3. Vector Autoregressive Moving Average (VARMA) Models

"In every action, we must look beyond the action at our past, present and future state, and at others whom it affects, and see the relations of all these things."
(Blaise Pascal, The Thoughts of Blaise Pascal)

In Chapter 2, we have focused our attention on univariate time series models. However, because markets and institutions are highly intercorrelated, in financial applications we often need to jointly model a number of different time series to study the dynamic relationship among them. Therefore, in this chapter, we introduce econometric models for **multivariate time series analysis**. Loosely speaking, instead of focusing on the time series realization of a single variable, as we did in Chapter 2, now we consider a set of variables (e.g., the log-returns of N assets or the yields of Treasury bonds for N different maturity buckets), $\mathbf{y}_t = \left[y_{1,t}, y_{2,t}, ..., y_{N,t} \right]'$ with $t = 1, 2, ..., T$, where T is the number of observations in the series. The resulting sequence is called a N-dimensional (discrete) vector stochastic process.

In particular, we devote most of our attention to the **vector autoregressive** (VAR) models popularized by Sims (1980) that have come to be commonly used in financial applications. These are very flexible models where a researcher needs to know very little *ex-ante* theoretical information about the relationship among the variables to guide the specification of the model and all variables are treated as a-priori endogenous. In fact, as we shall see throughout this chapter, a VAR allows each variable to depend not only on its own lags (and/or combinations of white noise terms) but also on the lags of the other variables in the model.

In the rest of the chapter, we proceed as follows. First, we generalize the concepts of (weak) stationarity to the case of N-dimensional vector time series and discuss how to compute the

first two moments of the resulting multivariate distribution. Second, we introduce VAR models in their structural and reduced forms and their applications, including impulse response function analysis and variance decomposition. Third, we introduce the concept of Granger causality and show how to test for it. Finally, we briefly introduce vector moving average (VMA) and vector autoregressive moving average (VARMA) models. In this chapter, we focus as much as possible on the intuition and on the applications and as little as possible on the algebra and related technicalities. Of course, these remain important to any rigorous approach: an in-depth review of the statistical theory underlying multivariate time series analysis can be found in Lütkepohl (2005) and Reinsel (1993).

1- Foundations of Multivariate Time Series Analysis

1.1 Weak Stationarity of Multivariate Time Series

In Chapter 2, we have introduced the concept of stationarity of a time series as a necessary condition to be able to use past observations of a variable to forecast its future realizations. In particular, we said that a time series is (strictly) stationary if its statistical properties do not change over time and that it is **weakly stationary** if its first two moments are time invariant. These definitions still apply when we generalize them to multivariate time series.

Definition 3.1. (Weak Stationarity) Consider a N-dimensional time series $\mathbf{y}_t = \left[y_{1,t}, y_{2,t}, ..., y_{N,t} \right]'$. Formally, this is said to be weakly stationary if its first two unconditional moments are finite and constant through time, i.e.,

- $E\left[\mathbf{y}_t\right] \equiv \boldsymbol{\mu} < \infty$ for all t;
- $E\left[\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)'\right] \equiv \boldsymbol{\Gamma}_0 < \infty$ for all t;
- $E\left[\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbf{y}_{t-h} - \boldsymbol{\mu}\right)'\right] \equiv \boldsymbol{\Gamma}_h$ for all t and h.

where the expectations are taken element-by-element over the joint distribution of $\mathbf{y}_t$. In particular, $\boldsymbol{\mu}$ is the vector of the means, $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_N]'$, and $\boldsymbol{\Gamma}_0$ is the $N \times N$ covariance matrix where

the *i*th diagonal element is the variance of $Y_{i,t}$ and the (*i*, *j*)th element is the covariance between $Y_{i,t}$ and $Y_{j,t}$. Finally, $\mathbf{\Gamma}_h$ is the cross-covariance matrix at lag *h*.

Of course, the definition of weak stationarity provided above is completely analogous to the one discussed in Chapter 2 (as it is a corresponding definition of strict stationarity that is omitted here to save space), but it requires the computation of cross-covariance and cross-correlation matrices, that we shall discuss in Section 1.2.

1.2 Cross-Covariance and Cross-Correlation Matrices

While a Reader should be familiar with the computation of the covariance matrix at lag zero, we provide a primer on how to get the correlation matrix (at lag zero) from the covariance matrix $\mathbf{\Gamma_0}$. Let $\mathbf{D}$ be a $N \times N$ diagonal matrix collecting (on its main diagonal) the standard deviations of $Y_{i,t}$ for $i = 1,...,N$. The concurrent (i.e., at lag zero), correlation matrix of $\mathbf{y}_t$ is defined as

$$\boldsymbol{\rho}_0 = \mathbf{D}^{-1}\mathbf{\Gamma}_0\mathbf{D}^{-1}, \tag{3.1}$$

where the (*i, j*)th element of $\boldsymbol{\rho}_0$ is the correlation coefficient between $Y_{i,t}$ and $Y_{j,t}$ at time *t*:

$$\rho_{i,j}(0) = \frac{Cov[Y_{i,t}, Y_{j,t}]}{\sigma_{i,t}\sigma_{j,t}}. \tag{3.2}$$

Because $\rho_{i,j}(0) = \rho_{j,i}(0)$, $-1 \le \rho_{i,j} \le 1$, and $\rho_{i,i} = 1$ for $1 \le i$ and $j \le N$, $\boldsymbol{\rho}_0$ is a symmetric matrix with unit diagonal elements.

We are now interested in computing the cross-covariance and cross-correlation matrices at lags different from 0. More specifically, the lag-*h* cross-covariance matrix of $\mathbf{y}_t$ is defined as:

$$\mathbf{\Gamma}_h = E\left[\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbf{y}_{t-h} - \boldsymbol{\mu}\right)'\right], \tag{3.3}$$

where $\boldsymbol{\mu}$ is the mean vector of $\mathbf{y}_t$. Therefore, the (*i, j*)th element of $\mathbf{\Gamma}_h$ is the covariance between $Y_{i,t}$ and $Y_{j,t-h}$. From Definition 3.1, for a weakly stationary time series, the cross-covariance matrix is

time-invariant, i.e., it only depends on the lag length *h* and not on the temporal index *t*.
The lag-*h* cross-correlation matrix is defined as

$$\boldsymbol{\rho}_h = \mathbf{D}^{-1}\boldsymbol{\Gamma}_h\mathbf{D}^{-1}, \tag{3.4}$$

where, as before, **D** is the diagonal matrix of standard deviations of the individual series $y_{i,t}$. Therefore, the (*i, j*)*th* element of $\boldsymbol{\rho}_h$ is the correlation coefficient between $y_{i,t}$ and $y_{j,t-h}$:

$$\rho_{i,j}(h) = \frac{Cov[y_{i,t}, y_{j,t}]}{\sigma_{i,t}\sigma_{j,t}}. \tag{3.5}$$

Interestingly, when $h > 0$, the correlation coefficient $\rho_{i,j}(h)$ measures the **linear dependence** of $y_{i,t}$ on $y_{j,t-h}$. Similarly, $\rho_{j,i}(h)$ measures the linear dependence of $y_{j,t}$ on $y_{i,t-h}$. Finally, the diagonal element $\rho_{i,i}(h)$ is simply the lag-*h* autocorrelation coefficient of $y_{i,t}$. Notably, one has to recognize that $\rho_{j,i}(h) \neq \rho_{i,j}(h)$ for any $i \neq j$, as these coefficients measure different linear relationships. Therefore, $\boldsymbol{\Gamma}_h$ and $\boldsymbol{\rho}_h$ do not need to be symmetric. In summary, the cross-correlation matrices of a weakly stationary vector time series summarize in a compact and easy-to-use way, the following information:

- if $\rho_{i,j}(0) \neq 0$, $y_{i,t}$ and $y_{j,t}$ are *contemporaneously linearly correlated*;
- if $\rho_{i,j}(h) = \rho_{j,i}(h) = 0$ for all $h \geq 0$, then $y_{i,t}$ and $y_{j,t}$ share no linear relationship;
- if $\rho_{i,j}(h) = 0$ and $\rho_{j,i}(h) = 0$ for all $h > 0$, then $y_{i,t}$ and $y_{j,t}$ are said to be linearly *uncoupled*;
- if $\rho_{i,j}(h) = 0$ for all $h > 0$, but $\rho_{j,i}(q) \neq 0$ for at least some $q > 0$, then there is a *unidirectional (linear) relationship* between $y_{i,t}$ and $y_{j,t}$ where $y_{i,t}$ does not depend on $y_{j,t}$, but $y_{j,t}$ depends on (some) lagged values of $y_{i,t}$;

- if $\rho_{i,j}(h) \neq 0$ for at least some $h > 0$, and $\rho_{j,i}(q) \neq 0$ for at least some $q > 0$ then there is a *linear feedback relationship* between $y_{i,t}$ and $y_{j,t}$.

The concepts of unidirectional vs. feedback linear relationships among variables will be further developed in the so-called Granger-Sims causality tests (see Chapter 3.3)

1.3 Sample Cross-Covariance and Cross-Correlation Matrices

Now that we have discussed what cross-covariance and cross-correlation matrices are, we are ready to discuss how they can be computed in practice from the data. In fact, as we already know from Chapter 2, we only observe empirical realizations of a time series and thus we can only compute **sample cross-covariances** and **cross-correlations**, which (under some conditions) will provide consistent but biased estimates of their true, unobserved counterparts (see Fuller, 1976, for a technical discussion of the asymptotic properties of sample cross-covariances and cross-correlations).

Given a sample $\{\mathbf{y}_t \mid t = 1,...,T\}$, the cross-covariance matrix can be estimated by

$$\hat{\boldsymbol{\Gamma}}_h = \frac{1}{T}\sum_{t=h+1}^{T}\left(\mathbf{y}_t - \bar{\mathbf{y}}\right)\left(\mathbf{y}_{t-h} - \bar{\mathbf{y}}\right) \text{ with } h \geq 0, \quad (3.6)$$

where $\bar{\mathbf{y}}$ is the vector of sample means, i.e., $\bar{\mathbf{y}} = \left[\bar{y}_1, \bar{y}_2, ..., \bar{y}_N\right]'$ and $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{i,t}$ with $i = 1,...,N$. The cross-correlation matrix can be then estimated as

$$\hat{\boldsymbol{\rho}}_h = \hat{\mathbf{D}}^{-1}\hat{\boldsymbol{\Gamma}}_h\hat{\mathbf{D}}^{-1}, \text{ with } h \geq 0, \quad (3.7)$$

where $\hat{\mathbf{D}}$ is the $N \times N$ diagonal matrix of the sample standard deviations of each of the component series.

Example 3.1. Consider the weekly yields of US one-month Treasury bills and ten-year Treasury bonds, for the sample January 1990 - December 2016, as plotted in Figure 3.1.

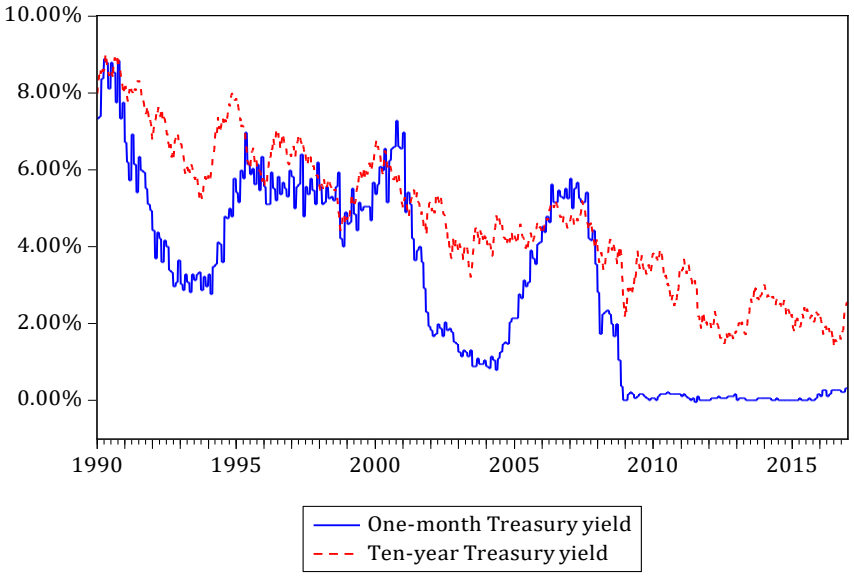

Figure 3.1 –Plot of weekly yields for one-month and ten-year U.S. Treasury bond

These yields form a bivariate time series $\mathbf{y}_t = \left[y_{1,t}, y_{2,t} \right]'$, where $y_{1,t}$ is the one-month Treasury bill yield and $y_{2,t}$ is the ten-year yield. First, we compute the vector of sample means of the series and the contemporaneous correlation matrix, which are reported in Table 3.1. All the values reported in Table 3.1, with the exceptions of the correlation coefficient (which is by construction pure numbers, i.e., without a scale), are percentages (e.g., 3.04 should be read as 3.04%). It is easy to see that the two series have a high contemporaneous correlation coefficient, $\rho_{1,2}(0) = 0.87$ and thus they are concurrently linearly correlated. However, cross-correlations at different lags can give us additional useful information about the dynamic relationship between the series.

	Mean	Standard Deviation	Skewness	Kurtosis	Minimum	Maximum
One-month Treasury yield	3.04	2.51	0.18	1.75	-0.05	8.89
Ten-year Treasury yield	4.74	1.89	0.14	2.15	1.38	9.02
(b) Correlation Matrix						
	One-month Treasury yield			Ten-year Treasury yield		
One-month Treasury yield	1					
Ten-year Treasury yield	0.87			1		

Table 3.1 – Descriptive statistics of the one-month and ten-year U.S. Treasury yield series

Table 3.2 shows the cross-correlations between the series. In particular, the first set of bins (in the first column) shows the correlations between the one-month Treasury yield and the lagged values of ten-year Treasury yields (for increasing lags, *h*). The set of bins in the second column shows the correlation between the one-month Treasury yield and the leading values of the ten-year Treasury yield, which are equivalent (because of the definition of stationarity) to the correlation between ten-year Treasury yields and lagged values of one-month bill yields (for increasing lags, *h*). According to the definition given above, the two series display a strong feedback relationship, as both $\rho_{i,j}(h) \neq 0$ and $\rho_{j,i}(q) \neq 0$ hold.

One-month yield,ten-year yield(-h)	One-month yield,ten-year yield(+h)	h	lag	lead
		0	0.8681	0.8681
		1	0.8672	0.8657
		2	0.8663	0.8631
		3	0.8653	0.8606
		4	0.8641	0.8581
		5	0.8627	0.8556
		6	0.8610	0.8530
		7	0.8590	0.8504
		8	0.8567	0.8476
		9	0.8537	0.8447
		10	0.8505	0.8416
		11	0.8474	0.8387
		12	0.8442	0.8360
		13	0.8411	0.8333
		14	0.8377	0.8304
		15	0.8342	0.8275
		16	0.8307	0.8242
		17	0.8272	0.8211
		18	0.8232	0.8182
		19	0.8191	0.8156
		20	0.8150	0.8131
		21	0.8109	0.8104
		22	0.8072	0.8075
		23	0.8035	0.8047
		24	0.7999	0.8019

Table 3.2 – Sample cross-correlations between one-month and ten-year Treasury yields

1.4 Multivariate Portmanteau Tests

In Chapter 2, we have introduced the Ljung and Box's (1978) Q-statistic to jointly test whether several (*m*) consecutive autocorrelation coefficients were equal to zero. As far as multivariate time series are concerned, we are interested in testing whether there are both no auto- and cross-correlations in a vector series $\mathbf{y}_t$. A simple, multivariate version of the Ljung-Box statistic

to test, the null hypothesis $H_0 : \boldsymbol{\rho}_1 = ... = \boldsymbol{\rho}_m = 0$ versus the alternative hypothesis $H_1 : \boldsymbol{\rho}_i \neq 0$ for some $i \in \{1,...,m\}$ is

$$Q(m) = T^2 \sum_{h=1}^{m} \frac{1}{T-h} tr\left(\hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Gamma}}_0^{-1} \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Gamma}}_0^{-1}\right), \qquad (3.8)$$

where T is the sample size, N is the dimension of $\mathbf{y}_t$, m is the maximum lag length that we wish to test and $tr(\mathbf{A})$ is the **trace** of some matrix $\mathbf{A}$, simply defined as the sum of the diagonal elements of $\mathbf{A}$. Under the null hypothesis, $Q(m)$ is asymptotically distributed as a χ^2 distribution with $N^2 m$ degrees of freedom. For practical purposes, it is important to note that the χ^2 approximation to the distribution of the test statistic may be misleading for small values of *m*. In addition, not knowing the small sample distribution is clearly a shortcoming, because infinite samples are not available. Using Monte Carlo techniques, it was found that in small samples the nominal size of the portmanteau test tends to be lower than the significance level chosen (see, e.g., Hosking, 1980). Moreover, the test has low power against many alternatives.

To overcome this drawback, both Hosking (1980, 1981) and Li and McLeod (1981) have proposed adjusted versions of the multivariate Ljung-Box statistic that, despite being asymptotically equivalent to the original one, have better finite sample performance. The test statistic proposed by Hosking (1980) has the expression

$$Q^*(m) = T(T+2) \sum_{h=1}^{m} \frac{1}{T-h} tr\left(\hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Gamma}}_0^{-1} \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Gamma}}_0^{-1}\right), \quad (3.9)$$

while the test statistic proposed by Li and McLeod (1981) is instead

$$Q^{**}(m) = T \sum_{h=1}^{m} \frac{1}{T-h} tr\left(\hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Gamma}}_0^{-1} \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Gamma}}_0^{-1}\right) + \frac{N^2 m(m+1)}{2T}.$$

Both Li and McLeod (1981) and Hosking (1981) provided simulation experiments to demonstrate the improvement of their suggested modified portmanteau test with respect to the original multivariate version of Ljung-Box statistic. Li (2004) has noted that a comparison of these two modified tests with the original one shows that both modifications work equally well and were better

than the original multivariate portmanteau test. Kheoh and McLeod (1992) have suggested that the variance of the Li-McLeod modified portmanteau test is less than (3.8).

1.5 Multivariate White Noise Process

Before we move on to the introduction of vector autoregressive models, we introduce the concept of multivariate white noise, which will be useful in the rest of the chapter to define a few classes of multivariate models.

Definition 3.2. **(Multivariate White Noise)** Let $\boldsymbol{z}_t = [z_{1,t}, z_{2,t}, \ldots z_{N,t}]$ be a $N \times 1$ vector of random variables. This multivariate time series is said to be a **multivariate white noise** if it is a stationary vector with zero mean, and if the values of $\mathbf{z}_t$ at different times are uncorrelated, i.e., $\boldsymbol{\Gamma}_h$ is an $N \times N$ matrix of zeros at all $h \neq 0$.

Definition 3.2 implies that each component of $\mathbf{z}_t$ simply behaves like a univariate white noise; additionally, the individual white noises are uncoupled in a linear sense. It is important to understand that the assumption that the values of $\mathbf{z}_t$ are uncorrelated does not necessarily imply that they are independent (while we know that independence implies zero correlation, see the Mathematical and Statistical Appendix at the end of the book). However, independence can be inferred by the lack of correlations at all leads and lags among the random variables that enter $\mathbf{z}_t$, when the random vector follows a multivariate normal distribution.

2- Introduction to VAR Analysis

2.1 From Structural to Reduced-Form VARs

Vector autoregressive (VAR) models are a natural generalization of the univariate AR model already discussed in Chapter 2. In practice, a VAR is a system regression model that treats all the variables as endogenous and allows each of them to depend on *p* lagged values of itself and of all the other variables in the system. Formally, a VAR(*p*) model can be defined as follows.

Definition 3.3. (Vector Autoregressive Model) A Vector Autoregressive model of order *p* (in short VAR(p)) is a process that can be represented as

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + ... + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t = \mathbf{a}_0 + \sum_{j=1}^{p}\mathbf{A}_j\mathbf{y}_{t-j} + \mathbf{u}_t,$$

where $\mathbf{y}_t$ is a $N \times 1$ vector containing *N* endogenous variables, $\mathbf{a}_0$ is a $N \times 1$ vector of constants, $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_p$ are the p $N \times N$ matrices of autoregressive coefficients, and $\mathbf{u}_t$ is a $N \times 1$ vector of uncorrelated, white noise disturbances.

In order to help the Reader familiarize with the concepts, we start our discussion introducing a bivariate VAR(1) model, while in Section 2.3 we generalize it to a VAR(*p*) model with *N* endogenous variables (hence, equations). Consider the following bivariate, first-order Markovian system

$$y_{1,t} = b_{1,0} - b_{1,2}y_{2,t} + \varphi_{1,1}y_{1,t-1} + \varphi_{1,2}y_{2,t-1} + \varepsilon_{1,t} \quad (3.12)$$

$$y_{2,t} = b_{2,0} - b_{2,1}y_{1,t} + \varphi_{2,1}y_{1,t-1} + \varphi_{2,2}y_{2,t-1} + \varepsilon_{2,t} \quad (3.13)$$

where both the variables $y_{1,t}$ and $y_{2,t}$ are assumed to be stationary and the **structural error terms** $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are uncorrelated white-noise disturbances with standard deviation σ_1 and σ_2, respectively. The system in (3.12) - (3.13) can also be rewritten in a more compact form using matrix notation:

$$\begin{bmatrix} 1 & b_{1,2} \\ b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix} + \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix}\begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \quad (3.14)$$

or,

$$\mathbf{B}\mathbf{y}_t = \mathbf{Q}_0 + \mathbf{Q}_1\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (3.15)$$

where

$$\mathbf{B} \equiv \begin{bmatrix} 1 & b_{1,2} \\ b_{2,1} & 1 \end{bmatrix}, \quad \mathbf{y}_t \equiv \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix}, \quad \mathbf{Q}_0 = \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix}, \quad \mathbf{Q}_1 = \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix},$$

$$\mathbf{y}_{t-1} \equiv \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} \text{ and } \boldsymbol{\varepsilon}_t \equiv \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}.$$

In this system, that is also known as a **structural VAR** (or **VAR in primitive form**), $y_{1,t}$ depends on its own lag and on one lag of $y_{2,t}$, but also on the current value of $y_{2,t}$; similarly, $y_{2,t}$ depends on its own lag and on one lag of $y_{1,t}$, but also on the current value of $y_{1,t}$. Therefore, a VAR in its structural form captures **contemporaneous feedback effects**: $-b_{1,2}$ measures the contemporaneous effect of a unit change of $y_{2,t}$ on $y_{1,t}$ and $-b_{2,1}$ measures the contemporaneous effect of a unit change of $y_{1,t}$ on $y_{2,t}$.

Unfortunately, structural VARs are not very practical for applied purposes because standard estimation techniques require the regressors to be uncorrelated with the error terms, which is clearly not the case of the VAR in its structural form. This is due to the presence of contemporaneous feedback effects: obviously, each contemporaneous variable is correlated with its own error term. From (3.12) and (3.13), it is clear that from the first equation, when $-b_{1,2}$ is non-zero, $y_{1,t}$ depends on $y_{2,t}$ from the second equation and therefore on $\varepsilon_{2,t}$, and it will be correlated with it; from the second equation, when $-b_{2,1}$ is non-zero, $y_{2,t}$ depends on $y_{1,t}$ from the first equation and therefore on $\varepsilon_{1,t}$. As an additional drawback of the structural model, contemporaneous terms cannot be used in forecasting, i.e., exactly where VAR models tend to be largely popular. As a result, in time series analysis, it is common to manipulate the VAR in its structural form to make it more directly useful. Pre-multiplying both sides of (3.15) by $\mathbf{B}^{-1}$ we obtain

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{u}_t, \tag{3.16}$$

where $\mathbf{a}_0 = \mathbf{B}^{-1}\mathbf{Q}_0$, $\mathbf{A}_1 = \mathbf{B}^{-1}\mathbf{Q}_1$ and $\mathbf{u}_t = \mathbf{B}_t^{-1}\boldsymbol{\varepsilon}_t$. Denoting by $a_{i,0}$ the element in row *i* of the vector $\mathbf{a}_0$, by $a_{i,j}$ the element in row *i* and column *j* of the matrix $\mathbf{A}_1$, and by $u_{i,t}$ the element in row *i* of the vector $\mathbf{u}_t$, we can rewrite (3.16) in the equivalent form:

$$y_{1,t} = a_{1,0} + a_{1,1} y_{1,t-1} + a_{1,2} y_{2,t-1} + u_{1,t} \tag{3.17}$$

$$y_{2,t} = a_{2,0} + a_{2,1} y_{1,t-1} + a_{2,2} y_{2,t-1} + u_{2,t}. \quad (3.18)$$

This system is called **reduced-form** VAR or, alternatively, it is said to describe a VAR in its **standard form**. The model in (3.16) only features *lagged* endogenous variables (i.e., it does not contain contemporaneous feedback terms) and it can be estimated equation by equation using OLS (as we shall see in detail in Section 2.4). Clearly, the new, **reduced-form error terms**, $u_{1,t}$ and $u_{2,t}$, are composites of the two original (also called pure or structural) shocks $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$. This is easy to see if we solve $\mathbf{u}_t = \mathbf{B}^{-1}\boldsymbol{\varepsilon}_t$ to get:

$$u_{1,t} = \frac{\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t}}{1 - b_{1,2}b_{2,1}} \quad (3.19)$$

$$u_{2,t} = \frac{\varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}}{1 - b_{1,2}b_{2,1}}. \quad (3.20)$$

Recalling that $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are white noise processes, we can easily derive the properties of the reduced form errors $u_{1,t}$ and $u_{2,t}$. First, taking the expected value of (3.19) and (3.20) (and recalling that, based on the definition of a white noise, $E\left[\varepsilon_{1,t}\right] = 0$ and $E\left[\varepsilon_{2,t}\right] = 0$), we obtain that

$$E\left[u_{1,t}\right] = E\left[\frac{\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t}}{1 - b_{1,2}b_{2,1}}\right] = 0 \quad (3.21)$$

$$E\left[u_{2,t}\right] = E\left[\frac{\varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}}{1 - b_{1,2}b_{2,1}}\right] = 0. \quad (3.22)$$

In addition, because $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are uncorrelated, i.e., $Cov\left[\varepsilon_{1,t}, \varepsilon_{2,t}\right] = 0$, we find that the variance of $u_{1,t}$ is

$$Var\left[u_{1,t}\right] = \frac{Var\left[\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t}\right]}{\left(1 - b_{1,2}b_{2,1}\right)^2} = \frac{Var\left[\varepsilon_{1,t}\right] + b_{1,2}^2 Var\left[\varepsilon_{2,t}\right] - 2b_{1,2}Cov\left[\varepsilon_{1,t}, \varepsilon_{2,t}\right]}{\left(1 - b_{1,2}b_{2,1}\right)^2}$$
$$= \frac{\sigma_{\varepsilon,1}^2 + b_{1,2}^2\sigma_{\varepsilon,2}^2}{\left(1 - b_{1,2}b_{2,1}\right)^2} \quad (3.23)$$

and, similarly,

$$Var\left[u_{2,t}\right]=\frac{\sigma_{\varepsilon,2}^2+b_{2,1}^2\sigma_{\varepsilon,1}^2}{\left(1-b_{1,2}b_{2,1}\right)^2}. \tag{3.24}$$

It is easy to see that the variances of $u_{1,t}$ and $u_{2,t}$ are constant over time. Finally, the covariance between the two structural errors is equal to

$$Cov\left[u_{1,t},u_{2,t}\right]=\frac{E\left[\left(\varepsilon_{1,t}-b_{1,2}\varepsilon_{2,t}\right)\left(\varepsilon_{2,t}-b_{2,1}\varepsilon_{1,t}\right)\right]}{\left(1-b_{1,2}b_{2,1}\right)^2}=\frac{-\left(b_{2,1}\sigma_{\varepsilon,1}^2+b_{1,2}\sigma_{\varepsilon,2}^2\right)}{\left(1-b_{1,2}b_{2,1}\right)^2} \tag{3.25}$$

Noticeably, while the reduced-form error terms remain serially uncorrelated (i.e., autocorrelations are equal to zero) as the structural errors were, they are cross-correlated unless $b_{1,2}=b_{2,1}=0$ (i.e., there are no contemporaneous effects of $y_{1,t}$ on $y_{2,t}$ and vice versa). The variances and covariances of the reduced-form errors can be collected in the matrix $\boldsymbol{\Sigma}_u$:

$$\boldsymbol{\Sigma}_u=\begin{bmatrix} Var\left[u_{1,t}\right] & Cov\left[u_{1,t},u_{2,t}\right] \\ Cov\left[u_{1,t},u_{2,t}\right] & Var\left[u_{2,t}\right] \end{bmatrix}=\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}.$$

The reduced form VAR in (3.17)-(3.18) is very practical and easy to estimate (this can be done by simple OLS), but it is important to understand that, *in general*, it is not possible to **identify** the structural parameters and errors (i.e., the sample estimates of the coefficients and the residuals of the primitive system) from the OLS estimates of the parameters and the residuals of the standard form VAR. This lack of identification (because the model is linear, the problem is both local and global, see Chapter 8 for a differentiation of the two concepts) may be overcome if one is prepared to impose appropriate restrictions on the primitive system. This is unsurprising: the structural VAR in (3.12)-(3.13) contains eight coefficients and two variances of the error terms, for a total of ten parameters; the VAR in its standard form only contains nine parameters (six coefficients, two variances and one covariance of the error terms). Therefore, and this occurs for a rather intuitive accounting, back-of-the-envelope reason, it is not possible to recover all the information that was present in the primitive system unless we are able to restrict one of its parameters. To this

purpose, a popular identification scheme is the one proposed by Sims (1980), based on a **recursive Choleski triangularization**.
Suppose that you are willing to impose a restriction on the primitive system in (3.12)-(3.13) such that $b_{1,2}$ is equal to zero, meaning that $y_{1,t}$ has a contemporaneous effect on $y_{2,t}$, but $y_{2,t}$ only affects $y_{1,t}$ with a one period lag:

$$y_{1,t} = b_{1,0} + \varphi_{1,1} y_{1,t-1} + \varphi_{1,2} y_{2,t-1} + \varepsilon_{1,t} \quad (3.27)$$

$$y_{2,t} = b_{2,0} - b_{2,1} y_{1,t} + \varphi_{2,1} y_{1,t-1} + \varphi_{2,2} y_{2,t-1} + \varepsilon_{2,t} \quad .(3.28)$$

This corresponds to imposing a Choleski decomposition on the covariance matrix of the residuals of the VAR in its standard form. Indeed, now we can re-write the relationship between the pure shocks (from the structural VAR) and the regression residuals as

$$u_{1,t} = \varepsilon_{1,t} \quad (3.29)$$

$$u_{2,t} = \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t} \, . \quad (3.30)$$

Practically, imposing the restriction $b_{1,2} = 0$ means that $\mathbf{B}^{-1}$ is given by

$$\mathbf{B}^{-1} \equiv \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix},$$

and thus, pre-multiplication of the primitive system (3.12)-(3.13) by the lower diagonal matrix $\mathbf{B}^{-1}$ yields

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \quad (3.31)$$

which results in

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} b_{1,0} \\ b_{2,0} - b_{1,0} b_{2,1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} - b_{2,1}\varphi_{1,1} & \varphi_{2,2} - b_{2,1}\varphi_{1,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t} \end{bmatrix} \quad (3.32)$$

.

The system has now only nine parameters that can be identified using the OLS estimates from (3.17)–(3.18). Indeed, using simple algebra we can see that: $a_{1,0} = b_{1,0}$; $a_{2,0} = b_{2,0} - b_{1,0} b_{2,1}$; $a_{1,1} = \varphi_{1,1}$; $a_{1,2} = \varphi_{1,2}$; $a_{2,1} = \varphi_{2,1} - b_{2,1}\varphi_{1,1}$; $a_{2,2} = \varphi_{2,2} - b_{2,1}\varphi_{1,2}$. In

addition, since we know from (3.29)-(3.30) that $u_{1,t} = \varepsilon_{1,t}$ and $u_{2,t} = \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}$, we can compute:

$$\sigma_1^2 \equiv Var\left[u_{1,t}\right] = \sigma_{\varepsilon,1}^2, \tag{3.33}$$

$$\sigma_2^2 \equiv Var\left[u_{2,t}\right] = \sigma_{\varepsilon,2}^2 + b_{2,1}^2\sigma_{\varepsilon,1}^2, \tag{3.34}$$

$$Cov\left[u_{1,t}, u_{2,t}\right] = -b_{2,1}\sigma_{\varepsilon,1}^2. \tag{3.35}$$

The implication of the identification restriction that we just imposed is that, while both the $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ shocks affect the contemporaneous value of $y_{2,t}$, only $\varepsilon_{1,t}$ impacts the contemporaneous value of $y_{1,t}$. In practice, the observed values of $u_{1,t}$ are completely attributed to pure (structural) shocks to $y_{1,t}$. This technique of decomposing the residuals in a triangular fashion is indeed called Choleski decomposition (or triangularization). Put in other words, we see that the covariance matrix of the residuals is forced to be equal to

$$\mathbf{\Sigma}_u = \mathbf{W}\mathbf{\Sigma}\mathbf{W}' = \mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}^{1/2})', \tag{3.36}$$

where $\mathbf{W} = \mathbf{B}^{-1}$, $\mathbf{\Sigma}$ is the diagonal covariance matrix of the structural innovations, and $\mathbf{\Sigma}^{1/2}$ is the triangular "square root" of the covariance matrix $\mathbf{\Sigma}_u$. Equation (3.36) is easily checked:

$$\begin{aligned}\mathbf{\Sigma}_u &= \begin{bmatrix}1 & 0\\ -b_{2,1} & 1\end{bmatrix}\begin{bmatrix}\sigma_{\varepsilon,1}^2 & 0\\ 0 & \sigma_{\varepsilon,2}^2\end{bmatrix}\begin{bmatrix}1 & 0\\ -b_{2,1} & 1\end{bmatrix}'\\ &= \begin{bmatrix}1 & 0\\ -b_{2,1} & 1\end{bmatrix}\begin{bmatrix}\sigma_{\varepsilon,1}^2 & 0\\ 0 & \sigma_{\varepsilon,2}^2\end{bmatrix}\begin{bmatrix}1 & -b_{2,1}\\ 0 & 1\end{bmatrix}\\ &= \begin{bmatrix}\sigma_{\varepsilon,1}^2 & -b_{2,1}\sigma_{\varepsilon,1}^2\\ -b_{2,1}\sigma_{\varepsilon,1}^2 & \sigma_{\varepsilon,2}^2 + b_{2,1}\sigma_{\varepsilon,1}^2\end{bmatrix}\end{aligned}, \tag{3.37}$$

which is exactly what we found in equations (3.33)-(3.35). The decomposition in (3.36) is what we call the Choleski decomposition of the symmetric matrix $\mathbf{\Sigma}_u$. Needless to say, the task that one usually wants to accomplish is to go back from the estimated $\mathbf{\Sigma}_u$ to the original (and unobserved) diagonal matrix $\mathbf{\Sigma}$. With a little bit of algebra, we understand that this is equivalent to

$$\mathbf{\Sigma} = \mathbf{W}^{-1}\mathbf{\Sigma}_u(\mathbf{W}^{'})^{-1}. \tag{3.38}$$

This technique can be generalized to a VAR system with any number N of equations. In particular, in a N-variate VAR, exact identification requires us to impose $(N^2 - N)/2$ restrictions in order to retrieve the N structural shocks from the residual of the OLS estimate. Being based on a triangular structure, a Choleski decomposition forces exactly $(N^2 - N)/2$ values of the matrix $\mathbf{B}$ to be zero (or to some other constant).

Let us pause for a moment to understand the meaning (and the implications) of the Choleski decomposition for a less simplistic model, for instance a VAR(1) with three endogenous variables (and therefore three equations). The parameters in the structural model consist of three intercept terms, six (two for each equation) coefficients that map the contemporaneous effect of each variable on the other two, nine autoregressive coefficients (contained in a 3×3 matrix) and the three variance coefficients of the error terms, for a total of 21 parameters. The VAR in its reduced form contains 12 estimated coefficients (three intercepts and nine autoregressive coefficients), three variances and three covariances, for a total of 18 coefficients. Therefore, we shall need to impose three restrictions to identify the parameters of the primitive system from the OLS estimates of the VAR in its standard form, which is exactly $(3^2-3)/2=3$ restrictions. Indeed, imposing a triangular (Choleski) decomposition on the structural residuals is equivalent to pre-multiplying the structural VAR by the lower triangular matrix

$$\boldsymbol{B}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -b_{2,1} & 1 & 0 \\ b_{2,1}b_{3,2} - b_{3,1} & 1 & 1 \end{bmatrix}, \tag{3.39}$$

which yields the reduced form residuals:

$$\boldsymbol{u}_t = \boldsymbol{B}^{-1}\boldsymbol{\varepsilon}_t = \begin{bmatrix} 1 & 0 & 0 \\ -b_{2,1} & 1 & 0 \\ b_{2,1}b_{3,2} & -b_{3,1} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} =$$

$$= \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t} \\ \varepsilon_{1,t}(b_{2,1}b_{3,2} - b_{3,1}) - b_{3,1}\varepsilon_{2,t} + \varepsilon_{3,t} \end{bmatrix} \tag{3.39}$$

Because the Choleski decomposition is based on pre-multiplying by a (lower) triangular matrix, it follows that when we decide the ordering of the variables in a VAR system, we are also deciding which kind of restrictions the decomposition will impose on the contemporaneous effects of each variable on the others. For example, in the tri-variate case of (3.39) above, $b_{1,2}, b_{1,3}$, and $b_{2,3}$ are set to zero, meaning that the first variable in the system is forced not to be contemporaneously affected by shocks to any of the other variables; the second variable in the system is only contemporaneously affected by shocks to the first variable; the last variable is contemporaneously affected by the shocks to both the other variables. It is easy to generalize this reasoning to the *N*-variable case.

It should be evident that there are as many Choleski decompositions as all the possible orderings of the variables, which are therefore a combinatorial factor of *N*. Therefore, we shall need to be aware that any time that we apply a Choleski triangular identification scheme to a VAR model that results in a specific ordering, we will be introducing a number of (potentially arbitrary) assumptions on the contemporaneous relationships among the variables. Therefore, despite being very practical, Choleski decompositions are quite deliberate in the restrictions that they place and tend not to be based on any theoretical assumptions regarding the nature of the economic relationships among the variables. Alternative identification schemes are possible (although they are more popular in the macroeconomics literature than in applied finance). A review of some commonly used restriction schemes to achieve identification based on a theoretical background can be found in Lütkepohl (2005, Chapter 9).

2.2 Stationarity Conditions and the Population Moments of a VAR(1) Process

Let us now discuss the properties of a reduced-form, standard VAR(1) model such as the one in (3.16). Assume that $\mathbf{y}_t$, $\mathbf{a}_0$, $\mathbf{y}_{t-1}$ and $\mathbf{u}_t$ are *N x 1* vectors and $\mathbf{A}_1$ is a $N \times N$ matrix and that the process is weakly stationary, according to Definition 3.1. By taking the expectation of $\mathbf{y}_t$ and using the fact that $E\left[\mathbf{u}_t\right] = \mathbf{0}$, we obtain:

$$E\left[\mathbf{y}_{\mathrm{t}}\right]=\mathbf{a}_0+\mathbf{A}_1E\left[\mathbf{y}_{\mathrm{t\text{-}1}}\right]. \tag{3.40}$$

Because we are assuming stationarity, $E\left[\mathbf{y}_t\right]$ is time-invariant so that $E\left[\mathbf{y}_t\right]=E\left[\mathbf{y}_{t-1}\right]$ and thus

$$\boldsymbol{\mu}\equiv E\left[\mathbf{y}_t\right]=\left(\mathbf{I}_N-\mathbf{A}_1\right)^{-1}\mathbf{a}_0, \tag{3.41}$$

provided that the matrix $\mathbf{I}_N-\mathbf{A}_1$ is non-singular, where $\mathbf{I}_N$ is the $N\times N$ identity matrix. Clearly, the unconditional mean vector $\boldsymbol{\mu}$ in (3.42) must be contrasted with the conditional mean vector:

$$\boldsymbol{\mu}_{t|t-1}\equiv E\left[\mathbf{y}_t\,|\,\Im_{t-1}\right]=E\left[\mathbf{y}_t\,|\,\mathbf{y}_{t-1}\right]=\mathbf{a}_0+\mathbf{A}_1\mathbf{y}_{t-1}. \tag{3.42}$$

Therefore, using $\mathbf{a}_0=\left(\mathbf{I}_N-\mathbf{A}_1\right)\boldsymbol{\mu}$, the VAR(1) model can be rewritten as

$$\mathbf{y}_t-\boldsymbol{\mu}=\mathbf{A}_1(\mathbf{y}_{t-1}-\boldsymbol{\mu})+\mathbf{u}_t. \tag{3.43}$$

If we let $\tilde{\mathbf{y}}_{\mathrm{t}}\equiv\mathbf{y}_{\mathrm{t}}\text{-}\boldsymbol{\mu}$ be the mean-corrected time-series, or equivalently the vector process expressed in deviations from its unconditional mean, we can write the model as:

$$\tilde{\mathbf{y}}_{\mathrm{t}}=\mathbf{A}_1\tilde{\mathbf{y}}_{\mathrm{t\text{-}1}}+\mathbf{u}_{\mathrm{t}}. \tag{3.44}$$

Clearly, it is possible to substitute $\tilde{\mathbf{y}}_{\mathrm{t\text{-}1}}=\mathbf{A}_1\tilde{\mathbf{y}}_{\mathrm{t\text{-}2}}+\mathbf{u}_{\mathrm{t\text{-}1}}$ in the expression (3.45), obtaining

$$\tilde{\mathbf{y}}_t=\mathbf{A}_1\left(\mathbf{A}_1\tilde{\mathbf{y}}_{t-2}+\mathbf{u}_{t-1}\right)+\mathbf{u}_t=\mathbf{A}_1^2\tilde{\mathbf{y}}_{t-2}+\mathbf{A}_1\mathbf{u}_{t-1}+\mathbf{u}_t. \tag{3.45}$$

We can now substitute $\tilde{\mathbf{y}}_{t-2}=\mathbf{A}_1\tilde{\mathbf{y}}_{t-3}+\mathbf{u}_{t-2}$ in the expression (3.46), and then keep iterating till we obtain:

$$\tilde{\mathbf{y}}_t=\mathbf{u}_t+\mathbf{A}_1\mathbf{u}_{t-1}+\mathbf{A}_1^2\mathbf{u}_{t-2}+\mathbf{A}_1^3\mathbf{u}_{t-3}+...=\sum_{i=1}^{\infty}\mathbf{A}_1^i\mathbf{u}_{t-i}+\mathbf{u}_t. \tag{3.46}$$

Notice that $\tilde{\mathbf{y}}_t=\mathbf{y}_t-\boldsymbol{\mu}$, so that (3.47) can also be re-written as

$$\mathbf{y}_t=\boldsymbol{\mu}+\sum_{i=1}^{\infty}\mathbf{A}_1^i\mathbf{u}_{t-i}+\mathbf{u}_t. \tag{3.47}$$

If we define $\boldsymbol{\Theta}_i\equiv\mathbf{A}_1^i$, we can rewrite (3.48) as

$$\mathbf{y}_t=\boldsymbol{\mu}+\sum_{i=1}^{\infty}\boldsymbol{\Theta}_i\mathbf{u}_{t-i}+\mathbf{u}_t, \tag{3.48}$$

which is the **vector moving average (VMA) infinite representation** of the VAR(1) model and that it is immediately

useful to discuss its properties. First, because $\mathbf{u}_t$ is serially uncorrelated, it is also uncorrelated with the past values of $\mathbf{y}_t$, i.e., $Cov[\mathbf{u}_t, \mathbf{y}_{t-1}] = \mathbf{0}$. For this reason, $\mathbf{u}_t$ is often referred to as the vector of **innovations** of the series at time *t*. Second, post-multiplying the expression (3.48) by $\mathbf{u}'_t$, taking the expectation, and exploiting again the fact that $\mathbf{u}_t$ is serially uncorrelated, we obtain $Cov\left[\mathbf{y}_t, \mathbf{u}_t\right] = \mathbf{\Sigma}_u$. Third, (3.48) implies that in a VAR(1) model, $\mathbf{y}_t$ depends on the past innovations $\mathbf{u}_{t-j}$ with a coefficient matrix $\mathbf{A}_1^j$, i.e., with coefficients that are collected in increasing powers of the VAR(1) matrix. For such dependence to fade progressively away as the time distance between $\mathbf{y}_t$ and past innovations grows—which seems to be a sensible condition, in the sense that in the VAR(1) model past shocks are gradually forgotten in a typical geometric decaying fashion— $\mathbf{A}_1^j$ must converge to zero as *j* goes to infinity. In practice, this means that all the *N* eigenvalues of the matrix $\mathbf{A}_1$ must be less than 1 in modulus, in order to avoid that $\mathbf{A}_1^j$ will either explode or converge to a nonzero matrix as *j* goes to infinity. Therefore, provided that the covariance matrix of $\mathbf{u}_t$ exists, the requirement that all the eigenvalues of $\mathbf{A}_1$ are less than one in modulus is a necessary and sufficient condition for $\mathbf{y}_t$ to be **stable** (and, thus, **stationary**, as stability implies stationarity as discussed in Chapter 2), that is:

$$\det(\mathbf{I}_N - \mathbf{A}_1 z) \neq 0 \text{, for } |z| \leq 1.^1 \qquad (3.49)$$

Of course, you will recognize that under (3.48), (3.49) represents the multivariate extension of the **Wold's representation** theorem already stated in Chapter 2 for univariate stationary time series. Finally, using expression (3.48), we have that

---

[1] The condition in (3.50) is simply an alternative way to state that all the eigenvalues of the matrix A must be less than one in modulus. In fact, all the eigenvalues of matrix $\mathbf{A}_1$ are less than one in modulus if and only if the polynomial $\det(\mathbf{I}_N - \mathbf{A}_1 z)$ has no roots in and on the complex unit circle.

$$Cov\left[\mathbf{y}_t\right] \equiv \Gamma_0 = \mathbf{\Sigma}_u + \mathbf{A}_1\mathbf{\Sigma}_u\mathbf{A}'_1 + \mathbf{A}_1^2\mathbf{\Sigma}_u\left(\mathbf{A}_1^2\right)' + ... = \sum_{i=0}^{\infty} A_1^i\mathbf{\Sigma}_u\left(\mathbf{A}_1^i\right)',$$

where $\mathbf{A}_1^0$ is a $N \times N$ identity matrix $\mathbf{I}_N$. Also in this case, this is to be contrasted with the conditional covariance matrix for $\mathbf{y}_t$:

$$Cov[\mathbf{y}_t \mid \Im_{t-1}] = Cov[\mathbf{y}_t \mid \mathbf{y}_{t-1}] = \mathbf{A}_1 Cov[\mathbf{y}_{t-1} \mid \mathbf{y}_{t-1}]\mathbf{A}'_1 + \mathbf{\Sigma}_u = \mathbf{\Sigma}_u,$$

because $Cov[\mathbf{y}_{t-1} \mid \mathbf{y}_{t-1}] = O$. Interestingly, while the *unconditional* covariance matrix is a complex function of both the covariance matrix of the residuals, $\mathbf{\Sigma}_u$, and of the matrix of vector autoregressive coefficients $\mathbf{A}_1$, conditioning on past information, the covariance matrix of $\mathbf{y}_t$ is the same as the covariance matrix of the residuals, $\mathbf{\Sigma}_u$; therefore, when the residuals are simultaneously uncorrelated (i.e., $\mathbf{\Sigma}_u$ is diagonal), then also $Cov[\mathbf{y}_t \mid \mathbf{y}_{t-1}]$ will be diagonal.

To find a more useful expression in place of (3.51), note that it can alternatively be written as

$$\Gamma_0 = \sum_{i=0}^{\infty} \mathbf{\Theta}_i\mathbf{\Sigma}_u\mathbf{\Theta}_i', \tag{3.52}$$

where the coefficients $\mathbf{\Theta}_i$ are simply the coefficients of the moving average representations of the VAR. This way of representing (3.51) is quite convenient because these coefficients can be easily recovered once we write the VAR(1) process in *lag operator* notation, that is,

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{A}(L)\mathbf{y}_t + \mathbf{u}_t, \tag{3.53}$$

or, alternatively,

$$\tilde{\mathbf{A}}(L)\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{u}_t, \tag{3.54}$$

where L is the lag operator discussed in Chapter 2 and $\tilde{\mathbf{A}}(L) \equiv \mathbf{I}_N - \mathbf{A}(L)$. At this point, let $\mathbf{\Theta}(L) \equiv \sum_{i=0}^{\infty} \mathbf{\Theta}_i L^i$ be an operator such that $\mathbf{\Theta}(L)\tilde{\mathbf{A}}(L) = \mathbf{I}_N$ and pre-multiply (3.55) by $\mathbf{\Theta}(L)$ to obtain

$$\mathbf{y}_t = \mathbf{\Theta}(L)\boldsymbol{\mu} + \mathbf{\Theta}(L)\mathbf{u}_t, \tag{3.55}$$

that is,

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \boldsymbol{\Theta}_i \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Theta}_i \mathbf{u}_{t-i}. \tag{3.56}$$

This means that the operator $\boldsymbol{\Theta}(L)$ is the inverse of $\tilde{\mathbf{A}}(L)$. With a modicum of additional but tedious algebra (that the interested Reader can find in Lütkepohl, 2005), it is possible to prove that

$$\boldsymbol{\Theta}_i = \sum_{j=1}^{i} \boldsymbol{\Theta}_{i-j} \mathbf{A}_1, \tag{3.57}$$

where $\boldsymbol{\Theta}_0 = \mathbf{I}_N$. Finally, post-multiplying by $\tilde{\mathbf{y}}'_{t-h}$ in equation (3.45), taking expectation, and exploiting the fact that $Cov\left[\mathbf{u}_t, \mathbf{y}_{t-j}\right] = E\left[\mathbf{u}_t \mathbf{y}'_{t-j}\right] = \mathbf{0}$ for $j > 0$, we obtain

$$E(\tilde{\mathbf{y}}_t \tilde{\mathbf{y}}'_{t+1-h}) = A_1 E(\tilde{\mathbf{y}}_t \tilde{\mathbf{y}}'_{t-h}) \quad \text{for } h > 0. \tag{3.58}$$

Therefore, the cross-covariance matrices $\boldsymbol{\Gamma}_h$ can be computed as

$$\boldsymbol{\Gamma}_h = \mathbf{A}_1 \boldsymbol{\Gamma}_{h-1} \text{ for } h > 0. \tag{3.59}$$

By repeated substitution, it is easy to show that

$$\boldsymbol{\Gamma}_h = \mathbf{A}_1^h \boldsymbol{\Gamma}_0 \text{ for } h > 0, \tag{3.60}$$

and thus, once $\boldsymbol{\Gamma}_0$ has been computed, all the other cross-covariance matrix for $h > 0$ can be calculated by recursive substitution.

Finally, by pre- and post-multiplying (3.60) by $\mathbf{D}^{-1/2}$ we can also work out the expression of the cross-correlation matrix, that is,

$$\boldsymbol{\rho}_h = \mathbf{D}^{-1/2} \mathbf{A}_1 \boldsymbol{\Gamma}_{h-1} \mathbf{D}^{-1/2} = \mathbf{D}^{-1/2} \mathbf{A}_1 \mathbf{D}^{1/2} \mathbf{D}^{-1/2} \boldsymbol{\Gamma}_{h-1} \mathbf{D}^{-1/2} = \boldsymbol{\Psi} \boldsymbol{\rho}_{h-1}, \tag{3.61}$$

where $\boldsymbol{\Psi} = \mathbf{D}^{-1/2} \mathbf{A}_1 \mathbf{D}^{-1/2}$. Again, by recursive iteration we obtain

$$\boldsymbol{\rho}_h = \boldsymbol{\Psi}^h \boldsymbol{\rho}_0 \text{ for } h > 0, \tag{3.62}$$

and thus, once $\boldsymbol{\rho}_0$ has been computed, it is trivial to obtain all the other correlation matrices.

Example 3.2. Let us suppose that we have estimated the following VAR(1) model for the one-month and the ten-year Treasury yields that were already plotted in Example 3.1. Without entering into the details of the estimation, that we shall discuss in Section 2.4 (we shall provide a complete sample output in Example 3.3), we only report the estimated coefficients (*t*-statistics are in square brackets),

$$\begin{bmatrix} y_{1M,t} \\ y_{10Y,t} \end{bmatrix} = \begin{bmatrix} \underset{[-2.5382]}{-0.0490} \\ \underset{[0.8711]}{0.0080} \end{bmatrix} + \begin{bmatrix} \underset{[210.6540]}{0.9819} & \underset{[0.4077]}{0.0209} \\ \underset{[3.3784]}{0.0009} & \underset{[240.0320]}{0.9970} \end{bmatrix} \begin{bmatrix} y_{1M,t-1} \\ y_{10Y,t-1} \end{bmatrix} + \begin{bmatrix} u_{1M,t} \\ u_{10Y,t} \end{bmatrix},$$

and the estimated covariance matrix of the reduced-form residuals is:

$$\hat{\boldsymbol{\Sigma}}_u = \begin{bmatrix} 0.0476 & 0.0013 \\ 0.0013 & 0.0110 \end{bmatrix}.$$

We also compute the unconditional first and second moments of the series. Let us start from the mean, that can be computed quite easily by applying the formula in (3.42):

$$\boldsymbol{\mu} = \left(\mathbf{I}_2 - \mathbf{A}_1\right)^{-1} \mathbf{a}_0 = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.9819 & 0.0209 \\ 0.0009 & 0.9970 \end{bmatrix} \right)^{-1} \begin{bmatrix} -0.0490 \\ 0.0080 \end{bmatrix} =$$

$$\left(\begin{bmatrix} 0.1801 & -0.0209 \\ -0.0009 & 0.0030 \end{bmatrix} \right)^{-1} \begin{bmatrix} -0.0490 \\ 0.0080 \end{bmatrix} = \begin{bmatrix} 84.53 & 588.90 \\ 25.36 & 510.00 \end{bmatrix} \begin{bmatrix} -0.0490 \\ 0.0080 \end{bmatrix} = \begin{bmatrix} 0.5692 \\ 2.8374 \end{bmatrix}$$

.

Therefore, the one-month Treasury yield has an unconditional mean of approximately 57 bps, while the ten-year Treasury yield has an unconditional mean of approximately 284 bps, which implies an average riskless yield spread of 227 bps per year. Knowing that the one-month Treasury yield on Sept. 30, 2016 was 0.16%, and the ten-year Treasury yield was 1.58%, we can also compute their **conditional** expectations:

$$\boldsymbol{\mu}_{t|09/30/16} = E\left[\mathbf{y}_t \mid \mathbf{y}_{09/30/16}\right] = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{09/30/16} = \begin{bmatrix} -0.0490 \\ 0.0080 \end{bmatrix} + \begin{bmatrix} 0.9819 & 0.0209 \\ 0.0009 & 0.9970 \end{bmatrix} \begin{bmatrix} 0.16 \\ 1.58 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1411 \\ 1.5834 \end{bmatrix}$$

For completeness, we note that, at least in hindsight, on October 7, 2016, i.e., one period (week) later, the one-month Treasury yield turned out to be 0.21% and the ten-year Treasury yield was 1.70%. The differences between the conditional expectations of the yields and their realized value, approximately 7 and 12 bps, respectively, are the forecast errors, that we shall discuss in Section 2.6.

We now compute the unconditional covariance matrix of the two series, $\boldsymbol{\Gamma}_0$:

$$\widehat{\boldsymbol{\Gamma}}_0 = \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{A}}'_1 + \widehat{\boldsymbol{\Sigma}}_u \Rightarrow vec(\widehat{\boldsymbol{\Gamma}}_0) = vec(\widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{\Gamma}}_0\widehat{\boldsymbol{A}}'_1) + vec(\widehat{\boldsymbol{\Sigma}}_u)$$
$$vec(\widehat{\boldsymbol{\Gamma}}_0) = (\widehat{\boldsymbol{A}}_1 \otimes \widehat{\boldsymbol{A}}_1)vec(\widehat{\boldsymbol{\Gamma}}_0) + vec(\widehat{\boldsymbol{\Sigma}}_u)$$
$$\Rightarrow \boldsymbol{I}_4 - (\widehat{\boldsymbol{A}}_1 \otimes \widehat{\boldsymbol{A}}_1)vec(\widehat{\boldsymbol{\Gamma}}_0) = vec(\widehat{\boldsymbol{\Sigma}}_u) \Rightarrow vec(\widehat{\boldsymbol{\Gamma}}_0) =$$
$$= \left[\boldsymbol{I}_4 - (\widehat{\boldsymbol{A}}_1 \otimes \widehat{\boldsymbol{A}}_1)\right]^{-1} vec(\widehat{\boldsymbol{\Sigma}}_u)$$

Plugging in the estimates reported above, we find:

$$vec\left(\hat{\Gamma}_0\right) = \left[\mathbf{I}_4 - \hat{\mathbf{A}}_1 \otimes \hat{\mathbf{A}}_1\right]^{-1} vec\left(\hat{\Sigma}_u\right) = \left(I_4 - \begin{bmatrix} 0.9641 & 0.0205 & 0.0205 & 0.0004 \\ 0.0009 & 0.9790 & 1.88e-05 & 0.0208 \\ 0.0009 & 1.88e-05 & 0.9790 & 0.0208 \\ 8.10e-0.7 & 0.0009 & 0.0009 & 0.9940 \end{bmatrix} \right)^{-1}$$
$$\times \begin{bmatrix} 0.0476 \\ 0.0013 \\ 0.0013 \\ 0.0110 \end{bmatrix} = \begin{bmatrix} 29.9370 & 41.6850 & 41.6850 & 292.152 \\ 1.7951 & 60.0539 & 12.5807 & 252.762 \\ 1.7951 & 12.5807 & 60.0539 & 252.762 \\ 0.5418 & 10.8845 & 10.8845 & 242.671 \end{bmatrix} \begin{bmatrix} 0.0476 \\ 0.0013 \\ 0.0013 \\ 0.0110 \end{bmatrix} = \begin{bmatrix} 4.7461 \\ 2.9602 \\ 2.9602 \\ 2.7235 \end{bmatrix}$$

,

which gives the unconditional covariance matrix:

$$\widehat{\boldsymbol{\Gamma}}_0 \begin{bmatrix} 4.746 & 2.960 \\ 2.960 & 2.724 \end{bmatrix} \neq \widehat{\boldsymbol{\Sigma}}_u = \begin{bmatrix} 0.0476 & 0.0013 \\ 0.0013 & 0.0110 \end{bmatrix}.$$

Clearly, conditional ($\hat{\boldsymbol{\Sigma}}_u$) and unconditional second moments are radically different: the residuals, also because both series are highly serially correlated, have very low variances and a correlation of 0.057 (= 0.0013/(0.0476x0.0110)$^{1/2}$). In unconditional terms, one-month and ten-year rates are characterized by rather large standard deviations (2.179 and 1.650 percent per year) and a correlation of 0.823 (= 2.960/(4.746x2.724)$^{1/2}$). The latter is more in line with reality and asset pricing expectations, of course.

2.3 Generalization to a VAR(p) Model

Now that we have analyzed the properties of a VAR(1) model, their generalization to the VAR(p) model $\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + ... + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t$, that we presented in (3.11) should be quite obvious.

Using again the lag operator *L* as we did for the VAR(1), (3.11) can be rewritten as

$$\left(\mathbf{I}_N - A_1L - ... - A_pL^p\right)\mathbf{y}_t = \mathbf{a}_0 + \mathbf{u}_t, \qquad (3.63)$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. More compactly, (3.64) can be rewritten as

$$\tilde{\mathbf{A}}(L)\mathbf{y}_t = \mathbf{a}_0 + \mathbf{u}_t, \tag{3.64}$$

where now $\tilde{\mathbf{A}}(L) = \mathbf{I}_N - \mathbf{A}_1 L - ... - \mathbf{A}_p L^p$. Assuming that $\mathbf{y}_t$ is weakly stationary, we obtain that

$$\boldsymbol{\mu} = E\left[\mathbf{y}_t\right] = \left(\mathbf{I}_N - \mathbf{A}_1 - ... - \mathbf{A}_p\right)^{-1} \mathbf{a}_0, \tag{3.65}$$

provided that the inverse of the matrix $\left(\mathbf{I}_N - \mathbf{A}_1 - ... - \mathbf{A}_p\right)$ exists. Also in this case, the conditional mean vector has expression $\boldsymbol{\mu}_{t|t-1} \equiv E\left[\mathbf{y}_t \,|\, \mathbf{y}_{t-1}\right] = \mathbf{a}_0 + \sum_{j=1}^{p} \mathbf{A}_j \mathbf{y}_{t-j}$. Again, for notational convenience, we can transform equation (3.11) by defining $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \boldsymbol{\mu}$:

$$\tilde{\mathbf{y}}_t = \mathbf{A}_1 \tilde{\mathbf{y}}_{t-1} + \mathbf{A}_2 \tilde{\mathbf{y}}_{t-2} + ... + \mathbf{A}_p \tilde{\mathbf{y}}_{t-p} + \mathbf{u}_t. \tag{3.66}$$

Using this equation and applying the same techniques that we have applied in the case of the VAR(1) in Section 2.2, it is possible to show that:

- $\mathrm{cov}\left[\mathbf{y}_t, \mathbf{u}_t\right] = \boldsymbol{\Sigma}_u$, the covariance matrix of $\mathbf{u}_t$;
- $\mathrm{cov}\left[\mathbf{y}_{t-h}, \mathbf{u}_t\right] = \mathbf{0}$ for any $h > 0$;
- $\boldsymbol{\Gamma}_h = \mathbf{A}_1 \boldsymbol{\Gamma}_{h-1} + ... + \mathbf{A}_p \boldsymbol{\Gamma}_{h-p}$ for $h > p$;
- $\boldsymbol{\rho}_h = \boldsymbol{\Psi}_1 \boldsymbol{\rho}_{h-1} + ... + \boldsymbol{\Psi}_p \boldsymbol{\rho}_{h-p}$ for $h > p$, where $\boldsymbol{\Psi}_i = \mathbf{D}^{-1/2} \mathbf{A}_i \mathbf{D}^{1/2}$.

Naturally, all the considerations that we have expressed with references to a VAR(1) can easily be generalized to a VAR(p) model. Such an effort simplifies if we consider that a VAR(p) model can be represented as a Kp-dimensional VAR(1). To this end, define

$$\underset{(Np)\times 1}{\boldsymbol{\xi}_t} \equiv \begin{bmatrix} \tilde{\mathbf{y}}'_t \\ \tilde{\mathbf{y}}'_{t-1} \\ \vdots \\ \tilde{\mathbf{y}}'_{t-p+1} \end{bmatrix}, \quad \underset{(Np\times Np)}{\mathbf{F}_1} \equiv \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_p \\ \mathbf{I}_N & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \tag{3.67}$$

which is also known as the *companion matrix* of the VAR(p) system, and $\underset{(Kp)\times 1}{\mathbf{U}_t} \equiv [\mathbf{u}_t \;\; \mathbf{0} \;\; ... \;\; \mathbf{0}]'$. Then a VAR($p$) model can be written as

$$\boldsymbol{\xi}_t = \mathbf{F}_1 \boldsymbol{\xi}_{t-1} + \mathbf{U}_t, \tag{3.68}$$

where

$$E\left[\mathbf{U}_t \mathbf{U}'_t\right] = \begin{bmatrix} \boldsymbol{\Sigma}_u & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix} \text{ and } E\left[\mathbf{U}_t \mathbf{U}'_{t-h}\right] = \mathbf{0} \text{ for } h > 0.$$

Clearly, (3.69) can be represented in the same form of equation (3.48), i.e., in its VMA representation,

$$\boldsymbol{\xi}_t = \mathbf{U}_t + \mathbf{F}_1 \mathbf{U}_{t-1} + \mathbf{F}_1^2 \mathbf{U}_{t-2} + \dots = \sum_{i=1}^{\infty} \mathbf{F}_1^i \mathbf{U}_{t-i} + \mathbf{U}_t, \tag{3.70}$$

that is, denoting $\boldsymbol{\Pi}_i \equiv \mathbf{J}\mathbf{F}_1^i\mathbf{J}'$, where $\mathbf{J} \equiv \left[\mathbf{I}_N, \mathbf{0}, \dots, \mathbf{0}\right]'$, we have:

$$\boldsymbol{\xi}_t = \sum_{i=1}^{\infty} \boldsymbol{\Pi}_i \mathbf{U}_{t-i} + \mathbf{U}_t. \tag{3.71}$$

It follows that a VAR(p) model is stable (and thus stationary) as long as the eigenvalues of the **companion matrix** $\mathbf{F}_1$ defined in (3.68) are all less than one in modulus, which, implies

$$\det(\mathbf{I}_N - \mathbf{A}_1 z - \dots - \mathbf{A}_p z^p) \neq 0, \text{ for } |z| \leq 1 . \tag{3.72}$$

This condition states that the roots of the characteristic polynomial associated with the matrix should all exceed one in modulus (i.e., they should lie *outside the unit circle*) or, equivalently, that the (inverse) roots from the characteristic polynomial should all lie inside the unit circle, as we already discussed in Chapter 2 for univariate AR models.

2.4 Estimation of a VAR(p) Model

Let us consider an unrestricted, stationary VAR(p) model similar to the one specified in (3.11) and suppose that we want to estimate its parameters.[2] Following the notation in Lütkepohl (2005), we can write (3.11) as

$$\mathbf{Y} = \mathbf{B}\mathbf{Z} + \mathbf{U}, \tag{3.73}$$

[2] A model is said to be unrestricted when the estimation process is allowed to determinate any possible value for the unknown parameters; on the contrary, a model is restricted if the estimation procedure restricts the parameters in some way (for instance, by imposing that some of them is equal to constant values).

where $\boldsymbol{Y} \equiv [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$, $\mathbf{B} \equiv \left[\mathbf{a}_0, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p\right]$ $\mathbf{U} \equiv \left[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_T\right]$, and $\mathbf{Z} \equiv \left[\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_{T-1}\right]$ with $\mathbf{Z}_t \equiv [\mathbf{1}', \mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-p+1}]'$. Also consider that $\mathbf{y} \equiv vec(\mathbf{Y})$, $\boldsymbol{\beta} \equiv vec(\mathbf{B})$, and $\mathbf{u} \equiv vec(\mathbf{U})$, where "*vec*" is the column stacking operator that stacks the columns of a matrix in a column vector. Also recall that the covariance matrix of the residuals is $\boldsymbol{\Sigma}_u$.

The **multivariate LS estimator** (here a GLS estimator) of $\boldsymbol{\beta}$ minimizes the quantity:

$$S(\boldsymbol{\beta}) = \boldsymbol{u}'(\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u)^{-1}\boldsymbol{u} \tag{3.74}$$

Although we shall skip the details of the computation of the estimator (which the interested Reader may find in Lütkepohl, 2005), it is useful to report the solution to the problem:

$$\hat{\boldsymbol{\beta}} = \left((\mathbf{Z}\mathbf{Z}')^{-1} \otimes \boldsymbol{\Sigma}_u\right)\left(\mathbf{Z} \otimes \boldsymbol{\Sigma}_u^{-1}\right)\mathbf{y} = \left((\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z} \otimes \mathbf{I}_N\right)\mathbf{y} \tag{3.75}$$

Notably, the GLS estimator in (3.76) is identical to the OLS estimator obtained by minimizing:

$$S(\boldsymbol{\beta}) = \mathbf{u}'\mathbf{u} = \left[\mathbf{y} - \left(\mathbf{Z}' \otimes \mathbf{I}_N\right)\boldsymbol{\beta}\right]'\left[\mathbf{y} - \left(\mathbf{Z}' \otimes \mathbf{I}_N\right)\boldsymbol{\beta}\right], \tag{3.76}$$

as demonstrated by Zellner (1962). Therefore, as mentioned before, a **standard, unrestricted VAR(*p*) can be simply estimated equation by equation by OLS**. We shall call such an estimator $\hat{\mathbf{B}}$: by construction, being obtained by stacking rows of $\hat{\boldsymbol{\beta}}$ OLS estimators obtained equation-by-equation, $\hat{\mathbf{B}}$ is a N x (p + 1) matrix.

The finite-sample properties of the LS estimator are difficult to derive analytically given the complexity of the expression in (3.76) and therefore we only discuss its asymptotic properties here. Under standard assumptions (see Lütkepohl, 2005, for details), the OLS estimator $\hat{\mathbf{B}}$ is consistent and asymptotically normally distributed,

$$\sqrt{T}\,\mathrm{vec}(\hat{\mathbf{B}} - \mathbf{B}) \xrightarrow{D} N(0, \boldsymbol{\Sigma}_{\hat{\mathbf{B}}}), \tag{3.77}$$

where the *vec* of $\hat{\mathbf{B}}$ needs to be taken to turn the estimator into a vector. This result can also be written more intuitively as

$$vec(\hat{\boldsymbol{B}}) \overset{a}{\sim} N(vec(\boldsymbol{B}), \boldsymbol{\Sigma}_{\hat{\boldsymbol{B}}}/T), \tag{3.78}$$

where the "a" on the top of distribution symbol means "asymptotically distributed as" and $\boldsymbol{\Sigma}_{\hat{\mathbf{B}}} = \text{plim}(\mathbf{ZZ}'/T)^{-1} \otimes \boldsymbol{\Sigma}_u$. Intuitively, this means that as $T \to \infty$, the covariance matrix of the OLS estimator converges (in the sense that deviations from the right-hand side of the formula carry a very small probability) to a complex web of inverse average cross-products between lagged values of the endogenous variables, $\text{plim}(\mathbf{ZZ}'/T)^{-1}$, multiplied by each of the elements of the covariance matrix of the structural residuals, $\boldsymbol{\Sigma}_u$. A few Readers will note the analogy with the $\sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$ expression in Chapter 1.

The covariance matrix $\boldsymbol{\Sigma}_u$ can be estimated as

$$\hat{\boldsymbol{\Sigma}}_u = \frac{1}{T - Np}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}'_t \quad \text{or} \quad \tilde{\boldsymbol{\Sigma}}_u = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}'_t. \quad (3.79)$$

where $\widehat{\boldsymbol{u}}_t - \boldsymbol{y}_t - \widehat{\boldsymbol{B}}\boldsymbol{Z}_{t-1}$. Both estimators are consistent and asymptotically normally distributed independently of $\hat{\mathbf{B}}$. The first estimator is sometimes referred to as the "degree-of-freedom adjusted" version of the covariance matrix estimator.

Alternatively, one may estimate a VAR(p) model using maximum likelihood methods. Given a sample of T observations on the N-variate variable $\mathbf{Y}$ defined as above and a pre-sample of p initial conditions $y_{-p+1}, y_{-p+2}, ..., y_0$, under the assumption that the process is stationary and that innovations are a Gaussian multivariate white noise, the variables $\mathbf{Y} = \left[\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\right]'$ will also be jointly normally distributed. In addition, because the multivariate white noise is assumed to be Gaussian, the innovations at different times will be independent (which allows for considerable simplification when computing the likelihood function). The noise error terms are assumed to be independent with covariance matrix $\boldsymbol{\Sigma}_u$ and, as an implication, $\mathbf{u}$ (that is the vectorization of $\mathbf{U}$ as discussed above) has a covariance matrix $\boldsymbol{\Sigma}_U = \mathbf{I}_T \otimes \boldsymbol{\Sigma}_u$. As a cumulative result of all these assumptions, $\mathbf{u}$ has the following NT-variate normal density:

$$f_{\mathbf{u}}(\mathbf{u}) = (2\pi)^{-\frac{NT}{2}} \left|\mathbf{I}_T \otimes \boldsymbol{\Sigma}_u\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{u}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_u^{-1})\mathbf{u}\right).$$

The density function in (3.81) can also be expressed in terms of the endogenous variables:

$$f_y(\mathbf{y}) = (2\pi)^{-\frac{NT}{2}} \left|\mathbf{I}_T \otimes \boldsymbol{\Sigma}_u\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{Y}-\mathbf{BZ})'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_u^{-1})(\mathbf{Y}-\mathbf{BZ})\right). \tag{3.81}$$

Therefore, the log-likelihood that should be maximized can be represented as follows:

$$\ell(\mathbf{B},\boldsymbol{\Sigma}_u;\mathbf{Y},\mathbf{Z}) = \ln f_y(\mathbf{Y}) = -\frac{NT}{2}\ln(2\pi) - \frac{T}{2}\ln\left|\boldsymbol{\Sigma}_u\right| - \frac{1}{2}(\mathbf{Y}-\mathbf{BZ})'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}_u^{-1})(\mathbf{Y}-\mathbf{BZ})$$

$$= -\frac{NT}{2}\ln(2\pi) - \frac{T}{2}\ln\left|\boldsymbol{\Sigma}_u\right| - \frac{1}{2}\mathrm{tr}\left(\mathbf{U}'\boldsymbol{\Sigma}_u^{-1}\mathbf{U}\right) \tag{3.82}$$

.

Importantly, under the assumption of Gaussian innovations, the OLS estimator in (3.76) is equivalent (conditional on the initial values, i.e., the equivalence is in fact to a quasi-ML because of this form of conditioning, see Chapter 5 for additional details) to the ML estimator of the coefficients. Moreover, the ML estimator of the matrix $\boldsymbol{\Sigma}_u$ is

$$\tilde{\boldsymbol{\Sigma}}_u = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t', \tag{3.83}$$

which is nothing else than the average cross- vector product of the OLS residuals. Substituting the expression for the matrix $\tilde{\boldsymbol{\Sigma}}_u$ that maximizes the likelihood, in the class of all symmetric positive definite matrices, back into (3.83), we obtain

$$\ell(\mathbf{B},\boldsymbol{\Sigma}_u;\mathbf{Y},\mathbf{Z}) = -\frac{NT}{2}\ln(2\pi) - \frac{T}{2}\ln\left|\tilde{\boldsymbol{\Sigma}}_u\right| - \frac{1}{2}NT \,. \tag{3.84}$$

This object is also known as the concentrated log-likelihood of the VAR(*p*) model. Optimizing (3.83) in one pass or maximizing over (3.84)-(3.85) iterating between the two objects until convergence is achieved, will return identical results. Example 3.3 shows the typical estimation outputs of OLS estimation of a VAR model.

Example 3.3. Consider the weekly yields of the one-month, one-year, five-year, and ten-year US Treasury bonds between January 1990 and December 2016 (for a total of 1,408 observations). Suppose that we specify a VAR(1) model for the series. Using Eviews, we have estimated the following model:

$$\begin{bmatrix} y_{1M,t} \\ y_{1Y,t} \\ y_{5Y,t} \\ y_{10Y,t} \end{bmatrix} = \begin{bmatrix} \underset{(-0.775)}{-0.008} \\ \underset{(0.042)}{0.021} \\ \underset{(0.527)}{0.009} \\ \underset{(0.217)}{0.016} \end{bmatrix} + \begin{bmatrix} \underset{(0.000)}{0.835} & \underset{(0.000)}{0.219} & \underset{(0.035)}{-0.083} & \underset{(0.288)}{0.032} \\ \underset{(0.000)}{-0.031} & \underset{(0.000)}{1.023} & \underset{(0.028)}{0.034} & \underset{(0.010)}{-0.031} \\ \underset{(0.012)}{-0.016} & \underset{(0.059)}{0.022} & \underset{(0.000)}{0.993} & \underset{(0.960)}{-0.001} \\ \underset{(0.260)}{-0.007} & \underset{(0.456)}{0.008} & \underset{(0.703)}{0.008} & \underset{(0.000)}{0.988} \end{bmatrix} \begin{bmatrix} y_{1M,t-1} \\ y_{1Y,t-1} \\ y_{5Y,t-1} \\ y_{10Y,t-1} \end{bmatrix} + \begin{bmatrix} u_{1M,t} \\ u_{1Y,t} \\ u_{5Y,t} \\ u_{10Y,t} \end{bmatrix}$$

where *p*-values are reported in brackets. The estimated covariance matrix of the residuals is:

$$\hat{\boldsymbol{\Sigma}}_u = \begin{bmatrix} 0.043 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.007 & 0.007 & 0.005 \\ 0.001 & 0.007 & 0.012 & 0.011 \\ 0.001 & 0.005 & 0.011 & 0.012 \end{bmatrix}.$$

The complete estimation output is reported in Table 3.3. Below each estimated coefficient, the Reader finds the standard errors and the associated *p*-values (in brackets). The coefficients that are statistically significant at a size of the test lower or equal to 5% are boldfaced.

	Yield 1M	Yield 1Y	Yield 5Y	Yield 10Y
Yield 1M (-1)	**0.835**	**-0.031**	**-0.016**	-0.007
	0.012	0.005	0.007	0.006
	(0.000)	(0.000)	(0.012)	(0.260)
Yield 1Y (-1)	**0.219**	**1.023**	0.022	0.008
	0.022	0.009	0.012	0.011
	(0.000)	(0.000)	(0.059)	(0.456)
Yield 5Y (-1)	**-0.083**	**0.034**	**0.993**	0.008
	0.039	0.015	0.021	0.020
	(0.035)	(0.028)	(0.000)	(0.703)
Yield 10Y (-1)	0.032	**-0.031**	-0.001	**0.988**
	0.030	0.012	0.016	0.015
	(0.288)	(0.010)	(0.960)	(0.000)
C	-0.008	0.021	0.009	0.016
	0.027	**0.010**	0.014	0.013
	(0.775)	(0.042)	(0.527)	(0.217)
R-squared	0.993	0.999	0.997	0.997
Adj. R-squared	0.993	0.999	0.997	0.997
Sum sq. resids	59.956	9.306	16.509	15.014
S.E. equation	0.207	0.081	0.108	0.103
F-statistic	51568.215	303363.793	139899.774	117486.468
Log likelihood	224.179	1535.678	1132.122	1198.978
Akaike AIC	-0.311	-2.174	-1.601	-1.696
Schwarz SBC	-0.293	-2.156	-1.582	-1.677
Mean dependent	3.035	3.166	4.155	4.735
S.D. dependent	2.512	2.393	2.166	1.893
Log likelihood		6215.082		
Akaike AIC		-8.800		
Schwarz SBC		-8.725		

Table 3.3 – Estimation output of a VAR(1) model for the one-month, one-, five-, and ten-year yields of the U.S. Treasury bonds

Each column of Table 3.3 represents one equation of the system; because usually equations are written as rows, this implies that they have been flipped around to populate the columns. For instance, the first column corresponds to the first equation of the VAR(1):

$$y_{1M,t} = \underset{(-0.775)}{-0.008} + \underset{(0.000)}{0.835}\, y_{1M,t-1} + \underset{(0.022)}{0.219}\, y_{1Y,t-1} - \underset{(0.039)}{0.083}\, y_{5Y,t-1} + \underset{(0.288)}{0.032}\, y_{10Y,t-1} + u_t$$

As we have discussed, each equation can be estimated separately by OLS. Therefore, the second panel of Table 3.3 presents standard OLS regression statistics for each equation (including the R-square and the adjusted R-square), to which we can attribute the same meaning that has been attached to them in Chapter 1. For example, the *F*-statistic refers to the null hypothesis that all the lags of the endogenous variables are jointly non-significant in each of the system equations. The numbers at the very bottom of the table are instead the summary statistics for the VAR system as a whole. For instance, because an overall, multivariate R-square statistic is not obviously defined, while for each single equation we do report one R-square, in overall terms it makes sense to report the maximized log-likelihood, also because we know that the OLS and ML estimators are identical when the errors are multivariate normal.
In this example, we have assumed that one lag of the endogenous variables was sufficient to explain the key features of the data. However, this assumption was rather arbitrary. Therefore, in Section 2.5, we shall discuss how we can decide the appropriate lag length for a general VAR model.

Before we move on, we shall summarize below two extremely important results that we have discussed (although we have not provided the proofs) in this section:

- when a reduced-form VAR is unconstrained, the GLS estimator is the same as the OLS estimator and therefore an **unconstrained VAR** can be estimated equation by equation by OLS;
- for an unconstrained VAR, the ML and OLS estimators are the same **under the assumption of Gaussian innovations** (further discussion of this topic is provided in on-line supplementary material).

2.5 Specification of a VAR Model and Hypothesis Testing

In Section 2.4, we have discussed how to estimate a VAR model of order *p*, but we have not explained how a researcher may go about deciding the appropriate number of lags to be included. In general, increasing the order of a VAR model reduces the (absolute) size of the residuals and improves the fit of the model, but also its forecasting power. Equivalently, as it is often the case in applied econometrics, by increasing the number of parameters of the model, we generally improve its in-sample accuracy, at expenses of

its out-of-sample predictive power. This occurs because in a VAR, long lag lengths quickly consume degrees of freedom in the individual regression equations (i.e., the number of observations minus the number of parameters to be estimated): if the lag length is specified to be p, each of the N equations will contain Np coefficients plus the intercept term. Therefore, appropriate lag selection is usually crucial to the usefulness of VAR(p) models. In the following, we discuss the selection of the common lag length parameter p to apply to all equations of the VAR model. This prevents us from considering the case of **restricted, standard VAR** models in which the structure and number of lags included in each equation may vary across different equations. These models can be useful but tend to be less frequently used in applied finance.[3]

A first method that can be used to select the appropriate lag length is the **likelihood ratio (LR) test**. In order to understand how this works when applied to the selection of p, suppose that we want to test the hypothesis that a set of variables was generated from a Gaussian VAR with p_0 lags against the alternative specification of $p_1 > p_0$ lags. For instance, assume that we aim at testing whether 4 lags are appropriate, against an alternative specification with 5 lags. Under the assumption of normally distributed shocks entertained earlier (or when the VAR is assumed to be correctly specified under the quasi-MLE principle), the likelihood ratio statistic is

$$LRT(p_0, p_1) = T\left(\ln\left|\tilde{\boldsymbol{\Sigma}}_u^{p_0}\right| - \left|\tilde{\boldsymbol{\Sigma}}_u^{p_1}\right|\right), \qquad (3.85)$$

where T is the number of usable observations, $\left|\tilde{\boldsymbol{\Sigma}}_u^{p_0}\right|$ is the determinant of the covariance matrix estimated under the hypothesis that the VAR model includes p_0 (say, 4) lags of all the

[3] This has a simple justification: when the VAR includes restrictions, then the numerical equivalence between ML, GLS, and OLS estimators breaks down, and consistent estimation needs to be performed jointly using ML methods applied to the full multivariate model. As for their specification, the number of lags in each of the individual equations is often specified using simple t- or F-tests to either go general-to-simple, or simple-to-general. Moreover, there is an inner incoherence between estimating a multivariate model by MLE and performing lag length specification tests at an equation-by-equation level.

variables and $\left|\tilde{\boldsymbol{\Sigma}}_u^{p_1}\right|$ is the determinant of the covariance matrix estimated under the alternative hypothesis that the VAR model contains p_1 (say, 5) lags.

As an alternative, Sims (1980) has proposed a small sample modification of the LR statistic in (3.87) that consists of using $T-(Np+1)$ rather than T as its scale factor, where Np + 1 is the number of parameters per equation under the alternative hypothesis:

$$LRT'(p_0,p_1)=(T-Np-1)\left(\ln\left|\tilde{\boldsymbol{\Sigma}}_u^{p_0}\right|-\left|\tilde{\boldsymbol{\Sigma}}_u^{p_1}\right|\right). \quad (3.86)$$

Both statistics have an asymptotic χ^2 distribution with degrees of freedom equal to the number of restrictions in the system, N(p_1 – p_0). In our example, there are N restrictions in each of the N equation, for a total number of N^2 restrictions. Large values of the test statistics in (3.86)-(3.87) trigger a rejection of the null hypothesis that p_0 lags are sufficient to capture the key features of the (conditional mean function of the) data. On the contrary, if the calculated value of the statistic is less than the critical value of the χ^2 corresponding to the specified size of the test, we will not be able to reject the null that p_0 lags are sufficient. When this occurs, we may think of restricting the model even more, and calculate the likelihood ratio statistic under the null that less than p_0 (say, $p_0-1=3\geq 1$) are adequate, against the alternative of p_0 and to iterate this procedure until we can reject the null hypothesis. This way of specifying the model by sequential LR testing the lag order of a VAR(p) is said to represent a **general-to-simple approach**.[4]

On the one hand, LR tests are quite intuitive, and they are applicable to any type of cross- and within-equation restrictions. For instance, let $\boldsymbol{\Sigma}_u^U$ and $\boldsymbol{\Sigma}_u^R$ be the covariance matrices of the residuals of the unrestricted system and of the restricted one, respectively, for whatever types of restrictions (e.g., that the

[4] Technically, a general-to-simple approach should impose that the size of the tests be adjusted because—being based on a common sample—the tests fail to be independent. However, this issue tends to be disregarded in practice.

covariances between alternative pairs of reduced-form residuals be identical). Then the statistic

$$T\left(\ln\left|\tilde{\boldsymbol{\Sigma}}_u^R\right|-\left|\tilde{\boldsymbol{\Sigma}}_u^U\right|\right) \tag{3.87}$$

can be compared to a χ^2 distribution characterized by a number of degrees of freedom equal to the number of restrictions in the system. In case the resulting sample statistic is less than the critical value under a χ^2 at the specified size level, we shall not reject the null hypothesis that the restricted model is adequate to fit the data. On the other hand, LR tests can only be used to perform a pairwise comparison of two VAR systems, one that is obtained as a restricted version of the other. We also say that the smaller VAR with fewer lags is **nested inside** the bigger VAR with a larger number of lags. As a consequence, if we want to determine the appropriate number of lags that are needed to best characterize a sample, we have first to specify the largest VAR and then proceed to pair it down until we can reject the null hypothesis, meaning that while in some applications going simple-to-general may be logically appealing, sequential LR testing is inconsistent with it. A further drawback of the LR test approach is that, as already emphasized, the χ^2 test will be valid asymptotically only under the assumption that errors from each equation are normally distributed. In general, without distributional assumptions, it is unclear whether performing LR tests may have any merit. Finally, when the sample size is small, it remains unclear whether LR tests may display reasonable power without being subject to substantial size distortions (see Hoffman and Schlagenhauf, 1982, for a discussion).

An alternative approach to the selection of the appropriate lag length is to minimize a multivariate version of the information criteria that were firstly presented in Chapter 2, namely:

$$(M)AIC=\ln\left|\tilde{\boldsymbol{\Sigma}}_u\right|+2\frac{K}{T}, \tag{3.88}$$

$$(M)SBC=\ln\left|\tilde{\boldsymbol{\Sigma}}_u\right|+\frac{K}{T}\ln(T), \tag{3.89}$$

$$(M)HQIC=\ln\left|\tilde{\boldsymbol{\Sigma}}_u\right|+2\frac{K}{T}\ln(\ln(T)), \tag{3.90}$$

where (M) stands for multivariate (to signal that this is a multivariate generalization of the univariate versions proposed in

Chapter 2), $\tilde{\boldsymbol{\Sigma}}_u$ is the estimated covariance matrix of the residuals, T is the number of observations in the sample, and K is the total number of regressors across all equations in the VAR(p) (that is, N^2p+N, where N is the number of equations and p is the number of lags).[5] The intuition behind the criteria and the properties that we discussed in Chapter 2 fully apply to their multivariate generalizations.

Finally, it is interesting to introduce one additional criterion to determine the model order/lag length proposed by Akaike (1969), namely, the *final prediction error* (FPE) measure:

$$FPE(p)=\left[\frac{T+Np+1}{T-Np+1}\right]^N\left|\tilde{\boldsymbol{\Sigma}}_u\right|, \tag{3.91}$$

where $\left|\tilde{\boldsymbol{\Sigma}}_u\right|$ is the determinant of the estimated covariance matrix of the residuals from a given VAR(p) model. Example 3.4 shows how these criteria can be used and compared to select the best fitting VAR(p) model.

Example 3.4. In Example 3.3, we have specified a VAR(1) model for the weekly yields of one-month, one year, five-year, and ten-year US Treasury bonds. However, we have failed to check whether a larger VAR model could be more appropriate to fit the data.

Table 3.4 shows the values of the information criteria that we have just discussed for a number of lags ranging between 0 and 15. It also reports the maximized log-likelihood associated to each model and the sequential (modified, in the sense that it is computed applying Sims' small sample adjusted in (3.87)) log-likelihood test outcomes. Therefore, the second row reports the LR test of $p=1$ versus the alternative $p=2$, the third row tests the null of $p=2$ versus the alternative $p=3$, and so on. In the general, the kth row

[5] When a VAR is estimated equation-by-equation by OLS, the covariance matrix is just computed residually as a result of the estimation process, as we know from Chapter 1. However, when a VAR model is estimated by MLE (and this must occur when it is restricted and thus OLS is not a consistent estimator), we should in principle take into account also the $N(N+1)/2$ elements of the covariance matrix as parameters to be estimated.

shall use a LR statistic to test the null of $p = k - 1$ versus the alternative of $p = k$.

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-4979.73	NA	0.015	7.150	7.1653	7.1559
1	6156.06	22191.71	1.77E-09	-8.804	-8.728	-8.775
2	6258.44	203.42	1.56E-09	-8.927	-8.792*	-8.877*
3	6278.01	38.77	1.55E-09	-8.933	-8.737	-8.859
4	6293.26	30.13	1.55E-09	-8.932	-8.676	-8.836
5	6348.24	108.32	1.47E-09	-8.987	-8.672	-8.869
6	6369.05	40.87	1.46E-09	-8.994	-8.618	-8.854
7	6384.72	30.68	1.46E-09	-8.994	-8.558	-8.831
8	6400.32	30.47	1.46E-09	-8.993	-8.497	-8.808
9	6413.26	25.20	1.47E-09	-8.989	-8.433	-8.781
10	6436.94	45.96	1.45E-09	-9.000	-8.383	-8.769
11	6454.65	34.26	1.45e-09*	-9.002*	-8.326	-8.749
12	6467.35	24.53	1.45E-09	-8.998	-8.261	-8.722
13	6484.71	33.39	1.45E-09	-9.000	-8.203	-8.702
14	6498.44	26.331*	1.46E-09	-8.996	-8.139	-8.676
15	6509.84	21.82	1.47E-09	-8.990	-8.073	-8.647

Table 3.4 – VAR selection criteria applied to one-month, one-, five- and ten-year U.S. Treasury yields

Unsurprisingly, as we have already observed in Chapter 2, different criteria may lead to different lag selections. In this case, the AIC and the FPE select quite a large VAR(11) model, while the Schwarz and the HQ criteria favor a more parsimonious VAR(2) model. However, a VAR(11) model for the four Treasury yield series requires the estimation of a 180 parameters ($N^2 p + N = 4^2 \times 11 + 4 = 180$) with a **saturation ratio** (that is, the number of observations available across the entire model per each parameter that has to be estimated) of only 7.8. Instead, a VAR(2) model implies the estimation of only 36 parameters, with a much safer saturation ratio of 39.1 parameters. Therefore, we elect to specify and estimate the VAR(2) model below (*p*-values are in parentheses):

$$\begin{bmatrix} y_{1M,t} \\ y_{1Y,t} \\ y_{5Y,t} \\ y_{10Y,t} \end{bmatrix} = \begin{bmatrix} \underset{(0.725)}{-0.001} \\ \underset{(0.168)}{0.014} \\ \underset{(0.793)}{0.004} \\ \underset{(0.236)}{0.016} \end{bmatrix} + \begin{bmatrix} \underset{(0.000)}{\mathbf{0.877}} & \underset{(0.110)}{0.170} & \underset{(0.515)}{-0.118} & \underset{(0.796)}{0.042} \\ \underset{(0.039)}{\mathbf{-0.021}} & \underset{(0.000)}{\mathbf{1.185}} & \underset{(0.149)}{0.010} & \underset{(0.456)}{-0.046} \\ \underset{(0.400)}{-0.012} & \underset{(0.800)}{-0.014} & \underset{(0.000)}{\mathbf{1.270}} & \underset{(0431)}{0.066} \\ \underset{(0.727)}{-0.005} & \underset{(0.327)}{-0.051} & \underset{(0.157)}{0.126} & \underset{(0.000)}{\mathbf{1.098}} \end{bmatrix} \begin{bmatrix} y_{1M,t-1} \\ y_{1Y,t-1} \\ y_{5Y,t-1} \\ y_{10Y,t-1} \end{bmatrix} +$$

$$+ \begin{bmatrix} \underset{(0.057)}{0.049} & \underset{(0.591)}{0.059} & \underset{(0.865)}{0.031} & \underset{(0.965)}{-0.007} \\ \underset{(0.000)}{\mathbf{-0.044}} & \underset{(0.000)}{\mathbf{-0.167}} & \underset{(0.240)}{-0.081} & \underset{(0.649)}{-0.028} \\ \underset{(0.068)}{-0.024} & \underset{(0.507)}{-0.037} & \underset{(0.002)}{\mathbf{1.270}} & \underset{(0.357)}{0.077} \\ \underset{(0.454)}{-0.010} & \underset{(0.258)}{-0.061} & \underset{(0.153)}{-0.127} & \underset{(0.191)}{-0.104} \end{bmatrix} \begin{bmatrix} y_{1M,t-2} \\ y_{1Y,t-2} \\ y_{5Y,t-2} \\ y_{10y,t-2} \end{bmatrix} + \begin{bmatrix} u_{1M,t} \\ u_{1Y,t} \\ u_{5Y,t} \\ u_{10y,t} \end{bmatrix},$$

with estimated covariance matrix of residuals equal to

$$\hat{\boldsymbol{\Sigma}}_u = \begin{bmatrix} 0.043 & 0.001 & 0.001 & 0.001 \\ 0.001 & 0.006 & 0.006 & 0.005 \\ 0.001 & 0.006 & 0.011 & 0.010 \\ 0.001 & 0.005 & 0.010 & 0.010 \end{bmatrix}.$$

The coefficients that are significant at a confidence level lower or equal than 5% have been highlighted.

2.6 Forecasting with a VAR model

Similarly to what we have discussed in Chapter 2 for AR models, one obvious application of VAR models is forecasting. Analogously to what we have discussed with reference to univariate models, also in the context of VAR models, loss functions that lead to the minimization of the mean squared forecast error (MSFE) are the most widely used. Evidence in favor of using the MSFE as key forecasting index are given, for instance, by Granger (1969b) and Granger and Newbold (1986), who show that minimum MSFE forecasts also minimize a range of loss functions other than the MSFE. Moreover, for many loss functions, the optimal prediction function is a simple function of minimum MSFE predictions.

Consider a (stationary) N-dimensional VAR(p) process similar to the one in (3.11). Assume that $\mathbf{u}_t$ is an independent multivariate white noise, such that $\mathbf{u}_t$ and $\mathbf{u}_s$ are independent for $t \neq s$ and

thus $E_t\left[\mathbf{u}_{t+h} \mid \Im_t\right]=0$ for $h>0$. The minimum time t MSFE prediction at a forecast horizon *h* is the conditional expected value

$$E_t\left[\mathbf{y}_{t+h} \mid \Im_t\right]=E_t\left[\mathbf{y}_{t+h} \mid \left\{\mathbf{y}_s \mid s \le t\right\}\right], \qquad (3.92)$$

where $\Im_t$ is the information set containing the variables up to and including period *t*. This prediction minimizes the MSFE of each component of the vector $\mathbf{y}_t$.Therefore,

$$E_t\left[\mathbf{y}_{t+h} \mid \Im_t\right]=\mathbf{a}_0+\mathbf{A}_1 E_t\left[\mathbf{y}_{t+h-1} \mid \Im_t\right]+\ldots+\mathbf{A}_p E_t\left[\mathbf{y}_{t+h-p} \mid \Im_t\right],$$

is the optimal *h*-step-ahead predictor of a VAR(*p*) process. The formula in (3.94) can be used recursively to compute *h*-step-ahead predictions starting with $h=1$. For instance, let us consider the case of a VAR(1) model. The one-step-ahead forecast of $\mathbf{y}_t$ with origin at time *t* is

$$E_t\left[\mathbf{y}_{t+1} \mid \Im_t\right]=\mathbf{a}_0+\mathbf{A}_1 E_t\left[\mathbf{y}_t \mid \Im_t\right]=\mathbf{a}_0+\mathbf{A}_1\mathbf{y}_t, \qquad (3.94)$$

where $E_t\left[\mathbf{y}_t \mid \Im_t\right]=\mathbf{y}_t$, given that we are at time *t*. Then, in order to obtain the two-step-ahead forecast we can simply use the value $E_t\left[\mathbf{y}_{t+1} \mid \Im_t\right]$ that we have just computed. Through this iterative process, we can compute the *h*-step-ahead forecast. The conditional expectation that turns out to provide the minimum MSFE has the following properties:

- it is an unbiased predictor, meaning that $E\left[\mathbf{y}_{t+h}-E_t\left[\mathbf{y}_{t+h} \mid \Im_t\right]\right]=0$;
- if $\mathbf{u}_t$ is an independent white noise vector (that, as we shall recall, is a stronger assumption than being uncorrelated), $MSFE\left[E_t\left[\mathbf{y}_{t+h}\right]\right]=MSFE\left[E_t\left[\mathbf{y}_{t+h} \mid \mathbf{y}_t, \mathbf{y}_{t-1} \ldots\right]\right]$, meaning that MSFE of the prediction equals the conditional MSFE given $\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots$.

In case $\mathbf{u}_t$ is not an independent white noise, additional assumptions are required to find the optimal prediction of a VAR(*p*) process. However, without these assumptions it is still possible to find the minimum MSFE predictor among those that are linear functions of $\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots$. Without going into the details of the proof (which can be found in Lütkepohl, 2005), it can be shown that the best linear predictor in terms of MSFE minimization is:

$$E_t\left[\mathbf{y}_{t+h} \mid \Im_t\right] = \mathbf{a}_0 + \mathbf{A}_1 E_t\left[\mathbf{y}_{t+h-1} \mid \Im_t\right] + \ldots + \mathbf{A}_p E_t\left[\mathbf{y}_{t+h-p} \mid \Im_t\right]. \tag{3.95}$$

For the sake of simplicity we analyze again the case of a VAR(1) model, where the prediction function is

$$E_t\left[\mathbf{y}_{t+h} \mid \Im_t\right] = \mathbf{a}_0 + \mathbf{A}_1 E\left[\mathbf{y}_{t+h-1} \mid \Im_t\right]. \tag{3.96}$$

The one-step forecast error $\breve{\mathbf{u}}_t(1)$ is simply:

$$\breve{\mathbf{u}}_t(1) = \mathbf{y}_{t+1} - E_t\left[\mathbf{y}_{t+1} \mid \Im_t\right] = \mathbf{u}_{t+1}, \tag{3.97}$$

and the associated covariance matrix of forecast errors is $\mathbf{\Sigma}_u$. By iterating over this formula, we can obtain that the *h*-step forecast error $\breve{\mathbf{u}}_t(h)$ as

$$\breve{\mathbf{u}}_t(h) = \mathbf{y}_{t+h} - E_t\left[\mathbf{y}_{t+h} \mid \Im_t\right] = \sum_{i=0}^{h-1} \mathbf{A}_1^i \mathbf{u}_{t+h-i}, \tag{3.98}$$

where $\mathbf{A}^0 = \mathbf{I}_N$. The covariance matrix of the forecast errors is therefore $\sum_{i=0}^{h-1} \mathbf{A}_1^i \mathbf{\Sigma} (\mathbf{A}_1^i)'$. The generalization to a VAR(*p*) model is straightforward, although computations are non-trivial (the interested Reader is referred to Lütkepohl, 2005). Example 3.5 shows VAR models in action when it comes to prediction.

Example 3.5. Figure 3.2 shows the one-week ahead forecasts of the one-month, one-, five-, and ten-year U.S. Treasury bond yields obtained from the VAR(2) model estimated in Example 3.4.

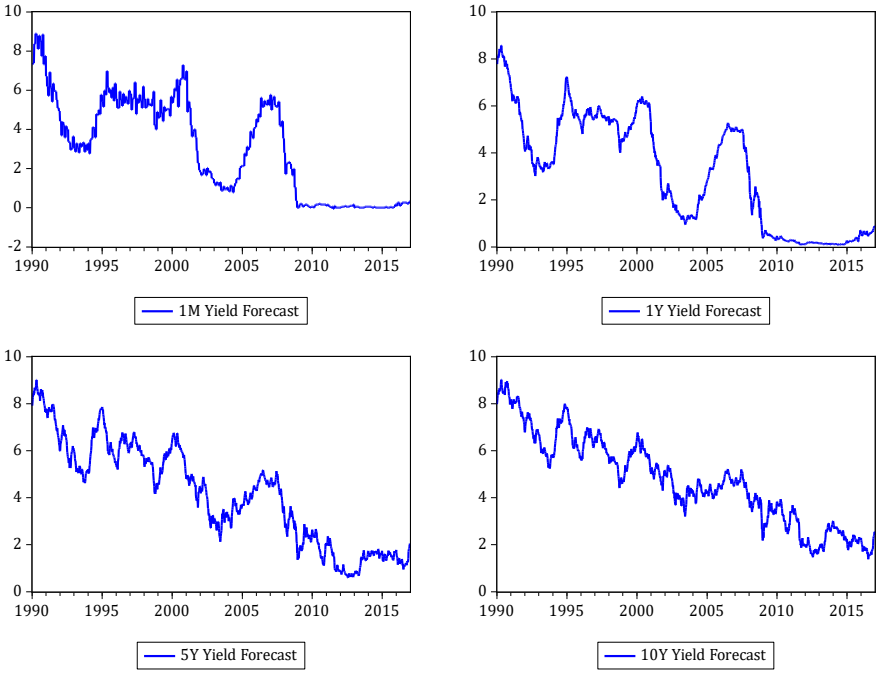

Figure 3.2 –One-week-ahead forecasts of one-month, one-, five- and ten-year U.S. Treasury yields from a VAR(2)

Table 3.5 reports the forecast accuracy measures that have been discussed in Chapter 2, namely, the root mean squared error (RMSE, which is just the square root of the mean square forecast error), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). Clearly, the lower these prediction error measures are, the higher the practical usefulness of a model.

Variable	Inc. obs.	RMSE	MAE	MAPE
1M Yield	1409	0.21	0.10	35.18
1Y Yield	1409	0.08	0.05	3.63
5Y Yield	1409	0.11	0.08	2.74
10Y Yield	1409	0.10	0.08	1.99

RMSE: Root Mean Square Error
MAE: Mean Absolute Error
MAPE: Mean Absolute Percentage Error

Table 3.5 – Forecast accuracy measures for a VAR(2) model of one-month, one-,five- and ten-year U.S. Treasury yields

Obviously, these accuracy measures are useful when we would like to compare the predictive power of different models. For example, we may want to compare the forecast accuracy of the VAR(2) vs. the VAR(11) model that was selected by the FPE and AIC criteria in Example 3.4. Table 3.6 displays the forecast accuracy measures for the VAR(11) model. It is evident that the VAR(2) and the VAR(11) models display very similar predictive power, although the VAR(11) slightly outperforms the VAR(2) according to some specific indicators.

Variable	Inc. obs.	RMSE	MAE	MAPE
1M Yield	1409	0.19	0.10	48.40
1Y Yield	1409	0.08	0.05	3.96
5Y Yield	1409	0.10	0.08	2.79
10Y Yield	1409	0.10	0.08	1.99

RMSE: Root Mean Square Error
MAE: Mean Absolute Error
MAPE: Mean Absolute Percentage Error

Table 3.6 – Forecast accuracy measures for a VAR(11) model of 1-month, 1-,5- and 10-year U.S. Treasury yields

3- Structural Analysis with VAR Models

3.1 Impulse Response Functions

In Section 2, we have discussed the statistical properties of a VAR(p) model, how it can be estimated, and how it can be used in forecasting applications. However, VAR models are often used in practice with the goal of understanding the **dynamic relationships** between the variables of interest. For instance, in Example 3.4, we have estimated a VAR(2) model for the one-month, one-, five-, and ten-year U.S. Treasury yield series and then, in Example 3.5, we have computed and assessed one-step-ahead forecasts. However, a researcher may also be interested in studying the effects that a sudden increase (decrease) in the 1-month rate, for instance as a result of a tight (expansive) monetary policy, may have on the other yields in the system (when these four specific maturity buckets are used to summarize the term

structure of the Treasury yield curve). In other words, a researcher may be interested in the effects that a shock to one (or more) variable(s) produces over the others. Therefore, in this section we introduce **impulse response functions** (in short IRFs). A general definition of impulse response function is as follows.

Definition 3.4. (Impulse Response Function) In the context of a VAR model, an impulse response functions trace out the time path of the effects of an exogenous shock to one (or more) of the endogenous variables on some or all of the other variables in a VAR system.

To simplify, let us start our discussion from the simple VAR(1) model (written in reduced form) discussed in Section 2.1, namely:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{1,0} \\ a_{2,0} \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}. \quad (3.99)$$

We already know that a stationary VAR(p) model has a moving average representation, and, in particular this also applies to the VAR(1), i.e., using a compact notation,

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{u}_t, \quad (3.100)$$

can be rewritten as

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{A}_1^i \mathbf{u}_{t-i}, \quad (3.101)$$

or, alternatively, recalling the algebraic steps that we have discussed in Section 2.2, to

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Theta}_i \mathbf{u}_{t-i}, \quad (3.102)$$

that is,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \theta_{1,1(i)} & \theta_{1,2(i)} \\ \theta_{2,1(i)} & \theta_{2,2(i)} \end{bmatrix} \begin{bmatrix} u_{1,t-i} \\ u_{2,t-i} \end{bmatrix}. \quad (3.103)$$

You will also recall from our discussion in Section 2.1 that the two error processes, $\{u_{1,t}\}$ and $\{u_{2,t}\}$ can be also represented in terms of the two sequences $\{\varepsilon_{1,t}\}$ and $\{\varepsilon_{2,t}\}$, i.e., the structural (or pure), unobserved innovations:

$$\begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} = \frac{1}{1 - b_{1,2} b_{2,1}} \begin{bmatrix} 1 & -b_{1,2} \\ -b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}. \quad (3.104)$$

Therefore, plugging (3.105) into (3.104), we obtain a 2x2 matrix $\mathbf{\Phi}_i$ equal to

$$\mathbf{\Phi}_i = \frac{\mathbf{A}_1^i}{1-b_{1,2}b_{2,1}}\begin{bmatrix} 1 & -b_{1,2} \\ -b_{2,1} & 1 \end{bmatrix} = \frac{\mathbf{\Theta}_i}{1-b_{1,2}b_{2,1}}\begin{bmatrix} 1 & -b_{1,2} \\ -b_{2,1} & 1 \end{bmatrix},$$

and therefore,

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{\Phi}_i \boldsymbol{\varepsilon}_{t-i} \,. \tag{3.106}$$

It is now easy to see how the moving average representation of the VAR can be useful: the coefficients of the matrix $\mathbf{\Phi}_i$, i.e., each $\phi_{i,j}$ can be used to generate the effects of shocks to the innovations $\varepsilon_{1,t}$, $\varepsilon_{2,t}$ on the entire time path of the $\{y_{1,t}\}$ and $\{y_{2,t}\}$ series. In other words, the four coefficients $\phi_{1,1(i)}$, $\phi_{1,2(i)}$, $\phi_{2,1(i)}$ and $\phi_{2,2(i)}$ for each *i* can be regarded as **impact multipliers**. For instance, $\phi_{1,2(0)}$ represents the instantaneous impact on $y_{1,t}$ of a one-unit change in $\varepsilon_{2,t}$ (i.e., the structural innovation to $y_{2,t}$), while $\phi_{1,2(1)}$ is the one-period response of $y_{1,t}$ to the same unit change in $\varepsilon_{2,t-1}$. The cumulative effects of a one-unit shock (or impulse) to $\varepsilon_{2,t}$ on the variable $y_{1,t}$ after *H* periods can then be obtained by computing the sum $\sum_{i=0}^{H} \phi_{1,2(i)}$. Clearly, the same result holds for the cumulative effects of a unit shock to $\varepsilon_{1,t}$ on $y_{2,t}$, which can be computed as $\sum_{i=0}^{H} \phi_{2,1(i)}$, and so on. Interestingly, if we let the horizon H approach to infinity, we obtain the so-called **long-run multipliers**. Indeed, as the sequences $\{y_{1,t}\}$ and $\{y_{2,t}\}$ are assumed to be stationary, it follows that $\sum_{i=0}^{\infty} \phi_{j,k(i)}$ for *j*, *k* = 1, 2, …, *N*, is finite. Put into other words, because a VAR model can be easily generalized to contain *N* variables instead of two, the element $\phi_{j,k(i)}$, i.e., the (*j*,*k*)th of the matrix $\mathbf{\Phi}_i$ represents the

reaction of the *j*th variable of the system to a one-unit shock in a variable *k*, *i* periods ago. Therefore, given Definition 3.4, the set of elements $\phi_{j,k(i)}$ with $i = 1,2,...,H$ can be easily seen as the **impulse response function** of the *j*th variable of the system, up to the period H. The sum $\sum_{i=0}^{H} \phi_{j,k(i)}$, that represents the cumulative effects of a shock to variable *k* on the variable *j* after *H* periods, is also known as the **cumulative response** of the variable *j* to a shock to the variable *k*.

What is the problem with the VAR representation in (3.107)? If the VAR system were identified, i.e., if it were possible to recover all the parameters of the structural VAR model from the estimates of the VAR in its standard form, it would be possible to trace out the effects of a shock to one (or more) of the structural innovations to the variables. However, we already know from Section 2.1 that a VAR in its reduced form is under-identified by construction and therefore we are not able to compute the coefficients $\phi_{j,k(i)}$ from the OLS estimates of the VAR in its standard form unless we do not impose adequate restrictions. As we have seen in Section 2.1, one method to place these restrictions consists of the application of a Choleski decomposition. In practice, by using a Choleski decomposition, we can re-write the VMA representation of a VAR(1) in (3.103) (note that this also applies to a VAR(*p*), because a VAR(*p*) can be rewritten as a VAR(1)) such that

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Theta}_i \mathbf{W}\mathbf{W}^{-1}\mathbf{u}_{t-i}, \tag{3.107}$$

where $\boldsymbol{\Sigma}_u = \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}'$, $\boldsymbol{\varepsilon}_t = \mathbf{W}^{-1}\mathbf{u}_{t-1}$, and $\boldsymbol{\Phi}_i = \boldsymbol{\Theta}_i\mathbf{W}$. It shall be easy to recognize that (3.108) is equivalent to (3.107). However, it should be already clear from Section 2.1 that a Choleski decomposition allows only the shock to the first variable to contemporaneously affect all the other variables in the system. A shock to the second variable will produce a contemporaneous effect on all the variables in the system, but the first one (this may of course be impacted in the subsequent period, through the transmission effects mediated by the autoregressive coefficients). A shock to the third variable will affect all the variables in the system, but the first two, and so on. Therefore, it is important to recognize that this identification scheme forces a potentially important **identification asymmetry**

on the system that is typical of Choleski ordering schemes. For instance, in our initial bivariate example, a shock to $\varepsilon_{1,t}$ has a contemporaneous effect on both $u_{1,t}$ and $u_{2,t}$ (and thus on $y_{1,t}$ and $y_{2,t}$), but a shock to $\varepsilon_{2,t}$ does not contemporaneously impact $u_{1,t}$ (and thus $y_{1,t}$) . For this reason, $y_{1,t}$ is said to be "casually prior" to $y_{2,t}$, a bit of language that will be better explained later on. Of course, as already emphasized in Section 2.1, a different ordering of the variables in the system would have been possible, implying a reverse ordering of the shocks and that $y_{2,t}$ would have been "casually prior" to $y_{1,t}$. To make our reasoning clearer, in Example 3.6 we see how the decomposition works in practice.

Example 3.6. Let us consider a VAR(1) for the one-month U.S. Treasury bill and the ten-year Treasury bond yields (the same series for the January 1990 - December 2016 sample that we have estimated in Example 3.2):

$$\begin{bmatrix} y_{1M,t} \\ y_{10Y,t} \end{bmatrix} = \begin{bmatrix} \underset{[-2.5382]}{-0.0490} \\ \underset{[0.8711]}{0.0080} \end{bmatrix} + \begin{bmatrix} \underset{[210.6540]}{0.9819} & \underset{[0.4077]}{0.0209} \\ \underset{[3.3784]}{0.0009} & \underset{[240.0320]}{0.9970} \end{bmatrix} \begin{bmatrix} y_{1M,t-1} \\ y_{10Y,t-1} \end{bmatrix} + \begin{bmatrix} u_{1M,t} \\ u_{10Y,t} \end{bmatrix},$$

with estimated covariance matrix of the reduced-form residuals:

$$\hat{\boldsymbol{\Sigma}}_u = \begin{bmatrix} 0.0476 & 0.0013 \\ 0.0013 & 0.011 \end{bmatrix}.$$

As we shall recall from Section 2.1, applying a Choleski decomposition we get that $\text{var}[u_{1M,t}] = \sigma_1^2$, $var[u_{10Y,t}] = \sigma_2^2 - b_{2,1}^2\,\sigma_1^2$, $\text{cov}[u_{1M,t}, u_{10Y,t}] = -b_{2,1}\sigma_1^2$. Therefore, $b_{2,1}$ is equal to

$$b_{2,1} = -\frac{\hat{\sigma}_{1,2}}{\hat{\sigma}_1^2} = -\frac{0.0013}{0.047} = -0.027,$$

and equations (3.29)-(3.30) become

$$u_{1M,t} = \varepsilon_{1,t},$$

$$u_{10Y,t} = \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t} = \varepsilon_{2,t} - 0.027\varepsilon_{1,t}.$$

This means that a shock to $\varepsilon_{1.t}$ equal to one-standard deviation (0.218, that is $\sqrt{0.0476}$) causes an immediate change by 0.218 in $u_{1M,t}$ (and thus in $y_{1M,t}$); in addition, it will also cause an immediate increase (albeit very small) of $0.218 \times 0.027 = 0.006$ in $u_{10Y,t}$ (and thus in the 10-year Treasury yield) because of the implicit correlation structure that is admissible under the selected Choleski scheme. At time $t+1$, the lagged value of the one-month yield enters the first equation with a coefficient 0.9819 and thus after one period the one-month yield will grow by $0.9819 \times 0.218 = 0.214$ (i.e., approximately 21 basis points, henceforth bps) above what it would have been without the shock. The ten-year yield would have been $0.9970 \times 0.006 = 0.00598$ higher because of the effect of its own lag. In addition, the lagged value of the 1-month yield also enters the second equation with a coefficient 0.0009, and thus the 10-year Treasury yield will rise by an additional $0.0009 \times 0.218 = 0.000196$; in total, the 10-year Treasury yield would be approximately 0.0062 higher with respect to what it would have been without a shock to the 1-month yield. Therefore, one period after the one standard deviation shock to the one-month Treasury yield has occurred, the **cumulative response** of the one-month Treasury yield to its own shock would have been $0.218 + 0.214 = 0.432$, that is, 43 bps. In addition, the accumulated response of the ten-year Treasury yield to the one standard deviation shock to the one-month Treasury yield would have been $0.006 + 0.0062 = 0.0123$. The process then progresses further over subsequent rounds of impulse and reaction.

Alternatively, it is easy to see what happens if we give a one standard deviation shock to $\varepsilon_{2,t}$ (equal to 0.105): $u_{10Y,t}$ immediately increases by 0.105 (and so does $y_{10Y,t}$), but nothing happens to $u_{1M,t}$. Therefore, at time $t+1$ the 10-year yield would be higher by $0.9970 \times 0.105 = 0.10469$ (i.e., approximately 10 bps) because of the effect of its own lag (for an accumulated response of 0.209). In addition, the lag of the ten-year yield now affects the 1-month yield with a coefficient of 0.0209 and therefore the one-month Treasury yield will be $0.0209 \times 0.105 = 0.0022$ higher than

it would have been without a shock happening to the 10-year Treasury yield.

Figure 3.3 depicts the impulse response functions to a one-standard deviation shock to the 1-month yield and to the 10-year yield on the basis of a Choleski triangular scheme that places the one-month yield at the top of the variable ordering.

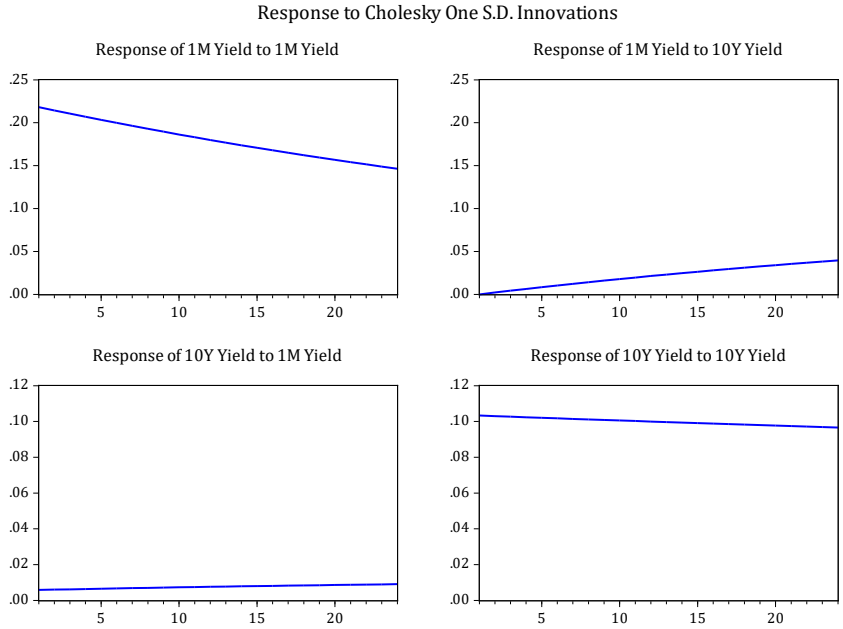

Figure 3.3 – Impulse response functions to shocks to one-month and ten-year yields, ordered on the basis of a Choleski triangular scheme that places shocks to the one-month at the top of the ordering

Notably, as we have seen in Example 3.6, it is not compulsory to give simple one standard deviation shocks. A researcher is free to give to the system all kinds of shocks that she is interested in or that she feels are economically plausible. However, it is quite common in practice to study the effects of a shock equal to one standard deviation, especially when the variables have different scales. Such a rescaling may sometimes give a better picture of the dynamic relationships among variables because the average scale of the innovations occurring in a system depends on their standard deviation.

Summing up, two points should be clear:

- a reduced-form VAR, although commonly employed in

applied finance, is under-identified and it is not possible to recover the structural parameters from its estimates unless we impose some restrictions, i.e., identification forces the researcher to impose some structure on the system;

- Choleski decompositions provide a minimal set of restrictions concerning the simultaneous relationships among variables that can be used to identify the structural model (however other identification schemes, based on a theoretical background, are of course possible but appear to be less common in finance).

It should be also clear that under a Choleski decomposition the ordering of the variables in the system is important: it is indeed crucial which of the variables is placed first and which one is placed second, and so on. In addition, the relevance of the ordering depends on the magnitude of the correlation coefficients between the innovations $u_{1,t}$, $u_{2,t}$,..., $u_{N,t}$: in our example, when $cov[u_{1,t}, u_{2,t}] \simeq 0$, it must be that $b_{1,2} \simeq 0$, in which case none of the variables is simultaneously associated and the reduced-form VAR is practically isomorphic to the structural VAR, so that all standard shocks are also structural shocks. When the reduced-form shocks are instead highly correlated, as it is often the case, unfortunately, the ordering of the variables cannot be determined with statistical methods but has to be selected by the researcher. Therefore, as suggested by Sims (1981), it is often warmly suggested that a researcher tries different orderings of the variables to understand what are the implications to choose some restrictions instead of others in terms of the resulting estimates of the IRFs.

Another important issue with IRFs is that they are constructed using the estimated coefficients. Given that each coefficient is estimated with uncertainty (due to a variety of factors, such as small sample sizes and measurement error), the IRFs will contain sampling error as well, i.e., they will be highly nonlinear transformations of the sample parameter estimates. Therefore, it is often advisable, after having computed and plotted the IRFs of interest, to also construct confidence intervals around them to account for the uncertainty that derives from parameter estimation. Although under some assumptions, confidence bands can be constructed relying on asymptotic theory that implies that OLS (equal to ML) parameter estimates are normally distributed, recently it has become common to use **bootstrapping methods**,

see the Mathematical and Statistical Appendix at the end of the book for a brief introduction. The bootstrap is based on resampling from either the distribution of parameter estimates obtained from the true, original data (*parametric* bootstrap), or directly from the data with replacement to obtain blocks of consecutive observations (*nonparametric* bootstrap); in both cases, the final goal is to generate a large number of alternative pseudo-samples then used to approximate the distribution of one or more sample statistics of interest—for instance, the IRFs of a VAR—computed across the pseudo-samples (see, for instance, Efron and Tibshirani, 1986). When applied to IRFs, bootstrapping techniques have two major advantages: first, they produce confidence intervals that are more reliable than those based on asymptotic theory (see Kilian, 1998); second, this methodology avoids the computation of exact expressions for the asymptotic variance of the IRF coefficients, which is otherwise rather complex (see Lütkepohl, 1991). The bootstrap method consists in the implementation of the following steps.

- Each equation is estimated by OLS/MLE and the vector series $\left\{\mathbf{u}_t^b\right\}$ of *T* errors (with *T* equal to the original sample size) is constructed by randomly sampling with replacement from the estimated residuals. Random sampling with replacement from an initial dataset means that *T* observations are drawn, randomly from the original sample. After each drawn the observation is replaced in the sample, so that any observation can be drawn more than once. Importantly, when drawing the observations, one has to properly consider the fact that the error terms are correlated across the equations, which implies that horizontal blocks of *N* different structural residuals are jointly drawn.
- The series $\left\{\mathbf{u}_t^b\right\}$ and the estimated coefficients are then used to construct a pseudo-vector of endogenous variable series, $\left\{\mathbf{y}_t^b\right\}$.
- The coefficients used to generate $\left\{\mathbf{y}_t^b\right\}$ are discarded and new coefficients are estimated from $\left\{\mathbf{y}_t^b\right\}$. The impulse response functions are computed from the newly estimated

coefficients and saved, also indexed by the bootstrap iteration *b*.

When this procedure is repeated a sufficiently large number of times, $b = 1, 2, \ldots, B$, the resulting impulse response functions can be used to construct the confidence bands. As an example, a 95% confidence interval is the one that excludes the highest and the lowest bootstrapped, re-sampled 5% observations: for each horizon $h = 1, 2, \ldots, H$ the lowest (highest) 2.5% IRFs are excluded, and the interval is set to contain the remaining 95% IRFs. An impulse response function is considered to be statistically significant if zero is not included in the bootstrapped confidence interval.

Example 3.7. We are now ready to return to Example 3.4. In case of a positive shock to the short end of the yield curve (a tightening of conventional monetary policy), what can we expect to happen to the rest of the curve, on average? Let us consider the VAR(2) model estimated in Example 3.4 and compute the IRFs to a one standard deviation positive shock (equal to approximately 21 bps) to the 1-month yield. Figure 3.4 shows the responses of each of the variables in the system over 52 weeks, i.e., for $h = 1, 2, \ldots, 52$. The dotted lines represent the 95% bootstrapped confidence intervals.

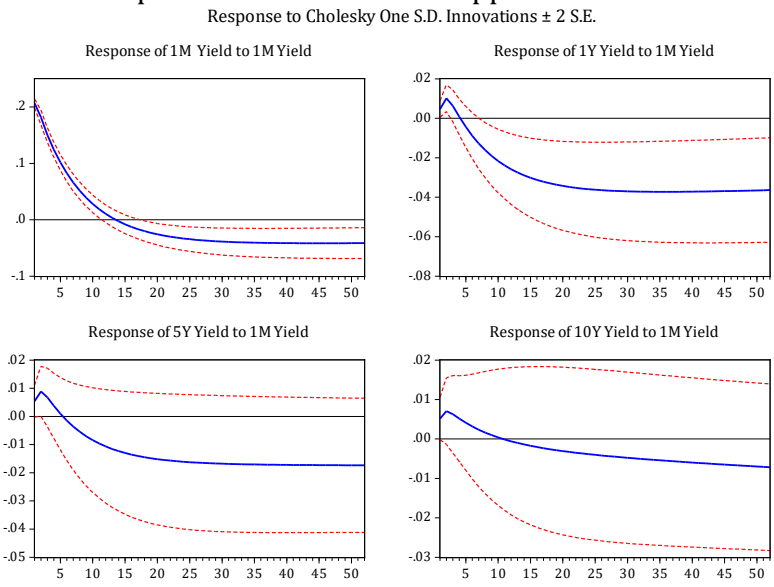

Figure 3.4 – IRFs to a one standard deviation positive shocks to the 1-month Treasury yield

Unsurprisingly, the response of the one-month yield to its own shock is positive and quite persistent. However, after approximately 18 weeks such an effect turns negative and statistically significant. Conversely, the other Treasury yields display weak or no effects: the responses of the five- and ten-year yields to the shock are never significant (as zero is always in the confidence interval, the null hypothesis that the IRF is equal to zero cannot be rejected). The 1-year yield is mildly positively affected by the shocks, but after two weeks the response turns negative (small, but significant). If one takes a one-month U.S. T-bills positive shock as indicative of a monetary policy tightening, the figure gives rather attenuated indications of policy transmission to longer-term riskless rates.

3.2 Variance Decompositions

In Section 2.6, we have discussed how a VAR model can be used in forecasting. However, irrespective of the actual accuracy of the predictions, understanding the properties of forecast errors is helpful in order to assess the interrelationships among the variables in the system. In (3.99), we have provided the formula to compute the forecast error for a VAR(1) model. It is possible to re-formulate such an equation exploiting the VMA representation of the model, so that the *h*-step-ahead forecast error is

$$\mathbf{u}_t(h) = \mathbf{y}_{t+h} - E_t\left[\mathbf{y}_{t+h}\right] = \sum_{i=0}^{h-1} \mathbf{\Phi}_i \boldsymbol{\varepsilon}_{t+h-i}. \qquad (3.108)$$

To help our understanding, we apply (3.109) to the bivariate VAR model that we have discussed in Section 3.1 and, focusing only on the series $\left\{y_{1,t}\right\}$, we note that

$$u_{y1}(h) = y_{1,t+h} - E\left[y_{1,t+h|t}\right] = \phi_{1,1}(0)\varepsilon_{1,t+h} + \phi_{1,1}(1)\varepsilon_{1,t+h-1} + ... + \phi_{1,1}(h-1)\varepsilon_{1,t+1} + \phi_{1,2}(0)\varepsilon_{2,t+h} + \phi_{1,2}(1)\varepsilon_{2,t+h-1} + ... + \phi_{1,2}(h-1)\varepsilon_{2,t+1} \qquad (3.109)$$

Consequently, if we denote by $\sigma^2_{y1}(h)$ the *h*-step-ahead variance of the forecast of $y_{1,t+h}$, we obtain:

$$\sigma^2_{y1}(h) = \sigma^2_{y1}\left[\phi^2_{1,1}(0) + \phi^2_{1,1}(1) + ... + \phi^2_{1,1}(h-1)\right] + \sigma^2_{y2}\left[\phi^2_{1,2}(0) + \phi^2_{1,2}(1) + ... + \phi^2_{1,2}(h-1)\right] \qquad (3.110)$$

Interestingly, because all the coefficients $\phi_{j,k}^2$ are non-negative as they are squared, the variance of the forecast error increases as the forecast horizon *h* increases.

It is possible to decompose the *h*-step-ahead forecast error variance in (3.111) into the proportion due to each of the (structural) shocks. In particular, the proportion of the forecast error variance due to the shocks in the sequence $\{\varepsilon_{1,t}\}$ is

$$\frac{\sigma_{y1}^2\left[\phi_{1,1}^2(0)+\phi_{1,1}^2(1)+...+\phi_{1,1}^2(h-1)\right]}{\sigma_{y1}^2(h)}, \qquad (3.111)$$

while the proportion of forecast error variance due to the shocks in the sequence $\{\varepsilon_{2,t}\}$ is

$$\frac{\sigma_{y2}^2\left[\phi_{1,2}^2(0)+\phi_{1,2}^2(1)+...+\phi_{1,2}^2(h-1)\right]}{\sigma_{y1}^2(h)}. \qquad (3.112)$$

It is easy to see how this result can be generalized to a VAR including *N* variables instead of the two in our example. The computation of the proportion of the forecast error variance due to each shock is often referred to as **forecast error variance decomposition**. In practice, variance decompositions determine how much of the *h*-step-ahead forecast error variance of a given variable is explained by innovations to each explanatory variable for $h=1,2,....$ For instance, in our bi-variate example, if the $\varepsilon_{2,t}$ shocks explain none of the forecast variance of $y_{1,t}$ at all forecast horizons, we would say that the series $\{y_{1,t}\}$ is **exogenous**, that is, it evolves independently of the $\varepsilon_{2,t}$ shocks and of the $\{y_{2,t}\}$ sequence. Conversely, if $\varepsilon_{2,t}$ shocks explain all the forecast variance of $\{y_{1,t}\}$ at all forecast horizons, then $\{y_{1,t}\}$ is said to be completely endogenous. In most practical applications, it is common for $\varepsilon_{1,t}$ ($\varepsilon_{2,t}$) to explain most of the forecast variance of $y_{1,t}$ ($y_{2,t}$) at short-term horizons, while the importance of shocks to $y_{2,t}$ ($y_{1,t}$) on the forecast variance of $y_{1,t}$ ($y_{2,t}$) grows with the forecast horizon.

Importantly, like in IRF analysis, forecast error variance decompositions of reduced-form VARs require identification (because otherwise we would be unable to go from the coefficients in $\boldsymbol{\Theta}_i$ to their counterparts in $\boldsymbol{\Phi}_i$), therefore, Choleski decompositions (or other restriction schemes) are typically imposed. As we shall recall from Section 3.1, in the bivariate model examined in this section, this means that all the one-period forecast error variance of $y_{1,t}$ is attributed to $\varepsilon_{1,t}$. It is important to emphasize again that assuming a particular ordering is necessary to compute the impulse responses and variance decompositions from a VAR, although the restrictions underlying the ordering used may not be supported by the data because they may be decided by the researcher on an a-priori basis. As already discussed in the case of IRFs, when possible economic theory should give some guidance on what is a plausible ordering of the variables (i.e., to point out that when the movement in a variable is likely to temporally precede rather than follow the movements by the other variables this variable should be placed at the top of the ordering). Once more, however, the lower the pairwise cross-correlations among the errors are, the weaker the impact of the ordering on the results.

In conclusion, forecast error variance decomposition and impulse response function analyses both entail similar information from the time series under analysis and are often used in combination (such a combined approach is called **innovation accounting**) to uncover the dynamic interrelationships among the endogenous variables.

Example 3.8. We present the variance decompositions for the forecast error variance of the one-month, one-, five-, and ten-year Treasury yields from the VAR(2) estimated in Example 3.4 at forecast horizons between 1 and 12 weeks. In particular, panel (a) of Table 3.7 shows in which proportion the innovations to each variable in the system contribute to the forecast error variance of the one-month T-bill yield at different horizons. The variance decompositions of the one-year, five-year and ten-year Treasury yields can be found in panels (b), (c), and (d), respectively.

Variance Decomposition of 1M Yield:

Period	S.E.	1M Yield	1Y Yield	5Y Yield	10Y Yield
1	0.207	100.000	0.000	0.000	0.000
2	0.275	99.897	0.061	0.039	0.003
3	0.315	99.377	0.504	0.110	0.008
4	0.340	98.241	1.544	0.199	0.016
5	0.359	96.425	3.254	0.299	0.022
6	0.373	93.943	5.625	0.405	0.027
7	0.386	90.867	8.591	0.512	0.030
8	0.398	87.309	12.045	0.615	0.031
9	0.409	83.405	15.854	0.710	0.030
10	0.421	79.300	19.879	0.793	0.029
11	0.433	75.127	23.985	0.861	0.027
12	0.445	70.998	28.060	0.915	0.028

Panel (a)

Variance Decomposition of 1Y Yield:

Period	S.E.	1M Yield	1Y Yield	5Y Yield	10Y Yield
1	0.079	0.348	99.652	0.000	0.000
2	0.126	0.784	99.110	0.092	0.015
3	0.164	0.609	99.201	0.153	0.037
4	0.197	0.425	99.325	0.188	0.063
5	0.226	0.358	99.341	0.210	0.091
6	0.254	0.406	99.244	0.228	0.122
7	0.280	0.543	99.057	0.243	0.156
8	0.304	0.741	98.806	0.258	0.195
9	0.328	0.978	98.512	0.274	0.236
10	0.351	1.237	98.191	0.291	0.282
11	0.373	1.506	97.855	0.308	0.330
12	0.394	1.778	97.513	0.328	0.382

Panel (b)

Variance Decomposition of 5Y Yield:

Period	S.E.	1M Yield	1Y Yield	5Y Yield	10Y Yield
1	0.106	0.261	53.622	46.117	0.000
2	0.166	0.391	53.547	46.045	0.017
3	0.212	0.341	53.959	45.674	0.026
4	0.251	0.267	54.526	45.180	0.027
5	0.284	0.209	55.120	44.646	0.025
6	0.314	0.174	55.695	44.109	0.022
7	0.341	0.158	56.235	43.588	0.019
8	0.366	0.159	56.737	43.087	0.017
9	0.390	0.173	57.200	42.612	0.015
10	0.412	0.197	57.629	42.161	0.013
11	0.433	0.227	58.026	41.735	0.012
12	0.454	0.263	58.393	41.333	0.011

Panel (c)

Variance Decomposition of 10Y Yield:

Period	S.E.	1M Yield	1Y Yield	5Y Yield	10Y Yield
1	0.101	0.260	37.242	51.653	10.845
2	0.158	0.306	36.977	52.897	9.820
3	0.202	0.285	37.053	53.316	9.345
4	0.237	0.253	37.265	53.389	9.093
5	0.268	0.221	37.527	53.311	8.941
6	0.296	0.194	37.805	53.163	8.839
7	0.321	0.170	38.086	52.981	8.763
8	0.344	0.150	38.363	52.784	8.703
9	0.365	0.134	38.633	52.580	8.653
10	0.385	0.120	38.896	52.375	8.608
11	0.404	0.109	39.151	52.171	8.568
12	0.422	0.100	39.399	51.971	8.530

Panel (d)

Table 3.7 – Forecast error variance decomposition of one-month, one-, five-, ten-year Treasury yields when the Choleski ordering is one-month, one-, five-, ten-year yields

Notably, the forecast error variance of the one-month yield at a one-week horizon is entirely explained by its own innovations. By construction, this derives from the specific Choleski triangularization that entails placing the one-month yield on the top of the ordering. However, even at a forecast horizon of 12

weeks, the own innovations continue to contribute as much as 70% of the forecast variance of the one-month yield.

Interestingly, the movements in the ten-year yield seem to explain little of the forecast error variance of the other riskless yield series and even of its own error variance. However, in order to understand why the ordering of the variable is often crucial, Table 3.8 shows how the variance decomposition of the ten-year yield changes when the ten-year yield is placed at the top of the ordering in a different Choleski identification scheme.

Variance Decomposition of 10Y Yield

Period	S.E.	10Y Yield	1M Yield	1Y Yield	5Y Yield
1	0.101	100.000	0.000	0.000	0.000
2	0.158	99.939	0.003	0.001	0.057
3	0.202	99.902	0.002	0.001	0.096
4	0.237	99.879	0.004	0.001	0.116
5	0.268	99.860	0.010	0.005	0.125
6	0.296	99.840	0.020	0.010	0.130
7	0.321	99.818	0.032	0.018	0.132
8	0.344	99.794	0.047	0.027	0.132
9	0.365	99.767	0.063	0.039	0.132
10	0.385	99.738	0.079	0.051	0.132
11	0.404	99.707	0.097	0.065	0.131
12	0.422	99.676	0.115	0.080	0.130

Table 3.8 – Forecast error variance decomposition of the ten-year Treasury yields (ten-year yield on the top of the Choleski ordering)

Under this new ordering, most of the forecast error variance at all the 12 horizons considered is explained by its own ten-year yield innovations. Indeed, the estimated correlation coefficients among the innovations of the variables in the VAR(2) are:

$$\widehat{\boldsymbol{\rho}} = \begin{matrix} 1M \\ 1Y \\ 5Y \\ 10Y \end{matrix} \begin{bmatrix} 1 & 0.051 & 0.051 & 0.050 \\ 0.051 & 1 & 0.734 & 0.612 \\ 0.051 & 0.734 & 1 & 0.938 \\ 0.050 & 0.612 & 0.938 & 1 \end{bmatrix}.$$

Interestingly, the correlation between the innovations to ten-year and the five-year yields is very close to one, while for a few additional pairs of reduced-form yield residuals display substantial correlations. As we have learned, when correlation coefficients between the innovation are high, the ordering that a researcher selects to achieve identification may be of crucial importance.

3.3 Granger Causality

Another tool that is useful in order to investigate the dynamic relationships among the variables in a VAR system is **Granger causality** (see Granger, 1969a). Formally, the definition of Granger causality is as follows.

Definition 3.5. **(Granger Causality)** Let $\Im_t$ be the information set containing all the relevant information available up to and including time *t*. In addition, let $\mathbf{y}_t\left(h \mid \Im_t\right)$ be the optimal (minimum MSFE) *h*-step-ahead prediction of the process $\left\{\mathbf{y}_t\right\}$ at the forecast origin *t*, based on the information set $\Im_t$. The vector time series process $\left\{\mathbf{x}_t\right\}$ is said to (Granger-) cause $\left\{\mathbf{y}_t\right\}$ in a Granger sense if and only if $MSE_{yt}\left(h \mid \Im_t\right) < MSE_{yt}\left(h \mid \Im_t \setminus \left\{\mathbf{x}_s \mid s \le t\right\}\right)$.

Alternatively, it is possible to define Granger causality using "its complement" (or lack thereof), i.e., $\left\{\mathbf{x}_t\right\}$ does not cause $\left\{\mathbf{y}_t\right\}$ in a Granger sense at horizon *h*, if taking into account present and past values of $\left\{\mathbf{x}_t\right\}$ does not improve the accuracy of the *h*-step ahead prediction of the future realizations of $\left\{\mathbf{y}_t\right\}$. Finally, if and only if $\left\{\mathbf{x}_t\right\}$ causes $\left\{\mathbf{y}_t\right\}$ and $\left\{\mathbf{y}_t\right\}$ causes $\left\{\mathbf{x}_t\right\}$, then the joint process $\left\{\mathbf{x}'_t, \mathbf{y}'_t\right\}'$ is said to represent a **feedback system**.

Notably, because the information set $\Im_t$ of all the existent relevant information is rarely available to the forecaster, the optimal prediction given $\Im_t$ cannot be determined. Therefore, instead of considering the entire set $\Im_t$, we only consider the information in the past and present values of the process under examination. In addition, instead of comparing optimal predictors, we compare the optimal linear predictors that we have discussed in Section 2.6. Therefore, we can re-write Definition 3.5 as follows.

Definition 3.6. **(Granger Causality - Restricted)** Let $\mathbf{y}_t\left(h \mid \left\{\mathbf{x}_s, \mathbf{y}_s \mid s \le t\right\}\right)$ be the optimal linear (minimum MSFE) *h*-step-ahead prediction function

of the process $\{\mathbf{y}_t\}$ at the forecast origin *t*, based on the information $\{\mathbf{x}_s, \mathbf{y}_s \mid s \le t\}$. The process $\{\mathbf{x}_t\}$ is said to Granger cause $\{\mathbf{y}_t\}$ if

$$MSFE(E[\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \dots, \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \dots]) < MSFE(E[\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \dots])$$

.

Notably, Granger causality is different from exogeneity: indeed, for $\mathbf{y}_t$ to be exogenous it is required that it is not affected by the contemporaneous value of $\mathbf{x}_t$, while Granger causality refers to the effects of the past values of $\{\mathbf{x}_t\}$ on the current value of $\mathbf{y}_t$.

In order to discuss the Granger causal relationship among the variables in a VAR system, let us go back to our bivariate example, and, in particular, to its VMA representation:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \theta_{1,1(i)} & \theta_{1,2(i)} \\ \theta_{2,1(i)} & \theta_{2,2(i)} \end{bmatrix} \begin{bmatrix} u_{1,t-i} \\ u_{2,t-i} \end{bmatrix}. \quad (3.113)$$

It can be proven (see Lütkepohl, 2005), that

$$y_{1,t}\left(1 \mid \{y_{1,s}, y_{2,s} \mid s \le t\}\right) = y_{1,t}\left(1 \mid \{y_{1,s} \mid s \le t\}\right) \Leftrightarrow \theta_{1,2}(i) = 0$$

for $i = 1,2,\dots$. In addition, equality of the one-step-ahead predictors implies the equality of the *h*-step-ahead predictors, for $h = 2,3,\dots$. Therefore, the fact that $\theta_{1,2(i)} = 0$, for $i = 1,2,\dots$ provides a necessary and sufficient condition for $y_{1,s}$ *not* being caused by $y_{2,t}$ in a Granger sense. Therefore, the lack of Granger causality can be easily verified from the VMA representation of the model. In addition, it is worthwhile noting that for a stationary, stable VAR(*p*) process

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{1,0} \\ a_{2,0} \end{bmatrix} + \begin{bmatrix} a_{1,1(1)} & a_{1,2(1)} \\ a_{2,1(1)} & a_{2,2(1)} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \dots + \begin{bmatrix} a_{1,1(p)} & a_{1,2(p)} \\ a_{2,1(p)} & a_{2,2(p)} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \quad (3.115)$$

the condition in (3.115) is satisfied if and only if $a_{1,2(i)} = 0$ for $i = 1, 2, \dots, p$. This implies that the lack of causality can be assessed simply by looking at the representation of the VAR in its standard

form. This means that in this context, the lack of Granger causality can be easily verified by performing a standard *F*-test (like the one discussed in Chapter 1) of the restriction $a_{1,2(1)} = a_{1,2(2)} = ... = a_{1,2(p)} = 0$.

A multivariate generalization of Granger causality leads to **block-exogeneity tests** (or **block-causality tests**, a slightly more precise definition) which are useful to check whether adding a variable into a VAR may increase the accuracy of the forecasts produced by the model. In other words, the test aims at verifying whether one variable, call it $y_{n,t}$, Granger causes any other variables in the system, that is, whether taking into account the lagged value of $y_{n,t}$ helps forecasting any of the other variables in the VAR.

From a practical point of view, block-causality tests simply consist of likelihood ratio tests like the one discussed in Section 2.5:

$$(T-m)\left(\ln\left|\tilde{\boldsymbol{\Sigma}}_u^R\right| - \left|\tilde{\boldsymbol{\Sigma}}_u^U\right|\right) \tag{3.116}$$

where $\tilde{\boldsymbol{\Sigma}}_u^R$ is the covariance matrix of the residuals from a model that has been restricted to have all the coefficients of the lags of the variable $y_{n,t}$ set to zero and $\tilde{\boldsymbol{\Sigma}}_u^U$ is the residual covariance matrix of the unrestricted model. For instance, let us consider a tri-variate VAR(1) model

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} a_{1,0} \\ a_{2,0} \\ a_{3,0} \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \end{bmatrix}. \tag{3.117}$$

Suppose that we want to test whether $y_{3,t}$ Granger causes either $y_{2,t}$ or $y_{1,t}$. In practice, we need to test the restricted model

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} a_{1,0} \\ a_{2,0} \\ a_{3,0} \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{1,2} & 0 \\ a_{2,1} & a_{2,2} & 0 \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \end{bmatrix}, \tag{3.118}$$

vs. the unrestricted model in (3.118) using a LR test. Failure to reject the null hypothesis that the restricted model is sufficient to fit the data (that is, if the calculated value of the statistic is less than the critical value of the χ^2 at a pre-specified size level) means that $y_{3,t}$ Granger does not Granger causes any of the other two

variables in the system. Of course, additional tests may be implemented to separately test whether $y_{3,t}$ Granger causes $y_{1,t}$, $y_{2,t}$, or both.

Example 3.9. To conclude the analysis of the VAR(2) model for the one-month, one-, five-, and ten-year Treasury yields that we have estimated in previous examples, we test Granger causality for all the variables in the model. In particular, Table 3.9 considers one dependent variable at a time and tests whether the lags of each of the other variables help to predict it. In other words, in this case, the *chi*-square statistics refer to a test in which the null is that the lagged coefficients of the "excluded" variable are equal to zero (i.e., the "excluded" variable does not help to forecast the selected dependent variable).

Dependent variable: 1M Yield

Excluded	Chi-sq	df	Prob.
1Y Yield	102.054	2	0.000
5Y Yield	4.965	2	0.084
10Y Yield	1.309	2	0.520
All	180.123	6	0.000

Panel (a)

Dependent variable: 1Y Yield

Excluded	Chi-sq	df	Prob.
1M Yield	33.950	2	0.000
5Y Yield	3.236	2	0.198
10Y Yield	2.714	2	0.257
All	43.161	6	0.000

Panel(b)

Dependent variable: 5Y Yield

Excluded	Chi-sq	df	Prob.
1M Yield	5.630	2	0.060
1Y Yield	3.976	2	0.137
10Y Yield	1.238	2	0.539
All	7.535	6	0.274

Panel (c)

Dependent variable:10Y Yield

Excluded	Chi-sq	df	Prob.
1M Yield	0.940	2	0.625
1Y Yield	1.638	2	0.441
5Y Yield	2.051	2	0.359
All	4.579	6	0.599

Panel (d)

<INSERT TABLE 3.9 HERE>

Table 3.9 – Granger causality tests

Notably, the only lead-lag interactions that seem to be significant at conventional size levels are the following:

- the one-year yield (and five-year yield, at 10% confidence level) Granger causes the 1-month yield;
- the one-month yield Granger causes the one-year and the five-year yields.

Therefore, there is a feedback effect (or two-way causality) between one-month and one-year Treasury yields; the five-year yield and the one-month yield form a feedback system using a p-value of 0.10.

4- Vector Moving Average and Vector Autoregressive Moving Average Models

4.1 Vector Moving Average Models

Although less common in financial applications, a researcher could also specify a vector moving average (VMA) model,

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{u}_t + \boldsymbol{\Theta}_1 \mathbf{u}_{t-1} + \boldsymbol{\Theta}_2 \mathbf{u}_{t-2} + \ldots + \boldsymbol{\Theta}_q \mathbf{u}_{t-q}. \quad (3.119)$$

where $\mathbf{y}_t = \left[y_{1,t} y_{2,t} \ldots y_{K,t} \right]'$, $\mathbf{u}_t$ is a zero-mean multivariate white noise with non-singular covariance matrix $\boldsymbol{\Sigma}_u$ and $\mu = \left[\mu_1 \mu_2 \ldots \mu_K \right]'$ is the mean vector of $\mathbf{y}_t$. It is possible to verify that, exactly as a VAR has an infinite VMA representation, a VMA model potentially has an infinite-order VAR representation. For concreteness, let us consider a VMA(1) model with zero mean (i.e., $\boldsymbol{\mu} = \mathbf{0}$):

$$\mathbf{y}_t = \mathbf{u}_t + \boldsymbol{\Theta}_1 \mathbf{u}_{t-1}. \quad (3.120)$$

It follows that

$$\mathbf{u}_t = \mathbf{y}_t - \boldsymbol{\Theta}_1 \mathbf{u}_{t-1}, \quad (3.121)$$

and thus

$$\mathbf{u}_{t-1} = \mathbf{y}_{t-1} - \boldsymbol{\Theta}_1 \mathbf{u}_{t-2}. \quad (3.122)$$

Therefore, we can rewrite (3.121) as

$$\mathbf{y}_t = \mathbf{u}_t + \boldsymbol{\Theta}_1 \left(\mathbf{y}_{t-1} - \boldsymbol{\Theta}_1 \mathbf{u}_{t-2} \right). \quad (3.123)$$

By iterative substitution we eventually show that

$$\mathbf{y}_t = -\sum_{i=1}^{\infty} \left(-\boldsymbol{\Theta}_1 \right)^i \mathbf{y}_{t-i} + \mathbf{u}_t, \quad (3.124)$$

which is the infinite-order VAR representation of the process. Note that this is only potentially infinite, because it may be that $(-\boldsymbol{\Theta}_1)^i$ may be equal to zero for some *i* greater than some finite number *p*, so that the VAR representation may in fact turn out to be of finite order *p*. For this representation to be meaningful, $\boldsymbol{\Theta}_1^i$ must approach zero as *i* approaches to infinity, which requires that the eigenvalues of $\boldsymbol{\Theta}_1$ are less than one in modulus, that is:

$$\det\left(\mathbf{I}_K - (-\boldsymbol{\Theta}_1) z \right) = \det\left(\mathbf{I}_K + \boldsymbol{\Theta}_1 z \right) \neq 0, \text{ for } z \in \mathbb{C}, \; |z| \leq 1. \quad (3.125)$$

This is the same condition that we have discussed for the stability of a VAR(1) model.

In general, a VMA(q) process similar to the one in (3.120) with $\boldsymbol{\mu}=\mathbf{0}$ has a pure VAR representation,

$$\mathbf{y}_t=\sum_{i=1}^{\infty}\Pi_i\mathbf{y}_{t-i}+\boldsymbol{\varepsilon}_t, \tag{3.126}$$

if $\det\left(\mathbf{I}_K+\boldsymbol{\Theta}_1 z+\ldots+\boldsymbol{\Theta}_q z^q\right)\neq 0$, for $z\in\mathbb{C}$, $|z|\leq 1$. Such a VMA(q) is said to be **invertible**.

We can also examine the first and second moments of a VMA(q). As the multivariate white noise $\boldsymbol{\varepsilon}_t$ has zero mean vector, the mean of $\mathbf{y}_t$ is simply the vector $\boldsymbol{\mu}=\left[\mu_1,\mu_2\ldots,\mu_N\right]$. For the sake of simplicity, in what follows we assume $\boldsymbol{\mu}=\mathbf{0}$. The autocovariance matrices are then

$$\boldsymbol{\Gamma}(h)=E\left(\mathbf{y}_t\mathbf{y}'_{t-h}\right)=\sum_{i=0}^{q-h}\boldsymbol{\Theta}_{i+h}\boldsymbol{\Sigma}_{\varepsilon}\boldsymbol{\Theta}'_i, \text{ for } h=0,1,\ldots,q \tag{3.127}$$

and $\mathbf{0}$ for $h>q$. Clearly, $\boldsymbol{\Gamma}(0)$ is simply the covariance matrix of the series.

Unlikely VAR models, VMA processes can never be simply estimated equation by equation by OLS. One way to estimate them is the maximum likelihood approach, more precisely by a maximum conditional-likelihood (that assumes $\mathbf{u}_t$ to be equal to zero for $t\leq 0$) or alternatively by exact-likelihood (that treats $\mathbf{u}_t$ for $t\leq 0$ as additional parameters of the model). However, a detailed review of these methods is out of the scope of this book. The interested Reader can find a treatment in Lütkepohl (2005).

4.2 Vector Autoregressive Moving Average Models

For the sake of completeness, we finally introduce vector autoregressive moving average (VARMA) processes, that are VAR models that are allowed to include finite order MA process. The general form of a VARMA(p,q) process with VAR order p and MA order q is

$$\mathbf{y}_t=\mathbf{a}_0+\mathbf{A}_1\mathbf{y}_{t-1}+\ldots+\mathbf{A}_p\mathbf{y}_{t-p}+\mathbf{u}_t+\boldsymbol{\Theta}_1\mathbf{u}_{t-1}+\boldsymbol{\Theta}_2\mathbf{u}_{t-2}+\ldots+\boldsymbol{\Theta}_q\mathbf{u}_{t-q},$$

where $\boldsymbol{u}_t$ is a white noise process with non-singular covariance matrix Σ_u.

A little bit of algebraic manipulation may be worthy in order to better understand the nature of this process. Let us now define $\boldsymbol{\upsilon}_t$ such as

$$\boldsymbol{\upsilon}_t = \mathbf{u}_t + \boldsymbol{\Theta}_1 \mathbf{u}_{t-1} + \boldsymbol{\Theta}_2 \mathbf{u}_{t-2} + ... + \boldsymbol{\Theta}_q \mathbf{u}_{t-q}. \qquad (3.129)$$

If we substitute (3.130) into (3.129), we obtain:

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1 y_{t-1} + ... + \mathbf{A}_p y_{t-p} + \boldsymbol{\upsilon}_t. \qquad (3.130)$$

If this process is stable, that is, if $\det\left(\mathbf{I}_K + \mathbf{A}_1 z + ... + \mathbf{A}_p z^p\right) \neq 0$ for $|z| \leq 1$, it is also stationary and can be re-written in its infinite VMA representation as

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \mathbf{D}_i \boldsymbol{\upsilon}_{t-i} = \boldsymbol{\mu} + \sum_{i=0}^{\infty} \boldsymbol{\Theta}_i \mathbf{u}_{t-i}, \qquad (3.131)$$

that is, a pure VMA process where $\boldsymbol{\mu} = \left(\mathbf{I}_K - \mathbf{A}_1 - ... - \mathbf{A}_p\right)^{-1} \mathbf{a}_0$.

Again, to compute the autocovariance matrices of a VARMA model, we will assume that $\boldsymbol{\mu} = 0$ to simplify the algebra; then we post-multiply (3.131) by $\mathbf{y}'_{t-h}$ and taking its expectation, we have

$$E\left[\mathbf{y}_t \mathbf{y}'_{t-h}\right] = \mathbf{A}_1 E\left[\mathbf{y}_{t-1} \mathbf{y}'_{t-h}\right] + ... + \mathbf{A}_p E\left[\mathbf{y}_{t-p} \mathbf{y}'_{t-h}\right] + E\left[\mathbf{u}_t \mathbf{y}'_{t-h}\right] + \boldsymbol{\Theta}_1 E\left[\mathbf{u}_{t-1} \mathbf{y}'_{t-h}\right] +$$

given that $E\left[\mathbf{u}_{t-1} \mathbf{y}'_s\right] = \mathbf{O}$ for any $s < t$. Hence, for $h > q$ we can show that:

$$\boldsymbol{\Gamma}(h) = \mathbf{A}_1 \boldsymbol{\Gamma}(h-1) + ... + \mathbf{A}_p \boldsymbol{\Gamma}(h-p). \qquad (3.133)$$

If $p > q$ and $\boldsymbol{\Gamma}(0), ..., \boldsymbol{\Gamma}(p-1)$ are known, the relationship in (3.134) can be used to compute the autocovariance matrices recursively from $h = p, p+1, ...$.

Noticeably, as for the VMA model, also a VARMA model cannot be simply estimated by OLS, but it requires maximum likelihood estimation. The interested Reader may find more details about the estimation of VARMA models in Lütkepohl (2005).

References

Akaike, H., Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21, 243-247, 1969.

Efron, B., and Tibshirani, R., Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54-75, 1986.

Fuller, W. A., *Introduction to Statistical Time Series*, John Wiley, New York, 1976.

Granger, C. W. J., Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, 37, 424-438, 1969a.

Granger, C. W. J., Prediction with a generalized cost of error function, *Operations Research Quarterly*, 20, 199-207, 1969b.

Granger, C. W. J. and Newbold, P., *Forecasting Economic Time Series*, 2nd edition, Academic Press, New York, 1986.

Hoffman, D., L., and Schlagenhauf D., An econometric investigation of the monetary neutrality and rationality propositions from an international perspective. *Review of Economics and Statistics*, 64, 562-571, 1982.

Kheoh, T.S. & McLeod, A.I., Comparison of two modified portmanteau tests for model adequacy, *Computational Statistics and Data Analysis*, 14, 99-106, 1992.

Kilian, L., Small-sample confidence intervals for impulse response functions, *Review of Economics and Statistics*, 80, 218-230, 1998.

Li, W.K., *Diagnostic Checks in Time Series*, New York: Chapman and Hall/CRC, 2004.

Ljung, G. M. and Box, G. E. P, On a measure of lack of fit in time series models. *Biometrika*, 65, 297-303, 1978.

Hosking, J. R. M., The multivariate portmanteau statistic, *Journal of the American Statistical Association*, 75, 602-608, 1980.

Hosking, J. R. M., Equivalent forms of the multivariate portmanteau statistic, *Journal of the Royal Statistical Society*, B43, 261–262, 1981a.

Li, W. K. and McLeod, A. I., Distribution of the residual autocorrelations in multivariate ARMA time series models, *Journal of the Royal Statistical Society*, B43, 231-239, 1981.

Lütkepohl, H., *Introduction to Multiple Time Series Analysis*. Springer, Berlin, 1991.

Lütkepohl, H., *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, 2005.

Sims, C. A., Macroeconomics and reality, *Econometrica*, 48:1–48, 1980.

Sims, C. A., An autoregressive index model for the U.S. 1948-1975, in J. Kmenta & J. B. Ramsey (Editions), *Large-Scale Macro-Econometric Models*, North-Holland, Amsterdam, pp. 283–327, 1981.

Reinsel, G. C., *Elements of Multivariate Time Series Analysis*, Springer, New York, 1993.

Zellner, A., An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias, *Journal of the American Statistical Association* 57: 348–368, 1962.