Vector autoregressions

Robert M. Kunst

September 2007

This course is exclusively based on the book "New Introduction to Multiple Time Series" by Helmut Lütkepohl. While the book's title indicates some greater generality, we will restrict focus to vector autoregressions as the basic tool of linear multiple time-series analysis. This may even include some nonlinear models with asymmetric, breaking or threshold behavior that are variants of linear vector autoregressions.

1 Introduction

In this section, Lütkepohl departs from the aim of *predicting* economic variables, which is certainly an important but not the only conceivable task of time-series modelling. Note the slight error in the basic definition on page 2, as a *multiple time series* is not a set but a vector of time series.

Rigorously, a *time series* is a finite sequence of observations on a variable, for example of y, at time points $t = 1, \ldots, T$. A typical member of the sequence is denoted as y_t. Then, a *multiple time series* consists of observations y_{kt} for variables $k = 1, \ldots, K$ and for time points $1, \ldots, T$. The word 'time series' refers to data. T denotes the *sample size* or the length of the time series.

In contrast, a 'time-series process' is a *stochastic process* and therefore a statistical, constructed entity. The basic idea of time-series analysis is that the observed time series is a (partial) realization of such a stochastic process. We remember that univariate time-series analysis uses processes of the type

$$y : Z \times \Omega \to \mathbb{R},$$

where Z denotes an index set and Ω forms a probability space, formally denoted as $(\Omega, \mathcal{F}, \Pr)$ with a σ–algebra $\mathcal{F}$ and a probability measure Pr. For *discrete-time* processes, which are exclusively used here, Z is supposed to be 'at most countable' and may be the integers or the positive integers.

We also remember that, for fixed $t \in Z$, y_t is a (real-valued) *random variable*, while for fixed $\omega \in \Omega$, $y(\omega)$ is a sequence of real values that is called a *realization* of the process or sometimes a *trajectory.* A stochastic process may therefore be viewed as a random variable with values that are sequences, or alternatively as a sequence of real-valued random variables that are defined on a common probability space.

Now, in multivariate time-series analysis, the only change is that we focus on processes that are formally defined as

$$y : Z \times \Omega \to \mathbb{R}^K,$$

where the random variables are K-vector-valued. Again, for fixed $t \in Z$, we have a vector-valued random variable, and for fixed ω, we have a *realization* of K possibly infinite sequences of real numbers.

Like in all time-series textbooks, the same symbol y is used for time-series, i.e. data, and for time-series processes, i.e. stochastic entities. The expression *data generation process (DGP)* is used for the process that has or may have generated the observed data, such that the observed data is interpreted as a (partial) realization of the DGP. Of course, we know for sure that the DGP is a DGP only in cases where the data have been generated artificially by using a random generator. Otherwise, the concept of the DGP is an abstraction.

Finally, a *vector autoregressive process* is defined as a special case of multivariate time-series process, such that y_t, a K–vector, depends on its past via the formula

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

where ν is a constant K–vector for the intercept and all A_j, $j = 1, \ldots, p$ are $K \times K$–matrices, while u_t denotes a *multivariate white noise* process. In this scheme, any of the component variables $y_{k,t}$, $k = 1, \ldots, K$ depends on p lags of itself and of the other $K - 1$ component variables.

For the moment, the properties of the white noise u_t are not defined more rigorously. A convenient backdrop case is that this is an independent sequence of multivariate Gaussian random variables with identical expectation of 0 and identical variance matrix Σ_u. Not all results hold for the most general case of uncorrelated variables with identical zero mean and identical variance matrix, which would be the usual white-noise definition.

The section closes with the traditional Box-Jenkins flowchart for applied time-series analysis, where time-series models are specified and estimated and then subjected to 'model checking'. If the models are found to be 'good', one may continue with forecasting and other uses ('structural analysis'). If the models are found to be bad, one is to return to the specification stage. While the flowchart is commonly followed in principle, many users may specify and estimate a set of models right from the start and compare them rather than check for a single model's validity. For example, it may make sense to compare a set of valid and invalid models in a forecast evaluation.

2 Stable VAR Processes

The word 'stable' is used to describe a property of the coefficient structure, notwithstanding the influence of starting conditions. Thus, it corresponds to 'asymptotically stationary' with other authors. Note that some authors use 'stable' as a synonym for 'stationary'.

2.1 Basic assumptions and properties

A VAR(1) process

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

with white noise u_t (i.e., $\mathrm{E}u_t = 0$ and $\mathrm{var}u_t = \Sigma_u$ and uncorrelated over time) is called *stable* if all eigenvalues of A_1 have modulus less than one. This assumption corresponds to the assumption of a coefficient of modulus less than one in a univariate AR(1) process and allows to continue the following iterative substitution:

$$\begin{aligned} y_t &= \nu + A_1(\nu + A_1 y_{t-2} + u_{t-1}) + u_t, \\ &\quad ... \\ y_t &= (I_K + A_1 + ... + A_1^{t-1})\nu + A_1^t y_0 + \sum_{j=0}^{t-1} A_1^j u_{t-j} \end{aligned}$$

to infinity, assuming the process was started in the infinite past, such that one obtains

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

with μ defined as $(I_K - A_1)^{-1}\nu$, the formal limit of the geometric sum of the matrices A_1^j. It can be shown that this assumption is sufficient for the convergence of the sum of stochastic variables in mean square, such that y_t is a well-defined stochastic process.

Invoking a result from matrix algebra, it is stated that the condition on the eigenvalues of A_1 is equivalent to the condition

$$\det(I_K - A_1 z) \neq 0 \quad \text{for} \quad |z| \leq 1.$$

This is straightforward, as eigenvalues are defined as solutions to $\det(A_1 - \lambda I_K) = 0$. This equation can be divided through λ. Therefore, the eigenvalues are less than one if and only if the solutions to the above condition are larger than one.

What about higher-order VAR(p) processes? The main result is that the generalized condition

$$\det(I_K - A_1 z - A_2 z^2 - \ldots - A_p z^p) \neq 0 \quad \text{for} \quad |z| \leq 1$$

yields stability.

This stability condition can be derived from the following ideas. Firstly, any VAR(p) can be written as a VAR(1) if one uses the matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} A_1 & A_2 & \ldots & A_{p-1} & A_p \\ I_K & 0 & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \ldots & I_K & 0 \end{bmatrix},$$

in the form $Y_t = \nu + \mathbf{A}Y_{t-1} + U_t$, where Y_t is a Kp–vector that contains p successive observation vectors y_t, stacked on top of each other:

$$Y_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \ldots \\ y_{t-p+1} \end{bmatrix},$$

while ν and U_t contain non-zero elements in their first K rows only:

$$\nu = \begin{bmatrix} \nu \\ 0 \\ \ldots \\ 0 \end{bmatrix}, \quad U_t = \begin{bmatrix} u_t \\ 0 \\ \ldots \\ 0 \end{bmatrix}.$$

Meditating on this so-called state-space representation, one finds that the original VAR(p) equation is contained in the first block of K rows, while the remaining rows are mere identities $y_{t-1} = y_{t-1}$ etc. Anyway, it is a VAR(1) model and the above condition

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

guarantees stability. By some matrix algebra (properties of determinants of block matrices), it is fairly easy to establish that the two versions of the stability condition for AR(p) models are equivalent.

Similarly, the result $\mu = (I_{Kp} - \mathbf{A})^{-1}\nu$ must hold, where only the first K rows of the Kp–vector μ are of interest. Formally, this part can be obtained as

$$\mu = J\mu = J(I_{Kp} - \mathbf{A})^{-1}\nu,$$

where J is a $K \times Kp$ block matrix of the form

$$J = [I_K : 0 : \ldots : 0] .$$

Similarly, one may formally represent the first part of the Y_t vector as

$$y_t = JY_t = J\mu + J\sum_{j=0}^{\infty} \mathbf{A}^j U_{t-j}.$$

Once such an infinite-order moving-average representation of a VAR process has been established, one may also calculate its autocovariance function. For example, $y_t = \mu + \sum_{j=0}^{\infty} A_1^j u_{t-j}$ (for a VAR(1) process) implies

$$\begin{aligned} \Gamma_y(h) &= \mathrm{E}\,(y_t - \mu)\,(y_{t-h} - \mu)' \\ &= \lim_{n\to\infty} \sum_{i=0}^{n}\sum_{j=0}^{n} A_1^i \mathrm{E}\left(u_{t-i}u'_{t-h-j}\right)\left(A_1^j\right)' \\ &= \sum_{j=0}^{\infty} A_1^{h+j}\Sigma_u A_1^{j\prime}, \end{aligned}$$

which uses the white-noise property in the last equality. Similarly, for a general VAR(p) process, one may write

$$\Gamma_Y(h) = \sum_{j=0}^{\infty} \mathbf{A}^{h+j}\Sigma_U\left(\mathbf{A}^j\right)' .$$

This is a $Kp \times Kp$–matrix, with the interesting part $\Gamma_y(h)$ forming the northwest $K \times K$ block, which again may be retrieved formally using J:

$$\Gamma_y(h) = \sum_{j=0}^{\infty} J\mathbf{A}^{h+j}\Sigma_U\left(\mathbf{A}^j\right)' J'.$$

One may also note that $\Gamma_Y(0)$ contains all $\Gamma_y(j)$ with $0 \le j \le p-1$ as building blocks.

A more traditional moving-average representation is obtained using $u_t = JU_t$:

$$\begin{aligned} y_t &= J\mu + J\sum_{j=0}^{\infty} \mathbf{A}^j U_{t-j} = \mu + \sum_{j=0}^{\infty} J\mathbf{A}^j J' J U_{t-j} \\ &= \mu + \sum_{j=0}^{\infty} \Phi_j u_{t-j}, \end{aligned}$$

where $\Phi_j = J\mathbf{A}^j J'$, which looks simpler than it is. The large matrix $\mathbf{A}$ must be taken to its j-th power first, before one can focus on its northwest corner by multiplying J into it. Thus, these operations will rarely be done on paper. They serve as a basis for computer programming in matrix languages and for theoretical proofs.

From these manipulations, the formula

$$\Gamma_y(h) = \sum_{j=0}^{\infty} \Phi_{h+j} \Sigma_u \Phi_j'$$

follows immediately.

An alternative way of conducting the manipulations is by using the lag operator L of $Ly_t = y_{t-1}$. Using

$$A(L) = I_K - A_1 L - \ldots - A_p L^p,$$

we may write the AR(p) model as

$$A(L)\, y_t = \nu + u_t$$

or

$$y_t = A^{-1}(1)\,\nu + A^{-1}(L)\, u_t.$$

Apparently, the book by LÜTKEPOHL does not use the lag operator calculus as much as others and it mainly relies on matrix algebra.

At this point, the expressions *stationary* and *stable* are explained in some more detail. In line with the time-series literature, a process is called *stationary* when its mean and its autocovariance function are time-constant. If the joint distribution of successive observations is time-constant, the process is called *strictly stationary.* This terminology is equivalent to the one known from univariate time-series analysis. Of more interest is the relation of stationarity and stability. Stable processes are always stationary when started in the infinite past (Proposition 2.1). Stationary processes must be stable if their law of generation can be interpreted in the usual direction of time. Non-degenerate processes with roots on the unit circle cannot be stationary.

Is every stationary multivariate process a VAR process? No, but one may justify VAR modelling as follows. According to *Wold's Theorem*, every stationary process has a moving-average representation. In 'most' cases, multivariate moving-average processes can be approximated by VAR processes.

If the world is so complicated, can one use VAR models with a low dimension? If a K–variate process is stationary, then also any k–variate subprocess of it with $k < K$ is stationary. The k–variate subprocess is never 'misspecified', there is no 'omitted variable bias'.

The section closes with some further algebraic results. For example, while it is easy to derive the mean μ from the coefficient matrices and the intercept ν, the matrix $\Gamma_y(0)$, which just describes the variances and covariances of the vector variable y_t, i.e. $\text{var}(y_t)$, is determined by the formula

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

for the VAR(1) and by an analogous formula for VAR(p). Such a closed-form evaluation may be preferable to the infinite sum that was introduced above. However, it requires the implementation of Kronecker products. Similarly, h–order autocorrelations $R_y(h)$ can be formally obtained from autocovariances as

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

using the diagonal matrix D that contains square roots of the diagonal elements of $\Gamma_y(0)$ along its diagonal, i.e. standard deviations of y_t. Such properties are convenient tools for programming.

2.2 Forecasting

Because forecasting is one of the main aims of VAR modelling, it is worth while to spend some thoughts on forecasting. However, it appears that many economists are not so much interested in forecasting as in structural analysis. That topic is covered in the next subsection.

Lütkepohl defines the task of forecasting as follows. In time point t (the *forecast origin*), an information set is available, denoted as Ω_t. Formally, this may be a σ–algebra built from the past of the process $\{y_s, s \leq t\}$. The forecaster tries to approximate a 'future' y_{t+h}, $h > 0$, as close as possible. This approximation, formally measurable with respect to Ω_t, is called *h–step predictor* and h is called the *forecast horizon.* It is debatable whether the forecaster typically wants to be 'close' to the random variable y_{t+h} or rather to the realized value $y_{t+h}(\omega)$. In real life, the latter target is more likely.

Traditionally, the forecasting literature distinguishes *point prediction*, if the aim of prediction is a number $\hat{y}_{t+h}$, *interval prediction*, where the aim of prediction is a confidence interval, *variance prediction*, for example in forecasting volatility, and *distribution prediction*, where the complete distribution of the random variable y_{t+h} is forecast(ed). The book focuses on point and interval prediction.

A basic result in forecasting is that the expected squared error $\mathrm{E}\left(\hat{y}_{t+h} - y_{t+h}\right)^2$ is minimized among all Ω_t–measurable predictors $\hat{y}_{t+h}$ by using the conditional expectation

$$\hat{y}_{t+h} = \mathrm{E}\left(y_{t+h}|\Omega_t\right).$$

It follows that conditional expectations are 'optimal' predictors if the forecaster's *loss* is well approximated by squares. In a VAR scheme

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

the conditional expectation is evaluated easily, if u_t can be assumed to be *independent* or more generally fulfills some martingale-difference property. If u_t is merely uncorrelated white noise, conditional expectation may differ from the following simple construction. Even in that case, however, this construction yields the best *linear* predictor. Some textbooks even introduce a separate notation for linear conditional expectation.

If $\mathrm{E}(u_{t+1}|\Omega_t) = 0$, then clearly

$$\mathrm{E}\left(y_{t+1}|\Omega_t\right) = \nu + A_1 y_t + \ldots + A_p y_{t-p+1},$$

and similarly

$$\mathrm{E}\left(y_{t+2}|\Omega_t\right) = \nu + A_1 \mathrm{E}\left(y_{t+1}|\Omega_t\right) + A_2 y_t + \ldots + A_p y_{t-p+2}.$$

By iteration, h–step predictors can be obtained from $(h-1)$–step predictors for any h. Basically, this is how forecasting works within a time-series computer program. Usually, all parameters ν, A_j, $j = 1, \ldots, p$, Σ_u, must be estimated from data. If only values within Ω_t are used for this estimation, the forecast is usually called *out-of-sample.* This is the acid test for good forecasting properties. Economists tend to write $\mathrm{E}_t\left(y_{t+h}\right)$ for $\mathrm{E}\left(y_{t+h}|\Omega_t\right)$ if the information set Ω_t is clear from the context.

Using the state-space representation $Y_t = \nu + \mathbf{A}Y_{t-1} + U_t$ makes forecasting notation particularly simple, as the VAR(1) formula

$$\mathrm{E}_t\left(Y_{t+1}\right) = \nu + \mathbf{A}Y_t$$

yields

$$\mathrm{E}_t\left(Y_{t+h}\right) = \left(I_{Kp} + \mathbf{A} + \mathbf{A}^2 + \ldots + \mathbf{A}^{h-1}\right)\nu + \mathbf{A}^h Y_t$$

by iteration, and this $\mathrm{E}_t\left(Y_{t+h}\right)$ contains the forecast $\mathrm{E}_t\left(y_{t+h}\right)$ as its first K elements.

It is easily shown that the prediction error can be represented as a moving average

$$y_{t+h}-\mathrm{E}_t\left(y_{t+h}\right)=\sum_{j=0}^{h-1}\Phi_j u_{t+h-j},$$

where the Φ_j are the Wold-type moving-average matrices that were introduced above. This allows some crude evaluation of the variance of the error

$$\Sigma_y\left(h\right)=\sum_{j=0}^{h-1}\Phi_j\Sigma_u\Phi_j'.$$

As h increases, this forecast-error variance converges to the variance of y, i.e. Σ_y. This means that the forecast becomes less informative, until its error variance becomes as large as the total variance.

In practice, these formulae do not hold exactly, as the parameters are unknown and their estimation uncertainty adds to the above. Similarly, the expression on the right-hand side of the formula can only be approximated from data. Particularly, the mean squared error in a forecasting experiment need not increase monotonically in h. Even if all parameters were known, it would just be an *estimate* of $\Sigma_y\left(h\right)$.

Genuine interval forecasts are less common in economic applications. Note that some interval constructions use the assumption of Gaussian (normal) errors that is not always fulfilled in empirical applications.

2.3 Structural analysis with VAR models

This subsection is of primary interest for many economic applications. Lütkepohl focuses on three aspects: Granger causality and variants, impulse response analysis, and forecast error variance decompositions. These aspects are related.

2.3.1 Granger causality

In 1969, the 2003 Nobel laureate Granger introduced a then new statistical concept of causality, already within the VAR framework. According to his definition, a time-series variable x is said to *cause* z if (according to Lütkepohl)

$$\Sigma_z\left(h|\Omega_t\right)<\Sigma_z\left(h|\Omega_t\backslash\{x_s,s\leq t\}\right)$$

for at least one $h > 0$. In words, the forecast error variance of predicting z_{t+h} from a complete universe that contains the past of z and x is smaller than the forecast error variance from a universe without the past of x. In short, according to the idea, the forecast for z_{t+h} improves by accounting for x. A problem with the definition is the subtraction symbol. If x and z are closely related, the condition may be fulfilled trivially. This critique may be countered by *adding* the past of x to the left-hand and omitting it from the right-hand side, instead of *subtracting*. In line with typical usage in multivariate analysis, it is to be noted that the symbol $<$ denotes that the difference right minus left is non-negative definite and non-zero.

In practice, the conditional expectations behind the formula will often be replaced by linear expectations (see above) and possible h will only be evaluated within reasonable limits.

In the 1970s, the concept of Granger causality was discussed fiercely. It is now generally agreed that it constitutes an important concept of its own right, while this does not exclude the possibility that other causality concepts may be more appropriate in certain circumstances. Note that z causing x does not exclude the possibility that x causes z, a feature called *feedback*.

The power of the concept relies on the fact that the formula has a simple characterization within a VAR. To present that one, assume that the $K-$variate process y_t consists of the two components z and x of dimension M and $K-M$. Formally, $y_t' = (z_t', x_t')$. Further, assume that y_t has a convergent infinite-order MA representation

$$y_t = \mu + \sum_{j=0}^{\infty} \Phi_j u_{t-j} = \mu + \Phi(L) u_t, \quad \Phi_0 = I_K.$$

Then, the MA(∞) model may also be written in its partitioned form

$$y_t = \begin{bmatrix} z_t \\ x_t \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \Phi_{11}(L) & \Phi_{12}(L) \\ \Phi_{21}(L) & \Phi_{22}(L) \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}.$$

It is shown easily, using evaluations of conditional expectation and uniqueness arguments, that x does not Granger-cause z if and only if $\Phi_{12} \equiv 0$, i.e. if $\Phi_{12,j} = 0$ for all $j = 1, \ldots, \infty$. This is exactly Lütkepohl's Proposition 2.2. Of course, an equivalent property holds for z not causing x and for Φ_{21}.

This result is of interest for theoretical considerations and for impulse response analysis, while for most practical purposes it is more of concern how (non-)causality is expressed in VAR models. Fortunately, it holds that for all finite-order VAR processes, a completely identical property holds, i.e. in the partitioned VAR system

$$\begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} z_t \\ x_t \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix},$$

x will not Granger-cause z if and only if the operator $A_{12} \equiv 0$. Likewise, z will not Granger-cause x if and only if $A_{21} \equiv 0$. The proof of equivalence uses the facts that the inverse of a block triangular matrix is again block triangular and that the leading matrix $A_0 = I_K$.

If the vector variable is partitioned into three parts and one wishes to investigate causality from one part to another one, keeping the third part as given, one may conjecture that zero operator blocks again convey the key information. The section in LÜTKEPOHL's book on pp. 49-51 is very valuable in voicing a warning against this naive conjecture. One can show that it holds for moving-average systems but that it is not generally true for vector autoregressions. Even when a block that formally appears to correspond to x causing z with y given may be zero, x may still cause z at larger horizons $h > 1$. Exact conditions are a bit involved.

It is the basic idea of Granger's causality that the causal direction can be seen from the cause regularly *preceding* the effect. Nevertheless, Granger added the concept of *instantaneous causality* to his 1969 paper. A variable x is said to cause z instantaneously if the forecast for z_{t+1} using the past of z and of x and some other information can be 'improved' by taking current x_{t+1} into account. It can be shown that instantaneous causality has no direction, in the sense that x causes z instantaneously if and only if z causes x instantaneously.

It can also be shown that instantaneous causality is seen in the error covariance matrix of a traditional VAR or multivariate MA system. If an off-diagonal $M \times (K - M)$–block in Σ_u is zero, i.e.

$$\Sigma_u = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix},$$

then there is no instantaneous causality between the first M and the remaining $K - M$ variables.

Alternatively, however, one may consider a VAR system of the form

$$\begin{aligned} &\begin{bmatrix} A_{11}(0) & A_{12}(0) \\ 0 & A_{22}(0) \end{bmatrix} \begin{bmatrix} z_t \\ x_t \end{bmatrix} \\ = &\begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \begin{bmatrix} A_{11}(L) - A_{11}(0) & A_{12}(L) - A_{12}(0) \\ A_{21}(L) & A_{22}(L) - A_{22}(0) \end{bmatrix} \begin{bmatrix} z_t \\ x_t \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix}, \end{aligned}$$

with triangular leading matrix $A(0)$ and diagonal Σ_v. It is easy to transform a system of the u–type into the v–form by splitting Σ_u into its 'roots' $\Sigma_u = LL'$. In the v–system, instantaneous causality is characterized by $A(0)$ being block diagonal. Intuitively, this means that z does not depend on current x but note that intuition can be tricky. While x does not appear to depend on current z in any case, we know that the instantaneous causality property is symmetric. The VAR model cannot be estimated with admitting a completely unrestricted $A(0)$—intuitively, current z as explaining x and *vice versa*—and an unrestricted error covariance matrix, as that system would not be identified.

2.3.2 Impulse response analysis

Because impulse response analysis (IRA) appears to correspond to the economic concept of multiplier analysis, thus pretending to answer the question how much a variable x changes at time point $t+h$, if another variable z is changed at t, it enjoys a tremendous popularity. The basic problem, however, is that this economic question cannot be answered within a VAR. IRA can only inform us by how much a variable x changes at time point $t+h$ if an error term changes at time point t. Because there are many different observationally equivalent representations of a VAR, there is no unanimous match between error terms and variables.

In this section, Lütkepohl tries to circumvent the problem by defining impulse responses as responses of mean forecasts to changes in starting conditions. It was outlined that in a VAR(1)—for the ease of notation, in this section we consider models without intercepts—mean forecasts can be obtained by $\hat{y}_t(h) = A_1^h y_t$ and $\hat{y}_0(h) = A_1^h y_0$. If the process is started in $t = 0$, one may use $y_0 = u_0$, fill for example the starting vector y_0 with a unit vector, and evaluate the vectors

$$A_1^h \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad h = 0, 1, \ldots,$$

which then give information on the reaction of forecasts in all y variables at successive time horizons. The same exercise can be repeated for all unit vectors, and one obtains a matrix of vectors that is conveniently represented as a matrix of time-series graphs. Lütkepohl calls this the *forecast error impulse responses*.

It is of main interest that the thus defined scheme is nothing else than a map of the moving-average representation of the VAR system, in above symbols $\Phi_0 = I_K, \Phi_1, \Phi_2, \ldots$ Thus, it follows immediately that blocks and parts will be zero whenever there is an event of non-causality. Non-causal directions appear as flat lines at zero in the usual matrix of plots. Because the MA representation must converge for a stable VAR, all impulse responses must approach zero as $h \to \infty$. Because a VAR(p) can always be represented as a VAR(1) with matrix $\mathbf{A}$, the same arguments must hold for higher-order VAR systems.

While a single impulse response curve always starts at one for the diagonals and at zero for the off-diagonals and approaches zero as $h \to \infty$, it need not be monotonic. However, a Proposition 2.4 states that, if it is zero for all $h \leq p(K-1)$, there will be no non-zero response for any larger h.

Some researchers prefer inspecting accumulated impulse responses. The sum of all MA coefficients must converge for a stable VAR, and hence these curves will converge to the respective elements of $I_K+\Phi_1+\Phi_2+\ldots$ as $h \to \infty$. Formally, these limits are easily evaluated as

$$\Psi_\infty = \Phi(1) = (I_K - A_1 - \ldots - A_p)^{-1},$$

as $\Phi(1)$ just denotes the (infinite) MA power series $\sum \Phi_j z^j$ evaluated at $z=1$.

Usually, this is not the kind of IRA that is performed in empirical studies. Rather, instead of the original MA representation

$$y_t = \sum_{j=0}^{\infty} \Phi_j u_{t-j}, \quad \mathrm{var} u_t = \Sigma_u,$$

one uses an orthogonalized representation. Assume that $\Sigma_u = PP'$, with lower triangular P, which representation is always possible due to the non-negative definiteness of a variance matrix with P. Then, we have

$$y_t = \sum_{j=0}^{\infty} \Phi_j PP^{-1} u_{t-j} = \sum_{j=0}^{\infty} \Theta_j w_{t-j}, \quad \mathrm{var} w_t = I_K,$$

where $\Theta_j = \Phi_j P$ and $w_t = P^{-1}u_t$. The matrices Θ_j are also MA coefficient matrices and may also define an 'impulse response'. The errors are uncorrelated in this representation, and therefore this is called an *orthogonal impulse response.*

Similarly, one may use a variant of the Cholesky factorization $\Sigma_u = FDF'$ with diagonal D with positive elements and lower-triangular F with the special property that F^{-1} has unit diagonal. This permits considering the VAR system

$$F^{-1}y_t = F^{-1}A_1 y_{t-1} + ... + F^{-1}A_p y_{t-p} + F^{-1}u_t,$$

which, after renaming $F^{-1}u_t = \varepsilon_t$ and $A_j^* = F^{-1}A_j$, $j = 1, \ldots, p$, $A_0^* = I_K - F^{-1}$ becomes

$$y_t = A_0^* y_t + A_1^* y_{t-1} + ... + A_p^* y_{t-p} + \varepsilon_t, \quad \mathrm{var}\varepsilon_t = D.$$

Here, the errors are orthogonal but variables depend on current observations of other system variables. However, note that by construction, as A_0^* is lower triangular with a zero diagonal, the first variable depends on no current values, the second variable depends on current values of the first one only, and so on. The representation is dependent on the ordering of variables, and it may make sense to put the variable that is least likely to react to current influences—according to economic prior information—into the first position.

These transformations and orthogonalizations play a key role in the so-called *structural VAR analysis* that apparently attracts much interest of economic researchers. In short, there is a multitude of different possible restrictions on A_0^* matrices, and $A_0^* = 0$ and the Cholesky form of a triangular matrix with zero diagonal are just two out of these many representations. Such restrictions cannot be obtained from the data—they are not 'over-identifying'—but they may follow considerations from economic theory. Some properties of the VAR system are immune to transformations—for example, stability conditions—while others are not—for example, IRA.

Contrary to LÜTKEPOHL's derivation, the Θ_j derived above do not correspond exactly to the A_j^* but to a re-scaled version. In other words, the Θ_j postulate unit-variance errors, while the errors in the A_j^* VAR have variances according to the D diagonal. Otherwise, the two representations are equivalent. Computer packages usually evaluate IRA according to the Θ_j representation.

Because of the system transformation, the orthogonalized IRA does not always reflect events of Granger (non-)causality. Nevertheless, LÜTKEPOHL shows that, also for orthogonal IRA, if an entry in the matrix of impulse responses is zero up to horizon $p(K-1)$, it will also be zero for larger horizons.

2.3.3 Forecast error variance decompositions

A different and also quite popular summary of the dynamic properties of a VAR is the so-called *variance decomposition.* LÜTKEPOHL defines a statistic $\omega_{jk,h}$ by

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

The meaning of the symbols is as follows. h is the forecast horizon, thus the formula describes a reaction after h lags. The subscripts j and k express the fact that the reaction of a variable j to a shock in w_k is under investigation, i.e. a reaction to the k–the error component. We note that this is not necessarily the same as the k–th variable. Θ_i is the i–th MA matrix of the (better: an) orthogonalized representation, and $\theta_{jk,i}$ is its element at position (j,k). Thus, the numerator accumulates the squared impulse responses in the (j,k) diagram up to a horizon h, while the denominator accumulates squared impulses of variable j to *all* influences.

Obviously, the expressions $\omega_{jk,h}$ sum up to one for different k. Therefore, they can be interpreted as percentages that show how much of the (forecast) error variance is due to a certain source. The expression 'forecast error variance decomposition' derives from the fact that the denominator represents the variance of the h–step forecast error from the VAR for the j–th variable. Note that the calculation requires orthogonal errors and it is invalid for the original Φ_j and u_t. Because all $\theta_{jk,i} \to 0$ as $i \to \infty$, the ω will not change anymore after a certain h. Usually, the error variance decomposition is summarized in tables rather than in graphs.

3 Estimation of VAR Processes

It is suggested that less emphasis is put on this chapter, as succeeding sections may be more relevant and interesting. Matrix manipulations are demanding. Apart from the usage of Kronecker products and of all their algebraic rules, particularly in conjunction with vectorization operators, Lütkepohl makes extensive use of selection and duplication matrices. Full command of the possibilities of this matrix manipulations requires some basic reading of the work of Magnus and Neudecker, for example.

3.1 Multivariate least squares

This is not to be confounded with *multiple* least squares. In multiple least squares, a vector variable depends on another vector variable. In the context of VAR estimation, we consider

$$Y = BZ + U,$$

where Y is a $K \times T$–matrix that contains the observations of the vector process y_t for $t = 1, \ldots, T$. In this notation—related to the so-called Dutch notation of econometrics—time points are columns and variables are rows. B is the coefficient matrix of dimension $K \times (Kp + 1)$, with the first column containing intercepts, ν' in the book's notation. Z contains the regressor observations, dimension is $(Kp + 1) \times T$, with a first row of ones for the intercept. The next K rows contain the first lag of Y, then comes the second lag, and so on. U is the matrix of errors and corresponds to Y.

This representation requires so-called *pre-sample* values for the time series at $t = 0, \ldots, -p + 1$, otherwise the lags matrix Z could not be filled. Clearly, if data are available from 1964, say, and $p = 2$ lags are modelled, $t = 1$ must correspond to the year of 1966, in this notation.

The vec operator stacks matrices into vectors, column by column. Thus, vec(Y) will be a KT–vector, starting with the $t = 1$ observations on the y variable, then the $t = 2$ observations, and so on.

The rules of vectorization imply that the basic equation can be re-written as

$$\mathrm{vec}(Y) = (Z' \otimes I_K)\mathrm{vec}(B) + \mathrm{vec}(U),$$

which looks like a (multiple) regression model. $Z' \otimes I_K$ is a large $KT \times K(Kp + 1)$–matrix whose building blocks are scalar diagonal matrices with the regressor elements. If the errors covariance matrix is Σ_u, then the error matrix of this regression will be

$$\Sigma_{\mathbf{u}} = I_T \otimes \Sigma_u,$$

which is block diagonal and repeats Σ_u along the diagonal.

For this 'long' regression model, the OLS estimate is calculated straightforward as

$$\hat{\beta} = \{(ZZ')^{-1}Z \otimes I_K\}\mathbf{y},$$

using bold face for the vectorized variables $\beta = \text{vec}(B)$ and $\mathbf{y} = \text{vec}\,(Y)$. Due to Dutch notation, transpose marks are the opposite of the familiar, i.e. not $Z'Z$ but ZZ'. Without any vectorization, we obtain the OLS representation in pure matrix form

$$\hat{B} = YZ'(ZZ')^{-1},$$

which is compact but less accessible. This is the way that multivariate least squares would be programmed in a matrix language.

In order to derive the asymptotic properties of OLS, the usual white-noise assumptions on errors u_t must be strengthened. The book suggests to define *standard white noise* by requiring *independence* over time instead of zero correlation and bounded fourth moments. Note that this is still weaker than *iid*, as higher moments are allowed to be time-varying within certain limits.

If U contains standard white noise, it can be shown (but is not even proved in the book) that the matrix ZZ'/T converges to a nonsingular—i.e. positive definite—limit matrix Γ and that

$$\frac{1}{\sqrt{T}}\text{vec}(UZ') \Rightarrow N(0, \Gamma \otimes \Sigma_u)$$

in distribution. These are typical central limit theorem assumptions for least-squares estimation in the stochastic regression model.

Using these additional assumptions, it is easily shown that the least-squares estimator $\hat{B}$ is consistent and that

$$\sqrt{T}(\hat{\beta} - \beta) = \sqrt{T}\text{vec}(\hat{B} - B) \Rightarrow N(0, \Gamma^{-1} \otimes \Sigma_u),$$

which again denotes convergence in distribution. OLS is root-T consistent, and all standard errors can be approximated by square roots of the diagonal of the estimated matrix to the right.

The unknown matrix Σ_u can be estimated consistently by

$$\tilde{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t \hat{u}_t',$$

if $\hat{u}$ are OLS residuals. Consistency is unaffected by degrees-of-freedom corrections. For example, a denominator $T - Kp - 1$ may take care of the coefficients to be estimated in every component equation. It remains debatable whether this is a good idea. Corrected standard deviations may show less bias, while uncorrected ones correspond to the maximum-likelihood estimation concept.

If coefficients and standard errors can be estimated, t–ratios can be formed. Note that it is not guaranteed that these statistics follow Student distributions in finite samples. It can only be shown that they converge in distribution to standard normal as $T \to \infty$. Based on conjectures and some simulations, one may consider t distributions with $T-Kp-1$ or $K(T-Kp-1)$ degrees of freedom. LÜTKEPOHL demonstrates that not much is to be gained by such modifications. Better stick to the asymptotic standard normal.

3.2 Mean adjustment and Yule-Walker estimation

The relevance of this section has been dictated by the tradition of time-series computer programs. Classical computer codes tend to 'simplify' time-series estimation by two features. Firstly, all data are corrected to zero means right at the beginning and calculation continues with mean-zero data. Secondly, autoregressions are estimated by moment estimates rather than by least squares.

Usually, mean adjustment does not do much harm. It can be shown that it does not matter at all in asymptotics, i.e. large samples. Of course, if VAR processes tend to reach the boundary of their stability region, it will matter more and will become less attractive. Note that least squares works for integrated (unstable) processes, while mean adjustment does not.

Moment estimates tend to simplify calculations, as the matrix ZZ' contains many instances of very similar parts. For example, cross products of y_t and y_{t-1} are not too different from cross products of y_{t-1} and y_{t-2}. Why not exploit this and estimate these parts by the same estimate? When computer time was still a matter of major concern, this simplification was an important device.

It was found that this so-called *Yule-Walker* estimate has advantages and disadvantages. It is advantageous that it always yields stable structures, i.e. the estimate $\hat{B}$ will correspond to a stable VAR. Least squares has a positive probability of creating unstable (explosive) structures, even when the DGP is stable. It is, however, a major disadvantage that Yule-Walker estimates are simply worse than OLS estimates. They have a larger bias in small samples and are less 'robust' to many specific features. Anyway, the asymptotic efficiency can be proved to be the same, as the aforementioned limit law also holds for Yule-Walker estimates.

3.3 Maximum Likelihood estimation

Because of Kruskal's Theorem, least-squares is maximum likelihood (ML) in the multivariate regression model, as long as all equations have the same regressors. This assumption is violated if some coefficients are restricted at zero, maybe following a first-round OLS estimation and an attempt at creating a more parsimonious VAR. In that case, ML can be approximated by some SUR estimator.

In the regular case, however, likelihood maximization and OLS have identical asymptotic properties. If ML relies on a nonlinear optimization procedure, it may face convergence properties and its careless application is generally discouraged. Asymptotic properties were already given for the OLS estimator. An exception is the variance matrix of the $\hat{\Sigma}_u$ estimate whose asymptotic variance matrix can only be expressed using special duplication matrices.

3.4 Forecasting, causality testing, impulse response inference

These sections contain very valuable material, particularly the compilation of results for the IRA inference is rather unique in time-series textbooks. At first reading, we may wish to skip them, however.

4 VAR Order Selection and Checking Model Adequacy

This is one of the strong parts of the book. Many researchers prefer to cite Lütkepohl, when it comes to the usage of information criteria, rather than the original literature or genuine time-series textbooks.

4.1 Introduction

The problem is how to determine the lag order p in a VAR

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

when observing a time series y_t, $t = 1, \ldots, T$. Traditionally, lag order selection for time series has been performed by any of the four following devices: visual inspection of correlograms etc. (the Box-Jenkins suggestion); information criteria; hypothesis tests, for example on whether a lag is zero; diagnostic tests. The last class of methods—'choose the lag such that the residuals pass diagnostics'—has turned out to be the least reliable, and the first one may require an expert in pattern recognition.

4.2 Sequences of tests

The section starts with the important truth that, if a VAR(p) model is 'correct', then also a VAR($p+1$) model will be correct, as a VAR(p) is a VAR($p+1$) with a zero coefficient matrix A_{p+1}. However, it is not advisable to estimate parameters that are really zero, as this may deteriorate forecasting performance. Even when the parameters are very small and the sample size is not too large, the gain in using additional parameters is dominated by far by the loss in working with a coefficient with a sizeable estimation error. 'Parsimonious' models will forecast best, as can be demonstrated easily by some Monte Carlo simulation.

A test for

$$H_0 : A_{p+1} = 0, \quad H_A : A_{p+1} \neq 0,$$

is a special case of a restriction test on the coefficients in a VAR($p+1$) model. It can be shown that the likelihood-ratio (LR) test statistic

$$\lambda_{LR} = 2\{\log l(\tilde{\delta}) - \log l(\tilde{\delta}_r)\}$$

is under H_0 asymptotically distributed as chi-square with N degrees of freedom. Here, $\tilde{\delta}$ is the ML estimate of a parameter vector δ under H_A—the unrestricted estimate—and $\tilde{\delta}_r$ is the estimate under H_0—the 'restricted' model.

N is the number of independent restrictions that define H_0, in the example of testing for A_{p+1} we have $N = K^2$. Finally, $\log l$ is the log-likelihood.

If estimation has been conducted based on likelihood maximization, typically l is available and the statistic can be calculated as given above. Otherwise, an alternative form may be convenient, such as the form

$$T(\log|\tilde{\Sigma}_u| - \log|\tilde{\Sigma}_u^r|),$$

where $\tilde{\Sigma}_u$ and $\tilde{\Sigma}_u^r$ are estimates of the error variance matrix from the unrestricted and restricted models. Absolute value signs denote determinant operators. The likelihood ratio depends on these variance matrix estimates only.

Other feasible test statistics are approximations to the LR statistic, such as Wald statistics, Lagrange-multiplier (LM) statistics or approximations thereof. The asymptotic distribution of these variants is identical under H_0. Particularly for the LM statistic, some researchers have suggested variants whose null distribution may be approximated by an $F(N, T - Kp - 1)$ distribution: so-called LMF tests. Other researchers use that distribution even for the LR statistic after division by N.

In any case, once the test tool is available, it can be used as follows. Either one estimates a VAR(M) with a large M that is the maximum of lag orders that one wishes to consider. The last matrix A_M is tested against zero and, if the test rejects, the selected order is M. Otherwise, one may consider testing for $A_{M-1} = 0$ in a VAR($M - 1$), and so on. This is an occasion of the 'general-to-specific' modeling principle.

Similarly, one may start from a VAR(0) or VAR(1) model and test whether an additional matrix $A_1 = 0$ or $A_2 = 0$. One may extend the lag order until one gets the first acceptance and then one stops. This strategy is more in line with LM testing. Traditionally, it is discouraged and the general-to-specific sequence is preferred. This preference is basically due to the feature that all maintained models are 'valid', as long as the largest VAR(M) is correctly specified. In contrast, most maintained models in a specific-to-general search will be 'invalid'. Consider, for example, the situation that the true lag order is 3, while we test VAR(1) against VAR(2).

However, even in a general-to-specific search will the selected p depend critically on the imposed significance level. While many may choose 5%, there is no coercive reason to do so. Some suggest letting that level go to zero, as $T \to \infty$. Also, the significance level of a single test in a sequence may be wildly different from the nominal level, as the outcome also depends on previous tests in the chain. The technique of information criteria answers most of these complaints.

4.3 Information criteria for VAR order selection

Information criteria (IC) are statistics that measure the distance between observations and model classes. If the IC value is small, the distance is small and the model class contains a good descriptor of the DGP.

Typical criteria consist of two additive (or multiplicative) parts. The first one is a naive goodness-of-fit measure, such as (minus) the maximized likelihood within a given model class or a residual variance matrix. It becomes smaller as the model becomes more sophisticated. The second part is a penalty that increases with the model's complexity. In simple criteria, this second part is just a monotonic function of the number of estimated ('free') parameters.

The most popular IC is the AIC due to Akaike

$$AIC(m) = \log|\tilde{\Sigma}_u(m)| + \frac{2}{T}m,$$

if m denotes the number of free parameters and $\tilde{\Sigma}$ denotes the ML estimate of the error variance matrix based on using the given model class with m free parameters.

In principle, VAR(p) models have $m = K^2p + K + K(K+1)/2$ free parameters. However, we are merely interested in finding p and we will keep the dimension K constant and we will not impose restrictions on the error variance matrix. Neither will we discard the intercept. Therefore, we act as if the VAR(p) model had $m = pK^2$ parameters, as this is the only part that is affected by the lag order p. Thus, for a VAR(p) process, the AIC is defined as

$$AIC(p) = \log|\tilde{\Sigma}_u(p)| + \frac{2pK^2}{T},$$

where the change of notation from the general $AIC(m)$ to $AIC(p)$ should be obvious.

Lütkepohl provides an ingenuous derivation of the factor 2 in the AIC by approximating the mean squared error (MSE) of a VAR-based one-step prediction. This motivates that the main strength of the AIC is due to the forecasting properties of the selected model. Some authors also maintain that AIC implies a fairly good model choice in smaller samples under different aspects. However, there is also a result ascribed to Paulsen that only criteria of the form

$$Cr(m) = \log|\tilde{\Sigma}_u(m)| + \frac{mc_T}{T}$$

can achieve *consistency* in the sense that, as $T \to \infty$, the selected $\hat{p}$ will be the true p with probability one. Here, c_T is a function of T that fulfills the properties $c_T \to \infty$ and $c_T/T \to 0$ as $T \to \infty$. Obviously, a constant does not fulfill the former condition.

The conditions ascribed to PAULSEN are fulfilled by the Hannan-Quinn (HQ) criterion

$$HQ(m) = \log|\tilde{\Sigma}_u(m)| + \frac{2m \log\log T}{T}$$

and by the SCHWARZ version of the BIC

$$SC(m) = \log|\tilde{\Sigma}_u(m)| + \frac{m \log T}{T}.$$

Obviously, HQ imposes a milder penalty by using a very slowly growing function and will admit slightly larger models, while SC is very strict and selects comparatively small lag orders.

If applied to vector autoregressions of order p, SC has the form

$$SC(p) = \log|\tilde{\Sigma}_u(p)| + \frac{pK^2 \log T}{T}.$$

Again, note that not all parameters are included in the matrices A_j, $j = 1, \ldots, p$, which is one of the reasons why these criteria vary across authors and software. Another source of discrepancies is the estimate $\tilde{\Sigma}$ that is supposed to be a ML estimate but is occasionally replaced by a bias-corrected estimate with a correction for degrees of freedom. For the AIC, these modifications bear the name AICC and are preferred by some time-series analysts due to better small-sample properties.

4.4 Residual analysis: autocorrelation

It is helpful to note that *residuals* from estimated VAR models are *not* white noise, while errors from the true VAR are white noise. Nevertheless, all tests for misspecification rely on residuals. Roughly, what is checked is whether the residuals are 'too non-white'.

The book by LÜTKEPOHL does not provide the properties of ACF estimates in the general case, it does so for the case that this is the ACF for a white noise. We have formally that

$$\sqrt{T}\mathbf{c}_h \Rightarrow N(0, I_h \otimes \Sigma_u \otimes \Sigma_u),$$

where $\Longrightarrow$ denotes convergence in distribution, and that

$$\sqrt{T}\mathbf{r}_h \Rightarrow N(0, I_h \otimes R_u \otimes R_u).$$

Here, $\mathbf{c}_h$ denotes $\mathrm{vec}(C_1, \ldots, C_h)$, which is formed from h empirical autocovariance matrices

$$C_j = \frac{1}{T} \sum_{t=j+1}^{T} u_t u'_{t-i},$$

while $\mathbf{r}_h$ denotes a similar vectorization of the first h empirical ACF matrices (Proposition 4.4). Correlations are formed simply from the autocovariance and variance matrix estimates. In the formal notation suggested in the book

$$R_j = D^{-1} C_j D^{-1},$$

the diagonal matrices D contain standard deviation estimates of the K components, i.e. square roots of the variance estimates.

The double Kronecker product looks frightening, and the asymptotic variance matrices are large with dimension $hK^2 \times hK^2$. Of more concern is the feature that the formulae are literally only valid for true white noise, and residuals are not white noise. Therefore, it is important that similar results hold for the estimated autocovariances of residuals from a fitted valid VAR model. It can be shown (Proposition 4.5) that

$$\sqrt{T}\hat{\mathbf{c}}_h \Rightarrow N(0, I_h \otimes \Sigma_u \otimes \Sigma_u - \left(\bar{G}' \otimes I_K\right) \left\{\Gamma_Y (0)^{-1} \otimes \Sigma_u\right\} \left(\bar{G} \otimes I_K\right)).$$

This means that the formula for true errors holds only up to a correction term. The correction term contains approximately known elements, such as Σ_u and the variance matrix of the process Y. Moreover, there appears a $\bar{G}$ matrix, which is defined as

$$\bar{G} = \begin{bmatrix} \Sigma_u & \Phi_1 \Sigma_u & \ldots & \Phi_{h-1} \Sigma_u \\ 0 & \Sigma_u & & \Phi_{h-2} \Sigma_u \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \Phi_{h-p} \Sigma_u \end{bmatrix}.$$

This matrix has dimension $Kp \times Kh$, and it is not quadratic. The above representation assumes $h > p$. We remember that the matrices Φ_j stem from the—usually infinite—MA representation.

A similar result holds for the—even more important—estimates of the errors autocorrelation function based on residuals $\hat{\mathbf{r}}_h$. The correction matrix has a similar structure but it is slightly more complex (Proposition 4.6).

Note that correction terms are non-negative definite, therefore autocovariance and autocorrelation functions defined from estimation residuals tend to be *smaller* than those defined from the unobserved errors. Failure to adjust results in low rejection power.

Usually, the hypothesis of whiteness of errors is tested using so-called portmanteau or Q tests. In analogy to the univariate case, Q statistics are defined as

$$Q_h = T \sum_{j=1}^{h} \mathrm{tr} \left(\hat{R}_j' \hat{R}_u^{-1} \hat{R}_j \hat{R}_u^{-1} \right),$$

where $\hat{R}_j$ are estimates of the autocorrelation function of the errors at lag j and $\hat{R}_u$ is an estimate of the errors correlation matrix $R_u = R_0$. It is fairly easy to show that the formula yields the same Q_h if autocovariance estimates replace autocorrelation estimates.

Under the null hypothesis of $R_1 = R_2 = \ldots = R_h = 0$, the asymptotic distribution of Q_h can be shown to be χ^2 with $K^2(h-p)$ degrees of freedom. It is important to remark, as Lütkepohl does here, that 'asymptotic' means $T \rightarrow \infty$ as well as $h \rightarrow \infty$ and that therefore the χ^2 approximation will be invalid for small h.

Many authors prefer to use the modified portmanteau statistic

$$\bar{Q}_h = T^2 \sum_{j=1}^{h} (T-j)^{-1} \mathrm{tr} \left(\hat{R}_j' \hat{R}_u^{-1} \hat{R}_j \hat{R}_u^{-1} \right),$$

whose asymptotic distribution under the null is identical to that of Q.

If one intends to test for only a few errors autocorrelations, it is recommended to use a Lagrange multiplier (LM) test statistic instead of the portmanteau—although Q is occasionally described as an approximation to an LM statistic. We remember from univariate analysis that an LM test for autocorrelation is constructed using the following steps:

1. Estimate the main model by OLS;

2. Regress the OLS residual of the main model on *all* regressors of the main model *and* on h lags of the OLS residuals;

3. the value TR^2 of the second, auxiliary regression is the LM statistic and it is distributed as χ^2 with h degrees of freedom under the null.

In a multivariate situation, only the R^2 has to be redefined, the remainder of the principle continues to hold, including the auxiliary regression which is now multivariate. Instead of the TR^2 form—which is even in the univariate context only an approximation to the LM statistic—Lütkepohl presents a direct form that relies on the coefficient estimates on $\hat{u}_{t-1}, \ldots, \hat{u}_{t-h}$ and their approximate covariance matrix. The number of these coefficients is hK^2, and this is the degrees of freedom of the χ^2 distribution that the statistic obtains asymptotically.

The section closes with a warning that the asymptotic approximation can be poor in smaller samples, as is the case for most residual statistics.

4.5 Residual analysis: normality

Non-normal errors may violate some of the assumptions for the variance properties of a VAR. They may also indicate that the linear VAR can be improved upon by taking care of nonlinear effects, particularly as linear forecasts are not optimal in the presence of non-linearity. They may also point to outliers and structural breaks. Thus, non-normal errors *may* sound a warning on VAR results, while they need not necessarily invalidate the VAR.

The most common check for non-normality in univariate models is the *Jarque-Bera test*, for which priority is acknowledged to LOMNICKI. This test checks whether skewness (third moment) is zero and whether kurtosis (normalized fourth moment) is 3. Thus, its null hypothesis contains non-Gaussian distributions. Under its null, the statistic is asymptotically distributed as χ^2 with 2 degrees of freedom.

The multivariate version of the Jarque-Bera test statistic suggested by LÜTKEPOHL is simply the sum of K univariate statistics that are calculated from K orthogonalized VAR residuals. The orthogonalization is performed via a Cholesky split identical to the one that is needed for the IRA. It is fairly obvious that this statistic is distributed as χ^2 with $2K$ degrees of freedom under its null. Orthogonalization is important, as only the sum of uncorrelated χ^2 random variables will yield a χ^2 distribution. Note that the software EViews offers two alternative suggestions of how to overcome the problem of correlation across the original VAR errors. If the LÜTKEPOHL option is selected, EViews estimates the errors covariance matrix by dividing through $T-1-Kp$, in order to correct for degrees of freedom. Other software, like STATA, uses simple T weighting, which implies an increased tendency to reject the normal null.

4.6 Residual analysis: structural change

LÜTKEPOHL considers two tests for the alternative that some characteristics of the VAR have changed over time. Firstly, the multivariate version of the popular univariate CHOW test assumes a given possible time point, at which we suspect that the process may have changed, as for example a major innovation in the economic environment. Alternatively, a stability test checks whether the linear VAR forecast behaves as it should under the assumption of normal errors and a well-specified time-constant VAR.

The implementation of the Chow test is fairly trivial. Because it essentially tests for the significance of $pK^2 + K$ regressors—all VAR regressors including the intercept multiplied with a dummy that is 0 before a break point and one afterwards—the Chow statistic has an asymptotic χ^2 distribution with $K + pK^2$ degrees of freedom under its null. It corresponds to a regression F for this special situation and may also be compared to the F distribution.

The forecast test departs from the observation that the (theoretical, i.e. for known VAR coefficients) h–step forecast error from a VAR model at time T has the property

$$e_T(h) \sim N(0, \Sigma_y(h)),$$

where

$$\Sigma_y(h) = \Sigma_{j=0}^{h-1} \Phi_j \Sigma_u \Phi_j'.$$

We remember that Φ_j denotes the matrices from the infinite MA representation of the VAR. In practice, the entities on the right side must be replaced by estimates. Then, one may hope that

$$\bar{\tau}_h = e_T(h)' \hat{\Sigma}_y(h)^{-1} e_T(h)$$

is distributed as χ^2 with K degrees of freedom under the null. Unfortunately, this is not quite true, as the naive estimate for Σ_y underestimates the variance of the prediction error. It is recommended to add an adjustment term to $\hat{\Sigma}_y$ that cares for the fact that the VAR parameters must be estimated and are not known to the forecaster.

5 Parameter constraints and Bayesian estimation

Most of this section is relatively specialized and can be omitted at first reading. The main difficulty with parameter constraints in a VAR system is that they usually violate the conditions for the validity of Kruskal's Theorem. Therefore, least squares is no longer equivalent to ML estimation and SUR methods yield efficiency gains.

LÜTKEPOHL concentrates on a SUR method under the name of EGLS, which is probably pronounced as 'eagles'. We remember that SUR can be viewed as a GLS (*generalized least squares*) procedure. In practice this in turn has to be substituted by 'feasible' GLS.

EGLS is a two-step method:

1. In the first step, the errors variance matrix Σ_u is estimated from the residuals of a standard unrestricted multivariate least-squares regression.

2. In the second step, the VAR is estimated using all restrictions and the estimate $\hat{\Sigma}_u$ instead of Σ_u by SUR (GLS) regression.

Numerical maximization of the likelihood (ML estimation) has few advantages over EGLS, as it has the same asymptotic efficiency and may be time-consuming and sensitive to problems in the nonlinear optimization algorithm.

An interesting theoretical result concerns residual specification testing in restricted VAR models. The degrees of freedom in the portmanteau Q have to be modified by subtracting the number of restrictions from the standard K^2h, while those for the LM test remain unchanged. LÜTKEPOHL warns anyway that these asymptotic distributions can be crude approximations in finite samples.

6 Vector error-correction models

In the second part of the book, the focus shifts from stationary processes to integrated processes. Integrated processes appear when some of the roots of the characteristic polynomial have a modulus of one and the remaining roots are outside the unit circle. While roots with a modulus of one but not equal to one—such as -1 or $\pm i$—are interesting in the analysis of seasonal data, LÜTKEPOHL restricts focus to the possibility of unit roots at one. It is advisable *not* to call such processes simply 'non-stationary', as there is an unlimited amount of possibilities for deviations from stationarity/stability. Integrated processes are only an important class of non-stationary processes.

6.1 Univariate integrated processes

Because the considered processes are not stable and their infinite-order MA representations do not converge, they will only be considered for the index set $\mathbb{N}$, that is, they must be started at $t = 0$, for example.

For univariate autoregressions with $K = 1$, the situation is quite simple. To an AR(p) process corresponds a polynomial of order p, instead of a determinant of a matrix polynomial. If it has d roots of one and all other roots larger than one, there are exactly $p - d$ 'nice and stable' roots. In this case, the process is said to be *integrated of order* d or $I(d)$.

The simplest $I(1)$ process is the *random walk*. An AR(1) process with a unit root has no other roots and can be written as

$$y_t = y_{t-1} + u_t = y_0 + \sum_{j=1}^{t} u_j.$$

It has the obvious properties

$$\mathrm{E}(y_t) = y_0$$

and

$$\mathrm{var}(y_t) = t\sigma_u^2,$$

if we set $\mathrm{var}(u_t) = \sigma_u^2$. It is easily shown that

$$\mathrm{corr}(y_t, y_{t+h}) = \frac{t}{\sqrt{t^2 + th}} \to 1,$$

as $t \to \infty$. This strong correlation among observations gives rise to the sometimes confusing name 'stochastic trend'. Generally, integrated processes are said to 'have a stochastic trend', which is not a trend line.

If the AR(1) equation has an intercept

$$y_t = \nu + y_{t-1} + u_t = y_0 + \nu t + \sum_{j=1}^{t} u_j,$$

the process is called a *random walk with drift.* We see that it is composed of a starting value, the deterministic linear time trend νt, and a 'stochastic trend'. Clearly,

$$\mathrm{E}(y_t) = y_0 + \nu t,$$

while the variance properties of the random walk remain unaffected.

All $I(d)$ processes have the property that, while not being stationary themselves, they become stable after taking first differences d times. This property may also serve as a definition for $I(d)$ in the sense that a process is called $I(d)$ if $\Delta^d y_t$ has a convergent MA representation but $\Delta^{d-1} y_t$ is not stable. This definition generalizes the previous one that was restricted to autoregressions. To rule out some notorious cases, LÜTKEPOHL demands that the MA representation converges in the sense of $\sum_{j=0}^{\infty} j|\theta_j| < \infty$, where θ_j are the Wold-type coefficients that are otherwise convened as Φ_j for $K < 1$.

At this point, an important ingredient of the derivation is the so-called *Beveridge-Nelson decomposition* for an $I(1)$ process. If Δy_t has an infinite-order MA representation with coefficients θ_j, one can show that

$$y_t = y_0 + \theta(1) \sum_{j=1}^{t} u_j + \sum_{j=0}^{\infty} \theta_j^* u_{t-j} - w_0^*,$$

where the first and the last term collect some starting value influence. The second term is a 'pure' random walk, while the third term is a stable/stationary component. $\theta(1)$ is simply $\sum_{j=0}^{\infty} \theta_j$, while the coefficients θ_j^* are defined via

$$\theta_j^* = -\sum_{k=j+1}^{\infty} \theta_k, j \geq 0.$$

In plain words, any $I(1)$ processes can be written as the sum of starting conditions, a pure random walk, and a stable/stationary component. In this part, the handling of the negatively indexed u_t is not always clean, as y_t need not be defined for $t < 0$. The decomposition is correct anyway.

Derivation: If $\Delta y_t = \theta(L) u_t$, then

$$y_t - y_0 = \theta(1) \sum_{j=1}^{t} u_j + (\theta(L) - \theta(1)) \sum_{j=1}^{t} u_j.$$

The first term is a random walk. For the second term, we consider the expansion

$$\theta(z) = \theta(1) + (z-1)\theta^*(z).$$

To determine $\theta^*(z)$, consider

$$\begin{aligned}\theta(z) - \theta(1) &= \sum_{j=0}^{\infty} \theta_j z^j - \sum_{j=0}^{\infty} \theta_j = \sum_{j=1}^{\infty} \theta_j (z^j - 1) \\ &= (z-1) \sum_{j=1}^{\infty} \theta_j \sum_{k=0}^{j-1} z^k = (z-1) \sum_{j=0}^{\infty} \theta_j^* z^j,\end{aligned}$$

where the last equality requires some further manipulation.□

6.2 VAR processes with integrated variables

We remember from linear algebra that the inverse of a matrix is the same as the product of the inverted determinant and the so-called *adjoint* of the matrix, in symbols

$$M^{-1} = |M|^{-1} M^{adj}.$$

This result can be applied to the formal representation of a VAR

$$\begin{aligned} A(L) y_t &= u_t, \\ |A(L)| y_t &= A(L)^{adj} u_t. \end{aligned}$$

If $A(L)$ is a polynomial matrix, then $A\left(L\right)^{adj}$ will also be a polynomial matrix, while its determinant is a scalar polynomial in the lag operator L. Thus, we obtain a representation with a scalar autoregressive polynomial on the left and a matrix moving-average of finite order on the right. If there are unit roots and integrated components, the scalar polynomial will show them. The real difficulty is, however, that even if all components of y are integrated of order d, $|A(z)|$ does not necessarily have the factor $(1-z)^{Kd}$. The unit root may have a lower multiplicity, and this is what defines *cointegration*.

Similarly, if the VAR model contains an intercept, the adjoint polynomial matrix will operate on that intercept and modify it in the ARMA–type representation.

6.3 Cointegration and error correction

Cointegrated variables have been much in the focus of economic interest since the 1980s. It is natural to view cointegration in a multivariate system framework, although the literature offers an alternative regression-based viewpoint also. The original definition of cointegration was not entirely system-based. According to this concept, a vector of variables is called *cointegrated* if all individual variables are $I(d)$ and there is a linear combination of the variables

$$\beta' y_t = \beta_1 y_{1t} + \ldots + \beta_K y_{Kt} = z_t,$$

such that z_t is integrated of order b or I(b), and $b < d$. The most interesting and most common application is $d = 1$ and $b = 0$, such that a linear combination of first-order integrated variables is stable/stationary. This type of cointegration is then sometimes denoted as $CI(d, b)$. Note that linear combinations are defined properly only if at least one coefficient of β is non-zero. In this concept, there must be at least *two* such non-zero coefficients. The vector β is called the *cointegrating vector*.

This basic concept is not quite sufficient for multivariate analysis, as it neglects the not uncommon issue of different integration orders among the variable vector components. Lütkepohl suggests to call a vector variable y_t integrated of order d when $\Delta^d y_t$ is stable/stationary but $\Delta^{d-1} y_t$ is not. Given this definition, he proceeds to defining cointegration $CI(d, b)$ if a linear combination exists such that the implied z_t is $I(b)$ with $b < d$. While integration orders are defined for the vector variable, cointegration is defined via the scalar variable z_t.

A simple example may serve to explain the issues at stake. According to the classical definition, a bivariate system consisting of a random walk and a stationary variable—for simplicity, assume the two components are independent—is *not* cointegrated. According to the Lütkepohl definition, it is cointegrated $CI(1, 1)$ with the cointegrating vector $(0, 1)'$. The second component is already the required z_t and is $I(0)$.

The concept of cointegration was introduced in a path-breaking article by Engle&Granger in 1987 under the title 'Cointegration and Error Correction'. The older concept of *error correction* denotes that a variable in differences Δy_t reacts to past changes but also to the difference between the past level y_t and another variable x_t. If $y - x$ is interpreted as a kind of 'equilibrium' term—or rather disequilibrium term—with a long-run constant mean or 'equilibrium' $\mathrm{E}\,(y - x)$, such a reaction can be viewed as Δy_t reacting to deviations from $\mathrm{E}\,(y - x)$ in the levels and therefore as 'error correction'. Part of the literature suggested replacing 'error correction' by 'equilibrium correction' but Lütkepohl sticks to the more widely used terminology. In a nutshell, Engle&Granger showed that the new concept of cointegration is equivalent to the known concept of error correction.

It is of more concern to the main issue that Engle&Granger also showed that a $CI(1, 1)$ cointegrated VAR(p) can always be written in the so-called *vector error-correction model* (VECM) representation

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

where the matrix $\mathbf{\Pi}$ has rank $r < K$. It is easy to show that there is indeed a simple way to re-write any VAR(p) that has been given in the original form

$$y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t$$

as a VECM and to derive the one-one relations of the two representations. We have

$$\begin{array}{rcl} \mathbf{\Pi} & = & -I_K + A_1 + \ldots + A_p, \\ \mathbf{\Gamma}_i & = & -(A_{i+1} + \ldots + A_p), \quad i = 1, \ldots, p-1, \end{array}$$

and, in the other direction,

$$\begin{array}{rcl} A_1 & = & I_K + \mathbf{\Pi} + \mathbf{\Gamma}_1, \\ A_j & = & \mathbf{\Gamma}_j - \mathbf{\Gamma}_{j-1}, \quad 2 \leq j \leq p-1, \\ A_p & = & -\mathbf{\Gamma}_{p-1}. \end{array}$$

It is less straight forward to show that $\mathbf{\Pi}$ must have reduced rank for the $CI(1,1)$ case. Clearly, if y_t has a unit root, then by definition **Pi** must be singular, as the characteristic polynomial matrix $A(1)$ is singular for $z = 1$ and is identical to $-$**Pi**.

If the matrix $\mathbf{\Pi}$ has reduced rank, we have the result from linear algebra that it can be represented in the form $\mathbf{\Pi} = \alpha\beta'$ with $(K \times r)$–matrices α and β of 'full' lower rank r. The representation is not unique, as any nonsingular $(r \times r)$–matrix M can be used to achieve another representation by observing

$$\mathbf{\Pi} = \alpha\beta' = \alpha M M^{-1} \beta' = \alpha^* \beta^{*\prime}$$

for $\alpha^* = \alpha M$ and $\beta^* = \beta M^{-1\prime}$. This fact is of crucial importance here, as β can be shown to contain the cointegrating vectors, while α contains the coefficients that determine how Δy_t reacts to deviations from the mean of the cointegrating variables $\beta' y_{t-1}$. Note that this mean must be zero here, as the VAR contains no constants. There exist various names for the matrix α and its column vectors. LÜTKEPOHL prefers to call them the *loading matrix* and the *loading vectors*. We note that neither the loading matrix nor the cointegrating vectors can be unique. The rank r, however, is unaffected by the manipulations.

The main properties are summarized in the famous *Granger Representation Theorem*:

If y_t is a K–dimensional $I(1)$ VAR(p) process that is $CI(1,1)$ with the representation

$$\Delta y_t = \alpha\beta' y_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p} + u_t,$$

with all values set at zero for $t \leq 0$, and that the determinant of the characteristic polynomial has only stable roots and exactly $K - r$ roots equal one,while α and β are full-rank matrices of dimension $K \times r$. Then, y_t has the representation

$$y_t = \mathbf{\Xi} \sum_{j=1}^{t} u_j + \mathbf{\Xi}^* (L)\, u_t + y_0^*,$$

where

$$\mathbf{\Xi} = \beta_\perp \left\{ \alpha'_\perp \left(I_K - \sum_{j=1}^{p-1} \mathbf{\Gamma}_j \right) \beta_\perp \right\}^{-1} \alpha'_\perp;$$

$\mathbf{\Xi}^* (L)\, u_t$ is $I\,(0)$, and y_0^* contains some starting conditions.

The subscript $\perp$ is used to denote the *orthogonal complement* of a matrix, that is the matrix that completes a rectangular matrix to a quadratic one by adding some vectors that are orthogonal to the existing ones. In short, the theorem yields a multivariate counterpart to the Beveridge-Nelson decomposition by re-writing the multivariate y_t process as the sum of a reduced-rank random walk, a stationary process, and some starting conditions.

The literature offers several versions of the theorem's proof. The one by Lütkepohl is probably not the most accessible one. The first component can be viewed as the *common trends* in the system. If the matrix $\mathbf{\Pi}$ has full rank, the orthogonal complements are empty and there are no trends and also no common ones. This case is excluded by assuming y_t to be $I\,(1)$. If the matrix in curly brackets is singular and cannot be inverted, it can be shown that there is integration of higher order. Also this case is excluded.

6.4 Deterministic terms in cointegrated processes

In the previous subsection, constants and similar terms were deliberately omitted. There are two ways to introduce such terms. Either one considers them as intercepts

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

possibly adding even a trend term to the intercept if one wishes to do so; or one may consider

$$y_t = \mu_t + x_t,$$

where μ_t contains everything deterministic and x_t corresponds to the process of the last subsection. LÜTKEPOHL considers the latter variant. It is uncertain whether this is a very wise decision. While the two viewpoints must result in equivalent models, the first one may yield a more intuitive interpretation. For a good exposition of the former viewpoint, see JOHANSEN.

If $\mu_t = \mu_0$, a simple constant, its first differences are 0 and one may consider $x_t = y_t - \mu_0$ and insert in the non-deterministic VECM

$$\Delta y_t = \alpha\beta' \left(y_{t-1} - \mu_0\right) + \mathbf{\Gamma}_1 \Delta y_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t.$$

To achieve a more compact form, one may extend β by another row that contains $-\beta'\mu_0$ and extend y_{t-1} by another element containing a simple 1. These constructs can be named β^o and y_{t-1}^o, which yields

$$\Delta y_t = \alpha\beta^{o\prime} y_{t-1}^o + \mathbf{\Gamma}_1 \Delta y_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t.$$

The drawback is that β^o does no more contain cointegrating vectors but cointegrating vectors plus a constant. This may lead us to believe that the last row is necessary to make $\beta' y$ stationary, which of course is not the case.

In this first case, the above Granger Representation Theorem continues to hold. Still, y_t is the sum of a reduced-rank random walk, a stationary process, and some starting conditions. The stationary process does have a non-zero mean now. In most of the literature, this case is called the 'restricted constant' or the 'no-trend restriction'. It may be an appropriate framework for the analysis of variables that are known to be integrated but not systematically trending, such as interest rates.

If $\mu_t = \mu_0 + \mu_1 t$, a linear time trend, first differences are a constant μ_1. One may consider $x_t = y_t - \mu_0 - \mu_1 t$ and again insert into the VECM

$$\begin{aligned} \Delta y_t &= \mu_0 + \alpha\beta' \left\{y_{t-1} - \mu_0 - \mu_1 (t-1)\right\} + \mathbf{\Gamma}_1 \left(\Delta y_{t-1} - \mu_1\right) + \ldots & (1) \\ & \quad + \mathbf{\Gamma}_{p-1} \left(\Delta y_{t-p+1} - \mu_1\right) + u_t. & (2) \end{aligned}$$

All deterministic terms except those in the curly bracket can be collected to yield

$$\Delta y_t = \nu + \alpha\beta' \left\{y_{t-1} - \mu_0 - \mu_1 (t-1)\right\} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t,$$

where ν is a function of μ_0, μ_1, and all $\mathbf{\Gamma}_j$ matrices. Again, one may extend β by two elements $-\mu_0$ and $-\mu_1$ and extend y_t by a constant 1 and a trend term t. This yields formally

$$\Delta y_t = \nu + \alpha\beta^{+\prime} y_{t-1}^+ + \mathbf{\Gamma}_1 \Delta y_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t.$$

The interpretation of this model, however, is even more difficult, as it appears that a trend term is required to make y stationary. If this is indeed the case, $\beta' y$ can hardly be called stationary, as it is trending, and the idea of cointegration becomes pretty weak. Only if the loading matrix α achieves a deletion of the contribution of the linear trend term in $\beta^{+\prime} y^{+}$, can the basic concept of cointegration be saved. If this deletion occurs, the system can be written as

$$\Delta y_t = \nu + \alpha \beta^{o\prime} y^{o}_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \ldots + \mathbf{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t.$$

This is the specification of central interest in empirical cointegration research. In the notation of LÜTKEPOHL, it occurs as a special restriction, while in the intercept notation it evolves naturally. The Granger Representation Theorem is no more valid in its above form but a multivariate trend term must be added to the Beveridge-Nelson–type decomposition. y_t becomes the sum of a rank-deficient multivariate random walk, a stationary process, a trend term, and some starting conditions. Alternatively, one may consider drifting random walks instead of the non-deterministic forms and achieves a three-component representation again.

6.5 Granger causality in cointegrated VARs

In theory, the differences to the stable case are not very pronounced. As outlined for the stable case, Granger causality can be checked either on the VAR or on the infinite MA representation.

With respect to the VAR representation, one may also consider the VECM form, where it is to be noted that dynamic influences from a subvector to another subvector can run via the $\mathbf{\Gamma}_j$ matrices or via $\mathbf{\Pi} = \alpha \beta'$. In the first case, causal effects should fade out as $t \to \infty$, while the latter effects may persist. It is important to check both sets of coefficients in order to obtain reliable results on Granger non-causality.

With respect to the infinite MA representation, one should note that it literally does not exist, as it does not converge for unstable systems. Nevertheless, finite partial sums do exist and should reflect all causal events. Similarly, the errors covariance matrix may be used for investigating instantaneous causality.

This good news, however, is to be contrasted with severe difficulties in empirical inference, as the asymptotic χ^2 laws for relevant test statistics may become incorrect. In anticipation of Section 7 of LÜTKEPOHL's book, one may recommend that the main trick is to 'augment' the previously identified VAR(p) by one lag, such that it becomes a VAR($p+1$). The lags at $p+1$

do not need testing, as they are assumedly zero anyway. Tests on the lags 1 to p in the estimated VAR$(p+1)$ are valid Granger causality tests with the asymptotically correct degrees of freedom.

6.6 Impulse responses in cointegrated VARs

In analogy to the Granger causality tests, impulse response analysis can be conducted by simply generalizing the procedure known from the analysis of stable systems. The MA representation does not converge, as the Φ_j matrices do not converge to zero. Therefore, impulse response functions do not converge to 0, as $t \to \infty$. For $I(1)$ processes, they typically converge to a fixed and finite value that represents the long-run impact of a shock. Empirical studies of GDP data for various countries have resulted in values that are scattered around one, which would be the typical value for a pure random walk. For processes integrated of higher order than one, also these limits will diverge.

The $CI(1,1)$ cointegrated variables $\beta' y_t$ should behave like variables in a stable VAR. Their impulse responses approach zero as $t \to \infty$, and their accumulated impulse response sequences approach a finite value. Theses properties should be visible from transforming the original impulse responses for the cointegrated system by β'. For example, if $y_1 - y_2$ is stationary in a bivariate VAR, the difference of the individual impulse response functions must converge to zero.

Even when impulse responses do not converge, the forecast error variance decomposition statistics do, as they evolve from bounded fractions.