INTRODUCTION TO TIME SERIES ANALYSIS

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

May 25, 2010

Lecture Outline

1. Basic Concepts
   - Stationarity
   - Difference Equations
   - Back-Shift Operator
2. Univariate Time Series
   - Moving Average Series
   - Autoregressive Series
   - Autoregressive Moving Average Series
   - Invertibility of MA Series
3. Forecasting
   - Forecasts: Infinite Observations
   - Forecasts: Finite Observations

Lecture Outline (cont'd)

4. Box-Jenkins Approach
   - Differencing
   - Identification
   - Model Estimation
   - Asymptotic Properties of the QMLE
   - Model Diagnostic Tests
   - Model Selection Criteria
5. Multivariate Time Series
   - Vector AR Series
   - Model Estimation
   - Asymptotic Properties of the QMLE
   - Impulse Response Functions
   - Structural VAR Models

Lecture Outline (cont'd)

6 Non-Stationary Time Series
- Functional Central Limit Theorem
- $I(1)$ Series
- Autoregression of $I(1)$ Series
- Tests of Unit Root
- Tests of Stationarity against $I(1)$

7 Models with Integrated Time Series
- Spurious Regressions
- Co-Integration
- Co-Integrating Regressions
- Fully-Modified Estimation
- Johansen's Maximum Likelihood Procedure

Lecture Outline (cont'd)

8. Volatility Models
   - ARCH Models
   - GARCH Models
   - EGARCH Models
   - GJR-GARCH Models
   - Estimating GARCH Models
   - Stochastic Volatility Models
   - Realized Volatility

Weak Stationarity

- $\{y_t\}$ is said to be weakly stationary or covariance stationary if its mean is time invariant: $\mathbb{E}(y_t) = \mu$, and its autocovariances:

  $$\mathbb{E}[(y_t - \mu)(y_{t-j} - \mu)] = \gamma_j, \qquad j = 0, \pm 1, \pm 2, \ldots,$$

  depend on j but not on t; $\gamma_0 = \text{var}(y_t)$ is also time invariant.

- As a result, the autocorrelations of y_t,

  $$\rho_j = \gamma_j / \gamma_0, \qquad j = 0, \pm 1, \pm 2, \ldots,$$

  are also independent of t, and $\rho_j = \rho_{-j}$.

- Example: A white noise is a series with zero mean, constant variance, and zero autocorrelations. Hence, it is weakly stationary.

Strict Stationarity

- $\{y_t\}$ is said to be strictly stationary if its finite dimensional distributions are invariant under time displacements, i.e., for each s,

$$F_{t_1,\ldots,t_n}(c_1,\ldots,c_n) = F_{t_1+s,\ldots,t_n+s}(c_1,\ldots,c_n).$$

  A strict stationary series need not be weak stationary, unless it has finite second moment.

  - i.i.d. random variables are strictly stationary, but i.i.d. Cauchy (or $t(2)$) random variables are not. (Why?)
  - A series is Gaussian if its finite dimensional distributions are all Gaussian. A weakly stationary, Gaussian series is strictly stationary.

Difference Equations

Consider the first-order difference equation:

$$y_t = \psi_1 y_{t-1} + u_t, \quad t = 0, 1, 2, \ldots$$

By recursive substitution,

$$\begin{aligned} y_t &= \psi_1 \big(\psi_1 y_{t-2} + u_{t-1}\big) + u_t \\ &= \psi_1^2 y_{t-2} + \psi_1 u_{t-1} + u_t \\ &= \psi_1^3 y_{t-3} + \psi_1^2 u_{t-2} + \psi_1 u_{t-1} + u_t \\ \vdots & \qquad\qquad \vdots \\ &= \psi_1^{t+1} y_{-1} + \psi_1^t u_0 + \psi_1^{t-1} u_1 + \cdots + \psi_1 u_{t-1} + u_t. \end{aligned}$$

Similarly, $y_{t+j} = \psi_1^{j+1} y_{t-1} + \psi_1^j u_t + \psi_1^{j-1} u_{t+1} + \cdots + \psi_1 u_{t+j-1} + u_{t+j}$.

- The impulse response (dynamic multiplier) of y_t is the effect of one unit change of u_t to the future observation y_{t+j}:

  $$\frac{\partial y_{t+j}}{\partial u_t} = \psi_1^j,$$

  which depends only on j but not on t.

- A system is said to be stable if the impulse response eventually vanishes as j tends to infinity. It is explosive if the impulse response diverges.

- The first-order difference equation is stable (explosive) when $|\psi_1| < 1$ $(|\psi_1| > 1)$.

- The accumulated response (interim multiplier) of y_t is

$$\sum_{i=0}^{j} \frac{\partial y_{t+j}}{\partial u_{t+i}} = \psi_1^j + \psi_1^{j-1} + \cdots + \psi_1 + 1.$$

- For $|\psi_1| < 1$,

$$\lim_{j \to \infty} \sum_{i=0}^{j} \psi_1^{j-i} = \frac{1}{1-\psi_1},$$

  which represents the long-run effect (total multiplier) of a permanent change in u.

The *p*th-order difference equation,

$$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p} + u_t,$$

can be expressed as a first-order vector difference equation:

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}}_{\boldsymbol{\eta}_t} = \underbrace{\begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_{p-1} & \psi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}}_{\mathbf{F}} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{bmatrix}}_{\boldsymbol{\eta}_{t-1}} + \underbrace{\begin{bmatrix} u_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\boldsymbol{\nu}_t}.$$

That is, $\boldsymbol{\eta}_t = \mathbf{F}\boldsymbol{\eta}_{t-1} + \boldsymbol{\nu}_t$. Then,

$$\boldsymbol{\eta}_{t+j} = \mathbf{F}^{j+1}\boldsymbol{\eta}_{t-1} + \mathbf{F}^{j}\boldsymbol{\nu}_t + \mathbf{F}^{j-1}\boldsymbol{\nu}_{t+1} + \cdots + \mathbf{F}\boldsymbol{\nu}_{t+j-1} + \boldsymbol{\nu}_{t+j}.$$

- The impulse response of $\boldsymbol{\eta}_t$ is

$$\nabla_{\boldsymbol{\nu}_t} \boldsymbol{\eta}_{t+j} = \mathbf{F}^j,$$

and the impulse response of y_{t+j} is $\partial y_{t+j} / \partial u_t = f_{11}^j$, the $(1,1)$ element of $\mathbf{F}^j$.

- The long-run effect of $\boldsymbol{\nu}$ is

$$\lim_{j \to \infty} \sum_{i=0}^{j} \mathbf{F}^{j-i} = (\mathbf{I}_p - \mathbf{F})^{-1},$$

and its $(1,1)$ element is

$$\frac{1}{1 - \psi_1 - \cdots - \psi_p},$$

which is the long-run effect of a change of u on y.

- By diagonalization, $\mathbf{C}^{-1}\mathbf{F}\mathbf{C} = \mathbf{\Lambda}$, where $\mathbf{C}$ is nonsingular and $\mathbf{\Lambda}$ is diagonal with all the eigenvalues of $\mathbf{F}$ on its main diagonal. Then,

$$\mathbf{F}^j = (\mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1})(\mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1})\cdots = \mathbf{C}\mathbf{\Lambda}^j\mathbf{C}^{-1},$$

  which converges to a zero matrix when all the eigenvalues of $\mathbf{F}$ are less than one in modulus (inside the unit circle).

- A *p*th-order difference equation is stable if all the eigenvalues of $\mathbf{F}$ are less than one in modulus. It is explosive if there is at least one eigenvalue greater than one in modulus.

- The eigenvalues of $\mathbf{F}$ are the roots of the characteristic equation:

$$\lambda^p - \psi_1\lambda^{p-1} - \cdots - \psi_{p-1}\lambda - \psi_p = 0,$$

  and hence are also known as characteristic roots.

Example: Consider the second-order difference equation with

$$\mathbf{F} = \begin{bmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{bmatrix}.$$

Its eigenvalues are the roots of:

$$\det(\mathbf{F} - \lambda \mathbf{I}_2) = -(\psi_1 - \lambda)\lambda - \psi_2 = \lambda^2 - \psi_1\lambda - \psi_2 = 0.$$

These two roots are

$$\lambda_1 = \frac{\psi_1 + \sqrt{\psi_1^2 + 4\psi_2}}{2}, \qquad \lambda_2 = \frac{\psi_1 - \sqrt{\psi_1^2 + 4\psi_2}}{2}.$$

And $\lambda = a + bi$ is less than one in modulus if $|\lambda| = (a^2 + b^2)^{1/2} < 1$; that is, λ is inside the unit circle on the complex plane.

Back-Shift Operator

- Theback-shift operator $\mathcal{B}$: $\mathcal{B}y_t = y_{t-1}$, $\mathcal{B}^2 y_t = \mathcal{B}(\mathcal{B}y_t) = y_{t-2}$, etc.
- First-order difference equation is

$$y_t = \psi_1 \mathcal{B} y_t + u_t, \quad \text{or} \quad (1 - \psi_1 \mathcal{B}) y_t = u_t.$$

- Pre-multiplying both sides of this equation by $(1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t)$ we have

$$\begin{aligned} &(1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t) u_t \\ &\qquad = (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t)(1 - \psi_1 \mathcal{B}) y_t \\ &\qquad = (1 - \psi_1^{t+1} \mathcal{B}^{t+1}) y_t, \end{aligned}$$

which is approximately y_t when t is large and $|\psi_1| < 1$.

- Passing to the limit we can define, for $|\psi_1| < 1$,

$$(1 - \psi_1 \mathcal{B})^{-1} = \lim_{t \to \infty} (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots + \psi_1^t \mathcal{B}^t),$$

  so that $(1 - \psi_1 \mathcal{B})(1 - \psi_1 \mathcal{B})^{-1} = \mathcal{I}$, the identity operator.
- Recall: $(1 - \psi_1 \mathcal{B} - \psi_2 \mathcal{B}^2) y_t = u_t$ is stable if $\lambda^2 - \psi_1 \lambda - \psi_2 = 0$ has all the roots inside the unit circle.
  - Setting $\lambda = z^{-1}$ and multiplying both sides by z^2, we obtain

$$(1 - \psi_1 z - \psi_2 z^2) = (1 - \lambda_1 z)(1 - \lambda_2 z) = 0,$$

    which has roots: $z_1 = 1/\lambda_1$ and $z_2 = 1/\lambda_2$.
  - A second-order difference equation is stable if all the roots of $(1 - \psi_1 z - \psi_2 z^2) = 0$ are outside the unit circle.
- A pth-order difference equation is stable if all the roots of $1 - \psi_1 z - \ldots - \psi_p z^p = 0$ are outside the unit circle.

Moving Average Series

- $\{y_t\}$ is a moving average (MA) series if

$$y_t = \mu + \Pi(\mathcal{B})\varepsilon_t,$$

where $\{\varepsilon_t\}$ is a white noise with mean zero and variance σ_ε^2.

- For the MA(1) series: $y_t = \mu + \varepsilon_t - \pi_1 \varepsilon_{t-1}$, $\mathbb{E}(y_t) = \mu$ and

$$\begin{aligned}
\gamma_0 &= \mathbb{E}[(\varepsilon_t - \pi_1\varepsilon_{t-1})^2] = (1+\pi_1^2)\sigma_\varepsilon^2, \\
\gamma_1 &= \mathbb{E}[(\varepsilon_t - \pi_1\varepsilon_{t-1})(\varepsilon_{t-1} - \pi_1\varepsilon_{t-2})] = -\pi_1\sigma_\varepsilon^2, \\
\gamma_j &= 0, \qquad j = 2, 3, \ldots.
\end{aligned}$$

Hence, $\rho_1 = -\pi_1/(1+\pi_1^2)$ and $\rho_j = 0$ for $j = 2, 3, \ldots$.

- It is weakly stationary regardless of the value of π_1.

Moving Average Series

- $\{y_t\}$ is a moving average (MA) series if

$$y_t = \mu + \Pi(\mathcal{B})\varepsilon_t,$$

where $\{\varepsilon_t\}$ is a white noise with mean zero and variance σ_ε^2.

- For the MA(1) series: $y_t = \mu + \varepsilon_t - \pi_1 \varepsilon_{t-1}$, $\mathbb{E}(y_t) = \mu$ and

$$\begin{aligned}
\gamma_0 &= \mathbb{E}[(\varepsilon_t - \pi_1 \varepsilon_{t-1})^2] = (1 + \pi_1^2)\sigma_\varepsilon^2, \\
\gamma_1 &= \mathbb{E}[(\varepsilon_t - \pi_1 \varepsilon_{t-1})(\varepsilon_{t-1} - \pi_1 \varepsilon_{t-2})] = -\pi_1 \sigma_\varepsilon^2, \\
\gamma_j &= 0, \qquad j = 2, 3, \ldots.
\end{aligned}$$

Hence, $\rho_1 = -\pi_1/(1 + \pi_1^2)$ and $\rho_j = 0$ for $j = 2, 3, \ldots$.

- It is weakly stationary regardless of the value of π_1.

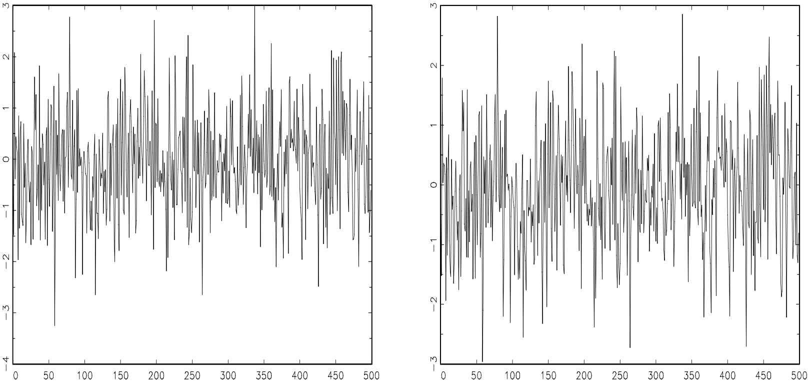

Figure: White noise (left) and MA series with $\pi_1 = 0.2$ (right).

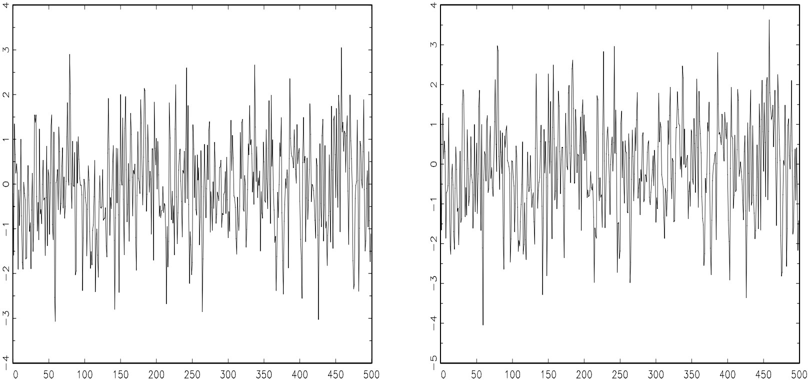

Figure: MA series with $\pi_1 = 0.5$ (left) and $\pi_1 = 0.8$ (right).

- Extending to the MA(q) series:

$$y_t = \mu + \varepsilon_t - \pi_1 \varepsilon_{t-1} - \pi_2 \varepsilon_{t-2} - \cdots - \pi_q \varepsilon_{t-q},$$

  we have $\mathbb{E}(y_t) = \mu$, $\gamma_0 = (1 + \pi_1^2 + \cdots + \pi_q^2)\sigma_\varepsilon^2$, and

$$\gamma_j = \left(\sum_{k=1}^{q-j} \pi_k \pi_{k+j} - \pi_j \right) \sigma_\varepsilon^2,$$

  with $\pi_0 = -1$ and $\pi_j = 0$ if $j > q$.

- As $\gamma_j = 0$ and $\rho_j = 0$ for $j = q+1, q+2, \ldots$, an MA(q) series has only a fixed "memory" of q periods.

- It is also weakly stationary regardless of the values of its MA coefficients.

- Consider now the MA(∞) series:

$$y_t = \mu + \varepsilon_t - \sum_{j=1}^{\infty} \pi_j \varepsilon_{t-j}.$$

  Then, $\mathbb{E}(y_t) = \mu$ and

$$\gamma_j = \left(\sum_{k=1}^{\infty} \pi_k \pi_{k+j} - \pi_j \right) \sigma_{\varepsilon}^2, \quad j = 0, 1, 2, \ldots$$

- By the Cauchy-Schwartz inequality,

$$\sum_{k=0}^{\infty} \pi_k \pi_{k+j} \leq \left(\sum_{k=0}^{\infty} \pi_k^2 \right)^{1/2} \left(\sum_{k=0}^{\infty} \pi_{k+j}^2 \right)^{1/2} < \infty.$$

  Hence, all γ_j are well defined when $\sum_{j=0}^{\infty} \pi_j^2 < \infty$.

- MA(∞) series is weakly stationary provided that its MA coefficients π_j are square summable.

- Consider now the MA(∞) series:

$$y_t = \mu + \varepsilon_t - \sum_{j=1}^{\infty} \pi_j \varepsilon_{t-j}.$$

Then, $\mathbb{E}(y_t) = \mu$ and

$$\gamma_j = \left(\sum_{k=1}^{\infty} \pi_k \pi_{k+j} - \pi_j \right) \sigma_\varepsilon^2, \quad j = 0, 1, 2, \ldots$$

- By the Cauchy-Schwartz inequality,

$$\sum_{k=0}^{\infty} \pi_k \pi_{k+j} \le \left(\sum_{k=0}^{\infty} \pi_k^2 \right)^{1/2} \left(\sum_{k=0}^{\infty} \pi_{k+j}^2 \right)^{1/2} < \infty.$$

Hence, all γ_j are well defined when $\sum_{j=0}^{\infty} \pi_j^2 < \infty$.

- MA(∞) series is weakly stationary provided that its MA coefficients π_j are square summable.

Autoregressive Series

- An autoregressive (AR) series is:

$$\Psi(\mathcal{B}) y_t = c + \varepsilon_t,$$

  where $\{\varepsilon_t\}$ is again a white noise.

- For $\Psi(\mathcal{B}) = 1 - \psi_1 \mathcal{B}$, $y_t = c + \psi_1 y_{t-1} + \varepsilon_t$. When $|\psi_1| < 1$, this AR(1) series has an MA(∞) representation:

$$\begin{aligned} y_t &= (1 - \psi_1 \mathcal{B})^{-1}(c + \varepsilon_t) \\ &= (1 + \psi_1 + \psi_1^2 + \cdots)c + (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots)\varepsilon_t \\ &= c/(1 - \psi_1) + (1 + \psi_1 \mathcal{B} + \psi_1^2 \mathcal{B}^2 + \cdots)\varepsilon_t, \end{aligned}$$

  where $1/(1 - \psi_1)$ is just $\Psi(1)^{-1}$.

- The AR(1) series with $|\psi_1| < 1$ is weakly stationary. (why?)
- The weakly stationary AR(1) series with $\mathbb{E}(y_t) = c/(1 - \psi_1)$. Utilizing the result

$$\gamma_j = \left(\sum_{k=1}^{\infty} \pi_k \pi_{k+j} - \pi_j \right) \sigma_\varepsilon^2,$$

and noting $\pi_k = -\psi_1^k$, we obtain

$$\gamma_j = \psi_1^j \frac{\sigma_\varepsilon^2}{1 - \psi_1^2} = \psi_1 \gamma_{j-1} = \psi_1^j \gamma_0, \quad j = 0, 1, 2, \ldots,$$

and hence $\rho_j = \psi_1 \rho_{j-1} = \psi_1^j$. The autocovariances and autocorrelations of an AR(1) series have the same AR(1) structure.

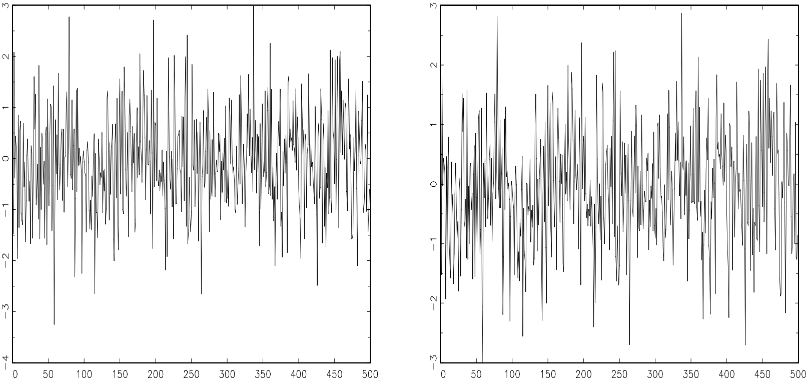

Figure: White noise (left) and AR series with $\psi_1 = 0.2$ (right).

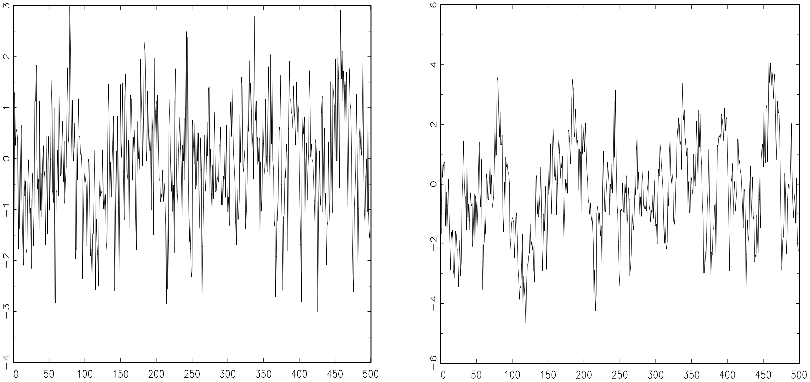

Figure: AR series with $\psi_1 = 0.5$ (left) and $\psi_1 = 0.8$ (right).

- An AR(p) series with $\Psi(\mathcal{B}) = 1 - \psi_1\mathcal{B} - \psi_2\mathcal{B}^2 - \cdots - \psi_p\mathcal{B}^p$ is

$$y_t = c + \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p} + \varepsilon_t.$$

It is weakly stationary if $\Psi(z) = 0$ has all roots outside the unit circle.

- A weakly stationary AR(p) series has an MA(∞) representation:

$$y_t = \Psi(1)^{-1}c + \Psi(\mathcal{B})^{-1}\varepsilon_t.$$

so that $\mu = \Psi(1)^{-1}c = c/(1 - \psi_1 - \psi_2 - \cdots - \psi_p)$, and

$$\gamma_j = \psi_1\gamma_{j-1} + \psi_2\gamma_{j-2} + \cdots + \psi_p\gamma_{j-p}, \quad j = 1, 2, \ldots,$$

and $\gamma_0 = \psi_1\gamma_1 + \psi_2\gamma_2 + \cdots + \psi_p\gamma_p + \sigma_\varepsilon^2$.

- For autocorrelations, we have the Yule-Walker equations:

$$\rho_j = \psi_1\rho_{j-1} + \psi_2\rho_{j-2} + \cdots + \psi_p\rho_{j-p}, \quad j = 1, 2, \ldots,$$

which form a pth-order difference equation in ρ_j.

Autoregressive Moving Average Series

- An ARMA(p, q) series is:

  $$\Psi(\mathcal{B})y_t = c + \Pi(\mathcal{B})\varepsilon_t,$$

  where $\Psi(\mathcal{B})$ and $\Pi(\mathcal{B})$ are pth- and qth-order polynomial in $\mathcal{B}$. It is weakly stationary if all the roots of $\Psi(z) = 0$ are outside the unit circle.

- Letting $\Phi(\mathcal{B}) = \Psi(\mathcal{B})^{-1}\Pi(\mathcal{B})$, we have

  $$y_t = \Psi(1)^{-1}c + \Phi(\mathcal{B})\varepsilon_t = \Psi(1)^{-1}c + \sum_{j=0}^{\infty}\phi_j\varepsilon_{t-j},$$

  with $\phi_0 = 1$. Its mean is $\mu = c/(1 - \psi_1 - \cdots - \psi_p)$.

- The autocovariances and autocorrelations are of complex forms; We omit the details.
- In terms of the deviations from the mean, we have

$$\Psi(B)(y_t - \mu) = \Pi(B)\varepsilon_t.$$

For $j = q+1, q+2, \ldots,$

$$\Psi(B)(y_t - \mu)(y_{t-j} - \mu) = \Pi(B)\varepsilon_t(y_{t-j} - \mu),$$

and $\mathbb{E}[\Pi(B)\varepsilon_t(y_{t-j} - \mu)] = 0$, so that

$$\gamma_j = \psi_1\gamma_{j-1} + \cdots + \psi_p\gamma_{j-p}.$$

That is, the autocovariances for $j > q$ obey the AR(p) structure.

Invertibility of MA Series

- The MA series $y_t = \mu + \Pi(\mathcal{B})\varepsilon_t$ is invertible if all the roots of $\Pi(z) = 0$ are outside the unit circle.
- An invertible MA(1) series has the following AR(∞) representation:

$$(1 - \pi_1 \mathcal{B})^{-1}(y_t - \mu) = \sum_{j=0}^{\infty} \pi_1^j \mathcal{B}^j (y_t - \mu) = \varepsilon_t.$$

  Each innovation ε_t can be expressed as a weighted sum of current and all past observations y_t.

- Similarly, each innovation ε_t of an invertible MA(q) series can also be expressed as a weighted sum of current and all past y_t.

- An MA(1) series with $|\pi_1| > 1$ is non-invertible, as $(1 + \pi_1 \mathcal{B} + \pi_1^2 \mathcal{B}^2 + \cdots)$ can not be defined as $(1 - \pi_1 \mathcal{B})^{-1}$.
- Consider the forward-shift operator $\mathcal{B}^{-1}$, where $\mathcal{B}^{-1}$ is such that $\mathcal{B}^{-1} y_t = y_{t+1}$ and $\mathcal{B}^{-1}\mathcal{B} = \mathcal{I}$. Then, $(1 - \pi_1^{-1}\mathcal{B}^{-1})$ has all the roots inside the unit circle, and its inverse is

$$(1 - \pi_1^{-1}\mathcal{B}^{-1})^{-1} = (1 + \pi_1^{-1}\mathcal{B}^{-1} + \pi_1^{-2}\mathcal{B}^{-2} + \cdots).$$

- Straightforward calculation shows that

$$-\pi_1^{-1}\mathcal{B}^{-1}(1 + \pi_1^{-1}\mathcal{B}^{-1} + \pi_1^{-2}\mathcal{B}^{-2} + \cdots)(1 - \pi_1 \mathcal{B}) = \mathcal{I},$$

where the second term in parentheses is $(1 - \pi_1^{-1}\mathcal{B}^{-1})^{-1}$.

- For $|\pi_1| > 1$, we can define

$$(1 - \pi_1 \mathcal{B})^{-1} = -\big(1 - \pi_1^{-1} \mathcal{B}^{-1}\big)^{-1} \big(\pi_1^{-1} \mathcal{B}^{-1}\big),$$

  which is in terms of the forward-shift operator.

- A non-invertible MA(1) series is thus

$$-\pi_1^{-1} \mathcal{B}^{-1} (1 + \pi_1^{-1} \mathcal{B}^{-1} + \pi_1^{-2} \mathcal{B}^{-2} + \cdots)(y_t - \mu) = \varepsilon_t,$$

  so that ε_t is a weighted sum of all future y_t.

- For a non-invertible MA(q) series, each innovation ε_t also depends on all future observations y_{t+j}, $j > 0$. As far as forecasting is concerned, a non-invertible MA series does not make much sense.

Orthogonal Projection

Let $\mathbf{x}_t = (1, y_t, y_{t-1}, \ldots, y_{t-m+1})'$, an $(m+1) \times 1$ vector.

- $\mathbb{E}(y_{t+1}|\mathbf{x}_t)$ minimizes the mean squared error (MSE):

  $$\mathbb{E}[y_{t+1} - g(\mathbf{x}_t)]^2,$$

  among all measurable functions of $\mathbf{x}_t$, such that

  $$\mathbb{E}\big\{[y_{t+1} - \mathbb{E}(y_{t+1}|\mathbf{x}_t)]g(\mathbf{x}_t)\big\} = 0.$$

- $\mathbb{E}(y_{t+1}|\mathbf{x}_t)$ is the orthogonal projection of y_{t+1} onto the space of functions of $\mathbf{x}_t$ in the MSE sense, also known as the best L_2-predictor of y_{t+1}.

Linear Projection

- $\widehat{P}(y_{t+1}|\mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\alpha}$ is the linear projection of y_{t+1} if it minimizes

  $$\mathbb{E}[y_{t+1} - \ell(\mathbf{x}_t)]^2,$$

  among all linear functions of $\mathbf{x}_t$, such that

  $$\mathbb{E}[\mathbf{x}_t(y_{t+1} - \mathbf{x}_t'\boldsymbol{\alpha})] = \mathbf{0}.$$

- Analogous to the OLS estimator, $\boldsymbol{\alpha} = [\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')]^{-1}\,\mathbb{E}(\mathbf{x}_t y_{t+1})$.
- Note that the orthogonal projection need not be a linear function of $\mathbf{x}_t$ and that the linear projection is not the orthogonal projection, except in some special cases.

Forecasts: Infinite Observations

Consider the MA(∞) series: $y_t = \mu + \varepsilon_t - \sum_{j=1}^{\infty} \pi_j \varepsilon_{t-j}$, so that

$$
\begin{aligned}
y_{t+s} &= \mu + \varepsilon_{t+s} - \sum_{j=1}^{\infty} \pi_j \varepsilon_{t+s-j} \\
&= \underbrace{\varepsilon_{t+s} - \pi_1 \varepsilon_{t+s-1} - \cdots - \pi_{s-1} \varepsilon_{t+1}}_{y_{t+s} - \widehat{P}(y_{t+s} | \varepsilon_j, j \leq t)} \\
&\qquad \underbrace{+\mu - \pi_s \varepsilon_t - \pi_{s+1} \varepsilon_{t-1} - \pi_{s+2} \varepsilon_{t-2} - \cdots}_{\widehat{P}(y_{t+s} | \varepsilon_j, j \leq t)},
\end{aligned}
$$

provided we observe all ε_j, $j \leq t$, and know μ and all π_j. Clearly, $y_{t+s} - \widehat{P}(y_{t+s} | \varepsilon_j, j \leq t)$ is uncorrelated with any ε_j, $j \leq t$, and

$$
\mathrm{MSE}\big(\widehat{P}(y_{t+s} | \varepsilon_j, j \leq t)\big) = (1 + \pi_1^2 + \cdots + \pi_{s-1}^2)\sigma_\varepsilon^2.
$$

As an example, suppose that y_t is an MA(q) series. We have

$$\widehat{P}(y_{t+s}|\varepsilon_j, j \leq t) = \mu - \sum_{j=s}^{q} \pi_j \varepsilon_{t+s-j}, \quad s = 1, 2, \ldots, q,$$

and $\widehat{P}(y_{t+s}|\varepsilon_j, j \leq t) = \mu$, $s = q+1, q+2, \ldots$. The resulting MSEs are:

$$\begin{array}{ll} \sigma_\varepsilon^2, & s = 1, \\ (1 + \pi_1^2 + \cdots + \pi_{s-1}^2)\sigma_\varepsilon^2, & s = 2, 3, \ldots, q, \\ (1 + \pi_1^2 + \cdots + \pi_q^2)\sigma_\varepsilon^2, & s = q+1, q+2, \ldots \end{array}$$

Thus, to predict an MA(q) series more than q periods ahead, the optimal linear forecast is the unconditional mean μ, and the MSE remains constant. (What is the intuition?)

For MA(∞) series,

$$\Pi(\mathcal{B}) = 1 - \pi_1 \mathcal{B} - \pi_2 \mathcal{B}^2 - \cdots - \pi_s \mathcal{B}^s - \pi_{s+1} \mathcal{B}^{s+1} - \cdots .$$

Letting $[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+$ denote the polynomial with only non-negative power of $\mathcal{B}$ we have

$$[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ = -\pi_s - \pi_{s+1}\mathcal{B} - \pi_{s+2}\mathcal{B}^2 - \cdots .$$

It follows that

$$\widehat{P}(y_{t+s} | \varepsilon_j, j \leq t) = \mu + [\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ \varepsilon_t .$$

When the MA series is invertible, we have $\Psi(\mathcal{B})(y_t - \mu) = \varepsilon_t$ such that $\Psi(\mathcal{B}) = [\Pi(\mathcal{B})]^{-1}$. Hence, ε_t can be constructed from current and lagged y_t, so that

$$\begin{aligned}\widehat{P}(y_{t+s}|\varepsilon_j, j \leq t) &= \widehat{P}(y_{t+s}|y_j, j \leq t) \\ &= \mu + [\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ \Psi(\mathcal{B})(y_t - \mu) \\ &= \mu + [\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ [\Pi(\mathcal{B})]^{-1}(y_t - \mu).\end{aligned}$$

The last expression is known as the Wiener-Kolmogorov prediction formula which relies solely on the observed values: y_t and its lagged values.

Example: Forecasting AR(1) Series

Consider $(1-\psi_1\mathcal{B})(y_t-\mu)=\varepsilon_t$ so that

$$\Pi(\mathcal{B})=1+\psi_1\mathcal{B}+\psi_1^2\mathcal{B}^2+\cdots=[1-\psi_1\mathcal{B}]^{-1}.$$

Then,

$$[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+=\psi_1^s+\psi_1^{s+1}\mathcal{B}+\psi_1^{s+2}\mathcal{B}^2+\cdots=\psi_1^s/(1-\psi_1\mathcal{B})=\psi_1^s\Pi(\mathcal{B}).$$

Consequently, the Wiener-Kolmogorov prediction formula is

$$\widehat{P}(y_{t+s}|y_j,j\leq t)=\mu+[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+[\Pi(\mathcal{B})]^{-1}(y_t-\mu)=\mu+\psi_1^s(y_t-\mu),$$

which tends toward μ geometrically as s increases.

Example: Forecasting MA(1) Series

Consider $(y_t - \mu) = (1 - \pi_1 \mathcal{B})\varepsilon_t$. For $s = 1$, $[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ = -\pi_1$ and

$$\begin{aligned}\widehat{P}(y_{t+s}|y_j, j \leq t) &= \mu + [\mathcal{B}^{-s}\Pi(\mathcal{B})]_+[\Pi(\mathcal{B})]^{-1}(y_t - \mu) \\ &= \mu - \pi_1(y_t - \mu) - \pi_1^2(y_{t-1} - \mu) - \cdots .\end{aligned}$$

For $s = 2, 3, \ldots$, $[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ = 0$ and $\widehat{P}(y_{t+s}|y_j, j \leq t) = \mu$.

Writing $\varepsilon_t = (1 - \pi_1 \mathcal{B})^{-1}(y_t - \mu)$, we can express ε_t using a recursion:

$$\hat{\varepsilon}_t = (y_t - \mu) + \pi_1 \hat{\varepsilon}_{t-1} = (y_t - \mu) + \pi_1(y_{t-1} - \mu) + \cdots .$$

It follows that for $s = 1$,

$$\widehat{P}(y_{t+s}|y_j, j \leq t) = \mu - \pi_1 \hat{\varepsilon}_t.$$

Example: Forecasting MA(q) Series

Consider $(y_t - \mu) = (1 - \pi_1 \mathcal{B} - \cdots - \pi_q \mathcal{B}^q)\varepsilon_t$. For $s = 1, 2, \ldots, q$,

$$[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ = -\pi_s - \pi_{s+1}\mathcal{B} - \cdots - \pi_q \mathcal{B}^{q-s}$$

Using the recursion: $\hat{\varepsilon}_t = (y_t - \mu) + \pi_1 \hat{\varepsilon}_{t-1} + \cdots + \pi_q \hat{\varepsilon}_{t-q}$, The Wiener-Kolmogorov prediction formula now reads

$$\widehat{P}(y_{t+s}|y_j, j \leq t) = \mu - \pi_s \hat{\varepsilon}_t - \pi_{s+1}\hat{\varepsilon}_{t-1} - \cdots - \pi_q \hat{\varepsilon}_{t+s-q}.$$

For $s = q+1, q+2, \ldots$, $[\mathcal{B}^{-s}\Pi(\mathcal{B})]_+ = 0$ and $\widehat{P}(y_{t+s}|y_j, j \leq t) = \mu$.

Forecasts: Finite Observations

Given finitely many observations, we may approximate $\widehat{P}(y_{t+s}|y_j, j \leq t)$ by

$$\widehat{P}(y_{t+s}|y_t, \ldots, y_{t-m+1}, \varepsilon_{t-m} = 0, \varepsilon_{t-m-1} = 0, \ldots).$$

For an MA(q) series, set $\hat{\varepsilon}_{t-m} = \hat{\varepsilon}_{t-m-1} = \cdots = \hat{\varepsilon}_{t-m-q+1} = 0$. Using the recursion, $\hat{\varepsilon}_t = (y_t - \mu) + \pi_1 \hat{\varepsilon}_{t-1} + \cdots + \pi_q \hat{\varepsilon}_{t-q}$, we obtain

$$\hat{\varepsilon}_{t-m+1} = (y_{t-m+1} - \mu),$$

$$\hat{\varepsilon}_{t-m+2} = (y_{t-m+2} - \mu) + \pi_1 \hat{\varepsilon}_{t-m+1},$$

$$\hat{\varepsilon}_{t-m+3} = (y_{t-m+3} - \mu) + \pi_1 \hat{\varepsilon}_{t-m+2} + \pi_2 \hat{\varepsilon}_{t-m+1},$$

and so on. These $\hat{\varepsilon}_t$ values are then used to compute the approximation.

Exact Finite-Sample Forecasts

We may also compute the projection of y_{t+s} on m lagged values. For $s = 1$, we want to compute:

$$\mathbf{x}_t'\boldsymbol{\alpha}^{(m)} = \alpha_1^{(m)}(y_t - \mu) + \alpha_2^{(m)}(y_{t-1} - \mu) + \cdots + \alpha_m^{(m)}(y_{t-m+1} - \mu).$$

The projection coefficients are

$$\begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \vdots \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}.$$

The results are readily generalized for $s = 2, 3, \ldots$. We want to compute

$$
\begin{aligned}
&\widehat{P}(y_{t+s} | y_t, \ldots, y_{t-m+1}) \\
&\quad = \alpha_1^{(m,s)}(y_t - \mu) + \alpha_2^{(m,s)}(y_{t-1} - \mu) + \cdots + \alpha_m^{(m,s)}(y_{t-m+1} - \mu),
\end{aligned}
$$

where the projection coefficients are

$$
\begin{bmatrix} \alpha_1^{(m,s)} \\ \alpha_2^{(m,s)} \\ \vdots \\ \alpha_m^{(m,s)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_s \\ \gamma_{s+1} \\ \vdots \\ \gamma_{s+m-1} \end{bmatrix}.
$$

Wold's Decomposition

Any zero-mean, covariance-stationary series y_t can be expressed as:

$$y_t = \sum_{j=0}^{\infty} \pi_j \varepsilon_{t-j} + \kappa_t,$$

with $\pi_0 = 1$ and π_j square summable, where ε_t is a white noise, and κ_t can be predicted arbitrarily well using lagged y_t.

- y_t is the sum of two components: an MA(∞) component and a linearly deterministic component.
- We need to find (identify) a model that can properly represent the MA(∞) component.

The standard Box-Jenkins approach consists of the following steps:

1. Transform the original time series to a covariance stationary series.
2. Identify a preliminary ARMA(p, q) model for the transformed series.
3. Estimate unknown parameters in this preliminary model.
4. Conduct diagnostic tests to check model adequacy and re-estimate an ARMA model when the preliminary model is found inappropriate.

The steps 2–4 may be repeated until a suitable model is found.

Differencing

When the series η_t exhibits trending pattern, its trend may be removed by differencing:

$$y_t = \eta_t - \eta_{t-1} = (1 - \mathcal{B})\eta_t.$$

In other words, η_t are integrated y_t.

- If y_t is an ARMA series, η_t is known as an ARIMA (autoregressive, integrated, moving average) series.
- If y_t is a sequence of i.i.d. random variables, η_t is an ARIMA$(0, 1, 0)$ series and also known as a random walk.

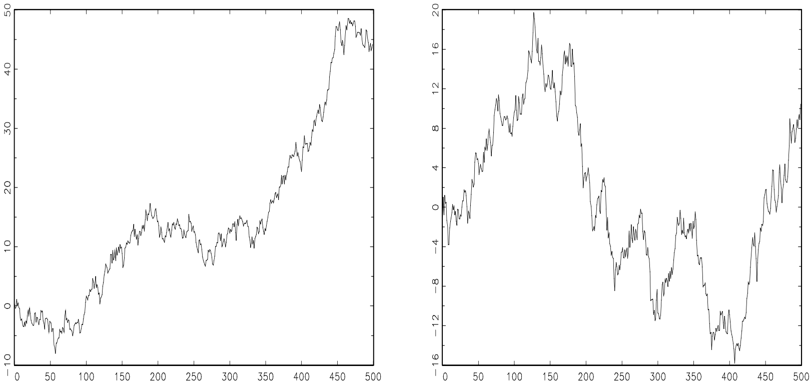

Figure: The time paths of a Gaussian random walk.

- When $(1-\mathcal{B})\eta_t = y_t$ and y_t is a weakly stationary AR(1) series: $y_t = \psi_1 y_{t-1} + \varepsilon_t$, η_t is also an ARMA$(2,0)$ series:

  $$(1-\psi_1\mathcal{B})(1-\mathcal{B})\eta_t = [1-(1+\psi_1)+\psi_1\mathcal{B}^2]\eta_t = \varepsilon_t.$$

  Here, the AR polynomial $\Psi(z)=0$ has a root on the unit circle, also known as a unit root.

- Similarly, when y_t is a stationary ARMA(p,q) series, η_t is an ARIMA$(p,1,q)$ series or an ARMA$(p+1,q)$ series with an AR unit root. An ARIMA$(p,1,q)$ series is also known as an integrated series, or simply an $I(1)$ series.

- When y_t are obtained by differencing η_t twice:

$$y_t = (\eta_t - \eta_{t-1}) - (\eta_{t-1} - \eta_{t-2}) = (\eta_t - 2\eta_{t-1} + \eta_{t-2}),$$

  η_t is an ARIMA$(p, 2, q)$ series or an $I(2)$ series.

- An ARIMA(p, d, q) series is an $I(d)$ series, and it must be differenced d times to yield a stationary ARMA representation.

- Seasonal pattern may be eliminated by seasonal differencing:

$$y_t = \eta_t - \eta_{t-4} = (1 - \mathcal{B}^4)\eta_t.$$

  Note that $(1 - z^4) = 0$ contains four unit roots because

$$(1 - z^4) = (1 - z^2)(1 + z^2) = (1 - z)(1 + z)(1 + iz)(1 - iz);$$

  each unit root accounts for the behavior of η_t at some frequency.

Other approaches for removing deterministic components:

- Eliminating a deterministic trend by regressing η_t on the time trend variable t and/or higher orders of t: $t, t^2, \ldots, t^p$.
- Eliminating quarterly pattern by regressing η_t on quarterly dummies: $D_{1,t} = 1$ if t is in the first quarter and $D_{1,t} = 0$ otherwise; $D_{2,t} = 1$ if t is in the second quarter and $D_{2,t} = 0$ otherwise; $D_{3,t} = 1$ if t is in the third quarter and $D_{3,t} = 0$ otherwise.
- Eliminating the day-of-week effect by regressing η on a daily dummy.
- There are other "filters" in the literature or even in some statistics softwares.

Identification

- Sample autocovariances:

$$\hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y}),$$

  with $\bar{y} = \sum_{t=1}^{T} y_t / T$, and Sample autocorrelations:

$$\hat{\rho}_j = \hat{\gamma}_j / \hat{\gamma}_0.$$

- Under regularity conditions,

$$\hat{\rho}_j \xrightarrow{\mathbb{P}} \gamma_j / \gamma_0 = \rho_j,$$

  and $\sqrt{T}\hat{\rho}_j$ are asymptotically normally distributed.

- $\hat{\rho}_j$ can help to identify an MA model because ρ_j of an MA(q) series has an abrupt cut-off at $j = q$ with $\rho_j = 0$ for $j > q$.
- For those j such that $\rho_j \approx 0$,

$$\operatorname{var}(\hat{\rho}_j) \approx \frac{1}{T} \sum_{i=-\infty}^{\infty} \rho_i^2.$$

In particular, for an MA(q) series,

$$\operatorname{var}(\hat{\rho}_j) \approx \frac{1}{T}\left(1 + 2\rho_1^2 + \cdots + 2\rho_q^2\right), \quad j = q+1, q+2, \ldots.$$

- The 95% confidence intervals of $\hat{\rho}_j$, $j = q+1, q+2, \ldots$, are

$$\pm \frac{1.96}{\sqrt{T}}\left(1 + 2\hat{\rho}_1^2 + \cdots + 2\hat{\rho}_q^2\right)^{1/2}.$$

- For a white noise,

$$\text{var}(\hat{\rho}_j) \approx 1/T, \qquad j = 1, 2, \ldots,$$

  so that $\pm 1.96/\sqrt{T}$ form the 95% confidence interval.

- Many programs plot $\hat{\rho}_j$ against j and use $\pm 1.96/\sqrt{T}$ (or $\pm 2/\sqrt{T}$) as the 95% confidence interval. This is appropriate only for checking the autocorrelations of a white noise, however.

- Even it is appropriate, this confidence interval is for checking a single sample autocorrelation but not for checking m sample autocorrelations jointly.

Partial Autocorrelations

- The partial autocorrelation $\alpha_m^{(m)}$ of y_t are

  $$\begin{aligned} \alpha_1^{(1)} &= \operatorname{corr}(y_t, y_{t-1}) = \rho_1, \\ \alpha_m^{(m)} &= \operatorname{corr}\big[y_t - \widehat{P}(y_t \mid \mathcal{Y}_{t-m+1}^{t-1}),\ y_{t-m} - \widehat{P}(y_{t-m} \mid \mathcal{Y}_{t-m+1}^{t-1})\big], \end{aligned}$$

  for $m = 2, 3, \ldots$. By the Frisch-Waugh-Lovell Theorem, $\alpha_m^{(m)}$ is also the last coefficient of the linear projection of y_t on $1, y_{t-1}, \ldots, y_{t-m}$.
- For an AR(p) series y_t, it is correlated with y_{t-m} for $m \leq p$, so that the last coefficient of the linear projection of y_t on $1, y_{t-1}, \ldots, y_{t-m}$ should be different from zero. For $m > q$, the last coefficient of the linear projection above must be zero. Thus, the partial autocorrelations can help to identify an AR(p) model.

The first m sample partial autocorrelations $\hat{\alpha}_m^{(m)}$ are the coefficient estimates of $a_m^{(m)}$ of the following regressions:

$$\begin{aligned}
y_t &= a_0^{(1)} + a_1(1)y_{t-1} + e_t, \\
y_t &= a_0^{(2)} + a_1^{(2)}y_{t-1} + a_2^{(2)}y_{t-2} + e_t, \\
y_t &= a_0^{(3)} + a_1^{(3)}y_{t-1} + a_2^{(3)}y_{t-2} + a_3^{(3)}y_{t-3} + e_t, \\
y_t &= a_0^{(m)} + a_1^{(m)}y_{t-1} + a_2^{(m)}y_{t-2} + \cdots + a_m^{(m)}y_{t-m} + e_t.
\end{aligned}$$

It can be shown that $\mathrm{var}\big(\hat{\alpha}_m^{(m)}\big) \approx 1/T$ and for an AR(p) series:

$$\sqrt{T}\hat{\alpha}_m^{(m)} \xrightarrow{D} \mathcal{N}(0,1). \quad m = p+1, p+2, \ldots.$$

We may plot $\hat{\alpha}_m^{(m)}$ against m and use the confidence interval $(\pm 1.96/\sqrt{T})$ to evaluate $\hat{\alpha}_m^{(m)}$.

Model Estimation

When a preliminary ARMA(p, q) model $\Psi(\mathcal{B})y_t = c + \Pi(\mathcal{B})\varepsilon_t$ is chosen, we must estimate $\boldsymbol{\theta}$, the parameter vector that includes $\psi_1, \ldots, \psi_p$ in the AR polynomial, $\pi_1, \ldots, \pi_q$ in the MA polynomial, c, and the variance σ_ε^2.

- Quasi-Maximum Likelihood Estimation (QMLE): We maximize a postulated (log-) likelihood function with respect to $\boldsymbol{\theta}$. This likelihood function, which may or may not be correctly specified for the true density function underlying the data, is known as a quasi-likelihood function; the resulting maximizer is thus know as a QMLE, $\tilde{\boldsymbol{\theta}}_T$.
- The log-likelihood function for ARMA models is nonlinear in parameters in general and hence must be solved via some nonlinear optimization algorithms.

Estimation of AR Models

Consider the AR(p) model: $\Psi(\mathcal{B})y_t = c + \varepsilon_t$. Let $\mathbf{Y}^t = \{y_1, y_2, \ldots, y_t\}$ and $f(\mathbf{Y}^t; \boldsymbol{\theta})$ be its joint density function. Also let $f(y_t \mid \mathbf{Y}^{t-1}; \boldsymbol{\theta})$ be the conditional density. The joint quasi-likelihood function of $\mathbf{Y}^t$ is

$$\begin{aligned}
L_T(\mathbf{Y}^T; \boldsymbol{\theta}) &= f(y_T \mid \mathbf{Y}^{T-1}; \boldsymbol{\theta}) f(\mathbf{Y}^{T-1}; \boldsymbol{\theta}) \\
&= f(y_T \mid \mathbf{Y}^{T-1}; \boldsymbol{\theta}) f(y_{T-1} \mid \mathbf{Y}^{T-2}; \boldsymbol{\theta}) f(\mathbf{Y}^{T-2}; \boldsymbol{\theta}) \\
&= \cdots = \Bigg(\prod_{j=p+1}^{T} f(y_j \mid \mathbf{Y}^{j-1}; \boldsymbol{\theta}) \Bigg) f(\mathbf{Y}^p; \boldsymbol{\theta}).
\end{aligned}$$

and $\mathcal{L}_T(\mathbf{Y}^T; \boldsymbol{\theta})$, the average of the log-quasi-likelihood function, is

$$\frac{1}{T} \ln L_T(\mathbf{Y}^T; \boldsymbol{\theta}) = \frac{1}{T} \Bigg(\ln f(\mathbf{Y}^p; \boldsymbol{\theta}) + \sum_{j=p+1}^{T} \ln f(y_j \mid \mathbf{Y}^{j-1}; \boldsymbol{\theta}) \Bigg).$$

Conditional QMLE for AR Models

Assuming conditional normality for $f(y_t \mid \mathbf{Y}^{t-1}; \boldsymbol{\theta})$:

$$\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(\frac{-(y_t - c - \psi_1 y_{t-1} - \cdots - \psi_p y_{t-p})^2}{2\sigma_\varepsilon^2} \right).$$

Take the initial $y_1, \ldots, y_p$ as given, we can ignore $f(\mathbf{Y}^p; \boldsymbol{\theta})$ and maximize

$$\begin{aligned} \mathcal{L}_T^c(\mathbf{Y}^T; \boldsymbol{\theta}) &= \frac{1}{T} \sum_{j=p+1}^{T} \ln f(y_j \mid \mathbf{Y}^{j-1}; \theta) \\ &= -\frac{T-p}{2T} \log(2\pi) - \frac{T-p}{2T} \log \sigma_\varepsilon^2 \\ &\qquad - \frac{1}{T} \sum_{j=p+1}^{T} \frac{\left(y_t - c - \psi_1 y_{t-1} - \cdots - \psi_p y_{t-p}\right)^2}{2\sigma_\varepsilon^2}. \end{aligned}$$

Conditional on the initial values $y_1, \ldots, y_p$, the QMLEs of $c, \psi_1, \ldots, \psi_p$ are the OLS estimators based on the data $y_{p+1}, \ldots, y_T$. The QMLE of σ_ε^2 is

$$\tilde{\sigma}^2 = \frac{1}{T-p} \sum_{j=p+1}^{T} \hat{\mathrm{e}}_t^2,$$

with $\hat{\mathrm{e}}_t$ the OLS residuals. Such estimators are known as the conditional QMLEs of the AR(p) model.

Exact QMLE for AR Models

With the joint normality assumption on $\mathbf{Y}^p$,

$$f(\mathbf{Y}^p;\boldsymbol{\theta}) = (2\pi\sigma_\varepsilon^2)^{-p/2}\det(\mathbf{V}_p)^{-1/2} \exp\left[\frac{-1}{2\sigma_\varepsilon^2}\left(\mathbf{Y}^p - \frac{c}{1-\psi_1-\cdots-\psi_p}\boldsymbol{\ell}\right)'\mathbf{V}_p^{-1} \left(\mathbf{Y}^p - \frac{c}{1-\psi_1-\cdots-\psi_p}\boldsymbol{\ell}\right)\right],$$

where $\mathbf{V}_p = \mathrm{var}(bY^p)$ and $\boldsymbol{\ell}$ is the p-dimensional vector of ones.

Letting γ_j denote the autocovariances as before, we have

$$\sigma_\varepsilon^2 \mathbf{V}_p = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{pmatrix}.$$

Note that $\mathbf{V}_p^{-1}$ can be expressed in terms of AR parameters; see Hamilton (1994, p. 125). For example, for $p = 1$, $\mathbf{V}_p^{-1} = 1 - \psi_1^2$; for $p = 2$,

$$\mathbf{V}_p^{-1} = \begin{pmatrix} (1 - \psi_2^2) & -(\psi_1 + \psi_1\psi_2) \\ -(\psi_1 + \psi_1\psi_2) & (1 - \psi_2^2) \end{pmatrix}.$$

The log-likelihood function $\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta})$ now reads:

$$
\begin{aligned}
&-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\sigma_\varepsilon^2+\frac{1}{2}\log\big(\det(\mathbf{V}_p^{-1})\big)\\
&-\frac{1}{2T\sigma_\varepsilon^2}\Big(\mathbf{Y}^p-\frac{c}{1-\psi_1-\cdots-\psi_p}\boldsymbol{\ell}\Big)'\mathbf{V}_p^{-1}\Big(\mathbf{Y}^p-\frac{c}{1-\psi_1-\cdots-\psi_p}\boldsymbol{\ell}\Big)\\
&-\frac{1}{T}\sum_{j=p+1}^{T}\frac{\big(y_t-c-\psi_1 y_{t-1}-\cdots-\psi_p y_{t-p}\big)^2}{2\sigma_\varepsilon^2},
\end{aligned}
$$

which is now a complex nonlinear function in parameters. The resulting maximizers are the exact QMLEs.

Estimation of MA Models

Consider the MA(1) model: $y_t = \mu + \varepsilon_t - \pi_1 \varepsilon_{t-1}$. Then,

$$\varepsilon_t = y_t - \mu + \pi_1 \varepsilon_{t-1},$$

so that for $\varepsilon_0 = 0$, $\varepsilon_1 = y_1 - \mu$, $\varepsilon_2 = y_2 - \mu + \pi_1(y_1 - \mu)$, and so on. Given

$$\begin{aligned} f(y_t|\mathbf{Y}^{t-1}, \varepsilon_0 = 0; \boldsymbol{\theta}) &= f(y_t|\varepsilon_{t-1}, \varepsilon_0 = 0; \boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(\frac{-(y_t - \mu + \pi_1 \varepsilon_{t-1})^2}{2\sigma_\varepsilon^2} \right), \end{aligned}$$

the quasi-log-likelihood function conditional on $\varepsilon_0 = 0$ is

$$\begin{aligned} &\mathcal{L}(\mathbf{Y}^T|\varepsilon_0 = 0; \boldsymbol{\theta}) \\ &\quad = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_\varepsilon^2) - \frac{1}{T}\sum_{t=1}^{T} \frac{(y_t - \mu + \pi_1 \varepsilon_{t-1})^2}{2\sigma_\varepsilon^2}. \end{aligned}$$

Q: Why $\varepsilon_0 = 0$?

Ans: It is common to set ε_0 to its expected value 0.

Plugging the recursive formulae of ε_t into $\mathcal{L}(\mathbf{Y}^T | \varepsilon_0 = 0; \boldsymbol{\theta})$ results in a highly nonlinear function in parameters. The maximizer of this log-likelihood function is the conditional QMLE of the MA(1) model.

Similarly, the conditional QMLE of the MA(q) model is obtained conditional on $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0$. Here, ε_t are computed via the following recursions:

$$\varepsilon_t = y_t - \mu + \pi_1 \varepsilon_{t-1} + \cdots + \pi_q \varepsilon_{t-q}.$$

Estimation of ARMA Models

To compute the conditional QMLE of the ARMA(p,q), we need p initial values of $y_0, y_{-1}, \ldots, y_{-p+1}$ and q initial values of $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1}$. Here, ε_t are computed via

$$\varepsilon_t = y_t - c - \psi_1 y_{t-1} - \cdots - \psi_p y_{t-p} + \pi_1 \varepsilon_{t-1} + \cdots + \pi_q \varepsilon_{t-q}.$$

It is typical to set the initial ε's to zero and the initial y's to the expected value $c/(1 - \psi_1 - \cdots - \psi_p)$.

Remark: The exact and conditional QMLEs have different ways to handle initial values. Under weak stationarity, the effect of initial values eventually dies out, so that these two QMLEs are asymptotically equivalent.

Asymptotic Properties of the QMLE

The QMLE $\tilde{\boldsymbol{\theta}}_T$ maximizes the average of the quasi-log-likelihood function $\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta})$. Let $\boldsymbol{\theta}^*$ denote the unknown parameter vector that maximizes $\mathbb{E}\left[\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta})\right]$.

Consistency: Under suitable conditions, $\tilde{\boldsymbol{\theta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$. When $\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta})$ is "close" to $\mathbb{E}\left[\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta})\right]$ on the parameter space in a proper sense, the maximizer of the former, $\tilde{\boldsymbol{\theta}}_T$, will also be "close" to the maximizer of the latter, $\boldsymbol{\theta}^*$.

Asymptotic Normality: As $\tilde{\boldsymbol{\theta}}_T$ solves the average of the score: $\nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}) = \mathbf{0}$, we have from the mean-value expansion that

$$\mathbf{0} = \nabla\mathcal{L}_T(\mathbf{Y}^T;\tilde{\boldsymbol{\theta}}_T) = \nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}^*) + \nabla^2\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}^\dagger)(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*).$$

Under a weak uniform law of large number ensures,

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}^*) + o_{\mathbb{P}}(1),$$

where $\mathbf{H}_T(\boldsymbol{\theta}) = \mathbb{E}[\nabla^2\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta})]$. This shows that $T^{1/2}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)$ and $-\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}^*)$ are asymptotically equivalent. If $\sqrt{T}\nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}^*)$ obeys a central limit theorem such that

$$\mathbf{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\mathbf{0},\mathbf{I}),$$

with $\mathbf{B}_T(\boldsymbol{\theta}) = \mathrm{var}(\sqrt{T}\nabla\mathcal{L}_T(\mathbf{Y}^T;\boldsymbol{\theta}))$, we immediately have

$$\mathbf{C}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\mathbf{0},\mathbf{I}),$$

where $\mathbf{C}_T(\boldsymbol{\theta}^*) = \mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{B}_T(\boldsymbol{\theta}^*)\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}$.

When the likelihood is correctly specified, we have the information matrix equality:

$$\mathbf{H}_T(\boldsymbol{\theta}^*) + \mathbf{B}_T(\boldsymbol{\theta}^*) = \mathbf{0}.$$

Then, $\mathbf{C}_T(\theta^*) = -\mathbf{H}_T(\mathbf{Y}^T; \boldsymbol{\theta}^*)^{-1} = \mathbf{B}_T(\boldsymbol{\theta}^*)^{-1}$.

Remark: The information matrix equality may break down when the conditional normality assumption is invalid and/or when important dynamic structures are ignored in model specification.

Without the information matrix equality, $\mathbf{C}_T(\boldsymbol{\theta}^*)$ can not be simplified, and both $\mathbf{H}_T(\boldsymbol{\theta}^*)$ and $\mathbf{B}_T(\boldsymbol{\theta}^*)$ must be estimated. Clearly, $\mathbf{H}_T(\boldsymbol{\theta}^*)$ can be estimated using its sample counterpart: $\tilde{\mathbf{H}}_T = \nabla^2 \mathcal{L}_T(\mathbf{Y}^T; \tilde{\boldsymbol{\theta}}_T)$. For $\mathbf{B}_T(\boldsymbol{\theta}^*)$, a Newey-West-type estimator is usually needed to accommodate potential correlations and heterogeneity in the data.

Model Diagnostic Tests

Consider a joint test of $\boldsymbol{\rho}_m = (\rho_1, \ldots, \rho_m)' = \mathbf{0}$, where ρ_i are the autocorrelations of the raw series y_t. Let $\hat{\boldsymbol{\rho}}_m = (\hat{\rho}_1, \ldots, \hat{\rho}_m)'$ be the vector of m sample autocorrelations. Under general conditions,

$$\sqrt{T}(\hat{\boldsymbol{\rho}}_m - \boldsymbol{\rho}_m) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

so that

$$T(\hat{\boldsymbol{\rho}}_m - \boldsymbol{\rho}_m)'\mathbf{V}^{-1}(\hat{\boldsymbol{\rho}}_m - \boldsymbol{\rho}_m) \xrightarrow{D} \chi^2(m).$$

As shown in Lobato et al. (2001), the (i,j) th element of $\mathbf{V}$ is

$$v_{ij} = \frac{1}{\gamma_0^2}\left[c_{i+1,j+1} - \rho_i c_{1,j+1} - \rho_j c_{1,i+1} + \rho_i \rho_j c_{1,1}\right].$$

In the previous expression,

$$c_{i+1,j+1} = \sum_{k=-\infty}^{\infty} \mathbb{E}\big[(y_t - \mu)(y_{t+i} - \mu)(y_{t+k} - \mu)(y_{t+k+j} - \mu)\big] - \mathbb{E}\big[(y_t - \mu)(y_{t+i} - \mu)\big]\, \mathbb{E}\big[(y_{t+k} - \mu)(y_{t+k+j} - \mu)\big].$$

- Under the null, $\mathbf{V}$ simplifies and $v_{ij} = c_{i+1,j+1}/\gamma_0^2$ with

$$c_{i+1,j+1} = \sum_{k=-\infty}^{\infty} \mathbb{E}\big[(y_t - \mu)(y_{t+i} - \mu)(y_{t+k} - \mu)(y_{t+k+j} - \mu)\big].$$

- When y_t are in fact serially independent, $c_{i+1,j+1} = \gamma_0^2$ for $i = j$ and zero otherwise. Thus, $\mathbf{V} = \mathbf{I}$, and $\sqrt{T}\hat{\boldsymbol{\rho}}_m \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Q Tests

- Using the result above, Box-Pierce's Q test is

$$\mathcal{Q}_T = T\hat{\boldsymbol{\rho}}_m'\hat{\boldsymbol{\rho}}_m = T\sum_{i=1}^{m}\hat{\rho}_i^2 \xrightarrow{D} \chi^2(m).$$

- A finite-sample correction of $\mathcal{Q}_T$ is Ljung-Box's Q test:

$$\widetilde{\mathcal{Q}}_T = T^2\sum_{i=1}^{m}\frac{\hat{\rho}_i^2}{T-i} \xrightarrow{D} \chi^2(m).$$

Fuller (1976, p. 242) shows that, when y_t are serially independent with mean zero, variance σ^2, and finite 6 th moment,

$$\operatorname{cov}\left(\sqrt{T}\hat{\rho}_i, \sqrt{T}\hat{\rho}_j\right) = \begin{cases} \frac{T-i}{T} + O(T^{-1}), & i = j \neq 0, \\ O(T^{-1}), & i \neq j. \end{cases}$$

- Instead of serial independence, assume

$$\mathbb{E}\big[(y_t-\mu)(y_{t+i}-\mu)(y_{t+k}-\mu)(y_{t+k+j}-\mu)\big]=0,$$

for each k when $i\neq j$ and for $k\neq 0$ when $i=j$, we find $c_{i+1,j+1}=0$ when $i\neq j$, and

$$c_{i+1,j+1}=\mathbb{E}\big[(y_t-\mu)^2(y_{t+i}-\mu)^2\big],\quad i=j.$$

Then, $\mathbf{V}$ is diagonal with the diagonal element $v_{ii}=c_{i+1,i+1}/\gamma_0^2$.

- v_{ii} can be consistently estimated by

$$\hat{v}_{ii}=\frac{\frac{1}{T}\sum_{t=1}^{T-i}(y_t-\bar{y})^2(y_{t+i}-\bar{y})^2}{[\frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y})^2]^2}.$$

The Q^* test due to Lobato et al. (2001) is

$$\mathcal{Q}_T^*=T\sum_{i=1}^{m}\hat{\rho}_i^2/\hat{v}_{ii}\xrightarrow{D}\chi^2(m).$$

Remarks:

1. The Box-Pierce and Ljung-Box Q tests are not really tests of serial uncorrelatedness because they are based on the assumption of serial independence. They are in fact tests of a stronger null hypothesis.
2. Under conditional homoskedasticity, it can be seen that $c_{i+1,j+1} = \mathbb{E}\big[(y_t - \mu)^2 (y_{t+i} - \mu)^2\big]$ would be γ_0^2 for $i = j$. Q^* test does not require this condition but instead relies on the estimates of $c_{i+1,j+1}$. Thus, this test ought to be more robust to conditional heteroskedasticity.
3. When the Q-type tests are applied to the residuals of an ARMA(p,q) model, the asymptotic null distribution becomes $\chi^2(m - p - q)$ or $\chi^2(m - p - q - 1)$ if the model contains a constant term.

Spectral Tests

In contrast with Q tests, we are interested in testing all autocorrelations:

$$H_0 \colon \rho_1 = \rho_2 = \rho_3 = \cdots = 0.$$

- The spectral density is the Fourier transform of the autocorrelations:

$$f(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \rho_j e^{-ij\omega}, \quad \omega \in [-\pi, \pi],$$

  where ω denotes frequency.

- Periodogram is the sample counterpart of $f(\omega)$:

$$I_T(\omega) = \frac{1}{2\pi} \sum_{j=-(T-1)}^{T-1} \hat{\rho}_j e^{-ij\omega}.$$

- Under the null, $f(\omega) = (2\pi)^{-1}$ for all ω. We can then base a test on the difference between $I_T(\omega)$ and $(2\pi)^{-1}$:

$$\frac{1}{2\pi}\Bigg(\sum_{j=-(T-1)}^{T-1} \hat{\rho}_j e^{-ij\omega} - 1\Bigg) = \frac{1}{\pi}\sum_{j=1}^{T-1} \hat{\rho}_j \cos(j\omega),$$

because $\exp(-ij\omega) = \cos(j\omega) - i\,\sin(j\omega)$, sin is an odd function, and cos is an even function.

- Integrating this function with respect to ω on $[0, a]$, $0 \leq a \leq \pi$,

$$\frac{1}{\pi}\sum_{j=1}^{T-1} \hat{\rho}_j \frac{\sin(ja)}{j},$$

which should also be "close" to zero for all a under the null.

- The spectral test of Durlauf (1991) is based on:

$$D_T(t) = \frac{\sqrt{2T}}{\pi} \sum_{j=1}^{m(T)} \hat{\rho}_j \, \frac{\sin(j\pi t)}{j},$$

where $\pi t = a$ and $m(T)$ grows with T but at a slower rate.

- A standard Brownian motion B can be approximated by

$$W_T(t) = \epsilon_0 t + \frac{\sqrt{2}}{\pi} \sum_{j=1}^{T} \epsilon_j \, \frac{\sin(j\pi t)}{j} \Rightarrow B(t), \quad t \in [0,1],$$

where ϵ_t are i.i.d. $\mathcal{N}(0,1)$ and $\Rightarrow$ stands for weak convergence. Then,

$$W_T(t) - tW_T(1) = \frac{\sqrt{2}}{\pi} \sum_{j=1}^{T} \epsilon_j \, \frac{\sin(j\pi t)}{j} \Rightarrow B^0(t), \quad t \in [0,1],$$

where B^0 denotes the Brownian bridge.

- Recall $T^{1/2}\hat{\rho}_i \xrightarrow{D} \mathcal{N}(0,1)$ under serial independence. Thus,

$$D_T(t) \Rightarrow B^0(t), \qquad t \in [0,1].$$

The spectral tests are based on various functionals of D_T.

1. Anderson-Darling test:

$$\mathrm{AD}_T = \int_0^1 \frac{[D_T(t)]^2}{t(1-t)}\,\mathrm{d}t \Rightarrow \int_0^1 \frac{[B^0(t)]^2}{t(1-t)}\,\mathrm{d}t.$$

2. Cramér-von Mises test:

$$\mathrm{CvM}_T = \int_0^1 [D_T(t)]^2\,\mathrm{d}t \Rightarrow \int_0^1 [B^0(t)]^2\,\mathrm{d}t.$$

3. Kolmogorov-Smirnov test:

$$\mathrm{KS}_T = \sup|D_T(t)| \Rightarrow \sup|B^0(t)|.$$

4. Kuiper test:

$$\mathrm{Ku}_T = \sup_{s,t}|D_T(t) - D_T(s)| \Rightarrow \sup|B^0(t) - B^0(s)|.$$

- Deo (2000) notes that when y_t are conditionally heteroskedastic, the asymptotic variance of $T^{1/2}\hat{\rho}_j$ is $\mathbb{E}(y_t^2 y_{t-j}^2)/\gamma(0)^2$. Hence, D_T is not properly normalized and may converge to a different limit.
- Similar to Q^* test, Deo (2000) proposes a modification of D_T:

$$D_T^c(t) = \frac{\sqrt{2T}}{\pi} \sum_{j=1}^{m(T)} \frac{\hat{\rho}_j}{\sqrt{\hat{v}_{jj}}} \frac{\sin(j\pi t)}{j},$$

$$\sqrt{\hat{v}_{jj}} = \frac{1}{\hat{\gamma}(0)} \left(\frac{1}{T-j} \sum_{t=1}^{T-j} (y_t - \bar{y})^2 (y_{t+j} - \bar{y})^2 \right)^{1/2}.$$

- The modified Cramér-von Mises test is

$$\mathrm{CvM}_T^c = \int_0^1 [D_T^c(t)]^2 \,\mathrm{d}t \Rightarrow \int_0^1 [B^0(t)]^2 \,\mathrm{d}t.$$

Variance-Ratio Test

The variance-ratio test of Cochrane (1988) is designed to check if the series η_t is a random walk, or equivalently, if $y_t = \eta_t - \eta_{t-1}$ are i.i.d.

- Suppose y_t $(t = 0, 1, \ldots, kT)$ have mean zero and variance σ^2 and $\sigma_k^2 = \mathrm{var}(y_t + \cdots + y_{t-k+1})$. Under the null, $\sigma_k^2 = k\sigma^2$.
- Let $\bar{y} = \frac{1}{kT}\sum_{t=1}^{kT}(\eta_t - \eta_{t-1})$. The sample variance of y_t is

$$\hat{\sigma}^2 = \frac{1}{kT}\sum_{t=1}^{kT}\big(\eta_t - \eta_{t-1} - \bar{y}\big)^2,$$

which is consistent and asymptotically efficient for σ^2 under the null.

- Given $\sigma_k^2 = \text{var}(\eta_t - \eta_{t-k})$, we can treat each block of k random variables (i.e., $\eta_{kt} - \eta_{kt-k}$) as a whole and estimate σ_k^2 by

$$\tilde{\sigma}_k^2 = \frac{1}{T}\sum_{t=1}^{T}\big(\eta_{kt} - \eta_{kt-k} - k\bar{y}\big)^2 = \frac{1}{T}\sum_{t=1}^{T}\big[k(\bar{y}_t - \bar{y})\big]^2,$$

  where $\bar{y}_t = \sum_{kt-k+1}^{kt} y_i/k$. Clearly, $\tilde{\sigma}_k^2/k$ is consistent for σ^2 under the null, but it is not asymptotically efficient. (Why?)
- The variance-ratio test check if the ratio of $\tilde{\sigma}_k^2/k$ to $\hat{\sigma}^2$ is sufficiently close to one.

- We now write

$$\begin{aligned}\frac{1}{\sqrt{k}}\sqrt{T}(\tilde{\sigma}_k^2 - k\sigma^2) &= \sqrt{kT}\left(\frac{\tilde{\sigma}_k^2}{k} - \sigma^2\right)\\ &= \sqrt{kT}\left(\frac{\tilde{\sigma}_k^2}{k} - \hat{\sigma}^2\right) + \sqrt{kT}(\hat{\sigma}^2 - \sigma^2).\end{aligned}$$

  - By Hausman (1978), the two terms on the right-hand side are asymptotically uncorrelated.
  - Under the null, $\sqrt{kT}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{D} \mathcal{N}(0,\, 2\sigma^4)$. (Check!)
  - Similarly, as $\sigma_k^2 = k\sigma^2$ under the null,

$$\sqrt{T}(\tilde{\sigma}_k^2 - k\sigma^2) \xrightarrow{D} \mathcal{N}(0,\, 2k^2\sigma^4),$$

    so that the left-hand side converges to $\mathcal{N}(0,\, 2k\sigma^4)$.

- It follows that

$$\sqrt{kT}\left(\frac{\tilde{\sigma}_k^2}{k} - \hat{\sigma}^2\right) \xrightarrow{D} \mathcal{N}(0,\, 2(k-1)\sigma^4),$$

or alternatively,

$$\sqrt{kT}\left(\frac{\tilde{\sigma}_k^2}{k\hat{\sigma}^2} - 1\right) \xrightarrow{D} \mathcal{N}(0,\, 2(k-1)).$$

- Setting $\mathrm{VR}(k) := \tilde{\sigma}_k^2/(k\hat{\sigma}^2)$, we have

$$\sqrt{kT}[\mathrm{VR}(k) - 1]/\sqrt{2(k-1)} \xrightarrow{D} \mathcal{N}(0,\, 1).$$

Clearly, this test depends on the choice of k.

Model Selection Criteria

- Model selection criteria are usually the Gaussian log-likelihood values penalized by model complexity (in terms of number of parameters).
  - Akaike Information Criterion (AIC):

$$\mathrm{AIC} = \ln \tilde{\sigma}_T^2 + \frac{2(p+q+1)}{T}.$$

  - Schwartz Information Criterion (SIC):

$$\mathrm{SIC} = \ln \tilde{\sigma}_T^2 + \frac{(p+q+1)\ln T}{T}.$$

    The SIC is dimensionally consistent, in the sense that it can select the correct ARMA orders when the sample is sufficiently large.

- In practice, we usually estimate an array of ARMA models and choose the one with the smallest AIC or SIC as the "best" model.

Vector AR (VAR) Series

Let ε_t $(d \times 1)$ be a vector time series with mean zero, the covariance matrix $\mathbf{\Sigma}_\varepsilon$, and $\operatorname{cov}(\varepsilon_t, \varepsilon_s) = \mathbf{0}$ for $t \neq s$. A VAR series $\{\mathbf{y}_t\}$ is:

$$\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \mathbf{c} + \varepsilon_t,$$

where $\mathbf{\Psi}(\mathcal{B}) = \mathbf{I}_d - \mathbf{\Psi}_1\mathcal{B} - \mathbf{\Psi}_2\mathcal{B}^2 - \ldots$ is a matrix polynomial in $\mathcal{B}$.

For a VAR(1) series, $\mathbf{\Psi}(\mathcal{B}) = \mathbf{I}_d - \mathbf{\Psi}_1\mathcal{B}$. It is weakly stationary if all the characteristic roots of $\mathbf{\Psi}_1$ are inside the unit circle. Let $(\mathbf{I}_d - \mathbf{\Psi}_1\mathcal{B})^{-1} = \mathbf{I}_d + \mathbf{\Psi}_1\mathcal{B} + \mathbf{\Psi}_1^2\mathcal{B}^2 + \cdots$. The MA($\infty$) representation is

$$\mathbf{y}_t = (\mathbf{I} - \mathbf{\Psi}_1)^{-1}\mathbf{c} + \sum_{j=0}^{\infty} \mathbf{\Psi}_1^j \varepsilon_{t-j}.$$

VAR(p) Series

The VAR(p) series $\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \mathbf{c} + \boldsymbol{\varepsilon}_t$ can be expressed as:

$$\underbrace{\begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{bmatrix}}_{\mathbf{Y}_t} = \underbrace{\begin{bmatrix} \mathbf{c} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathbf{C}} + \underbrace{\begin{bmatrix} \mathbf{\Psi}_1 & \mathbf{\Psi}_2 & \cdots & \mathbf{\Psi}_{p-1} & \mathbf{\Psi}_p \\ \mathbf{I}_d & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_d & \mathbf{0} \end{bmatrix}}_{\mathbf{F}} \underbrace{\begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \mathbf{y}_{t-3} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix}}_{\mathbf{Y}_{t-1}} + \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathbf{E}_t}.$$

That is, $\mathbf{Y}_t = \mathbf{C} + \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{E}_t$, and it is weakly stationary if all the characteristic roots of $\mathbf{F}$ ($pd \times pd$) are inside the unit circle.

Given VAR(1) series $\mathbf{y}_t = \mathbf{c} + \mathbf{\Psi}_1 \mathbf{y}_{t-1} + \varepsilon_t$, $\mathbb{E}(\mathbf{y}_t) = (\mathbf{I}_d - \mathbf{\Psi}_1)^{-1}\mathbf{c}$, and the autocovariances are

$$\mathbf{\Gamma}_j = \operatorname{cov}(\mathbf{y}_t, \mathbf{y}_{t-j}) = \sum_{i=0}^{\infty} \mathbf{\Psi}_1^{i+j} \mathbf{\Sigma}_\varepsilon \mathbf{\Psi}_1^{i\prime}, \quad j = 0, 1, 2, \ldots,$$

with $\mathbf{\Gamma}_0 = \operatorname{var}(\mathbf{y}_t) = \sum_{i=0}^{\infty} \mathbf{\Psi}_1^{i} \mathbf{\Sigma}_\varepsilon \mathbf{\Psi}_1^{i\prime}$. Note that $\mathbf{\Gamma}_j = \mathbf{\Gamma}_{-j}'$. For $\mathbf{\Gamma}_0$, its k th diagonal element is $\gamma_{kk,0}$, the variance of $y_{k,t}$, and its (h,k) th element is $\gamma_{hk,0}$, the contemporaneous covariance of $y_{h,t}$ and $y_{k,t}$. (Explain the elements of $\mathbf{\Gamma}_j$.)

The multivariate Yule-Walker equations:

$$\mathbf{\Gamma}_j = \mathbf{\Psi}_1 \mathbf{\Gamma}_{j-1}, \quad j = 1, 2, \ldots,$$

and $\mathbf{\Gamma}_0 = \mathbf{\Psi}_1 \mathbf{\Gamma}_1' + \mathbf{\Sigma}_\varepsilon$.

Let $\mathbf{D}$ denote the diagonal matrix with the k th diagonal element $\gamma_{kk,0}$. The autocorrelations of $\mathbf{y}_t$ are

$$\mathbf{R}_j = \mathbf{D}^{-1/2}\mathbf{\Gamma}_j\mathbf{D}^{-1/2}, \quad j = 0, 1, 2, \ldots$$

For the VAR(p) series $\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \mathbf{c} + \boldsymbol{\varepsilon}_t$, $\mathbb{E}(\mathbf{y}_t) = \mathbf{\Psi}(1)^{-1}\mathbf{c}$, and the multivariate Yule-Walker equations of autocovariances are

$$\mathbf{\Gamma}_j = \mathbf{\Psi}_1\mathbf{\Gamma}_{j-1} + \mathbf{\Psi}_2\mathbf{\Gamma}_{j-2} + \cdots + \mathbf{\Psi}_p\mathbf{\Gamma}_{j-p}, \qquad j = 1, 2, \ldots,$$

and $\mathbf{\Gamma}_0 = \mathbf{\Psi}_1\mathbf{\Gamma}_1' + \mathbf{\Psi}_2\mathbf{\Gamma}_2' + \cdots + \mathbf{\Psi}_p\mathbf{\Gamma}_p' + \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$.

Model Estimation

Consider a VAR(p) model: $\mathbf{y}_t = \mathbf{c} + \boldsymbol{\Psi}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Psi}_p \mathbf{y}_{t-p} + \varepsilon_t$. Given p initial values: $\mathbf{y}_1, \ldots, \mathbf{y}_p$ and the conditional normality assumption,

$$\begin{aligned}\mathcal{L}_T(\boldsymbol{\theta}) &= \frac{1}{T} \sum_{j=p+1}^{T} \ln f(\mathbf{y}_j \mid \mathbf{Y}^{j-1}; \boldsymbol{\theta}) \\ &= -\frac{(T-p)d}{2T} \ln(2\pi) + \frac{(T-p)}{2T} \ln(\det(\boldsymbol{\Sigma}^{-1})) \\ &\qquad - \frac{1}{2T} \sum_{j=p+1}^{T} (\mathbf{y}_j - \mathbf{P}' \eta_j)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_j - \mathbf{P}' \eta_j),\end{aligned}$$

where $\eta_j = (1 \ \mathbf{y}'_{j-1} \ \ldots \ \mathbf{y}'_{j-p})'$ is $(pd+1) \times 1$ and $\mathbf{P} = (\mathbf{c} \ \boldsymbol{\Psi}_1 \ \boldsymbol{\Psi}_2 \ \ldots \ \boldsymbol{\Psi}_p)'$ is $(pd+1) \times d$.

QMLE of $\mathbf{P}$

The QMLE of $\mathbf{P}$ is:

$$\widehat{\mathbf{P}}_T = \left(\sum_{j=p+1}^{T} \boldsymbol{\eta}_j \boldsymbol{\eta}_j' \right)^{-1} \left(\sum_{j=p+1}^{T} \boldsymbol{\eta}_j \mathbf{y}_j' \right),$$

and the k th column of $\widehat{\mathbf{P}}$ is the OLS estimates of regressing of $y_{k,t}$ on $\boldsymbol{\eta}_t$:

$$\left(\sum_{j=p+1}^{T} \boldsymbol{\eta}_j \boldsymbol{\eta}_j' \right)^{-1} \left(\sum_{j=p+1}^{T} \boldsymbol{\eta}_j y_{k,j} \right).$$

That is, the coefficients of a VAR(p) model can be estimated by separately estimating each autoregression via OLS.

Let Let $\hat{\mathbf{e}}_t = \mathbf{y}_t - \widehat{\mathbf{P}}_T' \eta_t$ denote the residuals. We can write

$$\begin{aligned}
&\sum_{j=p+1}^{T} (\mathbf{y}_j - \mathbf{P}'\eta_j)' \mathbf{\Sigma}^{-1} (\mathbf{y}_j - \mathbf{P}'\eta_j) \\
&\quad = \sum_{j=p+1}^{T} \big[\hat{\mathbf{e}}_j + \big(\widehat{\mathbf{P}}_T - \mathbf{P}\big)'\eta_j\big]' \mathbf{\Sigma}^{-1} \big[\hat{\mathbf{e}}_j + \big(\widehat{\mathbf{P}}_T - \mathbf{P}\big)'\eta_j\big] \\
&\quad = \sum_{j=p+1}^{T} \hat{\mathbf{e}}_j' \mathbf{\Sigma}^{-1} \hat{\mathbf{e}}_j + \sum_{j=p+1}^{T} \eta_j' \big(\widehat{\mathbf{P}}_T - \mathbf{P}\big) \mathbf{\Sigma}^{-1} \big(\widehat{\mathbf{P}}_T - \mathbf{P}\big)' \eta_j.
\end{aligned}$$

because $\sum_{j=p+1}^{T} \eta_j' \big(\widehat{\mathbf{P}}_T - \mathbf{P}\big) \mathbf{\Sigma}^{-1} \hat{\mathbf{e}}_j = 0$. (Why?) The RHS can be minimized when the second term is zero (i.e., $\mathbf{P} = \widehat{\mathbf{P}}_T$).

QMLE of $\boldsymbol{\Sigma}$

To estimate $\boldsymbol{\Sigma}$, we maximize $\mathcal{L}_T(\widehat{\mathbf{P}}_T, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}^{-1}$ and obtain

$$\frac{\partial \mathcal{L}_T(\widehat{\mathbf{P}}_T, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{(T-p)}{2T}\boldsymbol{\Sigma}' - \frac{1}{2T}\sum_{j=p+1}^{T} \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j'.$$

This yields the QMLE of $\boldsymbol{\Sigma}$:

$$\widehat{\boldsymbol{\Sigma}}_T = \frac{1}{T-p}\sum_{j=p+1}^{T} \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j',$$

with the i th diagonal element and (i,j) th off-diagonal element:

$$\hat{\sigma}_{ii} = \frac{1}{T-p}\sum_{t=p+1}^{T} \hat{e}_{i,t}^2, \qquad \hat{\sigma}_{ij} = \frac{1}{T-p}\sum_{t=p+1}^{T} \hat{e}_{i,t}\hat{e}_{j,t}.$$

Asymptotic Properties

- Consistency: Under suitable conditions, $\widehat{\mathbf{P}}_T \xrightarrow{\mathbb{P}} \mathbf{P}$ and $\widehat{\mathbf{\Sigma}}_T \xrightarrow{\mathbb{P}} \mathbf{\Sigma}$.
- Asymptotic Normality: Let $\mathbf{p} = \operatorname{vec}(\mathbf{P})$ and $\mathbf{Q} = \mathbb{E}(\boldsymbol{\eta}_t \boldsymbol{\eta}_t')$. Then,

  $$\sqrt{T}(\hat{\mathbf{p}}_T - \mathbf{p}) \xrightarrow{D} \mathcal{N}(\mathbf{0},\ \mathbf{\Sigma} \otimes \mathbf{Q}^{-1}),$$

  when the information matrix equality holds. For the i th regression,

  $$\sqrt{T}(\hat{\mathbf{p}}_{i,T} - \mathbf{p}_i) \xrightarrow{D} \mathcal{N}(\mathbf{0},\ \sigma_{ii}^2 \mathbf{Q}^{-1}), \quad i = 1, \ldots, d.$$

- Wald Test: Under the null hypothesis $\mathbf{R}\mathbf{p} = \mathbf{r}$,

  $$\mathcal{W}_T = T(\mathbf{R}\hat{\mathbf{p}}_T - \mathbf{r})'[\mathbf{R}(\widehat{\mathbf{\Sigma}}_T \otimes \mathbf{Q}^{-1})\mathbf{R}]^{-1}(\mathbf{R}\hat{\mathbf{p}}_T - \mathbf{r}) \xrightarrow{D} \chi^2(q).$$

Impulse Response Functions

Consider the VAR(p) series $\boldsymbol{\Psi}(\mathcal{B})\mathbf{y}_t = \mathbf{c} + \boldsymbol{\varepsilon}_t$. Its MA representation is

$$\mathbf{y}_t = \boldsymbol{\Phi}(1)\mathbf{c} + \boldsymbol{\Phi}(\mathcal{B})\boldsymbol{\varepsilon}_t = \boldsymbol{\Phi}(1)\mathbf{c} + \sum_{j=0}^{\infty} \boldsymbol{\Phi}_j \boldsymbol{\varepsilon}_{t-j},$$

where the polynomial $\boldsymbol{\Phi}(\mathcal{B}) = \boldsymbol{\Psi}(\mathcal{B})^{-1}$ with $\boldsymbol{\Phi}_0 = \mathbf{I}_d$.

- The impulse response of $\mathbf{y}_t$ to one unit shock of $\varepsilon_{i,t-j}$ is the i th column of $\boldsymbol{\Phi}_j$, i.e., $\boldsymbol{\Phi}_j\mathbf{e}_i$, with $\mathbf{e}_i$ the i th Cartesian unit vector.
- The accumulated response over n periods is

$$\mathbf{A}_n\mathbf{e}_i = \Big(\sum_{j=0}^{n} \boldsymbol{\Phi}_j\Big)\mathbf{e}_i;$$

the long-run effect is $\mathbf{A}_\infty\mathbf{e}_i$, where $\mathbf{A}_\infty = \boldsymbol{\Phi}(1) = \boldsymbol{\Psi}(1)^{-1}$.

Orthogonalized Impulse Response

- $\boldsymbol{\Sigma}_{\varepsilon}$ is not a diagonal matrix in general. When $\varepsilon_{i,t-j}$ and $\varepsilon_{k,t-j}$ are correlated for some i, k, a shock of $\varepsilon_{i,t-j}$ may come with shocks of other innovations. As such, the impulse response of $\mathbf{y}_t$ to a shock of $\varepsilon_{i,t-j}$ may involve the response to shocks of other innovations.
- Cholesky decomposition: $\boldsymbol{\Sigma}_{\varepsilon} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a lower triangular matrix with non-zero diagonal elements. The elements of $\mathbf{v}_t = \mathbf{L}^{-1}\varepsilon_t$ are uncorrelated and $\mathrm{var}(\mathbf{v}_t) = \mathbf{L}^{-1}\boldsymbol{\Sigma}_{\varepsilon}\mathbf{L}^{-1\prime} = \mathbf{I}_d$. These are known as orthogonalized innovations.
- Alternatively, write $\boldsymbol{\Sigma}_{\varepsilon} = \mathbf{L}\mathbf{D}\mathbf{L}'$, where $\mathbf{D}$ is diagonal and $\mathbf{L}$ is lower triangular with the diagonal elements being 1s. Then, $\mathbf{v}_t = \mathbf{L}^{-1}\varepsilon_t$ are also orthogonalized innovations because $\mathrm{var}(\mathbf{v}_t) = \mathbf{D}$.

- The MA representation of the VAR(p) series $\mathbf{y}_t$ in terms of $\mathbf{v}_t$ is

$$\mathbf{y}_t = \mathbf{\Phi}(1)\mathbf{c} + \sum_{j=0}^{\infty} \mathbf{\Phi}_j \mathbf{L}\, \mathbf{v}_{t-j} = \mathbf{\Phi}(1)\mathbf{c} + \sum_{j=0}^{\infty} \mathbf{\Theta}_j\, \mathbf{v}_{t-j},$$

  where $\mathbf{\Theta}_j = \mathbf{\Phi}_j \mathbf{L}$. The i th column of $\mathbf{\Theta}_j$, $\mathbf{\Theta}_j \mathbf{e}_i$, is the orthogonalized impulse response of $\mathbf{y}_t$ to one unit shock of $v_{i,t-j}$. This impulse response is not contaminated by the effect of other innovations.

- Advantage: There is no "scaling" problem, because $\text{var}(\mathbf{v}_t) = \mathbf{I}_d$ so that one unit shock is also a shock of one standard deviation.

- Drawback: The orthogonalized impulse responses are not uniquely defined because the decomposition of $\mathbf{\Sigma}_\varepsilon$ depends on the ordering of the elements of $\mathbf{y}_t$.

Generalized Impulse Response

Pesaran and Shin (1998) define the generalized impulse response of $\mathbf{y}_t$ to the shock $\varepsilon_{i,t-j} = \delta$ as

$$\mathbb{E}\big(\mathbf{y}_t \mid \varepsilon_{i,t-j} = \delta, \mathcal{F}^{t-j-1}\big) - \mathbb{E}\big(\mathbf{y}_t \mid \mathcal{F}^{t-j-1}\big),$$

where $\mathcal{F}^t$ is the info. set up to time t. From the MA representation,

$$\begin{aligned}\mathbb{E}\big(\mathbf{y}_t|\varepsilon_{i,t-j} = \delta, \mathcal{F}^{t-j-1}\big) = \mathbf{\Phi}(1)\mathbf{c} + \sum_{k=j+1}^{\infty} \mathbf{\Phi}_k \boldsymbol{\varepsilon}_{t-k} \\ + \mathbf{\Phi}_j\, \mathbb{E}(\boldsymbol{\varepsilon}_{t-j}|\varepsilon_{i,t-j} = \delta).\end{aligned}$$

This differs from $\mathbb{E}\big(\mathbf{y}_t|\mathcal{F}^{t-j-1}\big)$ by the last term, $\mathbf{\Phi}_j\, \mathbb{E}(\boldsymbol{\varepsilon}_{t-j}|\varepsilon_{i,t-j} = \delta)$, which is the generalized impulse response to the shock $\varepsilon_{i,t-j} = \delta$.

- When $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution,

$$\mathbb{E}(\varepsilon_{k,t}|\varepsilon_{i,t}=\delta)=\frac{\sigma_{ki}}{\sigma_{ii}}\,\delta,$$

  where σ_{ki} is the (k,i) th element of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

- Using this result we have the generalized impulse response:

$$\boldsymbol{\Phi}_j\,\mathbb{E}(\boldsymbol{\varepsilon}_{t-j}|\varepsilon_{i,t-j}=\delta)=\boldsymbol{\Phi}_j\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{e}_i\delta/\sigma_{ii}.$$

  Setting $\delta=\sigma_{ii}^{1/2}$, a shock of one standard deviation to the i th equation, the generalized impulse response of $\mathbf{y}_t$ is $\boldsymbol{\Phi}_j\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\mathbf{e}_i/\sigma_{ii}^{1/2}$.

- This impulse response does not depend on the ordering of the elements of $\mathbf{y}_t$ but requires the normality assumption on $\boldsymbol{\varepsilon}_t$.

Forecast Error Variance Decomposition

- As $\mathbf{y}_t = \mathbf{\Phi}(1)\mathbf{c} + \sum_{j=0}^{\infty} \mathbf{\Theta}_j \, \mathbf{v}_{t-j}$, the optimal h-step ahead forecast is

$$\hat{\mathbf{y}}_t(h) := \mathbb{E}\left(\mathbf{y}_{t+h}|\mathcal{F}^t\right) = \mathbf{\Phi}(1)\mathbf{c} + \sum_{j=h}^{\infty} \mathbf{\Theta}_j \, \mathbf{v}_{t+h-j},$$

  and the forecast error is $\sum_{j=0}^{h-1} \mathbf{\Theta}_j \, \mathbf{v}_{t+h-j}$.

- The h-step forecast error variance of $y_{i,t+h}$ is defined as

$$\mathbb{E}\left[y_{i,t+h} - \hat{y}_{i,t}(h)\right]^2 = \sum_{j=0}^{h-1} \sum_{k=1}^{d} \theta_{ik,j}^2 = \sum_{k=1}^{d} \sum_{j=0}^{h-1} (\mathbf{e}_i' \mathbf{\Theta}_j \mathbf{e}_k)^2,$$

  where $\sum_{j=0}^{h-1} (\mathbf{e}_i' \mathbf{\Theta}_j \mathbf{e}_k)^2$ is the forecast error variance due to the k th innovation.

- Noting that $\sum_{k=1}^{d} \mathbf{e}_k \mathbf{e}_k' = \mathbf{I}$, the forecast error variance of $y_{i,t+h}$ is

$$\sum_{j=0}^{h-1} \sum_{k=1}^{d} \mathbf{e}_i' \boldsymbol{\Theta}_j \mathbf{e}_k \mathbf{e}_k' \boldsymbol{\Theta}_j \mathbf{e}_i = \sum_{j=0}^{h-1} \mathbf{e}_i' \boldsymbol{\Theta}_j \boldsymbol{\Theta}_j' \mathbf{e}_i.$$

In terms of the original coefficient matrices, we have

$$\mathbb{E}\left[y_{i,t+h} - \hat{y}_{i,t}(h)\right]^2 = \sum_{j=0}^{h-1} \mathbf{e}_i' \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Phi}_j' \mathbf{e}_i.$$

- The orthogonalized forecast error variance decomposition is the proportion of the total forecast error variance of $y_{i,t+h}$ that can be attributed to the kth innovation:

$$\frac{\sum_{j=0}^{h-1} (\mathbf{e}_i' \boldsymbol{\Theta}_j \mathbf{e}_k)^2}{\sum_{j=0}^{h-1} \sum_{k=1}^{d} (\mathbf{e}_i' \boldsymbol{\Theta}_j \mathbf{e}_k)^2} = \frac{\sum_{j=0}^{h-1} (\mathbf{e}_i' \boldsymbol{\Theta}_j \mathbf{e}_k)^2}{\sum_{j=0}^{h-1} \mathbf{e}_i' \boldsymbol{\Phi}_j \boldsymbol{\Sigma}_\varepsilon \boldsymbol{\Phi}_j' \mathbf{e}_i}.$$

- The ratios of the decomposition above sum to one (over k). Hence, each ratio signifies the relative importance of a particular (orthogonalized) innovation.
- As the generalized impulse response of $\mathbf{y}_{i,t}$ to the shock of k th innovation is $\mathbf{e}_i \mathbf{\Phi}_j \mathbf{\Sigma}_\varepsilon \mathbf{e}_k / \sigma_{kk}^{1/2}$, we can define the generalized forecast error variance decomposition as

$$\frac{\sum_{j=0}^{h-1} (\mathbf{e}_i' \mathbf{\Phi}_j \mathbf{\Sigma}_\varepsilon \mathbf{e}_k)^2 / \sigma_{kk}}{\sum_{j=0}^{h-1} \mathbf{e}_i' \mathbf{\Phi}_j \mathbf{\Sigma}_\varepsilon \mathbf{\Phi}_j' \mathbf{e}_i}.$$

  Note that these ratios do not sum to one. That is, each ratio does not represent the relative importance of an innovation.

Structural VAR Models

Consider a money demand function:

$$(m_t - p_t) = \beta_0 + \beta_1 y_t + \beta_2 r_t + \beta_3 (m_{t-1} - p_{t-1}) + v_t,$$

where m_t is the log of nominal money balance held by the public, p_t is the log of price level, y_t is the log of GNP, and r_t is nominal interest rate.

- With the assumption $v_t = \rho v_{t-1} + u_t$,

$$\begin{aligned}(m_t - p_t) = {} & (1-\rho)\beta_0 + \beta_1 y_t - \beta_1 \rho y_{t-1} + \beta_2 r_t - \beta_2 \rho r_{t-1} \\ & + (\beta_3 + \rho)(m_{t-1} - p_{t-1}) - \beta_3 \rho (m_{t-2} - p_{t-2}) + u_t.\end{aligned}$$

- This can be obtained from the model of $(m_t - p_t)$ on y_t, y_{t-1}, r_t, r_{t-1}, $(m_{t-1} - p_{t-1})$, and $(m_{t-2} - p_{t-2})$ with parameter restrictions.

- The aforementioned general model can not be consistently estimated by OLS because there is simultaneity bias when y_t and r_t are present.
- Consider now a general structural VAR model for $\mathbf{y}_t = (m_t \; p_t \; y_t \; r_t)'$:

$$\mathbf{B}_0 \mathbf{y}_t = \mathbf{k} + \mathbf{B}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \mathbf{u}_t,$$

  where the diagonal elements of $\mathbf{B}_0$ are all 1s and $\text{var}(\mathbf{u}_t) = \mathbf{D}$.
- When $\mathbf{B}_0$ is invertible, we have the following VAR(p) model:

$$\mathbf{y}_t = \mathbf{c} + \boldsymbol{\Psi}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Psi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

  with $\mathbf{c} = \mathbf{B}_0^{-1}\mathbf{k}$, $\boldsymbol{\Psi}_i = \mathbf{B}_0^{-1}\mathbf{B}_i$, and $\boldsymbol{\varepsilon}_t = \mathbf{B}_0^{-1}\mathbf{u}_t$. This is a reduced form, in the sense that $\mathbf{y}_t$ depends only on predetermined variables.

- The reduced form does not depend on current variables and hence can be estimated by OLS.
- Q: Can the structural parameters $\mathbf{k}$, $\mathbf{B}_i$ and $\mathbf{D}$ be identified from the parameter estimates of the VAR(p) model:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Psi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Psi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t?$$

- Q: Can we identify the impulse response to the structural innovations $\mathbf{u}_t$? By construction, each element of $\boldsymbol{\varepsilon}_t$ is a linear combination of $\mathbf{u}_t$. Hence, evaluating $\partial \mathbf{y}_{t+s} / \partial \varepsilon_{i,t}$ or even orthogonzalized impulse response may not be meaningful.

Recursive Model

Consider the structural model:

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{G}\boldsymbol{\eta}_t + \mathbf{u}_t,$$

where $\mathbf{G} = (\mathbf{k}\ \mathbf{B}_1\ \ \ldots\ \ \mathbf{B}_p)$, $\mathbf{B}_0$ is lower triangular with 1s on the principal diagonal, and $\mathbf{D} = \mathrm{var}(\mathbf{u}_t)$ is diagonal with positive entries. This structural model is recursive in the sense that the the elements of $\mathbf{y}_t$ enter the model recursively.

The reduced form is: $\mathbf{y}_t = \mathbf{P}'\boldsymbol{\eta}_t + \varepsilon_t$, where $\mathbf{P}' = \mathbf{B}_0^{-1}\mathbf{G}$ contains elements: $\mathbf{c} = \mathbf{B}_0^{-1}\mathbf{k}$ and $\mathbf{\Psi}_i = \mathbf{B}_0^{-1}\mathbf{B}_i$, and $\varepsilon_t = \mathbf{B}_0^{-1}\mathbf{u}_t$ with

$$\mathrm{var}(\varepsilon_t) = \mathbf{\Sigma}_\varepsilon = \mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime}.$$

The system is just identified.

- $\mathbf{B}_0$ is nonsingular such that $\mathbf{B}_0^{-1}$ is also lower triangular with 1s on the principal diagonal.
- Given that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ can be decomposed as $\mathbf{L}\widetilde{\mathbf{D}}\mathbf{L}'$, the structural parameters that satisfy $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime}$ do exist.
- The structural parameters $\mathbf{B}_i$ can be uniquely obtained as $\mathbf{B}_i = \mathbf{B}_0\boldsymbol{\Psi}_i$.

Full Information Maximum Likelihood (FIML)

FIML estimation when the system is just identified:

1. Estimate $\mathbf{P}$ in the reduced form via OLS and estimate $\mathbf{\Sigma}_\varepsilon$ using the OLS residuals. Let the resulting estimates be $\widehat{\mathbf{P}}_T$ and $\widehat{\mathbf{\Sigma}}_T$.
2. Triangularizing $\widehat{\mathbf{\Sigma}}_T$ to get $\widehat{\mathbf{L}}\widehat{\mathbf{D}}\widehat{\mathbf{L}}'$.
3. Computing the estimated structural parameters as $\widehat{\mathbf{k}} = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{c}}$ and $\widehat{\mathbf{B}}_i = \widehat{\mathbf{L}}^{-1}\widehat{\mathbf{\Psi}}_i$.
4. The estimated orthogonalized impulse response coefficients computed from the estimated VAR model describes the dynamic response to structural innovations $\mathbf{u}_t = \mathbf{L}^{-1}\varepsilon$.

Non-Recursive Model

Without restriction, the log-likelihood function of the reduced from is

$$\mathcal{L}_T(\mathbf{B}_0, \mathbf{D}, \mathbf{P}) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\big(\det(\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime})\big) - \frac{1}{2T}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{P}'\boldsymbol{\eta}_t)'\big[\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime})\big]^{-1}(\mathbf{y}_t - \mathbf{P}'\boldsymbol{\eta}_t).$$

Let $\hat{\varepsilon}_t = \mathbf{y}_t - \widehat{\mathbf{P}}'_T\boldsymbol{\eta}_t$ be the OLS residuals. Then,

$$\mathcal{L}_T(\mathbf{B}_0, \mathbf{D}, \widehat{\mathbf{P}}_T) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\big(\det(\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime})\big) - \frac{1}{2T}\sum_{t=1}^{T}\hat{\varepsilon}'_t\big[\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime}\big]^{-1}\hat{\varepsilon}_t.$$

It can be verified that

$$\sum_{t=1}^{T} \hat{\varepsilon}_t' \big[\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime})\big]^{-1} \hat{\varepsilon}_t = \operatorname{trace}\left(\sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t' \big[\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime})\big]^{-1} \right)$$

$$= \operatorname{trace}\left(T\widehat{\mathbf{\Sigma}}_T \mathbf{B}_0' \mathbf{D}^{-1} \mathbf{B}_0 \right),$$

and that $\ln\big(\det(\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime})\big) = -\log(\det(\mathbf{B}_0)^2) + \ln(\det(\mathbf{D}))$. Hence,

$$\mathcal{L}_T(\mathbf{B}_0, \mathbf{D}, \widehat{\mathbf{P}}_T) = -\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln\big(\det(\mathbf{B}_0)^2\big) - \frac{1}{2}\ln\big(\det(\mathbf{D})\big)$$

$$- \frac{1}{2}\operatorname{trace}\left(\widehat{\mathbf{\Sigma}}_T \mathbf{B}_0' \mathbf{D}^{-1} \mathbf{B}_0 \right).$$

If there exist unique $\mathbf{B}_0$ and $\mathbf{D}$ satisfying $\mathbf{B}_0^{-1}\mathbf{D}\mathbf{B}_0^{-1\prime} = \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ (i.e., $\mathbf{B}_0$ and $\mathbf{D}$ are identified), maximizing $\mathcal{L}_T(\mathbf{B}_0, \mathbf{D}, \widehat{\mathbf{P}}_T)$ yilds the QMLEs for $\mathbf{B}_0$ and $\mathbf{D}$.

Identification

How can the parameters in $\mathbf{B}_0$ and $\mathbf{D}$ be identified?

- Order condition: The number of parameters in $\mathbf{B}_0$ and $\mathbf{D}$ is no more than that of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. As $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ has $d(d+1)/2$ parameter and $\mathbf{D}$ is diagonal with d parameters, the order condition requires $\mathbf{B}_0$ to have at most $d(d-1)/2$ free parameters.
- Rank condition: (to be completed)

Brownian motion

The process $\{w(t),\ t \in [0, \infty)\}$ is the standard Wiener process (standard Brownian motion) if it has continuous sample paths almost surely and satisfies:

1. $\mathbb{P}\big(w(0) = 0\big) = 1.$
2. For $0 \leq t_0 \leq t_1 \leq \cdots \leq t_k$,

   $$\mathbb{P}\big(w(t_i) - w(t_{i-1}) \in B_i,\ i \leq k\big) = \prod_{i \leq k} \mathbb{P}\big(w(t_i) - w(t_{i-1}) \in B_i\big),$$

   where B_i are Borel sets.
3. For $0 \leq s < t$, $w(t) - w(s) \sim \mathcal{N}(0, t - s)$.

Note: w has independent and Gaussian increments.

- $w(t) \sim \mathcal{N}(0, t)$ such that for $r \leq t$,

$$\operatorname{cov}\big(w(r),\, w(t)\big) = \mathbb{E}\big[w(r)\big(w(t) - w(r)\big)\big] + \mathbb{E}\big[w(r)^2\big] = r.$$

- The sample paths of w are a.s. continuous but highly irregular (nowhere differentiable).

  To see this, note $w_c(t) = w(c^2 t)/c$ for $c > 0$ is also a standard Wiener process. (Why?) Then, $w_c(1/c) = w(c)/c$. For a large c such that $w(c)/c > 1$, $\frac{w_c(1/c)}{1/c} = w(c) > c$. That is, the sample path of w_c has a slope larger than c on a very small interval $(0, 1/c)$.

- The difference quotient:

$$[w(t+h) - w(t)]/h \sim \mathcal{N}(0,\ 1/|h|)$$

  can not converge to a finite limit (as $h \rightarrow 0$) with a positive prob.

- The d-dimensional, standard Wiener process $\mathbf{w}$ consists of d mutually independent, standard Wiener processes, so that for $s < t$, $\mathbf{w}(t) - \mathbf{w}(s) \sim \mathcal{N}(\mathbf{0},\ (t-s)\,\mathbf{I}_d)$.

  1. $\mathbf{w}(t) \sim \mathcal{N}(\mathbf{0},\ t\,\mathbf{I}_d)$.
  2. $\operatorname{cov}(\mathbf{w}(r), \mathbf{w}(t)) = \min(r,t)\,\mathbf{I}_d$.

- The Brownian bridge $\mathbf{w}^0$ on $[0,1]$ is $\mathbf{w}^0(t) = \mathbf{w}(t) - t\mathbf{w}(1)$. Clearly, $\mathbb{E}[\mathbf{w}^0(t)] = \mathbf{0}$, and for $r < t$,

$$\begin{aligned} \operatorname{cov}\big(\mathbf{w}^0(r), \mathbf{w}^0(t)\big) &= \operatorname{cov}\big(\mathbf{w}(r) - r\mathbf{w}(1), \mathbf{w}(t) - t\mathbf{w}(1)\big) \\ &= r(1-t)\,\mathbf{I}_d. \end{aligned}$$

Weak Convergence

$\mathbb{P}_n$ converges weakly to $\mathbb{P}$, denoted as $\mathbb{P}_n \Rightarrow \mathbb{P}$, if for every bounded, continuous real function f on S,

$$\int f(s)\, \mathrm{d}\mathbb{P}_n(s) \to \int f(s)\, \mathrm{d}\, \mathbb{P}(s),$$

where $\{\mathbb{P}_n\}$ and $\mathbb{P}$ are probability measures on $(S, \mathcal{S})$.

- When $\mathbf{z}_n$ and $\mathbf{z}$ are all $\mathbb{R}^d$-valued random variables, $\mathbb{P}_n \Rightarrow \mathbb{P}$ reduces to the usual notion of convergence in distribution: $\mathbf{z}_n \xrightarrow{D} \mathbf{z}$.
- When $\mathbf{z}_n$ and $\mathbf{z}$ are d-dimensional stochastic processes with the distributions induced by $\mathbb{P}_n$ and $\mathbb{P}$, $\mathbf{z}_n \xrightarrow{D} \mathbf{z}$, also denoted as $\mathbf{z}_n \Rightarrow \mathbf{z}$, implies that all the finite-dimensional distributions of $\mathbf{z}_n$ converge to the corresponding distributions of $\mathbf{z}$.

Continuous Mapping Theorem

Continuous Mapping Theorem

Let $g\colon \mathbb{R}^d \mapsto \mathbb{R}$ be a function continuous almost everywhere on $\mathbb{R}^d$, except for at most countably many points. If $\mathbf{z}_n \Rightarrow \mathbf{z}$, then $g(\mathbf{z}_n) \Rightarrow g(\mathbf{z})$.

Proof: Let S and S' be two metric spaces with Borel σ-algebras $\mathcal{S}$ and $\mathcal{S}'$ and $g\colon S \mapsto S'$ be a measurable mapping. For $\mathbb{P}$ on $(S, \mathcal{S})$, define $\mathbb{P}^*$ on $(S', \mathcal{S}')$ as

$$\mathbb{P}^*(A') = \mathbb{P}(g^{-1}(A')), \qquad A' \in \mathcal{S}'.$$

For every bounded, continuous f on S', $f \circ g$ is also bounded and continuous on S. $\mathbb{P}_n \Rightarrow \mathbb{P}$ now implies that

$$\int f \circ g(s)\, \mathrm{d}\,\mathbb{P}_n(s) \to \int f \circ g(s)\, \mathrm{d}\,\mathbb{P}(s),$$

which is equivalent to $\int f(a)\, \mathrm{d}\,\mathbb{P}_n^*(a) \to \int f(a)\, \mathrm{d}\,\mathbb{P}^*(a)$, proving $\mathbb{P}_n^* \Rightarrow \mathbb{P}^*$.

Functional Central Limit Theorem (FCLT)

- ζ_i are i.i.d. with mean zero and variance σ^2. Let $s_n = \zeta_1 + \cdots + \zeta_n$ and $z_n(i/n) = (\sigma\sqrt{n})^{-1} s_i$.
- For $t \in [(i-1)/n, i/n)$, the constant interpolations of $z_n(i/n)$ is

$$z_n(t) = z_n((i-1)/n) = \frac{1}{\sigma\sqrt{n}}\, s_{[nt]},$$

where $[nt]$ is the the largest integer less than or equal to nt.

- From Lindeberg-Lévy's CLT,

$$\frac{1}{\sigma\sqrt{n}}\, s_{[nt]} = \left(\frac{[nt]}{n}\right)^{1/2} \frac{1}{\sigma\sqrt{[nt]}}\, s_{[nt]} \xrightarrow{D} \sqrt{t}\,\mathcal{N}(0,\ 1),$$

which is just $\mathcal{N}(0,\ t)$, the distribution of $w(t)$.

- For $r < t$, we have

$$(z_n(r),\, z_n(t) - z_n(r)) \xrightarrow{D} \big(w(r),\, w(t) - w(r)\big),$$

and hence $(z_n(r),\, z_n(t)) \xrightarrow{D} (w(r),\, w(t))$. This is easily extended to establish convergence of any finite-dimensional distributions and leads to the functional central limit theorem (or invariance principle).

Donsker's Invariane Principle

Let ζ_t be i.i.d. with mean μ_o and variance $\sigma_o^2 > 0$ and

$$z_T(r) = \frac{1}{\sigma_o\sqrt{T}} \sum_{t=1}^{[Tr]} (\zeta_t - \mu_o), \quad r \in [0,1].$$

Then, $z_T \Rightarrow w$ as $T \to \infty$.

- Non-i.i.d. random variables: Let ζ_t be r.v.s with mean μ_t and variance σ_t^2. $\{\zeta_t\}$ is said to obey an FCLT if

$$z_T(r) = \frac{1}{\sigma_* \sqrt{T}} \sum_{t=1}^{[Tr]} (\zeta_t - \mu_t) \Rightarrow w(r), \quad r \in [0,1],$$

where σ_*^2 is the long-run variance of ζ_t:

$$\sigma_*^2 = \lim_{T \to \infty} \operatorname{var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \zeta_t \right).$$

- Note that σ_*^2 accommodates the correlations among ζ_t. When ζ_t are i.i.d. with variance σ_o^2, $\sigma_*^2 = \sigma_o^2$.

- Multivariate random variables: Let $\boldsymbol{\zeta}_t$ be r.v.s with mean $\boldsymbol{\mu}_t$ and variance $\boldsymbol{\Sigma}_t^2$. $\{\boldsymbol{\zeta}_t\}$ obeys an FCLT if

$$\mathbf{z}_T(r) = \frac{1}{\sqrt{T}} \boldsymbol{\Sigma}_*^{-1/2} \sum_{t=1}^{[Tr]} (\boldsymbol{\zeta}_t - \boldsymbol{\mu}_t) \Rightarrow \mathbf{w}(r), \quad r \in [0,1],$$

where $\mathbf{w}$ is the d-dimensional, standard Wiener process, and $\boldsymbol{\Sigma}_*$ is the long-run covariance matrix:

$$\boldsymbol{\Sigma}_* = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\left(\sum_{t=1}^{T} (\boldsymbol{\zeta}_t - \boldsymbol{\mu}_t) \right) \left(\sum_{t=1}^{T} (\boldsymbol{\zeta}_t - \boldsymbol{\mu}_t) \right)' \right].$$

Example: $y_t = y_{t-1} + u_t$, $t = 1, 2, \ldots$, with $y_0 = 0$, where u_t are i.i.d. with mean zero and variance σ_u^2. By Donsker's FCLT, the partial sum $y_{[Tr]} = \sum_{t=1}^{[Tr]} u_t$ is such that

$$\frac{1}{T^{3/2}} \sum_{t=1}^{T} y_t = \sigma_u \sum_{t=1}^{T} \int_{t/T}^{(t+1)/T} \frac{1}{\sqrt{T}\sigma_u} y_{[Tr]} \, \mathrm{d}r \Rightarrow \sigma_u \int_0^1 w(r) \, \mathrm{d}r,$$

Similarly,

$$\frac{1}{T^2} \sum_{t=1}^{T} y_t^2 = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{y_t}{\sqrt{T}} \right)^2 \Rightarrow \sigma_u^2 \int_0^1 w(r)^2 \, \mathrm{d}r,$$

so that $\sum_{t=1}^{T} y_t^2$ is $O_{\mathbb{P}}(T^2)$.

$I(1)$ Series

$\{y_t\}$ is said to be an $I(1)$ (integrated of order 1) series if $y_t = y_{t-1} + \epsilon_t$, with ϵ_t satisfying:

[C1] $\{\epsilon_t\}$ is a weakly stationary series with mean zero and variance σ_ϵ^2 and obeys an FCLT:

$$\frac{1}{\sigma_* \sqrt{T}} \sum_{t=1}^{[Tr]} \epsilon_t = \frac{1}{\sigma_* \sqrt{T}} y_{[Tr]} \Rightarrow w(r), \qquad 0 \leq r \leq 1,$$

where w is standard Wiener process, and σ_*^2 is the long-run variance of ϵ_t:

$$\sigma_*^2 = \lim_{T \to \infty} \operatorname{var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \epsilon_t \right).$$

- Partial sums of an $I(0)$ series (e.g., $\sum_{i=1}^{t} \epsilon_i$) form an $I(1)$ series, while taking first difference of an $I(1)$ series (e.g., $y_t - y_{t-1}$) yields an $I(0)$ series.
    - A random walk is $I(1)$ with i.i.d. ϵ_t and $\sigma_*^2 = \sigma_\epsilon^2$.
    - When $\epsilon_t = y_t - y_{t-1}$ is a stationary ARMA(p,q) series, y is an $I(1)$ series and known as an ARIMA$(p,1,q)$ series.
- An $I(1)$ series y_t has mean zero and variance increasing linearly with t, and its autocovariances cov(y_t, y_s) do not decrease when $|t-s|$ increases.
- Many macroeconomic and financial time series are (or behave like) $I(1)$ series.

ARIMA vs. ARMA Series

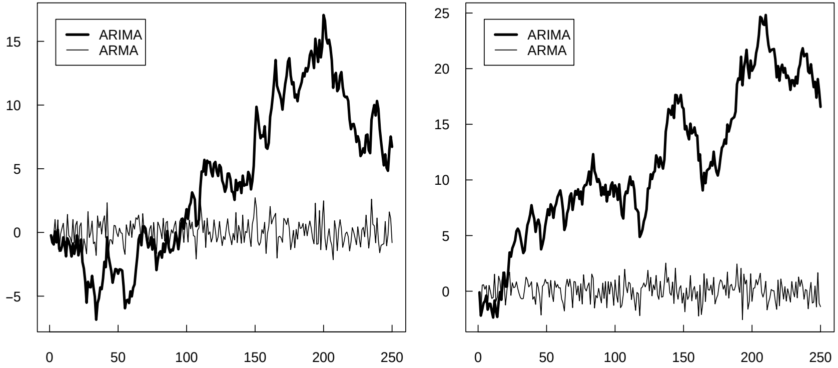

Figure: Sample paths of ARIMA and ARMA series.

$I(1)$ vs. Trend Stationarity

Trend stationary series: $y_t = a_o + b_o t + \epsilon_t$, where ϵ_t are $I(0)$.

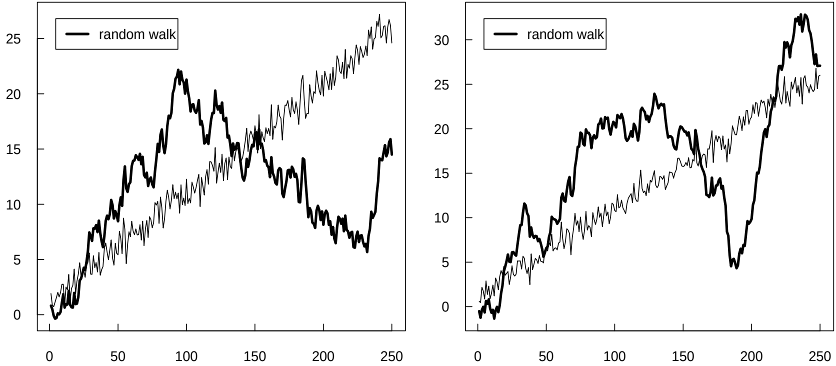

Figure: Sample paths of random walk and trend stationary series.

Autoregression of $I(1)$ Series

Suppose $\{y_t\}$ is a random walk such that $y_t = \alpha_o y_{t-1} + \epsilon_t$ with $\alpha_o = 1$ and ϵ_t i.i.d. random variables with mean zero and variance σ_ϵ^2.

- $\{y_t\}$ does not obey a LLN, and $\sum_{t=2}^T y_{t-1}\epsilon_t = O_{\mathbb{P}}(T)$ and $\sum_{t=2}^T y_{t-1}^2 = O_{\mathbb{P}}(T^2)$.
- Given the specification: $y_t = \alpha y_{t-1} + e_t$, the OLS estimator of α is:

$$\hat{\alpha}_T = \frac{\sum_{t=2}^T y_{t-1}y_t}{\sum_{t=2}^T y_{t-1}^2} = 1 + \frac{\sum_{t=2}^T y_{t-1}\epsilon_t}{\sum_{t=2}^T y_{t-1}^2} = 1 + O_{\mathbb{P}}(T^{-1}),$$

which is T-consistent. This is also known as a super consistent estimator.

Asymptotic Properties of the OLS Estimator

Limits of Partial Sums: I

Let $y_t = y_{t-1} + \epsilon_t$ be an $I(1)$ series with ϵ_t satisfying [C1]. Then,

(i) $T^{-3/2} \sum_{t=1}^{T} y_{t-1} \Rightarrow \sigma_* \int_0^1 w(r)\,\mathrm{d}r$;

(ii) $T^{-2} \sum_{t=1}^{T} y_{t-1}^2 \Rightarrow \sigma_*^2 \int_0^1 w(r)^2\,\mathrm{d}r$;

(iii) $T^{-1} \sum_{t=1}^{T} y_{t-1}\epsilon_t \Rightarrow$

$$\frac{1}{2}[\sigma_*^2 w(1)^2 - \sigma_\epsilon^2] = \sigma_*^2 \int_0^1 w(r)\,\mathrm{d}w(r) + \frac{1}{2}(\sigma_*^2 - \sigma_\epsilon^2),$$

where w is the standard Wiener process.

Note: When y_t is a random walk, $\sigma_*^2 = \sigma_\epsilon^2$.

Model without a Constant Term

Let $y_t = y_{t-1} + \epsilon_t$ be an $I(1)$ series with ϵ_t satisfying [C1]. Given the specification $y_t = \alpha y_{t-1} + e_t$, the normalized OLS estimator of α is:

$$T(\hat{\alpha}_T - 1) = \frac{\sum_{t=2}^{T} y_{t-1}\epsilon_t / T}{\sum_{t=2}^{T} y_{t-1}^2 / T^2} \Rightarrow \frac{\frac{1}{2}\left[w(1)^2 - \sigma_\epsilon^2/\sigma_*^2\right]}{\int_0^1 w(r)^2 \, \mathrm{d}r}.$$

where w is the standard Wiener process. When y_t is a random walk,

$$T(\hat{\alpha}_T - 1) \Rightarrow \frac{\frac{1}{2}\left[w(1)^2 - 1\right]}{\int_0^1 w(r)^2 \, \mathrm{d}r},$$

which does not depend on σ_ϵ^2 and σ_*^2 and is asymptotically pivotal.

Limits of Partial Sums: II

Let $y_t = y_{t-1} + \epsilon_t$ be an $I(1)$ series with ϵ_t satisfying [C1]. Then,

(i) $T^{-2} \sum_{t=1}^{T} (y_{t-1} - \bar{y}_{-1})^2 \Rightarrow \sigma_*^2 \displaystyle\int_0^1 w^*(r)^2 \, \mathrm{d}r$;

(ii) $T^{-1} \sum_{t=1}^{T} (y_{t-1} - \bar{y}_{-1}) \epsilon_t \Rightarrow \sigma_*^2 \displaystyle\int_0^1 w^*(r) \, \mathrm{d}w(r) + \frac{1}{2}(\sigma_*^2 - \sigma_\epsilon^2)$,

where w is the standard Wiener process and $w^*(t) = w(t) - \int_0^1 w(r) \, \mathrm{d}r$.

Model with a Constant Term

Let $y_t = y_{t-1} + \epsilon_t$ be an $I(1)$ series with ϵ_t satisfying [C1]. Given the specification $y_t = c + \alpha y_{t-1} + e_t$, the normalized OLS estimators of α and c are:

$$T(\hat{\alpha}_T - 1) \Rightarrow \frac{\int_0^1 w^*(r)\,\mathrm{d}w(r) + \frac{1}{2}(1 - \sigma_\epsilon^2/\sigma_*^2)}{\int_0^1 w^*(r)^2\,\mathrm{d}r} =: A,$$

$$\sqrt{T}\hat{c}_T \Rightarrow A\left(\sigma_* \int_0^1 w(r)\,\mathrm{d}r\right) + \sigma_* w(1).$$

In particular, when y_t is a random walk,

$$T(\hat{\alpha}_T - 1) \Rightarrow \frac{\int_0^1 w^*(r)\,\mathrm{d}w(r)}{\int_0^1 w^*(r)^2\,\mathrm{d}r}.$$

- The limiting results for autoregressions with an $I(1)$ series are not invariant to model specification.
- All the results here are based on the data with DGP: $y_t = y_{t-1} + \epsilon_t$. intercept. These results would break down if the DGP is $y_t = c_o + y_{t-1} + \epsilon_t$ with a non-zero c_o; such series are said to be $I(1)$ with drift.
- For an $I(1)$ series with a drift:

$$y_t = c_o + y_{t-1} + \epsilon_t = c_o\, t + \sum_{i=1}^{t} \epsilon_i,$$

which contains a deterministic trend and an $I(1)$ series without drift.

Tests of Unit Root

1. Given the specification $y_t = \alpha y_{t-1} + e_t$, the unit root hypothesis is $\alpha_o = 1$, and a leading unit-root test is the t test:

$$\tau_0 = \frac{\left(\sum_{t=2}^{T} y_{t-1}^2\right)^{1/2} (\hat{\alpha}_T - 1)}{\hat{\sigma}_{T,1}},$$

   where $\hat{\sigma}_{T,1}^2 = \sum_{t=2}^{T} (y_t - \hat{\alpha}_T y_{t-1})^2 / (T-2)$.

2. Given the specification $y_t = c + \alpha y_{t-1} + e_t$, a unit-root test is

$$\tau_c = \frac{\left[\sum_{t=2}^{T} (y_{t-1} - \bar{y}_{-1})^2\right]^{1/2} (\hat{\alpha}_T - 1)}{\hat{\sigma}_{T,2}},$$

   where $\hat{\sigma}_{T,2}^2 = \sum_{t=2}^{T} (y_t - \hat{c}_T - \hat{\alpha}_T y_{t-1})^2 / (T-3)$.

Dickey-Fuller Tests

Dickey-Fuller Tests: Random Walk

Let y_t be generated as a random walk. Then,

$$\tau_0 \Rightarrow \frac{\frac{1}{2}[w(1)^2 - 1]}{\left[\int_0^1 w(r)^2 \,\mathrm{d}r\right]^{1/2}},$$

$$\tau_c \Rightarrow \frac{\int_0^1 w^*(r) \,\mathrm{d}w(r)}{\left[\int_0^1 w^*(r)^2 \,\mathrm{d}r\right]^{1/2}}.$$

- For the specification with a time trend variable:

$$y_t = c + \alpha y_{t-1} + \beta\left(t - \frac{T}{2}\right) + e_t,$$

the t-statistic of $\alpha_o = 1$ is denoted as τ_t.

Dickey-Fuller distributions

Table: Some percentiles of the Dickey-Fuller distributions.

Test	1%	2.5%	5%	10%	50%	90%	95%	97.5%	99%
τ_0	-2.58	-2.23	-1.95	-1.62	-0.51	0.89	1.28	1.62	2.01
τ_c	-3.42	-3.12	-2.86	-2.57	-1.57	-0.44	-0.08	0.23	0.60
τ_t	-3.96	-3.67	-3.41	-3.13	-2.18	-1.25	-0.94	-0.66	-0.32

- These distributions are not symmetric about zero and assume more negative values.
- τ_c assumes negatives values about 95% of times, and τ_t is virtually a non-positive random variable.

The Dickey-Fuller Distributions

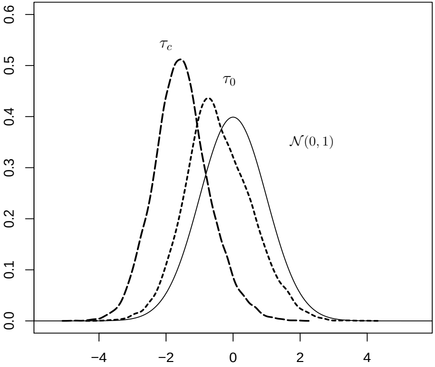

Figure: The limiting distributions of the Dickey-Fuller τ_0 and τ_c tests.

Implementation

In practice, we estimate one of the following specifications:

1. $\Delta y_t = \theta y_{t-1} + e_t$.
2. $\Delta y_t = c + \theta y_{t-1} + e_t$.
3. $\Delta y_t = c + \theta y_{t-1} + \beta\big(t - \frac{T}{2}\big) + e_t$.

The unit-root hypothesis $\alpha_o = 1$ is now equivalent to $\theta_o = 0$.

- The weak limits of the normalized estimators $T\hat{\theta}_T$ are the same as the respective limits of $T(\hat{\alpha}_T - 1)$ under the null hypothesis.
- The unit-root tests are now computed as the *t*-ratios of these specifications.

Phillips-Perron Tests

Note: The Dickey-Fuller tests check only the random walk hypothesis and are invalid for testing general $I(1)$ series.

Dickey-Fuller Tests: General $I(1)$ Series

Let $y_t = y_{t-1} + \epsilon_t$ be an $I(1)$ series with ϵ_t satisfying [C1]. Then,

$$\tau_0 \Rightarrow \frac{\sigma_*}{\sigma_\epsilon} \left(\frac{\frac{1}{2}[w(1)^2 - \sigma_\epsilon^2/\sigma_*^2]}{\left[\int_0^1 w(r)^2 \, \mathrm{d}r\right]^{1/2}} \right),$$

$$\tau_c \Rightarrow \frac{\sigma_*}{\sigma_\epsilon} \left(\frac{\int_0^1 w^*(r) \, \mathrm{d}w(r) + \frac{1}{2}(1 - \sigma_\epsilon^2/\sigma_*^2)}{\left[\int_0^1 w^*(r)^2 \, \mathrm{d}r\right]^{1/2}} \right),$$

- Let $\hat{e}_t$ denote the OLS residuals and s^2_{Tn} a Newey-West type estimator of σ^2_* based on $\hat{e}_t$:

$$s^2_{Tn} = \frac{1}{T-1}\sum_{t=2}^{T}\hat{e}_t^2 + \frac{2}{T-1}\sum_{s=1}^{T-2}\kappa\left(\frac{s}{n}\right)\sum_{t=s+2}^{T}\hat{e}_t\hat{e}_{t-s},$$

with κ a kernel function and $n = n(T)$ its bandwidth.

- Phillips (1987) proposed the following modified τ_0 and τ_c statistics:

$$Z(\tau_0) = \frac{\hat{\sigma}_T}{s_{Tn}}\tau_0 - \frac{\frac{1}{2}(s^2_{Tn} - \hat{\sigma}^2_T)}{s_{Tn}\left(\sum_{t=2}^{T} y^2_{t-1}/T^2\right)^{1/2}},$$

$$Z(\tau_c) = \frac{\hat{\sigma}_T}{s_{Tn}}\tau_c - \frac{\frac{1}{2}(s^2_T - \hat{\sigma}^2_T)}{s_{Tn}\left[\sum_{t=2}^{T}(y_{t-1} - \bar{y}_{-1})^2\right]^{1/2}};$$

see also Phillips and Perron (1988).

The Phillips-Perron tests eliminate the nuisance parameters by suitable transformations of τ_0 and τ_c and have the same limits as those of the Dickey-Fuller tests.

Phillips-Perron Tests

Let $y_t = y_{t-1} + \epsilon_t$ be an $I(1)$ series with ϵ_t satisfying [C1]. Then,

$$Z(\tau_0) \Rightarrow \frac{\frac{1}{2}\left[w(1)^2 - 1\right]}{\left[\int_0^1 w(r)^2 \, \mathrm{d}r\right]^{1/2}},$$

$$Z(\tau_c) \Rightarrow \frac{\int_0^1 w^*(r) \, \mathrm{d}w(r)}{\left[\int_0^1 w^*(r)^2 \, \mathrm{d}r\right]^{1/2}}.$$

Augmented Dickey-Fuller (ADF) Tests

Said and Dickey (1984) suggest "filtering out" the correlations in a weakly stationary series by a linear AR model with a proper order. The "augmented" specifications are:

1. $\Delta y_t = \theta y_{t-1} + \sum_{j=1}^{k} \gamma_j \Delta y_{t-j} + e_t$.
2. $\Delta y_t = c + \theta y_{t-1} + \sum_{j=1}^{k} \gamma_j \Delta y_{t-j} + e_t$.
3. $\Delta y_t = c + \theta y_{t-1} + \beta\left(t - \frac{T}{2}\right) + \sum_{j=1}^{k} \gamma_j \Delta y_{t-j} + e_t$.

Note: This approach avoids non-parametric kernel estimation of σ_*^2 but requires choosing a proper lag order k for the augmented specifications (say, by a model selection criteria, such as AIC or SIC).

KPSS Tests

$\{y_t\}$ is trend stationary if it fluctuates around a deterministic trend:

$$y_t = a_o + b_o\, t + \epsilon_t,$$

where ϵ_t satisfy [C1]. When $b_o = 0$, it is level stationary. Kwiatkowski, Phillips, Schmidt, and Shin (1992) proposed testing stationarity by

$$\eta_T = \frac{1}{T^2\, s_{Tn}^2} \sum_{t=1}^{T} \left(\sum_{i=1}^{t} \hat{e}_i \right)^2,$$

where s_{Tn}^2 is a Newey-West estimator of σ_*^2 based on $\hat{e}_t$.

- To test the null of trend stationarity, $\hat{e}_t = y_t - \hat{a}_T - \hat{b}_T\, t$.
- To test the null of level stationarity, $\hat{e}_t = y_t - \bar{y}$.

The partial sums of $\hat{e}_t = y_t - \bar{y}$ are such that

$$\sum_{t=1}^{[Tr]} \hat{e}_t = \sum_{t=1}^{[Tr]} (\epsilon_t - \bar{\epsilon}) = \sum_{t=1}^{[Tr]} \epsilon_t - \frac{[Tr]}{T} \sum_{t=1}^{T} \epsilon_t, \quad r \in (0, 1].$$

Then by a suitable FCLT,

$$\frac{1}{\sigma_* \sqrt{T}} \sum_{t=1}^{[Tr]} \hat{e}_t \Rightarrow w(r) - rw(1) = w^0(r).$$

Similarly, given $\hat{e}_t = y_t - \hat{a}_T - \hat{b}_T\, t$,

$$\frac{1}{\sigma_* \sqrt{T}} \sum_{t=1}^{[Tr]} \hat{e}_t \Rightarrow w(r) + (2r - 3r^2) w(1) - (6r - 6r^2) \int_0^1 w(s)\, \mathrm{d}s,$$

which is a "tide-down" process (it is zero at $r = 1$ with prob. one).

KPSS Tests

1. Let $y_t = a_o + b_o\, t + \epsilon_t$ with ϵ_t satisfying [C1]. Then, η_T computed from $\hat{e}_t = y_t - \hat{a}_T - \hat{b}_T\, t$ is:

$$\eta_T \Rightarrow \int_0^1 f(r)^2 \,\mathrm{d}r,$$

where $f(r) = w(r) + (2r - 3r^2)w(1) - (6r - 6r^2)\int_0^1 w(s)\,\mathrm{d}s$.

2. Let $y_t = a_o + \epsilon_t$ with ϵ_t satisfying [C1]. Then, η_T computed from $\hat{e}_t = y_t - \bar{y}$ is:

$$\eta_T \Rightarrow \int_0^1 w^0(r)^2 \,\mathrm{d}r,$$

where w^0 is the Brownian bridge.

Table: Some percentiles of the distributions of the KPSS test.

Test	1%	2.5%	5%	10%
level stationarity	0.739	0.574	0.463	0.347
trend stationarity	0.216	0.176	0.146	0.119

- These tests have power against $I(1)$ series because η_T would diverge under $I(1)$ alternatives.
- KPSS tests also have power against other alternatives, such as stationarity with mean changes and trend stationarity with trend breaks. Thus, rejecting the null of stationarity does not imply that the series must be $I(1)$.

The KPSS Distributions

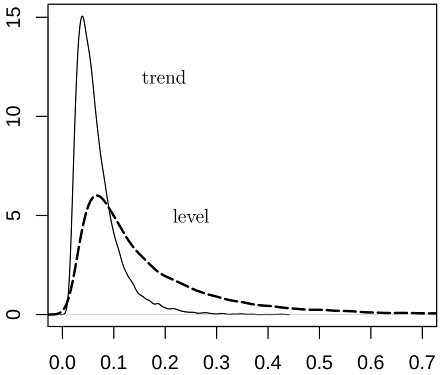

Figure: The limiting distributions of the KPSS tests.

Spurious Regressions

Given two independent random walks, Granger and Newbold (1974) found that regressing one on the other typically yields a significant t-ratio. This is known as the problem of spurious regression.

Let $y_t = y_{t-1} + u_t$ and $x_t = x_{t-1} + v_t$ be $I(1)$ series, where u_t and v_t are mutually independent series satisfying the following condition.

[C2] u_t and v_t are two weakly stationary series with mean zero and variances σ_u^2 and σ_v^2. They obey FCLT with the long-run variances:

$$\sigma_y^2 = \lim_{T\to\infty} \frac{1}{T} \mathbb{E} \left(\sum_{t=1}^{T} u_t \right)^2, \qquad \sigma_x^2 = \lim_{T\to\infty} \frac{1}{T} \mathbb{E} \left(\sum_{t=1}^{T} v_t \right)^2.$$

Consider the regression: $y_t = \alpha + \beta x_t + e_t$. Let $\hat{\alpha}_T$ and $\hat{\beta}_T$ denote the OLS estimators and $t_\alpha = \hat{\alpha}_T / s_\alpha$ and $t_\beta = \hat{\beta}_T / s_\beta$ their t-ratios, where s_α and s_β are the OLS standard errors.

The results below are immediate:

$$\frac{1}{T^{3/2}} \sum_{t=1}^{T} y_t \Rightarrow \sigma_y \int_0^1 w_y(r)\, \mathrm{d}\, r, \quad \frac{1}{T^2} \sum_{t=1}^{T} y_t^2 \Rightarrow \sigma_y^2 \int_0^1 w_y(r)^2\, \mathrm{d}\, r,$$

where w_y is a standard Wiener processes. Similarly,

$$\frac{1}{T^{3/2}} \sum_{t=1}^{T} x_t \Rightarrow \sigma_x \int_0^1 w_x(r)\, \mathrm{d}\, r, \quad \frac{1}{T^2} \sum_{t=1}^{T} x_t^2 \Rightarrow \sigma_x^2 \int_0^1 w_x(r)^2\, \mathrm{d}\, r,$$

where w_x is a standard Wiener process which is independent of w_y.

It is also easy to show:

$$\frac{1}{T^2}\sum_{t=1}^{T}(y_t-\bar{y})^2 \Rightarrow \sigma_y^2\int_0^1 w_y(r)^2\,\mathrm{d}r-\sigma_y^2\left(\int_0^1 w_y(r)\,\mathrm{d}r\right)^2 =: \sigma_y^2 m_y,$$

$$\frac{1}{T^2}\sum_{t=1}^{T}(x_t-\bar{x})^2 \Rightarrow \sigma_x^2\int_0^1 w_x(r)^2\,\mathrm{d}r-\sigma_x^2\left(\int_0^1 w_x(r)\,\mathrm{d}r\right)^2 =: \sigma_x^2 m_x,$$

where $w_y^*(t)=w_y(t)-\int_0^1 w_y(r)\,\mathrm{d}r$ and $w_x^*(t)=w_x(t)-\int_0^1 w_x(r)\,\mathrm{d}r$ are two mutually independent, "de-meaned" Wiener processes. Similarly,

$$\begin{aligned}&\frac{1}{T^2}\sum_{t=1}^{T}(y_t-\bar{y})(x_t-\bar{x}_t)\\&\qquad\Rightarrow \sigma_y\sigma_x\left(\int_0^1 w_y(r)w_x(r)\,\mathrm{d}r-\int_0^1 w_y(r)\,\mathrm{d}r\int_0^1 w_x(r)\,\mathrm{d}r\right)\\&\qquad=: \sigma_y\sigma_x m_{yx}.\end{aligned}$$

Phillips (1986)

For the specification $y_t = \alpha + \beta x_t + e_t$, we have:

1. $\hat{\beta}_T \Rightarrow \dfrac{\sigma_y \, m_{yx}}{\sigma_x \, m_x}$,
2. $T^{-1/2}\hat{\alpha}_T \Rightarrow \sigma_y \left(\displaystyle\int_0^1 w_y(r)\,\mathrm{d}r - \frac{m_{yx}}{m_x} \int_0^1 w_x(r)\,\mathrm{d}r \right)$,
3. $T^{-1/2}\, t_\beta \Rightarrow \dfrac{m_{yx}}{(m_y m_x - m_{yx}^2)^{1/2}}$,
4. $T^{-1/2}\, t_\alpha \Rightarrow \dfrac{m_x \int_0^1 w_y(r)\,\mathrm{d}r - m_{yx} \int_0^1 w_x(r)\,\mathrm{d}r}{\left[(m_y m_x - m_{yx}^2) \int_0^1 w_x(r)^2\,\mathrm{d}r\right]^{1/2}}$,

where w_x and w_y are mutually independent, standard Wiener processes.

Remark: As t_α and t_β both diverge, it is easy to obtain large t-ratios and to conclude that these coefficients are significantly different from zero based on the critical values from $\mathcal{N}(0,1)$.

- Spurious correlation: When y_t and x_t are mutually independent $I(1)$ series, their sample correlation coefficient does not converge in probability to zero but converges weakly to a random variable. (Show the result!)
- Spurious trend: Nelson and Kang (1984) also showed that, given the time trend specification for a random walk y_t:

  $$y_t = a + b\,t + e_t,$$

  it is likely to draw a false inference that the time trend is significant in explaining y_t. Phillips and Durlauf (1986) show the F test of $b_o = 0$ diverges at the rate T; this explains why spurious trend may arise.

Co-Integration

Consider an equilibrium relation $ay - bx = 0$. With real data (y_t, x_t), $z_t := ay_t - bx_t$ are equilibrium errors. When y_t and x_t are both $I(1)$, a linear combination of them is, in general, also $I(1)$. Then, z_t have growing variance and wander away from zero (the equilibrium condition). As such, the equilibrium condition places no empirical restriction on z_t.

The equilibrium condition is empirically relevant when a linear combination of $I(1)$ series is $I(0)$. Suppose y_t and x_t involve the same random walk q_t: $y_t = q_t + u_t$ and $x_t = cq_t + v_t$, where u_t and v_t are two $I(0)$ series. A linear combination of y_t and x_t that annihilates the common trend is

$$z_t = cy_t - x_t = cu_t - v_t,$$

which is clearly $I(0)$.

- Granger (1981) and Engle and Granger (1987): Let $\mathbf{y}_t$ $(d \times 1)$ be a vector $I(1)$ series. The elements of $\mathbf{y}_t$ are co-integrated if there exists a $d \times 1$ vector, $\boldsymbol{\alpha}$, such that $z_t = \boldsymbol{\alpha}'\mathbf{y}_t$ is $I(0)$. We say the elements of $\mathbf{y}_t$ are CI(1,1). The vector $\boldsymbol{\alpha}$ is a co-integrating vector (CIV).
- If $\boldsymbol{\alpha}$ is a CIV, so is $c\boldsymbol{\alpha}$ for any $c \neq 0$. Hence, we are interested in CIVs that are linearly independent. The space spanned by linearly independent CIVs is the co-integrating space, and its dimension is the co-integrating rank.
- If the co-integrating rank is $r < d$, putting the r CIVs together we have the $d \times r$ matrix $\mathbf{A}$ such that $\mathbf{z}_t = \mathbf{A}'\mathbf{y}_t$ is a vector $I(0)$ series. The co-integrating rank is at most $d - 1$. (Why?)
- More generally, let $\mathbf{y}_t$ be $I(n)$, if there exist $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}\mathbf{y}_t$ is $I(n-m)$, the elements of $\mathbf{y}_t$ are said to be $\mathrm{CI}(n, m)$.

Characterization of Co-Integration

Suppose that $\mathbf{y}_t$ $(d \times 1)$ is an $I(n)$ series such that $\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \epsilon_t$, where $\mathbf{\Psi}(z)$ is a matrix of polynomials and $\det(\mathbf{\Psi}(z)) = 0$ has solutions on or outside the unit circle. Let $\operatorname{adj}(\mathbf{\Psi}(z))$ denote the adjoint matrix of $\mathbf{\Psi}(z)$. It is well known that $\operatorname{adj}(\mathbf{\Psi}(z))\mathbf{\Psi}(z) = \det(\mathbf{\Psi}(z))\,\mathbf{I}_d$, so that

$$\det(\mathbf{\Psi}(\mathcal{B}))\mathbf{y}_t = \operatorname{adj}(\mathbf{\Psi}(\mathcal{B}))\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \operatorname{adj}(\mathbf{\Psi}(\mathcal{B}))\epsilon_t.$$

Co-Integration: I

Given $\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \epsilon_t$, where $\mathbf{y}_t$ is $I(n)$, any vector α such that $\alpha'\operatorname{adj}(\mathbf{\Psi}(1)) = \mathbf{0}$ is a CIV, and the CIV space is the null space (kernel) of $\operatorname{adj}(\mathbf{\Psi}(1))'$. If the co-integrating rank is r, then $\mathrm{rank}(\operatorname{adj}(\mathbf{\Psi}(1))) = d - r$.

If $\mathbf{\Psi}(z)$ has n unit roots, $\det(\mathbf{\Psi}(z)) = (1-z)^n J(z)$ with $J(1) \neq 0$. Then, $J(\mathcal{B})^{-1}$ is well defined and

$$(1-\mathcal{B})^n \mathbf{y}_t = J(\mathcal{B})^{-1} \mathrm{adj}(\mathbf{\Psi}(\mathcal{B}))\boldsymbol{\epsilon}_t$$

is an $I(0)$ series. If there is an $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{\alpha}' \mathrm{adj}(\mathbf{\Psi}(1)) = \mathbf{0}$, then for some $m > 0$ we can write $\boldsymbol{\alpha}' \mathrm{adj}(\mathbf{\Psi}(z)) = (1-z)^m \mathbf{h}(z)'$, where $\mathbf{h}$ is a vector polynomial with $\mathbf{h}(1) \neq \mathbf{0}$. It follows that

$$(1-\mathcal{B})^n \boldsymbol{\alpha}' \mathbf{y}_t = J(\mathcal{B})^{-1} \boldsymbol{\alpha}' \mathrm{adj}(\mathbf{\Psi}(\mathcal{B}))\boldsymbol{\epsilon}_t = J(\mathcal{B})^{-1} (1-\mathcal{B})^m \mathbf{h}(\mathcal{B})' \boldsymbol{\epsilon}_t,$$

or equivalently,

$$(1-\mathcal{B})^{n-m} \boldsymbol{\alpha}' \mathbf{y}_t = J(\mathcal{B})^{-1} \mathbf{h}(\mathcal{B})' \boldsymbol{\epsilon}_t,$$

which is $I(0)$. That is, $\boldsymbol{\alpha}$ is a CIV and $\mathbf{y}_t \sim CI(n, m)$.

Conversely, if $\boldsymbol{\alpha}$ is a CIV,

$$(1-\mathcal{B})^n \boldsymbol{\alpha}' \mathbf{y}_t = \boldsymbol{\alpha}' J(\mathcal{B})^{-1} \mathrm{adj}(\boldsymbol{\Psi}(\mathcal{B})) \boldsymbol{\epsilon}_t,$$

such that the right-hand side is integrated of order less than n. As $J(z)$ does not have a unit root, it must be the case that $\boldsymbol{\alpha}'\mathrm{adj}(\boldsymbol{\Psi}(z))$ have at least one unit root, i.e., $\boldsymbol{\alpha}'\mathrm{adj}(\boldsymbol{\Psi}(1)) = \mathbf{0}$.

Co-Integration: II

Given $\boldsymbol{\Psi}(\mathcal{B})\mathbf{y}_t = \epsilon_t$, where $\mathbf{y}_t$ is $I(n)$, the CIV space is the row space of $\boldsymbol{\Psi}(1)$, and $\mathrm{rank}(\boldsymbol{\Psi}(1))$ is the co-integrating rank.

Given that $\boldsymbol{\Psi}(1)\mathrm{adj}(\boldsymbol{\Psi}(1)) = \det(\boldsymbol{\Psi}(1))\,\mathbf{I}_d = \mathbf{0}$, $\boldsymbol{\Psi}(1)$ is also in the null space of $\mathrm{adj}(\boldsymbol{\Psi}(1))'$ with rank r, the co-integrating rank, because $\mathrm{rank}(\mathrm{adj}(\boldsymbol{\Psi}(1))) = d - r$. Note that the jth row of $\boldsymbol{\Psi}(1)$ can be written as

$$c_{j1}\boldsymbol{\alpha}_1' + c_{j2}\boldsymbol{\alpha}_2 + \cdots + c_{jr}\boldsymbol{\alpha}_r = \mathbf{c}_j'\mathbf{A}'.$$

It follows that $\boldsymbol{\Psi}(1) = \boldsymbol{\Gamma}\mathbf{A}'$, where $\boldsymbol{\Gamma}$ is the matrix with the jth row $\mathbf{c}_j'$.

Remark: If $\mathbf{y}_t \sim CI(1,1)$, estimating an unrestricted VAR model will suffer from efficiency loss because the matrix of AR coefficients, $\boldsymbol{\Psi}(1)$, should satisfy the restriction of singularity.

Example: Consider $\boldsymbol{\Psi}(\mathcal{B})\mathbf{y}_t = \boldsymbol{\epsilon}_t$, where

$$\boldsymbol{\Psi}(z) = \begin{bmatrix} 1-(1-\psi)z & -\psi z \\ \psi z & 1-(1+\psi)z \end{bmatrix}, \quad \boldsymbol{\Psi}(1) = \begin{bmatrix} \psi & -\psi \\ \psi & -\psi \end{bmatrix}.$$

Because $\det(\boldsymbol{\Psi}(z)) = (1-z)^2$, we have

$$(1-\mathcal{B})^2\mathbf{y}_t = \text{adj}(\boldsymbol{\Psi}(B))\boldsymbol{\epsilon}_t = \begin{bmatrix} 1-(1+\psi)\mathcal{B} & \psi\mathcal{B} \\ -\psi\mathcal{B} & 1-(1-\psi)\mathcal{B} \end{bmatrix}\boldsymbol{\epsilon}_t\,.$$

That is, $\mathbf{y}_t$ is $I(2)$. As

$$\text{adj}(\boldsymbol{\Psi}(1)) = \begin{bmatrix} -\psi & \psi \\ -\psi & \psi \end{bmatrix},$$

we have $\boldsymbol{\alpha} = [1 \;\; -1]'$ such that $\boldsymbol{\alpha}'\text{adj}(\boldsymbol{\Psi}(1)) = \mathbf{0}$.

Writing $\boldsymbol{\Psi}(z) = \boldsymbol{\Psi}(1) + (1-z)\boldsymbol{\Psi}(z)^*$,

$$\begin{bmatrix} 1-(1-\psi)z & -\psi z \\ \psi z & 1-(1+\psi)z \end{bmatrix} = \begin{bmatrix} \psi & -\psi \\ \psi & -\psi \end{bmatrix} + (1-z) \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

It can be verified that $a_{11} = 1-\psi$, $a_{12} = \psi$, $a_{21} = -\psi$, and $a_{22} = 1+\psi$. That is,

$$\boldsymbol{\Psi}(z)^* = \begin{bmatrix} 1-\psi & \psi \\ -\psi & 1+\psi \end{bmatrix}.$$

Writing $\text{adj}(\boldsymbol{\Psi}(z)) = \text{adj}(\boldsymbol{\Psi}(1)) + (1-z)\text{adj}(\boldsymbol{\Psi}(z))^*$, we have

$$\text{adj}(\boldsymbol{\Psi}(z))^* = \begin{bmatrix} 1+\psi & -\psi \\ \psi & 1-\psi \end{bmatrix}.$$

It follows that

$$\begin{aligned}\boldsymbol{\alpha}'\operatorname{adj}(\boldsymbol{\Psi}(z)) &= \boldsymbol{\alpha}'\operatorname{adj}(\boldsymbol{\Psi}(1)) + (1-z)\boldsymbol{\alpha}'\operatorname{adj}(\boldsymbol{\Psi}(z))^* \\ &= (1-z)[1 \quad -1].\end{aligned}$$

Thus,

$$(1-\mathcal{B})^2\boldsymbol{\alpha}'\mathbf{y}_t = (1-\mathcal{B})\boldsymbol{\alpha}'\boldsymbol{\epsilon}_t,$$

showing that $\mathbf{y}_t \sim CI(2,1)$.

For $(1-\mathcal{B})\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}(\mathcal{B})\boldsymbol{\epsilon}_t$, using $\mathbf{C}(z) = \mathbf{C}(1) + (1-z)\mathbf{C}(z)^*$ we have

$$\mathbf{y}_t = \boldsymbol{\mu} t + \mathbf{C}(1)\sum_{j=1}^{t}\boldsymbol{\epsilon}_j + \mathbf{C}(\mathcal{B})^*\boldsymbol{\epsilon}_t.$$

Thus, if $\boldsymbol{\alpha}$ is a CIV, we have $\boldsymbol{\alpha}'\boldsymbol{\mu} = 0$ and $\boldsymbol{\alpha}'\mathbf{C}(1) = \mathbf{0}$.

Co-Integration: III

Suppose that $\mathbf{y}_t$ is $CI(1,1)$ and $(1-\mathcal{B})\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}(\mathcal{B})\boldsymbol{\epsilon}_t$. The CIV space is the null space of $\mathbf{C}(1)'$, and $\mathrm{rank}(\mathbf{C}(1)) = d - r$ if the co-integrating rank is r.

Remark: If $\mathbf{y}_t \sim CI(1,1)$ and has an MA representation, the matrix of MA coefficients, $\mathbf{C}(1)$, must also be singular.

Other Representations

- Common-trend representation of Stock & Watson (1988): When $\mathbf{y}_t \sim CI(1,1)$ with co-integrating rank r, the elements of $\mathbf{y}_t$ contains only $d - r$ common (stochastic and deterministic) trends. Moreover, a CIV that eliminates unit roots must also eliminate the time trend.
- ARMA representation of Engle & Granger (1987): When $\mathbf{y}_t \sim CI(1,1)$, there exists a finite vector ARMA representation

  $$\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = d(\mathcal{B})\epsilon_t,$$

  where $d(\mathcal{B})$ is a scalar polynomial with $d(1)$ finite, $\mathbf{\Psi}(0) = \mathbf{I}$, $\operatorname{rank}(\mathbf{\Psi}(1)) = r$, and $\mathbf{\Psi}(1) = \mathbf{\Gamma}\mathbf{A}'$ for some $\mathbf{\Gamma}$. If $d(\mathcal{B}) = 1$, this is just a VAR representation.

Error Correction Model

An error correction model (ECM): The change of one variable is related to past changes of all variables in the system and past equilibrium errors:

$$\boldsymbol{\Phi}(\mathcal{B})(1-\mathcal{B})\mathbf{y}_t = -\boldsymbol{\Gamma}\mathbf{z}_{t-1} + \mathbf{v}_t,$$

where $\mathbf{z}_t = \mathbf{A}'\mathbf{y}_t$, $\boldsymbol{\Gamma} \neq \mathbf{0}$, $\mathbf{v}_t$ is stationary, and $\boldsymbol{\Phi}(0) = \mathbf{I}$. As the levels and differences of $\mathbf{y}_t$ appear in both sides of the equation, an ECM is appropriate only if co-integration exists. Conversely, if $\mathbf{y}_t \sim CI(1,1)$, using the ARMA representation $\boldsymbol{\Psi}(\mathcal{B})\mathbf{y}_t = d(\mathcal{B})\epsilon_t$ we have

$$(\boldsymbol{\Psi}(1) + \boldsymbol{\Psi}(\mathcal{B})^*)(1-\mathcal{B})\mathbf{y}_t = -\boldsymbol{\Psi}(1)\mathbf{y}_{t-1} + d(\mathcal{B})\epsilon_t,$$

by the fact that $\boldsymbol{\Psi}(z) = \boldsymbol{\Psi}(1) + (1-z)\boldsymbol{\Psi}(z)^*$.

Letting $\mathbf{\Phi}(z) = \mathbf{\Psi}(1) + \mathbf{\Psi}(z)^*$, we obtain

$$\mathbf{\Phi}(\mathcal{B})(1-\mathcal{B})\mathbf{y}_t = -\mathbf{\Gamma}\mathbf{A}'\mathbf{y}_{t-1} + d(\mathcal{B})\epsilon_t.$$

As $\mathbf{I} = \mathbf{\Psi}(0) = \mathbf{\Psi}(1) + \mathbf{\Psi}(0)^*$, and hence

$$\mathbf{\Phi}(0) = \mathbf{\Psi}(1) + \mathbf{\Psi}(0)^* = \mathbf{I}.$$

Error Correction Representation

If $\mathbf{y}_t \sim CI(1,1)$, there exists an ECM: $\mathbf{\Phi}(\mathcal{B})(1-\mathcal{B})\mathbf{y}_t = -\mathbf{\Gamma}\mathbf{z}_{t-1} + \mathbf{v}_t$ with $\mathbf{z}_t = \mathbf{A}'\mathbf{y}_t$ and $\mathbf{\Phi}(0) = \mathbf{I}$.

Remark: If $\mathbf{y}_t \sim CI(1,1)$, a VAR model in differences is misspecified because it ignores the long-run equilibrium relationship. As a result, the OLS estimates of such VAR model are inconsistent.

Suppose that $\Delta \mathbf{y}_t = \mathbf{u}_t$ has a bounded and continuous spectral density matrix $\mathbf{f}_{\mathbf{u}}(\lambda)$. If $\mathbf{y}_t \sim CI(1,1)$, let the spectral density of $\boldsymbol{\alpha}'\mathbf{y}_t = z_t$ be $f_z(\lambda)$. The following spectral characterization is due to Phillips and Ouliaris (1988).

Spectral Characterization

$\mathbf{y}_t \sim CI(1,1)$ is equivalent to $\boldsymbol{\alpha}'\mathbf{f}_{\mathbf{u}}(\lambda)\boldsymbol{\alpha} = |1 - e^{i\lambda}|^2 f_z(\lambda)$ (only if $\boldsymbol{\alpha}' f_{\mathbf{u}}(0)\boldsymbol{\alpha} = 0$).

Co-integration thus implies that the long-run variance,

$$\boldsymbol{\Sigma} = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\left(\sum_{t=1}^{T} \mathbf{u}_t\right)\left(\sum_{t=1}^{T} \mathbf{u}_t\right)'\right] = 2\pi\, \mathbf{f}_{\mathbf{u}}(0),$$

is singular.

Recall that for a stationary series v_t with the spectral density of f_v, the spectral density of $w_t = \Phi(\mathcal{B})v_t$ is

$$f_w(\lambda) = \Phi(e^{i\lambda}) f_v(\lambda) \Phi(e^{-i\lambda}).$$

In the context of co-integration, $\boldsymbol{\alpha}'\mathbf{u}_t = \boldsymbol{\alpha}'(\Delta\mathbf{y}_t) = z_t - z_{t-1}$, and

$$\boldsymbol{\alpha}'\mathbf{f}_{\mathbf{u}}(\lambda)\boldsymbol{\alpha} = |1 - e^{i\lambda}|^2 f_z(\lambda).$$

For $\lambda = 0$, $\boldsymbol{\alpha}'\mathbf{f}_{\mathbf{u}}(0)\boldsymbol{\alpha} = 0$. Let $\mathbf{Z}_{\mathbf{u}}(\lambda)$ be the spectral process of $\mathbf{u}$ which is of orthogonal increments, with

$$\mathbb{E}(\mathbf{Z}_{\mathbf{u}}(\lambda)) = \mathbf{0}, \qquad \mathbb{E}(\mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\lambda)\mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\mu)^*) = \delta_{\lambda,\mu}\mathbf{F}_{\mathbf{u}}(d\lambda),$$

where δ is Kronecker delta, $\mathbf{Z}_{\mathbf{u}}^*$ is the complex conjugate of $\mathbf{Z}_{\mathbf{u}}$, and $\mathbf{F}_{\mathbf{u}}$ is the spectral distribution function of $\mathbf{u}$.

The spectral representation of $\mathbf{u}_t$ is:

$$\mathbf{u}_t = \int_{-\pi}^{\pi} e^{it\lambda}\, \mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\lambda).$$

It follows that

$$z_t - z_{t-1} = \boldsymbol{\alpha}'\mathbf{u}_t = \int_{-\pi}^{\pi} e^{it\lambda}\, \boldsymbol{\alpha}'\mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\lambda),$$

and

$$z_t = \int_{-\pi}^{\pi} e^{it\lambda}\, \frac{\boldsymbol{\alpha}'\mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\lambda)}{1 - e^{i\lambda}}.$$

This shows $\int_{-\pi}^{\pi} f_z(\lambda)\,\mathrm{d}\lambda$ is

$$\mathbb{E}(z_t^2) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\boldsymbol{\alpha}'\,\mathbb{E}(\mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\lambda)\mathbf{Z}_{\mathbf{u}}(\,\mathrm{d}\mu)^*)\boldsymbol{\alpha}}{|1-e^{i\lambda}||1-e^{i\mu}|} = \int_{-\pi}^{\pi} \frac{\boldsymbol{\alpha}'\mathbf{f}_{\mathbf{u}}(\lambda)\boldsymbol{\alpha}}{|1-e^{i\lambda}|^2}\,\mathrm{d}\lambda.$$

Engle-Granger Two-Step Procedure

A natural way to estimate the CIV $\boldsymbol{\alpha}$ is to minimize the sample variation of $\boldsymbol{\alpha}'\mathbf{y}_t$. The two-step procedure of Engle and Granger (1987) is:

1. Co-integrating regression: $y_{1t} = c + \mathbf{a}'\mathbf{y}_{2t} + \zeta_t$, with the normalized CIV $\boldsymbol{\alpha} = [1 \quad -\mathbf{a}']'$. The OLS estimator $\hat{\mathbf{a}}_T$ is T-consistent. Note that the choice of dependent variable here is arbitrary.
2. For estimating ECM, replac ζ_{t-1} with lagged OLS residuals $\hat{\zeta}_{t-1}$:

$$\boldsymbol{\Phi}(\mathcal{B})(1-\mathcal{B})\mathbf{y}_t = -\boldsymbol{\Gamma}\hat{\zeta}_{t-1} + v_t.$$

The asymptotic properties of the resutling ECM coefficient estimates are the same as those obtained in the ECM with ζ_{t-1}.

- Simultaneity: When the elements of $\mathbf{y}_t$ are co-integrated, they must be determined jointly, so that ζ_t are correlated with $\mathbf{y}_{2,t}$. This correlation does not affect OLS consistency in co-integrating regression but may cause finite-sample bias and efficiency loss.
- Saikkonen (1991) considers the following projections:

$$\zeta_t = \sum_{j=-\infty}^{\infty} \mathbf{u}'_{2,t-j}\mathbf{b}_j + e_t, \qquad \mathbf{u}_{2,t} = \Delta\mathbf{y}_{2,t},$$

  and estimates the modified co-integrating regression:

$$y_{1,t} = c + \mathbf{a}'\mathbf{y}_{2,t} + \sum_{j=-k}^{k} \Delta\mathbf{y}'_{2,t-j}\mathbf{b}_j + e_t.$$

  The resulting estimates are asymptotically efficient in the sense of Saikkonen (1991, Definition 2.2).

For $\mathbf{u}_t = \Delta \mathbf{y}_t$, let

$$\mathbf{\Sigma} = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\left(\sum_{t=1}^{T} \mathbf{u}_t \right) \left(\sum_{t=1}^{T} \mathbf{u}_t \right)' \right] = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

Denote the "correlation" between $y_{1,t}$ and $\mathbf{y}_{2,t}$ as:

$$\rho^2 = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}/\mathbf{\Sigma}_{11},$$

assuming that $\mathbf{\Sigma}_{22}$ is p.d. (i.e., the elements of $\mathbf{y}_{2,t}$ are not co-integrated).

- If $y_{1,t}$ and $\mathbf{y}_{2,t}$ are co-integrated, then it is necessary that $\rho^2 = 1$. This suggests that the choice of the dependent variable in a co-integrating regression does not matter asymptotically.
- If $\rho^2 < 1$, $y_{1,t}$ and $\mathbf{y}_{2,t}$ are not co-integrated, and the regression of $y_{1,t}$ on $\mathbf{y}_{2,t}$ is spurious in the sense of Granger & Newbold (1974).

To see this, write

$$\boldsymbol{\Sigma} = \mathbf{L}'\mathbf{L} = \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{bmatrix}' \begin{bmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{bmatrix},$$

where $\mathbf{L}_{22} = \boldsymbol{\Sigma}_{22}^{1/2}$, $\mathbf{L}_{21} = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}$, and

$$\mathbf{L}_{11} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{1/2} =: \boldsymbol{\Sigma}_{11.2}^{1/2}.$$

Clearly, $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11}(1 - \rho^2)$. As

$$\det(\boldsymbol{\Sigma}) = (\det(\mathbf{L}))^2 = \mathbf{L}_{11}^2(\det(\mathbf{L}_{22}))^2 = \boldsymbol{\Sigma}_{11.2}\det(\boldsymbol{\Sigma}_{22}),$$

$\det(\boldsymbol{\Sigma}) = 0$ if and only if $\boldsymbol{\Sigma}_{11.2} = 0$, or equivalently, $\rho^2 = 1$.

Tests for Co-Integration

Engle and Granger (1987) suggest testing the null of no co-integration by applying unit-root tests to the residuals $\hat{\zeta}_t$ of co-integrating regression. That is, test $\rho_0 = 0$ in the regressions below:

$$\Delta\hat{\zeta}_t = \hat{\rho}_T \hat{\zeta}_{t-1} + \hat{e}_t,$$

$$\Delta\hat{\zeta}_t = \hat{\rho}_T \hat{\zeta}_{t-1} + \sum_{j=1}^{p} \hat{\phi}_{jT}\, \Delta\hat{\zeta}_{t-j} + \hat{e}_t;$$

other DF-type models can also be used.

Remark: The OLS residuals of a co-integrating regression are results of a minimization problem and hence behave like a stationary series. Thus, the DF critical values should not be used; see Engle and Granger (1987) and Engle and Yoo (1987) for proper critical values.

Similarly, by performing the auxiliary regression: $\hat{\zeta}_t = \hat{\rho}_T \hat{\zeta}_{t-1} + \hat{e}_t$, we can apply the Z-type tests for unit root:

$$Z(\hat{\rho}_T) = T(\hat{\rho}_T - 1) + \frac{\frac{1}{2}(s_{Tn}^2 - s_e^2)}{T^{-2}\sum_{t=1}^{T}\hat{\zeta}_{t-1}^2},$$

$$Z(\tau_\rho) = \left(\sum_{t=1}^{T}\hat{\zeta}_{t-1}^2\right)^{1/2}\frac{\hat{\rho}_T - 1}{s_{Tn}} + \frac{\frac{1}{2}(s_{Tn}^2 - s_e^2)}{s_{Tn}(T^{-2}\sum_{t=1}^{T}\hat{\zeta}_{t-1}^2)^{1/2}},$$

where $s_e^2 = T^{-1}\sum_{t=1}^{T}\hat{e}_t^2$ and

$$s_{Tn}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2 + \frac{2}{T}\sum_{\tau=1}^{n} w_{\tau n}\sum_{t=\tau+1}^{T}\hat{e}_t\hat{e}_{t-\tau}.$$

Empirical critical values of these tests are in Phillips & Ouliaris (1990).

Phillips & Ouliaris (1990):

- Under the alternative hypothesis of co-integration, the ADF τ-test and $Z(\tau_\rho)$ are $O_p(T^{1/2})$, whereas $Z(\hat{\rho}_T)$ is $O_p(T)$. That is, t-type tests diverge slower under the alternative than other statistics.
- If $\Delta\hat{\zeta}_t$ are used to compute s^2_{Tn} in Z-type of tests, the resulting test statistics are $O_p(1)$ and hence inconsistent, under the alternative hypothesis. Therefore, residuals, rather than differences, should be used to construct these test statistics.
- If the null hypothesis is co-integration, we need to test whether Σ is singular. They show that tests in this direction are inconsistent and dependent on data; hence testing co-integration is not recommended.

Fully-Modified Estimation

Digression: Canonical Correlation

Let $\mathbf{y}_t$ $(n \times 1)$ and $\mathbf{x}_t$ $(m \times 1)$ be two stationary variables with zero mean and covariance matrices $\boldsymbol{\Sigma}_{\mathbf{yy}}$ and $\boldsymbol{\Sigma}_{\mathbf{xx}}$. Also, $\boldsymbol{\Sigma}_{\mathbf{yx}} = \mathbb{E}(\mathbf{y}_t \mathbf{x}_t') = \boldsymbol{\Sigma}_{\mathbf{xy}}'$. For $\mathbf{A}$ $(n \times k)$ and $\mathbf{B}$ $(m \times k)$ with $k = \min(n, m)$, consider

$$\eta_t = \mathbf{A}'\mathbf{y}_t, \qquad \xi_t = \mathbf{B}'\mathbf{x}_t.$$

$\mathbf{A}$ and $\mathbf{B}$ are such that

$$\mathbf{A}'\boldsymbol{\Sigma}_{\mathbf{yy}}\mathbf{A} = \mathbf{I}_k, \quad \mathbf{B}'\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{B} = \mathbf{I}_k, \quad \mathbf{A}'\boldsymbol{\Sigma}_{\mathbf{yx}}\mathbf{B} = \mathbf{R},$$

where $\mathbf{R}$ is a diagonal matrix with $r_i = \mathrm{corr}(\boldsymbol{\eta}_{i,t}, \boldsymbol{\xi}_{i,t})$ being the diagonal elements. Without loss of generality, the elements of $\mathbf{y}$ and $\mathbf{x}$ are arranged such that r_i in $\mathbf{R}$ are in descending order. Note that $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$ are known as canonical covariates and that r_i is the i th canonical correlation.

For $\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}$, let λ_i, $i = 1, \ldots, n$, denote its eigenvalues in descending order, and $\mathbf{a}_i$ the corresponding eigenvectors, i.e.,

$$\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_i = \lambda_i \mathbf{a}_i.$$

We normalize a_i (in the metric of $\mathbf{\Sigma}_{\mathbf{yy}}$) such that $\mathbf{a}_i'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{a}_i = 1$, instead of $\mathbf{a}_i'\mathbf{a}_i = 1$. Note that the eigenvalue problem is to solve

$$\det\big(\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}} - \lambda \mathbf{I}\big) = \det\big(\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\big)\det\big(\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}} - \lambda \mathbf{\Sigma}_{\mathbf{yy}}\big) = 0,$$

which is equivalent to solving $\det\big(\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}} - \lambda \mathbf{\Sigma}_{\mathbf{yy}}\big) = 0$.

Similarly, for $\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}$, let ν_i, $i = 1, \ldots, n$, denote its eigenvalues in descending order, and $\mathbf{b}_i$ the corresponding eigenvectors, normalized as $\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xx}}\mathbf{b}_i = 1$.

For $\mathbf{A} = [\mathbf{a}_1 \ \ldots \ \mathbf{a}_n]$ and $\mathbf{B} = [\mathbf{b}_1 \ \ldots \ \mathbf{b}_n]$, we have $0 \leq \lambda_i = \nu_i < 1$ and

$$\mathbf{A}'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{A} = \mathbf{I}_k, \quad \mathbf{B}'\mathbf{\Sigma}_{\mathbf{xx}}\mathbf{B} = \mathbf{I}_k, \quad \mathbf{A}'\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{B} = \mathbf{R},$$

with $\mathbf{R}^2 = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is the diagonal matrix with λ_i as the diagonal elements. Thus, λ_i is the squared canonical correlations r_i^2.

Proof: To see $\lambda_i = \nu_i$, note that

$$\mathbf{\Sigma}_{\mathbf{xy}}\big(\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\big)\mathbf{a}_i = \lambda_i\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_i.$$

This shows that λ_i are also eigenvalues of $\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}$ with eigenvectors $\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_i$.

To see $\mathbf{A}'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{A} = \mathbf{I}_k$, note that

$$\mathbf{a}_j'\mathbf{\Sigma}_{\mathbf{yy}}\big(\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\big)\mathbf{a}_i = \lambda_i\mathbf{a}_j'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{a}_i.$$

Also, $\mathbf{a}_i'\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j = \lambda_j\mathbf{a}_i'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{a}_j$.
These indicate that $(\lambda_i - \lambda_j)\mathbf{a}_j'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{a}_i = 0$. For $i \neq j$, $\mathbf{a}_j'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{a}_i = 0$, so that $\mathbf{a}_i$ and $\mathbf{a}_j$ are orthogonal in the metric of $\mathbf{\Sigma}_{\mathbf{yy}}$. For $i = j$, the normalization gives $\mathbf{a}_i'\mathbf{\Sigma}_{\mathbf{yy}}\mathbf{a}_i = 1$.

To show $\mathbf{A}'\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{B} = \mathbf{R}$ with $\mathbf{R}^2 = \mathbf{\Lambda}$, note

$$\big(\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\big)\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j = \lambda_i\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j,$$

$$\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xy}}\big(\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j\big) = \lambda_j\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j.$$

Thus, $(\lambda_i - \lambda_j)\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j = 0$, implying $\mathbf{b}_i'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{a}_j = 0$ for $i \neq j$.

For $i = j$,

$$\mathbf{b}_i' \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \boldsymbol{\Sigma}_{\mathbf{yx}} \mathbf{b}_i = \mathbf{b}_i' \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \boldsymbol{\Sigma}_{\mathbf{yx}} (\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \boldsymbol{\Sigma}_{\mathbf{xx}}) \mathbf{b}_i = \lambda_i (\mathbf{b}_i' \boldsymbol{\Sigma}_{\mathbf{xx}} \mathbf{b}_i) = \lambda_i.$$

Suppose $k = n$, $\mathbf{A}$ is nonsingular, so that $\mathbf{A}' \boldsymbol{\Sigma}_{\mathbf{yy}} \mathbf{A} = \mathbf{I}_k$ implies $\boldsymbol{\Sigma}_{\mathbf{yy}} = (\mathbf{A}')^{-1} \mathbf{A}^{-1}$. We thus have

$$\mathbf{b}_i' \boldsymbol{\Sigma}_{\mathbf{xy}} (\mathbf{A}\mathbf{A}') \boldsymbol{\Sigma}_{\mathbf{yx}} \mathbf{b}_i = \lambda_i.$$

As $\mathbf{b}_i' \boldsymbol{\Sigma}_{\mathbf{xy}} \mathbf{a}_j = 0$ for $i \neq j$,

$$r_i^2 = (\mathbf{b}_i' \boldsymbol{\Sigma}_{\mathbf{xy}} \mathbf{a}_i)^2 = \lambda_i.$$

Johansen's Maximum Likelihood Procedure

$\mathbf{y}_t$ $(d \times 1)$ is $CI(1,1)$ with co-integrating rank r and has a VAR(p) representation: $\mathbf{\Psi}(\mathcal{B})\mathbf{y}_t = \boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t$ are i.i.d. $N(\mathbf{0}, \mathbf{S})$. Writing

$$\begin{aligned}\mathbf{y}_t = {} & \mathbf{\Psi}_1(\Delta\mathbf{y}_{t-1} + \cdots + \Delta\mathbf{y}_{t-p+1} + \mathbf{y}_{t-p}) + \\ & \qquad \mathbf{\Psi}_2(\Delta\mathbf{y}_{t-2} + \cdots + \Delta\mathbf{y}_{t-p+1} + \mathbf{y}_{t-p}) + \ \cdots \ + \mathbf{\Psi}_p\,\mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t,\end{aligned}$$

we then have

$$\Delta\mathbf{y}_t = \mathbf{\Pi}_1\,\Delta\mathbf{y}_{t-1} + \mathbf{\Pi}_2\,\Delta\mathbf{y}_{t-2} + \cdots + \mathbf{\Pi}_{p-1}\,\Delta\mathbf{y}_{t-p+1} + \mathbf{\Pi}_p\mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{\Pi}_i = -\mathbf{I} + \mathbf{\Psi}_1 + \ldots + \mathbf{\Psi}_i$, $i = 1, \ldots, p$. Note that $-\mathbf{\Pi}_p = \mathbf{\Psi}(1) = \mathbf{\Gamma}\mathbf{A}'$ is subject to the restriction of singularity, but $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_{p-1}$, $\mathbf{\Gamma}$, $\mathbf{A}$, and $\mathbf{S}$ are not.

Given $\mathbf{\Gamma}$ and $\mathbf{A}$, the parameters $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_{p-1}$ can be estimated from the regression of $\Delta \mathbf{y}_t + \mathbf{\Gamma}\mathbf{A}'\mathbf{y}_{t-p}$ on $\Delta \mathbf{y}_{t-1}, \ldots, \Delta \mathbf{y}_{t-p+1}$. The residuals of this regression are $\mathbf{r}_t^*$ and can be expressed as $\mathbf{r}_{0t} + \mathbf{\Gamma}\mathbf{A}'\mathbf{r}_{pt}$.

- Let J_0 denote the regression of of $\Delta \mathbf{y}_t$ on $\Delta \mathbf{y}_{t-1}, \ldots, \Delta \mathbf{y}_{t-p+1}$ with the residuals $\mathbf{r}_{0t}$.
- Let J_p be the regression of $\mathbf{y}_{t-p}$ on $\Delta \mathbf{y}_{t-1}, \ldots, \Delta \mathbf{y}_{t-p+1}$ with the residuals $\mathbf{r}_{pt}$.

Given $\mathbf{\Gamma}$, $\mathbf{A}$, and $\mathbf{S}$, the concentrated (Gaussian) likelihood is

$$L_T(\mathbf{\Gamma}, \mathbf{A}, \mathbf{S}) = \det(\mathbf{S})^{-T/2}$$

$$\exp\left(-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{r}_{0t} + \mathbf{\Gamma}\mathbf{A}'\mathbf{r}_{pt})'\mathbf{S}^{-1}(\mathbf{r}_{0t} + \mathbf{\Gamma}\mathbf{A}'\mathbf{r}_{pt})\right).$$

For each $\mathbf{A}$, we can maximize over $\boldsymbol{\Gamma}$ and $\mathbf{S}$ and obtain

$$\begin{aligned}
\boldsymbol{\Gamma}(\mathbf{A}) &= -\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{0t}\mathbf{r}_{pt}'\mathbf{A}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{A}'\mathbf{r}_{pt}\mathbf{r}_{pt}'\mathbf{A}\right)^{-1} \\
&= -\mathbf{M}_{0p}\mathbf{A}(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A})^{-1}, \\
\mathbf{S}(\mathbf{A}) &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{0t}\mathbf{r}_{0t}'\right) - \\
&\qquad \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{0t}\mathbf{r}_{pt}'\mathbf{A}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{A}'\mathbf{r}_{pt}\mathbf{r}_{pt}'\mathbf{A}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{A}'\mathbf{r}_{pt}\mathbf{r}_{0t}'\right) \\
&= \mathbf{M}_{00} - \mathbf{M}_{0p}\mathbf{A}(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A})^{-1}\mathbf{A}'\mathbf{M}_{p0},
\end{aligned}$$

where $\mathbf{M}_{ij} = T^{-1}\sum_{t=1}^{T}\mathbf{r}_{it}\mathbf{r}_{jt}'$ for $i, j = 0, p$.

Substituting $\mathbf{\Gamma}(\mathbf{A})$ and $\mathbf{S}(\mathbf{A})$ into $L_T(\mathbf{\Gamma}, \mathbf{A}, \mathbf{S})$, the concentrated likelihood is now proportional to $\det(\mathbf{S}(\mathbf{A}))^{-T/2}$. Note that

$$\begin{aligned}
&\det\left(\begin{bmatrix} \mathbf{M}_{00} & \mathbf{M}_{0p}\mathbf{A} \\ \mathbf{A}'\mathbf{M}_{p0} & \mathbf{A}'\mathbf{M}_{pp}\mathbf{A} \end{bmatrix}\right) \\
&\quad = \det(\mathbf{M}_{00}) \det(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A} - \mathbf{A}'\mathbf{M}_{p0}\mathbf{M}_{00}^{-1}\mathbf{M}_{0p}\mathbf{A}) \\
&\quad = \det(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A}) \det(\mathbf{M}_{00} - \mathbf{M}_{0p}\mathbf{A}(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A})^{-1}\mathbf{A}'\mathbf{M}_{p0}).
\end{aligned}$$

It follows that

$$\det(\mathbf{S}(\mathbf{A})) = \det(\mathbf{M}_{00}) \frac{\det(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A} - \mathbf{A}'\mathbf{M}_{p0}\mathbf{M}_{00}^{-1}\mathbf{M}_{0p}\mathbf{A})}{\det(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A})}.$$

We would like to minimize the 2nd term on the right with respect to $\mathbf{A}$.

Let $\mathbf{\Lambda}$ denote the matrix of ordered eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$ of $\mathbf{M}_{p0}\mathbf{M}_{00}^{-1}\mathbf{M}_{0p}$ in the metric of $\mathbf{M}_{pp}$, i.e., the solutions of

$$\det(\lambda \mathbf{M}_{pp} - \mathbf{M}_{p0}\mathbf{M}_{00}^{-1}\mathbf{M}_{0p}) = 0.$$

Also let $\mathbf{C}$ denote the matrix of corresponding eigenvectors normalized as $\mathbf{C}'\mathbf{M}_{pp}\mathbf{C} = \mathbf{I}$. Then,

$$\mathbf{\Lambda} = \mathbf{C}'\mathbf{M}_{p0}\mathbf{M}_{00}^{-1}\mathbf{M}_{0p}\mathbf{C}.$$

By setting $\mathbf{A} = \mathbf{CU}$ ($\mathbf{U}$ is $d \times r$),

$$\frac{\det(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A} - \mathbf{A}'\mathbf{M}_{p0}\mathbf{M}_{00}^{-1}\mathbf{M}_{0p}\mathbf{A})}{\det(\mathbf{A}'\mathbf{M}_{pp}\mathbf{A})} = \frac{\det(\mathbf{U}'\mathbf{U} - \mathbf{U}'\mathbf{\Lambda}\mathbf{U})}{\det(\mathbf{U}'\mathbf{U})},$$

which should be minimized with respect to $\mathbf{U}$.

Analogous to Raleigh's quotient, $\det(\mathbf{U}'\mathbf{U} - \mathbf{U}'\mathbf{\Lambda}\mathbf{U})/\det(\mathbf{U}'\mathbf{U})$ is minimized when $\mathbf{U}$ is the matrix of the first r Cartesian unit vectors. As such, $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{U}'\mathbf{\Lambda}\mathbf{U}$ is the diagonal matrix with r largest eigenvalues on the principal diagonal. The resulting minimum is

$$\prod_{i=1}^{r}(1 - \lambda_i).$$

The MLE $\widehat{\mathbf{A}} = \mathbf{C}\mathbf{U}$ is the matrix of the first r eigenvectors in $\mathbf{C}$ and is the coefficient matrix of the first r canonical covariates of $\mathbf{r}_p$ with respect to $\mathbf{r}_0$, corresponding to the eigenvalues λ_i, $i = 1, \ldots, r$. These eigenvalues are squares of the r largest canonical correlations.

Estimation of Co-Integrating Space

The MLE of span[$\mathbf{A}$] is the space spanned by r canonical covariates corresponding to the r largest squared canonical correlations between the residuals from the regressions J_p and J_0.

Clearly, $\mathbf{A}'\mathbf{M}_{pp}\mathbf{A} = \mathbf{U}'\mathbf{C}'\mathbf{M}_{pp}\mathbf{C}\mathbf{U} = \mathbf{I}$. $\boldsymbol{\Gamma}$ and $\mathbf{S}$ can be estimated as

$$\hat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}(\widehat{\mathbf{A}}) = -\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{0t}\mathbf{r}_{pt}'\right)\widehat{\mathbf{A}},$$

$$\hat{\mathbf{S}} = \mathbf{S}(\widehat{\mathbf{A}}) = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{0t}\mathbf{r}_{0t}'\right) - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}',$$

and $\widehat{\boldsymbol{\Psi}}(1) = \widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{A}}'$. This analysis is closely related to the "reduced rank regression" of Ahn & Reinsel (1987).

Likelihood Ratio Test

- When there are r CIVs and only $r' < r$ CIVs are used in the ECM, relevant equilibrium error terms are left out in the model.
- If $r' > r$ "CIVs" are used in the ECM, the model in effect contains linear combinations of $\mathbf{y}_t$ that are still $I(1)$. Conventional inference would be invalid in this case.

The maximum of the concentrated likelihood is $L_T(\widehat{\mathbf{A}})$ such that

$$L_T(\widehat{\mathbf{A}})^{-2/T} = \det(\widehat{\mathbf{S}}) = \det(\mathbf{M}_{00}) \prod_{i=1}^{r} (1 - \lambda_i).$$

Without the constraint that $\mathrm{rank}\,\mathbf{\Psi}(1) \leq r$, the unconstrained likelihood is

$$L_T(\widetilde{\mathbf{A}})^{-2/T} = \det(\mathbf{M}_{00}) \prod_{i=1}^{d} (1 - \lambda_i).$$

Null hypothesis: There are at most r CIVs. The LR test is

$$\mathcal{LR}_T(r) = -2(\log L_T(\widehat{\mathbf{A}}) - \log L_T(\widetilde{\mathbf{A}})) = -T \sum_{i=r+1}^{d} \log(1-\lambda_i).$$

That is, we test whether the last $d-r$ eigenvalues (squared canonical correlations) are sufficiently close to zero simultaneously.

Likelihood Ratio Test: Johansen (1988)

Under the null, the estimates of $\mathbf{A}$, $\mathbf{S}$, and $\mathbf{\Psi}(1)$ are consistent, and

$$\mathcal{LR}_T(r) \xrightarrow{D}$$
$$\operatorname{trace}\left(\int_0^1 \mathrm{d}\mathbf{w}(s)\mathbf{w}(s)' \left(\int_0^1 \mathbf{w}(s)\mathbf{w}(s)'\,\mathrm{d}s\right)^{-1} \int_0^1 \mathbf{w}(s)\ \mathrm{d}\mathbf{w}(s)'\right),$$

where $\mathbf{w}$ is a $(d-r)$-dimensional standard Wiener process.

Remarks:

- If $r = d - 1$, then the limit becomes

$$\frac{\left(\int_0^1 \mathbf{w}(r)\, \mathrm{d}\mathbf{w}(r)\right)^2}{\int_0^1 \mathbf{w}(r)^2 \,\mathrm{d}r} = \frac{[\frac{1}{2}(\mathbf{w}(1)^2 - 1)]^2}{\int_0^1 \mathbf{w}(r)^2 \,\mathrm{d}r},$$

  which is the square of the limit of Dickey-Fuller τ test.

- Johansen (1988) tabulates the empirical distribution of the LR statistic for $d - r = 1, \ldots, 5$. This distribution can be approximated by $c\chi^2(q)$ for suitable values of c and q; setting $q = 2(d-r)^2$, Johansen suggests using $c = 0.85 - 0.58/q$.

- Johansen (1991) allows a constant term and seasonal dummies in the VAR model. The corresponding empirical distribution is tabulated in Johansen & Juselius (1990); see also Osterwald-Lenum (1992).

Remarks (Cont'd):

- In practice, one may sequentially perform LR tests:

$$\mathcal{LR}_T(1), \ \ldots, \ \mathcal{LR}_T(d-1).$$

  That is, we first test at most one co-integrating relation, at most 2 co-integrating relations, and so on. Note that we need to control the correct significance level of this sequential testing procedure.

- Based on the same idea, one may construct a test of r co-integrating relationships against the alternative of $r+1$ co-integrating relations using the statistic: $-T\log(1-\lambda_{r+1})$.

Summary of Johansen's procedure

1. Perform regressions J_0 and J_p to obtain residuals $\mathbf{r}_{0t}$ and $\mathbf{r}_{pt}$ and compute cross-moment matrices $\mathbf{M}_{00}$, $\mathbf{M}_{0p}$, and $\mathbf{M}_{pp}$ based on $\mathbf{r}_{0t}$ and $\mathbf{r}_{pt}$.
2. Find the coefficient matrix of r canonical covariates of $\mathbf{r}_p$ with respect to $\mathbf{r}_0$, corresponding to the r largest squared canonical correlations. This gives $\widehat{\mathbf{A}}$, the estimates of (the space of) CIVs, from which we obtain $\widehat{\mathbf{\Gamma}}$, $\widehat{\mathbf{S}}$, and $\widehat{\mathbf{\Psi}}(1)$.
3. Compute LR statistic: $-T\sum_{i=r+1}^{d}\log(1-\lambda_i)$ to check if there are at most r CIVs, where λ_i is the i th squared canonical correlation.
4. Regress $\Delta\mathbf{y}_t + \widehat{\mathbf{\Gamma}}\widehat{\mathbf{A}}'\mathbf{y}_{t-p}$ on $\Delta\mathbf{y}_{t-1},\ldots\Delta\mathbf{y}_{t-p+1}$ to get estimates $\widehat{\mathbf{\Pi}}_1,\ldots,\widehat{\mathbf{\Pi}}_{p-1}$ in the ECM system.

Some Stylized Facts

- Financial time series usually exhibit volatility clustering, in the sense that large (small) changes are followed by large (small) changes, in either sign.
- A financial time series may have rather weak serial correlations, but a function of this series (e.g., taking square or absolute value) may exhibit much stronger correlations.
- The number of outliers of these variables are more than what a normal distribution can describe. That is, the marginal distributions have thicker tails than a normal distribution.
- Volatility asymmetry and changing volatility patterns are also quite common in practice.

ARCH Models

The autoregressive conditional heteroskedasticity (ARCH) model of Engle (1982):

- ARCH(1): $y_t = \sqrt{h_t}\, u_t$, where u_t are i.i.d. with mean zero and variance one, and

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2, \quad \alpha_0 > 0,\ \alpha_1 \geq 0.$$

- The conditional mean of y_t is

$$\mathbb{E}(y_t \mid \mathcal{F}^{t-1}) = \sqrt{h_t}\, \mathbb{E}(u_t \mid \mathcal{F}^{t-1}) = \sqrt{h_t}\, \mathbb{E}(u_t) = 0,$$

  and the conditional variance is

$$\mathbb{E}(y_t^2 \mid \mathcal{F}^{t-1}) = h_t\, \mathbb{E}(u_t^2 \mid \mathcal{F}^{t-1}) = h_t\, \mathbb{E}(u_t^2) = h_t.$$

- Note that y_t is a white noise with $\mathbb{E}(y_t) = \mathbb{E}[\mathbb{E}(y_t \mid \mathcal{F}^{t-1})] = 0$,

$$\operatorname{var}(y_t) = \mathbb{E}(h_t) = \alpha_0 + \alpha_1 \operatorname{var}(y_{t-1}) = \alpha_0/(1-\alpha_1),$$

  and

$$\mathbb{E}(y_t y_{t-j}) = \mathbb{E}\left[\sqrt{h_t h_{t-j}}\, u_{t-j}\, \mathbb{E}(u_t \mid \mathcal{F}^{t-1})\right] = 0, \quad j = 1, 2, \ldots .$$

- Yet, y_t^2 are serially correlated with the AR(1) representation:

$$y_t^2 = h_t + (y_t^2 - h_t) = \alpha_0 + \alpha_1 y_{t-1}^2 + h_t(u_t^2 - 1),$$

  where $h_t(u_t^2 - 1)$ are innovations with $\mathbb{E}[h_t(u_t^2 - 1)] = 0$ and

$$\mathbb{E}[h_t h_{t-j}(u_t^2 - 1)(u_{t-j}^2 - 1)] = \mathbb{E}(h_t h_{t-j})\, \mathbb{E}(u_t^2 - 1)\, \mathbb{E}(u_{t-j}^2 - 1) = 0.$$

- Under conditional normality, $\mathbb{E}(y_t^4 \mid \mathcal{F}^{t-1}) = 3h_t^2$, and

$$\begin{aligned} m_4 &= 3\big[\alpha_0^2 + 2\alpha_0\alpha_1\,\mathbb{E}(h_t) + \alpha_1^2\,\mathbb{E}(y_{t-1}^4)\big] \\ &= 3\alpha_0^2\Big(1 + \frac{2\alpha_1}{1-\alpha_1}\Big) + 3\alpha_1^2 m_4 \\ &= \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}, \end{aligned}$$

  where we write $m_4 := \mathbb{E}(y_t^4)$. Thus, $0 \leq \alpha_1^2 < 1/3$.

- The kurtosis coefficient of y_t is

$$\frac{m_4}{\text{var}(y_t)^2} = 3\,\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3,$$

  and the marginal distribution of y_t has thicker tails than a normal distribution.

- ARCH(p): $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2, \quad \alpha_0 > 0,\ \alpha_1, \ldots, \alpha_p \geq 0.$$

- AR(p) representation of y_t^2:

$$y_t^2 = h_t + (y_t^2 - h_t) = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2 + h_t(u_t^2 - 1).$$

- $\{y_t\}$ is a white noise with $\mathrm{var}(y_t) = \alpha_0/(1 - \alpha_1 - \cdots - \alpha_p)$.
- AR(p_1)-ARCH(p_2): $y_t = c + \psi_1 y_{t-1} + \cdots + \psi_{p_1} y_{t-p_1} + \varepsilon_t$, where $\varepsilon_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_{p_2} \varepsilon_{t-p_2}^2, \quad \alpha_0 > 0,\ \alpha_1, \ldots, \alpha_{p_2} \geq 0.$$

GARCH Models

The generalized ARCH (GARCH) model of Bollerslev (1986):

- GARCH(1,1): $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}, \quad \alpha_0 > 0,\ \alpha_1, \beta_1 \geq 0.$$

- ARMA(1,1) representation of y_t^2:

$$\begin{aligned} y_t^2 &= h_t + (y_t^2 - h_t) \\ &= \alpha_0 + (\alpha_1 + \beta_1) y_{t-1}^2 + h_t(u_t^2 - 1) - \beta_1 h_{t-1}(u_{t-1}^2 - 1), \end{aligned}$$

with serially uncorrelated innovations $h_t(u_t^2 - 1)$.

- y_t have mean zero and

$$\begin{aligned}\operatorname{var}(y_t) = \mathbb{E}(h_t) &= \alpha_0 + \alpha_1\,\mathbb{E}(y_{t-1}^2) + \beta_1\,\mathbb{E}(h_{t-1}) \\ &= \frac{\alpha_0}{1-(\alpha_1+\beta_1)}.\end{aligned}$$

  Thus, $\alpha_1 + \beta_1$ must be less than one to ensure a finite variance.

- The autocovariances of y_t and y_{t-j}, $j = 1, 2, \ldots$, are also zero, so that $\{y_t\}$ is still a white noise.
- The kurtosis coefficient is, under conditional normality,

$$\frac{m_4}{\operatorname{var}(y_t)^2} = 3\,\frac{1-(\alpha_1+\beta_1)^2}{1-(\alpha_1+\beta_1)^2-2\alpha_1^2} > 3,$$

  provided that $1-(\alpha_1+\beta_1)^2-2\alpha_1^2 > 0$.

- GARCH(p, q): $y_t = \sqrt{h_t}\, u_t$, with the conditional variance:

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}, \quad \alpha_0 > 0,\ \alpha_i, \beta_j \geq 0.$$

- ARMA representation of y_t^2:

$$y_t^2 = \alpha_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) y_{t-i}^2 + h_t(u_t^2 - 1) - \sum_{j=1}^{q} \beta_j h_{t-j}(u_{t-j}^2 - 1),$$

where we set $\alpha_i = 0$ if $i > p$ and $\beta_i = 0$ if $i > q$.

- We also have $\mathbb{E}(y_t) = 0$,

$$\mathrm{var}(y_t) = \frac{\alpha_0}{1 - \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i)},$$

and zero autocovariances.

- AR(p_1)-GARCH(p_2, q): $y_t = c + \psi_1 y_{t-1} + \cdots + \psi_{p_1} y_{t-p_1} + \varepsilon_t$, where $\varepsilon_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \sum_{i=1}^{p_2} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}, \qquad \alpha_0 > 0,\ \alpha_i, \beta_j \geq 0.$$

  Extension to AR(p_1, q_1)-GARCH(p_2, q_2) is also possible.

- For GARCH(1,1), it is quite common to observe that the sum of the estimated α_1 and β_1 is close to one.

- Integrated GARCH (IGARCH): $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + (1 - \beta_1) y_{t-1}^2 + \beta_1 h_{t-1}, \qquad \alpha_0 > 0,\ 0 < \beta_1 < 1.$$

  In this case, var(y_t) is unbounded, and y_t^2 has a unit root.

GARCH-in-Mean Models

GARCH-in-Mean (GARCH-M) model of Engle, Lilien, and Robins (1987): By noting that asset returns may also depend on their volatility, they propose

$$y_t = c + \gamma h_t + \varepsilon_t,$$

with $\varepsilon_t = \sqrt{h_t}\, u_t$ and

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}, \quad \alpha_0 > 0,\ \alpha_1, \beta_1 \geq 0,$$

where γ is called the risk premium parameter.

Drawbacks of GARCH Models

- GARCH models are unable to represent volatility asymmetry, because the positive and negative values of the lagged innovations exert the same effect on the conditional variance. Black (1976) observed that the volatility of stock returns tends to increase (decrease) when there is "bad news" ("good news").
- To ensure positiveness of h_t in the GARCH model, non-negative constraints are imposed on the coefficients in the variance equation. These constraints are convenient, yet they are not necessary.

EGARCH Models

- Weighted innovations:

$$g(u_t) = \theta_1 u_t + \gamma_1(|u_t| - \mathbb{E}\,|u_t|),$$

  where $|u_t| - \mathbb{E}\,|u_t|$ are also i.i.d. random variables with mean zero, so that $g(u_t)$ have mean zero. When u_t are normally distributed, for example, we have $\mathbb{E}\,|u_t| = \sqrt{2/\pi}$.

- $g(u_t)$ can be represented as a threshold function:

$$g(u_t) = \begin{cases} (\theta_1 + \gamma_1)u_t - \gamma_1\,\mathbb{E}\,|u_t|, & u_t \geq 0, \\ (\theta_1 - \gamma_1)u_t - \gamma_1\,\mathbb{E}\,|u_t|, & u_t < 0. \end{cases}$$

  It should be noted that the asymmetric response of g to u_t is due to θ_1, rather than γ_1. (Why?)

- Exponential GARCH (EGARCH) model of Nelson (1992): h_t is an exponential function of lagged h_t and the weighted innovation $g(u_{t-1})$. An EGARCH(1,1) process is $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \exp\left[\alpha_0 + \beta_1 \ln(h_{t-1}) + \left(\theta_1 \frac{y_{t-1}}{\sqrt{h_{t-1}}} + \gamma_1 \left|\frac{y_{t-1}}{\sqrt{h_{t-1}}}\right|\right)\right].$$

- θ_1 is usually interpreted as a measure of the "leverage" effect of u_{t-1}, while γ_1 is interpreted as the "magnitude" effect. The estimate of θ_1 is usually found to be negative, while γ_1 is found to be positive. This shows that positive shocks have less impact on volatility.
- Due to exponential function, an innovation with larger magnitude has much larger impact on h_t. Moreover, there is no constraint on the coefficients in h_t.

- News impact curve of Engle and Ng (1993): The relationship between the conditional variance h_t and u_{t-1}, holding constant the information on and before time $t-2$ (lagged conditional variances are evaluated at the unconditional variance). It is easy to see that the news impact curve of a GARCH process is symmetric, but that of an EGARCH process is asymmetric.
- EGARCH(p, q): $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \exp\left[\alpha_0 + \sum_{i=1}^{q} \beta_i \ln(h_{t-i}) + \sum_{j=1}^{p} \left(\theta_j \frac{y_{t-j}}{\sqrt{h_{t-j}}} + \gamma_j \left|\frac{y_{t-j}}{\sqrt{h_{t-j}}}\right|\right)\right],$$

where θ_j and γ_j characterize the asymmetry and magnitude effects of the shock u_{t-j} on the volatility h_t.

GJR-GARCH Models

Focusing on possibly different impacts of positive and negative shocks on conditional variance, Glosten, Jegannathan, and Runkle (1993) propose a threshold-type GARCH model, now known as the GJR-GARCH model.

- GJR-GARCH(1,1): $y_t = \sqrt{h_t}\, u_t$, with

$$h_t = \alpha_0 + \beta_1 h_{t-1} + (\alpha_1 + \theta_1 D_{t-1}) y_{t-1}^2,$$

  where $D_{t-1} = 1$ when $y_{t-1} < 0$ and $D_{t-1} = 0$ otherwise.
- This model is capable of capturing both volatility clustering and volatility asymmetry without imposing the exponential function. Non-negativity constraints on the coefficients in h_t are still needed.
- Extending to AR(p_1)-GJR-GARCH(p_2, q) models is straightforward.

Estimating GARCH Models

- To estimate GARCH models, one must make assumption on the conditional distribution of y_t (or ε_t). Conditional normality results in a leptokurtic marginal distribution, but it can not fully account for the outliers in real data.
- WE may also postulate a $t(\nu)$ distribution which has variance $\nu/(\nu-2)$ when $\nu > 2$. Normalizing y_t to have conditional variance one yields the density:

$$f(u) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{(\nu-2)\pi}} \left(1 + \frac{u^2}{\nu-2}\right)^{-(\nu+1)/2},$$

  where Γ is the Gamma function such that $\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} \, \mathrm{d}y$.
- Other flexible conditional distributions may also be employed.

- We may estimate an ARCH(p) model by maximizing

$$\mathcal{L}_T = -\frac{T-p}{2T}\ln(h_t) - \frac{1}{2T}\sum_{i=p+1}^{T}\frac{y_t^2}{h_t},$$

where $h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2$.

- For an AR(p_1)-ARCH(p_2) model, we maximize

$$\begin{aligned}\mathcal{L}_T = & -\frac{T-p^*}{2T}\ln(h_t) \\ & - \frac{1}{2T}\sum_{i=p^*+1}^{T}\frac{(y_t - c - \psi_1 y_{t-1} - \cdots - \psi_{p_1} y_{t-p_1})^2}{h_t},\end{aligned}$$

where $p^* = \max(p_1, p_2)$ and $h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p_2}^2$.

- We may substitute the t density function for the normal density.

Another commonly used distribution is the generalized error distribution. We may normalize y_t to have conditional mean zero and conditional variance one and obtain the following density of u_t:

$$f(u) = \frac{\nu \exp[-|u/\lambda|^{\nu}/2]}{\lambda\, 2^{1+1/\nu}\, \Gamma(1/\nu)}, \qquad \nu > 0,$$

where ν is the parameter characterizing the thickness of tails and

$$\lambda = \left[2^{-2/\nu}\, \Gamma(1/\nu)\, \Gamma(3/\nu)\right]^{1/2}.$$

It is standard normal when $\nu = 2$; for $\nu < 2$ ($\nu > 2$), the tails of this distribution are thicker (thinner) than the standard normal distribution. For example, it is double exponential when $\nu = 1$ and uniform on $[-\sqrt{3}, \sqrt{3}]$ when $\nu \to \infty$; see Nelson (1991).

Stochastic Volatility Models

A simple stochastic volatility (SV) process is $y_t = \sqrt{h_t}\, u_t$, with

$$\ln(h_t) = \alpha_0 + \alpha_1 \ln(h_{t-1}) + v_t, \qquad |\alpha_1| < 1,$$

where v_t are random variables such that $\{v_t\}$ and $\{u_t\}$ are independent of each other. The inclusion of new innovations v_t admits more flexibility in the model but also renders model estimation much more difficult.

Assume u_t are independent $\mathcal{N}(0,1)$ and v_t are independent $\mathcal{N}(0,\sigma_v^2)$. Then,

$$\ln(h_t) \sim \mathcal{N}\left(\frac{\alpha_0}{1-\alpha_1}, \frac{\sigma_v^2}{1-\alpha_1^2}\right).$$

Clearly, $\mathbb{E}(y_t) = 0$. Knowing the mean and variance of the lognormal random variable, we can calculate

$$\mathbb{E}(y_t^2) = \mathbb{E}(h_t)\,\mathbb{E}(u_t^2) = \exp\left(\frac{\alpha_0}{1-\alpha_1} + \frac{\sigma_v^2}{2(1-\alpha_1^2)}\right),$$

$$\mathbb{E}(y_t^4) = \mathbb{E}(h_t^2)\,\mathbb{E}(u_t^4) = 3\,\exp\left(\frac{2\alpha_0}{1-\alpha_1} + \frac{2\sigma_v^2}{1-\alpha_1^2}\right).$$

Thus, y_t are leptokurtic because

$$\mathbb{E}(y_t^4)/[\mathbb{E}(y_t^2)]^2 = 3\,\exp\left(\frac{\sigma_v^2}{1-\alpha_1^2}\right) > 3.$$

The estimation of an SV model is typically cumbersome; see e.g., Jacquier, Polson, and Rossi (1994), Harvey, Ruiz, and Shephard (1994), and Harvey and Shephard (1996). Let Y^T denote the collection of all y_t and h^T the collection of all conditional variances h_t. Then, the density of Y^T is

$$P(Y^T) = \int P(Y^T, h^T)\,\mathrm{d}\,h^T = \int P(Y^T | h^T)\, P(h^T)\,\mathrm{d}\,h^T,$$

which is a mixture over the density of h^T. Difficulty in estimation arises because a T-dimensional integral must be evaluated. The Markov chain Monte Carlo (MCMC) method suggested by Jacquier, Polson, and Rossi (1994) avoids this difficulty. There are other estimation methods, e.g., the method of quasi-maximum likelihood and the generalized method of moment.

Realized Volatility

A major difficulty in studying volatility (conditional variance) is that it is not observable.

- Without a benchmark, it is difficult to determine the true volatility pattern.
- It is hard to compare the performance of different parametric models.

Note that squares y_t or squared residuals can not serve as benchmark. As such, a model-free estimate of conditional variance (volatility) is highly desirable. The realized volatility (realized variance) proposed by Andersen, Bollerslev, Diebold, and Ebens (2001) is such an estimate.

A standard diffusion model:

$$\mathrm{d}\,p_t = \mu_t \,\mathrm{d}\,t + \sigma_t \,\mathrm{d}\,W_t,$$

where μ_t is the drift term, σ_t is the diffusion, and W is a standard Wiener process. Let $r_{t,m} = p_t - p_{t-m}$; the conditional distribution of $r_{t+1,1}$ is

$$\mathcal{N}\left(\int_0^1 \mu_{t+s}\,\mathrm{d}\,s,\ \int_0^1 \sigma_{t+s}^2\,\mathrm{d}\,s\right).$$

Partition the time between t and $t+1$ into $m = [1/\delta]$ non-overlapping sub-periods, each with the length δ (say, 1 min or 5 mins). For example, $r_{t+1,1}$ is the one-day return and the sum of m δ-period returns:

$$r_{t+1,1} = r_{t+\delta,\delta} + r_{t+2\delta,\delta} + \cdots + r_{t+1,\delta} = \sum_{j=1}^{[1/\delta]} r_{t+j\delta,\delta}.$$

When the partition become finer (i.e., $\delta \to 0$, or $m \to \infty$), the quadratic variations of r are such that

$$\sum_{j=1}^{[1/\delta]} r_{t+j\delta,\delta}^2 \to \int_0^1 \sigma_{t+s}^2 \xrightarrow{\text{a.s.}} s.$$

This suggests that $\sum_{j=1}^{[1/\delta]} r_{t+j\delta,\delta}^2$ serves as a natural estimate of the conditional variance (integrated variance) of $r_{t+1,1}$. For example, the realized daily volatility can be computed as the sum of squared returns of intraday data at a higher frequency (say, squared 5-minute returns). When $\mathbf{r}_{t+1,1}$ are vectors of asset returns, one may also define the "realized variance-covariance matrix" as

$$\sum_{j=1}^{[1/\delta]} (\mathbf{r}_{t+j\delta,\delta})(\mathbf{r}_{t+j\delta,\delta})'.$$