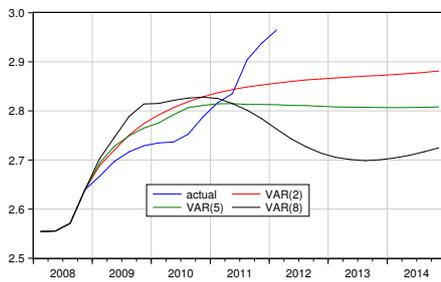
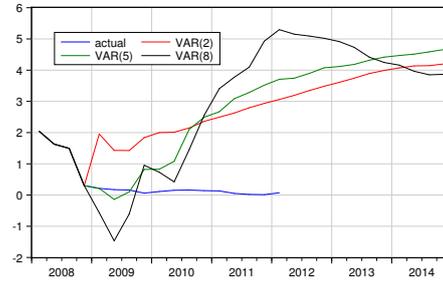
(a)  $\log Y_t$ (b)  $\log P_t$ (c)  $\log M_t$ (d)  $R_t$ 

Figure 14.1: Forecast Comparison of Alternative Models

Table 14.2: Forecast evaluation of alternative VAR models

	VAR(2)	VAR(5)	VAR(8)
	log $Y_t$		
RMSE	8.387	6.406	5.678
bias proportion	0.960	0.961	0.951
variance proportion	0.020	0.010	0.001
covariance proportion	0.020	0.029	0.048
MAE	8.217	6.279	5.536
	log $P_t$		
RMSE	3.126	1.064	1.234
bias proportion	0.826	0.746	0.001
variance proportion	0.121	0.001	0.722
covariance proportion	0.053	0.253	0.278
MAE	2.853	0.934	0.928
	log $M_t$		
RMSE	5.616	6.780	9.299
bias proportion	0.036	0.011	0.002
variance proportion	0.499	0.622	0.352
covariance proportion	0.466	0.367	0.646
MAE	4.895	5.315	7.762
	$R_t$		
RMSE	2.195	2.204	2.845
bias proportion	0.367	0.606	0.404
variance proportion	0.042	0.337	0.539
covariance proportion	0.022	0.057	0.057
MAE	2.125	1.772	2.299

RMSE and MAE for log  $Y_t$ , log  $P_t$ , and log  $M_t$  are multiplied by 100.

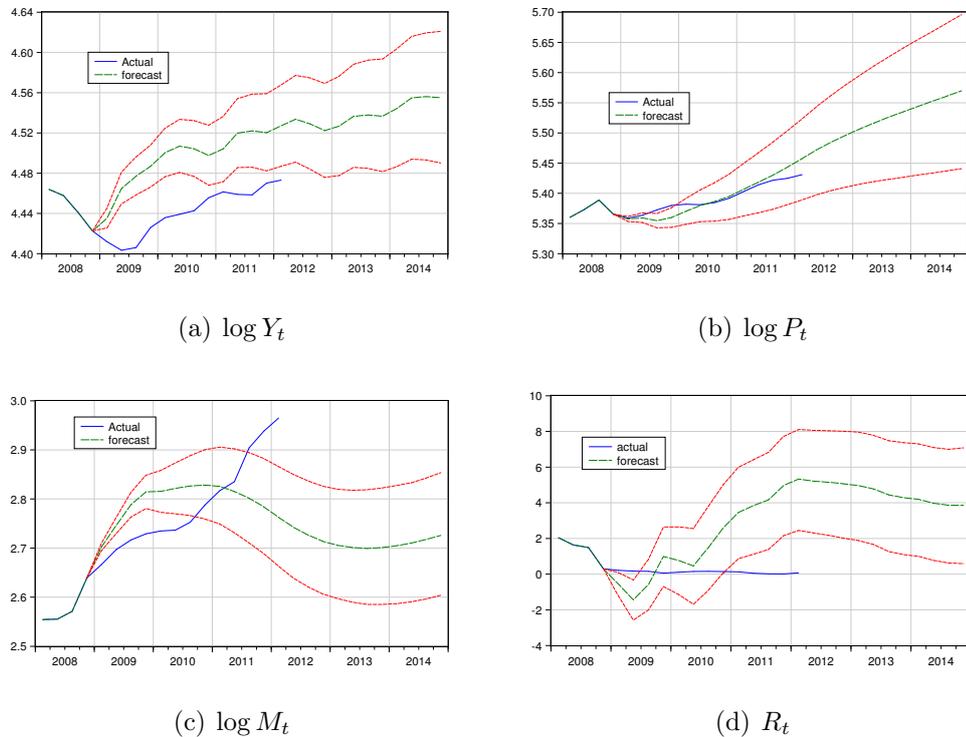


Figure 14.2: Forecast of VAR(8) model and 80 percent confidence intervals (red dotted lines)

confidence intervals. One straightforward way to construct such intervals is by computing the matrix of mean-squared-errors MSE using equation (14.5). The diagonal elements of this matrix can be interpreted as a measure of the forecast error variances for each variable. Under the assumption that the innovations  $\{Z_t\}$  are Gaussian, such confidence intervals can be easily computed. However, in practice this assumption is likely to be violated. This problem can be circumvented by using the empirical distribution function of the residuals to implement a bootstrap method similar to the computation of the Value-at-Risk in Section 8.4. Figure 14.2 plots the forecasts of the VAR(8) model together with a 80 percent confidence interval computed from the bootstrap approach. It shows that, with the exception of the logged price level, the actual realizations fall out of the confidence interval despite the fact that the intervals are already relatively large. This documents the uniqueness of the financial crisis and gives a hard time for any forecasting model.

Instead of computing a confidence interval, one may estimate the prob-

ability distribution of possible future outcomes. This provides a complete description of the uncertainty related to the prediction problem (Christoffersen, 1998; Diebold et al., 1998; Tay and Wallis, 2000; Corradi and Swanson, 2006). Finally, one should be aware that the innovation uncertainty is not the only source of uncertainty. As the parameters of the model are themselves estimated, there is also a coefficient uncertainty. In addition, we have to face the possibility that the model is misspecified.

The forecasting performance of the VAR models may seem disappointing at first. However, this was only be a first attempt and further investigations are usually necessary. These may include the search for *structural breaks* (See Bai et al., 1998; Perron, 2006). Another reason for the poor forecasting may be due to the *over-parametrization* of VAR models. The VAR(8) model, for example, 32 lagged dependent variables plus a constant in each of the four equations which leads to a total 132 parameters. This problem can be dealt with by applying Bayesian shrinkage techniques. This approach, also known as Bayesian VAR (BVAR), was particularly successful when using the so-called Minnesota prior (See Doan et al., 1984; Litterman, 1986; Kunst and Neusser, 1986; Banbura et al., 2010)

Besides these more fundamental issues, one may rely on more technical remedies. One such remedy is the use of direct rather iterated forecasts. This difference is best explained in the context of the VAR(1) model  $X_t = \Phi X_{t-1} + Z_t$ ,  $Z_t \sim \text{WN}(0, \Sigma)$ . The *iterated* forecast for  $X_{T+h}$  uses the OLS-estimate  $\hat{\Phi}$  to compute the forecast  $\hat{\Phi}^h X_T$  (see Chapter 14). Alternatively, one may estimate instead of the VAR(1), the model  $X_t = \Upsilon X_{t-h} + Z_t$  and compute the *direct* forecast for  $X_{T+h}$  as  $\hat{\Upsilon} X_T$ . Although  $\hat{\Upsilon}$  has larger variance than  $\hat{\Phi}$  if the VAR(1) is correctly specified, it is robust to misspecification (see Bhansali, 1999; Schorfheide, 2005; Marcellino et al., 2006).

Another interesting and common device is intercept correction or residual adjustment. Thereby the constant terms are adjusted in such a way that the residuals of the most recent observation become zero. The model is thereby set back on track. In this way the forecaster can guard himself against possible structural breaks. Residual adjustment can also serve as a device to incorporate anticipated events, like announced policies, which are not yet incorporated into the model. See Clements and Hendry (1996, 2006) for further details and additional forecasting devices.



# Chapter 15

## Interpretation and Identification of VAR Models

Although the estimation of VAR models poses no difficulties as outlined in the previous chapter, the individual coefficients are almost impossible to interpret. On the one hand, there are usually many coefficients, a VAR(4) model with three variables, for example, already has twelve coefficients per equation and thus 36 coefficients in total to interpret; on the other hand, there is in general no unambiguous relation of the VAR coefficients to the coefficients of a particular model. The last problem is known as the *identification* problem. To overcome this identification problem, many techniques have been developed which should allow to give the estimated VAR model an explicit economic interpretation.

### 15.1 Wiener-Granger Causality

As a first technique for the understanding of VAR processes, we analyze the concept of *causality* which was introduced by Granger (1969). The concept is also known as *Wiener-Granger causality* because Granger's idea goes back to the work of Wiener (1956). Take a multivariate time series  $\{X_t\}$  and consider the forecast of  $X_{1,T+h}$ ,  $h \geq 1$ , given  $X_T, X_{T-1}, \dots$  where  $\{X_t\}$  has not only  $X_{1t}$  as a component, but also another variable or group of variables  $X_{2t}$ .  $X_t$  may contain even further variables than  $X_{1,t}$  and  $X_{2,t}$ . The mean-squared forecast error is denoted by  $\text{MSE}_1(h)$ . Consider now an alternative forecast of  $X_{1,T+h}$  given  $\tilde{X}_T, \tilde{X}_{T-1}, \dots$  where  $\{\tilde{X}_t\}$  is obtained from  $\{X_t\}$  by eliminating the component  $X_{2,t}$ . The mean-squared error of this forecast is denoted by  $\widetilde{\text{MSE}}_1(h)$ . According to Granger, we can say that the second

variable  $X_{2,t}$  *causes* or *is causal for*  $X_{1,t}$  if and only if

$$\text{MSE}(h)_1 < \widetilde{\text{MSE}}_1(h) \quad \text{for some } h \geq 1.$$

This means that the information contained in  $\{X_{2t}\}$  and its past improves the forecast of  $\{X_{1t}\}$  in the sense of the mean-squared forecast error. Thus the concept of Wiener-Granger causality makes only sense for purely non-deterministic processes and rest on two principles:<sup>1</sup>

- The future cannot cause the past. Only the past can have a causal influence on the future.<sup>2</sup>
- A specific cause contains information not available otherwise.

The concept of Wiener-Granger causality played an important role in the debate between monetarists and Keynesians over the issue whether the money stock has an independent influence on real activity. It turned out that this question can only be resolved within a specific context. Sims (1980a), for example, showed that the relationship between the growth rate of the money stock and changes in real activity depends on whether a short interest rate is accounted for in the empirical analysis or not. Another problem of the concept is that it is not unambiguously possible to infer a causal relationship just from the chronology of two variables as demonstrated by Tobin (1970). This and other conceptual issues (see Zellner (1979) and the discussion in the next chapter) and econometric problems (Geweke, 1984) led to a decline in the practical importance of this concept.

We propose two econometric implementations of the causality concept. The first one is based on a VAR, the second one is non-parametric and uses the cross-correlations. In addition, we propose an interpretation in terms of the causal representation, respectively the Wold Decomposition Theorem (see Chapter 14)

### 15.1.1 VAR Approach

If one restricts oneself to linear least-squares forecasts, the above definition can be easily operationalized in the context of VAR models with only two

---

<sup>1</sup>Compare this to the concept of a causal representation developed in Sections 12.3 and 2.3.

<sup>2</sup>Sometimes also the concept of *contemporaneous causality* is considered. This concept is, however, controversial and has therefore not gained much success in practice and will, therefore, not be pursued.

variables (see also Sims, 1972). Consider first a VAR(1) model. Then according to the explanations in Chapter 14 the one-period forecast is:

$$\mathbb{P}_T X_{T+1} = \begin{pmatrix} \mathbb{P}_T X_{1,T+1} \\ \mathbb{P}_T X_{2,T+1} \end{pmatrix} = \Phi X_T = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \begin{pmatrix} X_{1,T} \\ X_{2,T} \end{pmatrix}$$

and therefore

$$\mathbb{P}_T X_{1,T+1} = \phi_{11} X_{1T} + \phi_{12} X_{2T}.$$

If  $\phi_{12} = 0$  then the second variable does not contribute to the one-period forecast of the first variable and can therefore be omitted:  $\text{MSE}_1(1) = \widetilde{\text{MSE}}_1(1)$ . Note that

$$\Phi^h = \begin{pmatrix} \phi_{11}^h & 0 \\ * & \phi_{22}^h \end{pmatrix},$$

where  $*$  is a placeholder for an arbitrary number. Thus the second variable is not only irrelevant for the one-period forecast, but for any forecast horizon  $h \geq 1$ . Thus, the second variable is *not causal* for the first variable in the sense of Wiener-Granger causality.

These arguments can be easily extended to VAR(p) models. According to equation (14.4) we have that

$$\mathbb{P}_T X_{1,T+1} = \phi_{11}^{(1)} X_{1T} + \phi_{12}^{(1)} X_{2T} + \dots + \phi_{11}^{(p)} X_{1,T-p+1} + \phi_{12}^{(p)} X_{2,T-p+1}$$

where  $\phi_{ij}^{(k)}$  denotes  $(i, j)$ -th element,  $i = 1, 2$ , of the matrix  $\Phi_k$ ,  $k = 1, \dots, p$ . In order for the second variable to have no influence on the forecast of the first one, we must have that  $\phi_{12}^{(1)} = \phi_{12}^{(2)} = \dots = \phi_{12}^{(p)} = 0$ . This implies that all matrices  $\Phi_k$ ,  $k = 1, \dots, p$ , must be lower triangular, i.e. they must be of the form  $\begin{pmatrix} * & 0 \\ * & * \end{pmatrix}$ . As the multiplication and addition of lower triangular matrices is again a lower triangular matrix, the second variable is irrelevant in forecasting the first one at any forecast horizon. This can be seen by computing the corresponding forecast function recursively as in Chapter 14.

Based on this insight it is straightforward to test the null hypothesis that the second variable does not cause the first one within the VAR(p) context:

$$H_0: \{X_{2t}\} \text{ does not cause } \{X_{1t}\}.$$

In terms of the VAR model this hypothesis can be stated as:

$$H_0: \phi_{12}^{(1)} = \phi_{12}^{(2)} = \dots = \phi_{12}^{(p)} = 0.$$

The alternative hypothesis is that the null hypothesis is violated. As the method of least-squares estimation leads on quite general conditions to asymptotically normal distributed coefficient estimates, it is straightforward to test the above hypothesis by a Wald-test (F-test). In the context of a VAR(1) model a simple t-test is also possible.

If more than two variables are involved the concept of Wiener-Granger causality is no longer so easy to implement. Consider for expositional purposes a VAR(1) model in three variables with coefficient matrix:

$$\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} & 0 \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix}.$$

The one-period forecast function of the first variable then is

$$\mathbb{P}_T X_{1,T+1} = \phi_{11} X_{1T} + \phi_{12} X_{2T}.$$

Thus, the third variable  $X_{3T}$  is irrelevant for the one-period forecast of the first variable. However, as the third variable has an influence on the second variable,  $\phi_{23} \neq 0$ , and because the second variable has an influence on the first variable,  $\phi_{12} \neq 0$ , the third variable will provide indirectly useful information for the forecast of the first variable for forecasting horizons  $h \geq 2$ . Consequently, the concept of causality cannot immediately be extended from two to more than two variables.

It is, however, possible to merge variables one and two, or variables two and three, into groups and discuss the hypothesis that the third variable does not cause the first two variables, seen as a group; likewise that the second and third variable, seen as a group, does not cause the first variable. The corresponding null hypotheses then are:

$$H_0 : \phi_{23} = \phi_{13} = 0 \quad \text{or} \quad H_0 : \phi_{12} = \phi_{13} = 0.$$

Under these null hypotheses we get again lower (block-) triangular matrices:

$$\begin{pmatrix} \phi_{11} & \phi_{12} & \vdots & 0 \\ \phi_{21} & \phi_{22} & \vdots & 0 \\ \dots & \dots & \vdots & \dots \\ \phi_{31} & \phi_{32} & \vdots & \phi_{33} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \phi_{11} & \vdots & 0 & 0 \\ \dots & \dots & \dots & \dots \\ \phi_{21} & \vdots & \phi_{22} & \phi_{23} \\ \phi_{31} & \vdots & \phi_{32} & \phi_{33} \end{pmatrix}.$$

Each of these hypotheses can again be checked by a Wald-test (F-test).

### 15.1.2 Wiener-Granger Causality and Causal Representation

We can get further insights into the concept of causality by considering a bivariate VAR,  $\Phi(L)X_t = Z_t$ , with causal representation  $X_t = \Psi(L)Z_t$ . Partitioning the matrices according to the two variables  $\{X_{1t}\}$  and  $\{X_{2t}\}$ , Theorem 12.1 of Section 12.3 implies that

$$\begin{pmatrix} \Phi_{11}(z) & \Phi_{12}(z) \\ \Phi_{21}(z) & \Phi_{22}(z) \end{pmatrix} \begin{pmatrix} \Psi_{11}(z) & \Psi_{12}(z) \\ \Psi_{21}(z) & \Psi_{22}(z) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where the polynomials  $\Phi_{12}(z)$  and  $\Phi_{21}(z)$  have no constant terms. The hypothesis that the second variable does not cause the first one is equivalent in this framework to the hypothesis that  $\Phi_{12}(z) = 0$ . Multiplying out the above expression leads to the condition

$$\Phi_{11}(z)\Psi_{12}(z) = 0.$$

Because  $\Phi_{11}(z)$  involves a constant term, the above equation implies that  $\Psi_{12}(z) = 0$ . Thus the causal representation is lower triangular. This means that the first variable is composed of the first shock,  $\{Z_{1t}\}$  only whereas the second variable involves both shocks  $\{Z_{1t}\}$  and  $\{Z_{2t}\}$ . The univariate causal representation of  $\{X_{1t}\}$  is therefore the same as the bivariate one.<sup>3</sup> Finally, note the similarity to the issue of the identification of shocks discussed in subsequent sections.

### 15.1.3 Cross-correlation Approach

In the case of two variables we also examine the cross-correlations to test for causality. This non-parametric test has the advantage that one does not have to rely on an explicit VAR model. This advantage becomes particularly relevant, if a VMA model must be approximated by a high order AR model. Consider the cross-correlations

$$\rho_{12}(h) = \text{corr}(X_{1t}, X_{2,t-h}).$$

If  $\rho_{12}(h) \neq 0$  for  $h > 0$ , we can say that the past values of the second variable are useful for forecasting the first variable such that the second variable causes the first one in the sense of Wiener and Granger. Another terminology is that the second variable is a *leading indicator* for the first one. If in addition,

---

<sup>3</sup>As we are working with causal VAR's, the above arguments also hold with respect to the Wold Decomposition.

$\rho_{12}(h) \neq 0$ , for  $h < 0$ , so that the past values of the first variable help to forecast the second one, we have causality in both directions.

As the distribution of the cross-correlations of two independent variables depends on the autocorrelation of each variable, see Theorem 11.4, Haugh (1976) and Pierce and Haugh (1977) propose a test based on the filtered time series. Analogously to the test for independence (see Section 11.2), we proceed in two steps:

**Step 1:** Estimate in the first step a univariate AR(p) model for each of the two time series  $\{X_{1t}\}$  and  $\{X_{2t}\}$ . Thereby chose  $p$  such that the corresponding residuals  $\{\hat{Z}_{1t}\}$  and  $\{\hat{Z}_{2t}\}$  are white noise. Note that although  $\{\hat{Z}_{1t}\}$  and  $\{\hat{Z}_{2t}\}$  are both not autocorrelated, the cross-correlations  $\rho_{Z_1, Z_2}(h)$  may still be non-zero for arbitrary orders  $h$ .

**Step 2:** As  $\{\hat{Z}_{1t}\}$  and  $\{\hat{Z}_{2t}\}$  are the forecast errors based on forecasts which rely only on the own past, the concept of causality carries over from the original variables to the residuals. The null hypothesis that the second variable does not cause the first variable in the sense of Wiener and Granger can then be checked by the *Haugh-Pierce statistic*:

$$\text{Haugh-Pierce statistic: } T \sum_{h=1}^L \hat{\rho}_{Z_1, Z_2}^2(h) \sim \chi_L^2. \quad (15.1)$$

Thereby  $\hat{\rho}_{Z_1, Z_2}^2(h)$ ,  $h = 1, 2, \dots$ , denotes the squared estimated cross-correlation coefficients between  $\{\hat{Z}_{1t}\}$  and  $\{\hat{Z}_{2t}\}$ . Under the null hypothesis that the second variable does not cause the first one, this test statistic is distributed as a  $\chi^2$  distribution with  $L$  degrees of freedom.

## 15.2 Structural and Reduced Form

### 15.2.1 A Prototypical Example

The discussion in the previous section showed that the relation between VAR models and economic models is ambiguous. In order to better understand the quintessence of the problem, we first analyze a simple macroeconomic example. Let  $\{y_t\}$  and  $\{m_t\}$  denote the output and the money supply of an economy<sup>4</sup> and suppose that the relation between the two variables is

<sup>4</sup>If one is working with actual data, the variables are usually expressed in log-differences to achieve stationarity.

represented by the following *simultaneous equation system*:

$$\begin{aligned} \text{AD-curve:} & & X_{1t} = y_t &= a_1 m_t + \gamma_{11} y_{t-1} + \gamma_{12} m_{t-1} + v_{yt} \\ \text{policy reaction curve:} & & X_{2t} = m_t &= a_2 y_t + \gamma_{21} y_{t-1} + \gamma_{22} m_{t-1} + v_{mt} \end{aligned}$$

Note the contemporaneous dependence of  $y_t$  on  $m_t$  in the AD-curve and a corresponding dependence of  $m_t$  on  $y_t$  in the policy reaction curve. These equations are typically derived from economic reasoning and may characterize a model explicitly derived from economic theory. In statistical terms the simultaneous equation system is called the *structural form*. The error terms  $\{v_{yt}\}$  and  $\{v_{mt}\}$  are interpreted as demand shocks and money supply shocks, respectively. They are called *structural shocks* and are assumed to follow a multivariate white noise process:

$$V_t = \begin{pmatrix} v_{yt} \\ v_{mt} \end{pmatrix} \sim \text{WN}(0, \Omega) \quad \text{with } \Omega = \begin{pmatrix} \omega_y^2 & 0 \\ 0 & \omega_m^2 \end{pmatrix}.$$

Note that the two structural shocks are assumed to be contemporaneously uncorrelated which is reflected in the assumption that  $\Omega$  is a diagonal matrix. This assumption in the literature is uncontroversial. Otherwise, there would remain some unexplained relationship between them. The structural shocks can be interpreted as the statistical analog of an experiment in the natural sciences. The experiment corresponds in this case to a shift of the AD-curve due to, for example, a temporary non-anticipated change in government expenditures or money supply. The goal of the analysis is then to trace the reaction of the economy, in our case represented by the two variables  $\{y_t\}$  and  $\{m_t\}$ , to these isolated and autonomous changes in aggregate demand and money supply. The structural equations imply that the reaction is not restricted to contemporaneous effects, but is spread out over time. We thus represent this reaction by the impulse response function.

We can write the system more compactly in matrix notation:

$$\begin{pmatrix} 1 & -a_1 \\ -a_2 & 1 \end{pmatrix} \begin{pmatrix} y_t \\ m_t \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ m_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_{yt} \\ v_{mt} \end{pmatrix}$$

or

$$AX_t = \Gamma X_{t-1} + BV_t$$

$$\text{where } A = \begin{pmatrix} 1 & -a_1 \\ -a_2 & 1 \end{pmatrix}, \Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Assuming that  $a_1a_2 \neq 1$ , we can solve the above simultaneous equation system for the two endogenous variables  $y_t$  and  $m_t$  to get the *reduced form* of the model:

$$\begin{aligned} X_{1t} = y_t &= \frac{\gamma_{11} + a_1\gamma_{21}}{1 - a_1a_2}y_{t-1} + \frac{\gamma_{12} + a_1\gamma_{22}}{1 - a_1a_2}m_{t-1} + \frac{v_{yt}}{1 - a_1a_2} + \frac{a_1v_{mt}}{1 - a_1a_2} \\ &= \phi_{11}y_{t-1} + \phi_{12}m_{t-1} + Z_{1t} \\ X_{2t} = m_t &= \frac{\gamma_{21} + a_2\gamma_{11}}{1 - a_1a_2}y_{t-1} + \frac{\gamma_{22} + a_2\gamma_{12}}{1 - a_1a_2}m_{t-1} + \frac{a_2v_{yt}}{1 - a_1a_2} + \frac{v_{mt}}{1 - a_1a_2} \\ &= \phi_{21}y_{t-1} + \phi_{22}m_{t-1} + Z_{2t}. \end{aligned}$$

Thus, the reduced form is a VAR(1) model with error term  $\{Z_t\} = \{(Z_{1t}, Z_{2t})'\}$ . The reduced form can also be expressed in matrix notation as:

$$\begin{aligned} X_t &= A^{-1}\Gamma X_{t-1} + A^{-1}BV_t \\ &= \Phi X_{t-1} + Z_t \end{aligned}$$

where

$$Z_t \sim \text{WN}(0, \Sigma) \quad \text{with} \quad \Sigma = A^{-1}B\Omega B'A^{-1}.$$

Whereas the structural form represents the inner economic relations between the variables (economic model), the reduced form given by the VAR model summarizes their outer directly observable characteristics. As there is no unambiguous relation between the reduced and structural form, it is impossible to infer the inner economic relationships from the observations alone. This is known in statistics as the *identification problem*. Typically, a whole family of structural models is compatible with a particular reduced form. The models of the family are thus *observationally equivalent* to each other as they imply the same distribution for  $\{X_t\}$ . The identification problem can be overcome if one is willing to make additional *a priori assumptions*. The nature and the type of these assumption and their interpretation is subject of the rest of this chapter.

In our example, the parameters characterizing the structural and the reduced form are

$$\{a_1, a_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \omega_y^2, \omega_m^2\}$$

and

$$\{\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}, \sigma_1^2, \sigma_{12}, \sigma_2^2\}.$$

As there are eight parameters in the structural form, but only seven parameters in the reduced form, there is no one-to-one relation between structural and reduced form. The VAR(1) model delivers estimates for the seven reduced form parameters, but there is no way to infer from these estimates the parameters of the structural form. Thus, there is a fundamental identification problem.

The simple counting of the number of parameters in each form tells us that we need at least one additional restriction on the parameters of the structural form. The simplest restriction is a zero restriction. Suppose that  $a_2$  equals zero, i.e. that the central bank does not react immediately to current output. This seems reasonable because national accounting figures are usually released with some delay. With this assumption, we can infer the structural parameters from reduced ones:

$$\begin{aligned}\gamma_{21} &= \phi_{21}, & \gamma_{22} &= \phi_{22}, \\ \nu_{mt} = Z_{2t} &\Rightarrow \omega_m^2 = \sigma_2^2, & \Rightarrow a_1 &= \sigma_{12}/\sigma_2^2, & \Rightarrow \omega_y^2 &= \sigma_1^2 - \sigma_{12}^2/\sigma_2^2 \\ \gamma_{11} &= \phi_{11} - (\sigma_{12}/\sigma_2^2)\phi_{21}, & \gamma_{12} &= \phi_{12} - (\sigma_{12}/\sigma_2^2)\phi_{22}.\end{aligned}$$

**Remark 15.1.** *Note that, because  $Z_t = A^{-1}BV_t$ , the reduced form disturbances  $Z_t$  are a linear combination of the structural disturbances, in our case the demand disturbance  $v_{yt}$  and the money supply disturbance  $v_{mt}$ . In each period  $t$  the endogenous variables output  $y_t$  and money supply  $m_t$  are therefore hit simultaneously by both shocks. It is thus not possible without further assumptions to assign the movements in  $Z_t$  and consequently in  $X_t$  to corresponding changes in the fundamental structural shocks  $v_{yt}$  and  $v_{mt}$ .*

**Remark 15.2.** *As Cooley and LeRoy (1985) already pointed out, the statement “money supply is not causal in the sense of Wiener and Granger for real economic activity”, which, in our example is equivalent to  $\phi_{12} = 0$ , is not equivalent to the statement “money supply does not influence real economic activity” because  $\phi_{12}$  can be zero without  $a_1$  being zero. Thus, the notion of causality is not very meaningful in inferring the inner (structural) relationships between variables.*

## 15.2.2 Identification: the General Case

We now present the general identification problem in the context of VAR.<sup>5</sup> The starting point of the analysis consists of a linear model, derived ideally

---

<sup>5</sup>A thorough treatment of the identification problem in econometrics can be found in Rothenberg (1971), and in the VAR context in Rubio-Ramírez et al. (2010).

from economic theory, in its *structural form*:

$$AX_t = \Gamma_1 X_{t-1} + \dots + \Gamma_p X_{t-p} + BV_t \quad (15.2)$$

where  $V_t$  are the structural disturbances. These disturbances usually have an economic interpretation, for example as demand or supply shocks.  $A$  is a  $n \times n$  matrix which is normalized such that the diagonal consists of ones only. The matrix  $B$  is also normalized such that its diagonal contains only ones. The process of structural disturbances,  $\{V_t\}$ , is assumed to be a multivariate white noise process with a diagonal covariance matrix  $\Omega$ :

$$V_t \sim \text{WN}(0, \Omega) \quad \text{with} \quad \Omega = \mathbb{E}V_t V_t' = \begin{pmatrix} \omega_1^2 & 0 & \dots & 0 \\ 0 & \omega_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_n^2 \end{pmatrix}.$$

The assumption that the structural disturbance are uncorrelated with each other is not viewed as controversial as otherwise there would be unexplained relationships between them. In the literature one encounters an alternative completely equivalent normalization which leaves the coefficients in  $B$  unrestricted but assumes the covariance matrix of  $V_t$ ,  $\Omega$ , to be equal to the identity matrix  $I_n$ .

The *reduced form* is obtained by solving the equation system with respect to  $X_t$ . Assuming that  $A$  is nonsingular, the premultiplication of equation (15.2) by  $A^{-1}$  leads to the reduced form which corresponds to a VAR(p) model:

$$\begin{aligned} X_t &= A^{-1}\Gamma_1 X_{t-1} + \dots + A^{-1}\Gamma_p X_{t-p} + A^{-1}BV_t \\ &= \Phi_1 X_{t-1} + \dots + \Phi_p X_{t-p} + Z_t. \end{aligned} \quad (15.3)$$

The relation between the structural disturbances  $V_t$  and the reduced form disturbances  $Z_t$  is in the form of a simultaneous equation system:

$$AZ_t = BV_t. \quad (15.4)$$

While the structural disturbances are not directly observed, the reduced form disturbances are given as the residuals of the VAR and can thus be considered as given. The relation between the lagged variables is simply

$$\Gamma_j = A\Phi_j, \quad j = 1, 2, \dots, p.$$

Consequently, once  $A$  and  $B$  have been identified, not only the coefficients of the lagged variables in the structural form are identified, but also the impulse

response functions (see Section 15.4.1). We can therefore concentrate our analysis of the identification problem on equation (15.4).

With these preliminaries, it is now possible to state the identification problem more precisely. Equation (15.4) shows that the structural form is completely determined by the parameters  $(A, B, \Omega)$ . Taking the normalization of  $A$  and  $B$  into account, these parameters can be viewed as points in  $\mathbb{R}^{n(2n-1)}$ . These parameters determine the distribution of  $Z_t = A^{-1}BV_t$  which is completely characterized by the covariance matrix of  $Z_t$ ,  $\Sigma$ , as the mean is equal to zero.<sup>6</sup> Thus, the parameters of the reduced form, i.e. the independent elements of  $\Sigma$  taking the symmetry into account, are points in  $\mathbb{R}^{n(n+1)/2}$ . The relation between structural and reduced form can therefore be described by a function  $g : \mathbb{R}^{n(2n-1)} \rightarrow \mathbb{R}^{n(n+1)/2}$ :

$$\Sigma = g(A, B, \Omega) = A^{-1}B\Omega B' A'^{-1}. \quad (15.5)$$

Ideally, one would want to find the inverse of this function and retrieve, in this way, the structural parameters  $(A, B, \Omega)$  from  $\Sigma$ . This is, however, in general not possible because the dimension of the domain space of  $g$ ,  $n(2n-1)$ , is strictly greater, for  $n \geq 2$ , than the dimension of its range space,  $n(n+1)/2$ . This discrepancy between the dimensions of the domain and the range space of  $g$  is known as the *identification problem*. To put it in another way, there are only  $n(n+1)/2$  (nonlinear) equations for  $n(2n-1)$  unknowns.<sup>7</sup>

To overcome the identification problem, we have to bring in additional information. A customary approach is to impose a priori assumptions on the structural parameters. The Implicit Function Theorem tells us that we need

$$3n(n-1)/2 = n(2n-1) - n(n+1)/2 \quad (15.6)$$

such restrictions, so-called identifying restrictions, to be able to invert the function  $g$ . Note that this is only a *necessary condition* and that the identification problem becomes more severe as the dimension of the VAR increases because the number of restrictions grows at a rate proportional to  $n^2$ .

This result can also be obtained by noting that the function  $g$  in equation (15.5) is invariant to the following transformation  $h$ :

$$h : (A, B, \Omega) \longrightarrow (RA, RB\Omega^{1/2}Q\Omega^{-1/2}, D\Omega D^{-1})$$

---

<sup>6</sup>As usual, we concentrate on the first two moments only.

<sup>7</sup>Note also that our discussion of the identification problem focuses on local identification, i.e. the invertibility of  $g$  in an open neighborhood of  $\Sigma$ . See Rothenberg (1971) and Rubio-Ramírez et al. (2010) for details on the distinction between local and global identification.

where  $R$ ,  $Q$  and  $D$  are arbitrary invertible matrices such that  $R$  respects the normalization of  $A$  and  $B$ ,  $Q$  is an orthogonal matrix, and  $D$  is a diagonal matrix. It can be verified that

$$(g \circ h)(A, B, \Omega) = g(A, B, \Omega).$$

The dimensions of the matrices  $R$ ,  $Q$ , and  $D$  are  $n^2 - 2n$ ,  $n(n - 1)/2$ , and  $n$ , respectively. Summing up gives  $3n(n - 1)/2 = n^2 - 2n + n(n - 1)/2 + n$  degrees of freedom as before.

The applied economics literature proposed several identification schemes:

- (i) Short-run restrictions (among many others, Sims, 1980b; Blanchard, 1989; Blanchard and Watson, 1986; Christiano et al., 1999)
- (ii) Long-run restrictions (Blanchard and Quah, 1989; Galí, 1992)
- (iii) Sign restrictions: this method restricts the set of possible impulse response functions (see Section 15.4.1) and can be seen as complementary to the other identification schemes (Faust, 1998; Uhlig, 2005; Fry and Pagan, 2011; Kilian and Murphy, 2012).
- (iv) Identification through heteroskedasticity (Rigobon, 2003)
- (v) Restrictions derived from a dynamic stochastic general equilibrium (DSGE) model. Typically, the identification issue is overcome by imposing a priori restrictions via a Bayesian approach (Negro and Schorfheide (2004) among many others).
- (vi) Identification using information on global versus idiosyncratic shocks in the context of multi-country or multi-region VAR models (Canova and Ciccarelli, 2008; Dees et al., 2007)
- (vii) Instead of identifying all parameters, researchers may be interested in identifying only one equation or a subset of equations. This case is known as *partial identification*. The schemes presented above can be extended in a straightforward manner to the partial identification case.

These schemes are not mutually exclusive, but can be combined with each other. In the following we will only cover the identification through short- and long-run restrictions, because these are by far the most popular ones. The economic importance of these restrictions for the analysis of monetary policy has been emphasized by Christiano et al. (1999).

### 15.2.3 Identification: the Case $n = 2$

Before proceeding further, it is instructive to analyze the case  $n = 2$  in more detail.<sup>8</sup> Assume for simplicity  $A = I_2$ , then the equation system (15.5) can be written explicitly as

$$\begin{aligned}\sigma_1^2 &= \omega_1^2 + (B)_{12}^2 \omega_2^2 \\ \sigma_{12} &= (B)_{21} \omega_1^2 + (B)_{12} \omega_2^2 \\ \sigma_2^2 &= (B)_{21}^2 \omega_1^2 + \omega_2^2\end{aligned}$$

with unknowns  $(B)_{12}$ ,  $(B)_{21}$ ,  $\omega_1^2$ , and  $\omega_2^2$ . Note that the assumption of  $\Sigma$  being positive definite implies that  $(B)_{12}(B)_{21} \neq 1$ . Thus, we can solve out  $\omega_1^2$  and  $\omega_2^2$  and reduce the three equations to only one:

$$((B)_{12} - b_{21}^{-1}) ((B)_{21} - b_{12}^{-1}) = r_{12}^{-2} - 1 \quad (15.7)$$

where  $b_{21}$  and  $b_{12}$  denote the least-squares regression coefficients of  $Z_{2t}$  on  $Z_{1t}$ , respectively of  $Z_{1t}$  on  $Z_{2t}$ , i.e.  $b_{21} = \sigma_{12}/\sigma_1^2$  and  $b_{12} = \sigma_{12}/\sigma_2^2$ .  $r_{12}$  is the correlation coefficient between  $Z_{2t}$  and  $Z_{1t}$ , i.e.  $r_{12} = \sigma_{12}/(\sigma_1\sigma_2)$ . Note that imposing a zero restriction by setting  $(B)_{12}$ , for example, equal to zero, implies that  $(B)_{21}$  equals  $b_{21}$ ; and vice versa, setting  $(B)_{21} = 0$ , implies  $(B)_{12} = b_{12}$ . As a final remark, the right hand side of equation (15.7) is always positive as the inverse of the squared correlation coefficient is bigger than one. This implies both product terms must be of the same sign.

Equation (15.7) delineates all possible combinations of  $(B)_{12}$  and  $(B)_{21}$  which are compatible with a given covariance matrix  $\Sigma$ . Its graph represents a rectangular hyperbola in the parameter space  $((B)_{12}, (B)_{21})$  with center  $C = (b_{21}^{-1}, b_{12}^{-1})$  and asymptotes  $(B)_{12} = b_{21}^{-1}$  and  $(B)_{21} = b_{12}^{-1}$  and is plotted in Figure 15.1. The hyperbola consist of two disconnected branches with a pole at the center  $C = (b_{21}^{-1}, b_{12}^{-1})$ . At this point, the relation between the two parameters changes sign. The Figure also indicates the two possible zero restrictions  $(B)_{12} = 0$  and  $(B)_{21} = 0$ , called short-run restrictions. These two restrictions are connected and its path completely falls within one quadrant. Thus, along this path the sign of the parameters remain unchanged.

Suppose that instead of fixing a particular parameter, we only want to restrict its sign. Assuming that  $(B)_{12} \geq 0$  implies that  $(B)_{21}$  must lie in one of the two disconnected intervals  $(-\infty, b_{21}]$  and  $(b_{12}^{-1}, +\infty)$ .<sup>9</sup> Although not very explicit, some economic consequences of this topological particularity are discussed in Fry and Pagan (2011). Alternatively, assuming  $(B)_{12} \leq 0$

<sup>8</sup>The exposition is inspired by Leamer (1981).

<sup>9</sup>That these two intervals are disconnected follows from the fact that  $\Sigma$  is positive definite.

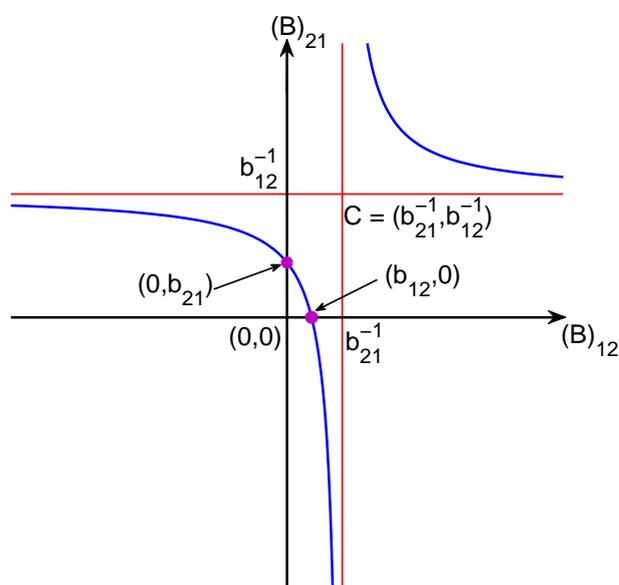


Figure 15.1: Identification in a two-dimensional structural VAR with  $\sigma_{12} > 0$

implies  $(B)_{21} \in [b_{21}, b_{12}^{-1})$ . Thus,  $(B)_{21}$  is unambiguously positive. Sign restrictions for  $(B)_{21}$  can be discussed in the same way.

### 15.3 Identification via Short-run Restrictions

Short-run restrictions represent the most common identification scheme encountered in practice. They impose direct linear restrictions on the structural parameters  $A$  and  $B$  and restrict in this way the contemporaneous effect of the structural shocks on the variables of the system. The most common type of such restrictions are *zero restrictions* which set certain coefficients a priori to zero. These zero restrictions are either derived from an explicit economic theory or are based on some ad hoc arguments. As explained above, it is necessary to have at least  $3n(n-1)/2$  restrictions at hand. If there are more restrictions, we have an *overidentified* system. This is, however, rarely the case in practice because the number of necessary restrictions grows at a rate proportional to  $n^2$ . The case of overidentification is, thus, not often encountered and as such is not treated.<sup>10</sup> The way to find appropriate restrictions in a relatively large system is documented in Section 15.4.5. If the number of

<sup>10</sup>The case of overidentification is, for example, treated in Bernanke (1986).

restrictions equals  $3n(n-1)/2$ , we say that the system is *exactly identified*.

Given the necessary number of a priori restrictions on the coefficients  $A$  and  $B$ , there are two ways to infer  $A$ ,  $B$ , and  $\Omega$ . The first one views the relation (15.4) as a simultaneous equation system in  $Z_t$  with error terms  $V_t$  and to estimate the coefficients by instrumental variables as in Blanchard and Watson (1986).<sup>11</sup> The second way relies on the method of moments and solves the nonlinear equation system (15.5) as in Bernanke (1986). In the case of exact identification both methods are numerically equivalent.

In our example treated of Section 15.2.1, we had  $n = 2$  so that three restrictions were necessary (six parameters, but only three equations). These three restrictions were obtained by setting  $B = I_2$  which gives two restrictions (i.e.  $b_{12} = b_{21} = 0$ ). The third restriction was to set the immediate reaction of money supply to a demand shock to zero, i.e. to set  $a_2 = 0$ . We then showed that these three restrictions are also sufficient to solve the nonlinear equation system (15.5).

Sims (1980b) proposed in his seminal article the VAR approach as an adequate alternative to then popular structural simultaneous approach. In particular, he suggested a simple recursive identification scheme. This scheme takes  $A = I_n$  so that the equation system (15.5) simplifies to:

$$\Sigma = B\Omega B'$$

Next we assume that  $B$  is a lower triangular matrix:

$$B = \begin{pmatrix} 1 & 0 & \dots & 0 \\ * & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & 1 \end{pmatrix}$$

where  $*$  is just a placeholder. The matrices  $B$  and  $\Omega$  are uniquely determined by the *Cholesky decomposition* of the matrix  $\Sigma$ . The Cholesky decomposition factorizes a positive-definite matrix  $\Sigma$  uniquely into the product  $B\Omega B'$  where  $B$  is a lower triangular matrix with ones on the diagonal and a diagonal matrix  $\Omega$  with strictly positive diagonal entries (Meyer, 2000). As  $Z_t = BV_t$ , Sims' identification gives rise to the following interpretation.  $v_{1t}$  is the only structural shock which has an effect on  $X_{1t}$  in period  $t$ . All other shocks have no contemporaneous effect. Moreover,  $Z_{1t} = v_{1t}$  so that the residual from the first equation is just equal to the first structural shock and that  $\sigma_1^2 = \omega_1^2$ . The second variable  $X_{2t}$  is contemporaneously only affected by  $v_{1t}$  and  $v_{2t}$ , and not by the remaining shocks  $v_{3t}, \dots, v_{nt}$ . In particular,  $Z_{2t} = b_{21}v_{1t} + v_{2t}$

<sup>11</sup>A version of the instrumental variable (IV) approach is discussed in Section 15.5.2.

so that  $b_{21}$  can be retrieved from the equation  $\sigma_{21} = b_{21}\omega_1^2$ . This identifies the second structural shock  $v_{2t}$  and  $\omega_2^2$ . Due to the triangular nature of  $B$ , the system is recursive and all structural shocks and parameters can be identified successively. The application of the Cholesky decomposition as an identification scheme rests crucially on the *ordering* of the variables  $(X_{1t}, X_{2t}, \dots, X_{nt})'$  in the system.

Sims' scheme, although easy to implement, becomes less plausible as the number of variables in the system increases. For this reason the more general scheme with  $A \neq I_n$  and  $B$  not necessarily lower triangular are more popular. However, even for medium sized systems such as  $n = 5, 30$  restrictions are necessary which stresses the imagination even of brilliant economists as the estimation of Blanchard's model in Section 15.4.5 shows.

Focusing on the identification of the matrices  $A$  and  $B$  brings also an advantage in terms of estimation. As shown in Chapter 13, the OLS-estimator of the VAR coefficient matrices  $\Phi_1, \dots, \Phi_p$  equals the GLS-estimator independently of  $\Sigma$ . Thus, the estimation of the structural parameters can be broken down into two steps. In the first step, the coefficient matrices  $\Phi_1, \dots, \Phi_p$  are estimated using OLS. The residuals are then used to estimate  $\Sigma$  which leads to an estimate of the covariance matrix (see equation (13.6)). In the second step, the coefficients of  $A$ ,  $B$ , and  $\Omega$  are then estimated given the estimate of  $\Sigma$ ,  $\hat{\Sigma}$ , by solving the nonlinear equation system (15.5) taking the specific identification scheme into account. Thereby  $\Sigma$  is replaced by its estimate  $\hat{\Sigma}$ . As  $\sqrt{T} \left( \text{vech}(\hat{\Sigma}) - \text{vech}(\Sigma) \right)$  converges in distribution to a normal distribution with mean zero,  $\hat{A}$ ,  $\hat{B}$  and  $\hat{\Omega}$  are also asymptotically normal because they are obtained by a one-to-one mapping from  $\hat{\Sigma}$ .<sup>12</sup> The Continuous Mapping Theorem further implies that  $\hat{A}$ ,  $\hat{B}$  and  $\hat{\Omega}$  converge to their true means and that their asymptotic covariance matrix can be obtained by an application of the delta-method (see Theorem E.1 in the Appendix E). Further details can be found in Bernanke (1986), Blanchard and Watson (1986), Giannini (1991), Hamilton (1994a), and Sims (1986).

## 15.4 Interpretation of VAR Models

### 15.4.1 Impulse Response Functions

The direct interpretation of VAR models is rather difficult because it is composed of many coefficients so that it becomes difficult to understand the

<sup>12</sup>The vech operator transforms a symmetric  $n \times n$  matrix  $\Sigma$  into a  $\frac{1}{2}n(n+1)$  vector by stacking the columns of  $\Sigma$  such that each element is listed only once.

dynamic interactions between the variables. It is therefore advantageous to simulate the dynamic effects of the different structural shocks by computing the *impulse response functions*. They show the effect over time of the structural shocks on the variables at issue. These effects can often be related to the underlying economic model and are thus at the heart of the VAR analysis. The examples in Section 15.4.4 and 15.4.5 provide some illustrations of this statement.

The impulse response functions are derived from the causal representation<sup>13</sup> of the VAR process (see Section 12.3):

$$\begin{aligned} X_t &= Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \dots \\ &= A^{-1} B V_t + \Psi_1 A^{-1} B V_{t-1} + \Psi_2 A^{-1} B V_{t-2} + \dots \end{aligned}$$

The effect of the  $j$ -th structural disturbance on the  $i$ -th variable after  $h$  periods, denoted by  $\frac{\partial X_{i,t+h}}{\partial v_{jt}}$  is thus given by the  $(i, j)$ -th element of the matrix  $\Psi_h A^{-1} B$ :

$$\frac{\partial X_{i,t+h}}{\partial v_{jt}} = [\Psi_h A^{-1} B]_{i,j}.$$

Clearly, the impulse response functions depends on the identification scheme chosen. There are  $n^2$  impulse response functions if the system consists of  $n$  variables. Usually, the impulse response functions are represented graphically as a plot against  $h$ .

## 15.4.2 Variance Decomposition

Another instrument for the interpretation of VAR models is the *forecast error variance decomposition* (“FEVD”) or variance decomposition for short which decomposes the total forecast error variance of a variable into the variances of the structural shocks. It is again based on the causal representation of the VAR(p) model. According to equation (14.3) in Chapter 14 the variance of the forecast error or mean squared error (MSE) is given by:

$$\begin{aligned} \text{MSE}(h) &= \mathbb{E}(X_{t+h} - \mathbb{P}_t X_{t+h})(X_{t+h} - \mathbb{P}_t X_{t+h})' \\ &= \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j' = \sum_{j=0}^{h-1} \Psi_j A^{-1} B \Omega B' A'^{-1} \Psi_j'. \end{aligned}$$

Given a specific identification scheme and estimates of the structural parameters, it is possible to attribute the MSE to the variance of the structural

<sup>13</sup>Sometimes the causal representation is called the MA( $\infty$ ) representation.

disturbances. Thereby it is customary to write the contribution of each disturbance as a percentage of the total variance. For this purpose let us write the  $\text{MSE}(h)$  explicitly as

$$\text{MSE}(h) = \begin{pmatrix} m_{11}^{(h)} & * & \dots & * \\ * & m_{22}^{(h)} & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & m_{nn}^{(h)} \end{pmatrix}.$$

Our interest lies exclusively on the variances,  $m_{ii}^{(h)}$ ,  $i = 1, \dots, n$ , so that we represent the uninteresting covariance terms by the placeholder  $*$ . These variances can be seen as a linear combination of the  $\omega_i^2$ 's because the covariance matrix of the structural disturbances  $\Omega = \text{diag}(\omega_1^2, \dots, \omega_n^2)$  is a diagonal matrix:

$$m_{ii}^{(h)} = d_{i1}^{(h)}\omega_1^2 + \dots + d_{in}^{(h)}\omega_n^2$$

where the weights  $d_{ij}^{(h)}$ ,  $i, j = 1, \dots, n$  and  $h = 1, 2, \dots$ , are strictly positive. They can be computed as

$$d_{ij}^{(h)} = \left( \sum_{k=0}^{h-1} [\Psi_k A^{-1} B]_{ij}^2 \right)$$

In order to arrive at the percentage value of the contribution of the  $j$ -th disturbance to the MSE of the  $i$ -th variable at forecast horizon  $h$ , we divide each summand in the above expression by the total sum:

$$100 \times \frac{d_{ij}^{(h)}\omega_j^2}{m_{ii}^{(h)}}, \quad i, j = 1, \dots, n, \text{ for } h = 0, 1, 2, \dots$$

Usually, these numbers are either displayed graphically as a plot against  $h$  or in table form (see the example in Section 15.4.5). These numbers show which percentage of the forecast variance at horizon  $h$  can be attributed to a particular structural shock and thus measures the contribution of each of these shocks to the overall fluctuations of the variables in question.

### 15.4.3 Confidence Intervals

The impulse response functions and the variance decomposition are the most important tools for the analysis and interpretation of VAR models. It is, therefore, of importance not only to estimate these entities, but also to provide corresponding confidence intervals to underpin the interpretation from

a statistical perspective. In the literature two approaches have been established: an analytic and a bootstrap approach. The analytic approach relies on the fact that the coefficient matrices  $\Psi_h$ ,  $h = 1, 2, \dots$ , are continuously differentiable functions of the estimated VAR coefficients:  $\text{vec}(\Psi_h) = F_h(\beta)$  where  $\beta$  denotes as in Chapter 13 the vectorized form of the VAR coefficient matrices  $\Phi_1, \dots, \Phi_p$ .<sup>14</sup> The relation between VAR coefficients and the causal representation (MA( $\infty$ ) representation) was established in Section 12.3. This discussion shows that the functions  $F_h : \mathbb{R}^{pn^2} \rightarrow \mathbb{R}^{n^2}$  are highly nonlinear. In Chapter 13 it was shown that  $\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma \otimes \Gamma_p^{-1})$  so that we can apply the Continuous Mapping Theorem (see theorem E.1 or Serfling (1980, 122-124)), sometimes also called the Delta method to get

$$\sqrt{T}(F_h(\hat{\beta}) - F_h(\beta)) \xrightarrow{d} N\left(0, \left(\frac{\partial F_h(\beta)}{\partial \beta'}\right) (\Sigma \otimes \Gamma_p^{-1}) \left(\frac{\partial F_h(\beta)}{\partial \beta'}\right)'\right)$$

where in practice the covariance matrix is estimated by replacing  $\beta$  and  $\Sigma$  by their estimate. The computation of the gradient matrices  $\frac{\partial F_h(\beta)}{\partial \beta'}$  is rather involved, especially when  $h$  becomes large. Details can be found in Lütkepohl (1990, 2006) and Mittnik and Zadrozny (1993).

The use of this asymptotic approximation has two problems. First, the complexity of the relationship between the  $\Phi_i$ 's and the  $\Psi_h$ 's augments with  $h$  so that the quality of the approximation diminishes with  $h$  for any given sample size. This is true even when  $\hat{\beta}$  is exactly normally distributed. Second, the distribution of  $\hat{\beta}$  is approximated poorly by the normal distribution. This is particularly relevant when the roots of  $\Phi(L)$  are near the unit circle. In this case, the bias towards zero can become substantial (see the discussion in Section 7.2). These two problems become especially relevant as  $h$  increases. For these reasons the analytic approach has become less popular.

The bootstrap approach (Monte Carlo or Simulation approach), as advocated by Runkle (1987), Kilian (1998) and Sims (1999), has become the most favored approach. This is partly due to the development of powerful computer algorithms, and the increased speed in computations. The so-called naive bootstrap approach consists of several steps.<sup>15</sup>

**First step:** Using a random number generator new disturbances are created.

This can be done in two ways. The first one assumes a particular distribution for  $V_t$ :  $V_t \sim N(0, \hat{\Omega})$ , for example. The realizations  $V_1, \dots, V_T$

<sup>14</sup>Recall that the  $\text{vec}$  operator stacks the columns of a  $n \times m$  matrix to get one  $nm$  vector.

<sup>15</sup>The bootstrap is a resampling method. Efron and Tibshirani (1993) provide a general introduction to the bootstrap.

are then independent draws from this distribution. The second one, takes random draws with replacement from the identified realizations  $\hat{V}_1, \dots, \hat{V}_T$ .<sup>16</sup> The second way has the advantage that non explicit distributional assumption is made which results in a better approximation of the true distribution of  $\hat{V}_t$ .

**Second step:** Given the fixed starting values  $X_{-p+1}, \dots, X_0$ , the estimated coefficients matrices  $\hat{\Phi}_1, \dots, \hat{\Phi}_p$  and the new disturbances drawn in step one, a new realization of the time series for  $\{X_t\}$  is generated.

**Third step:** Estimate the VAR model, given the newly generated realizations for  $\{X_t\}$ , to obtain new estimates for the coefficient matrices.

**Fourth step:** Generate a new set of impulse response functions given the new estimates, taking the identification scheme as fixed.

The steps one to four are repeated several times to generate a whole family of impulse response functions which form the basis for the computation of the confidence bands. In many applications, these confidence bands are constructed in a naive fashion by connecting the confidence intervals for individual impulse responses at different horizons. This, however, ignores the fact that the impulse responses at different horizons are correlated which implies that the true coverage probability of the confidence band is different from the presumed one. Thus, the joint probability distribution of the impulse responses should serve as the basis of the computation of the confidence bands. Recently, several alternatives have been proposed which take this feature in account. Lütkepohl et al. (2013) provides a comparison of several methods.

The number of repetitions should be at least 500. The method can be refined somewhat if the bias towards zero of the estimates of the  $\Phi$ 's is taken into account. This bias can again be determined through simulation methods (Kilian, 1998). A critical appraisal of the bootstrap can be found in Sims (1999) where additional improvements are discussed. The bootstrap of the variance decomposition works in similar way.

#### 15.4.4 Example 1: Advertisement and Sales

In this example we will analyze the dynamic relationship between advertisement expenditures and sales by a VAR approach. The data we will use are the famous data from the Lydia E. Pinkham Medicine Company which cover

---

<sup>16</sup>The draws can also be done blockwise. This has the advantage that possible remaining temporal dependences are taken in account.

yearly observations from 1907 to 1960. These data were among the first ones which have been used to quantify the effect of advertisement expenditures on sales. The data are taken from Berndt (1991, Chapter 8) where details on the specificities of the data and a summary of the literature can be found.

We denote the two-dimensional logged time series of advertisement expenditures and sales by  $\{X_t\} = \{(\ln(\text{advertisement}_t), \ln(\text{sales}_t))'\}$ . We consider VAR models of order one to six. For each VAR( $p$ ),  $p = 1, 2, \dots, 6$ , we compute the corresponding information criteria AIC and BIC (see Section 14.3). Whereas AIC favors a model of order two, BIC proposes the more parsimonious model of order one. To be on the safe side, we work with a VAR model of order two whose estimates are reported below:

$$X_t = \begin{pmatrix} 0.145 \\ (0.634) \\ 0.762 \\ (0.333) \end{pmatrix} + \begin{pmatrix} 0.451 & 0.642 \\ (0.174) & (0.302) \\ -0.068 & 1.245 \\ (0.091) & (0.159) \end{pmatrix} X_{t-1} \\ + \begin{pmatrix} -0.189 & 0.009 \\ (0.180) & (0.333) \\ -0.176 & -0.125 \\ (0.095) & (0.175) \end{pmatrix} X_{t-2} + Z_t,$$

where the estimated standard deviations of the coefficients are reported in parenthesis. The estimate  $\hat{\Sigma}$  of the covariance matrix  $\Sigma$  is

$$\hat{\Sigma} = \begin{pmatrix} 0.038 & 0.011 \\ 0.011 & 0.010 \end{pmatrix}.$$

The estimated VAR(2) model is taken to be the reduced form model. The structural model contains two structural shocks: a shock to advertisement expenditures,  $V_{At}$ , and a shock to sales,  $V_{St}$ . The disturbance vector of the structural shock is thus  $\{V_t\} = \{(V_{At}, V_{St})'\}$ . It is related to  $Z_t$  via relation (15.4), i.e.  $AZ_t = BV_t$ . To identify the model we thus need 3 restrictions.<sup>17</sup> We will first assume that  $A = I_2$  which gives two restrictions. A plausible further assumption is that shocks to sales have no contemporaneous effects on advertisement expenditures. This zero restriction seems justified because advertisement campaigns have to be planned in advance. They cannot be produced and carried out immediately. This argument then delivers

<sup>17</sup>Formula (15.6) for  $n = 2$  gives 3 restrictions.

the third restriction as it implies that  $B$  is a lower triangular. This lower triangular matrix can be obtained from the Cholesky decomposition of  $\widehat{\Sigma}$ :

$$\widehat{B} = \begin{pmatrix} 1 & 0 \\ 0.288 & 1 \end{pmatrix} \quad \text{and} \quad \widehat{\Omega} = \begin{pmatrix} 0.038 & 0 \\ 0 & 0.007 \end{pmatrix}.$$

The identifying assumptions then imply the impulse response functions plotted in Figure 15.2.

The upper left figure shows the response of a sudden transitory increase in advertisement expenditures by one percent (i.e. of a one-percent increase of  $V_{At}$ ) to itself. This shock is positively propagated to the future years, but is statistically zero after four years. After four years the shock even changes to negative, but statistically insignificant, expenditures. The same shock produces an increase of sales by 0.3 percent in the current and next year as shown in the lower left figure. The effect then deteriorates and becomes even negative after three years. The right hand figures display the reaction of a sudden transitory increase of sales by one percent. Again, we see that the shock is positively propagated. Thereby the largest effect is reached after two years and then declines monotonically. The reaction of advertisement expenditures is initially equal to zero by construction as it corresponds to the identifying assumption with regard to  $B$ . Then, the effect starts to increase and reaches a maximum after three years and then declines monotonically. After 15 years the effect is practically zero. The 95-percent confidence intervals are rather large so that all effects are no longer statistically significant after a few number of years.

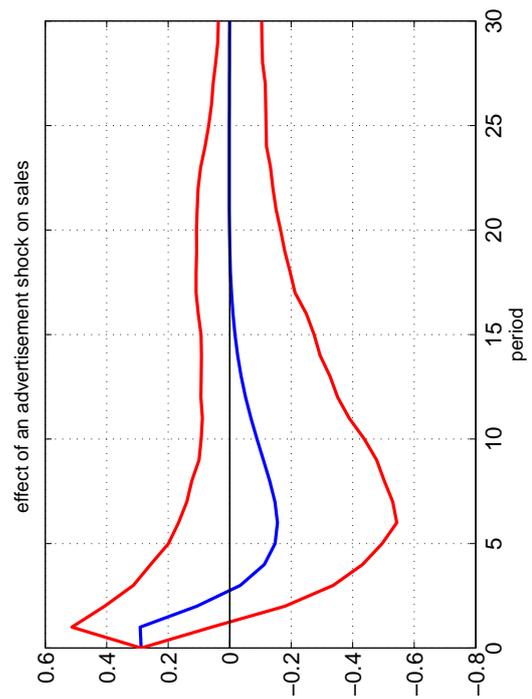
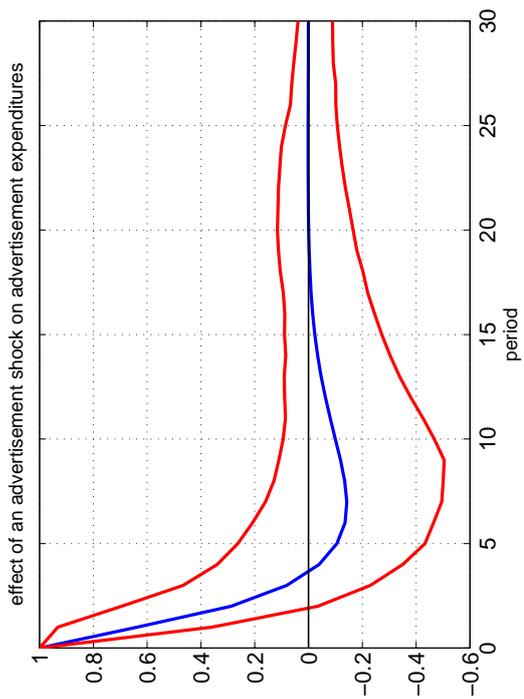
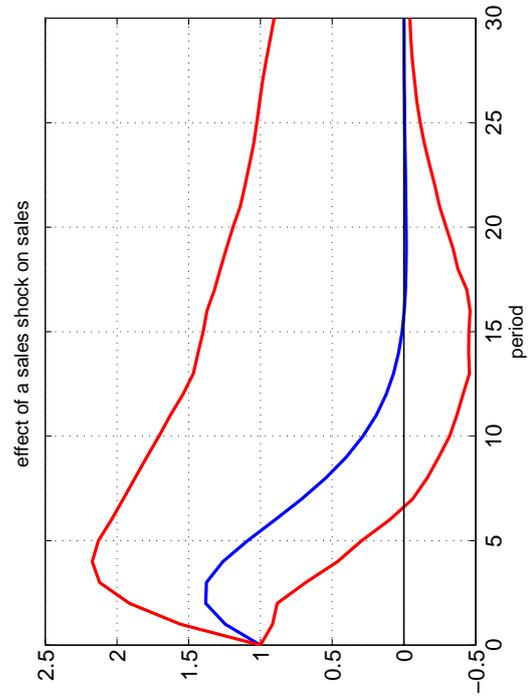
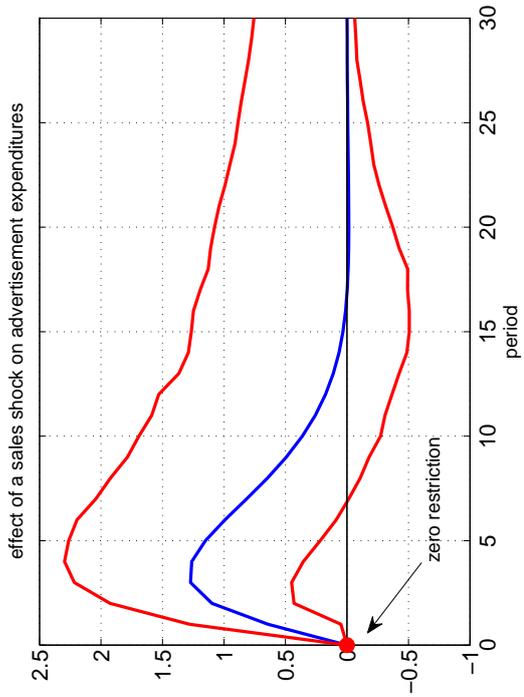
### 15.4.5 Example 2: IS-LM Model with Phillips Curve

In this example we replicate the study of Blanchard (1989) which investigates the US business cycle within a traditional IS-LM model with Phillips curve.<sup>18</sup> The starting point of his analysis is the VAR(p) model:

$$X_t = \Phi_1 X_{t-1} + \dots + \Phi_p X_{t-p} + C D_t + Z_t$$

---

<sup>18</sup>The results do not match exactly those of Blanchard (1989), but are qualitatively similar.



where  $\{X_t\}$  is a five-dimensional time series  $X_t = (Y_t, U_t, P_t, W_t, M_t)'$ . The individual elements of  $X_t$  denote the following variables:

- $Y_t$  ... growth rate of real GDP
- $U_t$  ... unemployment rate
- $P_t$  ... inflation rate
- $W_t$  ... growth rate of wages
- $M_t$  ... growth rate of money stock.

The VAR has attached to it a disturbance term  $Z_t = (Z_{yt}, Z_{ut}, Z_{pt}, Z_{wt}, Z_{mt})'$ . Finally,  $\{D_t\}$  denotes the deterministic variables of the model such as a constant, time trend or dummy variables. In the following, we assume that all variables are stationary.

The business cycle is seen as the result of five structural shocks which impinge on the economy:

- $V_{dt}$  ... aggregate demand shock
- $V_{st}$  ... aggregate supply shock
- $V_{pt}$  ... price shock
- $V_{wt}$  ... wage shock
- $V_{mt}$  ... money shock.

We will use the IS-LM model to rationalize the restrictions so that we will be able to identify the structural form from the estimated VAR model. The disturbance of the structural and the reduced form models are related by the simultaneous equation system:

$$AZ_t = BV_t$$

where  $V_t = (V_{yt}, V_{st}, V_{pt}, V_{wt}, V_{mt})'$  and where  $A$  and  $B$  are  $5 \times 5$  matrices with ones on the diagonal. Blanchard (1989) proposes the following specification:

$$\begin{aligned} \text{(AD):} \quad & Z_{yt} = V_{dt} + b_{12}V_{st} \\ \text{(OL):} \quad & Z_{ut} = -a_{21}Z_{yt} + V_{st} \\ \text{(PS):} \quad & Z_{pt} = -a_{34}Z_{wt} - a_{31}Z_{yt} + b_{32}V_{st} + V_{pt} \\ \text{(WS):} \quad & Z_{wt} = -a_{43}Z_{pt} - a_{42}Z_{ut} + b_{42}V_{st} + V_{wt} \\ \text{(MR):} \quad & Z_{mt} = -a_{51}Z_{yt} - a_{52}Z_{ut} - a_{53}Z_{pt} - a_{54}Z_{wt} + V_{mt}. \end{aligned}$$

In matrix notation the above simultaneous equation system becomes:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ a_{21} & 1 & 0 & 0 & 0 \\ a_{31} & 0 & 1 & a_{34} & 0 \\ 0 & a_{42} & a_{43} & 1 & 0 \\ a_{51} & a_{52} & a_{53} & a_{54} & 1 \end{pmatrix} \begin{pmatrix} Z_{yt} \\ Z_{ut} \\ Z_{pt} \\ Z_{wt} \\ Z_{mt} \end{pmatrix} = \begin{pmatrix} 1 & b_{12} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & b_{32} & 1 & 0 & 0 \\ 0 & b_{42} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} V_{dt} \\ V_{st} \\ V_{pt} \\ V_{wt} \\ V_{mt} \end{pmatrix}.$$

The first equation is interpreted as an aggregate demand (AD) equation where the disturbance term related to GDP growth,  $Z_{yt}$ , depends on the demand shock  $V_{dt}$  and the supply shock  $V_{st}$ . The second equation is related to Okun's law (OL) which relates the unemployment disturbance  $Z_{ut}$  to the demand disturbance and the supply shock. Thereby an increase in GDP growth reduces unemployment in the same period by  $a_{21}$  whereas a supply shock increases it. The third and the fourth equation represent a price (PS) and wage setting (WS) system where wages and prices interact simultaneously. Finally, the fifth equation (MR) is supposed to determine the money shock (MR). No distinction is made between money supply and money demand shocks. A detailed interpretation of these equations is found in the original article by Blanchard (1989).

Given that the dimension of the system is five (i.e.  $n = 5$ ), formula (15.6) instructs us that we need  $3 \times (5 \times 4)/2 = 30$  restrictions. Counting the number of zero restrictions implemented above, we see that we only have 28 zeros. Thus we lack two additional restrictions. We can reach the same conclusion by counting the number of coefficients and the number of equations. The coefficients are  $a_{21}, a_{31}, a_{34}, a_{42}, a_{43}, a_{51}, a_{52}, a_{53}, a_{54}, b_{12}, b_{32}, b_{42}$  and the diagonal elements of  $\Omega$ , the covariance matrix of  $V_t$ . We therefore have to determine 17 unknown coefficients out of  $(5 \times 6)/2 = 15$  equations. Thus we find again that we are short of two restrictions. Blanchard discusses several possibilities among which the restrictions  $b_{12} = 1.0$  and  $a_{34} = 0.1$  seem most plausible.

The sample period runs through the second quarter in 1959 to the second quarter in 2004 encompassing 181 observations. Following Blanchard, we include a constant in combination with a linear time trend in the model. Whereas BIC suggests a model of order one, AIC favors a model of order two. As a model of order one seems rather restrictive, we stick to the VAR(2)

model whose estimated coefficients are reported below:<sup>19</sup>

$$\begin{aligned}\widehat{\Phi}_1 &= \begin{pmatrix} 0.07 & -1.31 & 0.01 & 0.12 & 0.02 \\ -0.02 & 1.30 & 0.03 & -0.00 & -0.00 \\ -0.07 & -1.47 & 0.56 & 0.07 & 0.03 \\ 0.07 & 0.50 & 0.44 & 0.07 & 0.06 \\ -0.10 & 1.27 & -0.07 & 0.04 & 0.49 \end{pmatrix} \\ \widehat{\Phi}_2 &= \begin{pmatrix} 0.05 & 1.79 & -0.41 & -0.13 & 0.05 \\ -0.02 & -0.35 & 0.00 & 0.01 & -0.00 \\ -0.04 & 1.38 & 0.28 & 0.05 & -0.00 \\ 0.07 & -0.85 & 0.19 & 0.10 & -0.04 \\ -0.02 & -0.77 & -0.07 & 0.11 & 0.17 \end{pmatrix} \\ \widehat{C} &= \begin{pmatrix} 2.18 & -0.0101 \\ 0.29 & 0.0001 \\ 0.92 & -0.0015 \\ 4.06 & -0.0035 \\ -0.98 & -0.0025 \end{pmatrix} \\ \widehat{\Sigma} &= \begin{pmatrix} 9.94 & -0.46 & -0.34 & 0.79 & 0.29 \\ -0.46 & 0.06 & -0.02 & -0.05 & -0.06 \\ -0.34 & -0.02 & 1.06 & 0.76 & 0.13 \\ 0.79 & -0.05 & 0.76 & 5.58 & 0.76 \\ 0.29 & -0.06 & 0.13 & 0.76 & 11.07 \end{pmatrix}.\end{aligned}$$

The first column of  $\widehat{C}$  relates to the constants, whereas the second column gives the coefficients of the time trend. From these estimates and given the identifying restrictions established above, the equation  $\widehat{\Sigma} = A^{-1}B\Omega B'A'^{-1}$

---

<sup>19</sup>To save space, the estimated standard errors of the coefficients are not reported.

uniquely determines the matrices  $A$ ,  $B$  and  $\Omega$ :

$$\hat{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.050 & 1 & 0 & 0 & 0 \\ 0.038 & 0 & 1 & 0.1 & 0 \\ 0 & 1.77 & -0.24 & 1 & 0 \\ 0.033 & 1.10 & 0.01 & -0.13 & 1 \end{pmatrix}$$

$$\hat{B} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -1.01 & 1 & 0 & 0 \\ 0 & 1.55 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\hat{\Omega} = \begin{pmatrix} 9.838 & 0 & 0 & 0 & 0 \\ 0 & 0.037 & 0 & 0 & 0 \\ 0 & 0 & 0.899 & 0 & 0 \\ 0 & 0 & 0 & 5.162 & 0 \\ 0 & 0 & 0 & 0 & 10.849 \end{pmatrix}$$

In order to give better interpretation of the results we have plotted the impulse response functions and their 95-percent confidence bands in Figures 15.3. The results show that a positive demand shock has only a positive and statistically significant effect on GDP growth in the first three quarters, after that the effect becomes even slightly negative and vanishes after sixteen quarters. The positive demand shock reduces unemployment significantly for almost fifteen quarters. The maximal effect is achieved after three to four quarters. Although the initial effect is negative, the positive demand shock also drives inflation up which then pushes up wage growth. The supply shock also has a positive effect on GDP growth, but it takes more than four quarters before the effect reaches its peak. In the short-run the positive supply shock even reduces GDP growth. In contrast to the demand shock, the positive supply shock increases unemployment in the short-run. The effect will only reduce unemployment in the medium- to long-run. The effect on price and wage inflation is negative.

Finally, we compute the forecast error variance decomposition according to Section 15.4.2. The results are reported in Table 15.1. In the short-run, the identifying restrictions play an important role as reflected by the plain zeros. The demand shock accounts for almost all the variance of GDP growth in the short-run. The value of 99.62 percent for forecast horizon of one quarter, however, diminishes as  $h$  increases to 40 quarters, but still remains with a value 86.13 very high. The supply shock on the contrary does not explain much of the variation in GDP growth. Even for a horizon of 40 quarters,

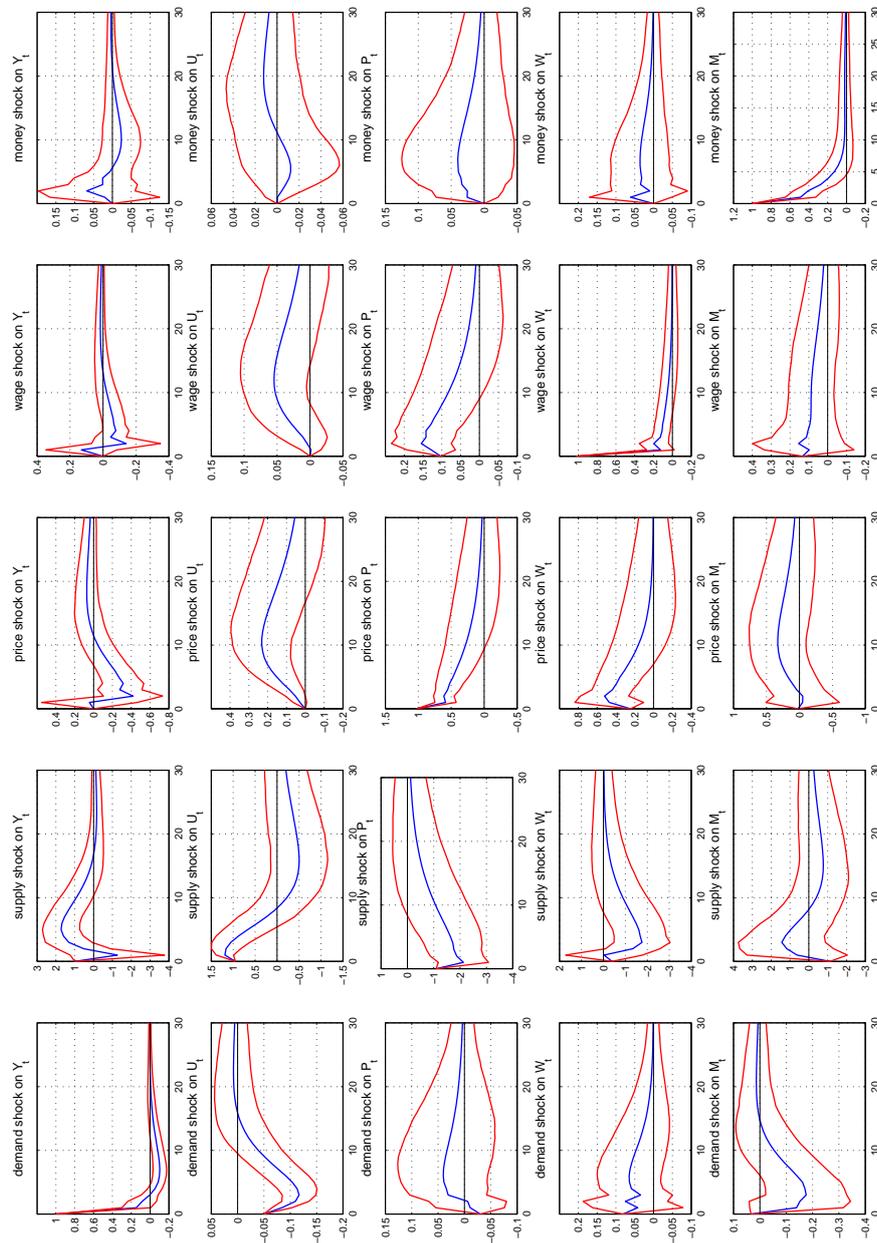


Figure 15.3: Impulse response functions for the IS-LM model with Phillips curve with 95-percent confidence intervals computed using the bootstrap procedure (compare with Blanchard (1989))

it contributes only 6.11 percent. The supply shock is, however, important for the variation in the unemployment rate, especially in the short-run. It explains more than 50 percent whereas demand shocks account for only 42.22 percent. Its contribution diminishes with the increase of the forecast horizon giving room for price and wage shocks. The variance of the inflation rate is explained in the short-run almost exclusively by price shocks. However, as the the forecast horizon is increased supply and wage shocks become relatively important. The money growth rate does not interact much with the other variables. Its variation is almost exclusively explained by money shocks.

## 15.5 Identification via Long-Run Restrictions

### 15.5.1 A Prototypical Example

Besides short-run restrictions, essentially zero restrictions on the coefficients of  $A$  and/or  $B$ , Blanchard and Quah (1989) proposed *long-run restrictions* as an alternative option. These long-run restrictions have to be seen as complementary to the short-run ones as they can be combined. Long-run restrictions constrain the long-run effect of structural shocks. This technique makes only sense if some integrated variables are involved, because in the stationary case the effects of all shocks vanish eventually. To explain this, we discuss the two-dimensional example given by Blanchard and Quah (1989).

They analyze a two-variable system consisting of logged real GDP denoted by  $\{Y_t\}$  and the unemployment rate  $\{U_t\}$ . Logged GDP is typically integrated of order one (see Section 7.3.4 for an analysis for Swiss GDP) whereas  $U_t$  is considered to be stationary. Thus they apply the VAR approach to the stationary process  $\{X_t\} = \{(\Delta Y_t, U_t)'\}$ . Assuming that  $\{X_t\}$  is already demeaned and follows a causal VAR process, we have the following representations:

$$\begin{aligned} X_t &= \begin{pmatrix} \Delta Y_t \\ U_t \end{pmatrix} = \Phi_1 X_{t-1} + \dots + \Phi_p X_{t-p} + Z_t \\ &= \Psi(L)Z_t = Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \dots \end{aligned}$$

For simplicity, we assume that  $A = I_2$ , so that

$$Z_t = BV_t = \begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} v_{dt} \\ v_{st} \end{pmatrix}$$

where  $V_t = (v_{dt}, v_{st})' \sim \text{WN}(0, \Omega)$  with  $\Omega = \text{diag}(\omega_d^2, \omega_s^2)$ . Thereby  $\{v_{dt}\}$  and  $\{v_{st}\}$  denote demand and supply shocks, respectively. The causal representation of  $\{X_t\}$  implies that the effect of a demand shock in period  $t$  on GDP

Table 15.1: Forecast error variance decomposition (FEVD) in terms of demand, supply, price, wage, and money shocks (percentages)

<i>growth rate of real GDP</i>					
horizon	demand	supply	price	wage	money
1	99.62	0.38	0	0	0
2	98.13	0.94	0.02	0.87	0.04
4	93.85	1.59	2.13	1.86	0.57
8	88.27	4.83	3.36	2.43	0.61
40	86.13	6.11	4.29	2.58	0.89
<i>unemployment rate</i>					
horizon	demand	supply	price	wage	money
1	42.22	57.78	0	0	0
2	52.03	47.57	0.04	0.01	0.00
4	64.74	33.17	1.80	0.13	0.16
8	66.05	21.32	10.01	1.99	0.63
40	39.09	16.81	31.92	10.73	0.89
<i>inflation rate</i>					
horizon	demand	supply	price	wage	money
1	0.86	4.18	89.80	5.15	0
2	0.63	13.12	77.24	8.56	0.45
4	0.72	16.79	68.15	13.36	0.97
8	1.79	19.34	60.69	16.07	2.11
40	2.83	20.48	55.84	17.12	3.74
<i>growth rate of wages</i>					
horizon	demand	supply	price	wage	money
1	1.18	0.10	0.97	97.75	0
2	1.40	0.10	4.30	93.50	0.69
4	2.18	2.75	9.78	84.49	0.80
8	3.80	6.74	13.40	74.72	1.33
40	5.11	8.44	14.19	70.14	2.13
<i>growth rate of money stock</i>					
horizon	demand	supply	price	wage	money
1	0.10	0.43	0.00	0.84	98.63
2	1.45	0.44	0.02	1.02	97.06
4	4.22	1.09	0.04	1.90	92.75
8	8.31	1.55	0.81	2.65	86.68
40	8.47	2.64	5.77	4.55	78.57

growth in period  $t + h$  is given by:

$$\frac{\partial \Delta Y_{t+h}}{\partial v_{dt}} = [\Psi_h B]_{11}$$

where  $[\Psi_h B]_{11}$  denotes the upper left hand element of the matrix  $\Psi_h B$ .  $Y_{t+h}$  can be written as  $Y_{t+h} = \Delta Y_{t+h} + \Delta Y_{t+h-1} + \dots + \Delta Y_{t+1} + Y_t$  so that the effect of the demand shock on the *level* of logged GDP is given by:

$$\frac{\partial Y_{t+h}}{\partial v_{dt}} = \sum_{j=0}^h [\Psi_j B]_{11} = \left[ \sum_{j=0}^h \Psi_j B \right]_{11}.$$

Blanchard and Quah (1989) propose, in accordance with conventional economic theory, to restrict the long-run effect of the demand shock on the level of logged GDP to zero:

$$\lim_{h \rightarrow \infty} \frac{\partial Y_{t+h}}{\partial v_{dt}} = \sum_{j=0}^{\infty} [\Psi_j B]_{11} = 0.$$

This implies that

$$\sum_{j=0}^{\infty} \Psi_j B = \left( \sum_{j=0}^{\infty} \Psi_j \right) B = \Psi(1)B = \begin{pmatrix} 0 & * \\ * & * \end{pmatrix},$$

where  $*$  is a placeholder. This restriction is sufficient to infer  $b_{21}$  from the relation  $[\Psi(1)]_{11}b_{11} + b_{21}[\Psi(1)]_{12} = 0$  and the normalization  $b_{11} = 1$ :

$$b_{21} = -\frac{[\Psi(1)]_{11}}{[\Psi(1)]_{12}} = -\frac{[\Phi(1)^{-1}]_{11}}{[\Phi(1)^{-1}]_{12}}.$$

The second part of the equation follows from the identity  $\Phi(z)\Psi(z) = I_2$  which gives  $\Psi(1) = \Phi(1)^{-1}$  for  $z = 1$ . The long-run effect of the supply shock on  $Y_t$  is left unrestricted and is therefore in general nonzero. Note that the implied value of  $b_{21}$  depends on  $\Phi(1)$ , and thus on  $\Phi_1, \dots, \Phi_p$ . The results are therefore, in contrast to short-run restrictions, much more sensitive to correct specification of the VAR.

The relation  $Z_t = BV_t$  implies that

$$\Sigma = B \begin{pmatrix} \omega_d^2 & 0 \\ 0 & \omega_s^2 \end{pmatrix} B'$$

or more explicitly

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} \omega_d^2 & 0 \\ 0 & \omega_s^2 \end{pmatrix} \begin{pmatrix} 1 & b_{21} \\ b_{12} & 1 \end{pmatrix}.$$

Taking  $b_{21}$  as already given from above, this equation system has three equations and three unknowns  $b_{12}, \omega_d^2, \omega_s^2$  which is a necessary condition for a solution to exist.

### Analytic solution of the system

Because  $b_{21}$  is already determined from the long-run restriction, we rewrite the equation system<sup>20</sup> explicitly in terms of  $b_{21}$ :

$$\begin{aligned}\sigma_1^2 &= \omega_d^2 + b_{12}^2 \omega_s^2 \\ \sigma_{12} &= b_{21} \omega_d^2 + b_{12} \omega_s^2 \\ \sigma_2^2 &= b_{21}^2 \omega_d^2 + \omega_s^2.\end{aligned}$$

Using the last two equations, we can express  $\omega_d^2$  and  $\omega_s^2$  as functions of  $b_{12}$ :

$$\begin{aligned}\omega_d^2 &= \frac{\sigma_{12} - b_{12} \sigma_2^2}{b_{21} - b_{12} b_{21}^2} \\ \omega_s^2 &= \frac{\sigma_2^2 - b_{21} \sigma_{12}}{1 - b_{12} b_{21}}.\end{aligned}$$

These expressions are only valid if  $b_{21} \neq 0$  and  $b_{12} b_{21} \neq 1$ . The case  $b_{21} = 0$  is not interesting with regard to content. It would simplify the original equation system strongly and would result in  $b_{12} = \sigma_{12}/\sigma_2^2$ ,  $\omega_d^2 = (\sigma_1^2 \sigma_2^2 - \sigma_{12}^2)/\sigma_2^2 > 0$  and  $\omega_s^2 = \sigma_2^2 > 0$ . The case  $b_{12} b_{21} = 1$  contradicts the assumption that  $\Sigma$  is a positive-definite matrix and can therefore be disregarded.<sup>21</sup>

Inserting the solutions of  $\omega_d^2$  and  $\omega_s^2$  into the first equation, we obtain a quadratic equation in  $b_{12}$ :

$$(b_{21} \sigma_2^2 - b_{21}^2 \sigma_{12}) b_{12}^2 + (b_{21}^2 \sigma_1^2 - \sigma_2^2) b_{12} + (\sigma_{12} - b_{21} \sigma_1^2) = 0.$$

The discriminant  $\Delta$  of this equation is:

$$\begin{aligned}\Delta &= (b_{21}^2 \sigma_1^2 - \sigma_2^2)^2 - 4 (b_{21} \sigma_2^2 - b_{21}^2 \sigma_{12}) (\sigma_{12} - b_{21} \sigma_1^2) \\ &= (b_{21}^2 \sigma_1^2 + \sigma_2^2)^2 - 4 b_{21} \sigma_{12} (b_{21}^2 \sigma_1^2 + \sigma_2^2) + 4 b_{21}^2 \sigma_{12}^2 \\ &= (b_{21}^2 \sigma_1^2 + \sigma_2^2 - 2 b_{21} \sigma_{12})^2 > 0.\end{aligned}$$

<sup>20</sup>This equation system is similar to the one analyzed in Section 15.2.3.

<sup>21</sup>If  $b_{12} b_{21} = 1$ ,  $\det \Sigma = \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 = 0$ . This implies that  $Z_{1t}$  and  $Z_{2t}$  are perfectly correlated, i.e.  $\rho_{Z_{1t}, Z_{2t}}^2 = \sigma_{12}^2 / (\sigma_1^2 \sigma_2^2) = 1$ .

The positivity of the discriminant implies that the quadratic equation has two real solutions  $b_{12}^{(1)}$  and  $b_{12}^{(2)}$ :

$$b_{12}^{(1)} = \frac{\sigma_2^2 - b_{21}\sigma_{12}}{b_{21}\sigma_2^2 - b_{21}^2\sigma_{12}} = \frac{1}{b_{21}},$$

$$b_{12}^{(2)} = \frac{\sigma_{12} - b_{21}\sigma_1^2}{\sigma_2^2 - b_{21}\sigma_{12}}.$$

The first solution  $b_{12}^{(1)}$  can be excluded because it violates the assumption  $b_{12}b_{21} \neq 1$  which stands in contradiction to the positive-definiteness of the covariance matrix  $\Sigma$ . Inserting the second solution back into the solution for  $\omega_d^2$  and  $\omega_s^2$ , we finally get:

$$\omega_d^2 = \frac{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}{b_{21}^2\sigma_1^2 - 2b_{21}\sigma_{12} + \sigma_2^2} > 0$$

$$\omega_s^2 = \frac{(\sigma_2^2 - b_{21}\sigma_{12})^2}{b_{21}^2\sigma_1^2 - 2b_{21}\sigma_{12} + \sigma_2^2} > 0.$$

Because  $\Sigma$  is a symmetric positive-definite matrix,  $\sigma_1^2\sigma_2^2 - \sigma_{12}^2$  and the denominator  $b_{21}^2\sigma_1^2 - 2b_{21}\sigma_{12} + \sigma_2^2$  are strictly positive. Thus, both solutions yield positive variances and we have found the unique admissible solution.

### 15.5.2 The General Approach

The general case of long-run restrictions has a structure similar to the case of short-run restrictions. Take as a starting point again the structural VAR (15.2) from Section 15.2.2:

$$AX_t = \Gamma_1 X_{t-1} + \dots + \Gamma_p X_{t-p} + BV_t, \quad V_t \sim \text{WN}(0, \Omega),$$

$$A(L)X_t = BV_t$$

where  $\{X_t\}$  is stationary and causal with respect to  $\{V_t\}$ . As before the matrix  $A$  is normalized to have ones on its diagonal and is assumed to be invertible,  $\Omega$  is a diagonal matrix with  $\Omega = \text{diag}(\omega_1^2, \dots, \omega_n^2)$ , and  $B$  is a matrix with ones on the diagonal. The matrix polynomial  $A(L)$  is defined as  $A(L) = A - \Gamma_1 L - \dots - \Gamma_p L^p$ . The reduced form is given by

$$\Phi(L)X_t = Z_t, \quad Z_t \sim \text{WN}(0, \Sigma)$$

where  $AZ_t = BV_t$  and  $A\Phi_j = \Gamma_j$ ,  $j = 1, \dots, p$ , respectively  $A\Phi(L) = A(L)$ .

The long-run variance of  $\{X_t\}$ ,  $J$ , (see equation (11.1) in Chapter 11) can be derived from the reduced as well as from the structural form, which gives the following expressions:

$$\begin{aligned} J &= \Phi(1)^{-1}\Sigma\Phi(1)^{-1'} = \Phi(1)^{-1}A^{-1}B\Omega B'A'^{-1}\Phi(1)^{-1'} = A(1)^{-1}B\Omega B'A(1)^{-1'} \\ &= \Psi(1)\Sigma\Psi(1)' = \Psi(1)A^{-1}B\Omega B'A'^{-1}\Psi(1)' \end{aligned}$$

where  $X_t = \Psi(L)Z_t$  denotes the causal representation of  $\{X_t\}$ . The long-run variance  $J$  can be estimated by adapting the methods in Section 4.4 to the multivariate case. Thus, the above equation system has a similar structure as the system (15.5) which underlies the case of short-run restrictions. As before, we get  $n(n+1)/2$  equations with  $2n^2 - n$  unknowns. The nonlinear equation system is therefore undetermined for  $n \geq 2$ . Therefore,  $3n(n-1)/2$  additional equations or restrictions are necessary to achieve identification. Hence, conceptually we are in a similar situation as in the case of short-run restrictions.<sup>22</sup>

In practice, it is customary to achieve identification through zero restrictions where some elements of  $\Psi(1)A^{-1}B$ , respectively  $\Phi(1)^{-1}A^{-1}B$ , are set a priori to zero. Setting the  $ij$ -th element  $[\Psi(1)A^{-1}B]_{ij} = [\Phi(1)^{-1}A^{-1}B]_{ij}$  equal to zero amounts to set the *cumulative* effect of the  $j$ -th structural disturbance  $V_{jt}$  on the  $i$ -th variable equal to zero. If the  $i$ -th variable enters  $X_t$  in first differences, as was the case for  $Y_t$  in the previous example, this zero restriction restrains the long-run effect on the level of that variable.

An interesting simplification arises if one assumes that  $A = I_n$  and that  $\Psi(1)B = \Phi(1)^{-1}B$  is a lower triangular matrix. In this case,  $B$  and  $\Omega$  can be estimated from the Cholesky decomposition of the estimated long-run variance  $\hat{J}$ . Let  $\hat{J} = \hat{L}\hat{D}\hat{L}'$  be the Cholesky decomposition where  $\hat{L}$  is lower triangular matrix with ones on the diagonal and  $\hat{D}$  is a diagonal matrix with strictly positive diagonal entries. As  $\hat{J} = \hat{\Phi}(1)^{-1}B\hat{\Omega}B'\hat{\Phi}(1)^{-1'}$ , the matrix of structural coefficients can then be estimated as  $\hat{B} = \hat{\Phi}(1)\hat{L}\hat{U}^{-1}$ . The multiplication by the inverse of the diagonal matrix  $U = \text{diag}(\hat{\Phi}(1)\hat{L})$  is necessary to guarantee that the normalization of  $\hat{B}$  (diagonal elements equal to one) is respected.  $\Omega$  is then estimated as  $\hat{\Omega} = \hat{U}\hat{D}\hat{U}$ .

Instead of using a method of moments approach, it is possible to use an instrumental variable (IV) approach. For this purpose we rewrite the reduced form of  $X_t$  in the Dickey-Fuller form (see equations (7.1) and (16.3)):

$$\Delta X_t = -\Phi(1)X_{t-1} + \tilde{\Phi}_1\Delta X_{t-1} + \dots + \tilde{\Phi}_{p-1}\Delta X_{t-p+1} + Z_t,$$

where  $\tilde{\Phi}_j = -\sum_{i=j+1}^p \Phi_i$ ,  $j = 1, 2, \dots, p-1$ . For the ease of exposition, we

<sup>22</sup>See Rubio-Ramírez et al. (2010) for a unified treatment of both type of restrictions.

assume  $B = I_n$  so that  $AZ_t = V_t$ . Multiplying this equation by  $A$  yields:

$$A\Delta X_t = -A\Phi(1)X_{t-1} + A\tilde{\Phi}_1\Delta X_{t-1} + \dots + A\tilde{\Phi}_{p-1}\Delta X_{t-p+1} + V_t. \quad (15.8)$$

Consider for simplicity the case that  $A\Phi(1)$  is a lower triangular matrix. This implies that the structural shocks  $V_{2t}, V_{3t}, \dots, V_{nt}$  have no long-run impact on the first variable  $X_{1t}$ . It is, therefore, possible to estimate the coefficients  $A_{12}, A_{13}, \dots, A_{1n}$  by instrumental variables taking  $X_{2,t-1}, X_{3,t-1}, \dots, X_{n,t-1}$  as instruments.

For  $n = 2$  the Dickey-Fuller form of the equation system (15.8) is:

$$\begin{pmatrix} 1 & A_{12} \\ A_{21} & 1 \end{pmatrix} \begin{pmatrix} \Delta\tilde{X}_{1t} \\ \Delta\tilde{X}_{2t} \end{pmatrix} = - \begin{pmatrix} [A\Phi(1)]_{11} & 0 \\ [A\Phi(1)]_{21} & [A\Phi(1)]_{22} \end{pmatrix} \begin{pmatrix} \tilde{X}_{1,t-1} \\ \tilde{X}_{2,t-1} \end{pmatrix} + \begin{pmatrix} V_{1t} \\ V_{2t} \end{pmatrix},$$

respectively

$$\begin{aligned} \Delta\tilde{X}_{1t} &= -A_{12}\Delta\tilde{X}_{2t} - [A\Phi(1)]_{11}\tilde{X}_{1,t-1} + V_{1t} \\ \Delta\tilde{X}_{2t} &= -A_{21}\Delta\tilde{X}_{1t} - [A\Phi(1)]_{21}\tilde{X}_{1,t-1} - [A\Phi(1)]_{22}\tilde{X}_{2,t-1} + V_{2t}. \end{aligned}$$

Thereby  $\Delta\tilde{X}_{1t}$  and  $\Delta\tilde{X}_{2t}$  denote the OLS residuals from a regression of  $\Delta X_{1t}$ , respectively  $\Delta X_{2t}$  on  $(\Delta X_{1,t-1}, \Delta X_{2,t-1}, \dots, \Delta X_{1,t-p+1}, \Delta X_{2,t-p+1})$ .  $\tilde{X}_{2,t-1}$  is a valid instrument for  $\Delta\tilde{X}_{2t}$  because this variable does not appear in the first equation. Thus,  $A_{12}$  can be consistently estimated by the IV-approach. For the estimation of  $A_{21}$ , we can use the residuals from the first equation as instruments because  $V_{1t}$  and  $V_{2t}$  are assumed to be uncorrelated. From this example, it is easy to see how this recursive method can be generalized to more than two variables. Note also that the IV-approach can also be used in the context of short-run restrictions.

The issue whether a technology shock leads to a reduction of hours worked in the short-run, led to a vivid discussion on the usefulness of long-run restrictions for structural models (Galí, 1999; Christiano et al., 2003, 2006; Chari et al., 2008). From an econometric point of view, it turned out, on the one hand, that the estimation of  $\Phi(1)$  is critical for the method of moments approach. The IV-approach, on the other hand, depends on the strength or weakness of the instrument used (Pagan and Robertson, 1998; Gospodinov, 2010).

It is, of course, possible to combine both short- and long-run restrictions simultaneously. An interesting application of both techniques was presented by Galí (1992). In doing so, one must be aware that both type of restrictions are consistent with each other and that counting the number of restrictions gives only a necessary condition.

**Example 3: Identifying Aggregate Demand and Supply Shocks**

In this example, we follow Blanchard and Quah (1989) and investigate the behavior of the growth rate of real GDP and the unemployment rate for the US over the period from the first quarter 1979 to the second quarter 2004 (102 observations). The AIC and the BIC suggest models of order two and one, respectively. Because some coefficients of  $\widehat{\Phi}_2$  are significant at the 10 percent level, we prefer to use the VAR(2) model which results in the following estimates:<sup>23</sup>

$$\begin{aligned}\widehat{\Phi}_1 &= \begin{pmatrix} 0.070 & -3.376 \\ -0.026 & 1.284 \end{pmatrix} \\ \widehat{\Phi}_2 &= \begin{pmatrix} 0.029 & 3.697 \\ -0.022 & -0.320 \end{pmatrix} \\ \widehat{\Sigma} &= \begin{pmatrix} 7.074 & -0.382 \\ -0.382 & 0.053 \end{pmatrix}.\end{aligned}$$

These results can be used to estimate  $\Phi(1) = I_2 - \Phi_1 - \Phi_2$  and consequently also  $\Psi(1) = \Phi(1)^{-1}$ :

$$\begin{aligned}\widehat{\Phi}(1) &= \begin{pmatrix} 0.901 & -0.321 \\ 0.048 & 0.036 \end{pmatrix} \\ \widehat{\Psi}(1) &= \widehat{\Phi}(1)^{-1} = \begin{pmatrix} 0.755 & 6.718 \\ -1.003 & 18.832 \end{pmatrix}.\end{aligned}$$

Assuming that  $Z_t = BV_t$  and following the argument in Section 15.5.1 that the demand shock has no long-run impact on the level of real GDP, we can retrieve an estimate for  $b_{21}$ :

$$\hat{b}_{21} = -[\widehat{\Psi}(1)]_{11}/[\widehat{\Psi}(1)]_{12} = -0.112.$$

The solution of the quadratic equation for  $b_{12}$  are -8.894 and 43.285. As the first solution results in a negative variance for  $\omega_2^2$ , we can disregard this solution and stick to the second one. The second solution makes also sense economically, because a positive supply shock leads to positive effects on GDP. Setting  $b_{12} = 43.285$  gives the following estimates for covariance matrix of the structural shocks  $\Omega$ :

$$\widehat{\Omega} = \begin{pmatrix} \widehat{\omega}_d^2 & 0 \\ 0 & \widehat{\omega}_s^2 \end{pmatrix} = \begin{pmatrix} 4.023 & 0 \\ 0 & 0.0016 \end{pmatrix}.$$

<sup>23</sup>The results for the constants are suppressed to save space.

The big difference in the variance of both shocks clearly shows the greater importance of demand shocks for business cycle movements.

Figure 15.4 shows the impulse response functions of the VAR(2) identified by the long-run restriction. Each figure displays the dynamic effect of a demand and a supply shock on real GDP and the unemployment rate, respectively, where the size of the initial shock corresponds to one standard deviation. The result conforms well with standard economic reasoning. A positive demand shock increases real GDP and lowers the unemployment rate in the short-run. The effect is even amplified for some quarters before it declines monotonically. After 30 quarters the effect of the demand has practically vanished so that its long-run effect becomes zero as imposed by the restriction. The supply shock has a similar short-run effect on real GDP, but initially increases the unemployment rate. Only when the effect on GDP becomes stronger after some quarters will the unemployment rate start to decline. In the long-run, the supply shock has a positive effect on real GDP but no effect on unemployment. Interestingly, only the short-run effects of the demand shock are statistically significant at the 95-percent level.

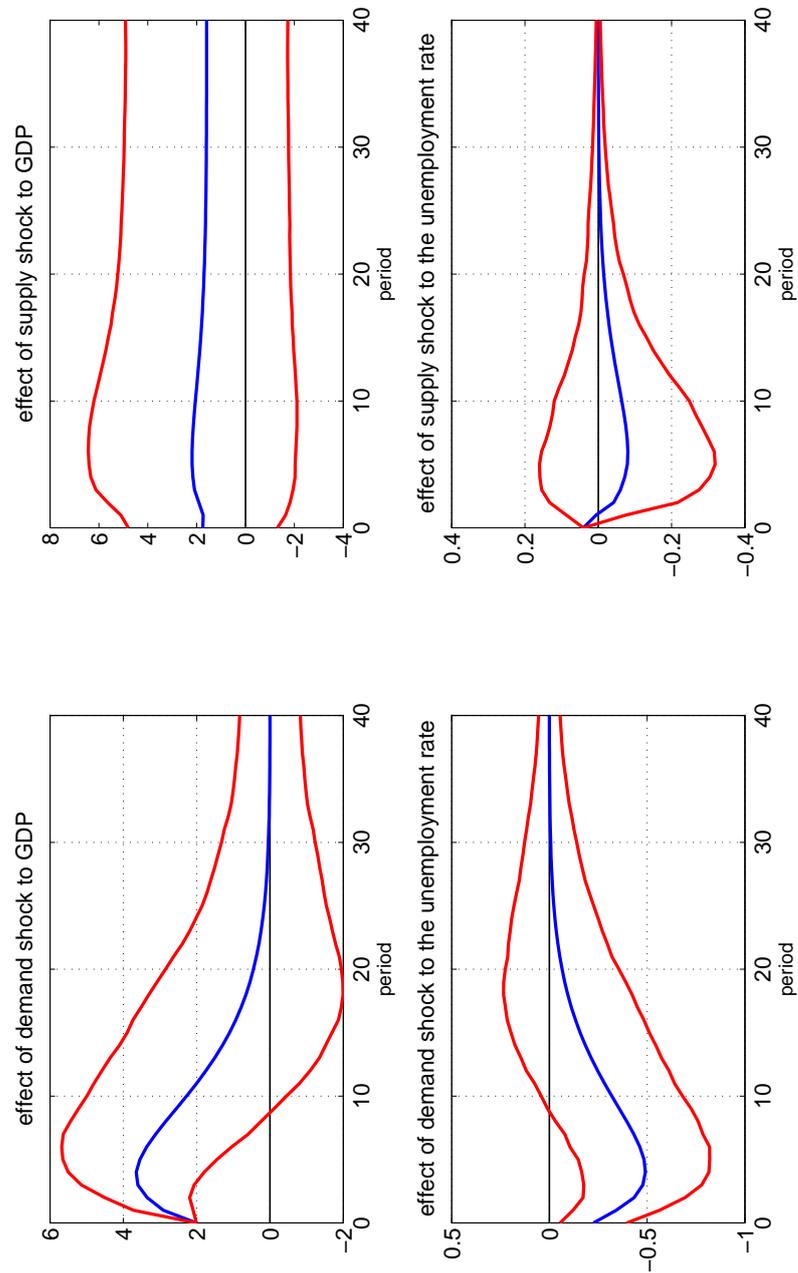


Figure 15.4: Impulse response functions of the Blanchard-Quah model (Blanchard and Quah, 1989) with 95-percent confidence intervals computed using the bootstrap procedure

# Chapter 16

## Cointegration

As already mentioned in Chapter 7, many raw economic time series are non-stationary and become stationary only after some transformation. The most common of these transformations is the formation of differences, perhaps after having taken logs. In most cases first differences are sufficient to achieve stationarity. The stationarized series can then be analyzed in the context of VAR models as explained in the previous chapters. However, many economic theories are formalized in terms of the original series so that we may want to use the VAR methodology to infer also the behavior of the untransformed series. Yet, by taking first differences we lose probably important information on the levels. Thus, it seems worthwhile to develop an approach which allows us to take the information on the levels into account and at the same time take care of the nonstationary character of the variables. The concept of *cointegration* tries to achieve this double requirement.

In the following we will focus our analysis on variables which are integrated of order one, i.e. on time series which become stationary after having taken first differences. However, as we have already mentioned in Section 7.5.1, a regression between integrated variables may lead to spurious correlations which make statistical inferences and interpretations of the estimated coefficients a delicate issue (see Section 7.5.3). A way out of this dilemma is presented by the theory of *cointegrated processes*. Loosely speaking, a multivariate process is cointegrated if there exists a linear combination of the processes which is stationary although each process taken individually may be integrated. In many cases, this linear combination can be directly related to economic theory which has made the analysis of cointegrated processes an important research topic. Although the bivariate case has already been dealt with in Section 7.5.1, we will analyze the issue of cointegration more in depth in this chapter.

The concept of cointegration goes back to the work of Engle and Granger

(1987) which is itself based on the precursor study of Davidson et al. (1978). In the meantime the literature has grown tremendously. Good introductions can be found in Banerjee et al. (1993), Watson (1994) or Lütkepohl (2006). For the more statistically inclined reader Johansen (1995) is a good reference.

## 16.1 A Theoretical Example

Before we present the general theory of cointegration within the VAR context, it is instructive to introduce the concept in the well-known class of present discounted value models. These models relate some variable  $X_t$  to present discounted value of another variable  $Y_t$ :

$$X_t = \gamma(1 - \beta) \sum_{j=0}^{\infty} \beta^j \mathbb{P}_t Y_{t+j} + u_t, \quad 0 < \beta < 1,$$

where  $u_t \sim \text{WN}(0, \sigma_u^2)$  designates a preference shock. Thereby,  $\beta$  denotes the subjective discount factor and  $\gamma$  is some unspecified parameter. The present discounted value model states that the variable  $X_t$  is proportional to the sum of future  $Y_{t+j}$ ,  $j = 0, 1, 2, \dots$ , discounted by the factor  $\beta$ . We can interpret  $X_t$  and  $Y_t$  as the price and the dividend of a share, as the interest rate on long- and short-term bonds, or as consumption and income. In order to operationalize this model, we will assume that forecasts are computed as linear mean-squared error forecasts. The corresponding forecast operator is denoted by  $\mathbb{P}_t$ . Furthermore, we will assume that the forecaster observes  $Y_t$  and its past  $Y_{t-1}, Y_{t-2}, \dots$ . The goal of the analysis is the investigation of the properties of the bivariate process  $\{(X_t, Y_t)'\}$ . The analysis of this important class models presented below is based on Campbell and Shiller (1987).<sup>1</sup>

The model is closed by assuming some specific time series model for  $\{Y_t\}$ . In this example, we will assume that  $\{Y_t\}$  is an integrated process of order one (see Definition 7.1 in Section 7.1) such that  $\{\Delta Y_t\}$  follows an AR(1) process:

$$\Delta Y_t = \mu(1 - \phi) + \phi \Delta Y_{t-1} + v_t, \quad |\phi| < 1 \text{ and } v_t \sim \text{WN}(0, \sigma_v^2).$$

This specification of the  $\{Y_t\}$  process implies that  $\mathbb{P}_t \Delta Y_{t+h} = \mu(1 - \phi^h) + \phi^h \Delta Y_t$ . Because  $\mathbb{P}_t Y_{t+h} = \mathbb{P}_t \Delta Y_{t+h} + \dots + \mathbb{P}_t \Delta Y_{t+1} + Y_t$ ,  $h = 0, 1, 2, \dots$ , the present discounted value model can be manipulated to give:

$$X_t = \gamma(1 - \beta) [Y_t + \beta \mathbb{P}_t Y_{t+1} + \beta^2 \mathbb{P}_t Y_{t+2} + \dots] + u_t$$

<sup>1</sup>A more recent interesting application of this model is given by the work of Beaudry and Portier (2006).

$$\begin{aligned}
&= \gamma(1 - \beta) [ Y_t \\
&\quad + \beta Y_t + \beta \mathbb{P}_t \Delta Y_{t+1} \\
&\quad + \beta^2 Y_t + \beta^2 \mathbb{P}_t \Delta Y_{t+1} + \beta^2 \mathbb{P}_t \Delta Y_{t+2} \\
&\quad + \beta^3 Y_t + \beta^3 \mathbb{P}_t \Delta Y_{t+1} + \beta^3 \mathbb{P}_t \Delta Y_{t+2} + \beta^3 \mathbb{P}_t \Delta Y_{t+3} \\
&\quad + \dots ] + u_t \\
&= \gamma(1 - \beta) \left[ \frac{1}{1 - \beta} Y_t + \frac{\beta}{1 - \beta} \mathbb{P}_t \Delta Y_{t+1} + \frac{\beta^2}{1 - \beta} \mathbb{P}_t \Delta Y_{t+2} + \dots \right] + u_t
\end{aligned}$$

This expression shows that the integratedness of  $\{Y_t\}$  is transferred to  $\{X_t\}$ . Bringing  $Y_t$  to the left we get the following expression:

$$S_t = X_t - \gamma Y_t = \gamma \sum_{j=1}^{\infty} \beta^j \mathbb{P}_t \Delta Y_{t+j} + u_t.$$

The variable  $S_t$  is occasionally referred to as the spread. If  $\gamma$  is greater than zero, expected increases in  $\Delta Y_{t+j}$ ,  $j \geq 1$ , have a positive impact on the spread today. For  $\gamma = 1$ ,  $S_t$  can denote the log of the price-dividend ratio of a share, or minus the logged savings ratio as in the permanent income model of Campbell (1987). If investors expect positive (negative) change in the dividends tomorrow, they want to buy (sell) the share thereby increasing (decreasing) its price already today. In the context of the permanent income hypothesis expected positive income changes lead to a reduction in today's saving rate. If, on the contrary, households expect negative income changes to occur in the future, they will save already today ("saving for the rainy days").

Inserting for  $\mathbb{P}_t \Delta Y_{t+j}$ ,  $j = 0, 1, \dots$ , the corresponding forecast equation  $\mathbb{P}_t \Delta Y_{t+h} = \mu(1 - \phi^h) + \phi^h \Delta Y_t$ , we get:

$$S_t = \frac{\beta\gamma\mu(1 - \phi)}{(1 - \beta)(1 - \beta\phi)} + \frac{\beta\gamma\phi}{1 - \beta\phi} \Delta Y_t + u_t.$$

The remarkable feature about this relation is that  $\{S_t\}$  is a stationary process because both  $\{\Delta Y_t\}$  and  $\{u_t\}$  are stationary, despite the fact that  $\{Y_t\}$  and  $\{X_t\}$  are both an integrated processes of order one. The mean of  $S_t$  is:

$$\mathbb{E}S_t = \frac{\beta\gamma\mu}{1 - \beta}.$$

From the relation between  $S_t$  and  $\Delta Y_t$  and the AR(1) representation of  $\{\Delta Y_t\}$

we can deduce a VAR representation of the joint process  $\{(S_t, \Delta Y_t)'\}$ :

$$\begin{pmatrix} S_t \\ \Delta Y_t \end{pmatrix} = \mu(1 - \phi) \begin{pmatrix} \frac{\beta\gamma}{(1-\beta)(1-\beta\phi)} + \frac{\beta\gamma\phi}{1-\beta\phi} \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & \frac{\beta\gamma\phi^2}{1-\beta\phi} \\ 0 & \phi \end{pmatrix} \begin{pmatrix} S_{t-1} \\ \Delta Y_{t-1} \end{pmatrix} \\ + \begin{pmatrix} u_t + \frac{\beta\gamma\phi}{1-\beta\phi}v_t \\ v_t \end{pmatrix}.$$

Further algebraic transformation lead to a VAR representation of order two for the level variables  $\{(X_t, Y_t)'\}$ :

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = c + \Phi_1 \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \Phi_2 \begin{pmatrix} X_{t-2} \\ Y_{t-2} \end{pmatrix} + Z_t \\ = \mu(1 - \phi) \begin{pmatrix} \frac{\gamma}{(1-\beta)(1-\beta\phi)} \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & \gamma + \frac{\gamma\phi}{1-\beta\phi} \\ 0 & 1 + \phi \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} \\ + \begin{pmatrix} 0 & \frac{-\gamma\phi}{1-\beta\phi} \\ 0 & -\phi \end{pmatrix} \begin{pmatrix} X_{t-2} \\ Y_{t-2} \end{pmatrix} + \begin{pmatrix} 1 & \frac{\gamma}{1-\beta\phi} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_t \\ v_t \end{pmatrix}$$

Next we want to check whether this stochastic difference equation possesses a stationary solution. For this purpose, we must locate the roots of the equation  $\det \Phi(z) = \det(I_2 - \Phi_1 z - \Phi_2 z^2) = 0$  (see Theorem 12.1). As

$$\det \Phi(z) = \det \begin{pmatrix} 1 & \left(-\gamma - \frac{\gamma\phi}{1-\beta\phi}\right)z + \frac{\gamma\phi}{1-\beta\phi}z^2 \\ 0 & 1 - (1 + \phi)z + \phi z^2 \end{pmatrix} = 1 - (1 + \phi)z + \phi z^2,$$

the roots are  $z_1 = 1/\phi$  and  $z_2 = 1$ . Thus, only the root  $z_1$  lies outside the unit circle whereas the root  $z_2$  lies on the unit circle. The existence of a *unit root* precludes the existence of a stationary solution. Note that we have just one unit root, although each of the two processes taken by themselves are integrated of order one.

The above VAR representation can be further transformed to yield a representation of process in first differences  $\{(\Delta X_t, \Delta Y_t)'\}$ :

$$\begin{pmatrix} \Delta X_t \\ \Delta Y_t \end{pmatrix} = \mu(1 - \phi) \begin{pmatrix} \frac{\gamma}{(1-\beta)(1-\beta\phi)} \\ 1 \end{pmatrix} - \begin{pmatrix} 1 & -\gamma \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} \\ + \begin{pmatrix} 0 & \frac{\gamma\phi}{1-\beta\phi} \\ 0 & \phi \end{pmatrix} \begin{pmatrix} \Delta X_{t-1} \\ \Delta Y_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & \frac{\gamma}{1-\beta\phi} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_t \\ v_t \end{pmatrix}.$$

This representation can be considered as a generalization of the Dickey-Fuller regression in first difference form (see equation (7.1)). In the multivariate

case, it is known as the vector error correction model (VECM) or vector error correction representation. In this representation the matrix

$$\Pi = -\Phi(1) = \begin{pmatrix} -1 & \gamma \\ 0 & 0 \end{pmatrix}$$

is of special importance. This matrix is singular and of rank one. This is not an implication which is special to this specification of the present discounted value model, but arises generally as shown in Campbell (1987) and Campbell and Shiller (1987). In the VECM representation all variables except  $(X_{t-1}, Y_{t-1})'$  are stationary by construction. This implies that  $-\Pi(X_{t-1}, Y_{t-1})'$  must be stationary too, despite the fact that  $\{(X_t, Y_t)'\}$  is not stationary as shown above. Multiplying  $-\Pi(X_{t-1}, Y_{t-1})'$  out, one obtains two linear combinations which define stationary processes. However, as  $\Pi$  has only rank one, there is just one linearly independent combination. The first one is  $X_{t-1} - \gamma Y_{t-1}$  and equals  $S_{t-1}$  which was already shown to be stationary. The second one is degenerate because it yields zero. The phenomenon is called *cointegration*.

Because  $\Pi$  has rank one, it can be written as the product of two vectors  $\alpha$  and  $\beta$ :

$$\Pi = \alpha\beta' = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ -\gamma \end{pmatrix}'.$$

Clearly, this decomposition of  $\Pi$  is not unique because  $\tilde{\alpha} = a\alpha$  and  $\tilde{\beta} = a^{-1}\beta$ ,  $a \neq 0$ , would also qualify for such a decomposition as  $\Pi = \tilde{\alpha}\tilde{\beta}'$ . The vector  $\beta$  is called a *cointegration vector*. It has the property that  $\{\beta'(X_t, Y_t)'\}$  defines a stationary process despite the fact that  $\{(X_t, Y_t)'\}$  is non-stationary. The cointegration vector thus defines a linear combination of  $X_t$  and  $Y_t$  which is stationary. The matrix  $\alpha$ , here only a vector, is called the *loading matrix* and its elements the *loading coefficients*.

The VAR and the VECM representations are both well suited for estimation. However, if we want to compute the impulse responses, we need a causal representation. Such a causal representation does not exist due to the unit root in the VAR process for  $\{(X_t, Y_t)'\}$  (see Theorem 12.1). To circumvent this problem we split the matrix  $\Phi(z)$  into the product of two matrices  $M(z)$  and  $V(z)$ .  $M(z)$  is a diagonal matrix which encompasses all unit roots on its diagonal.  $V(z)$  has all its roots outside the unit circle so that  $V^{-1}(z)$  exists for  $|z| < 1$ . In our example, we get:

$$\begin{aligned} \Phi(z) &= M(z)V(z) \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1-z \end{pmatrix} \begin{pmatrix} 1 & \left(-\gamma - \frac{\gamma\phi}{1-\beta\phi}\right)z + \frac{\gamma\phi}{1-\beta\phi}z^2 \\ 0 & 1-\phi z \end{pmatrix}. \end{aligned}$$

Multiplying  $\Phi(z)$  with  $\widetilde{M}(z) = \begin{pmatrix} 1-z & 0 \\ 0 & 1 \end{pmatrix}$  from the left, we find:

$$\widetilde{M}(z)\Phi(z) = \widetilde{M}(z)M(z)V(z) = (1-z)I_2V(z) = (1-z)V(z).$$

The application of this result to the VAR representation of  $\{(X_t, Y_t)\}$  leads to a causal representation of  $\{(\Delta X_t, \Delta Y_t)\}$ :

$$\begin{aligned} \Phi(L) \begin{pmatrix} X_t \\ Y_t \end{pmatrix} &= M(L)V(L) \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = c + Z_t \\ \widetilde{M}(L)\Phi(L) \begin{pmatrix} X_t \\ Y_t \end{pmatrix} &= (1-L)V(L) \begin{pmatrix} X_t \\ Y_t \end{pmatrix} \\ &= \mu(1-\phi) \begin{pmatrix} 1-L & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\gamma}{(1-\beta)(1-\beta\phi)} \\ 1 \end{pmatrix} + \begin{pmatrix} 1-L & 0 \\ 0 & 1 \end{pmatrix} Z_t \\ V(L) \begin{pmatrix} \Delta X_t \\ \Delta Y_t \end{pmatrix} &= \begin{pmatrix} 0 \\ \mu(1-\phi) \end{pmatrix} + \begin{pmatrix} 1-L & 0 \\ 0 & 1 \end{pmatrix} Z_t \\ \begin{pmatrix} \Delta X_t \\ \Delta Y_t \end{pmatrix} &= \mu \begin{pmatrix} \gamma \\ 1 \end{pmatrix} + V^{-1}(L) \begin{pmatrix} 1-L & 0 \\ 0 & 1 \end{pmatrix} Z_t \\ \begin{pmatrix} \Delta X_t \\ \Delta Y_t \end{pmatrix} &= \mu \begin{pmatrix} \gamma \\ 1 \end{pmatrix} + \Psi(L)Z_t. \end{aligned}$$

The polynomial matrix  $\Psi(L)$  can be recovered by the method of undetermined coefficients from the relation between  $V(L)$  and  $\Psi(L)$ :

$$V(L)\Psi(L) = \begin{pmatrix} 1-L & 0 \\ 0 & 1 \end{pmatrix}$$

In this exposition, we abstain from the explicit computation of  $V^{-1}(L)$  and  $\Psi(L)$ . However, the following relation holds:

$$V(1) = \begin{pmatrix} 1 & -\gamma \\ 0 & 1-\phi \end{pmatrix} \implies V^{-1}(1) = \begin{pmatrix} 1 & \frac{\gamma}{1-\phi} \\ 0 & \frac{1}{1-\phi} \end{pmatrix}.$$

Implying that

$$\Psi(1) = V^{-1}(1) \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = (1-\phi)^{-1} \begin{pmatrix} 0 & \gamma \\ 0 & 1 \end{pmatrix}.$$

The cointegration vector  $\beta = (1, -\gamma)'$  and loading matrix  $\alpha = (-1, 0)'$  therefore have the following properties:

$$\beta'\Psi(1) = (0 \ 0) \quad \text{and} \quad \Psi(1)\alpha = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Like in the univariate case (see Theorem 7.1 in Section 7.1.4), we can also construct the Beveridge-Nelson decomposition in the multivariate case. For this purpose, we decompose  $\Psi(L)$  as follows:

$$\Psi(L) = \Psi(1) + (L - 1)\tilde{\Psi}(L)$$

with  $\tilde{\Psi}_j = \sum_{i=j+1}^{\infty} \Psi_i$ . This result can be used to derive the multivariate Beveridge-Nelson decomposition (see Theorem 16.1 in Section 16.2.3):

$$\begin{aligned} \begin{pmatrix} X_t \\ Y_t \end{pmatrix} &= \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} + \mu \begin{pmatrix} \gamma \\ 1 \end{pmatrix} t + \Psi(1) \sum_{j=1}^t Z_j + \text{stationary process} \\ &= \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} + \mu \begin{pmatrix} \gamma \\ 1 \end{pmatrix} t + \frac{1}{1-\phi} \begin{pmatrix} 0 & \gamma \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{\gamma}{1-\beta\phi} \\ 0 & 1 \end{pmatrix} \sum_{j=1}^t \begin{pmatrix} u_j \\ v_j \end{pmatrix} \\ &\quad + \text{stationary process} \\ &= \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} + \mu \begin{pmatrix} \gamma \\ 1 \end{pmatrix} t + \frac{1}{1-\phi} \begin{pmatrix} 0 & \gamma \\ 0 & 1 \end{pmatrix} \sum_{j=1}^t \begin{pmatrix} u_j \\ v_j \end{pmatrix} \\ &\quad + \text{stationary process.} \end{aligned} \tag{16.1}$$

The Beveridge-Nelson decomposition represents the integrated process  $\{(X_t, Y_t)'\}$  as a sum of three components: a linear trend, a multivariate random walk and a stationary process. Multiplying the Beveridge-Nelson decomposition from the left by the cointegration vector  $\beta = (1, -\gamma)'$ , we see that both the trend and the random walk component are eliminated and that only the stationary component remains.

Because the first column of  $\Psi(1)$  consists of zeros, only the second structural shock, namely  $\{v_t\}$ , will have a long-run (permanent) effect. The long-run effect is  $\gamma/(1-\phi)$  for the first variable,  $X_t$ , and  $1/(1-\phi)$  for the second variable,  $Y_t$ . The first structural shock (preference shock)  $\{u_t\}$  has non long-run effect, its impact is of a transitory nature only. This decomposition into permanent and transitory shocks is not typical for this model, but can be done in general as part of the so-called common trend representation (see Section 16.2.4).

Finally, we will simulate the reaction of the system to a unit valued shock in  $v_t$ . Although this shock only has a temporary influence on  $\Delta Y_t$ , it will have a permanent effect on the level  $Y_t$ . Taking  $\phi = 0.8$ , we get long-run effect (persistence) of  $1/(1-\phi) = 5$  as explained in Section 7.1.3. The present discounted value model then implies that this shock will also have a permanent effect on  $X_t$  too. Setting  $\gamma = 1$ , this long-run effect is given by  $\gamma(1-\beta) \sum_{j=0}^{\infty} \beta^j (1-\phi)^{-1} = \gamma/(1-\phi) = 5$ . Because this long-run effect

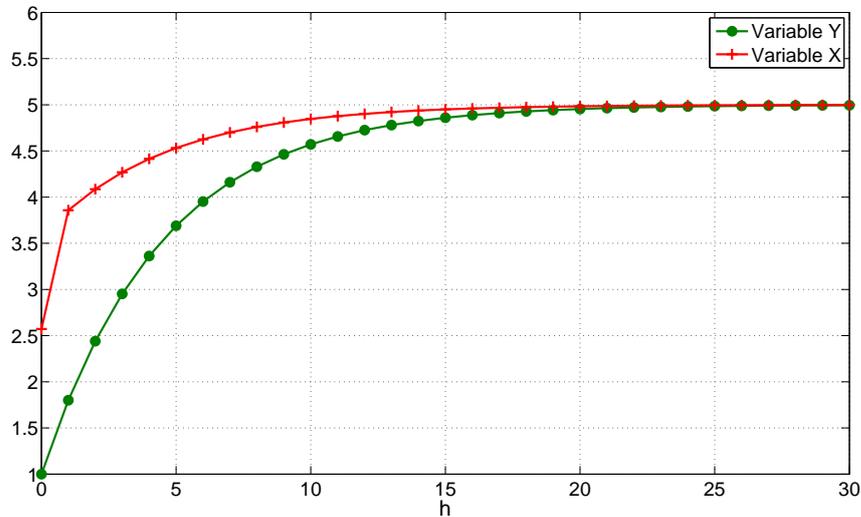


Figure 16.1: Impulse response functions of the present discounted value model after a unit shock to  $Y_t$  ( $\gamma = 1, \beta = 0.9, \phi = 0.8$ )

is anticipated in period  $t$ , the period of the occurrence of the shock,  $X_t$  will increase by more than one. The spread turns, therefore, into positive. The error correction mechanism will then dampen the effect on future changes of  $X_t$  so that the spread will return steadily to zero. The corresponding impulse responses of both variables are displayed in Figure 16.1.

Figure 16.2 displays the trajectories of both variables after a stochastic simulation where both shocks  $\{u_t\}$  and  $\{v_t\}$  are drawn from a standard normal distribution. One can clearly discern the non-stationary character of both series. However, as it is typically for cointegrated series, they move more or less in parallel to each other. This parallel movement is ensured by the error correction mechanism. The difference between both series which is equal to the spread under this parameter constellation is mean reverting around zero.

## 16.2 Definition and Representation of Cointegrated Processes

### 16.2.1 Definition

We now want to make the concepts introduced earlier more precise and give a general definition of cointegrated processes and derive the different repre-

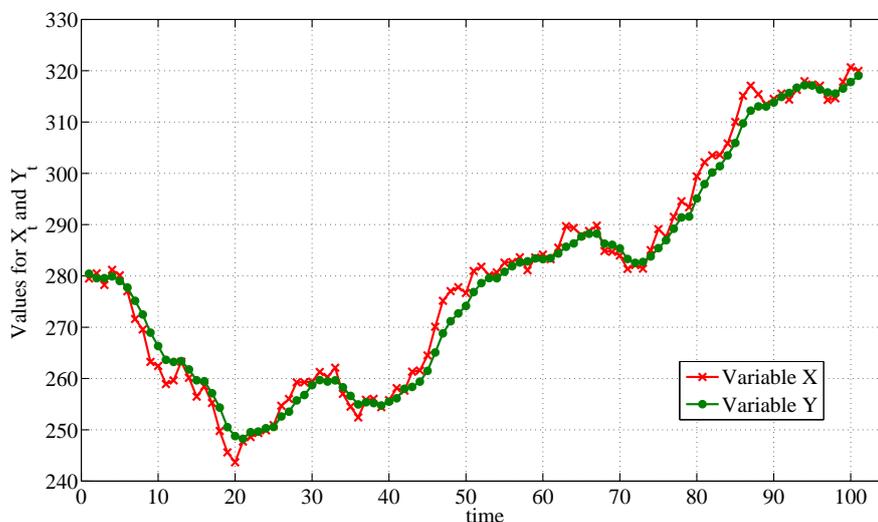


Figure 16.2: Stochastic simulation of the present discounted value model under standard normally distributed shocks ( $\gamma = 1, \beta = 0.9, \phi = 0.8$ )

sentations we have seen in the previous section. Given an arbitrary regular (purely non-deterministic) stationary process  $\{U_t\}_{t \in \mathbb{Z}}$  of dimension  $n$ ,  $n \geq 1$ , with mean zero and some distribution for the starting random variable  $X_0$ , we can define recursively a process  $\{X_t\}$ ,  $t = 0, 1, 2, \dots$  as follows:

$$X_t = \mu + X_{t-1} + U_t, \quad t = 1, 2, \dots$$

Thereby,  $\mu$  denotes an arbitrary constant vector of dimension  $n$ . If  $U_t \sim \text{WN}(0, \Sigma)$ , then  $\{X_t\}$  is a multivariate random walk with drift  $\mu$ . In general, however,  $\{U_t\}$  is autocorrelated and possesses a Wold representation  $U_t = \Psi(L)Z_t$  (see Section 14.1.1) such that

$$\Delta X_t = \mu + U_t = \mu + \Psi(L)Z_t = \mu + Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \dots, \quad (16.2)$$

where  $Z_t \sim \text{WN}(0, \Sigma)$  and  $\sum_{j=0}^{\infty} \|\Psi_j\|^2 < \infty$  with  $\Psi_0 = I_n$ . We now introduce the following definitions.

**Definition 16.1.** A regular stationary process  $\{U_t\}$  with mean zero is integrated of order zero,  $I(0)$ , if and only if it can be represented as

$$U_t = \Psi(L)Z_t = Z_t + \Psi_1 Z_{t-1} + \Psi_2 Z_{t-2} + \dots$$

such that  $Z_t \sim \text{WN}(0, \Sigma)$ ,  $\sum_{j=0}^{\infty} j \|\Psi_j\| < \infty$ , and  $\Psi(1) = \sum_{j=0}^{\infty} \Psi_j \neq 0$ .

**Definition 16.2.** A stochastic process  $\{X_t\}$  is integrated of order  $d$ ,  $I(d)$ ,  $d = 0, 1, 2, \dots$ , if and only if  $\Delta^d(X_t - \mathbb{E}(X_t))$  is integrated of order zero.

In the following we concentrate on  $I(1)$  processes. The definition of an  $I(1)$  process implies that  $\{X_t\}$  equals  $X_t = X_0 + \mu t + \sum_{j=1}^t U_j$  and is thus non-stationary even if  $\mu = 0$ . The condition  $\Psi(1) \neq 0$  corresponds to the one in the univariate case (compare Definition 7.1 in Section 7.1). On the one hand, it precludes the case that a trend-stationary process is classified as an integrated process. On the other hand, it implies that  $\{X_t\}$  is in fact non-stationary. Indeed, if the condition is violated so that  $\Psi(1) = 0$ , we could express  $\Psi(L)$  as  $(1-L)\tilde{\Psi}(L)$ . Thus we could cancel  $1-L$  on both sides of equation (16.2) to obtain a stationary representation of  $\{X_t\}$ , given some initial distribution for  $X_0$ . This would then contradict our primal assumption that  $\{X_t\}$  is non-stationary. The condition  $\sum_{j=0}^{\infty} j\|\Psi_j\| < \infty$  is stronger than  $\sum_{j=0}^{\infty} \|\Psi_j\|^2 < \infty$  which follows from the Wold's Theorem. It guarantees the existence of the Beveridge-Nelson decomposition (see Theorem 16.1 below).<sup>2</sup> In particular, the condition is fulfilled if  $\{U_t\}$  is a causal ARMA process which is the prototypical case.

Like in the univariate case, we can decompose an  $I(1)$  process additively into several components.

**Theorem 16.1** (Beveridge-Nelson decomposition). *If  $\{X_t\}$  is an integrated process of order one, it can be decomposed as*

$$X_t = X_0 + \mu t + \Psi(1) \sum_{j=1}^t Z_j + V_t,$$

where  $V_t = \tilde{\Psi}(L)Z_0 - \tilde{\Psi}(L)Z_t$  with  $\tilde{\Psi}_j = \sum_{i=j+1}^{\infty} \Psi_i$ ,  $j = 0, 1, 2, \dots$  and  $\{V_t\}$  stationary.

*Proof.* Following the proof of the univariate case (see Section 7.1.4):

$$\Psi(L) = \Psi(1) + (L-1)\tilde{\Psi}(L)$$

---

<sup>2</sup>This condition could be relaxed and replaced by the condition  $\sum_{j=0}^{\infty} j^2\|\Psi_j\|^2 < \infty$ . In addition, this condition is an important assumption for the application of the law of large numbers and for the derivation of the asymptotic distribution (Phillips and Solo, 1992).

with  $\tilde{\Psi}_j = \sum_{i=j+1}^{\infty} \Psi_i$ . Thus,

$$\begin{aligned} X_t &= X_0 + \mu t + \sum_{j=1}^t U_j = X_0 + \mu t + \sum_{j=1}^t \Psi(L)Z_j \\ &= X_0 + \mu t + \sum_{j=1}^t \left( \Psi(1) + (L-1)\tilde{\Psi}(L) \right) Z_j \\ &= X_0 + \mu t + \Psi(1) \sum_{j=1}^t Z_j + \sum_{j=1}^t (L-1)\tilde{\Psi}(L)Z_j \\ &= X_0 + \mu t + \Psi(1) \sum_{j=1}^t Z_j + \tilde{\Psi}(L)Z_0 - \tilde{\Psi}(L)Z_t. \end{aligned}$$

The only point left is to show that  $\tilde{\Psi}(L)Z_0 - \tilde{\Psi}(L)Z_t$  is stationary. Based on Theorem 10.2, it is sufficient to show that the coefficient matrices are absolutely summable. This can be derived by applying the triangular inequality and the condition for integrated processes:

$$\sum_{j=0}^{\infty} \|\tilde{\Psi}_j\| = \sum_{j=0}^{\infty} \left\| \sum_{i=j+1}^{\infty} \Psi_i \right\| \leq \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} \|\Psi_i\| = \sum_{j=1}^{\infty} j \|\Psi_j\| < \infty.$$

□

The process  $\{X_t\}$  can therefore be viewed as the sum of a linear trend,  $X_0 + \mu t$ , with stochastic intercept, a multivariate random walk,  $\Psi(1) \sum_{j=0}^t Z_t$ , and a stationary process  $\{V_t\}$ . Based on this representation, we can then define the notion of cointegration (Engle and Granger, 1987).

**Definition 16.3** (Cointegration). *A multivariate stochastic process  $\{X_t\}$  is called cointegrated if  $\{X_t\}$  is integrated of order one and if there exists a vector  $\beta \in \mathbb{R}^n$ ,  $\beta \neq 0$ , such that  $\{\beta'X_t\}$ , is integrated of order zero, given a corresponding distribution for the initial random variable  $X_0$ .  $\beta$  is called the cointegrating or cointegration vector. The cointegrating rank is the maximal number,  $r$ , of linearly independent cointegrating vectors  $\beta_1, \dots, \beta_r$ . These vectors span a linear space called the cointegration space.*

The Beveridge-Nelson decomposition implies that  $\beta$  is a cointegrating vector if and only if  $\beta'\Psi(1) = 0$ . In this case the random walk component  $\sum_{j=1}^t Z_j$  is annihilated and only the deterministic and the stationary component remain.<sup>3</sup> For some issues it is of interest whether the cointegration

<sup>3</sup>The distribution of  $X_0$  is thereby chosen such that  $\beta'X_0 = \beta'\tilde{\Psi}(L)Z_0$ .

vector  $\beta$  also eliminates the trend component. This would be the case if  $\beta'\mu = 0$ .

The cointegration vectors are determined only up to some basis transformations. If  $\beta_1, \dots, \beta_r$  is a basis for the cointegration space then  $(\beta_1, \dots, \beta_r)R$  is also a basis for the cointegration space for any nonsingular  $r \times r$  matrix  $R$  because  $((\beta_1, \dots, \beta_r)R)'\Psi(1) = 0$ .

### 16.2.2 Vector Autoregressive (VAR) and Vector Error Correction Models (VECM)

Although the Beveridge-Nelson decomposition is very useful from a theoretical point of view, in practice it is often more convenient to work with alternative representations. Most empirical investigations of integrated processes start from a VAR(p) model which has the big advantage that it can be easily estimated:

$$X_t = c + \Phi_1 X_{t-1} + \dots + \Phi_p X_{t-p} + Z_t, \quad Z_t \sim \text{WN}(0, \Sigma)$$

where  $\Phi(L) = I_n - \Phi_1 L - \dots - \Phi_p L^p$  and  $c$  is an arbitrary constant. Subtracting  $X_{t-1}$  on both sides of the difference equation, the VAR model can be rewritten as:

$$\Delta X_t = c + \Pi X_{t-1} + \Gamma_1 \Delta X_{t-1} + \dots + \Gamma_{p-1} \Delta X_{t-p+1} + Z_t \quad (16.3)$$

where  $\Pi = -\Phi(1) = -I_n + \Phi_1 + \dots + \Phi_p$  and  $\Gamma_i = -\sum_{j=i+1}^p \Phi_j$ . We will make the following assumptions:

- (i) All roots of the polynomial  $\det \Phi(z)$  are outside the unit circle or equal to one, i.e.

$$\det \Phi(z) = 0 \implies \begin{cases} |z| > 1 & \text{or} \\ z = 1, \end{cases}$$

- (ii) The matrix  $\Pi$  is singular with rank  $r$ ,  $1 \leq r < n$ .

- (iii)  $\text{Rank}(\Pi) = \text{Rank}(\Pi^2)$ .

Assumption (i) makes sure that  $\{X_t\}$  is an integrated process with order of integration  $d \geq 1$ . Moreover, it precludes other roots on the unit circles than one. The case of seasonal unit roots is treated in Hylleberg et al. (1990) and Johansen and Schaumburg (1998).<sup>4</sup> Assumption (ii) implies that there exists

<sup>4</sup>The seasonal unit roots are the roots of  $z^s - 1 = 0$  where  $s$  denotes the number of seasons. These roots can be expressed as  $\cos(2k\pi/s) + \iota \sin(2k\pi/s)$ ,  $k = 0, 1, \dots, s-1$ .

at least  $n - r$  unit roots and two  $n \times r$  matrices  $\alpha$  and  $\beta$  with full column rank  $r$  such that

$$\Pi = \alpha\beta'.$$

The columns of  $\beta$  thereby represent the cointegration vectors whereas  $\alpha$  denotes the so-called loading matrix. The decomposition of  $\Pi$  in the product of  $\alpha$  and  $\beta'$  is not unique. For every non-singular  $r \times r$  matrix  $R$  we can generate an alternative decomposition  $\Pi = \alpha\beta' = (\alpha R'^{-1})(\beta R)'$ . Finally, assumption (iii) implies that the order of integration is exactly one and not greater. The number of unit roots is therefore exactly  $n - r$ .<sup>5</sup> This has the implication that  $\Phi(z)$  can be written as

$$\Phi(z) = U(z)M(z)V(z)$$

where the roots of the matrix polynomials  $U(z)$  and  $V(z)$  are all outside the unit circle and where  $M(z)$  equals

$$M(z) = \begin{pmatrix} (1-z)I_{n-r} & 0 \\ 0 & I_r \end{pmatrix}.$$

This representation of  $\Phi(z)$  is a special form of the Smith-McMillan factorization of polynomial matrices (see Kailath (1980) and Yoo (1987)). This factorization isolates the unit roots in one simple matrix so that the system can be analyzed more easily.

These assumptions will allow us to derive from the VAR(p) model several representations where each of them brings with it a particular interpretation. Replacing  $\Pi$  by  $\alpha\beta'$  in equation (16.3), we obtain the *vector error correction representation* or *vector error correction model* (VECM):

$$\Delta X_t = c + \alpha\beta'X_{t-1} + \Gamma_1\Delta X_{t-1} + \dots + \Gamma_{p-1}\Delta X_{t-p+1} + Z_t. \quad (16.4)$$

Multiplying both sides of the equation by  $(\alpha'\alpha)^{-1}\alpha'$  and solving for  $\beta'X_{t-1}$ , we get:

$$\beta'X_{t-1} = (\alpha'\alpha)^{-1}\alpha' \left( \Delta X_t - c - \sum_{j=1}^{p-1} \Gamma_j \Delta X_{t-j} - Z_t \right).$$

$\alpha$  has full column rank  $r$  so that  $\alpha'\alpha$  is a non-singular  $r \times r$  matrix. As the right hand side of the equation represents a stationary process, also the left hand side must be stationary. This means that the  $r$ -dimensional process  $\{\beta'X_{t-1}\}$  is stationary despite the fact that  $\{X_t\}$  is integrated and has potentially a unit root with multiplicity  $n$ .

<sup>5</sup>For details see Johansen (1995), Neusser (2000) and Bauer and Wagner (2003).

The term error correction was coined by Davidson et al. (1978). They interpret the mean of  $\beta' X_t$ ,  $\mu^* = \mathbb{E}\beta' X_t$ , as the long-run equilibrium or steady state around which the system fluctuates. The deviation from equilibrium (error) is therefore given by  $\beta' X_{t-1} - \mu^*$ . The coefficients of the loading matrix  $\alpha$  should then guarantee that deviations from the equilibrium are corrected over time by appropriate changes (corrections) in  $X_t$ .

### An illustration

To illustrate the concept of the error correction model, we consider the following simple system with  $\alpha = (\alpha_1, \alpha_2)'$ ,  $\alpha_1 \neq \alpha_2$ , and  $\beta = (1, -1)'$ . For simplicity, we assume that the long-run equilibrium  $\mu^*$  is zero. Ignoring higher order lags, we consider the system:

$$\begin{aligned}\Delta X_{1t} &= \alpha_1(X_{1,t-1} - X_{2,t-1}) + Z_{1t} \\ \Delta X_{2t} &= \alpha_2(X_{1,t-1} - X_{2,t-1}) + Z_{2t}.\end{aligned}$$

The autoregressive polynomial of this system is:

$$\Phi(z) = \begin{pmatrix} 1 - (1 + \alpha_1)z & \alpha_1 z \\ -\alpha_2 z & 1 - (1 - \alpha_2)z \end{pmatrix}.$$

The determinant of this polynomial is  $\det \Phi(z) = 1 - (2 + \alpha_1 - \alpha_2)z + (1 + \alpha_1 - \alpha_2)z^2$  with roots equal to  $z = 1$  and  $z = 1/(1 + \alpha_1 - \alpha_2)$ . This shows that assumption (i) is fulfilled. As  $\Pi = \begin{pmatrix} \alpha_1 & -\alpha_1 \\ \alpha_2 & -\alpha_2 \end{pmatrix}$ , the rank of  $\Pi$  equals one which implies assumption (ii). Finally,

$$\Pi^2 = \begin{pmatrix} \alpha_1^2 - \alpha_1\alpha_2 & -\alpha_1^2 + \alpha_1\alpha_2 \\ -\alpha_2^2 + \alpha_1\alpha_2 & \alpha_2^2 - \alpha_1\alpha_2 \end{pmatrix}.$$

Thus, the rank of  $\Pi^2$  is also one because  $\alpha_1 \neq \alpha_2$ . Hence, assumption (iii) is also fulfilled.

We can gain an additional insight into the system by subtracting the second equation from the first one to obtain:

$$X_{1t} - X_{2t} = (1 + \alpha_1 - \alpha_2)(X_{1,t-1} - X_{2,t-1}) + Z_{1t} - Z_{2t}.$$

The process  $\beta' X_t = X_{1t} - X_{2t}$  is stationary and causal with respect to  $Z_{1t} - Z_{2t}$  if and only if  $|1 + \alpha_1 - \alpha_2| < 1$ , or equivalently if and only if  $-2 < \alpha_1 - \alpha_2 < 0$ . Note the importance of the assumption that  $\alpha_1 \neq \alpha_2$ . It prevents that  $X_{1t} - X_{2t}$  becomes a random walk and thus a non-stationary (integrated) process. A sufficient condition is that  $-1 < \alpha_1 < 0$  and  $0 < \alpha_2 < 1$ .

which imply that a positive (negative) error, i.e.  $X_{1,t-1} - X_{2,t-1} > 0 (< 0)$ , is corrected by a negative (positive) change in  $X_{1t}$  and a positive (negative) change in  $X_{2t}$ . Although the shocks  $Z_{1t}$  and  $Z_{2t}$  push  $X_{1t} - X_{2t}$  time and again away from its long-run equilibrium, the error correction mechanism ensures that the variables are adjusted in such a way that the system moves back to its long-run equilibrium.

### 16.2.3 The Beveridge-Nelson Decomposition

We next want to derive from the VAR representation a causal representation or MA( $\infty$ ) representation for  $\{\Delta X_t\}$ . In contrast to a normal causal VAR model, the presence of unit roots precludes the simple application of the method of undetermined coefficients, but requires an additional effort. Multiplying the VAR representation in equation (16.7),  $\Phi(L)X_t = U(L)M(L)V(L)X_t = c + Z_t$ , from the left by  $U^{-1}(L)$  we obtain:

$$M(L)V(L)X_t = U^{-1}(1)c + U^{-1}(L)Z_t.$$

Multiplying this equation by  $\widetilde{M}(L) = \begin{pmatrix} I_{n-r} & 0 \\ 0 & (1-L)I_r \end{pmatrix}$  leads to:

$$V(L)\Delta X_t = \widetilde{M}(1)U^{-1}(1)c + \widetilde{M}(L)U^{-1}(L)Z_t$$

which finally leads to

$$\begin{aligned} \Delta X_t &= V^{-1}(1)\widetilde{M}(1)U^{-1}(1)c + V^{-1}(L)\widetilde{M}(L)U^{-1}(L)Z_t \\ &= \mu + \Psi(L)Z_t. \end{aligned}$$

This is the MA( $\infty$ ) representation of  $\{\Delta X_t\}$  and corresponds to equation (16.2).

Because  $\Pi = -\Phi(1) = -U(1)M(1)V(1)$ , the following relation holds for the partitioned matrices:

$$\Phi(1) = \begin{pmatrix} U_{11}(1) & U_{12}(1) \\ U_{21}(1) & U_{22}(1) \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & I_r \end{pmatrix} \begin{pmatrix} V_{11}(1) & V_{12}(1) \\ V_{21}(1) & V_{22}(1) \end{pmatrix} = \begin{pmatrix} U_{12}(1) \\ U_{22}(1) \end{pmatrix} (V_{21}(1) \quad V_{22}(1)).$$

This implies that we can define  $\alpha$  and  $\beta$  as

$$\alpha = - \begin{pmatrix} U_{12}(1) \\ U_{22}(1) \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} V_{21}(1)' \\ V_{22}(1)' \end{pmatrix}.$$

$U(1)$  and  $V(1)$  are non-singular so that  $\alpha$  and  $\beta$  have full column rank  $r$ . Based on this derivation we can formulate the following lemma.

**Lemma 16.1.** *The columns of the so-defined matrix  $\beta$  are the cointegration vectors for the process  $\{X_t\}$ . The corresponding matrix of loading coefficients is  $\alpha$  which fulfills  $\Psi(1)\alpha = 0$ .*

*Proof.* We must show that  $\beta'\Psi(1) = 0$  which is the defining property of cointegration vectors. Denoting by  $(V^{(ij)}(1))_{i,j=1,2}$  the appropriately partitioned matrix of  $V(1)^{-1}$ , we obtain:

$$\begin{aligned}\beta'\Psi(1) &= \beta' \begin{pmatrix} V^{(11)}(1) & V^{(12)}(1) \\ V^{(21)}(1) & V^{(22)}(1) \end{pmatrix} \begin{pmatrix} I_{n-r} & 0 \\ 0 & 0 \end{pmatrix} U^{-1}(1) \\ &= (V_{21}(1) \quad V_{22}(1)) \begin{pmatrix} V^{(11)}(1) & 0 \\ V^{(21)}(1) & 0 \end{pmatrix} U^{-1}(1) \\ &= \begin{pmatrix} V_{21}(1)V^{(11)}(1) + V_{22}(1)V^{(21)}(1) & : 0 \end{pmatrix} U^{-1}(1) = 0_n\end{aligned}$$

where the last equality is a consequence of the property of the inverse matrix.

With the same arguments, we can show that  $\Psi(1)\alpha = 0$ .  $\square$

The equivalence between the VEC and the MA representation is known as Granger's representation theorem in the literature. Granger's representation theorem immediately implies the Beveridge-Nelson decomposition:

$$X_t = X_0 + \Psi(1)ct + \Psi(1) \sum_{j=1}^t Z_j + V_t \quad (16.5)$$

$$= X_0 + V^{-1}(1)\widetilde{M}(1)U^{-1}(1)ct + V^{-1}(1)\widetilde{M}(1)U^{-1}(1) \sum_{j=1}^t Z_j + V_t \quad (16.6)$$

where the stochastic process  $\{V_t\}$  is stationary and defined as  $V_t = \widetilde{\Psi}(L)Z_0 - \widetilde{\Psi}(L)Z_t$  with  $\widetilde{\Psi}_j = \sum_{i=j+1}^{\infty} \Psi_i$  and  $\Psi(L) = V^{-1}(L)\widetilde{M}(L)U^{-1}(L)$ . As  $\beta'\Psi(1) = \beta'V^{-1}(1)\widetilde{M}(1)U^{-1}(1) = 0$ ,  $\beta$  eliminates the stochastic trend (random walk),  $\sum_{j=1}^t Z_j$ , as well as the deterministic linear trend  $\mu t = V^{-1}(1)\widetilde{M}(1)U^{-1}(1)ct$ .

An interesting special case is obtained when the constant  $c$  is a linear combination of the columns of  $\alpha$ , i.e. if there exists a vector  $g$  such that  $c = \alpha g$ . Under this circumstance,  $\Psi(1)c = \Psi(1)\alpha g = 0$  and the linear trend vanishes and we have  $\mathbb{E}\Delta X_t = 0$ . In this case, the data will exhibit no trend although the VAR model contains a constant. A similar consideration can be made if the VAR model is specified to contain a constant and a linear time trend  $dt$ . The Beveridge-Nelson decomposition would then imply that the data should follow a quadratic trend. However, in the special case that  $d$  is a linear combination of the columns of  $\alpha$ , the quadratic trend disappears and only the linear remains because of the constant.

### 16.2.4 Common Trend and Triangular Representation

The  $\Psi(1)$  in the Beveridge-Nelson decomposition is singular. This implies that the multivariate random walk  $\Psi(1) \sum_{j=1}^{\infty} Z_j$  does not consist of  $n$  independent univariate random walks. Instead only  $n - r$  independent random walks make up the stochastic trend so that  $\{X_t\}$  is driven by  $n - r$  stochastic trends. In order to emphasize this fact, we derive from the Beveridge-Nelson decomposition the so-called *common trend representation* (Stock and Watson, 1988b).

As  $\Psi(1)$  has rank  $n - r$ , there exists a  $n \times r$  matrix  $\gamma$  such that  $\Psi(1)\gamma = 0$ . Denote by  $\gamma^\perp$  the  $n \times (n - r)$  matrix whose columns are orthogonal to  $\gamma$ , i.e.  $\gamma'\gamma^\perp = 0$ . The Beveridge-Nelson decomposition can then be rewritten as:

$$\begin{aligned} X_t &= X_0 + \Psi(1) \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix} \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix}^{-1} c t \\ &\quad + \Psi(1) \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix} \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix}^{-1} \sum_{j=1}^t Z_j + V_t \\ &= X_0 + \left( \Psi(1)\gamma^\perp & : & 0 \right) \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix}^{-1} c t \\ &\quad + \left( \Psi(1)\gamma^\perp & : & 0 \right) \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix}^{-1} \sum_{j=1}^t Z_j + V_t \\ &= X_0 + \left( \Psi(1)\gamma^\perp & : & 0 \right) \tilde{c} t + \left( \Psi(1)\gamma^\perp & : & 0 \right) \sum_{j=1}^t \tilde{Z}_j + V_t \end{aligned}$$

where  $\tilde{c} = \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix}^{-1} c$  and  $\tilde{Z}_j = \begin{pmatrix} \gamma^\perp & : & \gamma \end{pmatrix}^{-1} Z_j$ . Therefore, only the first  $n - r$  elements of the vector  $\tilde{c}$  are relevant for the deterministic linear trend. The remaining elements are multiplied by zero and are thus irrelevant. Similarly, for the multivariate random walk only the first  $n - r$  elements of the process  $\{\tilde{Z}_t\}$  are responsible for the stochastic trend. The remaining elements of  $\tilde{Z}_t$  are multiplied by zero and thus have no permanent, but only a transitory influence. The above representation decomposes the shocks orthogonally into *permanent* and *transitory* ones (Gonzalo and Ng, 2001). The previous lemma shows that one can choose for  $\gamma$  the matrix of loading coefficients  $\alpha$ .

Summarizing the first  $n - r$  elements of  $\tilde{c}$  and  $\tilde{Z}_t$  to  $\tilde{c}_1$  and  $\tilde{Z}_{1t}$ , respectively, we arrive at the *common trend representation*:

$$X_t = X_0 + B\tilde{c}_1 t + B \sum_{j=1}^t \tilde{Z}_{1j} + V_t$$

where the  $n \times (n - r)$  matrix  $B$  is equal to  $\Psi(1)\gamma^\perp$ .

Applying these results to our introductory example, we arrive at

$$B = \frac{1}{1 - \phi} \begin{pmatrix} \gamma \\ 1 \end{pmatrix}, \quad \tilde{c} = \mu(1 - \phi), \quad \text{and} \quad \tilde{Z}_{1t} = u_t.$$

This again demonstrates that the trend, the linear as well as the stochastic trend, are exclusively stemming from the nonstationary variables  $\{Y_t\}$  (compare with equation (16.1)).

Finally, we want to present a triangular representation which is well suited to deal with the nonparametric estimation approach advocated by Phillips (1991) and Phillips and Hansen (1990). In this representation we normalize the cointegration vector such  $\beta = (I_r, -b)'$ . In addition, we partition the vector  $X_t$  into  $X_{1t}$  and  $X_{2t}$  such that  $X_{1t}$  contains the first  $r$  and  $X_{2t}$  the last  $n - r$  elements.  $X_t = (X'_{1t}, X_{2t})'$  can then be expressed as:

$$\begin{aligned} X_{1t} &= b'X_{2t} + \pi_1 D_t + u_{1t} \\ \Delta X_{2t} &= \pi_2 \Delta D_t + u_{2t} \end{aligned}$$

where  $D_t$  summarizes the deterministic components such as constant and linear time trend.  $\{u_{1t}\}$  and  $\{u_{2t}\}$  denote potentially autocorrelated stationary time series.

### 16.3 Johansen's Test for Cointegration

In Section 7.5.1 we have already discussed a regression based test for cointegration among two variables. It was based on a unit root of the residuals from a bivariate regression of one variable against the other. In this regression, it turned out to be irrelevant which of the two variables was chosen as the regressor and which one as the regressand. This method can, in principle, be extended to more than two variables. However, with more than two variables, the choice of the regressand becomes more crucial as not all variables may be part of the cointegrating relation. Moreover, more than one independent cointegrating relation may exist. For these reasons, it is advantageous to use a method which treats all variables symmetrically. The *cointegration test* developed by Johansen fulfills this criterion because it is based on a VAR which does not single out a particular variable. This test has received wide recognition and is most often used in practice. The test serves two purposes. First, we want to determine the number  $r$  of cointegrating relationships. Second, we want to test properties of the cointegration vector  $\beta$  and the loading matrix  $\alpha$ .

The exposition of the Johansen test follows closely the work of Johansen where the derivations and additional details can be found (Johansen, 1988,

1991, 1995). We start with a VAR(p) model with constant  $c$  in VEC form (see equation (16.3)):

$$\Delta X_t = c + \Pi X_{t-1} + \Gamma_1 \Delta X_{t-1} + \dots + \Gamma_{p-1} \Delta X_{t-p+1} + Z_t, \quad t = 1, 2, \dots, T \quad (16.7)$$

where  $Z_t \sim \text{IIDN}(0, \Sigma)$  and given starting values  $X_0 = x_0, \dots, X_{-p+1} = x_{-p+1}$ . The problem can be simplified by regressing  $\Delta X_t$  as well as  $X_{t-1}$  against  $c, \Delta X_{t-1}, \dots, \Delta X_{t-p+1}$  and working with the residuals from these regressions.<sup>6</sup> This simplification results in a VAR model of order one. We therefore start our analysis without loss of generality with a VAR(1) model without constant term:

$$\Delta X_t = \Pi X_{t-1} + Z_t$$

where  $Z_t \sim \text{IIDN}(0, \Sigma)$ .<sup>7</sup>

The phenomenon of cointegration manifests itself in the singularity of the matrix  $\Pi$ . In particular, we want to determine the rank of  $\Pi$  which gives the number of linearly independent cointegrating relationships. Denoting by  $r$ , the rank of  $\Pi$ ,  $0 \leq r \leq n$ , we can formulate a sequence of hypotheses:

$$H(r) : \text{rank}(\Pi) \leq r, \quad r = 0, 1, \dots, n.$$

Hypothesis  $H(r)$ , thus, implies that there exists *at most*  $r$  linearly independent cointegrating vectors. The sequence of hypotheses is nested in the sense that  $H(r)$  implies  $H(r+1)$ :

$$H(0) \subseteq H(1) \subseteq \dots \subseteq H(n).$$

The hypothesis  $H(0)$  means that  $\text{rank}(\Pi) = 0$ . In this case,  $\Pi = 0$  and there are no cointegration vectors.  $\{X_t\}$  is thus driven by  $n$  independent random walks and the VAR can be transformed into a VAR model for  $\{\Delta X_t\}$  which in our simplified version just means that  $\Delta X_t = Z_t \sim \text{IIDN}(0, \Sigma)$ . The hypothesis  $H(n)$  places no restriction on  $\Pi$  and includes in this way the case that the level of  $\{X_t\}$  is already stationary. Of particular interest are the hypotheses between these two extreme ones where non-degenerate cointegrating vectors are possible. In the following, we not only want to test for the number of linearly independent cointegrating vectors,  $r$ , but we also want to test hypotheses about the structure of the cointegrating vectors summarized in  $\beta$ .

<sup>6</sup>If the VAR model (16.7) contains further deterministic components besides the constant, these components have to be accounted for in these regressions.

<sup>7</sup>This two-stage least-squares procedure is also known as partial regression and is part of the Frisch-Waugh-Lowell Theorem (Davidson and MacKinnon, 1993, 19-24).

Johansen's test is conceived as a likelihood-ratio test. This means that we must determine the likelihood function for a sample  $X_1, X_2, \dots, X_T$  where  $T$  denotes the sample size. For this purpose, we assume that  $\{Z_t\} \sim \text{IIDN}(0, \Sigma)$  so that logged likelihood function of the parameters  $\alpha$ ,  $\beta$ , and  $\Sigma$  conditional on the starting values is given by :

$$\begin{aligned} \ell(\alpha, \beta, \Sigma) = & -\frac{Tn}{2} \ln(2\pi) + \frac{T}{2} \ln \det(\Sigma^{-1}) \\ & - \frac{1}{2} \sum_{t=1}^T (\Delta X_t - \alpha\beta'X_{t-1})' \Sigma^{-1} (\Delta X_t - \alpha\beta'X_{t-1}) \end{aligned}$$

where  $\Pi = \alpha\beta'$ . For a fixed given  $\beta$ ,  $\alpha$  can be estimated by a regression of  $\Delta X_t$  on  $\beta'X_{t-1}$ :

$$\hat{\alpha} = \hat{\alpha}(\beta) = S_{01}\beta(\beta'S_{11}\beta)^{-1}$$

where the moment matrices  $S_{00}, S_{11}, S_{01}$  and  $S_{10}$  are defined as:

$$\begin{aligned} S_{00} &= \frac{1}{T} \sum_{t=1}^T (\Delta X_t)(\Delta X_t)' \\ S_{11} &= \frac{1}{T} \sum_{t=1}^T X_{t-1}X_{t-1}' \\ S_{01} &= \frac{1}{T} \sum_{t=1}^T (\Delta X_t)X_{t-1}' \\ S_{10} &= S_{01}' \end{aligned}$$

The covariance matrix of the residuals then becomes:

$$\hat{\Sigma} = \hat{\Sigma}(\beta) = S_{00} - S_{01}\beta(\beta'S_{11}\beta)^{-1}\beta'S_{10}.$$

Using these results, we can concentrate the log-likelihood function to obtain:

$$\begin{aligned} \ell(\beta) = \ell(\hat{\alpha}(\beta), \beta, \hat{\Sigma}(\beta)) &= -\frac{Tn}{2} \ln(2\pi) - \frac{T}{2} \ln \det(\hat{\Sigma}(\beta)) - \frac{Tn}{2} \\ &= -\frac{Tn}{2} \ln(2\pi) - \frac{Tn}{2} - \frac{T}{2} \ln \det (S_{00} - S_{01}\beta(\beta'S_{11}\beta)^{-1}\beta'S_{10}). \quad (16.8) \end{aligned}$$

The expression  $\frac{Tn}{2}$  in the above equation is derived as follows:

$$\begin{aligned} & \frac{1}{2} \sum_{t=1}^T (\Delta X_t - \hat{\alpha} \beta' X_{t-1})' \hat{\Sigma}^{-1} (\Delta X_t - \hat{\alpha} \beta' X_{t-1}) \\ &= \frac{1}{2} \text{tr} \left( \sum_{t=1}^T (\Delta X_t - \hat{\alpha} \beta' X_{t-1}) (\Delta X_t - \hat{\alpha} \beta' X_{t-1})' \hat{\Sigma}^{-1} \right) \\ &= \frac{1}{2} \text{tr} \left( (T S_{00} - T \hat{\alpha} \beta' S_{10} - T S_{01} \beta \hat{\alpha}' + T \hat{\alpha} \beta' S_{11} \beta \hat{\alpha}') \hat{\Sigma}^{-1} \right) \\ &= \frac{T}{2} \text{tr} \left( \underbrace{(S_{00} - \hat{\alpha} \beta' S_{10})}_{=\hat{\Sigma}} \hat{\Sigma}^{-1} \right) = \frac{Tn}{2}. \end{aligned}$$

The log-likelihood function is thus maximized if

$$\begin{aligned} \det(\hat{\Sigma}(\beta)) &= \det(S_{00} - S_{01} \beta (\beta' S_{11} \beta)^{-1} \beta' S_{10}) \\ &= \det S_{00} \frac{\det(\beta' (S_{11} - S_{10} S_{00}^{-1} S_{01}) \beta)}{\det(\beta' S_{11} \beta)} \end{aligned}$$

is minimized over  $\beta$ .<sup>8</sup> The minimum is obtained by solving the following generalized eigenvalue problem (Johansen, 1995):

$$\det(\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}) = 0.$$

This eigenvalue problem delivers  $n$  eigenvalues

$$1 \geq \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_n \geq 0$$

with corresponding  $n$  eigenvectors  $\hat{\beta}_1, \dots, \hat{\beta}_n$ . These eigenvectors are normalized such that  $\hat{\beta}' S_{11} \hat{\beta} = I_n$ . Therefore we have that

$$\text{argmin}_{\beta} \det(\hat{\Sigma}(\beta)) = (\det S_{00}) \prod_{i=1}^n \hat{\lambda}_i.$$

**Remark 16.1.** *In the case of cointegration,  $\Pi$  is singular with  $\text{rank} \hat{\Pi} = r$ . To estimate  $r$ , it seems natural to investigate the number of nonzero*

---

<sup>8</sup>Thereby we make use of the following equality for partitioned matrices:

$$\det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det A_{11} \det(A_{22} - A_{21} A_{11}^{-1} A_{12}) = \det A_{22} \det(A_{11} - A_{12} A_{22}^{-1} A_{21})$$

where  $A_{11}$  and  $A_{22}$  are invertible matrices (see for example Meyer, 2000, p. 475).

eigenvalues of  $\hat{\Pi} = S_{01}S_{11}^{-1}$ . However, because eigenvalues may be complex, it is advantageous not to investigate the eigenvalues of  $\hat{\Pi}$  but those of  $\hat{\Pi}'\hat{\Pi}$  which are all real and positive due to the symmetry of  $\hat{\Pi}'\hat{\Pi}$ . These eigenvalues are called the singular values of  $\hat{\Pi}$ .<sup>9</sup> Noting that

$$\begin{aligned} 0 &= \det(\lambda S_{11} - S_{10}S_{00}^{-1}S_{01}) = \det S_{11} \det(\lambda I_n - S_{11}^{-1/2}S_{10}S_{00}^{-1}S_{01}S_{11}^{-1/2}) \\ &= \det S_{11} \det(\lambda I_n - (S_{00}^{-1/2}S_{01}S_{11}^{-1/2})'(S_{00}^{-1/2}S_{01}S_{11}^{-1/2})), \end{aligned}$$

the generalized eigenvalue problem above therefore just determines the singular values of  $S_{00}^{-1/2}S_{01}S_{11}^{-1/2} = S_{00}^{-1/2}\hat{\Pi}S_{11}^{1/2}$ .

**Remark 16.2.** Based on the observation that, for  $n = 1$ ,  $\lambda = \frac{S_{01}S_{10}}{S_{11}S_{00}}$  equals the squared empirical correlation coefficient between  $\Delta X_t$  and  $X_{t-1}$ , we find that the eigenvalues  $\lambda_j$ ,  $j = 1, \dots, n$ , are nothing but the squared canonical correlation coefficients (see Johansen, 1995; Reinsel, 1993). Thereby, the largest eigenvalue,  $\hat{\lambda}_1$ , corresponds to the largest squared correlation coefficient that can be achieved between linear combinations of  $\Delta X_1$  and  $X_{t-1}$ . Thus  $\beta_1$  gives the linear combination of the integrated variable  $X_{t-1}$  which comes closest in the sense of correlation to the stationary variable  $\{\Delta X_t\}$ . The second eigenvalue  $\lambda_2$  corresponds to the maximal squared correlation coefficient between linear combinations of  $\Delta X_t$  and  $X_{t-1}$  which are orthogonal to the linear combination corresponding to  $\lambda_1$ . The remaining squared canonical correlation coefficients are obtained by iterating this procedure  $n$  times.

If the dimension of the cointegrating space is  $r$  then  $\hat{\beta}$  consists of those eigenvectors which correspond to the  $r$  largest eigenvalues  $\hat{\lambda}_1, \dots, \hat{\lambda}_r$ . The remaining eigenvalues  $\lambda_{r+1}, \dots, \lambda_n$  should be zero. Under the null hypothesis  $H(r)$ , the log-likelihood function (16.8) can be finally expressed as:

$$\ell(\hat{\beta}_r) = -\frac{Tn}{2} \ln \pi - \frac{Tn}{2} - \frac{T}{2} \ln \det S_{00} - \frac{T}{2} \sum_{i=1}^r \ln(1 - \lambda_i).$$

The expression for the optimized likelihood function can now be used to construct the Johansen likelihood-ratio test. There are two versions of the test depending on the alternative hypothesis:

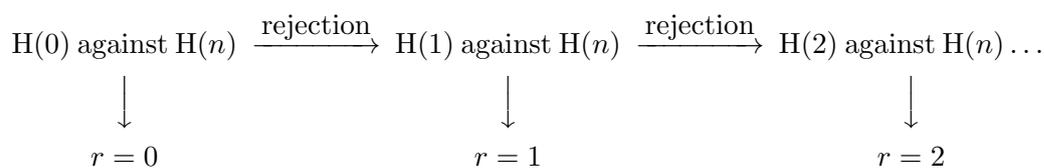
$$\begin{aligned} \text{trace test: } & H_0 : H(r) \text{ against } H(n), \\ \text{max test: } & H_0 : H(r) \text{ against } H(r+1). \end{aligned}$$

<sup>9</sup>An appraisal of the singular values of a matrix can be found in Strang (1988) or Meyer (2000).

The corresponding likelihood ratio test statistics are therefore:

$$\begin{aligned} \text{trace test: } 2(\ell(\hat{\beta}_n) - \ell(\hat{\beta}_r)) &= -T \sum_{j=r+1}^n \ln(1 - \hat{\lambda}_j) \approx T \sum_{j=r+1}^n \hat{\lambda}_j, \\ \text{max test: } 2(\ell(\hat{\beta}_{r+1}) - \ell(\hat{\beta}_r)) &= -T \ln(1 - \hat{\lambda}_{r+1}) \approx T \hat{\lambda}_{r+1}. \end{aligned}$$

In practice it is useful to adopt a sequential test strategy based on the trace test. Given some significance level, we test in a first step the null hypothesis  $H(0)$  against  $H(n)$ . If, on the one hand, the test does not reject the null hypothesis, we conclude that  $r = 0$  and that there is no cointegrating relation. If, on the other hand, the test rejects the null hypothesis, we conclude that there is at least one cointegrating relation. We then test in a second step the null hypothesis  $H(1)$  against  $H(n)$ . If the test does not reject the null hypothesis, we conclude that there exists one cointegrating relation, i.e. that  $r = 1$ . If the test rejects the null hypothesis, we examine the next hypothesis  $H(2)$ , and so on. In this way we obtain a test sequence. If in this sequence, the null hypothesis  $H(r)$  is not rejected, but  $H(r + 1)$  is, we conclude that exist  $r$  linearly independent cointegrating relations as explained in the diagram below.



If in this sequence we do not reject  $H(r)$  for some  $r$ , it is useful to perform the max test  $H(r)$  against  $H(r + 1)$  as a robustness check.

The asymptotic distributions of the test statistics are, like in the Dickey-Fuller unit root test, nonstandard and depend on the specification of the deterministic components. Based on the VAR model

$$\Delta X_t = c_0 + c_1 t + \Pi X_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta X_{t-j} + Z_t$$

with  $\Pi = \alpha\beta'$ , we can distinguish five different cases:<sup>10</sup>

**Case I:** The time series shows no trend and the cointegrating relation involves no constant. This corresponds to

$$c_0 = 0 \quad \text{and} \quad c_1 = 0.$$

<sup>10</sup>It is instructive to compare these cases to those of the unit root test (see Section 7.3.1).

**Case II:** The time series shows no trend, but the cointegrating relations involve some constants. This corresponds to

$$c_0 = \alpha\gamma_0 \quad \text{and} \quad c_1 = 0$$

with  $\gamma_0 \neq 0$ .

**Case III:** The time series shows a linear trend, but there are no constants in the cointegrating relations. This corresponds to

$$c_0 \neq 0 \quad \text{and} \quad c_1 = 0.$$

**Case IV:** The time series shows a linear trend and in the cointegrating relations involve both constants and linear trends. This corresponds to

$$c_0 \neq 0 \quad \text{and} \quad c_1 = \alpha\gamma_1$$

where  $\gamma_1 \neq 0$ .

**Case V:** The time series shows a quadratic trend and the cointegrating relation involves both constants and linear trends. This corresponds to

$$c_0 \neq 0 \quad \text{and} \quad c_1 \neq 0.$$

The corresponding asymptotic distributions of the trace as well as the max test statistic in these five cases are tabulated in Johansen (1995), MacKinnon et al. (1999), and Osterwald-Lenum (1992).<sup>11</sup> The finite sample properties of these tests can be quite poor. Thus, more recently, bootstrap methods have been proven to provide a successful alternative in practice (Cavaliere et al., 2012).

As mentioned previously, the cointegrating vectors are not unique, only the cointegrating space is. This makes the cointegrating vectors often difficult to interpret economically, despite some basis transformation. It is therefore of interest to see whether the space spanned by the cointegrating vectors summarized in the columns of  $\hat{\beta}$  can be viewed as a subspace spanned by some hypothetical vectors  $H = (h_1, \dots, h_s)$ ,  $r \leq s < n$ . If this hypothesis is true, the cointegrating vectors should be linear combinations of the columns of  $H$  so that the null hypothesis can be formulated as

$$H_0 : \beta = H\varphi \tag{16.9}$$

---

<sup>11</sup>The tables by MacKinnon et al. (1999) allow for the possibility of exogenous integrated variables.

for some  $s \times r$  matrix  $\varphi$ . Under this null hypothesis, this amounts to solve an analogous general eigenvalue problem:

$$\det(\varphi H' S_{11} H - H' S_{10} S_{00}^{-1} S_{01} H) = 0.$$

The solution of this problem is given by the eigenvalues  $1 > \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \tilde{\lambda}_s > 0$  with corresponding normalized eigenvectors. The likelihood ratio test statistic for this hypothesis is then

$$T \sum_{j=1}^r \ln \frac{1 - \tilde{\lambda}_j}{1 - \hat{\lambda}_j}.$$

This test statistic is asymptotically distributed as a  $\chi^2$  distribution with  $r(n - s)$  degrees of freedom.

With similar arguments it is possible to construct a test of the null hypothesis that the cointegrating space spanned by the columns of  $\hat{\beta}$  contains some hypothetical vectors  $K = (h_1, \dots, h_s)$ ,  $1 \leq s \leq r$ . The null hypothesis can then be formulated as

$$H_0 : K\varphi = \beta \quad (16.10)$$

for some  $s \times r$  matrix  $\varphi$ . Like in the previous case, this hypothesis can also be tested by the corresponding likelihood ratio test statistic which is asymptotically distributed as a  $\chi^2$  distribution with  $s(n - r)$  degrees of freedom. Similarly, it is possible to test hypotheses on  $\alpha$  and joint hypotheses on  $\alpha$  and  $\beta$  (see Johansen, 1995; Kunst and Neusser, 1990; Lütkepohl, 2006).

## 16.4 An Example

This example reproduces the study by Neusser (1991) with actualized data for the United States over the period first quarter 1950 to fourth quarter 2005. The starting point is a VAR model which consists of four variables: real gross domestic product (Y), real private consumption (C), real gross investment (I), and the ex-post real interest rate (R). All variables, except the real interest rate, are in logs. First, we identify a VAR model for these variables where the order is determined by Akaike's (AIC), Schwarz' (BIC) or Hannan-Quinn' (HQ) information criteria. The AIC suggests seven lags whereas the other criteria propose a VAR of order two. As the VAR(7) consists of many statistically insignificant coefficients, we prefer the more

parsimonious VAR(2) model which produces the following estimates:

$$X_t = \begin{pmatrix} Y_t \\ C_t \\ I_t \\ R_t \end{pmatrix} = \begin{pmatrix} 0.185 \\ (0.047) \\ 0.069 \\ (0.043) \\ 0.041 \\ (0.117) \\ -0.329 \\ (0.097) \end{pmatrix} + \begin{pmatrix} 0.951 & 0.254 & 0.088 & 0.042 \\ (0.086) & (0.091) & (0.033) & (0.032) \\ 0.157 & 0.746 & 0.065 & -0.013 \\ (0.079) & (0.084) & (0.031) & (0.030) \\ 0.283 & 0.250 & 1.304 & 0.026 \\ (0.216) & (0.229) & (0.084) & (0.081) \\ 0.324 & -0.536 & -0.024 & 0.551 \\ (0.178) & (0.189) & (0.069) & (0.067) \end{pmatrix} X_{t-1} + \begin{pmatrix} -0.132 & -0.085 & -0.089 & -0.016 \\ (0.085) & (0.093) & (0.033) & (0.031) \\ -0.213 & 0.305 & -0.066 & 0.112 \\ (0.078) & (0.085) & (0.031) & (0.029) \\ -0.517 & 0.040 & -0.364 & 0.098 \\ (0.214) & (0.233) & (0.084) & (0.079) \\ -0.042 & 0.296 & 0.005 & 0.163 \\ (0.176) & (0.192) & (0.069) & (0.065) \end{pmatrix} X_{t-2} + Z_t$$

where the estimated standard errors of the corresponding coefficients are reported in parenthesis. The estimate covariance matrix  $\hat{\Sigma}$ , is

$$\hat{\Sigma} = 10^{-4} \begin{pmatrix} 0.722 & 0.428 & 1.140 & 0.002 \\ 0.428 & 0.610 & 1.026 & -0.092 \\ 1.140 & 1.026 & 4.473 & -0.328 \\ 0.002 & -0.092 & -0.328 & 3.098 \end{pmatrix}.$$

The sequence of the hypotheses starts with  $H(0)$  which states that there exists no cointegrating relation. The alternative hypothesis is always  $H(n)$  which says that there are  $n$  cointegrating relations. According to Table 16.1 the value of the trace test statistic is 111.772 which is clearly larger than the 5 percent critical value of 47.856. Thus, the null hypothesis  $H(0)$  is rejected and we consider next the hypothesis  $H(1)$ . This hypothesis is again clearly rejected so that we move on to the hypothesis  $H(2)$ . Because  $H(3)$  is not rejected, we conclude that there exists 3 cointegrating relations. To check this result, we test the hypothesis  $H(2)$  against  $H(3)$  using the max test. As

Table 16.1: Evaluation of the results of Johansen's cointegration test

null hypothesis	eigenvalue	trace statistic		max statistic	
		value of test statistic	critical value	value of test statistic	critical value
H(0) : $r = 0$	0.190	111.772	47.856	47.194	27.584
H(1) : $r \leq 1$	0.179	64.578	29.797	44.075	21.132
H(2) : $r \leq 2$	0.081	20.503	15.495	18.983	14.265
H(3) : $r \leq 3$	0.007	1.520	3.841	1.520	3.841

critical 5 percent values are taken from MacKinnon et al. (1999).

this test also rejects H(2), we can be pretty confident that there are three cointegrating relations given as:

$$\hat{\beta} = \begin{pmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \\ -258.948 & -277.869 & -337.481 \end{pmatrix}.$$

In this form, the cointegrating vectors are economically difficult to interpret. We therefore ask whether they are compatible with the following hypotheses:

$$\beta_C = \begin{pmatrix} 1.0 \\ -1.0 \\ 0.0 \\ 0.0 \end{pmatrix}, \quad \beta_I = \begin{pmatrix} 1.0 \\ 0.0 \\ -1.0 \\ 0.0 \end{pmatrix}, \quad \beta_R = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 1.0 \end{pmatrix}.$$

These hypotheses state that the log-difference (ratio) between consumption and GDP, the log-difference (ratio) between investment and GDP, and the real interest rate are stationary. They can be rationalized in the context of the neoclassical growth model (see King et al., 1991; Neusser, 1991). Each of them can be brought into the form of equation (16.10) where  $\beta$  is replaced by its estimate  $\hat{\beta}$ . The corresponding test statistics for each of the three cointegrating relations is distributed as a  $\chi^2$  distribution with one degree of freedom,<sup>12</sup> which gives a critical value of 3.84 at the 5 percent significance level. The corresponding values for the test statistic are 12.69, 15.05 and 0.45,

<sup>12</sup>The degrees of freedom are computed according to the formula:  $s(n-r) = 1(4-3) = 1$ .

respectively. This implies that we must reject the first two hypotheses  $\beta_C$  and  $\beta_I$ . However, the conjecture that the real interest is stationary, cannot be rejected. Finally, we can investigate the joint hypothesis  $\beta_0 = (\beta_C, \beta_I, \beta_R)$  which can be represented in the form (16.9). In this case the value of the test statistic is 41.20 which is clearly above the critical value of 7.81 inferred from the  $\chi^2_3$  distribution.<sup>13</sup> Thus, we must reject this joint hypothesis.

---

<sup>13</sup>The degrees of freedom are computed according to the formula:  $r(n-s) = 3(4-3) = 3$ .

# Chapter 17

## State-Space Models and the Kalman Filter

The state space representation is a flexible technique originally developed in automatic control engineering to represent, model, and control dynamic systems. Thereby we summarize the unobserved or partially observed state of the system in period  $t$  by an  $m$ -dimensional vector  $X_t$ . The evolution of the state is then described by a VAR of order one usually called the state equation. A second equation describes the connection between the state and the observations given by a  $n$ -dimensional vector  $Y_t$ . Despite its simple structure, state space models encompass a large variety of model classes: VARMA, respectively VARIMA models,<sup>1</sup> unobserved-component models, factor models, structural time series models which decompose a given time series into a trend, a seasonal, and a cyclical component, models with measurement errors, etc.

From a technical point of view, the main advantage of state space modeling is the unified treatment of estimation, forecasting, and smoothing. At the center of the analysis stands the Kalman-filter named after its inventor Rudolf Emil Kálmán (Kalman, 1960, 1963). He developed a projection based algorithm which recursively produces a statistically optimal estimate of the state. The versatility and the ease of implementation have made the Kalman filter an increasingly popular tool also in the economically oriented times series analysis. Here we present just an introduction to the subject and refer to Anderson and Moore (1979), Brockwell and Davis (1991, Chapter 12), Brockwell and Davis (1996, Chapter 8), Hamilton (1994a, Chapter 13), Hamilton (1994b), Hannan and Deistler (1988) or Harvey (1989), for further details.

---

<sup>1</sup>VARIMA models stand for vector autoregressive integrated moving-average models.

## 17.1 The State Space Representation

The state space representation or model consists of a *state equation* which describes the dynamics of the system and an *observation equation* which connects the observations to the state. The state may be unobservable or partly observable and the relation between observations and state may be subject to measurement errors. In the case of time invariant coefficients<sup>2</sup> we can set up these two equations as follows:

$$\text{state equation:} \quad X_{t+1} = FX_t + V_{t+1}, \quad t = 1, 2, \dots \quad (17.1)$$

$$\text{observation equation:} \quad Y_t = A + GX_t + W_t, \quad t = 1, 2, \dots \quad (17.2)$$

Thereby  $X_t$  denotes an  $m$ -dimensional vector which describes the state of the system in period  $t$ . The evolution of the state is represented as a vector autoregressive model of order one with coefficient matrix  $F$  and disturbances  $V_{t+1}$ .<sup>3</sup> As we assume that the state  $X_t$  is unobservable or at least partly unobservable, we need a second equation which relates the state to the observations. In particular, we assume that there is a linear time-invariant relation given by  $A$  and  $G$  of the  $n$ -dimensional vector of observations,  $Y_t$ , to the state  $X_t$ . This relation may be contaminated by measurement errors  $W_t$ . The system is initialized in period  $t = 1$ .

We make the following simplifying assumption of the state space model represented by equations (17.1) and (17.2).

- (i)  $\{V_t\} \sim \text{WN}(0, Q)$  where  $Q$  is a constant nonnegative definite  $m \times m$  matrix.
- (ii)  $\{W_t\} \sim \text{WN}(0, R)$  where  $R$  is a constant nonnegative definite  $n \times n$  matrix.
- (iii) The two disturbances are uncorrelated with each other at all leads and lags, i.e.:

$$\mathbb{E}(W_s V_t') = 0, \quad \text{for all } t \text{ and } s.$$

- (iv)  $V_t$  and  $W_t$  are multivariate normally distributed.
- (v)  $X_1$  is uncorrelated with  $V_t$  as well as with  $W_t$ ,  $t = 1, 2, \dots$

<sup>2</sup>For the ease of exposition, we will present first the time-invariant case and analyze the case of time-varying coefficients later.

<sup>3</sup>In control theory the state equation (17.1) is amended by an additional term  $HU_t$  which represents the effect of control variables  $U_t$ . These exogenous controls are used to regulate the system.

**Remark 17.1.** *In a more general context, we can make both covariance matrices  $Q$  and  $R$  time-varying and allow for contemporaneous correlations between  $V_t$  and  $W_t$  (see example 17.4.1).*

**Remark 17.2.** *As both the state and the observation equation may include identities, the covariance matrices need not be positive definite. They can be non-negative definite.*

**Remark 17.3.** *Neither  $\{X_t\}$  nor  $\{Y_t\}$  are assumed to be stationary.*

**Remark 17.4.** *The specification of the state equation and the normality assumption imply that the sequence  $\{X_1, V_1, V_2, \dots\}$  is independent so that the conditional distribution  $X_{t+1}$  given  $X_t, X_{t-1}, \dots, X_1$  equals the conditional distribution of  $X_{t+1}$  given  $X_t$ . Thus, the process  $\{X_t\}$  satisfies the Markov property. As the dimension of the state vector  $X_t$  is arbitrary, it can be expanded in such a way as to encompass every component  $X_{t-1}$  for any  $t$  (see, for example, the state space representation of a VAR( $p$ ) model with  $p > 1$ ). However, there remains the problem of the smallest dimension of the state vector (see Section 17.3.2).*

**Remark 17.5.** *The state space representation is not unique. Defining, for example, a new state vector  $\tilde{X}_t$  by multiplying  $X_t$  with an invertible matrix  $P$ , i.e.  $\tilde{X}_t = PX_t$ , all properties of the system remain unchanged. Naturally, we must redefine all the system matrices accordingly:  $\tilde{F} = PFP^{-1}$ ,  $\tilde{Q} = PQP'$ ,  $\tilde{G} = GP^{-1}$ .*

Given  $X_1$ , we can iterate the state equation forward to arrive at:

$$X_t = F^{t-1}X_1 + \sum_{j=1}^{t-1} F^{j-1}V_{t+1-j}, \quad t = 1, 2, \dots$$

$$Y_t = A + GF^{t-1}X_1 + \sum_{j=1}^{t-1} GF^{j-1}V_{t+1-j} + W_t, \quad t = 1, 2, \dots$$

The state equation is called stable or causal if all eigenvalues of  $F$  are inside the unit circle which is equivalent that all roots of  $\det(I_m - Fz) = 0$  are outside the unit circle (see Section 12.3). In this case the state equation has a unique stationary solution:

$$X_t = \sum_{j=0}^{\infty} F^{j-1}V_{t+1-j}. \quad (17.3)$$

The process  $\{Y_t\}$  is therefore also stationary and we have:

$$Y_t = A + \sum_{j=0}^{\infty} GF^{j-1}V_{t+1-j} + W_t. \quad (17.4)$$

In the case of a stationary state space model, we may do without an initialization period and take  $t \in \mathbb{Z}$ .

In the case of a stable state equation, we can easily deduce the covariance function for  $\{X_t\}$ ,  $\Gamma_X(h)$ ,  $h = 0, 1, 2, \dots$ . According to Section 12.4 it holds that:

$$\begin{aligned} \Gamma_X(0) &= F\Gamma_X(0)F' + Q, \\ \Gamma_X(h) &= F^h\Gamma_X(0), \quad h = 1, 2, \dots \end{aligned}$$

where  $\Gamma_X(0)$  is uniquely determined given the stability assumption. Similarly, we can derive the covariance function for the observation vector,  $\Gamma_Y(h)$ ,  $h = 0, 1, 2, \dots$ :

$$\begin{aligned} \Gamma_Y(0) &= G\Gamma_X(0)G' + R, \\ \Gamma_Y(h) &= GF^h\Gamma_X(0)G', \quad h = 1, 2, \dots \end{aligned}$$

### 17.1.1 Examples

The following examples should illustrate the versatility of the state space model and demonstrate how many economically relevant models can be represented in this form.

#### VAR(p) process

Suppose that  $\{Y_t\}$  follows a  $n$ -dimensional VAR(p) process given by  $\Phi(L)Y_t = Z_t$ , respectively by  $Y_t = \Phi_1 Y_{t-1} + \dots + \Phi_p Y_{t-p} + Z_t$ , with  $Z_t \sim \text{WN}(0, \Sigma)$ . Then the companion form of the VAR(p) process (see Section 12.2) just represents the state equation (17.1):

$$\begin{aligned} X_{t+1} &= \begin{pmatrix} Y_{t+1} \\ Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \dots & 0 & 0 \\ 0 & I_n & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_n & 0 \end{pmatrix} \begin{pmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{pmatrix} + \begin{pmatrix} Z_{t+1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= FX_t + V_{t+1}, \end{aligned}$$

with  $V_{t+1} = (Z'_{t+1}, 0, 0, \dots, 0)'$  and  $Q = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$ . The observation equation is just an identity because all components of  $X_t$  are observable:

$$Y_t = (I_n, 0, 0, \dots, 0)X_t = GX_t.$$

Thus,  $G = (I_n, 0, 0, \dots, 0)$  and  $R = 0$ . Assuming that  $X_t$  is already mean adjusted,  $A = 0$ .

### ARMA(1,1) process

The representation of ARMA processes as a state space model is more involved when moving-average terms are involved. Suppose that  $\{Y_t\}$  is an ARMA(1,1) process defined by the stochastic difference equation  $Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}$  with  $Z_t \sim \text{WN}(0, \sigma^2)$  and  $\phi\theta \neq 0$ .

Let  $\{X_t\}$  be the AR(1) process defined by the stochastic difference equation  $X_t - \phi X_{t-1} = Z_t$  and define the state vector  $\mathbf{X}_t$  by  $\mathbf{X}_t = (X_t, X_{t-1})'$ , then we can write the observation equation as:

$$Y_t = (1, \theta)\mathbf{X}_t = G\mathbf{X}_t$$

with  $R = 0$ . The state equation is then

$$\mathbf{X}_{t+1} = \begin{pmatrix} X_{t+1} \\ X_t \end{pmatrix} = \begin{pmatrix} \phi & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X_t \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} Z_{t+1} \\ 0 \end{pmatrix} = F\mathbf{X}_t + V_{t+1},$$

where  $Q = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix}$ . It is easy to verify that the so defined process  $\{Y_t\}$  satisfies the stochastic difference equation  $Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}$ . Indeed  $Y_t - \phi Y_{t-1} = (1, \theta)\mathbf{X}_t - \phi(1, \theta)\mathbf{X}_{t-1} = X_t + \theta X_{t-1} - \phi X_{t-1} - \theta \phi X_{t-2} = (X_t - \phi X_{t-1}) + \theta(X_{t-1} - \phi X_{t-2}) = Z_t + \theta Z_{t-1}$ .

If  $|\phi| < 1$ , the state equation defines a causal process  $\{X_t\}$  so that the unique stationary solution is given by equation (17.3). This implies a stationary solution for  $\{Y_t\}$  too. It is thus easy to verify if this solution equals the unique solution of the ARMA stochastic difference equation.

The state space representation of an ARMA model is not unique. An alternative representation in the case of a causal system is given by:

$$\begin{aligned} X_{t+1} &= \phi X_t + (\phi + \theta)Z_t = FX_t + V_{t+1} \\ Y_t &= X_t + Z_t = X_t + W_t. \end{aligned}$$

Note that in this representation the dimension of the state vector is reduced from two to one. Moreover, the two disturbances  $V_{t+1} = (\phi + \theta)Z_t$  and  $W_t = Z_t$  are perfectly correlated.

**ARMA(p,q) process**

It is straightforward to extend the above representation to ARMA(p,q) models.<sup>4</sup> Let  $\{Y_t\}$  be defined by the following stochastic difference equation:

$$\Phi(L)Y_t = \Theta(L)Z_t \quad \text{with } Z_t \sim \text{WN}(0, \sigma^2) \text{ and } \phi_p \theta_q \neq 0.$$

Define  $r$  as  $r = \max\{p, q + 1\}$  and set  $\phi_j = 0$  for  $j > p$  and  $\theta_j = 0$  for  $j > q$ . Then, we can set up the following state space representation with state vector  $\mathbf{X}_t$  and observation equation

$$Y_t = (1, \theta_1, \dots, \theta_{r-1})\mathbf{X}_t$$

where the state vector equals  $\mathbf{X}_t = (X_t, \dots, X_{t-r+2}, X_{t-r+1})'$  and where  $\{X_t\}$  follows an AR(p) process  $\Phi(L)X_t = Z_t$ . The AR(p) process can be transformed into companion form to arrive at the state equation:

$$\mathbf{X}_{t+1} = \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{r-1} & \phi_r \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \mathbf{X}_t + \begin{pmatrix} Z_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

**Missing Observations**

The state space approach is best suited to deal with missing observations. However, in this situation the coefficient matrices are no longer constant, but time-varying. Consider the following simple example of an AR(1) process for which we have only observations for the periods  $t = 1, \dots, 100$  and  $t = 102, \dots, 200$ , but not for period  $t = 101$  which is missing. This situation can be represented in state space form as follows:

$$\begin{aligned} X_{t+1} &= \phi X_t + Z_t \\ Y_t &= G_t X_t + W_t \\ G_t &= \begin{cases} 1, & t = 1, \dots, 100, 102, \dots, 200; \\ 0, & t = 101. \end{cases} \\ R_t &= \begin{cases} 0, & t = 1, \dots, 100, 102, \dots, 200; \\ c > 0, & t = 101. \end{cases} \end{aligned}$$

---

<sup>4</sup>See also exercise 17.2

This means that  $W_t = 0$  and that  $Y_t = X_t$  for all  $t$  except for  $t = 101$ . For the missing observation, we have  $G_{101} = Y_{101} = 0$ . The variance for this observation is set to  $R_{101} = c > 0$ .

The same idea can be used to obtain quarterly data when only yearly data are available. This problem typically arises in statistical offices which have to produce, for example, quarterly GDP data from yearly observations incorporating quarterly information from indicator variables (see Section 17.4.1). More detailed analysis for the case of missing data can be found in Harvey and Pierce (1984) and Brockwell and Davis (1991, Chapter 12.3).

### Structural Time Series Analysis

An important application of the state space representation in economics is the decomposition of a given time series into several components: trend, cycle, season and irregular component. This type of analysis is usually coined structural time series analysis (See Harvey, 1989; Mills, 2003). Consider, for example, the additive decomposition of a time series  $\{Y_t\}$  into a trend  $T_t$ , a seasonal component  $S_t$ , a cyclical component  $\{C_t\}$ , and an irregular or cyclical component  $W_t$ :

$$Y_t = T_t + S_t + C_t + W_t.$$

The above equation relates the observed time series to its unobserved components and is called the *basic structural model* (BSM) (Harvey, 1989).

The state space representation is derived in several steps. Consider first the case with no seasonal and no cyclical component. The trend is typically viewed as a random walk with time-varying drift  $\delta_{t-1}$ :

$$\begin{aligned} T_t &= \delta_{t-1} + T_{t-1} + \varepsilon_t, & \varepsilon_t &\sim \text{WN}(0, \sigma_\varepsilon^2) \\ \delta_t &= \delta_{t-1} + \xi_t, & \xi_t &\sim \text{WN}(0, \sigma_\xi^2). \end{aligned}$$

The second equation models the drift as a random walk. The two disturbances  $\{\varepsilon_t\}$  and  $\{\xi_t\}$  are assumed to be uncorrelated with each other and with  $\{W_t\}$ . Defining the state vector  $X_t^{(T)}$  as  $X_t^{(T)} = (T_t, \delta_t)'$ , the state and the observation equations become:

$$\begin{aligned} X_{t+1}^{(T)} &= \begin{pmatrix} T_{t+1} \\ \delta_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} T_t \\ \delta_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{t+1} \\ \xi_{t+1} \end{pmatrix} = F^{(T)} X_t^{(T)} + V_{t+1}^{(T)} \\ Y_t &= (1, 0) X_t^{(T)} + W_t \end{aligned}$$

with  $W_t \sim \text{WN}(0, \sigma_W^2)$ . This representation is called the local linear trend (LLT) model and implies that  $\{Y_t\}$  follows an ARIMA(0,2,2) process (see exercise 17.5.1).

In the special case of a constant drift equal to  $\delta$ ,  $\sigma_\xi^2 = 0$  and we have that  $\Delta Y_t = \delta + \varepsilon_t + W_t - W_{t-1}$ .  $\{\Delta Y_t\}$  therefore follows a MA(1) process with  $\rho(1) = -\sigma_W^2/(\sigma_\varepsilon^2 + 2\sigma_W^2) = -(2 + \kappa)^{-1}$  where  $\kappa = \sigma_\varepsilon^2/\sigma_W^2$  is called the signal-to-noise ratio. Note that the first order autocorrelation is necessarily negative. Thus, this model is not suited for time series with positive first order autocorrelation in its first differences.

The seasonal component is characterized by  $S_t = S_{t-d}$  and  $\sum_{t=1}^d S_t = 0$  where  $d$  denotes the frequency of the data.<sup>5</sup> Given starting values  $S_1, S_0, S_{-1}, \dots, S_{-d+3}$ , the following values can be computed recursively as:

$$S_{t+1} = -S_t - \dots - S_{t-d+2} + \eta_{t+1}, \quad t = 1, 2, \dots$$

where a noise  $\eta_t \sim \text{WN}(0, \sigma_\eta^2)$  is taken into account.<sup>6</sup> The state vector related to the seasonal component,  $X_t^{(S)}$ , is defined as  $X_t^{(S)} = (S_t, S_{t-1}, \dots, S_{t-d+2})'$  which gives the state equation

$$X_{t+1}^{(S)} = \begin{pmatrix} -1 & -1 & \dots & -1 & -1 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} X_t^{(S)} + \begin{pmatrix} \eta_{t+1} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = F^{(S)} X_t^{(S)} + V_{t+1}^{(S)}$$

with  $Q^{(S)} = \text{diag}(\sigma_\eta^2, 0, \dots, 0)$ .

Combining the trend and the seasonal model to an overall model with state vector  $X_t$  given by  $X_t = (X_t^{(T)'} , X_t^{(S)'})'$ , we arrive at the state equation:

$$X_{t+1} = \begin{pmatrix} F^{(T)} & 0 \\ 0 & F^{(S)} \end{pmatrix} X_t + \begin{pmatrix} V_{t+1}^{(T)} \\ V_{t+1}^{(S)} \end{pmatrix} = F X_t + V_{t+1}$$

with  $Q = \text{diag}(\sigma_\varepsilon^2, \sigma_\delta^2, \sigma_\eta^2, 0, \dots, 0)$ . The observation equation then is:

$$Y_t = (1 \ 0 \ 1 \ 0 \ \dots \ 0) X_t + W_t$$

with  $R = \sigma_W^2$ .

Finally, we can add a cyclical component  $\{C_t\}$  which is modeled as a harmonic process (see section 6.2) with frequency  $\lambda_C$ , respectively periodicity  $2\pi/\lambda_C$ :

$$C_t = A \cos \lambda_C t + B \sin \lambda_C t$$

<sup>5</sup>Four in the case of quarterly and twelve in the case of monthly observations.

<sup>6</sup>Alternative seasonal models can be found in Harvey (1989) and Hylleberg (1986).

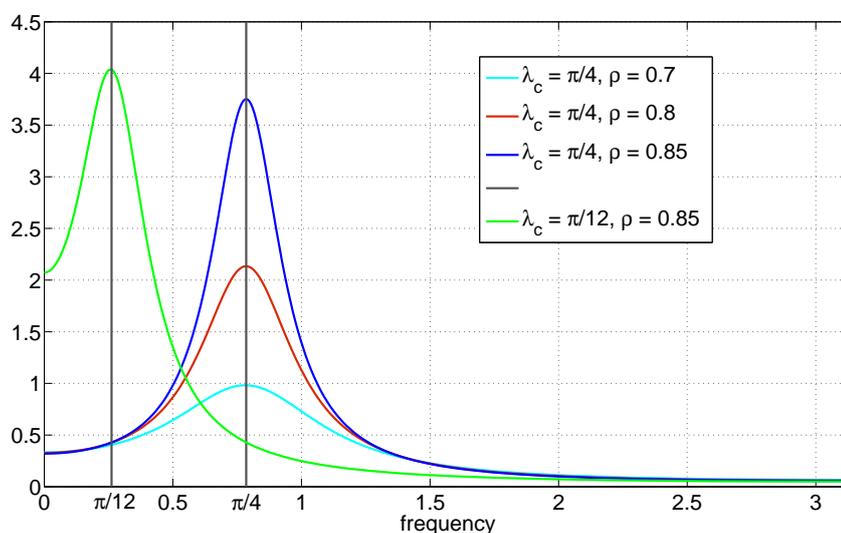


Figure 17.1: Spectral density of the cyclical component for different values of  $\lambda_C$  and  $\rho$

Following Harvey (1989, p.39), we let the parameters  $A$  and  $B$  evolve over time by introducing the recursion

$$\begin{pmatrix} C_{t+1} \\ C_{t+1}^* \end{pmatrix} = \rho \begin{pmatrix} \cos \lambda_C & \sin \lambda_C \\ -\sin \lambda_C & \cos \lambda_C \end{pmatrix} \begin{pmatrix} C_t \\ C_t^* \end{pmatrix} + \begin{pmatrix} V_{1,t+1}^{(C)} \\ V_{2,t+1}^{(C)} \end{pmatrix}$$

where  $C_0 = A$  and  $C_0^* = B$  and where  $\{C_t^*\}$  is an auxiliary process. The dampening factor  $\rho$  allows for additional flexibility in the specification.  $\{V_{1,t}^{(C)}\}$  and  $\{V_{2,t}^{(C)}\}$  are two mutually uncorrelated white-noise processes. It is instructive to examine the spectral density (see section 6.1) of the cyclical component in figure 17.1. It can be shown (see exercise 17.5.2) that  $\{C_t\}$  follows an ARMA(2,1) process.

The cyclical component can be incorporated into the state space model above by augmenting the state vector  $X_{t+1}$  by the cyclical components  $C_{t+1}$  and  $C_{t+1}^*$  and the error term  $V_{t+1}$  by  $\{V_{1,t+1}^{(C)}\}$  and  $\{V_{2,t+1}^{(C)}\}$ . The observations equation has to be amended accordingly. Section 17.4.2 presents an empirical application of this approach.

### Dynamic Factor Models

Dynamic factor models are an interesting approach when it comes to modeling simultaneously a large cross-section of time series. The concept was introduced into macroeconomics by Sargent and Sims (1977) and was then developed further and popularized by Quah and Sargent (1993), Reichlin (2003) and Breitung and Eickmeier (2006), among others. The idea is to view each time series  $Y_{it}$ ,  $i = 1, \dots, n$ , as the sum of a linear combination of some joint unobserved factors  $f_t = (f_{1t}, \dots, f_{rt})'$  and an idiosyncratic component  $\{W_{it}\}$ ,  $i = 1, \dots, n$ . Dynamic factor models are particularly effective when the number of factors  $r$  is small compared to the number of time series  $n$ . In practice, several hundred time series are related to a handful factors. In matrix notation we can write the observation equation for the dynamic factor model as follows:

$$Y_t = \Lambda_0 f_t + \Lambda_1 f_{t-1} + \dots + \Lambda_q f_{t-q} + W_t$$

where  $\Lambda_i$ ,  $i = 0, 1, \dots, q$ , are  $n \times r$  matrices. The state vector  $X_t$  equals  $(f_t', \dots, f_{t-q}')'$  if we assume that the idiosyncratic component is white noise, i.e.  $W_t = (W_{1t}, \dots, W_{nt})' \sim \text{WN}(0, R)$ . The observation equation can then be written compactly as:

$$Y_t = GX_t + W_t$$

where  $G = (\Lambda_0, \Lambda_1, \dots, \Lambda_q)$ . Usually, we assume that  $R$  is a diagonal matrix. The correlation between the different time series is captured exclusively by the joint factors.

The state equation depends on the assumed dynamics of the factors. One possibility is to model  $\{f_t\}$  as a VAR(p) process with  $\Phi(L)f_t = e_t$ ,  $e_t \sim \text{WN}(0, \Sigma)$ , and  $p \leq q + 1$ , so we can use the state space representation of the VAR(p) process from above. For the case  $p = 2$  and  $q = 2$  we get:

$$X_{t+1} = \begin{pmatrix} f_{t+1} \\ f_t \\ f_{t-1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 & 0 \\ I_r & 0 & 0 \\ 0 & I_r & 0 \end{pmatrix} \begin{pmatrix} f_t \\ f_{t-1} \\ f_{t-2} \end{pmatrix} + \begin{pmatrix} e_{t+1} \\ 0 \\ 0 \end{pmatrix} = FX_t + V_{t+1}$$

and  $Q = \text{diag}(\Sigma, 0, 0)$ . This scheme can be easily generalized to the case  $p > q + 1$  or to allow for autocorrelated idiosyncratic components, assuming for example that they follow autoregressive processes.

### Real Business Cycle Model (RBC Model)

State space models are becoming increasingly popular in macroeconomics, especially in the context of dynamic stochastic general equilibrium (DSGE)

models. These models can be seen as generalizations of the real business cycle (RBC) models.<sup>7</sup> In these models a representative consumer is supposed to maximize the utility of his consumption stream over his infinite life time. Thereby, the consumer has the choice to consume part of his income or to invest his savings (part of his income which is not consumed) at the market rate of interest. These savings can be used as a mean to finance investment projects which increase the economy wide capital stock. The increased capital stock then allows for increased production in the future. The production process itself is subject to a random shocks called technology shocks.

The solution of this optimization problem is a nonlinear dynamic system which determines the capital stock and consumption in every period. Its local behavior can be investigated by linearizing the system around its steady state. This equation can then be interpreted as the state equation of the system. The parameters of this equation  $F$  and  $Q$  are related, typically in a nonlinear way, to the parameters describing the utility and the production function as well as the process of technology shocks. Thus, the state equation summarizes the behavior of the theoretical model.

The parameters of the state equation can then be estimated by relating the state vector, given by the capital stock and the state of the technology, via the observation equation to some observable variables, like real GDP, consumption, investment, or the interest rate. This then completes the state space representation of the model which can be analyzed and estimated using the tools presented in Section 17.3.<sup>8</sup>

## 17.2 Filtering and Smoothing

As we have seen, the state space model provides a very flexible framework for a wide array of applications. We therefore want to develop a set of tools to handle this kind of models in terms of interpretation and estimation. In this section we will analyze the problem of inferring the unobserved state from the data given the parameters of the model. In Section 17.3 we will then investigate the estimation of the parameters by maximum likelihood.

In many cases the state of the system is not or only partially observable. It is therefore of interest to infer from the data  $Y_1, Y_2, \dots, Y_T$  the state vector

---

<sup>7</sup>Prototypical models can be found in King et al. (1988) or Woodford (2003). Canova (2007) and Dejong and Dave (2007) present a good introduction to the analysis of DSGE models.

<sup>8</sup>See Sargent (2004) or Fernandez-Villaverde et al. (2007) for systematic treatment of state space models in the context of macroeconomic models. In this literature the use of Bayesian methods is widespread (see An and Schorfheide, 2007; Dejong and Dave, 2007).

$X_t$ . We can distinguish three types of problems depending on the information used:

- (i) estimation of  $X_t$  from  $Y_1, \dots, Y_{t-1}$ , known as the *prediction* problem;
- (ii) estimation of  $X_t$  from  $Y_1, \dots, Y_t$ , known as the *filtering* problem;
- (iii) estimation of  $X_t$  from  $Y_1, \dots, Y_T$ , known as the *smoothing* problem.

For the ease of exposition, we will assume that the disturbances  $V_t$  and  $W_t$  are normally distributed. The recursive nature of the state equation implies that  $X_t = F^{t-1}X_1 + \sum_{j=0}^{t-2} F^j V_{t-j}$ . Therefore,  $X_t$  is also normally distributed for all  $t$ , if  $X_1$  is normally distributed. From the observation equation we can infer also that  $Y_t$  is normally distributed, because it is the sum of two normally distributed random variables  $A + GX_t$  and  $W_t$ . Thus, under these assumptions, the vector  $(X'_1, \dots, X'_T, Y'_1, \dots, Y'_T)'$  is jointly normally distributed:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_T \\ Y_1 \\ \vdots \\ Y_T \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Gamma_X & \Gamma_{YX} \\ \Gamma_{XY} & \Gamma_Y \end{pmatrix} \right)$$

where the covariance matrices  $\Gamma_X, \Gamma_{YX}, \Gamma_{XY}$  and  $\Gamma_Y$  can be retrieved from the model given the parameters.

For the understanding of the rest of this section, the following theorem is essential (see standard textbooks, like Amemiya, 1994; Greene, 2008).

**Theorem 17.1.** *Let  $Z$  be a  $n$ -dimensional normally distributed random variable with  $Z \sim N(\mu, \Sigma)$ . Consider the partitioned vector  $Z = (Z'_1, Z'_2)'$  where  $Z_1$  and  $Z_2$  are of dimensions  $n_1 \geq 1$  and  $n_2 \geq 1$ ,  $n = n_1 + n_2$ , respectively. The corresponding partitioning of the covariance matrix  $\Sigma$  is*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where  $\Sigma_{11} = \mathbb{V}Z_1$ ,  $\Sigma_{22} = \mathbb{V}Z_2$ , and  $\Sigma_{12} = \Sigma'_{21} = \text{cov}(Z_1, Z_2) = \mathbb{E}(Z_1 - \mathbb{E}Z_1)'(Z_2 - \mathbb{E}Z_2)$ . Then the partitioned vectors  $Z_1$  and  $Z_2$  are normally distributed. Moreover, the conditional distribution of  $Z_1$  given  $Z_2$  is also normal with mean and variance

$$\begin{aligned} \mathbb{E}(Z_1|Z_2) &= \mathbb{E}Z_1 + \Sigma_{12}\Sigma_{22}^{-1}(Z_2 - \mathbb{E}Z_2), \\ \mathbb{V}(Z_1|Z_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

This formula can be directly applied to figure out the mean and the variance of the state vector given the observations. Setting  $Z_1 = (X'_1, \dots, X'_t)'$  and  $Z_2 = (Y'_1, \dots, Y'_{t-1})'$ , we get the predicted values; setting  $Z_1 = (X'_1, \dots, X'_t)'$  and  $Z_2 = (Y'_1, \dots, Y'_t)'$ , we get the filtered values; setting  $Z_1 = (X'_1, \dots, X'_t)'$  and  $Z_2 = (Y'_1, \dots, Y'_T)'$ , we get the smoothed values.

### AR(1) Process with Measurement Errors

We illustrate the above ideas by analyzing a univariate AR(1) process with measurement errors:<sup>9</sup>

$$\begin{aligned} X_{t+1} &= \phi X_t + v_{t+1}, & v_t &\sim \text{IIDN}(0, \sigma_v^2) \\ Y_t &= X_t + w_t, & w_t &\sim \text{IIDN}(0, \sigma_w^2). \end{aligned}$$

For simplicity, we assume  $|\phi| < 1$ . Suppose that we only have observations  $Y_1$  and  $Y_2$  at our disposal. The joint distribution of  $(X_1, X_2, Y_1, Y_2)'$  is normal. The covariances can be computed by applying the methods discussed in Chapter 2:

$$\begin{pmatrix} X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \frac{\sigma_v^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & 1 & \phi \\ \phi & 1 & \phi & 1 \\ 1 & \phi & 1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2} & \phi \\ \phi & 1 & \phi & 1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2} \end{pmatrix} \right)$$

The smoothed values are obtained by applying the formula from Theorem 17.1:

$$\begin{aligned} \mathbb{E} \left( \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \middle| Y_1, Y_2 \right) &= \\ &= \frac{1}{(1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2})^2 - \phi^2} \begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix} \begin{pmatrix} 1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2} & -\phi \\ -\phi & 1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \end{aligned}$$

Note that for the last observation,  $Y_2$  in our case, the filtered and the smoothed values are the same. For  $X_1$  the filtered value is

$$\mathbb{E}(X_1|Y_1) = \frac{1}{1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2}} Y_1.$$

An intuition for this result can be obtained by considering some special cases. For  $\phi = 0$ , the observations are not correlated over time. The filtered

<sup>9</sup>Sargent (1989) provides an interesting application showing the implications of measurement errors in macroeconomic models.

value for  $X_1$  therefore corresponds to the smoothed one. This value lies between zero, the unconditional mean of  $X_1$ , and  $Y_1$  with the variance ratio  $\sigma_w^2/\sigma_v^2$  delivering the weights: the smaller the variance of the measurement error the closer the filtered value is to  $Y_1$ . This conclusion holds also in general. If the variance of the measurement error is relatively large, the observations do not deliver much information so that the filtered and the smoothed values are close to the unconditional mean.

For large systems the method suggested by Theorem 17.1 may run into numerical problems due to the inversion of the covariance matrix of  $Y$ ,  $\Sigma_{22}$ . This matrix can become rather large as it is of dimension  $nT \times nT$ . Fortunately, there exist recursive solutions to this problem known as the Kalman filter, and also the Kalman smoother.

### 17.2.1 The Kalman Filter

The Kalman filter circumvents the problem of inverting a large  $nT \times nT$  matrix by making use of the Markov property of the system (see Remark 17.4). The distribution of  $X_t$  given the observations up to period  $t$  can thereby be computed recursively from the distribution of the state in period  $t - 1$  given the information available up to period  $t - 1$ . Starting from some initial distribution in period 0, we can in this way obtain in  $T$  steps the distribution of all states. In each step only an  $n \times n$  matrix must be inverted. To describe the procedure in detail, we introduce the following notation:

$$\begin{aligned}\mathbb{E}(X_t|Y_1, \dots, Y_h) &= X_{t|h} \\ \mathbb{V}(X_t|Y_1, \dots, Y_h) &= P_{t|h}.\end{aligned}$$

Suppose, we have already determined the distribution of  $X_t$  conditional on the observations  $Y_1, \dots, Y_t$ . Because we are operating in a framework of normally distributed random variables, the distribution is completely characterized by its conditional mean  $X_{t|t}$  and variance  $P_{t|t}$ . The goal is to carry forward these entities to obtain  $X_{t+1|t+1}$  and  $P_{t+1|t+1}$  having observed an additional data point  $Y_{t+1}$ . This problem can be decomposed into a *forecasting* and an *updating* step.

**Step 1: Forecasting step** The state equation and the assumption about the disturbance term  $V_{t+1}$  imply:

$$\begin{aligned}X_{t+1|t} &= FX_{t|t} \\ P_{t+1|t} &= FP_{t|t}F' + Q\end{aligned}\tag{17.5}$$

The observation equation then allows to compute a forecast of  $Y_{t+1}$  where we assume for simplicity that  $A = 0$ :

$$Y_{t+1|t} = GX_{t+1|t} \quad (17.6)$$

**Step 2: Updating step** In this step the additional information coming from the additional observation  $Y_{t+1}$  is processed to update the conditional distribution of the state vector. The joint conditional distribution of  $(X'_{t+1}, Y'_{t+1})'$  given  $Y_1, \dots, Y_t$  is

$$\begin{pmatrix} X_{t+1} \\ Y_{t+1} \end{pmatrix} \Big| Y_1, \dots, Y_t \sim N \left( \begin{pmatrix} X_{t+1|t} \\ Y_{t+1|t} \end{pmatrix}, \begin{pmatrix} P_{t+1|t} & P_{t+1|t}G' \\ GP_{t+1|t} & GP_{t+1|t}G' + R \end{pmatrix} \right)$$

As all elements of the distribution are available from the forecasting step, we can again apply Theorem 17.1 to get the distribution of the filtered state vector at time  $t + 1$ :

$$X_{t+1|t+1} = X_{t+1|t} + P_{t+1|t}G'(GP_{t+1|t}G' + R)^{-1}(Y_{t+1} - Y_{t+1|t}) \quad (17.7)$$

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}G'(GP_{t+1|t}G' + R)^{-1}GP_{t+1|t} \quad (17.8)$$

where we replace  $X_{t+1|t}$ ,  $P_{t+1|t}$ , and  $Y_{t+1|t}$  by  $FX_{t|t}$ ,  $FP_{t|t}F' + Q$ , and  $GFX_{t|t}$ , respectively, which have been obtained from the forecasting step.

Starting from given values for  $X_{0|0}$  and  $P_{0|0}$ , we can therefore iteratively compute  $X_{t|t}$  and  $P_{t|t}$  for all  $t = 1, 2, \dots, T$ . Only the information from the last period is necessary at each step. Inserting equation (17.7) into equation (17.5) we obtain as a forecasting equation:

$$X_{t+1|t} = FX_{t|t-1} + FP_{t|t-1}G'(GP_{t|t-1}G' + R)^{-1}(Y_t - GX_{t|t-1})$$

where the matrix

$$K_t = FP_{t|t-1}G'(GP_{t|t-1}G' + R)^{-1}$$

is known as the (*Kalman*) *gain matrix*. It prescribes how the innovation  $Y_t - Y_{t|t-1} = Y_t - GX_{t|t-1}$  leads to an update of the predicted state.

**Initializing the algorithm** It remains to determine how to initialize the recursion. In particular, how to set the starting values for  $X_{0|0}$  and  $P_{0|0}$ . If  $X_t$  is stationary and causal with respect to  $V_t$ , the state equation has the solution  $X_0 = \sum_{j=0}^{\infty} F^j V_{t-j}$ . Thus,

$$X_{0|0} = \mathbb{E}(X_0) = 0$$

$$P_{0|0} = \mathbb{V}(X_0)$$

where  $P_{0|0}$  solves the equation (see Section 12.4)

$$P_{0|0} = FP_{0|0}F' + Q.$$

According to equation (12.4), the solution of the above matrix equation is:

$$\text{vec}(P_{0|0}) = [I - F \otimes F]^{-1} \text{vec}(Q).$$

If the process is not stationary, we can set  $X_{0|0}$  to zero and  $P_{0|0}$  to infinity. In practice, a very large number is sufficient.

### 17.2.2 The Kalman Smoother

The Kalman filter determines the distribution of the state at time  $t$  given the information available up to this time. In many instances, we want, however, make an optimal forecast of the state given all the information available, i.e. the whole sample. Thus, we want to determine  $X_{t|T}$  and  $P_{t|T}$ . The Kalman filter determines the smoothed distribution for  $t = T$ , i.e.  $X_{T|T}$  and  $P_{T|T}$ . The idea of the Kalman smoother is again to determine the smoothed distribution in a recursive manner. For this purpose, we let the recursion run backwards. Starting with the last observation in period  $t = T$ , we proceed back in time by letting  $t$  take successively the values  $t = T - 1, T - 2, \dots$  until the first observation in period  $t = 1$ .

Using again the linearity of the equations and the normality assumption, we get:

$$\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \Big| Y_1, Y_2, \dots, Y_t \sim N \left( \begin{pmatrix} X_{t|t} \\ FX_{t|t} \end{pmatrix}, \begin{pmatrix} P_{t|t} & P_{t|t}F' \\ FP_{t|t} & P_{t+1|t} \end{pmatrix} \right)$$

This implies that

$$\mathbb{E}(X_t | Y_1, \dots, Y_t, X_{t+1}) = X_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}(X_{t+1} - X_{t+1|t}).$$

The above mean is only conditional on all information available up to time  $t$  and on the information at time  $t + 1$ . The Markov property implies that this mean also incorporates the information from the observations  $Y_{t+1}, \dots, Y_T$ . Thus, we have:

$$\begin{aligned} \mathbb{E}(X_t | Y_1, \dots, Y_T, X_{t+1}) &= \mathbb{E}(X_t | Y_1, \dots, Y_t, X_{t+1}) \\ &= X_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}(X_{t+1} - X_{t+1|t}) \end{aligned}$$

Applying the law of iterated expectations or means (see, for example, Amemiya, 1994, p. 78), we can derive  $X_{t|T}$ :

$$\begin{aligned} X_{t|T} &= \mathbb{E}(X_t|Y_1, \dots, Y_T) = \mathbb{E}(\mathbb{E}(X_t|Y_1, \dots, Y_T, X_{t+1})|Y_1, \dots, Y_T) \\ &= \mathbb{E}(X_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}(X_{t+1} - X_{t+1|t})|Y_1, \dots, Y_T) \\ &= X_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}(X_{t+1|T} - X_{t+1|t}). \end{aligned} \quad (17.9)$$

The algorithm can now be implemented as follows. In the first step compute  $X_{T-1|T}$  according to equation (17.9) as

$$X_{T-1|T} = X_{T-1|T-1} + P_{T-1|T-1}F'P_{T|T-1}^{-1}(X_{T|T} - X_{T|T-1}).$$

All entities on the right hand side can readily be computed by applying the Kalman filter. Having found  $X_{T-1|T}$ , we can again use equation (17.9) for  $t = T - 2$  to evaluate  $X_{T-2|T}$ :

$$X_{T-2|T} = X_{T-2|T-2} + P_{T-2|T-2}F'P_{T-1|T-2}^{-1}(X_{T-1|T} - X_{T-1|T-2}).$$

Proceeding backward through the sample we can derive a complete sequence of smoothed states  $X_{T|T}, X_{T-1|T}, X_{T-2|T}, \dots, X_{1|T}$ . These calculations are based on the computations of  $X_{t|t}, X_{t+1|t}, P_{t|t}$ , and  $P_{t+1|t}$  which have already been obtained from the Kalman filter. The smoothed covariance matrix  $P_{t|T}$  is given as (see Hamilton, 1994a, Section 13.6):

$$P_{t|T} = P_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}(P_{t+1|T} - P_{t+1|t})P_{t+1|t}^{-1}F'P_{t|t}.$$

Thus, we can compute also the smoothed variance with the aid of the values already determined by the Kalman filter.

### AR(1) Process with Measurement Errors (continued)

We continue our illustrative example of an AR(1) process with measurement errors and just two observations. First, we determine the filtered values for the state vector with the aid of the Kalman filter. To initialize the process, we have to assign a distribution to  $X_0$ . For simplicity, we assume that  $|\phi| < 1$  so that it makes sense to assign the stationary distribution of the process as the distribution for  $X_0$ :

$$X_0 \sim N\left(0, \frac{\sigma_V^2}{1 - \phi^2}\right)$$

Then we compute the forecasting step as the first step of the filter (see equation (17.5)):

$$\begin{aligned} X_{1|0} &= \phi X_{0|0} = 0 \\ P_{1|0} &= \phi^2 \frac{\sigma_v^2}{1 - \phi^2} + \sigma_v^2 = \frac{\sigma_v^2}{1 - \phi^2} \\ Y_{1|0} &= 0. \end{aligned}$$

$P_{1|0}$  was computed by the recursive formula from the previous section, but is, of course, equal to the unconditional variance. For the updating step, we get from equations (17.7) and (17.8):

$$\begin{aligned} X_{1|1} &= \left( \frac{\sigma_v^2}{1 - \phi^2} \right) \left( \frac{\sigma_v^2}{1 - \phi^2} + \sigma_w^2 \right)^{-1} Y_1 = \frac{1}{1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2}} Y_1 \\ P_{1|1} &= \left( \frac{\sigma_v^2}{1 - \phi^2} \right) - \left( \frac{\sigma_v^2}{1 - \phi^2} \right)^2 \left( \frac{\sigma_v^2}{1 - \phi^2} + \sigma_w^2 \right)^{-1} \\ &= \frac{\sigma_v^2}{1 - \phi^2} \left( 1 - \frac{1}{1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2}} \right) \end{aligned}$$

These two results are then used to calculate the next iteration of the algorithm. This will give the filtered values for  $t = 2$  which would correspond to the smoothed values because we just have two observations. The forecasting step is:

$$\begin{aligned} X_{2|1} &= \frac{\phi}{1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2}} Y_1 \\ P_{2|1} &= \frac{\phi^2 \sigma_v^2}{1 - \phi^2} \left( 1 - \frac{1}{1 + \frac{\sigma_w^2(1-\phi^2)}{\sigma_v^2}} \right) + \sigma_v^2 \end{aligned}$$

Next we perform the updating step to calculate  $X_{2|2}$  and  $P_{2|2}$ . It is easy to verify that this leads to the same results as in the first part of this example.

An interesting special case is obtained when we assume that  $\phi = 1$  so that the state variable is a simple random walk. In this case the unconditional variance of  $X_t$  and consequently also of  $Y_t$  are no longer finite. As mentioned previously, we can initialize the Kalman filter by  $X_{0|0} = 0$  and  $P_{0|0} = \infty$ . This implies:

$$\begin{aligned} Y_{1|0} &= X_{1|0} = X_{0|0} = 0 \\ P_{1|0} &= P_{0|0} + \sigma_v^2 = \infty. \end{aligned}$$

Inserting this result in the updating equations (17.7) and (17.8), we arrive at:

$$X_{1|1} = \frac{P_{1|0}}{P_{1|0} + \sigma_W^2} (Y_1 - Y_{1|0}) = \frac{P_{0|0} + \sigma_V^2}{P_{0|0} + \sigma_V^2 + \sigma_W^2} Y_1$$

$$P_{1|1} = P_{1|0} - \frac{P_{1|0}^2}{P_{1|0} + \sigma_W^2} = (P_{0|0} + \sigma_V^2) \left( 1 - \frac{P_{1|0}}{P_{1|0} + \sigma_W^2} \right) = \frac{(P_{0|0} + \sigma_V^2) \sigma_W^2}{P_{0|0} + \sigma_V^2 + \sigma_W^2}.$$

Letting  $P_{0|0}$  go to infinity, leads to:

$$X_{1|1} = Y_1$$

$$P_{1|1} = \sigma_W^2.$$

This shows that the filtered variance is finite for  $t = 1$  although  $P_{1|0}$  was infinite.

## 17.3 Estimation of State Space Models

Up to now we have assumed that the parameters of the system are known and that only the state is unknown. In most economic applications, however, also the parameters are unknown and have therefore to be estimated from the data. One big advantage of the state space models is that they provide an integrated approach to forecasting, smoothing and estimation. In particular, the Kalman filter turns out to be an efficient and quick way to compute the likelihood function. Thus, it seems natural to estimate the parameters of state space models by the method of maximum likelihood.

### 17.3.1 The Likelihood Function

The joint unconditional density of the observations  $(Y_1', \dots, Y_T)'$  can be factorized into the product of conditional densities as follows:

$$f(Y_1, \dots, Y_T) = f(Y_T | Y_1, \dots, Y_{T-1}) f(Y_1, \dots, Y_{T-1})$$

$$= \vdots$$

$$= f(Y_T | Y_1, \dots, Y_{T-1}) f(Y_{T-1} | Y_1, \dots, Y_{T-2}) \dots f(Y_2 | Y_1) f(Y_1)$$

Each conditional density is the density of a normal distribution and is therefore given by:

$$f(Y_t | Y_1, \dots, Y_{t-1}) = (2\pi)^{-n/2} (\det \Delta_t)^{-1/2} \exp \left[ -\frac{1}{2} (Y_t - Y_{t|t-1})' \Delta_t^{-1} (Y_t - Y_{t|t-1}) \right]$$

where  $\Delta_t = GP_{t|t-1}G' + R$ . The Gaussian likelihood function  $L$  is therefore equal to:

$$L = (2\pi)^{-(Tn)/2} \left( \prod_{t=1}^T \det(\Delta_t) \right)^{-1/2} \exp \left[ -\frac{1}{2} \sum_{t=1}^T (Y_t - Y_{t|t-1})' \Delta_t^{-1} (Y_t - Y_{t|t-1}) \right].$$

Note that all the entities necessary to evaluate the likelihood function are provided by the Kalman filter. Thus, the evaluation of the likelihood function is a byproduct of the Kalman filter. The maximum likelihood estimator (MLE) is then given by the maximizer of the likelihood function, or more conveniently the log-likelihood function. Usually, there is no analytic solution available so that one must resort to numerical methods. An estimation of the asymptotic covariance matrix can be obtained by evaluating the Hessian matrix at the optimum. Under the usual assumptions, the MLE is consistent and delivers asymptotically normally distributed estimates (Greene, 2008; Amemiya, 1994).

The direct maximization of the likelihood function is often not easy in practice, especially for large systems with many parameters. The expectation-maximization algorithm, EM algorithm for short, represents a valid, though slower alternative. As the name indicates, it consists of two steps which have to be carried out iteratively. Based on some starting values for the parameters, the first step (expectation step) computes estimates,  $X_{t|T}$ , of the unobserved state vector  $X_t$  using the Kalman smoother. In the second step (maximization step), the likelihood function is maximized taking the estimates of  $X_t$ ,  $X_{t|T}$ , as additional observations. The treatment of  $X_{t|T}$  as additional observations, allows to reduce the maximization step to a simple multivariate regression. Indeed, by treating  $X_{t|T}$  as if they were known, the state equation becomes a simple VAR(1) which can be readily estimated by linear least-squares to obtain the parameters  $F$  and  $Q$ . The parameters  $A$ ,  $G$  and  $R$  are also easily retrieved from a regression of  $Y_t$  on  $X_{t|T}$ . Based on these new parameter estimates, we go back to step one and derive new estimates for  $X_{t|T}$  which are then used in the maximization step. One can show that this procedure maximizes the original likelihood function (see Dempster et al., 1977; Wu, 1983). A more detailed analysis of the EM algorithm in the time series context is provided by Brockwell and Davis (1996).<sup>10</sup>

<sup>10</sup>The analogue to the EM algorithm in the Bayesian context is given by the Gibbs sampler. In contrast to the EM algorithm, we compute in the first step not the expected value of the states, but we draw a state vector from the distribution of state vectors given the parameters. In the second step, we do not maximize the likelihood function, but draw a parameter from the distribution of parameters given the state vector drawn previously. Going back and forth between these two steps, we get a Markov chain in the parameters

Sometimes it is of interest not only to compute parameter estimates and to derive from them estimates for the state vector via the Kalman filter or smoother, but also to find confidence intervals for the estimated state vector to take the uncertainty into account. If the parameters are known, the methods outlined previously showed how to obtain these confidence intervals. If, however, the parameters have to be estimated, there is a double uncertainty: the uncertainty from the filter and the uncertainty arising from the parameter estimates. One way to account for this additional uncertainty is by the use of simulations. Thereby, we draw a given number of parameter vectors from the asymptotic distribution and compute for each of these draws the corresponding estimates for the state vector. The variation in these estimates is then a measure of the uncertainty arising from the estimation of the parameters (see Hamilton, 1994a, Section 13.7).

### 17.3.2 Identification

As emphasized in Remark 17.5 of Section 17.1, the state space representations are not unique. See, for example, the two alternative representations of the ARMA(1,1) model in Section 17.1. This non-uniqueness of state space models poses an identification problem because different specifications may give rise to observationally equivalent models.<sup>11</sup> This problem is especially serious if all states are unobservable. In practice, the identification problem gives rise to difficulties in the numerical maximization of the likelihood function. For example, one may obtain large differences for small variations in the starting values; or one may encounter difficulties in the inversion of the matrix of second derivatives.

The identification of state space models can be checked by transforming them into VARMA models and by investigating the issue in this reparameterized setting (Hannan and Deistler, 1988). Exercise 17.5.6 invites the reader to apply this method to the AR(1) model with measurement errors. System identification is a special field in systems theory and will not be pursued further here. A systematic treatment can be found in the textbook by Ljung (1999).

---

and the states whose stationary distribution is exactly the distribution of parameters and states given the data. A detailed description of Bayesian methods and the Gibbs sampler can be found in Geweke (2005). Kim and Nelson (1999) discuss this method in the context of state space models.

<sup>11</sup>Remember that, in our context, two representations are equivalent if they generate the same mean and covariance function for  $\{Y_t\}$ .

## 17.4 Examples

### 17.4.1 Estimating Quarterly GDP data from Yearly Ones

The official data for quarterly GDP are released in Switzerland by the State Secretariat for Economic Affairs (SECO). They estimate these data taking the yearly values provided by the Federal Statistical Office (FSO) as given. This division of tasks is not uncommon in many countries. One of the most popular methods for disaggregation of yearly data into quarterly ones was proposed by Chow and Lin (1971).<sup>12</sup> It is a regression based method which can take additional information in the form of indicator variables (i.e. variables which are measured at the higher frequency and correlated at the lower frequency with the variable of interest) into account. This procedure is, however, rather rigid. The state space framework is much more flexible and ideally suited to deal with missing observations. Applications of this framework to the problem of disaggregation were provided by Bernanke et al. (1997:1) and Cuche and Hess (2000), among others. We will illustrate this approach below.

Starting point of the analysis are the yearly growth rates of GDP and indicator variables which are recorded at the quarterly frequency and which are correlated with GDP growth. In our application, we will consider the growth of industrial production ( $IP$ ) and the index on consumer sentiment ( $C$ ) as indicators. Both variables are available on a quarterly basis from 1990 onward. For simplicity, we assume that the annualized quarterly growth rate of GDP,  $\{Q_t\}$ , follows an AR(1) process with mean  $\mu$ :

$$Q_t - \mu = \phi(Q_{t-1} - \mu) + w_t, \quad w_t \sim \text{WN}(0, \sigma_w^2)$$

In addition, we assume that GDP is related to industrial production and consumer sentiment by the following two equations:

$$\begin{aligned} IP_t &= \alpha_{IP} + \beta_{IP}Q_t + v_{IP,t} \\ C_t &= \alpha_C + \beta_CQ_t + v_{C,t} \end{aligned}$$

where the residuals  $v_{IP,t}$  and  $v_{C,t}$  are uncorrelated. Finally, we define the relation between quarterly and yearly GDP growth as:

$$J_t = \frac{1}{4}Q_t + \frac{1}{4}Q_{t-1} + \frac{1}{4}Q_{t-2} + \frac{1}{4}Q_{t-3}, \quad t = 4, 8, 12 \dots$$

---

<sup>12</sup>Similarly, one may envisage the disaggregation of yearly data into monthly ones or other forms of disaggregation.

We can now bring these equations into state space form. Thereby the observation equation is given by

$$Y_t = A_t + G_t X_t + W_t$$

with observation and state vectors

$$Y_t = \begin{cases} \begin{pmatrix} J_t \\ IP_t \\ C_t \end{pmatrix}, & t = 4, 8, 12, \dots; \\ \begin{pmatrix} 0 \\ IP_t \\ C_t \end{pmatrix}, & t \neq 4, 8, 12, \dots \end{cases}$$

$$X_t = \begin{pmatrix} Q_t - \mu \\ Q_{t-1} - \mu \\ Q_{t-2} - \mu \\ Q_{t-3} - \mu \end{pmatrix}$$

and time-varying coefficient matrices

$$A_t = \begin{cases} \begin{pmatrix} \mu \\ \alpha_{IP} \\ \alpha_C \end{pmatrix}, & t = 4, 8, 12, \dots; \\ \begin{pmatrix} 0 \\ \alpha_{IP} \\ \alpha_C \end{pmatrix}, & t \neq 4, 8, 12, \dots \end{cases}$$

$$G_t = \begin{cases} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \beta_{IP} & 0 & 0 & 0 \\ \beta_C & 0 & 0 & 0 \end{pmatrix}, & t = 4, 8, 12, \dots; \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta_{IP} & 0 & 0 & 0 \\ \beta_C & 0 & 0 & 0 \end{pmatrix}, & t \neq 4, 8, 12, \dots \end{cases}$$

$$R_t = \begin{cases} \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_{IP}^2 & 0 \\ 0 & 0 & \sigma_C^2 \end{pmatrix}, & t = 4, 8, 12, \dots; \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_{IP}^2 & 0 \\ 0 & 0 & \sigma_C^2 \end{pmatrix}, & t \neq 4, 8, 12, \dots \end{cases}$$

The state equation becomes:

$$X_{t+1} = FX_t + V_{t+1}$$

where

$$F = \begin{pmatrix} \phi & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$Q = \begin{pmatrix} \sigma_w^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

On my homepage <http://www.neusser.ch/> you will find a MATLAB code which maximizes the corresponding likelihood function numerically. Figure 17.2 plots the different estimates of GDP growth and compares them with the data released by SECO.

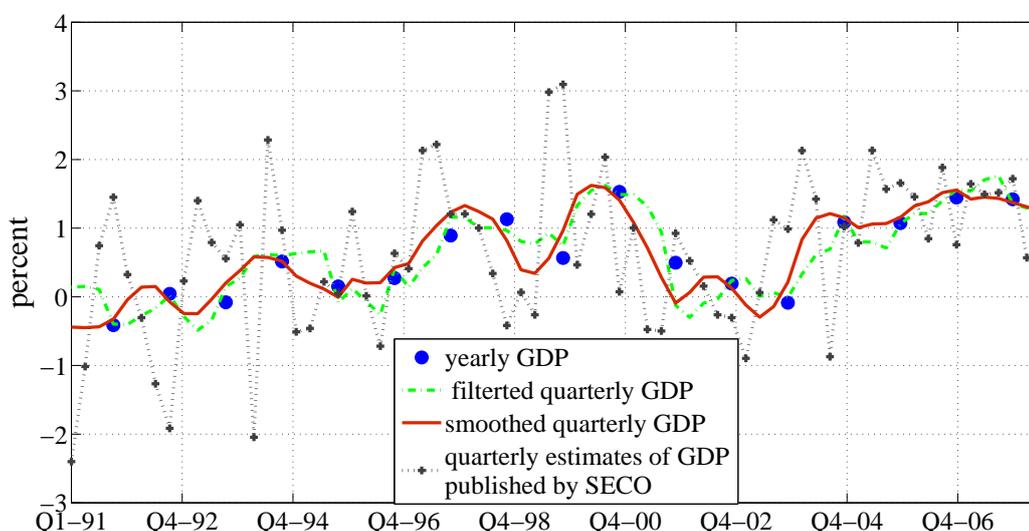


Figure 17.2: Estimates of quarterly GDP growth rates for Switzerland

### 17.4.2 Structural Time Series Analysis

A customary practice in business cycle analysis is to decompose a time series into several components. As an example, we estimate a structural time series model which decomposes a times series additively into a local linear trend, a business cycle component, a seasonal component, and an irregular component. This is the specification studied as the basic structural model (BSM) in subsection 17.1.1. We carry over the specification explained there to apply it to quarterly real GDP of Switzerland. Figure 17.3 shows the smoothed estimates of the various components. In the left upper panel the demeaned logged original series (see figure 17.3(a)) is plotted. One clearly discern the trend and the seasonal variations. The right upper panel shows the local linear trend (LLT). As one can see the trend is not a straight line, but exhibits pronounced waves of low frequency. The business cycle component showed in figure 17.3(c) is much more volatile. The large drop of about 2.5 percent in 2008/09 corresponds to the financial markets. The lower right panel plots the seasonal component (see figure 17.3(d)). From a visual inspections, one can infer that the volatility of the seasonal component is much larger than the cyclical component (compare the scale of the two components) so that movements in GDP are dominated by seasonal fluctuations.<sup>13</sup> Moreover, the

<sup>13</sup>The irregular component which is not shown has only very small variance.

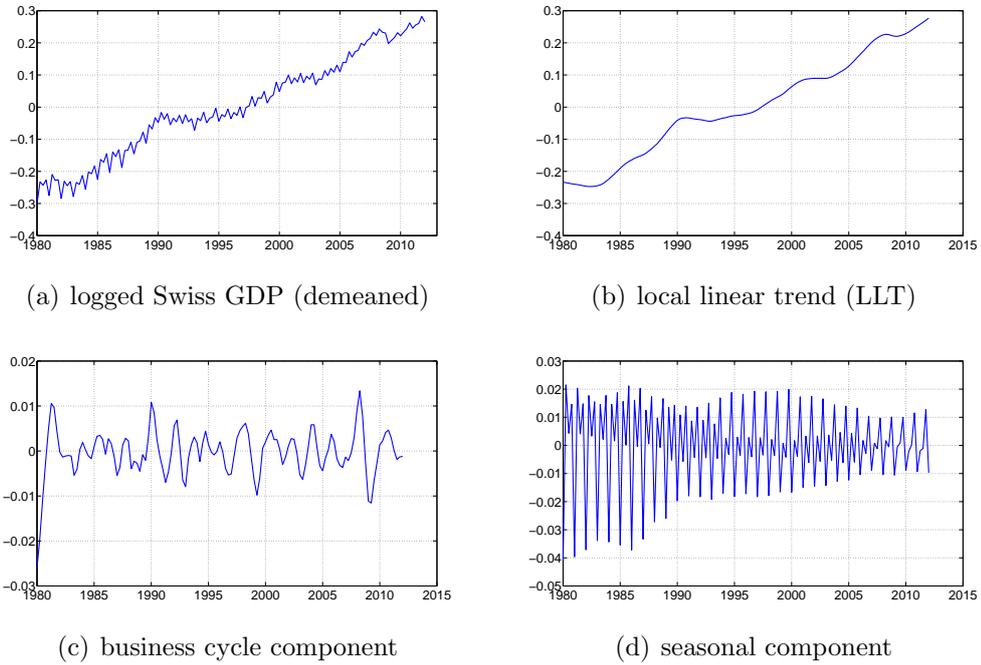


Figure 17.3: Components of the Basic Structural Model (BSM) for real GDP of Switzerland

seasonal component changes its character over time.

## 17.5 Exercises

**Exercise 17.5.1.** Consider the basic structural time series model for  $\{Y_t\}$ :

$$\begin{aligned} Y_t &= T_t + W_t, & W_t &\sim \text{WN}(0, \sigma_W^2) \\ T_t &= \delta_{t-1} + T_{t-1} + \varepsilon_t, & \varepsilon_t &\sim \text{WN}(0, \sigma_\varepsilon^2) \\ \delta_t &= \delta_{t-1} + \xi_t, & \xi_t &\sim \text{WN}(0, \sigma_\xi^2) \end{aligned}$$

where the error terms  $W_t$ ,  $\varepsilon_t$  and  $\xi_t$  are all uncorrelated with other at all leads and lags.

(i) Show that  $\{Y_t\}$  follows an  $\text{ARIMA}(0,2,2)$  process.

(ii) Compute the autocorrelation function of  $\{\Delta^2 Y_t\}$ .

**Exercise 17.5.2.** If the cyclical component of the basic structural model for  $\{Y_t\}$  is:

$$\begin{pmatrix} C_t \\ C_t^* \end{pmatrix} = \rho \begin{pmatrix} \cos \lambda_C & \sin \lambda_C \\ -\sin \lambda_C & \cos \lambda_C \end{pmatrix} \begin{pmatrix} C_{t-1} \\ C_{t-1}^* \end{pmatrix} + \begin{pmatrix} V_{1,t}^{(C)} \\ V_{2,t}^{(C)} \end{pmatrix}$$

where  $\{V_{1,t}^{(C)}\}$  and  $\{V_{2,t}^{(C)}\}$  are mutually uncorrelated white-noise processes.

- (i) Show that  $\{C_t\}$  follows an ARMA(2,1) process with ACF given by  $\gamma_h(h) = \rho^h \cos \lambda_C h$ .

**Exercise 17.5.3.** Write the ARMA( $p, q$ ) process  $Y_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$  as a state space model such that the state vector  $X_t$  is given by:

$$X_t = \begin{pmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p} \\ Z_{t-1} \\ Z_{t-2} \\ \vdots \\ Z_{t-q} \end{pmatrix}.$$

**Exercise 17.5.4.** Show that  $X_t$  and  $Y_t$  have a unique stationary and causal solution if all eigenvalues of  $F$  are absolutely strictly smaller than one. Use the results from Section 12.3.

**Exercise 17.5.5.** Find the Kalman filter equations for the following system:

$$\begin{aligned} X_t &= \phi X_{t-1} + w_t \\ Y_t &= \lambda X_t + v_t \end{aligned}$$

where  $\lambda$  and  $\phi$  are scalars and where

$$\begin{pmatrix} v_t \\ w_t \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & \sigma_{vw} \\ \sigma_{vw} & \sigma_w^2 \end{pmatrix} \right).$$

**Exercise 17.5.6.** Consider the state space model of an AR(1) process with measurement error analyzed in Section 17.2 :

$$\begin{aligned} X_{t+1} &= \phi X_t + v_{t+1}, & v_t &\sim \text{IIDN}(0, \sigma_v^2) \\ Y_t &= X_t + w_t, & w_t &\sim \text{IIDN}(0, \sigma_w^2). \end{aligned}$$

For simplicity assume that  $|\phi| < 1$ .

- (i) Show that  $\{Y_t\}$  is an ARMA(1,1) process given by  $Y_t - \phi Y_{t-1} = Z_t + \theta Z_{t-1}$  with  $Z_t \sim \text{WN}(0, \sigma_Z^2)$ .

- (ii) Show that the parameters of the state space,  $\phi, \sigma_v^2, \sigma_w^2$  and those of the ARMA(1,1) model are related by the equation

$$\begin{aligned}\theta\sigma_z^2 &= -\phi\sigma_w^2 \\ \frac{1}{1+\theta^2} &= \frac{-\phi\sigma_w^2}{\sigma_v^2 + (1+\phi^2)\sigma_w^2}\end{aligned}$$

- (iii) Why is there an identification problem?

# Appendix A

## Complex Numbers

The simple quadratic equation  $x^2 + 1 = 0$  has no solution in the field of real numbers,  $\mathbb{R}$ . Thus, it is necessary to envisage the larger field of complex numbers  $\mathbb{C}$ . A complex number  $z$  is an ordered pair  $(a, b)$  of real numbers where ordered means that we regard  $(a, b)$  and  $(b, a)$  as distinct if  $a \neq b$ . Let  $x = (a, b)$  and  $y = (c, d)$  be two complex numbers. Then we endow the set of complex numbers with an addition and a multiplication in the following way:

$$\begin{aligned} \text{addition:} \quad & x + y = (a, b) + (c, d) = (a + c, b + d) \\ \text{multiplication:} \quad & xy = (a, b)(c, d) = (ac - bd, ad + bc). \end{aligned}$$

These two operations will turn  $\mathbb{C}$  into a field where  $(0, 0)$  and  $(1, 0)$  play the role of 0 and 1.<sup>1</sup> The real numbers  $\mathbb{R}$  are embedded into  $\mathbb{C}$  because we identify any  $a \in \mathbb{R}$  with  $(a, 0) \in \mathbb{C}$ .

The number  $i = (0, 1)$  is of special interest. It solves the equation  $x^2 + 1 = 0$ , i.e.  $i^2 = -1$ . The other solution being  $-i = (0, -1)$ . Thus any complex number  $(a, b)$  may be written as  $(a, b) = a + ib$  where  $a, b$  are arbitrary real numbers.<sup>2</sup>

---

<sup>1</sup>Subtraction and division can be defined accordingly:

$$\begin{aligned} \text{subtraction:} \quad & (a, b) - (c, d) = (a - c, b - d) \\ \text{division:} \quad & (a, b)/(c, d) = \frac{(ac + bd, bc - ad)}{(c^2 + d^2)}, \quad c^2 + d^2 \neq 0. \end{aligned}$$

<sup>2</sup>A more detailed introduction of complex numbers can be found in Rudin (1976) or any other mathematics textbook.

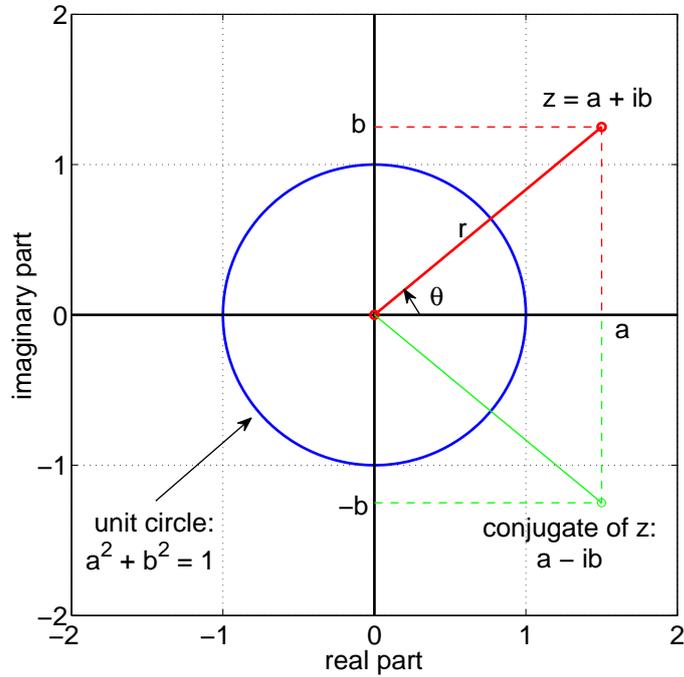


Figure A.1: Representation of a complex number

An element  $z$  in this field can be represented in two ways:

$$\begin{array}{ll} z = a + ib & \text{Cartesian coordinates} \\ = re^{i\theta} = r(\cos \theta + i \sin \theta) & \text{polar coordinates.} \end{array}$$

In the representation in Cartesian coordinates  $a = \operatorname{Re}(z) = \Re(z)$  is called the *real part* whereas  $b = \operatorname{Im}(z) = \Im(z)$  is called the *imaginary part* of  $z$ .

A complex number  $z$  can be viewed as a point in the two-dimensional Cartesian coordinate system with coordinates  $(a, b)$ . This geometric interpretation is represented in Figure A.1.

The *absolute value* or *modulus* of  $z$ , denoted by  $|z|$ , is given by  $r = \sqrt{a^2 + b^2}$ . Thus, the absolute value is nothing but the distance of  $z$  viewed as a point in the complex plane (the two-dimensional Cartesian coordinate system) to the origin (see Figure A.1).  $\theta$  denotes the angle to the positive real axis (x-axis) measured in radians. It is denoted by  $\theta = \arg z$ . It holds that  $\tan \theta = \frac{b}{a}$ . Finally, the conjugate of  $z$ , denoted by  $\bar{z}$ , is defined by  $\bar{z} = a - ib$ .

Setting  $r = 1$  and  $\theta = \pi$ , gives the following famous formula:

$$e^{i\pi} + 1 = (\cos \pi + i \sin \pi) + 1 = -1 + 1 = 0.$$

This formula relates the most famous numbers in mathematics.

From the definition of complex numbers in polar coordinates, we immediately derive the following implications:

$$\begin{aligned}\cos \theta &= \frac{e^{i\theta} + e^{-i\theta}}{2} = \frac{a}{r}, \\ \sin \theta &= \frac{e^{i\theta} - e^{-i\theta}}{2i} = \frac{b}{r}.\end{aligned}$$

Further implications are de Moivre's formula and Pythagoras' theorem (see Figure A.1):

$$\begin{array}{ll}\text{de Moivre's formula} & (re^{i\theta})^n = r^n e^{in\theta} = r^n (\cos n\theta + i \sin n\theta) \\ \text{Pythagoras' theorem} & 1 = e^{i\theta} e^{-i\theta} = (\cos \theta + i \sin \theta)(\cos \theta - i \sin \theta) \\ & = \cos^2 \theta + \sin^2 \theta\end{array}$$

From Pythagoras' theorem it follows that  $r^2 = a^2 + b^2$ . The representation in polar coordinates allows to derive many trigonometric formulas.

Consider the polynomial  $\Phi(z) = \phi_0 - \phi_1 z - \phi_2 z^2 - \dots - \phi_p z^p$  of order  $p \geq 1$  with  $\phi_0 = 1$ .<sup>3</sup> The *fundamental theorem of algebra* then states that every polynomial of order  $p \geq 1$  has exactly  $p$  roots in the field of complex numbers. Thus, the field of complex numbers is algebraically complete. Denote these roots by  $\lambda_1, \dots, \lambda_p$ , allowing that some roots may appear several times. The polynomial can then be factorized as

$$\Phi(z) = (1 - \lambda_1^{-1}z) (1 - \lambda_2^{-1}z) \dots (1 - \lambda_p^{-1}z).$$

This expression is well-defined because the assumption of a nonzero constant ( $\phi_0 = 1 \neq 0$ ) excludes the possibility of roots equal to zero. If we assume that the coefficients  $\phi_j$ ,  $j = 0, \dots, p$ , are real numbers, the complex roots appear in conjugate pairs. Thus if  $z = a + ib$ ,  $b \neq 0$ , is a root then  $\bar{z} = a - ib$  is also a root.

---

<sup>3</sup>The notation with “ $-\phi_j z^j$ ” instead of “ $\phi_j z^j$ ” was chosen to conform to the notation of AR-models.



# Appendix B

## Linear Difference Equations

Linear difference equations play an important role in time series analysis. We therefore summarize the most important results.<sup>1</sup> Consider the following linear difference equation of order  $p$  with constant coefficients. This equation is defined by the recursion:

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p}, \quad \phi_p \neq 0, t \in \mathbb{Z}.$$

Thereby  $\{X_t\}$  represents a sequence of real numbers and  $\phi_1, \dots, \phi_p$  are  $p$  constant coefficients. The above difference equation is called *homogeneous* because it involves no other variable than  $X_t$ . A solution to this equation is a function  $F : \mathbb{Z} \rightarrow \mathbb{R}$  such that its values  $F(t)$  or  $F_t$  reduce the difference equation to an identity.

It is easy to see that if  $\{X_t^{(1)}\}$  and  $\{X_t^{(2)}\}$  are two solutions than  $\{c_1 X_t^{(1)} + c_2 X_t^{(2)}\}$ , for any  $c_1, c_2 \in \mathbb{R}$ , is also a solution. The set of solutions is therefore a linear space (vector space).

**Definition B.1.** A set of solutions  $\{\{X_t^{(1)}\}, \dots, \{X_t^{(m)}\}\}$ ,  $m \leq p$ , is called linearly independent if

$$c_1 X_t^{(1)} + \dots + c_m X_t^{(m)} = 0, \quad \text{for } t = 0, 1, \dots, p-1$$

implies that  $c_1 = \dots = c_m = 0$ . Otherwise we call the set linearly dependent.

Given arbitrary starting values  $x_0, \dots, x_{p-1}$  for  $X_0, \dots, X_{p-1}$ , the difference equation determines all further through the recursion:

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} \quad t = p, p+1, \dots$$

---

<sup>1</sup>For more detailed presentations see Agarwal Agarwal (2000), Elaydi Elaydi (2005) or Neusser Neusser (2009).

Similarly for  $X_t$  mit  $t = -1, -2, \dots$ . Suppose we have  $p$  linearly independent solutions  $\{\{X_t^{(1)}\}, \dots, \{X_t^{(p)}\}\}$  then there exists exactly  $p$  numbers  $c_1, \dots, c_p$  such that the solution

$$X_t = c_1 X_t^{(1)} + c_2 X_t^{(2)} + \dots + c_p X_t^{(p)}$$

is compatible with arbitrary starting values  $x_0, \dots, x_{p-1}$ . These starting values then determine uniquely all values of the sequence  $\{X_t\}$ . Thus  $\{X_t\}$  is the only solution compatible with starting values. The goal therefore consists in finding  $p$  linearly independent solutions.

We guess that the solutions are of the form  $X_t = z^{-t}$  where  $z$  may be a complex number. If this guess is right then we must have for  $t = 0$ :

$$1 - \phi_1 z - \dots - \phi_p z^p = 0.$$

This equation is called the *characteristic equation*.<sup>2</sup> Thus  $z$  must be a root of the polynomial  $\Phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ . From the fundamental theorem of algebra we know that there are exactly  $p$  roots in the field of complex numbers. Denote these roots by  $z_1, \dots, z_p$ .

Suppose that these roots are different from each other. In this case  $\{\{z_1^{-t}\}, \dots, \{z_p^{-t}\}\}$  constitutes a set of  $p$  linearly independent solutions. To show this it is sufficient to verify that the determinant of the matrix

$$W = \begin{pmatrix} 1 & 1 & \dots & 1 \\ z_1^{-1} & z_2^{-1} & \dots & z_p^{-1} \\ z_1^{-2} & z_2^{-2} & \dots & z_p^{-2} \\ \vdots & \vdots & & \vdots \\ z_1^{-p+1} & z_2^{-p+1} & \dots & z_p^{-p+1} \end{pmatrix}$$

is different from zero. This determinant is known as Vandermonde's determinant and is equal to  $\det W = \prod_{1 \leq i < j \leq p} (z_i - z_j)$ . This determinant is clearly different from zero because the roots are different from each other. The general solution to the difference equation therefore is

$$X_t = c_1 z_1^{-t} + \dots + c_p z_p^{-t} \tag{B.1}$$

where the constants  $c_1, \dots, c_p$  are determined from the starting values (initial conditions).

In the case where some roots of the characteristic polynomial are equal, the general solution becomes more involved. Let  $z_1, \dots, z_r$ ,  $r < p$ , be the

---

<sup>2</sup>Sometimes one can find  $z^p - \phi_1 z^{p-1} - \dots - \phi_p = 0$  as the characteristic equation. The roots of the two characteristic equations are then reciprocal to each other.

roots which are different from each other and denote their corresponding multiplicities by  $m_1, \dots, m_r$ . It holds that  $\sum_{j=1}^r m_j = p$ . The general solution is then given by

$$X_t = \sum_{j=1}^r (c_{j0} + c_{j1}t + \dots + c_{jm_j-1}t^{m_j-1}) z_j^{-t} \quad (\text{B.2})$$

where the constants  $c_{ji}$  are again determined from the starting values (initial conditions).



# Appendix C

## Stochastic Convergence

This appendix presents the relevant concepts and theorems from probability theory. The reader interested in more details should consult corresponding textbooks, for example Billingsley (1986), Brockwell and Davis (1991), Hogg and Craig (1995), or Kallenberg (2002) among many others.

In the following, all real random variables or random vectors  $X$  are defined with respect to some probability space  $(\Omega, \mathfrak{A}, \mathbf{P})$ . Thereby,  $\Omega$  denotes an arbitrary space with  $\sigma$ -field  $\mathfrak{A}$  and probability measure  $\mathbf{P}$ . A random variable, respectively random vector,  $X$  is then defined as a measurable function from  $\Omega$  to  $\mathbb{R}$ , respectively  $\mathbb{R}^n$ . The probability space  $\Omega$  plays no role as it is introduced just for the sake of mathematical rigor. The interest rather focuses on the distributions induced by  $\mathbf{P} \circ X^{-1}$ .

We will make use of the following important inequalities.

**Theorem C.1** (Cauchy-Bunyakovskii-Schwarz inequality). *For any two random variables  $X$  and  $Y$ ,*

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}.$$

*The equality holds if and only if  $X = \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}Y$ .*

**Theorem C.2** (Minkowski's inequality). *Let  $X$  and  $Y$  be two random variables with  $\mathbb{E}|X|^2 < \infty$  and  $\mathbb{E}|Y|^2 < \infty$ , then*

$$(\mathbb{E}|X + Y|^2)^{1/2} \leq (\mathbb{E}|X|^2)^{1/2} + (\mathbb{E}|Y|^2)^{1/2}.$$

**Theorem C.3** (Chebyshev's inequality). *If  $\mathbb{E}|X|^r < \infty$  for  $r \geq 0$  then for every  $r \geq 0$  and any  $\varepsilon > 0$*

$$\mathbf{P}[|X| \geq \varepsilon] \leq \varepsilon^{-r} \mathbb{E}|X|^r.$$

**Theorem C.4** (Borel-Cantelli lemma). *Let  $A_1, A_2, \dots \in \mathfrak{A}$  be an infinite sequence of events in some probability space  $(\Omega, \mathfrak{A}, \mathbf{P})$  such that  $\sum_{k=1}^{\infty} \mathbf{P}(A_k) < \infty$ . Then,  $\mathbf{P}\{A_k \text{ i.o.}\} = 0$ . The event  $\{A_k \text{ i.o.}\}$  is defined by  $\{A_k \text{ i.o.}\} = \limsup_k \{A_k\} = \bigcap_{k=1}^{\infty} \bigcup_{j=k}^{\infty} A_j$  where i.o. stands for infinitely often.*

On several occasions it is necessary to evaluate the limit of a sequence of random variables. In probability theory several concepts of convergence are discussed: *almost sure convergence*, *convergence in probability*, *convergence in  $r$ -th mean* (*convergence in quadratic mean*), *convergence in distribution*. We only give definitions and the most important theorems leaving an in-depth discussion to the relevant literature. Although not explicitly mentioned, many of the theorem below also hold in an analogous way in a multidimensional context.

**Definition C.1** (Almost Sure Convergence). *For random variables  $X$  and  $\{X_t\}$  defined on the same probability space  $(\Omega, \mathfrak{A}, \mathbf{P})$ , we say that  $\{X_t\}$  converges almost surely or with probability one to  $X$  if*

$$\mathbf{P} \left\{ \omega \in \Omega : \lim_{t \rightarrow \infty} X_t(\omega) = X(\omega) \right\} = 1.$$

*This fact is denoted by  $X_t \xrightarrow{\text{a.s.}} X$  or  $\lim X_t = X$  a.s.*

**Theorem C.5** (Kolmogorov's Strong Law of Large Numbers (SLLN)). *Let  $X, X_1, X_2, \dots$  be identically and independently distributed random variables. Then, the arithmetic average  $\bar{X}_T = \frac{1}{T} \sum_{t=1}^T X_t$  converges almost surely to  $\mathbb{E}X$  if and only if  $\mathbb{E}|X| < \infty$ .*

**Definition C.2** (Convergence in Probability). *For random variables  $X$  and  $\{X_t\}$  defined on the same probability space, we say that  $\{X_t\}$  converges in probability to  $X$  if*

$$\lim_{t \rightarrow \infty} \mathbf{P}[|X_t - X| > \varepsilon] = 0 \quad \text{for all } \varepsilon > 0.$$

*This fact is denoted by  $X_t \xrightarrow{p} X$  or  $\text{plim } X_t = X$ .*

**Remark C.1.** *If  $X$  and  $\{X_t\}$  are real valued random vectors, we replace the absolute value in the definition above by the Euclidean norm  $\|\cdot\|$ . This is, however, equivalent to saying that every component  $X_{it}$  converges in probability to  $X_i$ , the  $i$ -th component of  $X$ .*

**Definition C.3** (Convergence in  $r$ -th mean). *A sequence  $\{X_t\}$  of random variables converges in  $r$ -th mean to a random variable  $X$  if*

$$\lim_{t \rightarrow \infty} \mathbb{E}(|X_t - X|^r) = 0 \quad \text{for } r > 0.$$

We denote this fact by  $X_t \xrightarrow{r} X$ . If  $r = 1$  we say that the sequence converges absolutely; and if  $r = 2$  we say that the sequence converges in mean square which is denoted by  $X_t \xrightarrow{m.s.} X$ .

**Remark C.2.** In the case  $r = 2$ , the corresponding definition for random vectors is

$$\lim_{t \rightarrow \infty} \mathbb{E}(\|X_t - X\|^2) = \lim_{t \rightarrow \infty} \mathbb{E}(X_t - X)'(X_t - X) = 0.$$

**Theorem C.6** (Riesz-Fisher). Let  $\{X_t\}$  be a sequence of random variables such  $\sup_t \mathbb{E}|X_t|^2 < \infty$ . Then there exists a random variable  $X$  with  $\mathbb{E}|X|^2 < \infty$  such that

$$X_t \xrightarrow{m.s.} X \quad \text{if and only if} \quad \mathbb{E}|X_t - X_s|^2 \rightarrow 0 \quad \text{for } t, s \rightarrow \infty.$$

This version of the Riesz-Fisher theorem provides a condition, known as the Cauchy criterion, which is often easier to verify when the limit is unknown.

**Definition C.4** (Convergence in Distribution). A sequence  $\{X_t\}$  of random vectors with corresponding distribution functions  $\{F_{X_t}\}$  converges in distribution, if there exists an random vector  $X$  with distribution function  $F_X$  such that

$$\lim_{t \rightarrow \infty} F_{X_t}(x) = F_X(x) \quad \text{for all } x \in \mathcal{C}$$

where  $\mathcal{C}$  denotes the set of points for which  $F_X(x)$  is continuous. We denote this fact by  $X_t \xrightarrow{d} X$ .

Note that, in contrast to the previously mentioned modes of convergence, convergence in distribution does not require that all random vectors are defined on the same probability space. The convergence in distribution states that, for large enough  $t$ , the distribution of  $X_t$  can be approximated by the distribution of  $X$ .

The following Theorem relates the four convergence concepts.

**Theorem C.7.** (i) If  $X_t \xrightarrow{a.s.} X$  then  $X_t \xrightarrow{p} X$ .

(ii) If  $X_t \xrightarrow{p} X$  then there exists a subsequence  $\{X_{t_n}\}$  such that  $X_{t_n} \xrightarrow{a.s.} X$ .

(iii) If  $X_t \xrightarrow{r} X$  then  $X_t \xrightarrow{p} X$  by Chebyshev's inequality (Theorem C.3).

(iv) If  $X_t \xrightarrow{p} X$  then  $X_t \xrightarrow{d} X$ .

(v) If  $X$  is a fixed constant, then  $X_t \xrightarrow{d} X$  implies  $X_t \xrightarrow{p} X$ . Thus, the two concepts are equivalent under this assumption.

These facts can be summarized graphically:

$$\begin{array}{c}
 X_{t_n} \xrightarrow{a.s.} X \\
 \uparrow \\
 X_t \xrightarrow{a.s.} X \implies X_t \xrightarrow{p} X \implies X_t \xrightarrow{d} X \\
 \uparrow \\
 X_t \xrightarrow{r} X
 \end{array}$$

A further useful theorem is:

**Theorem C.8.** *If  $\mathbb{E}X_t \rightarrow \mu$  and  $\mathbb{V}X_t \rightarrow 0$  then  $X_t \xrightarrow{m.s.} \mu$  and consequently  $X_t \xrightarrow{p} \mu$ .*

**Theorem C.9** (Continuous Mapping Theorem). *For any continuous function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  and random vectors  $\{X_t\}$  and  $X$  defined on some probability space, the following implications hold:*

- (i)  $X_t \xrightarrow{a.s.} X$  implies  $f(X_t) \xrightarrow{a.s.} f(X)$ .
- (ii)  $X_t \xrightarrow{p} X$  implies  $f(X_t) \xrightarrow{p} f(X)$ .
- (iii)  $X_t \xrightarrow{d} X$  implies  $f(X_t) \xrightarrow{d} f(X)$ .

An important application of the Continuous Mapping Theorem is the so-called Delta method which can be used to approximate the distribution of  $f(X_t)$  (see Appendix E).

A further useful result is given by:

**Theorem C.10** (Slutzky's Lemma). *Let  $\{X_t\}$  and  $\{Y_t\}$  be two sequences of random vectors such that  $X_t \xrightarrow{d} X$  and  $Y_t \xrightarrow{d} c$ ,  $c$  constant, then*

- (i)  $X_t + Y_t \xrightarrow{d} X + c$ ,
- (ii)  $Y_t' X_t \xrightarrow{d} c' X$ .
- (iii)  $X_t/Y_t \xrightarrow{d} X/c$  if  $c$  is a nonzero scalar.

Like the (cumulative) distribution function, the characteristic function provides an alternative way to describe a random variable.

**Definition C.5** (Characteristic function). *The characteristic function of a real random vector  $X$ , denoted by  $\varphi_X$ , is defined by*

$$\varphi_X(s) = \mathbb{E}e^{i\lambda'X}, \quad \lambda \in \mathbb{R}^n,$$

where  $i$  is the imaginary unit.

If, for example,  $X \sim N(\mu, \sigma^2)$ , then  $\varphi_X(s) = \exp(is\mu - \frac{1}{2}\sigma^2 s^2)$ . The characteristic function uniquely determines the distribution of  $X$ . Thus, if two random variables have the same characteristic function, they have the same distribution. Moreover, convergence in distribution is equivalent to convergence of the corresponding characteristic functions.

**Theorem C.11** (Convergence of characteristic functions, Lévy). *Let  $\{X_t\}$  be a sequence of real random variables with corresponding characteristic functions  $\varphi_{X_t}$  then*

$$X_t \xrightarrow{d} X \quad \text{if and only if} \quad \lim_{t \rightarrow \infty} \varphi_{X_t}(\lambda) = \varphi_X(\lambda), \quad \text{for all } \lambda \in \mathbb{R}^n.$$

In many cases the limiting distribution is a normal distribution. In which case one refers to the asymptotic normality.

**Definition C.6** (Asymptotic Normality). *A sequence of random variables  $\{X_t\}$  with “means”  $\mu_t$  and “variances”  $\sigma_t^2 > 0$  is said to be asymptotically normally distributed if*

$$\sigma_t^{-1}(X_t - \mu_t) \xrightarrow{d} X \sim N(0, 1).$$

Note that the definition does not require that  $\mu_t = \mathbb{E}X_t$  nor that  $\sigma_t^2 = \mathbb{V}(X_t)$ . Asymptotic normality is obtained if the  $X_t$ 's are identically and independently distributed with constant mean and variance. In this case the Central Limit Theorem (CLT) holds.

**Theorem C.12** (Central Limit Theorem). *Let  $\{X_t\}$  be a sequence of identically and independently distributed random variables with constant mean  $\mu$  and constant variance  $\sigma^2$  then*

$$\sqrt{T} \frac{\bar{X}_T - \mu}{\sigma} \xrightarrow{d} N(0, 1),$$

where  $\bar{X}_T = T^{-1} \sum_{t=1}^T X_t$  is the arithmetic average.

It is possible to relax the assumption of identically distributed variables in various ways so that there exists a variety of CLT's in the literature. For our purpose it is especially important to relax the independence assumption. A natural way to do this is by the notion of *m-dependence*.

**Definition C.7** (m-Dependence). *A strictly stationary random process  $\{X_t\}$  is called m-dependent for some nonnegative integer  $m$  if and only if the two sets of random variables  $\{X_\tau, \tau \leq t\}$  and  $\{X_\tau, \tau \geq t+m+1\}$  are independent.*

Note that for such processes  $\Gamma(j) = 0$  for  $j > m$ . This type of dependence allows to prove the following generalized Central Limit Theorem (see Brockwell and Davis, 1991).

**Theorem C.13** (CLT for  $m$ -dependent processes). *Let  $\{X_t\}$  be a strictly stationary mean zero  $m$ -dependent process with autocovariance function  $\Gamma(h)$  such that  $\mathbf{V}_m = \sum_{h=-m}^m \Gamma(h) \neq 0$  then*

- (i)  $\lim_{T \rightarrow \infty} T\mathbb{V}(\bar{X}_T) = \mathbf{V}_m$  and
- (ii)  $\sqrt{T}\bar{X}_T$  is asymptotically normal  $N(0, \mathbf{V}_m)$ .

Often it is difficult to derive the asymptotic distribution of  $\{X_t\}$  directly. This situation can be handled by approximating the original process  $\{X_t\}$  by a process  $\{X_t^{(m)}\}$  which is easier to handle in terms of its asymptotic distribution and where the precision of the approximation can be “tuned” by the parameter  $m$ .

**Theorem C.14** (Basis Approximation Theorem). *Let  $\{X_t\}$  and  $\{X_t^{(m)}\}$  be two random vectors process such that*

- (i)  $X_t^{(m)} \xrightarrow{d} X^{(m)}$  as  $t \rightarrow \infty$  for each  $m = 1, 2, \dots$ ,
- (ii)  $X^{(m)} \xrightarrow{d} X$  as  $m \rightarrow \infty$ , and
- (iii)  $\lim_{m \rightarrow \infty} \limsup_{t \rightarrow \infty} \mathbf{P}[|X_t - X_t^{(m)}| > \epsilon] = 0$  for every  $\epsilon > 0$ .

Then

$$X_t \xrightarrow{d} X \quad \text{as } t \rightarrow \infty.$$

# Appendix D

## Beveridge-Nelson Decomposition

The Beveridge-Nelson decomposition proves to be an indispensable tool throughout this book. Based on the seminal paper by Phillips and Solo (1992), we prove the following Theorem for matrix polynomials where  $\|\cdot\|$  denotes the matrix norm (see Definition 10.6 in Chapter 10). The univariate version is then a special case with the absolute value replacing the norm.

**Theorem D.1.** *Any a lag polynomial  $\Psi(L) = \sum_{j=0}^{\infty} \Psi_j L^j$  where  $\Psi_j$  are  $n \times n$  matrices with  $\Psi_0 = I_n$  can be represented by*

$$\Psi(L) = \Psi(1) - (I_n - L)\tilde{\Psi}(L) \quad (\text{D.1})$$

where  $\tilde{\Psi}(L) = \sum_{j=0}^{\infty} \tilde{\Psi}_j L^j$  with  $\tilde{\Psi}_j = \sum_{i=j+1}^{\infty} \Psi_i$ . Moreover,

$$\sum_{j=1}^{\infty} j^2 \|\Psi_j\|^2 < \infty \quad \text{implies} \quad \sum_{j=0}^{\infty} \|\tilde{\Psi}_j\|^2 < \infty \quad \text{and} \quad \|\Psi(1)\| < \infty.$$

*Proof.* The first part of the Theorem is obtained by the algebraic manipulations below:

$$\begin{aligned} \Psi(L) - \Psi(1) &= I_n + \Psi_1 L + \Psi_2 L^2 + \dots \\ &\quad - I_n - \Psi_1 - \Psi_2 - \dots \\ &= \Psi_1(L - I_n) + \Psi_2(L^2 - I_n) + \Psi_3(L^3 - I_n) + \dots \\ &= (L - I_n)\Psi_1 + (L - I_n)\Psi_2(L + I_n) + (L - I_n)\Psi_3(L^2 + L + I_n) + \dots \\ &= -(I_n - L)\underbrace{(\Psi_1 + \Psi_2 + \Psi_3 + \dots)}_{\tilde{\Psi}_0} + \\ &\quad \underbrace{(\Psi_2 + \Psi_3 + \dots)}_{\tilde{\Psi}_1} L + \underbrace{(\Psi_3 + \dots)}_{\tilde{\Psi}_2} L^2 + \dots \end{aligned}$$

Taking any  $\delta \in (1/2, 1)$ , the second part of the Theorem follows from

$$\begin{aligned}
\sum_{j=0}^{\infty} \|\tilde{\Psi}_j\|^2 &= \sum_{j=0}^{\infty} \left\| \sum_{i=j+1}^{\infty} \Psi_j \right\|^2 \leq \sum_{j=0}^{\infty} \left( \sum_{i=j+1}^{\infty} \|\Psi_j\| \right)^2 \\
&= \sum_{j=0}^{\infty} \left( \sum_{i=j+1}^{\infty} i^\delta \|\Psi_j\| i^{-\delta} \right)^2 \\
&\leq \sum_{j=0}^{\infty} \left( \sum_{i=j+1}^{\infty} i^{2\delta} \|\Psi_j\|^2 \right) \left( \sum_{i=j+1}^{\infty} i^{-2\delta} \right) \\
&\leq (2\delta - 1)^{-1} \sum_{j=0}^{\infty} \left( \sum_{i=j+1}^{\infty} i^{2\delta} \|\Psi_i\|^2 \right) j^{1-2\delta} \\
&= (2\delta - 1)^{-1} \sum_{i=0}^{\infty} \left( \sum_{j=0}^{i-1} j^{1-2\delta} \right) i^{2\delta} \|\Psi_i\|^2 \\
&\leq [(2\delta - 1)(2 - 2\delta)]^{-1} \sum_{j=0}^{\infty} j^{2\delta} \|\Psi_i\|^2 j^{2-2\delta} \\
&= [(2\delta - 1)(2 - 2\delta)]^{-1} \sum_{j=0}^{\infty} j^2 \|\Psi_i\|^2 < \infty.
\end{aligned}$$

The first inequality follows from the triangular inequality for the norm. The second inequality is Hölder's inequality (see, for example, Naylor and Sell, 1982, p. 548) with  $p = q = 2$ . The third and the fourth inequality follow from the Lemma below. The last inequality, finally, follows from the assumption. The last assertion follows from

$$\begin{aligned}
\|\Psi(1)\| &\leq \sum_{j=0}^{\infty} \|\Psi_j\| = \|I_n\| + \sum_{j=1}^{\infty} j \|\Psi_j\| j^{-1} \\
&\leq \|I_n\| + \left( \sum_{j=1}^{\infty} j^2 \|\Psi_j\|^2 \right)^{1/2} \left( \sum_{j=1}^{\infty} j^{-2} \right)^{1/2} < \infty.
\end{aligned}$$

The last inequality is again a consequence of Hölder's inequality. The summability assumption then guarantees the convergence of the first term in the product. Cauchy's condensation test finally establishes the convergence of the last term.  $\square$

**Lemma D.1.** *The following results are useful:*

$$(i) \text{ For any } b > 0, \sum_{i=j+1}^{\infty} i^{-1-b} \leq b^{-1} j^{-b}.$$

(ii) For any  $c \in (0, 1)$ ,  $\sum_{j=1}^i j^{c-1} \leq c^{-1}i^c$ .

*Proof.* Let  $k$  be a number greater than  $j$ , then  $k^{-1-b} \leq j^{-1-b}$  and

$$k^{-1-b} \leq \int_{k-1}^k j^{-1-b} dj = b^{-1}(k-1)^{-b} - b^{-1}k^{-b}.$$

This implies that  $\sum_{k=j+1}^{\infty} k^{-1-b} \leq b^{-1}j^{-b}$ . This proves part (i) by changing the summation index back from  $k$  to  $j$ . Similarly,  $k^{c-1} \leq j^{c-1}$  and

$$k^{c-1} \leq \int_{k-1}^k j^{c-1} dj = c^{-1}k^c - c^{-1}(k-1)^c.$$

Therefore  $\sum_{k=1}^i k^{c-1} \leq c^{-1}i^c$  which proves part (ii) by changing the summation index back from  $k$  to  $j$ .  $\square$

**Remark D.1.** An alternative common assumption is  $\sum_{j=1}^{\infty} j\|\Psi_j\| < \infty$ . It is, however, easy to see that this assumption is more restrictive as it implies the one assumed in the Theorem, but not vice versa. See Phillips and Solo (1992) for more details.



# Appendix E

## The Delta Method

It is often the case that it is possible to obtain an estimate  $\hat{\beta}_T$  of some parameter  $\beta$ , but that one is really interested in a function  $f$  of  $\beta$ . The Continuous Mapping Theorem then suggests to estimate  $f(\beta)$  by  $f(\hat{\beta}_T)$ . But then the question arises how the distribution of  $\hat{\beta}_T$  is related to the distribution of  $f(\hat{\beta}_T)$ .

Expanding the function into a first order Taylor approximation allows to derive the following theorem.

**Theorem E.1.** *Let  $\{\hat{\beta}_T\}$  be a  $K$ -dimensional sequence of random variables with the property  $\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} N(0, \Sigma)$  then*

$$\sqrt{T} \left( f(\hat{\beta}_T) - f(\beta) \right) \xrightarrow{d} N \left( 0, \nabla f(\beta) \Sigma \nabla f(\beta)' \right),$$

where  $f : \mathbb{R}^K \rightarrow \mathbb{R}^J$  is a continuously differentiable function with Jacobian matrix (matrix of first order partial derivatives)  $\nabla f(\beta) = \partial f(\beta) / \partial \beta'$ .

*Proof.* See Serfling (Serfling, 1980, 122-124). □

**Remark E.1.** *In the one-dimensional case where  $\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} N(0, \sigma^2)$  and  $f : \mathbb{R} \rightarrow \mathbb{R}$  the above theorem becomes:*

$$\sqrt{T} \left( f(\hat{\beta}_T) - f(\beta) \right) \xrightarrow{d} N \left( 0, [f'(\beta)]^2 \sigma^2 \right)$$

where  $f'(\beta)$  is the first derivative evaluated at  $\beta$ .

**Remark E.2.** *The  $J \times K$  Jacobian matrix of first order partial derivatives is defined as*

$$\nabla f(\beta) = \partial f(\beta) / \partial \beta' = \begin{pmatrix} \frac{\partial f_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial f_1(\beta)}{\partial \beta_K} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_J(\beta)}{\partial \beta_1} & \cdots & \frac{\partial f_J(\beta)}{\partial \beta_K} \end{pmatrix}.$$

**Remark E.3.** In most applications  $\beta$  is not known so that one evaluates the Jacobian matrix at  $\hat{\beta}_T$ .

**Example: univariate**

Suppose we have obtained an estimate of  $\beta$  equal to  $\hat{\beta} = 0.6$  together with an estimate for its variance  $\hat{\sigma}_{\hat{\beta}}^2 = 0.2$ . We can then approximate the variance of  $f(\hat{\beta}) = 1/\hat{\beta} = 1.667$  by

$$\hat{V}(f(\hat{\beta})) = \left[ \frac{-1}{\hat{\beta}^2} \right]^2 \hat{\sigma}_{\hat{\beta}}^2 = 1.543.$$

**Example: multivariate**

In the process of computing the impulse response function of a VAR(1) model with  $\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix}$  one has to calculate  $\Psi_2 = \Phi^2$ . If we stack all coefficients of  $\Phi$  into a vector  $\beta = \text{vec}(\Phi) = (\phi_{11}, \phi_{21}, \phi_{12}, \phi_{22})'$  then we get:

$$f(\beta) = \text{vec} \Psi_2 = \text{vec} \Phi^2 = \begin{pmatrix} \psi_{11}^{(2)} \\ \psi_{21}^{(2)} \\ \psi_{12}^{(2)} \\ \psi_{22}^{(2)} \end{pmatrix} = \begin{pmatrix} \phi_{11}^2 + \phi_{12}\phi_{21} \\ \phi_{11}\phi_{21} + \phi_{21}\phi_{22} \\ \phi_{11}\phi_{12} + \phi_{12}\phi_{22} \\ \phi_{12}\phi_{21} + \phi_{22}^2 \end{pmatrix},$$

where  $\Psi_2 = [\psi_{ij}^{(2)}]$ . The Jacobian matrix then becomes:

$$\nabla f(\beta) = \begin{pmatrix} \frac{\partial \psi_{11}^{(2)}}{\partial \phi_{11}} & \frac{\partial \psi_{11}^{(2)}}{\partial \phi_{21}} & \frac{\partial \psi_{11}^{(2)}}{\partial \phi_{12}} & \frac{\partial \psi_{11}^{(2)}}{\partial \phi_{22}} \\ \frac{\partial \psi_{21}^{(2)}}{\partial \phi_{11}} & \frac{\partial \psi_{21}^{(2)}}{\partial \phi_{21}} & \frac{\partial \psi_{21}^{(2)}}{\partial \phi_{12}} & \frac{\partial \psi_{21}^{(2)}}{\partial \phi_{22}} \\ \frac{\partial \psi_{12}^{(2)}}{\partial \phi_{11}} & \frac{\partial \psi_{12}^{(2)}}{\partial \phi_{21}} & \frac{\partial \psi_{12}^{(2)}}{\partial \phi_{12}} & \frac{\partial \psi_{12}^{(2)}}{\partial \phi_{22}} \\ \frac{\partial \psi_{22}^{(2)}}{\partial \phi_{11}} & \frac{\partial \psi_{22}^{(2)}}{\partial \phi_{21}} & \frac{\partial \psi_{22}^{(2)}}{\partial \phi_{12}} & \frac{\partial \psi_{22}^{(2)}}{\partial \phi_{22}} \end{pmatrix} = \begin{pmatrix} 2\phi_{11} & \phi_{12} & \phi_{21} & 0 \\ \phi_{21} & \phi_{11} + \phi_{22} & 0 & \phi_{21} \\ \phi_{12} & 0 & \phi_{11} + \phi_{22} & \phi_{12} \\ 0 & \phi_{12} & \phi_{21} & 2\phi_{22} \end{pmatrix}.$$

In Section 15.4.4 we obtained the following estimate for a VAR(1) model for  $\{X_t\} = \{(\ln(A_t), \ln(S_t))'\}$ :

$$X_t = \hat{c} + \hat{\Phi}X_{t-1} + \hat{Z}_t = \begin{pmatrix} -0.141 \\ 0.499 \end{pmatrix} + \begin{pmatrix} 0.316 & 0.640 \\ -0.202 & 1.117 \end{pmatrix} X_{t-1} + \hat{Z}_t.$$

The estimated covariance matrix of  $\text{vec} \hat{\Phi}$ ,  $\hat{V}(\text{vec} \hat{\Phi})$ , was:

$$\hat{V}(\hat{\beta}) = \hat{V}(\text{vec} \hat{\Phi}) = \begin{pmatrix} 0.0206 & 0.0069 & -0.0201 & -0.0067 \\ 0.0069 & 0.0068 & -0.0067 & -0.0066 \\ -0.0201 & -0.0067 & 0.0257 & 0.0086 \\ -0.0067 & -0.0066 & 0.0086 & 0.0085 \end{pmatrix}.$$

We can then approximate the variance of  $f(\hat{\beta}) = \text{vec}(\hat{\Phi}^2)$  by

$$\hat{V}(f(\text{vec } \hat{\Phi})) = \hat{V}(\text{vec } \hat{\Phi}^2) = \nabla f(\text{vec } \Phi)|_{\Phi=\hat{\Phi}} \hat{V}(\text{vec } \hat{\Phi}) \nabla f(\text{vec } \Phi)'|_{\Phi=\hat{\Phi}}.$$

This leads :

$$\hat{V}(f(\text{vec } \hat{\Phi})) = \begin{pmatrix} 0.0245 & 0.0121 & -0.0245 & -0.0119 \\ 0.0121 & 0.0145 & -0.0122 & -0.0144 \\ -0.0245 & -0.0122 & 0.0382 & 0.0181 \\ -0.0119 & -0.0144 & 0.0181 & 0.0213 \end{pmatrix}$$



# Bibliography

- B. Abraham and J. Ledolter. *Statistical Methods for Forecasting*. John Wiley and Sons, New York, 1983.
- R. P. Agarwal. *Difference Equations and Inequalities*. Marcel Dekker, Inc., New York, 2nd edition, 2000.
- H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute for Statistical Mathematics*, 21:243–247, 1969.
- T. Amemiya. *Introduction to Statistics and Econometrics*. Harvard University Press, Cambridge, Massachusetts, 1994.
- S. An and F. Schorfheide. Bayesian analysis of DSGE models. *Econometric Reviews*, 26:113–172, 2007.
- B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Electrical Engineering Series. Prentice-Hall, Englewood Cliffs, New Jersey, 1979.
- D. W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858, 1991.
- D. W. K. Andrews and J. C. Monahan. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, 60:953–966, 1992.
- R. Ashley, C. W. J. Granger, and R. Schmalensee. Advertising and aggregate consumption: An analysis of causality. *Econometrica*, 48(5):1149–1168, 1980.
- J. Bai, R. L. Lumsdaine, and J. H. Stock. Testing for and dating common breaks in multivariate time series. *Review of Economic Studies*, 65:395–432, 1998.
- M. Banbura, D. Giannone, and L. Reichlin. Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, 25:71–92, 2010.

- A. Banerjee, J. Dolado, J. W. Galbraith, and D. F. Hendry. *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford University Press, Oxford, 1993.
- D. Bauer and M. Wagner. A canonical form for unit root processes on the state space framework. *Diskussionsschrift 3-12, Volkswirtschaftliches Institut, Universität Bern*, 2003.
- P. Beaudry and F. Portier. Stock prices, news and economic fluctuations. *American Economic Review*, 96(4):1293–1307, 2006.
- B. S. Bernanke. Alternative explanations of money-income correlation. In K. Brunner and A. Meltzer, editors, *Real Business Cycles, Real Exchange Rates, and Actual Policies*, number 25 in Carnegie-Rochester Conference Series on Public Policy, pages 49–100. North-Holland, Amsterdam, 1986.
- B. S. Bernanke, M. Gertler, and M. W. Watson. Systematic monetary policy and the effects of oil price shocks. *Brookings Papers on Economic Activity*, pages 91–142, 1997:1.
- E. R. Berndt. *The Practice of Econometrics*. Addison Wesley, Reading, Massachusetts, 1991.
- R. J. Bhansali. Parameter estimation and model selection for multistep prediction of a time series: A review. In S. Gosh, editor, *Asymptotics, Non-parametrics, and Time Series*, pages 201–225. Marcel Dekker, New York, 1999.
- P. Billingsley. *Probability and Measure*. John Wiley & Sons, New York, 2nd edition, 1986.
- O. J. Blanchard. A traditional interpretation of macroeconomic fluctuations. *American Economic Review*, 79:1146–1164, 1989.
- O. J. Blanchard and D. Quah. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review*, 79:655–673, 1989.
- O. J. Blanchard and M. W. Watson. Are business cycles all alike? In R. Gordon, editor, *The American Business Cycle: Continuity and Change*, pages 123–179, Chicago, 1986. University of Chicago Press.
- T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

- T. Bollerslev. On the correlation structure for the generalized autoregressive conditional heteroskedastic process. *Journal of Time Series Analysis*, 9: 121–131, 1988.
- T. Bollerslev, R. F. Engle, and D. B. Nelson. ARCH models. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, volume IV of *Handbook in Economics*, pages 2959–3038, Amsterdam, 1994. Elsevier Science B.V.
- P. Bougerol and N. Picard. Stationarity of GARCH processes and some nonnegative time series. *Journal of Econometrics*, 52:115–127, 1992.
- G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised edition, 1976.
- P. Brandner and K. Neusser. Business cycles in open economies: Stylized facts for Austria and Germany. *Weltwirtschaftliches Archiv*, 128:67–87, 1992.
- J. Breitung and S. Eickmeier. Dynamic factor models. In O. Hübler and J. Frohn, editors, *Modern Econometric Analysis*, chapter 3, pages 25–40. Springer Verlag, Berlin, 2006.
- P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, 2nd edition, 1991.
- P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag, New York, 1996.
- D. Burren and K. Neusser. The role of sectoral shifts in the decline of real GDP volatility. *Macroeconomic Dynamics*, 2012. doi: doi:10.1017/S1365100511000289.
- J. Y. Campbell. Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis. *Econometrica*, 55:1249–1273, 1987.
- J. Y. Campbell and N. G. Mankiw. Are output fluctuations transitory? *Quarterly Journal of Economics*, 102:857–880, 1987.
- J. Y. Campbell and P. Perron. Pitfalls and opportunities: What macroeconomists should know about unit roots. In O. J. Blanchard and S. Fischer, editors, *Macroeconomics Annual 1991*, volume 6, pages 141–201. MIT Press, 1991.

- J. Y. Campbell and R. J. Shiller. Cointegration and tests of present value models. *Journal of Political Economy*, 95:1062–1088, 1987.
- J. Y. Campbell, A. W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey, 1997.
- F. Canova. *Methods for Applied Macroeconomic Research*. Princeton University Press, Princeton, New Jersey, 2007.
- F. Canova and M. Ciccarelli. Estimating multi-country VAR models. Working Paper 603, European Central Bank, 2008.
- G. Cavaliere, A. Rahbek, and A. M. R. Taylor. Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica*, 80:1721–1740, 2012.
- V. V. Chari, P. J. Kehoe, and E. R. McGrattan. Are structural VARs with long-run restrictions useful in developing business cycle theory? *Journal of Monetary Economics*, 55:1337–1352, 2008.
- G. C. Chow and A. Lin. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *Review of Economics and Statistics*, 53:372–375, 1971.
- L. J. Christiano and M. Eichenbaum. Unit roots in real GNP: Do we know and do care? *Carnegie-Rochester Conference Series on Public Policy*, 32:7–62, 1990.
- L. J. Christiano, M. Eichenbaum, and C. L. Evans. *Monetary Policy Shocks: What have we Learned and to what End?*, volume 1A of *Handbook of Macroeconomics*, chapter 2, pages 65–148. North-Holland, Amsterdam, 1999.
- L. J. Christiano, M. Eichenbaum, and R. J. Vigfusson. What happens after a technology shock? Working Paper No. 9819, NBER, 2003.
- L. J. Christiano, M. Eichenbaum, and R. J. Vigfusson. Assessing structural VARs. International Finance Discussion Papers No. 866, Board of Governors of the Federal Reserve System, 2006.
- P. F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39:841–862, 1998.

- M. P. Clements and D. F. Hendry. Intercept corrections and structural change. *Journal of Applied Econometrics*, 11:475–494, 1996.
- M. P. Clements and D. F. Hendry. Forecasting with breaks. In *Handbook of Economic Forecasting*, volume 1, pages 605–657. Elsevier, Amsterdam, 2006.
- J. H. Cochrane. How big is the random walk in GNP? *Journal of Political Economy*, 96(5):893–920, 1988.
- T. F. Cooley and S. F. LeRoy. Atheoretical macroeconometrics - a critique. *Journal of Monetary Economics*, 16:283–308, 1985.
- V. Corradi and N. R. Swanson. Predictive density evaluation. In *Handbook of Economic Forecasting*, volume 1, pages 197–284. Elsevier, Amsterdam, 2006.
- N. A. Cuche and M. A. Hess. Estimating monthly GDP in a general Kalman filter framework: Evidence from Switzerland. *Economic and Financial Modelling*, 7:153–194, 2000.
- J. E. Davidson, D. F. Hendry, F. Srba, and S. Yeo. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, 88:661–692, 1978.
- R. Davidson and J. G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, Oxford, 1993.
- S. Dees, F. D. Mauro, M. H. Pesaran, and V. Smith. Exploring the international linkages of the Euro area: A global VAR analysis. *Journal of Applied Econometrics*, 22:1–38, 2007.
- M. Deistler and K. Neusser. Prognosen uni- und multivariater Zeitreihen. In P. Mertens, editor, *Prognoserechnung*, pages 225–256, Heidelberg, 2012. Physica-Verlag.
- D. N. Dejong and C. Dave. *Structural Macroeconomics*. Princeton University Press, Princeton, New Jersey, 2007.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- P. J. Dhrymes. *Introductory Econometrics*. Springer-Verlag, New York, 1978.

- D. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431, 1976.
- D. A. Dickey and W. A. Fuller. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49:1057–1072, 1981.
- F. X. Diebold, T. A. Gunther, and A. S. Tay. "evaluating density forecasts, with applications to financial risk management. *International Economic Review*, 39:863–883, 1998.
- T. Doan, R. B. Litterman, and C. A. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–100, 1984.
- J.-M. Dufour. Unbiasedness of predictions from estimated autoregressions. *Econometric Theory*, 1:387–402, 1985.
- J. Durbin. The fitting of time series models. *Revue de l'Institut International de Statistique*, 28:233–244, 1960.
- B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, 1993.
- S. Elaydi. *An Introduction to Difference Equations*. Springer Science+Business Media, New York, 3rd edition, 2005.
- J. Elder and P. E. Kennedy. Testing for unit roots: What should students be taught? *Journal of Economic Education*, 32:137–146, 2001.
- G. Elliott and A. Timmermann. Economic forecasting. *Journal of Economic Literature*, 46:3–56, 2008.
- R. F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica*, 50:987–1007, 1982.
- R. F. Engle. Risk and volatility: Econometric models and financial practice. *American Economic Review*, 94:405–420, 2004.
- R. F. Engle and T. Bollerslev. Modeling the persistence of conditional variances. *Econometric Reviews*, 5:1–50, 1986.

- R. F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55:251–276, 1987.
- R. F. Engle, D. Lilien, and R. Robins. Estimating time varying risk premia in the term structure: The ARCH–M model. *Econometrica*, 55:391–407, 1987.
- T. Engsted and T. Q. Pedersen. Bias–correction in vector autoregressive models: A simulation study. *Econometrics*, 2:45–71, 2014.
- R. J. Epstein. *A History of Econometrics*. North-Holland, Amsterdam, 1987.
- Eurostat. *ESS Guidelines on Seasonal Adjustment*. Eurostat, Luxembourg, 2009.
- J. Fan and Q. Yao. *Nonlinear Time Series*. Springer-Verlag, New York, 2003.
- J. Faust. The robustness of identified VAR conclusions about money. *Carnegie-Rochester Conference Series in Public Policy*, 49:207–244, 1998.
- J. Fernandez-Villaverde, J. F. Rubio-Ramírez, T. J. Sargent, and M. W. Watson. ABCs and (Ds) of understanding VARs. *American Economic Review*, 97:1021–1026, 2007.
- M. Friedman and A. J. Schwartz. *A Monetary History of the United States, 1867 – 1960*. Princeton University Press, Princeton, New Jersey, 1963.
- R. A. Fry and A. R. Pagan. Sign restrictions in structural vector autoregressions: A critical review. *Journal of Economic Literature*, 49:938–960, 2011.
- W. A. Fuller. *Introduction to Statistical Time Series*. John Wiley and Sons, New York, 1976.
- J. Galí. How well does the IS-LM model fit postwar U.S. data? *Quarterly Journal of Economics*, 107(2):709–738, 1992.
- J. Galí. Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations? *American Economic Review*, 89: 249–271, 1999.
- J. F. Geweke. Inference and causality in economic time series models. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometrics*, volume II, pages 1101–1144. Elsevier, Amsterdam, 1984.

- J. F. Geweke. *Contemporary Bayesian Econometrics and Statistics*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York, 2005.
- E. Ghysels and D. R. Osborn. *The Econometric Analysis of Seasonal Time Series*. Cambridge University Press, Cambridge, UK, 2001.
- C. Giannini. Topics in structural VAR econometrics. Quaderni di Ricerca 21, Università degli Studi di Ancona, Dipartimento di Economia, 1991.
- L. Giraitis, P. Kokoszka, and R. Leipus. Stationary ARCH models: Dependence structure and central limit theorem. *Econometric Theory*, 16:3–22, 2000.
- L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance*, 48:1779–1801, 1993.
- I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Academic Press, New York, 1982.
- V. Gómez and A. Maravall. Programs TRAMO and SEATS. Instructions for the user (with some updates). Working Paper 9628, Servicio de Estudios, Banco de España, 1996.
- J. Gonzalo and S. Ng. A systematic framework for analyzing the dynamic effects of permanent and transitory shocks. *Journal of Economic Dynamics and Control*, 25:1527–1546, 2001.
- N. Gospodinov. Inference in nearly nonstationary SVAR models with long-run identifying restrictions. *Journal of Business and Economic Statistics*, 28:1–12, 2010.
- C. Gouriéroux. *ARCH Models and Financial Applications*. Springer-Verlag, New York, 1997.
- C. W. J. Granger. *Spectral Analysis of Economic Time Series*. Princeton University Press, Princeton, New Jersey, 1964.
- C. W. J. Granger. The typical spectral shape of an economic variable. *Econometrica*, 34:150–161, 1966.
- C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

- C. W. J. Granger and P. Newbold. Spurious regression in econometrics. *Journal of Econometrics*, 2:111–120, 1974.
- W. H. Greene. *Econometric Analysis*. Prentice Hall, New Jersey, 7th edition, 2008.
- W. J. Haan and A. T. Levin. A practitioner's guide to robust covariance matrix estimation. In G. S. Maddala and C. R. Rao, editors, *Handbook of Statistics: Robust Inference*, volume 15, pages 299–342, New York, 1997. Elsevier.
- P. Hall and Q. Yao. Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica*, 71(1):285–317, 2003.
- R. E. Hall. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy*, 86:971–987, 1978.
- J. D. Hamilton. *Time Series Analysis*. Princeton University Press, New Jersey: Princeton, 1994a.
- J. D. Hamilton. State-Space models. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 50, pages 3039–3080. Elsevier, Amsterdam, 1994b.
- E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. John Wiley and Sons, New York, 1988.
- L. P. Hansen and T. J. Sargent. Two difficulties in interpreting vector autoregressions. In L. P. Hansen and T. J. Sargent, editors, *Rational Expectations Econometrics*, Underground Classics in Economics, pages 77–119. Westview, Boulder, Colorado, 1991.
- L. P. Hansen and T. J. Sargent. Seasonality and approximation errors in rational expectations models. *Journal of Econometrics*, 55:21–55, 1993.
- A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge, 1989.
- A. C. Harvey and A. Jaeger. Detrending, stylized facts and the business cycle. *Journal of Applied Econometrics*, 8:231–247, 1993.
- A. C. Harvey and R. G. Pierce. Estimating missing observations in economic time series. *Journal of the American Statistical Association*, 79:125–131, 1984.

- L. D. Haugh. Checking the independence of two covariance stationary time series: A univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71:378–385, 1976.
- M. A. Hauser, B. M. Pötscher, and E. Reschenhofer. Measuring persistence in aggregate output: ARMA models, fractionally integrated ARMA models and nonparametric procedures. *Empirical Economics*, 24:243–269, 1999.
- R. J. Hodrick and E. C. Prescott. Post-war U.S. business cycles: An empirical investigation. Discussion Paper 451, Carnegie-Mellon University, Pittsburgh, 1980.
- R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. Prentice-Hall, Upper Saddle River, New Jersey, 5th edition, 1995.
- E. P. Hong. The autocorrelation structure for the GARCH-M process. *Economics Letters*, 37:129–132, 1991.
- E. P. Howrey. A spectrum analysis of the long swing hypothesis. *International Economic Review*, 9:228–252, 1968.
- S. Hylleberg. *Seasonality in Regression*. Academic Press, Orlando, Florida, 1986.
- S. Hylleberg, R. F. Engle, C. W. J. Granger, and S. Yoo. Seasonal integration and cointegration. *Journal of Econometrics*, 44:215–238, 1990.
- S. T. Jensen and A. Rahbek. Asymptotic normality for non-stationary, explosive GARCH. *Econometric Theory*, 20(6):1203–1226, 2004.
- S. Johansen. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12:231–254, 1988.
- S. Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.
- S. Johansen. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, Oxford, 1995.
- S. Johansen and E. Schaumburg. Likelihood analysis of seasonal cointegration. *Journal of Econometrics*, 88:301–339, 1998.
- T. Kailath. *Linear Systems*. Prentice Hall, Englewood Cliffs, New Jersey, 1980.

- O. Kallenberg. *Foundations of Modern Probability*. Probability and its Applications. Springer-Verlag, New York, 2nd edition, 2002.
- R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions of the ASME Series D*, 82: 35–45, 1960.
- R. E. Kalman. New methods in Wiener filtering theory. In J. L. Bogdanoff and F. Kozin, editors, *Proceedings of the First Symposium of Engineering Applications of Random Function Theory and Probability*, pages 270–388. Wiley, New York, 1963.
- M. G. Kendall. Note on the bias in the estimation of autocorrelation. *Biometrika*, 41:403–404, 1954.
- L. Kilian. Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics*, 80:186–201, 1998.
- L. Kilian and D. P. Murphy. Why agnostic sign restrictions are not enough: Understanding the dynamics of oil market VAR models. *Journal of the European Economic Association*, 10:1166–1188, 2012.
- C.-J. Kim and C. R. Nelson. *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press, Massachusetts, 1999.
- R. G. King and S. T. Rebelo. Low frequency filtering and real business cycles. *Journal of Economic Dynamics and Control*, 17:201–237, 1993.
- R. G. King, C. I. Plosser, and S. Rebelo. Production, growth, and business cycles: I. The basic neoclassical model. *Journal of Monetary Economics*, 21:195–232, 1988.
- R. G. King, C. I. Plosser, J. H. Stock, and M. W. Watson. Stochastic trends and economic fluctuations. *American Economic Review*, 81:819–840, 1991.
- Klein. *Economic Fluctuations in United States 1921-1941*. Wiley, 1950.
- C. Klüppelberg, A. Lindner, and R. Maller. A continuous time GARCH process driven by a Lévy process: Stationarity and second order behaviour. *Journal of Applied Probability*, 41:601–622, 2004.
- D. Kristensen and O. Linton. A closed-form estimator for the GARCH(1,1)-model. *Econometric Theory*, 22:323–337, 2006.

- R. Kunst and K. Neusser. A forecasting comparison of some var techniques. *International Journal of Forecasting*, 2:447–456, 1986.
- R. Kunst and K. Neusser. Cointegration in a macroeconomic system. *Journal of Applied Econometrics*, 5:351–365, 1990.
- S. Kuznets. *Secular Movements in Production and Prices. Their Nature and their Bearing upon Cyclical Fluctuations*. Houghton Mifflin, Boston, 1930.
- D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54:159–178, 1992.
- E. E. Leamer. Is it a demand curve, or is it a supply curve? Partial identification through inequality constraints. *Review of Economics and Statistics*, 63:319–327, 1981.
- S.-W. Lee and B. E. Hansen. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, 10:29–52, 1994.
- M. Lippi and L. Reichlin. The dynamic effects of aggregate demand and supply disturbances: Comment. *American Economic Review*, 83:644–652, 1993.
- R. B. Litterman. Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business and Economic Statistics*, 4:25–38, 1986.
- L. Ljung. *System Identification: Theory for the User*. Prentice Hall, Englewood Cliffs, New Jersey, 2nd edition, 1999.
- R. E. Lucas. Econometric policy evaluation: A critique. In K. Brunner and A. H. Meltzer, editors, *The Phillips Curve and Labor Markets*, volume 1 of *Carnegie-Rochester Conference Series on Public Policy*, pages 19–46, Amsterdam, 1976. North-Holland.
- R. L. Lumsdaine. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica*, 64:575–596, 1986.
- H. Lütkepohl. Asymptotic distributions of impulse response functions and forecast error variance decomposition. *Review of Economics and Statistics*, 72:116–125, 1990.

- H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, 2006.
- H. Lütkepohl, A. Staszewska-Bystrova, and P. Winker. Comparison of methods for constructing joint confidence bands for impulse response functions. Discussion Paper 1292, DIW Berlin, 2013.
- J. G. MacKinnon. Critical values for co-integration tests. In R. F. Engle and C. W. J. Granger, editors, *Long-Run Economic Relationships*, pages 267–276, Oxford, 1991. Oxford University Press.
- J. G. MacKinnon. Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, 11:601–618, 1996.
- J. G. MacKinnon and A. A. Smith, Jr. Approximate bias correction in econometrics. *Journal of Econometrics*, 85:205–230, 1998.
- J. G. MacKinnon, A. Haug, and L. Michelis. Numerical distribution functions of the likelihood ratio test for cointegration. *Journal of Applied Econometrics*, 14:563–577, 1999.
- J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, Chichester, 1988.
- M. Marcellino, J. H. Stock, and M. W. Watson. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometric*, 135:499–526, 2006.
- F. H. C. Marriott and J. A. Pope. Bias in the estimation of autocorrelation. *Biometrika*, 41:393–402, 1954.
- P. Mertens and S. Rässler. Prognoserechnung - einföhrung und Überblick. In P. Mertens and S. Rässler, editors, *Prognoserechnung*, pages 1–37, Heidelberg, 2005. Physica-Verlag.
- C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 2000.
- T. C. Mills. *Modelling Trends and Cycles in Economic Time Series*. Palgrave Texts in Econometrics. Palgrave Macmillan, Hampshire, 2003.
- J. A. Miron. *The Economics of Seasonal Cycles*. MIT Press, Cambridge, Massachusetts, 1996.

- S. Mittnik and P. Zadrozny. Asymptotic distributions of impulse responses, step responses and variance decompositions of estimated linear dynamic models. *Econometrica*, 61:857–870, 1993.
- J. Muth. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55:299–306, 1960.
- A. W. Naylor and G. R. Sell. *Linear Operator Theory in Engineering and Science*, volume 40 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1982.
- M. D. Negro and F. Schorfheide. Priors from general equilibrium models for VARs. *International Economic Review*, 45:643–673, 2004.
- C. R. Nelson and C. I. Plosser. Trends and random walks in macro-economic time series: Some evidence and implications. *Journal of Monetary Economics*, 10:139–162, 1982.
- D. B. Nelson. Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6:318–334, 1990.
- D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1991.
- K. Neusser. Testing the long-run implications of the neoclassical growth model. *Journal of Monetary Economics*, 27:3–37, 1991.
- K. Neusser. An algebraic interpretation of cointegration. *Economics Letters*, 67:273–281, 2000.
- K. Neusser. Difference equations for economists. <http://www.neusser.ch/downloads/DifferenceEquations.pdf>, 2009.
- W. K. Newey and K. D. West. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61:631–653, 1994.
- S. Ng and P. Perron. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, 90:268–281, 1995.
- D. F. Nicholls and A. R. Pagan. Estimating predictions, prediction errors and their standard deviations using constructed variables. *Journal of Econometrics*, 24:293–310, 1984.

- M. Ogaki. An introduction to the generalized method of moments. Working Paper No. 314, University of Rochester, 1992.
- G. H. Orcutt and H. S. Winokur, Jr. First order autoregression inference, estimation, and prediction. *Econometrica*, 37:1–14, 1969.
- M. Osterwald-Lenum. A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics. *Oxford Bulletin of Economics and Statistics*, 54:461–471, 1992.
- A. R. Pagan. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25:183–209, 1984.
- A. R. Pagan and O. C. Robertson. Structural models of the liquidity effect. *Review of Economics and Statistics*, 80:202–217, 1998.
- P. Perron. The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 57:1361–1401, 1989.
- P. Perron. *Palgrave Handbooks of Econometrics*, volume Vol. 1, chapter Dealing with Structural Breaks, pages 278–352. Palgrave MacMillan, New York, 2006.
- P. C. Phillips. Time series regression with a unit root. *Econometrica*, 55:277–301, 1987.
- P. C. B. Phillips. Understanding spurious regressions in econometrics. *Journal of Econometrics*, 33:311–340, 1986.
- P. C. B. Phillips. Optimal inference in cointegrating systems. *Econometrica*, 59:283–306, 1991.
- P. C. B. Phillips. HAC estimation by automated regression. Cowles Foundation Discussion Paper 1470, Yale University, 2004.
- P. C. B. Phillips and B. E. Hansen. Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies*, 57:99–125, 1990.
- P. C. B. Phillips and S. Ouliaris. Asymptotic properties of residual based tests of cointegration. *Econometrica*, 58(1):165–193, 1990.
- P. C. B. Phillips and P. Perron. Testing for a unit root in time series regression. *Biometrika*, 75:335–346, 1988.

- P. C. B. Phillips and V. Solo. Asymptotics for linear processes. *Annals of Statistics*, 20:971–1001, 1992.
- D. A. Pierce and L. D. Haugh. Causality in temporal systems - characterization and survey. *Journal of Econometrics*, 5:265–293, 1977.
- S. M. Potter. Nonlinear impulse response functions. *Journal of Economic Dynamics and Control*, 24:1425–1446, 2000.
- H. Press and J. W. Tukey. Power spectral methods of analysis and their application to problems in airplane dynamics. In *Flight Test Manual*, pages 1–41. NATO Advisory Group for Aeronautical Research and Development, 1956.
- M. B. Priestley. *Spectral Analysis and Time Series*, volume 1&2. Academic Press, London, 1981.
- D. Quah. Permanent and transitory movements in labor income: An explanation for ‘excess smoothness’ in consumption. *Journal of Political Economy*, 98:449–475, 1990.
- D. Quah and T. J. Sargent. A dynamic index model for large cross sections. In J. H. Stock and M. W. Watson, editors, *Business Cycles, Indicators, and Forecasting*, chapter Chapter 7. University of Chicago Press, 1993.
- L. Reichlin. Factor models in large cross sections of time series. In M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, editors, *Advances in Econometrics, Theory and Applications*, volume III of *Econometric Society Monographs*, page 47–86. Cambridge University Press, Cambridge, UK, 2003.
- G. C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 1993.
- R. Rigobon. Identification through heteroskedasticity. *Review of Economics and Statistics*, 85:777–792, 2003.
- E. A. Robinson. A historical perspective of spectrum estimation. *Proceedings of the IEEE*, 70:885–907, 1982.
- M. Rosenblatt. *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. Springer-Verlag, New York, 2000.
- T. J. Rothenberg. Identification of parametric models. *Econometrica*, 39:577–591, 1971.

- J. F. Rubio-Ramírez, D. F. Waggoner, and T. Zha. Structural vector autoregressions: Theory of identification and algorithms for inference. *Review of Economic Studies*, 77:665–696, 2010.
- W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.
- W. Rudin. *Real and Complex Analysis*. McGraw-Hill, Boston, Massachusetts, 3rd edition, 1987.
- D. Runkle. Vector autoregressions and reality. *Journal of Business and Economic Statistics*, 5:437–442, 1987.
- S. E. Said and D. A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71:599–607, 1984.
- P. Samuelson. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6:41–49, 1965.
- T. J. Sargent. *Macroeconomic Theory*. Academic Press, Orlando, Florida, second edition, 1987.
- T. J. Sargent. Two models of measurements and the investment accelerator. *Journal of Political Economy*, 97:251–87, 1989.
- T. J. Sargent. *Recursive Macroeconomic Theory*. MIT Press, Cambridge, Massachusetts, 2nd edition, 2004.
- T. J. Sargent and C. A. Sims. Business cycle modelling without pretending to have too much *a priori* economic theory. In C. A. Sims, editor, *New Methods in Business Cycle Research*, pages 45–109. Federal Reserve Bank of Minneapolis, Minneapolis, 1977.
- F. Schorfheide. VAR forecasting under misspecification. *Journal of Econometrics*, 128:99–136, 2005.
- R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, New York, 1980.
- P. Shaman and R. A. Stine. The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association*, 83:842–848, 1988.
- B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.
- C. A. Sims. Money, income, and causality. *American Economic Review*, 62:540–552, 1972.

- C. A. Sims. Seasonality in regression. *Journal of the American Statistical Association*, 69:618–626, 1974.
- C. A. Sims. Comparison of interwar and postwar business cycles: Monetarism reconsidered. *American Economic Review*, 70(2):250–257, 1980a.
- C. A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–45, 1980b.
- C. A. Sims. Are forecasting models usable for policy analysis. *Federal Reserve Bank of Minneapolis Quarterly Review*, 10(1):2–16, 1986.
- C. A. Sims. rational expectations modeling with seasonally adjusted data. *Journal of Econometrics*, 55:9–19, 1993.
- C. A. Sims. Error bands for impulse responses. *Econometrica*, 67:1113–1155, 1999.
- C. A. Sims, J. H. Stock, and M. W. Watson. Inference in linear time series with some unit roots. *Econometrica*, 58:113–144, 1990.
- J. H. Stock. Unit roots, structural breaks and trends. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, volume IV of *Handbook in Economics*, pages 2739–2841, Amsterdam, 1994. Elsevier Science B.V.
- J. H. Stock and M. W. Watson. Variable trends in economic time series. *Journal of Economic Perspectives*, 2(3):147–174, 1988a.
- J. H. Stock and M. W. Watson. Testing for common trends. *Journal of American Statistical Association*, 83:1097–1107, 1988b.
- G. Strang. *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, San Diego, 3rd edition, 1988.
- D. Sul, P. C. B. Phillips, and C.-Y. Choi. Prewhitening bias in HAC estimation. *Oxford Bulletin of Economics and Statistics*, 67:517–546, 2005.
- A. S. Tay and K. F. Wallis. Density forecasting: A Survey. *Journal of Forecasting*, 19:235–254, 2000.
- J. Tinbergen. *Statistical Testing of Business Cycle Theories*. League of Nations, Genf, 1939.
- D. Tjøstheim and J. Paulsen. Bias of some commonly-used time series estimates. *Biometrika*, 70:389–399, 1983. Corrigendum (1984), 71, 656.

- J. Tobin. Money and income: Post hoc ergo propter hoc? *Quarterly Journal of Economics*, 84:310–317, 1970.
- H. Uhlig. What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics*, 52:381–419, 2005.
- H. Uhlig and M. Ravn. On adjusting the HP-filter for the frequency of observations. *Review of Economics and Statistics*, 84:371–376, 2002.
- T. J. Vogelsang. Wald-type tests for detecting breaks in the trend function of a dynamic time series. *Econometric Theory*, 13:818–849, 1997.
- M. W. Watson. *Vector Autoregressions and Cointegration*, volume 4 of *Handbook of Econometrics*, chapter 47, pages 2843–2915. North-Holland, Amsterdam, 1994.
- A. A. Weiss. Asymptotic theory for ARCH models: Estimation and testing. *Econometric Theory*, 2:107–131, 1986.
- H. White. A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.
- E. T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1923.
- N. Wiener. The theory of prediction. In E. F. Beckenbach, editor, *Modern Mathematics for Engineers*, Series 1, New York, 1956. McGraw-Hill.
- M. Woodford. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press, Princeton, New Jersey, 2003.
- C. F. J. Wu. On the convergence of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.
- B. S. Yoo. *Multi-Cointegrated Time Series and a Generalized Error Correction Model*. Ph.d., University of California, San Diego, 1987.
- G. U. Yule. Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society*, 89:1–64, 1926.
- G. U. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society, A*, 226:267–298, 1927.

- P. A. Zadrozny. Necessary and sufficient restrictions for existence of a unique fourth moment of a univariate GARCH(p,q) process. CESifo Working Paper No.1505, 2005.
- J.-M. Zakoïan. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, 18:931–955, 1994.
- A. Zellner. Causality and econometrics. In K. Brunner and A. Meltzer, editors, *Three Aspects of Policymaking: Knowledge, Data and Institution*, Carnegie-Rochester Conference Series on Public Policy, pages 9–54. North-Holland, Amsterdam, 1979.
- A. Zellner and F. Palm. Time series analysis and simultaneous equation econometric models. *Journal of Econometrics*, 2:17–54, 1974.
- E. Zivot and D. W. Andrews. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics*, 10:251–270, 1992.

# Index

- ACF, *see also* Autocorrelation function
- ADF-test, 157
- AIC, *see* Information criterion, 108, *see also* Information criterion, 263
- AR process, 28
  - autocorrelation function, 28
  - autocovariance function, 28
  - stationary solution, 28
- ARIMA process, 109, 142
- ARMA model
  - estimation, 93
  - identification, 93
- ARMA process, *see also* Autoregressive moving-average process
  - autocovariance function, 39
  - causality, 33
  - causality condition, 34
  - estimation, 101
  - invertibility, 38
  - invertibility condition, 38
  - maximum likelihood estimation, 101
  - state space representation, 344
- Autocorrelation function, 12
  - estimation, 77
    - asymptotic distribution, 78
    - Bartlett's formula, 78
    - confidence interval, 79
    - confidence interval for AR(1) process, 81
    - confidence interval for MA(q) process, 80
  - interpretation, 68
  - order, 12
  - properties, 20
  - random walk, 152
  - univariate, 12
- Autocorrelation function, partial, 65
  - AR process, 66
  - estimation, 82
  - interpretation, 68
  - MA process, 67
- Autocovariance function, 11
  - ARMA process, 39
  - estimation, 77
  - linear process, 131
  - MA(1) process, 20
  - multivariate, 216
  - order, 12
  - properties, 19
  - random walk, 152
  - univariate, 11
- Autoregressive conditional heteroskedasticity models, *see* Volatility
- Autoregressive final form, 239
- Autoregressive moving-average process, 25
- Autoregressive moving-average process mean, 26
- Back-shift operator, *see also* Lag operator
- Bandwidth, 86
- Bartlett's formula, 78
- Basic structural model, 345, 363
  - cyclical component, 346

- local linear trend model, 345
  - seasonal component, 346
- Bayesian VAR, 271
- Beveridge-Nelson decomposition, 146, 381
- Bias proportion, 266
- Bias, small sample, 98, 247
  - correction, 98, 247
- BIC, *see* Information criterion, 108, *see also* Information criterion, 263
- Borel-Cantelli lemma, 376
- Box-Pierce statistic, 80
- BSM, *see* Basic structural model
- Canonical correlation coefficients, 332
- Cauchy-Bunyakovskii-Schwarz inequality, 375
- Causal representation, 33
- Causality, *see also* Wiener-Granger causality, 341
  - Wiener-Granger causality, 273
- Central Limit Theorem
  - m-dependence, 379
- Characteristic function, 378
- Chebyshev's inequality, 375
- Cointegration, 172
  - Beveridge-Nelson decomposition, 320, 326
  - bivariate, 172
  - common trend representation, 327
  - definition, 321
  - Granger's representation theorem, 326
  - order of integration, 319
  - shocks, permanent and transitory, 327
  - Smith-McMillan factorization, 323
  - test
    - Johansen test, 328
    - regression test, 173
    - triangular representation, 328
  - VAR model, 322
    - assumptions, 322
  - VECM, 323
  - vector error correction, 323
- Companion form, 234
- Convergence
  - Almost sure convergence, 376
  - Convergence in  $r$ -th mean, 376
  - Convergence in distribution, 377
  - Convergence in probability, 376
- Correlation function, 216
  - estimator, 223
  - multivariate, 216
- Covariance function
  - estimator, 222
  - properties, 217
- covariance function, 216
- Covariance proportion, 266
- Covariance, long-run, 223
- Cross-correlation, 217
  - distribution, asymptotic, 224
- Cyclical component, 135, 346
- Dickey-Fuller distribution, 150
- Durbin-Levinson algorithm, 66
- Dynamic factor model, 348
- EM algorithm, 358
- Ergodicity, 8, 73
- Estimation
  - ARMA model, 101
  - order, 106
- Estimator
  - maximum likelihood estimator, 101, 102
  - method of moments
    - GARCH(1,1) model, 200
  - moment estimator, 95
  - OLS estimator, 97
    - process, integrated, 150
  - Yule-Walker estimator, 94

- Example
- AD-curve and Money Supply, 279
  - advertisement and sales, 292
  - ARMA processes, 35
  - consumption expenditure and advertisement, 227
  - demand and supply shocks, 308
  - estimation of long-run variance, 88
  - estimation of quarterly GDP data, 360
  - GDP and consumer sentiment index, 227
  - growth model, neoclassical, 337
  - inflation and short-term interest rate, 173
  - IS-LM model with Phillips curve, 294
  - modeling real GDP of Switzerland, 110
  - present discounted value model, 312
  - Swiss Market Index, 201
  - term structure of interest rate, 176
  - unit root test, 161
- Expectation, adaptive, 62
- Exponential smoothing, 60
- Factor model, dynamic, *see* Dynamic factor model
- FEVD, *see also* Forecast error variance decomposition
- Filter
- gain function, 133
  - Gibbs phenomenon, 135
  - high-pass, 135
  - Hodrick-Prescott filter, 135
  - HP-filter, 135
  - Kuznets filter, 133
  - low-pass, 134
  - phase function, 133
  - TRAMO-SEATS, 138
  - transfer function, 132
  - X-11 filter, 138
  - X-12-Filter, 138
- Filter, time invariant, 130
- Filtering problem, 350
- Final form, *see* Autoregressive final form
- Forecast error variance decomposition, 289
- Forecast evaluation
- Bias proportion, 266
  - Covariance proportion, 266
  - Mean-absolute-error, 266
  - Out-of-sample strategy, 267
  - Root-mean-squared-error, 266
  - Uncertainty, 267
  - Variance proportion, 266
- Forecast function, 47
- AR(p) process, 51
  - ARMA(1,1) process, 56
  - forecast error, 49
  - infinite past, 55
  - linear, 47
  - MA(q) process, 52
  - variance of forecast error, 50
- Forecast, direct, 271
- Forecast, iterated, 260, 267, 271
- Fourier frequencies, 126
- Fourier transform, discrete, 126
- FPE, *see* Information criterion, 264
- Frequency domain, 117
- Gain function, 133
- Gauss Markov theorem, 98
- Growth component, 135
- HAC variance, *see also* variance, heteroskedastic and autocorrelation consistent
- Harmonic process, 57, 123

- Hodrick-Prescott filter, 135
- HQC, see Information criterion, 108, see Information criterion, 263
- Identification
  - Kalman filter, 359
- Identification problem, 280
- Impulse response function, 33, 38
- Information criterion, 108, 263
  - AIC, 108, 263
  - BIC, 108, 263
  - Final prediction error, 264
  - FPE, 264
  - Hannan-Quinn, 263
  - HQC, 108
  - Schwarz, 108, 263
- Innovations, 59
- Integrated GARCH, 194
- Integrated process, 109
- Integrated regressors
  - rules of thumb, 175
- Integration, order of, 142
- Intercept correction, 271
- Invertibility, 38
- Johansen test
  - distribution, asymptotic, 334
  - hypothesis tests over  $\beta$ , 334
  - max test, 332
  - trace test, 332
- Kalman filter, 352
  - application
    - basic structural model, 363
    - estimation of quarterly GDP data, 360
  - AR(1) process, 351, 355
  - assumptions, 340
  - causal, 341
  - EM algorithm, 358
  - filtering problem, 350
  - forecasting step, 352
  - gain matrix, 353
  - identification, 359
  - initialization, 353
  - likelihood function, 357
  - Markov property, 341
  - measurement errors, 355
  - observation equation, 340
  - prediction problem, 350
  - smoother, 354
  - smoothing problem, 350
  - stable, 341
  - state equation, 340
  - stationarity, 341
  - updating step, 353
- Kalman smoother, 354
- Kernel function, 85
  - bandwidth, 86
    - optimal, 87
    - rule of thumb, 87
  - Bartlett, 85
  - boxcar, 85
  - Daniell, 85
  - lag truncation parameter, 86
    - optimal, 87
  - quadratic spectral, 85
  - Tukey-Hanning, 85
- Lag operator, 26
  - calculation rules, 26
  - definition, 26
  - polynomial, 27
- Lag polynomial, 27
- Lag truncation parameter, 86
- Lag window, 126
- Leading indicator, 227, 277
- Least-squares estimator, 97, 103
- Likelihood function, 101, 357
  - ARMA process, 101
  - Kalman filter, 357
- Ljung-Box statistic, 80
- Loading matrix

- definition, 323
- Local linear trend model, 345
- Long-run identification, 305
  - instrumental variables, 306
- m-dependence, 379
- MA process, 15, 27
  - autocorrelation function, 28
  - autocovariance function, 20, 28
- MAE, 266
- Matrix norm, 219
  - absolute summability, 220
  - quadratic summability, 220
  - submultiplicativity, 220
- Maximum likelihood estimator, 102, 197
  - ARMA(p,q) model, 101
  - asymptotic distribution, 104
    - AR process, 105
    - ARMA(1,1) process, 105
    - MA process, 105
  - GARCH(p,q) model, 199
- Maximum likelihood method, 101
- Mean, 71, 221
  - asymptotic distribution, 73, 75
  - distribution, asymptotic, 222
  - estimation, 71, 221
  - estimator, 222
- Mean reverting, 141
- Mean squared error matrix
  - estimated coefficients, 263
  - known coefficients, 260
- Measurement errors, 351
- Memory, short, 28
- Missing observations, 344
- Model, 9
- Normal distribution, multivariate
  - conditional, 350
- Normal equation, 48
- Observation equation, 340
- Observationally equivalent, 280
- OLS estimator, 97
  - distribution, asymptotic, 98
- Order of integration, 142
- Ordinary-least-squares estimator, 97
- Oscillation length, 119
- Overfitting, 106
- PACF, *see also* Autocorrelation function, partial
- Partial autocorrelation function
  - computation, 66
  - estimation, 82
- Period length, 119
- Periodogramm, 126
- Persistence, 146
- Phase function, 133
- Portmanteau test, 80
- PP-test, 158
- Prediction problem, 350
- Predictor, *see also* Forecast function
- Present discounted value model, 312
  - Beveridge-Nelson decomposition, 317
  - cointegration, 315
  - spread, 313
  - VAR representation, 314
  - vector error correction model, 315
- Prewhitening, 88
- Process, ARIMA, 142
- Process, stochastic, 7, 215
  - ARMA process, 25
  - branching process, 10
  - deterministic, 57
  - difference-stationary, 142
  - finite memory, 17
  - finite-range dependence, 17
  - Gaussian process, 13
  - harmonic process, 57
  - integrated, 109, 142, 319

- Beveridge-Nelson decomposition, 146
  - forecast, long-run, 143
  - impulse response function, 146
  - OLS estimator, 150
  - persistence, 146
  - variance of forecast error, 145
- linear, 219
- memory, 14
- moving-average process, 15
- multivariate, 215
- purely non-deterministic, 59
- random walk, 18
- random walk with drift, 18
- regular, 59
- spectral representation, 124
- trend-stationary, 142
  - forecast, long-run, 143
  - impulse response function, 145
  - variance of forecast error, 144
- white noise, 14
- Random walk, 9, 18
  - autocorrelation function, 152
  - autocovariance function, 152
- Random walk with drift, 18
- Real business cycle model, 349
- Realization, 8
- Restrictions
  - long-run, 301
  - short-run, 286
  - sign restrictions, 285
- RMSE, 266
- Seasonal component, 346
- Shock
  - permanent, 38
  - transitory, 37
- Shocks, structural, 279
- Short range dependence, *see also* Memory, short
  - Signal-to-noise ratio, 346
  - Singular values, 332
  - Smoothing, 354
  - Smoothing problem, 350
  - Smoothing, exponential, 60
  - Spectral average estimator, discrete, 127
  - Spectral decomposition, 117
  - Spectral density, 118, 123
    - ARMA process, 128
    - autocovariance function, 119
    - estimator, direct, 126
    - estimator, indirect, 125
    - Fourier coefficients, 119
    - spectral density, rational, 129
    - variance, long-run, 126
  - Spectral distribution function, 123
  - Spectral representation, 123, 124
  - Spectral weighting function, 127
  - Spectrum estimation, 117
  - Spurious correlation, 170
  - Spurious regression, 170
  - State equation, 340
  - State space, 8
  - State space representation, 234, 340
    - ARMA processes, 344
    - ARMA(1,1), 343
    - missing observations, 344
    - stationarity, 341
    - VAR process, 342
  - Stationarity, 11
    - multivariate, 216
    - strict, 13
    - weak, 11
  - Stationarity, strict
    - multivariate, 217
  - Strong Law of Large Numbers, 376
  - Structural break, 164
  - Structural change, 19
  - Structural time series analysis, 149
  - Structural time series model, 345, 363

- basic structural model, 345
- Structural breaks, 271
- Summability
  - absolute, 220
  - quadratic, 220
- Summability Condition, 381
- Superconsistency, 151
- Swiss Market Index (SMI), 201
- Test
  - autocorrelation, squared residuals, 196
  - cointegration
    - regression test, 173
  - Dickey-Fuller regression, 155
  - Dickey-Fuller test, 154, 156
    - augmented, 157
    - correction, autoregressive, 157
  - heteroskedasticity, 196
    - Engle's Lagrange-multiplier test, 197
  - Independence, 224
  - Johansen test, 328
    - correlation coefficients, canonical, 332
    - distribution, asymptotic, 334
    - eigenvalue problem, 331
    - hypotheses, 329
    - hypothesis tests over  $\beta$ , 334
    - likelihood function, 332
    - max test, 332
    - singular values, 332
    - trace test, 332
  - Kwiatkowski-Phillips-Schmidt-Shin-test, 168
  - Phillips-Perron test, 154, 158
  - stationarity, 168
  - uncorrelatedness, 224
  - unit root test
    - structural break, 164
    - testing strategy, 159
    - unit-root test, 154
    - white noise
      - Box-Pierce statistic, 80
      - Ljung-Box statistic, 80
      - Portmanteau test, 80
- Time, 7
- Time domain, 117
- Time series model, 9
- Times series analysis, structural, 149
- Trajectory, 8
- Transfer function, 132
- Transfer function form, 239
- Underfitting, 106
- Value-at-Risk, 206
- VaR, *see* Value-at-Risk
- VAR process
  - Bayesian VAR, 271
  - correlation function, 237
  - covariance function, 237
  - estimation
    - order of VAR, 263
    - Yule-Walker estimator, 254
  - forecast function, 257
    - mean squared error, 259, 260
  - form, reduced, 280, 282
  - form, structural, 279, 282
  - identification
    - long-run identification, 301, 305
    - short-run identification, 286
    - sign restrictions, 285
    - zero restrictions, 286
  - identification problem, 280, 283
    - Cholesky decomposition, 287
  - impulse response function, 289
    - bootstrap, 291
    - confidence intervals, 291
    - delta method, 291
  - state space representation, 342
  - Structural breaks, 271

- VAR(1) process, 232
  - stationarity, 232
  - variance decomposition, 289
    - confidence intervals, 291
- Variance proportion, 266
- Variance, heteroskedastic and auto-correlation consistent, 76
- Variance, long-run, 76, 223
  - estimation, 83, 88
    - prewhitening, 88
  - multivariate, 223
  - spectral density, 126
- VARMA process, 231
  - causal representation, 235
  - condition for causal representation, 235
- VECM, *see also* Cointegration
- Vector autoregressive moving-average process, *see also* VARMA process
- Vector autoregressive process, *see also* VAR process
- Volatility
  - ARCH(p) model, 185
  - ARCH-in-mean model, 189
  - EGARCH model, 189
  - Forecasting, 195
  - GARCH(1,1) model, 190
  - GARCH(p,q) model, 187
    - ARMA process, 187
    - heavy-tail property, 187
  - GARCH(p,q) model, asymmetric, 188
  - heavy-tail property, 184
  - IGARCH, 194
  - models, 185
  - TARCH(p,q) model, 188
- Weighting function, 85
- White noise, 14
  - multivariate, 218
  - univariate, 14
- Wiener-Granger causality, 273
  - test
    - F-test, 276
    - Haugh-Pierce test, 278
- Wold Decomposition Theorem, 57
  - multivariate, 261
  - univariate, 57
- Yule-Walker equations
  - multivariate, 238
  - univariate, 94
- Yule-Walker estimator, 94
  - AR(1) process, 95
  - asymptotic distribution, 95
  - MA process, 96