Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Het modelleren van multivariate economische tijdreeksen met vector autoregressieve processen

(Engelse titel: Modelling multivariate financial time series using vector autoregressive processes)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

Oskar Oostdam

Delft, Nederland
Juli 2019

BSc verslag TECHNISCHE WISKUNDE

"Het modelleren van multivariate economische tijdreeksen met vector autoregressieve processen"

(Engelse titel: "Modelling multivariate financial time series using vector autoregressive processes")

Oskar Oostdam

Technische Universiteit Delft

Begeleider

Dr. N. Parolya

Overige commissieleden

Drs. E.M. van Elderen

Dr. J.-J. Cai

Juli, 2019

Delft

Abstract

Time series analysis is used to predict future behaviour of processes and is widely used in the finance sector. In this paper we will analyse the modelling of multivariate time series of financial data using vector autoregressive processes. The goal is that the reader will understand the presented models and could theoretically perform time series analysis by himself.

Two specific models will be explained: the Vector Autoregressive model (VAR model) and the Vector Error Correction Model (VECM). We will describe various methods to analyse multivariate time series using these models, such as forecasting the process, variance decomposition of the forecast error, causality analysis and impulse response analysis. Examples of these models and analysis methods will be presented and investigated. Finally, we will perform a time series analysis with these models on Dutch indices and stock data. We conclude that real-world data often does not fit the VAR model and VECM requirements and that further improved models should be considered as well.

Contents

1 Introduction

Nowadays stock data, government indices, interest rates and other financial historical data of the past decades are easily obtainable for everyone. These historical data are widely used by companies and investors in order to obtain certain information of the data or to predict its future behaviour. Obtaining such information might be very useful to improve investing strategies, therefore one should definitely analyse historical data when striving for optimal returns. There are already many different existing models which am to gather as much information as possible of historical data. However, we will be looking at a specific group of models of so-called vector autoregressive processes. Most of the information about these models in this paper is based on Lütkepohl (2005). For simplicity we mostly applied the same notation as in this book.

In our models, we will be using multivariate time series. A univariate time series is a series of data points in some time order. However, a multivariate time series is a vector of combined univariate time series, in which each element of a multivariate time series vector represents a univariate time series. For multivariate time series we will mostly use the following notation throughout this paper:

y_t : Multivariate time series vector at time t.

u_t : Vector of errors made by a certain model at time t.

In the multivariate time series vector y_t we will be combining multiple univariate time series we want to investigate, which we will call the *variables of interest*. So if, $y_1, y_2, \ldots, y_K$ would be K different univariate time series, then we would create the K-dimensional multivariate time series vector

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Kt} \end{bmatrix}. \tag{1.1}$$

The models we will be looking at for analysing multivariate time series will be based on vector autoregressive processes. If we assume y_t to be a vector autoregressive process, then the value of each variable of interest at time t linearly depends on both the

1. previous values of the variable of interest,
2. previous values of the other variables of interest.

This means that if y_t has K variables of interest as in (1.1), then for the k-th variable of interest we could find constants $\nu_k, a_1, a_2, \ldots$ such that

$$y_{kt} = \nu_k + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}^T y_{t-1} + \begin{bmatrix} a_{K+1} \\ a_{K+2} \\ \vdots \\ a_{K+K} \end{bmatrix}^T y_{t-2} + \ldots \tag{1.2}$$

Note that $\nu_k, a_1, a_2, \ldots$ could be different for each variable of interest k. When we use models based on autoregressive processes, we will not only gather information of the individual variables of interest, but we could also gather information of the relationship with the other variables of interest. In this paper we will investigate two of these types of models, which are the Vector Autoregressive Model (VAR model) and the Vector Error Correction Model (VECM). In section 2 we will investigate the VAR model thoroughly. We will present everything that we need to apply the VAR model on multivariate time series. In addition, we will look at various methods to analyse a time series using the VAR model. In section 3 we will be presenting the VECM in a similar way. In section 4 we will perform a multivariate time series analysis with the presented models and methods on actual data of the Dutch stock market. Finally, we draw a conclusion of the strengths and the limitations of these time series models.

2 Vector Autoregressive Model

The VAR (Vector Autoregressive) model of a multivariate time series is based on the assumption that the time series is approximately a vector autoregressive process. If we assume y_t to be a VAR process of order p, then we assume y_t to approximately be a vector autoregressive process, where each variable of interest linearly depends on the previous p values of all variables of interest. Instead of the representation y_{kt} as in (1.2), the k-th variable of interest of a VAR process y_t of order p looks like

$$y_{kt} = \nu_k + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_K \end{bmatrix}^T y_{t-1} + \begin{bmatrix} a_{K+1} \\ a_{K+2} \\ \vdots \\ a_{2K} \end{bmatrix}^T y_{t-2} + \cdots + \begin{bmatrix} a_{(p-1)K+1} \\ a_{(p-1)K+2} \\ \vdots \\ a_{pK} \end{bmatrix}^T y_{t-p} + u_{kt},$$

where u_{kt} is the error made by the assumption that y_t is a vector autoregressive process. Now y_t is the vector that combines all variables of interest as in (1.1), hence we can write the VAR process y_t of order p, or the K-dimensional VAR(p) process as follows.

$$y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t, \quad t \in \mathbb{Z}, \tag{2.1}$$

with $y_t = (y_{1t}, y_{2t}, \ldots, y_{Kt})^T$ as the $(K \times 1)$ random vector with values of K variables of interest. Furthermore, the intercept ν is a fixed $(K \times 1)$ vector which can be used to express a non-zero mean of the process y_t and the matrices $A_1, A_2, \ldots, A_p$ are fixed $(K \times K)$ coefficient matrices. We also have the error terms, $u_t = (u_{1t}, u_{2t}, \ldots, u_{Kt})^T$, as a K-dimensional *white noise* process, which is defined as follows.

Definition 2.1. *The K-dimensional process u_t is called a white noise process if the following holds.*

1. $\mathbb{E}[u_t] = 0$,
2. $\mathbb{E}[u_t u_t^T] =: \Sigma_u$ *for all* t,
3. $\mathbb{E}[u_t u_s^T] = 0$ *for all* $s \neq t$.

We call Σ_u the covariance matrix of the process u_t and we assume throughout this paper Σ_u to be a nonsingular, i.e. Σ_u^{-1} exists. We also assume that all diagonal elements of Σ_u are non-zero. Mostly u_t is referred to as the *residuals* of the process in the literature.

2.1 Properties of the VAR model

In the following sections we will discuss important properties of the VAR model, which will give us a better understanding of the model. Later in this paper we will often refer back to these properties.

2.1.1 Stability

As stated in the VAR(p) model in (2.1), we have that $t \in \mathbb{Z}$ for a VAR(p) process y_t. At first it might look a little strange to let t also be able to have negative values, but this is simply because sometimes it is assumed that the starting point of the process happened in the infinite past. It will then obviously not be possible to start the time series at a certain finite time step. The question now is: what will happen with the stability of the VAR model when the starting point happened in the infinite past? Let us first consider the following lemma.

Lemma 2.1. *Let y_t be a K-dimensional VAR(1) process as in (2.1), where the process starts at y_{t-j-1}, then y_t can be generated as*

$$y_t \quad = (I_K + A_1 + \cdots + A_1^j)\nu + A_1^{j+1} y_{t-j-1} + \textstyle\sum_{i=0}^{j} A_1^i u_{t-i},$$

with I_K as the $(K \times K)$ identity matrix.

Proof. Let us first assume that y_t starts at y_0. From (2.1) we have that

$$\begin{aligned} y_1 &= \nu + A_1 y_0 + u_1, \\ y_2 &= \nu + A_1 y_1 + u_2, \\ &\vdots \\ y_t &= \nu + A_1 y_{t-1} + u_t. \end{aligned}$$

If we now substitute the generated y_1 in the equation of y_2, y_2 in the equation of y_3 and so on, we find

$$\begin{aligned} y_1 &= \nu + A_1 y_0 + u_1 \\ y_2 &= (I_K + A_1)\nu + A_1^2 y_0 + (A_1 u_1 + u_2) \\ &\vdots \\ y_t &= (I_K + A_1 + \cdots + A_1^{t-1})\nu + A_1^t y_0 + \sum_{i=0}^{t-1} A_1^i u_{t-i}. \end{aligned}$$

We see here that the equation of y_t can actually be rearranged into an equation with the variables $y_0, u_1, u_2, \ldots, u_t$. If we assume that our process did not specifically start at time 0, but at time $t-j+1$, it is easy to see that we find

$$y_t \quad = (I_K + A_1 + \cdots + A_1^j)\nu + A_1^{j+1} y_{t-j-1} + \textstyle\sum_{i=0}^{j} A_1^i u_{t-i}.$$

□

Now we can look what happens with y_t when we have the assumption that our information set contains an infinite amount of values of y, which is equivalent with having $j \to \infty$. We then find that y_t is stable when the following *stability condition* holds.

Theorem 2.1. *The VAR(1) process y_t is stable if the stability condition*

$$\det(I_K - A_1 z) \neq 0 \quad \textit{for } |z| \leqslant 1$$

holds.

Proof. In y_t, as in Lemma 2.1, we have three different terms that we sum up, which all contain j. It turns out that if all eigenvalues of A_1 have a modulus smaller than 1, then the sequence $(A_1^j)_{j \in \mathbb{N}_0}$ is absolutely summable and the sum of the sequence converges to $(I_K - A_1)^{-1}$ (Lütkepohl, 2005, p. 657). Using that all eigenvalues of A_1 have a modulus smaller than 1, it can be found that the term $\sum_{i=0}^{\infty} A_1^i u_{t-i}$ exists in mean square (Lütkepohl, 2005, p. 688) and obviously that $A_1^{j+1} y_{t-j-1}$ goes to zero. Now we find for $j \to \infty$ that

$$y_t = \mu + \sum_{i=0}^{\infty} A_1^i u_{t-i}, \tag{2.2}$$

where $\mu = (I_K - A_1)^{-1}\nu$. Hence, we find that y_t is stable when all eigenvalues of A_1 have a modulus smaller than 1. This condition is equivalent with

$$\det(I - A_1 z) \neq 0 \quad \text{for } |z| \leqslant 1,$$

which can simply be seen by using the fact that

$$\det(I_K - A_1 z) = z^K \det(\frac{I_K}{z} - A_1) \qquad \text{for } z \neq 0,$$

which is equal to 0 if z^{-1} is an eigenvalue of A_1. All eigenvalues of A_1 are smaller than 1, hence $\frac{1}{z} > 1$ results $z^K \det(\frac{I_K}{z} - A_1)$ to be 0. For the case $z = 0$ we have $\det(I_K - A_1 z) = 1$. □

We can also find the stability condition of a VAR(p) model. Let us first define the *companion form* of a VAR(p) process.

Definition 2.2. *The companion form of a VAR(p) process is*

$$Y_t \quad = \boldsymbol{\nu} + \boldsymbol{A}Y_{t-1} + U_t,$$

, where $Y_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix}$ *is a* $(Kp \times 1)$ *vector and* $U_t = \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ *is a* $(Kp \times 1)$ *vector. We define the* $(Kp \times 1)$ *vector* $\boldsymbol{\nu}$ *and the* $(Kp \times Kp)$ *matrix* $\boldsymbol{A}$ *as*

$$\boldsymbol{\nu} = \begin{bmatrix} \nu \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \boldsymbol{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I_K & 0 & \dots & 0 & 0 \\ 0 & I_K & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_K & 0 \end{bmatrix}.$$

We have defined Y_t *such that*

$$y_t = JY_t,$$

with $J := (I_K, 0, \dots, 0)$ *as* $(K \times Kp)$ *matrix.*

Now using the companion form a VAR(p) process, we can simply obtain the stability condition of y_t in a similar way we obtained Theorem 2.1.

Theorem 2.2. *The VAR(p) process* y_t *is stable if the stability condition*

$$\det(I_{Kp} - \boldsymbol{A}z) = \det(I_{Kp} - A_1 z - A_2 z^2 - \dots - A_p z^p) \neq 0 \quad \textit{for } |z| \leqslant 1$$

holds.

We call the stability condition in Theorem 2.2 the *reverse characteristic polynomial* and therefore we can say that the VAR(p) process is stable if the reverse characteristic polynomial has no roots in and on the complex unit circle. Again equivalent to the stability condition is only having eigenvalues of $\mathbf{A}$ with modulus smaller than 1.

Example 2.1. *We can look at an example of a bivariate VAR(2) process and check if it suffices the stability condition of a VAR(2) process. Suppose we have*

$$y_t = \nu + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} y_{t-2} + u_t.$$

The reverse characteristic polynomial of y_t *is*

$$\det\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} z - \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} z^2 \right) = 1 - z + 0.21z^2 - 0.025z^3.$$

The roots of the reverse characteristic polynomial are

$$z_1 = 1.3,$$
$$z_{2,3} \approx 3.55 \pm 4.26i,$$

where $|z_1| = 1.3$ *and* $|z_2| = |z_3| \approx 5.545$*. All roots of the reverse characteristic polynomial have modulus greater than 1, hence* y_t *is a stable process.*

2.1.2 Moving Average representation

The MA (moving average) representation of a VAR process is a very useful representation, which allows to rewrite the process into an infinte sum of some elements. The process Y_t as in Definition 2.2 is defined in such a way that it can be possible to rewrite Y_t into the MA representation as follows.

$$Y_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{A}^i U_{t-i}, \tag{2.3}$$

which we can obtain by a similar method we used to get (2.2). Here $\boldsymbol{\mu}$ represents the mean of Y_t. We assume here that the stability condition from Theorem 2.2 holds. Using (2.3), we can find the moving average representation of y_t as follows:

$$\begin{aligned} y_t &= JY_t \\ &= J\boldsymbol{\mu} + \sum_{i=0}^{\infty} J\mathbf{A}^i J^T J U_{t-i} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \end{aligned} \tag{2.4}$$

where we use $\Phi_i = J\mathbf{A}^i J^T$, the fact that $J^T J = I$ and $u_t = JU_t$. Since again $(\mathbf{A}^i)_{i\in\mathbb{N}_0}$ is absolutely summable by the stability condition, we have that $(\Phi_i)_{i\in\mathbb{N}_0}$ is absolutely summable as well. Now (2.4) is the moving average representation of y_t, which we will be frequently using later on.

It turns out a more direct way to determine the values of Φ_i as in (2.4) can be found by rewriting the VAR(p) model using the so-called *lag operator*, which is defined as follows.

Definition 2.3. *The lag operator L transforms an element of a time series to its previous time step, so*

$$Ly_t = y_{t-1}. \tag{2.5}$$

This operator is also sometimes referred to as the *backshift operator*. Using the lag operator, it will be possible to find the following values of Φ_i.

Theorem 2.3. *The values of Φ_i of a moving average representation can be represented as*

$$\begin{aligned} \Phi_0 &= I_K, \\ \Phi_i &= \sum_{j=1}^{i} \Phi_{i-j} A_j, \qquad i \in \mathbb{N}, \end{aligned}$$

where $A_j = 0$ for $j > p$.

Proof. With the lag operator we can rewrite the VAR(p) model (2.1) as

$$y_t = \nu + (A_1 L + A_2 L^2 + \cdots + A_p L^p) y_t + u_t. \tag{2.6}$$

If we now define $A(L)$ as

$$A(L) := (I - A_1 L - A_2 L^2 - \cdots - A_p L^p),$$

we can rewrite (2.6) to

$$A(L) y_t = \nu + u_t. \tag{2.7}$$

If we now define the infinite sum $\Phi(L)$ as

$$\Phi(L) := \sum_{i=0}^{\infty} \Phi_i L^i$$

and we define the following relationship between $A(L)$ and $\Phi(L)$ as

$$\Phi(L)A(L) = I_K, \tag{2.8}$$

then we can rewrite (2.7) as

$$\begin{aligned} y_t &= \Phi(L)\nu + \Phi(L)u_t \\ &= (\sum_{i=0}^{\infty} \Phi_i)\nu + \sum_{i=0}^{\infty} \Phi_i u_{t-i}. \end{aligned} \tag{2.9}$$

In the result above we again find a moving average representation of y_t as in (2.4) if we simply take $\mu = (\sum_{i=0}^{\infty} \Phi_i)\nu$, which is still a fixed term since ν is fixed. Now we can use the relationship (2.8) between $\Phi(L)$ and $A(L)$ to find the values of Φ_i. Writing out the relationship in terms of L results in

$$\begin{aligned} I_K &= (\Phi_0 + \Phi_1 L + \Phi_2 L^2 + \dots)(I_K - A_1 L - A_2 L^2 - \dots - A_p L^p) \\ &= \Phi_0 I_K + (\Phi_1 - \Phi_0 A_1)L + (\Phi_2 - \Phi_1 A_1 - \Phi_0 A_2)L^2 + \dots + \\ &\quad (\Phi_i - \sum_{j=1}^{i} \Phi_{i-j} A_j)L^i + \dots, \end{aligned} \tag{2.10}$$

which gives us now the following equality's

$$\begin{aligned} I_k &= \Phi_0 \\ 0 &= \Phi_1 - \Phi_0 A_1 \\ 0 &= \Phi_2 - \Phi_1 A_1 - \Phi_0 A_2 \\ &\vdots \\ 0 &= \Phi_i - \sum_{j=1}^{i} \Phi_{i-j} A_j \\ &\vdots \quad , \end{aligned}$$

with $A_j = 0$ if $j > p$. Now it is easy to see that the values for Φ_i can be written as

$$\begin{aligned} \Phi_0 &= I_K \\ \Phi_i &= \sum_{j=1}^{i} \Phi_{i-j} A_j, \qquad i \in \mathbb{N}. \end{aligned}$$

□

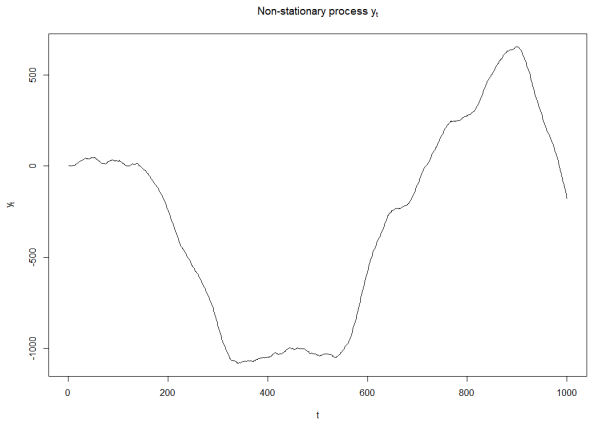
2.1.3 Stationarity

The stationarity (or non-stationarity) of a process is an important property that we will be using later on. Let us first look at the definition of stationarity of a process.

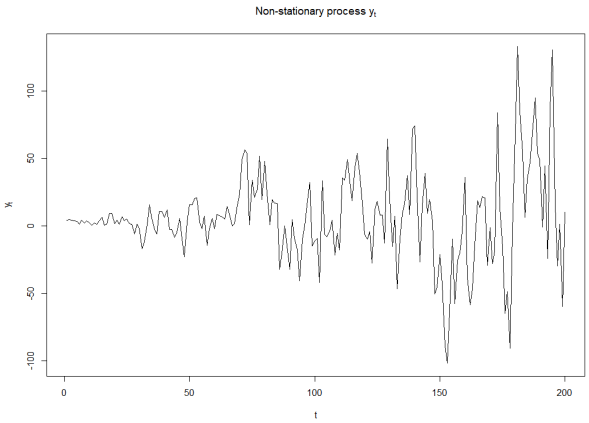
Definition 2.4. *We call a certain process stationary if the first and second moments are time invariant, which means that a process y_t is stationary when the following holds.*

1. $\mathbb{E}(y_t) = \mu \qquad \forall t,$
2. $\mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)^T] = \Gamma_y(h) \qquad \forall t, \forall h.$

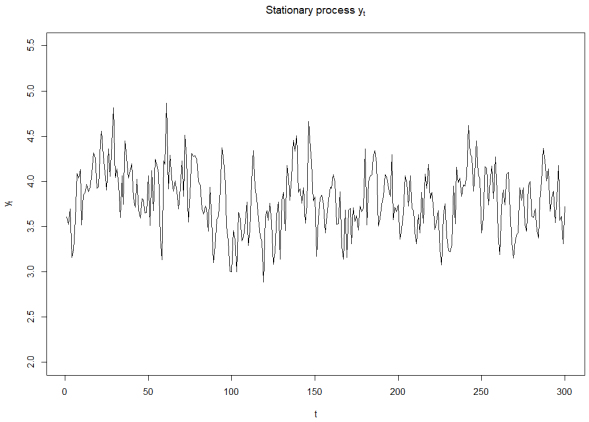
In the first condition of Definition 2.4 we have that the mean of y_t is the same vector μ for all possible time steps t. The second condition tells us that $\mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)^T]$ is a certain function Γ_y, that only depends on h and not on t. We call this function the *autocovariance* of y_t. In other words, the autocovariance of y_t does not depend on t, but only on the amount of time steps h between the two vectors y_t and y_{t-h}. Let us look at some examples of stationary and non-stationary univariate processes.



(a) Non-stationary process y_t.



(b) Non-stationary process y_t.



(c) Stationary process y_t.

Figure 2.1: Various stationary and non-stationary univariate processes.

In figure 2.1a we see a non-stationary process, since the first condition of Definition 2.4 does not hold. The mean of the process is obviously lower for time steps between 400 and 500 than for other time steps. In figure 2.1b we also see a non-stationary process, since the second condition of Definition 2.4 does not hold. The autocovariance of the process definitely seems to increase when t increases, which contradicts the condition. In figure 2.1c we clearly see a stationary process, since both conditions hold.

We would like to know when a VAR(p) proces y_t is actually stationary. If we look at a stable VAR(p) process y_t, we find that

$$\begin{aligned} \mathbb{E}[y_t] &= \mathbb{E}[JY_t] \\ &= J\mathbb{E}[Y_t] \\ &= J\boldsymbol{\mu} \\ &= (I - A_1 - A_2 - \cdots - A_p)^{-1}\nu, \end{aligned} \tag{2.11}$$

where $\boldsymbol{\mu} = (I_{Kp} - \mathbf{A})^{-1}\nu$, which can be found by using the same methodology we used to obtain μ in (2.2). It can also be found that

$$\begin{aligned} \Gamma_y(h) &= \mathbb{E}[(JY_t - \boldsymbol{\mu})(JY_{t-h} - \boldsymbol{\mu})^T] \\ &= J\sum_{i=0}^{\infty} \mathbf{A}^{h+i}\Sigma_U(\mathbf{A}^i)^T J^T, \end{aligned} \tag{2.12}$$

where $\Sigma_U = \mathbb{E}[U_t U_t^T]$ (Lütkepohl, 2005, pp. 688-689). We can see that the mean of a stable VAR(p)

process does not depend on t and that the autocovariance only depends on h, which means that a stable VAR(p) process is always stationary. This is why the stability condition of a VAR(p) process can also be referred to as the *stationarity condition*.

An important result of the stationarity of a process is Wold's theorem (Wold, 1938), which is as follows.

Theorem 2.4. *Every stationary process x_t can be written in the form:*

$$x_t = z_t + y_t,$$

where z_t is a deterministic process and y_t is a process uncorrelated with z_t that can be written in a moving average representation

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}.$$

With this theorem it is possible to show that we can rewrite y_t of the stationary process x_t as an infinite sum of coefficient matrices $A_1, A_2, \ldots$ as

$$y_t = \sum_{i=1}^{\infty} A_i y_{t-i} + u_t,$$

where we assumed Φ_i to be absolutely summable (Lütkepohl, 2005, p. 25). This means that every stationary process x_t has an infinite order VAR representation. Since the matrices Φ_i are absolutely summable, which means the matrices A_i are absolutely summable and will converge to 0, we can also see that a stationary process x_t can be approximated with a VAR representation of finite order. Hence, the stationarity of a process is a very strong property, since it implies that a finite order VAR process can be found.

2.1.4 Autocovariance

The autocovariance function is a function of a process that gives the covariance of the process between two different points in time, for example between y_t and y_{t-h}. We also refer to this as the covariance of the process at lag h. This function will depend on the value of the time t and the lag h. However, for stable VAR processes we found in (2.12) that the autocovariance only depends on h and not on t. In this section we will be looking at some properties of the autocovariance of a stable VAR process. We will use the notation $\Gamma_y(h)$ as the autocovariance function of a VAR process y_t, which is defined as follows.

Definition 2.5. *The autocovariance of a stationary process y_t at lag h is*

$$\Gamma_y(h) = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)^T],$$

with $\mu = \mathbb{E}[y_t]$. We can $\Gamma_y(h)$ the autocovariance function.

First, let us look at the autocovariance of a VAR(1) process y_t. We can obtain the following autocovariance function.

Lemma 2.2. *The autocovariance function of a VAR(1) process y_t is*

$$\Gamma_y(h) = \begin{cases} A_1 \Gamma_y(h-1) & h > 0 \\ A_1 \Gamma_y(1)^T + \Sigma_u & h = 0 \end{cases}$$

Proof. If we look at the process $y_t - \mu$, which is still a VAR(1) process, but with $\nu = 0$, we find

$$y_t - \mu = A_1(y_{t-1} - \mu) + u_t. \tag{2.13}$$

We can multiply both sides of (2.13) with $(y_{t-h} - \mu)^T$ and take the expectation, to obtain

$$\mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)^T] = A_1 \mathbb{E}[(y_{t-1} - \mu)(y_{t-h} - \mu)^T] + \mathbb{E}[u_t(y_{t-h} - \mu)^T],$$

or equivalently

$$\Gamma_y(h) = A_1\Gamma_y(h-1) + \mathbb{E}[u_t(y_{t-h} - \mu)^T].$$

When we take $h = 0$, we find that

$$\begin{aligned}\mathbb{E}[u_t(y_{t-h} - \mu)^T] &= \mathbb{E}[u_t(A_1(y_{t-1} - \mu) + u_t)^T] \\ &= \mathbb{E}[u_t u_t^T] \\ &= \Sigma_u\end{aligned}$$

and for $h > 0$

$$\begin{aligned}\mathbb{E}[u_t(y_{t-h} - \mu)^T] &= \mathbb{E}[u_t(A_1(y_{t-1} - \mu) + u_t)^T] \\ &= 0,\end{aligned}$$

since u_t is a white noise sequence and hence uncorrelated with $y_{t-1}, y_{t-2}, \ldots$ and $u_{t-1}, u_{t-2}, \ldots$ We now have the autocovariance function of y_t for different values of h as

$$\Gamma_y(h) = \begin{cases} A_1\Gamma_y(h-1) & h > 0 \\ A_1\Gamma_y(1)^T + \Sigma_u & h = 0 \end{cases}$$

We used the fact that $\Gamma_y(-1) = \Gamma_y(1)^T$, which follows straightforward from Definition (2.5). □

These equations are referred to as the *Yule-Walker equations* (Yule, 1927; Walker, 1931). If we assume that the matrices A_1 and Σ_u are known beforehand, then all we need to do to find the autocovariance of y_t for all values of h is to determine $\Gamma_y(0)$, because then we can use the Yule-Walker equations recursively to determine the autocovariances for all values of h.

We can determine $\Gamma_y(0)$ by combining the Yule-Walker equations of $\Gamma_y(0)$ and $\Gamma_y(1)$ as

$$\begin{aligned}\Gamma_y(0) &= A_1\Gamma_y(1)^T + \Sigma_u \\ &= A_1\Gamma_y(0)A_1^T + \Sigma_u. \end{aligned} \tag{2.14}$$

To obtain $\Gamma_y(0)$ from (2.14), we will first need to look at the definitions of the *vec operator* and the *Kronecker product*.

Definition 2.6. *If $A := (a_1, a_2, \ldots, a_n)$ is a $(m \times n)$ matrix with $a_1, a_2, \ldots, a_n$ as $(m \times 1)$ column vectors, then the vec operator returns the $(mn \times 1)$ vector*

$$\operatorname{vec}(A) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

Definition 2.7. *Suppose A is a $(m \times n)$ matrix, where a_{ij} is the i-th row j-th column element of A. If B is a $(p \times q)$ matrix, then the Kronecker product between these two matrices is defined as*

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \ldots & a_{1n}B \\ a_{21}B & a_{22}B & \ldots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \ldots & a_{mn}B \end{bmatrix}.$$

These two operators combined can give us the following useful lemma (Lütkepohl, 2005, pp. 661-662).

Lemma 2.3. *Let A, B and C be three matrices, where the product ABC is defined. Then*

$$\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec}(B).$$

We can now apply this lemma to find the following value for $\Gamma_y(0)$.

Corollary 2.1. *As a result of Lemma 2.2 we find that*

$$\text{vec}(\Gamma_y(0)) = (I_{K^2} - A_1 \otimes A_1)^{-1} \text{vec}(\Sigma_u),$$

where $\Gamma_y(0)$ can be found by reverting the vec operator back to a matrix.

Proof. When we apply Lemma 2.3 on (2.14), we find that

$$\begin{aligned} \text{vec}(\Gamma_y(0)) &= \text{vec}(A_1\Gamma_y(0)A_1^T + \Sigma_u) \\ &= \text{vec}(A_1\Gamma_y(0)A_1^T) + \text{vec}(\Sigma_u) \\ &= (A_1 \otimes A_1)\,\text{vec}(\Gamma_y(0)) + \text{vec}(\Sigma_u), \end{aligned}$$

where we used the fact that $\text{vec}(A + B) = \text{vec}(A) + \text{vec}(B)$. Solving the equation above allows us to find $\text{vec}(\Gamma_y(0))$ as

$$\text{vec}(\Gamma_y(0)) = (I_{K^2} - A_1 \otimes A_1)^{-1}\text{vec}(\Sigma_u),$$

where we now can find the original matrix $\Gamma_y(0)$ by simply reverting the vec operator back to the matrix itself. □

We can now look at the autocovariance of a VAR(p) process y_t. Using the same reasoning we used to obtain the Yule-Walker equations of a VAR(1) process in Lemma 2.2, we find that the VAR(p) process y_t has the following autocovariance function.

Theorem 2.5. *The autocovariance function of a VAR(p) process y_t is*

$$\Gamma_y(h) = \begin{cases} A_1\Gamma_y(h-1) + A_2\Gamma_y(h-2) + \cdots + A_p\Gamma_y(h-p) & h > 0 \\ A_1\Gamma_y(1)^T + +A_2\Gamma_y(2)^T + \cdots + A_p\Gamma_y(p)^T\Sigma_u & h = 0 \end{cases}$$

Again, we use the fact that $\Gamma_y(-h) = \Gamma_y(h)^T$ for all h, which follows straight from Definition (2.5). If we again assume that matrices $A_1, A_2, \ldots, A_p, \Sigma_u$ are known beforehand, we can see that $\Gamma_y(h)$ can be determined for $h \geqslant p$ if we know the values of $\Gamma_y(1), \Gamma_y(2), \ldots, \Gamma_y(p-1)$. We find that these matrices can be found using the following corollary.

Corollary 2.2. *As a result of Theorem 2.5 we find, using the companion form of a VAR(p) process, that*

$$vec(\Gamma_Y(0)) = (I_{(Kp)^2} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1} vec(\Sigma_U),$$

where $\Sigma_U = \mathbb{E}(U_t U_t^T)$ and

$$\Gamma_Y(0) = \begin{bmatrix} \Gamma_y(0) & \Gamma_y(1) & \ldots & \Gamma_y(p-1) \\ \Gamma_y(-1) & \Gamma_y(0) & \ldots & \Gamma_y(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_y(-p+1) & \Gamma_y(-p+2) & \ldots & \Gamma_y(0) \end{bmatrix}.$$

The matrices $\Gamma_y(1), \Gamma_y(2), \ldots, \Gamma_y(p-1)$ can be obtained by reverting the vec operator back to a matrix.

Proof. The VAR(p) process $y_t - \boldsymbol{\mu}$ can be written in companion form as in Definition 2.2, which results in

$$Y_t - \boldsymbol{\mu} = \mathbf{A}(Y_{t-1} - \boldsymbol{\mu}) + U_t,$$

where $\boldsymbol{\mu} = \mathbb{E}[Y_t]$. Using Definition 2.5 we find that the autocovariance of Y_t for $h = 0$ is

$$\begin{aligned}
\Gamma_Y(0) &:= \mathbb{E}[(Y_t - \boldsymbol{\mu})(Y_t - \boldsymbol{\mu})^T] \\
&= \mathbb{E}\left(\begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} \begin{bmatrix} y_t - \mu & y_{t-1} - \mu & \dots & y_{t-p+1} - \mu \end{bmatrix}^T \right) \\
&= \begin{bmatrix} \Gamma_y(0) & \Gamma_y(1) & \dots & \Gamma_y(p-1) \\ \Gamma_y(-1) & \Gamma_y(0) & \dots & \Gamma_y(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_y(-p+1) & \Gamma_y(-p+2) & \dots & \Gamma_y(0) \end{bmatrix}.
\end{aligned}$$

For similar reasons we used to obtain Corollary 2.1, we find that $\Gamma_Y(0)$ can be obtained by using

$$\text{vec}(\Gamma_Y(0)) = (I_{(Kp)^2} - \mathbf{A} \otimes \mathbf{A})^{-1} \text{vec}(\Sigma_U),$$

where $\Sigma_U = \mathbb{E}(U_t U_t^T)$. Hence, we can find the values for $\Gamma_y(0), \Gamma_y(1), \dots, \Gamma_y(p-1)$ by reverting the vec operator back to the matrix. □

Example 2.2. *Now we can try to find the autocovariances of the bivariate VAR(2) process y_t as in Example 2.1. We again have*

$$y_t = \nu + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} y_{t-2} + u_t$$

and we suppose the covariance matrix of u_t to be

$$\Sigma_u = \begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix}.$$

Then we have for the matrices $\boldsymbol{A}$ and Σ_U in the companion form of y_t as in Definition 2.2 that

$$\boldsymbol{A} = \begin{bmatrix} 0.5 & 0.1 & 0 & 0 \\ 0.4 & 0.5 & 0.25 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

and

$$\begin{aligned}
\Sigma_U &:= \mathbb{E}(U_t U_t^T) \\
&= \begin{bmatrix} \Sigma_u & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0.09 & 0 & 0 & 0 \\ 0 & 0.04 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}$$

Using Corollary 2.2, we find that

$$\begin{aligned}
\Gamma_Y(0) &= \begin{bmatrix} 0.131 & 0.066 & 0.072 & 0.051 \\ 0.066 & 0.181 & 0.104 & 0.143 \\ 0.072 & 0.104 & 0.131 & 0.066 \\ 0.051 & 0.143 & 0.066 & 0.181 \end{bmatrix} \\
&= \begin{bmatrix} \Gamma_y(0) & \Gamma_y(1) \\ \Gamma_y(1)^T & \Gamma_y(0) \end{bmatrix},
\end{aligned}$$

hence $\Gamma_y(0) = \begin{bmatrix} 0.131 & 0.066 \\ 0.066 & 0.181 \end{bmatrix}$ *and* $\Gamma_y(1) = \begin{bmatrix} 0.072 & 0.051 \\ 0.104 & 0.143 \end{bmatrix}$. *We now find the autocovariances for all* $h \geqslant 2$ *using the Yule-Walker equations of Theorem 2.5 as follows.*

$$\begin{aligned} \Gamma_y(2) &= A_1\Gamma_y(1) + A_2\Gamma_y(0) \\ &= \begin{bmatrix} 0.046 & 0.040 \\ 0.113 & 0.108 \end{bmatrix}, \\ \Gamma_y(3) &= A_1\Gamma_y(2) + A_2\Gamma_y(1) \\ &= (A_1^2 + A_2)\Gamma_y(1) + A_1A_2\Gamma_y(0) \\ &= \begin{bmatrix} 0.035 & 0.031 \\ 0.093 & 0.083 \end{bmatrix}, \\ &\vdots \end{aligned}$$

and so on.

2.1.5 Autocorrelation

Most of the time the autocorrelation function is being used instead of the autocovariance function, when analysing the linear correlation of the lags. The autocorrelation function of a process y_t at lag h returns the correlation between y_t and y_{t-h}. This function will again only depend on h and not on t when we work with stationary processes. For stationary processes, the autocorrelation is defined as follows:

Definition 2.8. *The autocorrelation of a stationary process* y_t *is*

$$R_y(h) = D^{-1}\Gamma_y(h)D^{-1}, \tag{2.15}$$

where $\Gamma_y(h)$ *is the autocorrelation for lag* h *and* D^{-1} *is defined as*

$$D^{-1} = \begin{bmatrix} \frac{1}{\sqrt{\gamma_{11}(0)}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\gamma_{22}(0)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\gamma_{KK}(0)}} \end{bmatrix}, \tag{2.16}$$

where $\gamma_{ij}(h)$ *is the* i*-th row and* j*-th column element of the autocovariance at lag* h.

With this definition, the i-th row and j-th column element of $R_y(h)$ is simply the correlation between $y_{i,t}$ and $y_{j,t-h}$. We can expand Example (2.2) to find the autocorrelation of y_t.

Example 2.3. *Continuing from Example (2.2), we can find the autocorrelation of* y_t *as follows. We have*

$$D^{-1} = \begin{bmatrix} \frac{1}{\sqrt{0.131}} & 0 \\ 0 & \frac{1}{\sqrt{0.181}} \end{bmatrix}, \tag{2.17}$$

hence

$$\begin{aligned} R_y(0) &= D^{-1}\Gamma_y(0)D^{-1} \\ &= \begin{bmatrix} 1 & 0.43 \\ 0.43 & 1 \end{bmatrix}, \\ R_y(1) &= D^{-1}\Gamma_y(1)D^{-1} \\ &= \begin{bmatrix} 0.55 & 0.33 \\ 0.68 & 0.79 \end{bmatrix}, \\ &\vdots \end{aligned} \tag{2.18}$$

and so on. We then can find the following autocorrelations between the variables.

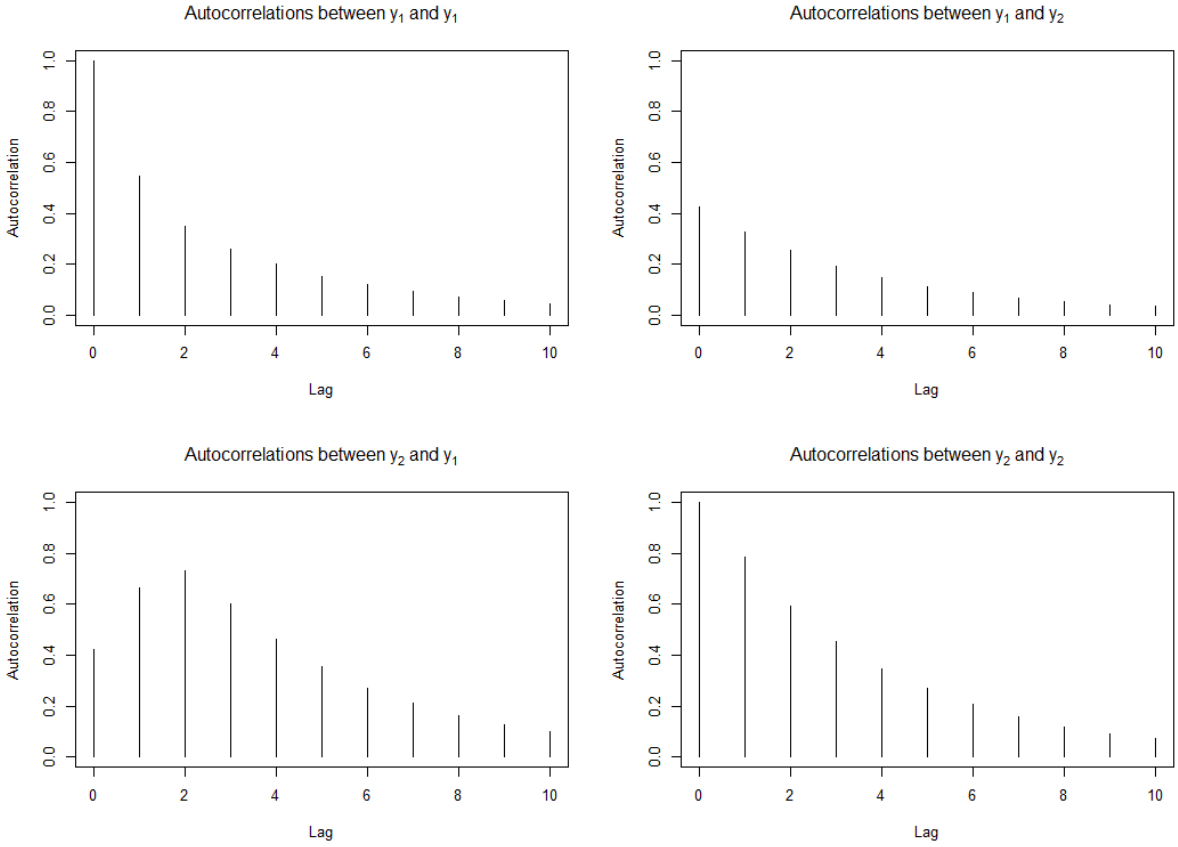


Figure 2.2: Autocorrelations between the variables for various lags.

2.2 Forecasting

When analysing financial time series, forecasting the process will be of great importance. It is a great way to predict the future behaviour of a process. When forecasting we can use all of the information that the time series has to predict future values of the process. Suppose we have a time series y_t with values up to time t, then we could try to forecast $t+1, t+2, \ldots$ up to maybe h steps ahead. We call period t the *forecast origin* and h the *forecast horizon*. If we have a function that gives us a forecast of a process h steps ahead of our forecast origin, then we call this function the *h-step predictor*. Our goal is to find a predictor of a stable VAR(p) process that provides us the best possible forecast. For VAR models it turns out that predictors that minimize the *mean squared error* are the most effective (Granger, 1969, pp. 199-207). The mean squared error of a predictor is defined as follows.

Definition 2.9. *If y_t is a stable VAR(p) process, with h-step predictor $\bar{y}_t(h)$, then the mean squared error (MSE) of this predictor is*

$$\mathrm{MSE}[\bar{y}_t(h)] := \mathbb{E}[(y_{t+h} - \bar{y}_t(h))(y_{t+h} - \bar{y}_t(h))^T]. \tag{2.19}$$

The value of $y_{t+h} - \bar{y}_t(h)$ is the error that is made by forecasting the process, hence it is referred to as the *forecast error*.

2.2.1 Zero mean VAR(1) models

To find the minimum MSE predictor for a VAR(p) process we will first analyse when we have a zero mean VAR(1) model y_t, which means

$$y_t = A_1 y_{t-1} + u_t.$$

As in Lemma 2.1 we can rewrite $y_t, y_{t+1}, \ldots, y_{t+h}$ in terms of A_1, y_t and the white noise term u_t. We find that

$$y_{t+h} = A_1^h y_t + \sum_{i=0}^{h-1} A_1^i u_{t+h-i}.$$

If we define a predictor of y_t with forecast origin t and forecast horizon h as $y_t(h)$, then this predictor should depend on $y_t, y_{t-1}, \ldots$ and so on. We can write $y_t(h)$ as

$$y_t(h) := B_0 y_t + B_1 y_{t-1} + \ldots,$$

where $B_0, B_1, \ldots$ are $(K \times K)$ matrices. Our goal is to find these matrices so that mean squared error of this predictor is minimized. Let us first look at the forecast error of this predictor, which is

$$\begin{aligned} y_{t+h} - y_t(h) &= A_1^h y_t + \sum_{i=0}^{h-1} A_1^i u_{t+h-i} - \sum_{i=0}^{\infty} B_i y_{t-i} \\ &= \sum_{i=0}^{h-1} A_1^i u_{t+h-i} + (A_1^h - B_0) y_t - \sum_{i=1}^{\infty} B_i y_{t-i}. \end{aligned}$$

Using the fact that $u_{t+1}, u_{t+1}, \ldots$ is uncorrelated with $y_t, y_{t-1}, \ldots$, we find the mean squared error as follows.

$$\begin{aligned} \text{MSE}[y_t(h)] &= \mathbb{E}[(y_{t+h} - y_t(h))(y_{t+h} - y_t(h))^T] \\ &= \mathbb{E}\left[\left(\sum_{i=0}^{h-1} A_1^i u_{t+h-i} \right) \left(\sum_{i=0}^{h-1} A_1^i u_{t+h-i} \right)^T \right] \qquad (2.20) \\ &+ \mathbb{E}\left[\left((A_1^h - B_0) y_t - \sum_{i=1}^{\infty} B_i y_{t-i} \right) \left((A_1^h - B_0) y_t - \sum_{i=1}^{\infty} B_i y_{t-i} \right)^T \right]. \qquad (2.21) \end{aligned}$$

From (2.20)-(2.21) we can see that the mean squared error of $y_t(h)$ is minimal when the second term is 0, which means when $B_0 = A_1^h$ and $B_i = 0$ for $i \in \mathbb{N}$. Now the minimum MSE predictor of a VAR(1) process with zero mean is

$$\begin{aligned} y_t(h) &= A_1^h y_t \\ &= A_1 y_t(h-1), \end{aligned}$$

with a forecast error of

$$\sum_{i=0}^{h-1} A_1^i u_{t+h-i}. \qquad (2.22)$$

2.2.2 Zero mean VAR(p) models

Now for a VAR(p) process y_t with zero mean we can write this process in the companion form as in Definition 2.2 to obtain

$$Y_t = \mathbf{A} Y_{t-1} + U_t. \qquad (2.23)$$

Using the same methodology to obtain the minimum MSE predictor as in (2.22), we find

$$\begin{aligned} Y_t(h) &= \mathbf{A}^h Y_t \\ &= \mathbf{A} Y_t(h-1), \end{aligned}$$

where $Y_t(h)$ is the minimum MSE predictor of Y_{t+h}. Using the definition of the vector Y_t, we can rewrite this predictor into the form

$$Y_t(h) = \begin{bmatrix} y_t(h) \\ y_t(h-1) \\ \vdots \\ y_t(h-p+1) \end{bmatrix},$$

where $y_t(h)$ is the h-step predictor of the VAR(p) process y_t. We use here that $y_t(j) := y_{t+j}$, where $j \leqslant 0$. In other words, the predictions of values of the time series at time steps before the forecast origin are simply the same values we already know. We can now find the minimum MSE predictor $y_t(h)$ as follows.

$$\begin{aligned} y_t(h) &= J Y_t(h) \\ &= J\mathbf{A} Y_t(h-1) \\ &= A_1 y_t(h-1) + A_2 y_t(h-2) + \cdots + A_p y_t(h-p). \end{aligned} \tag{2.24}$$

The forecast error of this predictor can also be determined. Using the companion form as in Definition 2.2, we find for a VAR(p) process with zero mean that

$$\begin{aligned} Y_{t+h} &= \mathbf{A} Y_{t+h-1} + U_{t+h} \\ &= \mathbf{A}^2 Y_{t+h-2} + \mathbf{A} U_{t+h-1} + U_{t+h} \\ &\vdots \\ &= \mathbf{A}^h Y_t + \sum_{i=0}^{h-1} \mathbf{A}^i U_{t+h-i}, \end{aligned}$$

which means our forecast error will look like

$$\begin{aligned} y_{t+h} - y_t(h) &= J Y_{t+h} - J Y_t(h) \\ &= J\left(\mathbf{A}^h Y_t + \sum_{i=0}^{h-1} \mathbf{A}^i U_{t+h-i} \right) - J\mathbf{A}^h Y_t \\ &= \sum_{i=0}^{h-1} J\mathbf{A}^i U_{t+h-i} \\ &= \sum_{i=0}^{h-1} J\mathbf{A}^i J^T J U_{t+h-i} \\ &= \sum_{i=0}^{h-1} \Phi_i u_{t+h-i}, \end{aligned} \tag{2.25}$$

where $\Phi_i = J\mathbf{A}^i J^T$ as in the moving average representation (2.4).

2.2.3 Non-zero mean VAR(p) models

Using the minimum MSE predictor of a VAR(p) process with zero mean (2.24) we can find the minimum MSE predictor of a VAR(p) process with a non-zero mean. If we have VAR(p) process with a non-zero mean y_t, then we define x_t to be

$$x_t := y_t - \mu, \tag{2.26}$$

which is a VAR(p) process with zero mean if we take $\mu := \mathbb{E}[y_t]$. Remember from (2.11) that we have

$$\mathbb{E}[y_t] = (I - A_1 - A_2 - \cdots - A_p)^{-1}\nu. \tag{2.27}$$

We found that the minimum MSE predictor of x_t is equal to

$$x_t(h) = A_1 x_t(h-1) + A_2 x_t(h-2) + \cdots + A_p x_t(h-p), \tag{2.28}$$

hence μ to both sides of the equation above and using (2.27) results in the minimum MSE predictor of the non-zero mean VAR(p) process

$$\begin{aligned} x_t(h) + \mu &= \mu + A_1 x_t(h-1) + \cdots + A_p x_t(h-p) \\ &= \mu + A_1(y_t(h-1) - \mu) + \cdots + A_p(y_t(h-p) - \mu) \\ &= \mu(I - A_1 - \cdots - A_p) + A_1 y_t(h-1) + \cdots + A_p y_t(h-p) \\ &= \nu + A_1 y_t(h-1) + \cdots + A_p y_t(h-p) \\ &= y_t(h). \end{aligned} \tag{2.29}$$

The forecast error of the minimum MSE predictor of a non-zero mean VAR(p) process is obviously the same as the forecast error of the non-zero mean process, since

$$\begin{aligned} y_{t+h} - y_t(h) &= x_{t+h} - \mu - x_t(h) + \mu \\ &= x_{t+h} - x_t(h) \\ &= \sum_{i=0}^{h-1} \Phi_i u_{t+h-i}. \end{aligned}$$

2.2.4 Forecast intervals

The forecast error we found for the minimum MSE predictor of a VAR(p) model shows us that this predictor does not perfectly predict future values. An assumption which is often made is that the white noise terms u_t are considered i.i.d. multivariate normally distributed ($u_t \sim \mathcal{N}(0, \Sigma_u)$). If we now look at the forecast error we found in (2.25), then we find, under the assumption of multivariate normal white noise terms, that the forecast error is a linear combination of the error terms. Hence, the forecast error is multivariate normally distributed as well, so

$$y_{t+h} - y_t(h) \sim \mathcal{N}(0, \Sigma_y(h)), \tag{2.30}$$

where $\Sigma_y(h)$ is the covariance matrix of the forecast error for h steps ahead. This covariance matrix can be written as follows using Definition 2.9 of the mean squared error and the fact that the error terms are uncorrelated.

$$\begin{aligned} \Sigma_y(h) := \mathrm{MSE}[y_t(h)] &= \mathbb{E}\left[\left(\sum_{i=0}^{h-1} \Phi_i u_{t+h-i}\right)\left(\sum_{i=0}^{h-1} \Phi_i u_{t+h-i}\right)^T\right] \qquad (2.31) \\ &= \mathbb{E}\left[\sum_{i=0}^{h-1} \Phi_i u_{t+h-i} u_{t+h-i}^T \Phi_i^T\right] \\ &= \sum_{i=0}^{h-1} \Phi_i \Sigma_u \Phi_i^T \\ &= \Sigma_y(h-1) + \Phi_{h-1}\Sigma_u\Phi_{h-1}^T. \qquad (2.32) \end{aligned}$$

Since $y_t(j) := y_{t+j}$ where $j \leqslant 0$, we see that the mean squared error of the predictor is 0 for $h \leqslant 0$, hence $\Sigma_y(h) = 0$ for $h \leqslant 0$. Now (2.30) tells us that the forecast errors of the variables of interest of y are normally distributed. We find for the k-th variable of interest that

$$\frac{y_{k,t+h} - y_{k,t}(h)}{\sigma_k(h)} \sim \mathcal{N}(0,1), \tag{2.33}$$

where $y_{k,t}(h)$ is the k-th component of $y_t(h)$ and $\sigma_k(h)$ is the square root of the k-th row k-th column element of $\Sigma_y(h)$. This means we can now set up confidence intervals of our prediction for each variable of interest. If we define z_α to be the value such that

$$\mathbb{P}(Z \leqslant z_\alpha) = 1 - \alpha, \tag{2.34}$$

where $Z \sim \mathcal{N}(0,1)$, then

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(-z_{\alpha/2} \leqslant \frac{y_{k,t+h} - y_{k,t}(h)}{\sigma_k(h)} \leqslant z_{\alpha/2}) \\ &= \mathbb{P}(y_{k,t}(h) - \sigma_k(h) z_{\alpha/2} \leqslant y_{k,t+h} \leqslant y_{k,t}(h) + \sigma_k(h) z_{\alpha/2}), \end{aligned} \tag{2.35}$$

which means our $(1-\alpha)100\%$ forecast interval of the k-th variable of interest for h steps ahead will be

$$\left[y_{k,t}(h) - \sigma_k(h) z_{\alpha/2}, y_{k,t}(h) + \sigma_k(h) z_{\alpha/2}\right]. \tag{2.36}$$

2.2.5 Forecasting example

Let us look again at the bivariate VAR(2) process

$$y_t = \nu + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} y_{t-2} + u_t, \tag{2.37}$$

with the covariance matrix of u_t

$$\Sigma_u = \begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix},$$

but now we choose the process to have a non-zero mean by assuming

$$\nu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad y_{-1} = \begin{bmatrix} 3.8 \\ 8.8 \end{bmatrix} \text{ and } y_0 \begin{bmatrix} 3.5 \\ 8.5 \end{bmatrix}.$$

We assume that our process y_t starts at y_{-1}. Using the values above we can now generate the bivariate VAR(p) process (2.37) from y_1 to y_N for a certain integer N, by taking random samples of the $\mathcal{N}(0, \Sigma_u)$ distribution for $u_1, u_2, \ldots, u_N$. For this example we take $N = 100$, which results in the following generated process.

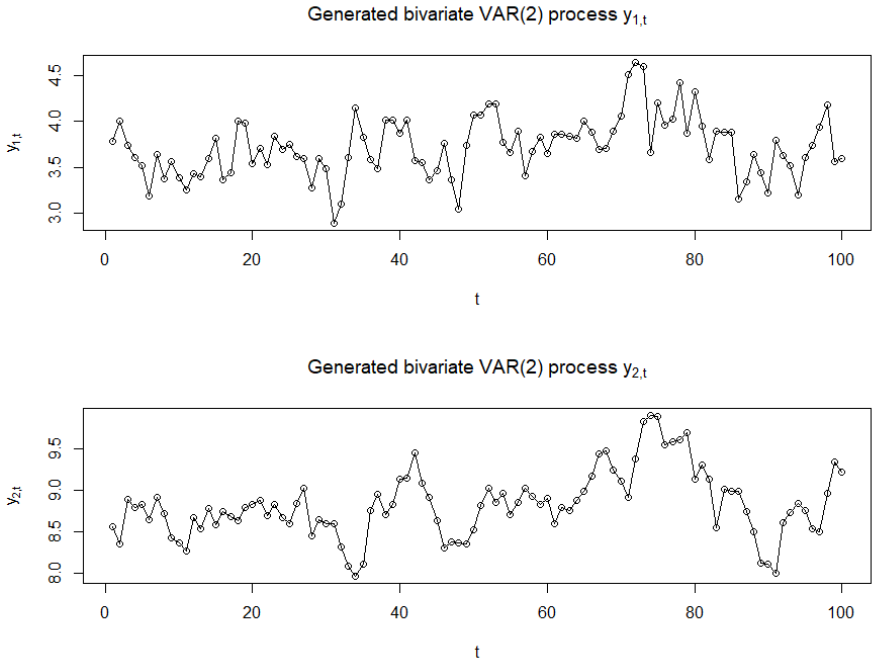


Figure 2.3: Generated process of (2.37) for $N = 100$.

We find that $y_{99} \approx \begin{bmatrix} 3.556 \\ 9.347 \end{bmatrix}$ and $y_{100} \approx \begin{bmatrix} 3.589 \\ 9.218 \end{bmatrix}$. In this example we would like to forecast this process up to 3 steps ahead of N. By using the minimum MSE predictor we found in (2.29), we find the following predictions.

$$\begin{aligned} y_N(1) &= \nu + A_1 y_N(0) + A_2 y_N(-1) \\ &\approx \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} \begin{bmatrix} 3.589 \\ 9.218 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} \begin{bmatrix} 3.556 \\ 9.347 \end{bmatrix} \\ &\approx \begin{bmatrix} 3.716 \\ 8.934 \end{bmatrix}, \end{aligned}$$

$$\begin{aligned} y_N(2) &= \nu + A_1 y_N(1) + A_2 y_N(0) \\ &\approx \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} \begin{bmatrix} 3.716 \\ 8.934 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} \begin{bmatrix} 3.589 \\ 9.218 \end{bmatrix} \\ &\approx \begin{bmatrix} 3.752 \\ 8.851 \end{bmatrix}, \end{aligned}$$

$$\begin{aligned} y_N(3) &= \nu + A_1 y_N(2) + A_2 y_N(1) \\ &\approx \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} \begin{bmatrix} 3.752 \\ 8.851 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} \begin{bmatrix} 3.716 \\ 8.934 \end{bmatrix} \\ &\approx \begin{bmatrix} 3.761 \\ 8.855 \end{bmatrix}. \end{aligned}$$

We can now look for a 95% forecast interval of these predictions for both variables of interest. First we will need to find the covariance matrices of the forecast errors for $h = 1, 2, 3$. From (2.32) we found that

$$\Sigma_y(h) = \Sigma_y(h-1) + \Phi_{h-1} \Sigma_u \Phi_{h-1}^T. \tag{2.38}$$

We now need to find the coefficient matrices Φ_0, Φ_1 and Φ_2 of the moving average representation. Using Theorem 2.3, we find

$$\Phi_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{2.39}$$

$$\begin{aligned} \Phi_1 &= \Phi_0 A_1 \\ &= \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix}, \end{aligned} \tag{2.40}$$

$$\begin{aligned} \Phi_2 &= \Phi_1 A_1 + \Phi_0 A_2 \\ &= \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix}^2 + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0.29 & 0.10 \\ 0.65 & 0.29 \end{bmatrix}. \end{aligned} \tag{2.41}$$

We now find

$$\begin{aligned}
\Sigma_y(1) &= \Phi_0\Sigma_u\Phi_0^T \\
&= \begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix}, \\
\Sigma_y(2) &= \Sigma_y(1) + \Phi_1\Sigma_u\Phi_1^T \\
&= \begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix}\begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix}\begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.5 \end{bmatrix} \\
&\approx \begin{bmatrix} 0.113 & 0.020 \\ 0.020 & 0.064 \end{bmatrix}, \\
\Sigma_y(3) &= \Sigma_y(2) + \Phi_2\Sigma_u\Phi_2^T \\
&\approx \begin{bmatrix} 0.113 & 0.020 \\ 0.020 & 0.064 \end{bmatrix} + \begin{bmatrix} 0.29 & 0.10 \\ 0.65 & 0.29 \end{bmatrix}\begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix}\begin{bmatrix} 0.29 & 0.65 \\ 0.10 & 0.29 \end{bmatrix} \\
&\approx \begin{bmatrix} 0.121 & 0.038 \\ 0.038 & 0.106 \end{bmatrix}.
\end{aligned}$$

Now using (2.36), we find for the first variable of interest

steps ahead	**forecast**	**lower bound**	**upper bound**	**interval length**
1	3.716	3.128	4.304	1.176
2	3.752	3.093	4.410	1.317
3	3.761	3.079	4.442	1.363

Table 1: The minimum MSE predictions for 1,2 and 3 steps of the first variable of interest and their 95% forecast intervals.

And for the second variable of interest we find

steps ahead	**forecast**	**lower bound**	**upper bound**	**interval length**
1	8.934	8.542	9.326	0.784
2	8.851	8.353	9.348	0.995
3	8.855	8.218	9.493	1.275

Table 2: The minimum MSE predictions for 1,2 and 3 steps of the second variable of interest and their 95% forecast intervals.

This whole process of forecasting 1,2 and 3 steps ahead can of course be expanded to forecast 4 or more steps ahead. This way we can find the following 10 step prediction.

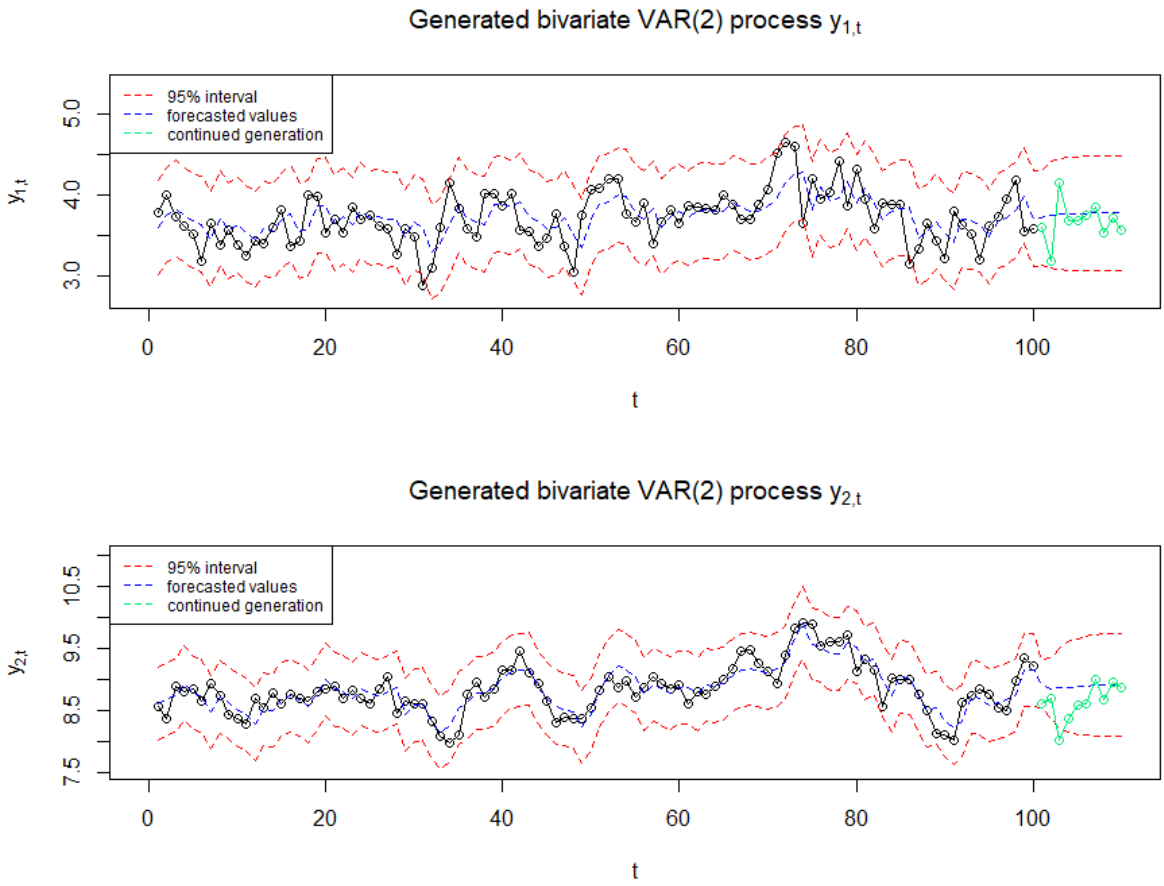


Figure 2.4: Prediction of the process 10 steps ahead for both variables of interest.

In the figure above we also see forecasted values and intervals for $t \leqslant 100$, even though those values of y_t are already known. What happens here is that for $t \leqslant 100$ we predict 1 step ahead, which creates a forecasted value with confidence intervals. Since the process is known for $t \leqslant 100$, we can check how well our predictor is performing. Looking at $t \leqslant 100$, we see that our process follows the forecasted values very well and it is nicely between the 95% interval most of the time.

Another way to check how well the predictor is performing is by simply continuing the generation and see whether our prediction for 10 steps ahead is accurate. We see that the continued generation pretty much follows the forecasted values and lies between the 95% interval, except for $y_{2,103}$. Since it is just a single value that lies outside the interval and since the process instantly corrects itself, we can conclude that this is not a big deal and that our predictor is still pretty accurate.

2.3 Analysis methods

In this section we will discuss various methods we can use to analyse VAR processes. These analysis methods all have their own strengths when it comes to time series analysis. The idea is that one should apply all these methods when performing time series analysis in order to gain as much information as possible.

2.3.1 Forecast error variance decomposition

The forecast error variance decomposition allows us to give a better interpretation of our forecast results. This method shows us how big the influence is of a certain variable of interest k on the error made by forecasting variable of interest j.

Let us first look at the definition of a *positive definite* matrix.

Definition 2.10. *Let A be a symmetric $(K \times K)$ matrix. Then A is positive definite when for all non-zero real $(K \times 1)$ vectors x the following holds:*

$$x^T A x > 0. \tag{2.42}$$

If we now look at the covariance matrix u_t of a VAR(p) process y_t, we find the following theorem.

Theorem 2.6. *The covariance matrix of u_t of a VAR(p) process y_t with K variables of interest is positive definite.*

Proof. Let x be a non-zero real $(K \times 1)$ vector, then

$$\begin{aligned} x^T \Sigma_u x &= \mathbb{E}[x^T u_t u_t^T x] \\ &= \mathbb{E}[(x^T u_t)(x^T u_t)^T] \\ &= \mathbb{E}[(x^T u_t)(x^T u_t)^T] \\ &> 0, \end{aligned}$$

since $x^T u_t$ is a constant and $u_t \neq 0$, since Σ_u is assumed not to be 0 in the VAR model. □

Now let us state the *Cholesky decomposition* (Brezinski, 1924) of the positive definite matrix.

Theorem 2.7. *Let A be a positive definite $(K \times K)$ matrix, then there exist a lower triangular matrix P with real and positive values on the diagonal such that*

$$A = PP^T. \tag{2.43}$$

Together with Theorem 2.6 we see that a matrix such as P also exists for the covariance matrix Σ_u. We can now rewrite the moving average representation of a VAR(p) process y_t as in (2.4) as the so-called *orthogonal representation* as follows.

$$\begin{aligned} y_t &= \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_i P P^{-1} u_{t-i} \\ &= \mu + \sum_{i=0}^{\infty} \Theta_i w_{t-i}, \end{aligned} \tag{2.44}$$

where $\Theta_i = \Phi_i P$ and $w_t = P^{-1} u_t$. Now using Theorem 2.7 we find that the covariance matrix of w_t is

$$\begin{aligned} \Sigma_w &= \mathbb{E}[w_t w_t^T] \\ &= P^{-1} \mathbb{E}[u_t u_t^T] (P^{-1})^T \\ &= P^{-1} \Sigma_u (P^{-1})^T \\ &= P^{-1} P P^T (P^{-1})^T \\ &= I_K. \end{aligned} \tag{2.45}$$

The w_t in the representation of (2.44) is often referred to as the *orthogonal residuals*, since (2.45) shows us that the residuals are uncorrelated between each variable of interest. We will be using this representation, since now we can determine how much one variable is influencing the covariance matrix of the forecast error. Previously this was not possible, since the variables of interest were correlated with each other. We now want to use the representation in (2.44) to rewrite the forecast error made by our minimum MSE predictor as in (2.25). We find the forecast error of

$$\begin{aligned} y_{t+h} - y_t(h) &= \sum_{i=0}^{h-1} \Phi_i u_{t+h-i} \\ &= \sum_{i=0}^{h-1} \Theta_i w_{t+h-i}. \end{aligned} \tag{2.46}$$

We can now look at the forecast error made by the predictor of just a single variable of interest. For notation we introduce $\theta_{mn,i}$ as m-th row and n-th column element of the matrix Θ_i in (2.46). To find the forecast error of the j-th variable of interest, we only use the j-th row of Θ_i in (2.46), which results in

$$\begin{aligned} y_{j,t+h} - y_{j,t}(h) &= \sum_{i=0}^{h-1} \theta_{j1,i} w_{1,t+h-i} + \cdots + \theta_{jK,i} w_{K,t+h-i} \\ &= \sum_{k=1}^{K} \left(\sum_{i=0}^{h-1} \theta_{jk,i} w_{k,t+h-i} \right). \end{aligned} \tag{2.47}$$

We see that the forecast error of the j-th variable of interest is formed with the orthogonal residuals of all K variables of interest. We now would like to look at the variance of the forecast error of the prediction h steps ahead for just a certain variable of interest j. This is the same as the j-th row j-th column element of the covariance matrix of the forecast error $\Sigma_y(h)$, which we can use to obtain the following theorem.

Theorem 2.8. *The variance of the forecast error of the j-th variable of interest equals*

$$\sum_{k=1}^{K} \sum_{i=0}^{h-1} \theta_{jk,i}^2. \tag{2.48}$$

Proof. The variance of the forecast error of the j-th variable of interest is the same as the j-th row j-th column element of $\Sigma_y(h)$, which is $\text{MSE}[y_{j,t}(h)]$. We now find

$$\begin{aligned} \text{MSE}[y_{j,t}(h)] &= \mathbb{E}[(y_{j,t+h} - y_{j,t}(h))^2] \\ &= \mathbb{E}\left[\left(\sum_{k=1}^{K} \left(\sum_{i=0}^{h-1} \theta_{jk,i} w_{k,t+h-i} \right) \right)^2 \right] && (2.49) \\ &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{i=0}^{h-1} \theta_{jk,i}^2 w_{k,t+h-i}^2 \right] + && (2.50) \\ &\quad \mathbb{E}\left[\sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{i=0}^{h-1} \theta_{jk,i} \theta_{jl,i} w_{k,t+h-i} w_{l,t+h-i} \mathbb{1}_{\{k \neq l\}} \right]. && (2.51) \end{aligned}$$

In the last step we split the square of (2.49) into 2 parts. In (2.50) we have the part where the orthogonal residuals w_t are multiplied with the same variable of interest and in (2.51) the orthogonal residuals are multiplied with a different variable of interest. We follow these steps because the equation in (2.51) equals 0. This is simply because the orthogonal residuals are uncorrelated between the variables of interest, which we found in (2.45). We now have

$$\begin{aligned} \text{MSE}[y_{j,t}(h)] &= \mathbb{E}\left[\sum_{k=1}^{K} \sum_{i=0}^{h-1} \theta_{jk,i}^2 w_{k,t+h-i}^2 \right] \\ &= \sum_{k=1}^{K} \sum_{i=0}^{h-1} \theta_{jk,i}^2 \mathbb{E}\left[w_{k,t+h-i}^2 \right]. \end{aligned} \tag{2.52}$$

In (2.45) we also found that $\mathbb{E}[w_{k,t}^2] = 1$ for all k, hence

$$\text{MSE}[y_{j,t}(h)] = \sum_{k=1}^{K} \sum_{i=0}^{h-1} \theta_{jk,i}^2 \tag{2.53}$$

□

We now have found the variance of the forecast error of the j-th variable of interest. Theorem 2.8 shows that the variance is formed with values of θ using all variables of interest. All we need to do to now is calculate how big the proportion of the forecast error variance is of a certain variable of interest k of the total forecast error variance or a variable of interest j. We call this proportion $\omega_{jk,h}$, which is defined in the following definition.

Definition 2.11. *The proportion of the forecast error variance of the k-th variable of interest by forecasting the j-th variable of interest h steps ahead is*

$$\omega_{jk,h} = \frac{\sum_{i=0}^{h-1} \theta_{jk,i}^2}{\text{MSE}[y_{j,t}(h)]}. \tag{2.54}$$

Example 2.4. *Continuing from the forecasting example in section 2.2.5 we can apply the forecast error variance decomposition for a better interpretation of the forecast results. In this example we will calculate the proportions of the forecast error variance of both variables of interest, based on the forecasts for 1, 2 and 3 steps ahead for both variables of interest. From Definition 2.11 we see that we need to determine matrices Θ_0, Θ_1 and Θ_2. Since $\Theta_i = \Phi_i P$, we will first need to determine the lower triangular matrix P from Theorem 2.7. In general we can apply the following method to find the matrix P.*

$$\begin{aligned} \Sigma_u = \begin{bmatrix} 0.09 & 0 \\ 0 & 0.04 \end{bmatrix} &= PP^T \\ &= \begin{bmatrix} p_{1,1} & 0 \\ p_{2,1} & p_{2,2} \end{bmatrix} \begin{bmatrix} p_{1,1} & p_{2,1} \\ 0 & p_{2,2} \end{bmatrix}, \end{aligned}$$

where $p_{1,1}$ and $p_{2,2}$ have to be positive and real valued. We now find

$$\begin{aligned} p_{1,1}^2 &= 0.09 \implies p_{1,1} = 0.03 \\ p_{1,1}p_{2,1} &= 0 \implies p_{2,1} = 0 \\ p_{2,1}^2 + p_{2,2}^2 &= 0.04 \implies p_{2,2} = 0.02, \end{aligned}$$

hence

$$P = \begin{bmatrix} 0.03 & 0 \\ 0 & 0.02 \end{bmatrix}.$$

Obviously the matrix P is straightforward when Σ_u is a diagonal matrix, but this does not always have to be the case. Now using matrices Φ_0, Φ_1 and Φ_2 we found earlier in (2.39 - 2.41), we find that

$$\begin{aligned} \Theta_0 &= \Phi_0 P \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.03 & 0 \\ 0 & 0.02 \end{bmatrix} \\ &= \begin{bmatrix} 0.03 & 0 \\ 0 & 0.02 \end{bmatrix}, \\ \Theta_1 &= \Phi_1 P \\ &= \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} \begin{bmatrix} 0.03 & 0 \\ 0 & 0.02 \end{bmatrix} \\ &= \begin{bmatrix} 0.15 & 0.02 \\ 0.12 & 0.10 \end{bmatrix}, \end{aligned}$$

$$\begin{aligned}\Theta_2 &= \Phi_2 P \\ &= \begin{bmatrix}0.29 & 0.10\\ 0.65 & 0.29\end{bmatrix}\begin{bmatrix}0.03 & 0\\ 0 & 0.02\end{bmatrix} \\ &= \begin{bmatrix}0.087 & 0.020\\ 0.195 & 0.058\end{bmatrix}.\end{aligned}$$

Now using Definition 2.11 and the covariance matrices of the forecast errors for 1,2 and 3 steps ahead (2.42 - 2.42), the proportions of the forecast error variance for the first variable of interest on itself will be

$$\begin{aligned}\omega_{11,1} &= \frac{\sum_{i=0}^{0}\theta_{11,i}^2}{\mathrm{MSE}[y_{1,t}(1)]} \\ &= \frac{0.3^2}{0.09} \\ &= 1 \\ \omega_{11,2} &= \frac{\sum_{i=0}^{1}\theta_{11,i}^2}{\mathrm{MSE}[y_{1,t}(2)]} \\ &= \frac{0.3^2+0.15^2}{0.1129} \\ &\approx 0.996 \\ \omega_{11,3} &= \frac{\sum_{i=0}^{2}\theta_{11,i}^2}{\mathrm{MSE}[y_{1,t}(3)]} \\ &\approx \frac{0.3^2+0.15^2+0.087^2}{0.121} \\ &\approx 0.993\end{aligned}$$

and the proportions of the forecast error variance for the first variable of interest on the second variable of interest will be

$$\begin{aligned}\omega_{21,1} &= \frac{\sum_{i=0}^{0}\theta_{21,i}^2}{\mathrm{MSE}[y_{2,t}(1)]} \\ &= \frac{0^2}{0.04} \\ &= 0 \\ \omega_{21,2} &= \frac{\sum_{i=0}^{1}\theta_{21,i}^2}{\mathrm{MSE}[y_{2,t}(2)]} \\ &= \frac{0^2+0.12^2}{0.0644} \\ &\approx 0.224 \\ \omega_{21,3} &= \frac{\sum_{i=0}^{2}\theta_{21,i}^2}{\mathrm{MSE}[y_{2,t}(3)]} \\ &\approx \frac{0^2+0.12^2+0.195^2}{0.106} \\ &\approx 0.496.\end{aligned}$$

Now for the proportions of the forecast error variance of the second variable of interest we can calculate them the same way as above, or simply take

$$\omega_{j2,i} = 1 - \omega_{j1,i} \quad \forall j, i, \tag{2.55}$$

since we are working with just 2 variables of interest. Of course using the same methodology, we can find the proportions of the forecast error variance for forecast horizons higher than 3. This will result into the following table.

Forecasted variable of interest	**Forecast horizon**	**Proportions of variable 1 on the forecast error variance**	**Proportions of variable 2 on the forecast error variance**
1	1	1	0
	2	0.996	0.004
	3	0.993	0.007
	4	0.992	0.008
	5	0.991	0.009
	10	0.989	0.011
2	1	0	1
	2	0.224	0.776
	3	0.496	0.504
	4	0.596	0.404
	5	0.637	0.363
	10	0.679	0.321

Table 3: Proportions of the forecast error variance of both variables of interest for forecasting both variables of interest with various forecast horizons.

We see that the proportions of a variable of interest for a forecast horizon of 1 both have a value of 1 on its own prediction. This is simply because in this case $\Theta_0^2 = PP = \Sigma_u$ and is a diagonal matrix, hence the forecast error variance of predicting one step ahead fully depends on its own variable of interest. Furthermore we see that forecast error variance of forecasting the first variable mostly depends on the first variable of interest. This means that the second variable of interest almost contributes no information to the first variable of interest. However, while forecasting the second variable of interest, its forecast error variance will in the long term depends more on the first variable of interest. We can conclude from this that in the long term the first variable of interest contributes more information to the second variable of interest than the second variable itself. These results do not come as a suprise, since the top right values of the coefficients matrices A_1 and A_2 have values close to 0, while the bottom left values have larger values.

2.3.2 Granger-causality

Granger-causality is a method to find out whether certain variables of interest are influencing each other. It is based on the idea that the prediction of certain variables of interest should be improved when other influencing variables of interests are added to the process. For example, if we know that variable x is affecting variable z, then the prediction of variable z would be better if we take variable x into account as well. This form of causality we call *Granger-causality*, which is defined as follows.

Definition 2.12. *Suppose x_t and z_t are multidimensional variables of interest. Let $\Sigma_z(h|\Omega_t)$ be the forecast mean squared error of the h-step minimum MSE predictor of z_t as in (2.25) based on the information set Ω_t, which contains all available information of all variables up until time t. Then we say that x_t Granger-causes z_t when*

$$\Sigma_z(h|\Omega_t) \neq \Sigma_z(h|\Omega_t \backslash \{x_s | s \leqslant t\}) \tag{2.56}$$

for at least one $h = 1, 2, \ldots$.

Let us now take a stable VAR(p) process y_t with K variables of interest. We define

$$y_t = \begin{bmatrix} z_t \\ x_t \end{bmatrix}, \tag{2.57}$$

where z_t and x_t are $(M \times 1)$ and $((K - M) \times 1)$ vectors respectively. If we now rewrite y_t into a moving

average representation as in (2.4), we get

$$\begin{aligned} y_t &= \begin{bmatrix} z_t \\ x_t \end{bmatrix} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i} \\ &:= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \Phi_{11,0} & \Phi_{12,0} \\ \Phi_{21,0} & \Phi_{22,0} \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + \begin{bmatrix} \Phi_{11,1} & \Phi_{12,1} \\ \Phi_{21,1} & \Phi_{22,1} \end{bmatrix} \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} + \dots, \end{aligned} \tag{2.58}$$

with appropriate dimensions such that

$$z_t = \mu_1 + \sum_{i=0}^{\infty} \Phi_{11,i} u_{1,t-i} + \sum_{i=0}^{\infty} \Phi_{12,i} u_{2,t-i}, \tag{2.59}$$

$$x_t = \mu_2 + \sum_{i=0}^{\infty} \Phi_{21,i} u_{1,t-i} + \sum_{i=0}^{\infty} \Phi_{22,i} u_{2,t-i}. \tag{2.60}$$

We can now obtain the following useful lemma.

Lemma 2.4. *Let y_t be as in (2.58), then x_t does not Granger-cause z_t if and only if*

$$\Phi_{12,i} = 0 \quad \text{for } i = 1, 2, \dots. \tag{2.61}$$

Proof. First we will show that the minimum MSE predictor of y_t can be rewritten into a moving average representation as in (2.4). We use that the forecast error of the minimum MSE predictor can be rewritten into the moving average representation, which has been shown in (2.25). Since we of course can rewrite y_{t+h} into a moving average representation as well, we find that $y_t(h)$ can also be rewritten into a moving average representation as follows. We use

$$\begin{aligned} y_{t+h} - y_t(h) &= \sum_{i=0}^{h-1} \Phi_i u_{t+h-i} \\ &\leftrightarrow \end{aligned}$$

such that

$$\begin{aligned} y_t(h) &= y_{t+h} - \sum_{i=0}^{h-1} \Phi_i u_{t+h-i} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_i u_{t+h-i} - \sum_{i=0}^{h-1} \Phi_i u_{t+h-i} \\ &= \mu + \sum_{i=h}^{\infty} \Phi_i u_{t+h-i} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_{i+h} u_{t-i}. \end{aligned} \tag{2.62}$$

Now let us first look at the case where $h = 1$. Using (2.62) we will try to find the 1 step prediction of z_t in a moving average representation based on the information set $\Omega_t = \{y_s | s \leqslant t\}$. We use the $(M \times K)$ matrix $Z = (I_M, 0, \dots, 0)$ such that $z_t = Z y_t$. For notation we use that the minimum MSE h step prediction of z_t with forecast origin t based on the information set Ω_t equals $z_t(h|\Omega_t)$, so the 1 step prediction of z_t is

$$\begin{aligned} z_t(1|\Omega_t) &= Z y_t(1) \\ &= Z \left(\mu + \sum_{i=0}^{\infty} \Phi_{i+1} u_{t-i} \right) \\ &= \mu_1 + \sum_{i=0}^{\infty} \Phi_{11,i+1} u_{1,t-i} + \sum_{i=0}^{\infty} \Phi_{12,i+i} u_{2,t-i}. \end{aligned} \tag{2.63}$$

Since

$$\begin{aligned} z_{t+1} &= Zy_{t+1} \\ &= \mu_1 + \sum_{i=0}^{\infty} \Phi_{11,i} u_{1,t+1-i} + \sum_{i=0}^{\infty} \Phi_{12,i} u_{2,t+1-i}, \end{aligned} \tag{2.64}$$

we have a forecast error of $z_t(h|\Omega_t)$ of

$$\begin{aligned} z_{t+h} - z_t(h|\Omega_t) &= \Phi_{11,0} u_{1,t+1} + \Phi_{12,0} u_{2,t+1} \\ &= u_{1,t+1}, \end{aligned} \tag{2.65}$$

because we found in Theorem 2.3 that $\Phi_0 = I_K$.

Secondly we will try to find the 1 step prediction of z_t based on $\Omega_t \backslash \{x_s | s \leqslant t\}$, which is equivalent to $\{z_s | s \leqslant t\}$. To find this prediction we will need an implication from Wold's theorem. This implication states that every subprocess of a stationary process also has a moving average representation, hence z_t has a moving average representation. This simply follows from the fact that

$$\begin{aligned} \mathbb{E}[z_t] &= \mathbb{E}[Zy_t] \\ &= Z\mu \end{aligned}$$

and

$$\begin{aligned} \Gamma_z(h) &= \mathbb{E}[z_t z_t^T] \\ &= \mathbb{E}[Zy_t (Zy_t)^T] \\ &= Z\Gamma_y(h) Z^T, \end{aligned}$$

hence z_t is stationary and has a moving average representation, which follows again from Wold's theorem.

Now we can rewrite z_t as

$$z_t = \mu_1 + \sum_{i=0}^{\infty} F_i v_{t-i}, \tag{2.66}$$

where F_i are some moving average coefficient matrices and v_t are the white noise terms. We can again use (2.62) to find a moving average representation of the minimum MSE 1 step predictor $z_t(1)$, which is

$$z_t(1|\Omega_t \backslash \{x_s | s \leqslant t\}) = \mu_1 + \sum_{i=0}^{\infty} F_{i+1} v_{t-i}. \tag{2.67}$$

Now using (2.66) for $t+1$ and the 1 step predictor (2.67), we find the following forecast error.

$$\begin{aligned} z_{t+1} - z_t(1|\Omega_t \backslash \{x_s | s \leqslant t\}) &= \mu_1 + \sum_{i=0}^{\infty} F_i v_{t+1-i} - \mu_1 - \sum_{i=0}^{\infty} F_{i+1} v_{t-i} \\ &= F_0 v_{t+1} \\ &= v_{t+1}, \end{aligned} \tag{2.68}$$

since again from Theorem 2.3 we have that $F_0 = I_K$.

Finally now all we need to do in order to show when x_t does not Granger-cause z_t is we have to see when the predictors in (2.63) and (2.67) are the same. This is equivalent to checking when the forecast errors in (2.65) and (2.68) are the same, therefore assume that

$$u_{1,t+1} = v_{t+1}. \tag{2.69}$$

Substituting (2.69) into (2.66) results into

$$z_t = \mu_1 + \sum_{i=0}^{\infty} F_i u_{1,t-i}. \tag{2.70}$$

Since from (2.59) we also have that

$$z_t = \mu_1 + \sum_{i=0}^{\infty} \Phi_{11,i} u_{1,t-i} + \sum_{i=0}^{\infty} \Phi_{12,i} u_{2,t-i}, \tag{2.71}$$

we see that (2.70) and (2.71) are the same if and only if $\Phi_{11,i} = F_i$ for $i \geqslant 0$ and $\Phi_{12,i} = 0$ for $i \geqslant 1$. Definition 2.12 now tells us that x_t does not Granger-cause z_t if and only if $\Phi_{12,i} = 0$ for $i = 1, 2, \ldots$ □

In the proof above we showed that the 1 step predictors in (2.63) and (2.67) are the same if and only if $\Phi_{12,i} = 0$ for $i = 1, 2, \ldots$. Using the same methodology used in this proof, it is possible to show that these predictors for h steps ahead are the same if and only if $\Phi_{12,i} = 0$ for $i = 1, 2, \ldots$ as well. Note that only one h has to be found where those predictors are different in order to have Granger-causality.

We can transform the condition that $\Phi_{12,i} = 0$ for $i = 1, 2, \ldots$ in Lemma 2.4 to a condition based on the coefficient matrices $A_1, A_2, \ldots, A_p$ of the stable VAR(p) process y_t as follows. Again we take y_t as in (2.57). We then let $A_{jk,i}$ be as follows.

$$\begin{aligned} y_t &= \begin{bmatrix} z_t \\ x_t \end{bmatrix} \\ &= \nu + A_1 + A_2 + \cdots + A_p \\ &= \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \begin{bmatrix} A_{11,1} & A_{12,1} \\ A_{21,1} & A_{22,1} \end{bmatrix} + \cdots + \begin{bmatrix} A_{11,p} & A_{12,p} \\ A_{21,p} & A_{22,p} \end{bmatrix}, \end{aligned} \tag{2.72}$$

with suitable dimensions for $A_{jk,i}$. We now get the following theorem.

Theorem 2.9. *Let y_t be a stable VAR(p) process as in (2.72) and $A_{jk,i}$ be the j-th row k-th column element of the coefficient matrix A_i, then x_t does not Granger-cause z_t if and only if*

$$A_{12,i} = 0 \quad \text{for } i = 1, 2, \ldots, p. \tag{2.73}$$

Alternatively z_t does not Granger-cause x_t if and only if

$$A_{21,i} = 0 \quad \text{for } i = 1, 2, \ldots, p. \tag{2.74}$$

Proof. Using Lemma 2.4 we have that x_t does not Granger-cause z_t if and only if $\Phi_{12,i} = 0$ for $i = 1, 2, \ldots$. Using this condition together with the Φ matrices we found in Theorem 2.3, we see that $A_{12,i}$ for $i = 1, 2, \ldots, p$ is an equivalent condition, hence x_t does not Granger-cause z_t if and only if

$$A_{12,i} = 0 \quad \text{for } i = 1, 2, \ldots, p. \tag{2.75}$$

The proof to show that z_t does not Granger-cause x_t can be done in a similar way. □

Example 2.5. *Let us again look at the bivariate VAR(2) process*

$$\begin{aligned} y_t &= \nu + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} y_{t-2} + u_t \\ &= \begin{bmatrix} z_t \\ x_t \end{bmatrix}, \end{aligned}$$

where $z_t = y_{1t}$ and $x_t = y_{2t}$. We see that $A_{12,1} = 0.1$ and $A_{12,2} = 0$ and we see that $A_{21,1} = 0.4$ and $A_{21,2} = 0.25$, hence using Theorem 2.9 we have that x_t does Granger-cause z_t and z_t also Granger-causes x_t. This means that a prediction of the process y_{1t} would be improved if the predictor takes the values of the process y_{2t} into account and a prediction of the process y_{2t} would be improved if the predictor takes the values of the process y_{1t} into account.

Example 2.6. *Let us now look at the following 3-dimensional stable VAR(1) process*

$$\begin{aligned} y_t &= \nu + \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_{t-1} + u_t \\ &=: \begin{bmatrix} z_t \\ x_t \end{bmatrix}, \end{aligned}$$

where $z_t = y_{1t}$ *and* $x_t = \begin{bmatrix} y_{2t} \\ y_{3t} \end{bmatrix}$. *We see that* $A_{12,1} = \begin{bmatrix} 0 & 0 \end{bmatrix}$ *and* $A_{21,1} = \begin{bmatrix} 0.1 \\ 0 \end{bmatrix}$, *hence* x_t *does not Granger-cause* z_t, *but* z_t *does Granger-cause* x_t. *We can calculate the forecast mean squared errors of the minimum MSE predictors of* x_t *and* z_t *with and without taking each other into account. In this example we will only be looking at the predictors that forecast 3 steps ahead. Using (2.32) we find*

$$\Sigma_y(3) \approx \begin{bmatrix} 2.953 & 0.146 & 0.011 \\ 0.146 & 1.161 & 0.663 \\ 0.011 & 0.663 & 0.943 \end{bmatrix},$$

which means

$$\begin{aligned} \Sigma_z(3|\Omega_t) &= \Sigma_z(3|\{y_s|s \leqslant t\}) \\ &\approx 2.953 \end{aligned} \tag{2.76}$$

and

$$\begin{aligned} \Sigma_x(3|\Omega_t) &= \Sigma_x(3|\{y_s|s \leqslant t\}) \\ &\approx \begin{bmatrix} 1.161 & 0.663 \\ 0.663 & 0.943 \end{bmatrix}. \end{aligned} \tag{2.77}$$

Now if the calculate the forecast mean squared error of x_t *and* z_t *without taking the other variable into account, we find*

$$\begin{aligned} \Sigma_z(3|\Omega_t \backslash \{x_s|s \leqslant t\}) &= \Sigma_z(3|\{z_s|s \leqslant t\}) \\ &\approx 2.953 \end{aligned} \tag{2.78}$$

and

$$\begin{aligned} \Sigma_x(3|\Omega_t \backslash \{z_s|s \leqslant t\}) &= \Sigma_x(3|\{x_s|s \leqslant t\}) \\ &\approx \begin{bmatrix} 1.131 & 0.661 \\ 0.661 & 0.942 \end{bmatrix}. \end{aligned} \tag{2.79}$$

We see that $\Sigma_z(3|\Omega_t) = \Sigma_z(3|\Omega_t \backslash \{x_s|s \leqslant t\})$, *which is what we expected since* z_t *did not Granger-cause* x_t. *If we would check for values of* h *other than 3, we would find the same result. We also see that* $\Sigma_x(3|\Omega_t) \neq \Sigma_x(3|\Omega_t \backslash \{z_s|s \leqslant t\})$, *hence* x_t *does indeed not Granger-cause* z_t.

2.3.3 Instantaneous causality

Instantaneous causality is a method to find out whether certain variables of interest will have a better 1-step prediction if the values of some other variables of interest of 1 step ahead are already known. Thus whenever adding x_{t+1} to the information set Ω_t is improving the prediction of z_t and conversely, then we speak of instantaneous causality between those variables. We get the following definition for instantaneous causality.

Definition 2.13. *Suppose* x_t *and* z_t *are multidimensional variables of interest. Let* $\Sigma_z(h|\Omega_t)$ *be the forecast mean squared error of the* h*-step minimum MSE predictor of* z_t *as in (2.25) based on the information set* Ω_t, *which contains all available information of all variables up until time* t. *Then we say that there is instantaneous causality between* x_t *and* z_t *when*

$$\Sigma_x(1|\Omega_t) \neq \Sigma_x(1|\Omega_t \cup \{z_{t+1}\}). \tag{2.80}$$

Note that we do not say that there is instantaneous causality from x_t to z_t or from z_t to x_t. We will show that both statements are equivalent, hence we call it instantaneous causality between x_t and z_t. To show whether there is no instantaneous causality between 2 variables, we can simply use the following theorem.

Theorem 2.10. *Let y_t a stable VAR(p) process as in (2.58), then there is instantaneous no causality between x_t and z_t if and only if*

$$\mathbb{E}[u_{1,t}u_{2,t}^T] = 0. \tag{2.81}$$

Proof. We will use the representation of y_t as in (2.44) to obtain

$$\begin{aligned} y_t &= \begin{bmatrix} z_t \\ x_t \end{bmatrix} \\ &= \mu + \sum_{i=0}^{\infty} \Theta_i w_{t-i} \\ &=: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \Theta_{11,0} & \Theta_{12,0} \\ \Theta_{21,0} & \Theta_{22,0} \end{bmatrix} \begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix} + \begin{bmatrix} \Theta_{11,1} & \Theta_{12,1} \\ \Theta_{21,1} & \Theta_{22,1} \end{bmatrix} \begin{bmatrix} w_{1,t-1} \\ w_{2,t-1} \end{bmatrix} + \dots, \end{aligned} \tag{2.82}$$

with appropriate dimensions such that

$$z_t = \mu_1 + \sum_{i=0}^{\infty} \Theta_{11,i} w_{1,t-i} + \sum_{i=0}^{\infty} \Theta_{12,i} w_{2,t-i}, \tag{2.83}$$

$$x_t = \mu_2 + \sum_{i=0}^{\infty} \Theta_{21,i} w_{1,t-i} + \sum_{i=0}^{\infty} \Theta_{22,i} w_{2,t-i}. \tag{2.84}$$

Using this notation we can look at the minimum MSE predictor 1 step ahead for x_t based on the information set $\{y_s | s \leqslant t\} \cup \{z_{t+1}\}$. From the representation in (2.82) we see that this information set is equivalent with $\{w_s | s \leqslant t\} \cup w_{1,t+1}$. Since the $w_{1,t+1}$ is uncorrelated with $\{w_s | s \leqslant t\}$, we find from (2.84) that

$$x_t(1|\{w_s|s \leqslant t\} \cup w_{1,t+1}) = x_t(1|\{w_s|s \leqslant t\}) + \Theta_{21,0} w_{1,t+1},$$

hence there is instantaneous causality between x_t and z_t if and only if $\Theta_{21,0} = 0$. We know from (2.44) that $\Theta_i = \Phi_i P$, where $PP^T = \Sigma_u$. Since from Theorem 2.3 we found that $\Phi_0 = I_K$, we see that $\Theta_0 = P$. We know that P is a lower triangular matrix, hence

$$\begin{aligned} \begin{bmatrix} \Theta_{11,0} & \Theta_{12,0} \\ \Theta_{21,0} & \Theta_{11,0} \end{bmatrix} &= P \\ &= \begin{bmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{bmatrix}, \end{aligned}$$

where P_{jk} and $\Theta_{jk,0}$ have the same dimensions. When $\Theta_{21,0} = 0$, then obviously $P_{21} = 0$. Since $\Sigma_u = PP^T$, we can see that $\text{Cov}(u_{mt}, u_{nt}) = 0$ for $m = 1, 2, \dots, M$ and $n = M+1, M+2, \dots, K$. Therefore there is instantaneous causality between x_t and z_t if and only if $\mathbb{E}[u_{1,t}u_{2,t}^T]$. □

Example 2.7. *Let us take a look again at the 3-dimensional VAR(1) process of Example 2.6, where we still have that $x_t = y_{1t}$ and $z_t = \begin{bmatrix} y_{2t} \\ y_{3t} \end{bmatrix}$. We define the covariance matrix of the white noise terms to be*

$$\Sigma_u = \begin{bmatrix} 2.25 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 0.74 \end{bmatrix}. \tag{2.85}$$

We see that $\mathbb{E}[u_{1,t}u_{2,t}^T] = 0$, hence there is no instantaneous causality between x_t and z_t, which means that the prediction of variable x_{t+1} will not improve if we take z_{t+1} into account.

2.3.4 Impulse Response analysis

With impulse response analysis we can investigate situations where suddenly one variable of interest increases in value. We call the sudden increase of a variable of interest an impulse, which is defined as follows.

Definition 2.14. *Suppose we have a stable VAR(p) process y_t, where we assume that y_t is equal to the mean μ for $t < 0$. Then we define the error terms for an impulse in the k-th variable of interest for $t = 0$ to be*

$$u_{k0} = 1 \quad \textit{and} \quad u_{j0} = 0 \quad \textit{for } j \neq k \tag{2.86}$$

and for $t > 0$ to be

$$u_1 = 0,\ u_2 = 0, \ldots.$$

For example, the impulse in the first variable of interest of a certain VAR(p) process will be

$$u_0 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ u_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\ u_2 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots$$

If we now create an impulse in a certain variable, we are able to see what happens with $y_1, y_2, \ldots$ and look at the effect that the impulse has on all of the variables of interest. We call this effect the *response* of the impulse.

Whenever the variables have different scales it can be useful to create a different impulse than in (2.86), since an impulse with $u_{k0} = 1$ might be a really small or a really big change in comparison with the k-th variable of interest. If that is the case then we can assume the value of u_{k0} to simply be the standard deviation of the k-th variable of interest. If we look for example at the impulse of the first variable of the bivariate VAR(2) process in Example 2.2, we could take u_{10} to be $\sqrt{0.09}$ instead of 1.

Since we are not really interested in the mean of the process, but only in the response on the impulses, we can take for simplicity $\mu = 0$ and $\nu = 0$. It turns out that now the responses caused by certain impulses are as follows.

Theorem 2.11. *Let y_t be a stable VAR(p) process, where $y_t = 0$ for $t < 0$ and $\nu = 0$. Then the response of an impulse in the k-th variable of interest for $t = i$ are the first K values of the k-th column of $\boldsymbol{A}^i$, where $\boldsymbol{A}$ is as in Definition 2.2.*

Proof. We can rewrite y_t in companion form as in Definition 2.2, hence

$$Y_t = \mathbf{A}Y_{t-1} + U_t.$$

An impulse in the k-th variable of y_t would result in $U_{k0} = 1$ and $U_{j0} = 0$ for $j \neq k$ and furthermore $U_t = 0$ for $t > 1$. Hence

$$\begin{aligned} Y_0 &= \mathbf{A}Y_{-1} + U_0 \\ &= U_0. \end{aligned}$$

We now have that

$$\begin{aligned} Y_1 &= \mathbf{A}Y_0 + U_1 \\ &= \mathbf{A}U_0, \\ Y_2 &= \mathbf{A}Y_1 + U_2 \\ &= \mathbf{A}^2U_0, \\ &\vdots \\ Y_i &= \mathbf{A}Y_{i-1} + U_i \\ &= \mathbf{A}^iU_0. \\ &\vdots \end{aligned}$$

Since U_0 is a vector of zeros, except for the k-th element, which is 1, we find that Y_i is the k-th column of $\mathbf{A}$, hence y_i are the first K values of the k-th column of $\mathbf{A}$. □

It is interesting to know when there is no response of a certain variable after an impulse. When this occurs we call this response a *zero impulse response*. Zero impulse responses occur when the variable of the impulse does not Granger-cause the other variables, because then the prediction of the response variables are not influenced by the impulse variable. From Lemma 2.4 it is now obvious that variable j has a zero impulse response from an impulse in variable $k \neq j$ when $\phi_{jk,i} = 0$ for $i = 1, 2, \ldots$, where $\phi_{jk,i} = 0$ is the j-th row k-th column element of Φ_i. In order to check if all $\phi_{jk,i}$ are 0, we do not have to find Φ_i for all values of i, since we have the following proposition (Lütkepohl, 2005, pp. 54-55).

Proposition 2.1. *Let y_t be a stable K-dimensional VAR(p) process, then for $j \neq k$ we have that*

$$\phi_{jk,i} = 0 \quad \textit{for } i = 1, 2, \ldots \tag{2.87}$$

is equivalent with

$$\phi_{jk,i} = 0 \quad \textit{for } i = 1, 2, \ldots, p(K-1). \tag{2.88}$$

This means that we simply have to look at the first $p(K-1)$ matrices of Φ_i to see whether a response of an impulse of a variable of interest is a zero impulse response.

Example 2.8. *Let us take a look at the 3-dimensional VAR(1) process as in Example 2.6, but for simplicity we take $\mu = 0$ and $\nu = 0$, hence $y_t = 0$ for $t < 0$. We get the following process.*

$$y_t = \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_{t-1} + u_t.$$

If we create an impulse in the first variable of interest, we have that

$$u_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \; u_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \; u_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \ldots,$$

hence

$$\begin{aligned} y_0 &= \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_{-1} + u_0 \\ &= \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

Now the responses of this impulse are

$$\begin{aligned} y_1 &= \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_0 + u_1 \\ &= \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 \\ 0.1 \\ 0 \end{bmatrix}, \end{aligned}$$

$$\begin{aligned} y_2 &= \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_1 + u_2 \\ &= \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 0.25 \\ 0.06 \\ 0.02 \end{bmatrix} \end{aligned}$$

and so on. Note that we can also use Theorem 2.11 to find that y_i is the first column of

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix}^i \tag{2.89}$$

and the responses at $t = i$ of the impulse in the second and third variable of interest are the second and third column of (2.89) respectively. We now obtain the following responses.

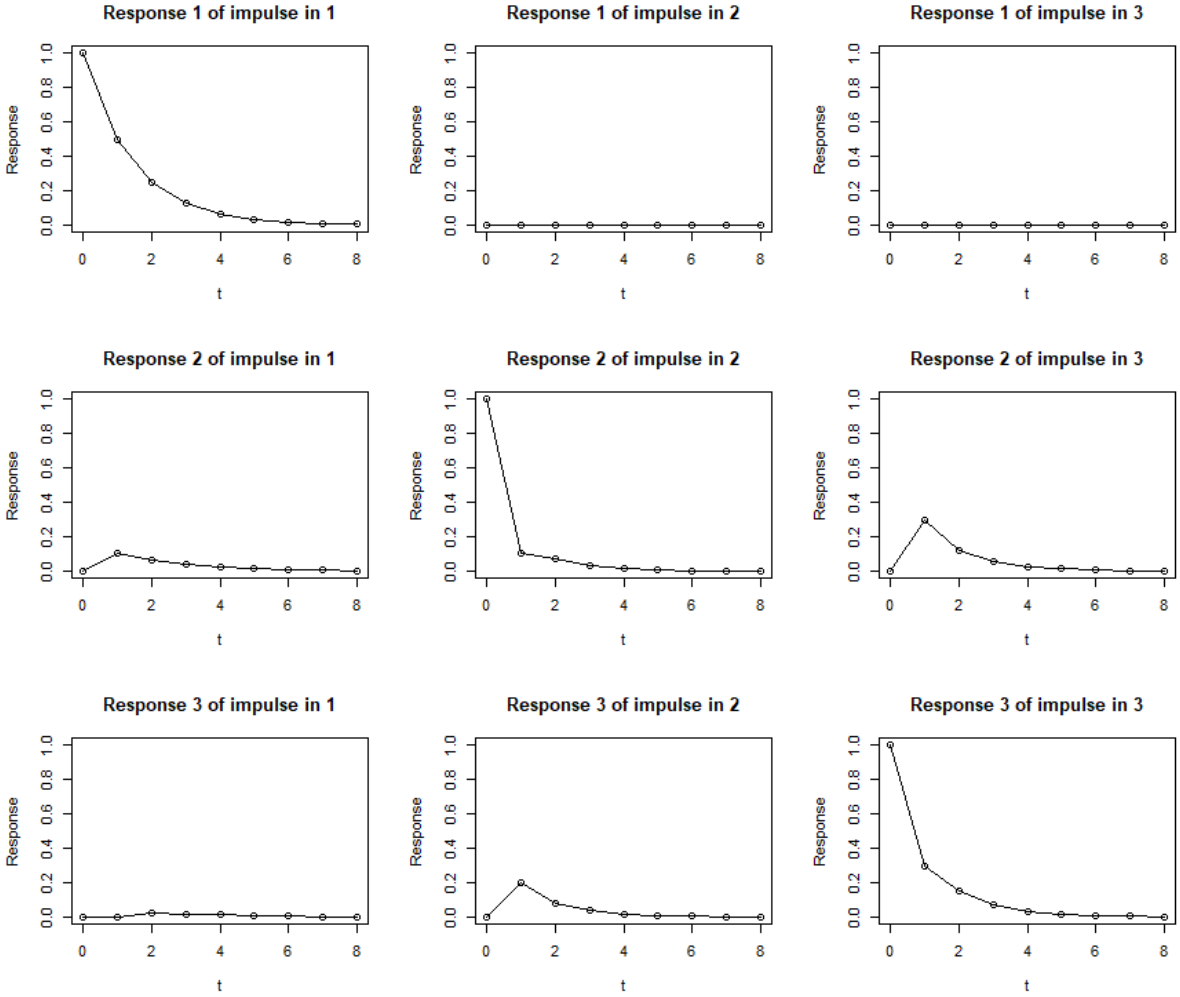


Figure 2.5: Impulse responses of our process.

Since from Example 2.6 we found that the second and third variable of interest do not Granger-cause the first variable of interest, it is obvious that the responses of the impulse in the first variable are 0, as we can see in the figure above. We could also calculate the moving average coefficients as in Theorem 2.3 to obtain

$$\Phi_1 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \tag{2.90}$$

$$\Phi_2 = \begin{bmatrix} 0.25 & 0 & 0 \\ 0.06 & 0.07 & 0.12 \\ 0.02 & 0.08 & 0.15 \end{bmatrix}. \tag{2.91}$$

We see that $\phi_{12,i} = \phi_{13,i} = 0$ for $i = 1, 2$, hence using Proposition 2.1 and the fact that $p(K-1) = 2$, we know that the responses of the second and third variable caused by the impulses of the first variable should indeed be 0.

2.3.5 Orthogonal Impulse Response analysis

The problem with impulse response analysis in the last section is that the impulse only happens in one variable at a time. Whenever the variables of interest are all uncorrelated, then this does not cause any problems, but when some variables are dependent then the responses may not be correct. When we

create an impulse in a variable of interest, we only assume that one error term changes in value at $t = 0$, but when that error term is dependent with an error term of another variable of interest, then in reality both error terms should change in value. Hence the impulse of correlated variables will result in responses that will likely not happen in reality. That is why we will be looking at the representation of a stable VAR(p) process y_t with uncorrelated error terms of the variables of interest as in (2.44). We will create an impulse in the k-th variable of interest that does not necessarily only have one non-zero error term, but which can have multiple non-zero values which depends on the correlation of the error terms. This special impulse will be called an *orthogonal impulse*. We can obtain the following useful theorem.

Theorem 2.12. *Let y_t be a stable VAR(p) process, where $y_t = 0$ for $t < 0$ and $\nu = 0$. Then the response of an orthogonalised impulse in the k-th variable of interest is for $t = i$ the k-th column of Θ_i, where Θ_i is as in (2.44).*

Proof. First, using the representation of y_t as in (2.44), we know that $\Sigma_u = PP^T$ for a lower triangular matrix P. We can rewrite this as

$$\begin{aligned} \Sigma_u &= PP^T \\ &= PD^{-1}DD^T(PD^{-1})^T \\ &= W\Sigma_\epsilon W^T, \end{aligned}$$

where we define D to be the diagonal matrix with the same diagonal as P. Furthermore $W := PD^{-1}$ and $\Sigma_\epsilon := DD^T$.

Remember from (2.1) that our VAR(p) process y_t with $\nu = 0$ looks like

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t. \tag{2.92}$$

Since of course $WW^{-1} = I$, we can rewrite (2.92) as

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + WW^{-1}u_t.$$

Since

$$\begin{aligned} \mathbb{E}[W^{-1}u_t(W^{-1}u_t)^T] &= W^{-1}\mathbb{E}[u_t u_t^T](W^{-1})^T \\ &= W^{-1}\Sigma_u(W^{-1})^T \\ &= W^{-1}P(W^{-1}P)^T \\ &= \Sigma_\epsilon, \end{aligned}$$

we can define $\epsilon_t := W^{-1}u_t$, hence

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + W\epsilon_t.$$

Since $\Sigma_\epsilon = DD^T$ and hence diagonal, it means that the ϵ_t are independent. Now with an impulse of the k-th variable of interest, we have e_{k0} as the value of its standard deviation or equivalently ϵ_0 is the k-th column of D, since $\Sigma_\epsilon = DD^T$ and D is a diagonal matrix. This means we get as response for the impulse in the k-th variable of interest at $t = 0$ that

$$\begin{aligned} y_0 &= W\epsilon_0 \\ &= PD^{-1}\epsilon_0 \\ &= Pe_k, \end{aligned}$$

where $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the unit vector with 1 at the k-th row. This means that y_0 is the k-th column of Θ_0, since from (2.44) we know that $\Theta_0 = P$.

Now for the response on the impulse in the k-th variable we have that

$$\begin{aligned} y_1 &= A_1 y_0 \\ &= \Phi_1 P e_k \\ &= \Theta_1 e_k, \end{aligned}$$

hence the response at time $t = 1$ of the impulse in the k-th variable of interest is the k-th column of Θ_1. Using the same methodology we can find that the response at time $t = i$ of the impulse in the k-th variable of interest is the k-th column of Θ_i. □

Again it is interesting to know when there is no response at all of a certain variable after a certain orthogonal impulse. When this occurs, we call this response a *zero orthogonal impulse response*. A zero orthoganal impulse response of the j-th variable caused by an orthogonal impulse in the k-th variable only happens when of course $\Theta_{jk,i} = 0$ for $i = 1, 2, \ldots$, where $\Theta_{jk,i}$ is the j-th row k-th column element of Θ_i. Again we do not need to find Θ_i for all values of i, since it is possible to find the following proposition (Lütkepohl, 2005, p. 61).

Proposition 2.2. *Let y_t be a stable K-dimensional VAR(p) process, then for $j \neq k$ we have that*

$$\Theta_{jk,i} = 0 \quad \textit{for } i = 1, 2, \ldots$$

is equivalent with

$$\Theta_{jk,i} = 0 \quad \textit{for } i = 1, 2, \ldots, p(K-1).$$

This means we only have to look at the first $p(K-1)$ matrices of Θ_i to see whether a certain response of a variable is a zero orthogonal impulse response.

Example 2.9. *If we continue with the same process as in Example 2.8, we can find Θ_0, Θ_1 and Θ_2 from 2.44 as follows. Since for*

$$P = \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 0.7 \end{bmatrix}$$

we have that $\Sigma_u = PP^T$, we find that

$$\begin{aligned} \Theta_0 &= P \\ &= \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0.5 & 0.7 \end{bmatrix}, \qquad (2.93) \\ \Theta_1 &= \Phi_1 P \\ &= \begin{bmatrix} 0.75 & 0 & 0 \\ 0.15 & 0.25 & 0.21 \\ 0 & 0.35 & 0.21 \end{bmatrix}, \qquad (2.94) \\ \Theta_2 &= \Phi_2 P \\ &= \begin{bmatrix} 0.375 & 0 & 0 \\ 0.090 & 0.130 & 0.084 \\ 0.030 & 0.155 & 0.105 \end{bmatrix}, \qquad (2.95) \end{aligned}$$

where $\Phi_0 = I_K$ which we found in Theorem 2.3 and Φ_1 and Φ_2 are as in (2.90) - (2.91). The Θ_i matrices for $i > 2$ can be found using the same methodology. Now using Theorem 2.12 we find the following responses.

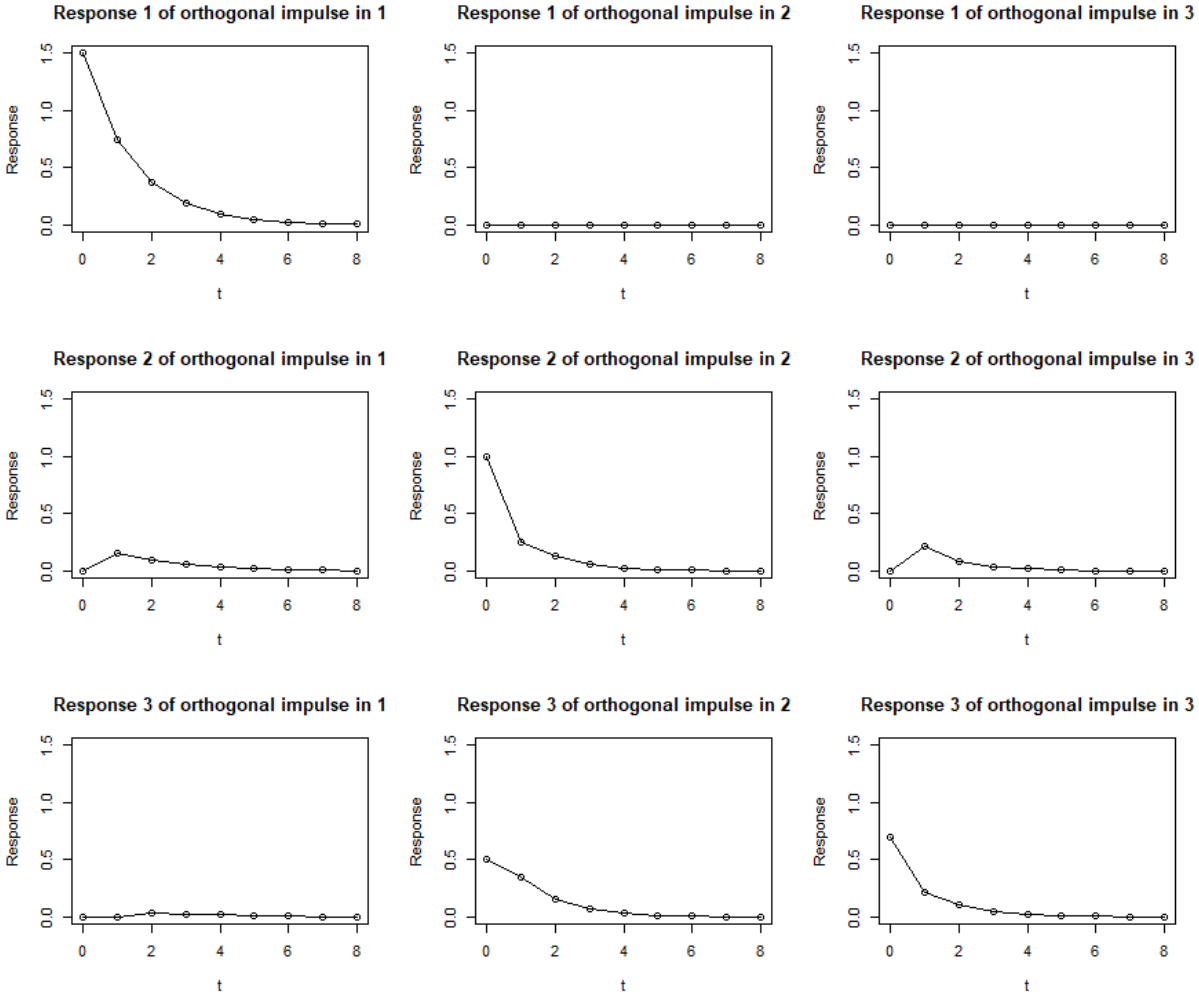


Figure 2.6: Orthogonal impulse responses of our process.

We see that there is no response for the second and third variable of interest on the orthogonal impulse in the first variable for $t = 0, 1, \ldots, 8$. *Using Proposition 2.2, the fact that* $p(K-1) = 2$ *and the matrices (2.93 - 2.95) we can confirm that the response for the second and third variable on the orthogonal impulse in the first variable are indeed zero orthogonal impulse responses.*

2.4 Estimators

In the previous sections we always assumed that we knew the intercept ν and the coefficient matrices $A_1, A_2, \ldots, A_p$ of a VAR(p) process. However in reality these matrices are not known beforehand and they have to be estimated. In this section we will discuss two different methods we can use to estimate these parameters

We will assume that we have a time series $y_1, y_2, \ldots, y_N$ with K variables of interest, hwere N is the sample size. We assume that for $t = 1, 2, \ldots, N$ the process can be fully generated by a VAR(p) process as in (2.1), so we assume that $y_{-p+1}, \ldots, y_{-1}, y_0$ is available as well. We call these values the presample values.

2.4.1 Ordinary Least Squares estimator

The first estimator we will investigate is the Ordinary Least Squares (OLS) estimator. For notation it will be useful to rewrite our time series using the following parameters.

Definition 2.15. *We define*

$$\begin{aligned}
Y &:= (y_1, y_2, \ldots, y_N) && (K \times N),\\
B &:= (\nu, A_1, A_2, \ldots, A_p) && (K \times (1 + Kp)),\\
Z_t &:= \begin{bmatrix} 1 \\ y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix} && ((1 + Kp) \times 1),\\
Z &:= (Z_0, Z_1, \ldots, Z_{N-1}) && ((1 + Kp) \times N),\\
U &:= (u_1, u_2, \ldots, u_N) && (K \times N),\\
\boldsymbol{y} &:= vec(Y) && (KN \times 1),\\
\boldsymbol{\beta} &:= vec(B) && ((K(1 + Kp)) \times 1),\\
\boldsymbol{b} &:= vec(B^T) && ((K(1 + Kp)) \times 1),\\
\boldsymbol{u} &:= vec(U) && (KN \times 1),
\end{aligned} \tag{2.96}$$

where our time series can now be rewritten compactly in the form

$$Y = BZ + U. \tag{2.97}$$

Since from (2.97) we have that

$$\text{vec}(Y) = \text{vec}(BZ) + \text{vec}(U), \tag{2.98}$$

we also find using Lemma 2.3

$$\mathbf{y} = (Z^T \otimes I_K)\boldsymbol{\beta} + \mathbf{u}, \tag{2.99}$$

where $\otimes$ is the Kronecker product as in Definition 2.7. The OLS estimator is the vector $\boldsymbol{\beta}$ in (2.99) that minimizes the sum of squared residuals, or equivalently the sum of squared error terms. In other words, the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes the function $S(\boldsymbol{\beta}) = \mathbf{u}^T\mathbf{u}$. This common problem has already been solved many times for vectors (Rice, 2007, p. 573) and the least squares estimator of (2.99) turns out to be

$$\hat{\boldsymbol{\beta}} = [(Z^T \otimes I_K)^T(Z^T \otimes I_K)]^{-1}(Z^T \otimes I_K)^T\mathbf{y} \tag{2.100}$$

We will simplify $\hat{\boldsymbol{\beta}}$ to find a least squares estimator for B in terms of Y and Z as in (2.98). To do so we will use Lemma (2.3) and the following lemmas (Lütkepohl, 2005, pp. 661-662) with the vec operator and the Kronecker product.

Lemma 2.5. *Let A and B be matrices, then*

$$(A \otimes B)^T = A^T \otimes B^T,$$

where $\otimes$ is the Kronecker product as in Definition 2.7.

Lemma 2.6. *Let A, B, C and D be matrices, then*

$$(A \otimes B)(C \otimes D) = AC \otimes BD,$$

where $\otimes$ is the Kronecker product as in Definition 2.7. We assume matrices A, B, C and D to have suitable dimensions for all matrix products used.

Now we can find the OLS estimator for B in the following theorem.

Theorem 2.13. *The least squares estimator for B in terms of Y and Z is*

$$\hat{B} = YZ^T(ZZ^T)^{-1} \tag{2.101}$$

Proof. From (2.100) we have that

$$\begin{aligned} \text{vec}(\hat{B}) &= \hat{\boldsymbol{\beta}} \\ &= [(Z^T \otimes I_K)^T(Z^T \otimes I_K)]^{-1}(Z^T \otimes I_K)^T\mathbf{y}. \end{aligned}$$

Using Lemma 2.5 we find that

$$\text{vec}(\hat{B}) = [(Z \otimes I_K)(Z^T \otimes I_K)]^{-1}(Z \otimes I_K)\mathbf{y}.$$

Now using Lemma 2.6 we find that

$$\text{vec}(\hat{B}) = (ZZ^T \otimes I_K)^{-1}(Z \otimes I_K)\mathbf{y}$$

Now finally using Lemma 2.3 twice, we find

$$\begin{aligned} \text{vec}(\hat{B}) &= (ZZ^T \otimes I_K)^{-1}\text{vec}(YZ^T) \\ &= \text{vec}(YZ^T(ZZ^T)^{-1}). \end{aligned}$$

Since both sides of the equation are matrices with the vec operator, it is obvious that

$$\hat{B} = YZ^T(ZZ^T)^{-1}.$$

□

2.4.2 Asymptotic properties of the Ordinary Least Squares estimator

The OLS estimator has some useful asymptotic properties including *consistency* and *asymptotic normality*. Before we look at these properties, let us first introduce definition of *standard white noise*.

Definition 2.16. *A standard white noise process is a white noise process u_t where all fourth moments exist and are bounded, hence for some constant c we have for all t that*

$$\mathbb{E}[u_{it}u_{jt}u_{kt}u_{mt}] \leqslant c \quad \text{for } i, j, k, m = 1, 2, \ldots, K.$$

The definition of the standard white noise allows us to create some restrictions on the residuals. Note that the assumption that u_t is i.i.d. normally distributed with a covariance matrix Σ_u is still a valid assumption, since it still suffices the condition of a standard white noise process. Now assuming we have a standard white noise process, it can be found that the OLS estimator has the following asymptotic properties (Lütkepohl, 2005, pp. 73-74).

Proposition 2.3. *Suppose y_t is a stable K-dimensionl VAR(p) process with standard white noise residuals, where y_t can be rewritten as in Definition 2.15. Let $\hat{B}$ be the least squares estimator of the VAR coefficients B, then the following asymptotic properties hold.*

1. *The least squares estimator $\hat{B}$ is consistent, i.e.*

$$\text{plim}\, \hat{B} = B.$$

2. *We have asymptotic normality of*

$$\sqrt{N}\ \text{vec}(\hat{B} - B) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_u),$$

where $\otimes$ is the Kronecker product as in Definition 2.7 and $\Gamma = \text{plim}\, \frac{ZZ^T}{N}$.

Note that plim in the first asymptotic property is equivalent with convergence in probability. In order to apply the second property in Proposition 2.3, we need to know the matrix $\Gamma \otimes \Sigma_u$. Since not all values of this matrix are known beforehand, we will have to estimate these matrices. The matrix Γ however is known beforehand, since Z and N are values that we can simply extract from our time series. Hence an estimator for Γ will be

$$\hat{\Gamma} = \frac{ZZ^T}{N}. \tag{2.102}$$

Since we know that $\Sigma_u = \mathbb{E}[u_t u_t^T]$, we know that a good estimator of Σ_u would simply be the the sample mean as

$$\begin{aligned} \tilde{\Sigma}_u &= \frac{1}{N}\sum_{t=1}^{N} \hat{u_t}\hat{u_t}^T \\ &= \frac{1}{N}\hat{U}\hat{U}^T, \end{aligned}$$

where $\hat{u_t}$ are the residuals of the VAR model with the estimated coefficients. Using (2.97) we find

$$\tilde{\Sigma}_u = \frac{1}{N}(Y - \hat{B}Z)(Y - \hat{B}Z)^T.$$

However Lütkepohl (2005, p. 75) suggested that the degrees of freedom should be taken into account, since this estimator might lead to a biased estimator of the covariance matrix. We then obtain the estimator

$$\hat{\Sigma}_u = \frac{1}{N - Kp - 1}(Y - \hat{B}Z)(Y - \hat{B}Z)^T, \tag{2.103}$$

since for each variable of interest $(Kp + 1)$ parameters have to be estimated.

Example 2.10. *In this example we want to take a look at the consistency of the OLS estimator. We will look at the error that this estimator is making while estimating the VAR coefficient A_1 of the following 3-dimensional VAR(1) process*

$$y_t = \nu + \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} y_{t-1} + u_t, \tag{2.104}$$

with multivariate normally distributed u_t, which has a covariance matrix

$$\Sigma_u = \begin{bmatrix} 2.25 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 0.74 \end{bmatrix}. \tag{2.105}$$

We choose

$$\nu = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \textit{and} \quad y_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

such that we can now generate our process from y_1 till y_N for an integer N by taking N samples of u_t. Using Theorem 2.13 we can find the least squares estimator $\hat{B}$ of our generated process.

The consistency of the least squares estimator tells us that

$$\operatorname{plim} \hat{B} = B, \tag{2.106}$$

where

$$B = \begin{bmatrix} 1 & 2.25 & 0 & 0 \\ 2 & 0 & 1 & 0.5 \\ 3 & 0 & 0.5 & 0.74 \end{bmatrix}.$$

Let us now generate y_1 till y_N an amount of 1000 times for some value of N and define the OLS estimation of the coefficient matrix A_1 of the i-th generation to be $\hat{A}_{(i)}$. We now also define $\widehat{\boldsymbol{A}}$ to be

$$\widehat{\boldsymbol{A}} = \frac{1}{1000} \sum_{i=1}^{1000} ||\hat{A}_{(i)} - A||_F,$$

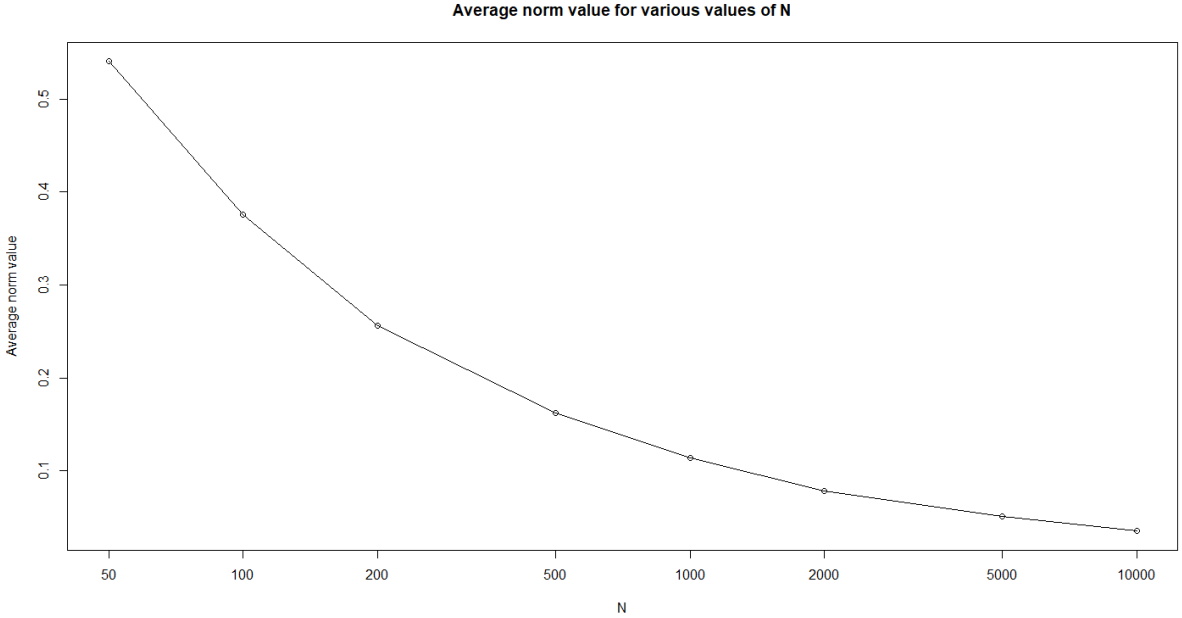
where

$$||\hat{A}_{(i)} - A||_F := \sqrt{\text{tr}\left[(\hat{A}_{(i)} - A)(\hat{A}_{(i)} - A)^T\right]}$$

is the Frobenius norm (Golub and van Loan, 1996). Then using the law of large numbers it is obvious that (2.106) implies that

$$\widehat{\boldsymbol{A}} \to 0,$$

for $N \to \infty$. We find the following values of $\widehat{\boldsymbol{A}}$ for various values of N.



Values of $\widehat{\boldsymbol{A}}$ for various N.

As expected, we see $\widehat{\boldsymbol{A}}$ moving towards 0 for $N \to \infty$. Also we see for $N \geqslant 1000$ that approximately $\widehat{\boldsymbol{A}} \leqslant 0.1$, hence we could conclude that for $N \geqslant 1000$ the OLS estimator of a VAR(1) process is approximately B.

2.4.3 *t*-Ratios

With t-ratios we can determine which values of the OLS estimator $\hat{B}$ are actually significant, i.e. not equal to 0. Let us define $\hat{\beta}_i$ and β_i to be the i-th element of $\text{vec}(\hat{B})$ and $\text{vec}(B)$ respectively. Also let $\hat{s}_i$ be the square root of the i-th row i-th column element of $(ZZ^T)^{-1} \otimes \hat{\Sigma}_u$. From the second asymptotic property in Proposition 2.3 we can now see that

$$\frac{\hat{\beta}_i - \beta_i}{\hat{s}_i} \sim \mathcal{N}(0,1) \quad \forall i. \tag{2.107}$$

This means that we can simply take $\beta_i = 0$ and thus divide the least squares estimator $\hat{B}$ by all corresponding values of $\hat{s}_i$ to obtain the *t-ratios*. Then we can look at the t-distribution with our degrees of freedom in order to look if our estimated coefficients are significant. Usually the sample size minus the

amount of parameters is taken as degrees of freedom, which is $KN - K(Kp+1)$, however Lütkepohl (2005, p. 77) argued that $N - (Kp+1)$ suffices as well. The t-distribution will be close to the standard normal distribution for large degrees of freedom, hence these differences between degrees of freedom will not differ so much anyways when we have a large data set.

Example 2.11. *Continuing with the same stable 3-dimensional VAR(1) process y_t as in Example 2.10, we can generate this process for $N = 1000$. We take $N = 1000$ since we showed in Example 2.10 that for this value the OLS estimator is a good approximation of the coefficients. We obtain the OLS estimator*

$$\begin{aligned} \hat{B} &= \begin{bmatrix} \hat{\nu} & \hat{A}_1 \end{bmatrix} \\ &\approx \begin{bmatrix} 1.049 & 0.451 & 0.007 & 0.007 \\ 1.766 & 0.127 & 0.067 & 0.357 \\ 3.013 & -0.003 & 0.198 & 0.304 \end{bmatrix}. \end{aligned} \tag{2.108}$$

When we calculate all values of $\hat{s}_i$, we can find the t-ratios corresponding to the coefficients of $\hat{B}$ as

$$\begin{bmatrix} 3.739 & 15.846 & 0.131 & 0.107 \\ 9.454 & 6.703 & 1.810 & 8.518 \\ 19.097 & -0.160 & 6.301 & 8.591 \end{bmatrix}. \tag{2.109}$$

When looking at the t-distribution with $N - (Kp+1) = 996$ degrees of freedom, we find that coefficients with t-ratios between approximately -1.962 and 1.962 will not be significant with a significance level of 5%. In (2.109) we see that only 4 coefficients have t-ratios between -1.962 and 1.962, but all other coefficients are significant. The t-ratios of the coefficients with estimations close to 0 will of course have low t-ratios. The second row second column element of A_1 is 0.1, hence it is not surprising that the t-ratio of this coefficient shows that it is not significant for $N = 1000$. However if we would take larger values for N, then we would see that only three t-ratios would not be significant, which are the coefficients of A_1 that are 0. This does not mean that we have chosen N wrong. We only need 1 significant value in the intercept and 1 in each of the coefficient matrices to show that the whole vector or matrix is not 0.

2.4.4 Maximum Likelihood estimator

The second method to estimate the intercept and the coefficient matrices of a VAR(p) model is the maximum likelihood estimator. To derive this estimator, we will be using the following notations.

Definition 2.17. *Define*

$$\begin{aligned} \mu &:= \mathbb{E}[y_t] && (K \times 1), \\ Y^0 &:= (y_1 - \mu, y_2 - \mu, \dots, y_N - \mu) && (K \times N), \\ A &:= (A_1, A_2, \dots, A_p) && (K \times Kp), \\ Y_t^0 &:= \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} && (Kp \times 1), \\ X &:= (Y_0^0, Y_1^0, \dots, Y_{N-1}^0) && (Kp \times N), \\ U &:= (u_1, u_2, \dots, u_N) && (K \times N), \\ \boldsymbol{\alpha} &:= \operatorname{vec}(A) && (K^2 p \times 1), \\ \boldsymbol{y} &:= \operatorname{vec}(Y) && (KN \times 1), \\ \boldsymbol{\mu}^* &= \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} && (KN \times 1), \end{aligned}$$

such that we can write the mean-adjusted form of the VAR(p) process

$$(y_t - \mu) = A_1(y_{t-1} - \mu) + A_2(y_{t-2} - \mu) + \cdots + A_p(y_{t-p} - \mu) + u_t$$

compactly as

$$Y^0 = AX + U. \tag{2.110}$$

The mean-adjusted form of the VAR(p) process is used so that the asymptotic normality we will find of the maximum likelihood estimators will be independent. The idea of the maximum likelihood estimator is to find the values of the parameters such that the log-likelihood function of the parameters $\mu, \boldsymbol{\alpha}, \Sigma_u$ is optimized, where we assume the residuals to be multivariate normally distributed. The methodology of the calculations we will use to find the log-likelihood function and the maximum likelihood parameters can be found in (Lütkepohl, 2005, pp. 87-90). We will simply present the results and analyse them.

We can find the log-likelihood function of

Proposition 2.4. *The log-likelihood function l of the parameters $\mu, \boldsymbol{\alpha}, \Sigma_u$ is*

$$l(\mu, \boldsymbol{\alpha}, \Sigma_u) = -\frac{KN}{2}\ln(2\pi) - \frac{N}{2}\ln(\det(\Sigma_u)) - \frac{1}{2}\operatorname{tr}\left[(Y^0 - AX)^T\Sigma_u^{-1}(Y^0 - AX)\right].$$

If we now would want to find the parameters $\tilde{\mu}, \tilde{\boldsymbol{\alpha}}$ and $\tilde{\Sigma}_u$ which maximize the log-likelihood function $l(\mu, \boldsymbol{\alpha}, \Sigma_u)$, we have to those values such that

$$\frac{\partial\, l(\mu, \boldsymbol{\alpha}, \Sigma_u)}{\partial \mu} = 0, \tag{2.111}$$

$$\frac{\partial\, l(\mu, \boldsymbol{\alpha}, \Sigma_u)}{\partial \boldsymbol{\alpha}} = 0, \tag{2.112}$$

$$\frac{\partial\, l(\mu, \boldsymbol{\alpha}, \Sigma_u)}{\partial \Sigma_u} = 0, \tag{2.113}$$

respectively holds. We then call $\tilde{\mu}, \tilde{\boldsymbol{\alpha}}$ and $\tilde{\Sigma}_u$ the *maximum likelihood estimators*. It turns out that these estimators can be found in the following proposition.

Theorem 2.14. *The maximum likelihood estimators $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\Sigma}_u$ which solves (2.111)-(2.113) can be found by solving the set of equations*

$$\begin{aligned}
\tilde{\mu} &= \frac{1}{N}\left(I_K - \sum_{i=1}^{p}\tilde{A}_i\right)^{-1}\sum_{t=1}^{N}\left(y_t - \sum_{i=1}^{p}\tilde{A}_i y_{t-i}\right), \\
\tilde{\boldsymbol{\alpha}} &= ((\tilde{X}\tilde{X}^T)^{-1}\tilde{X} \otimes I_K)(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}^*), \\
\tilde{\Sigma}_u &= \frac{1}{N}(\tilde{Y}^0 - \tilde{A}\tilde{X})(\tilde{Y}^0 - \tilde{A}\tilde{X})^T,
\end{aligned}$$

where $\tilde{Y}^0$ and $\tilde{X}$ are obtained from Definition 2.17 by using $\tilde{\mu}$ instead of μ. The matrices $\tilde{A}_i$ come from $\tilde{\boldsymbol{\alpha}} := \operatorname{vec}(\tilde{A})$, where $\tilde{A} := (\tilde{A}_1, \ldots, \tilde{A}_p)$.

It turns out that the maximum likelihood estimators $\tilde{\mu}$ and $\tilde{\boldsymbol{\alpha}}$ are actually the same as the least squares estimator $\hat{B}$ as in Theorem 2.13. One might wonder why we would also analyse the maximum likelihood estimator, since the result is the same. This is because our maximum likelihood estimator has some interesting asymptotic properties, which we obtain in the following section.

2.4.5 Asymptotic properties of the Maximum Likelihood estimator

The maximum likelihood estimators also have their own asymptotic properties, just as the OLS estimator. Before we take a look at the asymptotic properties, let us first look at the following definitions.

The so-called *vech operator* is almost the same as the vec operator from Definition 2.6. This operator is mostly used for symmetric matrices, since the vech operator does not collect the duplicates of the elements which are above the diagonal. It is defined as follows.

Definition 2.18. *If A is a $(m \times m)$ matrix, then the vech operator returns the $(\frac{m(m+1)}{2} \times 1)$ vector*

$$\text{vech}(A) := (a_{11}, \ldots . a_{m1}, a_{22}, \ldots, a_{m2}, \ldots, a_{(m-1)(m-1)}, a_{m(m-1)}, a_{mm})^T,$$

which are the stacked columns of A, but only with the elements that are on or below the diagonal.

Example 2.12. *For a (3×3) matrix we have*

$$\text{vech}\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{22} \\ a_{32} \\ a_{33} \end{bmatrix}.$$

Secondly, let us look at the *duplication matrix*.

Definition 2.19. *The duplication matrix $\boldsymbol{D}_K$ is a $(K^2 \times \frac{K(K+1)}{2})$ matrix such that for any $(K \times K)$ matrix A we have that*

$$\text{vec}(A) = \boldsymbol{D}_K \, \text{vech}(A).$$

Furthermore using these definitions we define $\mathbf{D}_K^+$ and $\boldsymbol{\sigma}$ as follows.

Definition 2.20. *We define*

$$\begin{aligned} \boldsymbol{D}_K^+ &:= (\boldsymbol{D}_K^T \boldsymbol{D}_K)^{-1} \boldsymbol{D}_K^T, \\ \boldsymbol{\sigma} &:= \text{vech}(\Sigma_u) \end{aligned}$$

where $\boldsymbol{D}_K$ is a duplication matrix.

Note that $\tilde{\boldsymbol{\sigma}}$ can be found the same way in the definition above, but with using $\tilde{\Sigma}_u$.

Using these definitions, we can find the asymptotic properties of the maximum likelihood estimators. Again, the methodology of the calculations of these properties and can be found in Lütkepohl, 2005, pp. 90-93. The following properties can be found.

Proposition 2.5. *Suppose y_t is a stable $K-$dimensional VAR(p) process with normally distributed error terms, then the following asymptotic properties hold.*

1. *The maximum likelihood estimators $\tilde{\mu}, \tilde{\boldsymbol{\alpha}}$ and $\tilde{\Sigma}_u$ are consistent estimators, i.e. they converge in probability to $\mu, \boldsymbol{\alpha}$ and Σ_u respectively.*
2. *We have asymptotic normality of*

$$\sqrt{N}\begin{bmatrix} \tilde{\mu} - \mu \\ \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{\tilde{\mu}} & 0 & 0 \\ 0 & \Sigma_{\tilde{\boldsymbol{\alpha}}} & 0 \\ 0 & 0 & \Sigma_{\tilde{\boldsymbol{\sigma}}} \end{bmatrix}\right),$$

where

$$\begin{aligned} \Sigma_{\tilde{\mu}} &= \left(I_K - \sum_{i=1}^{p} A_i\right)^{-1} \Sigma_u \left(I_K - \sum_{i=1}^{p} A_i^T\right)^{-1}, \\ \Sigma_{\tilde{\boldsymbol{\alpha}}} &= \Gamma_Y(0)^{-1} \otimes \Sigma_u, \\ \Sigma_{\tilde{\boldsymbol{\sigma}}} &= 2\boldsymbol{D}_K^+ (\Sigma_u \otimes \Sigma_u)(\boldsymbol{D}_K^+)^T, \end{aligned}$$

where $\Gamma_Y(0) := \text{plim} \, \frac{XX^T}{N}$.

The matrices $\Sigma_{\tilde{\mu}}, \Sigma_{\tilde{\boldsymbol{\alpha}}}$ and $\Sigma_{\tilde{\boldsymbol{\sigma}}}$ could be estimated with consistent estimators such that the asymptotic normality still holds. We therefore could use the consistent estimators $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\sigma}}$ and $\hat{\Gamma}_Y(0) := \frac{\tilde{X}\tilde{X}^T}{N}$ in $\Sigma_{\tilde{\mu}}, \Sigma_{\tilde{\boldsymbol{\alpha}}}$ and $\Sigma_{\tilde{\boldsymbol{\sigma}}}$ and obtain the same asymptotic normality.

2.4.6 Forecasting With Estimated Coefficients

Forecasting with estimated coefficients by an OLS or maximum likelihood estimation will be different than forecasting a known VAR process. Since we will be forecasting with estimated coefficients, it means that our h-step predictor will have to be estimated as well. This results into a different forecast error variance.

To show this occurrence, let us assume we estimated a VAR(p) process y_t. We then find using (2.29) that the estimated minimum MSE predictor is

$$\hat{y}_t(h) = \hat{\nu} + \hat{A}_1\hat{y}_t(h-1) + \cdots + \hat{A}_p\hat{y}_t(h-p), \tag{2.114}$$

where $\hat{y}_t(j) := y_{t+j}$ for $j \leqslant 0$. Using the result we found in (2.25) we find a forecast error of $\hat{y}_t(h)$ as

$$\begin{aligned} y_{t+h} - \hat{y}_t(h) &= (y_{t+h} - y_t(h)) + (y_t(h) - \hat{y}_t(h)) \\ &= \left(\sum_{i=0}^{h-1} \Phi_i u_{t+h-i}\right) + (y_t(h) - \hat{y}_t(h)). \end{aligned} \tag{2.115}$$

All we now need to find to start forecasting with forecast error intervals is the covariance matrix of the forecast error. Let us first look at the multivariate *delta method* in the following lemma (Doob, 1935).

Lemma 2.7. *Let $g(\beta) = (g_1(\beta), \ldots, g_K(\beta)^T$ be a continuous differentiable function and $\hat{\beta}$ be an estimator of the $(K \times 1)$ vector β with $\sqrt{N}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \Sigma)$. If $\frac{\partial g}{\partial \beta} \neq 0$, then*

$$\sqrt{N}(g(\hat{\beta}) - g(\beta)) \sim \mathcal{N}(0, \frac{\partial g(\beta)}{\partial \beta^T}\Sigma\frac{\partial g(\beta)^T}{\partial \beta}).$$

Using this lemma, we can obtain the following theorem.

Theorem 2.15. *The covariance matrix of the forecast error of $\hat{y}_t(h)$ is*

$$\Sigma_{\hat{y}}(h) = \Sigma_y(h) + \frac{\Omega(h)}{N},$$

with

$$\Omega(h) := \frac{\partial y(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}(\Gamma^{-1} \otimes \Sigma_u)\frac{\partial y(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}},$$

where $\boldsymbol{\beta} := \text{vec}(B)$ is defined as in Definition 2.15.

Proof. Using $\Sigma_y(h)$ from earlier in (2.32), we can find using the forecast error in (2.115) the following covariance matrix of the forecast error of $\hat{y}_t(h)$.

$$\begin{aligned} \Sigma_{\hat{y}}(h) &:= \text{MSE}[\hat{y}_t(h)] \\ &= \Sigma_y(h) + \text{MSE}[y_t(h) - \hat{y}_t(h)]. \end{aligned}$$

To obtain the MSE of $y_t(h) - \hat{y}_t(h)$ we will be using Lemma 2.7. Since y_t is obviously a continuous differentiable function with parameter $\boldsymbol{\beta}$, using this lemma we obtain

$$\sqrt{N}(\hat{y}_t(h) - y_t(h)) \sim \mathcal{N}(0, \frac{\partial y(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}(\Gamma^{-1} \otimes \Sigma_u)\frac{\partial y(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}}),$$

or

$$\hat{y}_t(h) - y_t(h) \sim \mathcal{N}(0, \frac{\Omega(h)}{N}). \tag{2.116}$$

Now (2.116) suggests that the covariance matrix of $y_t(h) - \hat{y}_t(h)$ is $\frac{\Omega(h)}{N}$, hence we can now fill in (2.116), which results in

$$\Sigma_{\hat{y}}(h) = \Sigma_y(h) + \frac{\Omega(h)}{N}.$$

□

Now we just need to find the function $\Omega(h)$, since the expression of Ω in Theorem 2.15 is definitely not trivial. It turns out that we can find following proposition (Lütkepohl, 2005, pp. 96-98).

Proposition 2.6. *The function $\Omega(h)$ in Theorem 2.15 can be expressed as*

$$\Omega(h) = \sum_{i=0}^{h-1} \sum_{j=0}^{h-1} \operatorname{tr}\left[(\boldsymbol{B}^T)^{h-1-i} \Gamma^{-1} \boldsymbol{B}^{h-1-j} \Gamma \right] \Phi_i \Sigma_u \Phi_j^T,$$

with

$$\boldsymbol{B} := \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ \nu & A_1 & A_2 & \dots & A_{p-1} & A_p \\ 0 & I_K & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & I_K & 0 \end{bmatrix}.$$

When we are estimating $\boldsymbol{\beta}$, we can not exactly determine $\Omega(h)$, but we will again have estimate it. We will have to find an estimate $\hat{\Omega}(h)$, which is defined the same way as in Proposition 2.6, but we use the estimated coefficients $\hat{\nu}, \hat{A}_1, \dots, \hat{A}_p$ and the estimators $\hat{\Gamma}$ and $\hat{\Sigma}_u$, which are all consistent estimators.

Now we are able to determine the covariance matrix of the forecast error of $\hat{y}_t(h)$. For similar reasons we earlier found the forecast intervals (2.36), we have a $(1-\alpha)100\%$ forecast interval of the k-th variable of interest of predicting h steps ahead of

$$\left[\hat{y}_{k,t}(h) - \hat{\sigma}_k(h) z_{\alpha/2}, \hat{y}_{k,t}(h) + \hat{\sigma}_k(h) z_{\alpha/2} \right], \tag{2.117}$$

where

$$\begin{aligned} y_{k,t}(h): &\quad k\text{-th element of } \hat{y}_t(h), \\ \hat{\sigma}_k(h): &\quad \text{square root of the } k\text{-th row } k\text{-th column element of } \hat{\Sigma}_{\hat{y}}(h), \\ z_\alpha: &\quad \text{value such that } \mathbb{P}(Z \leqslant z_\alpha) = 1-\alpha, \text{ where } Z \sim \mathcal{N}(0,1). \end{aligned}$$

Example 2.13. *Let us continue from Example 2.11 where we generated a 3-dimensional VAR(1) process for $N = 1000$. In this example we will forecast this generated process up to 2 steps ahead using (2.114), where we will need $\hat{\nu}, \hat{A}_1$ and $\hat{y}_{1000}(0)$. We can find $\hat{\nu}$ and $\hat{A}_1$ using the least square estimation we performed in (2.108). The value of $\hat{y}_{1000}(0)$ in our generation turns out to be*

$$\begin{aligned} \hat{y}_{1000}(0) &:= y_{1000} \\ &\approx \begin{bmatrix} 4.325 \\ 1.327 \\ 3.786 \end{bmatrix}. \end{aligned}$$

Using (2.114) allows us to find the forecasts

$$\begin{aligned} \hat{y}_{1000}(1) &= \hat{\nu} + \hat{A}_1 \hat{y}_{1000}(0) \\ &\approx \begin{bmatrix} 1.049 \\ 1.766 \\ 3.013 \end{bmatrix} + \begin{bmatrix} 0.451 & 0.007 & 0.007 \\ 0.127 & 0.067 & 0.357 \\ -0.003 & 0.198 & 0.304 \end{bmatrix} \begin{bmatrix} 4.325 \\ 1.327 \\ 3.786 \end{bmatrix} \\ &\approx \begin{bmatrix} 3.035 \\ 3.755 \\ 4.414 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned}\hat{y}_{1000}(2) &= \hat{\nu} + \hat{A}_1\hat{y}_{1000}(1) \\ &\approx \begin{bmatrix}1.049\\1.766\\3.013\end{bmatrix} + \begin{bmatrix}0.451 & 0.007 & 0.007\\0.127 & 0.067 & 0.357\\-0.003 & 0.198 & 0.304\end{bmatrix}\begin{bmatrix}3.035\\3.755\\4.414\end{bmatrix} \\ &\approx \begin{bmatrix}2.475\\3.978\\5.088\end{bmatrix}.\end{aligned}$$

Now from Theorem 2.15 we see that we can estimate the covariance matrix of the forecast error of $\hat{y}_{100}(h)$ *with*

$$\hat{\Sigma}_{\hat{y}}(h) = \hat{\Sigma}_y(h) + \frac{\hat{\Omega}(h)}{N}. \tag{2.118}$$

First, we see that we need to obtain $\hat{\Sigma}_y(1)$ *and* $\hat{\Sigma}_y(2)$*. From (2.32) we see that we therefore will need to find* $\hat{\Sigma}_u, \hat{\Phi}_0$ *and* $\hat{\Phi}_1$*.*

Using (2.103) we can find that

$$\begin{aligned}\hat{\Sigma}_u &= \frac{1}{N-Kp-1}(Y-\hat{B}Z)(Y-\hat{B}Z)^T \\ &\approx \begin{bmatrix}2.422 & -0.012 & 0.003\\-0.012 & 1.073 & 0.511\\0.003 & 0.511 & 0.765\end{bmatrix}.\end{aligned} \tag{2.119}$$

Also using Theorem 2.3, we find

$$\hat{\Phi}_0 := \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$$

and

$$\begin{aligned}\hat{\Phi}_1 &:= \hat{\Phi}_0\hat{A}_1 \\ &\approx \begin{bmatrix}0.451 & 0.007 & 0.007\\0.127 & 0.067 & 0.357\\-0.003 & 0.198 & 0.304\end{bmatrix}.\end{aligned}$$

Now filling in (2.32) results into

$$\begin{aligned}\hat{\Sigma}_y(1) &:= \hat{\Phi}_0\hat{\Sigma}_u\hat{\Phi}_0^T \\ &\approx \begin{bmatrix}2.422 & -0.012 & 0.003\\-0.012 & 1.073 & 0.511\\0.003 & 0.511 & 0.765\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\hat{\Sigma}_y(2) &:= \hat{\Sigma}_y(1) + \hat{\Phi}_1\hat{\Sigma}_u\hat{\Phi}_1^T \\ &\approx \begin{bmatrix}2.914 & 0.131 & 0.005\\0.131 & 1.239 & 0.654\\0.005 & 0.654 & 0.939\end{bmatrix}.\end{aligned}$$

Secondly, to find the covariance matrix as in (2.118), we will need to obtain $\hat{\Omega}(1)$ *and* $\hat{\Omega}(2)$*. To get these matrices, we see from Proposition 2.6 that we will need to find* $\hat{\boldsymbol{B}}$ *and* $\hat{\Gamma}$ *first.*

Using the least squares estimator $\hat{B}$ from (2.108), we find

$$\begin{aligned}\hat{\boldsymbol{B}} &:= \begin{bmatrix}1 & 0\\ \hat{\nu} & \hat{A}_1\end{bmatrix}\\ &\approx \begin{bmatrix}1 & 0 & 0 & 0\\ 1.049 & 0.451 & 0.007 & 0.007\\ 1.766 & 0.127 & 0.067 & 0.357\\ 3.013 & -0.003 & 0.198 & 0.304\end{bmatrix}.\end{aligned}$$

Now using (2.102), we find

$$\begin{aligned}\hat{\Gamma} &:= \frac{ZZ^T}{N}\\ &\approx \begin{bmatrix}1.000 & 2.030 & 4.285 & 5.532\\ 2.030 & 7.147 & 8.889 & 11.269\\ 4.285 & 8.889 & 19.654 & 24.419\\ 5.532 & 11.269 & 24.419 & 31.612\end{bmatrix}.\end{aligned}$$

Now using the matrices $\hat{\boldsymbol{B}}$ and $\hat{\Gamma}$ we found together with $\hat{\Sigma}_u$, $\hat{\Phi}_0$ and $\hat{\Phi}_1$, we find

$$\begin{aligned}\hat{\Omega}(1) &:= \operatorname{tr}\left[(\hat{\boldsymbol{B}}^T)^{1-1}\hat{\Gamma}^{-1}\hat{\boldsymbol{B}}^{1-1}\hat{\Gamma}\right]\hat{\Phi}_0\hat{\Sigma}_u\hat{\Phi}_0^T\\ &= \operatorname{tr}(I_{3*1+1})\hat{\Sigma}_u\\ &= 4\hat{\Sigma}_u\\ &\approx \begin{bmatrix}9.686 & -0.047 & 0.013\\ -0.047 & 4.292 & 2.044\\ 0.013 & 2.044 & 3.061\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\hat{\Omega}(2) &:= \sum_{i=0}^{1}\sum_{j=0}^{1}\operatorname{tr}\left[(\hat{\boldsymbol{B}}^T)^{1-i}\hat{\Gamma}^{-1}\hat{\boldsymbol{B}}^{1-j}\hat{\Gamma}\right]\hat{\Phi}_i\hat{\Sigma}_u\hat{\Phi}_j^T\\ &= \begin{bmatrix}9.550 & 1.125 & 0.016\\ 1.125 & 3.180 & 2.560\\ 0.016 & 2.560 & 3.048\end{bmatrix}.\end{aligned}$$

Now finally filling in (2.118) with $\hat{\Sigma}_y(0)$, $\hat{\Sigma}_y(1)$, $\hat{\Omega}(1)$ and $\hat{\Omega}(2)$ results in

$$\begin{aligned}\hat{\Sigma}_{\hat{y}}(1) &:= \hat{\Sigma}_y(1) + \frac{1}{N}\hat{\Omega}(1)\\ &\approx \begin{bmatrix}2.422 & -0.012 & 0.003\\ -0.012 & 1.073 & 0.511\\ 0.003 & 0.511 & 0.765\end{bmatrix} + \frac{1}{1000}\begin{bmatrix}9.686 & -0.047 & 0.013\\ -0.047 & 4.292 & 2.044\\ 0.013 & 2.044 & 3.061\end{bmatrix}\\ &\approx \begin{bmatrix}2.431 & -0.012 & 0.003\\ -0.012 & 1.077 & 0.513\\ 0.003 & 0.513 & 0.768\end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\hat{\Sigma}_{\hat{y}}(2) &:= \hat{\Sigma}_y(2) + \frac{1}{N}\hat{\Omega}(2)\\ &\approx \begin{bmatrix}2.914 & 0.131 & 0.005\\ 0.131 & 1.239 & 0.654\\ 0.005 & 0.654 & 0.939\end{bmatrix} + \frac{1}{1000}\begin{bmatrix}9.550 & 1.125 & 0.016\\ 1.125 & 3.180 & 2.560\\ 0.016 & 2.560 & 3.048\end{bmatrix}\\ &\approx \begin{bmatrix}2.924 & 0.132 & 0.005\\ 0.132 & 1.242 & 0.656\\ 0.005 & 0.656 & 0.942\end{bmatrix}.\end{aligned}$$

Now applying (2.117) gives us the following 95% forecast intervals. For the first variable of interest we find the following.

steps ahead	**forecast**	**lower bound**	**upper bound**	**interval length**
1	3.035	-0.021	6.091	6.112
2	2.475	-0.876	5.826	6.703

Table 4: The minimum MSE predictions for 1 and 2 steps of the first variable of interest and their 95% forecast intervals.

For the second variable of interest we find the following.

steps ahead	**forecast**	**lower bound**	**upper bound**	**interval length**
1	3.755	1.720	5.789	4.069
2	3.978	1.794	6.162	4.369

Table 5: The minimum MSE predictions for 1 and 2 steps of the second variable of interest and their 95% forecast intervals.

For the third variable of interest we find the following

steps ahead	**forecast**	**lower bound**	**upper bound**	**interval length**
1	4.414	2.380	6.132	3.752
2	5.088	2.903	6.990	4.087

Table 6: The minimum MSE predictions for 1 and 2 steps of the second variable of interest and their 95% forecast intervals.

Of course we can use the same methodology to find forecasts and their intervals for $h > 2$. If we would predict up to 10 steps ahead, we get the following figure.

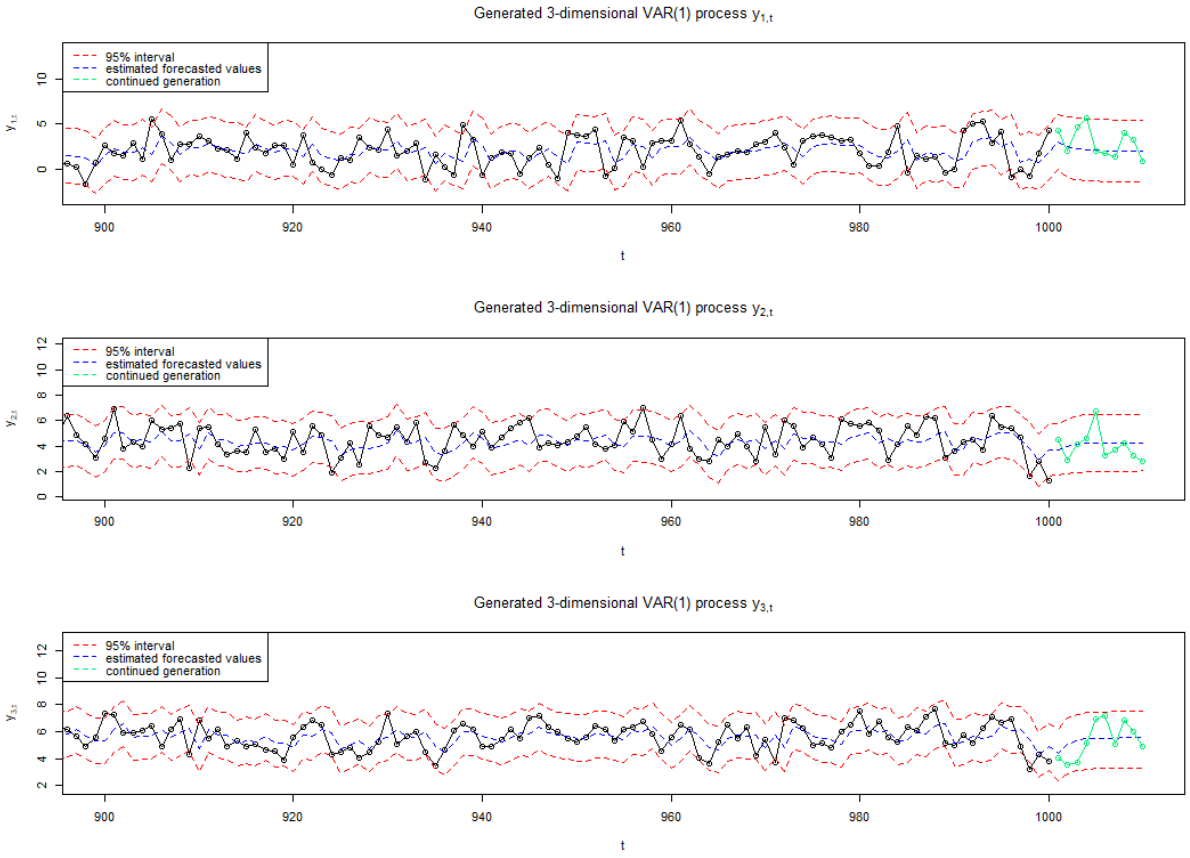


Figure 2.7: Estimated prediction of the estimated process 10 steps ahead for all 3 variables of interest.

We see that the generated process is following the forecasted values pretty well and it is lying between the 95% interval most of the times. The continued generation also looks to get nicely predicted.

2.5 Model tests

In this section we will describe the tests we can perform on a VAR process. First we will look at ways to test for existence of Granger-causality and Instantaneous causality between certain variables of interest. We previously found that causality occurs when certain values of the coefficient matrices or of the covariance matrix are non-zero. However when we estimate the coefficient matrices and the covariance matrix, we will not be able to find the original values of the matrices. Some values in those matrices might originally be 0, but the estimated values might not be 0, hence one can not determine causality between variables of interest when the process is estimated. Therefore tests should be used to test if some values of a matrix are significantly 0.

Secondly we will look at tests for autocorrelation and non-normality of the residuals. Performing these residual tests is also called *diagnostic checking*. Diagnostic checking is important when performing time series analysis, since we often assumed that the residuals are a white noise process, hence uncorrelated with each other. We also sometimes made the assumption that the residuals are normally distributed. Therefore it is important that these assumption are checked before performing any sort of analysis method.

2.5.1 Test for Granger-causality

From Theorem 2.9 we see that if we would want to know whether Granger-causality occurs or when it does not occur, we will have to test whether certain elements of the coefficient matrices are 0 or not. In

general we will be using the *Wald test* (Wald, 1939), which tests the null hypothesis

$$H_0 : C\boldsymbol{\beta} = c \tag{2.120}$$

against the alternative hypothesis

$$H_1 : C\boldsymbol{\beta} \neq c, \tag{2.121}$$

where C is a $(M \times (K(Kp+1))$ matrix of rank M and c is a $(M \times 1)$ vector. We choose matrix C such that we can test whether the elements in the coefficient matrices we want to look at are equal to c. Thus taking the right matrix C and taking $c = 0$ allows us to test no existence of Granger-causality against existence Granger-causality between certain variables.

In order to figure out what test statistic we will use to test H_0 against H_1, let us first look at the following lemma (Lütkepohl, 2005, pp. 692-693).

Lemma 2.8. *Let $\hat{\beta}$ be an estimator of the $(K \times 1)$ vector β with $\sqrt{N}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \Sigma)$. If C is a $(M \times K(Kp+1))$ matrix with $C \neq 0$, then the following holds.*

1. $\sqrt{N}(C\hat{\beta} - C\beta) \sim \mathcal{N}(0, C\Sigma C^T)$.
2. $N(\hat{\beta} - \beta)\Sigma^{-1}(\hat{\beta} - \beta) \sim \chi^2(K)$.

Now using this lemma, we can find the following test statistic, which is called the *Wald statistic*.

Theorem 2.16. *Let the null hypothesis in (2.120) be true and let $\hat{\boldsymbol{\beta}}$ be the OLS estimator of $\boldsymbol{\beta}$, then*

$$N(C\hat{\boldsymbol{\beta}} - c)^T \left[C(\Gamma^{-1} \otimes \Sigma_u)C^T\right]^{-1} (C\hat{\boldsymbol{\beta}} - c) \sim \chi^2(M).$$

Proof. Using the asymptotic normality of the OLS estimator as in Proposition 2.3, we have that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_u).$$

Using the first statement from Lemma 2.8 we see that

$$\sqrt{N}(C\hat{\boldsymbol{\beta}} - C\boldsymbol{\beta}) \sim \mathcal{N}(0, C(\Gamma^{-1} \otimes \Sigma_u)C^T).$$

Now since $C\hat{\boldsymbol{\beta}}$ and $C\boldsymbol{\beta}$ are $(M \times 1)$ vectors, we have using the second statement from Lemma 2.8 that

$$N(C\hat{\boldsymbol{\beta}} - C\boldsymbol{\beta}) \left[C(\Gamma^{-1} \otimes \Sigma_u)C^T\right]^{-1} (C\hat{\boldsymbol{\beta}} - c) \sim \chi^2(M).$$

□

In general Γ and Σ_u are not known in advance, hence we will need to use the estimators $\hat{\Gamma}$ and $\hat{\Sigma}_u$ from (2.102) and (2.103) respectively. Since these parameters will be estimated, Lütkepohl (2005, pp. 103) suggested it might be better to test with an adjusted distribution, which is derived from the following lemma.

Lemma 2.9. *Let d_1 and d_2 be certain degrees of freedom, then*

$$d_1 F(d_1, d_2) \sim \chi^2(d_1), \text{ when } d_2 \to \infty,$$

where $F(d_1, d_2)$ is an F random variable with d_1 and d_2 degrees of freedom.

We will be using the F-distribution in our test statistic, since this distribution has fatter tails, hence more room for errors in the estimation of Γ and Σ_u. Let us now define

$$\lambda_W := (C\hat{\boldsymbol{\beta}} - c)^T \left[C((ZZ^T)^{-1} \otimes \hat{\Sigma}_u)C^T\right]^{-1} (C\hat{\boldsymbol{\beta}} - c), \tag{2.122}$$

which is the Wald statistic, but with the estimated parameters $\hat{\Gamma}$ and $\hat{\Sigma}_u$. We know from Theorem 2.16 that also $\lambda_W \sim \chi^2(M)$, hence using Lemma 2.9 we also know that $\lambda_W \sim MF(M, d_2)$, for d_2 degrees of freedom. Most of the times d_2 will be taken to be the sample size minus the amount of unknown parameters. Lütkepohl (2005, p. 104) argued that $d_2 = N - (Kp + 1)$ suffices as well, since each individual restriction of the Wald test is applied on only one variable of interest. We can now come up with the following adjusted test statistic.

Proposition 2.7. *Let the null hypothesis in (2.120) be true and let $\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$, then*

$$\lambda_F := \frac{\lambda_W}{M} \sim F(M, N - (Kp + 1)).$$

Example 2.14. *Let us continue with the generated 3-dimensional VAR(1) process we used in Example 2.11. We found the least squares estimator $\hat{B}$ in (2.108). In Example 2.6 we have seen that in the original process the second and third variable of interest did not Granger-cause the first variable of interest. In this example we will be testing whether our generated process also shows no significant existence Granger-causality between these variables.*

We want to test the null hypothesis

$$H_0 : C\boldsymbol{\beta} = 0$$

against the alternative hypothesis

$$H_1 : C\boldsymbol{\beta} \neq 0,$$

where

$$C = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

is the (2×12) matrix. The rank of this matrix is obviously 2, hence $M = 2$. With this specific matrix we have that the hypotheses H_0 and H_1 are equivalent with testing whether the second and third element in the first row of the coefficient matrix are 0 or are not 0 respectively. Thus indeed Granger-causality is tested here. Together with $\hat{\Sigma}_u$ in (2.119) we find using (2.122) that

$$\begin{aligned} \lambda_W &= (C\hat{\boldsymbol{\beta}})^T \left[C((ZZ^T)^{-1} \otimes \hat{\Sigma}_u) C^T \right]^{-1} (C\hat{\boldsymbol{\beta}}) \\ &\approx 0.076, \end{aligned}$$

hence

$$\begin{aligned} \lambda_F &:= \frac{\lambda_W}{2} \\ &\approx 0.038. \end{aligned}$$

Now using Proposition 2.7, we see that λ_F is distributed as $F(2, 996)$. Now we find a p-value of approximately 0.963, hence we do not reject the null hypothesis if we take a significance level of 0.05 and we conclude that the second and third variable of interest do not Granger-cause the first variable of interest.

2.5.2 Test for instantaneous causality

To test for no instantaneous causality between certain variables, we will be testing when certain values of Σ_u are 0 or not. In general we will again be using the Wald test to test the null hypothesis

$$H_0 : C\boldsymbol{\sigma} = 0 \tag{2.123}$$

against the alternative hypothesis

$$H_1 : C\boldsymbol{\sigma} \neq 0, \tag{2.124}$$

where C is the $(M \times \frac{K(K+1)}{2})$ matrix of rank M, $\boldsymbol{\sigma}$ is again defined as $\text{vech}(\Sigma_u)$ and c is a $(M \times 1)$ vector. We again choose matrix C such that we can test whether certain elements on and below the diagonal of Σ_u are equal to c. When we take the right matrix C and we take $c = 0$, then we are testing no existence of instantaneous causality against the existence of instantaneous causality between certain variables.

Since we have from Proposition 2.5 the asymptotic property

$$\sqrt{N}(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \sim \mathcal{N}(0, \Sigma_{\tilde{\boldsymbol{\sigma}}}),$$

we can now obtain the following Wald statistic the same way we obtained Theorem 2.16.

Theorem 2.17. *Let the null hypothesis in (2.125) be true, then*

$$\lambda_W := N(C\tilde{\boldsymbol{\sigma}})^T \left[2C\boldsymbol{D}_K^+(\tilde{\Sigma}_u \otimes \tilde{\Sigma}_u)(C\boldsymbol{D}_K^+)^T\right]^{-1} C\tilde{\boldsymbol{\sigma}} \sim \chi^2(M).$$

Example 2.15. *Let us continue with the generated 3-dimensional VAR(1) process as in Example 2.11. In Example 2.7 we have seen that there is no instantaneous causality between the first variable of interest and the second variable combined with the third variable of interest. In this example we will be testing whether our generated process shows the same result.*

We want to test the null hypothesis

$$H_0 : C\boldsymbol{\sigma} = 0 \tag{2.125}$$

against the alternative hypothesis

$$H_1 : C\boldsymbol{\sigma} \neq 0, \tag{2.126}$$

where

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The rank of this matrix is 2, hence $M = 2$. We choose this specific matrix C such that the hypotheses H_0 and H_1 are equivalent with testing whether the second and third element of the first row and first column of Σ_u are 0 or not respectively. From Theorem 2.10 we have seen that this is equivalent with testing whether there is no instantaneous causality against testing whether there is instantaneous causality between the first variable and the second and third variable combined.

From Definition 2.19 we can find the following duplication matrix $\boldsymbol{D}_3$.

$$\boldsymbol{D}_3 := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

hence

$$\begin{aligned} \boldsymbol{D}_3^+ &:= (\boldsymbol{D}_K^T\boldsymbol{D}_K)^{-1}\boldsymbol{D}_K^T \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

We also have from Theorem 2.14 that

$$\tilde{\Sigma}_u = \frac{1}{N}(\tilde{Y}^0 - \tilde{A}\tilde{X})(\tilde{Y}^0 - \tilde{A}\tilde{X})^T.$$

Now since we know that $\tilde{\mu}$ and $\tilde{A}$ is the same as $\hat{\mu}$ and $\hat{A}$ from the least squares estimation respectively, we can find these values from the least squares estimator we found in (2.108). This way $\tilde{\Sigma}_u$ can be calculated and we find

$$\tilde{\Sigma}_u \approx \begin{bmatrix} 2.412 & -0.012 & 0.003 \\ -0.012 & 1.069 & 0.509 \\ 0.003 & 0.509 & 0.762 \end{bmatrix}.$$

Now using Theorem 2.17, we find

$$\begin{aligned}\lambda_W &= N(C\tilde{\boldsymbol{\sigma}})^T \left[2C\boldsymbol{D}_3^+(\tilde{\Sigma}_u \otimes \tilde{\Sigma}_u)(C\boldsymbol{D}_3^+)^T\right]^{-1} C\tilde{\boldsymbol{\sigma}} \\ &\approx 0.114.\end{aligned}$$

Since $\lambda_W \sim \chi(2)$, we find a p-value of approximately 0.944, which means we definitely do not reject H_0 if we take a significance level of 0.05, hence we have instantaneous causality between the first variable and the second combined with the third variable.

2.5.3 Test for residual autocorrelations

Testing for autocorrelations in the residuals is important when performing time series analysis with the VAR model. We assumed that a VAR model has white noise residuals, hence we should check for autocorrelation between the residuals before we perform analysis methods. If there exists autocorrelation between the residuals, then one might consider using another model.

Let us use the following notations.

Definition 2.21. *Let us define the estimator C_i of the autocovariance of the residuals at lag i to be*

$$C_i := \frac{1}{N}\sum_{t=i+1}^{N} u_t u_{t-i}^T$$

and define the estimator R_i of the autocorrelation of the residuals at lag i to be

$$R_i := D^{-1}C_i D^{-1},$$

where D is a diagonal $(K \times K)$ matrix with the same diagonal elements as C_0. Then we define $\boldsymbol{C}_h$ and $\boldsymbol{R}_h$ to be the following $(K \times Kh)$ matrices.

$$\begin{aligned}\boldsymbol{C}_h &:= (C_1, C_2, \ldots, C_h), \\ \boldsymbol{R}_h &:= (R_1, R_2, \ldots, R_h).\end{aligned}$$

In order to test for no autocorrelation of the residuals up to lag h we can use the so-called *portmanteau test* Castle and Hendry, 2010, which states to test the null hypothesis

$$H_0 : \mathbf{R}_h = 0 \tag{2.127}$$

against the alternative hypothesis

$$H_1 : \mathbf{R}_h \neq 0. \tag{2.128}$$

The test statistic can again be found with the use of an asymptotic property, this time of $\text{vec}(\hat{\mathbf{C}}_h)$. Note that $\hat{\mathbf{C}}_h$ can similarly be found by using $\hat{u}_t$. It turns out that $\sqrt{N}\,\text{vec}(\hat{\mathbf{C}}_h)$ is asymptotically normally distributed with mean 0 (Lütkepohl, 2005, pp. 165-166). The following test statistic can then be found (Lütkepohl, 2005, p. 169).

Theorem 2.18. *Let the null hypothesis in (2.127) be true, then*

$$Q_h := N\sum_{t=1}^{h} \text{tr}(\hat{C}_i^T \hat{C}_0^{-1} \hat{C}_i \hat{C}_0^{-1}) \sim \chi^2(K^2(h-p)).$$

Example 2.16. *Let us again continue with the 3-dimensional VAR(1) process from Example 2.11. We now do not assume the residuals u_t to be i.i.d. normally distributed, but we will modify u_t such that autocorrelation can occur. We will define u_t to be*

$$u_t = \rho u_{t-1} + \epsilon_t \quad \textit{for } \rho \in [-1, 1],$$

where ϵ_t is i.i.d. with $\mathcal{N}(0, \Sigma_u)$ and $u_0 = 0$. The idea is that autocorrelation between the residuals clearly occur when ρ is close to 1 or -1, but when ρ is close to 0 it might be harder for the portmanteau test to observe autocorrelation. If we generate our process 100 times with $\rho \in [-1, 1]$, calculate the corresponding p-value 100 times using Theorem 2.18 and calculate the percentage of the amount of rejected null hypotheses, we find the following figure.

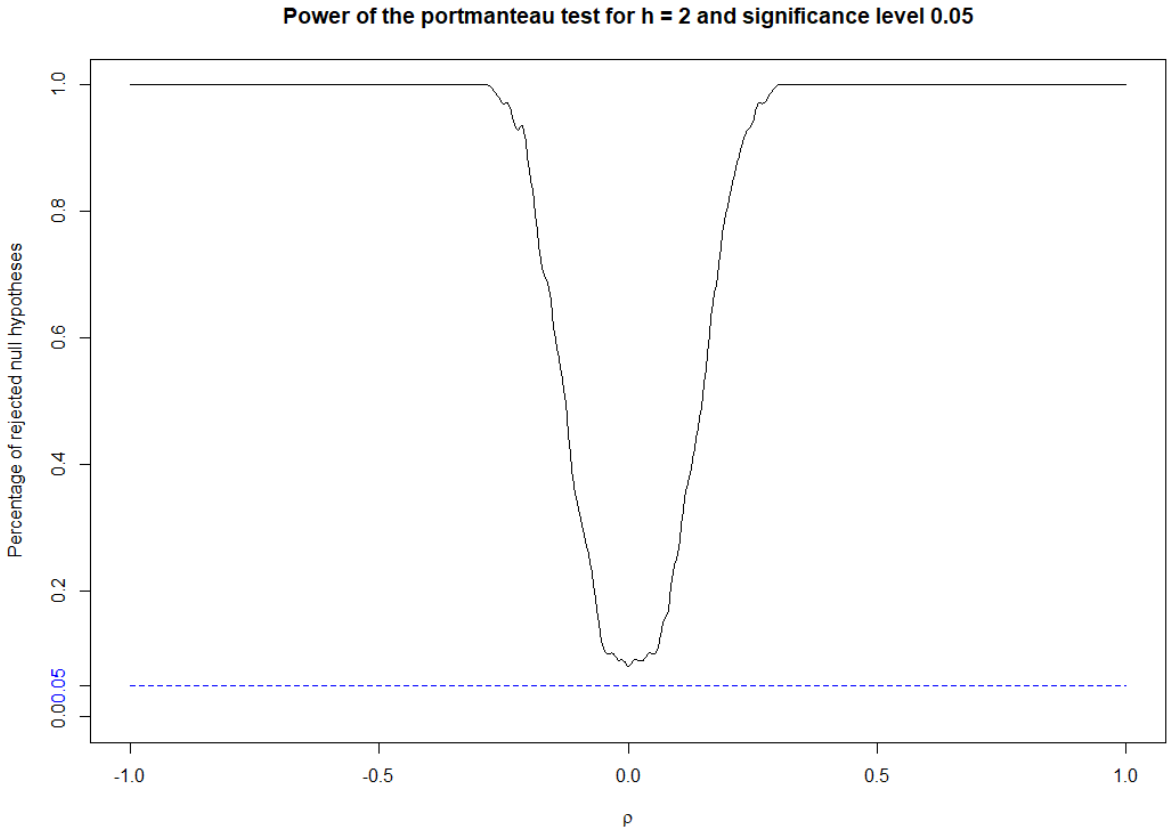


Figure 2.8: Percentage of the amount of rejected null hypotheses of the portmanteau test for $h = 2$, a significance level of 0.05 and for $\rho \in [-1, 1]$.

As expected, we see that the portmanteau test rejects the null hypothesis more often for rho closer to 0. The blue line in Figure 2.8 represents a 5% rejection rate of the null hypothesis. We see however that the rejection rate is always above 5%, even for $\rho = 0$, where we would actually expect a 0% rejection rate. A reason for this is that $N = 1000$ is probably not large enough for this test.

2.5.4 Test for non-normality of the error terms

Testing for non-normality of the residuals is an important test as well, since for some methods we made the assumption that the residuals are normally distributed, e.g. forecast intervals and the maximum likelihood estimation.

Let us assume we have a stable K-dimensional VAR(p) process with normally distributed white noise error terms, hence $u_t \sim \mathcal{N}(0, \Sigma_u)$ for some $(K \times K)$ matrix Σ_u. We can rewrite the multivariate normal distribution using the following lemma (Lütkepohl, 2005, pp. 174-175).

Lemma 2.10. *Let X be a K-dimensional multivariate normally distributed random variable with $X \sim \mathcal{N}(\mu, \Sigma)$, then we have that*

$$P^{-1}(X - \mu) \sim \mathcal{N}(0, I_K),$$

where P is the Cholesky decomposition of Σ from Theorem 2.7.

Now using Lemma 2.10 we find that

$$w_t = \begin{bmatrix} w_{1t} \\ w_{2t} \\ \vdots \\ w_{Kt} \end{bmatrix} := P^{-1}u_t \sim \mathcal{N}(0, I_K),$$

however when estimating results, our error terms are estimated, hence then we use

$$\hat{w}_t = \begin{bmatrix} \hat{w}_{1t} \\ \hat{w}_{2t} \\ \vdots \\ \hat{w}_{Kt} \end{bmatrix} := \hat{P}^{-1}\hat{u}_t \sim \mathcal{N}(0, I_K), \tag{2.129}$$

where $\hat{P}$ is the Cholesky decomposition of the consistent estimator $\hat{\Sigma}_u = \frac{\hat{U}\hat{U}^T}{N}$ and hence $\hat{P}$ is consistent, so (2.129) holds. Now using this result we want to find a test statistic for the normality of the residuals. The idea is to use the fact that skewness (third moment) and the kurthosis (fourth moment) of an univariate normal distribution are 0 and 3 respectively. Based on this idea we will test the following null hypothesis

$$H_0 : \mathbb{E}\begin{bmatrix} w_{1t}^3 \\ w_{2t}^3 \\ \vdots \\ w_{Kt}^3 \end{bmatrix} = 0 \text{ and } \mathbb{E}\begin{bmatrix} w_{1t}^4 \\ w_{2t}^4 \\ \vdots \\ w_{Kt}^4 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ \vdots \\ 3 \end{bmatrix} =: \mathbf{3}_K \tag{2.130}$$

against the alternative hypothesis

$$H_1 : \mathbb{E}\begin{bmatrix} w_{1t}^3 \\ w_{2t}^3 \\ \vdots \\ w_{Kt}^3 \end{bmatrix} \neq 0 \text{ and } \mathbb{E}\begin{bmatrix} w_{1t}^4 \\ w_{2t}^4 \\ \vdots \\ w_{Kt}^4 \end{bmatrix} \neq \begin{bmatrix} 3 \\ 3 \\ \vdots \\ 3 \end{bmatrix} =: \mathbf{3}_K. \tag{2.131}$$

to test normality against non-normality respectively. The well-known univariate case of this test called the *Jarque-Bera test* (Jarque and Bera, 1987). To determine the test statistic, we will be using the following lemma (Lütkepohl, 2005, pp. 175-177).

Lemma 2.11. *Define b_1 and b_2 to be*

$$b_1 = \begin{bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{K1} \end{bmatrix}, \quad \textit{with } b_{i1} := \frac{1}{T}\sum_{t=1}^{N} w_{it}^3, \quad \textit{for } i = 1, \ldots, K,$$

$$b_2 = \begin{bmatrix} b_{12} \\ b_{22} \\ \vdots \\ b_{K2} \end{bmatrix}, \quad \textit{with } b_{i2} := \frac{1}{T}\sum_{t=1}^{N} w_{it}^4, \quad \textit{for } i = 1, \ldots, K,$$

then

$$\sqrt{N}\begin{bmatrix} b_1 \\ b_2 - \mathbf{3}_K \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} 6I_K & 0 \\ 0 & 24I_K \end{bmatrix}).$$

Note that the same results with $\hat{w}_t$ can be formed, but then b_1 and b_2 will be called $\hat{b}_1$ and $\hat{b}_2$ respectively.

Now using this lemma we can easily find the test statistic as in the following theorem.

Theorem 2.19. *Let b_1 and b_2 be defined as in Lemma 2.11. If H_0 as in (2.130) is true, then*

$$\lambda_{sk} := N(\frac{b_1^T b_1}{6} + \frac{b_2^T b_2}{24}) \sim \chi(2K).$$

Proof. From Lemma 2.11 we see that the following holds.

$$\begin{aligned} \sqrt{N}\frac{b_1}{\sqrt{6}} &\sim \mathcal{N}(0, I_K), \\ \sqrt{N}\frac{(b_2 - \mathbf{3}_K)}{\sqrt{24}} &\sim \mathcal{N}(0, I_K). \end{aligned}$$

Hence we can find the test statistics

$$\begin{aligned} \lambda_s &:= N\frac{b_1^T b_1}{6} \sim \chi(K), \\ \lambda_k &:= N\frac{b_2^T b_2}{24} \sim \chi(K), \end{aligned}$$

which we will need to combine in order to test H_0 against H_1. Hence adding these test statistics results in the test statistic

$$\lambda_{sk} := \lambda_s + \lambda_k \sim \chi(2K).$$

□

Note that again this theorem holds for $\hat{b}_1$ and $\hat{b}_2$.

Example 2.17. *Let us continue with the generated 3-dimensional VAR(1) process as in Example 2.11. But now we will modify u_t again in such a way that normally distributed residuals will occur. We will take 100 samples where we assume u_t to be i.i.d. t-distributed for some degrees of freedom 1 till 50 and perform a non-normality test on each sample. Since the t-distribution will converge to the standard normal distribution when the degrees of freedom go to infinity, we will expect more rejections of the null hypothesis for t-distributed residuals with low degrees of freedom and less rejections of the null hypothesis for t-distributed residuals with high degrees of freedom. Now using Theorem 2.19 we can calculate the p-values of all 100 samples for all degrees of freedom and find the percentage of the amount of rejected null hypotheses. The following figure can then be found.*

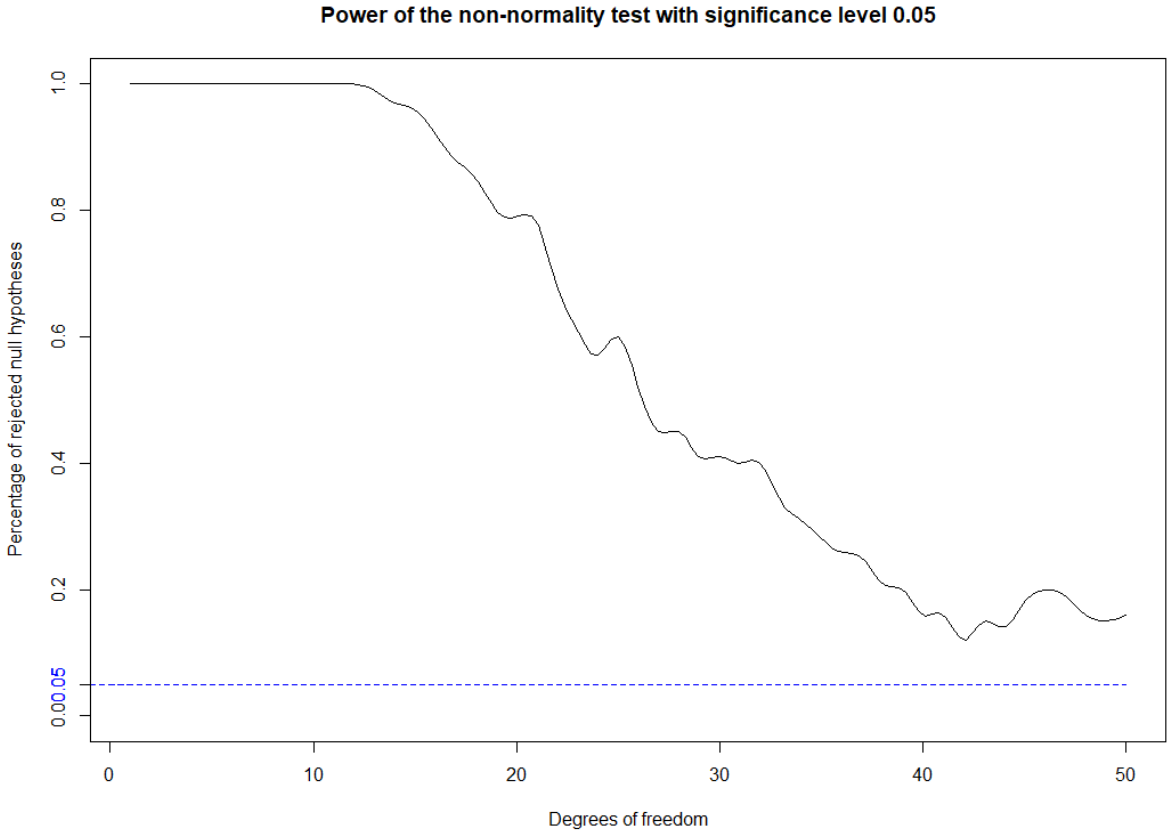


Figure 2.9: Rejection rate of the null hypothesis with a significance level of 0.05 for various degrees of freedom.

In Figure 2.9 we indeed find lower rejection rates for t-distributed residuals with higher degrees of freedom, which is what we expected. The blue line again represents a 5% rejection rate of the null hypothesis. Whenever we take infinite degrees of freedom, or normally distributed residuals, the rejection rate turns out to be 4.6% for 1000 normally distributed residuals. However, we would expect the rejection rate to go to 0% for normally distributed residuals. This results probably happens since $N = 1000$ is not large enough to show a 0% rejection rate. However, it is large enough if we agree that the rejection rate of normally distributed residuals should be below 5%.

2.6 Order selection

In all of the previous sections we have assumed that the VAR order p is known, however in reality we do not know p beforehand, thus also the VAR order will have to be estimated. In this section we will be investigating four different criteria that we can use on a process to estimate its VAR order $\hat{p}$.

2.6.1 FPE criterion

The *Final Prediction Error* (FPE) criterion is based on choosing an order m, such that the precision of the forecast is optimal. It is suggested in (Akaike, 1969, 1971) to base this criterion on the covariance matrix of the 1-step forecast error of $\hat{y}_t(1)$, which is $\Sigma_{\hat{y}}(1)$ as in Theorem 2.15. It is also easy to see that

$$\Sigma_{\hat{y}}(1) = \frac{N + Km + 1}{N}\Sigma_u.$$

To turn the covariance matrix into an order selection criterion, we simply look at $\Sigma_u(m)$, which is the covariance matrix of the error terms when we estimate a VAR(m) model. Again, since we are estimating

the process, we will be using the estimator

$$\hat{\Sigma}_u(m) \quad = \quad \frac{N}{N-Km-1}\tilde{\Sigma}_u(m),$$

where $\tilde{\Sigma}_u$ is the maximum likelihood estimator as in Theorem 2.14. In order to find the order such that the 1-step forecast is optimal, we look at the determinant of $\Sigma_{\hat{y}}(1)$ and find the order that minimizes this value. This way the FPE criterion is formed in the following proposition.

Proposition 2.8. *Define $FPE(m)$ to be*

$$\begin{aligned} FPE(m) \quad &:= \quad \det\left(\frac{N+Km+1}{N-Km-1}\frac{N}{N-Km-1}\tilde{\Sigma}_u(m)\right) \\ &= \quad \left(\frac{N+Km+1}{N-Km-1}\right)^K \det(\tilde{\Sigma}_u(m)), \end{aligned}$$

then we call the estimate of the VAR order based on the FPE criterion $\hat{p}(FPE)$ if

$$FPE(\hat{p}(FPE)) = \min\{FPE(m)|m = 0, 1, \ldots, M\}.$$

2.6.2 AIC, HQ criterion and SC

The Akaike's Information Criterion (AIC) (Akaike, 1973, 1974), the Hannan-Quinn (HQ) criterion (Hannan and Quinn, 1979; Quinn, 1980) and the Schwarz Criterion (SC) (Schwarz, 1978) are all based on optimising the precision of the forecast, but at the same time a penalty for the amount of unknown parameters will be taken into account. The following criterion will be used for these criteria.

$$Cr(m) := \ln(\det(\tilde{\Sigma}_u(m))) + m\frac{C_N}{N}, \tag{2.132}$$

where C_N is different for the AIC, HQ criterion and SC. We see that $\ln(\det(\tilde{\Sigma}_u(m)))$ corresponds to the precision of the forecast and $m\frac{C_N}{N}$ corresponds to the penalty for the amount of unknown parameters. In the following proposition we find the different criteria.

Proposition 2.9. *The estimate of the VAR order based on the AIC, HQ criterion and SC are $\hat{p}(AIC)$, $\hat{p}(HQ)$ and $\hat{p}(SC)$, where*

$$\begin{aligned} Cr(\hat{p}(AIC)) \quad &= \quad \min\{AIC(m)|m = 0, 1, \ldots, M\} \\ &= \quad \min\{Cr(m)|C_N = 2K^2 \wedge m = 0, 1, \ldots, M\}, \\ Cr(\hat{p}(HQ)) \quad &= \quad \min\{HQ(m)|m = 0, 1, \ldots, M\} \\ &= \quad \min\{Cr(m)|C_N = 2K^2\ln(\ln(N)) \wedge m = 0, 1, \ldots, M\}, \\ Cr(\hat{p}(SC)) \quad &= \quad \min\{SC(m)|m = 0, 1, \ldots, M\} \\ &= \quad \min\{Cr(m)|C_N = 2K^2\ln(N) \wedge m = 0, 1, \ldots, M\}. \end{aligned}$$

2.6.3 Consistency of the criteria

Now in total we have found 4 different criteria to estimate the VAR order. However it turns out that not all of these criteria are consistent, i.e. $\operatorname{plim}\hat{p} \neq p$ for $N \to \infty$, where p is the VAR order. It also turns out that some estimators are *strongly consistent*, which is defined as follows.

Definition 2.22. *The estimator $\hat{p}$ is called strongly consistent when*

$$\mathbb{P}(\lim_{N\to\infty}\hat{p} = p) = 1.$$

Using the following lemma (Hannan and Quinn, 1979; Quinn, 1980; Paulsen, 1984), we find that (strong) consistency of $\hat{p}$ is based on the convergence of functions with C_N, where $N \to \infty$.

Lemma 2.12. *The estimator $\hat{p}$ is consistent if and only if*

$$C_N \to \infty \quad \text{and} \quad \frac{C_N}{N} \to 0 \quad \text{when } N \to \infty \tag{2.133}$$

and the estimator $\hat{p}$ is strongly consistent if and only if (2.133) holds and

$$\frac{C_N}{2\ln(\ln(N))} > 1 \quad \text{when } N \to \infty. \tag{2.134}$$

Now using this lemma we can determine the consistency of all our criteria. We find the following theorems.

Theorem 2.20. *The following statements hold.*

1. *The estimators $\hat{p}(FPE)$ and $\hat{p}(AIC)$ are not consistent.*
2. *The estimator $\hat{p}(HQ)$ is consistent for $K = 1$ and strongly consistent for $K > 1$.*
3. *The estimator $\hat{p}(SC)$ is strongly consistent.*

Proof. (Statement 1) It can be shown that $\hat{p}(FPE) = \hat{p}(AIC)$ when $N \to \infty$, hence we only have to check $\hat{p}(AIC)$ with Lemma 2.12. It is obvious that C_N does not converge to ∞ with $N \to \infty$ since $C_N = 2K^2$, hence (2.133) does not hold, which means $\hat{p}(FPE)$ and $\hat{p}(AIC)$ are not consistent.

(Statement 2) For $C_N = 2K^2\ln(\ln(N))$ it is easy to see that (2.133) holds. We also see that $\frac{C_N}{2\ln(\ln(N))} = K^2$ when $N \to \infty$, which is bigger than 1 for $K > 1$. Using Lemma 2.12 shows us now that $\hat{p}(HQ)$ is consistent for $K = 1$ and strongly consistent for $K > 1$.

(Statement 3) For $C_N = 2K^2\ln(N)$ is it easy to see that both (2.133) and (2.134) hold, hence using Lemma 2.12 we see that $\hat{p}(SC)$ is strongly consistent. □

This theorem shows that the FPE criterion and the AIC will perform better when we have a small sample size and that the HQ criterion and the SC perform better when we have a big sample size. Sometimes it is hard to consider a certain sample size small or a big when performing analysis on a time series. In such situations we can apply all four criteria and simply take the highest resulting VAR order just to be sure.

Example 2.18. *If we again look at the generated 3-dimensional VAR(1) process of Example 2.11, we can apply all four criteria on the generated process and determine its VAR order, which we of course expect to be 1. We find the following values.*

VAR order	FPE	AIC	HQ	SC
1	**1.381**	**0.323**	**0.345**	**0.382**
2	1.391	0.330	0.370	0.434
3	1.400	0.337	0.393	0.485
4	1.414	0.347	0.420	0.540
5	1.428	0.357	0.447	0.594

Table 7: Various criteria values for VAR order 1 to 5.

Here the values of the criteria will keep increasing when we test for VAR orders higher than 5. Thus the bold values in Table 7 represent the minimum values of the criteria. All criteria suggest a VAR order of 1, as expected.

3 Vector Error Correction Model

In the previous sections we have always assumed that the time series we are analysing is stationary, however this does not always have to be the case. Many financial time series, e.g. stock prices, are non-stationary, hence the VAR model will fail on these financial time series. That is why the Vector Error Correction Model (VECM) is introduced. This model allows us to analyse some special cases of non-stationary time series, which are cointegrated processes. These type of time series are said to have variables of interest which are moving together or are driven by a common stochastic trend (Lütkepohl, 2005, p. 245). An example of a cointegrated bivariate time series can be seen in the figure below.

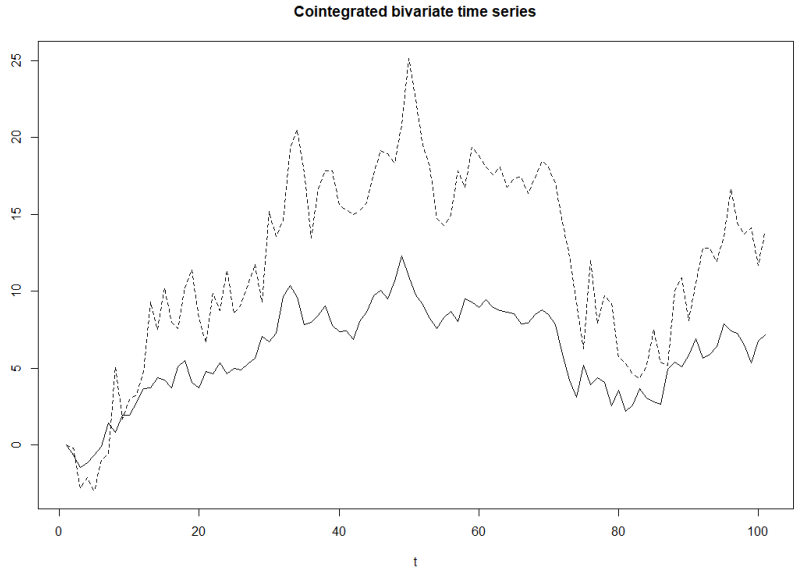


Figure 3.1: A cointegrated bivariate process.

We see that both variables of interest are indeed moving in a similar way. Let us first look at cointegrated processes in general.

3.1 Cointegrated processes

Let us first look at the definition of *differencing* a process.

Definition 3.1. *If we have a process y_t, which we want to difference d times, we obtain*

$$\begin{aligned} \Delta^d y_t &= \Delta^{d-1}(y_t - y_{t-1}) \\ &= \Delta^{d-1}(1-L)y_t \\ &= (1-L)^d y_t, \end{aligned}$$

where L is the lag operator operator from Definition 2.3.

The idea of the VECM is that we can somehow first make the time series stationary by differencing the process. Once it is stationary, we can apply the VAR model on the stationary process.

We now can define an *integrated process* as follows.

Definition 3.2. *A process y_t is called integrated of order d, or $y_t \sim I(d)$, when $\Delta^d y_t$ is stable, but $\Delta^{d-1} y_t$ is not stable.*

Since an integrated process of order d will be stable if we difference it d times, we will work with the stable process $\Delta^d y_t$ in order to form a VECM. But first we will need to introduce *equilibrium relationships*

between the variables of interest.

In financial data there are many relationships between different variables, for example between different stocks from the similar market sector. If such a relationship exists between different variables of interest of a process, then one could find a K dimensional vector $\boldsymbol{\beta}$ such that

$$\boldsymbol{\beta}^T y_t := \boldsymbol{\beta}_1 y_{1t} + \cdots + \boldsymbol{\beta}_K y_{Kt} = 0, \quad \text{with } \boldsymbol{\beta} \neq 0,$$

which we call a long-term equilibrium relationship. Obviously $\boldsymbol{\beta}$ will not have to be a unique vector, however up to $K-1$ independent linear combinations can be found in total, which we will discuss later.

In reality $\boldsymbol{\beta}^T y_t$ will not always be equal to 0 for all t, but it will be equal to a stochastic variable z_t, which represents the deviations of $\boldsymbol{\beta}^T y_t$. If the equilibrium relationship $\boldsymbol{\beta}^T y_t$ exists, then we may assume all variables of interests to move together, hence z_t is stable. With this knowledge we can now look at the definition of a *cointegrated process*.

Definition 3.3. *A K-dimensional process y_t is cointegrated of order (d,b), or $y_t \sim CI(d,b)$, if the following holds.*

1. *All variables of interest are integrated of order d.*
2. *There exist a linear combination $\boldsymbol{\beta}^T y_t = z_t$, where $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_K \end{bmatrix}$ and z_t is integrated of order $d-b$.*

3.1.1 The model

With the help of cointegrated processes we can derive the VECM. In this model we make an important assumption that all variables of interest of y_t are integrated of order 1. We will then be able to find a model for the first difference of y_t, or Δy_t.

Suppose we have a zero-mean K-dimensional, possible non-stationary, process y_t, with a single equilibrium relationship of $\boldsymbol{\beta}^T y_t$. In the VECM we will assume that Δy_t depends on 2 factors:

1. The linear combination $\boldsymbol{\beta}^T y_{t-1}$.
2. The previous $p-1$ differences $\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-p+1}$.

This means that for the k-th variable of interest we will have

$$\Delta y_{kt} = \boldsymbol{\alpha}_k \boldsymbol{\beta}^T y_{t-1} + \boldsymbol{\gamma}_{k,1}^T \Delta y_{t-1} + \cdots + \boldsymbol{\gamma}_{k,p-1}^T \Delta y_{t-p+1} + u_{kt}, \tag{3.1}$$

where $\boldsymbol{\gamma}_{k,1}^T, \ldots, \boldsymbol{\gamma}_{k,p-1}^T$ are $(K \times 1)$ vectors, $\boldsymbol{\alpha}_k$ is a constant and u_{kt} is a white noise process.

We assumed that there only existed 1 (independent) equilibrium relationship between the variables of interest, however up to $K-1$ equilibrium relationships might exist. Note that if K equilibrium relationships exist, then the equilibrium relation is 0 for all t instead of a stochastic variable z_t. This is quite unlikely when working with financial time series. Whenever r equilibrium relationships exist, we say that y_t is *cointegrated of rank r*. The variable $\boldsymbol{\alpha}_k$ will then turn into the $(1 \times r)$ vector and $\boldsymbol{\beta}$ will turn into the $(K \times r)$ matrix with columns as the equilibrium relationships.

We will now be able to write the VECM for Δy_t using (3.1) as

$$\Delta y_t = \boldsymbol{\Pi} y_{t-1} + \boldsymbol{\Gamma}_1 \Delta y_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t, \tag{3.2}$$

where $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_K \end{bmatrix} \boldsymbol{\beta}^T$ and $\boldsymbol{\Gamma}_i = \begin{bmatrix} \boldsymbol{\gamma}_1^T \\ \vdots \\ \boldsymbol{\gamma}_K^T \end{bmatrix}$ for $i = 1, 2, \ldots, p-1$ are $(K \times K)$ matrices and u_t is considered to be a K-dimensional white noise process.

Note that y_t being cointegrated of rank r will be equivalent with $\mathbf{\Pi}$ having rank r. Hence matrix $\mathbf{\Pi}$, or the *cointegration matrix*, will be 0 when y_t is cointegrated of rank 0, however when y_t is cointegrated of rank $r > 1$, then $\Pi \neq 0$.

Also note that for a cointegration rank $r > 1$ we can represent $\boldsymbol{\beta}$ into a normalised form as

$$\boldsymbol{\beta}^* := \begin{bmatrix} I_r \\ \boldsymbol{\beta}_{(K-r)} \end{bmatrix}, \tag{3.3}$$

where $\boldsymbol{\beta}_{(K-r)}$ is a $((K-r) \times r)$ matrix. This representation can simply be found by rearranging the variables.

Example 3.1. *Suppose* $\boldsymbol{\beta} = \begin{bmatrix} 2 & 0 \\ 2 & 1 \\ 0 & 2 \end{bmatrix}$*, then the normalised form in (3.3) is*

$$\boldsymbol{\beta}^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -2 & 2 \end{bmatrix}.$$

A nice result of the VECM is that we can rewrite the VECM representation (3.2) of Δy_t into a VAR(p) representation as

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p}, \tag{3.4}$$

where

$$\begin{aligned} A_1 &= I_K + \mathbf{\Pi} + \mathbf{\Gamma}_1, && (3.5) \\ A_i &= \mathbf{\Gamma}_i - \mathbf{\Gamma}_{i-1} \quad \text{for } i = 2, 3, \ldots, p-1, && (3.6) \\ A_p &= -\mathbf{\Gamma}_{p-1}. && (3.7) \end{aligned}$$

This VAR representation will be useful later on, e.g. forecasting the VECM.

Example 3.2. *Suppose we have the bivariate process* y_t *with an equilibrium relationship* $y_{1t} = y_{2t}$*, or equivalently* $\boldsymbol{\beta}^T y_t = 0$*, where* $\boldsymbol{\beta} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$*. Suppose we take the VECM representation of* Δy_t *to be*

$$\begin{aligned} \Delta y_t &= \boldsymbol{\alpha}\boldsymbol{\beta}^T y_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + u_t \\ &= \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Delta y_{t-1} + u_t \\ &= \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Delta y_{t-1} + u_t, \end{aligned} \tag{3.8}$$

where u_t *is i.i.d. bivariate normally distributed with covariance matrix*

$$\Sigma_u = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}.$$

Now using (3.4)-(3.7) we can rewrite our VECM in (3.8) as VAR(2) representation

$$y_t = \begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} y_{t-2} + u_t.$$

If we assume y_{-1} *and* y_0 *to be 0, then we can again generate the VAR process. Taking* $N = 100$ *results in the following generated process.*

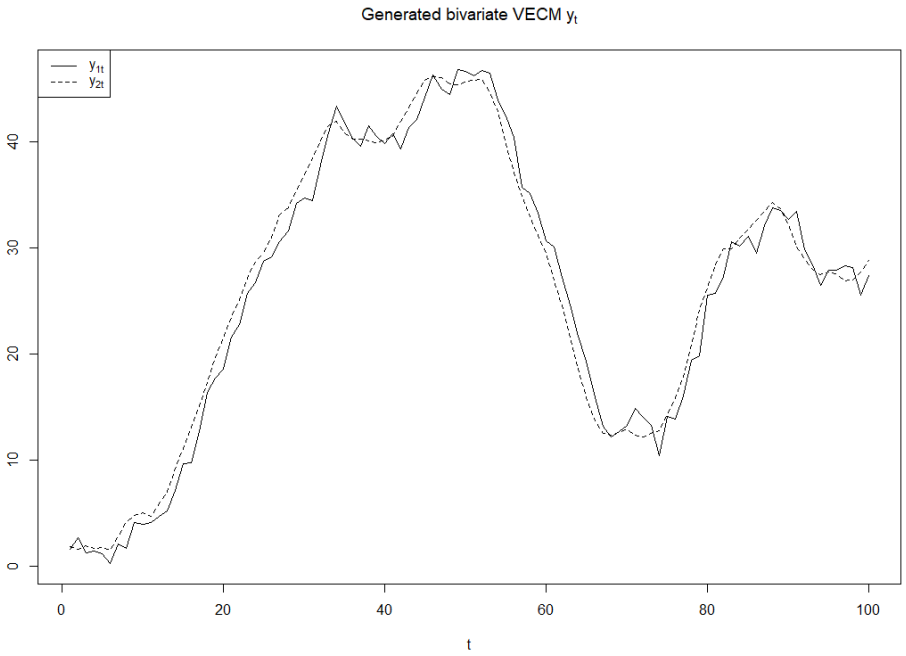


Figure 3.2: Generated VECM y_t.

We definitely see a non-stationary cointegrated process in Figure 3.2, hence the process is unstable. However, since the process is cointegrated, we can look at the equilibrium relationship $y_{1t} - y_{2t}$ and expect the result to be stable, hence stationary.

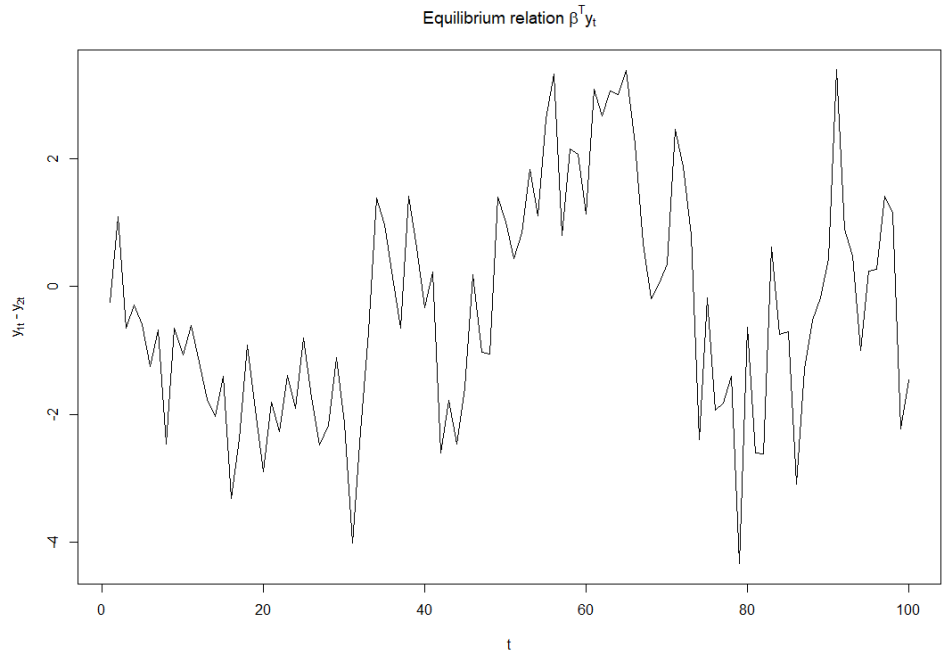


Figure 3.3: The equilibrium relationship of the generated VECM y_t.

We indeed find a stable process $y_{1t} - y_{2t}$ if we look at the Figure 3.3.

3.1.2 Non-zero mean VECM

So far we have assumed that we have VECM of a zero mean process y_t, however in reality we will mostly encounter processes that have a non-zero mean. The VECM in (3.2) will then have a different representation. We might actually consider a VECM with having a mean including linear trend. Then the VECM in (3.2) can be rewritten into the form

$$\Delta y_t = \nu_0 + \nu_1 t + \mathbf{\Pi} y_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t,$$

where ν_0 and ν_1 are some $(K \times 1)$ vectors. For simplicity of the calculations in this section we will assume that we have a zero mean, but one should be aware of the fact that a non-zero mean VECM could be considered as well.

3.1.3 Analysis methods

Just like we could forecast the VAR model and different analysis methods, we can also do the same for the VECM. Since we found in (3.4) - (3.7) that we can rewrite a VECM into a VAR model, most methods will remain the same when applied on the VAR representation. However a few differences might occur when the VAR process is unstable.

When forecasting an unstable VAR process, we will not necessarily have finite forecast intervals. We previously found that the (estimated) forecast error covariance matrix $\Sigma_y(h)$ contains (estimated) moving average coefficients Φ_i, where $\Phi_i \to 0$ when i goes to infinity. However when the process is unstable, then Φ_i does not necessarily converge to 0, hence some of the coefficients of the (estimated) forecast error covariance matrices will go to infinity when $h \to \infty$.

The forecast error variance decomposition and the conditions for instantaneous causality will of course have no reasons to change. Not even the test statistic for instantaneous causality will change, since we will later show in section 3.2.2 that the asymptotic normality of the maximum likelihood estimator for Σ_u in the VECM is the same as for the maximum likelihood estimator in the VAR model, which implies that the Wald statistic is the same.

The conditions for not having Granger-causality between variables also does not change. Remember from Theorem 2.9 that having $A_{12,i} = 0$ for $i = 1, 2, \ldots, p$ implied not having Granger-causality between certain variables and conversely. If we now simply use the VAR representation of a VECM, then it is easy to see that having $\mathbf{\Pi}_{12} = 0$ and $\mathbf{\Gamma}_{12,i} = 0$ for $i = 1, 2, \ldots, p-1$ implies not having Granger-causality between certain variables and conversely, where we partition $\mathbf{\Pi}$ and $\mathbf{\Gamma}$ the same way we partitioned $A_1, A_2, \ldots, A_p$ as in (2.72). However the Wald statistic we found earlier in section 2.5.1 will not always work correctly for the VECM, for reasons that the estimated Wald statistic on the coefficient matrices will not converge to the same χ^2 distribution. A more detailed description for this specific problem with testing Granger-causality can be found in (Toda and Phillips, 1993). A solution to this problem was proposed by (Dolado and Lütkepohl, 1996; Toda and Yamamoto, 1995). They suggested that the Wald statistic for the coefficient matrices did work if it was applied on a VAR model of order $p + 1$, called the *lag augmented* VAR model.

Finally the (orthogonal) impulse responses can be considered as well in the VECM. Remember from section 2.3.4 and 2.3.5 that the responses to a corresponding (orthogonal) impulse were represented by $(A^i)_{i \geqslant 1}$ (or $(\Theta_i)_{i \geqslant 1}$ for orthogonal impulses). The responses always converged to 0 when we had a stable VAR process, however when the process is not stable, these impulses do not necessarily converge to 0, hence a single impulse might result into a permanent effect in a certain variable.

3.2 Estimators

Just like in the VAR model we can find the ordinary least squares estimator and the maximum likelihood estimator of the VECM coefficient matrices. We will assume that we have a time series $y_1, y_2, \ldots, y_N$ with all necessary presample values and that u_t is i.i.d. multivariate normally distributed with covariance

matrix Σ_u. All of the theorems and propositions we will represent in this section can be found in (Lütkepohl, 2005, pp. 269-297).

3.2.1 Ordinary Least Squares estimator

We will use the following notations for the OLS estimator.

Definition 3.4. *We define*

$$\begin{aligned}
\Delta Y &:= (\Delta y_1, \Delta y_2, \dots, \Delta y_N) && (K \times N),\\
Y_{-1} &:= (y_0, y_1, \dots, y_{N-1}) && (K \times N),\\
\mathbf{\Gamma} &:= (\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \dots, \mathbf{\Gamma}_{p-1}) && (K \times KN),\\
\Delta X_t &:= \begin{bmatrix} \Delta y_t \\ \Delta y_{t-1} \\ \vdots \\ \Delta y_{t-(p-2)} \end{bmatrix} && (K(p-1) \times 1),\\
\Delta X &:= (\Delta X_0, \Delta X_1, \dots, \Delta X_{N-1}) && (K(p-1) \times N),\\
U &:= (u_1, u_2, \dots, u_N) && (K \times N),
\end{aligned}$$

such that we can rewrite the VECM into

$$\Delta Y = \mathbf{\Pi} Y_{-1} + \mathbf{\Gamma} \Delta X + U.$$

Using this notation it is possible to find the OLS estimator by applying the same methods we used in section 2.4.1 to obtain the OLS estimator for the VAR model. We then obtain the following estimator.

Theorem 3.1. *The ordinary least squares estimator for the VECM coefficient matrices* $\mathbf{\Pi}$ *and* $\mathbf{\Gamma}$ *is*

$$\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix} = \begin{bmatrix} \Delta Y Y_{-1}^T & \Delta Y \Delta X^T \end{bmatrix} \begin{bmatrix} Y_{-1} Y_{-1}^T & Y_{-1} \Delta X^T \\ \Delta X Y_{-1}^T & \Delta X \Delta X^T \end{bmatrix}^{-1}.$$

Again, asymptotic properties of the OLS estimator can be found. They are specified as follows.

Proposition 3.1. *Let* $\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix}$ *be the OLS estimator of a VECM of a process. Then the following asymptotic properties hold.*

1. *The least squares estimator* $\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix}$ *is consistent, i.e.*

$$\operatorname{plim} \begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{\Pi} & \mathbf{\Gamma} \end{bmatrix}.$$

2. *We have asymptotic normality of*

$$\sqrt{N} \operatorname{vec}(\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix} - \begin{bmatrix} \mathbf{\Pi} & \mathbf{\Gamma} \end{bmatrix}) \sim \mathcal{N}(0, \Sigma_{co}),$$

where

$$\Sigma_{co} = \left(\begin{bmatrix} \boldsymbol{\beta} & 0 \\ 0 & I_{K(p-1)} \end{bmatrix} \Omega^{-1} \begin{bmatrix} \boldsymbol{\beta}^T & 0 \\ 0 & I_{K(p-1)} \end{bmatrix} \right) \otimes \Sigma_u,$$

with

$$\Omega = \operatorname{plim} \frac{1}{N} \begin{bmatrix} \boldsymbol{\beta}^T Y_{-1} Y_{-1}^T \boldsymbol{\beta} & \boldsymbol{\beta}^T Y_{-1} \Delta X^T \\ \Delta X Y_{-1}^T \boldsymbol{\beta} & \Delta X \Delta X^T \end{bmatrix}.$$

The problem again with the asymptotic normality of the OLS estimator is that the matrix Σ_{co} is not known beforehand when estimating a process with unknown VECM coefficient matrices. That is why

we again have to find a consistent estimator of this matrix. It turns out that Σ_{co} can consistently be estimated by

$$\hat{\Sigma}_{\text{co}} = N \begin{bmatrix} Y_{-1}Y_{-1}^T & Y_{-1}\Delta X^T \\ \Delta X Y_{-1}^T & \Delta X \Delta X^T \end{bmatrix}^{-1} \otimes \hat{\Sigma}_u,$$

where

$$\hat{\Sigma}_u = \frac{1}{N - Kp} U U^T \tag{3.9}$$

for similar reasons as we found the estimator of the covariance matrix in (2.103) for the VAR model in section 2.4.2.

We can again also find the t-ratios of the estimator the same way we did for the OLS estimator of the VAR model in section 2.4.3. Now for similar reasons we can find using the asymptotic normality of $\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix}$ that

$$\frac{\text{vec}(\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix})_i - \text{vec}(\begin{bmatrix} \mathbf{\Pi} & \mathbf{\Gamma} \end{bmatrix})_i}{\hat{s}_i} \sim \mathcal{N}(0,1) \quad \forall i,$$

where $\text{vec}(\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix})_i$ and $\text{vec}(\begin{bmatrix} \mathbf{\Pi} & \mathbf{\Gamma} \end{bmatrix})_i$ are the i-th element of $\text{vec}(\begin{bmatrix} \hat{\mathbf{\Pi}} & \hat{\mathbf{\Gamma}} \end{bmatrix})$ and $\text{vec}(\begin{bmatrix} \mathbf{\Pi} & \mathbf{\Gamma} \end{bmatrix})$ respectively and $\hat{s}_i$ is the square root of the i-th row i-th column element of $\frac{1}{N}\hat{\Sigma}_{\text{co}}$.

3.2.2 Maximum Likelihood estimator

For the maximum likelihood estimator we will use the following notations.

Definition 3.5. *Using the same notations as in Definition 3.4, we define*

$$\begin{aligned}
M &:= I_N - \Delta X^T (\Delta X \Delta X^T)^{-1} \Delta X && (N \times N),\\
R_0 &:= \Delta Y M && (K \times N),\\
R_1 &:= Y_{-1} M && (K \times N),\\
S_{ij} &:= \frac{R_i R_j}{N}, \quad \textit{for } i = 0, 1 && (K \times K),\\
Y_{-p} &:= (y_{-p+1}, y_{-p+2}, \ldots, y_{N-p}) && (K \times N),\\
\boldsymbol{v} &:= (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K) && (K \times K),
\end{aligned}$$

where we define $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K$ to be the orthonormal eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_K$ of the $(K \times K)$ matrix

$$S_{11}^{-\frac{1}{2}} S_{10} S_{00}^{-1} S_{01} S_{11}^{-\frac{1}{2}}.$$

With these notations it is possible to find the following log-likelihood function of the VECM.

Proposition 3.2. *The log-likelihood function of the VECM is*

$$\begin{aligned}
l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{\Gamma}, \Sigma_u) &= -\frac{KN}{2}\ln(2\pi) - \frac{N}{2}\ln(\det(\Sigma_u)) \\
&\quad -\frac{1}{2}\operatorname{tr}\left[(\Delta Y - \boldsymbol{\alpha}\boldsymbol{\beta}^T Y_{-1} - \mathbf{\Gamma}\Delta X)^T (\Delta Y - \boldsymbol{\alpha}\boldsymbol{\beta}^T Y_{-1} - \mathbf{\Gamma}\Delta X)\right].
\end{aligned}$$

It can also be found that this log-likelihood function is maximized with the following maximum likelihood estimators.

Theorem 3.2. *The maximum likelihood estimators that maximize the log-likelihood function in Proposition 3.2 are*

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &:= \boldsymbol{v}^T S_{11}^{-\frac{1}{2}}, \\
\tilde{\boldsymbol{\alpha}} &:= S_{01}\tilde{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}^T S_{11} \tilde{\boldsymbol{\beta}})^{-1}, \\
\tilde{\mathbf{\Gamma}} &:= (\Delta Y - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}^T Y_{-1})\Delta X^T (\Delta X \Delta X^T)^{-1}, \\
\tilde{\Sigma}_u &:= \frac{1}{N}(\Delta Y - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}^T Y_{-1} - \tilde{\mathbf{\Gamma}}\Delta X)(\Delta Y - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}^T Y_{-1} - \tilde{\mathbf{\Gamma}}\Delta X)^T.
\end{aligned}$$

These maximum likelihood estimators again have asymptotic properties, which give some interesting results.

Proposition 3.3. *For the maximum likelihood estimators $\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Gamma}}$ and $\tilde{\Sigma}_u$ as in Theorem 3.2 the following holds.*

1. *The estimators $\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Gamma}}$ and $\tilde{\Sigma}_u$ are consistent estimators.*
2. *The matrix $\begin{bmatrix} \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}^T & \tilde{\boldsymbol{\Gamma}} \end{bmatrix}$ has the same asymptotic normality as the OLS estimator $\begin{bmatrix} \hat{\boldsymbol{\Pi}} & \hat{\boldsymbol{\Gamma}} \end{bmatrix}$, i.e.*

$$\sqrt{N}\,\mathrm{vec}(\begin{bmatrix} \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\beta}}^T & \tilde{\boldsymbol{\Gamma}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\alpha}\boldsymbol{\beta}^T & \boldsymbol{\Gamma} \end{bmatrix}) \sim \mathcal{N}(0, \Sigma_{co}),$$

   where Σ_{co} is defined as in Proposition 3.1.

3. *The estimator $\tilde{\boldsymbol{\sigma}} := \mathrm{vech}(\tilde{\Sigma}_u)$ has the same asymptotic normality as the maximum likelihood estimator of the VAR model for $\boldsymbol{\sigma} := \mathrm{vech}(\Sigma_u)$, i.e.*

$$\sqrt{N}(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) \sim \mathcal{N}(0, \Sigma_{\tilde{\boldsymbol{\sigma}}}),$$

   where

$$\Sigma_{\tilde{\boldsymbol{\sigma}}} = 2\boldsymbol{D}_K^+(\Sigma_u \otimes \Sigma_u)(\boldsymbol{D}_K^+)^T,$$

   with $\boldsymbol{D}_K$ defined as in Definition 2.20.

3.2.3 Obtaining the estimated equilibrium relations

With the maximum likelihood estimator in Theorem 3.2 we found a consistent estimator for $\boldsymbol{\beta}$, however with the OLS estimator we only found estimates for $\boldsymbol{\Pi}, \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_K$. When we use the OLS estimator, we know nothing about the estimated equilibrium relations $\hat{\boldsymbol{\beta}}$, other than $\hat{\boldsymbol{\Pi}} = \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\beta}}^T$. We do however know a straightforward estimator for $\boldsymbol{\alpha}$ if $\boldsymbol{\beta}$ is normalised as in (3.3). Since the first r columns of $\boldsymbol{\beta}^T$ will then be I_r, we find that a consistent estimator of $\boldsymbol{\alpha}$ will be the first r columns of $\boldsymbol{\Pi}$.

All that is left to find is the bottom $((K-r)\times r)$ matrix $\boldsymbol{\beta}_{(K-r)}$. It turns out that the following consistent estimator can be found.

$$\hat{\boldsymbol{\beta}}_{(K-r)} = (\hat{\boldsymbol{\alpha}}^T\hat{\Sigma}_u^{-1}\hat{\boldsymbol{\alpha}})^{-1}\hat{\boldsymbol{\alpha}}^T\hat{\Sigma}_u^{-1}\left(\sum_{i=1}^{N}(\Delta y_t - \hat{\boldsymbol{\alpha}}y_{t-1}^{(1)})(y_{t-1}^{(2)})^T\right)\left(\sum_{i=1}^{N}y_{t-1}^{(2)}(y_{t-1}^2)^T\right)^{-1},$$

where we divide y_t into $\begin{bmatrix} y_t^{(1)} \\ y_t^{(2)} \end{bmatrix}$, with $y_t^{(1)}$ the first r variables of y_t and $y_t^{(2)}$ the last $K-r$ variables of y_t. Furthermore we have $\hat{\boldsymbol{\alpha}}$ as the first r columns of the OLS estimator $\hat{\boldsymbol{\Pi}}$ and we have $\hat{\Sigma}_u$ as in (3.9). Finally with this estimator, we find the normalised estimator for $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} I_r \\ \hat{\boldsymbol{\beta}}_{(K-r)\cdot} \end{bmatrix}.$$

3.3 Cointegration rank selection

We now know how to estimate the coefficients of the VECM, however the lag order $p-1$ of the VECM and the cointegration rank $r := \mathrm{rk}(\boldsymbol{\Pi})$ are still unknown. Since we have shown in (3.4)-(3.7) that the VECM is basically another representation of the VAR model, we can simply apply the order selection criteria from section 2.6 on the differenced time series to find the lag order $p-1$. The final prediction error (FPE) criterion however will not work when working with unstable processes (Lütkepohl, 2005, p. 325), for reasons that involve the fact that some elements of the forecast error covariance matrix will approach infinity for the VECM.

In order to find the cointegration rank r we will test the following null hypothesis

$$H_0 \quad : \quad \text{rk}(\mathbf{\Pi}) = r \tag{3.10}$$

against the alternative hypothesis

$$H_1 \quad : \quad r < \text{rk}(\mathbf{\Pi}) < K. \tag{3.11}$$

Thus we will be testing whether the cointegration rank is r or if it is bigger than r. The idea of this test is to start test for $r = 0$, then keep testing for $r = 1, 2, \ldots$ until the null hypothesis is not rejected. The corresponding test statistic can be found to be the *LR statistic*, which we find in the following proposition (Lütkepohl, 2005, pp. 328-329).

Proposition 3.4. *The test statistic to test H_0 in (3.10) against H_1 in (3.11) is called the LR statistic and is defined as*

$$\begin{aligned} \lambda_{LR}(r, K) &:= 2\left[\ln(K) - \ln(r)\right] \\ &= -N \sum_{i=r+1}^{K} \ln(1 - \lambda_i), \end{aligned}$$

where λ_i for $i = 1, 2, \ldots, K$ are defined as in Definition 3.5 and $l(K)$ and $l(r)$ are the maximum log-likelihood functions of Proposition 3.2 where the maximum likelihood estimators are estimated with cointegration rank K and r respectively.

It has been found that the LR statistic is actually distributed as

$$\lambda_{LR}(r, K) \sim \text{tr}(\mathcal{D}),$$

where $\mathcal{D}$ is a specific function of $(K - r)$-dimensional Brownian motions (Johansen, 1988, 1995). This test statistic is also referred to as the *trace statistic*.

4 Application: Tech companies in the AEX

In this section we will perform a time series analysis, where we will apply all methods and techniques we have found in all of the sections before. The time series we will be looking at will be containing 4 variables of interest, which are the AEX-index, the Adyen stock closing prices, the ASML stock closing prices and the Philips stock closing prices from the 13th of June 2018 until the 13th of June 2019. The reason that we are not using more data before the 13th of June 2018 is since this was the IPO date of Adyen, which means that Adyen only started trading its stocks since that day. The AEX-index is a Dutch index that represents Dutch companies who trade on the Amsterdam Stock Exchange. The companies Adyen, ASML and Philips represent 3 tech companies that are part of the AEX-index. Our time series looks as follows.

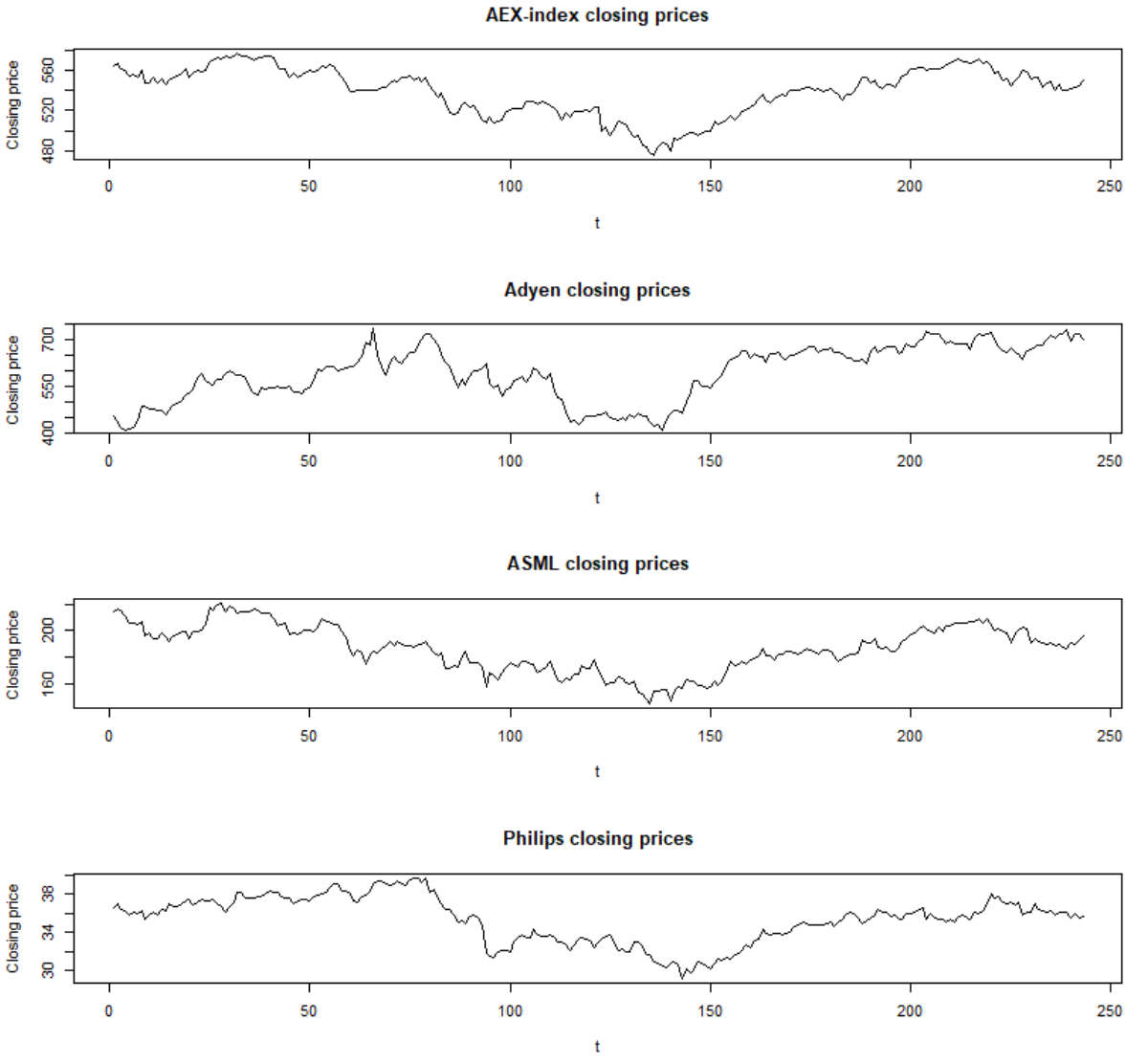


Figure 4.1: The time series of our variables of interest.

In Figure 4.1 we have $t = 1$ represents the data of 13th of June 2018 and so on. The time series will be represented as y_t through this whole section. We will be performing time series analysis on these variables and present our conclusions.

4.1 The model

First we will have to determine what model we will use for our time series. From Figure 4.1 it is obvious that our data is non-stationary and thus we should use a VECM. In order to use the VECM, we need to check if all variables of interest are integrated of order 1.

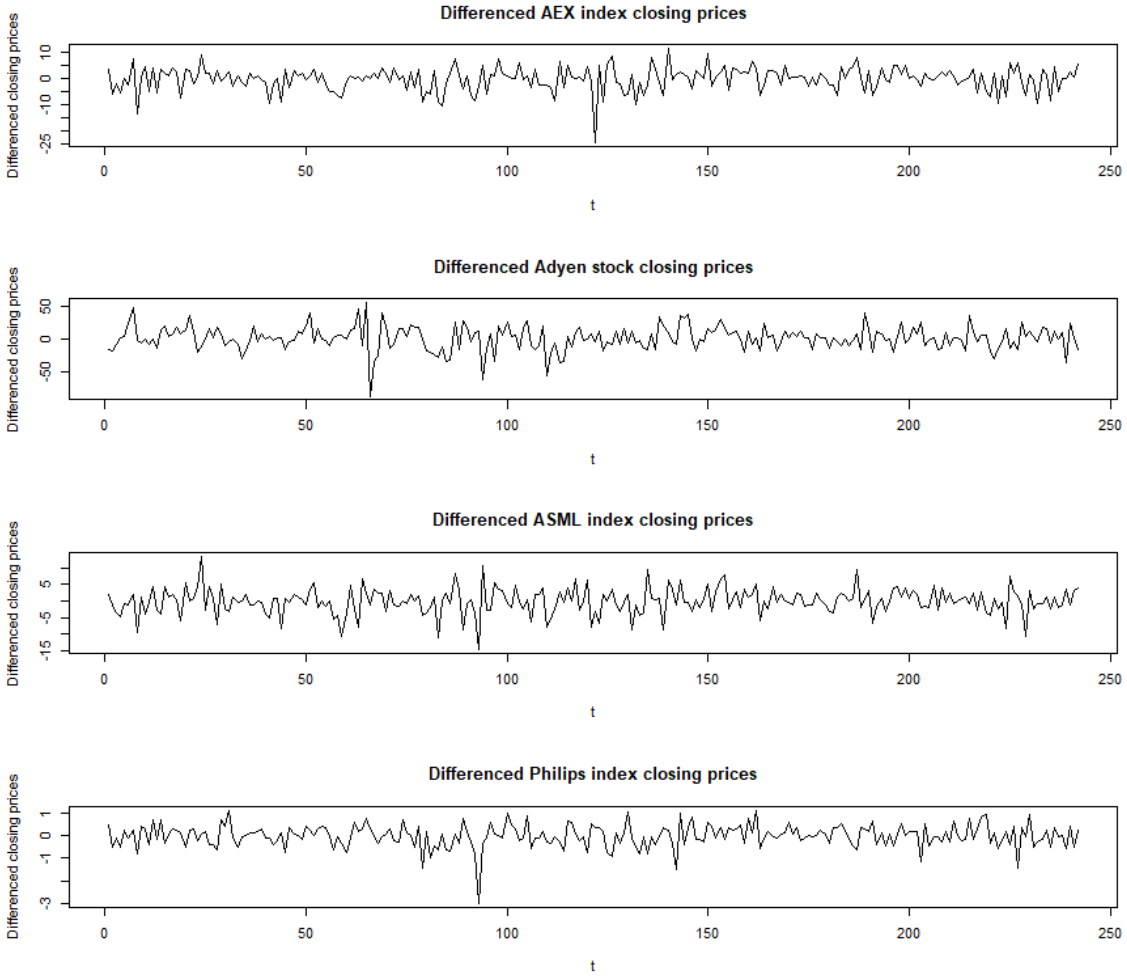


Figure 4.2: The differenced time series of our variables of interest

In Figure 4.2 the data seems to be stationary for all of the variables of interest. Since the original data is non-stationary, we can conclude that the variables of interest are integrated of order 1, hence we can apply the VECM. Before we start estimating our VECM coefficients, we first should determine the cointegration rank of our model. Let us observe the following trace statistics.

	VECM order 1	**VECM order 2**	**VECM order 3**			
r	$\lambda_{LR}(r,4)$	$\lambda_{LR}(r,4)$	$\lambda_{LR}(r,4)$	90%	**95%**	99%
3	2.92	3.45	2.87	10.49	**12.25**	16.26
2	14.93	17.31	14.73	22.76	**25.32**	30.45
1	**33.76**	**33.85**	32.38	39.06	**42.44**	48.45
0	70.99	64.11	**61.42**	59.14	**62.99**	70.05

Table 8: Trace statistics for various cointegration ranks and VECM orders. The last 3 columns show the critical values of the trace statistic.

Remember that the lowest cointegration rank that does not reject the null hypothesis is the cointegration rank we should choose. The bold test statistics in Table 8 have this property when we have a significance level of 0.05. The bold critical values on the right represent the critical values with a significance level of 0.05, which we will compare with the test statistics. We see that for a VECM of order 1 and 2 we have a cointegration rank of 1, however for a VECM of order 3 we have a cointegration rank of 0. One can argue based on these results that in general the cointegration rank should be 1, because the trace statistic of the VECM of order 3 lies between the 90% and 95% critical values and hence the trace statistic is almost significant.

If we want to make sure we choose the right cointegration rank, we should perform the order selection criteria on the differenced time series. We then find the following values of the criteria.

VECM order	AIC	HQ	SC
1	9.599	**9.719**	**9.896**
2	**9.538**	9.755	10.0727
3	9.560	9.872	10.333
4	9.619	10.026	10.629
5	9.709	10.213	10.957
6	9.805	10.404	11.29

Table 9: Values of various order selection criteria and VAR orders.

Note that we do not use the FPE criterion for reasons we discussed in section 3.3. In Table 9 we have the bold values representing the minimum values of the criteria. We will take the highest VECM order suggested by all of the criteria just to be safe, which is a VECM order of 2. Table 8 now suggests that we have a cointegration rank of 1.

We now have decided to use a VECM of order 2 and cointegration rank 1, but the mean of the process is still unspecified. Just to be safe, let us first assume that we have a VECM with a linear trend, which means we will use the model

$$\Delta y_t = \nu_0 + \nu_1 t + \mathbf{\Pi} y_{t-1} + \mathbf{\Gamma_1} \Delta y_{t-1} + \mathbf{\Gamma_1} \Delta y_{t-2},$$

where the rank of $\mathbf{\Pi}$ is 1. Now we can start estimating the coefficients of the model. We find

$$\mathbf{\Pi} = \begin{bmatrix} -0.063 & 0.010 & 0.076 & -0.335 \\ -0.073 & -0.072^{**} & 0.216 & 0.768 \\ 0.150^{*} & 0.002 & -0.207^{**} & -0.305 \\ 0.025^{**} & 0.0003 & -0.019^{\bullet} & -0.131^{***} \end{bmatrix},$$

$$\mathbf{\Gamma}_1 = \begin{bmatrix} -0.188^{\bullet} & 0.001 & 0.281^{**} & -0.754 \\ 0.166 & 0.109 & 0.335 & -1.714 \\ 0.143^{\bullet} & -0.007 & -0.147 & -0.573 \\ -0.008 & 0.001 & 0.013 & -0.021 \end{bmatrix}, \quad \mathbf{\Gamma}_2 = \begin{bmatrix} 0.069 & -0.021 & 0.010 & 1.493^{*} \\ 0.894^{*} & 0.102 & 0.034 & 0.061 \\ 0.142^{\bullet} & -0.016 & -0.161^{\bullet} & 1.362^{*} \\ 0.004 & 0.001 & 0.002 & 0.113^{\bullet} \end{bmatrix},$$

$$\nu_0 = \begin{bmatrix} 26.937 \\ 9.154 \\ -32.468^{\bullet} \\ -5.178^{*} \end{bmatrix}, \quad \nu_1 = \begin{bmatrix} -0.007 \\ 0.056^{\bullet} \\ -0.004 \\ -0.001^{\bullet} \end{bmatrix}.$$

We use here superscripts on the values of the matrices to show certain significance. The p-values of the t-ratios with superscript *** are lower than 0.001, with superscript ** between 0.001 and 0.01, with superscript * between 0.01 and 0.05 and with superscript $\bullet$ between 0.05 and 0.10. We use this notation to see which values are definitely significant and which values are close to being significant. We see that each coefficient matrix has at least one significance value, except for the trend vector v_1. However, we

could argue that the trend vector still might be significant, since we only have a small data set of 243 data points and 2 values have a p-value below 0.10.

Using this model we can find the following values our vector with cointegration relations $\boldsymbol{\beta}$.

$$\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 0.028 \\ -1.030 \\ -3.813 \end{bmatrix}.$$

This means we have an equilibrium relation of

$$y_{1t} + 0.028y_{2t} - 1.030y_{3t} - 3.813y_{4t} = 0,$$

where y_{1t}, y_{2t}, y_{3t} and y_{4t} represent the closing prices of the AEX-index, the Adyen stock, the ASML stock and the Philips stock respectively. The closing prices of the Adyen stock seems not to cointegrate well with the other variables. This probably happens since Adyen only started trading its stocks on the market at the start of our time series. Stock prices in general will have a different behaviour during the first few months after the IPO.

4.2 Diagnostic checking

Now that we have decided what model we will use, we can perform diagnostic checking, which is performing tests for the residuals of the VAR representation of the model. Let us look at the first test we presented for the residuals, which is the non-normality test.

Test	**Test statistic**	**95% quantile**	**P-value**
Non-normality test	569.01	15.51	0

Table 10: Non-normality test of the residuals.

The non-normality test shows us an incredibly high test statistic, hence our residuals are definitely not normally distributed. A reason why we have this high test statistic could be found using the ARCH-LM test (Engle, 1982).

Test	**Test statistic**	**95% quantile**	**P-value**
ARCH-LM test	650.12	553.13	$6.45 * 10^{-6}$

Table 11: ARCH-LM test of the residuals.

This test shows that an ARCH effect occurs in the residuals, which means that the squared residuals are correlated with each other. ARCH effects occur often in financial data. When one finds an ARCH effect in the residuals, then normally other models should be used. However, we will continue our analysis with the VECM and keep this result in mind.

Since our residuals or not normally distributed, there are some consequences. Our forecast intervals are based on the assumption that the residuals are normally distributed, but since the residuals are not normally distributed, we can not trust our forecast intervals.

The second test we presented for the residuals is the portmanteau test, where we test autocorrelation between the residuals. If we test for autocorrelation up to lag 10, we find the following.

Test	**Test statistic**	**95% quantile**	**P-value**
Portmanteau	121.56	142.14	0.34

Table 12: Portmanteau test of the residuals for $h = 10$.

The test shows that there seems to be no autocorrelation between the residuals up to lag 10. We now can still not conclude that the residuals are white noise, since we do not know if the residuals have a mean 0 and have equal variances for all t.

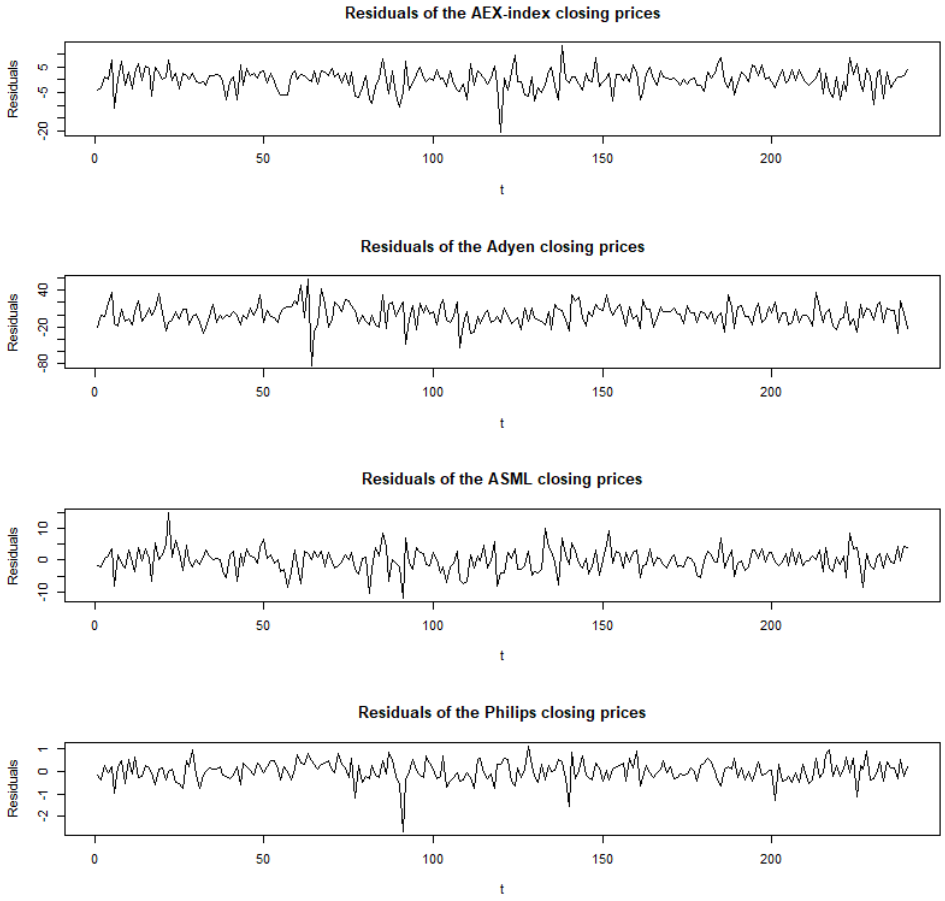


Figure 4.3: The residuals of the VAR representation for all variables of interest.

If we look at the residuals in Figure 4.3, then the assumption that the residuals have a mean 0 and that they have equal variances for all t seems to be a valid assumption we can make, hence we assume the residuals to be white noise. The estimated covariance matrix of the white noise residuals turns out to be

$$\Sigma_u = \begin{bmatrix} 17.991 & 5.462 & 9.688 & 0.183 \\ 5.462 & 307.554 & -1.289 & 2.191 \\ 9.688 & -1.289 & 13.695 & 0.122 \\ 0.183 & 2.191 & 0.122 & 0.214 \end{bmatrix}.$$

4.3 Causality tests

Now that we have performed diagnostic checking, we can apply causality tests on our model. Let us first look at the Granger-causality between our variables. Note that for Granger-causality, we will apply the Wald test on the lag augmented var representation, hence on the VAR(4) representation. The reason for this has been motivated in section 3.1.3. We will be testing if no Granger-causality exists between the variables against the existence of Granger-causality between the variables. We find the following p-values of our test statistics.

x_t	z_t	x_t **does not Granger-cause** z_t	z_t **does not Granger-cause** x_t
AEX-index	Adyen, ASML, Philips	$3.313 * 10^{-4}$	0.020
AEX-index, Adyen	ASML, Philips	0.0044	0.027
AEX-index, ASML	Adyen, Philips	$1.208 * 10^{-4}$	**0.066**
AEX-index, Philips	Adyen, ASML	0.0045	**0.116**
AEX-index, Adyen, ASML	Philips	$2.55 * 10^{-4}$	**0.326**
AEX-index, Adyen, Philips	ASML	0.0033	0.036
AEX-index, ASML, Philips	Adyen	0.019	**0.817**

Table 13: P-values of the Granger-causality tests.

The bold values in Table 13 represent the tests with p-values greater than 0.05, hence for these variables we have that z_t does not Granger-cause x_t. Whenever z_t does not Granger-cause x_t, it implies that the forecast of x_t will not be improved when z_t is added to the information set.

We see in Table 13 that all of the possible combinations of variables containing the AEX-index, Granger-causes the other variables of interest. Adding the AEX-index closing prices to the information set seems to improve the forecasts of the other variables in general. However, a few combinations of variables of interest with the Adyen stock closing prices seems to not Granger-cause the other variables. The p-value of testing if the Adyen stock closing prices do not Granger-causes the other variables has a really high value of 0.817. It looks like the Adyen stock closing prices is not really helping with forecasting the time series. Some combinations of variables containing the Philips stock closing prices do not Granger-cause the other variables as well. However, these test statistics have lower p-values than the combinations involving the Adyen stock closing prices.

The second type of causality we can investigate is instantaneous causality. The following p-values of the test statistics can be found.

x_t	z_t	**No instantaneous causality between** x_t **and** z_t
AEX-index	Adyen, ASML, Philips	$6.661 * 10^{-15}$
AEX-index, Adyen	ASML, Philips	0
AEX-index, ASML	Adyen, Philips	**0.206**
AEX-index, Philips	Adyen, ASML	0
AEX-index, Adyen, ASML	Philips	0.0018
AEX-index, Adyen, Philips	ASML	$7.327 * 10^{-15}$
AEX-index, ASML, Philips	Adyen	0.0016

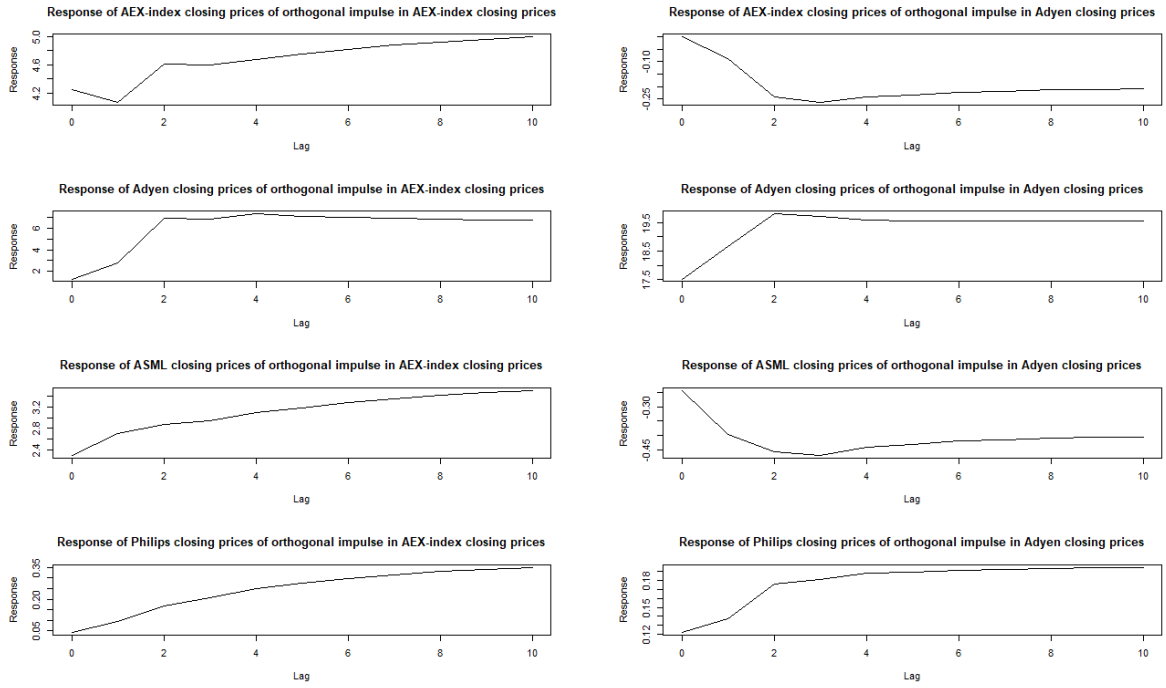
Table 14: P-values of the instantaneous causality tests.

Again, the bold value in Table 14 represents the p-value that is greater than 0.05. Remember that instantaneous causality between x_t and z_t implies that adding the values of the next time step of one of the variables improves the 1-step prediction of the other variable.

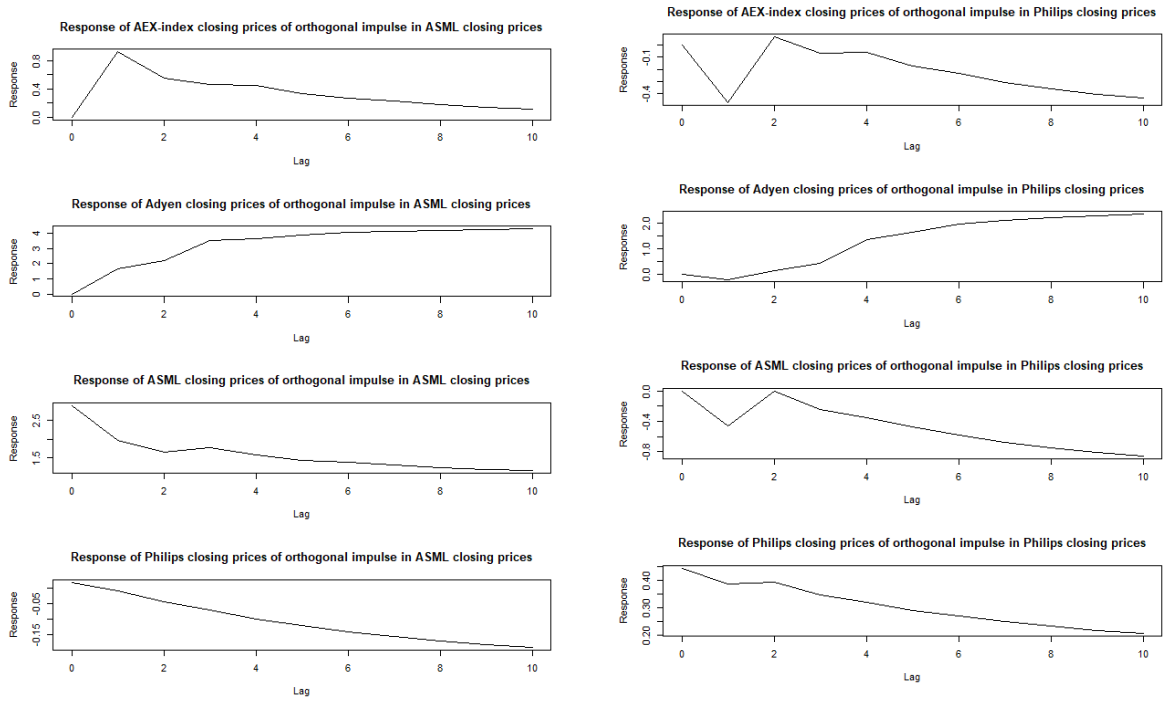
In Table 14 we find that only the AEX-index and the ASML stock closing prices have no instantaneous causality with the Adyen and the Philips stock closing prices. If we know the values of the next time step of one of the combinations, then we will find the same 1-step prediction we would have found without these values. We also see that the instantaneous causality between a combination of variables containing the AEX-index closing prices always has instantaneous causality with a combination of variables containing the ASML stock closing prices. These specific combinations show incredibly low p-values of the test statistics. Probably knowing the next value of the AEX-index closing prices helps improving the 1-step prediction of the ASML stock closing prices and vice versa.

4.4 Orthogonal impulse response functions

Let us now look at the orthogonal impulse response functions. Note that we use orthogonal impulses for reasons we discussed in section 2.3.5.



(a) Orthogonal impulse in the AEX-index closing prices. (b) Orthogonal impulse in the Adyen stock closing prices.



(c) Orthogonal impulse in the ASML stock closing prices. (d) Orthogonal impulse in the Philips stock closing prices.

Figure 4.4: Orthogonal impulse response functions of our time series.

In Figure 4.4b we see that the responses of the AEX-index closing prices and the ASML stock closing prices are really small compared to their responses on orthogonal impulses of other variables. It looks like these responses are zero orthogonal impulse responses, hence an orthogonal impulse of the Adyen stock closing prices results in almost no response in the AEX-index closing prices and the ASML stock closing prices. In Figure 4.4d we see that the same occurrence happens with an orthogonal impulse in the Philips stock closing prices. However, these responses of the AEX-index closing prices and the ASML stock closing prices seems to be higher than in Figure 4.4b.

We can find another interesting result in Figure 4.4a and Figure 4.4c. The orthogonal impulse of the AEX-index closing prices results into a high response of the ASML stock closing prices. The ASML stock closing prices seems to respond heavily when the AEX-index suddenly changes. However, the orthogonal impulse of the ASML stock closing prices results into a low response of the AEX-index stock closing prices. We find these results, since we are working with an index, where ASML is part of. A sudden change in the ASML stock closing prices will not effect the AEX-index closing prices that much, since the AEX-index is based on many other companies as well. A sudden change in the AEX-index closing prices suggest that the stock closing prices of all companies within the AEX-index will change on average as well, including ASML.

4.5 Forecast error variance decomposition

The forecast error variance decomposition can also be considered. We find the following proportions of the forecast error variance.

Forecasted variable of interest	**Forecast horizon**	**Proportions of AEX-index**	**Proportions of Adyen**	**Proportions of ASML**	**Proportions of Philips**
AEX-index	1	1	0	0	0
	2	0.970	$2.143 * 10^{-4}$	0.024	0.006
	3	0.975	0.001	0.020	0.004
	4	0.978	0.002	0.017	0.003
	5	0.980	0.002	0.016	0.002
	10	0.986	0.002	0.008	0.003
Adyen	1	0.005	0.995	0	0
	2	0.014	0.982	0.004	$6.821 * 10^{-5}$
	3	0.051	0.942	0.007	$5.751 * 10^{-5}$
	4	0.067	0.920	0.013	$1.546 * 10^{-4}$
	5	0.078	0.904	0.017	0.001
	10	0.093	0.874	0.028	0.005
ASML	1	0.381	0.004	0.615	0
	2	0.496	0.009	0.487	0.008
	3	0.571	0.012	0.411	0.006
	4	0.608	0.013	0.373	0.006
	5	0.642	0.014	0.338	0.006
	10	0.739	0.013	0.227	0.020
Philips	1	0.009	0.070	0.002	0.920
	2	0.028	0.086	0.001	0.884
	3	0.065	0.107	0.004	0.824
	4	0.102	0.120	0.009	0.768
	5	0.142	0.130	0.017	0.711
	10	0.297	0.148	0.065	0.490

Table 15: Proportions of the forecast error variance of our variables of interest for forecasting each of the variables of interest.

The proportions of the forecast error variance of forecasting the AEX-index closing prices and the Adyen stock closing prices seems to have a large proportion of the forecast error variance by its own variable.

This is not a surprise for the AEX-index closing prices, since it is an index of many more companies, hence individual companies have small influence.

The proportions of the forecast error variance of forecasting the ASML stock closing prices seems to be more spread out. The AEX-index closing prices proportions seems to increase in the long run and will eventually have a higher proportion than the ASML stock closing prices itself. The proportions of the Adyen and the Philips stock closing prices seems to be nearly 0 for all forecast horizons.

For the Philips stock closing prices it seems that for small forecast horizons, we have a large proportion of its own variable, however in the long run the proportions will be more spread out.

4.6 Forecasting

Finally, let us take a look at the forecasts of all variables of interest.

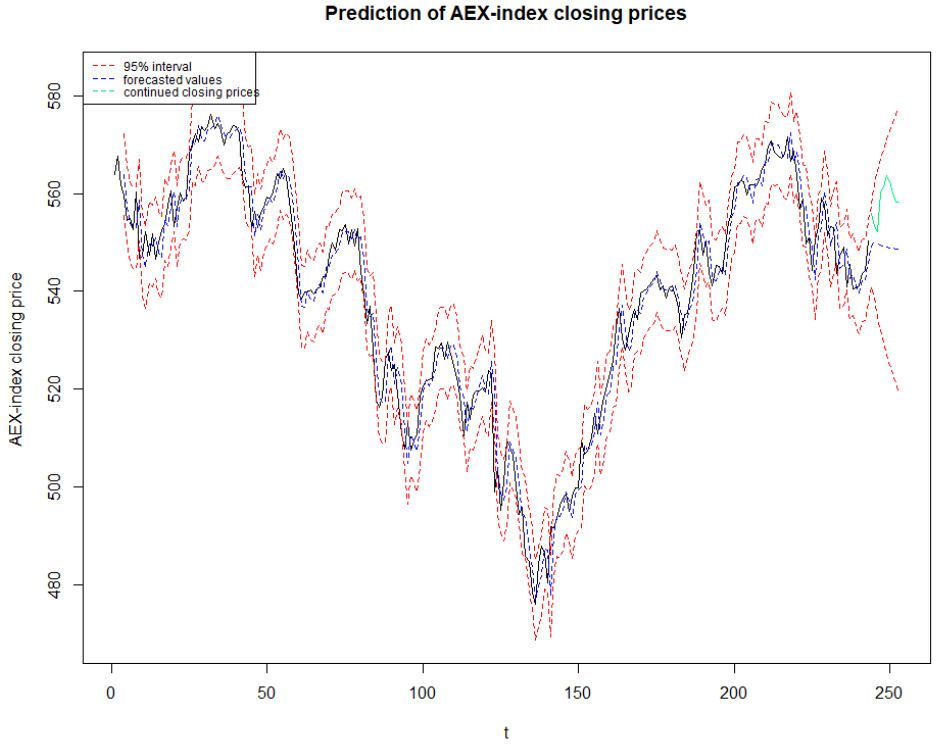


Figure 4.5: Predictions of the AEX-index closing prices.

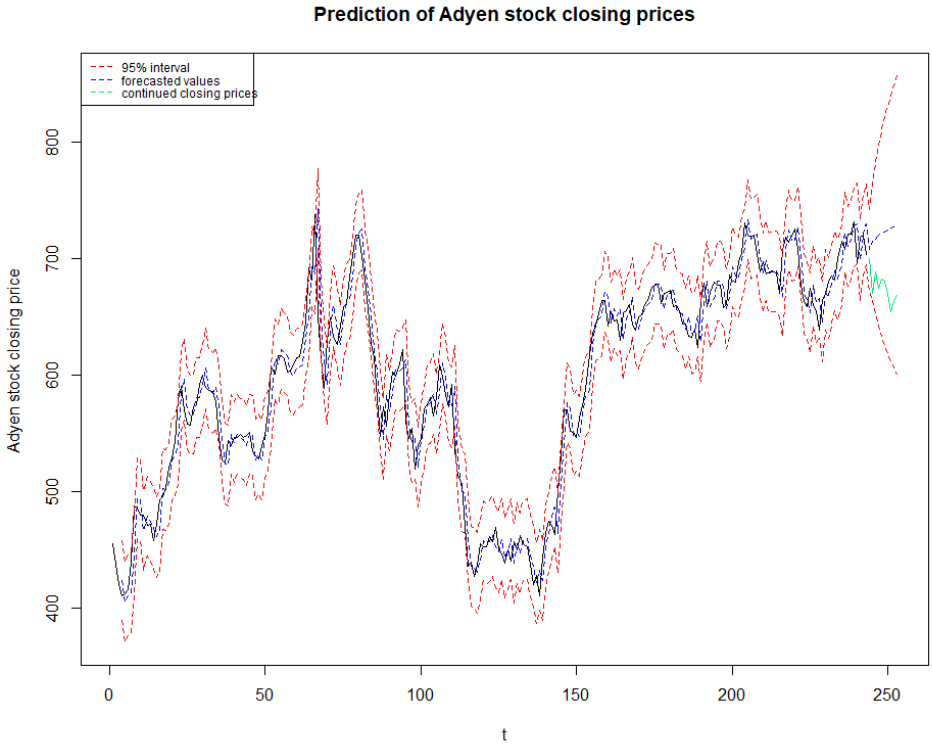


Figure 4.6: Predictions of the Adyen stock closing prices.

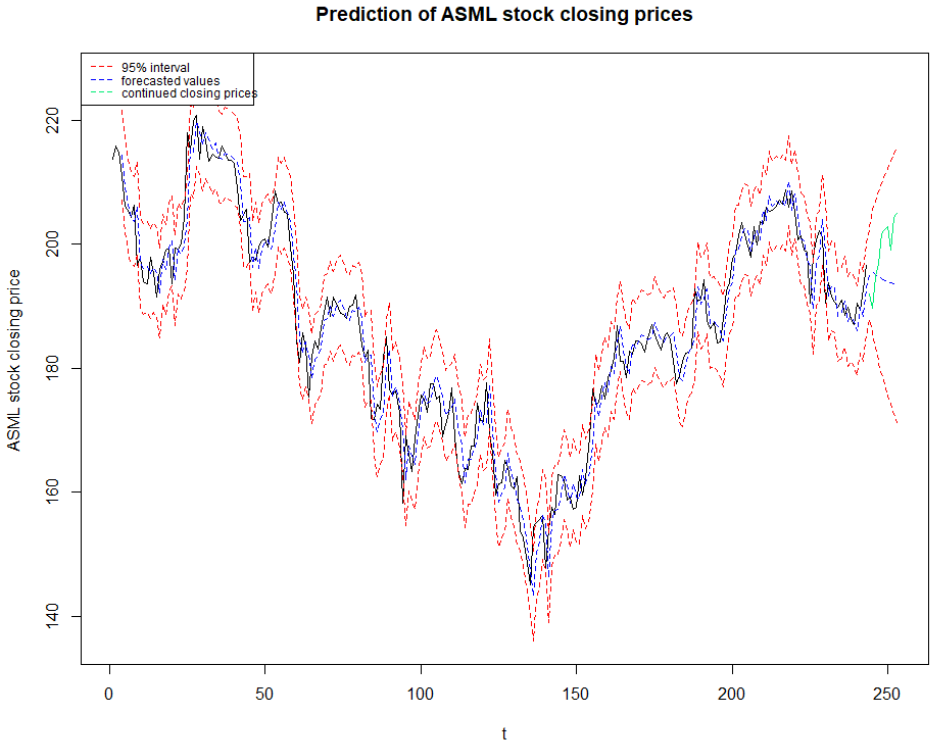


Figure 4.7: Predictions of the ASML stock closing prices.

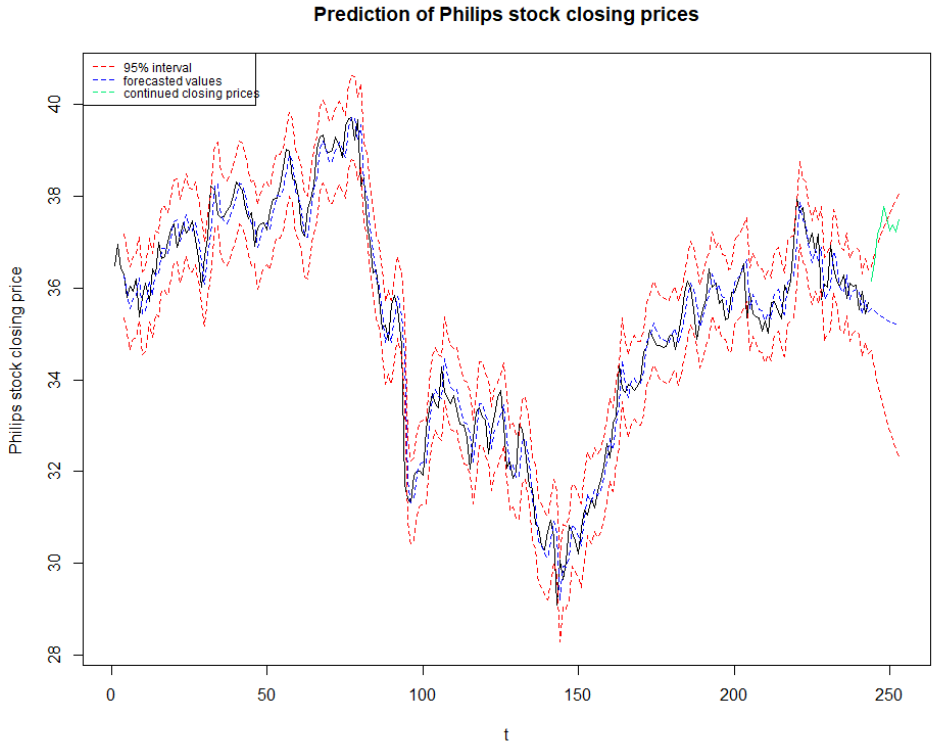


Figure 4.8: Predictions of the Philips stock closing prices.

Note that the continued closing prices represent the data past 13th of June 2019. In all figures we see that the 95% intervals seems to be too large for the variables of interest. These results are not unexpected, since we showed before that our residuals are not normally distributed. Our 95% intervals are based on the assumption that the residuals are normally distributed, hence we should indeed not trust these intervals. The 1-step forecasted values seems to perform well for $t \leqslant 243$, however, the h-step forecasted values for $t > 243$ do not seem to match the continued closing prices.

5 Conclusion

We have seen that we can use the VAR model and the VECM to model multivariate time series. Only when our time series is stationary, we found that we can use the VAR model, however, when the time series is non-stationary and is integrated of order 1, then we found that VECM can be used. For both models we have found estimators and order selection criteria, which we can apply to find the parameters of the models.

In addition, various analysis methods has been presented to perform time series analysis with these models, including

- forecasting,
- forecast error variance decomposition,
- causality analysis,
- (orthogonal) impulse response analysis.

The differences of these analysis methods between the VAR model and the VECM model has been discussed as well and they turn out to be really small. A reason for this was that the VECM has a VAR representation, hence the analysis methods can be applied on that representation.

Finally, we applied all of our presented methods on real-world financial data. We showed how one could perform time series analysis using only the VAR model and the VECM. Most of the analysis methods gave us a nice interpretation of our time series. However, only the forecast of the variables of interest seemed not to provide trustworthy results, since the residuals turned out not to be normally distributed.

In general it turns out that not all financial time series can be properly analysed with the VAR model or the VECM. These models contain a lot of assumptions and restrictions for the time series, hence many time series can not be properly analysed with these models. When the time series does not have these assumptions or restrictions, some of the analysis methods would give results we can not trust. Further improved models should then be applied on the time series.

6 Discussion

The model we found in our application did not have a lot of significant values. The reason for this to happen, was that our data set was too small. We previously showed that data sets with a sample size of approximately larger than 1000 samples would have nicely estimated coefficients. However, our time series only had 243 data points. The reason that we did not have a larger sample size, was because Adyen only started trading its stocks at $t = 1$, hence we simply did not have any more data available. In reality, most of the time our data is not perfect to work with, which our time series is a perfect example of. Whenever the data is not perfect, it does not imply that we can not perform time series analysis on that data at all. One should just perform time series analysis and argue which of the results can be trusted and which can not.

In addition, we assumed for the residuals of both models that the covariance matrix is constant for all t. However, often this assumption is wrong. For example, in the application we have seen that an ARCH effect occurred in the residuals, which implied that the squared residuals correlated with each other, hence the covariance matrix is not constant. We should then consider using further improved models, such as the multivariate ARCH model or the multivariate GARCH model. For further research, one could find more information of these models in (Lütkepohl, 2005, pp. 557-584).

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.

Akaike, H. (1971). Autoregressive model fitting for control. *Annals of the Institute of Statistical Mathematics*, 23(1):163–180.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in B. N. Petrov, F. Csáki (eds). *2nd International Symposium on Information Theory*, pages 267–281.

Akaike, H. (1974). A new look at the statistical model identication. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Brezinski, C. (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à celui des inconnues (Procédé du Commandant Cholesky)[note on a method for solving normal equations through the application of the least squares method to a system with fewer linear equations than unknowns (Method of Commandant Cholesky)]. *Bulletin Géodésique*, 2:67–77.

Castle, J. L. and Hendry, D. F. (2010). A low-dimension portmanteau test for non-linearity. *Journal of Econometrics*, 158(2):231–245.

Dolado, J. J. and Lütkepohl, H. (1996). Making wald tests work for cointegrated var systems. *Econometric Reviews*, 15(4):369–386.

Doob, J. L. (1935). The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 3rd edition.

Granger, C. W. J. (1969). Prediction with a generalized cost of error function. 20.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B*, 41(2):190–195.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163–172.

Johansen, S. (1988). Statistical analysis of cointegrated vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, Oxford.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Paulsen, J. (1984). Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis*, 5(2):115–127.

Quinn, B. G. (1980). Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society. Series B*, 42(2):182–185.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Cengage Learning, 3rd edition.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Toda, H. Y. and Phillips, P. C. B. (1993). Vector autoregressions and causality. *Econometrica*, 61(6):1367–1393.

Toda, H. Y. and Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66(1-2):225–250.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326.

Walker, G. (1931). On the periodicity in series of related terms. *Proceedings of the Royal Society of London*, 131(818):518–532.

Wold, H. (1938). *A Study in the Analysis of Stationary time series*. Almqvist Wiksells, Uppsala.

Yule, U. G. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London*, 226:267–298.