Time Series Analysis

3. Multivariate time series models

Andrew Lesniewski

Baruch College
New York

Fall 2019
Outline

1. Stationary VARMA models
2. General framework for cointegration
3. Multivariate stochastic volatility models
Multivariate time series analysis seeks to analyze several time series jointly.

The rationale behind this is the possible presence of interdependencies between the different time series.

These interdependencies, when quantified appropriately, could lead to improved reliability of forecasts.

For example, it is reasonable to expect various dependencies when considered jointly the following time series:

(i) (a) US GDP data, (b) S&P 500 data, and (c) oil prices.
(ii) (a) Mortgage applications, (b) interest rate data, and (c) unemployment rates.
(iii) (a) order flow data, (b) price levels, and (c) volatilities
(iv) ...

A. Lesniewski
Time Series Analysis
Consider a time series $X_t$, $-\infty < t < \infty$ of an $n$-dimensional state variable $X_t \in \mathbb{R}^n$. We represent $X_t$ as a column vector.

For modeling purposes, we assume that each $X_t$ is a random variable on a probability space, which is measurable with respect to the information set generated by all $X_s$ with $s \leq t$.

A time series is called covariance-stationary if:

(i) The mean $E(X_t) = \mu \in \mathbb{R}^n$ is a constant vector.

(ii) The autocovariance $\text{Cov}(X_t, X_s) = E((X_t - \mu)(X_s - \mu)^T) \in \text{Mat}_n(\mathbb{R})$ is time translation-invariant,

$$\text{Cov}(X_{t+\tau}, X_{s+\tau}) = \text{Cov}(X_t, X_s),$$

for all $\tau$. In particular, $\text{Var}(X_t)$ is a constant positive semidefinite matrix.

In the following, we will use the notation $\Gamma_{t-s} = \text{Cov}(X_t, X_s)$. By $R_{t-s}$ we will denote the autocorrelation matrix.
We claim that, as a consequence of stationarity, the following symmetry relation holds:

\[ \Gamma_t^T = \Gamma_{-t}, \]

(2)

which is a multivariate extension of the univariate property \( \Gamma_t = \Gamma_{-t} \).

The proof goes as follows:

\[
\Gamma_{t,ab}^T = \Gamma_{t,ba} \\
= \text{Cov}(X_{t,b}, X_{0,a}) \\
= \text{Cov}(X_{0,b}, X_{-t,a}) \quad \text{(time shift by \(-t\))} \\
= \text{Cov}(X_{-t,a}, X_{0,b}) \\
= \Gamma_{-t,ab}.
\]

Equation (2) implies the following symmetry property of the autocorrelation matrix:

\[ R_{t}^T = R_{-t}. \]

(3)
A vector autoregressive model VAR(1) is a linear time series model given by

$$X_t = a + BX_{t-1} + \varepsilon_t,$$

where $a \in \mathbb{R}^n$ is a constant vector, and $B \in \text{Mat}_n(\mathbb{R})$ is a constant matrix. The disturbances $\varepsilon \in \mathbb{R}^n$ are i.i.d. random vectors with $\varepsilon \sim N(0, \Omega)$, where $\Omega \in \text{Mat}_n(\mathbb{R})$ is a fixed (positive definite) covariance matrix,

$$\Omega = \mathbb{E}(\varepsilon_t \varepsilon_t^T).$$

We will now find the conditions under which the process (4) is covariance stationary.

Taking expected value of (4) yields

$$(I - B) \mu = a,$$

where $I$ is the $n \times n$ identity matrix.
This has a solution (for any $a$), if and only if 1 is not an eigenvalue of $B$, or

$$\det(I - B) \neq 0. \quad (6)$$

Then

$$\mu = (I - B)^{-1}a. \quad (7)$$

This is the first condition for covariance stationarity. In order to formulate the second condition, we will study the autocovariance of $X_t$. 
To this end, we use (7) to write (4) as

\[ X_t - \mu = B(X_{t-1} - \mu) + \varepsilon_t. \]  

(8)

Next, multiply (8) on the right by \((X_t - \mu)^T\) and take expectation. This yields the following equation for \(\Gamma_0\):

\[ \Gamma_0 = B\Gamma_0 B^T + \Omega. \]  

(9)

Similarly, multiplying (8) on the right by \((X_{t-k} - \mu)^T\) and taking expectation, we obtain the equation:

\[ \Gamma_k = B\Gamma_{k-1}. \]  

(10)

For a covariance stationary process, equation (9) has to have a solution. Unlike the univariate case, we cannot in general solve it in a closed form (the exception is when the matrices \(\Gamma_0\) and \(B\) commute, \(\Gamma_0 B = B\Gamma_0\)).
We can, nevertheless, draw from it some useful conclusions about covariance stationarity of the process (4). Let $\lambda$ be an eigenvalue of $B^T$,

$$B^T v = \lambda v, \text{ with } v \neq 0.$$ 

Multiplying (9) on the left by $v^T$ and on the right by $v$, we find that

$$v^T \Gamma_0 v = \lambda^2 v^T \Gamma_0 v + v^T \Omega v,$$

or

$$(1 - \lambda^2) v^T \Gamma_0 v = v^T \Omega v.$$ 

Since both matrices $\Omega$ and $\Gamma_0$ are positive definite, this is possible only if $|\lambda| < 1$. As a consequence, a necessary condition for the stationarity of the process is that all eigenvalues of $B^T$ have absolute values less than 1.
This condition can be rephrased in a more familiar form as follows. Consider the matrix-valued (linear) polynomial \( \psi(z) = I - Bz \), where \( I \) is the \( n \)-dimensional identity matrix.

Then the process is stationary if all the roots \( z_1, \ldots, z_n \) of the determinant \( \det(\psi(z)) \), or

\[
\det(I - Bz) = 0,
\]

lie outside of the unit circle.

This is clear, as the roots of this equation are given by \( 1/\lambda_1, \ldots, 1/\lambda_n \), where \( \lambda_i \) are the (nonzero) eigenvalues of \( B \).

Equations (9) and (10) form the Yule-Walker equations for a VAR(1) time series model.
As an example, consider the following VAR(1) model:

\[
X_{t,1} = 1.1 + 0.6X_{t-1,1} + 1.5X_{t-1,2} + \varepsilon_{t,1} \\
X_{t,2} = 2.0 + 0.1X_{t-1,1} + 0.4X_{t-1,2} + \varepsilon_{t,2}
\]  

(11)

with

\[
\Omega = \begin{pmatrix}
1.0 & 0.4 \\
0.4 & 1.0
\end{pmatrix}.
\]

Then

\[
B = \begin{pmatrix}
0.6 & 1.5 \\
0.1 & 0.4
\end{pmatrix}.
\]

and the eigenvalues are 0.9 and 0.1, and the process (11) is covariance stationary.

Notice that this is true (the eigenvalues have absolute values less than 1), even though the coefficient \(B_{12} = 1.5\) is greater than 1!
Let us go back to equation (9). Its solution can be written as an infinite series.

Namely, iterating the equation, we find that

\[
\Gamma_0 = B\Gamma_0 B^T + \Omega \\
= B^2 \Gamma_0 (B^2)^T + B\Omega B^T + \Omega \\
\cdots \\
= \sum_{j=0}^{\infty} B^j \Omega(B^j)^T. 
\]

One can show that, if all the eigenvalues have absolute values less than 1, this series converges.

This and equation (10) also imply that, for all \( k \geq 0 \),

\[
\Gamma_k = \sum_{j=0}^{\infty} B^{j+k} \Omega(B^j)^T. 
\]
In order to forecast a multivariate time series $X_t$ we minimize the expected MSE

$$E_t( (X_{t+k} - X^*_t|1:t)^\top (X_{t+k} - X^*_t|1:t) ) = \sum_{a=1}^{n} E_t( (X_{t+k,a} - X^*_t|1:t,a)^2 ). \quad (14)$$

Extending the reasoning of Lecture Notes #1, we find the optimal prediction is the conditional expectation

$$X^*_t|1:t = E_t(X_{t+k}). \quad (15)$$

This formula, applied to the one-period forecast in a $VAR(1)$ yields

$$X^*_{t+1|1:t} = E_t(X_{t+1})$$
$$= E_t(a + BX_t + \varepsilon_{t+1})$$
$$= a + BX_t. \quad (16)$$

The forecast error is $\varepsilon_{t+1}$, and so the variance of the forecast error is $\Omega$. 
In order to produce a $k$-period forecast with a VAR(1) model, we iterate:

$$X_{t+k} = a + BX_{t+k-1} + \varepsilon_{t+k}$$

$$= (I + B)a + B^2 X_{t+k-2} + \varepsilon_{t+k} + B\varepsilon_{t+k-1}$$

$$\ldots$$

$$= (I + B + \ldots + B^{k-1})a + B^k X_t + \varepsilon_{t+k} + B\varepsilon_{t+k-1} + \ldots + B^{k-1}\varepsilon_{t+1}.$$

This yields the optimal prediction:

$$X^*_{t+k|1:t} = (I + B + \ldots + B^{k-1})a + B^k X_t.$$  \hspace{1cm} (17)

The forecast error is

$$\varepsilon_{t+k} + B\varepsilon_{t+k-1} + \ldots + B^{k-1}\varepsilon_{t+1},$$

and so the forecast variance is

$$\Omega + B\Omega B^\top + \ldots + B^{k-1}\Omega(B^{k-1})^\top.$$
A \( \text{VAR}(p) \) model is an extension of the \( \text{VAR}(1) \) model to include \( p \) lags. It is specified by:

\[
X_t = a + B_1 X_{t-1} + \ldots + B_p X_{t-p} + \varepsilon_t,
\]

where \( a \in \mathbb{R}^n \) is a constant vector, and \( B_1, \ldots, B_p \in \text{Mat}_n(\mathbb{R}) \) are constant matrices.

The disturbances \( \varepsilon \in \mathbb{R}^n \) are i.i.d. random vectors with \( \varepsilon \sim N(0, \Omega) \).

Using the lag operator, we can write this process in the form

\[
\psi(L)X_t = a + \varepsilon_t,
\]

where

\[
\psi(z) = I - B_1 z - \ldots - B_p z^p
\]

is a polynomial in \( z \) taking on matrix values.
The requirement that $X_t$ is covariance stationary leads to the requirement that $E_t(X_t) = \mu$ is a constant, i.e.

$$(l - B_1 - \ldots - B_p)\mu = a.$$  \hspace{1cm} (21)

This is possible, if the matrix $\psi(1) = l - B_1 - \ldots - B_p$ is invertible, or

$$\det(\psi(1)) \neq 0.$$  \hspace{1cm} (22)

Then $\mu = \psi(L)^{-1}a$, and we can write

$$X_t = \mu + \psi(L)^{-1}\varepsilon_t.$$  \hspace{1cm} (23)
In order to find further conditions on the parameters of the process, we rewrite (18) in the form:

\[ \tilde{X}_t = B_1 \tilde{X}_{t-1} + \ldots + B_p \tilde{X}_{t-p} + \varepsilon_t, \]  

(24)

where

\[ \tilde{X}_t = X_t - \mu \]  

(25)

is the “de-meaned” process.

Using the same technique as in the case of \( \text{VAR}(1) \) we verify that

\[ \Gamma_0 = B_1 \Gamma_0 B_1^T + \ldots + B_p \Gamma_0 B_p^T + \Omega, \]

\[ \Gamma_k = B_1 \Gamma_{k-1} + \ldots + B_p \Gamma_{k-p}, \]  

(26)

which is the \( \text{VAR}(p) \) version of the Yule-Walker equation.
In order to state the second necessary condition for the covariance stationarity of the process (18), we note first that (18) can be written as a \( \text{VAR}(1) \) of a higher dimension. Namely, we define a \( pn \)-dimensional time series

\[
X_t = \begin{pmatrix}
\tilde{X}_t \\
\tilde{X}_{t-1} \\
\vdots \\
\tilde{X}_{t-p+1}
\end{pmatrix}
\]

We also define

\[
B = \begin{pmatrix}
B_1 & B_2 & \cdots & B_{p-1} & B_p \\
l & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & l & 0
\end{pmatrix}, \quad \text{and} \quad \eta_t = \begin{pmatrix}
\varepsilon_t \\
0 \\
\vdots \\
0
\end{pmatrix},
\]

where 0 denotes here the \( n \times n \) zero matrix.
With these definitions, the process (18) can be written as (verify!):

\[ X_t = BX_{t-1} + \eta_t. \]  

One can show, by extending the arguments presented in the case of \( \text{VAR}(1) \), that the is the requirement that all the roots of the equation

\[ \det(I - Bz) = 0 \]  

lie outside of the unit circle.

As in the case of \( \text{VAR}(1) \), we can find an expression for \( \Gamma_k \) in terms of an infinite series in the coefficients \( B_j \). This series converges under the stationarity condition formulated above.

All of this is fairly technical, and we will not study it in detail.
MLE for \( \text{VAR}(p) \)

\( \text{VAR}(p) \) models can be estimated using MLE very much the same way as in the case of the univariate \( \text{AR}(p) \) model. Let \( x_0, \ldots, x_T \in \mathbb{R}^n \) be a series of observations of an \( n \)-variate time series.

Denote by \( \theta = (a, B_1, \ldots, B_p, \Omega) \) the parameters of the model. The conditional PDF of the data can be written as:

\[
p(x_{p:T} | x_{0:p-1}, \theta) = \prod_{t=p}^{T} p(x_t | x_{0:t-1}, \theta) \\
= \prod_{t=p+1}^{T} p(\hat{\varepsilon}_t | x_{0:p-1}, \theta) \\
= \prod_{t=p+1}^{T} \frac{1}{(2\pi)^{n/2}(\text{det}(\Omega))^{1/2}} \exp \left( -\frac{1}{2} \hat{\varepsilon}_t^T \Omega^{-1} \hat{\varepsilon}_t \right).
\]
Hence the LLF is

\[
- \log L(\theta | x_0: T) = \frac{1}{2} (T - p) \log \det(\Omega) + \frac{1}{2} \sum_{t=p+1}^{T} \hat{\varepsilon}_t^T \Omega^{-1} \hat{\varepsilon}_t + \text{const.}
\]  

(29)

Notice that, if \( v_i \) are vectors and \( M \) is a matrix, then

\[
\sum_i v_i^T M v_i = \text{tr} \left( M \sum_i v_i v_i^T \right),
\]

where \( \text{tr} \) denotes the trace of a matrix (verify this!).

This allow us to rewrite the LLF above as:

\[
- \log L(\theta | x_0: T) = \frac{1}{2} (T - p) \log \det(\Omega) + \frac{1}{2} \text{tr} \left( \Omega^{-1} \sum_{t=p+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t^T \right) + \text{const.}
\]  

(30)
With a bit of multi-linear algebra, one can calculate the derivatives of the LLF with respect to the components of $\theta$ and find the critical points (the preferred values of the parameters $\hat{\theta}$). In particular, we find that the estimated covariance matrix $\hat{\Omega}$ is given by

$$
\hat{\Omega} = \frac{1}{T-p} \sum_{t=p+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t^T,
$$

where the $\hat{\varepsilon}_t$’s are calculated using the optimized values of $\hat{a}$ and $\hat{B}_j$.

We can thus write the maximized LLF in the following way:

$$
-\log L(\hat{\theta} | x_0:T) = \frac{1}{2} (T-p) \log \det(\hat{\Omega}) + \frac{1}{2} (T-p)n + \text{const.}
$$

We will use this expression later when discussing cointegration.
VAR\((p)\) model selection and checking

- As in the case of VAR\((p)\), a useful guide in the selection of the lag parameter \(p\) may be one of the information criteria, which in the case of VAR\((p)\) take the form:
  
  (i) Akaike information criterion,
  
  \[
  \text{AIC} = 2pn^2 - 2\log L(\hat{\theta} | x_{0:T}). \tag{33}
  \]

  (ii) Bayesian information criterion,

  \[
  \text{BIC} = \log(T)pn^2 - 2\log L(\hat{\theta} | x_{0:T}). \tag{34}
  \]

- It is always a good idea to check that the residuals \(\hat{\varepsilon}_t\) calculate with the optimized values of the parameters have the desired properties: independence for different \(t\)'s and stationarity. This can be done by running the appropriate statistics.

- The Python package `statsmodels.tsa.vector_ar` contains methods that are useful for modeling and analyzing multivariate time series using VAR and VECM specifications.
Sometimes, especially in the economic literature, one finds the distinction between two forms of the VAR\((p)\) model: structural and reduced form.

In the structural form, the covariance matrix \(\Omega\) of the shocks is diagonal (i.e. structural shocks are uncorrelated). The model is specified as follows:

\[
B_0 X_t = a + B_1 X_{t-1} + \ldots + B_p X_{t-p} + \epsilon_t, \tag{35}
\]

where \(B_0\) has 1’s on its diagonal.

In particular, this means that the current value of the variable \(X_{t,c}\) can be influenced by the lagged values of all the variables as well as the contemporaneous values of the variables \(X_{t,b}\) with \(b \neq c\).

From an economic point of view, a relationship represented by a structural VAR\((p)\) model reflects an underlying “structural” economic reality.

This is particularly valuable when:

(i) the drivers of various components of \(X_t\) are independent of each other;

(ii) different components of \(X_t\) can exert contemporaneous influence on each other.
Multiplying equation (35) on the left by $B_0^{-1}$ and renaming $B_0^{-1} a$ by $a$, $B_0^{-1} B_1$ by $B_1$, etc, we arrive at the reduced, i.e. standard, form (4) of the $VAR(p)$ model specification.

In the reduced form, the variables on the right hand side are predetermined at time $t$: no contemporaneous impact of one variable on another one is possible.

On the other hand, the covariance matrix $B_0^{-1} \Omega (B_0^{-1})^\top$ of the reduced form shocks is no longer diagonal, and thus the shocks are not independent. Different variables impact each other not through contemporaneous influences but through correlations among shocks.

For an in depth discussion of structural $VAR(p)$ models, see Lütkepohl’s book [2].
A VARMA\((p, q)\) model is an extension of the VAR(1) model to include \(p\) lags. It is specified by:

\[
\psi(L)X_t = a + \varphi(L)\varepsilon_t,
\]

(36)

where \(a \in \mathbb{R}^n\) is a constant vector, and \(\psi(z)\) and \(\varphi(z)\) are matrix-valued polynomials of an argument \(z\):

\[
\psi(z) = 1 - B_1 z - \ldots - B_p z^p,
\]

\[
\varphi(z) = 1 + \Theta_1 z + \ldots + \Theta_q z^q.
\]

(37)

Assuming that \(\psi(z)\) has no unit roots, we can write this model in the moving average form:

\[
X_t = \mu + \gamma(L)\varepsilon_t,
\]

(38)

where \(\gamma(L) = \psi(L)^{-1} \varphi(L)\).
Granger causality

- Multivariate time series analysis allows for a concept of causality that opens up new ways to determine whether one time series is useful in forecasting another one.

- This concept, the *Granger causality*, is a statistical test based on a VAR($p$) forecasting and is formulated as follows.

- Consider first a bivariate time series with components $X_t$, $Y_t$, which is assumed to follow an VAR(1) process:

$$
X_t = a_1 + B_{11} X_{t-1} + B_{12} Y_{t-1} + \varepsilon_{t,1},
$$

$$
Y_t = a_2 + B_{21} X_{t-1} + B_{22} Y_{t-1} + \varepsilon_{t,2}.
$$

(39)

- If $B_{12} = 0$, there is a directional asymmetry between the two time series: $X_t$ serves as an input variable, while $Y_t$ is an output variable. In other words, $Y_t$ does not influence $X_t$, while $X_t$ influences $Y_t$. 

A. Lesniewski

Time Series Analysis
Consider now two $k$-period forecasts of $X_t$ in the bivariate model above:

$$X_{t+k|f}^* = E(X_{t+k}|X_{1:t}, Y_{1:t}),$$
$$X_{t+k|p}^* = E(X_{t+k}|X_{1:t}).$$

In other words, $X_{t+k|f}^*$ is the forecast based on full information up to time $t$, while $X_{t+k|p}^*$ is based only on the information generated by the first series.

Let $\varepsilon_{k|f}$ and $\varepsilon_{k|p}$ denote the corresponding forecasting errors. Then, we say that $Y_t$ Granger-causes $X_t$, if

$$\text{Var}(\varepsilon_{k|f}) < \text{Var}(\varepsilon_{k|p}).$$  \hspace{1cm} (40)

In other words, $Y_t$ Granger-causes $X_t$, if the quality of the forecast of $X_t$ is improved by including the information about $Y_t$ (namely, the variance of the residuals declines).
The null hypothesis in the Granger test is *no Granger causality*, i.e. \( H_0 : B_{12} = 0 \) in (39). This is essentially an *F*-test on the regression coefficient (having no explanatory power).

One can generalize the discussion above to the VAR(\(p\)) setup with \( p \geq 2 \).

In this case, the null hypothesis cannot be rejected, if no lagged values of \( Y_t \) are retained in the regression.
Granger causality

- Note that Granger causality is not necessarily identical with physical causality. Granger causality is observed only indirectly, through a time series.

- If both component processes $X_t$ and $Y_t$ are driven by a common (but not explicitly included in the model) third process, one might still be likely not to reject the alternative hypothesis of Granger causality.

- The Granger test is designed to handle two variables, and may result in incorrect conclusions when the underlying relationship involves such *confounding variables* that causally affect both variables $X_t$ and $Y_t$. 
Cointegration is an expression of a long term equilibrium relationship between the components of a multivariate process $X_t$.

In Lecture Notes #2 we discussed the concept of cointegration in the context of two time series.

Using the concepts of multivariate time series analysis discussed above, we can now extend this concept to multiple cointegrating relations.

Let $X_t$ be a multivariate time series of dimension $n$, such that

(i) each component of $X_t$ is $I(1)$,
(ii) a cointegrating vector $a \in \mathbb{R}^n$, i.e. a non-zero vector such that the univariate time series $a^TX_t$ is $I(0)$ (i.e. stationary) exists.

A time series $X_t$ with these properties is said to be cointegrated of order 1.

If each component of $X_t$ is $I(d)$ and $a^TX_t$ is $I(d')$ with $d' < d$, we say that $X_t$ is cointegrated of order $d$. 
In general, there may exist more than one (linearly independent) cointegrating vectors.

Let \( r < n \) denote their number, and let \( \alpha \in \text{Mat}_{n,r}(\mathbb{R}) \) denote the matrix whose columns are these vectors.

This means that the matrix \( \alpha \) has the following properties:

(i) \( \alpha \) has (full) rank \( r \) (i.e. there are no linear dependencies among the columns of the matrix \( \alpha \)),

(ii) the \( r \)-dimensional time series \( Y_t = \alpha^T X_t \) is \( I(0) \).

We will call such a time series rank \( r \) cointegrated.
From linear algebra we know that there exists a matrix $\tilde{\alpha} \in \text{Mat}_{n,n-r}(\mathbb{R})$ of (maximal) rank $n - r$, such that $\alpha^T \tilde{\alpha} = 0$.

This fact follows from the rank-nullity theorem.

As a consequence, if $X_t$ is cointegrated of rank $r$, then the time series $Z_t = \tilde{\alpha}^T X_t$ is an $(n-r)$-dimensional unit root non-stationary time series.

Its components are called the common trends of $X_t$.

We shall now develop an MLE framework for testing for cointegration.
Cointegration

- We assume that $X_t$ follows a $VAR(p)$ specification given by (18). We wish to study it for possible cointegration.
- The method used to derive the ECM form of the $VAR(p)$ process in Lecture Notes #2 can be applied to (18). Namely, we define

\[ B = B_1 + \ldots + B_p, \]
\[ \Gamma_j = -(B_{j+1} + \ldots + B_p), \quad \text{for } j = 1, \ldots, p - 1. \] (41)

- After some algebra we can then write (18) in the form:

\[ (1 - BL - (\Gamma_1 L + \ldots + \Gamma_{p-1} L^{p-1})(1 - L)) X_t = a + \varepsilon_t. \] (42)

- Equivalently, this can be written as

\[ X_t = a + BX_{t-1} + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{p-1} \Delta X_{t-p+1} + \varepsilon_t. \] (43)

This form of the $VAR(p)$ process is referred to as the vector error correcting model (VECM).
For the purpose of testing for cointegration, we shall write the equation above in the following form:

\[ \Delta X_t = a + \Pi X_{t-1} + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{p-1} \Delta X_{t-p+1} + \epsilon_t, \]  

(44)

where the matrix \( \Pi \) is defined by

\[ \Pi = B - I. \]

This form of the equation makes it explicit that key to understanding the cointegration structure of \( X_t \) is the rank of the matrix \( \Pi \).
For example, the process (43) of Lecture Notes #2 can be written in the \textit{VECM} form as follows:

\[
\begin{pmatrix}
\Delta X_t \\
\Delta Y_t
\end{pmatrix} = \begin{pmatrix}
\alpha_1 + \gamma \alpha_2 \\
\alpha_2
\end{pmatrix} + \begin{pmatrix}
-1 & \gamma \\
0 & 0
\end{pmatrix} \begin{pmatrix}
X_{t-1} \\
Y_{t-1}
\end{pmatrix} + \begin{pmatrix}
\varepsilon_{t,1} + \gamma \varepsilon_{t,2} \\
\varepsilon_{t,2}
\end{pmatrix}
\]

Notice that the $2 \times 2$-matrix

\[
\Pi = \begin{pmatrix}
-1 & \gamma \\
0 & 0
\end{pmatrix}
\]

is of rank 1.

Define

\[
\alpha = \begin{pmatrix}
1 \\
-\gamma
\end{pmatrix}, \quad \beta = \begin{pmatrix}
-1 \\
0
\end{pmatrix}, \quad \tilde{\alpha} = \begin{pmatrix}
\gamma \\
1
\end{pmatrix}.
\]

Then $\alpha^T \tilde{\alpha} = 0$, and $\Pi = \beta \alpha^T$. 

General methodology for the testing for cointegration was developed by Johansen [1]. For a good textbook presentation, I recommend Tsay’s book [3].

A cointegration test is essentially a test on the rank $r$ of the matrix $\Pi$ in (44).

If $r = 0$, then $\Pi = 0$, and clearly there are no cointegrating vectors.

If, on the other hand, $r > 0$, then $X_t$ has $r$ cointegrating vectors and $n - r$ unit roots. In this case there is a cointegrating matrix $\alpha \in \text{Mat}_{n,r}(\mathbb{R})$ such that the $r$-dimensional process $\alpha^T X_t$ is $I(0)$.

The matrix $\Pi$ can be represented as $\Pi = \beta \alpha^T$, where $\beta \in \text{Mat}_{n,r}(\mathbb{R})$ has rank $r$. In other words, in the presence of $r$ cointegrating relations, only $r$ linear combinations of $X_{t-1}$ appear in (44).

The $(n - r)$-dimensional unit root process $\tilde{\alpha}^T X_t$, where $\tilde{\alpha} \in \text{Mat}_{n,n-r}(\mathbb{R})$, $\alpha^T \tilde{\alpha} = 0$, is the process of common trends (as discussed earlier).
Cointegration

- We reformulate the model (44) in terms of three linear regressions.

- First, we run the following two regressions:

\[
\Delta X_t = b + Q_1 \Delta X_{t-1} + \ldots + Q_p \Delta X_{t-p+1} + u_t,
\]

\[
X_{t-1} = c + R_1 \Delta X_{t-1} + \ldots + R_p \Delta X_{t-p+1} + v_t,
\]

(45)

where \( u_t \) and \( v_t \) are the residuals. These regressions are estimated (jointly) using the least squares (MLE).

- Next, given the estimated coefficients \( \hat{Q}_i \) and \( \hat{R}_i \) in (45), we calculate the residuals \( \hat{u}_t \) and \( \hat{v}_t \), and run the third regression:

\[
\hat{u}_t = \Pi \hat{v}_t + \varepsilon_t,
\]

(46)

with errors \( \varepsilon_t \). Its purpose is to estimate the matrix \( \Pi \).
A calculation shows that the coefficients $a$ and $\Gamma_i$ in (44) can be expressed in terms of the coefficients $b$, $c$, $Q_i$, $R_i$, and $\Pi$ in (45) and (46).

The same calculation shows that the residulas in (46) and (44) are identical, hence the notation.

Testing for the rank of the estimated matrix $\hat{\Pi}$ is thus equivalent to testing for correlations among the components of $\hat{u}$ and $\hat{v}$. 

Cointegration
Before we move on, we will take a break to review the basic ideas behind the likelihood ratio test under multivariate Gaussian distribution.

Consider an $n$-dimensional Gaussian random variable $X \sim N(0, \Sigma)$. We decompose it into two components $U$ and $V$,

$$X = \begin{pmatrix} U \\ V \end{pmatrix},$$  \hspace{1cm} (47)

where $U$ has dimension $p$, $U \sim N(0, \Sigma_U)$, and $V$ has dimension $q = n - p$, $V \sim N(0, \Sigma_V)$.

We write accordingly

$$\Sigma = \begin{pmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{pmatrix},$$  \hspace{1cm} (48)

where $\Sigma_{UV} = \Sigma_{VU}^T$ is the covariance matrix of $U$ and $V$. 
Our goal is to test the null hypothesis $H_0 : \Sigma_{UV} = 0$, i.e. $\text{Cov}(U, V) = 0$, versus the alternative $H_a : \Sigma_{UV} \neq 0$.

Let $x_i, i = 1, \ldots, N$, be a sample of observations, and let

$$
\hat{\Sigma}_U = \frac{1}{N} \sum_{i=1}^{N} u_i u_i^T,
$$

$$
\hat{\Sigma}_V = \frac{1}{N} \sum_{i=1}^{N} v_i v_i^T
$$

(49)

denote the MLE estimates of $\Sigma_U$ and $\Sigma_V$, respectively.

Under the null hypothesis $H_0$:

$$
\hat{\Sigma}_0 = \begin{pmatrix}
\hat{\Sigma}_U & 0 \\
0 & \hat{\Sigma}_V
\end{pmatrix}
$$

(50)
As we have already seen, the maximized likelihood function under the multivariate Gaussian distribution is of the form

$$\mathcal{L}(\hat{\Sigma}|x) = \text{const} \times \det(\hat{\Sigma})^{-N/2}$$

$$\propto \det(\hat{\Sigma})^{-N/2},$$

Therefore, under $H_0$,

$$\mathcal{L}(\hat{\Sigma}_0|x) \propto \det(\hat{\Sigma}_0)^{-N/2}$$

$$= \det(\hat{\Sigma}_U)^{-N/2} \det(\hat{\Sigma}_V)^{-N/2}, \quad (51)$$

On the other hand, under the alternative hypothesis $H_a$,

$$\hat{\Sigma}_a = \begin{pmatrix} \hat{\Sigma}_U & \hat{\Sigma}_{UV} \\ \hat{\Sigma}_{VU} & \hat{\Sigma}_V \end{pmatrix}. \quad (52)$$
We shall now use a property of the determinant of a matrix $M$ familiar from linear algebra. Namely, its value remains unchanged, if to a row of $M$ we add any linear combination of other rows of $M$.

In particular,

$$
\det \left( \begin{array}{cc}
\hat{\Sigma}_U & \hat{\Sigma}_{UV} \\
\hat{\Sigma}_{VU} & \hat{\Sigma}_V 
\end{array} \right) = \det \left( \begin{array}{cc}
\hat{\Sigma}_U & \hat{\Sigma}_{UV} \\
0 & \hat{\Sigma}_V - \hat{\Sigma}_{VU} \hat{\Sigma}_U^{-1} \hat{\Sigma}_{UV} 
\end{array} \right),
$$

where we have performed linear operations on $\hat{\Sigma}_a$ in such a way that $\hat{\Sigma}_{VU}$ is replaced by the zero matrix.

The resulting matrix has a block upper triangular structure, and its determinant is the product of the determinants of the diagonal blocks. Hence,

$$
\mathcal{L}(\hat{\Sigma}_a | \mathbf{x}) \propto \det(\hat{\Sigma}_a)^{-N/2} = \det(\hat{\Sigma}_U)^{-N/2} \det(\hat{\Sigma}_V - \hat{\Sigma}_{VU} \hat{\Sigma}_U^{-1} \hat{\Sigma}_{UV})^{-N/2}.
$$

(53)
The ratio statistics $R$ is defined as

$$R = \frac{\mathcal{L}(\hat{\Sigma}_0|x)}{\mathcal{L}(\hat{\Sigma}_a|x)}.$$  \hspace{1cm} (54)

The null hypothesis is rejected if $R$ is small.

Let $\lambda_i$, $i = 1, \ldots, q$, denote the eigenvalues of the matrix $\hat{\Sigma}_V^{-1}\hat{\Sigma}_V\hat{\Sigma}_U^{-1}\hat{\Sigma}_{UV}$. Then

$$R = \frac{\det(\hat{\Sigma}_V)^{-N/2}}{\det(\hat{\Sigma}_V - \hat{\Sigma}_V\hat{\Sigma}_U^{-1}\hat{\Sigma}_{UV})^{-N/2}} = \det(1 - \hat{\Sigma}_V^{-1}\hat{\Sigma}_V\hat{\Sigma}_U^{-1}\hat{\Sigma}_{UV})^{N/2} = \prod_{i=1}^{q} (1 - \lambda_i)^{N/2}.$$
It is customary to work with the log likelihood statistics normalized as follows:

\[ \mathcal{LR} = -2 \log \mathcal{R}. \]  

(55)

Explicitly,

\[ \mathcal{LR} = -N \log \prod_{i=1}^{q} (1 - \lambda_i) \]

(56)

\[ = -N \sum_{i=1}^{q} \log(1 - \lambda_i). \]

The null hypothesis is rejected if \( \mathcal{LR} \) is large.

We can now go back to the analysis of cointegrating time series.
Cointegration tests

- Going back to (46), we now wish to test the following hypothesis:

\[ H_0 : \text{rank}(\Pi) = r, \quad \text{against } H_1 : r < \text{rank}(\Pi) \leq r_1. \]  

(57)

- All residuals and estimated covariance matrices are computed according to the expressions discussed earlier (see (49)).

- The statistic used for testing the hypothesis (57) is given by the log likelihood ratio \( LR \) introduced above (with \( N \) replaced by \( T \), the length of the observed series).

- We use it as follows. Define

\[
LR(r, r_1) = LR(r_1) - LR(r) = T \left( - \sum_{j=1}^{r_1} \log(1 - \lambda_j) + \sum_{j=1}^{r} \log(1 - \lambda_j) \right)
\]

(58)

\[
= - T \sum_{j=r+1}^{r_1} \log(1 - \lambda_j).
\]
Cointegration tests

- Calculations, similar to the calculations we did in Lecture Notes #2, show that the asymptotic distributions of these statistics under the null hypothesis are as follows.
- Let $\mathcal{X}$ be the random matrix:

$$\mathcal{X} = \left( \int_0^1 W(s) dW(s)^T \right)^T \left( \int_0^1 W(s) W(s)^T ds \right)^{-1} \left( \int_0^1 W(s) dW(s)^T \right).$$

Here, $W(s)$ is the standard Brownian motion of dimension $n - r$. Then:

(i) \[ LR(r, n) \rightarrow \text{tr}(\mathcal{X}), \]  

(ii) \[ LR(r, r + 1) \rightarrow \lambda_{max}, \]  

where $\lambda_{max}$ denotes the largest eigenvalue of $\mathcal{X}$. 

\[ \text{(59)} \]
Cointegration

- In other words, the limit distributions of both statistics (under null hypothesis) are functionals of a multi-factor Brownian motion.
- Samples from these distributions can be easily simulated by standard Monte Carlo methods, and the critical values for the log likelihood ratio tests can be deduced.
- The critical values are tabulated in software packages.
- A Python implementation of this test is still work in progress. An R implementation is in the package `urca`.
- For these tests, the number of lags \( p \) has to be known. This can be chosen with the help of the AIC or BIC criteria discussed above.
Cointegration tests

There are two versions of the Johansen cointegration test:

(i) The *trace test*: in the trace test we work with the log likelihood ratio stat $\mathcal{L}\mathcal{R}(r, n)$ for testing the null hypothesis (57).

(ii) The *maximum eigenvalue test*: in the maximum eigenvalue test we work with the log likelihood ratio stat $\mathcal{L}\mathcal{R}(r, r + 1)$ for testing the null hypothesis (57).

These two tests may result in different inferences.

Generally, the trace test is regarded as stronger.
Cointegration

In order to determine the cointegrating rank we run a sequence of tests:

\[ H_0 : r = 0, \]
\[ H_0 : r = 1, \]
\[ \ldots, \]
\[ H_0 : r = n - 1, \]

and reject the null hypothesis the first time it is possible.

Both the trace test and maximum eigenvalue test can be used.

Example: take \( n = 3 \).

(i) We first test whether \( r = 0 \). If this cannot be rejected, we further analyze the corresponding VECM model.

(ii) If we can reject \( r = 0 \), we proceed to testing \( r = 1 \). If \( r = 1 \) cannot be rejected, we continue with analyzing the VECM model with \( r = 1 \).

(iii) Otherwise, we test \( r = 2 \). If this hypothesis cannot be rejected, we take \( r = 2 \) and analyze the VECM model.

(iv) Finally, if \( r = 2 \) cannot be rejected, it may mean that the VAR model is stationary.
When analyzing jointly the volatility of a number of financial time series, it may be important to take into account the dependencies between the shocks contributing to the volatilities of each of the individual time series.

Modeling each individual volatility series as a univariate GARCH (or other) model may be then inappropriate, as this approach does not take into account dependencies between different time series.

There exist a number of more or less sophisticated approaches to modeling joint stochastic volatility of a multivariate time series.

Here we provide a very brief introduction to some of them. For an extensive discussion, see [3].
The simplest and most parsimonious approach is the *exponentially weighted moving average* (EWMA) model which is a natural extension of the analogous univariate model discussed in Lecture Notes #2.

According to the general principle, we decompose a (multivariate) time series into the seasonality, trend, and disturbance parts (see formula (6) in Lecture Notes #1).

The disturbance model is specified as follows:

\[
\begin{align*}
\varepsilon_t &= \Omega_t^{1/2} z_t, \\
\Omega_t &= \lambda \Omega_{t-1} + (1 - \lambda) \varepsilon_{t-1} \varepsilon_{t-1}^T,
\end{align*}
\]

where \( z_t \) is a sequence of i.i.d. random vectors \( z_t \sim N(0, \mathbb{I}_n) \), and where \( \mathbb{I}_n \) denotes the \( n \)-dimensional identity matrix.
Here, $\Omega^{1/2}$ denotes the *square root* of the positive definite matrix $\Omega$, which is defined as the *unique positive define* matrix such that $(\Omega^{1/2})^2 = \Omega$.

Note that

$$\text{Cov}(\varepsilon_t, \varepsilon_t) = \mathbb{E}(\varepsilon_t \varepsilon_t^T)$$

$$= \mathbb{E}(\Omega_t^{1/2} z_t z_t^T \Omega_t^{1/2})$$

$$= \Omega_t^{1/2} I_n \Omega_t^{1/2}$$

$$= \Omega_t,$$

i.e. the process $\Omega_t$ is the covariance process of $\varepsilon_t$. 

(64)
The $BEKK(p, q)$ model (after Baba, Engle, Kraft, and Kroner) is a multivariate extension of the GARCH.

Its simplest version, $BEKK(1, 1)$ is specified as follows:

\[
\varepsilon_t = \Omega_t^{1/2} z_t, \\
\Omega_t = AA^T + B\varepsilon_{t-1}\varepsilon_{t-1}^T B^T + C\Omega_{t-1} C^T. 
\] (65)

Here, $A, B, C \in \text{Mat}_n(\mathbb{R})$ are square matrices. It is usually additionally assumed that $A$ is lower triangular.

The model has a vast number of parameters (even with the restriction on $A$): $n(n + 1)/2 + 2n^2$.

For example, without further restrictions, this model requires 11 parameters for a bivariate time series.
Another simple approach to modeling multivariate volatility is the *dynamic conditional correlation* (DCC) model, which is specified as follows.

Let $\Omega_t$ denote the covariance matrix of the shocks $\varepsilon_t$, and let $\sigma^2_t$ be its diagonal. The nonzero elements of $\sigma^2_t$ are the variances of the individual time series at time $t$, and so

$$\rho_t = \sigma^{-1}_t \Omega_t \sigma^{-1}_t$$

is the correlation matrix at time $t$.

The model is specified as follows:

$$q_t = (1 - \vartheta_1 - \vartheta_2) \bar{q} + \vartheta_1 q_{t-1} + \vartheta_2 \eta_{t-1} \eta_{t-1}^\top,$$

$$\rho_t = d^{-1}_t q_t d^{-1}_t.$$
Here:

(i) $\vartheta_1, \vartheta_2$ are real parameters such that $0 < \vartheta_1 + \vartheta_2 < 1$,
(ii) $\eta_{t,a} = \varepsilon_{t,a}/\sigma_{t,a}$ are the standardized disturbances,
(iii) $d_t$ is the positive square root of the diagonal of $q_t$ and is simply a normalizing factor.
(iv) $\bar{q}$ is the unconditional covariance matrix of the $\eta$'s.

Note that the model is specified so that $q_t$ is guaranteed to be positive definite.

In practice, the volatilities $\sigma_{t,a}$ can be estimated from univariate GARCH models.

The DCC model is thus very parsimonious, as it has two free parameters only. The impact of initial choices can be decayed by priming the model as explained in Lecture Notes #2.
References

