Contents

5 Stationary Multivariate Time Series

In many situations, analyzing a time series in isolation is reasonable. However, it can also be limiting. For example, Campbell (1996) links financially interesting variables, including stock returns and the default premium, together in a multivariate system allowing shocks to one variable to propagate to the others. A model which allows dependencies between state variables seems reasonable; investors constantly observe shocks in one asset which result in changed assessments of others. How do we include such "multivariate" reasoning in forecasting?

A naive approach is as follows. We write a model such as

$$Y_t = \alpha + \beta X_t + \epsilon_t, \tag{5.0.1}$$

that is a regression of Y_t on X_t. Then if we want to forecast Y_{t+h}, $h \geq 1$, we need to "plug" a forecast of X_{t+h}. This shows that we may need to forecast several variables simultaneously. The vector autoregression (VAR) is a mechanism used to link multiple stationary time series together. Their aim is to address a number of issues.

1. The optimal forecast of X_{t+h} may itself depend on Y_{t+j} for $j < h$, i.e. as the two variables are related, we may need to forecast each one using the other, by modeling their interactions.

2. When forecasting Y_{t+h} we need to make sure that ϵ_t is itself uncorrelated with the past variables.

3. In model (5.0.1), if $\mathbb{C}ov\left(X_t, \epsilon_t\right) \neq 0$, then the regression model does not yield an unbiased estimator of β.

4. We might want to conduct scenarios and see how changing X_t today (e.g. the price of an item) will have an impact on future Y_{t+h}, $h \geq 1$ (e.g. future sales). Thus we look for some notion of "intertemporal partial derivative". This is the purpose of the "impulse response functions". We might also want to see how a scenario for a path of future $(X_{t+1}, ..., X_{t+h}) = (x_{t+1}, ..., x_{t+h})$ affects future Y_{t+j}, i.e. conditional forecasts.

5. In model (5.0.1), if X_t or Y_t is non-stationary, then the estimates of (α, β) may be meaningless. In some situations, they may not converge to an interpretable quantity of interest as the sample size increases. We need to find a way to assess these estimators, this is the purpose of "cointegration"that we will see later in the course.

5.1 Vector Autoregressions

Vector autoregressions extend univariate autoregressions to multivariate (also called multidimensional) settings; the intuition behind most results carries over by simply replacing scalars with matrices, and scalar operations with matrix operations. The new concepts of VAR analysis are Granger causality and impulse response functions.

5.1.1 Definition

A pth order vector autoregression (VAR(p)) is defined as process with dynamics governed by

$$\mathbf{Y}_t = \mathbf{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{\Phi}_2 \mathbf{Y}_{t-2} + ... + \mathbf{\Phi}_p \mathbf{Y}_{t-p} + \mathbf{W}_t,$$

where $\mathbf{Y}_t$ is a K by 1 vector stochastic process, $\mathbf{\Phi}_j$, $j = 1, ..., p$ are K by K matrices and $\mathbf{W}_t$ is a vector white noise process, satisfying

$$\begin{aligned} \mathbb{E}\left[\mathbf{W}_t\right] &= \mathbf{0} \\ \mathbb{E}\left[\mathbf{W}_t \mathbf{W}_{t-s}'\right] &= \mathbf{0} \\ \mathbb{E}\left[\mathbf{W}_t \mathbf{W}_t'\right] &= \mathbf{\Sigma} \end{aligned}$$

where $\mathbf{\Sigma}$ is a positive definite finite matrix, and prime denotes transpose.

5.1.2 Properties of a VAR(1)

Stationarity. A VAR(1), given by,

$$\mathbf{Y}_t = \mathbf{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{W}_t,$$

is covariance stationary if the eigenvalues of $\mathbf{\Phi}_1$ are less than 1 in modulus.[1] In the univariate case, this is equivalent to the condition $|\varphi_1| < 1$. Backward substitution can be used to show that

$$\mathbf{Y}_t = \sum_{i=0}^{\infty} \mathbf{\Phi}_1^i \mathbf{W}_{t-i}.$$

Here $\mathbf{I}_K$ refers to a K by K identity matrix. The eigenvalue condition ensures that the infinite sum is well-defined.

Mean. In the VAR(1) above, the expectation of $\mathbf{Y}_t$ is zero, but in fact this can be extend to the case with an intercept

$$\mathbf{Y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{W}_t,$$

where $\mathbf{\Phi}_0$ is a vector of dimension K. Then, taking expectation

$$\mathbb{E}\left[\mathbf{Y}_t\right] = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbb{E}\left[\mathbf{Y}_{t-1}\right]$$

so under the assumption of stationarity (so $\mathbb{E}\left[\mathbf{Y}_t\right] = \mathbb{E}\left[\mathbf{Y}_{t-1}\right]$),

$$\mathbb{E}\left[\mathbf{Y}_t\right] = \left(\mathbf{I}_K - \mathbf{\Phi}_1\right)^{-1} \mathbf{\Phi}_0.$$

The eigenvalues play an important role in determining the mean. If an eigenvalue of $\mathbf{\Phi}_1$ is close to one, $\left(\mathbf{I}_K - \mathbf{\Phi}_1\right)^{-1}$ will contain large values and the unconditional mean will be large. Similarly, if $\mathbf{\Phi}_1 = \mathbf{0}$, then the mean is simply $\mathbf{\Phi}_0$ and the process is just white noise plus a constant. Let's denote $\mu = \mathbb{E}\left[\mathbf{Y}_t\right]$, then we can write

$$\mathbf{Y}_t - \mu = \mathbf{\Phi}_1 \left(\mathbf{Y}_{t-1} - \mu\right) + \mathbf{W}_t,$$

[1] **Definition** (**Eigenvalue**) λ is an eigenvalue of an n by n square matrix $\mathbf{A}$ if and only if $|\mathbf{A} - \lambda \mathbf{I}_n| = 0$. The crucial properties of eigenvalues for applications to VARs are given in the following theorem:
Theorem (**Matrix Power**) Let $\mathbf{A}$ be an n by n matrix. Then the following statements are equivalent:

- $\mathbf{A}^m \to 0$ as $m \to \infty$.
- All eigenvalues of $\mathbf{A}$, λ_i, $i = 1, 2, ..., n$, are less than 1 in modulus ($|\lambda_i| < 1$).
- The series $\sum_{i=0}^m \mathbf{A}^i \to (\mathbf{I}_n - \mathbf{A})^{-1}$ as $m \to \infty$.

so that we can disregard the issues related with nonzero $\mathbf{\Phi}_0$ below and assume throughout that $\mu = 0$.

Variance. Using the MA(∞) form, the long run variance can be shown to be

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbf{Y}_t - \mu\right)\left(\mathbf{Y}_t - \mu\right)'\right] = \mathbb{E}\left[\mathbf{Y}_t\mathbf{Y}_t'\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-i}\right)'\right] \\
&= \mathbb{E}\left[\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{W}_{t-i}'\mathbf{\Phi}_1^{i\prime}\right)\right] \\
&= \sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbb{E}\left[\mathbf{W}_{t-i}\mathbf{W}_{t-i}'\right]\mathbf{\Phi}_1^{i\prime} \text{ (covariances are zero)} \\
&= \sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{\Sigma}\mathbf{\Phi}_1^{i\prime}.
\end{aligned}$$

We can simplify the expression using the operator *vec* which transforms a matrix into a vector by stacking columns on top of one another

$$vec\left(\mathbb{E}\left[\mathbf{Y}_t\mathbf{Y}_t'\right]\right) = \sum_{i=0}^{\infty} vec\left(\mathbf{\Phi}_1^i\mathbf{\Sigma}\mathbf{\Phi}_1^{i\prime}\right)$$

where $\mu = \left(\mathbf{I}_K - \mathbf{\Phi}_1\right)^{-1}\mathbf{\Phi}_0$, and using the properties that $vec\left(\mathbf{ABC}\right) = \left(\mathbf{C}' \otimes \mathbf{A}\right) vec\left(\mathbf{B}\right)$, with $\otimes$ denoting the Kronecker product[2], we obtain

$$\begin{aligned}
vec\left(\mathbb{E}\left[\mathbf{Y}_t\mathbf{Y}_t'\right]\right) &= \sum_{i=0}^{\infty}\left(\mathbf{\Phi}_1^i \otimes \mathbf{\Phi}_1^i\right) vec\left(\mathbf{\Sigma}\right) \\
&= \sum_{i=0}^{\infty}\left(\mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1\right)^i vec\left(\mathbf{\Sigma}\right) \\
&= \left(\mathbf{I}_{K^2} - \mathbf{\Phi}_1 \otimes \mathbf{\Phi}_1\right)^{-1} vec\left(\mathbf{\Sigma}\right).
\end{aligned}$$

[2] If $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{k\ell})$ are matrices of dimensions, respectively, $n \times m$ and $p \times q$ then $\mathbf{A} \otimes \mathbf{B}$ is a matrix of dimensions $np \times mq$ whose elements are:

$$\mathbf{A} \otimes \mathbf{B} = \left[a_{ij}\mathbf{B}\right]_{(i,j)}$$

i.e.

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \ldots & a_{1m}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \ddots & a_{2m}\mathbf{B} \\ \vdots & & \ddots & \vdots \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \ldots & a_{nm}\mathbf{B} \end{bmatrix}.$$

Compared to the long run variance of a univariate autoregression, $\sigma^2/\left(1-\varphi_1^2\right)$, the similarities are less obvious. The difference arises from the noncommutative nature of the matrix product ($\mathbf{AB} \neq \mathbf{BA}$ in general). Once again the eigenvalues of $\mathbf{\Phi}_1$ play an important role. If any are close to 1, the variance will be large.

Autocovariance. The autocovariances of a stationary vector valued stochastic process are defined as

$$\mathbf{\Gamma}_s = \mathbb{E}\left[\left(\mathbf{Y}_t - \mu\right)\left(\mathbf{Y}_{t-s} - \mu\right)'\right],$$

where $s \in \mathbb{Z}$. Instead of being symmetric around t, as in the univariate case, they are symmetric in their transpose. Specifically: $\mathbf{\Gamma}_s = \mathbf{\Gamma}'_{-s}$. Computing the autocovariance is conveniently done using the MA(∞) representation,

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-s-i}\right)'\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{s-1}\mathbf{\Phi}_1^i\mathbf{W}_{t-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-s-i}\right)'\right] \\ &\quad + \mathbf{\Phi}_1^s\mathbb{E}\left[\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-s-i}\right)\left(\sum_{i=0}^{\infty}\mathbf{\Phi}_1^i\mathbf{W}_{t-s-i}\right)'\right] \\ &= \mathbf{\Phi}_1^s\mathbb{V}\left[\mathbf{Y}_t\right],\end{aligned}$$

where $\mathbb{V}\left[\mathbf{Y}_t\right]$ is the covariance matrix of the VAR(1). Like most properties of a VAR, this result is fundamentally similar to the autocovariance function of an AR(1): $\gamma_s = \varphi_1^s\mathbb{V}\left[Y_t\right]$.

In the previous notes we have seen that Markov chains of order p can be seen as Markov chains of order 1 after a change of parametrization. In a similar way, VAR(p) models can be seen as VAR(1) with an extended state space. Specifically, suppose $(\mathbf{Y}_t)$ follows a VAR(p) process,

$$\mathbf{Y}_t = \mathbf{\Phi}_1\mathbf{Y}_{t-1} + \mathbf{\Phi}_2\mathbf{Y}_{t-2} + ... + \mathbf{\Phi}_p\mathbf{Y}_{t-p} + \mathbf{W}_t.$$

By subtracting the mean and stacking p consecutives states of $\mathbf{Y}_t$ into a large column vector denoted

$\mathbf{Z}_t$, this VAR(p) can be transformed into a VAR(1), $\mathbf{Z}_t = \boldsymbol{\Upsilon}\mathbf{Z}_{t-1} + \xi_t$, with

$$\mathbf{Z}_t = \begin{bmatrix} \mathbf{Y}_t \\ \mathbf{Y}_{t-1} \\ \vdots \\ \mathbf{Y}_{t-p+1} \end{bmatrix}, \; \boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \boldsymbol{\Phi}_3 & \cdots & \boldsymbol{\Phi}_p \\ \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \end{bmatrix}, \; \xi_t = \begin{bmatrix} \mathbf{W}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix},$$

$$\mu = \left(\mathbf{I} - \sum_{k=1}^{p} \boldsymbol{\Phi}_k \right)^{-1} \boldsymbol{\Phi}_0.$$

This is known as the **companion form** and allows the statistical properties of VAR(p) processes to be derived from results on VAR(1).

Example: The interaction of stock and bond returns. Stocks and long term bonds are often thought to hedge one another. VARs provide a simple method to determine whether their returns are linked through time. Consider the VAR(1)

$$\begin{bmatrix} VWM_t \\ 10YR_t \end{bmatrix} = \begin{bmatrix} \phi_{01} \\ \phi_{02} \end{bmatrix} + \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ 10YR_{t-1} \end{bmatrix} + \begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix},$$

where VWM and $10YR$ stand, respectively for the Value Weighted Market returns, i.e., the return of the aggregate U.S. stock market where each stock return is weighted according to the value of the corresponding stock, and the return of the 10-year U.S. Treasury bonds, i.e., a measure of riskless long term interest rates. This can be written as two equations:

$$\begin{aligned} VWM_t &= \phi_{01} + \phi_{11,1} VWM_{t-1} + \phi_{12,1} 10YR_{t-1} + W_{1t}, \\ 10YR_t &= \phi_{02} + \phi_{21,1} VWM_{t-1} + \phi_{22,1} 10YR_{t-1} + W_{2t}. \end{aligned}$$

Since these models do not share any parameters, they can be separately estimated using OLS.[3] Using data on the VWM from the Center for Research on Security Prices (CRSP) and the 10 constant maturity treasury yield from the FRED database at the Federal Reserve Bank of Saint

[3] When W_{1t} and W_{2t} are uncorrelated, it is efficient to estimate the two equations separately via OLS. When these two noises are correlated, then joint estimation of the two equations is more efficient, in the sense that the variance of the estimators will be smaller.

Louis[4] from May 1953 until December 2004, a VAR(1) was estimated.

$$\begin{bmatrix} VWM_t \\ 10YR_t \end{bmatrix} = \begin{bmatrix} \underset{(0.00)}{0.996} \\ \underset{(0.68)}{0.046} \end{bmatrix} + \begin{bmatrix} \underset{(0.76)}{0.012} & \underset{(0.00)}{0.239} \\ \underset{(0.03)}{-0.058} & \underset{(0.00)}{0.334} \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ 10YR_{t-1} \end{bmatrix} + \begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix},$$

where the p-value for a test of the nullity of each coefficient is in parenthesis. A few things are worth noting. Stock returns are not predictable with their own lags but do appear to be predictable using lagged bond returns: positive bond returns lead to positive future returns in stocks. In contrast, positive returns in equities result in negative returns for future bond holdings. The long run mean can be computed as

$$\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.012 & 0.239 \\ -0.058 & 0.334 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0.996 \\ 0.046 \end{bmatrix} = \begin{bmatrix} 1.004 \\ -0.018 \end{bmatrix}$$

These values are similar to the sample means of 1.003 and -0.022.

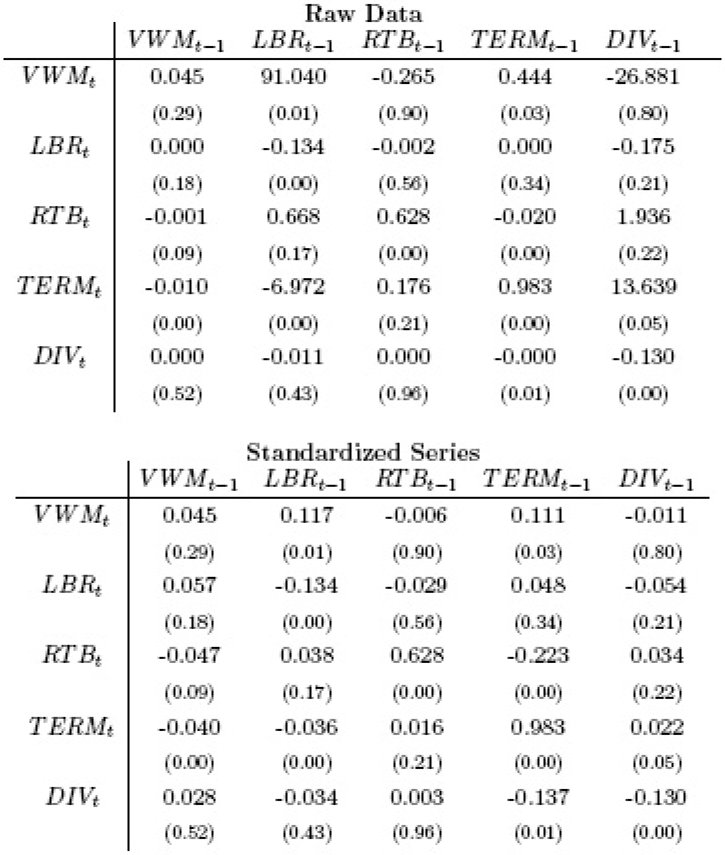
Example: Campbell's VAR. Campbell (1996) builds a theoretical model for asset prices where economically meaningful variables evolve according to a VAR, including stock returns, real labor income growth, the term premium, the relative t-bill rate and the dividend yield. The VWM series from CRSP is used for equity returns and real labor income is the log change in income from labor minus the log change in core inflation where both series are from FRED. The term premium is the difference between a 10 year constant maturity yield and a 3-month t-bill rate. Both series are from FRED. The relative t-bill rate is the current yield on a 1-month t-bill minus the average yield over the past 12 months and the data is available on Ken French's website. The dividend yield was computed as the difference in the VWM with and without dividends; both series are available from CRSP.

Using a VAR(1) specification, the model can be described as

$$\begin{bmatrix} VWM_t \\ LBR_t \\ RTB_t \\ TERM_t \\ DIV_t \end{bmatrix} = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \begin{bmatrix} VWM_{t-1} \\ LBR_{t-1} \\ RTB_{t-1} \\ TERM_{t-1} \\ DIV_{t-1} \end{bmatrix} + \begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix}.$$

Parameter estimates are reported in Table 1. Two sets of parameters are presented. The top panel contains estimates using non-scaled data. This produces some very large (in magnitude, not

[4] The yield is first converted to prices and then returns are computed as log differences in the prices.



Raw Data

	VWM_{t-1}	LBR_{t-1}	RTB_{t-1}	$TERM_{t-1}$	DIV_{t-1}
VWM_t	0.045	91.040	-0.265	0.444	-26.881
	(0.29)	(0.01)	(0.90)	(0.03)	(0.80)
LBR_t	0.000	-0.134	-0.002	0.000	-0.175
	(0.18)	(0.00)	(0.56)	(0.34)	(0.21)
RTB_t	-0.001	0.668	0.628	-0.020	1.936
	(0.09)	(0.17)	(0.00)	(0.00)	(0.22)
$TERM_t$	-0.010	-6.972	0.176	0.983	13.639
	(0.00)	(0.00)	(0.21)	(0.00)	(0.05)
DIV_t	0.000	-0.011	0.000	-0.000	-0.130
	(0.52)	(0.43)	(0.96)	(0.01)	(0.00)

Standardized Series

	VWM_{t-1}	LBR_{t-1}	RTB_{t-1}	$TERM_{t-1}$	DIV_{t-1}
VWM_t	0.045	0.117	-0.006	0.111	-0.011
	(0.29)	(0.01)	(0.90)	(0.03)	(0.80)
LBR_t	0.057	-0.134	-0.029	0.048	-0.054
	(0.18)	(0.00)	(0.56)	(0.34)	(0.21)
RTB_t	-0.047	0.038	0.628	-0.223	0.034
	(0.09)	(0.17)	(0.00)	(0.00)	(0.22)
$TERM_t$	-0.040	-0.036	0.016	0.983	0.022
	(0.00)	(0.00)	(0.21)	(0.00)	(0.05)
DIV_t	0.028	-0.034	0.003	-0.137	-0.130
	(0.52)	(0.43)	(0.96)	(0.01)	(0.00)

Table 1: Parameter estimates from Campbell's VAR. The top panel contains estimates using unscaled data while the bottom panel contains estimates from data that have been standardized to have unit variance. While the magnitudes of many coefficients change, the p-values and the eigenvalues of these two parameter matrices are identical. The standardized series have one slight advantage in that the parameters are roughly comparable since the series have approximately the same variance.

statistical significance) estimates which are the result of two variables having different scales. The bottom panel contains estimates from data which have been transformed by dividing each series by its standard deviation. This puts the coefficients on a roughly level playing field. Notice that the p-values are unchanged by the scaling. One less obvious feature of the two sets of estimates is that the eigenvalues of the two parameter matrices are identical; both estimates suggest the same persistence of shocks.

5.1.3 VAR forecasting.

Recall that the h-step ahead forecast from an AR(1) is given by,

$$\mathbb{E}\left[Y_{t+h}|Y_t\right] = \phi_1^h Y_t,$$

with associated forecast error

$$\begin{aligned} e_{t+h|t} &= Y_{t+h} - \mathbb{E}\left[Y_{t+h}|Y_t\right] = Y_{t+h} - \phi_1^h Y_t \\ &= \left[\phi_1^h Y_t + \sum_{j=0}^{h-1} \phi_1^j W_{t+h-j}\right] - \phi_1^h Y_t \\ &= \sum_{j=0}^{h-1} \phi_1^j W_{t+h-j}, \end{aligned}$$

so the mean square forecast error (MSFE) at horizon h is

$$\begin{aligned} \mathsf{MSFE}\left(h\right) &= \mathbb{E}\left[e_{t+h|t}^2|Y_t\right] = \mathbb{E}\left[\left(\sum_{j=0}^{h-1} \phi_1^j W_{t+h-j}\right)^2 |Y_t\right] \\ &= \mathbb{E}\left[\left(\sum_{j=0}^{h-1} \phi_1^j W_{t+h-j}\right)^2\right] \\ &= \sigma_W^2 \sum_{j=0}^{h-1} \phi_1^{2j} = \frac{1-\phi_1^{2h}}{1-\phi_1^2}\sigma_W^2, \end{aligned}$$

as was seen in Section 3.5 of these notes.

The h-step ahead forecast of a VAR(1) is essentially identical,

$$\mathbb{E}\left[\mathbf{Y}_{t+h}|\mathbf{Y}_t\right] = \mathbf{\Phi}_1^h \mathbf{Y}_t.$$

The associated forecast errors are

$$\mathbf{e}_{t+h|t} = \mathbf{Y}_{t+h} - \mathbf{\Phi}_1^h \mathbf{Y}_t = \sum_{j=0}^{h-1} \mathbf{\Phi}_1^j \mathbf{W}_{t+h-j},$$

with the MSFE that is now a matrix:

$$\mathsf{MSFE}\left(h\right) = \mathbb{E}\left[\mathbf{e}_{t+h|t}\mathbf{e}_{t+h|t}^{\prime}|Y_t\right] = \sum_{j=0}^{h-1} \mathbf{\Phi}_1^j \mathbf{\Sigma}_W \mathbf{\Phi}_1^{j\prime}.$$

The question is how to evaluate and compare such matrices? We could be only interested in the diagonal elements, i.e. only in how well we forecast the individual variables of $\mathbf{Y}_{t+h}$. These entries could be gathered into a scalar via the trace of the MSFE

$$tr\mathsf{MSFE}\left(h\right) = \sum_{k=1}^{K} \mathbb{E}\left[e_{k,t+h|t}^2|Y_t\right] = \sum_{j=0}^{h-1} tr\left[\mathbf{\Phi}_1^j \mathbf{\Sigma}_W \mathbf{\Phi}_1^{j\prime}\right],$$

where $e_{k,t+h|t}^2$ is the squared error made in forecasting individual variables $Y_{k,t+h}$ where $\mathbf{Y}_t = \left(Y_{1,t}, \ldots, Y_{K,t}\right)^{\prime}$. The drawback of the trace MSFE is that it does not take into account the covariances between the forecast errors so other functions (e.g. norms) of the MSFE can be used.

Example: The interaction of stock and bond returns. One important feature of VAR occurs when two series are related in time. Univariate forecasts cannot adequately capture the feedback between two series. To illustrate the differences, recursively estimated 1-step ahead forecasts were produced from both the stock-bond VAR(1),

$$\left[\begin{array}{c} VWM_t \\ 10YR_t \end{array}\right] = \left[\begin{array}{c} 0.996 \\ 0.046 \end{array}\right] + \left[\begin{array}{cc} 0.012 & 0.239 \\ -0.058 & 0.334 \end{array}\right] \left[\begin{array}{c} VWM_{t-1} \\ 10YR_{t-1} \end{array}\right] + \left[\begin{array}{c} W_{1t} \\ W_{2t} \end{array}\right],$$

and simple AR(1)'s for each series. The data set contains a total of 620 observations. Beginning with the first 381 observations the models (the VAR and the two ARs) were estimated using an expanding window of data — i.e., at each step increasing the number of observations used in estimation and forecasting one-step ahead until 619 observations were used to forecast the last of the sample — and 1-step ahead forecasts were computed. Figure 5.1.1 contains a graphical representation of the differences between AR(1) and VAR(1). The forecasts for the market are substantially different while the forecasts for the 10-year bond returns are not. The changes (or lack thereof) are simply a function of the model specification: the return on the 10-year bond has predictive power for both so the VAR(1) is a much better model than an AR(1) for stock returns, yet it is not much better

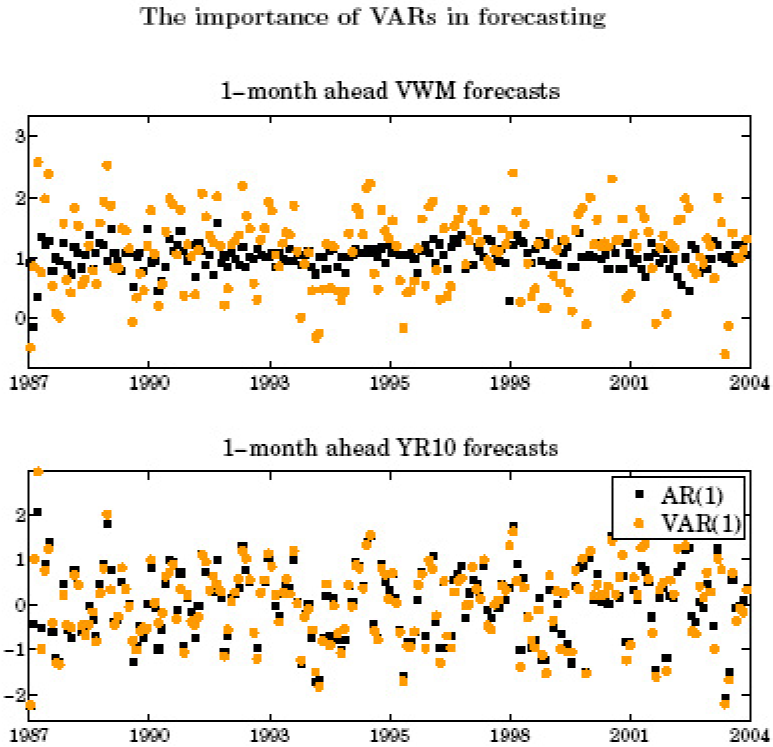


Figure 5.1.1: The figure contains 1-step ahead forecasts from both a VAR(1) and an AR(1) for both the value-weighted market returns and the return on a 10 year bond. These two pictures indicate that the return on the long bond has substantial predictive power for equity returns and the opposite is not true.

for bond returns.

Estimation and identification: they offer the first significant departures from the lessons learned in univariate models. In addition to ACFs and PACFs, vector stochastic processes also have cross-correlation functions (CCFs) and partial cross-correlation functions (PCCFs). The cross-correlations between two processes X_t and Y_t are defined as

$$\rho_{xy,s} = \frac{\mathbb{E}\left[\left(X_t - \mu_x\right)\left(Y_{t-s} - \mu_y\right)\right]}{\sqrt{\mathbb{V}\left[X_t\right]\mathbb{V}\left[y_t\right]}}.$$

It should be obvious that, unlike autocorrelations, cross-correlation are not symmetric so the order xy or yx matters. Partial cross-correlation are defined in a similar manner; the correlation between X_t and Y_{t-s} controlling for $Y_{t-1}, ..., Y_{t-s+1}$. To get a feel for the value of these two functions, Figure 5.1.2 contains the ACF and CCF of two VAR(1)with identical persistence. The top panel contains the functions for

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} .5 & .4 \\ .4 & .5 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix},$$

while the bottom contains the functions for a VAR

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} .9 & 0 \\ 0 & .9 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} W_{1t} \\ W_{2t} \end{bmatrix},$$

which is actually two separate AR(1) processes. The nontrivial VAR(1) demonstrates dependence with respect to both series while the AR-in-disguise shows no dependence between Y_t and X_{t-j} , $j > 0$.

With the new tools, CCFs and PCCFs, it would seem that the Box-Jenkins methodology could be directly extended to vector processes. However, in practice it is extraordinarily difficult to look at ACF, PACF, CCF and PCCF and to determine what type of model is needed or the appropriate coefficients. There are just too many possible interactions and too many possible models to choose from.

A solution is to take a “hands off” approach (as advocated by Sims). The initial VAR specification should include all variables which theory indicates are relevant to the problem at hand and a lag length should be chosen which has a high likelihood of capturing all of the dynamics. Once this value has been set, either a general-to-specific search can be conducted over the lag length or an information criteria should be used. In the VAR case, the AIC and BIC (Schwarz Criterion SC)

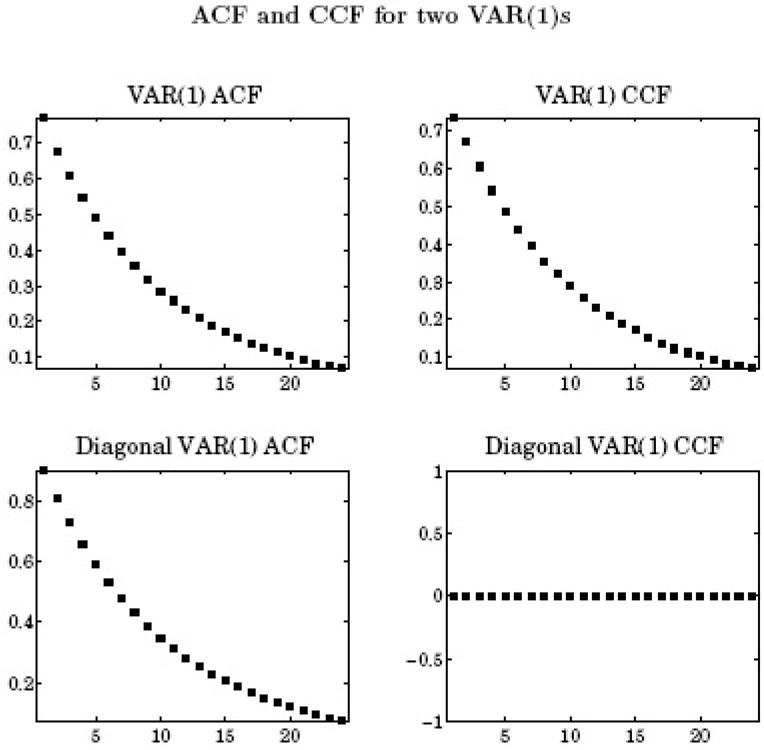


Figure 5.1.2: The top panel contains the ACF and CCF for a nontrivial VAR process where contemporaneous values depend on both series. The bottom contains the ACF and CCF for a trivial VAR which is simply composed of two AR(1)s.

are given by

$$\begin{aligned}\mathsf{AIC} : \ln\left|\mathbf{\Sigma}\left(p\right)\right| + \frac{2K^2p}{T},\\ \mathsf{BIC/SC} : \ln\left|\mathbf{\Sigma}\left(p\right)\right| + \frac{K^2p\ln T}{T},\end{aligned}$$

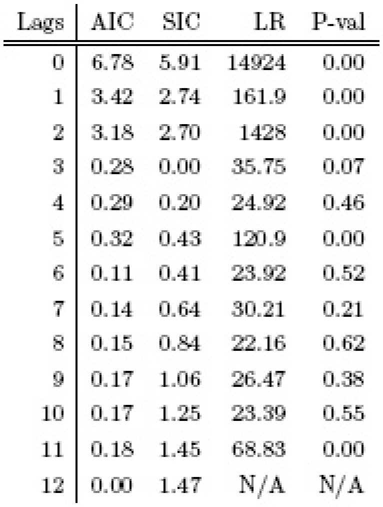
where $\mathbf{\Sigma}\left(p\right)$ is the covariance of the residuals using p lags and $|\cdot|$ indicates the determinant. The lag length should be chosen to minimize one of these criteria, and the SIC will always choose a (weakly) smaller model than the AIC.

To use a general-to-specific approach, a simple likelihood ratio test can be computed

$$\left(T - p_2K^2\right)\left(\ln\left|\mathbf{\Sigma}\left(p_1\right)\right| - \ln\left|\mathbf{\Sigma}\left(p_2\right)\right|\right) \to \chi^2_{(p_2-p_1)K^2},$$

where p_1 is the number of lags in the restricted (smaller) model, p_2 is the number of lags in the unrestricted (larger) model and K is the dimension of $\mathbf{Y}_t$. Since model 1 is a restricted version of model 2, its variance is larger which ensures this statistic is positive (a good thing since it has a χ^2 distribution).

Example: Campbell's VAR A lag length selection procedure was conducted using Campbell's VAR. The results are contained in table 2. This table contains both the AIC and SIC values for lags 0 through 12 as well as likelihood ratio test results for testing ℓ lags against $\ell + 1$. Note that the LR and p-value corresponding to lag ℓ is a test of the null of ℓ lags against an alternative of $\ell+1$ lags. Using the AIC, 12 lags would be selected since it produces the smallest value. If the initial lag length was less than 12, 6 lags would be selected. The SIC chooses 3 lags in an unambiguous manner. A general-to-specific procedure would choose 12 lags while a specific-to-general procedure would choose 4. The test statistic for a null $\mathsf{H}_0 : p = 11$ against an alternative $\mathsf{H}_1 : p = 12$ has a value of 68 and a p-value of 0. One final specification search was conducted. Rather than begin at the largest lag and work down one by one, a large specification search which evaluates models with every combination of lags up to 12 was computed. This required fitting 4096 regressions which fortunately only requires 9 seconds. For each possible combination of lags, the AIC and the BIC were computed. Using this methodology, the AIC search selected lags 1-4, 6, 10 and 12 while the BIC selected a smaller model with only lags 1, 3 and 12. Search procedures of this type are computationally viable for checking up to about 20 lags.



Lags	AIC	SIC	LR	P-val
0	6.78	5.91	14924	0.00
1	3.42	2.74	161.9	0.00
2	3.18	2.70	1428	0.00
3	0.28	0.00	35.75	0.07
4	0.29	0.20	24.92	0.46
5	0.32	0.43	120.9	0.00
6	0.11	0.41	23.92	0.52
7	0.14	0.64	30.21	0.21
8	0.15	0.84	22.16	0.62
9	0.17	1.06	26.47	0.38
10	0.17	1.25	23.39	0.55
11	0.18	1.45	68.83	0.00
12	0.00	1.47	N/A	N/A

Table 2: Normalized values for the AIC and SIC for Campbell's VAR. The AIC chooses 12 lags while the SIC chooses only 3. A general-to-specific search would stop at 12 lags since the likelihood ratio test of 12 lags against 11 rejects with a p-value of 0. If the initial number of lags was less than 12, the GeTS procedure would choose 6 lags. Note that the LR and $p$value corresponding to lag l is a test of the null of l lags against an alternative of $l+1$ lags.

5.2 Granger Causality

5.2.1 Definition

Granger causality is the first concept exclusive to vector analysis. Denoted "GC", it is the standard method to determine whether one variable is useful in predicting another and it is a good indicator of whether a VAR is needed. It was introduced by Clive Granger (2003 Economics Nobel Prize recipient for his contribution on 'cointegration'), hence the name.

Definition 1. *Granger causality is generally defined in the negative. A scalar random variable $\{X_t\}$ is said not to Granger cause $\{Y_t\}$ if*[5]

$$\mathbb{E}\left[Y_t|X_{t-1},Y_{t-1},X_{t-2},Y_{t-2},...\right]=\mathbb{E}\left[Y_t|Y_{t-1},Y_{t-2},...\right]$$

That is, $\{X_t\}$ does not Granger cause $\{Y_t\}$ if the forecast of Y_t is the same whether conditioned on past values of X_t or not.

[5] Technically, this definition is for Granger Causality in the mean. Other definition exist for Granger causality in the variance (replace conditional expectation with conditional variance) and distribution (replace conditional expectation with conditional distribution).

Granger causality can be simply illustrated in a bivariate VAR,

$$\left[\begin{array}{c} X_t \\ Y_t \end{array}\right] = \left[\begin{array}{cc} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{array}\right] \left[\begin{array}{c} X_{t-1} \\ Y_{t-1} \end{array}\right] + \left[\begin{array}{cc} \phi_{11,2} & \phi_{12,2} \\ \phi_{21,2} & \phi_{22,2} \end{array}\right] \left[\begin{array}{c} X_{t-2} \\ Y_{t-2} \end{array}\right] + \left[\begin{array}{c} W_{1t} \\ W_{2t} \end{array}\right].$$

In this model, if $\phi_{21,1} = \phi_{21,2} = 0$ then X_t does not Granger cause Y_t. If this is the case, one may be tempted to think that

$$Y_t = \phi_{22,1} Y_{t-1} + \phi_{22,2} Y_{t-2} + W_{2t}$$

is a correct specification of Y_t. However, it is not: W_{1t} and W_{2t} can be contemporaneously correlation. If it happens to be the case that X_t does not Granger cause Y_t and W_{1t} and W_{2t} have no contemporaneous correlation, then Y_t is said to be weakly exogenous and Y_t can be modeled completely independently of X_t. It is worth noting that $\{X_t\}$ not Granger causing $\{Y_t\}$ says nothing about wether $\{Y_t\}$ Granger causes $\{X_t\}$.

One limitation of GC is that it doesn't account for indirect effects. For example, suppose X_t and Y_t are both Granger caused by Z_t. When this is the case, X_t will usually Granger cause Y_t even when it has no effect once Z_t has been conditioned on. Specifically,

$$\mathbb{E}\left[Y_t | X_{t-1}, Y_{t-1}, Z_{t-1}, ...\right] = \mathbb{E}\left[Y_t | Y_{t-1}, Z_{t-1}, ...\right]$$

but

$$\mathbb{E}\left[Y_t | X_{t-1}, Y_{t-1}, ...\right] \neq \mathbb{E}\left[Y_t | Y_{t-1}, ...\right].$$

5.2.2 Testing

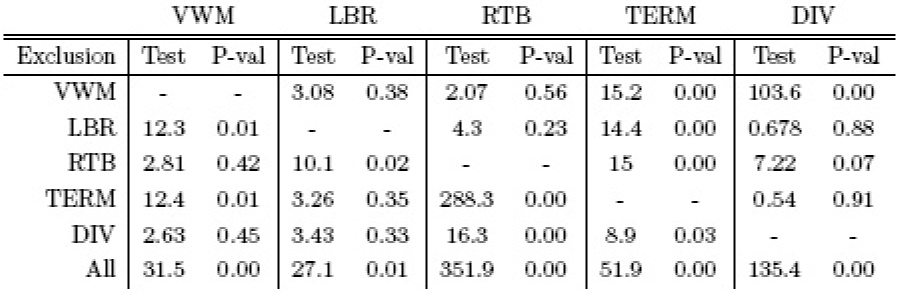
Testing for Granger causality in a VAR(p) is usually conducted using a likelihood ratio test. In this specification,

$$\mathbf{Y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{\Phi}_2 \mathbf{Y}_{t-2} + ... + \mathbf{\Phi}_p \mathbf{Y}_{t-p} + \mathbf{W}_t.$$

$\{Y_{jt}\}$ does not Granger cause $\{Y_{it}\}$ if $\phi_{ij,1} = \phi_{ij,2} = ... = \phi_{ij,p} = 0$. $\mathbf{\Phi}_0$ is always added in the regression to ensure that the errors have zero in-sample mean. The likelihood ratio test can be computed as

$$\left(T - pK^2 + K\right)\left(\ln|\mathbf{\Sigma}_r| - \ln|\mathbf{\Sigma}_u|\right) \to \chi^2_p$$

where $\mathbf{\Sigma}_r$ is the estimated residual covariance when the null of no Granger causation is imposed ($\mathsf{H}_0 : \phi_{ij,1} = \phi_{ij,2} = ... = \phi_{ij,p} = 0$) and $\mathbf{\Sigma}_u$ is the estimated covariance in the unrestricted VAR(p). If there is no Granger causation in your VAR, it's probably not a good idea to use VAR and univariate modeling might be sufficient.



	VWM		LBR		RTB		TERM		DIV	
Exclusion	Test	P-val	Test	P-val	Test	P-val	Test	P-val	Test	P-val
VWM	-	-	3.08	0.38	2.07	0.56	15.2	0.00	103.6	0.00
LBR	12.3	0.01	-	-	4.3	0.23	14.4	0.00	0.678	0.88
RTB	2.81	0.42	10.1	0.02	-	-	15	0.00	7.22	0.07
TERM	12.4	0.01	3.26	0.35	288.3	0.00	-	-	0.54	0.91
DIV	2.63	0.45	3.43	0.33	16.3	0.00	8.9	0.03	-	-
All	31.5	0.00	27.1	0.01	351.9	0.00	51.9	0.00	135.4	0.00

Table 3: Tests of Granger causality. This table contains tests where the variable on the left hand side is excluded from the regression for the variable along the top. Since the null is no GC, rejection indicates a relationship between past values of the variable on the left and contemporaneous values of variables on the top.

Example: Campbell's VAR Campbell's VAR will be used to illustrate testing for Granger causality. Table 3 contains the results of Granger causality tests from a VAR which included lags 1, 3 and 12 (as chosen by the BIC) for the 5 series in Campbell's VAR. Tests of past values of Y_t causing Y_t have been omitted as these aren't particularly informative in the multivariate context. The table tests whether the variables in the left hand column Granger cause the variables across the top. Remember that the null is no causality so that rejection (large test statistics and small $p$values) means there is a relationship. From the table, it can be seen that every variable causes at least one other variable since each row contains a $p$value indicating significance using standard test sizes (5 or 10%). It can also be seen that every variable is caused by another by examining the $p$values column by column.

5.3 Impulse Response Function

5.3.1 Definition

The second concept exclusive to vector analysis is the impulse response function. In the univariate world, the ACF was sufficient to understand shocks decay. When analyzing vector data, this is not longer the case. A shock to one series has an immediate effect but can also affect the other variables in a system which, in turn, can feed back into the original variable. After a few iterations of this cycle, it can be difficult to determine how a shock propagates even in a simple VAR(1).

The impulse response function of $Y_{i,t}$ with respect to a shock in $W_{j,t}$, for any j and i, is defined as the change in $Y_{i,t+s}$, $s \geq 0$ for a unit shock in $W_{j,t}$. This definition is somewhat hard to parse and the impulse response function can be clearly illustrated through a vector moving average (VMA).

As long as $\mathbf{Y}_t$ is covariance stationarity it must have a VMA representation,

$$\mathbf{Y}_t = \mathbf{W}_t + \mathbf{\Theta}_1 \mathbf{W}_{t-1} + \mathbf{\Theta}_2 \mathbf{W}_{t-2} + \dots$$

Using this VMA, the impulse response of Y_i with respect to a shock in W_j is the sequence $\left\{1, \mathbf{\Theta}_{1[ij]}, \mathbf{\Theta}_{2[ij]}, \mathbf{\Theta}_{3[ij]}, ...\right\}$. The difficult part is finding the matrices $\{\mathbf{\Theta}_l\}$, $l \geq 1$. In the simple VAR(1) model this is

$$\mathbf{Y}_t = \mathbf{W}_t + \mathbf{\Phi}_1 \mathbf{W}_{t-1} + \mathbf{\Phi}_1^2 \mathbf{W}_{t-2} + ...$$

However, in more complicated models, whether higher order VARs or VARMAs, determining the MA(∞) form can be tedious. One surprisingly simply, but completely correct, method to compute the elements of $\{\mathbf{\Theta}_j\}$ is to simulate the effect of a unit shock of $W_{j,t}$ directly. Suppose the model is a VAR(p),

$$\mathbf{Y}_t = \mathbf{\Phi}_1 \mathbf{Y}_{t-1} + \mathbf{\Phi}_2 \mathbf{Y}_{t-2} + ... + \mathbf{\Phi}_p \mathbf{Y}_{t-p} + \mathbf{W}_t.$$

The impulse responses can be computed by "shocking" $W_{j,t}$ by 1 unit and stepping the process forward. To use this procedure, set $\mathbf{Y}_{t-1} = \mathbf{Y}_{t-2} = \mathbf{Y}_{t-p} = \mathbf{0}$ and then begin the simulation by setting $W_{j,t} = 1$. The 0^{th} impulse will be $\mathbf{e}_j$, a vector with a 1 in the j^{th} position and zeros everywhere else. The second impulse will be,

$$\mathbf{\Theta}_1 = \mathbf{\Phi}_1 \mathbf{e}_j$$

while the second will be

$$\mathbf{\Theta}_2 = \mathbf{\Phi}_1^2 \mathbf{e}_j + \mathbf{\Phi}_2 \mathbf{e}_j$$

and the third is

$$\mathbf{\Theta}_3 = \mathbf{\Phi}_1^3 \mathbf{e}_j + \mathbf{\Phi}_1 \mathbf{\Phi}_2 \mathbf{e}_j + \mathbf{\Phi}_2 \mathbf{\Phi}_1 \mathbf{e}_j + \mathbf{\Phi}_3 \mathbf{e}_j.$$

The p^{th} lag contains the original shock while the other coefficients are capturing the complicated dynamics of the changes in $\mathbf{Y}_{t-s}, s < p$. While manual computation of an IR is tedious, it is trivial in computer packages.

5.3.2 Correlated shocks and non-unit variance: the structural VAR approach

The previous discussion has made use of unit shocks, $\mathbf{e}_j$ which represent a change of 1 in j^{th} error. This presents two problems: actual errors do not have unit variances and are often correlated. The solution to these problems is to use standardized residuals and/or correlated residuals. Suppose that the residuals in a VAR have a covariance of $\mathbf{\Sigma}_W$. To simulate the effect of a shock to element j, $W_{j,t}$ can be expressed as a linear combination of unitary and uncorrelated shocks, $\mathbf{W}_t = \mathbf{\Sigma}_W^{1/2} \varepsilon_t$

for some matrix $\mathbf{\Sigma}_W^{1/2}$ such that $\mathbf{\Sigma}_W^{1/2}\mathbf{\Sigma}_W^{1/2\prime} = \mathbf{\Sigma}_W$, and the impulses can be computed using the procedure previously outlined for the so called Structural VAR, SVAR: assume here a VAR(1)

$$\begin{aligned}\mathbf{Y}_t &= \mathbf{\Phi}_1\mathbf{Y}_{t-1} + \mathbf{W}_t = \mathbf{\Phi}_1\mathbf{Y}_{t-1} + \mathbf{\Sigma}_W^{1/2}\varepsilon_t \\ \varepsilon_t &\overset{iid}{\sim} \mathcal{N}\left(\mathbf{0}_{K\times 1}, \mathbf{I}_K\right)\end{aligned}$$

then

$$\begin{aligned}\mathbf{\Sigma}_W^{-1/2}\mathbf{Y}_t &= \left[\mathbf{\Sigma}_W^{-1/2}\mathbf{\Phi}_1\mathbf{\Sigma}_W^{1/2}\right]\mathbf{\Sigma}_W^{-1/2}\mathbf{Y}_{t-1} + \mathbf{\Sigma}_W^{-1/2}\mathbf{W}_t \\ &= \left[\mathbf{\Sigma}_W^{-1/2}\mathbf{\Phi}_1\mathbf{\Sigma}_W^{1/2}\right]\mathbf{\Sigma}_W^{-1/2}\mathbf{Y}_{t-1} + \varepsilon_t,\end{aligned}$$

i.e. a VAR(1) for the “new” variable $\mathbf{\Sigma}_W^{-1/2}\mathbf{Y}_t$ where the errors are not correlated and have unit variance.

Why do this? Essentially because we wish to compute *partial derivatives*, i.e. the impact of changing one variable only, so we want to be able to assume that we consider shocks that can happen on their own: since the variance of ε_t is $\mathbf{I}_K$, each component is uncorrelated with the others and can be shocked autonomously.

Why is it called “structural”? Essentially, because we need an extra assumption: that the shocks ε_t are meaningful. Indeed, while we can estimate $\mathbf{\Sigma}_W$ from the data, we cannot identify $\mathbf{\Sigma}_W^{1/2}$, i.e., it is not uniquely defined and we need to make a choice (possibly based on some external theory).

The choice of matrix square root, $\mathbf{\Sigma}_W^{1/2}$, matters. Two usual matrix square roots are the Cholesky and the spectral decomposition.

1. The Cholesky square root is a lower triangular matrix which imposes a natural order to the shocks. Shocking element j (using $\varepsilon_{j,t}$) has an effect of every series $1, 2, ..., j$, but not on $j+1, ..., K$.

2. By contrast the spectral decomposition is symmetric and a shock to the j^{th} error will generally effect every series instantaneously.

Unfortunately there is no right choice. If there is a natural ordering in a VAR where shocks to one series can be reasoned to have no contemporaneous effect on the other series, then the Cholesky is the correct choice. However, in many situations there is little theoretical guidance and the spectral decomposition is the natural choice. There are other possibilities, such as assuming that the long run effect of some shocks to specific variables is zero. The literature on Impulse Response Functions and SVARs is vast.[6]

[6] You may want to have a look at the slides on VARs by Ambrogio Cesa-Bianchi, https://sites.google.com/

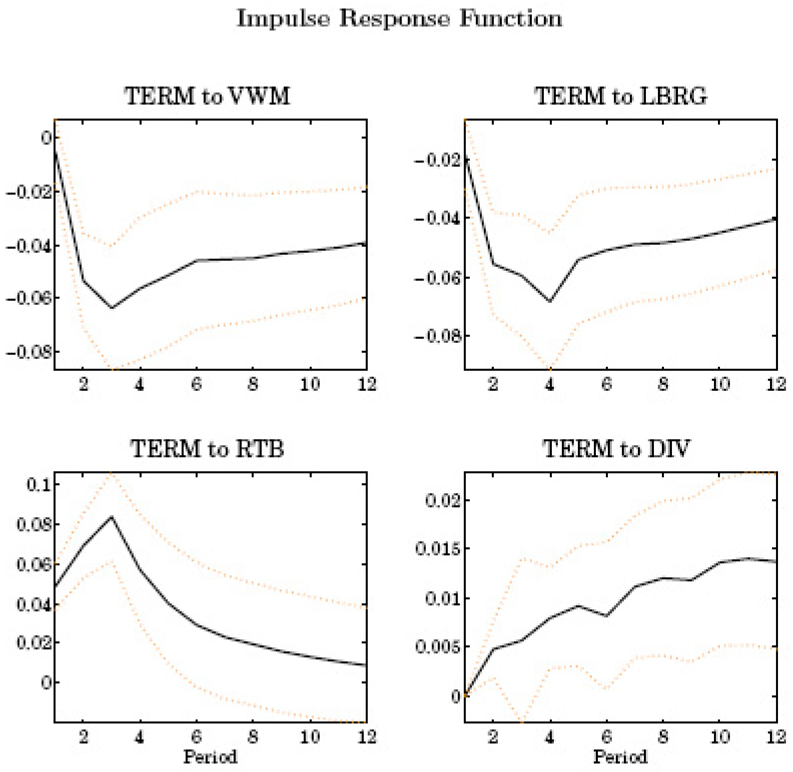


Figure 5.3.1: Impulse response function for 12 steps of the response of the term premium to equity returns, labor income growth, the relative T-bill rate and the dividend yield. The dotted lines represent 2 standard deviation (in each direction) confidence intervals.

Example: Impulse Response in Campbell's VAR Campbell's VAR will be used to illustrate impulse response functions. Figure 5.3.1 contains the impulse responses of the term premium to shocks in the four other variables: equity returns, labor income growth, the relative t-bill rate and the dividend rate. The dotted lines represent 2 standard deviation confidence intervals. The term premium decreases subsequent to positive shocks in the market or in labor income. Presumably this indicates that the economy is improving and there are inflationary pressures driving up the short end of the yield curve. Increases in the RTB lead to increases in the term premium as do shocks to the dividend yield.

`site/ambropo/LectureNotes`.

Confidence Intervals Impulse response functions, like the parameters of the VAR, are estimated quantities and subject to statistical variation. Hence, it is a good practice to place confidence bands around impulse response functions to allow anyone digesting your work to know whether an impulse response is large in a statistically meaningful way. Since the parameters of the VAR are asymptotically normal (as long as it is stationary and the innovations are white noise), the impulse responses will also be asymptotically normal by applying a technique known as the delta method (covered in the GMM notes). Unfortunately, the derivation is extremely tedious and has essentially no intuitive value. Interested readers can refer to 11.7 in Hamilton (1994).