$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_t + \mathbf{S}_t + \mathbf{W}_t$$

where

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-p+1} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I}_n & 0 & \dots & 0 & 0 \\ 0 & \mathbf{I}_n & \dots & 0 & 0 \\ 0 & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{I}_n & 0 \end{bmatrix}, \mathbf{S}_t = \begin{bmatrix} \mathbf{s}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{W}_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$\mathbf{X}_t$, $\mathbf{S}_t$, and $\mathbf{W}_t$ are $np \times 1$ vectors and $\mathbf{A}$ is a $np \times np$ square matrix. In order to compute explicit solutions of higher-order VAR processes, we have therefore only to consider VAR(1) models.

It can be demonstrated that the reverse characteristic equation of this VAR(1) system and that of the original VAR(p) system have the same roots.

Solving Stable VAR(1) Processes

We can now proceed to show how solutions to stable VAR models can be computed. Given the equivalence between VAR(1) and VAR(p) we will only consider VAR(1) models. We first consider stable processes that start in the infinite past and then move to possibly unstable processes that start at a given point in time from some initial conditions.

Consider an n-dimensional VAR(1) model,

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{v} + \varepsilon_t, \; t = 0, \pm 1, \pm 2, \dots$$

where the deterministic term, $\mathbf{v}$, is a constant vector. Suppose that the roots of its characteristic equation

$$\det(\mathbf{I}z - \mathbf{A}) = 0$$

lie inside the unit circle. The solutions of this equation are the eigenvalues of the matrix $\mathbf{A}$. Therefore, in the case of a stable process, all the eigenvalues of the matrix $\mathbf{A}$ have modulus less than one. Note that we here express the stability condition in terms of the characteristic equation, while in a previous section we used the reverse characteristic equation.

As the VAR operator is stable, the process is stationary and invertible. Being that it is a VAR(1) process, the infinite moving average polynomial is given by

$$(\mathbf{I} - \mathbf{A}L)^{-1} = \sum_{i=0}^{\infty} \mathbf{A}^i L^i \qquad \mathbf{A}^0 = \mathbf{I}$$

As we saw above in our discussion of stability, stationarity, and invertibility, an invertible process can be represented as follows:

$$\mathbf{x}_t = \mathbf{u} + \left(\sum_{i=0}^{\infty} \mathbf{A}^i L^i \right) \varepsilon_t$$

where

$$\mathbf{u} = \left(\sum_{i=0}^{\infty} \mathbf{A}^i \right) \mathbf{v}$$

is the constant mean of the process

$$\mathbf{u} = E[\mathbf{x}_t]$$

We now compute the autocovariances of the process. It can be demonstrated that the time-invariant autocovariances of the process are

$$\mathbf{\Gamma}_h = E[(\mathbf{x}_t - \mathbf{u})(\mathbf{x}_{t-h} - \mathbf{u})'] = \sum_{i=0}^{\infty} \mathbf{A}^{i+h} \mathbf{\Omega} (\mathbf{A}^i)'$$

where $\mathbf{\Omega}$ is the variance-covariance matrix of the noise term. This expression involves an infinite sum of matrices. While it is not convenient for practical computations, it can be demonstrated that the following recursive matrix equations hold:

$$\mathbf{\Gamma}_h = \mathbf{A}\mathbf{\Gamma}_{h-1}$$

These equations are called *Yule-Walker equations*. They are the *multivariate* equivalent of the Yule-Walker equations that we defined for univariate ARMA processes.

Yule-Walker equations can be used to compute the process autocovariances (see chapters 6 and 7 for the definition of these terms) recursively—provided that we know $\mathbf{\Gamma}_0$. Note that $\mathbf{\Gamma}_0$ is the variance-covariance matrix of the *process*, which is different from the variance-covariance matrix $\mathbf{\Omega}$ of the noise term. It can demonstrated that $\mathbf{\Gamma}_0$ satisfies

$$\mathbf{\Gamma}_0 = \mathbf{A}\mathbf{\Gamma}_0\mathbf{A}' + \mathbf{\Omega}$$

which allows to compute $\mathbf{\Gamma}_0$ via

$$\text{vec}(\mathbf{\Gamma}_0) = (\mathbf{I} - \mathbf{A} \otimes \mathbf{A})^{-1}\text{vec}(\mathbf{\Omega})$$

The vec operation and the Kronecker product $\otimes$ are defined in the appendix to this chapter.

To explicitly compute solutions, consider separately the case of distinct roots and the case of at least two coincident roots. Suppose first that the matrix $\mathbf{A}$ has distinct eigenvalues $(\lambda_1, \ldots, \lambda_n)$ (see the appendix to this chapter for a definition of eigenvalues) and distinct eigenvectors $(\xi_1, \ldots, \xi_n)$. The matrix $\mathbf{A}$ is thus nonsingular and can be represented as: $\mathbf{A} = \mathbf{\Xi}\mathbf{\Lambda}\mathbf{\Xi}^{-1}$ where $\Xi = [\xi_1, \ldots, \xi_n]$ is a nonsingular matrix whose columns are the eigenvectors and

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$.

Consider the process solution

$$\mathbf{x}_t = \mathbf{u} + \left(\sum_{i=0}^{\infty} \mathbf{A}^i L^i \right) \varepsilon_t$$

The infinite matrix series on the right-hand side converges as the eigenvalues of the matrix $\mathbf{A}$ have modulus less than one. In fact we can write

$$\mathbf{A} = \mathbf{\Xi}\Lambda\Xi^{-1}$$

$$\mathbf{A}^i = \overbrace{\mathbf{\Xi}\Lambda\Xi^{-1} \ldots \mathbf{\Xi}\Lambda\Xi^{-1}}^{i \text{ times}} = \mathbf{\Xi}\Lambda^i\Xi^{-1}$$

$$\mathbf{\Lambda}^i = \begin{pmatrix} \lambda_1^i & & 0 \\ & \ddots & \\ 0 & & \lambda_n^i \end{pmatrix}$$

and

$$(\mathbf{I} - \mathbf{A}L)^{-1} = \sum_{i=0}^{\infty} \mathbf{A}^i L^i = \sum_{i=0}^{\infty} \mathbf{\Xi}\Lambda^i \Xi^{-1} L^i$$

The process solution can therefore be written as

$$\mathbf{x}_t = \mathbf{u} + \left(\sum_{i=0}^{\infty} \mathbf{\Xi}\Lambda^i \Xi^{-1} L^i \right) \varepsilon_t$$

The process can be represented as a constant plus an infinite moving average of past noise terms weighted with exponential terms.

Solving Stable and Unstable Processes with Initial Conditions

In the previous section we considered stable, stationary systems defined on the entire time axis. In practice, however, most models start at a given time. If the system starts at a given moment with given initial conditions, it need be neither stable nor stationary. Consider an *n*-dimensional VAR(1) model

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{s}_t + \varepsilon_t, \quad t = 1, 2, \ldots$$

together with initial conditions.

Suppose the VAR(1) model starts at $t = 0$ and suppose that the initial conditions $\mathbf{x}_0$ are given. *The solution of the model is the sum of the general solution of the associated homogeneous system with the given initial conditions plus a particular solution.* The *general* solution can be written as follows:

$$\mathbf{x}_G(t) = \mathbf{\Xi}\mathbf{\Lambda}^t \mathbf{c} = c_1 \lambda_1^t \xi_1 + \ldots + c_n \lambda_n^t \xi_n$$

with constants *c* determined in function of initial conditions. A *particular* solution can be written as

$$\mathbf{x}_t = \sum_{i=0}^{t-1} \mathbf{\Xi}\mathbf{\Lambda}^i \mathbf{\Xi}^{-1} (\mathbf{s}_{t-i} + \boldsymbol{\varepsilon}_{t-i})$$

All solutions are a sum of a particular solution and the general solution. We can also see that the solution is a sum of the deterministic and stochastic parts:

$$\mathbf{x}_t = \overbrace{\boldsymbol{\Xi}\boldsymbol{\Lambda}^t\mathbf{c} + \sum_{i=0}^{t-1}\boldsymbol{\Xi}\boldsymbol{\Lambda}^i\boldsymbol{\Xi}^{-1}\mathbf{s}_{t-i}}^{\text{Deterministic part}} + \overbrace{\sum_{i=0}^{t-1}\boldsymbol{\Xi}\boldsymbol{\Lambda}^i\boldsymbol{\Xi}^{-1}\boldsymbol{\varepsilon}_{t-i}}^{\text{Stochastic part}}$$

From the above formulas, we can see that the modulus of eigenvalues dictates if past shocks decay, persist, or are amplified.

We now discuss the shape of the deterministic trend under the above assumptions. Recall that the deterministic trend is given by the mean of the process. Let us assume that the deterministic terms are either constant intercepts $\mathbf{s}_t = \mu$ or linear functions $\mathbf{s}_t = \gamma t + \mu$. Taking expectations on both sides of the above equation, we can write

$$E[\mathbf{x}_t] = \boldsymbol{\Xi}\boldsymbol{\Lambda}^t\mathbf{c} + \sum_{i=0}^{t-1}\boldsymbol{\Xi}\boldsymbol{\Lambda}^i\boldsymbol{\Xi}^{-1}\boldsymbol{\mu}$$

in the case of constant intercepts, and

$$E[\mathbf{x}_t] = \boldsymbol{\Xi}\boldsymbol{\Lambda}^t\mathbf{c} + \sum_{i=0}^{t-1}\boldsymbol{\Xi}\boldsymbol{\Lambda}^i\boldsymbol{\Xi}^{-1}(\gamma t + \boldsymbol{\mu})$$

in the case of a linear functions.

As the matrix $\boldsymbol{\Lambda}$ is diagonal, it is clear that the process deterministic trend can have different shapes in function of the eigenvalues. In both cases, the trend can be either a constant, a linear trend, or a polynomial of higher order. *If the process has only one unitary root, then a constant intercept produces a linear trend, while a linear function might produce a constant, linear, or quadratic trend.*

To illustrate the above, consider the following VAR(2) model where we replace the notation x_t with $x(t)$:

$$\begin{aligned} x(t) &= 0.6x(t-1) - 0.1y(t-1) - 0.7x(t-2) + 0.15y(t-2) + \varepsilon_x(t) \\ y(t) &= -0.12x(t-1) + 0.7y(t-1) + 0.22x(t-2) - 0.8y(t-2) + \varepsilon_y(t) \end{aligned}$$

with the following initial conditions at time $t = 1,2$:

$$x(1) = 1 \quad x(2) = 1.2 \quad y(1) = 1.5 \quad y(2) = -2$$

It can be transformed into a VAR(1) model as follows:

$$\begin{aligned}
x(t) &= 0.6x(t-1) - 0.1y(t-1) - 0.7z(t-1) + 0.15w(t-1) + \varepsilon_x(t) \\
y(t) &= -0.12x(t-1) + 0.7y(t-1) + 0.22z(t-1) - 0.8w(t-1) + \varepsilon_y(t) \\
z(t) &= x(t-1) \\
w(t) &= y(t-1)
\end{aligned}$$

with the following initial conditions:

$$x(2) = 1.2 \quad y(2) = -2 \quad z(2) = 1 \quad w(2) = 1.5$$

Note that now we have defined four initial conditions at $t = 2$.

The coefficient matrix

$$\mathbf{A} = \begin{pmatrix} 0.6 & -0.1 & -0.7 & 0.15 \\ -0.12 & 0.7 & 0.22 & -0.8 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

has four complex eigenvalues:

$$\mathbf{\Lambda} = \begin{bmatrix} 0.2654+0.7011i & 0 & 0 & 0 \\ 0 & 0.2654-0.7011i & 0 & 0 \\ 0 & 0 & 0.3846+0.8887i & 0 \\ 0 & 0 & 0 & 0.3846-0.8887i \end{bmatrix}$$

The corresponding eigenvector matrix (columns are the eigenvectors) is

$$\mathbf{\Xi} = \begin{bmatrix} 0.1571+0.4150i & 0.1571-0.4150i & -0.1311-0.3436i & -0.1311+0.3436i \\ 0.0924+0.3928i & 0.0924-0.3928i & 0.2346+0.5419i & 0.2346-0.5419i \\ 0.5920 & 0.5920 & -0.3794+0.0167i & -0.3794+0.0167i \\ 0.5337+0.0702i & 0.5337-0.0702i & 0.6098 & 0.6098 \end{bmatrix}$$

The general solution can be written as

$$\mathbf{x}_G = c_1 0.7497^t \cos(1.2090t + \rho_1) + c_2 0.9684^t \cos(1.1623t + \rho_2)$$

STATIONARY AUTOREGRESSIVE DISTRIBUTED LAG MODELS

An important extension of pure VAR models is given by the family of *autoregressive distributed lag* (ARDL) models. The ARDL model is essentially the coupling of a regression model and a VAR model. The ARDL model is written as follows:

$$\mathbf{y}_t = \mathbf{v} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Phi}_s \mathbf{y}_{t-s} + \mathbf{P}_0 \mathbf{x}_t + \cdots + \mathbf{P}_q \mathbf{x}_{t-q} + \boldsymbol{\eta}_t$$
$$\mathbf{x}_t = \mathbf{A}_1 \mathbf{x}_{t-1} + \cdots + \mathbf{A}_p \mathbf{x}_{t-p} + \boldsymbol{\varepsilon}_t$$

In the ARDL model, a variable $\mathbf{y}_t$ is regressed over its own lagged values and over the values of another variable $\mathbf{x}_t$, which follows a VAR(p) model. Both the $\boldsymbol{\eta}_t$ and the $\boldsymbol{\varepsilon}_t$ terms are assumed to be white noise with a time-invariant covariance matrix.

The previous ARDL model can be rewritten as a VAR(1) model as follows:

$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-s+2} \\ \mathbf{y}_{t-s+1} \\ \mathbf{x}_t \\ \vdots \\ \vdots \\ \vdots \\ \mathbf{x}_{t-p} \end{pmatrix} = \begin{pmatrix} \mathbf{v} \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \cdots & \mathbf{\Phi}_{s-1} & \mathbf{\Phi}_s & \mathbf{P}_0 & \mathbf{P}_1 & \cdots & \mathbf{P}_q & \cdots & 0 & 0 \\ \mathbf{I} & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I} & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{I} & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \mathbf{A}_1 & \cdots & \mathbf{A}_q & \cdots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ 0 & 0 & \cdots & 0 & 0 & 0 & \mathbf{I} & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & \mathbf{I} & 0 \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-s+1} \\ \mathbf{y}_{t-s} \\ \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-q} \\ \vdots \\ \mathbf{x}_{t-p} \\ \mathbf{x}_{t-p-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \boldsymbol{\varepsilon}_t \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

The estimation of the ARDL model can therefore be done with the methods used for VAR models. Coefficients can be estimated with OLS methods and the number of lags can be determined with the AIC or BIC criteria discussed in a previous section.

The ARDL model is quite important in financial econometrics: many models of stock returns are essentially ARDL models. In particular, all models where stock returns are regressed over a number of state variables that follow a VAR model are ARDL models.

VECTOR AUTOREGRESSIVE MOVING AVERAGE MODELS

Vector autoregressive moving average (VARMA) models combine an autoregressive part and a moving average part. In some cases, they can offer a more parsimonious modeling option than a pure VAR model. A VARMA(p,q) model without deterministic component is of the form:

$$\begin{aligned}\mathbf{A}(L)\mathbf{x}_t &= \mathbf{B}(L)\boldsymbol{\varepsilon}_t \\ \mathbf{A}(L) &= \mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_p L^p \\ \mathbf{B}(L) &= \mathbf{I} + \mathbf{B}_1 L + \mathbf{B}_2 L^2 + \cdots + \mathbf{B}_q L^q\end{aligned}$$

A VARMA model has two characteristic equations:

$$\begin{aligned}\det(\mathbf{A}(z)) &= 0 \\ \det(\mathbf{B}(z)) &= 0\end{aligned}$$

If the roots of the equation $\det(\mathbf{A}(z)) = 0$ are all strictly outside the unit circle, then the process is stable and can be represented as an infinite moving average. If the roots of the equation $\det(\mathbf{B}(z)) = 0$ are all strictly outside the unit circle, then the process is invertible and can be represented as an infinite autoregressive process. Both representations require the process to be defined for $-\infty < t < \infty$.

If the process starts at $t = 0$, the theory developed above in our discussion of VAR models can be applied. The process can be reduced to a VAR(1) model and then solved with the same methods.

Integrated Processes

Recall that a process is strictly stationary if the joint distribution of a finite collection x_t, x_{t-1},..., x_{t-k} does not vary with t; it is covariance-stationary if its first and second moments are time-invariant. Stationarity does not imply weak stationarity as distributions might have infinite

first or second moments. A process described by a *vector difference equation* (VDE) is stationary if the VDE is stable, that is, if the solutions of the characteristic equations lie strictly outside the unit circle.

A process is said to be *integrated* of order one if its first differences form a stationary process. Recursively, we can define a process integrated of order n if its first differences are integrated of order $n - 1$. An arithmetic random walk is a process integrated of order one as its differences are stationary. However, not all integrated processes are random walks as the definition of stationarity does not assume that processes are generated as IID sequences. In other words, a stationary process can exhibit autocorrelation.

Consider a multivariate process $\mathbf{x}_t$. The process $\mathbf{x}_t$ is said to be integrated of order d if we can write

$$(I - L)^d \mathbf{x}_t = \mathbf{y}_t$$

where $\mathbf{y}_t$ is a stationary process. Suppose that $\mathbf{x}_t$ can be represented by a VAR process,

$$(\mathbf{I} - \mathbf{A}_1 L - \mathbf{A}_2 L^2 - \cdots - \mathbf{A}_p L^p)\mathbf{x}_t = \Phi(\mathbf{A})\mathbf{x}_t = \boldsymbol{\varepsilon}_t$$

The process $\mathbf{x}_t$ is said to be integrated of order d if we can factorize Φ as follows:

$$\mathbf{A}(L)\mathbf{x}_t = (1 - L)^d \mathbf{C}(L)\mathbf{x}_t = \boldsymbol{\varepsilon}_t$$

where $\Psi(L)$ is a stable VAR process that can be inverted to yield

$$(1 - L)^d \mathbf{x}_t = \mathbf{C}(L)^{-1}\boldsymbol{\varepsilon}_t = \sum_{i=0}^{\infty} \mathbf{C}_i \boldsymbol{\varepsilon}_{t-i}$$

In particular, an integrated process with order of integration $d = 1$ admits the following representation:

$$\Delta \mathbf{x}_t = (1 - L)\mathbf{x}_t = \mathbf{C}(L)^{-1}\boldsymbol{\varepsilon}_t = \left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i L^i \right) \boldsymbol{\varepsilon}_t$$

The above definition can be generalized to allow for different orders of integration for each variable.

It is clear from the above definition that the characteristic equation of a process integrated of order d has d roots equal to 1.

Stochastic and Deterministic Trends

An integrated process is characterized by the fact that past shocks never decay. In more precise terms, we can demonstrate that an integrated process can be decomposed as the sum of three components: a deterministic trend, a stochastic trend, and a cyclic stationary process. To see this, consider first a process integrated of order 1 and without a constant intercept, given by

$$\Delta \mathbf{x}_t = \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t = \left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i L^i \right) \boldsymbol{\varepsilon}_t$$

Let's rewrite $\boldsymbol{\Psi}(L)$ as

$$\boldsymbol{\Psi}(L) = \boldsymbol{\Psi} + (1-L)\left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i^* L^i \right)$$

where $\boldsymbol{\Psi}(1) = \boldsymbol{\Psi}$. We can now write the process $\mathbf{x}_t$ as follows:

$$\Delta \mathbf{x}_t = (1-L)\mathbf{x}_t = \left[\boldsymbol{\Psi} + (1-L)\left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i^* L^i \right) \right] \boldsymbol{\varepsilon}_t$$

or, dividing by $(1 - L)$:

$$\mathbf{x}_t = \frac{\boldsymbol{\Psi}}{1-L}\boldsymbol{\varepsilon}_t + \left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i^* L^i \right) \boldsymbol{\varepsilon}_t$$

$$\mathbf{x}_t = \boldsymbol{\Psi} \sum_{i=1}^{t} \boldsymbol{\varepsilon}_i + \left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i^* L^i \right) \boldsymbol{\varepsilon}_t$$

The process $\mathbf{x}_t$ is thereby decomposed into a stochastic trend,

$$\boldsymbol{\Psi} \sum_{i=1}^{t} \boldsymbol{\varepsilon}_i$$

and a stationary component

$$\left(\sum_{i=0}^{\infty} \mathbf{\Psi}_i^* L^i\right)\boldsymbol{\varepsilon}_t$$

The difference between the two terms should be clearly stated: The stochastic term is a sum of shocks that never decay, while in the stationary term past shocks decay due to the weighting matrices $\mathbf{\Psi}_i^*$.

An eventual deterministic trend is added to the stochastic trend and to the stationary component. A constant intercept produces a linear trend or a constant. In fact, if we add a constant intercept $\mathbf{v}$ we can write

$$\Delta \mathbf{x}_t = (1-L)\mathbf{x}_t = \mathbf{v} + \left[\mathbf{\Psi} + (1-L)\left(\sum_{i=0}^{\infty} \mathbf{\Psi}_i^* L^i\right)\right]\boldsymbol{\varepsilon}_t$$

which implies

$$\mathbf{x}_t = \frac{\mathbf{\Psi}}{1-L}\boldsymbol{\varepsilon}_t + \left(\sum_{i=0}^{\infty} \mathbf{\Psi}_i^* L^i\right)\boldsymbol{\varepsilon}_t + \frac{\mathbf{\Psi}}{1-L}\mathbf{v}$$

$$\mathbf{x}_t = \mathbf{\Psi}\sum_{i=1}^{t} \boldsymbol{\varepsilon}_i + \left(\sum_{i=0}^{\infty} \mathbf{\Psi}_i^* L^i\right)\boldsymbol{\varepsilon}_t + t\mathbf{u}$$

where $\mathbf{u} = \Psi\mathbf{v}$. The term $\mathbf{u}$ can be zero even if the intercept $\mathbf{v}$ is different from zero.

A process $\mathbf{x}_t$ is called *trend stationary* if it is the sum of a deterministic trend plus a stationary component, that is if

$$\mathbf{x}_t = \mathbf{s}_t + \mathbf{\Psi}_t$$

A process is called *difference stationary* if it becomes stationary after differencing. A difference-stationary process is the sum of a stochastic trend plus a stationary process.

FORECASTING WITH VAR MODELS

One of the key objectives of financial modeling is forecasting. Forecasting entails a criterion for forecasting as we have to concentrate a proba-

bility distribution in a point forecast. A widely used criterion is the minimization of the *mean square error* (MSE). Suppose that a process $\mathbf{y}_t$ is generated by a VAR(p) process. It can be demonstrated that the optimal h-step ahead forecast according to the MSE criterion is the conditional expectation:

$$E_t(\mathbf{y}_{t+h}) \equiv E(\mathbf{y}_{t+h} | \mathbf{y}_s, s \leq t)$$

If the error terms are strict white noise, then the optimal forecast of a VAR model can be computed as follows:

$$E_t(\mathbf{y}_{t+h}) = \mathbf{v} + \mathbf{A}_1 E_t(\mathbf{y}_{t+h-1}) + \cdots + \mathbf{A}_p E_t(\mathbf{y}_{t+h-p})$$

This formula remains valid if the noise term is a martingale difference sequence (see Chapter 6 for a definition). If the error term is white noise, the above forecasting formula will be the best linear predictor.

APPENDIX: EIGENVECTORS AND EIGENVALUES

Consider a square $n \times n$ matrix $\mathbf{A}$ and a n-vector $\mathbf{x}$. We call *eigenvectors* of the matrix $\mathbf{A}$ those vectors such that the following relationship holds

$$\mathbf{Ax} = \lambda \mathbf{x}$$

for some real number λ. Given an eigenvector $\mathbf{x}$ the corresponding λ is called an eigenvalue. Zero is a trivial eigenvalue. Nontrivial eigenvalues are determined by finding the solutions of the equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

where $\mathbf{I}$ is the identity matrix. A $n \times n$ matrix has at most n distinct eigenvalues and eigenvectors.

Vectoring Operators and Tensor Products

We first define the *vec operator*. Given an $m \times n$ matrix,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

the vec operator, written as vec(**A**),[5] stacks the matrix columns in an $mn \times 1$ vector as follows:

$$\text{vec}(\mathbf{A}) = \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \\ \vdots \\ a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix}$$

Next it is useful to define the Kronecker product. Given the $m \times n$ matrix,

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

and the $p \times q$ matrix,

$$\mathbf{B} = \begin{pmatrix} b_{11} & \cdots & b_{1q} \\ \vdots & \ddots & \vdots \\ b_{p1} & \cdots & b_{pq} \end{pmatrix}$$

we define the Kronecker product $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$ as follows:

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}$$

The Kronecker product, also called the *direct product* or *the tensor product*, is an $(mp) \times (nq)$ matrix. It can be demonstrated that the tensor product satisfies the associative and distributive property and that, given any four matrices **A**, **B**, **C**, **D** of appropriate dimensions, the following properties hold:

$$(\mathbf{C}' \otimes \mathbf{A})\text{vec}(\mathbf{B}) = \text{vec}(\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})$$

[5] The vec operator should not be confused with the vech operator which is similar but not identical. The vech operator stacks the terms below and on the diagonal.

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$$

$$(\mathbf{A} \otimes \mathbf{B})' = (\mathbf{A}') \otimes (\mathbf{B}')$$

$$\text{Trace}(\mathbf{A'BCD'}) = (\text{vec}(\mathbf{A}))'(\mathbf{D} \otimes \mathbf{B})\text{vec}(\mathbf{C})$$

CONCEPTS EXPLAINED IN THIS CHAPTER (IN ORDER OF PRESENTATION)

Vector autoregressive (VAR) models
Wold decomposition theorem
Reverse characteristic equation
Characteristic equation
Stability conditions
Stable processes
Invertible processes
Infinite moving average representation
Absolutely summable sequences
Solutions of a VAR process
Equivalence VAR(p) and VAR(1)
Yule-Walker equations
  Computing autocovariances with Yule-Walker equations
General solutions of a homogeneous system
Particular solutions
Linear and quadratic trends
Autoregressive distributed lag (ARDL) models
Vector autoregressive moving average (VARMA) models
Integrated processes
Vector difference equations (VDE)
Stochastic trends
Trend stationary processes
Difference stationary processes
Forecasting
Mean square error
Eigenvalues and eigenvectors
Vec operator
Kronecker product
Tensor product

CHAPTER 10

Vector Autoregressive Models II

In this chapter we discuss estimation methods for vector autoregressive (VAR) models. We first consider estimation of stable systems. The key result here is that stable VAR systems can be conveniently estimated with least squares methods. We then proceed to the estimation of unstable systems.

ESTIMATION OF STABLE VAR MODELS

When discussing the estimation of regression models in Chapter 3, we introduced two main methods for estimating linear regressions: the least squares method and the maximum likelihood method. These methods apply immediately to unrestricted stable VAR models. Note that models are said to be "unrestricted" if the estimation process is allowed to determine any possible outcome, and "restricted" if the estimation procedure restricts parameters in some way.

Suppose that a time series is given and that the data generating process (DGP) of the series is the VAR(p) model:

$$\mathbf{x}_t = \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_1\mathbf{x}_{t-1} + \cdots + \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{v} + \boldsymbol{\varepsilon}_t$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{N,t})'$ is a N–dimensional stochastic time series in vector notation; $\mathbf{A}_i$ are deterministic $N \times N$ matrices; $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{N,t})'$ is a multivariate white noise with variance-covariance matrix Σ; and $\mathbf{v} = (v_1, \ldots, v_N)'$ is a vector of constants.

Let's first assume that stability condition

$$\det(\mathbf{A}(z)) \neq 0 \text{ for } |z| \leq 1$$

holds, that is, the roots of the reverse characteristic equation are strictly outside of the unit circle. The result is that the VAR(p) model is stable and the corresponding process stationary. We will consider processes that start at $t = 1$, assuming that p initial conditions: $\mathbf{x}_{-p+1}, \ldots, \mathbf{x}_0$ are given. In this case, stable VAR models yield asymptotically stationary processes.

Recall that the above N-dimensional VAR(p) model is equivalent to the following Np-dimensional VAR(1) model:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{V} + \mathbf{U}_t$$

where

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-p+1} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I}_N & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{I}_N & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{I}_N & 0 \end{bmatrix},$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{v} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{U}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Matrix $\mathbf{A}$ is called the *companion matrix* of the VAR(p) system.

Given that the VAR(p) model is unrestricted, it can be estimated as any linear regression model. As we consider only consistent estimators, the estimated parameters (in the limit of an infinite sample) satisfy the stability condition. However on a finite sample, the estimated parameters might not satisfy the stability condition.

We will first show how the estimation of a VAR(p) model and its VAR(1) equivalent can be performed with least squares and maximum likelihood methods. To do so we apply the estimation theory developed in Chapter 3, estimating the model coefficients either by the multivariate least squares method or by the maximum likelihood method.

Multivariate Least Squares Estimation

Conceptually, the multivariate *least squares* (LS) estimation method is equivalent to that of a linear regression (see Chapter 3); the notation, however, is more complex. This is because we are dealing with multiple time series and because noise terms are correlated. Similar to what we

did in estimating regressions (Chapter 3), we represent the autoregressive process as a single-matrix equation. We will introduce two different but equivalent notations.

Suppose that a sample of T observations of the N-variate variable $\mathbf{x}_t$, $t = 1, \ldots, T$ and a presample of p initial conditions $\mathbf{x}_{-p+1}, \ldots, \mathbf{x}_0$ are given. We first stack all observations $\mathbf{x}_t$, $t = 1, \ldots, T$ in a vector

$$\mathbf{x} = \begin{pmatrix} x_{1,1} \\ \vdots \\ x_{N,1} \\ \vdots \\ \vdots \\ x_{1,T} \\ \vdots \\ x_{N,T} \end{pmatrix}$$

Introducing a notation that will be useful later, we can also write

$$\mathbf{x} = \text{vec}(\mathbf{X})$$

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) = \begin{pmatrix} x_{1,1} & \cdots & x_{1,T} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,T} \end{pmatrix}$$

In other words, $\mathbf{x}$ is a $(NT \times 1)$ vector where all observations are stacked, while $\mathbf{X}$ is a $(N \times T)$ matrix where each column represents an N-variate observation.

Proceeding analogously with the noise terms, we stack the noise terms in a $(NT \times 1)$ vector as follows:

$$\mathbf{u} = \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{N,1} \\ \vdots \\ \vdots \\ \varepsilon_{1,T} \\ \vdots \\ \varepsilon_{N,T} \end{pmatrix}$$

We can represent this alternatively as follows:

$$\mathbf{u} = \text{vec}(\mathbf{U})$$

$$\mathbf{U} = \begin{pmatrix} \varepsilon_{1,1} & \cdots & \varepsilon_{1,T} \\ \vdots & \ddots & \vdots \\ \varepsilon_{N,1} & \cdots & \varepsilon_{N,T} \end{pmatrix}$$

where $\mathbf{U}$ is a $(N \times T)$ matrix such that each column represents an n-variate innovation term.

The noise terms are assumed to have a nonsingular covariance matrix,

$$\Sigma = [\sigma_{i,j}] = E[\varepsilon_{i,t}\varepsilon_{j,t}]$$

with $E[\varepsilon_{i,t}\,\varepsilon_{j,s}] = 0$, $\forall i, j, t \neq s$. The covariance matrix of $\mathbf{u}$, $\Sigma_{\mathbf{u}}$ can be written as

$$\Sigma_{\mathbf{u}} = \mathbf{I}_T \otimes \Sigma = \begin{pmatrix} \Sigma & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma \end{pmatrix}$$

In other words, the covariance matrix of $\mathbf{u}$ is a block-diagonal matrix where all diagonal blocks are equal to Σ. This covariance structure reflects the assumed white-noise nature of innovations that precludes autocorrelations and cross autocorrelations in the innovation terms.

Using the notation established above, we can now compactly write the VAR(p) model in two equivalent ways as follows:

$$\mathbf{X} = \mathbf{AW} + \mathbf{U}$$

$$\mathbf{x} = \mathbf{w}\beta + \mathbf{u}$$

The first is a matrix equation where the left and right sides are $N \times T$ matrices such that each column represents the VAR(p) equation for each observation. The second equation, which equates the two NT vectors on the left and right sides, can be derived from the first as follows, using the properties of the vec operator and the Kronecker product established in the appendix to Chapter 9

$$\text{vec}(\mathbf{X}) = \text{vec}(\mathbf{AW}) + \text{vec}\,(\mathbf{U})$$

$$\text{vec}(\mathbf{X}) = (\mathbf{W}' \otimes \mathbf{I}_N)\text{vec}(\mathbf{A}) + \text{vec}(\mathbf{U})$$

$$\mathbf{x} = \mathbf{w}\beta + \mathbf{u}$$

This latter equation is the equivalent of the regression equation established in Chapter 3.

EXHIBIT 10.1 The Regressor Matrix for a N-dimensional VAR(p) model

$$
\mathbf{w} = \begin{bmatrix}
1 & \cdots & 0 & x_{1,0} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,0} & \cdots & 0 & \cdots & \cdots & \cdots & x_{1,1-p} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,1-p} & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & x_{1,0} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,0} & \cdots & \cdots & \cdots & 0 & \cdots & x_{1,1-p} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,1-p} \\
1 & \cdots & 0 & x_{1,1} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,1} & \cdots & 0 & \cdots & \cdots & \cdots & x_{1,2-p} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,2-p} & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & x_{1,1} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,1} & \cdots & \cdots & \cdots & 0 & \cdots & x_{1,2-p} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,2-p} \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & & & & \vdots & \cdots & \vdots & & & & \vdots & \cdots & \vdots & & & & \vdots & \cdots & \vdots \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \ddots & & \vdots & \ddots & \vdots & & \ddots & & \vdots & \ddots & \vdots & & \ddots & & \vdots & \ddots & \vdots \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & & & & \vdots & \cdots & \vdots & & & & \vdots & \cdots & \vdots & & & & \vdots & \cdots & \vdots \\
1 & \cdots & 0 & x_{1,T-2} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,T-2} & \cdots & 0 & \cdots & \cdots & \cdots & x_{1,T-p-1} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,T-p-1} & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & x_{1,T-2} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,T-2} & \cdots & \cdots & \cdots & 0 & \cdots & x_{1,T-p-1} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,T-p-1} \\
1 & \cdots & 0 & x_{1,T-1} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,T-1} & \cdots & 0 & \cdots & \cdots & \cdots & x_{1,T-p} & \cdots & 0 & \cdots & \cdots & \cdots & x_{N,T-p} & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 1 & 0 & \cdots & x_{1,T-1} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,T-1} & \cdots & \cdots & \cdots & 0 & \cdots & x_{1,T-p} & \cdots & \cdots & \cdots & 0 & \cdots & x_{N,T-p}
\end{bmatrix}
$$

Matrix **w** is shown in Exhibit 10.1; matrices **W** and **A** and vector **B** are given by

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{T-2} & \mathbf{x}_{T-1} \\ \mathbf{x}_{-1} & \mathbf{x}_0 & \cdots & \mathbf{x}_{T-3} & \mathbf{x}_{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{x}_{1-p} & \mathbf{x}_{2-p} & \cdots & \mathbf{x}_{T-p-1} & \mathbf{x}_{T-p} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ x_{1,0} & x_{1,1} & \cdots & x_{1,T-2} & x_{1,T-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N,0} & x_{N,1} & \cdots & x_{N,T-2} & x_{N,T-1} \\ x_{1,-1} & x_{1,0} & \cdots & x_{1,T-3} & x_{1,T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N,-1} & x_{N,0} & \cdots & x_{N,T-3} & x_{N,T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{1,1-p} & x_{1,2-p} & \cdots & x_{1,T-p-1} & x_{1,T-p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{N,1-p} & x_{N,2-p} & \cdots & x_{N,T-p-1} & x_{N,T-p} \end{bmatrix}_{(Np+1)\times T}$$

$$\mathbf{A} = (\mathbf{v}, \mathbf{A}_1, \ldots, \mathbf{A}_p)$$

$$= \begin{pmatrix} v_1 & a_{11}^1 & \cdots & a_{1N}^1 & \cdots & \cdots & \cdots & a_{11}^p & \cdots & a_{1N}^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ v_N & a_{N1}^1 & \cdots & a_{NN}^1 & \cdots & \cdots & \cdots & a_{N1}^p & \cdots & a_{NN}^p \end{pmatrix}_{N\times(Np+1)}$$

$$\beta = \text{vec}(\mathbf{A}) = \begin{pmatrix} v_1 \\ \vdots \\ v_N \\ a_{11}^1 \\ \vdots \\ a_{N1}^1 \\ \vdots \\ a_{1N}^p \\ \vdots \\ a_{NN}^p \end{pmatrix}_{N(Np+1)\times 1}$$

To estimate the model, we have to write the sum of the squares of residuals as we did for the sum of the residuals in a regression (see Chapter 3). However, as already mentioned, we must also consider the multivariate nature of the noise terms and the presence of correlations.

Our starting point is the regression equation $\mathbf{x} = \mathbf{w}\beta + \mathbf{u}$, which implies $\mathbf{u} = \mathbf{x} - \mathbf{w}\beta$. As the innovation terms exhibit a correlation structure, we have to proceed as in the case of generalized least squares (GLS). We write the weighted sum of squared residuals as

$$S = \mathbf{u}'\Sigma_{\mathbf{u}}^{-1}\mathbf{u} = \sum_{t=1}^{T} \varepsilon_t'\Sigma^{-1}\varepsilon_t$$

For a given set of observations, the quantity S is a function of the model parameters $S = S(\beta)$. The function S admits the following alternative representation:

$$S(\beta) = \text{trace}[\mathbf{U}'\Sigma_{\mathbf{u}}^{-1}\mathbf{U}] = \text{trace}[(\mathbf{X} - \mathbf{A}\mathbf{W})'\Sigma_{\mathbf{u}}^{-1}(\mathbf{X} - \mathbf{A}\mathbf{W})]$$

Since

$$\begin{aligned} S = \mathbf{u}'\Sigma_{\mathbf{u}}^{-1}\mathbf{u} &= (\text{vec}(\mathbf{U}))'(\mathbf{I}_T \otimes \Sigma)^{-1}\text{vec}(\mathbf{U}) \\ &= (\text{vec}(\mathbf{X} - \mathbf{A}\mathbf{W}))'(\mathbf{I}_T \otimes \Sigma^{-1})\text{vec}(\mathbf{X} - \mathbf{A}\mathbf{W}) \\ &= \text{trace}[(\mathbf{X} - \mathbf{A}\mathbf{W})'\Sigma_{\mathbf{u}}^{-1}(\mathbf{X} - \mathbf{A}\mathbf{W})] \end{aligned}$$

The least squares estimate of the model parameters $\hat{\beta}$, are obtained by minimizing $S = S(\beta)$ with respect to beta requiring

$$\frac{\partial S(\beta)}{\partial \beta} = 0$$

Equating the vector of partial derivatives to zero yields the so-called *normal equations* of the LS method. From

$$\begin{aligned} S = \mathbf{u}'\Sigma_{\mathbf{u}}^{-1}\mathbf{u} &= (\mathbf{x} - \mathbf{w}\beta)'\Sigma_{\mathbf{u}}^{-1}(\mathbf{x} - \mathbf{w}\beta) \\ &= \mathbf{x}'\Sigma_{\mathbf{u}}^{-1}\mathbf{x} + \beta'\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{w}\beta - 2\beta'\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{x} \end{aligned}$$

it follows that the normal equations are given by

$$\frac{\partial S(\beta)}{\partial \beta} = 2\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{w}\beta - 2\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{x} = 0$$

The Hessian matrix turns out as

$$\frac{\partial^2 S(\beta)}{\partial \beta \partial \beta'} = 2\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{w}$$

and is positive definite given our assumptions. Consequently, the LS estimator is

$$\hat{\beta} = (\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{w})^{-1}\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{x}$$

This expression—which has the same form as the Aitkin GLS estimator—is a fundamental expression in LS methods. However, due to the structure of the regressors, further simplifications are possible for a VAR model, namely

$$\hat{\beta} = ((\mathbf{W}\mathbf{W}')^{-1}\mathbf{W} \otimes \mathbf{I}_N)\mathbf{x}$$

are possible. To demonstrate this point, consider the following derivation:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{w})^{-1}\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{x} \\
&= ((\mathbf{W}' \otimes \mathbf{I}_N)'(\mathbf{I}_T \otimes \Sigma)^{-1}(\mathbf{W}' \otimes \mathbf{I}_N))^{-1}(\mathbf{W} \otimes \mathbf{I}_N)(\mathbf{I}_T \otimes \Sigma)^{-1}\mathbf{x} \\
&= ((\mathbf{W} \otimes \mathbf{I}_N)(\mathbf{I}_T \otimes \Sigma^{-1})(\mathbf{W}' \otimes \mathbf{I}_N))^{-1}(\mathbf{W} \otimes \mathbf{I}_N)(\mathbf{I}_T \otimes \Sigma)^{-1}\mathbf{x} \\
&= ((\mathbf{W}\mathbf{I}_T) \otimes (\mathbf{I}_N\Sigma^{-1})(\mathbf{W}' \otimes \mathbf{I}_N))^{-1}(\mathbf{W}\mathbf{I}_T) \otimes (\mathbf{I}_N\Sigma^{-1})\mathbf{x} \\
&= ((\mathbf{W} \otimes \Sigma^{-1})(\mathbf{W}' \otimes \mathbf{I}_N))^{-1}(\mathbf{W} \otimes \Sigma^{-1})\mathbf{x} \\
&= ((\mathbf{W}\mathbf{W}')^{-1} \otimes (\Sigma^{-1}))(\mathbf{W} \otimes \Sigma)\mathbf{x} \\
&= ((\mathbf{W}\mathbf{W}')^{-1}\mathbf{W}) \otimes (\Sigma^{-1}\Sigma)\mathbf{x} \\
&= ((\mathbf{W}\mathbf{W}')^{-1}\mathbf{W} \otimes \mathbf{I}_N)\mathbf{x}
\end{aligned}$$

This derivation shows that, in the case of a stable unrestricted VAR process, the multivariate GLS estimator coincides with the ordinary least squares (OLS) estimator obtained by minimizing the quantity $S = \mathbf{u}'\mathbf{u}$.

We can therefore estimate VAR processes by OLS estimate equation by equation rather than the full N-dimensional system. Computationally, this entails a significant simplification.

We can also write another expression to estimate matrix $\mathbf{A}$. Using $\mathbf{X} = \mathbf{AW} + \mathbf{U}$, we have

$$\hat{\mathbf{A}} = \mathbf{X}\mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}$$

The relationship between $\hat{\mathbf{A}}$ and $\hat{\beta}$ is as follows:

$$\begin{aligned}\hat{\beta} &= ((\mathbf{W}\mathbf{W}')^{-1}\mathbf{W} \otimes \mathbf{I}_N)\mathbf{x} \\ \text{vec}\left(\hat{\mathbf{A}}\right) &= ((\mathbf{W}\mathbf{W}')^{-1}\mathbf{W} \otimes \mathbf{I}_N)\text{vec}(\mathbf{X}) \\ &= \text{vec}(\mathbf{X}\mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1})\end{aligned}$$

To summarize

1. Given a VAR(p) process, the multivariate GLS estimator coincides with the OLS estimator computed equation by equation.
2. The following three expressions for the estimator are equivalent:

$$\hat{\beta} = (\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{w})^{-1}\mathbf{w}'\Sigma_{\mathbf{u}}^{-1}\mathbf{x}$$

$$\hat{\beta} = ((\mathbf{W}\mathbf{W}')^{-1}\mathbf{W} \otimes \mathbf{I}_N)\mathbf{x}$$

$$\hat{\mathbf{A}} = \mathbf{X}\mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}$$

We next discuss the asymptotic distribution of these estimators.

The Asymptotic Distribution of LS Estimators

In Chapter 2 we stated that estimators depend on the sample and are therefore to be considered random variables. To assess the quality of the estimators, the distribution of the estimators must be determined.

It is difficult to calculate the finite sample distributions of the LS estimators of the stationary VAR model. Finite sample properties of a stationary VAR process can be approximately ascertained using Monte Carlo methods.

Significant simplifications arise as the sample size approaches infinity. The essential result is that the model estimators become normally distributed. The asymptotic properties of the LS estimators can be established under additional assumptions on the white noise. Suppose that the white-noise process has finite and bounded fourth moments and that noise variables at different times are independent and not merely uncorrelated as we have assumed thus far. (Note that these conditions are automatically satisfied by any Gaussian white noise.). Under these assumptions, it can be demonstrated that the following properties hold:

- The $((Np + 1) \times (Np + 1))$ matrix

$$\mathbf{\Gamma} := \text{plim}\frac{\mathbf{W}\mathbf{W}'}{T}$$

  exists and is nonsingular.

- The $(N(Np + 1) \times 1)$ vector $\hat{\boldsymbol{\beta}}$ of estimated model parameters is jointly normally distributed:

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N(0, \mathbf{\Gamma}^{-1} \otimes \mathbf{\Sigma})$$

  The $(N(Np + 1) \times N(Np + 1))$ matrix $\mathbf{\Gamma}^{-1} \otimes \mathbf{\Sigma}$ is the covariance matrix of the parameter distribution.

From this it follows that, for large samples, we can approximate matrices $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ by

$$\hat{\mathbf{\Gamma}} = \frac{\mathbf{W}\mathbf{W}'}{T}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{T}\mathbf{X}(\mathbf{I}_T - \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\mathbf{W})\mathbf{X}'$$

Note that these matrices are not needed to estimate the model parameters; they are required only for determining the distribution of the model parameters. If $N = 1$, these expressions are the same as those already established for multivariate regressions. The above estimator of the noise covariance matrix is biased. An unbiased estimator is obtained by multiplying the above by the factor $T/(T - Np - 1)$.

Estimating Demeaned Processes

In previous sections we assumed that the VAR(p) model has a constant intercept and that the process variables have, in general, a nonzero mean. Note that the mean and the intercept are not the same numbers. In fact, given that the process is assumed to be stationary, we can write

$$E(\mathbf{x}_t) = \mathbf{A}_1 E(\mathbf{x}_{t-1}) + \mathbf{A}_2 E(\mathbf{x}_{t-2}) + \cdots + \mathbf{A}_p E(\mathbf{x}_{t-p}) + \mathbf{v}$$

$$\boldsymbol{\mu} - \mathbf{A}_1\boldsymbol{\mu} - \mathbf{A}_2\boldsymbol{\mu} - \cdots - \mathbf{A}_p\boldsymbol{\mu} = \mathbf{v}$$

$$\boldsymbol{\mu} = (\mathbf{I}_N - \mathbf{A}_1 - \mathbf{A}_2 - \cdots - \mathbf{A}_p)^{-1}\mathbf{v}$$

We can recast the previous derivation in a different notation, assuming that the process variables are demeaned. In this case, we can rewrite the VAR process in the following form:

$$(\mathbf{x}_t - \boldsymbol{\mu}) = \mathbf{A}_1(\mathbf{x}_{t-1} - \boldsymbol{\mu}) + \mathbf{A}_2(\mathbf{x}_{t-2} - \boldsymbol{\mu}) + \cdots + \mathbf{A}_p(\mathbf{x}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t$$

Defining the demeaned vector $\mathbf{y}_t = \mathbf{x}_t - \mu$, the VAR process becomes

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t$$

The formulas previously established hold with some obvious changes. We will state them explicitly, as they will be used in the following sections. Defining

$$\begin{aligned}
\mathbf{Y} &= (y_1, \ldots, \mathbf{y}_T) \\
\mathbf{U} &= (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T) \\
\mathbf{y} &= \text{vec}(\mathbf{Y})
\end{aligned}$$

$$\begin{aligned}
\mathbf{u} &= \text{vec}(\mathbf{U}) \\
\boldsymbol{\Sigma}_{\mathbf{u}} &= \mathbf{I}_T \otimes \boldsymbol{\Sigma} \\
\mathbf{A} &= (\mathbf{A}_1, \ldots, \mathbf{A}_p) \\
\alpha &= \text{vec}(\mathbf{A})
\end{aligned}$$

$$\mathbf{Z} = \begin{pmatrix} \mathbf{y}_0 & \cdots & \mathbf{y}_{T-1} \\ \vdots & \ddots & \vdots \\ \mathbf{y}_{1-p} & \cdots & \mathbf{y}_{T-p} \end{pmatrix}$$

$$\mathbf{z} = (\mathbf{Z}' \otimes \mathbf{I}_N)$$

we have

$$\mathbf{y} = \mathbf{z}\alpha + \mathbf{u}$$

$$\mathbf{Y} = \mathbf{AZ} + \mathbf{U}$$

The LS estimators are

$$\hat{\boldsymbol{\alpha}} = (\mathbf{z}'\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{z})^{-1}\mathbf{z}'\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{y}$$

$$\hat{\boldsymbol{\alpha}} = ((\mathbf{ZZ}')^{-1}\mathbf{Z} \otimes \mathbf{I}_N)\mathbf{y}$$

$$\hat{\mathbf{A}} = \mathbf{YZ}'(\mathbf{ZZ}')^{-1}$$

It can be demonstrated that the sample mean,

$$\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$$

is a consistent estimator of the process mean and has a normal asymptotic distribution. If we estimate intercept $\hat{\mathbf{v}}$ from the original (nondemeaned) data, the mean $\hat{\boldsymbol{\mu}}$, is estimated by the mean can be estimated with the following estimator:

$$\hat{\boldsymbol{\mu}} = (\mathbf{I}_N - \mathbf{A}_1 - \mathbf{A}_2 - \cdots - \mathbf{A}_p)^{-1}\hat{\mathbf{v}}$$

This is consistent and follows an asymptotic normal distribution.

We now turn our attention to the maximum likelihood estimation of stable VAR models.

Maximum Likelihood Estimators

Under the assumption of Gaussian innovations, *maximum likelihood* (ML) estimation methods coincide with LS estimation methods when we condition on the first p observations. Recall from Chapter 2 that, given a known distribution, ML methods try to find the distribution parame-

ters that maximize the likelihood function (i.e., the joint distribution of the sample computed on the sample itself). In the case of a multivariate mean-adjusted VAR(p) process, the given sample data are T observations of the N-variate variable $\mathbf{y}_t$, $t = 1, \ldots, T$ and a presample of p initial conditions $\mathbf{y}_{-p+1}, \ldots, \mathbf{y}_0$. If we assume that the process is stationary and that innovations are Gaussian white noise, the variables $\mathbf{y}_t$, $t = 1, \ldots, T$ will also be jointly normally distributed. However, it is advantageous to express the joint distribution of the noise terms in function of the data. As the white noise is assumed to be Gaussian, the noise variables at different times are independent. As observed in **Chapter GG**, this allows considerable simplifications for computing the likelihood function.

The noise terms ($\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T$) are assumed to be independent with constant covariance matrix $\boldsymbol{\Sigma}$ and, therefore, $\mathbf{u} = \text{vec}(\mathbf{U})$ has covariance matrix $\boldsymbol{\Sigma}_\mathbf{u} = \mathbf{I}_T \otimes \boldsymbol{\Sigma}$. Under the assumption of Gaussian noise, $\mathbf{u}$ has the following NT-variate normal density:

$$
\begin{aligned}
f_\mathbf{u}(\mathbf{u}) &= (2\pi)^{-\frac{NT}{2}} \left|\mathbf{I}_T \otimes \boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{u}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{u}\right) \\
&= (2\pi)^{-\frac{NT}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\varepsilon}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_t\right)
\end{aligned}
$$

Using

$$
\begin{aligned}
\boldsymbol{\varepsilon}_1 &= \mathbf{y}_1 - \mathbf{A}_1\mathbf{y}_0 - \mathbf{A}_2\mathbf{y}_{-1} - \cdots - \mathbf{A}_p\mathbf{y}_{1-p} \\
\boldsymbol{\varepsilon}_2 &= \mathbf{y}_2 - \mathbf{A}_1\mathbf{y}_1 - \mathbf{A}_2\mathbf{y}_0 - \cdots - \mathbf{A}_p\mathbf{y}_{2-p} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\boldsymbol{\varepsilon}_p &= \mathbf{y}_p - \mathbf{A}_1\mathbf{y}_0 - \mathbf{A}_2\mathbf{y}_{p-2} - \cdots - \mathbf{A}_p\mathbf{y}_0 \\
\boldsymbol{\varepsilon}_{p+1} &= \mathbf{y}_{p+1} - \mathbf{A}_1\mathbf{y}_p - \mathbf{A}_2\mathbf{y}_{p-2} - \cdots - \mathbf{A}_p\mathbf{y}_1 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\boldsymbol{\varepsilon}_{T-1} &= \mathbf{y}_{T-1} - \mathbf{A}_1\mathbf{y}_{T-2} - \mathbf{A}_2\mathbf{y}_{T-3} - \cdots - \mathbf{A}_p\mathbf{y}_{T-p-1} \\
\boldsymbol{\varepsilon}_T &= \mathbf{y}_T - \mathbf{A}_1\mathbf{y}_{T-1} - \mathbf{A}_2\mathbf{y}_{T-2} - \cdots - \mathbf{A}_p\mathbf{y}_{T-p}
\end{aligned}
$$

written in matrix form

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \\ \varepsilon_{p+1} \\ \vdots \\ \varepsilon_{T-1} \\ \varepsilon_T \end{pmatrix} = \begin{pmatrix} \mathbf{I}_N & 0 & \cdots & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 & 0 & \\ -\mathbf{A}_1 & \mathbf{I}_N & \cdots & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 & 0 & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ -\mathbf{A}_p & -\mathbf{A}_{p-1} & \cdots & -\mathbf{A}_1 & \mathbf{I}_N & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & -\mathbf{A}_p & \cdots & -\mathbf{A}_2 & -\mathbf{A}_1 & \mathbf{I}_N & \cdots & \cdots & 0 & 0 & \cdots & 0 & 0 & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -\mathbf{A}_p & -\mathbf{A}_{p-1} & \cdots & \mathbf{I}_N & 0 & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -\mathbf{A}_p & \cdots & -\mathbf{A}_1 & \mathbf{I}_N & \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_p \\ \mathbf{y}_{p+1} \\ \vdots \\ \mathbf{y}_{T-p} \\ \vdots \\ \mathbf{y}_{T-1} \\ \mathbf{y}_T \end{pmatrix} + \begin{pmatrix} -\mathbf{A}_p & -\mathbf{A}_{p-1} & \cdots & -\mathbf{A}_1 \\ 0 & -\mathbf{A}_p & \cdots & -\mathbf{A}_2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\mathbf{A}_p \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mathbf{y}_{1-p} \\ \mathbf{y}_{2-p} \\ \mathbf{y}_{-1} \\ \mathbf{y}_0 \end{pmatrix}$$

and the model equation $\mathbf{y} = \mathbf{z}\alpha + \mathbf{u}$, we can express the density function in terms of the variables

$$f_y(\mathbf{y}) = \left|\frac{\partial \mathbf{u}}{\partial \mathbf{y}}\right| f_{\mathbf{u}}(\mathbf{u}) = (2\pi)^{-\frac{NT}{2}} \left|\mathbf{I}_T \otimes \boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{z}\alpha)'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{z}\alpha)\right)$$

We can now write the log-likelihood as follows:

$$\begin{aligned} \log(l) &= -\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{u}}\right| - \frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\varepsilon}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_t \\ &= -\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{u}}\right| - \frac{1}{2}(\mathbf{y} - \mathbf{z}\alpha)'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{z}\alpha) \\ &= -\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{u}}\right| - \frac{1}{2}\mathrm{trace}(\mathbf{U}'\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}\mathbf{U}) \\ &= -\frac{NT}{2}\log(2\pi) - \frac{T}{2}\log\left|\boldsymbol{\Sigma}_{\mathbf{u}}\right| - \frac{1}{2}\mathrm{trace}((\mathbf{Y} - \mathbf{A}\mathbf{Z})'\boldsymbol{\Sigma}_{\mathbf{u}}^{-1}(\mathbf{Y} - \mathbf{A}\mathbf{Z})) \end{aligned}$$

Equating the partial derivatives of this expression to zero, we obtain the very same estimators as with the LS method. In the case of Gaussian noise, LS/OLS methods and ML methods yield the same result.

ESTIMATING THE NUMBER OF LAGS

In the previous sections, we assumed that the order p of the model (i.e., the number of lags in the model) is known. The objective of this section is to establish criteria that allow determining *a priori* the correct number of lags. This idea has to be made more precise. We assume, as we did in the previous sections on the estimation of the model coefficients, that the true data generation process is a VAR(p) model. In this case, we expect that the correct model order is exactly p, that is, we expect to come out with a consistent estimate of the model order. This is not the same problem as trying to determine the optimal number of lags to fit a VAR model to a process that is not be generated by a VAR data generating process. We assume that the type of model is correctly specified and discuss methods to estimate the model order under this assumption.

In general, increasing the model order will reduce the size of residuals but tends to reduce the forecasting ability of the model. By increasing the number of parameters, we improve the in-sample accuracy but tend to worsen the out-of-sample forecasting ability. In this section we consider only linear models under the assumption that the data generation process is linear and autoregressive with unknown parameters.

To see how increasing the number of lags can reduce the forecasting ability of the model, consider that the forecasting ability of a linear VAR model can be estimated. Recall from Chapter 9 that the optimal forecast of a VAR model is the conditional mean. This implies that the optimal one-step forecast given the past p values of the process up to the present moment is

$$\hat{\mathbf{x}}_{t+1} = \mathbf{A}_1\mathbf{x}_t + \mathbf{A}_2\mathbf{x}_{t-1} + \cdots + \mathbf{A}_p\mathbf{x}_{t-p+1} + \mathbf{v}$$

The forecasting *mean square error* (MSE) can be estimated. It can be demonstrated that an approximate estimate of the one-step MSE is

$$\boldsymbol{\Sigma}_{\mathbf{x}}(1) = \frac{T + Np + 1}{T}\boldsymbol{\Sigma}(p)$$

where $\mathbf{\Sigma}(p)$ is the residual covariance matrix of a model of order p and $\mathbf{\Sigma}_{\mathbf{x}}(1)$ is the covariance matrix of the forecasting errors. Based on $\mathbf{\Sigma}_{\mathbf{x}}(1)$, Akaike suggested a criterion to estimate the model order.[1] First, we have to replace $\mathbf{\Sigma}(p)$ with its estimate. In the case of a zero-mean process, we can estimate $\mathbf{\Sigma}(p)$ as

$$\hat{\mathbf{\Sigma}}(p) = \frac{1}{T}\mathbf{X}(\mathbf{I}_T - \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\mathbf{W})\mathbf{X}'$$

The quantity

$$\text{FPE}(p) = \left[\frac{T + Np + 1}{T - Np + 1}\right]^N \det(\hat{\mathbf{\Sigma}}(p))$$

is called the *final prediction error* (FPE). In 1969, Akaike proposed to determine the model order by minimizing the FPE.[2] Four years later, he proposed a different criterion based on information theoretic considerations. The latter criterion, commonly called the *Akaike information criterion* (AIC), proposes to determine the model order by minimizing the following expression:

$$\text{AIC}(p) = \log\left|\hat{\mathbf{\Sigma}}(p)\right| + \frac{2pN^2}{T}$$

Neither the FPE nor the AIC estimators are consistent estimators in the sense that they determine the correct model order in the limit of an infinite sample. Different but consistent criteria have been proposed. Among them, the *Bayesian information criterion* (BIC) is quite popular. Proposed by Schwartz, the BIC chooses the model that minimizes the following expression:[3]

$$\text{BIC}(p) = \log\left|\hat{\mathbf{\Sigma}}(p)\right| + \frac{\log T}{T}2pN^2$$

[1] Hirotugu Akaike, "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics* 21 (1969), pp. 243–247.
[2] Hirotugu Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory* (Budapest: Akademiaio Kiado, 1973).
[3] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics* 6 (1978), pp. 461–464.

There is a vast literature on model selection criteria. The justification of each criterion impinges on rather complex considerations of information theory, statistics, and learning theory.[4]

AUTOCORRELATION AND DISTRIBUTIONAL PROPERTIES OF RESIDUALS

The validity of the LS method does not depend on the distribution of innovations provided that their covariance matrix exists. However, the LS method might not be optimal if innovations are not normally distributed. The ML method, in contrast, critically depends on the distributional properties of innovations. Nevertheless, both methods are sensitive to the autocorrelation of innovation terms. Distributional properties are critical in applications such as asset allocation, portfolio management, and risk management. Therefore, once the model order and parameters are estimated, it is important to check the absence of autocorrelation in the residuals and to ascertain deviations from normal distributions.

There is a range of tests for the autocorrelation and normality of residuals. In particular, the autocorrelation of residuals can be tested with the multivariate Ljung-Box test. The multivariate Ljung-Box test (or Q-test) is a generalization of the Ljung-Box test described in Chapter 7. The null hypothesis of the Ljung-Box test is that all noise terms at different lags up to lag s are uncorrelated. Given a n-dimensional VAR(p) model, the Q-test statistics, in the form introduced by Hosking,[5] is the following:

$$LB(s) = T(T+2)\sum_{j=1}^{s}\frac{1}{T-j}tr[\mathbf{C}_{0j}\mathbf{C}_{00}^{-1}\mathbf{C}_{0j}'\mathbf{C}_{00}^{-1}]$$

where

$$\mathbf{C}_{0j} = T^{-1}\sum_{t=j+1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t-j}'$$

[4] See, for example, D. P. Foster and R. A. Stine, "An Information Theoretic Comparison of Model Selection Criteria," Working Paper 1180, 1997, Northwestern University, Center for Mathematical Studies in Economics and Management Science.

[5] J.R.M. Hosking, "The Multivariate Portmanteau Statistic," *Journal of American Statistical Association* 75 (1980), pp. 602–608.

and T is the sample size. Under the null hypothesis, if $s > p$, the distribution of this statistic is approximately a Chi-square distribution with $n^2(s - p)$.

In the case of stationary models, the normality of distributions can be tested with one of the many tests for normality discussed in Chapter 7. These tests are available on most statistical computer packages.

VAR ILLUSTRATION

Let's now illustrate step by step the process of estimating a VAR model. As stated in Chapter 9, VAR models have been proposed to model asset returns. If asset returns, especially indexes, can be represented as VAR models, then future returns can be predicted from past returns. The objective of this exercise is not to investigate the econometric validity of this assumption but to show how to estimate VAR models and perform diagnostic checks. Specifically, we will show how to:

- Select the number of lags.
- Assess the significance of VAR regression equations (in particular the significance of individual coefficients).
- Assess the causal relationships among the variables.

This information is important both as a model diagnostic and as a tool to help the economic interpretation of the model.

We fit a VAR model to the monthly returns of three stock market indexes: Wilshire capitalization weighted (y_1), Wilshire equal weighted (y_2,), and S&P 500 (y_3,). The time period covered is from October 1989 to January 2003; the data set includes 160 monthly returns. We first model the three indexes as an unrestricted VAR process. Later, we explore the existence of cointegrating relationships. The three series of returns are shown in Exhibit 10.2.

Given that the return series are not integrated, the first task is to determine the number of lags of the VAR model. In practice one compares different models estimated with a different number of lags. To see how this is done, we compare models with one and two lags. We use the following notation:

Wilshire Capitalization Weighted (y_1):	WCW
Wilshire Equal Weighted (y_2):	WEW
S&P 500 (y_3):	S&P

A hyphen following the index abbreviation identifies the number of lags.

EXHIBIT 10.2 Monthly Returns for the Wilshire Capitalization Weighted, Wilshire Equal Weighted, and S&P 500: October 1989–January 2003

Month/ Year	Wilshire Cap Weighted (y_1)	Wilshire Equal Weighted (y_2)	S&P 500 (y_3)	Month/ Year	Wilshire Cap Weighted (y_1)	Wilshire Equal Weighted (y_2)	S&P 500 (y_3)
Oct-89	–2.92	–5.19	–2.33	Oct-92	1.21	2.34	0.36
Nov-89	1.77	–0.53	2.08	Nov-92	4.15	7.93	3.37
Dec-89	1.82	–1.18	2.36	Dec-92	1.78	4.55	1.31
Jan-90	–7.34	–5.10	–6.71	Jan-93	1.23	6.72	0.73
Feb-90	1.59	2.45	1.29	Feb-93	0.41	–1.12	1.35
Mar-90	2.50	2.74	2.63	Mar-93	2.57	3.32	2.15
Apr-90	–2.88	–2.70	–2.47	Apr-93	–2.76	–1.67	–2.45
May-90	9.13	5.21	9.75	May-93	3.13	4.32	2.70
Jun-90	–0.48	0.94	–0.70	Jun-93	0.46	1.36	0.33
Jul-90	–0.97	–2.78	–0.32	Jul-93	–0.01	1.34	–0.47
Aug-90	–9.41	–11.86	–9.03	Aug-93	3.86	4.01	3.81
Sep-90	–5.49	–7.87	–4.92	Sep-93	0.20	3.14	–0.74
Oct-90	–1.34	–6.53	–0.37	Oct-93	1.67	4.38	2.03
Nov-90	6.82	4.79	6.44	Nov-93	–1.62	–2.35	–0.94
Dec-90	3.17	0.58	2.74	Dec-93	1.80	1.40	1.23
Jan-91	4.86	8.63	4.42	Jan-94	3.15	5.01	3.35
Feb-91	7.78	15.58	7.16	Feb-94	–2.24	–0.43	–2.70
Mar-91	3.05	9.41	2.38	Mar-94	–4.53	–4.28	–4.35
Apr-91	0.32	3.83	0.28	Apr-94	0.96	–0.93	1.30
May-91	4.01	4.13	4.28	May-94	0.98	–0.18	1.63
Jun-91	–4.47	–3.03	–4.57	Jun-94	–2.67	–2.60	–2.47
Jul-91	4.70	3.71	4.68	Jul-94	2.97	1.37	3.31
Aug-91	2.76	3.58	2.35	Aug-94	4.42	4.08	4.07
Sep-91	–1.15	1.18	–1.64	Sep-94	–1.94	0.76	–2.41
Oct-91	1.84	3.40	1.34	Oct-94	1.63	0.13	2.29
Nov-91	–3.82	–2.17	–4.04	Nov-94	–3.66	–4.23	–3.67
Dec-91	10.98	4.80	11.43	Dec-94	1.35	–0.75	1.46
Jan-92	–0.20	17.04	–1.86	Jan-95	2.16	2.00	2.60
Feb-92	1.38	6.37	1.28	Feb-95	3.98	3.34	3.88
Mar-92	–2.48	–1.43	–1.96	Mar-95	2.64	1.92	2.96
Apr-92	1.34	3.75	2.91	Apr-95	2.48	2.52	2.91
May-92	0.61	0.94	0.54	May-95	3.39	1.99	3.95
Jun-92	–2.04	–2.80	–1.45	Jun-95	3.19	5.47	2.35
Jul-92	4.05	3.03	4.03	Jul-95	4.11	5.96	3.33
Aug-92	–2.11	–2.49	–2.02	Aug-95	0.97	3.41	0.27
Sep-92	1.19	1.53	1.15	Sep-95	3.81	2.98	4.19

EXHIBIT 10.2 (Continued)

Month/ Year	Wilshire Cap Weighted (y_1)	Wilshire Equal Weighted (y_2)	S&P 500 (y_3)	Month/ Year	Wilshire Cap Weighted (y_1)	Wilshire Equal Weighted (y_2)	S&P 500 (y_3)
Oct-95	–1.00	–4.51	–0.35	Dec-98	6.40	2.08	5.76
Nov-95	4.24	1.80	4.40	Jan-99	3.68	8.50	4.18
Dec-95	1.63	1.04	1.85	Feb-99	–3.62	–4.40	–3.11
Jan-96	2.68	3.02	3.44	Mar-99	3.86	–0.41	4.00
Feb-96	1.75	4.05	0.96	Apr-99	4.79	8.40	3.87
Mar-96	1.09	2.78	0.96	May-99	–2.19	4.30	–2.36
Apr-96	2.47	6.43	1.47	Jun-99	5.18	4.41	5.55
May-96	2.73	8.05	2.58	Jul-99	–3.21	0.82	–3.12
Jun-96	–0.82	–3.19	0.41	Aug-99	–0.93	–3.51	–0.50
Jul-96	–5.40	–8.41	–4.45	Sep-99	–2.61	–1.64	–2.74
Aug-96	3.20	4.55	2.12	Oct-99	6.36	–0.39	6.33
Sep-96	5.32	3.44	5.62	Nov-99	3.35	9.64	2.03
Oct-96	1.40	–2.23	2.74	Dec-99	7.59	8.62	5.89
Nov-96	6.63	2.03	7.59	Jan-00	–4.15	10.35	–5.02
Dec-96	–1.13	0.48	–1.96	Feb-00	2.24	15.95	–1.89
Jan-97	5.35	6.66	6.21	Mar-00	5.94	0.78	9.78
Feb-97	–0.05	–1.41	0.81	Apr-00	–5.21	–10.14	–3.01
Mar-97	–4.42	–4.86	–4.16	May-00	–3.49	–6.91	–2.05
Apr-97	4.36	–2.66	5.97	Jun-00	4.41	8.65	2.47
May-97	7.09	9.51	6.14	Jul-00	–2.04	–1.96	–1.56
Jun-97	4.59	4.89	4.46	Aug-00	7.26	5.64	6.21
Jul-97	7.69	5.25	7.94	Sep-00	–4.67	–3.91	–5.28
Aug-97	–3.76	3.54	–5.56	Oct-00	–2.12	–7.10	–0.42
Sep-97	5.90	8.97	5.48	Nov-00	–9.95	–13.18	–7.88
Oct-97	–3.33	–2.11	–3.34	Dec-00	1.78	–1.67	0.49
Nov-97	3.27	–2.02	4.63	Jan-01	3.83	26.88	3.55
Dec-97	1.85	–2.15	1.72	Feb-01	–9.48	–7.75	–9.12
Jan-98	0.54	1.75	1.11	Mar-01	–6.73	–7.28	–6.33
Feb-98	7.28	6.86	7.21	Apr-01	8.23	7.45	7.77
Mar-98	5.00	5.66	5.12	May-01	1.00	7.86	0.67
Apr-98	1.19	3.01	1.01	Jun-01	–1.68	1.28	–2.43
May-98	–2.66	–4.09	–1.72	Jul-01	–1.65	–2.96	–0.98
Jun-98	3.51	–2.69	4.06	Aug-01	–6.05	–4.00	–6.26
Jul-98	–2.19	–5.17	–1.06	Sep-01	–8.98	–13.02	–8.08
Aug-98	–15.57	–19.79	–14.46	Oct-01	2.54	8.39	1.91
Sep-98	6.53	3.80	6.41	Nov-01	7.65	7.61	7.67
Oct-98	7.44	3.62	8.13	Dec-01	1.80	6.19	0.88
Nov-98	6.30	8.68	6.06	Jan-02	–1.24	2.69	–1.46

EXHIBIT 10.2 (Continued)

Month/ Year	Wilshire Cap Weighted (y_1)	Wilshire Equal Weighted (y_2)	S&P 500 (y_3)
Feb-02	–2.06	–3.72	–1.93
Mar-02	4.38	8.97	3.76
Apr-02	–4.88	–0.22	–6.06
May-02	–1.18	–2.05	–0.74
Jun-02	–7.03	–6.79	–7.12
Jul-02	–8.07	–11.37	–7.80
Aug-02	0.59	0.74	0.66
Sep-02	–10.03	–8.46	–10.87
Oct-02	7.65	4.29	8.80
Nov-02	6.03	13.28	5.89
Dec-02	–5.54	–4.48	–5.88
Jan-03	–2.52	1.26	–2.62

We will first estimate and discuss each model separately and then compare the two models. Let's start with the one lag case, that is, represent our three index time series as the following VAR model:

$$\text{WCW-1} \quad y_1(t) = c_1 + a_{11,1}y_1(t-1) + a_{12,1}y_2(t-1) + a_{13,1}y_3(t-1) + \varepsilon_1(t)$$

$$\text{WEW-1} \quad y_2(t) = c_2 + a_{21,1}y_1(t-1) + a_{22,1}y_2(t-1) + a_{23,1}y_3(t-1) + \varepsilon_2(t)$$

$$\text{S\&P-1} \quad y_3(t) = c_3 + a_{31,1}y_1(t-1) + a_{32,1}y_2(t-1) + a_{33,1}y_3(t-1) + \varepsilon_3(t)$$

We treat each equation as a regression equation and compute the *t*-statistics for each coefficient to assess its significance. We also compute the *Q*-statistics to assess if residuals are autocorrelated and the Granger causality probability to assess causal links between variables.

Note that for a VAR(p) model and sample size T we have at most $T - p$ observations for estimation. Thus, with $p = 1$ and $T = 160$, we can use 159 observations.

For each equation, there are three independent regressors plus a constant so that $k = 4$ coefficients need to be estimated. Least-squares estimation results are reported for each equation in the three panels in Exhibit 10.3. The Granger causality probabilities are summarized in Exhibit 10.4. The three panels in Exhibit 10.5 show the predicted versus the actual values of the three variables.

Let us look at the estimated coefficients for equation WCW-1. Again, keep in mind that this is only an exercise on how to apply modeling tools, we do not claim any general validity for the results.

EXHIBIT 10.3 Estimates of the Three Equations of a VAR Model with One Lag

Panel A. Equation Estimated: WCW-1

Variable	Coefficient	*t*-statistic	*t*-probability
y_1 lag1	$a_{11,1} = 1.679106$	2.825664	0.005340
y_2 lag1	$a_{12,1} = -0.233356$	–2.148203	0.033251
y_3 lag1	$a_{13,1} = -1.489729$	–2.802706	0.005715
Constant	$c_1 = 1.040901$	2.881954	0.004513

$R^2 = 0.0497$
Adjusted $R^2 = 0.0313$
$\hat{\sigma}_1^2 = \hat{\mathbf{\varepsilon}}'\hat{\mathbf{\varepsilon}}/(n-k) = \hat{\mathbf{\varepsilon}}'\hat{\mathbf{\varepsilon}}/155 = 19.0604$
Q-statistic = 0.0323

Panel B. Equation Estimated: WEW-1

Variable	Coefficient	*t*-statistic	*t*-probability
y_1 lag1	$a_{21,1} = 3.382911$	4.445398	0.000017
y_2 lag1	$a_{22,1} = -0.312517$	–2.246509	0.026084
y_3 lag1	$a_{23,1} = -2.817606$	–4.139312	0.000057
Constant	$c_2 = 1.424466$	3.079692	0.002452

$R^2 = 0.1578$
Adjusted $R^2 = 0.1415$
$\hat{\sigma}_1^2 = \hat{\mathbf{\varepsilon}}'\hat{\mathbf{\varepsilon}}/(n-k) = \hat{\mathbf{\varepsilon}}'\hat{\mathbf{\varepsilon}}/155 = 31.2591$
Q-statistic = 0.0076

Panel C. Equation Estimated: S&P-1

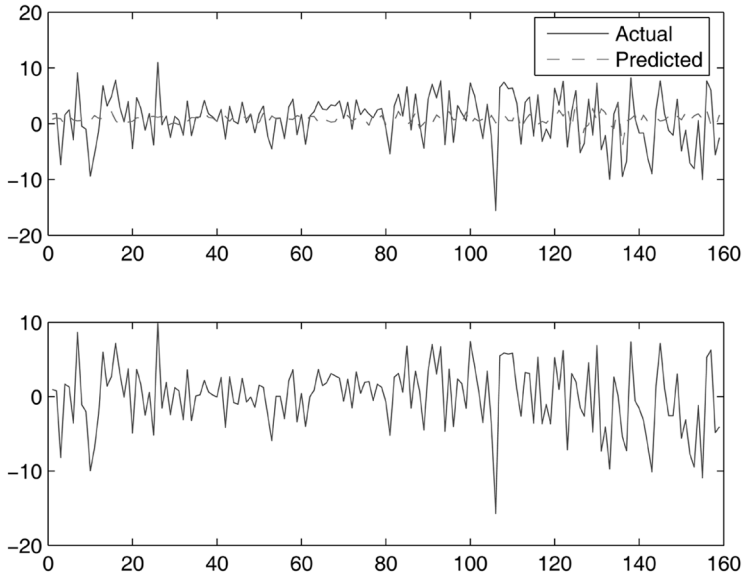
Variable	Coefficient	*t*-statistic	*t*-probability
y_1 lag1	$a_{31,1} = 1.667135$	2.847066	0.005011
y_2 lag1	$a_{32,1} = -0.237547$	–2.219168	0.027928
y_3 lag1	$a_{33,1} = -1.504514$	–2.872438	0.004644
Constant	$c_3 = 1.099824$	3.090190	0.002372

$R^2 = 0.0529$
Adjusted $R^2 = 0.0345$
$\hat{\sigma}_1^2 = \hat{\mathbf{\varepsilon}}'\hat{\mathbf{\varepsilon}}/(n-k) = \hat{\mathbf{\varepsilon}}'\hat{\mathbf{\varepsilon}}/155 = 18.5082$
Q-statistic = 0.0187

EXHIBIT 10.4 Granger Causality Probabilities

Variable	y_1	y_2	y_3
y_1	0.01	0.03	0.01
y_2	0.00	0.03	0.00
y_3	0.01	0.03	0.00

EXHIBIT 10.5 The Predicted versus the Actual Values of the Three Variables
Panel A: Actual versus Predicted Equation WEW–1. The second graph represents residuals.



Panel B: Actual versus Predicted Equation WMW–1. The second graph represents residuals.

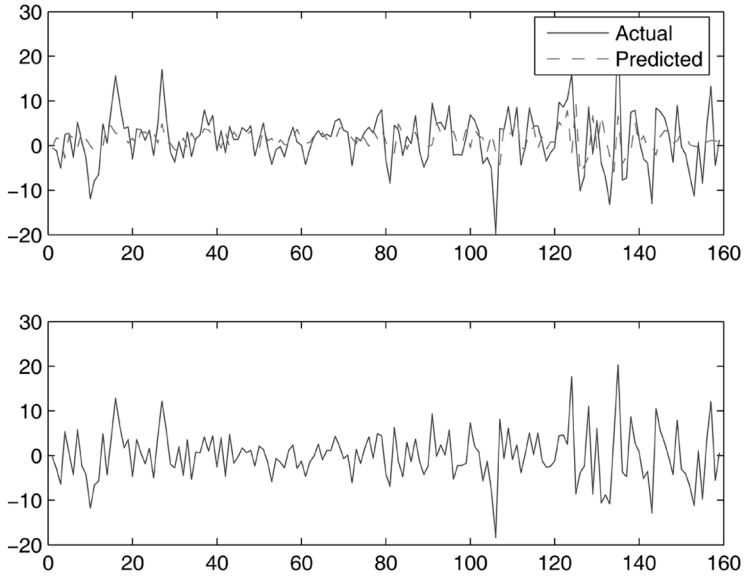
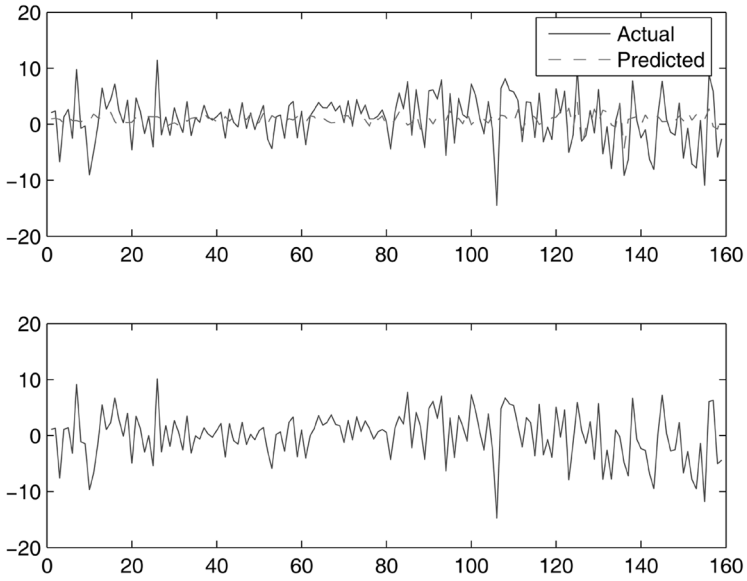


EXHIBIT 10.5 (Continued)
Panel C: Actual versus Predicted Equation S&P–1. The second graph represents residuals.



Recall that the *t*-statistics associated with each coefficient are obtained dividing each coefficient estimate by its respective estimated standard deviation (see Chapter 2) under the assumption that that coefficient is zero. As explained in Chapter 2, the *t*-statistic of an estimated coefficent represents how many standard deviations that coefficient is far from zero. For example, in equation WCW-1, the coefficient of is 1.679106 and its *t*-statistic is 2.825664. This means that this estimated coefficient is 2.825664 standard deviations from zero.

The *t*-probability[6] relative to a coefficient estimates the probability of the null hypothesis that that coefficient is zero; that is, it tests the significance of that coefficient. Small *t*-values correspond to statistically significant coefficients. For this exercise, we assume that coefficients are significant at 99% (i.e., *t*-values less than 0.01).The *t*-probability is the *p*-value of the *t*-statistics, that is, the probability of the tail beyond the observed value of the *t*-statistics of the Student-*t* distribution with $T - p$

[6] Recall from Chapter 2 that the *t*-probability is the probability that the *t*-statistic exceeds a given threshold. The *p*-probability is the probability of the null that all the coefficients be zero.

= 155 degrees of freedom. Exhibit 9.2 shows that only the coefficient of at lag 1 has a nonnegligible probability of being irrelevant. In fact, that coefficient is –2.148203 standard deviations from zero, which corresponds to a tail probability in excess of 3%.[7]

The overall usefulness of the WCW-1 equation can be assessed by the R^2 and the adjusted R^2. Though the null hypothesis that the coefficients of the equation WCW-1 are zero can be rejected, the R^2 shows that only 5% of the variance of variable is explained. The Q-statistic confirms that the residuals have a weak autocorrelation.

We can now repeat the same reasoning for equation WEW-1. Exhibit 9.2 shows that only the coefficient at lag 1 has a nonnegligible probability of being irrelevant. Note that the *t*-probabilities of coefficients of equation WEW-1 are lower than those of equation WCW-1.

The R^2 and the adjusted R^2 reveal that equation WEW-1 has more explanatory power than equation WCW-1. In fact, nearly 16% of the variance is explained. In addition, the Q-statistic shows that the autocorrelation of the residuals of equation WEW-1 is negligible. As explained in Chapter 7, the Q-statistic tests the autocorrelation coefficient at every lag by forming the sum of the squared autocorrelation coefficients.

The results for equation S&P-1 are very similar to those of equation WCW-1.

Exhibit 10.4 shows the Granger causality probabilities. All the probabilities are small, which implies that there are no clear causal links between the three indexes.

Finally, Exhibit 10.5 shows the predicted versus the actual values of the three variables.

Let's now discuss the VAR model with two lags. That is, we fit the following VAR model to the three index series:

$$\begin{aligned}\text{WEW-2}\quad y_1(t) = {} & c_1 + a_{11,1}y_1(t-1) + a_{12,1}y_2(t-1) + a_{13,1}y_3(t-1) \\ & + a_{11,2}y_1(t-2) + a_{12,2}y_2(t-2) + a_{13,2}y_3(t-2) + \varepsilon_1(t)\end{aligned}$$

$$\begin{aligned}\text{WMW-2}\quad y_2(t) = {} & c_2 + a_{21,1}y_1(t-1) + a_{22,1}y_2(t-1) + a_{23,1}y_3(t-1) \\ & + a_{21,2}y_1(t-2) + a_{22,2}y_2(t-2) + a_{23,2}y_3(t-2) + \varepsilon_2(t)\end{aligned}$$

$$\begin{aligned}\text{S\&P-2}\quad y_3(t) = {} & c_3 + a_{31,1}y_1(t-1) + a_{32,1}y_2(t-1) + a_{33,1}y_3(t-1) \\ & + a_{31,2}y_1(t-2) + a_{32,2}y_2(t-2) + a_{33,2}y_3(t-2) + \varepsilon_3(t)\end{aligned}$$

We can now perform the same exercise as with the VAR(1) model with two lags. The sample now includes 158 observations. One might object

[7] The threshold for statistical significance is subjective. In addition, some statisticians question the validity of tail probabilities to gauge significance. See Chapter 2.

that we compare models estimated on different data sets (158 versus159 data points). Though this is true, in practice we have effectively more data to estimate a VAR(1) model with respect to a VAR(2) model. Each regression equation now has six regressors, three variables for each lag.

The least squares estimates are reported for each equation in the three panels in Exhibit 10.6. The Granger causality probabilities are summarized in Exhibit 10.7 while the three panels in Exhibit 10.8 show the predicted versus the actual values of the three variables. Again, keep in mind that this is only an exercise on how to apply modeling tools.

In equation WCW-2, the *t*-statistics show that at a 99% confidence level we cannot reject the null hypothesis of zero coefficient for variable y_1 at lag 2, for variable y_2 at both lags, and for variable y_3 at lag 2. Equation WCW-2 has little explanatory power, with less than 7% of the variance of y_1 explained by the regression equation. This result is in agreement with the previous case of one lag.

For equation WEW-2, we find that for the coefficients $a_{21,2}$, $a_{22,2}$, $a_{23,2}$ the null hypothesis of zero cannot be rejected (i.e., the null hypothesis that these coefficients are not significant cannot be rejected).

As in the case with one lag (WEW-1), equation WEW-2 has a much higher explanatory power with an R^2 about 16%, but the residuals seem to be slightly autocorrelated.

In equation S&P-2 all first coefficients are significant. The overall significance of equation S&P-2 is similar to that for one lag in equation S&P-1.

EXHIBIT 10.6 **Estimates of the Three Equations of a VAR Model with Two Lags**
Panel A: Equation Estimated: WCW-2

Variable	Coefficient	*t*-statistic	*t*-probability
y_1 lag1	$a_{11,1} = 1.717375$	2.743612	0.006813
y_1 lag2	$a_{11,2} = 0.523684$	0.797806	0.426236
y_2 lag1	$a_{12,1} = -0.242473$	–1.946917	0.053400
y_2 lag2	$a_{12,2} = -0.178097$	–1.591905	0.113497
y_3 lag1	$a_{13,1} = -1.535214$	–2.764817	0.006406
y_3 lag2	$a_{13,2} = -0.343669$	–0.590118	0.555993
constant	$c_1 = 1.155683$	2.991625	0.003242

$R^2 = 0.0675$
Adjusted $R^2 = 0.0305$
$\hat{\sigma}_1^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k) = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/155 = 19.1921$
Q-statistic = 0.0284

EXHIBIT 10.6 (Continued)
Panel B: Equation Estimated: WEW-2

Variable	Coefficient	*t*-statistic	*t*-probability
y_1 lag1	$a_{21,1} = 3.402066$	4.225204	0.000041
y_1 lag2	$a_{21,2} = 0.550829$	0.652368	0.515156
y_2 lag1	$a_{22,1} = -0.309937$	–1.934661	0.054901
y_2 lag2	$a_{22,2} = -0.171607$	–1.192458	0.234952
y_3 lag1	$a_{23,1} = -2.859107$	–4.002905	0.000098
y_3 lag2	$a_{23,2} = -0.420667$	–0.561545	0.575259
constant	$c_2 = 1.581039$	3.181692	0.001778

$R^2 = 0.1659$
Adjusted $R^2 = 0.1328$
$\hat{\sigma}_1^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k) = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/155 = 31.7561$
Q-statistic = 0.0948

Panel C: Equation Estimated: S&P-2

Variable	Coefficient	*t*-statistic	*t*-probability
y_1 lag1	$a_{31,1} = 1.707183$	2.755778	0.006577
y_1 lag2	$a_{31,2} = 0.410345$	0.631660	0.528564
y_2 lag1	$a_{32,1} = -0.247446$	–2.007572	0.046473
y_2 lag2	$a_{32,2} = -0.130363$	–1.177396	0.240890
y_3 lag1	$a_{33,1} = -1.547094$	–2.815274	0.005524
y_3 lag2	$a_{33,2} = -0.275951$	–0.478782	0.632786
constant	$c_3 = 1.182225$	3.092252	0.002367

$R^2 = 0.0624$
*R*bar-squared = 0.0252
$\hat{\sigma}_1^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k) = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/155 = 18.7979$
Q-statistic = 0.0775

EXHIBIT 10.7 Granger Causality Probabilities

Variable	y_1	y_2	y_3
y_1	0.03	0.06	0.02
y_2	0.00	0.10	0.00
y_3	0.02	0.09	0.02

Exhibit 10.7 shows the Granger causality probabilities. With two lags, the Granger-causality probability exhibit a weak structure, which might indicate some causal links.

Exhibit 10.8 illustrates graphically the predicted versus the actual values of the three variables.

Let's now compare the two models. The variance of the residuals is slightly smaller in the case of two lags. To see if we should prefer the model with two lags to the model with one lag, let's use the AIC criterion. This criterion requires computing the following expression:

$$n \log \hat{\sigma}_{\varepsilon}^2 + 2k$$

where $\hat{\sigma}_{\varepsilon}^2$ is the variance of residuals, n is the number of data points and k the number of parameters. The model with the smallest AIC value has to be preferred. If we consider, for example, equations WCW-1 and WCW-2, in the case of one lag, there are 159 data points and four parameters to estimate, while in the case of two lags there are 158 data points and seven parameters to estimate. Computing the AIC coefficient yields:

EXHIBIT 10.8 Predicted Versus Actual Values of the Three Variables
Panel A: Actual versus Predicted Equation WEW-2. The second graph represents residuals.

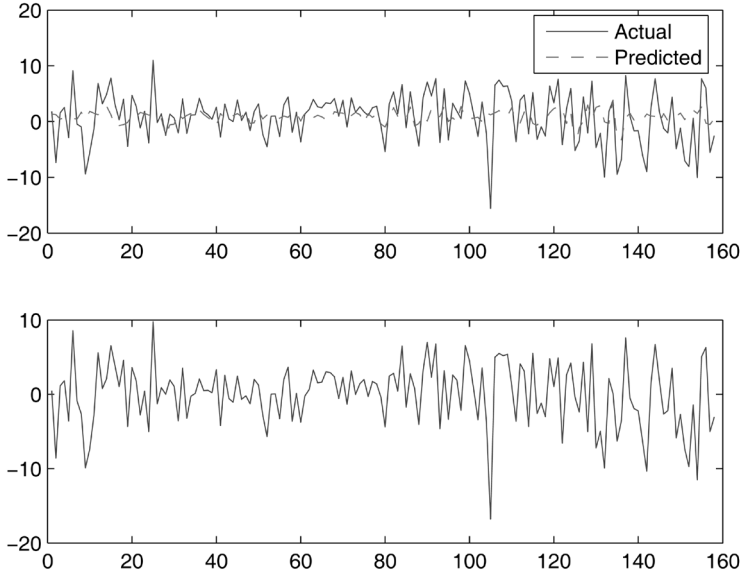
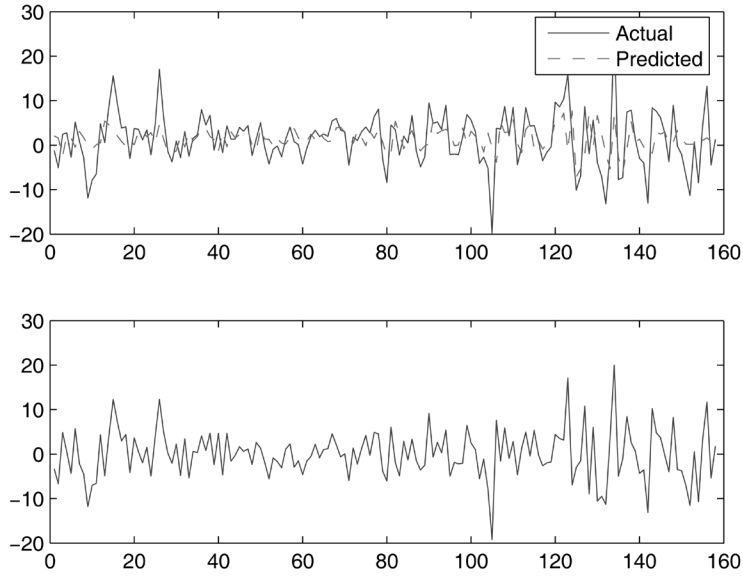
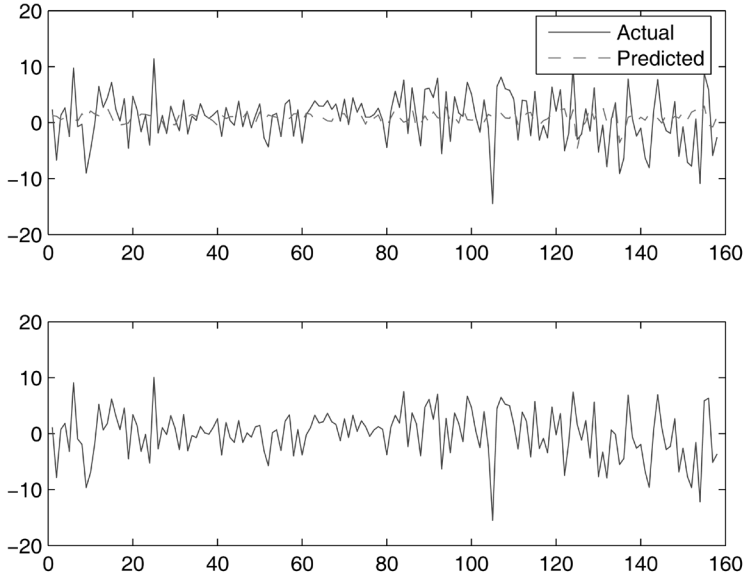


EXHIBIT 10.8 (Continued)

Panel B: Actual versus Predicted Equation WMW-2. The second graph represents residuals.



Panel C: Actual versus Predicted Equation S&P-2. The second graph represents residuals.



For one lag: $159 \times \log(19.0604) + 2 \times 4 = 476.6704$

For two lags: $158 \times \log(19.1921) + 2 \times 7 = 480.8108$

For the equations WEW-1 and WEW-2, we obtain respectively 555.3274 and 563.8355 and for the equations S&P-1 and S&P-2 471.9960 and 477.5317 respectively. Therefore there should be a slight preference for a model with only one lag.

CONCEPTS EXPLAINED IN THIS CHAPTER (IN ORDER OF PRESENTATION)

Restricted and unrestricted VAR models
Companion matrix
Multivariate least squares estimation
Vec operator
Kronecker product
Normal equations
LS estimators
Asymptotic distributions of LS estimators
Demeaned processes
ML estimators
Likelihood of the VAR model
Mean square error (MSE)
Final prediction error (FPE)
Akaike information criterion (AIC)
Bayesian information criterion (BIC)
Ljung-Box test
Q-test

CHAPTER 11

Cointegration and State Space Models

In this chapter, we introduce the concepts of cointegrated processes and state space models, as well as the relative estimation methods. State space models were introduced in the engineering literature in the 1960s especially through the work of Rudolf E. Kalman. Cointegration analysis is a more recent econometric tool. The first articles to introduce cointegrated models were penned by Engle and Granger in the second half of the 1980s.

Though vector autoregressive (VAR) processes and state space models are equivalent representations of the same processes, deeper insight into the relationship between state space models and cointegration was gained more recently when it was understood that cointegration implies a reduced number of common stochastic trends. The idea behind cointegration that there are feedback mechanisms that force processes to stay close together is therefore intimately related to the idea that the behavior of large sets of processes is driven by the dynamics of a smaller number of variables.

COINTEGRATION

Cointegration is one of the key concepts of modern econometrics. Let's start by giving an intuitive explanation of cointegration and its properties. Two or more processes are said to be *cointegrated* if they stay close to each other even if they "drift about" as individual processes. A colorful illustration is that of the drunken man and his dog: Both stumble about aimlessly but never drift too far apart. Cointegration is an impor-

tant concept both for economics and financial modeling. It implements the notion that there are feedbacks that keep variables mutually aligned. To introduce the notion of cointegration, recall the concepts of stationary processes and integrated processes.

Key Features of Cointegration

Let's first give an intuitive characterization to the concept of cointegration in the case of two stochastic processes. Cointegration can be understood in terms of its three key features:

- Reduction of order of integration
- Regression
- Common trends

First, consider *reduction of order of integration*. Two or more stochastic processes that are integrated of order one or higher are said to be cointegrated if there are linear combinations of the processes with a *lower* order of integration. In financial econometrics, cointegration is usually a property of processes integrated of order one that admit linear combinations integrated of order zero (stationary). As we will see, it is also possible to define fractional cointegration between fractionally integrated processes.

Second, the concept of cointegration can be also stated in terms of *linear regression*. Two or more processes integrated of order one are said to be cointegrated if it is possible to make a meaningful linear regression of one process on the other(s). In general, it is not possible to make a meaningful linear regression of one integrated process over another. However, regression is possible if the two processes are cointegrated. Cointegration is that property that allows one to meaningfully regress one integrated process on other integrated processes.

Finally, a property of cointegrated processes is the presence of integrated *common trends*. Given n processes with r cointegrating relationships, it is possible to determine n-r common trends. Common trends are integrated processes such that any of the n original processes can be expressed as a linear regression on the common trends. Cointegration entails dimensionality reduction insofar as common trends are the common drivers of a set of processes.

Long-Run Equilibrium

Given n processes integrated of order one, the processes are said to be cointegrated if there is a linear combination of the processes that is stationary. If the processes are stock prices, cointegration means that even

if the stock prices are individually integrated of order one—for example arithmetic random walks—there are portfolios that are stationary. The linear relationships that produce stationary processes are called *cointegrating* relationships.

Cointegrated processes are characterized by a short-term dynamics and a long-run equilibrium. Note that this latter property does not mean that cointegrated processes *tend* to a long-term equilibrium. On the contrary, the relative behavior is stationary. Long-run equilibrium is the static regression function, that is, the relationship between the processes after eliminating the short-term dynamics.

In general, there can be many linearly independent cointegrating relationships. Given n processes integrated of order one, there can be a maximum of $n - 1$ cointegrating relationships. Cointegrating relationships are not uniquely defined: In fact, any linear combination of cointegrating relationships is another cointegrating relationship.

More Rigorous Definition of Cointegration

Let's now define cointegration in more rigorous terms. The concept of cointegration was introduced by Granger in the second half of the 1980s.[1] It can be expressed in the following way. Suppose that n time series $x_{i,t}$, integrated of the same order d are given. If there is a linear combination of the series

$$\delta_t = \sum_{i=1}^{n} \beta_i x_{i,t}$$

that is integrated of order $e < d$, then the series are said to be cointegrated. Any linear combination as the one above is called a cointegrating relationship. The most commonly found concept of cointegration in financial econometrics is between processes integrated of order $d = 1$ that exhibit stationary linear combinations ($e = 0$).

The concept of cointegration can be extended to processes integrated of order d where d is a rational fraction. Such processes are called *fractionally integrated processes*. The reduction of the order of integration can be fractional too. For example, processes with order of integration $d = ½$ are cointegrated if they exhibit linear combinations that are stationary.

Given n time series, there can be from none to at most $n - 1$ cointegrating relationships. The cointegration vectors $[\beta_i]$ are not unique. In fact, given two cointegrating vectors $[\alpha_i]$ and $[\beta_i]$ such that

[1] Clive W.J. Granger, "Some Properties of Time Series Data and Their Use in Econometric Model Specification," *Journal of Econometrics* 16 (1981), pp. 121–130.

$$\sum_{i=1}^{n} \alpha_i X_i, \quad \sum_{i=1}^{n} \beta_i X_i$$

are integrated of order e, any linear combination of the cointegrating vectors is another cointegrating vector as the linear combination

$$A \sum_{i=1}^{n} \alpha_i X_i + B \sum_{i=1}^{n} \beta_i X_i$$

is integrated of order e.

Stochastic and Deterministic Cointegration

An important distinction has to be made between stochastic and deterministic cointegration. Following the definition of cointegration given above, a multivariate integrated process is cointegrated if there are stationary linear combinations of its components. Let us now look at how we define cointegration if the integrated process has a deterministic trend.

Suppose that the multivariate stochastic process $\mathbf{x}_t$ has a deterministic trend. The process $\mathbf{x}_t$ is said to be *stochastically cointegrated* if there are linear combinations of the process components, each including its own deterministic trend, that are trend stationary (i.e., stationary plus a deterministic trend). In other words, stochastic cointegration removes stochastic trends but not necessarily deterministic trends.

The process $\mathbf{x}_t$ is said to be *deterministically cointegrated* if there are linear combinations of the process components, each including its own deterministic trend, that are stationary without any deterministic trend. In other words, deterministic cointegration removes both stochastic trends and deterministic trends.

Common Trends

Suppose there are n time series $x_{i,t}$, $i = 1, \ldots, n$, and $k < n$ cointegrating relationships. It can be demonstrated that there are $n - k$ integrated time series $u_{j,t}$, $j = 1, \ldots, n - k$, called *common trends*, such that every time series can expressed as a linear combination of the common trends plus a stationary disturbance:

$$x_{i,t} = \sum_{j=1}^{n-k} \gamma_j u_{j,t} + \eta_{i,t}$$

In other words, each process can be regressed on the common trends. Common trends are integrated processes; they were first discussed by Stock and Watson.[2]

Let's now analyze how, in a set of cointegrated processes, each process can be expressed in terms of a reduced number of common stochastic trends. The exposition follows the original work of Stock and Watson.[3] Suppose that the n-variate process $\mathbf{x}_t$ has no deterministic trend, is integrated of order 1, and admits $n - k$ linearly independent cointegrating relationships. This means that there are $r = n - k$ vectors of coefficients $\beta_{i,j}$, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, r$ such that the processes

$$\sum_{i=1}^{n} \beta_{i,j} x_{i,t}$$

are stationary. Assuming that the process has no deterministic trend, we do not have to make any distinction between stochastic and deterministic cointegration. If $\mathbf{x}_t$ represent logarithms of stock prices, cointegration means that there are r portfolios that are stationary even if each individual price process is a random walk.

We arrange the cointegrating relationships in an $n \times r$ matrix:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,n-k} \\ \vdots & \ddots & \vdots \\ \beta_{n,1} & \cdots & \beta_{n,n-k} \end{pmatrix}$$

This matrix has rank r given that its columns are linearly independent. Therefore the r-variate process $\boldsymbol{\beta}'\mathbf{x}_t$ is stationary. Recall that the process can be represented as

$$\mathbf{x}_t = \boldsymbol{\Psi} \sum_{i=1}^{t} \boldsymbol{\varepsilon}_i + \left(\sum_{i=0}^{\infty} \boldsymbol{\Psi}_i^* L^i \right) \boldsymbol{\varepsilon}_t + \mathbf{x}_{-1}$$

where x_{-1} represents the constant term. It can be demonstrated that the assumption of r independent cointegrating relationships implies

[2] James H. Stock and Mark W. Watson, "Diffusion Indexes," NBER Working Paper W6702, 1998; James H. Stock and Mark W. Watson, "New Indexes of Coincident and Leading Economic Indications," in O.J. Blanchard and S. Fischer (eds.), *NBER Macroeconomics Annual 1989* (Cambridge, MA: MIT Press, 1989).

[3] James H. Stock and Mark W. Watson, "Testing for Common Trends," *Journal of the American Statistical Association* 83 (December 1988), pp. 1097–1107.