Università Commerciale
Luigi Bocconi

MSc. Finance/CLEFIN
2018/2019 Edition

FINANCIAL ECONOMETRICS AND EMPIRICAL FINANCE - MODULE 2

Exam – June 2019

Time Allowed: 85 minutes

Please answer all the questions by writing your answers <u>in the spaces provided</u>. No additional papers will be collected and therefore they will not be marked. You always need to carefully justify your answers and show your work. If you **feel that you **need** to make any assumptions to answer a question, please do so—your assumption will be evaluated along with your answer. The exam is closed book, closed notes. You can withdraw until 10 minutes before the due time.**

Question 1.A (6.5 points)

Consider a bivariate VAR(2) model for the yields of 1-month T-bills and of 10-year Treasury notes (y_t^{1M} and y_t^{10Y}, respectively). Write:

- The structural, unconstrained VAR(2) that includes contemporaneous effects between the two markets.
- The implied, unconstrained reduced-form VAR(2).

Explain through which steps it is possible to transform the structural VAR model into the reduced-form one (algebra is not required, unless it helps you to provide an efficient answer). How would/could you estimate the reduced-form model? Explain what issues/limitations are caused by the transformation of a structural VAR into a reduced-form model. Discuss how these limitations will affect the estimation of the impulse response functions (IRFs).

Debriefing: You were expected to write the following structural model (and not the generic VAR(p) model, be warned):

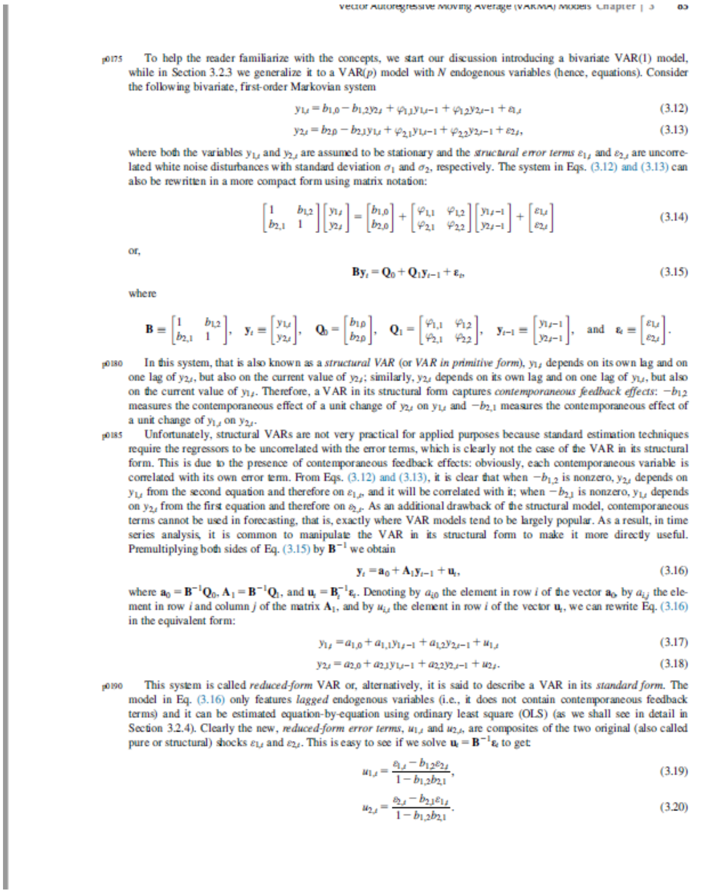
$$\mathbf{B}\mathbf{y}_t = \mathbf{Q}_0 + \mathbf{Q}_1\mathbf{y}_{t-1} + \mathbf{Q}_2\mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t$$

where $\mathbf{B} = \begin{bmatrix} 1 & \mathrm{b}_{1,2} \\ \mathrm{b}_{2,1} & 1 \end{bmatrix}$, $\mathbf{y}_t = \begin{bmatrix} y_t^{1M} \\ y_t^{10Y} \end{bmatrix}$, $\mathbf{Q}_0 = \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix}$, $\mathbf{Q}_1 = \begin{bmatrix} \varphi_{1,1,1} & \varphi_{1,2,1} \\ \varphi_{2,1,1} & \varphi_{2,2,1} \end{bmatrix}$, $\mathbf{Q}_2 = \begin{bmatrix} \varphi_{1,1,2} & \varphi_{1,2,2} \\ \varphi_{2,1,2} & \varphi_{2,2,2} \end{bmatrix}$

and $\boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$. In order to obtain the reduced form, the structural model needs to be multiplied by $\mathbf{B}^{-1}$, which leads to

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \mathbf{u}_t$$

where the errors, $\mathbf{u}_t$ are now a composite of the structural innovations and therefore they are not uncorrelated. See pages 83-86 of the book, copied below for your perusal.



p0175 To help the reader familiarize with the concepts, we start our discussion introducing a bivariate VAR(1) model, while in Section 3.2.3 we generalize it to a VAR(p) model with N endogenous variables (hence, equations). Consider the following bivariate, first-order Markovian system

$$y_{1,t} = b_{1,0} - b_{1,2}y_{2,t} + \varphi_{1,1}y_{1,t-1} + \varphi_{1,2}y_{2,t-1} + \varepsilon_{1,t} \tag{3.12}$$

$$y_{2,t} = b_{2,0} - b_{2,1}y_{1,t} + \varphi_{2,1}y_{1,t-1} + \varphi_{2,2}y_{2,t-1} + \varepsilon_{2,t}, \tag{3.13}$$

where both the variables $y_{1,t}$ and $y_{2,t}$ are assumed to be stationary and the *structural error terms* $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are uncorrelated white noise disturbances with standard deviation σ_1 and σ_2, respectively. The system in Eqs. (3.12) and (3.13) can also be rewritten in a more compact form using matrix notation:

$$\begin{bmatrix} 1 & b_{1,2} \\ b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix} + \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix}\begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \tag{3.14}$$

or,

$$\mathbf{B}\mathbf{y}_t = \mathbf{Q}_0 + \mathbf{Q}_1\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{3.15}$$

where

$$\mathbf{B} = \begin{bmatrix} 1 & b_{1,2} \\ b_{2,1} & 1 \end{bmatrix}, \quad \mathbf{y}_t = \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix}, \quad \mathbf{Q}_0 = \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix}, \quad \mathbf{Q}_1 = \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix}, \quad \mathbf{y}_{t-1} = \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}.$$

p0180 In this system, that is also known as a *structural VAR* (or *VAR in primitive form*), $y_{1,t}$ depends on its own lag and on one lag of $y_{2,t}$, but also on the current value of $y_{2,t}$; similarly, $y_{2,t}$ depends on its own lag and on one lag of $y_{1,t}$, but also on the current value of $y_{1,t}$. Therefore, a VAR in its structural form captures *contemporaneous feedback effects*: $-b_{1,2}$ measures the contemporaneous effect of a unit change of $y_{2,t}$ on $y_{1,t}$ and $-b_{2,1}$ measures the contemporaneous effect of a unit change of $y_{1,t}$ on $y_{2,t}$.

p0185 Unfortunately, structural VARs are not very practical for applied purposes because standard estimation techniques require the regressors to be uncorrelated with the error terms, which is clearly not the case of the VAR in its structural form. This is due to the presence of contemporaneous feedback effects: obviously, each contemporaneous variable is correlated with its own error term. From Eqs. (3.12) and (3.13), it is clear that when $-b_{1,2}$ is nonzero, $y_{2,t}$ depends on $y_{1,t}$ from the second equation and therefore on $\varepsilon_{1,t}$, and it will be correlated with it; when $-b_{2,1}$ is nonzero, $y_{1,t}$ depends on $y_{2,t}$ from the first equation and therefore on $\varepsilon_{2,t}$. As an additional drawback of the structural model, contemporaneous terms cannot be used in forecasting, that is, exactly where VAR models tend to be largely popular. As a result, in time series analysis, it is common to manipulate the VAR in its structural form to make it more directly useful. Premultiplying both sides of Eq. (3.15) by $\mathbf{B}^{-1}$ we obtain

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{u}_t, \tag{3.16}$$

where $\mathbf{a}_0 = \mathbf{B}^{-1}\mathbf{Q}_0$, $\mathbf{A}_1 = \mathbf{B}^{-1}\mathbf{Q}_1$, and $\mathbf{u}_t = \mathbf{B}_t^{-1}\boldsymbol{\varepsilon}_t$. Denoting by $a_{i,0}$ the element in row i of the vector $\mathbf{a}_0$, by $a_{i,j}$ the element in row i and column j of the matrix $\mathbf{A}_1$, and by $u_{i,t}$ the element in row i of the vector $\mathbf{u}_t$, we can rewrite Eq. (3.16) in the equivalent form:

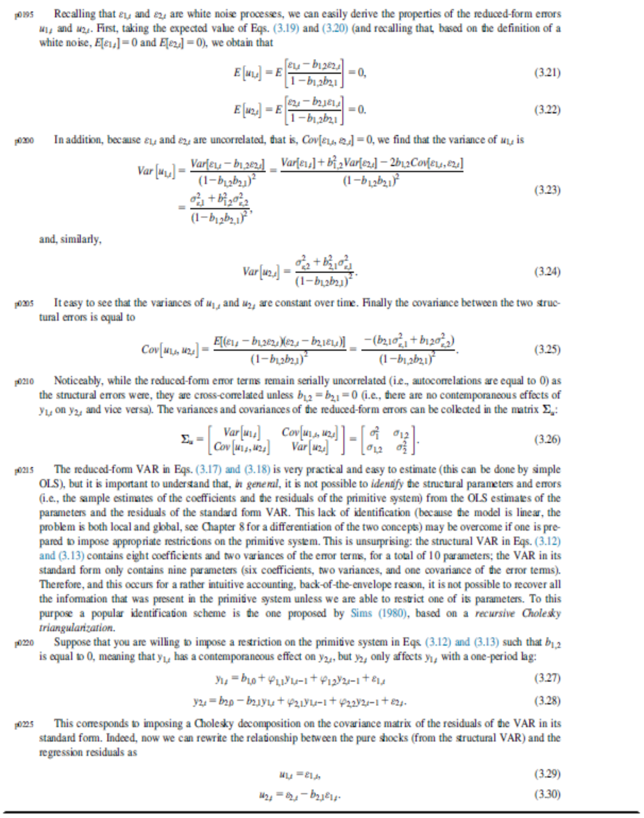
$$y_{1,t} = a_{1,0} + a_{1,1}y_{1,t-1} + a_{1,2}y_{2,t-1} + u_{1,t} \tag{3.17}$$

$$y_{2,t} = a_{2,0} + a_{2,1}y_{1,t-1} + a_{2,2}y_{2,t-1} + u_{2,t}. \tag{3.18}$$

p0190 This system is called *reduced-form* VAR or, alternatively, it is said to describe a VAR in its *standard form*. The model in Eq. (3.16) only features *lagged* endogenous variables (i.e., it does not contain contemporaneous feedback terms) and it can be estimated equation-by-equation using ordinary least square (OLS) (as we shall see in detail in Section 3.2.4). Clearly the new, *reduced-form error terms*, $u_{1,t}$ and $u_{2,t}$, are composites of the two original (also called pure or structural) shocks $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$. This is easy to see if we solve $\mathbf{u}_t = \mathbf{B}^{-1}\boldsymbol{\varepsilon}_t$ to get:

$$u_{1,t} = \frac{\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t}}{1 - b_{1,2}b_{2,1}}, \tag{3.19}$$

$$u_{2,t} = \frac{\varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}}{1 - b_{1,2}b_{2,1}}. \tag{3.20}$$



p0195 Recalling that $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are white noise processes, we can easily derive the properties of the reduced-form errors $u_{1,t}$ and $u_{2,t}$. First, taking the expected value of Eqs. (3.19) and (3.20) (and recalling that, based on the definition of a white noise, $E[\varepsilon_{1,t}] = 0$ and $E[\varepsilon_{2,t}] = 0$), we obtain that

$$E\left[u_{1,t}\right] = E\left[\frac{\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t}}{1 - b_{1,2}b_{2,1}}\right] = 0, \tag{3.21}$$

$$E\left[u_{2,t}\right] = E\left[\frac{\varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}}{1 - b_{1,2}b_{2,1}}\right] = 0. \tag{3.22}$$

p0200 In addition, because $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ are uncorrelated, that is, $Cov[\varepsilon_{1,t}, \varepsilon_{2,t}] = 0$, we find that the variance of $u_{1,t}$ is

$$\begin{aligned} Var\left[u_{1,t}\right] &= \frac{Var[\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t}]}{(1 - b_{1,2}b_{2,1})^2} = \frac{Var[\varepsilon_{1,t}] + b_{1,2}^2 Var[\varepsilon_{2,t}] - 2b_{1,2}Cov[\varepsilon_{1,t}, \varepsilon_{2,t}]}{(1 - b_{1,2}b_{2,1})^2} \\ &= \frac{\sigma_{\varepsilon,1}^2 + b_{1,2}^2\sigma_{\varepsilon,2}^2}{(1 - b_{1,2}b_{2,1})^2}, \end{aligned} \tag{3.23}$$

and, similarly,

$$Var\left[u_{2,t}\right] = \frac{\sigma_{\varepsilon,2}^2 + b_{2,1}^2\sigma_{\varepsilon,1}^2}{(1 - b_{1,2}b_{2,1})^2}. \tag{3.24}$$

p0205 It easy to see that the variances of $u_{1,t}$ and $u_{2,t}$ are constant over time. Finally the covariance between the two structural errors is equal to

$$Cov\left[u_{1,t}, u_{2,t}\right] = \frac{E[(\varepsilon_{1,t} - b_{1,2}\varepsilon_{2,t})(\varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t})]}{(1 - b_{1,2}b_{2,1})^2} = \frac{-(b_{2,1}\sigma_{\varepsilon,1}^2 + b_{1,2}\sigma_{\varepsilon,2}^2)}{(1 - b_{1,2}b_{2,1})^2}. \tag{3.25}$$

p0210 Noticeably, while the reduced-form error terms remain serially uncorrelated (i.e., autocorrelations are equal to 0) as the structural errors were, they are cross-correlated unless $b_{1,2} = b_{2,1} = 0$ (i.e., there are no contemporaneous effects of $y_{1,t}$ on $y_{2,t}$ and vice versa). The variances and covariances of the reduced-form errors can be collected in the matrix $\boldsymbol{\Sigma}_u$:

$$\boldsymbol{\Sigma}_u = \begin{bmatrix} Var\left[u_{1,t}\right] & Cov\left[u_{1,t}, u_{2,t}\right] \\ Cov\left[u_{1,t}, u_{2,t}\right] & Var\left[u_{2,t}\right] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}. \tag{3.26}$$

p0215 The reduced-form VAR in Eqs. (3.17) and (3.18) is very practical and easy to estimate (this can be done by simple OLS), but it is important to understand that, *in general*, it is not possible to *identify* the structural parameters and errors (i.e., the sample estimates of the coefficients and the residuals of the primitive system) from the OLS estimates of the parameters and the residuals of the standard form VAR. This lack of identification (because the model is linear, the problem is both local and global, see Chapter 8 for a differentiation of the two concepts) may be overcome if one is prepared to impose appropriate restrictions on the primitive system. This is unsurprising: the structural VAR in Eqs. (3.12) and (3.13) contains eight coefficients and two variances of the error terms, for a total of 10 parameters; the VAR in its standard form only contains nine parameters (six coefficients, two variances, and one covariance of the error terms). Therefore, and this occurs for a rather intuitive accounting, back-of-the-envelope reason, it is not possible to recover all the information that was present in the primitive system unless we are able to restrict one of its parameters. To this purpose a popular identification scheme is the one proposed by Sims (1980), based on a *recursive Cholesky triangularization*.

p0220 Suppose that you are willing to impose a restriction on the primitive system in Eqs. (3.12) and (3.13) such that $b_{1,2}$ is equal to 0, meaning that $y_{1,t}$ has a contemporaneous effect on $y_{2,t}$, but $y_{2,t}$ only affects $y_{1,t}$ with a one-period lag:

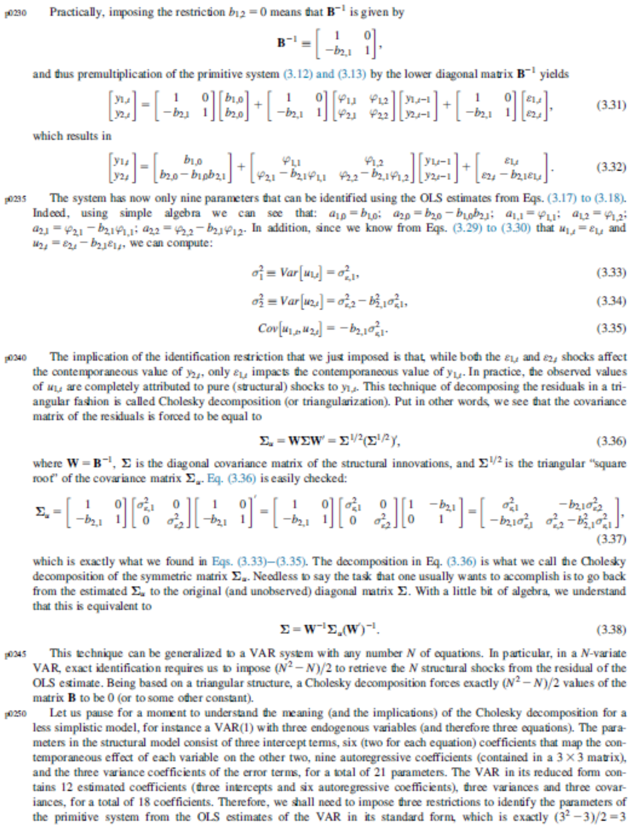
$$y_{1,t} = b_{1,0} + \varphi_{1,1}y_{1,t-1} + \varphi_{1,2}y_{2,t-1} + \varepsilon_{1,t} \tag{3.27}$$

$$y_{2,t} = b_{2,0} - b_{2,1}y_{1,t} + \varphi_{2,1}y_{1,t-1} + \varphi_{2,2}y_{2,t-1} + \varepsilon_{2,t}. \tag{3.28}$$

p0225 This corresponds to imposing a Cholesky decomposition on the covariance matrix of the residuals of the VAR in its standard form. Indeed, now we can rewrite the relationship between the pure shocks (from the structural VAR) and the regression residuals as

$$u_{1,t} = \varepsilon_{1,t}, \tag{3.29}$$

$$u_{2,t} = \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}. \tag{3.30}$$



p0230 Practically, imposing the restriction $b_{1,2} = 0$ means that $\mathbf{B}^{-1}$ is given by

$$\mathbf{B}^{-1} = \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix},$$

and thus premultiplication of the primitive system (3.12) and (3.13) by the lower diagonal matrix $\mathbf{B}^{-1}$ yields

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix}\begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}, \tag{3.31}$$

which results in

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} b_{1,0} \\ b_{2,0} - b_{1,0}b_{2,1} \end{bmatrix} + \begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} - b_{2,1}\varphi_{1,1} & \varphi_{2,2} - b_{2,1}\varphi_{1,2} \end{bmatrix}\begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t} \end{bmatrix}. \tag{3.32}$$

p0235 The system has now only nine parameters that can be identified using the OLS estimates from Eqs. (3.17) to (3.18). Indeed, using simple algebra we can see that: $a_{1,0} = b_{1,0}$; $a_{2,0} = b_{2,0} - b_{1,0}b_{2,1}$; $a_{1,1} = \varphi_{1,1}$; $a_{1,2} = \varphi_{1,2}$; $a_{2,1} = \varphi_{2,1} - b_{2,1}\varphi_{1,1}$; $a_{2,2} = \varphi_{2,2} - b_{2,1}\varphi_{1,2}$. In addition, since we know from Eqs. (3.29) to (3.30) that $u_{1,t} = \varepsilon_{1,t}$ and $u_{2,t} = \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t}$, we can compute:

$$\sigma_1^2 \equiv Var\left[u_{1,t}\right] = \sigma_{\varepsilon,1}^2, \tag{3.33}$$

$$\sigma_2^2 \equiv Var\left[u_{2,t}\right] = \sigma_{\varepsilon,2}^2 - b_{2,1}^2\sigma_{\varepsilon,1}^2, \tag{3.34}$$

$$Cov\left[u_{1,t}, u_{2,t}\right] = -b_{2,1}\sigma_{\varepsilon,1}^2. \tag{3.35}$$

p0240 The implication of the identification restriction that we just imposed is that, while both the $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ shocks affect the contemporaneous value of $y_{2,t}$, only $\varepsilon_{1,t}$ impacts the contemporaneous value of $y_{1,t}$. In practice, the observed values of $u_{1,t}$ are completely attributed to pure (structural) shocks to $y_{1,t}$. This technique of decomposing the residuals in a triangular fashion is called Cholesky decomposition (or triangularization). Put in other words, we see that the covariance matrix of the residuals is forced to be equal to

$$\boldsymbol{\Sigma}_u = \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}' = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})', \tag{3.36}$$

where $\mathbf{W} = \mathbf{B}^{-1}$, $\boldsymbol{\Sigma}$ is the diagonal covariance matrix of the structural innovations, and $\boldsymbol{\Sigma}^{1/2}$ is the triangular "square root" of the covariance matrix $\boldsymbol{\Sigma}_u$. Eq. (3.36) is easily checked:

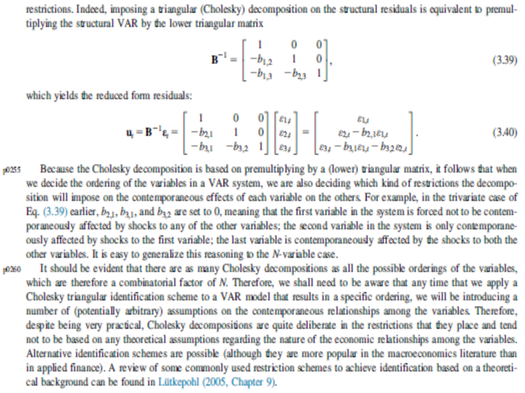
$$\boldsymbol{\Sigma}_u = \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} \sigma_{\varepsilon,1}^2 & 0 \\ 0 & \sigma_{\varepsilon,2}^2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix}' = \begin{bmatrix} 1 & 0 \\ -b_{2,1} & 1 \end{bmatrix}\begin{bmatrix} \sigma_{\varepsilon,1}^2 & 0 \\ 0 & \sigma_{\varepsilon,2}^2 \end{bmatrix}\begin{bmatrix} 1 & -b_{2,1} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_{\varepsilon,1}^2 & -b_{2,1}\sigma_{\varepsilon,2}^2 \\ -b_{2,1}\sigma_{\varepsilon,1}^2 & \sigma_{\varepsilon,2}^2 - b_{2,1}^2\sigma_{\varepsilon,1}^2 \end{bmatrix}, \tag{3.37}$$

which is exactly what we found in Eqs. (3.33)–(3.35). The decomposition in Eq. (3.36) is what we call the Cholesky decomposition of the symmetric matrix $\boldsymbol{\Sigma}_u$. Needless to say the task that one usually wants to accomplish is to go back from the estimated $\boldsymbol{\Sigma}_u$ to the original (and unobserved) diagonal matrix $\boldsymbol{\Sigma}$. With a little bit of algebra, we understand that this is equivalent to

$$\boldsymbol{\Sigma} = \mathbf{W}^{-1}\boldsymbol{\Sigma}_u(\mathbf{W}')^{-1}. \tag{3.38}$$

p0245 This technique can be generalized to a VAR system with any number N of equations. In particular, in a N-variate VAR, exact identification requires us to impose $(N^2 - N)/2$ to retrieve the N structural shocks from the residual of the OLS estimate. Being based on a triangular structure, a Cholesky decomposition forces exactly $(N^2 - N)/2$ values of the matrix $\mathbf{B}$ to be 0 (or to some other constant).

p0250 Let us pause for a moment to understand the meaning (and the implications) of the Cholesky decomposition for a less simplistic model, for instance a VAR(1) with three endogenous variables (and therefore three equations). The parameters in the structural model consist of three intercept terms, six (two for each equation) coefficients that map the contemporaneous effect of each variable on the other two, nine autoregressive coefficients (contained in a 3 × 3 matrix), and the three variance coefficients of the error terms, for a total of 21 parameters. The VAR in its reduced form contains 12 estimated coefficients (three intercepts and six autoregressive coefficients), three variances and three covariances, for a total of 18 coefficients. Therefore, we shall need to impose three restrictions to identify the parameters of the primitive system from the OLS estimates of the VAR in its standard form, which is exactly $(3^2 - 3)/2 = 3$



restrictions. Indeed, imposing a triangular (Cholesky) decomposition on the structural residuals is equivalent to premultiplying the structural VAR by the lower triangular matrix

$$\mathbf{B}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -b_{1,2} & 1 & 0 \\ -b_{1,3} & -b_{2,3} & 1 \end{bmatrix}, \tag{3.39}$$

which yields the reduced form residuals:

$$\mathbf{u}_t = \mathbf{B}^{-1}\boldsymbol{\varepsilon}_t = \begin{bmatrix} 1 & 0 & 0 \\ -b_{2,1} & 1 & 0 \\ -b_{3,1} & -b_{3,2} & 1 \end{bmatrix}\begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} - b_{2,1}\varepsilon_{1,t} \\ \varepsilon_{3,t} - b_{3,1}\varepsilon_{1,t} - b_{3,2}\varepsilon_{2,t} \end{bmatrix}. \tag{3.40}$$

p0255 Because the Cholesky decomposition is based on premultiplying by a (lower) triangular matrix, it follows that when we decide the ordering of the variables in a VAR system, we are also deciding which kind of restrictions the decomposition will impose on the contemporaneous effects of each variable on the others. For example, in the trivariate case of Eq. (3.39) earlier, $b_{2,1}$, $b_{3,1}$, and $b_{3,2}$ are set to 0, meaning that the first variable in the system is forced not to be contemporaneously affected by shocks to any of the other variables; the second variable in the system is only contemporaneously affected by shocks to the first variable; the last variable is contemporaneously affected by the shocks to both the other variables. It is easy to generalize this reasoning to the N-variable case.

p0260 It should be evident that there are as many Cholesky decompositions as all the possible orderings of the variables, which are therefore a combinatorial factor of N. Therefore, we shall need to be aware that any time that we apply a Cholesky triangular identification scheme to a VAR model that results in a specific ordering, we will be introducing a number of (potentially arbitrary) assumptions on the contemporaneous relationships among the variables. Therefore, despite being very practical, Cholesky decompositions are quite deliberate in the restrictions that they place and tend not to be based on any theoretical assumptions regarding the nature of the economic relationships among the variables. Alternative identification schemes are possible (although they are more popular in the macroeconomics literature than in applied finance). A review of some commonly used restriction schemes to achieve identification based on a theoretical background can be found in Lütkepohl (2005, Chapter 9).

Question 1.B (2 points)

Mitchell Dot Rink, a summer analyst at Gordon Socks, is estimating the VAR(2) model discussed in Question 1.A for the yields of 1-month T-bills and 10-year Treasury notes. He claims that he knows that, based on accepted theory, while shocks to 1-month T-bill yields do immediately affect 10-year Treasury yields, shocks to 10-year Treasury yields affect 1-month T-bill yield only with one lag. Therefore, so he claims, a Cholesky decomposition is not needed to identify such a model. Do you agree with Mitchell's conclusions? Carefully explain your answer.

Debriefing: MDR is not correct. The fact that he knows a theoretical relationship does not guarantee that he will find uncorrelated errors when he estimates the reduced form model. However, the theory tells him exactly which restriction he should impose, i.e., $b_{1,2} = 0$ (referring to the model specified in question 1.A. This is equivalent to a Cholesky decomposition where the 1-month T-bill comes first in the ordering.

Question 1.C (1.5 points)

Mitchell has now extended his VAR(2) model to include also 1- and 5-year Treasury yields and he would like to test Granger causality among the four series. Therefore, he has produced the output below. Looking at the table he has concluded that the 1-month yield is not Granger-caused by any other series, while it Granger-causes all of the others. After having briefly defined what Granger causality is, discuss whether you agree with Mitchell's conclusions. Clearly justify your answer.

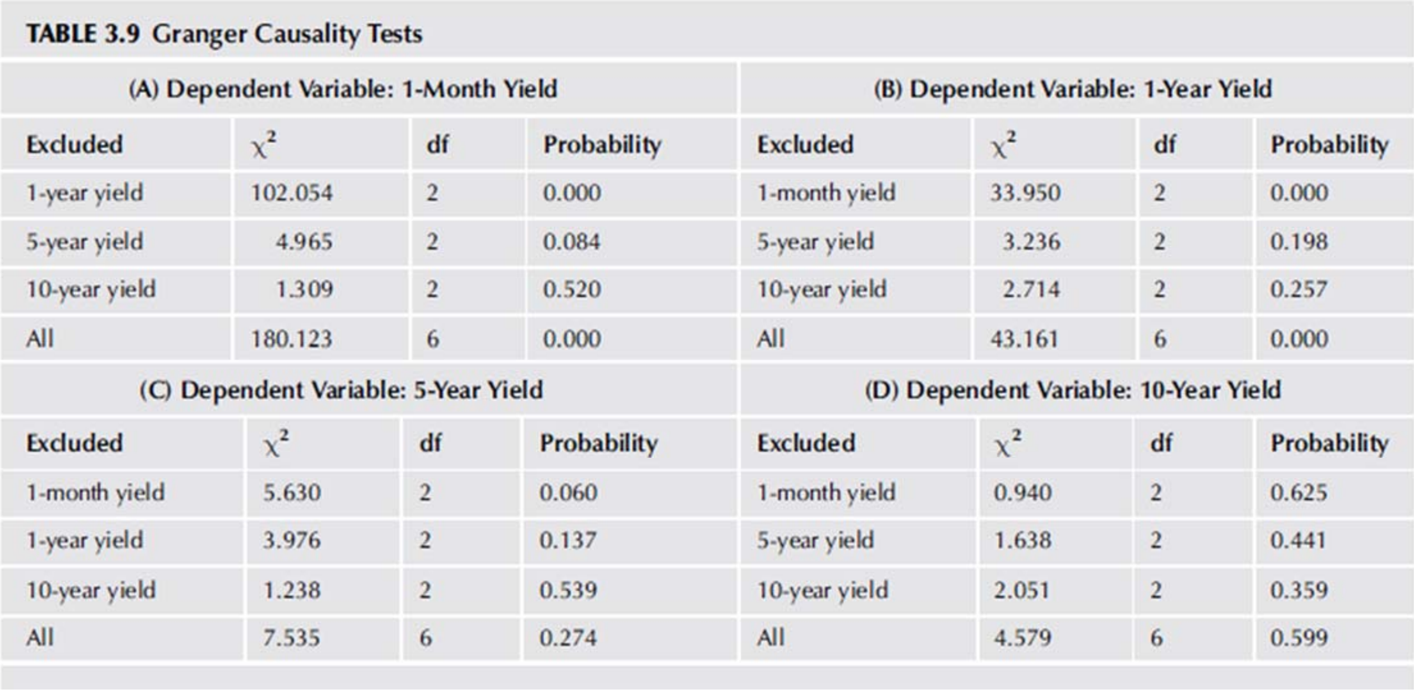


TABLE 3.9 Granger Causality Tests

(A) Dependent Variable: 1-Month Yield

Excluded	χ^2	df	Probability
1-year yield	102.054	2	0.000
5-year yield	4.965	2	0.084
10-year yield	1.309	2	0.520
All	180.123	6	0.000

(B) Dependent Variable: 1-Year Yield

Excluded	χ^2	df	Probability
1-month yield	33.950	2	0.000
5-year yield	3.236	2	0.198
10-year yield	2.714	2	0.257
All	43.161	6	0.000

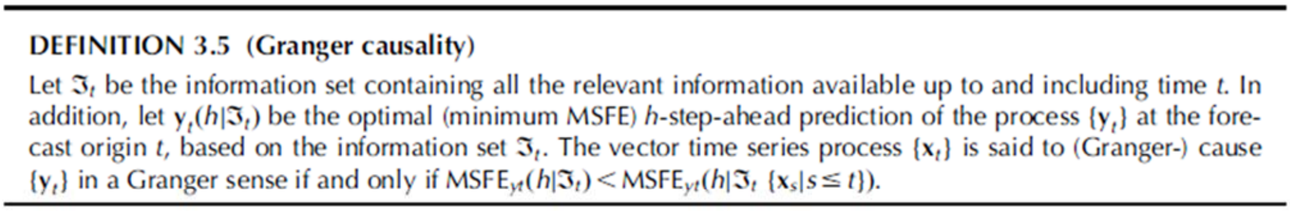
(C) Dependent Variable: 5-Year Yield

Excluded	χ^2	df	Probability
1-month yield	5.630	2	0.060
1-year yield	3.976	2	0.137
10-year yield	1.238	2	0.539
All	7.535	6	0.274

(D) Dependent Variable: 10-Year Yield

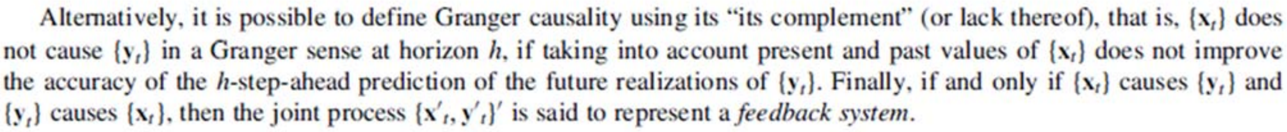
Excluded	χ^2	df	Probability
1-month yield	0.940	2	0.625
5-year yield	1.638	2	0.441
10-year yield	2.051	2	0.359
All	4.579	6	0.599

Debriefing: The output should look familiar as it comes from Example 3.9 in the book. As one can see, Mitchell is not right because the 1-month yield is Granger caused by the 1-year yield (and, to some extent, by the 5-year yield, at a 10% confidence level). In addition, it does not seem to Granger-cause the 10-year yield. Definition of Granger causality can be found at page 108 of the book (copied below).



DEFINITION 3.5 (Granger causality)

Let $\Im_t$ be the information set containing all the relevant information available up to and including time t. In addition, let $\mathbf{y}_t(h|\Im_t)$ be the optimal (minimum MSFE) h-step-ahead prediction of the process $\{\mathbf{y}_t\}$ at the forecast origin t, based on the information set $\Im_t$. The vector time series process $\{\mathbf{x}_t\}$ is said to (Granger-) cause $\{\mathbf{y}_t\}$ in a Granger sense if and only if $\text{MSFE}_{yt}(h|\Im_t) < \text{MSFE}_{yt}(h|\Im_t \ \{\mathbf{x}_s | s \leq t\})$.

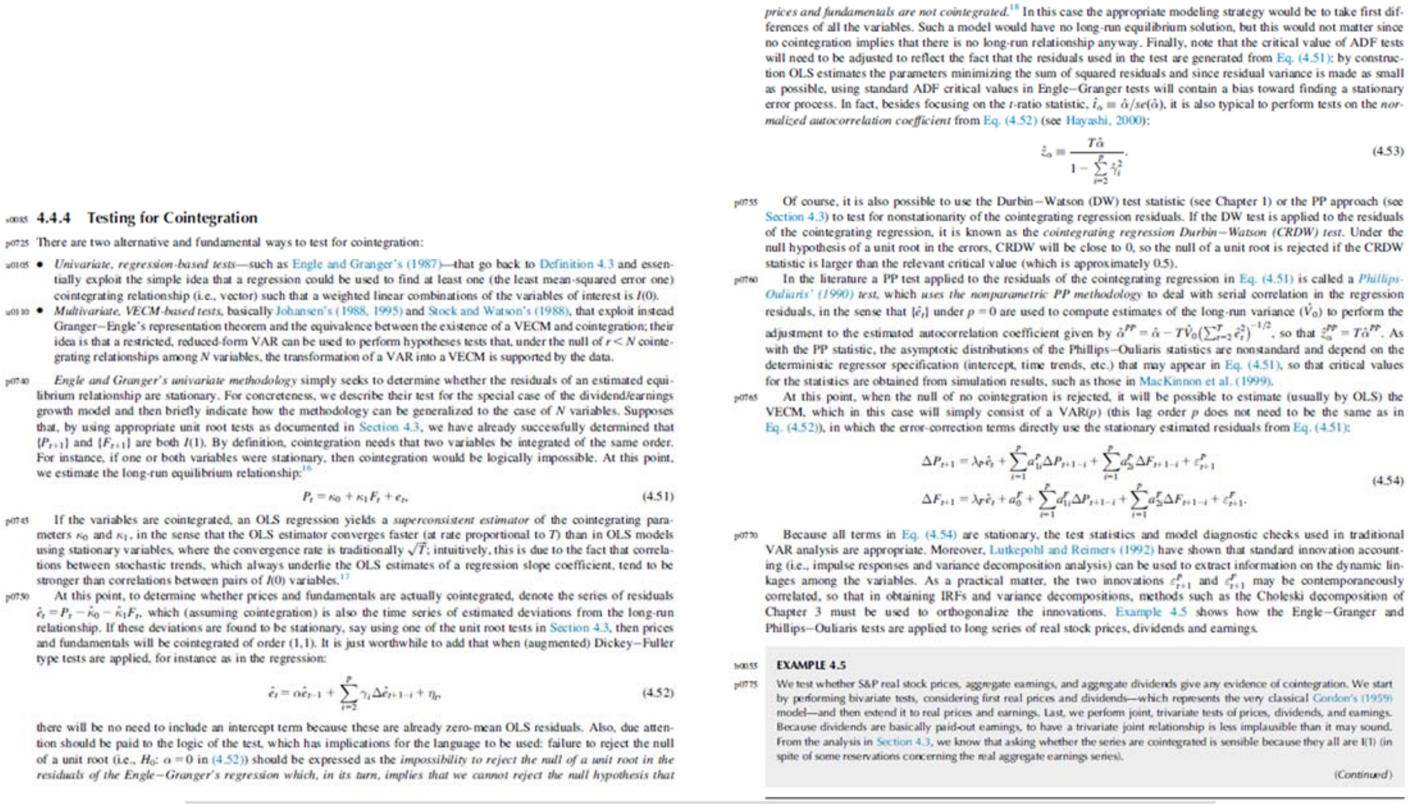


Alternatively, it is possible to define Granger causality using its "its complement" (or lack thereof), that is, $\{\mathbf{x}_t\}$ does not cause $\{\mathbf{y}_t\}$ in a Granger sense at horizon h, if taking into account present and past values of $\{\mathbf{x}_t\}$ does not improve the accuracy of the h-step-ahead prediction of the future realizations of $\{\mathbf{y}_t\}$. Finally, if and only if $\{\mathbf{x}_t\}$ causes $\{\mathbf{y}_t\}$ and $\{\mathbf{y}_t\}$ causes $\{\mathbf{x}_t\}$, then the joint process $\{\mathbf{x}'_t, \mathbf{y}'_t\}'$ is said to represent a *feedback system*.

Question 2.A (6.5 points)

Describe the two alternative (univariate and multivariate) ways to test for cointegration. In particular, be sure to discuss when each of the two tests is most appropriate and what is the rationale behind each of them, together with the steps that are required to implement the tests. Also discuss what are the main drawbacks of the Engle and Granger's test. What does it mean that Engle and Granger's test suffers from a "generated regressors problem"? Be sure to carefully justify your answer.

Debriefing:



s0085 **4.4.4 Testing for Cointegration**

p0725 There are two alternative and fundamental ways to test for cointegration:

u0105 - *Univariate, regression-based tests*—such as Engle and Granger's (1987)—that go back to Definition 4.3 and essentially exploit the simple idea that a regression could be used to find at least one (the least mean-squared error one) cointegrating relationship (i.e., vector) such that a weighted linear combinations of the variables of interest is $I(0)$.

u0110 - *Multivariate, VECM-based tests,* basically Johansen's (1988, 1995) and Stock and Watson's (1988), that exploit instead Granger–Engle's representation theorem and the equivalence between the existence of a VECM and cointegration; their idea is that a restricted, reduced-form VAR can be used to perform hypotheses tests that, under the null of $r < N$ cointegrating relationships among N variables, the transformation of a VAR into a VECM is supported by the data.

p0740 *Engle and Granger's univariate methodology* simply seeks to determine whether the residuals of an estimated equilibrium relationship are stationary. For concreteness, we describe their test for the special case of the dividend/earnings growth model and then briefly indicate how the methodology can be generalized to the case of N variables. Supposes that, by using appropriate unit root tests as documented in Section 4.3, we have already successfully determined that $\{P_{t+1}\}$ and $\{F_{t+1}\}$ are both $I(1)$. By definition, cointegration needs that two variables be integrated of the same order. For instance, if one or both variables were stationary, then cointegration would be logically impossible. At this point, we estimate the long-run equilibrium relationship:[16]

$$P_t = \kappa_0 + \kappa_1 F_t + e_t, \tag{4.51}$$

p0745 If the variables are cointegrated, an OLS regression yields a *superconsistent estimator* of the cointegrating parameters κ_0 and κ_1, in the sense that the OLS estimator converges faster (at rate proportional to T) than in OLS models using stationary variables, where the convergence rate is traditionally $\sqrt{T}$; intuitively, this is due to the fact that correlations between stochastic trends, which always underlie the OLS estimates of a regression slope coefficient, tend to be stronger than correlations between pairs of $I(0)$ variables.[17]

p0750 At this point, to determine whether prices and fundamentals are actually cointegrated, denote the series of residuals $\hat{e}_t = P_t - \hat{\kappa}_0 - \hat{\kappa}_1 F_t$, which (assuming cointegration) is also the time series of estimated deviations from the long-run relationship. If these deviations are found to be stationary, say using one of the unit root tests in Section 4.3, then prices and fundamentals will be cointegrated of order (1,1). It is just worthwhile to add that when (augmented) Dickey–Fuller type tests are applied, for instance as in the regression:

$$\hat{e}_t = \alpha \hat{e}_{t-1} + \sum_{i=2}^{p} \gamma_i \Delta \hat{e}_{t+1-i} + \eta_t, \tag{4.52}$$

there will be no need to include an intercept term because these are already zero-mean OLS residuals. Also, due attention should be paid to the logic of the test, which has implications for the language to be used: failure to reject the null of a unit root (i.e., H_0: $\alpha = 0$ in (4.52)) should be expressed as the *impossibility to reject the null of a unit root in the residuals of the Engle–Granger's regression which, in its turn, implies that we cannot reject the null hypothesis that prices and fundamentals are not cointegrated.*[18] In this case the appropriate modeling strategy would be to take first differences of all the variables. Such a model would have no long-run equilibrium solution, but this would not matter since no cointegration implies that there is no long-run relationship anyway. Finally, note that the critical value of ADF tests will need to be adjusted to reflect the fact that the residuals used in the test are generated from Eq. (4.51): by construction OLS estimates the parameters minimizing the sum of squared residuals and since residual variance is made as small as possible, using standard ADF critical values in Engle–Granger tests will contain a bias toward finding a stationary error process. In fact, besides focusing on the t-ratio statistic, $\hat{t}_\alpha \equiv \hat{\alpha}/se(\hat{\alpha})$, it is also typical to perform tests on the *normalized autocorrelation coefficient* from Eq. (4.52) (see Hayashi, 2000):

$$\hat{z}_\alpha \equiv \frac{T\hat{\alpha}}{1 - \sum_{i=2}^{p} \hat{\gamma}_i^2}. \tag{4.53}$$

p0755 Of course, it is also possible to use the Durbin–Watson (DW) test statistic (see Chapter 1) or the PP approach (see Section 4.3) to test for nonstationarity of the cointegrating regression residuals. If the DW test is applied to the residuals of the cointegrating regression, it is known as the *cointegrating regression Durbin–Watson (CRDW) test*. Under the null hypothesis of a unit root in the errors, CRDW will be close to 0, so the null of a unit root is rejected if the CRDW statistic is larger than the relevant critical value (which is approximately 0.5).

p0760 In the literature a PP test applied to the residuals of the cointegrating regression in Eq. (4.51) is called a *Phillips-Ouliaris' (1990) test,* which *uses the nonparametric PP methodology* to deal with serial correlation in the regression residuals, in the sense that $\{\hat{e}_t\}$ under $p = 0$ are used to compute estimates of the long-run variance ($\hat{V}_0$) to perform the adjustment to the estimated autocorrelation coefficient given by $\hat{\alpha}^{PP} = \hat{\alpha} - T\hat{V}_0\left(\sum_{t=2}^{T} \hat{e}_t^2\right)^{-1/2}$, so that $\hat{z}_\alpha^{PP} = T\hat{\alpha}^{PP}$. As with the PP statistic, the asymptotic distributions of the Phillips–Ouliaris statistics are nonstandard and depend on the deterministic regressor specification (intercept, time trends, etc.) that may appear in Eq. (4.51), so that critical values for the statistics are obtained from simulation results, such as those in MacKinnon et al. (1999).

p0765 At this point, when the null of no cointegration is rejected, it will be possible to estimate (usually by OLS) the VECM, which in this case will simply consist of a VAR(p) (this lag order p does not need to be the same as in Eq. (4.52)), in which the error-correction terms directly use the stationary estimated residuals from Eq. (4.51):

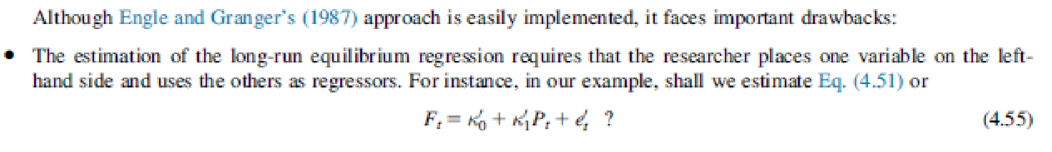
$$\begin{aligned} \Delta P_{t+1} &= \lambda_P \hat{e}_t + \sum_{i=1}^{p} a_{1i}^P \Delta P_{t+1-i} + \sum_{i=1}^{p} a_{2i}^P \Delta F_{t+1-i} + \varepsilon_{t+1}^P \\ \Delta F_{t+1} &= \lambda_F \hat{e}_t + a_0^F + \sum_{i=1}^{p} a_{1i}^F \Delta P_{t+1-i} + \sum_{i=1}^{p} a_{2i}^F \Delta F_{t+1-i} + \varepsilon_{t+1}^F. \end{aligned} \tag{4.54}$$

p0770 Because all terms in Eq. (4.54) are stationary, the test statistics and model diagnostic checks used in traditional VAR analysis are appropriate. Moreover, Lutkepohl and Reimers (1992) have shown that standard innovation accounting (i.e., impulse responses and variance decomposition analysis) can be used to extract information on the dynamic linkages among the variables. As a practical matter, the two innovations ε_{t+1}^P and ε_{t+1}^F may be contemporaneously correlated, so that in obtaining IRFs and variance decompositions, methods such as the Choleski decomposition of Chapter 3 must be used to orthogonalize the innovations. Example 4.5 shows how the Engle–Granger and Phillips–Ouliaris tests are applied to long series of real stock prices, dividends and earnings.

b0055 **EXAMPLE 4.5**

p0775 We test whether S&P real stock prices, aggregate earnings, and aggregate dividends give any evidence of cointegration. We start by performing bivariate tests, considering first real prices and dividends—which represents the very classical Gordon's (1959) model—and then extend it to real prices and earnings. Last, we perform joint, trivariate tests of prices, dividends, and earnings. Because dividends are basically paid-out earnings, to have a trivariate joint relationship is less implausible than it may sound. From the analysis in Section 4.3, we know that asking whether the series are cointegrated is sensible because they all are I(1) (in spite of some reservations concerning the real aggregate earnings series).

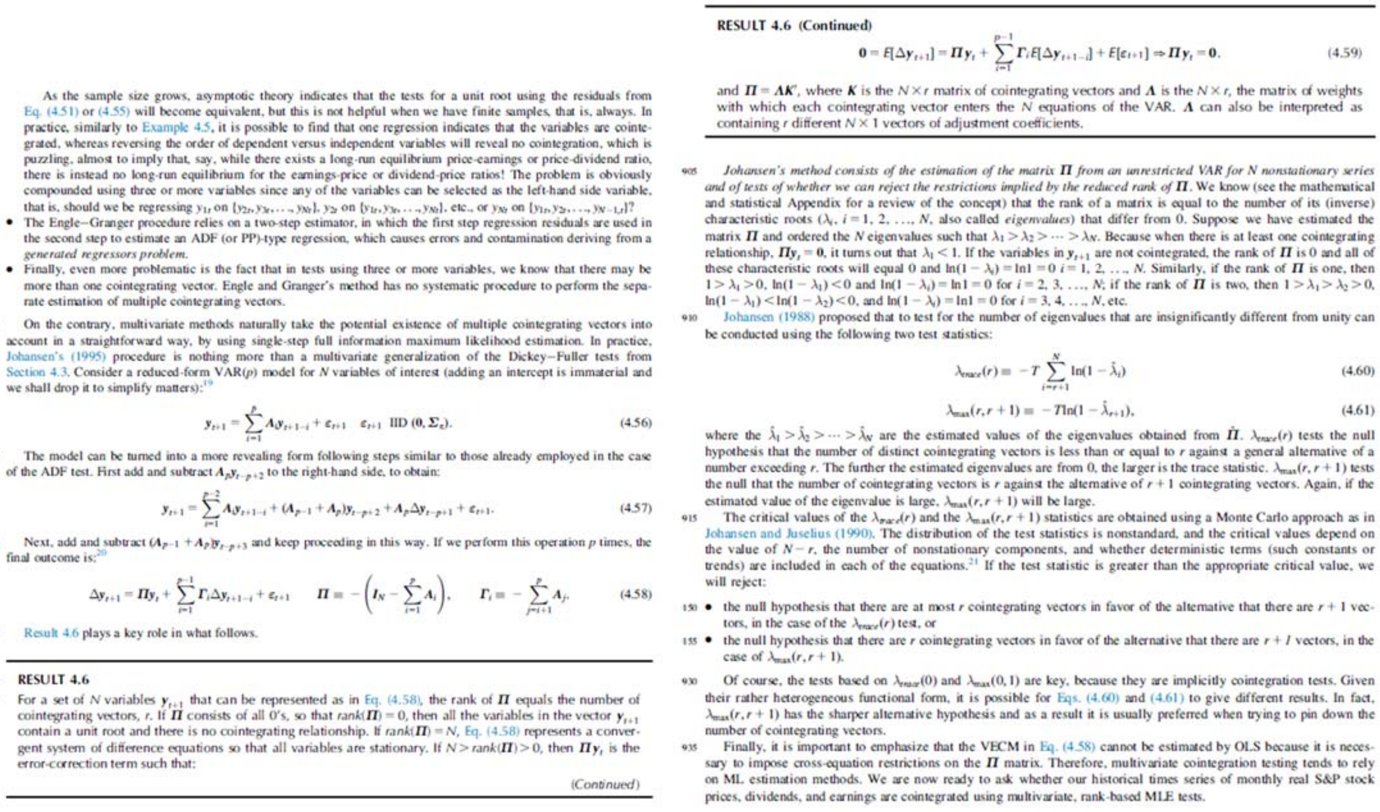
(Continued)



Although Engle and Granger's (1987) approach is easily implemented, it faces important drawbacks:

- The estimation of the long-run equilibrium regression requires that the researcher places one variable on the left-hand side and uses the others as regressors. For instance, in our example, shall we estimate Eq. (4.51) or

$$F_t = \kappa'_0 + \kappa'_1 P_t + e'_t \quad ? \tag{4.55}$$



As the sample size grows, asymptotic theory indicates that the tests for a unit root using the residuals from Eq. (4.51) or (4.55) will become equivalent, but this is not helpful when we have finite samples, that is, always. In practice, similarly to Example 4.5, it is possible to find that one regression indicates that the variables are cointegrated, whereas reversing the order of dependent versus independent variables will reveal no cointegration, which is puzzling, almost to imply that, say, while there exists a long-run equilibrium price-earnings or price-dividend ratio, there is instead no long-run equilibrium for the earnings-price or dividend-price ratios! The problem is obviously compounded using three or more variables since any of the variables can be selected as the left-hand side variable, that is, should we be regressing y_{1t} on $\{y_{2t}, y_{3t}, \ldots, y_{Nt}\}$, y_{2t} on $\{y_{1t}, y_{3t}, \ldots, y_{Nt}\}$, etc., or y_{Nt} on $\{y_{1t}, y_{2t}, \ldots, y_{N-1,t}\}$?

- The Engle–Granger procedure relies on a two-step estimator, in which the first step regression residuals are used in the second step to estimate an ADF (or PP)-type regression, which causes errors and contamination deriving from a *generated regressors problem*.
- Finally, even more problematic is the fact that in tests using three or more variables, we know that there may be more than one cointegrating vector. Engle and Granger's method has no systematic procedure to perform the separate estimation of multiple cointegrating vectors.

On the contrary, multivariate methods naturally take the potential existence of multiple cointegrating vectors into account in a straightforward way, by using single-step full information maximum likelihood estimation. In practice, Johansen's (1995) procedure is nothing more than a multivariate generalization of the Dickey–Fuller tests from Section 4.3. Consider a reduced-form VAR(p) model for N variables of interest (adding an intercept is immaterial and we shall drop it to simplify matters):[19]

$$\mathbf{y}_{t+1} = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{y}_{t+1-i} + \boldsymbol{\varepsilon}_{t+1} \quad \boldsymbol{\varepsilon}_{t+1} \ \text{IID}\ (\mathbf{0}, \boldsymbol{\Sigma}_{\varepsilon}). \tag{4.56}$$

The model can be turned into a more revealing form following steps similar to those already employed in the case of the ADF test. First add and subtract $\mathbf{A}_p \mathbf{y}_{t-p+2}$ to the right-hand side, to obtain:

$$\mathbf{y}_{t+1} = \sum_{i=1}^{p-2} \mathbf{A}_i \mathbf{y}_{t+1-i} + (\mathbf{A}_{p-1} + \mathbf{A}_p)\mathbf{y}_{t-p+2} + \mathbf{A}_p \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_{t+1}. \tag{4.57}$$

Next, add and subtract $(\mathbf{A}_{p-1} + \mathbf{A}_p)\mathbf{y}_{t-p+3}$ and keep proceeding in this way. If we perform this operation p times, the final outcome is:[20]

$$\Delta \mathbf{y}_{t+1} = \boldsymbol{\Pi} \mathbf{y}_t + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta \mathbf{y}_{t+1-i} + \boldsymbol{\varepsilon}_{t+1} \qquad \boldsymbol{\Pi} \equiv -\left(\mathbf{I}_N - \sum_{i=1}^{p} \mathbf{A}_i\right), \qquad \boldsymbol{\Gamma}_i \equiv -\sum_{j=i+1}^{p} \mathbf{A}_j. \tag{4.58}$$

Result 4.6 plays a key role in what follows.

RESULT 4.6

For a set of N variables $\mathbf{y}_{t+1}$ that can be represented as in Eq. (4.58), the rank of $\boldsymbol{\Pi}$ equals the number of cointegrating vectors, r. If $\boldsymbol{\Pi}$ consists of all 0's, so that $rank(\boldsymbol{\Pi}) = 0$, then all the variables in the vector $\mathbf{y}_{t+1}$ contain a unit root and there is no cointegrating relationship. If $rank(\boldsymbol{\Pi}) = N$, Eq. (4.58) represents a convergent system of difference equations so that all variables are stationary. If $N > rank(\boldsymbol{\Pi}) > 0$, then $\boldsymbol{\Pi}\mathbf{y}_t$ is the error-correction term such that:

(Continued)

RESULT 4.6 (Continued)

$$\mathbf{0} = E[\Delta \mathbf{y}_{t+1}] = \boldsymbol{\Pi} \mathbf{y}_t + \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i E[\Delta \mathbf{y}_{t+1-i}] + E[\boldsymbol{\varepsilon}_{t+1}] \Rightarrow \boldsymbol{\Pi} \mathbf{y}_t = \mathbf{0}. \tag{4.59}$$

and $\boldsymbol{\Pi} = \boldsymbol{\Lambda} \mathbf{K}'$, where $\mathbf{K}$ is the $N \times r$ matrix of cointegrating vectors and $\boldsymbol{\Lambda}$ is the $N \times r$, the matrix of weights with which each cointegrating vector enters the N equations of the VAR. $\boldsymbol{\Lambda}$ can also be interpreted as containing r different $N \times 1$ vectors of adjustment coefficients.

Johansen's method consists of the estimation of the matrix $\boldsymbol{\Pi}$ *from an unrestricted VAR for N nonstationary series and of tests of whether we can reject the restrictions implied by the reduced rank of* $\boldsymbol{\Pi}$. We know (see the mathematical and statistical Appendix for a review of the concept) that the rank of a matrix is equal to the number of its (inverse) characteristic roots (λ_i, $i = 1, 2, \ldots, N$, also called *eigenvalues*) that differ from 0. Suppose we have estimated the matrix $\boldsymbol{\Pi}$ and ordered the N eigenvalues such that $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Because when there is at least one cointegrating relationship, $\boldsymbol{\Pi}\mathbf{y}_t = \mathbf{0}$, it turns out that $\lambda_1 < 1$. If the variables in $\mathbf{y}_{t+1}$ are not cointegrated, the rank of $\boldsymbol{\Pi}$ is 0 and all of these characteristic roots will equal 0 and $\ln(1 - \lambda_i) = \ln 1 = 0$ $i = 1, 2, \ldots, N$. Similarly, if the rank of $\boldsymbol{\Pi}$ is one, then $1 > \lambda_1 > 0$, $\ln(1 - \lambda_1) < 0$ and $\ln(1 - \lambda_i) = \ln 1 = 0$ for $i = 2, 3, \ldots, N$; if the rank of $\boldsymbol{\Pi}$ is two, then $1 > \lambda_1 > \lambda_2 > 0$, $\ln(1 - \lambda_1) < \ln(1 - \lambda_2) < 0$, and $\ln(1 - \lambda_i) = \ln 1 = 0$ for $i = 3, 4, \ldots, N$, etc.

Johansen (1988) proposed that to test for the number of eigenvalues that are insignificantly different from unity can be conducted using the following two test statistics:

$$\lambda_{trace}(r) \equiv -T \sum_{i=r+1}^{N} \ln(1 - \hat{\lambda}_i) \tag{4.60}$$

$$\lambda_{max}(r, r+1) \equiv -T\ln(1 - \hat{\lambda}_{r+1}), \tag{4.61}$$

where the $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_N$ are the estimated values of the eigenvalues obtained from $\hat{\boldsymbol{\Pi}}$. $\lambda_{trace}(r)$ tests the null hypothesis that the number of distinct cointegrating vectors is less than or equal to r against a general alternative of a number exceeding r. The further the estimated eigenvalues are from 0, the larger is the trace statistic. $\lambda_{max}(r, r+1)$ tests the null that the number of cointegrating vectors is r against the alternative of $r+1$ cointegrating vectors. Again, if the estimated value of the eigenvalue is large, $\lambda_{max}(r, r+1)$ will be large.

The critical values of the $\lambda_{trace}(r)$ and the $\lambda_{max}(r, r+1)$ statistics are obtained using a Monte Carlo approach as in Johansen and Juselius (1990). The distribution of the test statistics is nonstandard, and the critical values depend on the value of $N - r$, the number of nonstationary components, and whether deterministic terms (such constants or trends) are included in each of the equations.[21] If the test statistic is greater than the appropriate critical value, we will reject:

- the null hypothesis that there are at most r cointegrating vectors in favor of the alternative that there are $r+1$ vectors, in the case of the $\lambda_{trace}(r)$ test, or
- the null hypothesis that there are r cointegrating vectors in favor of the alternative that there are $r + I$ vectors, in the case of $\lambda_{max}(r, r+1)$.

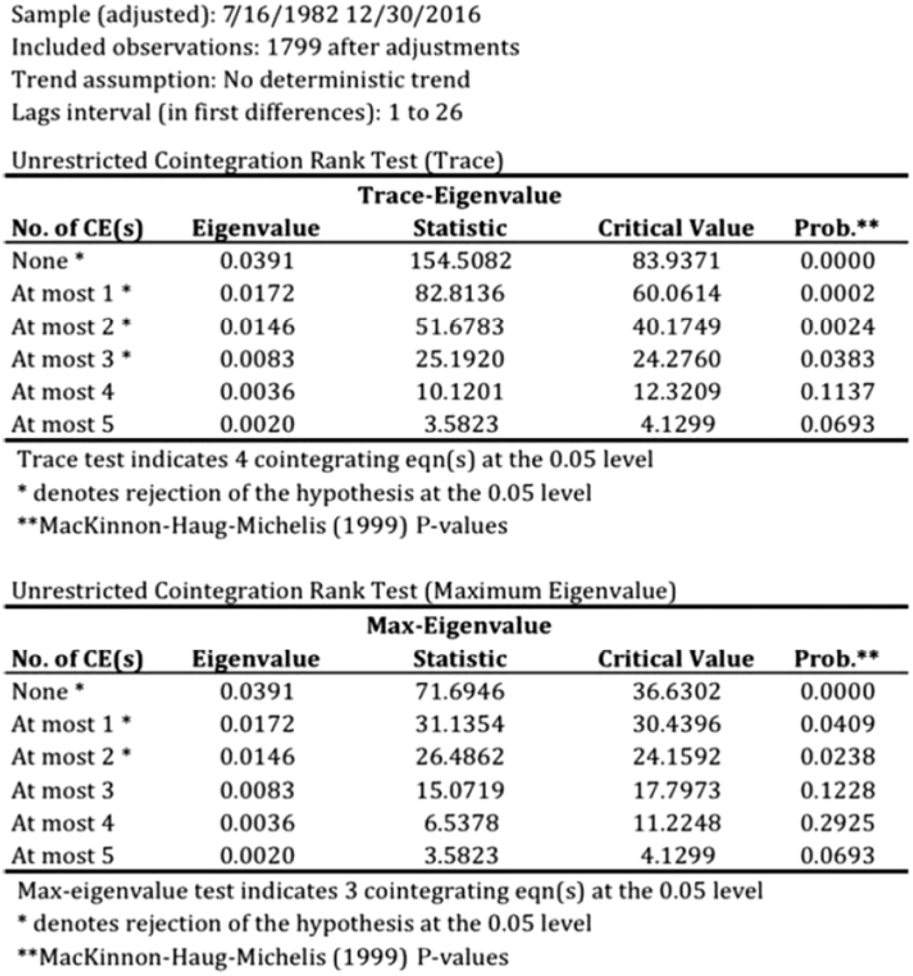
Of course, the tests based on $\lambda_{trace}(0)$ and $\lambda_{max}(0, 1)$ are key, because they are implicitly cointegration tests. Given their rather heterogeneous functional form, it is possible for Eqs. (4.60) and (4.61) to give different results. In fact, $\lambda_{max}(r, r+1)$ has the sharper alternative hypothesis and as a result it is usually preferred when trying to pin down the number of cointegrating vectors.

Finally, it is important to emphasize that the VECM in Eq. (4.58) cannot be estimated by OLS because it is necessary to impose cross-equation restrictions on the $\boldsymbol{\Pi}$ matrix. Therefore, multivariate cointegration testing tends to rely on ML estimation methods. We are now ready to ask whether our historical times series of monthly real S&P stock prices, dividends, and earnings are cointegrated using multivariate, rank-based MLE tests.

Question 2.B (2 points). Mango Bell, a junior analyst at Linova & Co, is studying the relationship between two series, namely a stock price and the associated earning-price ratio, which are both I(1). Mango has found out that the two series are cointegrated, therefore he decides to proceed to estimate a regression of the stock price over the price earning ratio. However, his boss claims that he has done a mistake since regressing two I(1) series one over the other leads to a spurious regression. Do you agree with Mango's boss? Carefully justify your answer; note that you are not asked to define cointegration.

Debriefing: Mango's boss is not correct: it is true that regressing two I(1) series one over the other may lead to a spurious regression. However, since the two series are cointegrated, not only the results from OLS estimates will be valid, but the OLS estimator will be super-consistent.

Question 2.C (2 points). With reference to weekly, constant-maturity US Treasury nominal rates (assumed to be I(1)) for the maturities 1- and 6-month, 1-, 3-, 7-, and 10-years and a January 8, 1982-December 30, 2016 sample, the following output shows the results of a standard Johansen's test.



Sample (adjusted): 7/16/1982 12/30/2016
Included observations: 1799 after adjustments
Trend assumption: No deterministic trend
Lags interval (in first differences): 1 to 26

Unrestricted Cointegration Rank Test (Trace)

		Trace-Eigenvalue		
No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None *	0.0391	154.5082	83.9371	0.0000
At most 1 *	0.0172	82.8136	60.0614	0.0002
At most 2 *	0.0146	51.6783	40.1749	0.0024
At most 3 *	0.0083	25.1920	24.2760	0.0383
At most 4	0.0036	10.1201	12.3209	0.1137
At most 5	0.0020	3.5823	4.1299	0.0693

Trace test indicates 4 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) P-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

		Max-Eigenvalue		
No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None *	0.0391	71.6946	36.6302	0.0000
At most 1 *	0.0172	31.1354	30.4396	0.0409
At most 2 *	0.0146	26.4862	24.1592	0.0238
At most 3	0.0083	15.0719	17.7973	0.1228
At most 4	0.0036	6.5378	11.2248	0.2925
At most 5	0.0020	3.5823	4.1299	0.0693

Max-eigenvalue test indicates 3 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) P-values

You care for cointegration because according to the expectation hypothesis (EH) of the term structure of interest rates, at least over the long-run it should happen that appropriately scaled sums of expected rates should equal the current rate minus a constant risk premium that rewards habitat effects and liquidity preference in favor or short-term bonds. Does the Johansen's test lead to identify one or more cointegrating relationships? What is the meaning of a number of cointegrating relationships exceeding one? What is the relationship between such findings on the existence of one or more cointegrating relationships and the fact the EH may hold on these data? Carefully explain your answer.

Debriefing: Even though the λ-trace and the λ-max tests point toward a different number of cointegrating vectors, there is no doubt that such number is at least two. This is consistent with the EH.
According to the EH, at least over the long-run, it should happen that appropriately scaled sums of expected rates should equal the current rate minus a constant risk premium that rewards habitat effects or a liquidity preference in favor or shorter-term bonds. Equivalently, all mispricings, i.e., deviations of long-term rates from weighted average of short-term rates, should be temporary and as such they should be I(0): if it were the case that the tested rates were all I(1), and if future short rates were easily predictable to the point to equal on average their future realized value, then the EH implies that one or more weighted sums of the I(1) rates

exist, such that the result is a I(0) variable plus a constant (the risk premium). However, note that cointegration between the rates is a necessary but not sufficient condition for the EH to be supported by the data. The validity of the EH would also require that a combination of rates is found to cointegrate with a cointegrating vector with structure $[1,\ \kappa_{7Y},\ \kappa_{3Y}, \ldots,\ \kappa_{1m}]'$, where the coefficients should be all negative and satisfy precise constraints. No such results or estimates have been reported for you to be able to decide on this matter.

Question 3.A (7 points) Consider the family of GARCH(p, q) models for asset returns. Define the persistence index and discuss what is the role that it plays in establishing the stationarity of a GARCH. Consider two alternative GARCH models for the same series of returns characterized by identical persistence index, but (i) the first model is characterized by a large $\sum_{i=1}^{p} \alpha_i$ and a small $\sum_{j=1}^{q} \beta_j$; (ii) the second model is characterized by a small $\sum_{i=1}^{p} \alpha_i$ and a large $\sum_{j=1}^{q} \beta_j$. What do you expect that the differences between the filtered, one-step ahead predicted variances from the two models will look like? Also consider the two cases that follow:

- You are a risk manager and you are considering calculating value-at-risk on the basis of a Gaussian homoskedastic model: is the mistake you are about to make larger under model (i) or model (ii)? Carefully explain why.
- You write and sell short-term options written on the underlying asset that you price using a tool that accounts for time varying volatility under GARCH: in the presence of large return shocks, will the mispricing be larger under model (i) or under model (ii)?

Debriefing:

Because the persistence index of a GARCH(p,q) model is given by $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j$, obviously large values of both the coefficients α_i $i = 1,\ldots,p$ and β_j $j = 1,\ldots,q$ act to increase the conditional volatility; however, they do so in different ways. The larger are the α_is, the larger is the response of $\sigma^2_{t+1|t}$ to new information; the larger are the β_js, the longer and stronger is the memory of conditional variance to past (forecasts of) variance. This means that for any given, fixed persistence index, it is possible for different stationary GARCH models to behave rather differently and therefore yield heterogeneous economic insights on the perceptions of uncertainty and risk, depending on the "fraction" of a fixed $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j$ brought about by $\sum_{i=1}^{p} \alpha_i$ and $\sum_{j=1}^{q} \beta_j$, respectively. Example 5.13 further elaborates on this point.

EXAMPLE 5.13

Let us fit a simple Gaussian ARMA(p_m, q_m)−GARCH(p,q) to monthly UK stock returns over a sample 1977−2016. To try an alternative strategy, we simply apply information criteria (with a preference for the BIC) to select the best model for a relatively wide range of choices of p_m, q_m, p, and q, including zeros and homoskedastic models. We obtain that a very simple constant mean ARMA(0,0)−GARCH(1,1) model brings the BIC down to 6.158, and this is sensibly inferior to the 6.170 of an AR(1)−GARCH(1,1) and the 6.172 of a more complex ARMA(1,1)−GARCH(1,1). In particular, ARCH models as well as bigger GARCH models never succeed in reducing the BIC below 6.158. The estimated model is then (p-values in parentheses):

$$R_{t+1} = \underset{(0.000)}{1.087} + \epsilon_{t+1} \quad \epsilon_{t+1} \quad \text{IID} \quad N(0, \sigma^2_{t+1|t})$$
$$\sigma^2_{t+1|t} = \underset{(0.095)}{0.540} + \underset{(0.000)}{0.086}\epsilon_t^2 + \underset{(0.000)}{0.898}\ \sigma^2_{t|t-1}.$$

The implied persistence index is then 0.984, which is in fact relatively high considering that these are monthly returns.

At this point, we perform the following experiment: holding constant the same time series of standardized residuals that we have obtained from the actual estimation of the model, we simulate four alternative series of both returns and variance predictions:

- α set to 1/4 of the estimated coefficient (0.022) with β set in such a way that their sum still equals 0.984, that is, to 0.962.
- α set to 1/2 of the estimated coefficient (0.043) with β set in such a way that their sum still equals 0.984, that is, to 0.943.
- α set to be 50% higher than the estimated coefficient (0.129) with β set in such a way that their sum still equals 0.984, that is, to 0.855.
- α set to be four times the estimated coefficient (0.344) with β set in such a way that their sum still equals 0.984, that is, to 0.640 (Fig. 5.9).

In all cases, the constant coefficient in the conditional variance model is held fixed to the estimated 0.540 and the series are initialized at the same value of 33.322, to favor comparability. Clearly, the first two cases illustrate the behavior of stationary, persistent GARCH models that react weakly to shocks; the last two cases concern GARCH models in which news tend to yield large impacts. Fig. 5.11 shows the resulting five implicit monthly *volatility* series, just because these enter more frequently the jargon of traders and researchers.

Of course, all series but one are just counter-factual simulations derived from the only, actual estimated process. However, the key role played by the structure of a GARCH model is easily detected. Even though, by construction, all processes have the same long-run, ergodic volatility of approximately 5.8% per month, the predicted volatility series characterized by high α and low β tend to fall most of the time below their long-run average, to then suddenly spike in short-lived volatility bursts that are however as high as three to four times the average. In our case, such models really make the turbulence of the 2007−09 Great Financial Crisis (henceforth GFC) look unprecedented. Note that risk managers and practitioners may then be induced into some degree of "Black Swan" type of complacency, when the numbers turn in the red or defaults occur. On the opposite, under GARCH models with low α and high β, volatility tends to be a rather smooth process that hovers around its long-run mean, to the point (in extreme cases in which β falls below 0.7) that extreme events such as the GFC may look hardly significant, in retrospect. In any event, the case stands: ML estimation has picked one and only one combination of α and β (among the infinite combinations such that $\alpha + \beta = 0.984$), because this one maximizes the likelihood that the sample of monthly UK returns does come from the model specified.

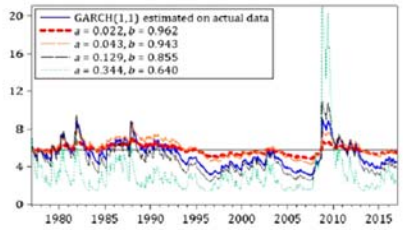
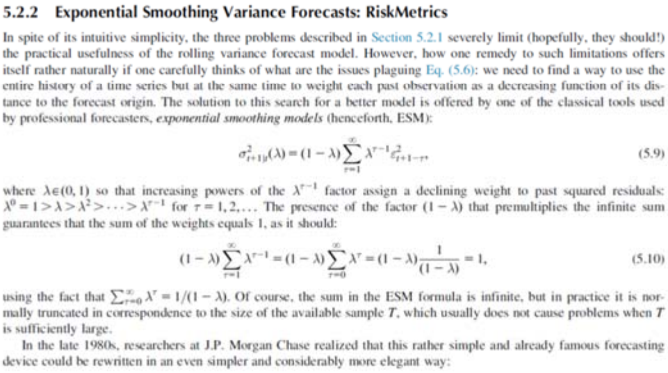


FIGURE 5.9 Predicted conditional volatility from a GARCH(1,1) model and four alternative simulated scenarios.

Question 3.B (2 points) Bruno Cerelli, an analyst at Reyer & So., has estimated a Gaussian

GARCH(1,1) model for FTSE MIB stock index returns and found that $\hat{\alpha} + \hat{\beta} = 1$; upon testing, he has not been able to reject the null hypothesis that $\alpha + \beta = 1$. Therefore, he has concluded that because the condition $\alpha + \beta < 1$ is violated, the GARCH model is non stationary so that a time-invariant unconditional distribution for the FTSE MIB returns does not exist and one cannot learn from past data to forecast returns. A colleague of his, Stefania Younot, has objected that this implication is unjustified, even though a GARCH with $\alpha + \beta = 1$ implies that variance follows a random walk with drift so that time t estimated variance is (in a mean-squared error sense) the best forecast for variance at time t + 1, t + 2, …, t + H for all $H \geq 1$. Which one of the two analysts at Reyer & So. is correct and why? Make sure to carefully explain your answer.

Debriefing:



5.2.2 Exponential Smoothing Variance Forecasts: RiskMetrics

In spite of its intuitive simplicity, the three problems described in Section 5.2.1 severely limit (hopefully, they should!) the practical usefulness of the rolling variance forecast model. However, how one remedy to such limitations offers itself rather naturally if one carefully thinks of what are the issues plaguing Eq. (5.6): we need to find a way to use the entire history of a time series but at the same time to weight each past observation as a decreasing function of its distance to the forecast origin. The solution to this search for a better model is offered by one of the classical tools used by professional forecasters, *exponential smoothing models* (henceforth, ESM):

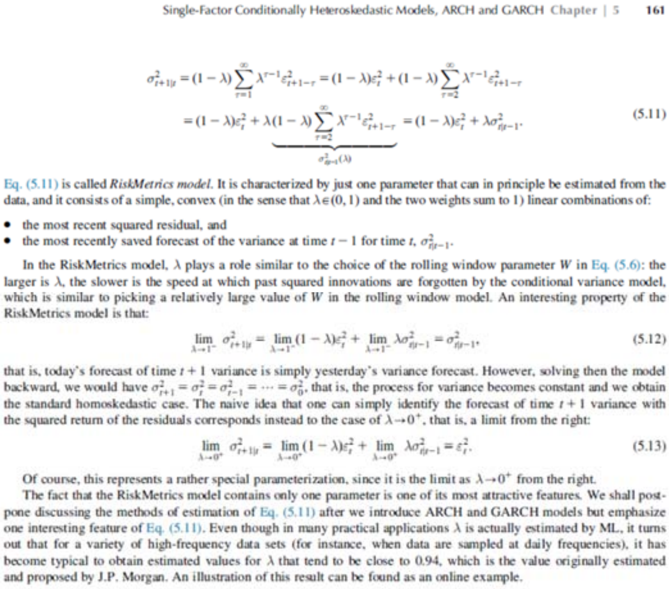
$$\sigma^2_{t+1|t}(\lambda) = (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1}\varepsilon^2_{t+1-\tau}, \qquad (5.9)$$

where $\lambda \in (0,1)$ so that increasing powers of the $\lambda^{\tau-1}$ factor assign a declining weight to past squared residuals: $\lambda^0 = 1 > \lambda > \lambda^2 > \cdots > \lambda^{\tau-1}$ for $\tau = 1, 2, \ldots$ The presence of the factor $(1-\lambda)$ that premultiplies the infinite sum guarantees that the sum of the weights equals 1, as it should:

$$(1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1} = (1-\lambda)\sum_{\tau=0}^{\infty}\lambda^{\tau} = (1-\lambda)\frac{1}{(1-\lambda)} = 1, \qquad (5.10)$$

using the fact that $\sum_{\tau=0}^{\infty}\lambda^{\tau} = 1/(1-\lambda)$. Of course, the sum in the ESM formula is infinite, but in practice it is normally truncated in correspondence to the size of the available sample T, which usually does not cause problems when T is sufficiently large.

In the late 1980s, researchers at J.P. Morgan Chase realized that this rather simple and already famous forecasting device could be rewritten in an even simpler and considerably more elegant way:



Single-Factor Conditionally Heteroskedastic Models, ARCH and GARCH Chapter | 5 161

$$\begin{aligned}\sigma^2_{t+1|t} &= (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1}\varepsilon^2_{t+1-\tau} = (1-\lambda)\varepsilon^2_t + (1-\lambda)\sum_{\tau=2}^{\infty}\lambda^{\tau-1}\varepsilon^2_{t+1-\tau} \\ &= (1-\lambda)\varepsilon^2_t + \lambda\underbrace{(1-\lambda)\sum_{\tau=2}^{\infty}\lambda^{\tau-1}\varepsilon^2_{t+1-\tau}}_{\sigma^2_{t|t-1}(\lambda)} = (1-\lambda)\varepsilon^2_t + \lambda\sigma^2_{t|t-1}.\end{aligned} \qquad (5.11)$$

Eq. (5.11) is called *RiskMetrics model*. It is characterized by just one parameter that can in principle be estimated from the data, and it consists of a simple, convex (in the sense that $\lambda \in (0,1)$ and the two weights sum to 1) linear combinations of:

- the most recent squared residual, and
- the most recently saved forecast of the variance at time $t-1$ for time t, $\sigma^2_{t|t-1}$.

In the RiskMetrics model, λ plays a role similar to the choice of the rolling window parameter W in Eq. (5.6): the larger is λ, the slower is the speed at which past squared innovations are forgotten by the conditional variance model, which is similar to picking a relatively large value of W in the rolling window model. An interesting property of the RiskMetrics model is that:

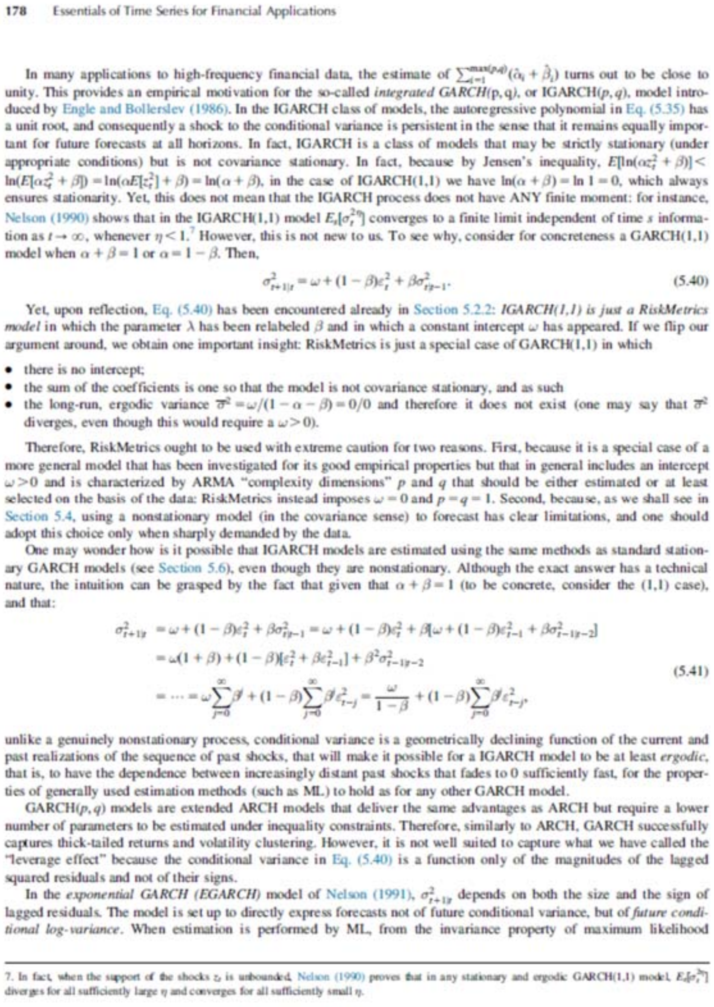
$$\lim_{\lambda\to 1^-}\sigma^2_{t+1|t} = \lim_{\lambda\to 1^-}(1-\lambda)\varepsilon^2_t + \lim_{\lambda\to 1^-}\lambda\sigma^2_{t|t-1} = \sigma^2_{t|t-1}, \qquad (5.12)$$

that is, today's forecast of time $t+1$ variance is simply yesterday's variance forecast. However, solving then the model backward, we would have $\sigma^2_{t+1} = \sigma^2_t = \sigma^2_{t-1} = \cdots = \sigma^2_0$, that is, the process for variance becomes constant and we obtain the standard homoskedastic case. The naive idea that one can simply identify the forecast of time $t+1$ variance with the squared return of the residuals corresponds instead to the case of $\lambda \to 0^+$, that is, a limit from the right:

$$\lim_{\lambda\to 0^+}\sigma^2_{t+1|t} = \lim_{\lambda\to 0^+}(1-\lambda)\varepsilon^2_t + \lim_{\lambda\to 0^+}\lambda\sigma^2_{t|t-1} = \varepsilon^2_t. \qquad (5.13)$$

Of course, this represents a rather special parameterization, since it is the limit as $\lambda \to 0^+$ from the right.

The fact that the RiskMetrics model contains only one parameter is one of its most attractive features. We shall postpone discussing the methods of estimation of Eq. (5.11) after we introduce ARCH and GARCH models but emphasize one interesting feature of Eq. (5.11). Even though in many practical applications λ is actually estimated by ML, it turns out that for a variety of high-frequency data sets (for instance, when data are sampled at daily frequencies), it has become typical to obtain estimated values for λ that tend to be close to 0.94, which is the value originally estimated and proposed by J.P. Morgan. An illustration of this result can be found as an online example.



178 Essentials of Time Series for Financial Applications

In many applications to high-frequency financial data, the estimate of $\sum_{i=1}^{\max(p,q)}(\hat{\alpha}_i + \hat{\beta}_i)$ turns out to be close to unity. This provides an empirical motivation for the so-called *integrated GARCH*(p, q), or IGARCH(p, q), model introduced by Engle and Bollerslev (1986). In the IGARCH class of models, the autoregressive polynomial in Eq. (5.35) has a unit root, and consequently a shock to the conditional variance is persistent in the sense that it remains equally important for future forecasts at all horizons. In fact, IGARCH is a class of models that may be strictly stationary (under appropriate conditions) but is not covariance stationary. In fact, because by Jensen's inequality, $E[\ln(\alpha z_t^2 + \beta)] < \ln(E[\alpha z_t^2 + \beta]) = \ln(\alpha E[z_t^2] + \beta) = \ln(\alpha + \beta)$, in the case of IGARCH(1,1) we have $\ln(\alpha + \beta) = \ln 1 = 0$, which always ensures stationarity. Yet, this does not mean that the IGARCH process does not have ANY finite moment: for instance, Nelson (1990) shows that in the IGARCH(1,1) model $E_s[\sigma_t^{2\eta}]$ converges to a finite limit independent of time s information as $t \to \infty$, whenever $\eta < 1$.[7] However, this is not new to us. To see why, consider for concreteness a GARCH(1,1) model when $\alpha + \beta = 1$ or $\alpha = 1 - \beta$. Then,

$$\sigma_{t+1|t}^2 = \omega + (1-\beta)\epsilon_t^2 + \beta\sigma_{t|t-1}^2. \tag{5.40}$$

Yet, upon reflection, Eq. (5.40) has been encountered already in Section 5.2.2: *IGARCH(1,1) is just a RiskMetrics model* in which the parameter λ has been relabeled β and in which a constant intercept ω has appeared. If we flip our argument around, we obtain one important insight: RiskMetrics is just a special case of GARCH(1,1) in which

- there is no intercept;
- the sum of the coefficients is one so that the model is not covariance stationary, and as such
- the long-run, ergodic variance $\overline{\sigma}^2 = \omega/(1-\alpha-\beta) = 0/0$ and therefore it does not exist (one may say that $\overline{\sigma}^2$ diverges, even though this would require a $\omega > 0$).

Therefore, RiskMetrics ought to be used with extreme caution for two reasons. First, because it is a special case of a more general model that has been investigated for its good empirical properties but that in general includes an intercept $\omega > 0$ and is characterized by ARMA "complexity dimensions" p and q that should be either estimated or at least selected on the basis of the data: RiskMetrics instead imposes $\omega = 0$ and $p = q = 1$. Second, because, as we shall see in Section 5.4, using a nonstationary model (in the covariance sense) to forecast has clear limitations, and one should adopt this choice only when sharply demanded by the data.

One may wonder how is it possible that IGARCH models are estimated using the same methods as standard stationary GARCH models (see Section 5.6), even though they are nonstationary. Although the exact answer has a technical nature, the intuition can be grasped by the fact that given that $\alpha + \beta = 1$ (to be concrete, consider the (1,1) case), and that:

$$\begin{aligned}\sigma_{t+1|t}^2 &= \omega + (1-\beta)\epsilon_t^2 + \beta\sigma_{t|t-1}^2 = \omega + (1-\beta)\epsilon_t^2 + \beta[\omega + (1-\beta)\epsilon_{t-1}^2 + \beta\sigma_{t-1|t-2}^2] \\ &= \omega(1+\beta) + (1-\beta)[\epsilon_t^2 + \beta\epsilon_{t-1}^2] + \beta^2\sigma_{t-1|t-2}^2 \\ &= \cdots = \omega\sum_{j=0}^{\infty}\beta^j + (1-\beta)\sum_{j=0}^{\infty}\beta^j\epsilon_{t-j}^2 = \frac{\omega}{1-\beta} + (1-\beta)\sum_{j=0}^{\infty}\beta^j\epsilon_{t-j}^2,\end{aligned} \tag{5.41}$$

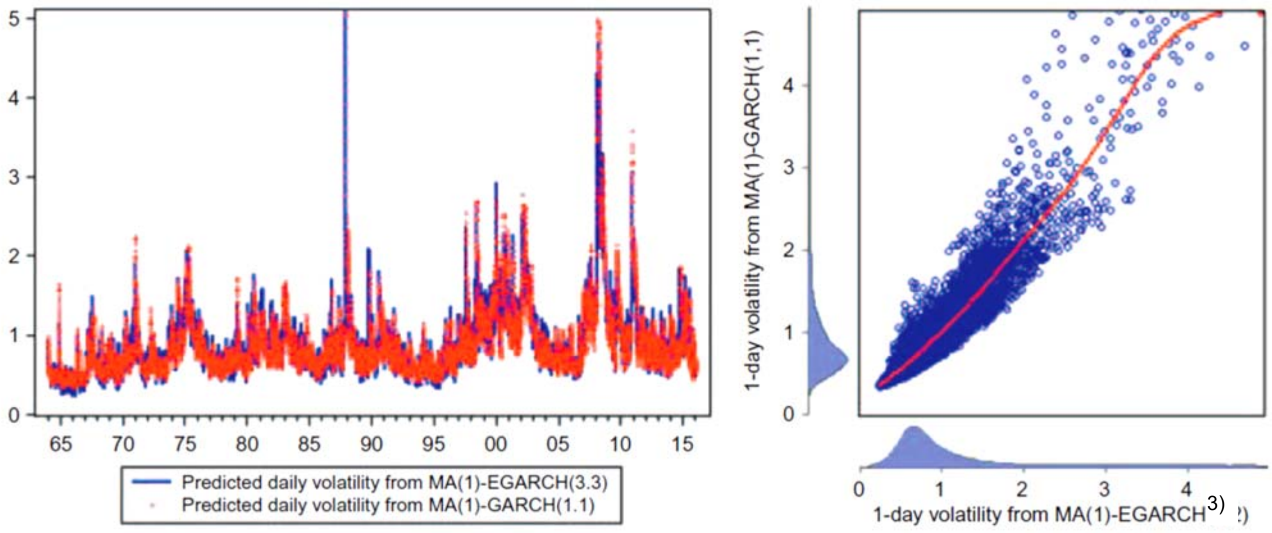
unlike a genuinely nonstationary process, conditional variance is a geometrically declining function of the current and past realizations of the sequence of past shocks, that will make it possible for a IGARCH model to be at least *ergodic*, that is, to have the dependence between increasingly distant past shocks that fades to 0 sufficiently fast, for the properties of generally used estimation methods (such as ML) to hold as for any other GARCH model.

GARCH(p, q) models are extended ARCH models that deliver the same advantages as ARCH but require a lower number of parameters to be estimated under inequality constraints. Therefore, similarly to ARCH, GARCH successfully captures thick-tailed returns and volatility clustering. However, it is not well suited to capture what we have called the "leverage effect" because the conditional variance in Eq. (5.40) is a function only of the magnitudes of the lagged squared residuals and not of their signs.

In the *exponential GARCH (EGARCH)* model of Nelson (1991), $\sigma_{t+1|t}^2$ depends on both the size and the sign of lagged residuals. The model is set up to directly express forecasts not of future conditional variance, but of *future conditional log-variance*. When estimation is performed by ML, from the invariance property of maximum likelihood

[7] In fact, when the support of the shocks z_t is unbounded, Nelson (1990) proves that in any stationary and ergodic GARCH(1,1) model, $E_s[\sigma_t^{2\eta}]$ diverges for all sufficiently large η and converges for all sufficiently small η.

Question 3.C (1.5 points) Using CRSP daily stock excess return data for a 1963-2016 sample, John, a quant researcher at Charles Thomas and Associates, has estimated two models: (i) a Gaussian MA(1)-GARCH(1,1) model, and (ii) a Gaussian MA(1)-EGARCH(3,3). The following plots compare in two different ways the predicted 1-day-ahead volatility filtered from the two different models. How can you describe the differences between the implied series of filtered/one-step ahead predicted variances from the two models? Suppose you are pricing securities the price of which monotonically increases with predicted variance (e.g., European puts and calls). Based on these two plots, what is the practical advantage that a EGARCH model may give over and above a simpler GARCH(1,1)? Make sure to carefully explain your answers.



Debriefing: Visibly, the EGARCH(3,3) is able to predict for the same day variances that are sometimes considerably higher and at other times visibly lower vs. those implied by a GARCH(1,1). The differences are particularly obvious in correspondence to October 1987 and September 2008, when the spikes in predicted variance differ across the two models, and in 1964-1966 when EGARCH(3,3) turned able to systematically forecast standard deviations that are 0.1-0.2% below GARCH(1,1). This can also be noted in the plot on the right, when for no value of volatility on the horizontal axis, the scatter plot reduced to a rather thin line close to the 45-degree line in red (the blue scatter plot always remains "thick" so to speak).

To a plain vanilla option pricer, EGARCH gives an additional layer of pricing flexibility, in the sense that both large and small shocks may predict rather heterogeneous, subsequent variances depending on the sign and the sequence of such shocks, given the fact that a EGARCH model is able to reflect complex patterns of leverage effects. This means that similar recent returns, depending on their sign and exact sequence may lead to different fair-value option prices and this may represent an advantage in terms of resulting P&L.