UNIVERSITY OF TÜBINGEN

SCHOOL OF BUSINESS AND ECONOMICS

FACULTY OF ECONOMETRICS, STATISTICS AND EMPIRICAL ECONOMICS

Advanced Time Series Analysis – Lecture Notes

Author:
Andreas LAY
E-Mail:
andreas.lay@student.uni-tuebingen.de

Course Instructor:
Dr. Thomas DIMPFL

February 8, 2014

Contents

1 Difference Equations

1.1 p*th*-order Difference Equations

A p*th*-order difference equation has the following form:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + w_t \tag{1.1.1}$$

Expressed in matrices:

$$\boldsymbol{\xi}_t = \boldsymbol{F}\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t \tag{1.1.2}$$

or

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} .$$

Equation (1.1.2) can be rewritten as

$$\boldsymbol{\xi}_t = \boldsymbol{F}^{t+1}\boldsymbol{\xi}_{-1} + \boldsymbol{F}^t\boldsymbol{v}_0 + \boldsymbol{F}^{t-1}\boldsymbol{v}_1 + \boldsymbol{F}^{t-2}\boldsymbol{v}_2 + \cdots + \boldsymbol{F}\boldsymbol{v}_{t-1} + \boldsymbol{v}_t \,. \tag{1.1.3}$$

Equation (1.1.3) can be generalized to

$$\boldsymbol{\xi}_{t+j} = \boldsymbol{F}^{j+1}\boldsymbol{\xi}_{t-1} + \boldsymbol{F}^j\boldsymbol{v}_t + \boldsymbol{F}^{j-1}\boldsymbol{v}_{t+1} + \boldsymbol{F}^{j-2}\boldsymbol{v}_{t+2} + \cdots + \boldsymbol{F}\boldsymbol{v}_{t+j-1} + \boldsymbol{v}_{t+j} \,. \tag{1.1.4}$$

where t is the starting point for a given vector of past values $\boldsymbol{\xi}_{t-1}$ and j the number of future periods starting in period t. It follows from equation (1.1.4) that

$$\begin{aligned} y_{t+j} =& f_{11}^{(j+1)} y_{t-1} + f_{12}^{(j+1)} y_{t-2} + \cdots + f_{1p}^{(j+1)} y_{t-p} \\ &+ f_{11}^{(j)} w_t + f_{11}^{(j-1)} w_{t+1} + f_{11}^{(j-2)} w_{t+2} + \cdots + f_{11}^{(1)} w_{t+j-1} + w_{t+j} \end{aligned} \tag{1.1.5}$$

which implies a marginal effect of a change in w_t on y_{t+j} of

$$\frac{\partial y_{t+j}}{\partial w_t} = f_{11}^{(j)} \tag{1.1.6}$$

with $f_{11}^{(j)}$ denoting the $(1,1)$ element of $\boldsymbol{F}^j$.

1.2 p*th*-order Difference Equations with Distinct Eigenvalues

The eigenvalues of matrix $\boldsymbol{F}$ are the values of λ for which

$$|\boldsymbol{F} - \lambda \boldsymbol{I}_p| = 0 \tag{1.2.1}$$

which implies that the eigenvalues are the values that satisfy

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p = 0 \,. \tag{1.2.2}$$

If the $(p \times p)$ matrix $\boldsymbol{F}$ has p distinct eigenvalues, this matrix (respectively matrix $\boldsymbol{F}^j$) can be expressed as

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{-1} \\ \Leftrightarrow \boldsymbol{F}^j &= \boldsymbol{T}\boldsymbol{\Lambda}^j\boldsymbol{T}^{-1} \end{aligned} \tag{1.2.3}$$

where $\boldsymbol{T}$ is a $(p \times p)$ matrix with the eigenvectors of matrix $\boldsymbol{F}$ as columns and $\boldsymbol{\Lambda}$ is a $(p \times p)$ matrix with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $\boldsymbol{F}$ along its principal diagonal and zeros elsewhere. As the eigenvalues are assumed to be distinct, the columns of $\boldsymbol{T}$ are linearly independent and thus $\boldsymbol{T}$ is invertible (if the eigenvalues of $\boldsymbol{F}$ are not distinct, the eigenvectors may or may not be linearly independent).

It follows that the marginal effect defined by equation (1.1.6) is a weighted average of the p distinct eigenvalues raised to the jth power:

$$\frac{\partial y_{t+j}}{\partial w_t} = f_{11}^{(j)} = c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j \tag{1.2.4}$$

with

$$c_i = \left[t_{1i}t^{i1}\right] \tag{1.2.5}$$

where t_{1i} is the ith element in the first row of $\boldsymbol{T}$ and t^{i1} is the ith element in the first column of $\boldsymbol{T}^{-1}$. It can be shown that

$$c_i = \frac{\lambda_i^{p-1}}{\prod\limits_{\substack{k=1,\\ k\neq i}}^{p} (\lambda_i - \lambda_k)} \,. \tag{1.2.6}$$

Example: A Second-Order Difference Equation with Distinct Eigenvalues

For $p = 2$, it follows from equation (1.2.2) that

$$\lambda^2 - \phi_1\lambda - \phi_2 = 0 \,. \tag{1.2.7}$$

Solving for λ results in two solutions:

$$\lambda_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} \,, \tag{1.2.8}$$

$$\lambda_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} \,. \tag{1.2.9}$$

The eigenvalues λ_1, λ_2 are complex whenever

$$\phi_1^2 + 4\phi_2 < 0 \,, \tag{1.2.10}$$

that is λ_1, λ_2 can be expressed as

$$\lambda_1 = a + bi \,, \tag{1.2.11}$$

$$\lambda_2 = a - bi \tag{1.2.12}$$

with

$$a = \frac{\phi_1}{2}\,, \tag{1.2.13}$$

$$b = \frac{\sqrt{-\phi_1^2 - 4\phi_2}}{2}\,. \tag{1.2.14}$$

Considering that

$$\cos(\theta) = \frac{a}{R}\,, \tag{1.2.15}$$

$$\sin(\theta) = \frac{b}{R} \tag{1.2.16}$$

equations (1.2.11) and (1.2.12) can be expressed as

$$\lambda_1 = R\left[\cos(\theta) + i\sin(\theta)\right] = R\left[e^{i\theta}\right]\,, \tag{1.2.17}$$

$$\lambda_2 = R\left[\cos(\theta) - i\sin(\theta)\right] = R\left[e^{-i\theta}\right] \tag{1.2.18}$$

with $R = \sqrt{a^2 + b^2}$.

2 Autoregressive Moving Average Models

2.1 Basic Concepts

We define a (vector) stochastic process $\{\boldsymbol{z}_i\}$ $(i = 1, 2, \dots)$ with

$$\boldsymbol{z}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix}$$

where x_{ik} $(k = 1, \dots, n)$ represent scalar stochastic processes. $\{\boldsymbol{z}_i\}$ is **strictly stationary** if the joint distribution of $(\boldsymbol{z}_i, \boldsymbol{z}_{i_1}, \boldsymbol{z}_{i_2}, \dots, \boldsymbol{z}_{i_r})$ depends only on $i_1 - i, i_2 - i, \dots, i_r - i$. A stochastic process is **weakly stationary** if

1. $E(\boldsymbol{z}_i)$ does not depend on i and
2. $\mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{z}_{i-j})$ exists, is finite and depends only on j but not on i.

The second condition can be formalized as

$$\mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{z}_{i-j}) \equiv \boldsymbol{\Gamma}_{ij} \stackrel{!}{=} \boldsymbol{\Gamma}_j \;\; \forall \;\; i \tag{2.1.1}$$

or expressed in matrices

$$\begin{aligned}\boldsymbol{\Gamma}_{ij} &= \begin{bmatrix} \mathrm{Cov}(x_{i1}, x_{(i-j)1}) & \mathrm{Cov}(x_{i1}, x_{(i-j)2}) & \dots & \mathrm{Cov}(x_{i1}, x_{(i-j)n}) \\ \mathrm{Cov}(x_{i2}, x_{(i-j)1}) & \mathrm{Cov}(x_{i2}, x_{(i-j)2}) & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_{in}, x_{(i-j)1}) & \dots & \dots & \mathrm{Cov}(x_{in}, x_{(i-j)n}) \end{bmatrix} \\ &\overset{!}{=} \begin{bmatrix} \gamma_j^{(11)} & \gamma_j^{(12)} & \dots & \gamma_j^{(1n)} \\ \gamma_j^{(21)} & \gamma_j^{(22)} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_j^{(n1)} & \dots & \dots & \gamma_j^{(nn)} \end{bmatrix}\end{aligned}$$

where $\gamma_j^{(kl)}$ represents the correlation between the random variables $x_{ik}, k \in \{1, 2, ..., n\}$ and $x_{(i-j)l}, l \in \{1, 2, ..., n\}$ resulting from (weakly) stationary stochastic processes. We might for example think of observing a time series of GDP as x_{i1} and inflation as x_{i2} which are both modelled by stochastic processes. In this case, $\gamma_2^{(12)}$ is the correlation between GDP today and inflation two periods ago. Each variable in each period is considered a random variable resulting from a stochastic process and the matrix $\boldsymbol{\Gamma}_j$ contains all pairwise correlations between those variables in period i and period $i - j$.

For weakly stationary processes, the following condition holds:

$$\boldsymbol{\Gamma}_{ij} = \boldsymbol{\Gamma}_{(i+j)j} \tag{2.1.2}$$

with

$$\boldsymbol{\Gamma}_{(i+j)j} = \begin{bmatrix} \mathrm{Cov}(x_{(i+j)1}, x_{i1}) & \dots & \mathrm{Cov}(x_{(i+j)1}, x_{il}) & \dots & \mathrm{Cov}(x_{(i+j)1}, x_{in}) \\ \vdots & \mathrm{Cov}(x_{(i+j)2}, x_{i2}) & \vdots & \vdots & \vdots \\ \mathrm{Cov}(x_{(i+j)k}, x_{i1}) & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(x_{(i+j)n}, x_{i1}) & \dots & \dots & \dots & \mathrm{Cov}(x_{(i+j)n}, x_{in}) \end{bmatrix},$$

which implies that

$$\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}'_{-j} \,. \tag{2.1.3}$$

A **white noise process** is defined as a covariance stationary process $\{\boldsymbol{z}_i\}$ with zero mean and no serial correlation:

$$\mathrm{E}(\boldsymbol{z}_i) = \mathbf{0}\,, \tag{2.1.4}$$

$$\mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{z}_{i-j}) = \mathbf{0} \text{ for } j \neq 0 \tag{2.1.5}$$

A stationary process is called **ergodic** if

$$\begin{aligned}&\lim_{n\to\infty} |\mathrm{E}\left[f\left(z_i, \dots, z_{i+k}\right) g\left(z_{i+n}, \dots, z_{i+n+l}\right)\right]| \\ &\quad = |\mathrm{E}\left[f\left(z_i, \dots, z_{i+k}\right)\right]| |\mathrm{E}\left[g\left(z_{i+n}, \dots, z_{i+n+l}\right)\right] \,.\end{aligned} \tag{2.1.6}$$

The **ergodic theorem** states that for a stationary and ergodic process $\boldsymbol{z}_i$ with $E(\boldsymbol{z}_i) = \boldsymbol{\mu}$,

$$\bar{\boldsymbol{z}}_n \equiv \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \overset{a.s.}{\to} \boldsymbol{\mu}\,. \tag{2.1.7}$$

Moreover, for any function $f(\cdot)$, $\{f(\boldsymbol{z}_i)\}$ is ergodic and stationary whenever $\{\boldsymbol{z}_i\}$ is.

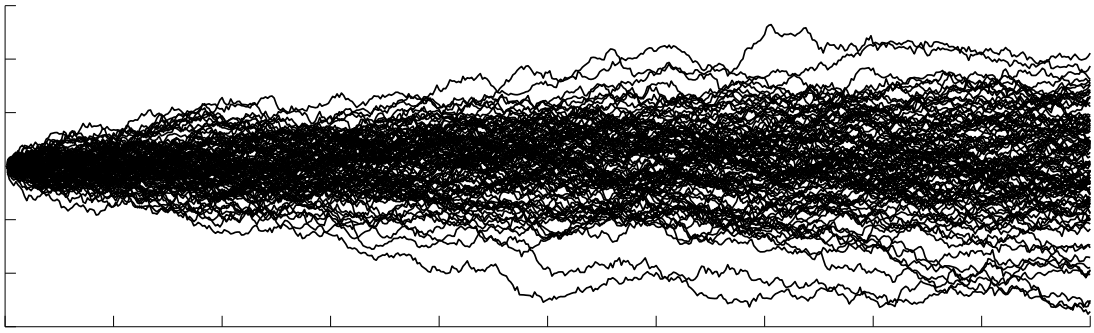

Figure 1: An example of a nonstationary process: Different realisations of a random walk.

2.2 Moving Average (MA) Processes

2.2.1 MA(1) Process

A **first-order MA process** (MA(1) process) is defined by

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \tag{2.2.1}$$

where $\{\epsilon_t\}$ represents a white noise process. The expected value of the process is constant for each period:

$$\mathrm{E}[Y_t] = \mu \tag{2.2.2}$$

The same is true for the variance:

$$\begin{aligned}\gamma_0 &= \mathrm{E}\left[(Y_t-\mu)^2\right] \\ &= \mathrm{E}\left[(\epsilon_t+\theta\epsilon_{t-1})^2\right] \\ &= \left(1+\theta^2\right)\sigma^2\end{aligned} \tag{2.2.3}$$

The first autocovariance is given by

$$\begin{aligned}\gamma_1 &= \mathrm{E}\left[(Y_t-\mu)(Y_{t-1}-\mu)\right] \\ &= \mathrm{E}\left[(\epsilon_t+\theta\epsilon_{t-1})(\epsilon_{t-1}+\theta\epsilon_{t-2})\right] \\ &= \theta\sigma^2\end{aligned} \tag{2.2.4}$$

All higher autocovariances γ_j with $j > 1$ are zero. The absolute value of the sum of the autocovariances is a finite number:

$$\sum_{j=0}^{\infty}|\gamma_j| = \left(1+\theta^2\right)\sigma^2 + |\theta\sigma^2| < \infty \tag{2.2.5}$$

Thus, if $\{\epsilon_t\}$ is Gaussian white noise, the MA(1) is ergodic for all moments.

We can express the MA(1) process in matrices as well:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} + \begin{bmatrix} \theta & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \theta & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \theta & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \theta & 1 \end{bmatrix} \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}$$
$$\Leftrightarrow \boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\epsilon} \tag{2.2.6}$$

where $\boldsymbol{y}, \boldsymbol{\mu}$ are $T \times 1$ column vectors, $\boldsymbol{B}$ a $T \times (T+1)$ matrix and $\boldsymbol{\epsilon}$ a $(T+1) \times 1$ column vector. The variance-covariance matrix of $\boldsymbol{y}$ is given by

$$\begin{aligned} \boldsymbol{\Sigma_y} \equiv \operatorname{Var}(\boldsymbol{y}) &= \operatorname{E}\left[(\boldsymbol{B}\boldsymbol{\epsilon})(\boldsymbol{B}\boldsymbol{\epsilon})'\right] \\ &= \operatorname{E}\left[\boldsymbol{B}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{B}'\right] \\ &= \boldsymbol{B}\boldsymbol{I}\boldsymbol{B}' \\ &= \boldsymbol{B}\boldsymbol{B}' \end{aligned} \tag{2.2.7}$$

as $\operatorname{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \operatorname{Var}(\boldsymbol{\epsilon}) = \boldsymbol{I}$ under the assumption that $\operatorname{Var}(\epsilon_t) = 1 \ \forall \ t \in \{0, 1, 2, \dots, T\}$. Thus, the $T \times T$ variance-covariance matrix of $\boldsymbol{y}$, $\boldsymbol{\Sigma_y}$, takes the following form:

$$\boldsymbol{\Sigma_y} = \begin{bmatrix} 1+\theta^2 & \theta & 0 & 0 & \dots & 0 & 0 \\ \theta & 1+\theta^2 & \theta & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \theta & 1+\theta^2 \end{bmatrix}$$

with the variances of y_t, $1 + \theta^2$, on the main diagonal.

2.2.2 MA(q) Process

Let's now take a look at a **q*th*-order moving average process**:

$$Y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q} \tag{2.2.8}$$

As in the MA(1) process, its expected value is given by μ:

$$\operatorname{E}[Y_t] = \mu \tag{2.2.9}$$

The MA(q) process' variance is

$$\begin{aligned} \gamma_0 &= \operatorname{Var}\left[\mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}\right] \\ &= \sigma^2 + \theta_1^2\sigma^2 + \theta_2^2\sigma^2 + \cdots + \theta_q^2\sigma^2 \\ &= \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma^2 \end{aligned} \tag{2.2.10}$$

as the white noise random variables are assumed to be independent and the variance of a sum of independent variables equals the sum of these variables' variances.

The covariance of a MA(q) process is zero for $j > q$ as Y_t and Y_{t-j} have no common random variables. For $j \leq q$, there are common random variables:

$$\begin{aligned}\epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_j\epsilon_{t-j} + \cdots + \theta_q\epsilon_{t-q} \\ \epsilon_{t-j} + \cdots + \theta_{q-j}\epsilon_{t-q} + \cdots + \theta_q\epsilon_{t-j-q}\end{aligned}$$

It follows that the autocovariance is given by

$$\begin{aligned}\gamma_j &= \mathrm{E}\big[\,(\epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_j\epsilon_{t-j} + \cdots + \theta_q\epsilon_{t-q}) \\ &\qquad (\epsilon_{t-j} + \cdots + \theta_{q-j}\epsilon_{t-q} + \cdots + \theta_q\epsilon_{t-j-q})\,\big] \\ &= (\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \cdots + \theta_q\theta_{q-j})\,\sigma^2 \end{aligned} \tag{2.2.11}$$

which implies that a MA(q) process is stationary and ergodic (if ϵ_t is Gaussian).

2.2.3 MA(∞) Process

A MA(∞) process is defined as

$$\begin{aligned}Y_t &= \mu + \psi_0\epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \ldots \\ &= \mu + \sum_{k=0}^{\infty}\psi_k\epsilon_{t-k}\,. \end{aligned} \tag{2.2.12}$$

For stationarity and ergodicity, we have to assume that

$$\sum_{k=0}^{\infty}|\psi_k| < \infty \tag{2.2.13}$$

which also implies that

$$\sum_{k=0}^{\infty}\psi_k^2 < \infty\,. \tag{2.2.14}$$

Then, the expected values equals μ as $\mathrm{E}[\epsilon_t] = 0\;\forall\, t$:

$$\mathrm{E}[Y_t] = \mu \tag{2.2.15}$$

Its variance and covariances are given by:

$$\begin{aligned}\gamma_0 &= \mathrm{Var}\left(\mu + \sum_{k=0}^{\infty}\psi_k\epsilon_{t-k}\right) \\ &= \left(\sum_{k=0}^{\infty}\psi_k^2\right)\sigma^2 \end{aligned} \tag{2.2.16}$$

$$\begin{aligned}\gamma_j &= \mathrm{Cov}\left(\mu + \sum_{k=0}^{\infty}\psi_k\epsilon_{t-k}, \mu + \sum_{k=0}^{\infty}\psi_k\epsilon_{t-j-k}\right) \\ &= \left(\sum_{k=0}^{\infty}\psi_{j+k}\psi_k\right)\sigma^2 \end{aligned} \tag{2.2.17}$$

Coefficients summable in absolute values imply absolutely summable autocovariances:

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty \tag{2.2.18}$$

If the ϵ's are Gaussian, a MA(∞) process is ergodic.

2.3 Autoregressive (AR) Processes

2.3.1 AR(1) Process

A first-order autoregressive process (AR(1) process) is defined as

$$Y_t = c + \phi Y_{t-1} + \epsilon_t \, . \tag{2.3.1}$$

$\{\epsilon_t\}$ is a white noise process with $\mathrm{E}[\epsilon_t] = 0, \mathrm{Var}(\epsilon_t) = \sigma^2 \ \forall \ t$ and $\mathrm{Cov}(\epsilon_t, \epsilon_\tau) = 0 \ \forall \ t \neq \tau$. The AR(1) process can be expressed as an MA(∞) process

$$\begin{aligned} Y_t &= (c + \epsilon_t) + \phi (c + \epsilon_{t-1}) + \phi^2 (c + \epsilon_{t-2}) + \dots \\ &= \frac{c}{1-\phi} + \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \dots \end{aligned} \tag{2.3.2}$$

with $\mu = c/(1-\phi)$ and $\psi_j = \phi^j$, under the assumption that $|\phi| < 1$. Thus, substituting $\psi_j = \phi^j$ into (2.2.15), (2.2.16) and (2.2.17) yields:

$$\mathrm{E}[Y_t] = \frac{c}{1-\phi} \tag{2.3.3}$$

$$\begin{aligned} \gamma_0 &= \left(1 + \left(\phi^1\right)^2 + \left(\phi^2\right)^2 + \left(\phi^3\right)^2 + \dots\right)\sigma^2 \\ &= \frac{\sigma^2}{1-\phi^2} \end{aligned} \tag{2.3.4}$$

$$\begin{aligned} \gamma_j &= \left(\phi^j + \phi^{j+2} + \phi^{j+4} + \dots\right)\sigma^2 \\ &= \phi^j \left(1 + \phi^2 + \phi^4 + \dots\right)\sigma^2 \\ &= \frac{\sigma^2}{1-\phi^2}\phi^j \left[= \gamma_0 \phi^j\right] \end{aligned} \tag{2.3.5}$$

2.3.2 AR(2) Process

Let's now take a look at an AR(2) process:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t \, . \tag{2.3.6}$$

We rewrite the process using the lag operator, L:

$$\left(1 - \phi_1 \mathrm{L} - \phi_2 \mathrm{L}^2\right) Y_t = c + \epsilon_t \tag{2.3.7}$$

Substituting the lag operator with a variable z and setting the resulting equation to zero yields

$$\left(1 - \phi_1 z - \phi_2 z^2\right) = 0 \, . \tag{2.3.8}$$

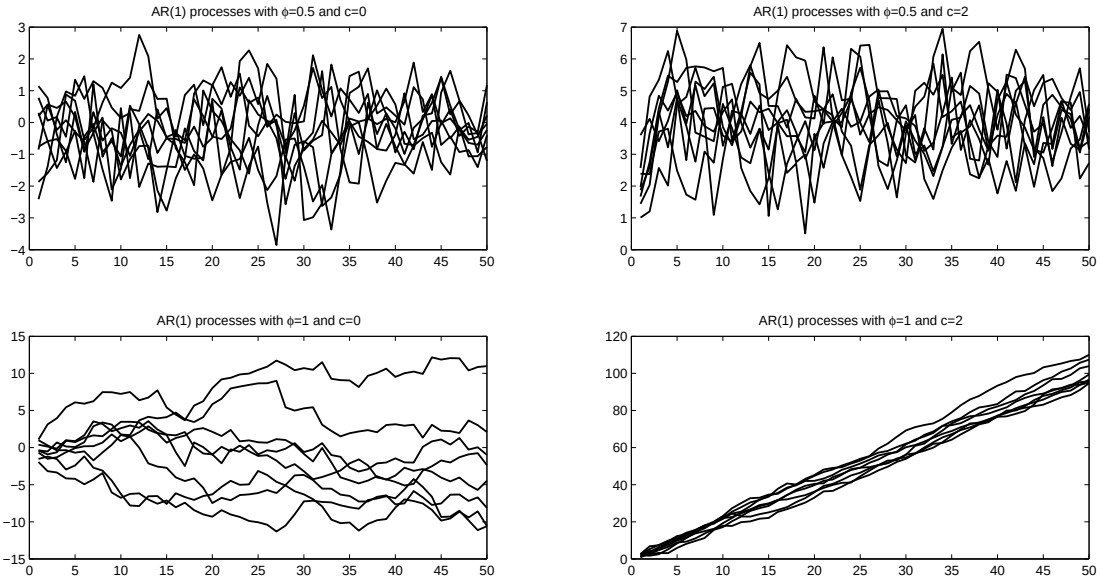

Figure 2: AR(1) processes with different parameters: The upper processes are stationary AR(1) processes, the lower ones non-stationary random-walks with and without drift.

Solving (2.3.8) for z is the same as finding the eigenvalues of the $\boldsymbol{F}$-matrix in (1.2.2) with $z_1 = \frac{1}{\lambda_1}, \ldots, z_p = \frac{1}{\lambda_p}$, that is the z_j's are the roots of the lag polynomial and the inverse of the eigenvalues λ_j. As the process is stable for $|\lambda_j| < 1$, the equivalent condition for **stationarity** is that $|z_j| > 1$. If the eigenvalues are less than one in absolute value, the lag polynomial is invertible:

$$\begin{aligned}\psi(\mathrm{L}) &\equiv \left(1 - \phi_1 \mathrm{L} - \phi_2 \mathrm{L}^2\right)^{-1} \\ &= \left(1 - \lambda_1 \mathrm{L}\right)^{-1} \left(1 - \lambda_2 \mathrm{L}\right)^{-1} \\ &= \left(1 + \lambda_1 \mathrm{L} + \lambda_1^2 \mathrm{L}^2 + \ldots\right) \left(1 + \lambda_2 \mathrm{L} + \lambda_2^2 \mathrm{L}^2 + \ldots\right) \\ &= 1 + \psi_1 L + \psi_2 L^2 + \ldots \end{aligned} \tag{2.3.9}$$

In other words: We can write the inverse of a lag polynomial of a stationary AR(2) process, $\psi(\mathrm{L})$, as a product of two infinite sums. The ψ_j's can be obtained by expanding (2.3.9) or just using equation (1.2.4) which was derived from a pth-order difference equation by taking the 11-element of the $\mathbf{F}$-matrix raised to the jth power.

Thus, given our original AR(2) process is stationary, we can rewrite it as

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{2.3.10}$$

which is the MA(∞) representation of the AR(2) process with $\mu = c/\left(1 - \phi_1 - \phi_2\right)$.

Substituting $c = \mu\left(1 - \phi_1 - \phi_2\right)$ into (2.3.6) yields

$$Y_t - \mu = \phi_1 \left(Y_{t-1} - \mu\right) + \phi_1 \left(Y_{t-2} - \mu\right) + \epsilon_t \,. \tag{2.3.11}$$

We can write the autocovariance γ_j by multiplying both sides with $Y_{t-j} - \mu$ and apply the expectation operator

$$\begin{aligned} \mathrm{E}\left[(Y_t - \mu)(Y_{t-j} - \mu)\right] &= \phi_1 \mathrm{E}\left[(Y_{t-1} - \mu)(Y_{t-j} - \mu)\right] \\ &+ \phi_2 \mathrm{E}\left[(Y_{t-2} - \mu)(Y_{t-j} - \mu)\right] \\ &+ \mathrm{E}\left[\epsilon_t (Y_{t-j} - \mu)\right] \end{aligned} \tag{2.3.12}$$

which can be expressed for $j > 0$ in terms of covariances and correlations as

$$\begin{aligned} \gamma_j &= \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2} \\ \Leftrightarrow \rho_j &= \phi_1 \rho_{j-1} + \phi_2 \rho_{j-2} \,. \end{aligned} \tag{2.3.13}$$

Using the fact that $\rho_0 = 1$ and $\rho_{-j} = \rho_j$, it is straightforward to derive formulas for the autocovariances.

2.3.3 AR(p) Process

The AR(p) process is given by

$$\begin{aligned} Y_t &= c + \sum_{j=1}^{p} \phi_j Y_{t-j} + \epsilon_t \\ &= c + \sum_{j=1}^{n} \phi_j \mathrm{L}^j Y_t + \epsilon_t \,. \end{aligned} \tag{2.3.14}$$

Again, the process is stationary if the eigenvalues of the **F**-matrix are less than one in absolute value or, equivalently, the roots of the Lag polynomial

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \tag{2.3.15}$$

exceed one in absolute value. In this case, the lag polynomial is invertible, that is it can be written as the product of p infinite sums:

$$\left(1 - \phi_1 \mathrm{L} - \phi_2 \mathrm{L}^2 - \cdots - \phi_p \mathrm{L}^p\right)^{-1} = \frac{1}{(1 - \lambda_1 \mathrm{L}) \ldots (1 - \lambda_p \mathrm{L})} \tag{2.3.16}$$

$$= \left(1 + \lambda_1 \mathrm{L} + \lambda_1^2 \mathrm{L}^2 + \ldots\right) \ldots \tag{2.3.17}$$

$$\left(1 + \lambda_p \mathrm{L} + \lambda_p^2 \mathrm{L}^2 + \ldots\right) \tag{2.3.18}$$

$$= 1 + \psi_1 \mathrm{L} + \psi_2 \mathrm{L}^2 + \ldots \tag{2.3.19}$$

With $\psi_j = c_1 \lambda_1^j + \cdots + c_p \lambda_p^j$. Thus, there is a MA($\infty$) representation for stationary AR(p) processes:

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{2.3.20}$$

with $\mu = c/\left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right)$.

Following the same logic as for the AR(2) process, the variance and the autocovariances are given by

$$\gamma_j = \begin{cases} \phi_1 \gamma_{j-1} + \phi_2 \gamma_{j-2} + \cdots + \phi_p \gamma_{j-p} & \text{for } j = 1, 2, \ldots \\ \phi_1 \gamma_1 + \phi_2 \gamma_2 + \cdots + \phi_p \gamma_p + \sigma^2 & \text{for } j = 0 \end{cases} .$$

2.4 ARMA(p,q) Processes

An ARMA(p,q) process combines the concepts of the MA(q) and AR(p) processes:

$$Y_t = c + \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{2.4.1}$$

Provided that the roots of the autoregressive part's lag-polynomial lie outside the unit circle, the polynomial is invertible and the process can be epxressed as

$$Y_t = \mu + \psi(\mathrm{L})\,\epsilon_t \tag{2.4.2}$$

with

$$\psi(\mathrm{L}) = \frac{(1 + \theta_1 \mathrm{L} + \theta_2 \mathrm{L}^2 + \cdots + \theta_q \mathrm{L}^q)}{(1 - \phi_1 \mathrm{L} - \phi_2 \mathrm{L}^2 - \cdots - \phi_p \mathrm{L}^p)}$$
$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}\,.$$

2.5 Maximum Likelihood Estimation

We want to estimate the model parameters $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)'$ of an ARMA process

$$Y_t = c + \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

by means of the **maximum likelihood** method. Therefore, we need to make assumptions about the distribution of ϵ_t. In general, we will assume that $\epsilon_t \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$. Our aim is to choose the parameters in way that maximizes the probability of drawing the sample we actually observe:

$$\begin{aligned}
&f_{Y_T, Y_{T-1}, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) \\
=&f_{Y_T|Y_{T-1}, \ldots, Y_1}(y_T|y_{T-1}, \ldots, y_1; \boldsymbol{\theta})\, f_{Y_{T-1}, Y_{T-2}, \ldots, Y_1}(y_{T-1}, y_{T-2}, \ldots, y_1; \boldsymbol{\theta}) \\
=&f_{Y_1}(y_1; \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta})
\end{aligned} \tag{2.5.1}$$

2.5.1 Gaussian AR(1) Estimation

For the AR(1) process

$$Y_t = c + \phi Y_{t-1} + \epsilon_t \tag{2.5.2}$$

we want to estimate $\boldsymbol{\theta} = (c, \phi, \sigma^2)'$. We know that for Gaussian ϵ_t, the first observation is normally distributed, $Y_1 \sim \mathcal{N}(c/(1-\phi), \sigma^2/(1-\phi^2))$. Moreover, the effects of $Y_{t-2}, \ldots, Y_1$ on Y_t work only through Y_{t-1}, thus

$$f_{Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta})\,. \tag{2.5.3}$$

Again, due to Gaussian ϵ_t, Y_t conditional on Y_{t-1} is normally distributed: $Y_t|Y_{t-1} \sim \mathcal{N}(c + \phi y_{t-1}, \sigma^2)$. With this information, we can easily set up the

likelihood function (2.5.1). The log likelihood function can be found by taking logs of the likelihood function:

$$\mathscr{L}(\boldsymbol{\theta}) = \log f_{Y_1}(y_1;\boldsymbol{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t, y_{t-1}|\boldsymbol{\theta}) \tag{2.5.4}$$

This function is also called the **exact maximum likelihood function** and yields consistent estimators under the assumption of stationarity. Moreover, numerical methods for estimation are required. Conditioning on the first observation, i.e. dropping the first summand of the right-hand side of (2.5.4), yields **conditional maximum likelihood** estimates which are consistent under non-stationarity and can be estimated by a simple OLS regression of Y_t on its first p lags. Substituting the Gaussian probability density function and maximizing (2.5.4) with respect to $\boldsymbol{\theta}$ yields estimates for $\boldsymbol{\theta}$:

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta}) = &-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log[\sigma^2/(1-\phi^2)] \\ &- \frac{\{y_1 - [c\,(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)} - [(T-1)/2]\log(2\pi) \\ &- [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right]\end{aligned} \tag{2.5.5}$$

2.5.2 Gaussian AR(p) Estimation

In order to estimate $\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \sigma^2)'$ for an AR(p) process

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t \tag{2.5.6}$$

we first rewrite the likelihood function (2.5.1):

$$\begin{aligned}&f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1;\boldsymbol{\theta}) \\ =&f_{Y_p,Y_{p-1},\ldots,Y_1}(y_p, y_{p-1}, \ldots, y_1;\boldsymbol{\theta}) \prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1}(y_t|y_{t-1}, \ldots, y_1;\boldsymbol{\theta})\end{aligned} \tag{2.5.7}$$

Let's collect the first p observations in a random vector $\boldsymbol{y_p} = (Y_1, \ldots, Y_p)'$. Its variance-covariance matrix is given by

$$\begin{aligned}\sigma^2\boldsymbol{V}_p \equiv &\operatorname{Var}(\boldsymbol{y_p}) \\ = &\begin{bmatrix} \mathrm{E}\,(Y_1-\mu)^2 & \mathrm{E}\,(Y_1-\mu)(Y_2-\mu) & \ldots & \mathrm{E}\,(Y_1-\mu)(Y_p-\mu) \\ \mathrm{E}\,(Y_2-\mu)(Y_1-\mu) & \mathrm{E}\,(Y_2-\mu)^2 & \ldots & \mathrm{E}\,(Y_2-\mu)(Y_p-\mu) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\,(Y_p-\mu)(Y_1-\mu) & \mathrm{E}\,(Y_p-\mu)(Y_2-\mu) & \ldots & \mathrm{E}\,(Y_p-\mu)^2 \end{bmatrix} \\ = &\begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix}\end{aligned} \tag{2.5.8}$$

Due to the assumption of Gaussian ϵ_t, the random vector $\boldsymbol{y}_p$ is normally distributed, $\boldsymbol{y}_p \sim \mathcal{N}(\boldsymbol{\mu}_p, \sigma^2 \boldsymbol{V}_p)$, where $\boldsymbol{\mu}_p$ is a $p \times 1$ vector containing the expected values of $Y_1, \ldots, Y_p$ which are given by $\mu = c/(1-\phi_1-\phi_2-\cdots-\phi_p)$. Thus, we have can substitute the first factor of (2.5.7) with a multivariate normal distribution.

Y_t does only depend on the previous p observations, so we can rewrite

$$f_{Y_t|Y_{t-1},\ldots,Y_1}(y_t|y_{t-1},\ldots,y_1;\boldsymbol{\theta})$$

as

$$f_{Y_t|Y_{t-1},\ldots,Y_{t-p}}(y_t|y_{t-1},\ldots,y_{t-p};\boldsymbol{\theta})$$

from which we know is a normal distribution as $Y_t|Y_{t-1},\ldots,Y_{t-p} \sim \mathcal{N}(c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}, \sigma^2)$. Again, it is straightforward to set up the log-likelihood function to maximize with respect to $\boldsymbol{\theta}$.

2.5.3 Gaussian MA(1) Estimation

Let's now set up the likelihood function for a MA(1) process

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \tag{2.5.9}$$

We do not observe the ϵ_t's, but it follows from

$$\epsilon_t = Y_t - \theta\epsilon_{t-1} - \mu \tag{2.5.10}$$

that if we set the value of ϵ_0 to zero, $\epsilon_0 = 0$, we can calculate the whole sequence of $\{\epsilon_1,\ldots,\epsilon_T\}$ in our sample. Thus, given our sample realisation of $\{y_1,\ldots,y_T\}$ and the initial condition $\epsilon_0 = 0$, all ϵ_t's are given as well. In the following, we're deriving the **conditional likelihood function** for a MA(1) process, specifically conditional on $\epsilon_0 = 0$:

$$\begin{aligned} &f_{Y_T,\ldots,Y_1|\epsilon_0=0}(y_T,\ldots,y_1|\epsilon_0=0) \\ =&f_{Y_T|Y_{T-1},\ldots,Y_1,\epsilon_0=0}(y_T|y_{T-1},\ldots,y_1,\epsilon_0=0) f_{Y_{T-1},\ldots,Y_1|\epsilon_0=0}(y_{T-1},\ldots,y_1|\epsilon_0=0) \end{aligned} \tag{2.5.11}$$

$$=f_{Y_1|\epsilon_0=0}(y_1|\epsilon_0=0)\prod_{t=2}^{T} f_{Y_t|Y_{t-1},\ldots,Y_1,\epsilon_0=0}(y_t|y_{t-1},\ldots,y_1,\epsilon_0=0) \tag{2.5.12}$$

Both $Y_1|\epsilon_0 = 0$ and $Y_t|Y_{t-1},\ldots,Y_1,\epsilon_0 = 0$ are normally distributed, $Y_1|\epsilon_0 = 0 \sim \mathcal{N}(\mu,\sigma^2)$ and $Y_t|Y_{t-1},\ldots,Y_1,\epsilon_0 = 0 \Leftrightarrow Y_t|\epsilon_{t-1} \sim \mathcal{N}(\mu+\theta\epsilon_{t-1},\sigma^2)$.

2.5.4 Gaussian MA(q) Estimation

For the MA(q) process

$$Y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q} \tag{2.5.13}$$

we first condition on the first q observations:

$$\epsilon_0 = \epsilon_{-1} = \cdots = \epsilon_{-q+1} = 0 \tag{2.5.14}$$

This allows us to iterate all ϵ_t by

$$\epsilon_t = y_t - \mu - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q} \tag{2.5.15}$$

for $t = 1, 2, \ldots, T$. Define $\boldsymbol{\epsilon}_0 \equiv (\epsilon_0, \epsilon_{-1}, \ldots, \epsilon_{-q+1})'$, then

$$\begin{aligned}
&f_{Y_t, Y_{T-1}, \ldots, Y_1 | \boldsymbol{\epsilon}_0 = \mathbf{0}}(y_T, y_{T-1}, \ldots, y_1 | \boldsymbol{\epsilon}_0 = \mathbf{0}; \boldsymbol{\theta}) \\
=&f_{\epsilon_T, \epsilon_{T-1}, \ldots, \epsilon_1 | \boldsymbol{\epsilon}_0 = \mathbf{0}}(\epsilon_T, \epsilon_{T-1}, \ldots, \epsilon_1 | \boldsymbol{\epsilon}_0 = \mathbf{0}; \boldsymbol{\theta}) \\
=&f_{\epsilon_T | \epsilon_{T-1}, \ldots, \epsilon_1, \boldsymbol{\epsilon}_0 = \mathbf{0}}(\epsilon_T | \epsilon_{T-1}, \ldots, \epsilon_1, \boldsymbol{\epsilon}_0 = \mathbf{0}; \boldsymbol{\theta}) \\
&\cdot f_{\epsilon_{T-1}, \ldots, \epsilon_1, \boldsymbol{\epsilon}_0 = \mathbf{0}}(\epsilon_{T-1}, \ldots, \epsilon_1, \boldsymbol{\epsilon}_0 = \mathbf{0}; \boldsymbol{\theta}) \\
=&\prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon_t^2/2\sigma^2}.
\end{aligned} \tag{2.5.16}$$

Thus, the log likelihood function is given by

$$\mathscr{L}(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\epsilon_t^2}{2\sigma^2}. \tag{2.5.17}$$

2.5.5 General Remarks on M-Estimators

The maximum likelihood estimator is an M-estimator, that is its objective function is a sample average:

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} m(\boldsymbol{w}_t; \boldsymbol{\theta}) \tag{2.5.18}$$

with $m(\boldsymbol{w}_t; \boldsymbol{\theta}) = \log f(\boldsymbol{w}_t; \boldsymbol{\theta})$ and thus $Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \log f(\boldsymbol{w}_t; \boldsymbol{\theta})$, given that there is no serial correlation. The $m \times 1$ vector $\boldsymbol{w}_t$ collects all of our m variables we observe in time t. Our estimator $\hat{\boldsymbol{\theta}}$ is the $k \times 1$ parameter vector that maximizes (2.5.18).

Let $\boldsymbol{\theta}_0$ denote the true parameter vector and suppose that $Q_n(\boldsymbol{\theta})$ is concave over the parameter space for any data $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t)$. If there is a function $Q_0(\boldsymbol{\theta})$ that is uniquely maximized at $\boldsymbol{\theta}_0$ (**identification**) and $Q_n(\boldsymbol{\theta})$ converges in probability to $Q_0(\boldsymbol{\theta})$ (**pointwise convergence**), then $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, i.e. the estimator $\hat{\boldsymbol{\theta}}$ is consistent.

If $\{\boldsymbol{w}_t\}$ is ergodic stationary, then $Q_n(\boldsymbol{\theta})$ converges to $\mathrm{E}[m(\boldsymbol{w}_t; \boldsymbol{\theta})]$, that is Q_0 is given by

$$Q_0 = \mathrm{E}[m(\boldsymbol{w}_t; \boldsymbol{\theta})] \tag{2.5.19}$$

Thus, the identification condition for consistency can be restated that, for a concave function $m(\boldsymbol{w}_t; \boldsymbol{\theta})$, $\mathrm{E}[m(\boldsymbol{w}_t; \boldsymbol{\theta})]$ is uniquely maximized by $\boldsymbol{\theta}_0$.

2.5.6 Asymptotic Normality of ML-Estimators

First of all, define symbols for the gradient (vector of first derivatives) and the Hessian (matrix of second derivatives) of the m function:

$$\underset{(k\times 1)}{\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta})} \equiv \frac{\partial m(\boldsymbol{w}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{2.5.20}$$

$$\underset{(k\times k)}{\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta})} \equiv \frac{\partial \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \tag{2.5.21}$$

The first order condition for maximum of equation (2.5.18) can be expressed as

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}}) = \mathbf{0}\,. \tag{2.5.22}$$

Equation (2.5.22) can be rewrittten using the mean value theorem as

$$\begin{aligned}\frac{\partial Q_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} &= \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 Q_n(\bar{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &= \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) + \left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})\right]\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{0}\,.\end{aligned} \tag{2.5.23}$$

Rearranging yields:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}})\right]^{-1} \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) \tag{2.5.24}$$

For ergodic stationary $\{\boldsymbol{w}_t\}$, we know that

$$\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{H}(\boldsymbol{w}_t;\bar{\boldsymbol{\theta}}) \xrightarrow{p} \mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right] \tag{2.5.25}$$

and for i.i.d. observations

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}) \tag{2.5.26}$$

which allows us to apply the Slutzky theorem:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \left(\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right]\right)^{-1} \boldsymbol{\Sigma} \left(\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right]\right)^{-1}) \tag{2.5.27}$$

That is, $\hat{\boldsymbol{\theta}}$ is **asymptotically normal** with

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \left(\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right]\right)^{-1} \boldsymbol{\Sigma} \left(\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right]\right)^{-1} \tag{2.5.28}$$

under the assumptions that

1. $\{\boldsymbol{w}_t\}$ is ergodic stationary,

2. $\boldsymbol{\theta}_0$ is in the interior of parameter space $\boldsymbol{\Theta}$,
3. $\boldsymbol{m}(\boldsymbol{w}_t;\boldsymbol{\theta})$ is twice continously differentiable in $\boldsymbol{\theta}$ for any $\boldsymbol{w}_t$,
4. $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$,
5. $\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right]$ is nonsingular.

If $\mathrm{E}\left[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right] = \mathbf{0}$, then

$$\boldsymbol{\Sigma} = \mathrm{E}\left[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)'\right] \tag{2.5.29}$$

which we assume to be equal to the expected information matrix times minus one:

$$\mathrm{E}\left[\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\boldsymbol{s}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)'\right] = -\mathrm{E}\left[\boldsymbol{H}(\boldsymbol{w}_t;\boldsymbol{\theta}_0)\right] \tag{2.5.30}$$

So equation (2.5.31) can be estimated by either

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}})^{(1)} = -\left[\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{H}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}})\right]^{-1} \tag{2.5.31}$$

or

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}})^{(2)} = \left[\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{s}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}})\boldsymbol{s}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}})'\right]^{-1}. \tag{2.5.32}$$

Note that for the variance of any unbiased estimator $\hat{\boldsymbol{\theta}}$

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) \geq \left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]^{-1} \tag{2.5.33}$$

has to hold, that is the minimum variance of the estimator is larger than or equal to the inverse of the Fisher information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ (**Cramer-Rao lower bound**) with

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) \equiv -\mathrm{E}[\boldsymbol{H}(\boldsymbol{w}_t;\hat{\boldsymbol{\theta}})] \tag{2.5.34}$$

and thus our estimator is – under the assumptions made here – **asymptotically efficient** as (2.5.31) converges to $\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]^{-1}$.

2.6 Unit Root Processes

2.6.1 First Differencing

The AR(p) process

$$\left(1 - \phi_1\mathrm{L} - \phi_2\mathrm{L}^2 - \cdots - \phi_p\mathrm{L}^p\right) y_t = \epsilon_t \tag{2.6.1}$$

can be rewritten as

$$\begin{aligned}&\left[(1-\rho\mathrm{L}) - \left(\zeta_1\mathrm{L} + \zeta_2\mathrm{L}^2 + \cdots + \zeta_{p-1}\mathrm{L}^{p-1}\right)(1-\mathrm{L})\right] y_t = \epsilon_t \\ \Leftrightarrow\ & y_t = \rho y_{t-1} + \zeta_1\Delta y_{t-1} + \zeta_2\Delta y_{t-2} + \cdots + \zeta_{p-1}\Delta y_{t-p+1} + \epsilon_t\end{aligned} \tag{2.6.2}$$

with

$$\rho \equiv \phi_1 + \phi_2 + \cdots + \phi_p$$
$$\zeta_j \equiv -(\phi_{j+1} + \phi_{j+2} + \cdots + \phi_p) \text{ for } j = 1, 2, \ldots, p-1.$$

Let's assume that we're dealing with a **unit root process**, i.e. exactly one root of the characteristic polynomial is equal to one and all other roots lie outside the unit circle. Thus,

$$1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0$$

which implies that $\rho = 1$. Moreover, under $H_0 : \rho = 1$:

$$\left(1 - \phi_1 \mathrm{z} - \phi_2 \mathrm{z}^2 - \cdots - \phi_p \mathrm{z}^p\right) =$$
$$\left(1 - \zeta_1 \mathrm{z} - \zeta_2 \mathrm{z}^2 - \cdots - \zeta_{p-1} \mathrm{z}^{p-1}\right)(1 - \mathrm{z})$$

equals zero for $z = 1$. It follows that the lag-polynomial on the left-hand side of

$$\left(1 - \zeta_1 \mathrm{L} - \zeta_2 \mathrm{L}^2 - \cdots - \zeta_{p-1} \mathrm{L}^{p-1}\right) \Delta y_t = \epsilon_t$$

is invertible, which implies that first differencing of a unit root process yields a stationary process.

2.6.2 Dickey-Fuller Test for Unit Roots

Case 1: True Process: Random Walk. Regression: No Constant, No Time Trend

We assume the **true process** follows a random walk:

$$y_t = y_{t-1} + u_t. \tag{2.6.3}$$

We estimate the parameter by a linear regression

$$y_t = \rho y_{t-1} + u_t \tag{2.6.4}$$

where u_t is i.i.d. The OLS estimate for ρ is given as

$$\hat{\rho}_T = \frac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2}. \tag{2.6.5}$$

Two test statistics with limiting distributions can be calculated under $H_0 : \rho = 1$:

$$T(\hat{\rho}_T - 1) \xrightarrow{d} \frac{(1/2)\{W(1)^2 - 1\}}{\int_0^1 W(r)^2 dr} \tag{2.6.6}$$

The second test statistic is the usual t test:

$$t_T = \frac{\hat{\rho} - 1}{\hat{\sigma}_{\hat{\rho}_T}} \tag{2.6.7}$$

where $\hat{\sigma}_{\hat{\rho}_T}$ is the OLS standard error for $\hat{\rho}_T$. Be aware that (2.6.7) does not have a limiting Gauss distribution.

Case 2: True Process: Random Walk. Regression: Constant, No Time Trend

As in case 1, the **true process** is assumed to follow a random walk:

$$y_t = y_{t-1} + u_t. \tag{2.6.8}$$

Our regression model is given as

$$y_t = \alpha + \rho y_{t-1} + u_t. \tag{2.6.9}$$

We can use the same test statistics as in case 1 – both have limiting distributions, though they differ from those in case 1.

Case 3: True Process: Random Walk with Drift. Regression: Constant, No Time Trend

In this case, the **true process** is assumed to follow a random walk with drift:

$$y_t = \alpha + y_{t-1} + u_t \tag{2.6.10}$$

We model the process as:

$$y_t = \alpha + \rho y_{t-1} + u_t \tag{2.6.11}$$

Be aware that if the true process follows a random walk with drift, the time series will show up a time trend. The alternative hypothesis does not include a time trend but only a constant though. Thus, if we reject $H_0 : \rho = 1$ against $H_1 : \rho < 1$, (2.6.11) still won't be appropriate if our non-unit-root time series shows up a time trend but we did not include it in this regression.

Case 4: True Process: Random Walk with Drift. Regression: Constant, Time Trend

As in case 3, the true process follows:

$$y_t = \alpha + y_{t-1} + u_t \tag{2.6.12}$$

This time, we include a time trend in our regression:

$$y_t = \alpha + \rho y_{t-1} + \delta t + u_t \tag{2.6.13}$$

Augmented Dickey Fuller Test

The augmented Dickey Fuller works under the null that the true process follows an AR(p) process with unit root. This may also be interpreted as allowing for serial correlation in the random walk case. Thus, we use the form (2.6.2) to express the AR(p) process

$$y_t = \rho y_{t-1} + \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \epsilon_t \tag{2.6.14}$$

and apply our usual OLS estimation (a time trend or constant may be included). Under $H_0 : \rho = 1$, i.e. the process is a unit root process, we can use the usual test statistics. We may equivalently restate (2.6.14) as

$$\Delta y_t = \theta y_{t-1} + \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \epsilon_t \tag{2.6.15}$$

with $\theta \equiv \rho - 1$ and calculate test statistic under $H_0 : \theta = 0$. In practice, given that the true process follows an AR(p) process, we do not know the number of lags p. In order to choose the number of lags, we can apply the Akaike information criterion (AIC)

$$\text{AIC} = 2k - 2\mathscr{L}(\hat{\boldsymbol{\theta}}) \tag{2.6.16}$$

where k is the number of estimated parameters. We calculate the AIC for different numbers of lags and chose the model with the lowest AIC. Alternatively, the Bayesian information criterion (BIC) or Schwarz criterion

$$\text{BIC} = \ln(n)k - 2\mathscr{L}(\hat{\boldsymbol{\theta}}) \tag{2.6.17}$$

where n is our sample size.

2.7 Forecasting

2.7.1 Conditional Expectation Forecast

Let $Y^*_{t+1|t}$ denote a forecast of Y_{t+1} based on $\boldsymbol{X}_t$. To evaluate the usefulness of this forecast, we need to set up a **loss function**. The **mean squared error** of a forecast is given by

$$\text{MSE}(Y^*_{t+1|t}) \equiv \text{E}(Y_{t+1} - Y^*_{t+1|t})^2. \tag{2.7.1}$$

It can be shown that among all forecasting rules, the expectation of Y_{t+1} conditional on $\boldsymbol{X}_t$

$$Y^*_{t+1|t} = \text{E}(Y_{t+1}|\boldsymbol{X}_t) \tag{2.7.2}$$

minimizes the MSE (2.7.1).

2.7.2 Linear Projection Forecast

We may just consider forecasts $Y^*_{t+1|t}$ that are a linear function of $\boldsymbol{X}_t$:

$$Y^*_{t+1|t} = \boldsymbol{\alpha}' \boldsymbol{X}_t \tag{2.7.3}$$

If there exists an $\boldsymbol{\alpha}$ such that the forecast error $(Y_{t+1} - \boldsymbol{\alpha}' \boldsymbol{X}_t)$ is uncorrelated with $\boldsymbol{X}_t$, i.e.

$$\text{E}[(Y_{t+1} - \boldsymbol{\alpha}' \boldsymbol{X}_t)\boldsymbol{X}_t] = \boldsymbol{0}', \tag{2.7.4}$$

then $\boldsymbol{\alpha}' \boldsymbol{X}_t$ is called **linear projection** of Y_{t+1} on $\boldsymbol{X}_t$. Among linear forecasting rules, the linear projection produces the smallest MSE.

2.7.3 Box-Jenkins Modeling Philosophy

The Box-Jenkings approach to modeling time series consists of four steps:

1. Transform the data until covariance stationarity is given.
2. Set initial values of p and q for an ARMA(p.q) model (check sample and partial autocorrelations).
3. Estimate $\phi(\mathrm{L})$ and $\theta(\mathrm{L})$.
4. Diagnostic analysis to check model consistency with respect to observed time series.

3 Vector Autoregressive Models

3.1 Structural and Standard Form of the VAR

Let's start with the following system of equations:

$$
\begin{aligned}
y_{1,t} &= k_1 + b_{12}^{(0)} y_{2,t} + b_{13}^{(0)} y_{3,t} + \cdots + b_{1n}^{(0)} y_{n,t} + \sum_{j=1}^{p} \sum_{k=1}^{n} b_{1k}^{(j)} y_{k,t-j} + u_{1,t} \\
y_{2,t} &= k_2 + b_{21}^{(0)} y_{1,t} + b_{23}^{(0)} y_{3,t} + \cdots + b_{2n}^{(0)} y_{n,t} + \sum_{j=1}^{p} \sum_{k=1}^{n} b_{2k}^{(j)} y_{k,t-j} + u_{2,t} \\
&\vdots \\
y_{n,t} &= k_n + b_{n1}^{(0)} y_{1,t} + b_{n2}^{(0)} y_{2,t} + \cdots + b_{n(n-1)}^{(0)} y_{(n-1),t} + \sum_{j=1}^{p} \sum_{k=1}^{n} b_{nk}^{(j)} y_{k,t-j} + u_{n,t}
\end{aligned}
$$

That is, at each point in time t we observe n variables which depend on the contemporaneous realisations of all the other $n-1$ variables as well as p lags of all n variables. The parameter $b_{ik}^{(j)}$ is the direct influence of the realisation of variable k in period $t-j$ on the contemporaneous value of variable i. As a next step, bring all contemporaneous observations to the left-hand side:

$$
\begin{aligned}
y_{1,t} - b_{12}^{(0)} y_{2,t} - b_{13}^{(0)} y_{3,t} - \cdots - b_{1n}^{(0)} y_{n,t} &= k_1 + \sum_{j=1}^{p} \sum_{k=1}^{n} b_{1k}^{(j)} y_{k,t-j} + u_{1,t} \\
y_{2,t} - b_{21}^{(0)} y_{1,t} - b_{23}^{(0)} y_{3,t} - \cdots - b_{2n}^{(0)} y_{n,t} &= k_2 + \sum_{j=1}^{p} \sum_{k=1}^{n} b_{2k}^{(j)} y_{k,t-j} + u_{2,t} \\
&\vdots \\
y_{n,t} - b_{n1}^{(0)} y_{1,t} - b_{n2}^{(0)} y_{2,t} - \cdots - b_{n(n-1)}^{(0)} y_{(n-1),t} &= k_n + \sum_{j=1}^{p} \sum_{k=1}^{n} b_{nk}^{(j)} y_{k,t-j} + u_{n,t}
\end{aligned}
$$

and rewrite the system in matrix notation:

$$
\boldsymbol{B}_0 \boldsymbol{y}_t = \boldsymbol{k} + \boldsymbol{B}_1 \boldsymbol{y}_{t-1} + \boldsymbol{B}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{B}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t \tag{3.1.1}
$$

with

$$\underset{(n\times 1)}{\boldsymbol{y}_t} \equiv \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{n,t} \end{bmatrix}, \underset{(n\times 1)}{\boldsymbol{k}} \equiv \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}, \underset{(n\times 1)}{\boldsymbol{u}_t} \equiv \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ \vdots \\ u_{n,t} \end{bmatrix},$$

$$\underset{(n\times n)}{\boldsymbol{B}_0} \equiv \begin{bmatrix} 1 & -b_{12}^{(0)} & \dots & -b_{1n}^{(0)} \\ -b_{21}^{(0)} & 1 & \dots & -b_{2n}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ -b_{n1}^{(0)} & -b_{n2}^{(0)} & \dots & 1 \end{bmatrix}, \underset{(n\times n)}{\boldsymbol{B}_j} \equiv \begin{bmatrix} b_{11}^{(j)} & b_{12}^{(j)} & \dots & b_{1n}^{(j)} \\ b_{21}^{(j)} & b_{22}^{(j)} & \dots & b_{2n}^{(j)} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}^{(j)} & b_{n2}^{(j)} & \dots & b_{nn}^{(j)} \end{bmatrix}.$$

If $\boldsymbol{B}_0$ is invertible, we can write (3.1.1) as

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\epsilon}_t \tag{3.1.2}$$

with

$$\begin{aligned} \boldsymbol{c} &\equiv \boldsymbol{B}_0^{-1}\boldsymbol{k}, \\ \boldsymbol{\Phi}_s &\equiv \boldsymbol{B}_0^{-1}\boldsymbol{B}_s, \\ \boldsymbol{\epsilon}_t &\equiv \boldsymbol{B}_0^{-1}\boldsymbol{u}_t. \end{aligned}$$

This is the **standard form of the VAR**. We can rewrite (3.1.2) using the lag operator as

$$\boldsymbol{\Phi}(\mathrm{L})\,\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\epsilon}_t \tag{3.1.3}$$

with $\boldsymbol{\Phi}(\mathrm{L}) \equiv (\boldsymbol{I}_n - \boldsymbol{\Phi}_1\mathrm{L} - \boldsymbol{\Phi}_2\mathrm{L}^2 - \cdots - \boldsymbol{\Phi}_p\mathrm{L}^p)$. The eigenvalues of $\boldsymbol{F}$ are the values that satisfy

$$|\boldsymbol{I}_n\lambda^p - \boldsymbol{\Phi}_1\lambda^{p-1} - \boldsymbol{\Phi}_2\lambda^{p-2} - \cdots - \boldsymbol{\Phi}_p| = 0 \tag{3.1.4}$$

and the VAR(p) is **stationary** iff $|\lambda| < 1$ for all λ satisfying (3.1.4) or, equivalently, iff all z satisfying

$$|\underbrace{\boldsymbol{I}_n - \boldsymbol{\Phi}_1 z - \boldsymbol{\Phi}_2 z^2 - \cdots - \boldsymbol{\Phi}_p z^p}_{=\Phi(z)}| = 0 \tag{3.1.5}$$

lie **outside the unit circle**.

If the vector process is covariance stationary, the vector of expected values is given as

$$\boldsymbol{\mu} = (\boldsymbol{I}_n - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2 - \cdots - \boldsymbol{\Phi}_p)^{-1}\boldsymbol{c} \tag{3.1.6}$$

and we can rewrite (3.1.2) in terms of deviations from its mean:

$$\boldsymbol{y}_t - \boldsymbol{\mu} = \boldsymbol{\Phi}_1(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\Phi}_2(\boldsymbol{y}_{t-2} - \boldsymbol{\mu}) + \cdots + \boldsymbol{\Phi}_p(\boldsymbol{y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{\epsilon}_t \tag{3.1.7}$$

Let's collect the last p observations of our n variables in a vector and define:

$$\underset{(np\times 1)}{\boldsymbol{\xi}_t} \equiv \begin{bmatrix} \boldsymbol{y}_t - \boldsymbol{\mu} \\ \boldsymbol{y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix}, \underset{(np\times np)}{\boldsymbol{F}} \equiv \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \dots & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_p \\ \boldsymbol{I}_n & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_n & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{I}_n & \boldsymbol{0} \end{bmatrix},$$

$$\underset{(np\times 1)}{\boldsymbol{v}_t} \equiv \begin{bmatrix} \boldsymbol{\epsilon}_t \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix}.$$

This allows us to write the VAR(p) model in a VAR(1) form:

$$\boldsymbol{\xi}_t = \boldsymbol{F}\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t \tag{3.1.8}$$

Forward iteration yields:

$$\begin{aligned} \boldsymbol{\xi}_t &= \boldsymbol{F}\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t \\ \boldsymbol{\xi}_{t+1} &= \boldsymbol{F}\boldsymbol{\xi}_t + \boldsymbol{v}_{t+1} \\ \boldsymbol{\xi}_{t+2} &= \boldsymbol{F}\left[\boldsymbol{F}\boldsymbol{\xi}_t + \boldsymbol{v}_{t+1}\right] + \boldsymbol{v}_{t+2} \\ &= \boldsymbol{v}_{t+2} + \boldsymbol{F}\boldsymbol{v}_{t+1} + \boldsymbol{F}^2\boldsymbol{\xi}_t \\ &\vdots \\ \boldsymbol{\xi}_{t+s} &= \boldsymbol{v}_{t+s} + \boldsymbol{F}\boldsymbol{v}_{t+s-1} + \boldsymbol{F}^2\boldsymbol{v}_{t+s-2} + \cdots + \boldsymbol{F}^{s-1}\boldsymbol{v}_{t+1} + \boldsymbol{F}^s\boldsymbol{\xi}_t \end{aligned} \tag{3.1.9}$$

The first n rows of (3.1.9) are given as

$$\begin{aligned} \boldsymbol{y}_{t+s} =& \boldsymbol{\mu} + \boldsymbol{\epsilon}_{t+s} + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_{t+s-1} + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_{t+s-2} + \cdots + \boldsymbol{\Psi}_{s-1}\boldsymbol{\epsilon}_{t+1} \\ &+ \boldsymbol{F}_{11}^{(s)}\left(\boldsymbol{y}_t - \boldsymbol{\mu}\right) + \boldsymbol{F}_{12}^{(s)}\left(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}\right) + \cdots + \boldsymbol{F}_{1p}^{(s)}\left(\boldsymbol{y}_{t-p+1} - \boldsymbol{\mu}\right) \end{aligned} \tag{3.1.10}$$

with $\boldsymbol{\Psi}_j = \boldsymbol{F}_{11}^{(j)}$ and $\boldsymbol{F}_{11}^{(j)}$ standing for the upper left block of the $\boldsymbol{F}$ matrix raised to the jth power. If all the eigenvalues of $\boldsymbol{F}$ lie inside the unit circle, it can be shown that

$$\lim_{s\to\infty} \boldsymbol{F}^s = \boldsymbol{0}$$

and thus $\boldsymbol{y}_t$ can be expressed as

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\mu} + \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_{t-2} + \dots \\ &= \boldsymbol{\mu} + \boldsymbol{\Psi}\left(\mathrm{L}\right)\boldsymbol{\epsilon}_t \end{aligned} \tag{3.1.11}$$

with $\boldsymbol{\Psi}\left(\mathrm{L}\right) \equiv \left(\boldsymbol{I}_n + \boldsymbol{\Psi}_1\mathrm{L} + \boldsymbol{\Psi}_2\mathrm{L}^2 + \dots\right)$. This is the VMA($\infty$) representation of the VAR(p) process. The relationships presented so far imply that

$$\boldsymbol{\Phi}\left(\mathrm{L}\right)\boldsymbol{\Psi}\left(\mathrm{L}\right) = \boldsymbol{I}_n \tag{3.1.12}$$

and hence

$$\begin{aligned} \boldsymbol{I}_n &= \left(\boldsymbol{I}_n + \boldsymbol{\Psi}_1\mathrm{L} + \boldsymbol{\Psi}_2\mathrm{L}^2 + \dots\right)\left(\boldsymbol{I}_n - \boldsymbol{\Phi}_1\mathrm{L} - \boldsymbol{\Phi}_2\mathrm{L}^2 - \cdots - \boldsymbol{\Phi}_p\mathrm{L}^p\right) \\ &= \boldsymbol{I}_n + \left(\boldsymbol{\Psi}_1 - \boldsymbol{\Phi}_1\right)\mathrm{L} + \left(\boldsymbol{\Psi}_2 - \boldsymbol{\Phi}_1\boldsymbol{\Psi}_1 - \boldsymbol{\Phi}_2\right)\mathrm{L}^2 + \dots \end{aligned}$$

which can be used to iterate $\boldsymbol{\Psi}_s$:

$$\boldsymbol{\Psi}_s = \boldsymbol{\Phi}_1\boldsymbol{\Psi}_{s-1} + \boldsymbol{\Phi}_2\boldsymbol{\Psi}_{s-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Psi}_{s-p} \tag{3.1.13}$$

for $s \geq 1$.

Differentiating $\boldsymbol{y}_{t+s}$ with respect to $\boldsymbol{\epsilon}_t'$ yields

$$\frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{\epsilon}_t'} = \boldsymbol{\Psi}_s, \tag{3.1.14}$$

i.e. the (i,j) element of $\boldsymbol{\Psi}_s$ is the effect of a shock $\epsilon_{j,t}$ on variable $y_{i,t+s}$, $i,j \in \{1,\ldots,n\}$. We may be interested in the effects of a shock in $u_{j,t}$ on $y_{i,t+s}$. Define $\boldsymbol{D} \equiv \mathrm{Var}(\boldsymbol{u}_t)$, $\boldsymbol{\Omega} \equiv \mathrm{Var}(\boldsymbol{\epsilon}_t)$ and $\boldsymbol{A} \equiv \boldsymbol{B}_0^{-1}$. Recall that $\boldsymbol{\epsilon}_t = \boldsymbol{A}\boldsymbol{u}_t$ from which follows

$$\frac{\partial \boldsymbol{y}_{t+s}}{\partial \boldsymbol{u}_t'} = \boldsymbol{\Psi}_s\boldsymbol{A}. \tag{3.1.15}$$

Cholesky Decomposition

That is, to find the response matrix to a shock in $u_{j,t}$ we first need to find $\boldsymbol{A}$. $\boldsymbol{\Omega}$ is given by

$$\begin{aligned}\boldsymbol{\Omega} &= \mathrm{Var}(\boldsymbol{A}\boldsymbol{u}_t) \\ &= \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}'\end{aligned} \tag{3.1.16}$$

If we assume $\boldsymbol{A}$ to be a lower-triangular matrix with ones on the main diagonal, then we can find exactly one matrix $\boldsymbol{A}$ and $\boldsymbol{D}$ as $\boldsymbol{\Omega}$ is a positive definite symmetric matrix (**triangular factorization**), where $\boldsymbol{D}$ is a matrix with $d_{jj} \neq 0$ and $d_{ij} = 0$ for $i \neq j$. Define $\boldsymbol{P} \equiv \boldsymbol{A}\boldsymbol{D}^{1/2}$, then (3.1.16) can be written as

$$\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}' = \boldsymbol{P}\boldsymbol{P}' \tag{3.1.17}$$

with

$$\boldsymbol{P} = \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \ldots & 0 \\ a_{21}\sqrt{d_{11}} & \sqrt{d_{22}} & 0 & \ldots & 0 \\ a_{31}\sqrt{d_{11}} & a_{32}\sqrt{d_{22}} & \sqrt{d_{33}} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1}\sqrt{d_{11}} & a_{n2}\sqrt{d_{22}} & a_{n3}\sqrt{d_{33}} & \ldots & \sqrt{d_{nn}} \end{bmatrix}$$

which is the so-called **Cholesky decomposition** from which we can easily derive our $\boldsymbol{A}$ and $\boldsymbol{D}$ matrices.

3.2 Forecast Error and Variance Decomposition

The forecast of $\boldsymbol{y}_{t+s}$ given $\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots$ is given as

$$\begin{aligned}\hat{\boldsymbol{y}}_{t+s|t} =& \boldsymbol{\mu} + \boldsymbol{F}_{11}^{(s)}(\boldsymbol{y}_t - \boldsymbol{\mu}) + \boldsymbol{F}_{12}^{(s)}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}) \\ &+ \cdots + \boldsymbol{F}_{1p}^{(s)}(\boldsymbol{y}_{t-p+1} - \boldsymbol{\mu})\end{aligned} \tag{3.2.1}$$

with $\hat{\boldsymbol{y}}_{t+s|t} \equiv \mathrm{E}(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots)$, and thus the forecast error is

$$\boldsymbol{y}_{t+s} - \hat{\boldsymbol{y}}_{t+s|t} = \boldsymbol{\epsilon}_{t+s} + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_{t+s-1} + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_{t+s-2} + \cdots + \boldsymbol{\Psi}_{s-1}\boldsymbol{\epsilon}_{t+1}\,. \tag{3.2.2}$$

Hence, the mean squared error of the forecast is

$$\begin{aligned}\text{MSE}\left(\hat{\boldsymbol{y}}_{t+s|t}\right) &= \text{E}\left[\left(\boldsymbol{y}_{t+s}-\hat{\boldsymbol{y}}_{t+s|t}\right)\left(\boldsymbol{y}_{t+s}-\hat{\boldsymbol{y}}_{t+s|t}\right)'\right] \\ &= \boldsymbol{\Omega}+\boldsymbol{\Psi}_1\boldsymbol{\Omega}\boldsymbol{\Psi}_1'+\boldsymbol{\Psi}_2\boldsymbol{\Omega}\boldsymbol{\Psi}_2'+\cdots+\boldsymbol{\Psi}_{s-1}\boldsymbol{\Omega}\boldsymbol{\Psi}_{s-1}'\,. \end{aligned} \tag{3.2.3}$$

Using the fact that

$$\boldsymbol{\epsilon}_t=\boldsymbol{A}\boldsymbol{u}_t=\boldsymbol{a}_1u_{1t}+\boldsymbol{a}_2u_{2t}+\cdots+\boldsymbol{a}_nu_{nt} \tag{3.2.4}$$

where $\boldsymbol{a}_j$ is the jth column of matrix $\boldsymbol{A}$ and the fact that the u_{jt}'s are uncorrelated, we can write

$$\begin{aligned}\boldsymbol{\Omega} &= \text{E}\left(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\right) \\ &= \boldsymbol{a}_1\boldsymbol{a}_1'\text{Var}\left(u_{1t}\right)+\boldsymbol{a}_2\boldsymbol{a}_2'\text{Var}\left(u_{2t}\right)+\cdots+\boldsymbol{a}_n\boldsymbol{a}_n'\text{Var}\left(u_{nt}\right)\,. \end{aligned} \tag{3.2.5}$$

Substituting (3.2.5) into (3.2.3), we can write the MSE as the sum of n terms:

$$\begin{aligned}\text{MSE}\left(\hat{\boldsymbol{y}}_{t+s|t}\right) = \sum_{j=1}^{n}[&\text{Var}(u_{jt})(\boldsymbol{a}_j\boldsymbol{a}_j'+\boldsymbol{\Psi}_1\boldsymbol{a}_j\boldsymbol{a}_j'\boldsymbol{\Psi}_1' \\ &+\boldsymbol{\Psi}_2\boldsymbol{a}_j\boldsymbol{a}_j'\boldsymbol{\Psi}_2'+\cdots+\boldsymbol{\Psi}_{s-1}\boldsymbol{a}_j\boldsymbol{a}_j'\boldsymbol{\Psi}_{s-1}')]\end{aligned} \tag{3.2.6}$$

Each summand of (3.2.6) is the contribution of the variance of u_{jt} to the MSE of the s-period-ahead-forecast.

4 Cointegration Models

4.1 Definition of Cointegration

An $(n\times 1)$ vector time series (vector stochastic process) $\boldsymbol{y}_t$ is **cointegrated** if each of the series $y_{it}, i\in\{1,2,\ldots,n\}$ is integrated of order 1 (I(1)), i.e. a unit root process, with a linear combination of the processes $\boldsymbol{a}'\boldsymbol{y}_t$ being stationary, i.e. I(0), for some nonzero $(n\times 1)$ vector $\boldsymbol{a}$. The vector $\boldsymbol{a}$ is called **cointegrating relation**. In general, the vector stochastic process $\boldsymbol{y}_t$ is said to be CI(d,b) if all scalar stochastic processes y_{it} of $\boldsymbol{y}_t$ are I(d) and there exists a vector $\boldsymbol{a}'$ such that $\boldsymbol{a}'\boldsymbol{y}_t$ is I(d-b).

4.2 Cointegrating Vectors

If there exists a cointegrating vector $\boldsymbol{a}$, then it is not unique, as if $\boldsymbol{a}'\boldsymbol{y}_t$ is stationary, then any stochastic process $b\boldsymbol{a}'\boldsymbol{y}_t$ for any scalar $b\neq 0$. Obviously, the cointegrating vectors $\boldsymbol{a}$ and $b\boldsymbol{a}$ are linearly dependent.

In general, in case of an n-variable vector time series, there can be at most $h<n$ linearly independent $(n\times 1)$ cointegrating vectors $\boldsymbol{a}_1,\boldsymbol{a}_2,\ldots,\boldsymbol{a}_h$ such that $\boldsymbol{A}'\boldsymbol{y}_t$ is a stationary vector time series, with $\boldsymbol{A}'$ defined as

$$\underset{(h\times n)}{\boldsymbol{A}'}\equiv\begin{bmatrix}\boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_h'\end{bmatrix}. \tag{4.2.1}$$

The matrix $\boldsymbol{A}'$ is the **basis** for the space of cointegrating vectors. Since $\boldsymbol{y}_t$ is assumed to be I(1), its first difference vector $\Delta\boldsymbol{y}_t$ is stationary with a vector of expected values $\boldsymbol{\delta} \equiv \mathrm{E}(\Delta\boldsymbol{y}_t)$. First of all, define

$$\boldsymbol{u}_t \equiv \Delta\boldsymbol{y}_t - \boldsymbol{\delta}. \tag{4.2.2}$$

Write $\boldsymbol{u}_t$ applying the Wold representation theorem:[1]

$$\boldsymbol{u}_t = \boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_{t-2} + \cdots = \boldsymbol{\Psi}(\mathrm{L})\boldsymbol{\epsilon}_t \tag{4.2.3}$$

with $\boldsymbol{\Psi}(\mathrm{L}) \equiv \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \mathrm{L}^j$ and $\boldsymbol{\Psi}_0 \equiv \boldsymbol{I}_n$. We suppose that $\boldsymbol{\epsilon}_t$ has a zero expected value and its elements are pairwise uncorrelated, both contemporaneously and over time.

Then, **two conditions for stationarity** of $\boldsymbol{A}'\boldsymbol{y}_t$ have to hold:

$$\boldsymbol{A}'\boldsymbol{\Psi}(1) = \boldsymbol{0}, \tag{4.2.4}$$

$$\boldsymbol{A}'\boldsymbol{\delta} = \boldsymbol{0} \tag{4.2.5}$$

with $\boldsymbol{\Psi}(1) = \boldsymbol{I}_n + \boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2 + \boldsymbol{\Psi}_3 + \ldots$. To see that these two conditions have to hold, rewrite (4.2.2) as

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{y}_0 + \boldsymbol{\delta}\cdot t + \boldsymbol{u}_1 + \boldsymbol{u}_2 + \cdots + \boldsymbol{u}_t \\ &= \boldsymbol{y}_0 + \boldsymbol{\delta}\cdot t + \boldsymbol{\Psi}(1)\cdot(\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2 + \cdots + \boldsymbol{\epsilon}_t) + \boldsymbol{\eta}_t - \boldsymbol{\eta}_0 \end{aligned} \tag{4.2.6}$$

which is known as the **Beveridge-Nelson decomposition**. It follows from (4.2.6) that for stationarity of $\boldsymbol{A}'\boldsymbol{y}_t$, (4.2.4) and (4.2.5) have to hold.

Moreover, (4.2.4) implies that $\boldsymbol{\Psi}(\mathrm{L})$ is not invertible as $|\boldsymbol{\Psi}(z)| = 0$ for $z = 1$, i.e. the first-differences of a cointegrated process only have a VMA but no VAR(p) representation. Thus, a VAR(p) model is not appropriate to model cointegrated time series in first-differences although these first-differences are stationary.

Phillips's Triangular Representation

The cointegrating base $\boldsymbol{A}'$ can be written as

$$\begin{aligned} \boldsymbol{A}' &= \begin{bmatrix} 1 & 0 & \ldots & 0 & -\gamma_{1,h+1} & -\gamma_{1,h+2} & \cdots & -\gamma_{1,n} \\ 0 & 1 & \ldots & 0 & -\gamma_{2,h+1} & -\gamma_{2,h+2} & \cdots & -\gamma_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 & -\gamma_{h,h+1} & -\gamma_{h,h+2} & \cdots & -\gamma_{h,n} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I}_h & -\boldsymbol{\Gamma}' \end{bmatrix} \end{aligned} \tag{4.2.7}$$

[1] The Wold representation theorem states that any covariance-stationary time series y_t can be written as the sum of two time series, with one summand being deterministic and the other one being stochastic:

$$y_t = \mu_t + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

with $\mathbf{\Gamma}'$ being an $(h \times g)$ coefficient matrix and $g \equiv n - h$. Define the (by construction stationary) $(h \times 1)$ residual-vector $\boldsymbol{z}_t \equiv \boldsymbol{A}'\boldsymbol{y}_t$ for a set of cointegrating relations. The mean of $\boldsymbol{z}_t$ is given by $\boldsymbol{\mu}_1^* \equiv \mathrm{E}(\boldsymbol{z}_t)$, the deviation of $\boldsymbol{z}_t$ from its mean by $\boldsymbol{z}_t^* \equiv \boldsymbol{z}_t - \boldsymbol{\mu}_1^*$. By partioning $\boldsymbol{y}_t$ as

$$\underset{(n\times 1)}{\boldsymbol{y}_t} = \begin{bmatrix} \underset{(h\times 1)}{\boldsymbol{y}_{1t}} \\ \underset{(g\times 1)}{\boldsymbol{y}_{2t}} \end{bmatrix} \tag{4.2.8}$$

we can express $\boldsymbol{z}_t$ as

$$\boldsymbol{z}_t^* + \boldsymbol{\mu}_1^* = \begin{bmatrix} \boldsymbol{I}_h & -\mathbf{\Gamma}' \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{1t} \\ \boldsymbol{y}_{2t} \end{bmatrix} \tag{4.2.9}$$

or after rearranging

$$\underset{(h\times 1)}{\boldsymbol{y}_{1t}} = \underset{(h\times g)}{\mathbf{\Gamma}'} \cdot \underset{(g\times 1)}{\boldsymbol{y}_{2t}} + \underset{(h\times 1)}{\boldsymbol{\mu}_1^*} + \underset{(h\times 1)}{\boldsymbol{z}_t^*} . \tag{4.2.10}$$

VAR Representation

Let $\boldsymbol{y}_t$ be represented by an VAR(p) process:

$$\mathbf{\Phi}(\mathrm{L})\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\epsilon}_t \tag{4.2.11}$$

Further suppose that $\Delta\boldsymbol{y}_t$ has the following Wold representation:

$$(1 - \mathrm{L})\boldsymbol{y}_t = \boldsymbol{\delta} + \mathbf{\Psi}(\mathrm{L})\boldsymbol{\epsilon}_t \tag{4.2.12}$$

Premultiplying (4.2.12) by $\mathbf{\Phi}(\mathrm{L})$ yields

$$(1 - \mathrm{L})\mathbf{\Phi}(\mathrm{L})\boldsymbol{y}_t = \mathbf{\Phi}(\mathrm{L})\boldsymbol{\delta} + \mathbf{\Phi}(\mathrm{L})\mathbf{\Psi}(\mathrm{L})\boldsymbol{\epsilon}_t \tag{4.2.13}$$

and after substitution of (4.2.11) into (4.2.13)

$$(1 - \mathrm{L})\boldsymbol{\epsilon}_t = \mathbf{\Phi}(\mathrm{L})\boldsymbol{\delta} + \mathbf{\Phi}(\mathrm{L})\mathbf{\Psi}(\mathrm{L})\boldsymbol{\epsilon}_t \tag{4.2.14}$$

as $(1 - \mathrm{L})\boldsymbol{\alpha} = \mathbf{0}$. For this equality to hold, the following relationships must hold:

$$\mathbf{\Phi}(1)\boldsymbol{\delta} = \mathbf{0}, \tag{4.2.15}$$

$$\mathbf{\Phi}(1)\mathbf{\Psi}(1) = \mathbf{0}. \tag{4.2.16}$$

Let $\boldsymbol{\pi}'$ denote any row of $\mathbf{\Phi}(1)$, then $\boldsymbol{\pi}'\mathbf{\Psi}(1) = \mathbf{0}'$ and $\boldsymbol{\pi}'\boldsymbol{\delta} = 0$ imply that $\boldsymbol{\pi}$ is a cointegrating vector, that is $\boldsymbol{\pi}' = \boldsymbol{b}'\boldsymbol{A}'$ which must hold for any row of $\mathbf{\Phi}(1)$ and thus

$$\mathbf{\Phi}(1) = \boldsymbol{B}\boldsymbol{A}', \tag{4.2.17}$$

i.e. there exists an $(n \times h)$ matrix $\boldsymbol{B}$ for which (4.2.17) holds.

Error-Correction Representation

Any VAR can be written as

$$\boldsymbol{y}_t = \boldsymbol{\zeta}_1 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{\zeta}_2 \Delta \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\zeta}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\alpha} + \boldsymbol{\rho} \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{4.2.18}$$

with

$$\boldsymbol{\rho} \equiv \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_2 + \cdots + \boldsymbol{\Phi}_p, \tag{4.2.19}$$

$$\boldsymbol{\zeta}_s \equiv -(\boldsymbol{\Phi}_{s+1} + \boldsymbol{\Phi}_{s+2} + \cdots + \boldsymbol{\Phi}_p) \tag{4.2.20}$$

for $s = 1, 2, \ldots, p-1$. First differencing yields

$$\Delta \boldsymbol{y}_t = \boldsymbol{\zeta}_1 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{\zeta}_2 \Delta \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\zeta}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\alpha} + \boldsymbol{\zeta}_0 \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \tag{4.2.21}$$

with $\boldsymbol{\zeta}_0 \equiv \boldsymbol{\rho} - \boldsymbol{I}_n = -\boldsymbol{\Phi}(1)$. If our vector process is cointegrated, we can express (4.2.21) as

$$\Delta \boldsymbol{y}_t = \boldsymbol{\zeta}_1 \Delta \boldsymbol{y}_{t-1} + \boldsymbol{\zeta}_2 \Delta \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\zeta}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\alpha} - \boldsymbol{B} \boldsymbol{z}_{t-1} + \boldsymbol{\epsilon}_t \tag{4.2.22}$$

with $\boldsymbol{z}_t \equiv \boldsymbol{A}' \boldsymbol{y}_{t-1}$ and $\boldsymbol{\zeta}_0$ substituted with $-\boldsymbol{B}\boldsymbol{A}'$. This is the **error-correction representation** of the cointegrated system.

4.3 Testing for Cointegrating Relationships

Engle-Granger Two-Step-Procedure

To test for a cointegrating relationship, in a **first step** regress the level of one variable on the level of all other variables, back out the estimated error term $\hat{\epsilon}_t$ and check the time series of estimated error terms for stationarity, i.e. using a standard Dickey-Fuller test. If the null hypothesis of non-stationarity is rejected, go on to the **second step**: Estimate the error-correction model (4.2.22) by replacing the $(h \times 1)$ vector $\boldsymbol{z}_{t-1}$ by $\hat{\epsilon}_t$. This procedure can only account for **one cointegrating relationship** and the result may thus depend on the ordering of the variables, that is – in case of $n > 2$ where more than 1 cointegrating relationship could exist – how the regression in step 1 is performed.

5 Identification of Structural VARs

5.1 Identification under Stationarity

Recall the VAR(p) process with $\boldsymbol{\epsilon}_t \equiv \boldsymbol{W} \boldsymbol{u}_t$

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{W} \boldsymbol{u}_t \tag{5.1.1}$$

where $\boldsymbol{W}$ is equal to our $\boldsymbol{A}$ matrix from the Cholesky decomposition if we set the main diagonal elements of $\boldsymbol{W}$ equal to 1 (which we will assume in the following). Now our reduced form shocks' variance-covariance matrix is given by $\boldsymbol{\Omega} = \boldsymbol{W}\boldsymbol{W}'$, where we have set $\mathrm{Var}(\boldsymbol{u}_t) = \boldsymbol{I}_n$.

To identify our $\boldsymbol{W}$ matrix in the n variables case, we need to impose $n(n-1)/2$ restrictions on the elements of $\boldsymbol{W}$. Setting the (i,j)-element of $\boldsymbol{W}$, w_{ij}, equal to zero implies no contemporaneous effect of shock u_{jt} on variable y_{it}. Setting elements of $\boldsymbol{W}$ equal to zero, $\boldsymbol{W}$ must maintain its full rank – as losing one or more ranks (e.g. by setting one row equal to zero) results in a singular matrix, but $\boldsymbol{W}$ has to be invertible in order to write the standard form VAR in its structural form.

Stationary VAR(p) processes have a VMA(∞) representation

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{u}_t + \boldsymbol{\Psi}_1\boldsymbol{W}\boldsymbol{u}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{W}\boldsymbol{u}_{t-2} + \dots, \tag{5.1.2}$$

that is the total long-term impact of structural innovation u_j on y_i is given by the (i,j)-element of matrix

$$\boldsymbol{L} \equiv \boldsymbol{W} + \boldsymbol{\Psi}_1\boldsymbol{W} + \boldsymbol{\Psi}_2\boldsymbol{W} + \cdots = \boldsymbol{\Psi}(1)\boldsymbol{W}. \tag{5.1.3}$$

Instead of imposing $n(n-1)/2$ restrictions on our $\boldsymbol{W}$ matrix, we may also impose m restrictions on $\boldsymbol{L}$ and the remaining $n(n-1)/2-m$ restrictions on $\boldsymbol{W}$. As our process is assumed to be stationary, $\boldsymbol{\Psi}(1)$ is just the inverse of $\boldsymbol{\Phi}(1)$.

5.2 Identification under Cointegration

In case of cointegration, the VAR(p) has no VMA(∞) representation in levels due to non-stationarity. Still, there is a VMA(∞) representation in first differences. The VAR in levels has the following Beveridge-Nelson decomposition

$$\begin{aligned}\boldsymbol{y}_t &= \boldsymbol{\delta}t + \boldsymbol{\Psi}\sum_{i=1}^{t}\boldsymbol{\epsilon}_i + \boldsymbol{\eta}_t + \boldsymbol{y}_0 - \boldsymbol{\eta}_0 \\ &= \boldsymbol{\delta}t + \boldsymbol{\Psi}\boldsymbol{W}\sum_{i=1}^{t}\boldsymbol{u}_i + \boldsymbol{\eta}_t + \boldsymbol{y}_0 - \boldsymbol{\eta}_0\end{aligned} \tag{5.2.1}$$

with $\boldsymbol{\Psi} \equiv \boldsymbol{\Psi}(1)$. The (i,j)-element of the impact matrix $\boldsymbol{P} \equiv \boldsymbol{\Psi}\boldsymbol{W}$ gives the effect of a random walk in u_j on variable y_i. Hence, restrictions concerning long-run effects of shocks in variable y_j on variable y_i are imposed by setting elements of the $\boldsymbol{P}$ matrix equal to zero.

5.3 Estimation Procedure

First of all, we estimate $\boldsymbol{\Psi}$ from the ML estimates of the VECM coefficients, that is estimate the VECM subject to $\boldsymbol{\zeta}_0 = -\boldsymbol{B}\boldsymbol{A}'$ and compute the orthogonal complement of $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{A}}_\perp$ and $\hat{\boldsymbol{B}}_\perp$. Then $\hat{\boldsymbol{\Psi}}$ can be computed as:

$$\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{A}}_\perp\left[\hat{\boldsymbol{B}}_\perp'\left(\boldsymbol{I}_n - \sum_{i=1}^{p-1}\hat{\boldsymbol{\zeta}}_{\boldsymbol{i}}\right)\hat{\boldsymbol{A}}_\perp\right]^{-1}\hat{\boldsymbol{B}}_\perp' \tag{5.3.1}$$

Let's now derive the likelihood function for our VAR(p) process. Assuming Gaussian $\boldsymbol{\epsilon}_t$, $\boldsymbol{\epsilon}_t$ has the following density function:

$$f_{\boldsymbol{\epsilon}_t}(\boldsymbol{\epsilon}_t) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{\epsilon}_t\right) \tag{5.3.2}$$

As the $\boldsymbol{\epsilon}_t$'s and the $\boldsymbol{y}_t$'s are just different sides of the same coin, we can write the log likelihood function as

$$\mathscr{L}(\boldsymbol{y};\boldsymbol{\Theta}) = -T\frac{n}{2}\ln(2\pi) - T\frac{1}{2}\ln|\boldsymbol{W}\boldsymbol{W}'| - \frac{1}{2}\sum_{t=1}^{T}\boldsymbol{\epsilon}_t(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{\epsilon}_t \tag{5.3.3}$$

6 Volatility Models

6.1 Autoregressive Conditional Heteroskedasticity (ARCH)

The AR(p) model

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t \tag{6.1.1}$$

with u_t white noise implies a constant unconditional variance σ^2 of u_t. Still, the conditional variance of u_t may change over time. A white noise process satisfying

$$u_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_m u_{t-m}^2 + w_t \tag{6.1.2}$$

with

$$\mathrm{E}(w_t) = 0,$$
$$\mathrm{E}(w_t w_\tau) = \begin{cases} \lambda^2 & \text{if } t = \tau \\ 0 & \text{if } t \neq \tau \end{cases},$$

i.e. the squared white noise process follows an AR(m) process, is called **autoregressive conditional heteroskedastic process** of order m. For u_t^2 to be covariance stationary, the roots of

$$1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_m z^m = 0 \tag{6.1.3}$$

have to lie outside the unit circle. For u_t^2 to be nonnegative, w_t has to be bounded from below by $-\zeta$ with $\zeta > 0$ and $\alpha_j \geq 0$ for $j = 1, 2, \ldots, m$. In this case, the unconditional white noise variance is given by

$$\sigma^2 = \mathrm{E}(u_t^2) = \zeta/(1 - \alpha_1 - \alpha_2 - \cdots - \alpha_m). \tag{6.1.4}$$

We may also specify that u_t satisfies

$$u_t = \sqrt{h_t} \cdot v_t \tag{6.1.5}$$

with $\{v_t\}$ an i.i.d. sequence with zero mean and unit variance. If h_t follows

$$h_t = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_m u_{t-m}^2 \tag{6.1.6}$$

then we can express u_t^2 as

$$h_t \cdot v_t^2 = h_t + w_t \tag{6.1.7}$$

implying that the conditional variance of w_t is not constant over time:

$$w_t = h_t \cdot (v_t^2 - 1) \tag{6.1.8}$$

6.2 Testing for ARCH Effects

Ljung-Box Statistics

The Ljung-Box statistics works under the null hypothesis $H_0 : u_t^2$ *is white noise* and is given by

$$Q = T(T+2)\sum_{\tau=1}^{L}\frac{\hat{\rho}_{\hat{u}^2}(\tau)}{T-\tau} \tag{6.2.1}$$

with $L \approx T/4$ and $Q \sim \chi^2(L)$. The τth-order error autocorrelations $\hat{\rho}_{\hat{u}^2}(\tau)$ are given by

$$\hat{\rho}_{\hat{u}^2}(\tau) \equiv \frac{\sum_{t=\tau+1}^{T}(\hat{u}_t^2-\hat{\sigma}^2)(\hat{u}_{t-\tau}^2-\hat{\sigma}^2)}{\sum_{t=1}^{T}(\hat{u}_t^2-\hat{\sigma}^2)^2} \tag{6.2.2}$$

where the $\hat{u}_t^2$ are the residuals from a ARMA(p,q) estimation.

Lagrange Multiplier Test

The Lagrange multiplier test works under the null $H_0 : u_t^2$ *is white noise* as well and can be obtained by first of all regressing the ARMA(p,q) residuals $\hat{u}_t^2$ on a constant and lagged residuals:

$$\hat{u}_t^2 = \alpha_0 + \alpha_1\hat{u}_{t-1}^2 + \alpha_2\hat{u}_{t-2}^2 + \cdots + \alpha_m\hat{u}_{t-m}^2 + \epsilon_t \tag{6.2.3}$$

Then our test statistics under $H_0 : \alpha_0 = \alpha_1 = \cdots = \alpha_m = 0$ is $\mathrm{TR}^2 \sim \chi^2(m)$.

Appendix

Probability Theory

A probability space is given by $(\Omega, \mathcal{F}, P)$ where Ω – the sample space – represents a set of all possible outcomes, where an outcome is defined as the result of a single execution of the underlying model. $\mathcal{F}$ is a σ-algebra: A σ-algebra is a set of subsets of Ω including the null set. Its elements are called events and thus $\mathcal{F}$ is called event space. The probability measure P is a function $P : \mathcal{F} \to [0,1]$, i.e. P maps an event (an element) of set $\mathcal{F}$ on a real number between 0 and 1.

Example

You draw a playing card from a set of four cards. The deck consists of the following 3 cards: 1×Ace, 1×King and 1×Queen. Ω is given by

$$\{\text{Ace}, \text{King}, \text{Queen}\}\ ,$$

$\mathcal{F}$ by

$$\{\{\}, \{\text{Ace}\}, \{\text{King}\}, \{\text{Queen}\}, \{\text{Ace}, \text{King}\},$$
$$\{\text{Ace}, \text{Queen}\}, \{\text{King}, \text{Queen}\}, \{\text{Ace}, \text{King}, \text{Queen}\}\}\ .$$

The probability of drawing Ace or King is given by $P(\{\text{Ace}, \text{King}\}) = 2/3$.

Convergence of Random Variables

A sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

where F and F_n are the cumulative distribution functions of X and X_n. This kind of convergence does **not** imply that the random variables' density functions converge as well. Convergence in distribution is also denoted as

$$X_n \xrightarrow{d} X \, .$$

Convergence in probability is defined as

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| \geq \epsilon) = 0$$

which is also denoted

$$X_n \xrightarrow{p} X \, .$$

Convergence in probability implies convergence in distribution, the opposite is not true.

A third kind of convergence is **almost sure convergence**:

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$