

Vector autoregression models

by Kevin Kotzé

Vector Autoregressive (VAR) models are widely used in time series research to examine the dynamic relationships that exist between variables that interact with one another. In addition, they are also important forecasting tools that are used by most macroeconomic or policy-making institutions.¹

The development of these models is the subject of much ongoing research that follows the seminal contribution of Sims (1972), which established a framework for modelling endogenous variables in a multivariate setting. Following the development of this framework, a number of statistical tests were derived to determine inter-dependencies and the nature of the dynamic relationships between variables. In addition, more recent developments include the use of structural decompositions, sign restrictions, the incorporation of time-varying parameters, structural breaks, stochastic volatility; to mention but a few.

The structure of VAR models enables one to explain the values of endogenous variables from their past observed values.² These models differ somewhat to structural vector autoregressive (SVAR) models, which allow for the explicit modelling of contemporaneous interdependence between the left-hand side variables. In addition, important contributions by Engle and Granger (1987) endowed econometricians with powerful tools for modelling cointegrated relationships between variables.

In what follows we introduce some of the key ideas and methods used in VAR analysis. Many of the concepts discussed in the current chapter are multivariate extensions of the tools and concepts that were applied to autoregressive models. These techniques are explained with the use matrices and matrix algebra.³ We also consider the stability properties of the model and show how one can derive the moving average representation of the VAR. Thereafter we discuss the issues related to specification, estimation and forecasting of the reduced-form VAR models. In the final section we introduce the important concept called Granger causality.⁴

1 The Reduced-Form VAR model

1.1 Notation

To describe the use of multivariate techniques, we need to introduce new notation. In its basic form, a VAR consists of a set of K endogenous variables $\mathbf{y}_t = (y_{1t}, \dots, y_{kt}, \dots, y_{Kt})$ for $k = 1, \dots, K$. After including p lags of the endogenous variables, the VAR(p) model may be defined as:

$$\mathbf{y}_t = A_1 \mathbf{y}_{t-1} + \dots + A_p \mathbf{y}_{t-p} + C D_t + \mathbf{u}_t \quad (1.1)$$

where A_i are $(K \times K)$ coefficient matrices for $i = 1, \dots, p$ and \mathbf{u}_t is a K -dimensional white noise process with time-invariant positive definite covariance matrix,⁵ $\mathbb{E} [\mathbf{u}_t \mathbf{u}_t'] = \Sigma_u$, such that

$$\begin{aligned} \mathbb{E} [\mathbf{u}_t] &= 0 \\ \mathbb{E} [\mathbf{u}_t \mathbf{u}_t'] &= \Sigma_u \text{ which is positive definite} \end{aligned}$$

The matrix C is the coefficient matrix of potentially deterministic regressors with dimension $(K \times M)$, and D_t is an $(M \times 1)$ column vector holding the appropriate deterministic regressors, such as a constant, trend, and dummy and/or seasonal dummy variables.

The expression in equation (1.1) clearly suggests that the VAR model relates the k 'th variable in vector \mathbf{y}_t to past values of itself and all other variables in the system. This is in contrast with the univariate autoregressive models, which only relate a given variable to past values of itself.

To consider an example, let us assume that there are no deterministic regressors, $K = 2$ and $P = 1$. Then equation (1.1) would simplify to that of a VAR(1) model that does not include any deterministic components,

$$\mathbf{y}_t = A_1 \mathbf{y}_{t-1} + \mathbf{u}_t$$

where \mathbf{y}_t , \mathbf{u}_t and the A_1 matrices that take the form

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix}, A_1 = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \text{ and } \mathbf{u}_t = \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \quad (1.2)$$

If we are to then assume that the specific elements of the A_1 matrix are given as,

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 1 & 0.2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \quad (1.3)$$

where after some matrix multiplications, we can write

$$\begin{aligned} y_{1,t} &= 0.5y_{1,t-1} + u_{1,t} \\ y_{2,t} &= 1y_{1,t-1} + 0.2y_{2,t-1} + u_{2,t} \end{aligned}$$

With this parametrisation, $y_{2,t}$ depends positively on past values of itself and on past values of $y_{1,t}$. On the other hand, the dynamics of $y_{1,t}$ depend on past values of itself only. These relationships are captured by the particular parametrisation of the model.

The variables that are to be included in a particular investigation will typically depend on the purpose of the study, where researchers may wish to incorporate those that are responsible for important dynamic interactions or a perceived causal relationships. In addition, economic theory could also be used to good effect when deciding upon the variables that are to be included. In many macroeconomic investigations it would be prudent to construct a model that does not only depend on past values of itself, but also on the past values of other variables. Hence, many VAR models typically include lags of all endogenous variables.

1.2 The companion-form

It is possible to express a VAR(p) model in a VAR(1) form, which turns out to be very useful in many practical situations (and for some technical derivations). This reformulation is done by expressing the VAR(p) model in the companion-form, which is accomplished as follows. Consider the model,

$$\mathbf{Z}_t = \Gamma_0 + \Gamma_1 \mathbf{Z}_{t-1} + \Upsilon_t$$

where we have defined

$$Z_t = \begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix}, \quad \Gamma_0 = \begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \Upsilon_t = \begin{bmatrix} \mathbf{u}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

So that the matrix notation is

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \mathbf{y}_{t-3} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The dimensions of the vectors Z_t , Γ_0 and Υ_t are of dimension $Kp \times 1$. The coefficient matrix, A_j for $j = 1, \dots, p$ will be of dimension $K \times K$, and Γ_1 is $Kp \times Kp$. In this case, Γ_1 is called the companion-form matrix.⁶

1.3 Stability in a VAR model

As in the case of the univariate autoregressive model, the VAR model is covariance-stationary when the effect of the shocks, \mathbf{u}_t , dissipate over time. This will be the case if the eigenvalues of the companion-form matrix are all less than one in absolute value. The eigenvalues of the matrix Γ_1 are represented by λ in the expression,

$$|\Gamma_1 - \lambda I| = 0 \quad (1.4)$$

To show how one could derive these eigenvalues, we make use of the simple VAR(1) example that we employed in (1.2), where the A_1 coefficient matrix is represented by what is contained in (1.3). In this case, $\Gamma_1 = A_1$, which may be inserted into expression (1.4) to provide,

$$\det \left[\begin{bmatrix} 0.5 & 0 \\ 1 & 0.2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right] = \det \left[\begin{bmatrix} 0.5 - \lambda & 0 \\ 1 & 0.2 - \lambda \end{bmatrix} \right] (0.5 - \lambda)(0.2 - \lambda) = 0$$

Such that,

$$\lambda_1 = 0.5, \lambda_2 = 0.2$$

Hence, the eigenvalues are $\lambda_1 = 0.5$ and $\lambda_2 = 0.2$, which are both less than 1, which would suggest that this VAR(1) process is stable (i.e. covariance-stationary). In the more general cases of VAR(p) models, the eigenvalues of Γ_1 may be easily computed with the aid of most standard software.

Some researchers consider the stability of a model by making use of the definition for *characteristic roots* rather than eigenvalues. In such cases, we could find the roots, z , of a matrix Γ_1 , from

$$|I - \Gamma_1 z|$$

When using this formulation, the interpretation is reversed, such that one would suggest that the VAR model is stable when the roots of the polynomial, $\det(I - \Gamma_1 z)$, lie outside of the unit circle (i.e. are bigger than one in absolute value).⁷ In what follows, we will seek to refer to the eigenvalues when investigating the stability of the variables in the model.

To simulate processes that may be described by a stable VAR(1) model, we could make use of equation (1.3), with $y_{k,0} = 0$, $\mu_k = 1$ for $k = [1, 2]$ and

$$\mathbf{u}_t \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} \right)$$

The result of this simulation is shown in Figure 1. As the process is stable by definition, both series $y_{1,t}$ and $y_{2,t}$ fluctuate around a constant mean. Note also that their variability does not change over time.

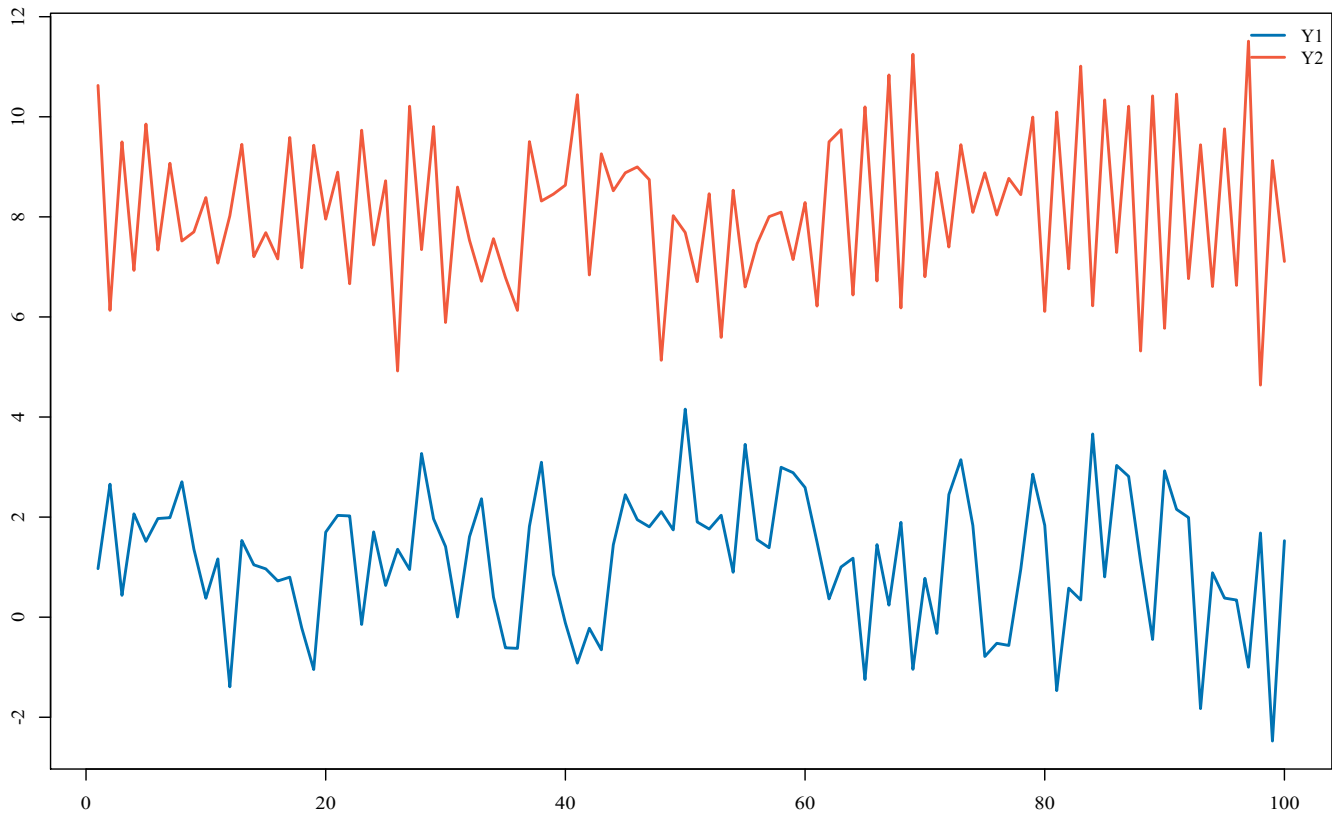


Figure 1: **Simulated VAR processes**

1.4 Wold decomposition

Just as the stable AR(p) model has a MA representation, the stable VAR(p) has a VMA representation. This proposition is provided by the Wold (1938) decomposition theorem, which states that every covariance-stationary time series y_t can be written as the sum of two uncorrelated processes, a deterministic component (which could be the mean) κ_t and a component that has an infinite moving average representation of $\sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$

$$y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} + \kappa_t \quad (1.5)$$

where we assume that $\theta_0 = 1$, $\sum_{j=0}^{\infty} |\theta_j| < \infty$ and the term ε_t is Gaussian white noise.

Finding the Wold representation involves fitting an infinite number of parameters $(\theta_1, \theta_2, \theta_3, \dots)$ to the data. With a finite number of observations, this is never possible. However, one can approximate an infinite order $\theta(L)$, by using models that have a finite number of parameters.

We already have implicitly used the Wold's decomposition theorem when deriving the moving average form of an AR(p) model. In what follows we will now derive the moving average form of a stable VAR(p).

2 Moving average representation of a VAR

If we assume that the VAR(p) model is a stable, we can derive its infinite moving average representation by using either recursive substitution, or lag operators. Since we can easily write any VAR(p) model as a VAR(1), with the aid of the companion-form, we will make use of the VAR model that was formulated in equation (1.1). Hence, when given

$$\mathbf{y}_t = \mu + A_1 \mathbf{y}_{t-1} + \mathbf{u}_t \quad (2.1)$$

and using recursive substitution, we can show that,

$$\begin{aligned} \mathbf{y}_t &= (1 + A_1 + A_1^2 + \dots + A_1^j) \mu + A_1^{j+1} \mathbf{y}_{t-(j+1)} \\ &\quad + (\mathbf{u}_t + A_1 \mathbf{u}_{t-1} + A_1^2 \mathbf{u}_{t-2} + \dots + A_1^j \mathbf{u}_{t-j}) \end{aligned} \quad (2.2)$$

Where $A_1^0 = I$. When the process is stable, $(I + A_1 + \dots + A_1^j) \mu \rightarrow (I - A_1)^{-1} \mu$ as $j \rightarrow \infty$. In addition, as $A_1^{j+1} \mathbf{y}_{t-j-1} \rightarrow 0$ it will be the case that $j \rightarrow \infty$. Hence, equation (2.2) reduces to,

$$\mathbf{y}_t = \varphi + \sum_{j=0}^{\infty} A_1^j \mathbf{u}_{t-j} \quad (2.3)$$

Where $\varphi = (I - A_1)^{-1} \mu$. Note that equation (2.3) is called the moving average representation of the VAR. This could be written in terms of the moving average coefficients as,

$$\mathbf{y}_t = \varphi + \sum_{j=0}^{\infty} B_j \mathbf{u}_{t-j}$$

where $B_j = A_1^j$, and $B_0 = I$.

The infinite moving average representation of the VAR(p) model can also be derived by employing the lag operator. To accomplish this objective, note that equation (2.1) could be written as,

$$(I - A_1 L) \mathbf{y}_t = \mu + \mathbf{u}_t$$

where $(I - A_1 L) = A(L)$, this expression could be written more conveniently as,

$$A(L) \mathbf{y}_t = \mu + \mathbf{u}_t$$

After multiplying through by $A(L)^{-1}$ we get the vector moving average representation,

$$\begin{aligned} \mathbf{y}_t &= A(L)^{-1} \mu + A(L)^{-1} \mathbf{u}_t \\ &= B(L) \mu + B(L) \mathbf{u}_t \\ &= \varphi + \sum_{j=0}^{\infty} B_j \mathbf{u}_{t-j} \end{aligned} \quad (2.4)$$

where we have used the geometric rule.

$$\begin{aligned}
A(L)^{-1} &= (I - A_1 L)^{-1} = \sum_{j=0}^p A_1^j L^j \\
&\equiv B(L) = \sum_{j=0}^{\infty} B_j L^j
\end{aligned} \tag{2.5}$$

with $B_0 = I$ and $\varphi = \left(\sum_{j=0}^{\infty} B_j \right) \mu$.

2.1 Finding the MA coefficients

To identify the moving average coefficients, B_j , we could make use of the relationship $I = B(L)A(L)$, which would hold due to the result in (2.5).⁸ This allows for the following derivation,

$$\begin{aligned}
I &= B(L)A(L) \\
I &= (B_0 + B_1 L + B_2 L^2 + \dots)(I - A_1 L - A_2 L^2 - \dots - A_p L^p) \\
&= [B_0 + B_1 L + B_2 L^2 + \dots] \\
&\quad - [B_0 A_1 L + B_1 A_1 L^2 + B_2 A_1 L^3 + \dots] \\
&\quad - [B_0 A_2 L^2 + B_1 A_2 L^3 + B_2 A_2 L^4 + \dots] - \dots \\
&= B_0 + (B_1 - B_0 A_1)L + (B_2 - B_1 A_1 - B_0 A_2)L^2 + \dots \\
&\quad + \left(B_p - \sum_{j=1}^p B_{p-j} A_j \right) L^p + \dots
\end{aligned}$$

Solving for the relevant lags (noting that $A_1 = 0$ for $j > p$), we get,

$$\begin{aligned}
B_0 &= I \\
B_1 &= B_0 A_1 \\
B_2 &= B_1 A_1 + B_0 A_2 \\
&\vdots \\
B_i &= \sum_{j=1}^i B_{i-j} A_j \quad \text{for } i = 1, 2, \dots
\end{aligned} \tag{2.6}$$

Hence, the B_j 's can be computed recursively from (2.6). To show how this may be applied in a practical example, consider the simple VAR(1) example in equation (2.1),

$$\mathbf{y}_t = \mu + A_1 \mathbf{y}_{t-1} + \mathbf{u}_t$$

Since we know from (2.6) that the coefficients will take the form,

$$\begin{aligned}
B_0 &= I \\
B_1 &= B_0 A_1 = A_1 \\
B_2 &= B_1 A_1 + B_0 A_2 = A_1^2 \\
&\vdots \\
B_p &= A_1^p
\end{aligned}$$

We are then able to plug in the known parameters from (1.3), to get the moving average coefficients,

$$\begin{aligned}
B_0 &= I \\
B_1 &= A_1 = \begin{bmatrix} 0.5 & 0 \\ 1 & 0.2 \end{bmatrix} \\
B_2 &= A_1^2 = \begin{bmatrix} 0.25 & 0 \\ 0.70 & 0.04 \end{bmatrix} \\
&\vdots \quad \vdots
\end{aligned}$$

In this way, B_i approaches zero as $p \rightarrow \infty$, which would arise if the processes in the VAR(p) model are stable (i.e. they do not have infinite memory).

2.2 Mean, variance and autocovariance of a VAR

Just as in the univariate case, the first-and second-order moments of the VAR(p) can easily be derived from the moving average representation in equation (2.4). Hence, given

$$\mathbf{y}_t = \varphi + \sum_{j=0}^{\infty} B_j \mathbf{u}_{t-j}$$

with $\varphi = \left(\sum_{j=0}^{\infty} B_j \right) \mu = (I - A_1)^{-1} \mu$. Firstly note that, due to Gaussian white noise assumption about the error terms, the expected mean value of the process is simply,

$$\mathbb{E}[\mathbf{y}_t] = \varphi = (I - A_1)^{-1} \mu$$

Sometimes this mean is described as the steady-state of the system. We will denote the covariance and autocovariances of the process by Ψ , which will be derived with the aid of Yule-Walker equations.

To complete this procedure we firstly write the VAR(p) in the mean-adjusted form, where we subtract the expected value from both sides of the equality sign,

$$\mathbf{y}_t - \varphi = A_1(\mathbf{y}_{t-1} - \varphi) + \mathbf{u}_t$$

Postmultiplying by $(\mathbf{y}_{t-s} - \varphi)'$ and taking the expectation gives,

$$\begin{aligned}
\mathbb{E} \left[(\mathbf{y}_t - \varphi) (\mathbf{y}_{t-s} - \varphi)' \right] &= A_1 \mathbb{E} \left[(\mathbf{y}_{t-1} - \varphi) (\mathbf{y}_{t-s} - \varphi)' \right] \\
&\quad + \mathbb{E} \left[\mathbf{u}_t (\mathbf{y}_{t-s} - \varphi)' \right]
\end{aligned}$$

Thus for $s = 0$,

$$\Psi_s = A_1 \Psi_{s-1} + \Sigma_u = A_1 \Psi'_1 + \Sigma_u \quad (2.7)$$

where after the second equality sign, we have used the fact that $\Psi_{-1} = \Psi'_1$. For $s > 0$ we would then have,

$$\Psi_s = A_1 \Psi_{s-1} \quad (2.8)$$

When A_1 and Σ_u are known, we can compute the autocovariance for $s = 0, \dots, S$ using equations (2.7) and (2.8). Hence, for $s = 1$, we use (2.8) to show that $\Psi_1 = A_1 \Psi_0$. Substituting this into equation (2.7) and noting that $(A_1 \Psi_0)' = \Psi'_0 A'_1$ by the rules of matrix algebra, we get,

$$\Psi_0 = A_1 \Psi_0 A_1' + \Sigma_u$$

This equation can be solved for Ψ_0 , by using the definition of the Kronecker product and the *vec* operator to derive,⁹

$$\text{vec}\Psi_0 = (I - A_1 \otimes A_1)^{-1} \text{vec}\Sigma_u$$

where *vec* denotes the column stacking operator that stacks the columns of a matrix in a column vector. It follows that once Ψ_0 has been derived in this manner, we can use equation (2.8) recursively to get the autocovariance for $s > 0$.

To get the autocorrelation function, we would then need to normalise the autocovariances such that they have ones on the diagonal at $s = 0$. Thus, we define the diagonal matrix, ϑ , whose diagonal elements are the square roots of the diagonal elements of Ψ_0 . Then the autocorrelation function for the VAR(p) is simply,

$$R_s = \vartheta^{-1} \Psi_h \vartheta^{-1}$$

To keep the notations as simple as possible, the above derivation assumed that the underlying model was a VAR(1). However, as noted earlier, a VAR(p) can always be written as a VAR(1) using companion-form of matrix. For the VAR(1), the companion-form matrix is of course equal to the A_1 matrix, and nothing in the previous derivations changes. For higher order VAR(p) models the companion-form, all of the expressions for the moving average representation, expected value and autocovariances do not change; but the size of the matrices will of course change. In addition, we would also need to make use a selection matrix to extract the values of interest. For a more detailed explanation of how this is done, see Lutkepohl (2005).

3 Estimating VAR parameters

The VAR system can be estimated equation-by equation using OLS. Estimating the VAR(p) using OLS would be consistent, and under the assumption of normality of the errors, efficient.¹⁰ More formally, we could start by assuming that we have a sample size of T , with $\mathbf{y}_1, \dots, \mathbf{y}_T$, for each of the K variables in our VAR(p) model, which may be defined in equation (1.1).

Following Lutkepohl (2005), we define $Y = [\mathbf{y}_1, \dots, \mathbf{y}_T]$, $A = [A_1, \dots, A_p]$, $U = [\mathbf{u}_1, \dots, \mathbf{u}_T]$ and $Z = [Z_0, \dots, Z_{T-1}]$, where:

$$Z_{t-1} = \begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix}$$

The model described in equation (1.1) could then be written as,

$$Y = AZ + U$$

And the OLS estimate of A is defined as,

$$\hat{A} = [\hat{A}_1, \dots, \hat{A}_p] = YZ'(ZZ')^{-1}$$

Under standard assumptions the OLS estimator \hat{A} is consistent and asymptotically normally distributed,

$$\sqrt{T}vec(\hat{A} - A) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_e) \quad (3.1)$$

where \xrightarrow{d} implies convergence in distribution and $ZZ'/T \xrightarrow{d} \Gamma$. Once again, vec denotes the column stacking operator that stacks the columns of a matrix in a column vector.

When interpreting the estimated VAR parameters, standard inference would apply. In addition, when there are no parameter restrictions that are employed on the VAR system, and for a normally distributed Gaussian stationary process, the OLS estimator defined that is defined in equation (3.1) is identical to the maximum likelihood (ML) estimator (conditional on the initial values).

3.1 Choice of variables and lags

When deciding on the variables and lags that should be included in the model, it is important to note that VAR can become heavily parameterised. This could result in problems where we have too many parameters, relative to observations in the data (i.e. degrees of freedom problem). For example, after including 6 variables in a model with 8 lags (i.e. two years of lags for quarterly data), then each equation in the VAR will contain $(6 \times 8)+1=49$ coefficients (including a constant). In many macroeconomic or financial investigations, we often find ourself in situations with limited observations, which would imply that our parameter estimates are imprecise (at best). Therefore, in most cases, including much more than 6 variables in the VAR that is estimated with classical techniques could result in several difficulties.

Therefore, since one cannot include all variables of potential interest, one has to make a careful choice by referring to economic theory or *a priori* knowledge when choosing variables. Having specified the model, the appropriate lag length of the VAR model also has to be chosen. As in the univariate case, one can use information criterion for this purpose (which would include the BIC or AIC).

4 VAR forecasts

VAR models are extremely popular tools forecasting the future values of variables. Although we have already derived expressions for the conditional expectations that were used to construct forecasts for univariate models, we repeat these formulas, where we use appropriate multivariate notation.

4.1 Point forecast

To generate the h -step ahead forecast, we could iterate the VAR model forward by a number of observations. Replacing the notation used for the univariate case with the VAR notation, the multivariate expression is,

$$\mathbf{y}_{t+h} = A_1^h \mathbf{y}_t + \sum_{i=0}^{h-1} A_1^i \mathbf{u}_{t+h-i} \quad (4.1)$$

While this forecasting framework has been described within the context of a VAR(1) model, we could express the VAR(p) model as a VAR(1) by using the companion-form, where the formula would remain largely unchanged. That is, $Z_{t+h} = \Gamma_1^h Z_t + \sum_{i=0}^{h-1} \Gamma_1^i \Upsilon_{t+h-i}$. Using the selector matrix

$\Xi = [I, 0, \dots, 0]$, which is of $K \times (Kp)$ dimension, the appropriate forecasts for \mathbf{y}_{t+h} can be extracted (i.e. $\mathbf{y}_{t+h} = \Xi \mathbf{Z}_{t+h}$). Moreover, adding a constant to the forecasting relationship is simply a multivariate extension of the univariate case. Therefore, continuing with the VAR(1) example, but adding a constant to the VAR, results in the following h -step ahead forecast formula,

$$\mathbf{y}_{t+h} = (I + A_1 + \dots A_1^{h-1})\mu + A_1^h \mathbf{y}_t + \sum_{i=0}^{h-1} A_1^i \mathbf{u}_{t+h-i} \quad (4.2)$$

By employing the conditional expectation to (4.2), we get the VAR point forecast,

$$\mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t] = (I + A_1 + \dots + A_1^{h-1})\mu + A_1^h \mathbf{y}_t$$

Which under the maintained assumption about stationarity, converges to the unconditional mean of the process when h goes to infinity,

$$\mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t] = \frac{\mu}{I - A_1} \quad \text{when } h \rightarrow \infty$$

Once again, these equations are just the multivariate extensions of the formulas that were derived previously.

4.2 Mean Squared Forecast Errors

The Mean Squared Forecast Errors (MSFE) are derived from the forecast function in equation (4.2). As before, the VAR forecast error at horizon h would be given as,

$$\mathbf{y}_{t+h} - \mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t] = \sum_{i=0}^{h-1} A_1^i \mathbf{u}_{t+h-i}$$

The expectation of this expression seeks to describe the expected forecast error. If the current assumptions about $\mathbb{E} [\mathbf{u}_t]$, are expected to hold in the future then the expected mean value for this expression is zero. Hence,

$$\mathbb{E} [\mathbf{y}_{t+h} - \mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t]] = \mathbb{E} [\mathbf{y}_{t+h}] - \mathbb{E} [\mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t]] = 0$$

Thus the predictions for $\mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t]$ would be unbiased when the sum of the errors is equal to zero. The MSFE is simply the forecast error variance. Therefore, in the multivariate VAR setting the MSFE, which we denote, σ_{t+h}^f , may be given as

$$\begin{aligned} \sigma_{t+h}^f &= \mathbb{E} \left[(\mathbf{y}_{t+h} - \mathbb{E} [\mathbf{y}_{t+h} | \mathbf{y}_t])^2 \right] \\ &= \mathbb{E} \left[\left(\sum_{i=0}^{h-1} A_1^i \mathbf{u}_{t+h-i} \right) \left(\sum_{i=0}^{h-1} A_1^i \mathbf{u}_{t+h-i} \right)' \right] \end{aligned}$$

By noting that the A_1 terms are constant, which would allow us to move them outside of the expectation, and that $\mathbb{E}(\mathbf{u}_{t+h-i}, \mathbf{u}_{t+h-i}') = \Sigma_u$ for all h , we are able to derive

$$\sigma_{t+h}^f = \sum_{i=0}^{h-1} \left(A_1^i \Sigma_u A_1^{i'} \right) \quad (4.3)$$

Given this expression, it is worth noting that there is an interesting relationship between the MSFE and the moving average form of the VAR. If we expand equation (2.4), and move it h -periods forward we get,

$$\mathbf{y}_{t+h} = \varphi + \mathbf{u}_{t+h} + B_1 \mathbf{u}_{t+h-1} + B_2 \mathbf{u}_{t+h-2} + \dots + B_h \mathbf{u}_t + B_{h+1} \mathbf{u}_{t-1} + \dots$$

Employing the conditional expectation to this expression, and computing the forecast error gives,

$$\mathbb{E}[\mathbf{y}_{t+h}|t] = \varphi + B_h \mathbf{u}_t + B_{h+1} \mathbf{u}_{t-1} + B_{h+2} \mathbf{u}_{t-2} + \dots$$

and

$$\mathbf{y}_{t+h} - \mathbb{E}[\mathbf{y}_{t+h}|t] = \mathbf{u}_{t+h} + B_1 \mathbf{u}_{t+h-1} + \dots + B_{h-1} \mathbf{u}_{t+1}$$

Which has variance equal to,

$$\sigma_{t+h}^f = \mathbb{E} \left[(\mathbf{y}_{t+h} - \mathbb{E}[\mathbf{y}_{t+h}|t]) (\mathbf{y}_{t+h} - \mathbb{E}[\mathbf{y}_{t+h}|t])' \right] = \sum_{j=0}^{h-1} (B_j \Sigma_u B_j')$$

This expression is of course equal to (4.3), but now stated using the moving average coefficients. Moreover, when $h \rightarrow \infty$ it is equal to $\Psi(0)$, the unconditional variance of the VAR process. In mathematical terms,

$$\sigma_{t+h}^f = \Psi(0) \quad \text{when } h \rightarrow \infty$$

Thus just as in the univariate case, the forecast variance converges to the unconditional variance of the process.

4.3 Uncertainty

When we assume that the errors of the VAR model are Gaussian and independent across time (i.e. $\mathbf{u}_t \sim \text{i.i.d } \mathcal{N}(0, \Sigma_u)$), it implies that the forecast errors for the k variables in the model are also normally distributed, such that,

$$\frac{\mathbf{y}_{k,t+h} - \mathbb{E}[\mathbf{y}_{k,t+h}|\mathbf{y}_{k,t}]}{\sigma_{k,t+h}^f} \sim \mathcal{N}(0, 1)$$

Where $\mathbf{y}_{k,t+h}$ and $\mathbb{E}[\mathbf{y}_{k,t+h}|\mathbf{y}_{k,t}]$ are the k 'th elements of the actual and predictor values, and $\sigma_{k,t+h}^f$ is the square root of the variance for the k 'th equation. That is, the square root of the k 'th element on the diagonal on the σ_{t+h}^f matrix. As such, forecast intervals can be generated around the VAR point forecasts using,

$$[\mathbb{E}[\mathbf{y}_{k,t+h}|\mathbf{y}_{k,t}] - z_{\alpha/2} \sigma_{k,t+h}^f, \mathbb{E}[\mathbf{y}_{k,t+h}|\mathbf{y}_{k,t}] + z_{\alpha/2} \sigma_{k,t+h}^f]$$

which is equivalent to what was derived in the previous discussion on forecasts

4.4 Forecast failure in macroeconomics

Small-scale VARs are widely used in macroeconomics. Since Sims (1980) pioneering work, examples of VARs used to forecast key economic variables such as output, prices, and the interest rates have been numerous. Yet, a body of recent work suggests that VAR models may be prone to instabilities. In

the face of such instabilities, a variety of estimation or forecasting methods might be used to improve the accuracy of forecasts from a VAR. These methods include using an intercept correction, time varying-parameter, differencing the data and model averaging, to mention but a few.¹¹

While many different structural factors could lead to instabilities in macroeconomic variables, breaks are regarded as the primary cause of poor forecast performance. The best way to deal with this problem often varies with the variable being forecasted, but some features emerge. According to Clements and Hendry (2011), forecasting in differences rather than in levels can provide some protection against mean shifts in the dependent variable.

Furthermore, Clark and McCracken (2008) argue that when the VAR in difference is estimated with Bayesian method or using some form of modelling averaging, the forecast consistency performs among the best methods.

5 Granger causality

VAR models are frequently employed in academic and practical settings for studying the relationships between a set of variables. One useful measure of the relationships is with the aid of the concept of Granger (1969) causality.

Using this notation of causality, the idea is that a cause cannot come after the effect. Hence, if variable $y_{2,t}$ Granger-causes behaviour in variable $y_{1,t}$, then $y_{2,t}$ should improve upon the predictions of $y_{1,t}$. Although this is very general description, following the customary assumption about an MSFE loss function and linear predictors, Granger causality in a VAR setting is easy to check. We just need to investigate the VAR representation of the system.

Consider again the VAR system in equation (1.3).

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} 0.5 & 0 \\ 1 & 0.2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$

Hence, $y_{2,t}$ does not help predict $y_{1,t}$. However, $y_{1,t}$ does help predict $y_{2,t}$. Therefore, following the definition of Granger causality, when placed in a forecasting framework, $y_{1,t}$ would Granger cause $y_{2,t}$, but $y_{2,t}$ does not Granger cause $y_{1,t}$.

Alternatively, if the coefficient matrix had been,

$$A_1 = \begin{bmatrix} 0.5 & 0.2 \\ 1 & 0.12 \end{bmatrix}$$

Both $y_{1,t}$ and $y_{2,t}$ would have Granger caused each other. In this case we often say that the system for y_t is a feedback system.

In practical situations, where the VAR representation has to be estimated, a test for Granger causality is easily implemented by employing a joint hypothesis test. That is, for each equation in the VAR we compute $K - 1$ restricted versions of the VAR model against the unrestricted version. The test, for each restricted version, is a joint hypothesis test that all the lags of the k 'th variable in the system are jointly significantly different from zero. As such, this is simply a standard F -test, where the null hypothesis is no Granger causality.

For example, if we have a VAR(2) model, with three variables, the first equation of the system would read,

$$y_{1,t} = \mu_1 + \alpha_{11}y_{1,t-1} + \alpha_{12}y_{2,t-1} + \alpha_{13}y_{3,t-1} + \dots \\ \alpha_{14}y_{1,t-2} + \alpha_{15}y_{2,t-2} + \alpha_{16}y_{3,t-2} + u_{1,t}$$

The test for no Granger causality from $y_{2,t}$ to $y_{1,t}$ would be an F -test with null, $\alpha_{12} = \alpha_{15} = 0$, and the test for no Granger causality from $y_{3,t}$ to $y_{1,t}$ would be an F -test with null, $\alpha_{13} = \alpha_{16} = 0$. Conducting similar test on the two other equations of the VAR representation would give a complete test for no Granger causality. A rejection of any of the null hypothesis indicates Granger causality.

Given these expressions, one should note that the lack of a Granger causality relationship from one variable to another variable cannot necessarily be interpreted as lack of a cause and effect relationship. As such, the name Granger causality is somewhat of a misnomer. Granger causality need not mean anything more than predictability.

6 Conclusion

The VAR model is a multivariate version of the AR model that was described previously. We noted that a VAR(p) can be written as a first-order VAR(1) model by making use of the companion-form matrix. If all eigenvalues of the companion-form matrix are less than unity in absolute value, then the VAR model is stable (covariance stationary). A stable VAR(p) model can be inverted and written as an infinite-order vector moving average model. The parameters in the VAR model can be estimated with OLS, using an equation-by-equation method. Under the standard assumption, the OLS estimator will be similar to the maximum likelihood estimator. Forecasting with a stable VAR model will converge towards the unconditional mean of the variables in the model, and the MSFE matrix of the forecast errors will converge towards the unconditional variance of the variables in the model. Density forecasts and forecast intervals could then be constructed based on the normality assumptions of the errors. Then lastly, we could use Granger causality test, which involve testing whether or not lagged values of a given variable in the VAR system help predict one of the other endogenous variables in the system. Such a test can simply be conducted using a F -test procedure, where the null hypothesis is no Granger causality.

7 Appendix: Short introduction to linear algebra

Many econometric and time series results, are more easily presented and understood with the use of linear algebra. There are many references that provide a thorough introduction to linear algebra, what is presented below represents a brief summary of a small selection of the techniques that may be applied. The exposition is far from complete and provides a basic definition of most of the terms that have been used.

For simplicity, we assume two $m \times n$ matrices, A and B ,

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}, \quad B = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}$$

where m is the number of rows and n the number of columns. In this case, $m = n = 2$ for both A and B .

The addition of the elements in two matrices is done entry-wise,

$$A + B = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} + \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} \alpha_{11} + \beta_{11} & \alpha_{12} + \beta_{12} \\ \alpha_{21} + \beta_{21} & \alpha_{22} + \beta_{22} \end{bmatrix}$$

Multiplying the matrix A by a scalar, c , provides

$$cA = \begin{bmatrix} c\alpha_{11} & c\alpha_{12} \\ c\alpha_{21} & c\alpha_{22} \end{bmatrix}$$

Multiplying the matrix A with B , corresponds to multiplying the rows in A with the columns in B ,

$$AB = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} = \begin{bmatrix} \alpha_{11}\beta_{11} + \alpha_{12}\beta_{21} & \alpha_{11}\beta_{12} + \alpha_{12}\beta_{22} \\ \alpha_{21}\beta_{11} + \alpha_{22}\beta_{21} & \alpha_{21}\beta_{12} + \alpha_{22}\beta_{22} \end{bmatrix}$$

Generally, the multiplication of an $m \times n$ matrix with another $n \times t$ matrix will result in an $m \times t$ matrix.

The transpose of a matrix is given by,

$$A' = \begin{bmatrix} \alpha_{11} & \alpha_{21} \\ \alpha_{12} & \alpha_{22} \end{bmatrix}$$

If $A = A'$, we say A is a symmetric matrix. In addition, taking the transpose of $(AB)'$ corresponds to performing the multiplication of $B'A'$.

The determinant and trace of a matrix are defined respectively as,

$$\det(A) = |A| = \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}$$

$$\text{trace}(A) = \alpha_{11} + \alpha_{22}$$

To compute the eigenvalues of a matrix, we could use the trace and determinant, where the λ 's are the eigenvalues of the matrix A .

$$\lambda^2 - \lambda \text{trace}(A) + \det(A) = 0$$

$$\lambda_1 = \frac{\text{trace}(A) + \sqrt{\text{trace}(A)^2 - 4 \det(A)}}{2}$$

$$\lambda_2 = \frac{\text{trace}(A) - \sqrt{\text{trace}(A)^2 - 4 \det(A)}}{2}$$

$$\lambda_1 + \lambda_2 = \text{trace}(A), \quad \lambda_1 \lambda_2 = \det(A)$$

If all eigenvalues of a matrix are non-negative, we sometimes say that the matrix is positive semi-definite. An example of a matrix that is always symmetric and positive semi-definite is the covariance matrix, Σ . Such a matrix can always be written as $\Sigma = PP'$, where P is a Cholesky decomposition of Σ (i.e. the lower triangular) matrix.

The inverse of a matrix could then be expressed as,

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} \alpha_{22} & -\alpha_{12} \\ -\alpha_{21} & \alpha_{11} \end{bmatrix}$$

If the determinant of a matrix equals zero, we say that the matrix is singular. For a singular matrix the inverse does not exist.

The rank of a matrix A is the maximum number of linear independent column vectors in A . For the matrix A , we have

$$\begin{aligned}rk(A) &= 2 && \text{if } \det(A) = \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} \neq 0 \\rk(A) &= 1 && \text{if } \det(A) = 0 \text{ but } A \neq 0 \\rk(A) &= 0 && \text{if } A = 0\end{aligned}$$

where $A = 0$ if all the elements of A are equal to zero. The rank of the system is often used when combining linear systems, such as $Ax = b$. Let A be an $n \times n$ matrix and b an $n \times 1$ vector of known values. Then, the $n \times 1$ vector of unknowns in x will have a unique solution, as long as $rk(A) = n$. We can also say that A has a reduced rank if $rk(A) < n$.

Employing the *vec* operator to A yields,

$$vec(A) = \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \\ \alpha_{12} \\ \alpha_{22} \end{bmatrix}$$

and the Kronecker product between A and B is,

$$A \otimes B = \begin{bmatrix} \alpha_{11}B & \alpha_{12}B \\ \alpha_{21}B & \alpha_{22}B \end{bmatrix}$$

Hence, the Kronecker product of an $m \times n$ matrix with an $r \times t$ matrix will result in as $mr \times nt$ matrix.

Finally, the identity matrix and a lower triangular matrix are particularly useful. The identity matrix of order n is an $(n \times n)$ matrix with ones along the principal diagonal and zeros elsewhere,

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

The elements above the principal diagonal of a lower triangular matrix are all equal to zero ($\alpha_{ij} = 0$ for $j > i$)

$$\begin{bmatrix} \alpha_{11} & 0 & 0 & \dots & 0 \\ \alpha_{21} & \alpha_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \alpha_{n3} & \dots & \alpha_{nn} \end{bmatrix}$$

8 References

- Allen, G. P., and R. Fildes. 2005. "Levels, Differences and Ecms - Principles for Improved Econometric Forecasting." *Oxford Bulletin of Economics and Statistics* 67 (1): 881–904.
- Banerjee, A., J. Dolado, J. Galbraith, and David F. Hendry. 1993. *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

- Clark, T. E., and M. W. McCracken. 2008. "Forecasting with Small Macroeconomic VARs in the Presence of Instabilities." *Frontiers of Economics and Globalization* 3: 93–147.
- Clements, Michael P., and David F. Hendry. 2011. *The Oxford Handbook of Economic Forecasting*. Oxford University Press, USA.
- Engle, Robert F., and Clive W. J. Granger. 1987. "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica* 55 (2): 251–76.
- Granger, Clive W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37(3): 424–38.
- Hamilton, James D. 1994. *Time Series Analysis*. Princeton: Princeton University Press.
- Hendry, David F. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.
- Johansen, Søren. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. *Journal of Economic Dynamics and Control*. Vol. 12. Advanced Texts in Econometrics 213. Oxford: Oxford University Press.
- Lutkepohl, H. 2005. *New Introduction to Multiple Time Series Analysis*. Heidelberg: Springer-Verlag.
- Sims, Christopher A. 1972. "Money, Income, and Causality." *American Economic Review* 62 (4): 540–52.
- . 1980. "Comparison of Interwar and Postwar Business Cycles." *American Economic Review* 70 (2): 250–57.
- Wold, H. 1938. "A Study in the Analysis of Stationary Time Series." PhD thesis, Uppsala.

-
1. For example, the large macroeconomic model of the Bureau of Economic Research and the core forecasting model of the South African Reserve Bank have underlying VAR structures.↵
 2. A number of deterministic regressors could also be included in this model.↵
 3. The appendix contains a summary of the essential tools that relate to aspects of matrix algebra that are utilised in this section.↵
 4. The interested reader is referred to Lutkepohl (2005), Hendry (1995), Johansen (1995), Hamilton (1994), Banerjee et al. (1993).↵
 5. Note that a symmetric matrix is defined to be positive definite if the real parts of all eigenvalues are positive. A non-symmetric matrix, X , is positive definite if all eigenvalues of $(X + X')/2$ are positive.↵
 6. Further details are available in Lutkepohl (2005).↵
 7. See, Hamilton (1994) for further details.↵
 8. Note that $A(L)^{-1} = B(L) \rightarrow I = B(L)A(L)$, since $A(L)^{-1}A(L) = I$.↵
 9. See Lutkepohl (2005) for details.↵

10. The resulting estimator has the same efficiency as the generalized LS (GLS) estimator.↵
11. See, Clements and Hendry (2011), Clark and McCracken (2008) and Allen and Fildes (2005) for detailed discussions.↵