2 Vector Autoregressive Models

Structural VAR analysis is based on the premise that the DGP is well approximated by a reduced-form VAR model. In applied work, it is therefore important to choose a suitable VAR specification, taking account of the properties of the data. This chapter is devoted to the question of how to specify and estimate reduced-form VAR models. In Section 2.1 stochastic and deterministic trends in the data are discussed. Section 2.2 outlines the basic linear VAR model and its properties. Section 2.3 examines the estimation of reduced-form VAR models. Section 2.4 discusses how to generate predictions from VAR models, and Section 2.5 introduces the concept of Granger causality. Lag-order selection and model diagnostics are discussed in Sections 2.6 and 2.7. Section 2.8 briefly reviews three classes of restricted reduced-form VAR models.

Given that the linear VAR model is one of the standard tools for empirical research in macroeconomics and finance, there are many previous good expositions of the topics covered in this chapter. Our discussion draws heavily on the material in Lütkepohl (2005, 2006, 2009, 2013)

2.1 Stationary and Trending Processes

We call a stochastic process covariance stationary or simply stationary if it has time invariant first and second moments. Similarly, an economic variable is referred to as covariance stationary if the underlying DGP is covariance stationary. More formally, the scalar process y_t, $t \in \mathbb{N}$ or $t \in \mathbb{Z}$, is covariance stationary if

$$\mathbb{E}(y_t) = \mu \quad \text{and} \quad \text{Cov}(y_t, y_{t+h}) = \gamma_h, \quad \forall t, h.$$

Note that μ and γ_h are constants that do not depend on t. This property is also known as second-order stationarity. If the joint distribution of $y_t, \ldots, y_{t+h}$ is time invariant, the process y_t is strictly stationary.

In practice, an economic variable being stationary is the exception rather than the rule. For example, often the raw data have to be transformed prior

to the analysis by taking natural logs to stabilize the variance of the variable. In addition, there are many variables that have trends that have to be removed or modeled explicitly to ensure stationarity. A trend in a time series variable is thought of as a systematic upward or downward movement over time. For example, a variable y_t may vary about a linear trend line of the form $y_t = \mu_0 + \mu_1 t + x_t$, where x_t is a zero mean stationary stochastic process. The straight line, $\mu_0 + \mu_1 t$, represents a simple deterministic trend function that captures the systematic upward or downward movement of many economic variables reasonably well.

Alternatively, a variable may be viewed as being driven by a stochastic trend. A simple example of a process with a stochastic trend is the univariate AR(1) process

$$y_t = a y_{t-1} + u_t$$

with coefficient $a = 1$ such that

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

This process is called a random walk. Its AR polynomial has a unit root, i.e.,

$$1 - az = 0 \quad \text{for } z = 1.$$

Its stochastic error u_t (also known as the innovation) is assumed to be a white noise process with mean 0 and variance σ_u^2. In other words, u_t and u_s are uncorrelated for $s \neq t$, $\mathbb{E}(u_t) = 0$, and $\mathbb{E}(u_t^2) = \sigma_u^2$. Given that $y_t - y_{t-1} = u_t$, it is easily seen that the effect of a random change in u_t on future values of y_t is not reversed in expectation. Thus, the effect of u_t on future values of y_t is permanent.

Successive substitution for lagged y_t variables in the defining equation of the random walk, $y_t = y_{t-1} + u_t$, yields

$$y_t = y_0 + \sum_{i=1}^{t} u_i. \tag{2.1.1}$$

Hence, assuming that the process is defined for $t \in \mathbb{N}$, we have

$$\mathbb{E}(y_t) = \mathbb{E}(y_0) \quad \text{and} \quad \text{Var}(y_t) = t\sigma_u^2 + \text{Var}(y_0).$$

In other words, even though $\text{Var}(y_0)$ is finite, the variance of a random walk tends to infinity. Moreover, the correlation

$$\text{Corr}(y_t, y_{t+h}) = \frac{\mathbb{E}\left[\left(\sum_{i=1}^{t} u_i\right)\left(\sum_{i=1}^{t+h} u_i\right)\right]}{[t\sigma_u^2(t+h)\sigma_u^2]^{1/2}} = \frac{t}{(t^2+th)^{1/2}} \underset{t\to\infty}{\longrightarrow} 1 \tag{2.1.2}$$

for any given integer h. Due to this property, even random variables y_t and y_s of the process far apart in time (such that s is much greater than t) are strongly correlated. This property indicates a strong persistence in the time series process. In fact, it turns out that the expected time between two crossings of zero is infinite. Such behaviour is associated with a trend in the data. Clearly, since u_t is stochastic, so is the trend.

A univariate AR(1) process with unit coefficient and a constant term,

$$y_t = \nu + y_{t-1} + u_t,$$

is called a random walk with drift. Successive substitution of lags of y_t shows that in this case

$$y_t = y_0 + t\nu + \sum_{i=1}^{t} u_i$$

and, hence, the process has a linear trend in the mean:

$$\mathbb{E}(y_t) = \mathbb{E}(y_0) + t\nu.$$

Higher-order AR processes such as

$$y_t = \nu + a_1 y_{t-1} + \cdots + a_p y_{t-p} + u_t,$$

where u_t is white noise as before, have stochastic trending properties similar to random walks if the AR polynomial $1 - a_1 z - \cdots - a_p z^p$ has a root for $z = 1$. The AR polynomial can be decomposed as

$$1 - a_1 z - \cdots - a_p z^p = (1 - \lambda_1 z) \times \cdots \times (1 - \lambda_p z), \tag{2.1.3}$$

where $\lambda_1, \ldots, \lambda_p$ are the reciprocals of the roots of the polynomial. If the process has only one unit root or, equivalently, only one of the λ_i roots is 1 and all the others are smaller than 1, the process behaves similarly to a random walk in that it follows a stochastic trend. More precisely, y_t can be decomposed into a random walk and a stationary component such that y_t varies about a stochastic trend generated by its random walk component.

The representation of the AR polynomial shows that the unit root can be removed by taking first differences of the process. Let $\Delta y_t \equiv (1 - L)y_t \equiv y_t - y_{t-1}$, where L is the lag operator such that $Ly_t \equiv y_{t-1}$, and Δ is the difference operator such that $\Delta \equiv 1 - L$ and hence $\Delta y_t = y_t - y_{t-1}$.

An AR(p) process with AR polynomial satisfying the condition

$$1 - a_1 z - \cdots - a_p z^p \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1, \tag{2.1.4}$$

is called stable. Here $|z|$ denotes the modulus of the complex number z. Put differently, $|z|$ is the distance from the origin of the complex plane. If, in addition, the mean of the AR process does not change over time deterministically, as would be the case in the presence of a deterministic time trend, if the error

term u_t has time-invariant variance σ_u^2, and if its first and second moments are bounded, then the AR process is stationary. Sometimes in the literature, condition (2.1.4) is rather imprecisely viewed as a condition ensuring stationarity. Of course, interpreting (2.1.4) as a stationarity condition implicitly assumes that there are no other deviations from stationarity such as a linear deterministic trend in the mean or an innovation variance changing over time.

To ensure finite moments, AR processes with unit roots are assumed to start at some fixed time period, say t_0, if not explicitly stated otherwise. For example, in the foregoing discussion we have assumed that $t_0 = 0$. In contrast, stable AR processes without unit roots are typically assumed to have started in the infinite past to ensure stationarity. Without that assumption they may only be asymptotically stationary in that the moments are not time-invariant, but converge to their limit values only for $t \to \infty$.

If the AR polynomial has $d \in \mathbb{N}$ unit roots and, hence, d of the λ_i roots in (2.1.3) are equal to 1, the process in called integrated of order d ($I(d)$). In that case, the process can be made stable by differencing it d times. For example, if $d = 1$, $\Delta y_t = y_t - y_{t-1}$ is stable. If $d = 2$, $\Delta^2 y_t = (1 - L)^2 y_t = y_t - 2y_{t-1} + y_{t-2}$ is stable, and so forth. If $d = 2$, the original y_t must be differenced twice. For example, if the log price level p_t is $I(2)$, then the inflation rate $\pi_t = \Delta p_t = p_t - p_{t-1}$ is $I(1)$, and the change in the inflation rate $\Delta\pi_t = \pi_t - \pi_{t-1} = p_t - p_{t-1} - (p_{t-1} - p_{t-2}) = p_t - 2p_{t-1} + p_{t-2}$ is $I(0)$. As before, initial values can be chosen such that $\Delta^d y_t = (1 - L)^d y_t$ is stationary, provided the conditions for the mean and for the innovation variance required for stationarity are satisfied.

Stable, stationary processes are referred to as $I(0)$ processes. Generally, for $d \in \mathbb{N}$, a stochastic process y_t is called $I(d)$, if $\Delta^d y_t \equiv z_t$ is a stationary process with infinite-order moving average (MA) representation, $z_t = \sum_{j=0}^{\infty} \theta_j u_{t-j} = \theta(L)u_t$, where the MA coefficients satisfy the condition $\sum_{j=0}^{\infty} j|\theta_j| < \infty$, $\theta(1) = \sum_{j=0}^{\infty} \theta_j \neq 0$, and $u_t \sim (0, \sigma_u^2)$ is white noise. For example, in the case of an $I(1)$ process, this condition implies that $y_t = y_{t-1} + z_t$ has the representation

$$\begin{aligned} y_t &= y_0 + z_1 + \cdots + z_t \\ &= y_0 + \theta(1)(u_1 + \cdots + u_t) + \sum_{j=0}^{\infty} \theta_j^* u_{t-j} - z_0^*, \end{aligned} \tag{2.1.5}$$

where $\theta_j^* = -\sum_{i=j+1}^{\infty} \theta_i$, $j = 0, 1, \ldots$, and $z_0^* = \sum_{j=0}^{\infty} \theta_j^* u_{-j}$ contains initial values. The variable y_t is decomposed into the sum of a random walk, $\theta(1)(u_1 + \cdots + u_t)$, a stationary process, $\sum_{j=0}^{\infty} \theta_j^* u_{t-j}$, and initial values, $y_0 - z_0^*$. The decomposition (2.1.5) is known as the Beveridge-Nelson decomposition (see Beveridge and Nelson 1981).

Of course, our primary interest is in systems of variables. Hence, it is useful to extend the $I(d)$ terminology to that setting as well. Accordingly, we call a vector process $y_t = (y_{1t}, \ldots, y_{Kt})'$ $I(d)$ if stochastic trends can be removed by

differencing y_t d times and if differencing $d-1$ times is not enough for trend removal. It is important to note, however, that in systems of variables even if only one of the variables is $I(d)$ individually, the whole system is viewed as $I(d)$. Moreover, it is possible that a single stochastic trend drives several of the variables jointly. This is the important case of cointegrated variables to be discussed in Chapter 3.

The $I(d)$ terminology has also been extended to non-integer, real numbers d. For general $d \in \mathbb{R}$ the so-called fractional differencing operator Δ^d is defined as a binomial expansion,

$$\begin{aligned}\Delta^d = (1-L)^d &= 1 - dL - \frac{d(1-d)}{2}L^2 - \frac{d(1-d)(2-d)}{6}L^3 - \cdots \\ &= \sum_{i=0}^{\infty}(-1)^i\binom{d}{i}L^i \\ &= \sum_{i=0}^{\infty}(-1)^i\frac{d(d-1)\times\cdots\times(d-i+1)}{1\times 2\times\cdots\times i}L^i.\end{aligned}$$

The infinite sum reduces to a finite sum for $d \in \mathbb{N}$. The process y_t is called fractional or fractionally integrated of order d if $\Delta^d y_t = z_t$ is $I(0)$ with MA representation $z_t = \theta(L)u_t$, $\theta(1) \neq 0$ (see, e.g., Johansen and Nielsen 2012). Such processes were introduced to the time series econometrics literature by Granger and Joyeux (1980) and Hosking (1981b). Fractionally integrated processes are often referred to as long-memory processes because for $d > 0$ they are more persistent and their autocorrelations taper off to zero more slowly than for $I(0)$ processes. Although fractionally integrated processes are not $I(0)$, they may be stationary. Stationarity of a fractionally integrated process requires $|d| < 0.5$.

Integer-valued differences often have a natural interpretation. For example, first differences of the logs of a variable represent growth rates. Such an easy interpretation is lost for fractionally differenced variables. Thus, it is perhaps not surprising that the concept of fractional integration to date has not been used much in structural VAR analysis. More importantly, reliable estimation of fractionally integrated processes requires larger samples than typically available in macroeconomics. Fractional processes therefore do not play an important role in this volume. In the remainder of this book, when we refer to $I(d)$ variables, we always mean non-negative integers d unless explicitly stated otherwise.

2.2 Linear VAR Processes

2.2.1 The Basic Model

Suppose that the relationship between a set of K time series variables, $y_t = (y_{1t}, \ldots, y_{Kt})'$, is of interest and that the DGP can be represented as the sum of

a deterministic part μ_t and a purely stochastic part x_t with mean zero such that

$$y_t = \mu_t + x_t. \tag{2.2.1}$$

In other words, the expected value of y_t is $\mathbb{E}(y_t) = \mu_t$. The deterministic term may contain a constant, polynomial trend terms, deterministic seasonal terms, and other dummy variables. For simplicity, μ_t is usually assumed to contain only a constant such that $\mu_t = \mu_0$. Occasionally a linear trend of the form $\mu_t = \mu_0 + \mu_1 t$ is considered. Generally the additive setup (2.2.1) makes it necessary to think about the deterministic terms at the beginning of the analysis and to allow for the appropriate polynomial order. In some applications trend adjustments are performed prior to a VAR analysis. This approach must be taken, for example, when the detrending procedure cannot be incorporated into the VAR specification. An example is the use of HP-filtered data. Further discussion of these alternative detrending methods can be found in Chapter 19. In that case there may be no deterministic term in the VAR representation in levels, i.e., $\mu_t = 0$ in expression (2.2.1) and $y_t = x_t$.

The purely stochastic part, x_t, of the DGP is assumed to follow a linear VAR process of order p (referred to as a VAR(p) model) of the form

$$x_t = A_1 x_{t-1} + \cdots + A_p x_{t-p} + u_t, \tag{2.2.2}$$

where the A_i, $i = 1, \ldots, p$, are $K \times K$ parameter matrices and the error process $u_t = (u_{1t}, \ldots, u_{Kt})'$ is a K-dimensional zero mean white noise process with covariance matrix $\mathbb{E}(u_t u_t') = \Sigma_u$ such that $u_t \sim (0, \Sigma_u)$. The white noise assumption rules out serial correlation in the errors but allows for conditional variance dynamics such as generalized autoregressive conditionally heteroskedastic (GARCH) errors (see e.g. Chapter 14). Sometimes it is useful to strengthen this assumption, for example, by postulating independent and identically distributed (iid) errors or by postulating that u_t is a martingale difference sequence.[1]

Expression (2.2.2) defines a system of equations. Each model variable in y_t is regressed on its own lags as well as lags of the other model variables up to a lag order p (see Chapter 1). To economize on notation, it is convenient to define the matrix polynomial in the lag operator $A(L) = I_K - A_1 L - \cdots - A_p L^p$ and write the process (2.2.2) as

$$A(L)x_t = u_t. \tag{2.2.3}$$

The observed variables y_t inherit the VAR structure of x_t. This can be seen easily by pre-multiplying (2.2.1) by $A(L)$ and considering $A(L)y_t = A(L)\mu_t +$

[1] The stochastic process v_t is called a martingale sequence if $\mathbb{E}(v_t | v_{t-1}, v_{t-2}, \ldots) = v_{t-1}$ $\forall t$. Then $u_t \equiv \Delta v_t$ is called a martingale difference if it has expectation $\mathbb{E}(u_t | v_{t-1}, v_{t-2}, \ldots) = 0$ $\forall t$. Unlike an iid white noise process, a white noise process that is a martingale difference sequence allows for conditional heteroskedasticity.

u_t. For instance, if the deterministic term is just a constant, i.e., $\mu_t = \mu_0$, then

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{2.2.4}$$

where $\nu = A(L)\mu_0 = A(1)\mu_0 = (I_K - \sum_{j=1}^{p} A_j)\mu_0$. In the terminology of the literature on simultaneous equations models, model (2.2.4) is a reduced form because all right-hand side variables are lagged and hence predetermined.

The VAR process x_t and, hence, y_t is stable if all roots of the determinantal polynomial of the VAR operator are outside the complex unit circle, i.e.,

$$\det(A(z)) = \det(I_K - A_1 z - \cdots - A_p z^p) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1, \tag{2.2.5}$$

where $\mathbb{C}$ denotes the set of complex numbers. Under common assumptions such as a constant mean and white noise innovations with time-invariant covariance matrix, a stable VAR process has time-invariant means, variances, and covariance structure and hence is stationary, as will be seen in the next subsection. Thus, condition (2.2.5) generalizes the stability condition (2.1.4) to the multivariate case.

For later reference we note that the K-dimensional VAR(p) process (2.2.4) can be written as a pK-dimensional VAR(1) process by stacking p consecutive y_t variables in a pK-dimensional vector, $Y_t = (y_t', \ldots, y_{t-p+1}')'$, and noting that

$$Y_t = \boldsymbol{\nu} + \mathbf{A} Y_{t-1} + U_t, \tag{2.2.6}$$

where

$$\boldsymbol{\nu} \equiv \underset{Kp\times 1}{\begin{bmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{bmatrix}}, \quad \mathbf{A} \equiv \underset{Kp\times Kp}{\begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0 & \cdots & 0 & 0 \\ 0 & I_K & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{bmatrix}}, \quad \text{and} \quad U_t \equiv \underset{Kp\times 1}{\begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}}.$$

The matrix $\mathbf{A}$ is referred to as the companion matrix of the VAR(p) process. Using the stability condition (2.2.5), Y_t is stable if

$$\det(I_{Kp} - \mathbf{A}z) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1, \tag{2.2.7}$$

which, of course, is equivalent to condition (2.2.5). It is easy to see that this condition is equivalent to all eigenvalues of $\mathbf{A}$ having modulus less than 1, which provides a convenient tool for assessing the stability of a VAR model and for computing the autoregressive roots. By construction, the eigenvalues of $\mathbf{A}$ are the reciprocals of the roots of the VAR lag polynomial (2.2.5).

2.2.2 The Moving Average Representation

A stable VAR(p) process y_t can be represented as the weighted sum of past and present innovations. This is easily seen for a VAR(1) process,

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

Successive substitution implies

$$y_t = \sum_{i=0}^{\infty} A_1^i \nu + \sum_{i=0}^{\infty} A_1^i u_{t-i} = (I_K - A_1)^{-1}\nu + \sum_{i=0}^{\infty} A_1^i u_{t-i}.$$

The sum on the right-hand side of this infinite-order representation exists if the eigenvalues of A_1 are all less than 1 in modulus. Similarly, a representation in terms of past and present innovations of a VAR(p) model can be obtained via the corresponding VAR(1) representation, resulting in

$$\begin{aligned} y_t &= A(L)^{-1}\nu + A(L)^{-1}u_t \\ &= A(1)^{-1}\nu + \sum_{i=0}^{\infty} J\mathbf{A}^i J' J U_{t-i} \\ &= \mu + \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \end{aligned} \tag{2.2.8}$$

where $J \equiv [I_K, 0_{K\times K(p-1)}]$ is a $K \times Kp$ matrix, $\mu = A(1)^{-1}\nu$ and the $K \times K$ coefficient matrices of the inverse VAR operator $A(L)^{-1} = \sum_{i=0}^{\infty} \Phi_i L^i$ are equal to $\Phi_i = J\mathbf{A}^i J'$, $i = 0, 1, \ldots$. These matrices can also be obtained recursively as

$$\Phi_0 = I_K, \quad \text{and} \quad \Phi_i = \sum_{j=1}^{i} \Phi_{i-j} A_j, \qquad i = 1, 2, \ldots,$$

with $A_j = 0$ for $j > p$ (see Lütkepohl 2005, chapter 2).

The existence of the inverse VAR operator is ensured by the stability of the process. The representation (2.2.8) is known as the moving average (MA) representation or more precisely the Wold MA representation or the prediction error MA representation. This qualifier is important because there are infinitely many MA representations of y_t. In fact, any nonsingular linear transformation of the white noise process u_t, say $v_t = Qu_t$, gives rise to a white noise process and can be used for an MA representation of y_t,

$$y_t = \mu + \sum_{i=0}^{\infty} \Theta_i v_{t-i}, \tag{2.2.9}$$

with $\Theta_i = \Phi_i Q^{-1}$, $i = 0, 1, \ldots$. A distinguishing feature of the Wold MA representation is that the weighting matrix Φ_0 of the unlagged error term

is the identity matrix, while Θ_0 is not an identity matrix for nontrivial transformations.

It follows immediately from the Wold MA representation that

$$\mathbb{E}(y_t) = \mu$$

and that

$$\Gamma_y(h) \equiv \text{Cov}(y_t, y_{t-h}) = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)'] = \sum_{i=0}^{\infty} \Phi_{h+i} \Sigma_u \Phi_i'. \tag{2.2.10}$$

Hence, the first and second moments of this VAR process are time invariant and the process is stationary (see Lütkepohl 2005, chapter 2).

2.2.3 VAR Models as an Approximation to VARMA Processes

An important result in this context is due to Wold (1938) who showed that every K-dimensional nondeterministic zero mean stationary process y_t has an MA representation

$$y_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}, \tag{2.2.11}$$

where $\Phi_0 = I_K$. This result follows from the Wold Decomposition Theorem and motivates the terminology used for the MA representation (2.2.8). This result is important because it illustrates the generality of the VAR model. Suppose the Φ_i are absolutely summable and that there exists an operator $A(L)$ with absolutely summable coefficient matrices satisfying $A(L)\Phi(L) = I_K$. Then $\Phi(L)$ is invertible [$A(L) = \Phi(L)^{-1}$] and y_t has a VAR representation of possibly infinite order that can be approximated arbitrarily well by a finite-order VAR(p) if p is sufficiently large.

In particular, under suitable conditions, a VAR(p) process may be used to approximate time series generated from vector autoregressive moving average (VARMA) models of the form

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_{p_0} y_{t-p_0} + u_t + M_1 u_{t-1} + \cdots + M_{q_0} u_{t-q_0},$$

where p_0 and q_0 denote the true autoregressive and moving average lag orders, provided the VAR lag order p is sufficiently large. If the VAR operator $A(z) = I_K - A_1 z - \cdots - A_{p_0} z^{p_0}$ of the VARMA process satisfies the stability condition (2.2.5) and, thus, the VAR operator has no roots in or on the complex unit circle, the VARMA process has a possibly infinite-order MA representation (2.2.11).

Moreover, if the determinant of the MA operator of the VARMA process has all its roots outside the unit circle, i.e.,

$$\det(M(z)) = \det(I_K + M_1 z + \cdots + M_{q_0} z^{q_0}) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1,$$

the process also has an equivalent pure VAR representation of possibly infinite order.[2] Unlike in the univariate case, the inverse of a finite-order operator may also be a finite-order operator in the multivariate case. In other words, $M(z)^{-1}$ may be a finite order operator if $M(z)$ has finite order. Hence, it is possible in the multivariate case that a finite-order MA process has an equivalent finite-order VAR representation and vice versa.

A detailed introductory exposition of VARMA processes is provided by Lütkepohl (2005), and a more advanced treatment can be found in Hannan and Deistler (1988). Since VARMA processes are much more difficult to deal with in practice, we focus on VAR models in the remainder of this book.

If the VAR process of interest has a unit root and, hence, the stability condition is not satisfied, the infinite-order MA representation (2.2.8) does not exist. However, we can still think of the process as starting from $Y_0 = (y_0', \ldots, y_{-p+1}')'$ and obtain a representation

$$y_t = \mu_t + \sum_{i=0}^{t-1} \Phi_i u_{t-i} + J\mathbf{A}^i Y_0$$

by successive substitution. For some purposes this representation is useful, but not for all. In particular, it obscures the long-run properties of the process. These are more easily understood using the so-called Granger representation discussed in Chapter 3.

2.2.4 Marginal Processes, Measurement Errors, Aggregation, Variable Transformations

The reduced-form MA representation is also a good point of departure for studying the implications of dropping variables from a VAR process. Consider a bivariate stationary process for two variables,

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{bmatrix} \phi_{11}(L) & \phi_{12}(L) \\ \phi_{21}(L) & \phi_{22}(L) \end{bmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \tag{2.2.12}$$

Thus, the first variable has the representation

$$y_{1t} = \mu_1 + \phi_{11}(L)u_{1t} + \phi_{12}(L)u_{2t}$$

[2] An MA representation with MA operator satisfying this invertibility condition is sometimes called a fundamental MA representation. In Chapter 17 we discuss nonfundamental MA representations that have roots inside the complex unit circle and, hence, do not satisfy the invertibility condition for the MA operator.

in terms of both innovation series. According to Wold's decomposition theorem, it also has an MA representation in terms of a scalar white noise process, v_t:

$$y_{1t} = \mu_1 + \sum_{i=0}^{\infty} \psi_i v_{t-i}.$$

This MA represents the marginal process of y_{1t} that is obtained by integrating out the second variable. If $\phi_{12}(L) \neq 0$, then $v_t \neq u_{1t}$ and $\psi_i \neq \phi_{11,i}$ in general. These facts are important to keep in mind for the analysis of impulse responses in Chapter 4. The point to remember is that, in general, dropping some variables from a multivariate time series process results in a lower-dimensional process with possibly quite different MA coefficients than the process for the original set of variables.

More generally, any transformation of the variables implies changes in the MA coefficients. Consider, for example, a nonsingular transformation matrix F and a transformed process

$$z_t = Fy_t = F\mu + \sum_{i=0}^{\infty} F\Phi_i F^{-1} F u_{t-i} = \mu_z + \sum_{i=0}^{\infty} \Psi_i v_{t-i}, \qquad (2.2.13)$$

where $\Psi_i = F\Phi_i F^{-1}$ and $v_t = Fu_t$. Obviously, such transformations change the MA coefficient matrices and white noise error term, and therefore also affect the lag order of the approximating autoregressive process. The same result also holds when F is not a square matrix. Suppose F is an $M \times K$ matrix of rank M. Then one may add $K - M$ rows to the matrix such that it becomes nonsingular, consider the resulting nonsingular transformation, and finally omit the last $K - M$ components of the transformed vector.

In short, linear transformations of a VAR process have MA representations quite different from that of the original process, but both representations are equally valid. For example, a researcher may be working with a VAR for the interest rate, inflation rate, and real GDP growth. These variables could alternatively be represented as autoregressive-moving average (ARMA) processes for each series separately, which in turn can be approximated by finite-order AR models. Linear transformations are also quite common when aggregating data across households and industries to form macroeconomic aggregates. Likewise, problems of temporal aggregation fall within this framework (see Lütkepohl 2005, chapter 11). For example, it is common to aggregate monthly inflation to quarterly inflation data, which involves taking a linear combination of monthly inflation rates.

In this context it is important to stress that different types of variables require different temporal aggregation methods. For flow variables such as GDP or industrial production, temporal aggregation to a lower frequency

involves accumulating the high-frequency observations over time. For example, quarterly industrial production is obtained by summing the monthly industrial production that has taken place within each quarter. In contrast, stock variables such as the number of unemployed workers or the population of a region are aggregated from monthly data to quarterly frequency by using, for example, the last monthly value of each quarter as the quarterly value and dropping the other monthly observations. In other words, temporal aggregation is performed by what is known as skip-sampling or systematic sampling. Alternatively, one may use the average of the monthly values as a quarterly value, depending on the economic context. The important point to note here is that different temporal aggregation schemes imply different changes in the DGP and hence in the MA representation of the variables. There is an extensive literature discussing these issues, in particular in the forecasting context. Early contributions include Tiao (1972), Amemiya and Wu (1972), Brewer (1973), Abraham (1982), and Wei (1981). A more recent systematic account of this literature is provided in Lütkepohl (1987). An alternative approach has been to combine time series observed at different frequencies within the same econometric model (see Foroni, Ghysels, and Marcellino 2013). Related issues in the context of structural modeling are taken up in Chapter 15.

Another example of a linear aggregation problem is additive measurement error in the data. Suppose that the variable of interest, say y_t^*, is measured with error and denote the measurement error by m_t such that the observed process is $y_t = y_t^* + m_t$. In other words, y_t is a linear transformation of the joint process

$$\begin{pmatrix} y_t^* \\ m_t \end{pmatrix}.$$

Then, assuming that the joint process is stationary, the previous discussion implies that the MA representation of y_t differs from that of y_t^*.

These considerations demonstrate that even linear transformations can have a substantial impact on the MA representation of a stationary process. Although these issues do not invalidate the reduced-form representation of VAR models, they may affect the structural interpretation and identification of VAR models, as discussed in later chapters. Our discussion in this section has been based on the MA representation and, hence, applies to stationary processes more generally. We now return to finite-order VAR processes and discuss parameter estimation within that model class.

2.3 Estimation of VAR Models

VAR models can be estimated by standard methods. Unrestricted least-squares (LS), generalized least-squares (GLS), bias-corrected least-squares, and maximum likelihood (ML) methods are discussed in Sections 2.3.1–2.3.4. Our main focus in this chapter is on stationary VAR processes. The properties of LS and

ML methods when there are integrated variables in the VAR model in levels are briefly discussed in Section 2.3.5. A more detailed dicussion of the estimation of integrated and cointegrated VAR processes can be found in Chapter 3. Bayesian estimation methods for VAR models are reviewed in Chapter 5.

2.3.1 Least-Squares Estimation

Consider the VAR(p) model (2.2.4) written in the more compact form

$$y_t = [\nu, A_1, \ldots, A_p]Z_{t-1} + u_t, \tag{2.3.1}$$

where $Z_{t-1} \equiv (1, y'_{t-1}, \ldots, y'_{t-p})'$ and u_t is assumed to be iid white noise with nonsingular covariance matrix, Σ_u, such that $u_t \overset{iid}{\sim} (0, \Sigma_u)$. Further deterministic terms may be handled analogously. Given a sample of size T, $y_1, \ldots, y_T$, and p presample vectors, $y_{-p+1}, \ldots, y_0$, ordinary LS for each equation separately results in efficient estimators. The LS estimator is

$$\hat{A} = [\hat{\nu}, \hat{A}_1, \ldots, \hat{A}_p] = \left(\sum_{t=1}^{T} y_t Z'_{t-1}\right)\left(\sum_{t=1}^{T} Z_{t-1}Z'_{t-1}\right)^{-1} = YZ'(ZZ')^{-1}, \tag{2.3.2}$$

where $Y \equiv [y_1, \ldots, y_T]$ is $K \times T$ and $Z \equiv [Z_0, \ldots, Z_{T-1}]$ is $(Kp+1) \times T$.

More precisely, stacking the columns of $A = [\nu, A_1, \ldots, A_p]$ in the $(pK^2 + K) \times 1$ vector $\boldsymbol{\alpha} = \text{vec}(A)$,

$$\sqrt{T}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\hat{\boldsymbol{\alpha}}}), \tag{2.3.3}$$

where $\Sigma_{\hat{\boldsymbol{\alpha}}} = \text{plim}\left(\frac{1}{T}ZZ'\right)^{-1} \otimes \Sigma_u$, if the process is stable. Under fairly general assumptions, the LS estimator has an asymptotic normal distribution (see Mann and Wald 1943). A sufficient condition for the consistency and asymptotic normality of $\hat{A}$ would be that u_t is a continuous iid random variable with four finite moments (see, e.g., Lütkepohl 2005, chapter 3). These assumptions may be relaxed to allow for conditional or unconditional heteroskedasticity, for example.

A consistent estimator of the innovation covariance matrix Σ_u under the assumption of iid innovations is

$$\widehat{\Sigma}_u = \frac{\widehat{U}\widehat{U}'}{T - Kp - 1}, \tag{2.3.4}$$

where $\widehat{U} = Y - \hat{A}Z$ are the LS residuals.

Thus, in large samples,

$$\text{vec}(\hat{A}) \overset{a}{\sim} \mathcal{N}\left(\text{vec}(A), \left(ZZ'\right)^{-1} \otimes \widehat{\Sigma}_u\right), \tag{2.3.5}$$

where $\overset{a}{\sim}$ denotes the approximate large-sample distribution. In other words, asymptotically the usual t-statistics can be used for testing restrictions on individual coefficients and for setting up confidence intervals.

Moreover, if multiple restrictions are of interest, Wald tests can be used. Suppose, for example, that we want to test the pair of linear hypotheses

$$\mathbb{H}_0 : R\boldsymbol{\alpha} = r \quad \text{versus} \quad \mathbb{H}_1 : R\boldsymbol{\alpha} \neq r,$$

where R is a given $N \times (pK^2 + K)$ matrix of rank N and r is a given N-dimensional vector. Then, if $\mathbb{H}_0$ is true, the result (2.3.3) implies that the statistic

$$W = T(R\hat{\boldsymbol{\alpha}} - r)'(R\widehat{\Sigma}_{\hat{\boldsymbol{\alpha}}}R')^{-1}(R\hat{\boldsymbol{\alpha}} - r),$$

where under our assumptions $\widehat{\Sigma}_{\hat{\boldsymbol{\alpha}}} = \left(\frac{1}{T}ZZ'\right)^{-1} \otimes \widehat{\Sigma}_u$ is a consistent estimator of $\Sigma_{\hat{\boldsymbol{\alpha}}}$, has an asymptotic χ^2 distribution with N degrees of freedom that can be used for testing $\mathbb{H}_0$. This test is known as a Wald test.

In VAR analysis, nonlinear functions of the parameters are often of interest. Examples are the structural impulse responses and forecast error variance decompositions considered in Chapter 4. Suppose that interest focuses on the nonlinear function $\phi : \mathbb{R}^{pK^2+K} \to \mathbb{R}^N$ that maps A on the N-dimensional vector $\phi(A)$. Estimation of ϕ may be based on the LS estimator of A, denoted by $\widehat{\phi} = \phi(\hat{A})$. Then, using result (2.3.3) and the delta method, it follows that

$$\sqrt{T}(\widehat{\phi} - \phi) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial \phi}{\partial \boldsymbol{\alpha}'} \Sigma_{\hat{\boldsymbol{\alpha}}} \frac{\partial \phi'}{\partial \boldsymbol{\alpha}}\right), \tag{2.3.6}$$

where $\partial\phi/\partial\boldsymbol{\alpha}'$ is the $N \times (pK^2 + K)$ matrix of partial derivatives of the nonlinear function ϕ with respect to the elements of $\boldsymbol{\alpha} \equiv \text{vec}(A)$. It is assumed that the matrix of partial derivatives is nonzero, when evaluated at the true parameter vector (see Serfling 1980). If the covariance matrix

$$\Sigma_{\widehat{\phi}} = \frac{\partial \phi}{\partial \boldsymbol{\alpha}'} \Sigma_{\hat{\boldsymbol{\alpha}}} \frac{\partial \phi'}{\partial \boldsymbol{\alpha}}$$

is nonsingular, result (2.3.6) can be used in the usual way for inference regarding ϕ. In other words, t-ratios may be used for significance tests of individual components of ϕ and Wald tests can be constructed for hypotheses related to more than one component of ϕ. For example, the null hypothesis $\mathbb{H}_0 : \phi(A) = \phi_0$ can be tested by using the statistic

$$W = T[\phi(\hat{A}) - \phi_0]'\widehat{\Sigma}_{\widehat{\phi}}^{-1}[\phi(\hat{A}) - \phi_0]$$

that has an asymptotic χ^2 distribution if $\mathbb{H}_0$ is true. If $\Sigma_{\widehat{\phi}}$ is singular, in contrast, Wald tests may not have the usual asymptotic χ^2 distribution anymore, whereas t-tests remain valid asymptotically, as long as all diagonal elements of $\Sigma_{\widehat{\phi}}$ are nonzero. General results for the case of a singular covariance matrix

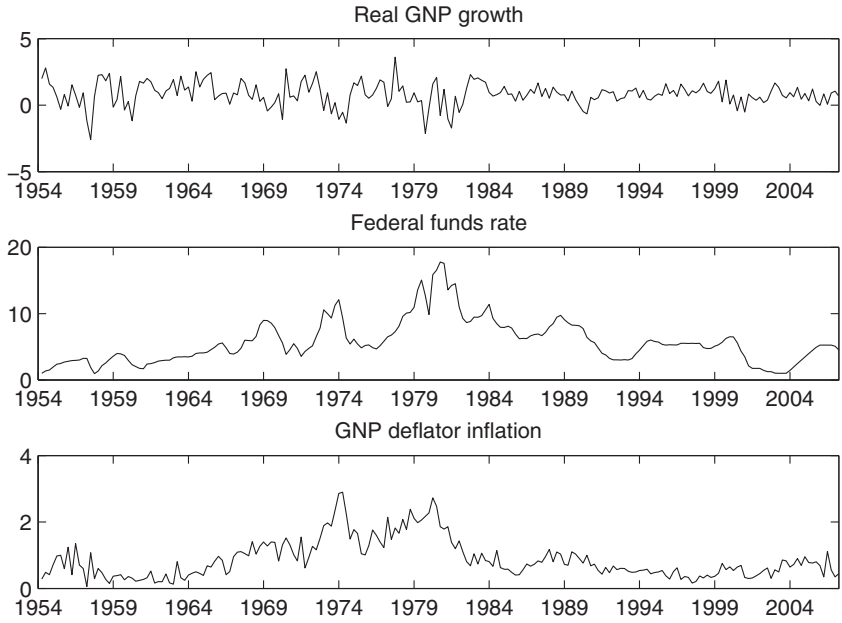

Figure 2.1. Quarterly U.S. data for real GNP growth, the federal funds rate, and the GNP deflator inflation for 1954q4–2007q4.

can be found in Andrews (1987). The issue of singularities is discussed in more detail in Chapter 12 in the context of the example of inference about structural impulse responses.

As an empirical illustration, consider a VAR(4) for $y_t = (\Delta gnp_t, i_t, \Delta p_t)'$, where gnp_t denotes the log of U.S. real GNP, p_t the corresponding GNP deflator in logs, and i_t the federal funds rate, averaged by quarter. The estimation period is restricted to 1954q4–2007q4. The data are shown in Figure 2.1. In this section we treat these data as $I(0)$.

The unrestricted LS estimates of the VAR(4) model with intercept are

$$\widehat{v} = \begin{bmatrix} 0.7240 \\ -0.3925 \\ 0.0067 \end{bmatrix},$$

$$\widehat{A}_1 = \begin{bmatrix} 0.2230 & 0.0097 & 0.3969 \\ 0.3147 & 1.0969 & 0.5979 \\ 0.0012 & 0.0636 & 0.4096 \end{bmatrix},$$

$$\widehat{A}_2 = \begin{bmatrix} 0.2143 & -0.3862 & 0.1360 \\ 0.1867 & -0.4860 & 0.5037 \\ -0.0174 & -0.0510 & 0.2350 \end{bmatrix},$$

$$\widehat{A}_3 = \begin{bmatrix} -0.0053 & 0.3407 & -0.5354 \\ 0.0275 & 0.4832 & -0.3212 \\ 0.0115 & -0.0052 & 0.0815 \end{bmatrix},$$

$$\widehat{A}_4 = \begin{bmatrix} -0.0411 & 0.0013 & -0.0268 \\ -0.0226 & -0.1642 & -0.3320 \\ 0.0667 & -0.0137 & 0.2463 \end{bmatrix},$$

and

$$\widehat{\Sigma}_u = \begin{bmatrix} 0.6031 & 0.0795 & -0.0214 \\ 0.0795 & 0.6565 & 0.0375 \\ -0.0214 & 0.0375 & 0.0684 \end{bmatrix}.$$

2.3.2 Restricted Generalized Least Squares

The LS estimator $\hat{\boldsymbol{\alpha}}$ in expression (2.3.2) is identical to the GLS estimator if no restrictions are imposed on the parameters. If there are parameter restrictions, LS estimation will be asymptotically inefficient and GLS estimation may be preferable. Suppose that there are linear restrictions on the parameters. For example, some of the lagged variables may be excluded from some of the equations. Such restrictions can be expressed by defining a suitable $(K^2p + K) \times M$ restriction matrix R of rank M such that

$$\boldsymbol{\alpha} = R\boldsymbol{\gamma}, \tag{2.3.7}$$

where $\boldsymbol{\gamma}$ is the $M \times 1$ vector of unrestricted parameters. The GLS estimator for $\boldsymbol{\gamma}$ then is

$$\widehat{\boldsymbol{\gamma}} = \left[R'\left(ZZ' \otimes \Sigma_u^{-1}\right)R\right]^{-1} R'\mathrm{vec}(\Sigma_u^{-1}YZ'), \tag{2.3.8}$$

where, as before, $Y \equiv [y_1, \ldots, y_T]$ is $K \times T$ and $Z \equiv [Z_0, \ldots, Z_{T-1}]$ is $(Kp + 1) \times T$. The GLS estimator has standard asymptotic properties under general conditions. In particular, under common assumptions it is consistent and asymptotically normally distributed. More precisely,

$$\sqrt{T}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{d} \mathcal{N}\left(0, (R'\Sigma_{\hat{\boldsymbol{\alpha}}}^{-1}R)^{-1}\right). \tag{2.3.9}$$

In practice, the white noise covariance matrix is unknown and has to be replaced by an estimator that may be based on unrestricted LS estimation of the model. Replacing Σ_u by a consistent estimator $\widehat{\Sigma}_u$ results in a feasible GLS estimator, denoted $\widehat{\widehat{\boldsymbol{\gamma}}}$, with the same asymptotic properties as the GLS estimator (e.g., Lütkepohl 2005, chapter 5). The corresponding feasible GLS estimator of $\boldsymbol{\alpha}$, $\widehat{\widehat{\boldsymbol{\alpha}}} = R\widehat{\widehat{\boldsymbol{\gamma}}}$, is also consistent and asymptotically normal such that

$$\sqrt{T}(\widehat{\widehat{\boldsymbol{\alpha}}} - \boldsymbol{\alpha}) \xrightarrow{d} \mathcal{N}\left(0, R(R'\Sigma_{\hat{\boldsymbol{\alpha}}}^{-1}R)^{-1}R'\right). \tag{2.3.10}$$

Alternatively, an iterated version of the feasible GLS estimator involves reestimating the white noise covariance matrix from the first round feasible GLS residuals and using that estimator in the next round. This procedure may be continued until convergence. The asymptotic properties of the resulting estimators for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are the same as without iteration.

The overall conclusion from this analysis is that standard estimation procedures can be used for restricted VAR models. They are valid asymptotically under the usual assumptions. Likewise, the asymptotic distribution of nonlinear functions of $\boldsymbol{\gamma}$ or $\boldsymbol{\alpha}$ may be obtained by the delta method as in (2.3.6).

As an illustration, consider restricting the lagged feedback from GNP deflator inflation to real GNP growth at all lags in the empirical example already considered in the previous section. In that case $M = 35$ and R is a 39×35 matrix with rows 10, 19, 28, and 37 being zero and all other rows having one unit element and consisting of zeros otherwise. The feasible GLS estimates are

$$\hat{\hat{\nu}} = \begin{bmatrix} 0.7517 \\ -0.3888 \\ 0.0058 \end{bmatrix},$$

$$\hat{\hat{A}}_1 = \begin{bmatrix} 0.2109 & 0.0219 & 0 \\ 0.3131 & 1.0985 & 0.5455 \\ 0.0017 & 0.0631 & 0.4237 \end{bmatrix},$$

$$\hat{\hat{A}}_2 = \begin{bmatrix} 0.1903 & -0.3771 & 0 \\ 0.1835 & -0.4848 & 0.4857 \\ -0.0166 & -0.0513 & 0.2398 \end{bmatrix},$$

$$\hat{\hat{A}}_3 = \begin{bmatrix} 0.0108 & 0.3243 & 0 \\ 0.0297 & 0.4810 & -0.2506 \\ 0.0109 & -0.0047 & 0.0625 \end{bmatrix},$$

$$\hat{\hat{A}}_4 = \begin{bmatrix} -0.0230 & -0.0127 & 0 \\ -0.0202 & -0.1661 & -0.3285 \\ 0.0660 & -0.0132 & 0.2454 \end{bmatrix}.$$

2.3.3 Bias-Corrected LS

The LS estimator of the VAR slope parameters, $\boldsymbol{\alpha}$, may be substantially biased in small samples. For given T, the small-sample bias depends on the specification of the deterministic regressors. If no deterministic regressors are needed for an adequate representation of the DGP, the bias tends to be negligible for realistic sample sizes. The inclusion of an intercept in the VAR model (or equivalently demeaning the data prior to the analysis) substantially increases

the bias for given T. Including an additional deterministic time trend further exacerbates the bias, again holding T fixed.

This LS bias may be estimated and corrected for. There is little interest in applied work in VAR models with a known mean of zero. Closed-form solutions for the asymptotic first-order mean bias in stationary VAR models with an intercept (but no other deterministic regressors) have been derived by Nicholls and Pope (1988) under the assumption of Gaussian iid innovations and by Pope (1990) without assuming Gaussianity. There are no closed-form solutions for the bias in VAR models including deterministic trends. For the latter case, Kilian (1998c) proposes a bootstrap estimator that can be easily adapted to any set of deterministic regressors in the VAR model. This bootstrap estimator is asymptotically as accurate as the closed-form solution when the latter exists, but can accommodate situations for which no closed-form bias estimates are currently available. Unless the VAR model is very large or it includes a deterministic time trend, the closed-form solution is preferred because of its lower computational cost.

Our derivation of the formula for the first-order mean bias for stationary VAR models follows Pope (1990) who postulates a VAR(p) model without a constant term for mean-adjusted data, which is equivalent to fitting a VAR(p) model with constant to the unadjusted data. Without loss of generality the underlying DGP can be written in VAR(1) form similar to (2.2.6),

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

The bias of the LS estimator $\widehat{\mathbf{A}}$ for $\mathbf{A}$ is

$$-\mathbb{B}_{\mathbf{A}}/T + O(T^{-3/2}), \tag{2.3.11}$$

where

$$\mathbb{B}_{\mathbf{A}} = \Sigma_U \Bigg[(I_{Kp} - \mathbf{A}')^{-1} + \mathbf{A}'(I_{Kp} - \mathbf{A}'^2)^{-1} + \sum_{\lambda} \lambda (I_{Kp} - \lambda \mathbf{A}')^{-1} \Bigg] \Gamma_Y(0)^{-1}, \tag{2.3.12}$$

$\Gamma_Y(0) = \mathbb{E}(Y_t Y_t')$, $\Sigma_U = \mathbb{E}(U_t U_t')$ and the sum is over the eigenvalues λ of $\mathbf{A}$, weighted by their multiplicities (see Pope 1990).

The direction of the bias in general depends on the autoregressive roots of the VAR process. For persistent processes, the bias tends to be downward. Obviously, this bias vanishes for $T \to \infty$. Although the bias of the LS estimator does not affect its asymptotic distribution, it may have a substantial effect on the LS estimates in samples of typical size. Using the closed-form solution (2.3.12), a bias-corrected LS estimator for $\mathbf{A}$ may be obtained by

substituting estimators for Σ_U, $\mathbf{A}$, its eigenvalues and $\Gamma_Y(0)$ in the formula for $\mathbb{B}_{\mathbf{A}}$ in expression (2.3.12) to obtain $\widehat{\mathbb{B}}_{\mathbf{A}}$ and adding $\widehat{\mathbb{B}}_{\mathbf{A}}/T$ to the LS estimator of $\mathbf{A}$. As estimators for Σ_U and $\Gamma_Y(0)$ the usual quantities $\widehat{\Sigma}_U = T^{-1}\sum_{t=1}^{T}\widehat{U}\widehat{U}'$ and $\widehat{\Gamma}_Y(0) = T^{-1}\sum_{t=1}^{T} Y_t Y_t'$ can be used. The first K rows of the bias-corrected estimator of $\mathbf{A}$ are the bias-corrected estimators of $A_1, \ldots, A_p$.[3]

A practical concern is that implementing bias adjustments may push the bias-corrected LS estimator into the explosive region (i.e., some of the eigenvalues of the bias-corrected estimator of $\mathbf{A}$ may be outside the complex unit circle), when the original estimate of $\mathbf{A}$ has eigenvalues close to the unit circle. Kilian (1998c) therefore proposes to shrink the bias adjustment such that the bias-corrected estimate remains stationary. If the original estimate of $\mathbf{A}$ is already unstable or explosive, no bias adjustment is carried out. This modification helps reduce the variance of the bias-corrected estimator while preserving the asymptotic normality of the LS estimator.

In practice, users of VAR models are typically not interested in the slope parameters themselves, but in smooth nonlinear functions of the VAR model parameters such as impulse responses. As demonstrated in Kilian (1998c), bias in the slope parameters tends to be associated with severe bias in the impulse response estimator, which often can be ameliorated by using bias-adjusted LS estimators in constructing the impulse responses. It is important to emphasize, however, that unbiasedness is not preserved under nonlinear transformation, so there is no reason for this alternative impulse response estimator to be unbiased in finite samples. Moreover, reductions in bias tend to be associated with increases in the variance, so it is not possible to prove that in general bias corrections will reduce the mean-squared error (MSE) of the impulse response estimator. Indeed, there is no consensus in the literature that impulse responses should be estimated based on bias-adjusted slope parameters rather than the original LS estimates. Simulation experiments, however, show that bias adjustments of the VAR slope parameters systematically and often substantially improve the coverage accuracy of bootstrap confidence intervals for impulse responses and related statistics (see, e.g., Kilian 1999b, 1998b). It is in the latter context that such bias adjustments have become an important tool in applied work. While bias adjustments almost invariably tend to improve the accuracy of bootstrap confidence intervals when the data are persistent, it is important to be aware that for samples of typical size bias adjustments may not be enough to ensure accurate inference in stationary VAR models, when the estimated model includes a deterministic time trend. A more extensive discussion of these issues is provided in Chapter 12.

[3] Iterating this bias correction tends to produce only small improvements in accuracy. Typically, the first-order estimate is quite close to the iterated bias estimate. Likewise, the effect of bias corrections on $\widehat{\Sigma}_U$ is of second order and can be ignored (see Kilian 1998c).

Continuing with the empirical example used in the two preceding subsections, we obtain

$$\widehat{A}_1^{BC} = \begin{bmatrix} 0.2358 & 0.0099 & 0.3947 \\ 0.3145 & 1.1119 & 0.5872 \\ 0.0013 & 0.0634 & 0.4210 \end{bmatrix},$$

$$\widehat{A}_2^{BC} = \begin{bmatrix} 0.2292 & -0.3841 & 0.1307 \\ 0.1820 & -0.4851 & 0.4846 \\ -0.0165 & -0.0517 & 0.2456 \end{bmatrix},$$

$$\widehat{A}_3^{BC} = \begin{bmatrix} 0.0016 & 0.3457 & -0.5544 \\ 0.0204 & 0.4875 & -0.3498 \\ 0.0121 & -0.0059 & 0.0824 \end{bmatrix},$$

$$\widehat{A}_4^{BC} = \begin{bmatrix} -0.0291 & 0.0085 & -0.0458 \\ -0.0333 & -0.1659 & -0.3633 \\ 0.0664 & -0.0121 & 0.2544 \end{bmatrix},$$

where the dominant root, as measured by the maximum of the modulus of the eigenvalues of **A**, increases from 0.9502 for the LS estimate in Section 2.3.1 to 0.9689 after the first-order mean bias adjustment.

2.3.4 Maximum Likelihood Estimation

If the sample distribution is known to have probability density function $f(y_1, \ldots, y_T)$, ML estimation is possible. Denoting the vector of all parameters including the innovation covariance parameters by θ and using the decomposition,

$$f(y_1, \ldots, y_T|\theta) = f_1(y_1) \times f_2(y_2|y_1) \times \cdots \times f_T(y_T|y_{T-1}, \ldots, y_1),$$

the log-likelihood is

$$\log l(\theta|y_1, \ldots, y_T) = \sum_{t=1}^{T} \log f_t(y_t|y_{t-1}, \ldots, y_1). \tag{2.3.13}$$

The maximizing vector $\tilde{\theta}$ is the ML estimator. Under conditions mirroring those required for the LS estimator, this conditional ML estimator has an asymptotic normal distribution,

$$\sqrt{T}(\tilde{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_a(\theta)^{-1}),$$

where $\mathcal{I}_a(\theta)$ is the asymptotic information matrix. Recall that the information matrix is defined as minus the expectation of the Hessian of the log-likelihood,

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log l}{\partial\theta\partial\theta'}\right].$$

The asymptotic information matrix is the limit of this matrix divided by the sample size,

$$\mathcal{I}_a(\theta) = \lim_{T\to\infty} \mathcal{I}(\theta)/T.$$

The asymptotic normality of the ML estimator permits the use of Wald tests as discussed earlier. It also facilitates the use of likelihood ratio (LR) and Lagrange multiplier (LM) tests. Consider testing the hypotheses

$$\mathbb{H}_0 : \varphi(\theta) = 0 \quad \text{versus} \quad \mathbb{H}_1 : \varphi(\theta) \neq 0, \tag{2.3.14}$$

where $\varphi : \mathbb{R}^M \to \mathbb{R}^N$ is a continuously differentiable function, $\varphi(\theta)$ is of dimension $N \times 1$, and the $N \times M$ matrix of first-order partial derivatives $\partial\varphi/\partial\theta'$ is assumed to have rank N when evaluated at the true parameter vector.

Then the LR test statistic is

$$LR = 2[\log l(\tilde{\theta}) - \log l(\tilde{\theta}_r)],$$

where $\tilde{\theta}$ and $\tilde{\theta}_r$ denote the unrestricted and restricted ML estimators, respectively. Under the usual conditions, it has a $\chi^2(N)$ distribution under $\mathbb{H}_0$.

The corresponding LM statistic for testing (2.3.14) is based on the score vector

$$s(\theta) = \frac{\partial \log l(\theta)}{\partial\theta},$$

which is zero when evaluated at the unrestricted ML estimator. The LM statistic measures the distance of this score vector from zero, when the score vector is evaluated at the restricted estimator. It is defined as

$$LM = s(\tilde{\theta}_r)'\mathcal{I}(\tilde{\theta}_r)^{-1}s(\tilde{\theta}_r) \tag{2.3.15}$$

and has an asymptotic $\chi^2(N)$ distribution under $\mathbb{H}_0$. It can be expressed equivalently as

$$LM = \widetilde{\lambda}' \left[\frac{\partial\varphi}{\partial\theta'}\bigg|_{\tilde{\theta}_r} \right] \mathcal{I}(\tilde{\theta}_r)^{-1} \left[\frac{\partial\varphi'}{\partial\theta}\bigg|_{\tilde{\theta}_r} \right] \widetilde{\lambda}, \tag{2.3.16}$$

where $\widetilde{\lambda}$ is the vector of Lagrange multipliers for which the Lagrange function has derivative zero when evaluated at the constrained estimator (see Lütkepohl 2005, appendix C.7).

Wald, LR, and LM test statistics have the same asymptotic distributions under the null hypothesis. They are based on the unrestricted estimator only, both the unrestricted and the restricted estimators, and the restricted estimator only, respectively. The Wald test has the disadvantage that it is not invariant under nonlinear transformations of the restrictions. Its small-sample power may be low, as shown by Gregory and Veall (1985) and Breusch and Schmidt (1988), for example.

ML theory is valid for large classes of distributions. In VAR analysis, it is common to postulate that the innovations, u_t, are iid $\mathcal{N}(0, \Sigma_u)$ random variables. This assumption implies that the y_t $(= \nu + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t)$ are also jointly normal and, for given initial values $y_{-p+1}, \ldots, y_0$,

$$f_t(y_t|y_{t-1}, \ldots, y_{-p+1}) = \left(\frac{1}{2\pi}\right)^{K/2} \det(\Sigma_u)^{-1/2} \exp\left(-\frac{1}{2} u_t' \Sigma_u^{-1} u_t\right). \tag{2.3.17}$$

Hence, the log-likelihood becomes

$$\begin{aligned} \log l = &- \frac{KT}{2} \log 2\pi - \frac{T}{2} \log(\det(\Sigma_u)) \\ &- \frac{1}{2} \sum_{t=1}^{T} (y_t - \nu - A_1 y_{t-1} - \cdots - A_p y_{t-p})' \Sigma_u^{-1} \\ &\qquad \times (y_t - \nu - A_1 y_{t-1} - \cdots - A_p y_{t-p}). \end{aligned} \tag{2.3.18}$$

Maximizing this function with respect to the unknown parameters yields the Gaussian ML estimator. If there are no restrictions on the parameters, this ML estimator for $\boldsymbol{\alpha} = \text{vec}[\nu, A_1, \ldots, A_p]$ is identical to the LS estimator and thus has the same asymptotic distribution as the LS estimator. Hence, in practice, one can rely on expression (2.3.2) when constructing the ML estimator of $\boldsymbol{\alpha}$. The corresponding ML estimator $\widetilde{\Sigma}_u = T^{-1} \widehat{U}\widehat{U}'$ may be obtained as $\widetilde{\Sigma}_u = \widehat{\Sigma}_u (T - Kp - 1)/T$.

The numerical equivalence of the LS and the Gaussian ML estimator of $\boldsymbol{\alpha}$ holds even when the true distribution of y_t is not Gaussian. In the latter case, the estimator is referred to as a quasi-ML or pseudo-ML estimator. It should be understood, however, that if the true error distribution is different from the normal distribution, it may be possible to obtain more precise ML estimators by imposing the true distribution in constructing the likelihood. In that case, maximizing the log-likelihood involves a nonlinear optimization problem and requires suitable iterative algorithms.

If there are additional restrictions on the parameters as in Section 2.3.2 and these restrictions are taken into account in the estimation, the Gaussian ML estimator may differ from the restricted GLS estimator, but again the same asymptotic properties are obtained. In this case, implementing the restricted ML estimator also requires numerical optimization methods.

Returning to the empirical example, the Gaussian ML estimator of the slope parameters and of the intercept are identical to the LS estimator. The corresponding ML estimator of Σ_u is

$$\widetilde{\Sigma}_u = \begin{bmatrix} 0.5656 & 0.0746 & -0.0201 \\ 0.0746 & 0.6157 & 0.0351 \\ -0.0201 & 0.0351 & 0.0642 \end{bmatrix}.$$

2.3.5 VAR Processes in Levels with Integrated Variables

One may also estimate a VAR process with integrated variables using the same techniques. The LS/ML estimator is consistent and asymptotically normal under general conditions and the same estimator may be used as in the stationary case (see Park and Phillips 1988, 1989; Sims, Stock, and Watson 1990; Lütkepohl 2005, chapter 7). More precisely, in VAR(p) models with $p > 1$, standard Gaussian inference on individual VAR slope parameters remains asymptotically valid even in the presence of $I(1)$ variables. The intuition is that the slope parameter matrices A_i in levels VARs can be written as linear combinations of estimators of other parameter matrices that are asymptotically Gaussian because they relate to stationary regressors.

Note that the VAR(p) model

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

is algebraically equivalent to the reparametrized specification

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

where $C = \sum_{i=1}^{p} A_i$ and $\mathbf{\Gamma}_i = -(A_{i+1} + \cdots + A_p)$, $i = 1, \ldots, p-1$. The latter representation may be viewed as the multivariate analogue of the augmented Dickey-Fuller (ADF) representation of the univariate AR model and is derived in the same manner. Estimating this reparameterized model by LS and computing estimates

$$\begin{aligned}
\widehat{A}_1 &= \widehat{C} + \widehat{\mathbf{\Gamma}}_1, \\
\widehat{A}_i &= \widehat{\mathbf{\Gamma}}_i - \widehat{\mathbf{\Gamma}}_{i-1}, \quad i = 2, \ldots, p-1, \\
\widehat{A}_p &= -\widehat{\mathbf{\Gamma}}_{p-1},
\end{aligned}$$

is equivalent to estimating the levels VAR representation directly by LS. Assuming that all variables in y_t are at most $I(1)$, the estimators $\widehat{\mathbf{\Gamma}}_i$, $i = 1, \ldots, p-1$, converge at rate $\sqrt{T}$ and have asymptotically normal marginal distributions because the lagged differences are $I(0)$ regressors. Since the asymptotic distribution of $\widehat{A}_i$, $i = 2, \ldots, p$, consists of linear combinations of asymptotically normal estimators, by construction they must be asymptotically normal as well. Likewise, $\widehat{A}_1$ can be shown to be asymptotically normal. Note that y_{t-1} is an $I(1)$ regressor, so the LS estimator $\widehat{C}$ may converge at rate T (see Chapter 3 for details). If so, it does not enter the limiting distribution of $\widehat{A}_1$ after $\sqrt{T}$-scaling. This fact allows us to ignore that $\widehat{C}$ may have a non-Gaussian limiting distribution after T-scaling. In other words, the nature of the asymptotic distribution of $\widehat{C}$ is irrelevant for the derivation of the asymptotic distribution of $\widehat{A}_1$. The latter distribution is dominated by the distribution of $\widehat{\mathbf{\Gamma}}_1$, which in turn is known to be asymptotically normal. To summarize, provided $p > 1$ in the VAR(p) model and an intercept is included in estimation, the LS estimator

of $\widehat{A}_i$, $i = 1, \ldots, p$, remains consistent and the marginal asymptotic distributions remain asymptotically normal even in the possible presence of a unit root or, for that matter, a near-unit root (see Chapter 3).

There is a difference in the joint asymptotic properties of the estimated VAR coefficients, however, that is worth emphasizing. If there are integrated variables, the covariance matrix $\Sigma_{\hat{\alpha}}$ of the asymptotic distribution is singular because some components of the estimator (or their linear combinations) converge at rate T rather than $\sqrt{T}$. As a result, standard tests of hypotheses involving several VAR parameters jointly may be invalid asymptotically (see Toda and Phillips 1993). Hence, caution is called for in conducting inference.

Toda and Yamamoto (1995) and Dolado and Lütkepohl (1996) show that a reparametrization of the model allows us to overcome the problems due to a nonsingular covariance matrix of the estimator. They show that if y_t consists of $I(0)$ and $I(1)$ variables only, it suffices to add an extra lag to the VAR process fitted to the data to obtain a nonsingular covariance matrix for the parameters associated with the first p lags. In other words, if the true DGP is a VAR(p) process and a lag-augmented VAR($p + 1$),

$$y_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + A_{p+1} y_{t-p-1} + u_t,$$

is fitted by LS, then the estimator of $[A_1, \ldots, A_p]$ has a nonsingular joint asymptotic distribution. Hence, testing hypotheses on linear combinations of these parameters with a Wald test with the usual χ^2 distribution is asymptotically valid. More generally, if $I(d)$ variables are present, the singularity problem of the covariance matrix can be resolved by augmenting the VAR model by d extra lags for estimation. Extensions of this result to unit root, local-to-unity, and long memory VAR processes of infinite order have been provided in Bauer and Maynard (2012) (see Chapter 3).

Of course, lag augmentation may involve a loss of efficiency in estimation, reducing the power of tests and inflating the width of confidence intervals, especially when the parameters of interest would have been estimated superefficiently in the absence of the redundant lag. Thus, there is a trade-off between robustness and efficiency.[4]

If there are parameter restrictions and GLS estimation is applied to $I(1)$ processes, a more detailed analysis of the integration and cointegration properties of the left-hand and right-hand side variables is called for to determine the asymptotic properties of the estimators and the associated inference procedures. It may then be preferable to estimate the process in vector error correction form. Suitable procedures are discussed in Chapter 3.

[4] An alternative estimation approach that is robust to moderate deviations from unit roots in either direction and that does not require lag augmentation is discussed in Magdalinos and Phillips (2009). It is not clear, however, what their asymptotic results imply for inference on structural impulse responses and related statistics.

Returning to our empirical example, one could alternatively interpret i_t and Δp_t as $I(1)$ variables (such that p_t is $I(2)$ rather than $I(1)$) with a cointegrating relationship between i_t and Δp_t such that the real interest rate $i_t - \Delta p_t$ is $I(0)$. The results of this section imply that using the LS estimates obtained earlier under the assumption of stationary data would remain justified in this case.

Alternatively, one could fit a lag-augmented VAR model. When augmenting the VAR(4) model already used as an empirical example earlier in this chapter by an additional lag and fitting a VAR(5) model to the data, we obtain the LS estimates

$$\widehat{A}_1 = \begin{bmatrix} 0.2114 & 0.0465 & 0.3993 \\ 0.3220 & 1.0879 & 0.4908 \\ 0.0044 & 0.0553 & 0.4234 \end{bmatrix},$$

$$\widehat{A}_2 = \begin{bmatrix} 0.1694 & -0.4387 & 0.1244 \\ 0.2091 & -0.4762 & 0.5008 \\ -0.0154 & -0.0458 & 0.2429 \end{bmatrix},$$

$$\widehat{A}_3 = \begin{bmatrix} 0.0199 & 0.3888 & -0.5773 \\ 0.0080 & 0.4586 & -0.2666 \\ 0.0083 & -0.0062 & 0.1125 \end{bmatrix},$$

$$\widehat{A}_4 = \begin{bmatrix} 0.0191 & -0.1489 & -0.2000 \\ -0.0706 & -0.1259 & -0.2355 \\ 0.0606 & -0.0019 & 0.3142 \end{bmatrix},$$

and

$$\widehat{\Sigma}_u = \begin{bmatrix} 0.5852 & 0.1034 & -0.0147 \\ 0.1034 & 0.6346 & 0.0339 \\ -0.0147 & 0.0339 & 0.0675 \end{bmatrix}.$$

The estimate of the augmented lag term, which under the maintained assumption of a VAR(4) DGP is known to be $0_{3\times 3}$ in population, is

$$\widehat{A}_5 = \begin{bmatrix} -0.0888 & 0.1173 & 0.2491 \\ 0.2161 & -0.0208 & 0.0171 \\ 0.0179 & -0.0053 & -0.1335 \end{bmatrix}.$$

The latter estimate is ignored in conducting further analysis.

2.3.6 Sieve Autoregressions

It is easy to construct plausible autoregressive DGPs that are not of finite lag order. For example, the autoregressive representation of an invertible VARMA process may be a VAR(∞) process. MA components arise naturally when data are aggregated over time, across households or across sectors, or when interest

centers on a subset of the variables in a higher-dimensional VAR process (see the discussion in Section 2.2.4). They also typically arise when the data are generated by a dynamic stochastic general equilibrium (DSGE) model (see Chapter 6).

Once we allow for the possibility that the DGP is a linear VAR(∞) process, we can reinterpret the fitted VAR model with lag order p_T as an approximation to the infinite-order DGP. The subscript is a reminder that the approximating lag order depends on T. The thought experiment is that the researcher fits a sequence of finite-order autoregressions, the lag order of which is assumed to increase at a suitable rate with the sample size. The fitted VAR model is viewed as an approximation to the possibly infinite-order autoregression, the quality of which improves with the sample size. In the limit, the approximation error becomes negligible. Hence, claims that VAR models suffer from omitted-variable bias when the DGP is a VARMA model are not correct in general (see, e.g., Braun and Mittnik 1993; Cooley and Dwyer 1998; Yao, Kam, and Vahid 2017).

Whereas the traditional finite-order VAR model has been developed as a parametric time series model, the VAR(p_T) model is a semiparametric model designed to approximate general linear time series models. Such methods are also commonly referred to as sieve autoregressions in the literature. Lewis and Reinsel (1985) establish that the LS estimator $\widetilde{\boldsymbol{\alpha}}_T$ remains consistent and asymptotically normal when fitting a VAR(p_T) model to a stationary VAR(∞) process, provided $p_T \to \infty$ at a rate that is not too fast and not too slow. In particular, we require the lower bound on the lag order to increase with T such that

$$\sqrt{T} \sum_{i=p_T+1}^{\infty} ||A_i|| \to 0$$

and the upper bound to increase such that $p_T^3/T \to 0$. Analogous results also apply to the estimation of the innovation variance Σ_u (Lütkepohl and Poskitt 1991; Lütkepohl 2005, section 15.2). Gonçalves and Kilian (2007) generalize this result to infinite-order processes with conditionally heteroskedastic martingale difference errors under the stronger assumption that $p_T^4/T \to 0$. The sieve approach has also proved useful in studying cointegrated processes, building on the framework developed by Saikkonen (1992) and Saikkonen and Lütkepohl (1996), who consider approximating cointegrated linear systems with iid innovations via autoregressive sieves. Thus, allowing for approximation error does not affect how the VAR parameters are estimated. The only difference is the interpretation of the fitted model and the fact that achieving a good approximation may require a larger lag order than commonly considered in applied work.

There are some subtle differences, however. For example, Onatski and Uhlig (2012) show that using the sieve approach may result in estimated processes

with roots near the unit circle, even if the actual underlying DGP is a stationary process with autoregressive roots well away from the unit circle. In fact, even if the DGP is white noise and the sieve approach is used for estimation, estimated roots near the unit circle will eventually be obtained. Thus, inferring from estimated roots near the unit circle that the underlying DGP is very persistent can be misleading.

Another important caveat is that the validity of the autoregressive sieve approximation to a VARMA process has not been investigated for VARMA DGPs, in which some roots of the autoregressive lag polynomial and of the moving-average lag polynomial nearly cancel. As noted earlier, these roots are functions of the model coefficients. For expository purposes consider the limiting case of a VARMA(1, 1) process, in which the values of the MA and AR coefficient matrices (and hence the roots) are exactly equal. In that case,

$$y_t = A_1 y_{t-1} + u_t - M_1 u_{t-1}$$

with white noise errors, u_t, where $A_1 = M_1$. This process is equivalent to the white noise process

$$y_t = u_t,$$

rendering the slope parameters A_1 and M_1 unidentified. A perhaps practically more relevant case is a VARMA(1, 1) model, in which some AR and MA roots are nearly equal in the sense that their difference is local to zero (see Andrews and Mikusheva 2015). This situation gives rise to a near-identification problem in the VARMA model that is likely to complicate estimation and inference for autoregressive sieve approximations. The theoretical properties of estimators of sieve autoregressions in this setting are not known. In practice, users of the sieve VAR approach effectively assume that there are no near-identification problems.

Extensions of the asymptotic properties of sieve estimators to smooth and differentiable functions of the slope parameters are straightforward, following the same line of reasoning as in Section 2.3.1. For example, Lütkepohl (1988) derives the asymptotic normal distribution of the estimated dynamic multipliers for the VAR(∞) model. Under standard assumptions estimators of impulse responses and related statistics will also be $\sqrt{T}$-consistent and asymptotically normal. There is one important difference, however. Notwithstanding the invariance of the distribution of the LS estimator of the VAR parameters, the asymptotic variance of the dynamic multipliers will differ, depending on whether the underlying DGP is a finite-order VAR model or a VAR(∞) model, forcing the user to choose between alternative asymptotic approximations. This choice may be avoided by the use of bootstrap methods of inference. For example, Paparoditis (1996) showed that an asymptotically valid bootstrap approximation of the distribution of dynamic multipliers can be constructed in exactly the same way whether the underlying DGP is a finite-order VAR

model or a VAR(∞) model. Further results for structural VAR models can be found in Inoue and Kilian (2002b). We will return to this point in Chapter 12 when discussing inference on structural impulse responses.

2.4 Prediction

It is useful to review some of the basic facts about VAR prediction. Reduced-form VAR models represent the conditional mean of a stochastic process given past observations. Hence, they are natural tools for prediction. Before we discuss how to predict future realizations of the data from estimated VAR processes, it is useful to treat the true process as known, allowing us to neglect estimation uncertainty.

2.4.1 Predicting from Known VAR Processes

Suppose y_t is generated by the K-dimensional VAR(p) process (2.2.4),

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

If the white noise process u_t is a martingale difference such that $\mathbb{E}(u_t|y_{t-1}, y_{t-2}, \dots) = \mathbb{E}(u_t|u_{t-1}, u_{t-2}, \dots) = 0$, then

$$\begin{aligned} y_{T+h|T} &\equiv \mathbb{E}(y_{T+h}|y_T, y_{T-1}, \dots) \\ &= \nu + A_1 y_{T+h-1|T} + \cdots + A_p y_{T+h-p|T}, \end{aligned} \tag{2.4.1}$$

where $y_{T+j|T} = y_{T+j}$ for $j \leq 0$, is the optimal, minimum mean squared error (MSE) h-step ahead prediction, given y_t, $t \leq T$. The martingale difference property of u_t is, for instance, satisfied if the u_t are a zero mean iid sequence (see Section 2.2.1). By iterating forward equation (2.4.1), the predictions can easily be computed recursively for $h = 1, 2, \dots$.

The prediction error associated with an h-step ahead prediction is

$$y_{T+h} - y_{T+h|T} = u_{T+h} + \Phi_1 u_{T+h-1} + \cdots + \Phi_{h-1} u_{T+1}, \tag{2.4.2}$$

where $\Phi_i = J\mathbf{A}^i J'$, $i = 0, 1, \dots$, as defined in Section 2.2. Equation (2.4.2) implies that the VAR innovation u_t is the prediction error for the 1-step ahead prediction as of period $t-1$. The errors have mean zero and, hence, the predictions are unbiased. The prediction error covariance or mean-squared prediction error (MSPE) matrix is

$$\Sigma_y(h) \equiv \mathbb{E}[(y_{T+h} - y_{T+h|T})(y_{T+h} - y_{T+h|T})'] = \sum_{j=0}^{h-1} \Phi_j \Sigma_u \Phi_j'. \tag{2.4.3}$$

In short, $y_{T+h} - y_{T+h|T} \sim (0, \Sigma_y(h))$. It is worth emphasizing that these expressions hold not only for stationary but also for integrated processes for any finite h.

If y_t is $I(0)$, then, as $h \to \infty$,

$$\Sigma_y(h) \to \Sigma_y \equiv \Gamma_y(0)$$

(see expression (2.2.10)). In other words, the prediction error covariance matrix converges to the unconditional covariance matrix of y_t when the prediction horizon goes to infinity and, hence, the prediction uncertainty remains bounded. In contrast, for integrated processes, the Φ_i do not converge to zero for $i \to \infty$ and the unconditional covariance matrix of y_t does not exist. Thus, for integrated processes, the prediction error covariance matrix is unbounded in the limit.

If the u_t are just uncorrelated and not martingale differences (i.e., the conditional mean, given past innovations, is not zero), the predictions obtained recursively from equation (2.4.1) are still the best linear predictions, but may not be minimum MSPE in a larger class of possibly nonlinear predictors.

Note that the prediction error does not depend on the deterministic term ν. In fact, all deterministic terms such as deterministic trend polynomials cancel in constructing the prediction error. Hence, they do not contribute to the prediction uncertainty. Many researchers find it implausible that trending behavior is not reflected in the uncertainty about long-term predictions. In the present setup, this is an implication of our assumption that the parameters of the true process are known. If the deterministic terms are estimated, they contribute to the variance of the predictions, as discussed in the next subsection.

2.4.2 Predicting from Estimated VAR Processes

In practice, the VAR process is unknown and estimated parameters are used in computing predictions. Denoting an h-step prediction based on the estimated process by $\hat{y}_{T+h|T}$, the prediction error is

$$y_{T+h} - \hat{y}_{T+h|T} = (y_{T+h} - y_{T+h|T}) + (y_{T+h|T} - \hat{y}_{T+h|T}). \tag{2.4.4}$$

Expression (2.4.2) shows that the first term on the right-hand side includes innovations u_t for $t > T$ only. In contrast, the second term on the right-hand side of expression (2.4.4) involves only observations y_t up to time T. Assuming that the model has been estimated on data up to time T, these two terms are independent (or at least uncorrelated), and the MSE matrix is

$$\begin{aligned}\Sigma_{\hat{y}}(h) &= \mathbb{E}[(y_{T+h} - \hat{y}_{T+h|T})(y_{T+h} - \hat{y}_{T+h|T})'] \\ &= \Sigma_y(h) + \mathrm{MSE}(y_{T+h|T} - \hat{y}_{T+h|T}),\end{aligned} \tag{2.4.5}$$

where the first term on the right-hand side denotes the population MSPE and the second term captures the estimation uncertainty. For a properly specified VAR model, the last term in this relation approaches zero, as the sample size grows, because, under standard assumptions, the difference $y_{T+h|T} - \hat{y}_{T+h|T}$ vanishes asymptotically in probability as $T \to \infty$. Thus, estimation uncertainty is negligible asymptotically if the fitted model is the DGP. In practice, only finite samples are available, and the precision of the predictions depends on the precision of the parameter estimators. Finite-sample correction factors for MSPEs and prediction intervals for stationary processes are provided in Baillie (1979), Reinsel (1980), Samaranayake and Hasza (1988), and Lütkepohl (2005, chapter 3).

There are several obvious extensions of this framework. First, one may consider explicitly the situation where the DGP is only approximated by the finite-order VAR model used for prediction. Second, if h-step ahead predictions are of interest, one may instead fit models specifically targeted at this horizon. Such extensions are discussed in Lütkepohl (2009).

If $u_t \overset{iid}{\sim} \mathcal{N}(0, \Sigma_u)$, the prediction errors are also multivariate normal, $y_{T+h} - y_{T+h|T} \sim \mathcal{N}(0, \Sigma_y(h))$, and prediction intervals may be set up in the usual way based on the MSPE matrices (2.4.5) (see Lütkepohl 2005). Constructing prediction intervals for non-Gaussian VAR processes subject to estimation uncertainty requires the use of bootstrap methods, as discussed in Chapter 12.

2.5 Granger Causality Analysis

The VAR model describes the joint DGP of the variables under consideration. A proposal for assessing the dynamic relationship between economic variables based on the VAR model was made by Granger (1969). His definition of what has become known as Granger causality is based on the notion of linear predictability. Granger calls a variable y_{2t} causal for a variable y_{1t} if the information in past and present values of y_{2t} helps reduce in expectation the squared prediction error for y_{1t}. Many researchers find it problematic to interpret a predictive relationship as a causal relationship. We provide a more extensive discussion of this issue in Chapter 7. Here we simply present a formal definition of what has become known as Granger causality.

Let Ω_t be the information set containing all information relevant to predicting y_1 available up to and including period t. Denote the optimal (minimum MSE) h-step prediction of y_{1t} at origin t, based on the information in Ω_t, by $y_{1,t+h|\Omega_t}$ and the corresponding MSPE by $\sigma^2_{y_1}(h|\Omega_t)$. Then the process y_{2t} is said to Granger cause y_{1t} if

$$\sigma^2_{y_1}(h|\Omega_t) < \sigma^2_{y_1}(h|\Omega_t \setminus \{y_{2s}|s \leq t\}) \quad \text{for at least one } h \in \{1, 2, \dots\}. \tag{2.5.1}$$

Here $\Omega_t \setminus \{y_{2s} | s \leq t\}$ denotes the set of all relevant information in the universe apart from the past and present information about the y_{2t} process. In other words, y_{2t} is Granger causal for y_{1t} if the latter variable can be predicted with lower mean squared error by taking into account the information in y_{2s}, $s \leq t$, in addition to all other relevant information.

This concept is easy to implement within a VAR framework if Ω_t is limited to past and present values of y_{1t} and y_{2t}. Suppose these two variables are generated by a bivariate VAR(p) process,

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} + \sum_{i=1}^{p} \begin{bmatrix} a_{11,i} & a_{12,i} \\ a_{21,i} & a_{22,i} \end{bmatrix} \begin{pmatrix} y_{1,t-i} \\ y_{2,t-i} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}.$$

Then y_{2t} is not Granger causal for y_{1t} if and only if $a_{12,i} = 0, i = 1, 2, \ldots, p$. In other words, no lags of y_{2t} appear in the y_{1t} equation of the model. In contrast, y_{2t} is Granger causal for y_{1t} if lags of the former variable appear in the y_{1t} equation with nonzero coefficients.

Given that Granger-noncausality is characterized by zero restrictions on the VAR coefficients, it can be tested using a standard Wald test if the process is stationary. If the process contains integrated variables, the testing problem becomes more complicated because the Wald test statistics will not have their usual asymptotic distributions anymore (see Toda and Phillips 1993). If the variables are known to be integrated but not cointegrated, one can appeal to standard asymptotics for the Granger causality test after differencing the data. If the model variables are known to be integrated and cointegrated, Granger causality may be assessed within the vector error correction framework considered in Chapter 3, but the distribution theory may be nonstandard and the asymptotic distributions depend on nuisance parameters.

If we are unsure about the integration and cointegration status of the model variables, an easy cure for this problem is to add a further lag to the VAR process and perform the tests on the first p lags of the lag-augmented VAR, as discussed in Section 2.3.5. More precisely, if the true process is a VAR(p), then a VAR($p+1$) is also a true process with $A_{p+1} = 0$. Since the last lag has coefficient matrix zero, the test for Granger causality can be performed on the first p lags. If the variables are at most $I(1)$, one extra lag is sufficient to ensure that the standard tests have standard asymptotic properties under the null hypothesis. More lags have to be added if the order of integration is higher than 1 (see Toda and Yamamoto 1995; Dolado and Lütkepohl 1996). Extensions to testing Granger causality in infinite-order VAR processes are discussed in Lütkepohl and Poskitt (1996) for stationary processes, in Saikkonen and Lütkepohl (1996) for integrated processes, and in Bauer and Maynard (2012) for local-to-unity processes and fractionally integrated processes (see Chapter 3). As noted in Section 2.3.5, the use of lag augmentation involves a loss of power. There is evidence, however, that the magnitude of this efficiency loss is small

in practice. For example, simulation results in Dolado and Lütkepohl (1996) for lag-augmented Wald tests of Granger causality suggest that the power loss is negligible, provided a reasonably large number of lags is used in the first place.[5]

The introduction of the concept of Granger causality has stimulated a heated debate about causality in econometrics and economics. Some of this discussion has focused on technical problems such as how to extend the information set. While it is straightforward to extend the concept of Granger causality to allow for vectors of variables y_{1t} and y_{2t} instead of scalars (e.g., Lütkepohl 2005, section 2.3.1), it is more difficult to assess the Granger causality for individual variables y_{2t} to y_{1t} in the presence of additional variables. In this case, the conditions for Granger causality become more complicated, even within the framework of finite-order VAR processes (see, e.g., Lütkepohl 1993; Dufour and Renault 1998). The Granger causal ordering of the variables may change if the information set changes. For example, if y_{2t} is Granger causal for y_{1t} in a bivariate model, it may not be Granger causal after adding a third variable to the VAR model, if that third variable is dynamically correlated with the two original variables. It is also possible that y_{2t} is not Granger causal for y_{1t} in a bivariate model and becomes Granger causal if the information set is extended to include other variables, as pointed out in Lütkepohl (1982a).

Another limitation of the concept of Granger causality is that it does not speak to the instantaneous relationships of the variables. This problem can be overcome by defining what has sometimes been referred to as instantaneous causality. In this context a variable y_{2t} is said to be instantaneously causal for y_{1t} if

$$\sigma^2_{y_1}(1|\Omega_t \cup \{y_{2,t+1}\}) < \sigma^2_{y_1}(1|\Omega_t). \tag{2.5.2}$$

In other words, taking into account $y_{2,t+1}$ improves the 1-step prediction of y_{1t}. It turns out that this concept is symmetric, i.e., instantaneous causality from y_{2t} to y_{1t} implies instantaneous causality in the reverse direction from y_{1t} to y_{2t} as well. In fact, in a bivariate VAR model this relation is equivalent to instantaneous correlation of the innovations u_{1t} and u_{2t} (Lütkepohl 2005, section 2.3.1). This observation highlights a fundamental problem with this concept. It fails to distinguish between correlation and causality and it does not speak to the direction of causality. A more extensive discussion of the causality issue is provided in Chapter 7.

[5] An alternative approach to testing Granger causality, when the model variables are potentially integrated and/or cointegrated, that does not involve lag augmentation has been proposed by Toda and Phillips (1993).

2.6 Lag-Order Selection Procedures

In Section 2.3, the lag order p of the VAR model was assumed to be known when estimating the model. In practice, p often has to be chosen on the basis of the available sample of data. Sequential testing procedures and information criteria are commonly used tools for deciding on an adequate VAR order. Both types of procedures require the specification of an upper bound and a lower bound on the range of admissible lag orders. The maximum lag order that is considered reasonable is denoted by $p_{\max}$ in the following. Typically, the corresponding $p_{\min}$ is set to zero. For sieve autoregressions, both the lower bound $p_{\min}$ and the upper bound $p_{\max}$ on the lag order are required to increase with the sample size T at a suitable rate.

2.6.1 Top-Down Sequential Testing

One possible sequential procedure involves testing the sequence of null hypotheses: $\mathbb{H}_0 : A_{p_{\max}} = 0$ vs. $\mathbb{H}_1 : A_{p_{\max}} \neq 0$, $\mathbb{H}_0 : A_{p_{\max}-1} = 0$ vs. $\mathbb{H}_1 : A_{p_{\max}-1} \neq 0, \ldots, \mathbb{H}_0 : A_1 = 0$ vs. $\mathbb{H}_1 : A_1 \neq 0$. The procedure continues as long as the null hypothesis is not rejected. If there is a rejection, the testing procedure terminates, and we conclude that we require as many lags as maintained under the last alternative hypothesis. If none of the null hypotheses can be rejected, we conclude that $p = 0$. Such procedures are known as top-down or general-to-specific procedures because they start with the largest model and sequentially reduce the lag order of the model. In implementing this sequential test, the usual Wald or LR tests for parameter restrictions can be employed. For example, the LR statistic for testing a VAR(m) against a VAR($m+1$) has the form

$$LR(m) = T[\log(\det(\widetilde{\Sigma}_u(m))) - \log(\det(\widetilde{\Sigma}_u(m+1)))],$$

where

$$\widetilde{\Sigma}_u(m) = T^{-1} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$$

is the ML estimator of the residual covariance matrix for a VAR(m) model. If $\mathbb{H}_0 : A_{m+1} = 0$ is tested against $\mathbb{H}_1 : A_{m+1} \neq 0$, the $LR(m)$ statistic has an asymptotic $\chi^2(K^2)$ distribution under $\mathbb{H}_0$.

In implementing this procedure it may be worth taking into account that the small-sample distributions of the test statistics can differ substantially from their asymptotic counterparts, in particular for larger models with many variables and lags (e.g., Lütkepohl 2005, section 4.3.4). A simple small-sample adjustment proposed by Sims (1980a) has the form

$$LR^{adj}(m) = (T - K(m+1)) \times [\log(\det(\widetilde{\Sigma}_u(m))) - \log(\det(\widetilde{\Sigma}_u(m+1)))].$$

The LR statistic $LR(m)$ is suitable for testing the VAR(m) model against the VAR($m+1$) model. It is also possible to test each lag length against the full VAR($p_{\max}$) model using the LR test statistic

$$LR^*(m) = T[\log(\det(\widetilde{\Sigma}_u(m))) - \log(\det(\widetilde{\Sigma}_u(p_{\max})))]$$

or a similar LR statistic including small-sample adjustments. An alternative approach, which may be viewed as a generalization of the sequential t-test for selecting the lag order of a univariate autoregression, is the sequential Wald test.

A common problem with all such sequential testing procedures is how to control the overall size of the test. A related point of concern is that using the same critical value at each stage of the test does not account for the fact that the tests are conditional on the outcomes of earlier tests (for further discussion, see Lütkepohl 2005, section 4.2.3).

2.6.2 Bottom-Up Sequential Testing

Alternatively, one may consider a bottom-up or specific-to-general approach by starting from the smallest model and adding lags only if residual autocorrelation tests suggest that the model does not fully capture the dynamic structure in the data. In other words, tests for residual autocorrelation are applied to the VAR(0) model, VAR(1) model, and so forth. The testing procedure terminates when no statistically significant autocorrelation is found. Thus, we do not have to fix the maximum VAR lag order in advance. Examples of sequential tests for residual autocorrelation in VAR models are the Portmanteau and LM tests.

Portmanteau Test for Residual Autocorrelation. The Portmanteau test for autocorrelation in the innovations evaluates the null hypothesis $\mathbb{H}_0$: $\mathbb{E}(u_t u'_{t-i}) = 0, i = 1, 2, \ldots$. The alternative hypothesis is that at least one autocovariance is nonzero. The test statistic is

$$Q_h = T \sum_{j=1}^{h} \operatorname{tr}(\hat{C}'_j \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1}), \tag{2.6.1}$$

where $\hat{C}_j = T^{-1} \sum_{t=j+1}^{T} \hat{u}_t \hat{u}'_{t-j}$ and the $\hat{u}_t$ are the LS residuals. In other words, the test statistic is based on the estimated residual autocovariances. If the VAR(p) process is stationary, no parameter restrictions are imposed and the errors are serially independent, the distribution of Q_h under $\mathbb{H}_0$ is approximately $\chi^2(K^2(h-p))$ when both T and h are large. The approximate χ^2 distribution is obtained, specifically, as $h/T \to 0$ for $T \to \infty$.

The degrees of freedom have to be adjusted if there are parameter restrictions. In that case, they are obtained as the difference between the number

of (non-instantaneous) autocovariances included in the statistic (K^2h) and the number of estimated VAR parameters (e.g., Ahn 1988; Hosking 1980, 1981a, 1981b; Li and McLeod 1981; Lütkepohl 2005, section 5.2.9). Likewise, if the VAR process contains integrated components, the degrees of freedom are affected (Brüggemann, Lütkepohl, and Saikkonen 2006). Since the degrees of freedom depend on the typically unknown cointegration properties of the process, the Portmanteau test cannot be recommended for nonstationary processes. Moreover, Francq and Raïssi (2007) show that the asymptotic distribution may be quite different from the usual χ^2 distribution if the errors are not autocorrelated but still contain some dependence structure in higher-order moments. Thus, the test has to be used with caution.

Even when the necessary conditions for deriving the asymptotic distribution hold, the approximate χ^2 distribution may be far from the actual distribution in small samples. To improve the match between actual and approximating distribution, the following modified statistic was proposed by Hosking (1980):

$$Q_h^* = T^2 \sum_{j=1}^{h} \frac{1}{T-j} \operatorname{tr}(\hat{C}_j' \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1}).$$

The number h of autocovariance terms in the test statistic should be considerably larger than p for a good approximation to the null distribution. Choosing h too large, however, may undermine the power of the test. In practice, usually a number of different values of h is considered. The Portmanteau test should only be applied to test for a large number of nonzero autocovariances. It is not suitable for testing the absence of low-order autocorrelation. For the latter purpose, the LM test is preferred.

LM Test for Residual Autocorrelation. The LM test for autocorrelation in the innovations was proposed by Breusch (1978) and Godfrey (1978). It may be viewed as a test for zero coefficient matrices in the model

$$u_t = D_1 u_{t-1} + \cdots + D_h u_{t-h} + e_t.$$

The quantity e_t denotes a white noise error term. Thus, testing

$$\mathbb{H}_0 : D_1 = \cdots = D_h = 0 \quad \text{versus}$$
$$\mathbb{H}_1 : D_i \neq 0 \text{ for at least one } i \in \{1, \ldots, h\}$$

is a test for autocorrelation in u_t.

An easy way of computing the LM statistic based on the residuals of the VAR(p) model is to consider the auxiliary model

$$\hat{u}_t = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p} + D_1 \hat{u}_{t-1} + \cdots + D_h \hat{u}_{t-h} + e_t^*, \tag{2.6.2}$$

where the $\hat{u}_t$ are the residuals from the original model, and the $\hat{u}_t$ with $t \leq 0$ are replaced by zero. The term e_t^* is an auxiliary error term. The LM statistic may be computed as

$$Q_{LM} = T\left(K - \text{tr}(\widetilde{\Sigma}_u^{-1}\widetilde{\Sigma}_e)\right),$$

where $\widetilde{\Sigma}_u = T^{-1}\sum_{t=1}^{T}\hat{u}_t\hat{u}_t'$, $\widetilde{\Sigma}_e = T^{-1}\sum_{t=1}^{T}\hat{e}_t^*\hat{e}_t^{*\prime}$ and $\hat{e}_t^*$ $(t = 1, \ldots, T)$ are the residuals from the estimated auxiliary model. Under the null hypothesis of no autocorrelation, the LM statistic has an asymptotic $\chi^2(hK^2)$ distribution.

In a Monte Carlo simulation study, Edgerton and Shukur (1999) find that this statistic may also have a small-sample distribution that differs substantially from its asymptotic χ^2 distribution and propose an F version with better small-sample properties. It is noteworthy that the asymptotic distribution remains valid under the null hypothesis even if there are integrated variables, as shown by Brüggemann, Lütkepohl, and Saikkonen (2006). As in the case of top-down sequential testing, the usual caveats about sequential testing apply.

2.6.3 Information Criteria

An alternative to sequential testing procedures is the use of information criteria for lag-order selection. Information criteria used for VAR lag-order selection have the general form

$$C(m) = \log(\det(\widetilde{\Sigma}_u(m))) + c_T\varphi(m), \tag{2.6.3}$$

where $\widetilde{\Sigma}_u(m) = T^{-1}\sum_{t=1}^{T}\hat{u}_t\hat{u}_t'$ is the residual covariance matrix estimator for a reduced-form VAR model of order m based on LS residuals $\hat{u}_t$, m is the candidate lag order at which the criterion function is evaluated, $\varphi(m)$ is a function of the order m that penalizes large lag orders, and c_T is a sequence of weights that may depend on the sample size.

The function $\varphi(m)$ corresponds to the total number of regressors in the system of VAR equations. Since there are mK lagged regressors in each equation and K equations in the VAR model, in the absence of any deterministic regressors, $\varphi(m) = mK^2$. More typically, when including an intercept, $\varphi(m) = mK^2 + K$.

Information criteria are based on the premise that there is a trade-off between the improved fit of the VAR model, as m increases, and the parsimony of the model. Given T, the fit of the model by construction improves with larger m, indicated by a reduction in $\log(\det(\widetilde{\Sigma}_u(m)))$. At the same time the penalty term, $c_T\varphi(m)$, unambiguously increases with m. We are searching for the value of m that balances the objectives of model fit and parsimony. The choice of the penalty term determines the trade-off between these conflicting objectives.

The three most commonly used information criteria for VAR models are known as the AIC, HQC, and SIC:

Akaike Information Criterion (AIC)

$$\mathrm{AIC}(m) = \log(\det(\widetilde{\Sigma}_u(m))) + \frac{2}{T}(mK^2 + K),$$

where $c_T = 2/T$. This criterion was proposed by Akaike (1973, 1974).

Hannan-Quinn Criterion (HQC)

$$\mathrm{HQC}(m) = \log(\det(\widetilde{\Sigma}_u(m))) + \frac{2\log(\log(T))}{T}(mK^2 + K),$$

where $c_T = 2\log(\log(T))/T$. This criterion is suggested by the work of Hannan and Quinn (1979) and Quinn (1980).

Schwarz Information Criterion (SIC)

$$\mathrm{SIC}(m) = \log(\det(\widetilde{\Sigma}_u(m))) + \frac{\log(T)}{T}(mK^2 + K),$$

where $c_T = \log(T)/T$. This criterion is due to Schwarz (1978) and Rissanen (1978). Because Schwarz (1978) derives this criterion in a Bayesian setting, it is also known as the Bayesian Information Criterion (BIC). It should be noted, however, that most Bayesian users of VAR models would object to the use of information criteria for lag-order selection (see Chapter 5).

The VAR order is chosen such that the respective criterion is minimized over the possible orders $m = 0, \ldots, p_{\max}$. Alternative information criteria have been proposed by Hurvich and Tsai (1993), Phillips and Ploberger (1996), Sin and White (1996), and Inoue and Kilian (2006), among others.

One concern with the use of sequential testing procedures is that the sequence of the tests matters. Moreover, as noted earlier, it can be challenging to control the size of sequential testing procedures. It can be shown that the use of information criteria is equivalent to simultaneously testing each candidate model against each other model, with the critical value being determined implicitly by the penalty function. Unlike in sequential testing, no one model is favored because it is chosen as the null hypothesis, and the order in which the criterion function is evaluated does not affect the lag-order choice. This advantage comes at the cost of the user no longer being able to control the effective size of the procedure in finite samples. By choosing an appropriate penalty term in the information criterion, one can, however, make sure that the size approaches zero asymptotically.

A key issue in implementing information criteria is the choice of the upper and lower bounds $p_{\max}$ and $p_{\min}$. In the context of a model of unknown finite lag order, the default is to set $p_{\min} = 0$ or sometimes $p_{\min} = 1$, reducing the problem to one of choosing a suitable upper bound. The value of $p_{\max}$ must be chosen long enough to allow for delays in the response of the model variables

to the shocks. In practice, common choices would be 12 or 24 lags for monthly data and 4 or 8 lags for quarterly data.

Occasionally users of VAR models implement information criteria for lag-order selection incorrectly. One common mistake relates to the choice of the evaluation period. When evaluating the fit of the model for a given lag order m, it is essential that we compute $\widetilde{\Sigma}_u(m)$ on exactly the same evaluation period, $t = p_{\max} + 1, \ldots, T$, for all m. If the evaluation period used in computing $\widetilde{\Sigma}_u(m)$ differs across $m \in \{1, \ldots, p_{\max}\}$, we risk that differences in the fit of different models are driven by the inclusion of additional observations rather than the inclusion of additional lags. In other words, we end up comparing apples and oranges, rendering the resulting ranking meaningless. For example, many canned software packages produce AIC values as a by-product of the regression output. It may seem that these values could be used to rank alternative VAR models according to their AIC values. This is not the case. The reason is that the relevant evaluation period for computing $\widetilde{\Sigma}_u(m)$ depends on $p_{\max}$. Without the user specifying $p_{\max}$, canned software packages cannot possibly compute the correct estimate $\widetilde{\Sigma}_u(m)$ or the corresponding AIC values. Instead, they display AIC values based on the evaluation period $t = m + 1, \ldots, T$, where $p_{\min} \leq m \leq p_{\max}$. It may seem that this mistake could not have important consequences, but Ng and Perron (2005) demonstrate by simulation that this mistake can result in severe distortions in the estimated order $\widehat{p}$.

Another potential mistake relates to the estimation of $\Sigma_u(m)$. It is essential that the criterion function be evaluated at the ML estimator $\widetilde{\Sigma}_u(m) = T^{-1} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$ rather than the LS estimator $\widehat{\Sigma}_u(m) = (T - Km - \kappa)^{-1} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$, where κ is the number of deterministic regressors in each equation. The use of information criteria is predicated on the trade-off between improved fit (measured by decline in $\det(\widetilde{\Sigma}_u(m))$ and the reduction in parsimony (measured by an increase in the penalty term), as more lags are added. There is no such trade-off when we estimate $\Sigma_u(m)$ by

$$\widehat{\Sigma}_u(m) = \frac{T}{T - Km - \kappa} \widetilde{\Sigma}_u(m).$$

Adding more lags lowers $\det(\widetilde{\Sigma}_u(m))$ by construction, but it also increases $T/(T - Km - \kappa)$, invalidating the model rankings.

Table 2.1 illustrates the implementation of the SIC, HQC, and AIC criteria in the context of the three-variable VAR model example already used earlier in this chapter to illustrate alternative estimation procedures. Let $m \in \{0, 1, \ldots, 4\}$. The lag-order values that minimize a given criterion function are shown in bold. Table 2.1 shows that the SIC favors a lag order of $p = 1$, whereas the HQC chooses a lag order of $p = 3$, and the AIC chooses $p = 4$.

Table 2.1. *Alternative Lag-Order Selection Criteria for VAR Models*

m	SIC(m)	HQC(m)	AIC(m)
0	0.4594	0.4308	0.4114
1	**−2.9750**	−3.0893	−3.1669
2	−2.9394	−3.1394	−3.2752
3	−2.9570	**−3.2428**	−3.4367
4	−2.8697	−3.2412	**−3.4933**

Note: Based on a VAR(m) model with intercept for $y_t = (\Delta gnp_t, i_t, \Delta p_t)'$, as defined in the earlier empirical example.

2.6.4 Recursive Mean-Squared Prediction Error Rankings

Yet another approach to selecting the lag order of a VAR model is to rank the candidate models based on their recursive mean-squared prediction error (MSPE) in simulated out-of-sample forecasts. The recursive MSPE is obtained as follows. Suppose we have a sample consisting of T observations and candidate VAR models of order $m \in \{p_{\min}, \ldots, p_{\max}\}$. Further suppose that we are interested in forecasting one of the VAR model variables. We initially estimate the models under consideration on the first R observations of the sample, $R < T$, and compare their predictions for the variable of interest in period $R + h$ to the realized value. We then expand the initial sample by estimating the models under consideration on the first $R + 1$ observations and construct the prediction error for $R + 1 + h$. We continue this procedure until the estimation window extends to $T - h$. This procedure results in a sequence of $T - R - h$ recursive prediction errors. The recursive MSPE of a given model is the average of its squared recursive prediction errors. We choose the lag order corresponding to the model with the lowest recursive MSPE. This procedure may be adapted to allow for forecasts of more than one VAR variable by evaluating the trace of the MSPE matrix.

This so-called predictive least-squares approach was originally introduced by Rissanen (1986) in the context of one-step ahead prediction. A complete asymptotic analysis is provided in Wei (1992). There is a close relationship between the predictive least-squares approach to lag-order selection and the use of information criteria. Rissanen studied the asymptotic properties of the simulated out-of-sample method under the assumption that the length of the initial recursive sample, R, is fixed with respect to T, where $R < T$ is the length of the initial window of data used for recursive estimation. Under these assumptions, Rissanen's predictive least-squares method can be shown to be asymptotically equivalent to the SIC, provided the true VAR model is included among the VAR models compared. In contrast, Inoue and Kilian (2006)

establish that under the alternative (and more conventional) assumption that R is a fixed fraction of T, the predictive least-squares method is asymptotically equivalent to the AIC.

Which of these asymptotic thought experiments provides a better small-sample approximation is not clear a priori, nor is it clear whether any of these asymptotic results are useful in practice. Simulation evidence for univariate autoregressions in Inoue and Kilian (2006) suggests that the AIC method tends to generate lower one-step ahead MSPEs than predictive least squares when both methods are feasible, and that both methods are less accurate than the SIC. There are no comparable simulation results for multivariate autoregressions nor are there simulation results for multi-step ahead predictions.

Given that information criteria are constructed on the basis of one-step ahead MSPEs rather than multi-step ahead MSPEs, one would expect the predictive least-squares method applied to the h-step ahead MSPE to be particularly useful for selecting the lag order of VAR models used to construct iterated forecasts. This line of inquiry recently has been explored by Ing (2003, 2004) for stationary autoregressive processes. Ing, Lin, and Yu (2009) provide an extension to integrated autoregressive processes. An empirical example of the use of predictive least squares for selecting the lag order of iterated VAR forecasting models is Baumeister and Kilian (2012).

2.6.5 The Relative Merits of Alternative Lag-Order Selection Tools

There are different ways of evaluating the relative merits of alternative lag-order selection criteria.

Consistency for the True VAR Lag Order. If the DGP is a VAR(p_0) model, one criterion is whether the lag-order estimator is consistent for the true lag order p_0, provided $p_{\min} \leq p_0 \leq p_{\max}$. Clearly, sequential tests will not be consistent for p_0 because of the positive probability of committing a type I error in statistical testing. In contrast, information criteria will be consistent for p_0 under suitable conditions. The reason is that the probability of committing a type I error vanishes asymptotically if we are careful about the choice of penalty function. As is easy to verify, the AIC is not consistent for p_0, whereas the SIC and HQC are (see Lütkepohl 2005, section 4.3.2). The HQC was designed to be the least parsimonious yet still consistent information criterion for VAR models. Although standard proofs of the consistency of the SIC and HQC and of the inconsistency of the AIC require the VAR model to be stationary, these results can be extended to VAR models with unit roots (see Paulsen 1984; Tsay 1984).

The price of obtaining a consistent lag-order estimator is greater parsimony in model selection. Parsimony here means that the criterion will favor VAR models with fewer lags. The larger the penalty term for given T, the more

parsimonious the lag-order estimate. It can be shown that for $T \geq 16$,

$$\widehat{p}^{\text{SIC}} \leq \widehat{p}^{\text{HQC}} \leq \widehat{p}^{\text{AIC}},$$

indicating that the SIC tends to be more parsimonious than the HQC, which in turn tends to be more parsimonious than the AIC. Of course, in a given application, all three criteria may suggest the same lag order.

Finite-Sample Properties of the Lag-Order Estimator. Even if we grant the premise that $p_{\min} \leq p_0 \leq p_{\max}$, one may not want to overrate the importance of the lag-order estimator being consistent. The convergence of $\widehat{p}$ toward p_0 can be very slow in practice, and in small samples consistent lag-order selection criteria tend to be strongly downbiased toward $p_{\min}$. For example, Kilian (2001) examines a stylized bivariate AR(4) data generating process and shows that for a sample size of $T = 80$, the SIC selects a lag order of 1 among $1 \leq p \leq 8$ with probability 92%, but the true lag order of 4 only with probability 2%. Even for $T = 160$, the probability of selecting $p = 1$ remains at 61% with a probability of only 28% of selecting the true lag order. Similar, if less extreme results hold for the HQC. In contrast, the AIC has a probability of selecting the true lag order of 57% for $T = 80$ and of 83% for $T = 160$. The probability of the AIC underestimating the lag order is 26% for $T = 80$ and 1% for $T = 160$, compared with 98% and 73% for the SIC. The finding that in small samples the distribution of the AIC lag-order estimates tends to be more balanced about the true lag order than for the SIC lag-order estimates is also consistent with simulation results in Nickelsburg (1985) and Lütkepohl (1985).

The high accuracy of the AIC in these simulation studies – even in large samples – may be surprising at first, but it reflects the asymptotic properties of the AIC. Although

$$\lim_{T\to\infty} \mathbb{P}\left(\widehat{p}^{\text{AIC}} > p_0\right) > 0,$$

reflecting the inconsistency of the AIC, it can be shown that

$$\lim_{T\to\infty} \mathbb{P}\left(\widehat{p}^{\text{AIC}} < p_0\right) = 0.$$

In other words, in the limit, the AIC will never select a lag order that is lower than p_0, but it will have a tendency to select with positive probability a lag order in excess of p_0. This point is important. Economists using VAR models have no inherent interest in the lag order of the process. They are interested in impulse responses, forecasts, and related statistics that can be written as smooth functions of VAR model parameters. These statistics of interest can be consistently estimated, as long as the lag order is not underestimated asymptotically, so there is little to choose between the SIC, HQC, and AIC on the grounds of consistency. The only potential concern is that in large samples the

AIC may choose an excessively large lag order and inflate the variance of, for example, the impulse response estimator.

This concern as well is largely unfounded. Paulsen and Tjøstheim (1985) establish that the probability of the AIC overfitting the VAR model is negligible asymptotically. Whereas for $K = 1$, the asymptotic probability of overfitting is about 30%, for $K = 2$ it drops to at most 12%, for $K = 3$ to 4%, and for $K = 4$ to 1%, so for most VAR applications, the efficiency loss is not a major concern. These asymptotic results are consistent with simulation evidence for large-dimensional VAR models in Gonzalo and Pitarakis (2002) who conclude that the AIC tends to be by far the most reliable estimator of p_0 compared with sequential tests as well as other information criteria.

Finite-sample considerations also favor the AIC. In related work, Kilian (2001) observes that in the context of impulse response analysis in finite samples overestimation of the lag order is costly only to the extent that the impulse response estimates are less precise, but underestimation tends to greatly distort the impulse response functions especially at longer horizons. Hence, users of VAR models ought to employ an asymmetric loss function, erring on the side of including too many lags. Kilian (2001) provides simulation evidence that VAR models based on the AIC lag-order estimate provide more accurate impulse response confidence intervals than VAR models based on more parsimonious lag-order selection criteria. He also shows that AIC-based impulse response estimates have lower MSE. This conclusion is further supported by simulation evidence in Kilian and Ivanov (2005) based on a wide range of monthly finite-order VAR models of the type used in empirical work. Kilian and Ivanov show that for monthly VAR models the AIC implies impulse response estimates with systematically lower MSE than more parsimonious criteria such as the SIC or HQC. For quarterly VAR models, which tend to imply smoother impulse responses, using the HQC generates the impulse response estimates with the lowest MSE.

There also is simulation evidence on the performance of sequential tests for lag-order selection. At least for univariate autoregressions, several studies have shown that the general-to-specific sequential-testing approach is preferred to the specific-to-general sequential testing approach. For VAR models neither approach is satisfactory for selecting the true lag order (see, e.g., Lütkepohl 1985; Nickelsburg 1985; Gonzalo and Pitarakis 2002). Moreover, the simulation evidence in Kilian and Ivanov (2005) shows that sequential testing procedures tend to produce impulse response estimators with systematically larger MSE than information criteria and cannot be recommended for applied work.

Lag-Order Selection in VAR(∞) Models. The consistency for p_0 becomes entirely irrelevant once we allow for the possibility that the DGP is a VAR(∞) process. In that case, one can reinterpret the VAR(p) model as an approximation to a VAR(∞) data generating process, as discussed in Section 2.3.6.

In such a setting, consistency for p_0 is not a meaningful concept. Instead, the objective of lag-order selection is to select the best approximating model. As a practical matter, the use of information criteria in this setting is subject to the same concerns regarding the finite-sample performance of lag-order selection criteria as in the case of finite lag-order VAR models. While the AIC may have a better asymptotic rationale than more parsimonious criteria in the VAR(∞) context, even the AIC underfits in small samples.

There are few optimality results regarding the lag-order choice for sieve autoregressions. It can be shown that the AIC is asymptotically efficient for the one-step ahead MSPE, provided the lower bound and upper bound on the lag order are chosen appropriately. This asymptotic efficiency result was originally obtained under the assumption that the processes used for estimation and for prediction are independent. This assumption is inappropriate for time series analysis, because the value to be predicted depends on past realizations. More recently, Ing and Wei (2003, 2005) established the optimality of the AIC for same-realization prediction from stationary AR(∞) processes, and Ing, Sin, and Yu (2010) extended this result to integrated AR(∞) processes. Of course, even this result applies only to one-step ahead prediction and is only asymptotic. It does not provide much guidance for applied econometric work based on sieve vector autoregressions.[6]

Moreover, much depends on the choice of the lower bound and upper bound, which also depend on T. Asymptotic theory implies that p has to grow at a suitable rate with the sample size T to preserve the consistency and asymptotic normality of the LS estimator of the approximating VAR(p) model. This insight does not tell us how to choose p for given T, however. In practice, this question has been addressed by simulating data from VARMA models fitted to actual data and searching for the lag order that provides the best sieve approximation for the statistic of interest (see, e.g., Berkowitz and Kilian 2000; Inoue and Kilian 2002b). These simulation studies suggest that the best approximating lag order tends to be larger than the lag-order choices commonly used in applied work and larger than the lag orders suggested by the AIC.

Implications of Data-Dependent Lag-Order Selection for Inference. An important question usually neglected in discussions of lag-order selection is that statistical inference about the VAR model parameters and inference about smooth functions of these parameters such as impulse responses is affected by the use of data-dependent VAR lag-order selection procedures. It is useful to start with the premise of a VAR(p_0) data generating process. In that case, one immediate problem is that the use of inconsistent lag-order selection procedures such as the AIC or sequential testing procedures affects the asymptotic

[6] For a comparison of the asymptotic properties of the AIC, SIC, and HQC and predictive least-squares approaches when the DGP is of infinite order, see Ing (2007).

distribution of the slope parameters conditional on the estimated lag orders. In particular, the post-model selection estimators are no longer asymptotically normal, invalidating the use of the delta method (see, e.g., Leeb and Pötscher 2005). This point is often ignored in applied work.

It may seem that this problem could be avoided by the use of consistent lag-order selection criteria such as the SIC or HQC. The traditional view in the literature for many years has been that consistent lag-order selection procedures allow one to employ the same asymptotic distributions that would be appropriate if the correct VAR(p_0) model were specified instead of VAR($\widehat{p}$) (see, e.g., Lütkepohl 1990; Kilian 1998c). This view has recently been shown to be mistaken by Leeb and Pötscher (2005). The reason is that the usual starting point in the literature has been an analysis of the pointwise asymptotic distribution of the slope parameters, where pointwise refers to the thought experiment of holding the true parameter fixed while letting the sample size diverge to infinity. Indeed the pointwise asymptotic distribution coincides with the usual asymptotic distributions derived for correctly specified models. As discussed in Leeb and Pötscher (2005), however, the finite-sample distribution of the slope parameters will not always be well approximated by the pointwise asymptotic normal distribution. Leeb and Pötscher provide examples in which this finite-sample distribution is bimodal rather than Gaussian. The reason is that the convergence of the finite-sample distribution to their pointwise limits is typically not uniform with respect to the underlying parameter values. Leeb and Pötscher demonstrate for a generic linear regression model that it may require arbitrarily large T for the standard pointwise asymptotics to become a good approximation.

This problem arises when the parameter in question is close to zero, but different from zero, and cannot simply be overcome by the use of bootstrap methods either (see Chapter 12). Although bootstrap methods of inference for VAR($\widehat{p}^{\text{AIC}}$) models appear reasonably accurate in large samples in the simulation studies of Kilian (1998a, 2001), there is no proof of the uniform validity of the bootstrap in this case. In fact, Leeb and Pötscher (2006) conjecture that no bootstrap method can solve the problem of nonuniformity. Furthermore, these problems do not vanish if we drop the fiction of a VAR(p_0) data generating process. Similar finite-sample problems arise even when the underlying data generating process is of infinite order. Leeb and Pötscher (2005) show that in the latter case even the pointwise asymptotic distributions are affected by the model selection procedure and no longer coincide with the usual pointwise asymptotic distributions that apply in the absence of model selection.

It is important to be clear that Leeb and Pötscher (2005, 2006) are concerned with the worst possible outcome when conducting inference, not the likely outcome. They establish that conventional inference may be unreliable in some cases. They do not establish that it will always be unreliable. In fact,

in many cases, standard inference will be perfectly adequate. It is difficult to determine how much we should be concerned with the worst possible outcome in practice. Without knowing the DGP it is impossible to know how practically relevant the situations described by Leeb and Pötscher are. Existing simulation results for DGPs based on actual data have not revealed any problems with the reliability of standard inference based on pointwise asymptotics (see, e.g., Kilian 1998a, 2001). Nevertheless, in light of the results in Leeb and Pötscher (2005, 2006), the use of all data-dependent lag-order selection procedures for VAR models must be reconsidered. An alternative approach is to estimate VAR models with fixed lag orders instead. For example, fitting a VAR($p_{\max}$) model, where $p_{\max}$ is a conservative bound on the range of possible lag orders, allows the use of standard asymptotic approximations for statistics such as impulse responses and circumvents the problems of inference described by Leeb and Pötscher.

Impulse Response Analysis. It is known (a) that the use of the AIC in some cases can be formally justified even when the underlying VAR is of infinite order; (b) that in a number of simulation studies the AIC performed better at selecting the true VAR order than alternative lag-order selection criteria such as sequential tests, the SIC, or the HQC; and (c) that AIC-based impulse response point estimates tend to have lower MSE than impulse response estimates based on other criteria, when working with monthly data (see Kilian and Ivanov 2005). At the same time, the use of the AIC tends to result in more accurate impulse response confidence intervals in monthly models, as demonstrated in Kilian (1998a, 2001). All this evidence does not necessarily mean that users of VAR models in practice should condition on the AIC lag-order estimate, however, because even the AIC has a tendency to underfit in small samples (see, e.g., Kilian 1998a, 2001; Gonzalo and Pitarakis 2002). Hence, in many applied situations, when estimating VAR impulse responses, a reasonable alternative approach is to impose a fixed lag order ex ante rather than to rely on lag-order selection criteria. In light of the asymmetric loss associated with overfitting and underfitting the model to be used for impulse response analysis, this lag order ought to be larger rather than smaller in case of doubt. The use of a reasonably large fixed lag order also avoids the problems of inference discussed in Leeb and Pötscher (2005, 2006).

It is useful to elaborate on how this fixed lag order may be chosen in practice. The reasoning underlying this choice is similar to the reasoning underlying the choice of $p_{\max}$ in implementing lag-order selection criteria. The appropriate number of autoregressive lags has nothing to do with the persistence of the data. Indeed, even a VAR(1) model is fully capable of capturing any degree of persistence in the data. Rather it has to do with the delays in the responses to shocks. One situation in which it is safer to allow for long lags

is when we model slowly building cycles in the data. A case in point are studies of the evolution of commodity prices. Inspection of commodity price data reveals cycles that may last for more than ten years. These cycles build and wane only gradually. A VAR model that truncates the lag order too early will miss these cycles, which are associated with gradual changes in the demand for commodities. This argument prompted Kilian (2009), for example, to allow for 24 monthly lags in a VAR model of the global oil market. It can be shown that the importance of global demand shocks in this model vanishes if the lag order is restricted to, say, 12 lags. The reason is that much of the response to the underlying demand shocks is small initially and accumulates only with a delay. Allowing for enough lags is crucial for detecting the role of demand shocks in this model.

This does not mean that such a large number of lags is always required. The appropriate lag-order choice depends on the economic context and on the number of variables included. The responses of typical U.S. macroeconomic aggregates, for example, can be reliably estimated with a much smaller number of lags. Many empirical VAR studies impose 12 monthly lags or 4 quarterly lags. These choices reflect the premise that much of the response occurs within the first year of a shock. Although lag-order selection criteria could be used to make the model more parsimonious, such criteria have to be viewed with caution, given the proclivity of even the least parsimonious lag order selection criteria to underfit in small samples. Edelstein and Kilian (2009), for example, in a study of the transmission of monthly energy price shocks to the U.S. economy since the 1970s, found that the AIC in some cases produces implausibly low lag-order estimates associated with counterintuitive estimates of the response functions. These problems vanished once a larger lag order of 6 was imposed throughout, and the VAR estimates remained qualitatively robust to using even higher lag orders between 6 and 12.

In practice, it is useful to assess the sensitivity of estimates to reasonable changes in the lag structure. The insistence on using a large number of lags does not mean that information criteria cannot provide useful information. For example, we may interpret the lag-order estimate implied by the AIC as a lower bound on the lag order to be imposed in estimation. The dangers of relying on lag-order selection criteria that are too parsimonious are also illustrated by Hamilton and Herrera (2004) who show that raising the lag order from 7, which is the estimate suggested by the AIC, to 12 or to 16 overturns the substantive findings in Bernanke, Gertler, and Watson (1997) regarding the effect of oil price shocks on U.S. real GDP growth. Specifically, Hamilton and Herrera find large and statistically significant direct effects of oil price shocks for $p = 12$ and $p = 16$ that were not apparent with $p = 7$. This example illustrates the importance of allowing for enough lags. Any empirical result that is reversed with higher lag orders must be considered suspect. On the other hand, results that remain intact with higher lag orders are likely to be trustworthy.

One objection to the strategy of erring on the side of including too many lags may be that allowing for so many lags may cause the impulse response estimates to become statistically insignificant. This is indeed a risk, although the inclusion of more lags need not increase the uncertainty about the impulse response estimates. In practice, the increase in uncertainty in the response estimates in many cases is smaller than one might have conjectured.

The risk of overfitting, moreover, does not mean that we are justified in imposing a lower lag order. Kilian (2001) illustrates by simulation that in this case we are indeed likely to obtain precisely estimated response functions with tight confidence intervals, except that these estimates bear little resemblance to the responses in the DGP. In other words, it is the nature of the economic problem rather than the data constraints faced by the researcher that dictates the appropriate lag order. In some cases, this may mean that a VAR model should not be estimated because the sample is too short to accommodate the required number of lags.

VAR Forecasts. One key difference between impulse response analysis and out-of-sample forecasting is that in the latter case we are not concerned with selecting the DGP, but with selecting the most accurate out-of-sample forecasting model. The usual criterion for measuring out-of-sample forecast accuracy is the MSPE. The well-known bias variance trade-off means that even if we knew the true lag order p_0, a VAR(p) model with $p < p_0$ may have lower out-of-sample MSPE in small samples. Thus consistency for p_0 is not a valid criterion for choosing between information criteria. Inoue and Kilian (2006) provide extensive discussion of the conditions under which information criteria will be consistent for the forecasting model with the lowest one-step ahead MSPE, even when one or more candidate models are misspecified. They show that the SIC is superior to the AIC in the context of choosing between a VAR(m) model and a VAR($m + 1$) model for one-step ahead prediction. The asymptotic validity of the SIC for selecting the lag order of the VAR model does not depend on the existence of p_0 or the inclusion of p_0 in the set $\{p_{\min}, \ldots, p_{\max}\}$. It would apply even if the underlying data generating process were a VAR(∞) model, for example, provided the set of candidate models does not depend on T. It is useful to keep in mind, however, that the results in Inoue and Kilian (2006) are asymptotic in nature. There is no guarantee that the SIC will select the most accurate one-step ahead VAR forecasting model in finite samples.

One of the limitations of standard lag-order selection criteria including the SIC is that they focus on the one-step ahead MSPE in assessing the fit of the model. The latter approach is appropriate if the model is to be used for one-step ahead forecasts, but in practice VAR models often are used for forecasting several steps ahead. For example, we may be concerned with forecasting the inflation rate not for the next month, but for six months from now. Even more

commonly, we will be interested in forecasting variables such as the cumulative inflation rate between today and some future date, based on a monthly VAR model. One option in that case is to employ a direct forecast of $p_{t+h} - p_t$, where h denotes the horizon over which inflation is to be forecast and p_t is the log of the price level. Unless $h = 1$, the construction of this direct forecast requires the forecaster to abandon the VAR framework we have considered so far. The other option is to recursively iterate forward the fitted VAR model, to obtain forecasts of the monthly inflation rate at horizons 1 through h, and to cumulate these forecasts to construct the implied forecast of the inflation rate between now and h periods from now.

It is immediately obvious that an iterated forecast based on a VAR model using lag orders selected by criteria designed to select the true lag order (or designed to select the VAR model with the smallest one-step ahead MSPE) will not necessarily produce optimal MSPEs at longer horizons, unless the VAR model is correctly specified (see Schorfheide 2005). Marcellino, Stock, and Watson (2006) and Pesaran, Pick, and Timmermann (2011) nevertheless show that, in practice, iterated forecasts tend to be more accurate in out-of-sample forecasting than direct forecasts, but only provided the lag order chosen is sufficiently large. Iterated forecasts based on VAR models with lag orders chosen by the SIC, for example, tend to be less accurate at longer horizons than VAR models with lag orders chosen by the AIC. This does not mean that the AIC lag order is optimal, of course.

Ing (2004) considers the problem of choosing the multi-step ahead forecast with the smallest possible MSPE for an autoregressive process of finite but unknown order. He provides several examples that illustrate that optimal multi-step forecasts are not guaranteed even when one is able to correctly identify the underlying autoregressive lag order. In response, Ing proposes an alternative lag-order selection criterion for multi-step ahead prediction that is valid for stationary autoregressions of finite lag order. A generalization of this idea to stationary VAR models was developed in Schorfheide (2005). Further generalizations to autoregressions of possibly infinite order are provided in Ing (2007), and extensions to possibly integrated autoregressions of possibly infinite order are provided in Ing, Lin, and Yu (2009) and Ing, Sin, and Yu (2010).

2.7 Model Diagnostics

There is a large set of tools for checking whether a given VAR model represents the DGP of the variables adequately. They range from informal graphical procedures such as residual plots to formal statistical specification tests of the adequacy of the assumptions underlying the model. In the following subsections selected formal specification tests are discussed. For the underlying theory and for related procedures see, e.g., Lütkepohl (2004, 2005). The presentation in this section partly follows Lütkepohl (2009).

2.7.1 Tests for Autocorrelation in the Innovations

A basic assumption for the VAR model is that the reduced-form innovation process u_t is white noise. In other words, the u_t are assumed to exhibit no serial correlation. As discussed in the previous section, the lag order of the model is typically chosen such that this condition is at least approximately satisfied. For a given VAR model it may still be desirable to check for residual autocorrelation. In that case, the Portmanteau and Breusch-Godfrey LM tests presented in Subsection 2.6.2 can be used.

2.7.2 Tests for Nonnormality

Although normality of the innovations of a VAR model and, hence, normality of the observed variables is not required for the validity of most asymptotic procedures related to VAR modeling, normality can still be a property of interest. For example, it facilitates predictive inference, as discussed earlier. Moreover, knowing that the distribution of the observations is far from normal is useful for assessing possible efficiency gains from using other estimation procedures. For example, there may be asymptotic efficiency gains if the true distribution were used instead of the Gaussian distribution in setting up the likelihood function.

In practice, such efficiency gains are not likely to be substantial in the small samples common in macroeconomic applications. If the residuals do not have a normal distribution, however, this may be a signal of other potential model defects. A rejection of normality, for example, may arise from unusually large residuals of three or even four times the size of the standard deviation, sometimes called outliers. Such outliers could be an indication that the VAR model is misspecified. Thus, a rejection of normality may suggest a closer look at the time series of residuals.

In this context, multivariate nonnormality tests may be applied to the full residual vector of the VAR model and univariate versions can be used for the errors of the individual equations. Lomnicki (1961) and Jarque and Bera (1987) have proposed nonnormality tests for univariate models that can be extended easily to multivariate models. The idea is to check whether the third and fourth moments of the residuals are conformable with those of a normal distribution. To use this approach, the residual vector of a VAR model is first transformed to make the individual components uncorrelated. Then the moments are compared with those of the normal distribution. For given residuals $\hat{u}_t$, $t = 1, \ldots, T$, of an estimated VAR process, the residual covariance matrix $\widetilde{\Sigma}_u$ is decomposed such that $\widetilde{\Sigma}_u = PP'$, where P is a suitable $K \times K$ matrix. The tests for nonnormality are then based on the skewness and kurtosis of the standardized residuals $\hat{u}_t^s = P^{-1}\hat{u}_t$ (see Lütkepohl 2005).

There are many possible P matrices that decompose $\widetilde{\Sigma}_u$, and the tests depend to some extent on the transformation matrix used. Doornik and Hansen

(1994) propose to use the square root matrix of $\widetilde{\Sigma}_u$ whereas Lütkepohl (2005, chapter 4) considers a Cholesky decomposition of the residual covariance matrix. Following the latter convention, let

$$\hat{b}_j = (\hat{b}_{1j}, \ldots, \hat{b}_{Kj})', \quad \text{where} \quad \hat{b}_{kj} = \frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{kt}^s)^j$$

for $j = 3, 4$. It can be shown that $\lambda_3 = \hat{b}_3'\hat{b}_3/6 \xrightarrow{d} \chi^2(K)$ can be used to test the symmetry of the error distribution and $\lambda_4 = (\hat{b}_4 - 3_K)'(\hat{b}_4 - 3_K)/24 \xrightarrow{d} \chi^2(K)$, where $3_K = (3, \ldots, 3)'$ is a $K \times 1$ vector, can be used to test for excess kurtosis relative the kurtosis of the Gaussian distribution. Moreover, the sum of the two test statistics is asymptotically χ^2 distributed with $2K$ degrees of freedom and can be viewed as an omnibus test of the null of Gaussianity. In small samples this test may be severely oversized. Small-sample corrections of the critical values based on bootstrap approximations are proposed in Kilian and Demiroglu (2000).

The literature on testing for nonnormality is extensive and many other tests are available. In the context of VAR modeling the tests based on the third and fourth moments are nevertheless the most popular tests. Nonnormality in the unconditional error distribution also arises if the innovations are conditionally heteroskedastic. Arguably more powerful tests for this feature are discussed in the next subsection.

2.7.3 Residual ARCH Tests

Unmodeled conditional heteroskedasticity in the VAR innovations does not invalidate the consistency of standard estimators of the VAR slope parameters, as long as the unconditional error variances remain finite. It undermines the efficiency of the estimator, however, and affects how we conduct inference about the parameters of interest (see Chapter 12). The consistent estimation of the unconditional error covariance matrix requires additional moment conditions, given that the assumption of iid errors is violated in the presence of autoregressive conditionally heteroskedastic (ARCH) errors, for example. Moreover, ARCH dynamics in the innovations may invalidate the assumption of a finite fourth moment required for asymptotic and bootstrap inference on structural impulse responses and related statistics, depending on the parameters of the model of conditional heteroskedasticity (see He and Teräsvirta 1999). Hence, knowing whether the innovations are conditionally heteroskedastic or not is important.

A multivariate ARCH model of order q for the VAR innovations u_t has the form

$$\text{vech}(\Sigma_{t|t-1}) = \delta_0 + D_1\text{vech}(u_{t-1}u_{t-1}') + \cdots + D_q\text{vech}(u_{t-q}u_{t-q}'),$$

where vech is the column-stacking operator for symmetric matrices that stacks the columns from the main diagonal downward, and $\Sigma_{t|t-1}$ is the conditional covariance matrix of u_t given $u_{t-1}, u_{t-2}, \ldots$. Moreover, δ_0 is a $\frac{1}{2}K(K+1)$-dimensional parameter vector and the D_j, $j = 1, \ldots, q$, are $\frac{1}{2}K(K+1) \times \frac{1}{2}K(K+1)$ coefficient matrices.

A test for ARCH or GARCH dynamics can be based on similar ideas as the LM test for residual autocorrelation by considering the pair of hypotheses

$$\mathbb{H}_0 : D_1 = \cdots = D_q = 0 \quad \text{versus}$$
$$\mathbb{H}_1 : D_i \neq 0 \text{ for at least one } i \in \{1, \ldots, q\}.$$

Clearly, there is no ARCH in the innovations if $\mathbb{H}_0$ is true. The relevant LM statistic can be obtained by estimating the auxiliary model

$$\text{vech}(\hat{u}_t \hat{u}_t') = \delta_0 + D_1 \text{vech}(\hat{u}_{t-1} \hat{u}_{t-1}') + \cdots + D_q \text{vech}(\hat{u}_{t-q} \hat{u}_{t-q}') + e_t \tag{2.7.1}$$

and computing the statistic

$$LM_{ARCH}(q) = \frac{1}{2} TK(K+1) \left(1 - \frac{2}{K(K+1)} \text{tr}(\widehat{\Omega} \widehat{\Omega}_0^{-1}) \right),$$

where $\widehat{\Omega}$ is the residual covariance matrix of the $\frac{1}{2}K(K+1)$-dimensional error term e_t of the regression model (2.7.1) with $q > 0$ and $\widehat{\Omega}_0$ is the corresponding matrix for the case $q = 0$. The statistic is similar to the one described by Doornik and Hendry (1997, section 10.9.2.4). It may be used with critical values from a $\chi^2(qK^2(K+1)^2/4)$ distribution, or finite sample approximations to critical values may be obtained by bootstrapping the statistic under the null of iid innovations.

One concern with multivariate tests for conditional heteroskedasticity is their low power in finite samples. An alternative approach is to test for GARCH in each VAR equation individually. The trade-off between these approaches, as measured by size and power, depends on whether there is conditional heteroskedasticity in all or only some of the VAR equations.

2.7.4 Time Invariance

An important assumption underlying standard VAR analysis is the time-invariance of the model. As defined earlier, stationarity requires time-invariant first and second unconditional moments. That assumption is violated not only if the stability condition is not satisfied; but it may also be violated if the parameters change over time. A wide range of procedures for checking the stability

or time-invariance of a given model exists (e.g., Doornik and Hendry 1997; Lütkepohl 2004, 2005, chapter 17). These procedures may be used to detect potential structural breaks during the sample period. The leading example of the class of tests of the null hypothesis of parameter stability is the Chow test, of which different versions have been proposed in the literature.

The central idea is to compare the null hypothesis of time-invariant parameters throughout the sample period against the possibility of a change in the parameter values at some date T_B. One version of the Chow test involves estimating the model on the full sample of T observations as well as on the first T_1 and the last T_2 observations, where $T_1 < T_B$ and $T_2 \leq T - T_B$. A formal LR test may be constructed from the Gaussian likelihood by comparing the maximum of the likelihood in the constant-parameter model to the maximum obtained after allowing for different parameter values before and after period T_B, possibly after dropping some observations around the break date.

Denoting the conditional log-density of the t^{th} observation vector by $l_t \equiv \log f_t(y_t|y_{t-1}, \ldots, y_{-p+1})$, the test statistic can be written as

$$LR_{\text{Chow}} = 2\left[\sup\left(\sum_{t=1}^{T_1} l_t\right) + \sup\left(\sum_{t=T-T_2+1}^{T} l_t\right) - \sup\left(\sum_{t=1}^{T_1} l_t + \sum_{t=T-T_2+1}^{T} l_t\right)\right]. \tag{2.7.2}$$

Under the null hypothesis of time-invariant parameters, the statistic has an asymptotic χ^2 distribution with degrees of freedom given by the number of restrictions implied by the assumption of a constant-coefficient model for the full sample period. In other words, the degrees of freedom are the difference between the sum of the number of free coefficients estimated in the first and last subperiods and the number of free coefficients in the full-sample model. It must be emphasized, however, that the distribution may be different if the process contains integrated components, even if the process is time invariant (see Lütkepohl 2005, section 8.4.3; Hansen 2003). For an investigation of the properties of Chow tests for VAR models, the reader is referred to Candelon and Lütkepohl (2001).

One practical question is how many observations to drop in between the two subsamples. Asymptotic arguments allow one to choose $T_1 = T_B - 1$ and $T_2 = T - T_B + 1$ without dropping any observations. If the parameter change is smooth or its exact timing is unknown, however, leaving out some observations may improve the small-sample power of the test.

Another version of a Chow test for multivariate time series models considers predictions from a model fitted to the first $T_B - 1$ observations and evaluates whether they are in line with the observed data (see Doornik and Hendry

1997). The latter authors also propose versions of these tests based on the F test statistic with the aim of improving the size of the test in small samples.

Candelon and Lütkepohl (2001) point out that, in particular for multivariate time series models, the asymptotic χ^2 distribution may be a poor approximation in small samples. Even F critical values may fail to correct the small-sample size distortions. Candelon and Lütkepohl therefore propose the use of bootstrap versions of Chow tests in order to improve their small-sample properties.

Chow-type tests can be generalized in various directions. For example, one can test for more than one break or one can assess the constancy of a subset of parameters keeping the remaining parameters fixed. It is also common to construct critical values for the null of a break at T_B under the premise that the test was performed repeatedly for a range of potential break points T_B. In this case, the critical values are derived for the test statistic $\sup_{T_B \in \mathcal{T}} LR_{\text{Chow}}$, where $\mathcal{T} \subset \{1, \ldots, T\}$ is the set of periods for which the test statistic is computed. In practice, it is necessary to trim some observations at the beginning and the end of the sample when defining the range of possible break dates to ensure that the test has power. The sup-test statistic does not have an asymptotic χ^2 distribution (see Andrews 1993; Andrews and Ploberger 1994; Hansen 1997a). Its asymptotic distribution depends on the fraction of the sample over which a search is performed. Critical values are given by Andrews (1993) for a range of situations. They can be derived analytically in some cases, but more commonly they are derived by bootstrap methods under the null hypothesis of no break (see Christiano 1992; Diebold and Chen 1996). The critical values obtained from this distribution guard against the danger of data mining across possible break points by considering only break dates that appear favorable to the rejection of the null of no break based on preliminary inspection of the data. Using conventional critical values in this case would result in spurious evidence of a structural break. Only if the break date can be pinned down uniquely based on non-sample information will conventional critical values for the Chow test apply.

Other tests for structural change are based on recursive residuals obtained by fitting models to samples of increasing length. In other words, starting from some sample size T_0, a recursive residual $\hat{u}_\tau^{(r)}$ is the last residual obtained by fitting a model to a sample for $t = 1, \ldots, \tau$ for $\tau \in \{T_0, \ldots, T\}$. Cumulated recursive residuals or squared residuals are then used for testing for structural change. These tests are known as CUSUM and CUSUM-of-squares tests and were originally proposed by Brown, Durbin, and Evans (1975) for testing structural change in linear regression models. Krämer, Ploberger, and Alt (1988) establish their validity in dynamic models. These tests are designed to test the structural stability of a model if the specific change point is not known. An application of this class of tests in the VAR context can be found in Lütkepohl (2004).

It should be noted that tests for structural change are prone to rejecting the null of no break in small samples, even when the null is true, whenever there are large transitory dynamics in the DGP. In the latter case, large transitory dynamics and permanent breaks tend to be observationally equivalent, and caution is called for in interpreting the results of these tests.

In this chapter the point of departure has been a VAR model with time-invariant parameters. Given this premise, it makes sense to treat the constant-parameter model as the null hypothesis. An alternative approach would be to start the specification search from a more general model class that allows for time-varying slope coefficients. Such models are considered in Chapter 18. For example, one may model the evolution of the model coefficients as a random walk process. Under the null hypothesis that the innovation variance of this random walk process is zero, the model coefficients are time-invariant. Suitable tests of this hypothesis have been discussed in Nyblom (1989) and Teräsvirta, Tjøstheim, and Granger (2010, section 6.4.4), among others.

2.8 Subset VAR Models, AVAR Models, and VARX Models

VAR models allow for unrestricted feedback between the model variables for a given number of lags. The motivation for this approach is that we are typically unable to derive exclusion restrictions on the reduced-form lag structure from economic theory. Such restrictions simply are not economically credible (see Sims 1980a). The advantage of unrestricted VAR model estimates is that they are more robust to misspecification than estimates from dynamic simultaneous equation models imposing more exclusion restrictions. The disadvantage of unrestricted VAR model estimates is that they tend to be less precise and hence may not be informative about the questions of interest. Almost from the inception of the VAR approach, therefore, researchers have tried to explore the use of statistical model selection criteria to reduce the perceived parameter profligacy of VAR(p) models and to improve the efficiency of VAR model estimates.

2.8.1 Subset VAR Models

The most common proposal is to allow for different lag structures in different equations of the model, rather than imposing the same lag order on all VAR model equations (see, e.g., Hsiao 1979, 1981). The resulting subset VAR models differ from the standard VAR models considered so far in that the regressors differ across equations. Subset VAR models may be estimated using equation-by-equation LS, but that estimator will not be efficient. Alternatives include (possibly iterated) feasible GLS estimation methods or constrained full information ML estimation methods, as discussed in Section 2.3.

One situation in which imposing different lag orders on the VAR equations may be plausible is the following example. Consider a bivariate VAR(1) model for real GDP growth and stock returns. Whereas real GDP growth exhibits some sluggishness, requiring the inclusion of at least one lag, stock returns are well approximated by a white noise process, suggesting that zero lags of the model variables in the second VAR equation may be adequate in practice. This argument is not based on a priori economic theory, but on the time series properties of the data. Rather than imposing this restriction ex ante, one may use statistical tests to determine whether to include the lagged VAR coefficients in the second equation. One testing procedure for determining subset VAR models is based on the t-ratios of individual model parameters. We sequentially eliminate the variables with the lowest t-ratios until all remaining variables have t-ratios greater than some threshold value, say 1.96. A more common approach is to evaluate information criteria for all plausible combinations of lag orders in the VAR equations, which allows us to compare all models at the same time. Compared with conventional lag-order selection in VAR(p) models, subset VAR model selection is more computationally demanding because there is no natural ordering between alternative subset models and the number of permutations to be considered may be as high as 2^{K^2p}. Moreover, the asymptotic theory motivating the use of information criteria for lag-order selection is derived under the premise that the number of candidate models is small relative to T (see, e.g., Inoue and Kilian 2006). One would not expect these criteria to be reliable when the number of subset models considered is large.

Brüggemann and Lütkepohl (2001) compare alternative strategies of selecting subset VAR models in a simulation study. Although none of these strategies was found to be reliable when it comes to detecting the DGP, using subset VAR models in some cases may result in models with improved forecast precision or with tighter impulse response confidence intervals than in the unrestricted VAR model. Nevertheless, subset VAR models have not played an important role in empirical research. One concern, already alluded to, is that pretests of this type may undermine subsequent inference about impulse responses or related statistics of interest. In addition, examples in which one would expect the appropriate lag order to differ a lot across equations are not common in empirical macroeconomics.

2.8.2 Asymmetric VAR Models

Keating (1993) introduces the closely related idea of asymmetric VAR (AVAR) models, in which the lag length differs across variables such that the lag length is the same for each variable in all equations, but may differ across variables.[7]

[7] The notion of asymmetry here is not related to the notion of asymmetric responses in Chapter 18, but merely to the lag structure.

This contrasts with subset VAR models in which the lag lengths of a given variable may differ across equations. There is no systematic evidence that asymmetric VAR models improve the accuracy of impulse response estimates or forecasts.

2.8.3 VARX Models

A third example of a restricted VAR model with different lag structures across equations is the VARX model (e.g., Lütkepohl 2005, chapter 10). VARX models are VAR models in which one or more variables are exogenous with respect to the remaining variables (see Chapter 7). This implies that there is no feedback from lagged endogenous variables to the exogenous variables in the system of reduced-form equations, allowing us to restrict these lag coefficients to zero. The equations for the exogenous variables include the same number of lags as the other equations but only lags of the exogenous variables. The reduced-form representation of the VARX model effectively imposes Granger noncausality from endogenous to exogenous variables. A VARX model may be estimated as a system of equations by restricted LS or ML methods (see Sections 2.3.3 and 2.3.4). Alternatively, the model may be estimated without separate equations for the exogenous variables.

The use of exogenous variables is not common in the VAR literature, but there are exceptions. For example, it has been argued that ocean temperatures are exogenous with respect to the global business cycle at least over the horizons considered in business cycle analysis. Similarly, one could make the case that the hours of sunlight per day in New York are exogenous with respect to the Dow Jones stock price index. Finding examples of exogenous economic variables is even harder. Until a few years ago, the price of crude oil was considered exogenous with respect to the U.S. economy. A number of recent studies, however, has shown that there is an important endogenous element in this price series (see Kilian 2008a). A better example is a small open economy that faces exogenous variation in world interest rates or in its terms of trade (see, e.g., Cushman and Zha 1997). Another application of VARX models involves models that include extraneous estimates of exogenous monetary and fiscal policy shocks. These shocks may be treated as exogenous variables that are subject to further exclusion restrictions on their own dynamics (see Chapters 6, 13, and 15).