Lecture Notes on Linear Time Series Models

Julian F. Ludwig *
Texas Tech University

April 27, 2024

*Texas Tech University, Department of Economics, P.O. Box 41014, Lubbock, TX 79409-1014, USA.

Contents

1 Introduction

Time series analysis involves predicting future values of a time series based on its past behavior, and two commonly used methods for this are the Autoregressive (AR) model and the Moving Average (MA) model. The AR model predicts future behavior by taking into account past values of the series, while the MA model forecasts future values by considering the errors in predicting the series' past behavior. The ARIMA (Autoregressive Integrated Moving Average) model is a more general framework that combines both the AR and MA models, as well as differencing to remove non-stationarity in the time series data. This allows for the modeling of a wider range of time series patterns.

To fully understand these models, it is essential to have a good understanding of polynomial math, including polynomial functions, their properties, and how they can be used to compute MA and AR representations of ARMA models. In these lecture notes, we will cover these topics in depth.

2 Polynomial Math

Polynomial math is a powerful tool for comprehending time series processes. In this section, I demonstrate the fundamental operations of polynomial mathematics that are necessary for working with ARMA processes in the subsequent sections.

2.1 Geometric Series Formula

The formula for the sum of a geometric series is given by:

$$\sum_{s=a}^{b} x^s = \begin{cases} \frac{x^a - x^{b+1}}{1-x} & \text{if } \ x \neq 1 \\ 1 + b - a & \text{if } \ x = 1 \end{cases}$$

When b approaches infinity, the sum remains finite if $|x| < 1$. This occurs because x^s approaches zero exponentially as s increases, ensuring the series converges even with infinitely

many terms. The formula for the infinite sum of a power series, known as the **geometric series formula**, is given as follows:

$$\sum_{s=a}^{\infty} x^s = \frac{x^a}{1-x} \qquad \text{if } \ |x| < 1$$

Assuming $a = 0$, we have $x^a = 1$ and thus $(1-x)^{-1} = 1 + x + x^2 + x^3 + ...$ In the subsequent sections, we typically utilize the geometric series formula to expand $(1-x)^{-1}$ into an infinite sum. Therefore, we can begin with the right-hand side expression of the geometric series formula and substitute it with the left-hand side expression.

Proof. For $x \neq 1$, we rearrange and simplify the series:

$$\sum_{s=a}^{b} x^s = \sum_{s=a}^{b} x^{s+1} + x^a - x^{b+1} = x\left(\sum_{s=a}^{b} x^s\right) + x^a - x^{b+1}$$

Now, the same sum appears on both the left- and right-hand sides, which implies that we can solve for it to obtain the formula:

$$(1-x)\sum_{s=a}^{b} x^s = x^a - x^{b+1} \quad \Rightarrow \quad \sum_{s=a}^{b} x^s = \frac{x^a - x^{b+1}}{1-x}$$

As $|x| < 1$, $\lim_{s\to\infty} x^s = 0$ justifies omitting x^{b+1} when b approaches infinity. For $x = 1$, the sum becomes:

$$\sum_{s=a}^{b} 1 = \underbrace{1}_{a} + \underbrace{1}_{a+1} + \underbrace{1}_{a+2} + \underbrace{1}_{a+3} + \ldots + \underbrace{1}_{a+(b-a)} = 1 + b - a$$

□

We can also employ the modified formula for the sum of a power series where each term

is multiplied by its corresponding power index:

$$\sum_{s=a}^{b} sx^s = \begin{cases} \frac{x^{1+a}-x^{2+b}}{(1-x)^2} + \frac{ax^a-(b+1)x^{1+b}}{1-x} & \text{if } x \neq 1 \\ a + \frac{(b-a)(b-a+1)}{2} & \text{if } x = 1 \end{cases}$$

This formula is derived by differentiating the earlier geometric series formula with respect to x. When $|x| < 1$, the terms x^{b+2} and x^{b+1} become zero as the series goes to infinity.

Proof. For $x \neq 1$, apply the derivative to the geometric series formula with respect to x and utilize the quotient rule:

$$\begin{aligned} \frac{d}{dx}\left(\sum_{s=a}^{b} x^s\right) &= \frac{d}{dx}\left(\frac{x^a - x^{b+1}}{1-x}\right) \\ \sum_{s=a}^{b} \frac{d}{dx} x^s &= \frac{\left[\frac{d}{dx}\left(x^a - x^{b+1}\right)\right](1-x) - \left(x^a - x^{b+1}\right)\left[\frac{d}{dx}(1-x)\right]}{(1-x)^2} \\ \sum_{s=a}^{b} sx^{s-1} &= \frac{\left[ax^{a-1} - (b+1)x^b\right](1-x) - \left(x^a - x^{b+1}\right)[-1]}{(1-x)^2} \\ \sum_{s=a}^{b} sx^s &= \frac{x^{1+a} - x^{2+b}}{(1-x)^2} + \frac{ax^a - (b+1)x^{1+b}}{1-x} \end{aligned}$$

For $x = 1$, the sum becomes:

$$\begin{aligned} \sum_{s=a}^{b} s &= (a+0) + (a+1) + (a+2) + \cdots + b \\ &= a + 1 + 2 + \cdots + (b-a) = a + \sum_{s=1}^{b-a} s \\ &= a + \frac{(b-a)(b-a+1)}{2} \end{aligned}$$

Here, the last equality uses the formula for the sum of the first $b-a$ integers, a topic that will be explored further when we discuss triangular numbers in Section 2.4. □

2.2 Polynomial, Roots, and Factored Form

A **polynomial of degree** n is defined as follows:

$$f(z) = a_0 + a_1 z^1 + \cdots + a_n z^n$$

The **roots** $\{r_1, \ldots, r_n\}$ of the polynomial are all values of z where the polynomial is equal to zero:

$$f(r) = a_0 + a_1 r^1 + \cdots + a_n r^n = 0, \qquad \forall r \in \{r_1, \ldots, r_n\}$$

We will show below that a polynomial of degree n has exactly n roots. To find the roots, we can write the polynomial in factored form as follows:

$$f(z) = a_n (z - r_1)(z - r_2) \cdots (z - r_n)$$

where $\{r_1, \ldots, r_n\}$ are the roots of the polynomial, because $f(r_i) = 0$, for all i.

For example, consider the polynomial:

$$f(z) = 18 + 15z + 3z^2 = 3(z + 2)(z + 3)$$

which implies that the roots are $\{r_1, r_2\} = \{-2, -3\}$.

To find the factors of a **quadratic polynomial** (i.e., a polynomial of degree $n = 2$), we can consider the general case:

$$\begin{aligned} f(z) &= a_2(x + p)(x + q) \\ &= \underbrace{a_2 pq}_{a_0} + \underbrace{a_2(p + q)}_{a_1} x + \underbrace{a_2}_{a_2} x^2 \end{aligned}$$

Thus, for the example above, we can divide by $a_2 = 3$ to get $f(z) = 6 + 5z + z^2$, and then we need to find p and q such that $pq = 6$ and $(p + q) = 5$, which is the case for $p = 2$ and

$q = 3$, and therefore $\{r_1, r_2\} = \{-p, -q\}$.

Alternatively, we can apply the **quadratic formula** to find the roots:

$$r = \frac{-a_1 \pm \sqrt{(a_1)^2 - 4(a_2)(a_0)}}{2(a_2)}$$

$$= \frac{-15 \pm \sqrt{(15)^2 - 4(3)(18)}}{2(3)} = \frac{-15 \pm \sqrt{9}}{6} = \begin{cases} -12/6 & = -2 \\ -18/6 & = -3 \end{cases}$$

Finally, the roots of a polynomial can also be computed in `R` using the `polyroot` function, which takes the coefficients $\{a_0, a_1, \ldots, a_n\}$ of the polynomial as input and produces the roots $\{r_1, r_2, \ldots, r_n\}$ as output: `{r} polyroot(c(18,15,3)) [1] -2+0i -3+0i`

Now, let's show that a polynomial of degree n has exactly n roots. For this we use the **Fundamental Theorem of Algebra**, which states that every non-constant polynomial has at least one (complex) root, and the **Polynomial Factor Theorem**, which states that if $f(z)$ is a polynomial of degree n, then there exists a complex number $r \in \mathbb{C}$ and a polynomial $g(z)$ of degree $n - 1$ such that:

$$f(z) = (z - r)g(z)$$

where r is a root, because $f(r) = 0$. Note that the root can be a complex number, which is a number that has both a real and an imaginary part, e.g. $r = 3 + 2i$. The imaginary part includes an imaginary unit i, defined as the object that satisfies $i^2 = -1$. When expanding the polynomial $f(z)$, the imaginary units cancel out, so complex numbers can be thought of as a useful tool to simplify mathematical calculations.

If we apply the Polynomial Factor Theorem n times we get

$$\begin{aligned} f(z) &= a_0 + a_1 z^1 + \cdots + a_n z^n \\ &= (z - r_1)\left(b_0 + b_1 z^1 + \cdots + b_{n-1} z^{n-1}\right) \\ &= (z - r_1)(z - r_2)\left(c_0 + c_1 z^1 + \cdots + c_{n-2} z^{n-2}\right) \\ &\vdots \\ &= (z - r_1)(z - r_2)(z - r_3)\cdots(z - r_n)d \end{aligned}$$

where $d = a_n$, because expansion of this polynomials results in a coefficient of d for z^n, which needs to be a_n according to $f(z)$.

Hence, once we derive the roots of the polynomial, we can write the polynomial in factored form (as opposed to its normal form):

$$\text{Normal Form}: \quad f(z) = a_0 + a_1 z^1 + \cdots + a_n z^n = a_0 + \sum_{k=1}^{n} a_k z^k$$

Factored Form :

$$\text{Slope} = 1: \quad f(z) = a_n(z - r_1)\cdots(z - r_n) = a_n \prod_{k=1}^{n}(z - r_k)$$

$$\text{Constant} = 1: \quad f(z) = a_0\left(1 - r_1^{-1} z\right)\cdots\left(1 - r_n^{-1} z\right) = a_0 \prod_{k=1}^{n}\left(1 - r_k^{-1} z\right)$$

The factored form with a normalized constant = 1 can be derived by pre-multiplying the original factored form by $(-r_1)\cdots(-r_n)$, and then using the fact that expansion has to result in an intercept of a_0 so that $(-r_1)\cdots(-r_n)\, a_n = a_0$ has to hold.

2.3 Reciprocal of a Polynomial

The reciprocal $f(z)^{-1}$ of a polynomial $f(z)$ is defined as the function that, when multiplied with the polynomial, yields the constant function 1:

$$f(z)\, f(z)^{-1} = \left(a_0 + a_1 z^1 + \cdots + a_n z^n\right)\left(c_0 + c_1 z^1 + c_2 z^2 + \cdots + c_{-1} z^{-1} + c_{-2} z^{-2} + \cdots\right) = 1$$

The reciprocal can be expressed as an **infinite series**, which is a function that can be expressed as an infinite sum of terms, i.e., a sum that continues forever. Specifically, we can express the reciprocal as a **Laurent series**, where the coefficients can be computed using the Cauchy integral formula, as discussed below. Note that an infinite series is not considered a polynomial because a polynomial has to have a finite degree.

It is important to note that the reciprocal of a polynomial is not the same as its inverse function, denoted as $f^{-1}(z)$, which satisfies $f\left(f^{-1}(z)\right) = z$. In general, $f\left(f(z)^{-1}\right) \neq z$.

To compute the reciprocal $f(z)^{-1}$, we can first write the factored form of the polynomial with the constant of the factors normalized to one, and take the inverse of each factor. Then, we can use the geometric series formula to expand each inverted factor $\left(1 - r_k^{-1}z\right)^{-1}$ into an infinite series $1 + r_k^{-1}z + \left(r_k^{-1}z\right)^2 + \ldots$. However, there are cases where the geometric series formula does not apply, such as when $\left|r_k^{-1}z\right| > 1$ for some k, in which case we need to rewrite the inverted factor as $\left(-r_k^{-1}z\right)^{-1}\left(1 - r_k z^{-1}\right)^{-1}$, and apply the geometric series formula to the latter term, where $\left|r_k z^{-1}\right| < 1$. The case where $\left|r_k^{-1}z\right| = 1$ will not be relevant for the study of time series processes.

Let's consider the case where the products of z with all inverted roots are inside the unit circle, i.e. $\left|r_k^{-1}z\right| < 1$ for all k. This is for example the case when the polynomial $f(z)$ is defined for $|z| \leq 1$, and all inverted roots are inside the unit circle, i.e. $\left|r_k^{-1}\right| < 1$, for all k. In this case, we can apply the geometric series formula directly to all inverted factors of the factored form, and obtain the reciprocal as follows:

$$
\begin{aligned}
f(z)^{-1} &= \frac{1}{a_0}\prod_{k=1}^{n}\left(1 - r_k^{-1}z\right)^{-1}, \qquad \left|r_k^{-1}z\right| < 1, \ \forall k \\
&= \frac{1}{a_0}\prod_{k=1}^{n}\left(1 + r_k^{-1}z + r_k^{-2}z^2 + r_k^{-3}z^3 + \cdots\right) \\
&= c_0 + c_1 z + c_2 z^2 + \cdots, \qquad c_s = \frac{1}{a_0}\sum_{\substack{j_1, j_2, \ldots, j_n \geq 0 \\ j_1 + j_2 + \cdots + j_n = s}} r_1^{-j_1} r_2^{-j_2} \cdots r_n^{-j_n}
\end{aligned}
$$

where the second equation follows from the geometric series formula, and the third equation expands the polynomial according to Cauchy's product formula.

Now let's consider the more general case where the products of z with the inverted roots are inside the unit circle for m roots, and outside the unit circle for the remaining $n-m$ roots. Without loss of generality, we can order the roots so that $\left|r_i^{-1}z\right| < 1$ for $i \leq m$, and $\left|r_l^{-1}z\right| > 1$, for $l > m$. Therefore, we need to rewrite the last $n-m$ inverted factors so that the geometric series formula applies. By putting everything together, the reciprocal of $f(z)$ can be expressed as:

$$
\begin{aligned}
f(z)^{-1} &= \frac{1}{a_0}\prod_{k=1}^{n}\left(1-r_k^{-1}z\right)^{-1}, \qquad \left|r_i^{-1}z\right| < 1 < \left|r_l^{-1}z\right|, \;\; i \leq m < l \\
&= \frac{1}{a_0}\prod_{i=1}^{m}\left(1-r_i^{-1}z\right)^{-1}\prod_{l=m+1}^{n}\left(-r_l^{-1}z\right)^{-1}\left(1-r_l z^{-1}\right)^{-1} \\
&= \frac{\prod_{l=m+1}^{n}(-r_l)}{a_0}z^{-(n-m)}\prod_{i=1}^{m}\left(1+r_i^{-1}z+r_i^{-2}z^2+\cdots\right)\prod_{l=m+1}^{n}\left(1+r_l z^{-1}+r_l^2 z^{-2}+\cdots\right) \\
&= \frac{(-r_{m+1})\cdots(-r_n)}{a_0}z^{-(n-m)}\left(1+b_1 z+b_2 z^2+\cdots\right)\left(1+d_1 z^{-1}+d_2 z^{-2}+\cdots\right) \\
&= c_0+c_1 z+c_2 z^2+\cdots+c_{-1}z^{-1}+c_{-2}z^{-2}+\cdots
\end{aligned}
$$

where the coefficients are obtained using Cauchy's product formula:

$$
\begin{aligned}
b_s &= \sum_{\substack{j_1,j_2,\ldots,j_m \geq 0 \\ j_1+j_2+\cdots+j_m = s}} r_1^{-j_1}r_2^{-j_2}\cdots r_m^{-j_m} \\
d_s &= \sum_{\substack{j_{m+1},j_{m+2},\ldots,j_n \geq 0 \\ j_{m+1}+j_{m+2}+\cdots+j_n = s}} r_{m+1}^{j_{m+1}}r_{m+2}^{j_{m+2}}\cdots r_n^{j_n} \\
c_s &= \frac{(-r_{m+1})\cdots(-r_n)}{a_0}\sum_{\substack{i,j \geq 0 \\ i-j = s+(n-m)}} b_i d_j
\end{aligned}
$$

Note that for the analysis of ARIMA processes, we typically consider "stable" polynomials $f(z)$ defined for $|z| \leq 1$, where all inverted roots are inside the unit circle, i.e. $\left|r_k^{-1}\right| < 1$, for all k. This ensures that the reciprocal $f(z)^{-1}$ can be expressed as an infinite series without any negative powers of z, allowing us to use the simpler formula derived earlier. However, when dealing with rational expectations models, the more general formula will be useful.

Now that we have covered the necessary polynomial math tools, let's move on to discussing ARIMA models.

2.4 Faulhaber's Formula

Faulhaber's formula, named after mathematician Johann Faulhaber, expresses the sum of the pth powers of the first n positive integers:

$$\sum_{k=1}^{n} k^p = 1^p + 2^p + 3^p + \cdots + n^p = \frac{1}{p+1} \sum_{k=0}^{p} \binom{p+1}{k} B_k n^{p-k+1}$$

Here, $\binom{p+1}{k} = \frac{(p+1)!}{k!\,(p+1-k)!}$ is the **binomial coefficient "$p+1$ choose k"**, where $k!$ $= k \times (k-1) \times (k-2) \times \cdots \times 2 \times 1$ represents the **factorial of** k, and B_k is the **Bernoulli number**, where $\{B_1, B_2, \dots\} = \left\{\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42}, 0, -\frac{1}{30}, \dots\right\}$.

For $p = 1$, we have the **triangular numbers**:

$$\sum_{k=1}^{n} k^1 = \frac{n(n+1)}{2} = \frac{1}{2}(n^2 + n)$$

and for $p = 2$, we have the **square pyramidal numbers**:

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} = \frac{1}{3}(n^3 + \tfrac{3}{2}n^2 + \tfrac{1}{2}n)$$

and the formulae for $p \geq 3$ will not be relevant for this lecture. However, we will use the fact that the sum of the pth powers of the first n positive integers is a polynomial in n of

order $p+1$.

3 Univariate Stationary Time Series

3.1 Autoregressive Moving Average (ARMA) Model

Consider a **moving average process of order** q, denoted as **MA(q)**:

$$Y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \sigma_\epsilon^2\right)$$

where $i.i.d.$ means that the residuals ϵ_t are independent and identically distributed over time t.

MA processes are known to have a short memory, meaning that information about current and past realizations of $\{Y_t\}$ is only useful for making forecasts at short horizons and becomes useless beyond horizon $q+1$. This is because beyond this horizon, Y_{t+q+1} and Y_t do not have any shocks in common and are thus independent, rendering the information about Y_t useless for predicting Y_{t+q+1}.

In contrast, consider an **autoregressive process of order** p, denoted as **AR(p)**:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \sigma_\epsilon^2\right)$$

where $i.i.d.$ means independent and identically distributed.

AR processes have a longer memory than MA processes because even though Y_t only directly depends on p lags, the lagged variables again depend on p lags, and so on, creating a chain of indirect dependencies. Therefore, there is a non-zero correlation between Y_t and Y_{t-l} for any finite integer l. However, the autocorrelation function of a stationary AR process decays exponentially, which means that past information becomes exponentially less relevant the further into the future the prediction goes.

Finally, consider the following generalization of an MA and AR process, an **autoregres-**

sive moving average process with p **AR lags** and q **MA lags**, denoted as **ARMA**(p, q):

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \sigma_\epsilon^2\right)$$

where $i.i.d.$ means independent and identically distributed.

ARMA processes combine the properties of MA and AR processes, making them a powerful forecasting tool.

3.2 Stationarity

A process is said to be **stationary** if its statistical properties do not change over time. Formally, stationarity means that the distribution of the random varibles Y_t in the stochastic process $\{Y_t\}$ is the same for all time periods t. A weaker form of stationarity is **weak stationarity**, which requires that the mean and covariance of the process are constant over time. More specifically, for a weakly stationary process $\{Y_t\}$, we have $E[Y_t] = \mu$, where μ is a constant, and $\mathrm{Cov}(Y_t, Y_{t-k}) = \gamma_k$, where γ_k depends only on the time lag k and not on time t.

Note that all finite-order MA processes are stationary. This is because MA processes solely depend on a constant and on shocks that have the same distribution over time. Infinite-order MA processes are also stationary as long as the coefficients decay fast enough, for example, if they decay exponentially. If the coefficients do not decay sufficiently fast, the variance of the infinite-order MA process goes to infinity, making it impossible for the process to be covariance stationary.

To determine whether AR processes are stationary, verify if the process can be rewritten to depend only on stationary shocks, which simplifies the task of determining whether the AR process is stationary. For instance, consider an AR(1) process. By repeatedly replacing the right-hand side variable over and over again, the AR(1) process can be expressed as

follows:

$$\begin{aligned}
Y_t &= c + \phi_1 Y_{t-1} + \epsilon_t \\
&= c + \phi_1 \left(c + \phi_1 Y_{t-2} + \epsilon_{t-1}\right) + \epsilon_t \\
&= c + \phi_1 \left(c + \phi_1 \left(c + \phi_1 Y_{t-3} + \epsilon_{t-2}\right) + \epsilon_{t-1}\right) + \epsilon_t \\
&\vdots \\
&= \left(1 + \phi_1 + \phi_1^2 + \cdots\right) c + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_1^2 \epsilon_{t-2} + \cdots + \lim_{s \to \infty} \phi_1^s Y_{t-s}
\end{aligned}$$

Note that when $|\phi_1| < 1$, then $\lim_{s \to \infty} \phi_1^s = 0$, so Y_t becomes a an infinite-order MA process, where the coefficients decay exponentially, i.e. the coefficient ϕ_1^k on ϵ_{t-k} goes to zero exponentially when k increases. Therefore, $|\phi_1| < 1$ implies that the AR(1) process is stationary.

An exponential decay in the MA parameters implies stationarity because it permits the use of the geometric series formula when computing the mean and covariances, ultimately resulting in constant values for these statistics:

$$\begin{aligned}
|\phi_1| < 1: \qquad E\left[Y_t\right] &= \left(1 + \phi_1 + \phi_1^2 + \cdots\right) c = \frac{c}{1 - \phi_1} \\
\operatorname{Cov}\left(Y_t, Y_{t-k}\right) &= \phi_1^k \left(1 + \phi_1^2 + \phi_1^4 + \cdots\right) \sigma_\epsilon^2 = \frac{\phi_1^k \sigma_\epsilon^2}{1 - \phi_1^2}
\end{aligned}$$

As the order of an AR(p) or an ARMA(p, q) process increases ($p \geq 2$), iterating on the process to express it as a function of shocks becomes increasingly difficult. This is where polynomial mathematics come in handy, which we will discuss further below. In short, the ARMA(p, q) process has an infinite-order MA representation, with coefficients that decay exponentially, and is thus stationary as long as the inverted roots of the AR lag polynomial are all inside the unit circle. This statement will make more sense once we define these objects.

Note that all stationary AR processes have an MA(∞) representation. More generally, **Wold's Decomposition Theorem**, also known as the **MA Representation Theorem**, states that all covariance-stationary time series processes, including non-linear processes, can be expressed as the sum of a deterministic component (such as an intercept or trend)

and a stochastic component represented by an MA(∞) process. Hence, to check for stationarity, it is sufficient to express the process in terms of shocks, verify whether all terms that do not belong to an MA(∞) cancel out, and ensure that the MA coefficients decay sufficiently fast so that the covariances are finite.

3.3 Invertibility

An **invertible** process is one where the residuals of the process are linear functions of current and past variables, $\{Y_t, Y_{t-1}, \dots\}$, as opposed to being a function of future variables or both future and past variables. An AR process is by definition invertible, as it relates current to past variables with a single residual ϵ_t, making the residual a linear function of current and past variables. Hence, a process is invertible if it can be rewritten as an (infinite-order) autoregressive (AR) process.

Invertibility is a desirable property in analyzing systems where the past has an effect on the future but not vice versa, which is often the case in real-world applications. If an ARMA process is not invertible, future variables would have an effect on current variables, making analysis more complex. Furthermore, for an invertible ARMA process, observing a time series up to time t, i.e. $\{y_0, y_1, \dots, y_t\}$, means that we also observe all the shocks that have occurred up to that point, i.e. $\{\epsilon_0, \epsilon_1, \dots, \epsilon_t\}$. This is not the case under non-invertibility, where past shocks may not be identifiable from the observed time series.

To demonstrate the necessary assumptions for invertibility of an MA process, consider an MA(1) process and use the process to replace the lagged shocks over and over again:

$$
\begin{aligned}
Y_t &= \mu + \theta_1 \epsilon_{t-1} + \epsilon_t \\
&= \mu + \theta_1 \left(Y_{t-1} - \mu - \theta_1 \epsilon_{t-2}\right) + \epsilon_t \\
&= \mu + \theta_1 \left(Y_{t-1} - \mu - \theta_1 \left(Y_{t-2} - \mu - \theta_1 \epsilon_{t-3}\right)\right) + \epsilon_t \\
&\vdots \\
&= \left(1 + (-\theta_1) + (-\theta_1)^2 + \cdots\right) \mu - (-\theta_1) Y_{t-1} - (-\theta_1)^2 Y_{t-2} - \cdots - \lim_{s \to \infty} (-\theta_1)^s \epsilon_{t-s} + \epsilon_t
\end{aligned}
$$

Note that when $|\theta_1| < 1$, then $\lim_{s \to \infty} (-\theta_1)^s = 0$, so that Y_t is a linear function of past variables $\{Y_{t-1}, Y_{t-2}, \dots\}$ and the shock ϵ_t. Hence, $|\theta_1| < 1$ implies that the MA(1) process can be expressed as an AR(∞) process and is thus invertible.

As the order of an MA(q) or an ARMA(p, q) process increases ($q \geq 2$), iterating on the process to express it as a function of current and past variables becomes increasingly difficult. This is where polynomial mathematics come in handy, which we will discuss further below. In short, the ARMA(p, q) process has an infinite-order AR representation, and is thus invertible as long as the inverted roots of the MA lag polynomial are all inside the unit circle. This statement will make more sense once we define these objects.

It turns out that all invertible MA(q) processes have an AR(∞) representation.

It is important to note that the direction of causality cannot be determined from the data alone. The data only provides the covariance between Y_t and Y_{t-1}, and it is not possible to determine whether the former depends on the latter or vice versa since $\text{Cov}(Y_t, Y_{t-1}) = \text{Cov}(Y_{t-1}, Y_t)$. This means that processes that relate current to past variables can produce identical moments as processes that relate current to future variables, using the same parameters. Therefore, two completely different processes can be observationally equivalent. When estimating MA or ARMA models using statistical software such as `R`, the program automatically chooses the invertible process instead of providing all the processes that produce the same moments.

To illustrate this point, consider two MA(1) processes (a) and (b), where $|\theta_1| < 1$:

$$
\begin{aligned}
(a) \qquad & Y_t = \mu + \theta_1 \epsilon_{t-1} + \epsilon_t, \quad \epsilon_t \overset{i.i.d.}{\sim} \text{N}\left(0, \sigma_\epsilon^2\right) \\
(b) \qquad & Y_t^* = \mu + \frac{1}{\theta_1} \epsilon_{t-1}^* + \epsilon_t^*, \quad \epsilon_t^* = \theta_1 \epsilon_{t+1}
\end{aligned}
$$

By iterating the former process backward and the latter forward, we get the AR(∞) process derived above. However, model (b) relates current to future variables instead of current to

past variables:

$$(a) \quad Y_t = \left(1 + (-\theta_1) + (-\theta_1)^2 + \cdots\right)\mu - (-\theta_1)\, Y_{t-1} - (-\theta_1)^2\, Y_{t-2} - \cdots + \epsilon_t$$

$$(b) \quad Y_t^* = \left(1 + (-\theta_1) + (-\theta_1)^2 + \cdots\right)\mu - (-\theta_1)\, Y_{t+1}^* - (-\theta_1)^2\, Y_{t+2}^* - \cdots + \epsilon_t$$

Even though both processes produce the same moments and are observationally equivalent, statistical software like `R` will not provide model (b), since model (a) is the one that is invertible. In contrast, when $|\theta_1| > 1$, then `R` would choose model (b), as it would be the one that is invertible in that case. The only case when R cannot produce an invertible process is when $|\theta_1| = 1$.

3.4 Lag Operator and Lag Polynomial

To analyze the properties of an ARMA process and compute its MA and AR representation, it is convenient to represent the process in terms of lag operators and lag polynomials, and then perform standard math operations on these polynomials.

The **lag operator** L, also known as the **backshift operator**, shifts the time period of a variable one period back, such that $LY_t = Y_{t-1}$. Repeated application of the lag operator k times gives $L^k Y_t = L^{k-1} Y_{t-1} = \ldots = Y_{t-k}$. Moreover, the inverse of the lag operator shifts the time period forward, such that $L^{-1} Y_t = Y_{t+1}$. In general, $L^{-h} Y_t = Y_{t+h}$. for any positive integer h.

Using the lag operator, we can express the ARMA process in terms of **lag polynomials**:

$$\phi(L)\, Y_t = c + \theta(L)\, \epsilon_t, \qquad \phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$$
$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$$

where $\phi(L)$ is the **AR lag polynomial** and $\theta(L)$ is the **MA lag polynomial**.

The benefit of using lag polynomials is that we can apply the same polynomial math tools to them as we do with ordinary polynomials, even though they're functions of an operator L rather than functions of a variable z. We simply replace the lag operator with

a variable z, define the variable range to $|z| \leq 1$, perform standard polynomial operations on it, and then replace z with L after the transformation.

To illustrate this idea, let's consider the AR(1) process and its MA(∞) representation that we derived earlier, given $|\phi_1| < 1$:

$$\begin{aligned} \text{AR}(1): \quad & (1 - \phi_1 L) Y_t = c + \epsilon_t, \\ \text{MA}(\infty): \quad & Y_t = \left(1 + \phi_1 L + \phi_1^2 L^2 + \phi_1^3 L^3 + \cdots\right)(c + \epsilon_t), \end{aligned}$$

where $L^k c = c$ implies an intercept of $\left(1 + \phi_1 + \phi_1^2 + \cdots\right) c$. Hence, the conversion between AR(1) and MA(∞) is as if we took the reciprocal of the AR(1) lag polynomial, where the lag operator is replaced with a variable z, where $|z| \leq 1$:

$$\phi(z)^{-1} = (1 - \phi_1 z)^{-1} = 1 + \phi_1 z + \phi_1^2 z^2 + \phi_1^3 z^3 + \cdots$$

This result holds if $|\phi_1 z| < 1$, by the geometric series formula, which is the case under stationarity $|\phi_1| < 1$, given $|z| \leq 1$. Hence, we can derive the MA representation of the AR(1) process much more quickly by using the geometric series formula on the AR lag polynomial, rather than by iteratively replacing past variables.

Similarly, let's consider the MA(1) process and its AR(∞) representation that we derived earlier, given $|\theta_1| < 1$:

$$\begin{aligned} \text{MA}(1): \quad & Y_t = \mu + (1 + \theta_1 L) \epsilon_t, \\ \text{AR}(\infty): \quad & \left(1 - \theta_1 L^1 + \theta_1^2 L^2 - \cdots\right) Y_t = \left(1 - \theta_1 L^1 + \theta_1^2 L^2 - \cdots\right) \mu + \epsilon_t, \end{aligned}$$

where $L^k \mu = \mu$ implies an intercept of $\left(1 - \theta_1 + \theta_1^2 - \cdots\right) \mu$. Hence, the conversion between MA(1) and AR(∞) is as if we took the reciprocal of the MA(1) lag polynomial, where the lag operator is replaced with a variable z, where $|z| \leq 1$:

$$\theta(z)^{-1} = (1 + \theta_1 z)^{-1} = 1 + (-\theta_1) z + (-\theta_1)^2 z^2 + (-\theta_1)^3 z^3 + \cdots$$

This result holds if $|(-\theta_1)\, z| < 1$, by the geometric series formula, which is the case under invertibility $|\theta_1| < 1$, given $|z| \leq 1$. Hence, we can derive the AR representation of the MA(1) process much more quickly by using the geometric series formula on the MA lag polynomial, rather than by iteratively replacing past variables.

In general, to rewrite an ARMA(p, q) process, we can first factorize the lag polynomials with a normalized constant of one and replace the lag operator with a variable $|z| \leq 1$. Then, we can use the geometric series formula to invert the factors:

$$\phi(L)\, Y_t = c + \theta(L)\, \epsilon_t$$

$$(1 - \phi_1 L - \cdots - \phi_p L^p)\, Y_t = c + (1 + \theta_1 L + \cdots + \theta_q L^q)\, \epsilon_t$$

$$\left(1 - r_1^{-1} L\right)\left(1 - r_2^{-1} L\right) \cdots \left(1 - r_p^{-1} L\right) Y_t = c + \left(1 - s_1^{-1} L\right)\left(1 - s_2^{-1} L\right) \cdots \left(1 - s_q^{-1} L\right) \epsilon_t$$

where $\{r_1, \ldots, r_p\}$ and $\{s_1, \ldots, s_q\}$ are the roots of the AR polynomial $\phi(z)$ and MA polynomial $\theta(z)$, respectively.

Note that an ARMA(p, q) process can be rewritten entirely as an MA(∞) process if the inverted roots of the AR lag polynomial lie inside the unit circle. This allows us to apply the geometric series formula to all AR factors, i.e. $\left(1 - r_i^{-1} z\right)^{-1} = 1 + r_i^{-1} z + r_i^{-2} z^2 + \ldots$, resulting in positive powers and dependence only on past shocks. The resulting MA coefficients decay exponentially, and the ARMA(p, q) process is stationary. Thus, we can test for stationarity by checking whether the inverted roots of the AR polynomial $\phi(z)$ lie inside the unit circle.

As a side note, it is worth mentioning that the property where all the inverted roots of the AR polynomial lie inside the unit circle is referred to as **stability**. A stable process is guaranteed to be stationary, but the converse is not necessarily true. A stationary process does not necessarily imply stability.

Similarly, an ARMA(p, q) process can be written entirely as an AR(∞) process if the inverted roots of the MA lag polynomial lie inside the unit circle. This allows us to apply the geometric series formula to all MA factors, i.e. $\left(1 - s_j^{-1} z\right)^{-1} = 1 + s_j^{-1} z + s_j^{-2} z^2 + \ldots$, resulting in positive powers and dependence only on past variables. The resulting AR

coefficients decay exponentially, and the ARMA(p,q) process is invertible. Thus, we can test for invertibility by checking whether the inverted roots of the MA polynomial $\theta(z)$ lie inside the unit circle.

Hence, under stationarity and invertibility, the ARMA(p,q) process has the following MA(∞) and an AR(∞) representation:

$$\begin{aligned}
\text{ARMA}(p,q): &\quad \phi(L)Y_t = c + \theta(L)\epsilon_t, \\
\text{MA}(\infty): &\quad Y_t = \phi(1)^{-1}c + \overline{\theta}(L)\epsilon_t, \qquad \overline{\theta}(z) = \phi(z)^{-1}\theta(z), \\
\text{AR}(\infty): &\quad \overline{\phi}(L)Y_t = \theta(1)^{-1}c + \epsilon_t, \qquad \overline{\phi}(z) = \theta(L)^{-1}\phi(z), \quad |z| \leq 1
\end{aligned}$$

Section 2 shows how to compute the reciprocal of a polynomial such as $\phi(z)^{-1}$ and $\theta(z)^{-1}$.

Example: Compute the MA(∞) representation of the following AR(2) process:

$$Y_t = 3 + 0.3Y_{t-1} + 0.4Y_{t-2} + \epsilon_t$$

Solution: Compute the factored form:

$$\begin{gathered}
\phi(L)Y_t = c + \theta(L)\epsilon_t \\
\Downarrow \\
\left(1 - 0.3L - 0.4L^2\right)Y_t = 3 + \epsilon_t \\
(1 - 0.8L)(1 + 0.5L)Y_t = 3 + \epsilon_t
\end{gathered}$$

where $r_1^{-1} = 0.8$ and $r_2^{-2} = -0.5$ are the inverted roots of the AR lag polynomial, which are both inside the unit circle; hence, the process is stationary and has an MA(∞) representation. To compute the MA(∞) process, apply the geometric series formula to the inverted

factors after replacing L with a variable z, where $|z| \leq 1$:

$$
\begin{aligned}
&Y_t = \phi(1)^{-1} c + \overline{\theta}(L)\,\epsilon_t, \qquad \overline{\theta}(z) = \phi(z)^{-1}\theta(z), \quad |z| \leq 1 \\
&\Downarrow \\
&Y_t = 10 + \overline{\theta}(L)\,\epsilon_t, \qquad \overline{\theta}(z) = (1-0.8z)^{-1}(1+0.5z)^{-1}, \quad |z| \leq 1 \\
&\qquad\qquad\qquad\qquad\quad = \left(1 + 0.8z + 0.8^2 z^2 + \cdots\right)\left(1 - 0.5z + 0.5^2 z^2 - \cdots\right)
\end{aligned}
$$

Here, the intercept is calculated as follows: $(1-0.8)^{-1}(1+0.5)^{-1}\,3 = 10$. To compute the lag polynomial $\overline{\theta}(z)$, we can use the Cauchy product formula as described in Section 2:

$$
Y_t = 10 + c_1\epsilon_{t-1} + c_2\epsilon_{t-2} + c_3\epsilon_{t-3} + \cdots + \epsilon_t, \qquad c_s = \sum_{j=0}^{s} 0.8^j\,(-0.5)^{s-j}
$$

and therefore we get the following (truncated) MA representation of the AR(2) process:

$$
\begin{aligned}
Y_t = 10 &+ 0.3\epsilon_{t-1} + 0.49\epsilon_{t-2} + 0.267\epsilon_{t-3} + 0.2761\epsilon_{t-3} + 0.18963\epsilon_{t-4} + 0.167329\epsilon_{t-5} \\
&+ 0.1260507\epsilon_{t-6} + 0.1047468\epsilon_{t-7} + 0.08184432\epsilon_{t-8} + 0.0664520209\epsilon_{t-9} + \cdots + \epsilon_t
\end{aligned}
$$

The formula for the MA coefficients c_s above reveals an exponential decay as s increases, indicating that the weights assigned to past shocks become extremely small in the MA(∞) process. Consequently, the process maintains a finite variance, and the mean, variances, and covariances remain constant due to the identically distributed shocks. Hence, the AR(2) process is stationary. However, if either of the inverted roots $r_1^{-1} = 0.8$ or $r_2^{-2} = -0.5$ were larger than one in magnitude, the coefficients c_s would not decay exponentially as s increases, resulting in an unstable process.

4 Univariate Non-Stationary Time Series

4.1 Deterministic and Stochastic Trends

A stationary process is characterized by a constant distribution, which implies constant mean and variance. However, many economic processes exhibit a mean that increases or decreases over time, such as GDP or hours worked, as well as variances that change over time. Despite these changes, the distribution of these processes typically does not fluctuate randomly but instead changes smoothly over time. This gradual change in the distribution is referred to as a **trend**.

The concepts of stationarity and trends are important because they impact our ability to analyze and make predictions using time series data. Under stationarity, the stochastic process $\{Y_t\}$ consists of random variables $\{Y_0, Y_1, \dots\}$ that all have the same distribution. A time series $\{y_t\}$ is an outcome of the stochastic process $\{Y_t\}$, which consists of only one observation per random variable $\{y_0, y_1, \dots, y_T\}$. For example, we observe only one data point for U.S. GDP in 2021, i.e., $y_{2021} = 23.32$ trillion USD. If the distribution of every random variable in the stochastic process is different at random, then time series data cannot be used to estimate the distribution of the stochastic process or make predictions about the future. This is why stationarity is a useful assumption, as it allows us to combine observations across time to learn more about the constant distribution of the stochastic process.

However, even without stationarity, there is still hope. If the process is non-stationary but exhibits a smooth change in distribution, such as having a trend, it is still possible to make inferences about the distribution of the stochastic process and make predictions. In this case, the trend can be estimated, whereas if the change in distribution were random, estimating how the distribution changes would be impossible.

We distinguish between deterministic and stochastic trends. A **deterministic trend** occurs when the change in the distribution depends entirely on the time period t. A deterministic trend in the mean of the distribution can thus be denoted with a function of t, $\delta(t)$:

$$\phi(L) Y_t = \delta(t) + \theta(L) \epsilon_t$$

where $\phi(L)$ and $\theta(L)$ are the AR and MA lag polynomials respectively. So far, we assumed that there is no time trend, i.e., $\delta(t) = c$ doesn't depend on t. Some common deterministic trend specifications are:

$$\begin{aligned}
\text{Linear Trend}: &\quad \delta(t) = \delta_0 + \delta_1 t \\
\text{Quadratic Trend}: &\quad \delta(t) = \delta_0 + \delta_1 t + \delta_2 t^2 \\
\text{Polynomial Trend of Degree } n: &\quad \delta(t) = \delta_0 + \delta_1 t + \delta_2 t^2 + \cdots + \delta_n t^n \\
\text{Exponential Trend}: &\quad \delta(t) = \delta_0 e^{\delta_1 t} \\
\text{Logistic Trend}: &\quad \delta(t) = \frac{\delta_2}{1 + \delta_0 e^{\delta_1 t}}
\end{aligned}$$

A process with a deterministic trend is called **trend-stationary**, because the detrended time series $\overline{Y}_t = Y_t - \delta(t)$ is stationary.

A **stochastic trend** occurs when the distribution of a time series changes over time, not due to an exogenous trend $\delta(t)$, but rather because the random events of the past remain relevant forever so that the process never reverts back to its original distribution. For example, the following random walk process has a stochastic trend, because unlike a stationary AR process, the weight on past shock doesn't go to zero:

$$Y_t = Y_{t-1} + \epsilon_t = Y_{t-2} + \epsilon_{t-1} + \epsilon_t = \cdots = Y_0 + \sum_{s=1}^{t} \epsilon_s$$

Hence, the variance of the process increases over time, because more and more shocks contribute to the outcome of the process. Such process is called **difference-stationary**, because the difference $\Delta Y_t = \epsilon_t$ is stationary.

More generally, a process that contains a stochastic trend is called a **unit root process**, indicating that at least one of the roots of the autoregressive (AR) lag polynomial is equal to one. In such cases, changes in past outcomes, denoted as $\Delta Y_{t-l} = Y_{t-l} - Y_{t-(l+1)} = (1 - L) Y_{t-l}$, remain relevant for today's outcome Y_t, even when considering changes that occurred in the infinite past as l approaches infinity. Consequently, past random events retain their relevance and do not lose importance, unlike when all inverted roots are strictly

smaller than one in magnitude.

To understand why a unit root in the AR lag polynomial implies this behavior, consider the following ARMA process with a unit root ($r_p = 1$):

$$\begin{aligned}
\text{ARMA}: &\quad \phi(L) Y_t = c + \theta(L) \epsilon_t \\
\text{Normal Form}: &\quad (1 - \phi_1 L - \cdots - \phi_p L^p) Y_t = c + \theta(L) \epsilon_t \\
\text{Factored Form}: &\quad \left(1 - r_1^{-1} L\right) \cdots \left(1 - r_{p-1}^{-1} L\right) (1 - L) Y_t = c + \theta(L) \epsilon_t \\
\text{Removal of Unit Root}: &\quad \left(1 - r_1^{-1} L\right) \cdots \left(1 - r_{p-1}^{-1} L\right) \Delta Y_t = c + \theta(L) \epsilon_t \\
\text{ARMA for 1st Difference}: &\quad \phi^*(L) \Delta Y_t = c + \theta(L) \epsilon_t
\end{aligned}$$

Hence, the first difference in the process follows an ARMA process and directly depends on the shocks. Since the level Y_t is simply the initial value plus the sum of the differences, it assigns equal weights to all past changes, resulting in the persistence of past shocks and the existence of a stochastic trend:

$$Y_t = Y_{t-1} + \Delta Y_t = Y_{t-2} + \Delta Y_{t-1} + \Delta Y_t = \cdots = Y_0 + \sum_{s=1}^{t} \Delta Y_s$$

Formally, a unit root process is denoted as $I(d)$, where d refers to the number of roots in the AR lag polynomial that are equal to one. Hence, if $d = 0$, there is no unit root and thus no stochastic trend, and if $d > 0$, the process has a stochastic trend. The notation $I(d)$ indicates that the process is **integrated of order** d. This means that if the time series is differenced d times, the resulting process no longer exhibits a unit root. Similarly, by taking the integral of a stationary process d times, we obtain an $I(d)$ process. To illustrate this, consider differencing a stationary process d times. Each differencing operation removes one unit root factor, resulting in a process that no longer exhibits a unit root. This is demonstrated in the above ARMA model, where $r_p = 1$ is removed by taking the first difference: $\left(1 - r_p^{-1} L\right) Y_t = \Delta Y_t$. Similarly, if $r_p = r_{p-1} = 1$, the unit roots are removed by taking the second difference: $\left(1 - r_{p-1}^{-1} L\right)\left(1 - r_p^{-1} L\right) Y_t = \Delta^2 Y_t$, where $\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$. Thus, differencing eliminates unit roots, while integration introduces unit

roots.

To summarize, a deterministic trend occurs when the change in the distribution is exogenous and depends solely on time t, while a stochastic trend arises from the accumulation of random shocks or innovations over time. It is important to note that predictions differ significantly depending on whether the trend is deterministic or stochastic. Predicting a deterministic trend requires data on past outcomes solely to estimate the function $\delta(t)$. In contrast, predicting a stochastic trend relies on the past outcomes of the time series, as the stochastic trend accumulates information over time.

4.2 Autoregressive Integrated Moving Average (ARIMA) Model

Section 3.1 introduces the ARMA model for stationary processes, and in Section 4.1, we learned that differencing eliminates unit roots and thus transforms a non-stationary process with a stochastic trend into a stationary one. When we apply the ARMA model to a differenced process, it is referred to as an **autoregressive integrated moving average** model, denoted as **ARIMA**(p, d, q):

$$\phi(L)\,\Delta^d Y_t = \delta^k(t) + \theta(L)\,\epsilon_t, \qquad \phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$$
$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$$

Here, d represents the number of unit roots in the AR polynomial, indicating the number of times the process needs to be differenced to achieve trend-stationarity, p represents the number of AR lags after differencing d times, and q represents the number of MA lags in the process. The deterministic trend is captured with $\delta^k(t) = \delta_0 + \delta_1 t + \cdots + \delta_k t^k$, which for simplicity is assumed to be a polynomial of degree k, but it can take any form.

After estimating the ARMA parameters, we discuss two ways to obtain the level series $\{Y_t\}$ from the differenced process $\{\Delta^d Y_t\}$. We discuss two ways. First, we can compute the non-stationary ARMA process for Y_t using $\Delta^d Y_t = (1-L)^d Y_t$ to rewrite the right-hand

side of the ARIMA model as follows:

$$\begin{aligned}\phi(L)\Delta^d Y_t &= \left(1-r_1^{-1}L\right)\cdots\left(1-r_p^{-1}L\right)(1-L)\cdots(1-L)\,Y_t \\ &= \left(1-r_1^{-1}L\right)\cdots\left(1-r_p^{-1}L\right)\left(1-r_{p+1}^{-1}L\right)\cdots\left(1-r_{p+d}^{-1}L\right)Y_t, \quad r_j=1, \ \ \forall j>p \\ &= \phi^*(L)\,Y_t\end{aligned}$$

where $\phi^*(L)$ is the AR lag polynomial of the non-stationary ARMA model for Y_t, and the MA lag polynomial $\theta(L)$ and the deterministic trend $\delta^k(t)$ remain the same. Note that the AR lag polynomial consists of $p+d$ lags; hence, the ARMA model has d additional AR lags compared to the ARIMA model. To summarize, to make forecasts about the level process $\{Y_t\}$, we rewrite the ARIMA model as a non-stationary ARMA model and then compute forecasts as in the ARMA model.

The second way to obtain the level series $\{Y_t\}$ from the differenced process $\left\{\Delta^d Y_t\right\}$ is to integrate the differenced series d times. For example, given ΔY_t for all t, and the initial value Y_0, we can compute Y_t by summing (or integrating) over all the differenced variables:

$$\Delta^1 Y_t = Y_0 + \sum_{s=1}^{t} \Delta^1 Y_s$$

and in general, the following holds:

$$\Delta^{j-1} Y_t = \Delta^{j-1} Y_{j-1} + \sum_{s=j}^{t} \Delta^j Y_s \quad j=1,2,\ldots$$

Hence, Y_t is derived by first computing $\Delta^{d-1}Y_t$ from $\Delta^d Y_t$, and then computing $\Delta^{d-2}Y_t$ from $\Delta^{d-1}Y_t$, and so on, until we arrive at $\Delta^0 Y_t = Y_t$.

Let's apply the above procedure to relate Y_t to $\Delta^4 Y_t$, and simplify the terms using

Faulhaber's formula from Section 2.4:

$$\begin{aligned}
Y_t &= Y_0 + \sum_{s_1=1}^{t} \Delta Y_{s_1} \\
&= Y_0 + \Delta Y_1 t + \sum_{s_1=2}^{t} \sum_{s_2=2}^{s_1} \Delta^2 Y_{s_2} \\
&= Y_0 + \left(\Delta Y_1 - \frac{1}{2}\Delta^2 Y_2\right) t + \left(\frac{1}{2}\Delta^2 Y_2\right) t^2 + \sum_{s_1=3}^{t} \sum_{s_2=3}^{s_1} \sum_{s_3=3}^{s_2} \Delta^3 Y_{s_3} \\
&= Y_0 + \left(\Delta Y_1 - \frac{1}{2}\Delta^2 Y_2 + \frac{11}{6}\Delta^3 Y_3\right) t + \left(\frac{1}{2}\Delta^2 Y_2 - \frac{1}{2}\Delta^3 Y_3\right) t^2 + \left(\frac{1}{6}\Delta^3 Y_3\right) t^3 \\
&\quad + \sum_{s_1=4}^{t} \sum_{s_2=4}^{s_1} \sum_{s_3=4}^{s_2} \sum_{s_4=4}^{s_3} \Delta^4 Y_{s_4}
\end{aligned}$$

The above calculations suggest that we can relate Y_t to $\Delta^d Y_t$ for any d as follows:

$$Y_t = \underbrace{\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_{d-1} t^{d-1}}_{\lambda^{d-1}(t)} + \sum_{s_1=d}^{t} \sum_{s_2=d}^{s_1} \sum_{s_3=d}^{s_2} \cdots \sum_{s_d=d}^{s_{d-1}} \Delta^d Y_{s_d}$$

where $\lambda^{d-1}(t)$ is a polynomial in t of degree $d-1$, and its coefficients depend on the initial values of the process: $\{Y_0, Y_1, \ldots, Y_{d-1}\}$. To summarize, to make forecasts about the level process $\{Y_t\}$ with this approach, first make forecasts of the differenced series $\left\{\Delta^d Y_t\right\}$ using the ARIMA model, and then use the above expression to compute the level series from the differenced series.

The advantage of the second approach is that it allows us to directly relate the level variables to the underlying shocks. This is not possible with the first approach, as the ARMA model for $\{Y_t\}$ has unit roots in the AR lag polynomial and therefore does not have an MA(∞) representation. Unlike the non-stationary ARMA model for $\{Y_t\}$, the ARIMA model for $\left\{\Delta^d Y_t\right\}$ has an MA(∞) representation as long as the inverted roots of

the AR lag polynomial $\phi(L)$ are all smaller than one in magnitude:

$$\begin{aligned}\phi(L)\Delta^d Y_t &= \delta^k(t) + \theta(L)\epsilon_t, \\ \Delta^d Y_t &= \gamma^k(t) + \overline{\theta}(L)\epsilon_t, \qquad \gamma^k(t) = \phi(1)^{-1}\delta^k(t) \\ &\qquad\qquad \overline{\theta}(z) = \phi(z)^{-1}\theta(z), \quad \forall |z| \leq 1\end{aligned}$$

where $\overline{\theta}(L) = 1 + \overline{\theta}_1 L + \overline{\theta}_2 L^2 + \cdots$ is an infinite series (see Section 3.4 for how to compute it), and $\gamma^k(t) = \gamma_0 + \gamma_1 t + \cdots + \gamma_k t^k$ represents the deterministic trend of the MA(∞) process.

To relate Y_t to the underlying shocks, we replace $\Delta^d Y_t$ with the above MA(∞) representation:

$$Y_t = \lambda^{d-1}(t) + \sum_{s_1=d}^{t}\sum_{s_2=d}^{s_1}\sum_{s_3=d}^{s_2}\cdots\sum_{s_d=d}^{s_{d-1}}\left(\gamma^k(s_d) + \overline{\theta}(L)\epsilon_{s_d}\right)$$

To find an expression for the dth integral of the deterministic trend $\gamma^k(t)$, which is a polynomial in t of degree k, we can use Faulhaber's formula from Section 2.4. The formula states that the sum of the kth powers of the first s positive integers is a polynomial in s of order $k+1$, and therefore,

$$\begin{aligned}\sum_{s_1=d}^{t}\sum_{s_2=d}^{s_1}\cdots\sum_{s_d=d}^{s_{d-1}}\gamma^k(s_d) &= \sum_{s_1=d}^{t}\sum_{s_2=d}^{s_1}\cdots\sum_{s_{d-1}=d}^{s_{d-2}}\sum_{s_d=d}^{s_{d-1}}\left(\gamma_0 + \gamma_1 s_d + \gamma_2 s_d^2 + \cdots + \gamma_k s_d^k\right) \\ &= \sum_{s_1=d}^{t}\sum_{s_2=d}^{s_1}\cdots\sum_{s_{d-1}=d}^{s_{d-2}}\left(\eta_0 + \eta_1 s_{d-1} + \eta_2 s_{d-1}^2 + \cdots + \eta_{k+1} s_{d-1}^{k+1}\right) \\ &\;\;\vdots \\ &= \sum_{s_1=d}^{t}\left(\psi_0 + \psi_1 s_1 + \psi_2 s_1^2 + \cdots + \psi_{k+d-1} s_1^{k+d-1}\right) \\ &= \omega_0 + \omega_1 t + \omega_2 t^2 + \cdots + \omega_{k+d} t^{k+d} = \omega^{k+d}(t)\end{aligned}$$

This reveals that integrating d times increases the degree of the polynomial of the deterministic trend by d. Hence, an intercept in the ARIMA(p,d,q) model, i.e., $\delta^0(t) = \delta_0$, implies that Y_t has a deterministic trend of the form of a polynomial of degree d.

This is important because it demonstrates that the ARIMA model produces not only a stochastic trend but also a deterministic trend in $\{Y_t\}$. If we observe a quadratic trend alongside a unit root $I(1)$ in the level series $\{Y_t\}$, it is essential to note that the trend in the ARIMA model should not be quadratic; instead, it should be linear. Introducing a quadratic trend in the ARIMA model would result in a cubic trend in $\{Y_t\}$. Therefore, it is necessary to either differentiate the series prior to determining the trend or account for the change in trend when transitioning from the differenced series to the level series. This ensures that the trend specification in the ARIMA model aligns accurately with the underlying behavior of the time series.

Using the above expression for the trend, we can relate the level variable to the deterministic trend and random shocks as follows:

$$Y_t = \lambda^{d-1}(t) + \omega^{k+d}(t) + \sum_{s_1=d}^{t}\sum_{s_2=d}^{s_1}\sum_{s_3=d}^{s_2}\cdots\sum_{s_d=d}^{s_{d-1}}\overline{\theta}(L)\,\epsilon_{s_d}$$

For example, for an ARIMA$(p,1,q)$, we have:

$$\begin{aligned} Y_t &= \lambda^0(t) + \omega^{k+1}(t) + \sum_{s=1}^{t}\overline{\theta}(L)\,\epsilon_s \\ &= \lambda_0 + \omega_0 + \omega_1 t + \omega_2 t^2 + \cdots + \omega_{k+1}t^{k+1} + \sum_{s=1}^{t}\sum_{l=1}^{\infty}\overline{\theta}_l\epsilon_{s-l} + \sum_{s=1}^{t}\epsilon_s \end{aligned}$$

where the intercept and the functions of t represent the deterministic trend, and the shocks represent both the stochastic trend and transitory effects.

Unlike in stationary ARMA models, where shocks only have a temporary effect on Y_t, the effect of a change in today's shock will persist until the infinite future in the ARIMA$(p,1,q)$ model:

$$\begin{aligned} &\text{Effect of } \epsilon_t \text{ on } \Delta Y_{t+\infty}: && \lim_{h\to\infty}\frac{\partial \Delta Y_{t+h}}{\partial \epsilon_t} = \lim_{h\to\infty}\overline{\theta}_h = 0 \\ &\text{Effect of } \epsilon_t \text{ on } Y_{t+\infty}: && \lim_{h\to\infty}\frac{\partial Y_{t+h}}{\partial \epsilon_t} = 1 + \sum_{l=1}^{\infty}\overline{\theta}_l = \overline{\theta}(1) \end{aligned}$$

where $\overline{\theta}(1)$ is the permanent or long-run effect. Therefore, $\overline{\theta}(1)\sum_{s=1}^{t}\epsilon_s$ represents the stochastic trend of $\{Y_t\}$, as it captures the permanent change in Y_t caused by past random shocks.

The **Beveridge-Nelson decomposition** (1981) separates this trend component from the temporary effects of the shocks. In particular, the Beveridge-Nelson decomposition of the ARIMA$(p,1,q)$ process is as follows:

$$Y_t = \underbrace{\lambda^0(t) + \omega^{k+1}(t) + \overline{\theta}(1)\sum_{s=1}^{t}\epsilon_s}_{g_t} + \underbrace{\tilde{\theta}(L)\epsilon_t}_{c_t}, \qquad \tilde{\theta}(L) = \left[\overline{\theta}(L) - \overline{\theta}(1)\right](1-L)^{-1}$$

where g_t is the trend component, and c_t captures the transitory effects of the shocks. More specifically, $\lambda^0(t)$ captures the permanent effects of the initial values, $\omega^{k+1}(t)$ represents the deterministic trend, $\overline{\theta}(1)\sum_{s=1}^{t}\epsilon_s$ represents the stochastic trend, and $\tilde{\theta}(L)\epsilon_t$ captures the temporary effects of the shocks.

To prove that the above Beveridge-Nelson decomposition indeed represents $\{Y_t\}$, we replace $\tilde{\theta}(L)$ with $\left[\overline{\theta}(L) - \overline{\theta}(1)\right](1-L)^{-1}$ and observe that we obtain the same expression for $\{Y_t\}$ as before. Note that $(1-L)^{-1}$ is the integral operator, i.e., $(1-L)^{-1}\epsilon_t = \sum_{s=1}^{t}\epsilon_s$. This is because $(1-L)^{-1}$ is the inverse of the difference operator $\Delta = (1-L)$, so we can verify this by multiplying both sides with the difference operator: $\epsilon_t = \sum_{s=1}^{t}\Delta\epsilon_s$ $= \epsilon_t - \epsilon_0 = \epsilon_t$, where $\epsilon_0 = 0$ is assumed.

By expanding the lag polynomial, we can find an expression for $\tilde{\theta}(L)$ as follows:

$$\begin{aligned}
\tilde{\theta}(L) &= \left[\overline{\theta}(L) - \overline{\theta}(1)\right](1-L)^{-1} \\
&= \left[\left(1 - \overline{\theta}(1)\right) + \overline{\theta}_1 L + \overline{\theta}_2 L^2 + \cdots\right]\left(1 + L + L^2 + \cdots\right) \\
&= 1 - \overline{\theta}(1) + \left(\overline{\theta}_1 + 1 - \overline{\theta}(1)\right)L + \left(\overline{\theta}_2 + \overline{\theta}_1 + 1 - \overline{\theta}(1)\right)L^2 + \cdots \\
&= -\sum_{i=1}^{\infty}\overline{\theta}_i - \sum_{j=2}^{\infty}\overline{\theta}_j L - \sum_{k=3}^{\infty}\overline{\theta}_k L^2 - \cdots \\
&= \tilde{\theta}_0 + \tilde{\theta}_1 L + \tilde{\theta}_2 L^2 + \cdots, \qquad \tilde{\theta}_j = -\sum_{i=j+1}^{\infty}\overline{\theta}_i
\end{aligned}$$

Here, we derive $(1-L)^{-1} = 1 + L + L^2 + \ldots$ by replacing the lag operator L with $|z| < 1$ and using the geometric series formula. After obtaining the series $1 + z + z^2 + \ldots$ in terms of z, we substitute z back with L to obtain $1 + L + L^2 + \ldots$.

To derive the Beveridge-Nelson decomposition of the ARIMA$(p, 1, q)$ process without relying on polynomial math, let's reconsider the following facts:

$$A. \qquad \epsilon_{s-l} = \epsilon_s - (\epsilon_s - \epsilon_{s-1}) - (\epsilon_{s-1} - \epsilon_{s-2}) - \cdots - \left(\epsilon_{s-(l-1)} - \epsilon_{s-l}\right) = \epsilon_s - \sum_{j=0}^{l-1} \Delta\epsilon_{s-j}$$

$$B. \qquad \sum_{s=1}^{t} \Delta\epsilon_{s-j} = (\epsilon_{t-j} - \epsilon_{t-1-j}) + \left(\epsilon_{t-1-j} - \epsilon_{t-1-(j+1)}\right) + \cdots + (\epsilon_{1-j} - \epsilon_{-j}) = \epsilon_{t-j} - \underbrace{\epsilon_{-j}}_{=0}$$

Using these facts, we can now rewrite the shocks of the process for $\{Y_t\}$ as follows:

$$\begin{aligned}
\sum_{s=1}^{t} \overline{\theta}(L)\,\epsilon_s &= \sum_{s=1}^{t}\sum_{l=1}^{\infty} \overline{\theta}_l \epsilon_{s-l} + \sum_{s=1}^{t} \epsilon_s \\
[A] \qquad &= \sum_{s=1}^{t}\sum_{l=1}^{\infty} \overline{\theta}_l \left(\epsilon_s - \sum_{j=0}^{l-1} \Delta\epsilon_{s-j}\right) + \sum_{s=1}^{t} \epsilon_s \\
&= \left(1 + \sum_{l=1}^{\infty} \overline{\theta}_l\right) \sum_{s=1}^{t} \epsilon_s - \sum_{s=1}^{t}\sum_{l=1}^{\infty} \overline{\theta}_l \sum_{j=0}^{l-1} \Delta\epsilon_{s-j} \\
&= \overline{\theta}(1) \sum_{s=1}^{t} \epsilon_s - \sum_{j=0}^{\infty} \left(\sum_{l=j+1}^{\infty} \overline{\theta}_l\right) \sum_{s=1}^{t} \Delta\epsilon_{s-j} \\
[B] \qquad &= \overline{\theta}(1) \sum_{s=1}^{t} \epsilon_s - \sum_{j=0}^{\infty} \left(\sum_{l=j+1}^{\infty} \overline{\theta}_l\right) \epsilon_{t-j} \\
&= \overline{\theta}(1) \sum_{s=1}^{t} \epsilon_s + \tilde{\theta}(L)\,\epsilon_t
\end{aligned}$$

In the last step, we introduce $\tilde{\theta}(L) = \tilde{\theta}_0 + \tilde{\theta}_1 L + \tilde{\theta}_2 L^2 + \ldots$, where $\tilde{\theta}_j = -\sum_{i=j+1}^{\infty} \overline{\theta}_i$. This yields the desired Beveridge-Nelson decomposition of the ARIMA$(p, 1, q)$ process.

Finally, we can recursively derive the Beveridge-Nelson decomposition of any ARIMA(p, d, q)

process as follows:

$$
\begin{aligned}
\Delta^{d} Y_t &= \gamma^{k}(t) + \overline{\theta}(L)\,\epsilon_t \\
&= \gamma^{k}(t) + \tilde{\theta}^{d}(L)\,\epsilon_t \\
\Delta^{d-1} Y_t &= \Delta^{d-1} Y_{d-1} + (1-L)^{-1}\,\Delta^{d} Y_t \\
&= \lambda^{0}(t) + \omega^{k+1}(t) + (1-L)^{-1}\,\tilde{\theta}^{d}(L)\,\epsilon_{s_d} \\
&= \lambda^{0}(t) + \omega^{k+1}(t) + \tilde{\theta}^{d}(1)\,(1-L)^{-1}\,\epsilon_t + \tilde{\theta}^{d-1}(L)\,\epsilon_t \\
\Delta^{d-2} Y_t &= \Delta^{d-2} Y_{d-2} + (1-L)^{-1}\,\Delta^{d-1} Y_t \\
&= \lambda^{1}(t) + \omega^{k+2}(t) + (1-L)^{-1}\,\tilde{\theta}^{d}(1)\,(1-L)^{-1}\,\epsilon_t + (1-L)^{-1}\,\tilde{\theta}^{d-1}(L)\,\epsilon_t \\
&= \lambda^{1}(t) + \omega^{k+2}(t) + \left[\tilde{\theta}^{d}(1)\,(1-L)^{-1} + \tilde{\theta}^{d-1}(1)\right](1-L)^{-1}\,\epsilon_t + \tilde{\theta}^{d-2}(L)\,\epsilon_t \\
\Delta^{d-3} Y_t &= \Delta^{d-3} Y_{d-3} + (1-L)^{-1}\,\Delta^{d-2} Y_t \\
&= \lambda^{2}(t) + \omega^{k+3}(t) + (1-L)^{-1}\left[\tilde{\theta}^{d}(1)\,(1-L)^{-1} + \tilde{\theta}^{d-1}(1)\right](1-L)^{-1}\,\epsilon_t + (1-L)^{-1}\,\tilde{\theta}^{d-2}(L)\,\epsilon_t \\
&= \lambda^{2}(t) + \omega^{k+3}(t) + \left[\tilde{\theta}^{d}(1)\,(1-L)^{-2} + \tilde{\theta}^{d-1}(1)\,(1-L)^{-1} + \tilde{\theta}^{d-2}(1)\right](1-L)^{-1}\,\epsilon_t + \tilde{\theta}^{d-3}(L)\,\epsilon_t \\
&\;\;\vdots \\
Y_t &= \underbrace{\lambda^{d-1}(t) + \omega^{k+d}(t) + \sum_{j=1}^{d} \tilde{\theta}^{j}(1)\,(1-L)^{-j}\,\epsilon_t}_{g_t} + \underbrace{\tilde{\theta}^{0}(L)\,\epsilon_t}_{c_t}
\end{aligned}
$$

Here, g_t represents the trend component, and c_t captures the transitory effects of the shocks. Specifically, $\lambda^{d-1}(t)$ captures the permanent effects of the initial values, $\omega^{k+d}(t)$ represents the deterministic trend, $\sum_{j=1}^{d} \tilde{\theta}^{j}(1)\,(1-L)^{-j}\,\epsilon_t$ represents the stochastic trend, and $\tilde{\theta}^{0}(L)\,\epsilon_t$ captures the temporary effects of the shocks. The integral operator $(1-L)^{-j}$ and the lag polynomial $\tilde{\theta}^{j}(L)$ are defined as follows:

$$
\begin{aligned}
(1-L)^{-j}\,\epsilon_t &= \sum_{s_1=d}^{t} \sum_{s_2=d}^{s_1} \cdots \sum_{s_{j-1}=d}^{s_{j-2}} \sum_{s_j=d}^{s_{j-1}} \epsilon_{s_j} \\
\tilde{\theta}^{j}(L) &= \tilde{\theta}^{j+1}(L)\,(1-L)^{-1} - \sum_{k=1}^{d-j} \tilde{\theta}^{j+k}(1)\,(1-L)^{-k} \qquad j = 0, \ldots, d-1
\end{aligned}
$$

and we have $\tilde{\theta}^{d}(L) = \overline{\theta}(L)$.

4.3 Dickey-Fuller Test

When analyzing time series data, it is important to determine the presence and nature of any underlying trend. However, it can be challenging to determine whether a time series is stationary, trend-stationary, or difference-stationary based solely on visual inspection. In such cases, the Dickey-Fuller test (1979) provides a statistical method to assess whether a time series exhibits a stochastic trend. This test helps in making informed decisions about the nature of the trend in the data.

The Dickey-Fuller test uses the insight that if a time series possesses a unit root, the differenced process ΔY_t should not depend on lagged level variables Y_{t-l}, but solely on lagged differenced variables ΔY_{t-l}. This is because when a unit root $r_p = 1$ exists, substituting $\left(1 - r_p^{-1} L\right) Y_t$ with ΔY_t eliminates the level variables entirely. Consequently, if Y_{t-1} is included in the regression, the coefficient on this variable should be zero if the process possesses a unit root.

To conduct the Dickey-Fuller test, we start with a potentially non-stationary ARMA(p, q) process:

$$\phi(L) Y_t = \delta^k(t) + \theta(L) \epsilon_t$$

$$Y_t = \delta^k(t) + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta(L) \epsilon_t$$

where $\phi(L)$ represents the AR lag polynomial, $\theta(L)$ represents the MA lag polynomial, and $\delta^k(t) = \delta_0 + \delta_1 t + \cdots + \delta_k t^k$ is a deterministic trend that, for simplicity, is assumed to be a polynomial in t of degree k. Note that if there is a unit root, by definition, one of the roots of the AR lag polynomial $\phi(L)$ is equal to one, i.e., $\phi(1) = 0$. This occurs when the AR coefficients add up to one, i.e., if $1 - \phi(1) = \sum_{l=1}^{p} \phi_l = 1$. In practice, the estimated coefficients will never exactly sum up to one. Therefore, a statistical test, such as the Dickey-Fuller test, is necessary to evaluate the hypothesis of a unit root.

Next, we relate the differenced variables to the lagged level variable Y_{t-1} by replacing the variables in the potentially non-stationary ARMA(p, q) process with the following functions

of Y_{t-1}:

$$Y_t = Y_{t-1} + \Delta Y_t$$

$$Y_{t-l} = Y_{t-1} - \Delta Y_{t-1} - \cdots - \Delta Y_{t-(l-1)} \quad l = 1, \ldots, p$$

which results in the following relationship between ΔY_t and Y_{t-1}:

$$\Delta Y_t = \delta^k(t) + \beta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \cdots + \gamma_{p-1} \Delta Y_{t-(p-1)} + \theta(L)\epsilon_t, \quad \beta = \phi_1 + \cdots + \phi_p - 1$$

$$\gamma_l = \phi_l + \cdots + \phi_p$$

where β represents the effect of Y_{t-1} on ΔY_t. If there is a unit root, i.e., $\sum_{l=1}^{p} \phi_l = 1$, then $\beta = 0$, indicating that ΔY_t does not depend on the level variable Y_t.

The above relationship suggests that we can reject the null hypothesis that there is a unit root if $\beta \neq 0$. However, that is not the case. As discussed in Section 4.2, a unit root not only produces a stochastic trend but also leads to a more sophisticated deterministic trend. If the time series is integrated of order d and follows a non-stationary ARMA model that contains a deterministic trend $\delta^k(t)$ modeled as a polynomial in t of degree k, the actual deterministic trend turns out to be a polynomial in t of degree $k + d$.

For instance, consider the case where the non-stationary ARMA model includes an intercept term. In this case, a unit root implies that the time series not only has a stochastic trend but also a linear deterministic trend. Therefore, when $\beta = 0$, it could either be a result of a stochastic trend (unit root) or a deterministic trend. In order to test for a stochastic trend only, the null hypothesis needs to be $H_0 : \beta = \delta_k = 0$. This formulation ensures that both the model under the null hypothesis and the alternative hypothesis $H_1 : \neg H_0$ have a deterministic trend that is a polynomial in t of degree k.

The Dickey-Fuller test can be applied to various specifications of deterministic trends. However, most software implementations typically consider three common cases:

1. $\delta^k(t) = 0$ with null hypothesis $H_0 : \beta = 0$,
2. $\delta^0(t) = \delta_0$ with null hypothesis $H_0 : \beta = \delta_0 = 0$, and
3. $\delta^1(t) = \delta_0 + \delta_1 t$ with null hypothesis $H_0 : \beta = \delta_1 = 0$.

These three cases allow for testing the presence of a unit root in the time series, assuming that the model has either no intercept, an intercept, or a linear trend.

For example, consider an AR(1) specification so that the test simplifies to the following regression:

$$Y_t = \delta_0 + \delta_1 t + \phi_1 Y_{t-1} + \epsilon_t$$

$$\Downarrow$$

$$\Delta Y_t = \delta_0 + \delta_1 t + \beta Y_{t-1} + \epsilon_t, \qquad \beta = \phi_1 - 1$$

Case 1 considers the case where $\delta_0 = \delta_1 = 0$, and compares the following two models:

$$H_1 : \beta \neq 0 \quad \Rightarrow \quad Y_t = \phi_1 Y_{t-1} + \epsilon_t = \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j} + \epsilon_t$$

$$H_0 : \beta = 0 \quad \Rightarrow \quad Y_t = Y_0 + \sum_{j=1}^{t} \Delta Y_j = Y_0 + \sum_{j=1}^{t} (\beta Y_{j-1} + \epsilon_j) = Y_0 + \sum_{j=1}^{t} \epsilon_j$$

Case 2 considers the case where $\delta_1 = 0$, and compares the following two models:

$$H_1 : \neg H_0 \quad \Rightarrow \quad Y_t = \frac{\delta_0}{1-\phi} + \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j} + \epsilon_t$$

$$H_0 : \beta = \delta_0 = 0 \quad \Rightarrow \quad Y_t = Y_0 + \sum_{j=1}^{t} (\delta_0 + \beta Y_{j-1} + \epsilon_j) = Y_0 + \sum_{j=1}^{t} \epsilon_j$$

Case 3 compares the following two models:

$$H_1 : \neg H_0 \quad \Rightarrow \quad Y_t = \frac{\delta_0(1-\phi) - \delta_1}{(1-\phi)^2} + \frac{\delta_1}{1-\phi} t + \sum_{j=1}^{\infty} \phi^j \epsilon_{t-j}$$

$$H_0 : \beta = \delta_1 = 0 \quad \Rightarrow \quad Y_t = Y_0 + \sum_{j=1}^{t} (\delta_0 + \delta_1 j + \beta Y_{j-1} + \epsilon_j) = Y_0 + \delta_0 t + \sum_{j=1}^{t} \epsilon_j$$

Hence, the choice of the appropriate case depends on the characteristics of the specific time series being analyzed. Case 1 should only be used if the time series has a mean of zero, which is typically not the case in economic data. Case 2 is applicable when the time series does not exhibit an upward or downward trend, and Case 3 allows for a linear trend in the

time series.

Under Case 1, the regression coefficient can be identified as follows:

$$\beta = \frac{\overline{\text{Cov}}(\Delta Y_t, Y_{t-1})}{\overline{\text{Var}}(Y_{t-1})} = \frac{\overline{\text{Cov}}(\beta Y_{t-1} + \epsilon_t, Y_{t-1})}{\overline{\text{Var}}(Y_{t-1})} = \frac{\beta \overline{\text{Var}}(Y_{t-1})}{\overline{\text{Var}}(Y_{t-1})} = \beta$$

Here, the bar on top of the variance (and covariance) indicates that it represents the sample variances (and covariances) across time. It is important to note that $\overline{\text{Var}}(Y_{t-1})$ does not compute the variance of the random variable Y_{t-1}, which may change over time if there is a unit root.

In the presence of a unit root, the variance of Y_{t-1} across time, denoted as $\overline{\text{Var}}(Y_{t-1})$, differs from the time-varying variance $\text{Var}_{t-1}(Y_{t-1})$. As a result, the t-statistic of $\hat{\beta}$ does not follow a standard Student-t distribution. Instead, it follows a nonstandard density function known as the **Dickey-Fuller distribution**, which does not have a tractable mathematical expression. Therefore, critical values for the t-statistic under the Dickey-Fuller distribution are typically obtained from tables or computed using simulation techniques. Researchers often refer to these tables to determine the critical values for specific significance levels, such as $\alpha = 0.05$, to test the null against the alternative hypothesis.

While the above test for an AR(1) model is known as the Dickey-Fuller test, the **augmented Dickey-Fuller test** allows for the inclusion of additional AR and MA lags. Incorporating additional AR lags can be achieved by including lagged differences ΔY_{t-l} in the regression equation. However, estimating MA coefficients can be more challenging in practice. To address this issue, a common approach is to eliminate the MA lag polynomial and introduce additional AR lags instead. This practice is justified by our earlier discussion in Section 3.4, where we demonstrated that an invertible ARMA process can be represented as an infinite-order AR process, with AR coefficients that exhibit exponential decay. Due to the exponential decay of the AR coefficients, truncating the AR lag polynomial at a large enough but finite number is unlikely to significantly affect the results. Consequently, replacing the MA coefficients with additional AR coefficients is expected to have a small

impact on the overall outcome. In practice, the determination of the number of AR lags is typically determined using an information criterion.

5 Linear Algebra

Before delving into multivariate time series models, it is essential to introduce linear algebra tools that complement the polynomial math tools we have used thus far.

5.1 Leveraging Linear Algebra for Computing the Reciprocal of a Polynomial

Let's reconsider a polynomial of degree p:

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_p z^p$$

Previously, we computed the reciprocal $f(z)^{-1}$ by factoring it with a normalized constant $= 1$:

$$f(z) = a_0 \left(1 - r_1^{-1} z\right) \cdots \left(1 - r_n^{-1} z\right) = a_0 \prod_{k=1}^{n} \left(1 - r_k^{-1} z\right)$$

We then applied the geometric series formula to each inverted factor and expanded the infinite sums using the Cauchy integral formula. This resulted in the Laurent series representation:

$$f(z)^{-1} = c_0 + c_1 z + c_2 z^2 + \cdots + c_{-1} z^{-1} + c_{-2} z^{-2} + \cdots$$

where $c_{-i} = 0$ for all $i \geq 1$ if the inverted roots have magnitudes smaller than one.

Now, we introduce a different method that utilizes linear algebra to derive the same reciprocal. The reason for introducing this method is that, unlike the previous approach, the linear algebra approach is applicable when working with matrix polynomials, which will be used to compute various aspects of multivariate time series models.

The approach is to utilize matrix algebra to represent the polynomial as a single factor

instead of multiple factors, one for each root. To achieve this, we can express the polynomial using matrix notation as follows:

$$\begin{bmatrix} f(z) \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} = a_0 \begin{bmatrix} 1 \\ z \\ \vdots \\ z^{p-2} \\ z^{p-1} \end{bmatrix} - a_0 \begin{bmatrix} -\frac{a_1}{a_0} & -\frac{a_2}{a_0} & \cdots & -\frac{a_{p-1}}{a_0} & -\frac{a_p}{a_0} \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} z \\ z^2 \\ \vdots \\ z^{p-1} \\ z^p \end{bmatrix}$$

$$\begin{aligned} \overline{f}(z) &= a_0\rho(z) - a_0 C z \rho(z) \\ &= a_0(I_p - Cz)\rho(z) \\ &= G(z)\rho(z), \qquad G(z) = a_0(I_p - Cz) \end{aligned}$$

where C is the **companion matrix**, characterized by parameters on the first row and an identity matrix starting on the second row and ending on the second-to-last column. The $p \times p$ matrix $G(z)$ represents the transformation of the $p \times 1$ polynomial vector $\rho(z)$ to generate the polynomial $f(z)$, and I_p is a $p \times p$ identity matrix. By expressing the polynomial as a transformation $G(z)$ of the polynomial vector $\rho(z)$, we can represent the polynomial $f(z)$ of degree p, which consists of p factors $\left(1 - r_i^{-1} z\right)$ for $i = 1, \ldots, p$, as a polynomial with a single factor $(I_p - Cz)$.

The reciprocal $f(z)^{-1}$, defined as $f(z)^{-1} f(z) = 1$, can then be computed as the top-left entry of the inverse matrix $G(z)^{-1}$:

$$\begin{aligned} \overline{f}(z) &= G(z)\rho(z), \qquad [\rho(z)]_{1,1} = 1 \\ G(z)^{-1}\overline{f}(z) &= \rho(z) \\ f(z)^{-1} &= \left[G(z)^{-1}\right]_{1,1} \end{aligned}$$

where $[M]_{i,j}$ is an operator that extracts the ith row and jth column of matrix M.

Under the assumption that all inverted roots of the polynomial $f(z)$ lie inside the unit circle and $|z| \leq 1$, we can compute the inverse matrix $G(z)^{-1}$ using the geometric series

formula applied to matrices:

$$\begin{aligned} G(z)^{-1} &= a_0^{-1}(I_p - Cz)^{-1} \\ &= a_0^{-1}\left(I_p + Cz + C^2z^2 + C^3z^3 + \cdots\right) \end{aligned}$$

which leads to the calculation of $f(z)^{-1}$ as described earlier.

However, when we extend this concept to matrix polynomials, we can no longer directly compute the inverted roots of $f(z)$. Instead, we need an alternative method to determine whether the geometric series formula applies. To do this, we introduce the concept of eigenvalues. We can then argue that if the eigenvalues of the companion matrix C are inside the unit circle, which in this example correspond to the inverted roots, then C^k approaches zero as k approaches infinity. Consequently, the geometric series formula can be applied.

5.2 Eigenvalues, Eigenvectors, and Determinant

Consider a $p \times p$ square matrix M. In linear algebra, we can interpret M as a linear transformation that acts on a $p \times 1$ vector v, resulting in a new vector $w = Mv$. If the vector v is aligned in a specific direction, the transformation Mv is equivalent to simply scaling the vector by a scalar λ. This special vector v is known as an **eigenvector** of M, and the scalar λ is referred to as the **eigenvalue** of M. It is defined by the following equation:

$$Mv = \lambda v$$

where $\lambda \in \mathbb{C}$ is a complex scalar and $v \in \mathbb{C}^p$ is a $p \times 1$ vector in complex space. Therefore, the matrix-vector product Mv yields the same result as scaling the eigenvector v by the scalar λ, which represents the eigenvalue. From this it follows that if the matrix is applied k times, then that corresponds to scaling the vector v by λ^k. The equation typically has multiple solutions, indicating that matrix M can have multiple eigenvalues and corresponding eigenvectors.

Moreover, if we apply the matrix M k times to the vector v, it corresponds to scaling the eigenvector v by the eigenvalue raised to the power of k

$$M^2v = M(Mv) = \lambda(\lambda v) = \lambda^2 v \quad \Rightarrow \quad M^k v = \lambda^k v$$

which property holds for each eigenvalue and its associated eigenvector.

Why is this relevant? Well, the behavior of eigenvalues and eigenvectors provides important insights into the properties of a matrix transformation. In particular, if all eigenvalues of matrix M are inside the unit circle, then all corresponding eigenvectors will be scaled down and approach the zero vector as the matrix is applied multiple times. In other words, if the eigenvalues and eigenvectors are inside the unit circle, the matrix M^k tends to zero as k approaches infinity. This property has important implications in understanding the long-term behavior and stability of systems described by matrix transformations.

In our context, the relevance of eigenvalues and eigenvectors lies in determining whether the geometric series formula can be applied. Specifically, if the companion matrix C satisfies the condition that applying it k times (C^k) tends to zero as k approaches infinity, then we can utilize the geometric series formula. Therefore, by examining the eigenvalues of C, we can determine whether the conditions for applying the geometric series formula are met.

It turns out that the eigenvalues of the companion matrix C correspond to the inverted roots of the polynomial $f(z)$, while the eigenvectors are the polynomial vectors $\rho(z)$ computed at those roots:

$$\begin{aligned} f(r) = 0 \quad \Rightarrow \quad \overline{f}(r) &= a_0(I_p - Cr)\rho(r) = 0 \\ C\rho(r) &= r^{-1}\rho(r) \\ Cv &= \lambda v \end{aligned}$$

Hence, the eigenvalues of C satisfy the equation $f\left(\lambda^{-1}\right) = 0$, and the corresponding eigenvectors are obtained as $v = \rho\left(\lambda^{-1}\right)$. Consequently, if all inverted roots of $f(z)$ lie inside the unit circle, it follows that all eigenvalues of C are also inside the unit circle. As a result,

C^k tends to zero as k approaches infinity.

While the computation of eigenvalues for a companion matrix is straightforward, as shown above, the process for a general matrix is more involved. In general, one can determine the eigenvectors of a matrix M by iteratively testing different vectors until finding one that remains within its own span when M is applied. The corresponding eigenvalues are then calculated by measuring the scaling factor applied to these eigenvectors. However, there is a more efficient method that allows for the direct computation of eigenvalues without initially knowing the eigenvectors. To do this, we introduce the concept of the **determinant** of a matrix M, denoted as $\det(M)$. The determinant quantifies the extent to which the transformation M stretches or compresses space. It represents the factor by which the area (or volume, mass) of a given region in space changes when subjected to the transformation M. A negative determinant indicates a reversal in orientation of the region, while a determinant of zero suggests that M compresses space into a lower dimension. Hence, if $\det(M) = 0$, it implies that the matrix M is not invertible, meaning that M^{-1} does not exist. This is because a transformation can map from a higher dimension to a lower dimension, but not vice versa.

Note that since scaling a region by 2 and then by 3 is the same as scaling a region by $2 \times 3 = 6$, we have that the determinant is **multiplicative**, i.e., $\det(AB) = \det(A)\det(B)$. Similarly, if the matrix M scales space by a factor of $\det(M)$, then the inverse transformation M^{-1} reverses the operation and thus scales space by a factor of $\det(M)^{-1}$. Hence, we can conclude that $\det\left(M^{-1}\right) = \det(M)^{-1}$.

Given that the eigenvector v cannot be the zero vector, we can compute the eigenvalues λ of M by setting the determinant of $M - \lambda I_p$ to zero:

$$Mv = \lambda v$$

$$(M - \lambda I_p)v = 0$$

$$\det(M - \lambda I_p) = 0$$

The last equation is derived from the fact that if $\det(M - \lambda I_p) \neq 0$, then $(M - \lambda I_p)$ is invertible, and therefore, the eigenvector v must be the zero vector, $v = (M - \lambda I_p)^{-1} 0 = 0$, which is not possible. Hence, we were able to find an equation that allows us to compute the eigenvalues without requiring knowledge of the eigenvectors. This provides a convenient and efficient method for determining the eigenvalues of a matrix.

Algebraically, the eigenvalues of a $p \times p$ matrix M are the p roots of its **characteristic polynomial** $\kappa(z)$, which is defined as the determinant of $(M - zI_p)$. The characteristic polynomial is a polynomial of degree p and can be written as:

$$\kappa(z) = \det(M - zI_p) = \alpha_0 + \alpha_1 z + \cdots + \alpha_p z^p \quad \Rightarrow \quad \kappa(\lambda) = 0$$

The coefficients of the characteristic polynomial have a specific relationship with the matrix M. In particular, α_0 is equal to the determinant of M, α_p is equal to $(-1)^p$, and α_{p-1} is equal to $(-1)^p$ times the **trace** of M, denoted by $\mathbf{tr}(M)$. The trace of a square matrix M is obtained by summing the diagonal entries of M.

For the specific case of a 2×2 matrix, the characteristic polynomial can be expressed as:

$$\kappa(z) = \det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} - zI_2\right) = \det\begin{bmatrix} a-z & b \\ c & d-z \end{bmatrix} = (a-z)(d-z) - bc$$

$$= \underbrace{ad - bc}_{\alpha_0} - \underbrace{(a+d)}_{\alpha_1} z + \underbrace{1}_{\alpha_2} z^2$$

By computing the roots of the characteristic polynomial $\kappa(z)$ through the equation $\kappa(\lambda) = 0$, we can find the two eigenvalues $\{\lambda_1, \lambda_2\}$ of the 2×2 matrix.

For the specific case of an upper (or lower) triangular matrix U, the determinant is the product of the main diagonal entries, and therefore the eigenvalues of an upper (or lower)

triangular matrix are precisely its diagonal entries:

$$\kappa(\lambda) = \det(U - \lambda I_p) = \det\left\{\begin{bmatrix} u_{11}-\lambda & u_{12} & \cdots & u_{1p} \\ 0 & u_{22}-\lambda & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{pp}-\lambda \end{bmatrix}\right\} = 0$$

$$(u_{11}-\lambda)(u_{22}-\lambda)\cdots(u_{pp}-\lambda) = 0$$

$$\lambda_i = u_{ii}, \quad i = 1, \ldots, p$$

Given this property for an upper (or lower) triangular matrix U, and the fact that the determinant of a matrix has the property of multiplicativity, one can compute the eigenvalues of a general matrix M by first performing the Schur decomposition $M = QUQ^{-1}$ (see next section for details), where U is an upper triangular matrix and Q is a unitary matrix. Then, the determinant of M can be expressed as $\det(M) = \det(Q)\det(U)\det(Q)^{-1} = \det(U)$, where $\det(U)$ is computed as described above.

5.3 Eigendecomposition, Jordan Decomposition, and Schur decomposition

We have observed that eigenvalues play a crucial role in determining the stability of a matrix, indicating whether it will diverge to infinity or not as it is applied repeatedly. Moreover, eigenvalues and eigenvectors enable us to perform an eigendecomposition of diagonalizable matrices. This decomposition provides efficient means to compute the kth power of a matrix and even handle fractional powers M^k when k is not an integer. In cases where a matrix is not diagonalizable, the Jordan or Schur decomposition offer valuable alternatives to the eigendecomposition.

Suppose the $p \times p$ matrix M is **diagonalizable** or **non-defective**, meaning that the p eigenvectors $\{v_1, \ldots, v_p\}$ are linearly independent. Then any $p \times 1$ vector w can be written

as a linear combination of the p eigenvectors:

$$w = \gamma_1 v_1 + \cdots + \gamma_p v_p = \begin{bmatrix} v_1 & \cdots & v_p \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix} = V\gamma \quad \Rightarrow \quad \gamma = V^{-1}w$$

Thus, applying transformation M^k to that vector w can be written as a linear combination of the eigenvectors scaled by the kth power of M's eigenvalues $\{\lambda_1, \ldots, \lambda_p\}$:

$$\begin{aligned} M^k w &= \gamma_1 M^k v_1 + \cdots + \gamma_p M^k v_p \\ &= \gamma_1 \lambda_1^k v_1 + \cdots + \gamma_p \lambda_p^k v_p = \begin{bmatrix} v_1 & \cdots & v_p \end{bmatrix} \begin{bmatrix} \lambda_1^k & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p^k \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix} = V\Lambda^k\gamma \end{aligned}$$

Finally, since $\gamma = V^{-1}w$, we can compute the **eigendecomposition** of M^k as follows:

$$M^k = V\Lambda^k V^{-1}, \quad V = \begin{bmatrix} v_1 & \cdots & v_p \end{bmatrix}, \quad \Lambda^k = \begin{bmatrix} \lambda_1^k & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p^k \end{bmatrix}$$

It is worth noting that k does not need to be an integer, as the kth power of the diagonal matrix Λ can be expressed as a diagonal matrix of eigenvalues, each raised to the power of k.

If the eigenvectors are not linearly independent, then V^{-1} doesn't exist, and therefore M does not have an eigendecomposition, that is, M is **defective**. In such cases, alternative decomposition methods like the Jordan or Schur decomposition are employed to analyze and compute the powers of the matrix. These methods provide a similar framework to the eigendecomposition but can handle matrices that are not diagonalizable. The **Jordan decomposition** involves transforming the matrix into its **Jordan canonical form** J,

which consists of blocks with eigenvalues on the diagonal and ones on the superdiagonal:

$$M = PJP^{-1}$$

where J is an upper diagonal matrix. If M is diagonalizable, the decomposition becomes the eigendecomposition with $P = V$ and $J = \Lambda$. The **Schur decomposition**, on the other hand, expresses the matrix as the product of a unitary matrix Q, an upper triangular matrix U, and the (conjugate) transpose Q':

$$M = QUQ', \quad Q'Q = I_p, \quad Q' = Q^{-1}$$

These decompositions allow for a more generalized analysis of matrix powers and properties, even for matrices that are defective. The advantage of the Jordan decomposition is that it simplifies to the eigendecomposition when the matrix is non-defective, and the advantage of the Schur decomposition is that it is faster to compute.

5.4 Reciprocal of a Matrix Polynomial

In Section 5.1, we explored the computation of the reciprocal of a polynomial using linear algebra, assuming that all the inverted roots of the polynomial are within the unit circle. In this section, we will expand our analysis to handle scenarios where some roots are inside the unit circle while others are outside. Additionally, we will extend our focus from scalar polynomials to matrix polynomials.

Let's consider a matrix polynomial of degree p:

$$F(z) = A_0 + A_1 z + A_2 z^2 + \cdots + A_p z^p$$

where A_i are $n \times n$ matrices of coefficients, and z is a scalar variable. As in Section 5.1, we use matrix algebra to represent the polynomial as a single factor. To achieve this, we

can express the polynomial using matrix notation as follows:

$$\begin{bmatrix} F(z) \\ 0 \\ \vdots \\ 0 \end{bmatrix} = (I_p \otimes A_0) \begin{bmatrix} I_n \\ zI_n \\ \vdots \\ z^{p-1}I_n \end{bmatrix} - (I_p \otimes A_0) \begin{bmatrix} -A_0^{-1}A_1 & -A_0^{-1}A_2 & \cdots & -A_0^{-1}A_p \\ I_n & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} zI_n \\ z^2I_n \\ \vdots \\ z^pI_n \end{bmatrix}$$

$$\begin{aligned} \overline{F}(z) &= (I_p \otimes A_0)\tau(z) - (I_p \otimes A_0)Cz\tau(z) \\ &= (I_p \otimes A_0)(I_{np} - Cz)\tau(z) \\ &= G(z)\tau(z), \qquad G(z) = (I_p \otimes A_0)(I_{np} - Cz) \end{aligned}$$

where C is the **multi-companion matrix** or **n-companion matrix**, characterized by parameters on the first n rows and an identity matrix starting on the $(n+1)$th row and ending on the $n(p-1)$th column. The $np \times np$ matrix $G(z)$ represents the transformation of the $np \times n$ matrix $\tau(z) = \rho(z) \otimes I_n$ to generate the matrix polynomial $F(z)$, and $(I_p \otimes A_0)$ is the Kronecker product of the $p \times p$ identity matrix I_p with the $n \times n$ matrix of coefficients A_0; thus, $(I_p \otimes A_0)$ is a block-diagonal matrix with p blocks, each being A_0. Therefore, by expressing the matrix polynomial as the transformation $G(z)$ applied to the matrix $\tau(z)$, we can represent it as a polynomial with a single factor $(I_{np} - Cz)$.

The reciprocal $F(z)^{-1}$, defined as $F(z)^{-1}F(z) = I_n$, can be computed as the top-left $n \times n$ block of the inverse matrix $G(z)^{-1}$:

$$\begin{aligned} \overline{F}(z) &= G(z)\tau(z), \qquad [\tau(z)]_{(1:n),(1:n)} = I_n \\ G(z)^{-1}\overline{F}(z) &= \tau(z) \\ F(z)^{-1} &= \left[G(z)^{-1}\right]_{(1:n),(1:n)} \end{aligned}$$

where $[M]_{(a:b),(c:d)}$ is an operator that extracts rows a to b, and columns c to d of matrix M.

Under the assumption that all eigenvalues of the multi-companion matrix C lie inside the unit circle and $|z| \leq 1$, the inverse matrix $G(z)^{-1}$ can be computed using the geometric

series formula:

$$\begin{aligned} G(z)^{-1} &= (I_p \otimes A_0)^{-1} (I_{np} - Cz)^{-1} \\ &= (I_p \otimes A_0)^{-1} \left(I_{np} + Cz + C^2 z^2 + C^3 z^3 + \cdots\right) \end{aligned}$$

This computation leads to the calculation of $F(z)^{-1}$ as described earlier.

In Section 5.2, we demonstrated that when the coefficients of the polynomial $F(z)$ are scalars, the eigenvalues of the companion matrix C correspond to the inverted roots of the polynomial $F(z)$. However, when the coefficients are $n \times n$ matrices, equating $F(z)$ to zero yields n^2 separate equations, which do not provide meaningful information. Nonetheless, we can establish that the eigenvalues of the companion matrix C correspond to the inverted roots of the polynomial representing the determinant of $F(z)$:

$$\det(F(z)) = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_{np} z^{np}$$

This polynomial in z has a degree of np since each entry in $F(z)$ is a polynomial of degree p. The determinant includes additive terms where up to n entries in $F(z)$ are multiplied with each other, resulting in powers of z up to np. This means that $\det(F(z))$ has np roots, and their inverses correspond to the np eigenvalues of $np \times np$ companion matrix C.

To prove that the inverses of the np roots of the polynomial $\det(F(z))$ correspond to the np eigenvalues of $np \times np$ companion matrix C, we use the fact that a determinant of zero $\det(F(r)) = 0$ implies that the transformation $F(r)$ compresses space into a lower dimension. Therefore, the null space of $F(r)$ is at least one-dimensional, indicating the existence of a non-zero vector $w(r)$ that transforms to the zero vector:

$$\det(F(r)) = 0 \quad \Rightarrow \quad F(r)\, w(r) = 0$$

The vector $w(r)$ depends on $F(r)$ and, consequently, on the root r.

Using this fact, we can compute the eigenvectors and eigenvalues of C as follows:

$$\begin{aligned}
F(r)\, w(r) &= 0 \\
\overline{F}(r)\,(1_n \otimes w(r)) &= 0 \\
(I_p \otimes A_0)\,(I_{np} - Cz)\,\tau(z)\,(1_n \otimes w(z)) &= 0 \\
(I_p \otimes A_0)\,(I_{np} - Cz)\,(\rho(r) \otimes w(r)) &= 0 \\
C\,(\rho(r) \otimes w(r)) &= r^{-1}\,(\rho(r) \otimes w(r)) \\
Cv = \lambda v, \quad \lambda = r^{-1}, \quad v &= \rho(r) \otimes w(r)
\end{aligned}$$

Hence, the eigenvalues of C satisfy the equation $\det\left(F\left(\lambda^{-1}\right)\right) = 0$, and the corresponding eigenvectors are obtained as $v = \rho\left(\lambda^{-1}\right) \otimes w\left(\lambda^{-1}\right)$. Consequently, if all inverted roots of $\det(F(z))$ lie inside the unit circle, it follows that all eigenvalues of C are also inside the unit circle. As a result, C^k tends to zero as k approaches infinity.

So far, we have shown how to compute the reciprocal $F(z)^{-1}$ when $|z| \leq 1$ and when the inverted roots of the polynomial $\det(F(z))$ are inside the unit circle. Under these assumptions, we can use the geometric series formula to compute the reciprocal as an infinite sum of positive powers of z. However, to generalize the computation for any z and any set of roots, we need to handle cases where the roots may lie both inside and outside the unit circle. To achieve this, we can apply the Jordan decomposition to the matrix C.

In the Jordan decomposition, the eigenvalues of C are ordered such that they increase in size. We can express the Jordan canonical form J as blocks corresponding to eigenvalues that satisfy $|\lambda_i z| < 1$ and $|\lambda_j z| > 1$:

$$C = PJP^{-1} = P \begin{bmatrix} J_{11} & J_{12} \\ 0 & J_{22} \end{bmatrix} P^{-1}$$

In this decomposition, the diagonals of the $m \times m$ matrix J_{11} correspond to eigenvalues λ_i that satisfy $|\lambda_i z| < 1$, while the diagonals of the $(np - m) \times (np - m)$ matrix J_{22} correspond to eigenvalues λ_j that satisfy $|\lambda_j z| > 1$.

To compute the reciprocal $F(z)^{-1}$, which is the top-left $n \times n$ block in $G(z)^{-1}$, we rewrite $G(z)^{-1}$ in terms of the Jordan canonical form J of C:

$$\begin{aligned} G(z)^{-1} &= (I_p \otimes A_0)^{-1} (I_{np} - Cz)^{-1} \\ &= (I_p \otimes A_0)^{-1} \left(I_{np} - PJP^{-1}z\right)^{-1} \\ &= (I_p \otimes A_0)^{-1} P^{-1} (I_{np} - Jz)^{-1} P \end{aligned}$$

Next, we need to determine the factor $(I_{np} - Jz)^{-1}$ in $G(z)^{-1}$. By writing the Jordan canonical form J as block-upper triangular matrix as defined above, we can compute $(I_{np} - Jz)^{-1}$ as an infinite sum of both positive and negative powers of z:

$$\begin{aligned} (I_{np} - Jz)^{-1} &= \left(I_{np} - \begin{bmatrix} J_{11} & J_{12} \\ 0 & J_{22} \end{bmatrix} z \right)^{-1} \\ &= \begin{bmatrix} I_m - J_{11}z & J_{12}z \\ 0 & I_{np-m} - J_{22}z \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (I_m - J_{11}z)^{-1} & -(I_m - J_{11}z)^{-1} J_{12}z (I_{np-m} - J_{22}z)^{-1} \\ 0 & (I_{np-m} - J_{22}z)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} (I_m - J_{11}z)^{-1} & (I_m - J_{11}z)^{-1} J_{12}z J_{22}^{-1} z^{-1} \left(I_{np-m} - J_{22}^{-1}z^{-1}\right)^{-1} \\ 0 & -J_{22}^{-1}z^{-1} \left(I_{np-m} - J_{22}^{-1}z^{-1}\right)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \left(I_m + J_{11}z + J_{11}^2z^2 + \cdots\right) & \left(I_m + J_{11}z + J_{11}^2z^2 + \cdots\right) J_{12}z \left(J_{22}^{-1}z^{-1} + J_{22}^{-2}z^{-2} + \cdots\right) \\ 0 & -\left(J_{22}^{-1}z^{-1} + J_{22}^{-2}z^{-2} + \cdots\right) \end{bmatrix} \end{aligned}$$

Here, the geometric series formula is used to obtain the infinite sum, and the following property for a block-upper triangular matrix is used to compute the inverse:

$$\begin{bmatrix} A & X \\ 0 & B \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}XB^{-1} \\ 0 & B^{-1} \end{bmatrix}$$

Hence, we have a solution for $G(z)^{-1}$, expressed as an infinite sum of both positive and negative powers of z. This allows us to compute the reciprocal $F(z)$ for any value of z and for any set of roots, even when the roots may lie both inside and outside the unit circle.

Note that if the companion matrix C is diagonalizable, then $J_{12} = 0$, and the Jordan decomposition simplifies to the eigendecomposition. It is also worth mentioning that the Jordan decomposition can be replaced with the Schur decomposition, where J is replaced with the corresponding upper triangular matrix U and P is replaced with a unitary matrix Q. The Schur decomposition can be computed faster than the Jordan decomposition, making it a more efficient option in most cases. In fact, any decomposition of the form $C = B_1 B_2 B_3$ can be used, as long as B_2 is an upper triangular matrix and the determinants of B_1 and B_3 satisfy $\det(B_1)\det(B_3) = 1$. Different decompositions may have their own advantages depending on the specific problem at hand.

6 Multivariate Stationary Time Series

6.1 Vector Autoregression (VAR)

A **vector autoregression of order p**, **VAR(p)**, for two stochastic processes $\{X_t\}$ and $\{Z_t\}$, is defined as a system of $n = 2$ equations where the regressors are the lagged values of $\{X_t\}$ and $\{Z_t\}$:

$$X_t = \delta_1 + \alpha_{11} X_{t-1} + \cdots + \alpha_{1p} X_{t-p} + \beta_{11} Z_{t-1} + \cdots + \beta_{1p} Z_{t-p} + v_t$$
$$Z_t = \delta_2 + \alpha_{21} X_{t-1} + \cdots + \alpha_{2p} X_{t-p} + \beta_{21} Z_{t-1} + \cdots + \beta_{2p} Z_{t-p} + w_t$$

where the errors v_t and w_t can be contemporaneously correlated, i.e. $\text{Cor}(v_t, w_t) \neq 0$. This correlation arises because Z_t does not appear as a regressor in the equation for X_t, and vice versa. Consequently, the unexplained fluctuations in Z_t, represented by the error term w_t, are not controlled for in the equation for X_t, leading to the presence of these fluctuations in the residual v_t of X_t. As a result, the residuals v_t and w_t can be correlated, indicating contemporaneous dependence between X_t and Z_t beyond their lagged relationships.

To write the VAR(p) more efficiently, define Y_t as a vector of n random variables, and

u_t as a vector of n residuals:

$$Y_t = \begin{bmatrix} Y_{1t} \\ \vdots \\ Y_{nt} \end{bmatrix}, \quad u_t = \begin{bmatrix} u_{1t} \\ \vdots \\ u_{nt} \end{bmatrix}, \qquad \text{e.g.:} \quad Y_t = \begin{bmatrix} X_t \\ Z_t \end{bmatrix}, \quad u_t = \begin{bmatrix} v_t \\ w_t \end{bmatrix}$$

and then write the system of equations using matrix notation:

$$Y_t = \delta + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + u_t$$

e.g. :

$$\begin{bmatrix} X_t \\ Z_t \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \beta_{11} \\ \alpha_{21} & \beta_{21} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Z_{t-1} \end{bmatrix} + \cdots + \begin{bmatrix} \alpha_{1p} & \beta_{1p} \\ \alpha_{2p} & \beta_{2p} \end{bmatrix} \begin{bmatrix} X_{t-p} \\ Z_{t-p} \end{bmatrix} + \begin{bmatrix} v_t \\ w_t \end{bmatrix}$$

In this representation, δ is an $n \times 1$ vector of coefficients, and Φ_l is an $n \times n$ matrix of coefficients for each lag l. This compact notation allows us to treat the VAR(p) system similarly to univariate AR(p) models, enabling the application of familiar tools and techniques for analysis.

The VAR(p) residual covariance matrix Ω is an $n \times n$ matrix that captures the linear dependencies among the contemporaneous residuals:

$$\Omega = \begin{bmatrix} \operatorname{Var}(u_{1t}) & \operatorname{Cov}(u_{1t}, u_{2t}) & \cdots & \operatorname{Cov}(u_{1t}, u_{nt}) \\ \operatorname{Cov}(u_{2t}, u_{1t}) & \operatorname{Var}(u_{2t}) & \cdots & \operatorname{Cov}(u_{2t}, u_{nt}) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(u_{nt}, u_{1t}) & \operatorname{Cov}(u_{nt}, u_{2t}) & \cdots & \operatorname{Var}(u_{nt}) \end{bmatrix}$$

e.g. :

$$\Omega = \begin{bmatrix} \operatorname{Var}(v_t) & \operatorname{Cov}(v_t, w_t) \\ \operatorname{Cov}(w_t, v_t) & \operatorname{Var}(w_t) \end{bmatrix}$$

The covariance between two random variables is not affected by the order of the variables, i.e. $\operatorname{Cov}(u_{it}, u_{jt}) = \operatorname{Cov}(u_{jt}, u_{it})$. This property ensures that the covariance matrix Ω is

symmetric, meaning that the number in the ith row and jth column is equal to the number in the jth row and ith column. The off-diagonal entries in Ω measure the contemporaneous relationships between variables, capturing dependencies that go beyond the lagged VAR regressors.

The **companion form of a VAR(p)** expresses the VAR(p) as a VAR(1) as follows:

$$\mathbf{Y}_t = c + C\mathbf{Y}_{t-1} + \mathbf{u}_t$$

$$\begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-(p-2)} \\ Y_{t-(p-1)} \end{bmatrix} = \begin{bmatrix} \delta \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ I_n & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & I_n & 0 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-(p-1)} \\ Y_{t-p} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

Here, I_n represents an $n \times n$ identity matrix, and c is an $np \times 1$ vector of coefficients, where n is the number of variables, and p is the number of lags. The matrix C, known as the **multi-companion matrix** or **n-companion matrix**, has a size of $np \times np$. It is characterized by parameters on the first n rows and an identity matrix starting from the $(n+1)$th row and ending at the $n\,(p-1)$th column. This compact form provides an equivalent representation of the VAR(p) model in terms of a VAR(1) model with coefficient matrices c and C.

The lag operator L can be extended to apply to vectors, allowing us to express lagged values of a vector Y_t. Specifically, $LY_t = Y_{t-1}$ represents shifting all variables in Y_t one period back, and $L^k Y_t = Y_{t-k}$ represents shifting them k periods back. Using the lag operator, the VAR(p) model can be expressed using lag polynomials, similar to how we represent an AR(p) model:

$$\Phi\,(L)\,Y_t = \delta + u_t, \qquad \Phi\,(L) = I_n - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p,$$
$$C\,(L)\,\mathbf{Y}_t = c + \mathbf{u}_t, \qquad C\,(L) = I_{np} - CL$$

Here, $\Phi(L)$ and $C(L)$ are matrix polynomials in the lag operator L. To analyze and compute properties of $\Phi(L)$ and $C(L)$, we can replace the lag operator with a scalar variable z belonging to the complex numbers $\mathbb{C}$. By performing polynomial operations on $\Phi(L)$ and $C(L)$, such as computing the reciprocal, we can derive various properties and alternative representations of the VAR(p) model. After completing the polynomial operations, we can substitute z back with the lag operator L to express the results in terms of lagged variables.

Similar to deriving the MA representation of an AR model, let's iterate on the companion form of the VAR(p) and try to compute an infinite-order **vector moving average process**, **VMA(∞)**:

$$
\begin{aligned}
\mathbf{Y}_t &= c + C\mathbf{Y}_{t-1} + \mathbf{u}_t \\
&= c + C\left(c + C\mathbf{Y}_{t-2} + \mathbf{u}_{t-1}\right) + \mathbf{u}_t \\
&= c + C\left(c + C\left(c + C\mathbf{Y}_{t-3} + \mathbf{u}_{t-2}\right) + \mathbf{u}_{t-1}\right) + \mathbf{u}_t \\
&\vdots \\
&= \left(I_n + C + C^2 + C^3 + \cdots\right) c + C\mathbf{u}_{t-1} + C^2\mathbf{u}_{t-2} + C^3\mathbf{u}_{t-3} + \cdots + \mathbf{u}_t + \lim_{k\to\infty} C^k\mathbf{Y}_{t-k} \\
&\Downarrow \\
Y_t &= \mu + \Theta u_{t-1} + \Theta^2 u_{t-2} + \Theta^3 u_{t-3} + \cdots + u_t + \lim_{k\to\infty} \Theta^k Y_{t-k}
\end{aligned}
$$

Here, Θ is defined as the $n \times n$ upper-left matrix of C, and μ is defined as the first n elements in $\left(I_n + C + C^2 + C^3 + \cdots\right) c$. We observe that the VMA(∞) representation exists if C^k approaches zero as k goes to infinity, indicating no dependence on past observables. Additionally, C^k should converge to zero sufficiently fast for the infinite sum $\sum_{l=0}^{\infty} C^{2l}$ to converge, ensuring that $\mathbf{Y}_t$ has a finite mean and variance. This condition is known as **square summability**, which is a weaker condition than **absolute summability**. Absolute summability is defined as the convergence of $\sum_{l=0}^{\infty} \left|C^l\right|$. Absolute summability and thus square summability occur when all eigenvalues of the companion matrix C have moduli smaller than one. In such cases, the geometric series formula can be applied, resulting

in $\sum_{l=0}^{\infty} \left|C^l\right| \leq \sum_{l=0}^{\infty} |C|^l = (I_n - |C|)^{-1}$, which is finite.

From Section 5.4, we learned that the eigenvalues of the companion matrix C are the inverted roots of the determinant of the matrix polynomials $C(z) = (I_n - Cz)$ or $\Phi(z) = I_n - \Phi_1 z - \cdots - \Phi_p z^p$. This property is known as **stability**. In other words, a VAR(p) is considered **stable** if all the roots of the determinant of the polynomial $\Phi(z)$ have a modulus greater than one:

$$\det\left(I_n - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p\right) \neq 0, \quad \text{for } |z| \leq 1$$

Just like with AR models, stability implies stationarity, as a stable VAR has a VMA(∞) representation that depends on residuals with a constant distribution. However, it's important to note that stationarity does not always imply stability.

6.2 VAR as a Linear Regression

Before expressing the VAR model as a linear regression, let's consider the standard linear regression model for cross-sectional data:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + u_i \\ &= X_i \beta + u_i, \end{aligned} \qquad \begin{aligned} X_i &= \begin{bmatrix} 1 & X_{i,1} & \cdots & X_{i,k} \end{bmatrix} \\ \beta &= \begin{bmatrix} \beta_0 & \cdots & \beta_k \end{bmatrix}' \\ i &= 1, 2, \ldots, m \end{aligned}$$

In this model, we have an observation indexed by i, where Y_i and $X_{i,j}$ are scalar random variables describing the observation, and u_i represents the part of Y_i that cannot be explained with the regressors.

We can combine the random variables of all m observations into $m \times 1$ vectors, where

each element represents a different observation:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix}, \quad X^{(j)} = \begin{bmatrix} X_{1,j} \\ \vdots \\ X_{m,j} \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}$$

Moreover, let's combine the constant and all k regressors into an $m \times k$ matrix:

$$X = \begin{bmatrix} 1_m & X^{(1)} & X^{(2)} & \cdots & X^{(k)} \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{m,1} & X_{m,2} & \cdots & X_{m,k} \end{bmatrix}$$

where 1_m is an $m \times 1$ vector of ones. Then we can combine all observations of the linear regression as follows:

$$Y = X\beta + u, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$$

Hence, assuming no correlation between regressors and residuals, we obtain the **least square (LS) estimator** $\hat{\beta}$:

$$\begin{aligned} E\left[X_{i,j} u_i\right] &= 0, \quad \forall i, \ \forall j \\ E\left[X' u\right] &= 0 \\ E\left[X'\left(Y - X\beta\right)\right] &= 0 \\ \beta &= E\left[X'X\right]^{-1} E\left[X'Y\right] \\ \hat{\beta} &= \left(X'X\right)^{-1}\left(X'Y\right) \end{aligned}$$

Here, β is a constant parameter vector, and $\hat{\beta}$ is a random vector, as it is a function of the random variables X and Y.

Now let's extend this regression framework to a VAR(p) for the stochastic n-dimensional

vector process $\{Y_t\}$:

$$\begin{aligned} Y_t &= \delta + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + u_t \\ &= BV_t + u_t, \end{aligned} \qquad \begin{aligned} V_t &= \begin{bmatrix} 1 & Y'_{t-1} & \cdots & Y'_{t-p} \end{bmatrix}' \\ B &= \begin{bmatrix} \delta & \Phi_1 & \cdots & \Phi_p \end{bmatrix} \\ t &= 1, 2, \ldots, T \end{aligned}$$

Here, Y_t represents the n-dimensional random vector at time t, δ is a constant vector, Φ_l is the VAR coefficient matrix at lag l, V_t is the stacked lagged vector at time t, and u_t is the error vector at time t.

While the standard linear regression model has just one equation, the VAR model has n equations. We discuss two ways to deal with this. First, we can express the VAR as a **general linear model**, also known as a **general multivariate regression model**, where n regressions are performed simultaneously. Second, we can rewrite the VAR so that it becomes a single equation, allowing us to estimate it as the standard linear regression model discussed above.

To express the VAR as a general multivariate regression model, we can combine the n-dimensional random vectors of all T observations into $T \times n$ matrices, where each row represents a different observation:

$$W = \begin{bmatrix} Y_{1,1} & \cdots & Y_{1,n} \\ \vdots & \cdots & \vdots \\ Y_{T,1} & \cdots & Y_{T,n} \end{bmatrix}, \quad V^{(l)} = \begin{bmatrix} Y_{1-l,1} & \cdots & Y_{1-l,n} \\ \vdots & \cdots & \vdots \\ Y_{T-l,1} & \cdots & Y_{T-l,n} \end{bmatrix}, \quad S = \begin{bmatrix} u_{1,1} & \cdots & u_{1,n} \\ \vdots & \cdots & \vdots \\ u_{T,1} & \cdots & u_{T,n} \end{bmatrix}$$

Here, W represents the dependent variables, $V^{(l)}$ refers to the lth lag of W, and S represents the stacked error matrix. Next, we combine the constant and all np regressors into a

$T \times n\,(p+1)$ matrix:

$$V = \begin{bmatrix} 1_T & V^{(1)} & \cdots & V^{(p)} \end{bmatrix} = \begin{bmatrix} 1 & Y_{1-1,1} & \cdots & Y_{1-1,n} & \cdots & Y_{1-p,1} & \cdots & Y_{1-p,n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 1 & Y_{T-1,1} & \cdots & Y_{T-1,n} & \cdots & Y_{T-p,1} & \cdots & Y_{T-p,n} \end{bmatrix}$$

where 1_T represents a $T \times 1$ vector of ones. Now, we can combine all observations of the VAR(p) as follows:

$$W = VB' + S, \qquad B = \begin{bmatrix} \delta & \Phi_1 & \cdots & \Phi_p \end{bmatrix}$$

The assumption that there is no correlation between regressors and residuals leads to the **multivariate LS estimator** $\hat{B}'$:

$$\begin{aligned} E\left[Y_{t-l,i} u_{t,j}\right] &= 0, \qquad t = 1, \ldots, T, \quad l = 1, \ldots, p, \quad i, j = 1, \ldots, n \\ E\left[V'S\right] &= 0 \\ E\left[V'\left(W - VB'\right)\right] &= 0 \\ B' &= E\left[V'V\right]^{-1} E\left[V'W\right] \\ \hat{B}' &= \left(V'V\right)^{-1} V'W \end{aligned}$$

Here, $\hat{B}$ is a matrix of random variables, as it is a function of the random matrices V and W, whereas B is a constant matrix of parameters.

The second approach involves expressing the VAR model as a single equation using the Kronecker product and the vec operator. The **Kronecker product**, denoted as $A \otimes B$, multiplies each element of the first matrix by the entire second matrix, while the **vec**

operator, represented as **vec**(A), stacks the columns of a matrix into a vector:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \otimes \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} = \begin{bmatrix} a \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} & b \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \\ c \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} & d \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \end{bmatrix}, \qquad \text{vec} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix}$$

These operators satisfy the following properties:

$$\begin{aligned} (A+B)\otimes(C+D) &= A\otimes C + A\otimes D + B\otimes C + B\otimes D \\ (A\otimes B)(C\otimes D) &= (AC)\otimes(BD) \\ (A\otimes B)^{-1} &= A^{-1}\otimes B^{-1} \\ (A\otimes B)' &= A'\otimes B' \\ \text{vec}(aA+bB) &= a\text{vec}(A) + b\text{vec}(B) \\ \text{vec}(ABC) &= (C'\otimes A)\,\text{vec}(B) \end{aligned}$$

Applying the vec operator to the VAR model, we have:

$$\begin{aligned} \text{vec}\,(Y_t') &= \text{vec}\,(V_t'B') + \text{vec}\,(u_t') \\ Y_t &= (I_n \otimes V_t')\,\text{vec}\,(B') + u_t \\ Y_t &= X_t\beta + u_t, \qquad & X_t &= I_n \otimes V_t' \\ & & \beta &= \text{vec}\,(B') \end{aligned}$$

Thus, we obtain the standard least squares (LS) estimator $\hat{\beta}$ as in the cross-sectional case. One advantage of this form is that standard linear regression formulas for covariance matrix estimation can be utilized, such as the ones discussed in White (2000).

6.3 Vector Autoregressive Moving Average (VARMA) Model

When working with univariate time series in Section 3, we not only had AR models for stationary time series but also MA and ARMA models. For multivariate time series, these models are called **vector moving average models** of order q, **VMA(q)**, and **vector autoregressive moving average models** of order p and q, **VARMA(p, q)**. However, the vector versions of these time series models are not always identified, meaning that it's possible to have two VARMA models with different AR and MA lags that produce exactly the same data.

To illustrate the identification problem, consider the example of distinguishing a VAR(1) from a VMA(1), respectively a VARMA$(1, 0)$ from a VARMA$(0, 1)$:

$$\begin{aligned} Y_t = \Phi_1 Y_{t-1} + \epsilon_t \quad &\Rightarrow \quad Y_t = \Phi_1 \epsilon_{t-1} + \Phi_1^2 \epsilon_{t-2} + \Phi_1^3 \epsilon_{t-3} + \cdots + \epsilon_t \\ Y_t = \Theta_1 \epsilon_{t-1} + \epsilon_t \quad &\Rightarrow \quad Y_t = \Theta_1 \epsilon_{t-1} + (0)\epsilon_{t-2} + (0)\epsilon_{t-3} + \cdots + \epsilon_t \end{aligned}$$

The two models have the same VMA(∞) representation if the parameters satisfy the following condition:

$$\Phi_1^j = \begin{cases} \Theta_1 & j = 1 \\ 0 & j \geq 2 \end{cases}$$

In the univariate case, when Φ_1 and Θ_1 are scalars, the above condition is only satisfied if $\Phi_1 = \Theta_1 = 0$. This makes them white noise processes rather than AR and MA processes and therefore AR and MA processes cannot generate the same data.

However, in the multivariate case when Φ_1 and Θ_1 are $n \times n$ matrices, it is possible for this condition to hold even when $\Phi_1 \neq 0$. Specifically, the condition holds when Φ_1 is a **nilpotent matrix**, meaning that $\Phi_1 \neq 0$ and $\Phi_1^2 = 0$. For example, consider the following

nilpotent matrix:

$$\Phi_1 = \begin{bmatrix} 0.5 & -0.3 & 0.2 \\ 1.5 & -0.9 & 0.6 \\ 1 & -0.6 & 0.4 \end{bmatrix} \quad \Rightarrow \quad \Phi_1^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

In this case, the VAR(1) and VMA(1) model with $\Theta_1 = \Phi_1$ exhibit the same VMA(∞) representation. As a result, these models are observationally equivalent, meaning they generate the same data and cannot be distinguished based on the available information.

There are several techniques discussed in the literature to handle this identification problem, and VARMA models are gaining attention. However, in macroeconomics literature, most researchers rely on VAR models instead. VAR models have the advantage that they are easy to estimate, and by limiting the model to only autoregressive components, they avoid this identification problem.

6.4 Structural Vector Autoregression (SVAR)

Because the VAR model is multivariate, it allows us to measure the causal effects between variables. In a cross-sectional setting, we can measure the causal effect of variable X on Y as the coefficient β_1 in a linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

However, in a time-series setting, the parameters of a VAR do not have a straightforward causal interpretation. For example, an exogenous change in variable X at time $t-2$ not only affects Y at time t, but it also affects X_{t-1}, which in turn has an effect on Y_t, as well. Thus, the overall effect of X_{t-2} on Y_t involves multiple parameters and cannot be attributed to a single coefficient.

To measure causal effects in a time-series setting, **impulse response functions (IRFs)** are used. IRFs measure the (causal) response of variables over time to an exogenous change

in their values, a so-called impulse. For example, consider a VAR of two stochastic processes: $\{X_t\}$ and $\{Z_t\}$. The IRF of $\{X_t\}$ to an impulse $d = (d_1, d_2)$ measures the causal response of X_{t+h}, for $h = 0, 1, 2, \ldots$, to an exogenous increase in X_t and Z_t by d_1 and d_2 respectively.

Formally, the IRF of $\{X_t\}$ to an impulse d at horizon h is the change in the forecasted value when X_t and Z_t increases by d:

$$
\begin{aligned}
IRF^X_{h,t} &= E[X_{t+h}|\hat{I}_t] - [X_{t+h}|I_t] \\
\hat{I}_t &= \sigma(X_t + d_1, X_{t-1}, X_{t-2}, \ldots, Z_t + d_2, Z_{t-1}, Z_{t-2}, \ldots), \\
I_t &= \sigma(X_t, X_{t-1}, X_{t-2}, \ldots, Z_t, Z_{t-1}, Z_{t-2}, \ldots) \qquad\qquad h = 0, 1, 2, \ldots \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = 1, \ldots, T
\end{aligned}
$$

In the above equation, $\sigma(\cdot)$ represents an operator that constructs an information set based on random variables, but its details are not relevant here.

In a linear system, the effect of a change is independent of the initial values before the change is applied. This means that increasing X_t and Z_t by d_1 and d_2 respectively has the same effect on the forecasted value of X_{t+h} regardless of the initial values X_t and Z_t. Therefore, in a VAR model, the IRFs are **time-independent**, i.e., $IRF^X_{h,t} = IRF^X_h$ for all t.

The challenge is to create a meaningful impulse vector d. For example, when studying the causal effect of interest rates X_t on inflation Z_t, one might be tempted to set $d = (1, 0)$ to represent an exogenous increase in X_t while keeping Z_t constant. However, this impulse is not meaningful because it is impossible to change the interest rate exogenously without simultaneously affecting inflation. In practice, when the interest rate changes, firms are likely to adjust prices immediately, leading to a non-zero impact on inflation. Therefore, $d = (1, 0)$ does not represent a meaningful impulse vector.

To establish meaningful impulses, we connect the vector autoregression (VAR) model with economic theory, resulting in a structural vector autoregression (SVAR). In construct-

ing an SVAR, we start with a system of simultaneous equations that captures all causal dependencies among variables. Unlike a VAR, X_t directly responds to Z_t and vice versa:

$$X_t = \tau_1 + \lambda_{10} Z_t + \kappa_{11} X_{t-1} + \cdots + \kappa_{1p} X_{t-p} + \lambda_{11} Z_{t-1} + \cdots + \lambda_{1p} Z_{t-p} + \sigma_1 \epsilon_{1t}$$

$$Z_t = \tau_2 + \kappa_{20} X_t + \kappa_{21} X_{t-1} + \cdots + \kappa_{2p} X_{t-p} + \lambda_{21} Z_{t-1} + \cdots + \lambda_{2p} Z_{t-p} + \sigma_2 \epsilon_{2t}$$

The residuals ϵ_{1t} and ϵ_{2t} are uncorrelated since the system captures all interdependencies among the variables. Specifically, we have $\text{Var}(\epsilon_{1t}) = 1$, $\text{Var}(\epsilon_{2t}) = 1$, and $\text{Cov}(\epsilon_{1t}, \epsilon_{2t}) = 0$. This is different from the error terms in a VAR, where $\text{cor}(u_{1t}, u_{2t}) \neq 0$.

The above system of simultaneous equations should reflect meaningful relationships derived from economic theory. Taking the example of interest rates and inflation, the first equation could describe the conduct of monetary policy by the central bank, while the second equation could capture how firms set prices. In these equations, each parameter holds economic significance and represents a causal effect. This system of equations with economic interpretations is referred to as the **structural form**, whereas the VAR represents the **reduced form**. **Identification** is the process of deriving the meaningful structural form from the reduced form. It is worth noting that in economics, the structural form may take the form of a **rational expectations model**, incorporating additional terms that capture individuals' expectations.

The residuals in the structural form, denoted as ϵ_{1t} and ϵ_{2t}, represent changes in the variables X_t and Z_t that are completely independent of the relationships between the two stochastic processes X_t and Z_t. As these structural form residuals are exogenous, they are commonly referred to as shocks or structural shocks. Consequently, the impulse response functions (IRFs) of Y_t and X_t to an increase in the shocks ϵ_{1t} and ϵ_{2t} carry meaning, as these shocks exogenously shift X_t and Y_t.

To compute IRFs to the meaningful shocks ϵ_{1t} and ϵ_{2t}, we first need to determine the contemporaneous effects of these shocks on X_t and Y_t. This will allow us to derive the impulse vectors we are interested in. To do this, we substitute X_t and Y_t on the right-hand side of the structural form and reorganize the equations to match the VAR. Let's consider

the simplified case where there is only one lag in the system, i.e., $p = 1$. We obtain the following VAR model:

$$X_t = \underbrace{\frac{\tau_1 + \lambda_{10} f_2}{1 - \lambda_{10}\kappa_{20}}}_{\delta_1} + \underbrace{\frac{\kappa_{11} + \lambda_{10}\kappa_{21}}{1 - \lambda_{10}\kappa_{20}}}_{\alpha_{11}} X_{t-1} + \underbrace{\frac{\lambda_{11} + \lambda_{10}\lambda_{21}}{1 - \lambda_{10}\kappa_{20}}}_{\beta_{11}} Z_{t-1} + \underbrace{\frac{\sigma_1}{1 - \lambda_{10}\kappa_{20}}\epsilon_{1t} + \frac{\lambda_{10}\sigma_2}{1 - \lambda_{10}\kappa_{20}}\epsilon_{2t}}_{u_{1t}}$$

$$Z_t = \underbrace{\frac{\tau_2 + \kappa_{20} f_1}{1 - \kappa_{20}\lambda_{10}}}_{\delta_2} + \underbrace{\frac{\kappa_{21} + \kappa_{20}\kappa_{11}}{1 - \kappa_{20}\lambda_{10}}}_{\alpha_{21}} X_{t-1} + \underbrace{\frac{\lambda_{21} + \kappa_{20}\lambda_{11}}{1 - \kappa_{20}\lambda_{10}}}_{\beta_{21}} Z_{t-1} + \underbrace{\frac{\kappa_{20}\sigma_1}{1 - \kappa_{20}\lambda_{10}}\epsilon_{1t} + \frac{\sigma_2}{1 - \kappa_{20}\lambda_{10}}\epsilon_{2t}}_{u_{2t}}$$

Hence, we have derived a VAR model, but now the residuals u_{1t} and u_{2t}, and thus X_t and Y_t, depend on meaningful shocks ϵ_{1t} and ϵ_{2t}.

The impulse vectors $d^{(1)}$ and $d^{(2)}$ represent the effects of the shocks ϵ_{1t} and ϵ_{2t} on contemporaneous Y_t and X_t, respectively. These impulse vectors can be derived by computing the following partial derivatives:

$$d^{(1)} = \begin{bmatrix} \frac{\partial Y_t}{\partial \epsilon_{1t}} \\ \frac{\partial X_t}{\partial \epsilon_{1t}} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_{1t}}{\partial \epsilon_{1t}} \\ \frac{\partial u_{2t}}{\partial \epsilon_{1t}} \end{bmatrix} = \begin{bmatrix} \frac{\sigma_1}{1 - \lambda_{10}\kappa_{20}} \\ \frac{\kappa_{20}\sigma_1}{1 - \kappa_{20}\lambda_{10}} \end{bmatrix}, \quad d^{(2)} = \begin{bmatrix} \frac{\partial Y_t}{\partial \epsilon_{2t}} \\ \frac{\partial X_t}{\partial \epsilon_{2t}} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_{1t}}{\partial \epsilon_{2t}} \\ \frac{\partial u_{2t}}{\partial \epsilon_{2t}} \end{bmatrix} = \begin{bmatrix} \frac{\kappa_{20}\sigma_1}{1 - \kappa_{20}\lambda_{10}} \\ \frac{\sigma_2}{1 - \kappa_{20}\lambda_{10}} \end{bmatrix}$$

Since the impulse vectors depend solely on how the VAR residuals depend on the shocks, we can disregard the other terms and focus on relating the VAR residuals to the structural shocks:

$$\begin{aligned} u_{1t} &= a_{11}\epsilon_{1t} + a_{12}\epsilon_{2t}, \quad a_{11} = \frac{\sigma_1}{1 - \lambda_{10}\kappa_{20}}, \quad a_{12} = \frac{\kappa_{20}\sigma_1}{1 - \kappa_{20}\lambda_{10}}, \\ u_{2t} &= a_{21}\epsilon_{1t} + a_{22}\epsilon_{2t}, \quad a_{21} = \frac{\kappa_{20}\sigma_1}{1 - \kappa_{20}\lambda_{10}}, \quad a_{22} = \frac{\sigma_2}{1 - \kappa_{20}\lambda_{10}}, \end{aligned}$$

This can be further represented as:

$$u_t = A\varepsilon_t$$

where $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ is the **impact matrix**. Note that the ith column of A represents the impulse vector of the ith shock; hence, in the bivariate example we have that $A = \begin{bmatrix} d^{(1)} & d^{(2)} \end{bmatrix}$.

The structural vector autoregression (SVAR) is defined as a VAR where the residuals

are replaced by the meaningful shocks:

$$
\begin{aligned}
X_t &= \delta_1 + \alpha_{11} X_{t-1} + \cdots + \alpha_{1p} X_{t-p} + \beta_{11} Z_{t-1} + \cdots + \beta_{1p} Z_{t-p} + a_{11}\epsilon_{1t} + a_{12}\epsilon_{2t} \\
Z_t &= \delta_2 + \alpha_{21} X_{t-1} + \cdots + \alpha_{2p} X_{t-p} + \beta_{21} Z_{t-1} + \cdots + \beta_{2p} Z_{t-p} + a_{21}\epsilon_{1t} + a_{22}\epsilon_{2t} \\
&\Downarrow \\
Y_t &= \delta + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + A\epsilon_t
\end{aligned}
$$

Hence, the only difference between a VAR and a SVAR is that u_t is replaced with $A\epsilon_t$. This definition also applies to the general case where Y_t, u_t, and ϵ_t are n-dimensional vectors of random variables, and A is an $n \times n$ impact matrix.

The impulse response functions (IRFs) to the n structural shocks are then defined as follows:

$$
\gamma_h^{(k)} = E[Y_{t+h} | \sigma\left(Y_t + \delta^{(k)}, Y_{t-1}, Y_{t-2}, \dots\right)] - E[Y_{t+h} | \sigma(Y_t, Y_{t-1}, Y_{t-2}, \dots)] \qquad h = 0, 1, 2, \dots
$$

Combining these IRFs, we obtain:

$$
\Gamma_h = \begin{bmatrix} \gamma_h^{(1)} & \dots & \gamma_h^{(n)} \end{bmatrix} = \Theta_h A \qquad h = 0, 1, 2, \dots
$$

Here, the ith row and jth column of Γ_h represents the IRF of the ith variable $Y_{i,t+h}$ to the jth shock $\epsilon_{j,t}$. Note that Θ_h is the $n \times n$ matrix of coefficients of the VMA(∞) representation at lag h, and A is the $n \times n$ impact matrix.

While data can be used to estimate the VAR parameters $\{\delta, \Phi_1, \dots, \Phi_p, \Omega\}$, the impact matrix A needs to be derived using additional assumptions from economic theory, which are called *identifying restrictions*. The reason for requiring additional restrictions is that the covariance matrix Ω of the VAR residuals u_t is symmetric and provides only $\frac{n(n+1)}{2}$ parameters, whereas the impact matrix A has n^2 parameters. Therefore, we need $\frac{n(n-1)}{2}$ additional equations derived from economic theory to estimate A. Note that the number of restrictions required increases rapidly with n; for example, when $n = 8$, we have $\frac{n(n-1)}{2} = 28$

restrictions.

In particular, to estimate A when $n = 2$, we can relate the estimated variances and covariances of the VAR residuals to A as follows:

$$\begin{aligned}
\text{Var}(u_{1t}) &= \text{Var}(a_{11}\epsilon_{1t} + a_{12}\epsilon_{2t}) = (a_{11})^2 + (a_{12})^2 \\
\text{Var}(u_{2t}) &= \text{Var}(a_{21}\epsilon_{1t} + a_{22}\epsilon_{2t}) = (a_{21})^2 + (a_{22})^2 \\
\text{Cov}(u_{1t}, u_{2t}) &= \text{Cov}(a_{11}\epsilon_{1t} + a_{12}\epsilon_{2t}, a_{21}\epsilon_{1t} + a_{22}\epsilon_{2t}) = a_{11}a_{21} + a_{12}a_{22}
\end{aligned}$$

However, we have only three equations for four unknowns, namely a_{11}, a_{12}, a_{21}, and a_{22}. Thus, we need one additional equation to identify the SVAR, as expected since $n = 2$ implies $\frac{n(n-1)}{2} = 1$. For example, one could use economic theory to argue that an exogenous increase in the interest rate of one percentage point increases the inflation rate by one-half percentage point, which adds a fourth equation $a_{21} = \frac{1}{2}a_{11}$, resulting in a unique impact matrix A.

It is common to identify a structural VAR using the **recursive approach**, which assumes that A is a lower-triangular matrix where $a_{ik} = 0$ for all $k > i$:

$$A = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}$$

This leads to $\frac{1}{2}n(n-1)$ identifying restrictions, which are sufficient for SVAR identification. For those interested in technical details, this restriction implies that B can be obtained as the *Cholesky decomposition* of the residual covariance matrix Ω, where $\Omega = BB'$.

The recursive approach implies that the first shock of the SVAR can have a contemporaneous effect on all variables in the model, while the second shock only affects the second to last variables on impact, the third shock only affects the third to last variable, and so

on. Therefore, the order of variables matters when using the recursive approach. However, note that the ordering of variables in a VAR matters only for impulse response functions (IRFs), as for forecasting purposes, ordering does not matter.

While we have only imposed identifying restrictions on contemporaneous effects, it is also possible to identify an SVAR by imposing restrictions on the impulse response functions (IRFs) or on the long-run relationships between cumulated variables. There is a vast literature on SVAR that provides a comprehensive discussion of the different types of identifying restrictions that can be used.

When variables have been transformed to achieve stationarity, the IRFs may not provide a comprehensive understanding of the original variable. For instance, while the IRF of a growth rate variable indicates the growth rate at each time horizon, it doesn't reveal the overall effect. To overcome this limitation, the **cumulative impulse response function (CIRF)** is used. The CIRF calculates the total effect by cumulating the IRFs at each horizon:

$$CIRF_h = \sum_{k=0}^{h} IRF_k$$

The CIRF of a differenced variable is equivalent to the IRF of the original time series. For example, the CIRF of the change in unemployment is the IRF of the unemployment level, and the CIRF of net immigration is the IRF of the total population due to migration. The CIRF of a growth rate variable is obtained by multiplying the IRF of the logarithm of the original series by 100. For instance, the CIRF of GDP growth is the IRF of 100 times the logarithm of GDP, and the CIRF of inflation is the IRF of 100 times the logarithm of the price level. It is important to note that the CIRFs are meaningful only for differenced series. For example, the CIRF of unemployment has no interpretation, and the same applies to the CIRF of an interest rate.

7 Multivariate Non-Stationary Time Series

7.1 Spurious Regression

In Section 4.1, we discussed the importance of stationarity in time series analysis, as it allows us to combine observations across time to make inferences about the underlying constant distribution. An alternative assumption is that the process could be non-stationary, yet the transition in distribution isn't random. Instead, it changes gradually over time, indicating the presence of a trend in the time series. Incorporating this trend into the model effectively controls for the changes in the distribution. As a result, the detrended or differenced process is stationary.

Estimating this gradual change in distribution can be achieved through the presumption that the mean of the time series varies according to a predefined deterministic function, or by utilizing tests such as the Dickey-Fuller test to discern the number of unit roots in a process. Stationarity allows us to exploit data properties, like autocorrelation and partial autocorrelation functions, for the selection of an appropriate model. However, trends differ as they span the entire time series without repetition. Therefore, trend patterns, such as linear trends, appear only once, unlike stationary patterns like business or seasonal cycles, which recur across the time series. Due to this, relying solely on data for determining the correct trend specification is less enlightening than selecting a stationary model, and researchers' assumptions about the trend profoundly influence the results.[1]

The challenge of a trend extending across the entire time series becomes especially noticeable when investigating causal relationships between time series. For instance, if two time series are independent but both display an upward trend, a regression of one on the other without taking the trends into account might erroneously suggest a strong causal relationship. This error arises as both time series demonstrate an increasing pattern over time, leading to the false impression that the escalation in one time series triggered the

[1] A potential approach to addressing the issue of observing a trend only once per time series sample involves the use of panel data. This data type comprises time series for multiple entities, like tracking technological progress over the past two decades for fifty different countries. By presuming each entity follows the same trend specification, we can obtain as many trend observations as there are entities. This enables us to evaluate which trend specification is appropriate for technological progress. For example, it allows us to discern whether technology is trend-stationary or difference-stationary.

rise in the other. Similarly, if both time series display a stochastic trend, their persistence could be high, resulting in both time series primarily moving in one direction throughout the sample period. This could result in a false impression of a robust causal relationship. This effect is known as **spurious regression**, first introduced by Granger and Newbold (1974). While the regression of two stationary time series on each other would yield a zero coefficient if they are independent, non-stationary time series may produce highly significant and non-zero coefficients.

Specifically, conducting a spurious regression of one random walk onto another independent random walk results in a coefficient that doesn't converge to its true value - zero. Instead, it follows a non-degenerate distribution. Consequently, the t-value of this regression often surpasses the critical values of a normal distribution, implying a significant relationship. Furthermore, the R-squared value is usually high because the omitted trend contributes substantially to the sample variance, which is erroneously ascribed to the regressor.

Interestingly, this still happens even when comparing two random walks that don't have a drift. One would expect, especially with infinite data, the series would sometimes move together, but just as often move in opposite directions, balancing out any apparent relationship. But here's the catch - the variance of a random walk increases over time. This means that the importance of early observations compared to later ones shrinks to nothing. So, even though one might technically have an infinite number of observations, the ones that actually matter - those contributing to the overall variance - remain finite.

Therefore, the takeaway is that the mere movement of two time series in tandem doesn't imply a relationship between them. One sign of a spurious regression is a high persistence of the regression residual, suggesting that the two time series follow different trends. Consequently, a spurious regression that accounts for one trend doesn't fully control for the other, leading to regression residuals that still exhibit a trend. Another indicator is a dramatic change in results upon modifying the regression specification, such as incorporating additional lags of the dependent variable. These lagged variables can help account for some

of the trend, thereby reducing the attributed impact of the trend to the other time series.

7.2 Cointegration

The prevous section revealed highlights the risk of spurious regressions: the erroneous regression of two independent, non-stationary time series on each other often results in a regression coefficient that is significantly different from zero. In contrast, this section explores the regression of two non-stationary time series on each other when they are not independent, but actually related.

We refer to such time series as **cointegrated** if they are non-stationary due to a stochastic trend, but the stochastic trend of some of the time series is either driven by other time series or multiple time series share the same stochastic trend. Formally, the n-dimensional vector of stochastic processes $\{Y_t\}$ is considered cointegrated if all n time series have a unit root, but there exists a linear combination $\beta' Y_t = \beta_1 Y_{1t} + \cdots + \beta_n Y_{nt}$ that is stationary.

For example, consider the following system of processes:

$$
\begin{aligned}
Y_{1t} &= \gamma_1 Y_{2t} + \gamma_2 Y_{3t} + \sigma_1 \epsilon_{1t} \\
Y_{2t} &= \gamma_3 Y_{3t} + \sigma_2 \epsilon_{2t} \\
Y_{3t} &= Y_{3,t-1} + \sigma_3 \epsilon_{3t}
\end{aligned}
$$

where ϵ_{1t}, ϵ_{2t}, and ϵ_{3t} are i.i.d. shocks. In this system, Y_{3t} may represent technological progress, which is often modeled as a random walk due to its cumulative nature: technological progress cannot be reversed, and it continues to accumulate over time. Y_{2t} represents hours worked, which could potentially decrease with advancements in technology, i.e. $\gamma_3 < 0$. Y_{1t} represents output per capita, which depends positively on both technological progress and hours worked, i.e. $\gamma_1 > 0$ and $\gamma_2 > 0$. Note that in the system described, there is only one source of stochastic trend caused by the technology shock ϵ_{3t}, as it is the only shock with a permanent effect.

All three processes, Y_{1t}, Y_{2t}, and Y_{3t}, are individually integrated of order one, denoted

as $I(1)$. This is because Y_t exhibits a stochastic trend, while the first difference, ΔY_t, is stationary. However, in this case, differencing the variables is not the appropriate approach, as ΔY_t would follow a VAR(∞). Instead, we can directly identify the system with its cointegrated relationships.

The cointegrated relationships can be represented as $\beta_1' = \begin{bmatrix} 1 & -\gamma_1 & -\gamma_2 \end{bmatrix}$ and $\beta_2' = \begin{bmatrix} 0 & 1 & -\gamma_3 \end{bmatrix}$, which result in two stationary processes: $\beta_1' Y_t = \sigma_1 \epsilon_{1t}$ and $\beta_2' Y_t = \sigma_2 \epsilon_{2t}$. These cointegrated relationships can be identified by performing a regression using the first or second equation. That's possible because In the presence of cointegration, the estimated coefficients of the regression model are (super-)consistent and have meaningful interpretations, unlike a spurious regression where the coefficients are spurious and do not converge to their true values.

While the residuals of a spurious regression have a unit root, the residuals of a regression with cointegrated variables are stationary. This is because any deviations from the cointegrated relationship are expected to be temporary. As a result, while the coefficients of spurious regressions do not converge to zero, the coefficients of a regression with cointegrated variables are *superconsistent*. This means that they converge to the true parameter value at a rate faster than the square root of the sample size.

The reason behind this faster convergence is as follows: In spurious regressions, the residual exhibits a unit root, leading to an increasing residual variance over time. On the other hand, in the presence of cointegration, the residual is stationary, resulting in a constant residual variance. Furthermore, the variance of the dependent and explanatory variables continues to increase with the sample size, making the contribution of the residual less relevant. As a result, there is very little uncertainty in the parameter estimates, leading to the rapid convergence of the coefficients.

Numerous economic theories suggest cointegration. For instance:

1. **Consumption and Income:** According to the Permanent Income Hypothesis, consumers' spending habits are influenced not only by their current income, but also by their future expected income. Therefore, although both consumption and income

may individually follow random walks, they move together in the long run and are likely cointegrated.

2. **Investment and Savings:** Economic theory posits that investment and savings should be equal in the long run. Even if savings and investment each follow a random walk, the difference between the two should be stationary, implying that these two series are cointegrated.

3. **Money Demand:** In monetary economics, the demand for money is said to depend on income and interest rates. Thus, if these variables are non-stationary, then the money demand might also be non-stationary, and these variables could be cointegrated.

4. **Purchasing Power Parity:** According to this theory, the exchange rate between two countries' currencies is determined by the price levels in the two countries. Therefore, if the price levels are non-stationary, then the exchange rate might also be non-stationary, and these variables could be cointegrated.

The cointegrated relationships play a crucial role in capturing the long-term equilibrium among a vector of cointegrated time series. However, it is also important to understand the dynamics of the temporary deviations from those relationships. The vector error correction model (VECM), which will be introduced in Section 7.4, encompasses both aspects.

7.3 Consistency of Regression Coefficients Under Non-Stationarity

This section explores how non-stationarity affects the convergence of regression coefficients, illustrated through the following **data generating process** (**DGP**):

$$X_t = \rho X_{t-1} + \eta_t \qquad \eta_t \overset{i.i.d.}{\sim} N(0, \sigma_{\eta\eta})$$
$$Y_t = \phi Y_{t-1} + \gamma X_t + \epsilon_t \qquad \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_{\epsilon\epsilon})$$

where the parameters $|\rho| \leq 1$, $|\phi| \leq 1$, and γ determine whether or not the stochastic processes $\{Y_t\}$ and $\{X_t\}$ exhibit a stochastic trend. Specifically, $\{Y_t\}$ and $\{X_t\}$ are both $I(1)$ with separate stochastic trends if $\rho = \phi = 1$ and $\gamma = 0$. This represents the scenario of **spurious regression** when regressing Y_t on X_t likely yields a significant non-zero coefficient instead of $\gamma = 0$. In contrast, if $\rho = 1$, $\phi = 0$, and $\gamma \neq 0$, then $\{Y_t\}$ and $\{X_t\}$ are both $I(1)$ but share the same stochastic trend. This is the scenario of **cointegration**, where regressing Y_t on X_t produces an unbiased estimate of γ, and the p-values are too high, indicating that the precision of the estimate is higher than standard regression results would suggest.

Mathematically, the challenge with spurious regression is that the regression coefficient does not converge to zero even as the number of observations increases, which is problematic when no true relationship exists between the variables. Conversely, with cointegration, the benefit is that the regression coefficient converges to the true value more rapidly than it would under conditions of stationarity, a phenomenon referred to as **superconsistency**. To illustrate this, consider the following linear regression model:

$$Y_t = \beta X_t + u_t$$

where u_t is the regression residual. By definition of a linear regression, u_t satisfies the zero conditional mean (ZCM) assumption, that is, $E[u_t|X_t] = E[u_t] = 0$ for all t.

The regression residual u_t is an unobserved variable, also known as **latent variable**, meaning there is no actual data available for u_t. Therefore, to estimate β, we must formulate the estimation process without directly including u_t since it does not exist in the data set. This is accomplished by employing the Zero Conditional Mean (ZCM) assumption, which implies $E[X_t u_t] = 0$, since u_t inside the expectation operator can be replaced by $E[u_t|X_t] = 0$ using the law of iterated expectations. Substituting the residual with the regression equation leads to $E[X_t(Y_t - \beta X_t)] = 0$. Estimating β then involves replacing

the expectation operator with the sample mean across T periods:

$$\frac{1}{T}\sum_{t=1}^{T} X_t(Y_t - \hat{\beta}X_t) = 0 \quad \Rightarrow \quad \hat{\beta} = \frac{\frac{1}{T}\sum_{t=1}^{T} X_t Y_t}{\frac{1}{T}\sum_{t=1}^{T} X_t^2}$$

Thus, we have obtained an estimate $\hat{\beta}$ of β.

The regression estimate $\hat{\beta}$ is dependent on random variables, making it a random variable itself, whereas the true coefficient β is a fixed constant. To understand their relationship, substitute Y_t with its regression model in the formula for $\hat{\beta}$:

$$\hat{\beta} = \frac{\frac{1}{T}\sum_{t=1}^{T} X_t(\beta X_t + u_t)}{\frac{1}{T}\sum_{t=1}^{T} X_t^2} = \beta + \frac{\frac{1}{T}\sum_{t=1}^{T} X_t u_t}{\frac{1}{T}\sum_{t=1}^{T} X_t^2} = \beta + \frac{\hat{Q}_{Xu}}{\hat{Q}_{XX}}$$

Here, $\hat{Q}_{XX}$ and $\hat{Q}_{Xu}$ are the sample estimates of $E[X_t^2]$ and $E[X_t u_t]$, respectively. This holds under the assumption of stationarity. However, if the series are non-stationary, the expectation for different time periods may differ; hence, a more accurate characterization of $\hat{Q}_{XX}$ and $\hat{Q}_{Xu}$ is that they are the sample estimates of $\frac{1}{T}\sum_{t=1}^{T} E[X_t^2]$ and $\frac{1}{T}\sum_{t=1}^{T} E[X_t u_t]$, respectively, which appropriately account for the potential variations across different time periods t.

What does β measure? To verify whether β accurately captures γ, we must examine the regression in the context of the two processes and understand that $\beta = \gamma$ implies $u_t = \phi Y_{t-1} + \epsilon_t$. Therefore, the regression correctly identifies $\beta = \gamma$ as long as the corresponding ZCM condition is met, i.e., $E[u_t|X_t] = E[\phi Y_{t-1} + \epsilon_t|X_t] = 0$. Note that ϕY_{t-1} depends on X_{t-1} via γ, and X_t depends on X_{t-1} through ρ. Thus, the ZCM assumption requires that one of the parameters ρ, ϕ, or γ must be zero for the regression to be an unbiased estimator of γ, i.e. ZCM requires $\rho\phi\gamma = 0$.

The process of determining what the parameter β measures is known as **identification**. Identification involves expressing the empirical parameter β in terms of the theoretical parameters γ, ϕ, and ρ of the data generating process. The regression model with parameter β is referred to as **reduced form**, which can be directly estimated from data. In contrast,

the data generating process characterized by parameters γ, ϕ, and ρ is known as the **structural form**, which may require additional assumptions and steps for identification. The steps to identify β are similar to those for calculating $\hat{\beta}$ but involve using the expectation operator:

$$\frac{1}{T}\sum_{t=1}^{T} E[X_t(Y_t - \hat{\beta}X_t)] = 0 \quad \Rightarrow \quad \beta = \frac{\frac{1}{T}\sum_{t=1}^{T} E[X_t Y_t]}{\frac{1}{T}\sum_{t=1}^{T} E\left[X_t^2\right]}$$

Next, noting that $E[X_t] = 0$ implies $E\left[X_t^2\right] = Var(X_t)$, we substitute for X_t and Y_t using the data generating process, which results in the following expression for β:

$$\beta = \gamma + \gamma \frac{\frac{1}{T}\sum_{t=1}^{T}\sum_{l=1}^{t-1} \phi^l \rho^l Var(X_t)}{\frac{1}{T}\sum_{t=1}^{T} Var(X_t)}$$

The variance of X_t depends on the persistence parameter ρ as follows:

$$Var(X_t) = \begin{cases} \frac{\sigma_{\eta\eta}}{1-\rho^2} & \text{if } |\rho| < 1 \\ \sigma_{\eta\eta} t & \text{if } \rho = 1 \end{cases}$$

Substituting for $Var(X_t)$ and applying the geometric series formula then yields the following values for β:

$$\beta = \begin{cases} \gamma + \frac{\gamma\phi\rho}{1-\phi\rho}(1 - o(1)) & \text{if } |\phi| < 1 \\ \left(\frac{T+2}{3}\right)\gamma & \text{if } \rho = \phi = 1 \end{cases}$$

where $o(1)$ denotes an expression that goes to zero as T approaches infinity. This confirms that the regression is an unbiased estimator for γ if $\phi\gamma\rho = 0$.

Checking whether a regression parameter is unbiased is not sufficient for identification. We must also ensure that the estimated coefficient $\hat{\beta}$ converges to its expected value as the sample size increases, i.e., $\hat{\beta} \underset{p}{\rightarrow} \beta$ as $T \rightarrow \infty$. This concept is referred to as the **consistency** of $\hat{\beta}$. To demonstrate that $\hat{\beta}$ is consistent, one approach is to show that its variance diminishes to zero as T increases, implying that $\hat{\beta}$ stabilizes to a constant value. Since $E[\hat{\beta}] = \beta$, this constant would equal β. This behavior is based on **Chebyshev's Inequality**, which states that if the variance of a sequence of random variables approaches

zero, the sequence will converge in probability to its expected value.

Use the formula for $\hat{\beta}$ and the property that shifting a random variable by a constant β does not affect its variance to derive the following equation:

$$Var(\hat{\beta}) = Var\left(\frac{\hat{Q}_{Xu}}{\hat{Q}_{XX}}\right)$$

where $\hat{Q}_{Xu}$ is the sample covariance between X_t and u_t over $t = 1, \ldots, T$, and $\hat{Q}_{XX}$ is the sample variance of X_t over the same period. A dataset provides only one outcome of the random variables $\hat{\beta}$, $\hat{Q}_{Xu}$, and $\hat{Q}_{XX}$. Therefore, direct computation of the variance is not feasible without additional assumptions.

A valuable approach to calculating the variance is to use the **law of total variance**, also known as **Eve's law**, which decomposes the variance into the expectation of the conditional variance and the variance of the conditional expectation:

$$Var(A) = E[Var(A|B)] + Var(E[A|B])$$

Applying this to the variance of $\hat{\beta}$, while conditioning on the stochastic process $\{X_t\}_{t=1}^{T}$, allows us to isolate the conditional variance of $\hat{Q}_{Xu}$:

$$Var(\hat{\beta}) = E\left[\frac{Var\left(\hat{Q}_{Xu}\middle|\{X_t\}_{t=1}^{T}\right)}{\hat{Q}_{XX}^2}\right] + Var\left(\frac{E\left[\hat{Q}_{Xu}\middle|\{X_t\}_{t=1}^{T}\right]}{\hat{Q}_{XX}}\right)$$

If $E\left[u_t|\{X_s\}_{s=1}^{T}\right] = 0$ for all t, then the second term becomes zero. However, the zero conditional mean assumption, $E[u_t|X_t] = 0$, to ensure unbiasedness is not sufficient for this condition to hold. The residuals u_t could still be influenced by past values of the explanatory variable X_{t-l}, yet still provide an unbiased estimate of β.

To simplify the variance, consider these assumptions:

$$
\begin{aligned}
\text{Zero Conditional Mean (ZCM)}: &\quad E[u_t|X_t] = E[u_t] = 0, && \forall t \\
\text{Strict ZCM}: &\quad E\left[u_t|\{X_s\}_{s=1}^T\right] = E[u_t] = 0, && \forall t \\
\text{Homoskedasticity}: &\quad Var(u_t|X_t) = Var(u_t), && \forall t \\
\text{Strict Homoskedasticity}: &\quad Cov\left(u_t, u_s|\{X_t\}_{t=1}^T\right) = Cov(u_t, u_s), && \forall t, s \\
\text{No Serial Correlation}: &\quad Cov(u_t, u_s) = 0, && \forall t \neq s \\
\text{Covariance Stationarity}: &\quad Cov(u_t, u_{t-l}) = Cov(u_s, u_{s-l}), && \forall t, s, l
\end{aligned}
$$

Hence, strong ZCM ensures the elimination of the second term in the variance. Violations of the homoskedasticity assumption are termed heteroskedasticity. **Heteroskedasticity** occurs when the variability in the response variable Y_t varies at different levels of the explanatory variable X_t. **Serial correlation**, or autocorrelation, arises when residuals in a regression model are correlated across time. If the residual process is independent and identically distributed (i.i.d.), it inherently satisfies the conditions of strong homoskedasticity, no serial correlation, and stationarity.

Consider the conditional variance of the first term under these assumptions:

$$
\begin{aligned}
Var\left(\hat{Q}_{Xu}\middle|\{X_t\}_{t=1}^T\right) &= Var\left(\frac{1}{T}\sum_{t=1}^T X_t u_t\middle|\{X_t\}_{t=1}^T\right) \\
&= \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T Cov\left(X_t u_t, X_s u_s\middle|\{X_t\}_{t=1}^T\right) \\
&= \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T X_t X_s Cov\left(u_t, u_s\middle|\{X_t\}_{t=1}^T\right) \\
\text{[Strong Homoskedasticity]} &= \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T X_t X_s Cov(u_t, u_s) \\
\text{[No Serial Correlation]} &= \frac{1}{T^2}\sum_{t=1}^T X_t^2 Var(u_t) \\
\text{[Covariance Stationarity]} &= \frac{1}{T}\hat{Q}_{XX} Var(u_t)
\end{aligned}
$$

Thus, these assumptions simplify the variance of $\hat{\beta}$ as follows:

$$Var(\hat{\beta})\Big|_{\substack{\text{Strong ZCM}\\ \text{Strong Homoskedasticity}\\ \text{No Serial Correlation}\\ \text{Stationary Residuals}}} = \frac{1}{T} Var(u_t) E\left[\frac{1}{\hat{Q}_{XX}}\right]$$

where $Var(u_t)$ can be estimated using the sample residuals $\hat{u}_t$, and $E\left[\hat{Q}_{XX}^{-1}\right]$ can be approximated with $\hat{Q}_{XX}^{-1}$, although not unbiasedly, since $E\left[\hat{Q}_{XX}^{-1}\right] \neq E[\hat{Q}_{XX}]^{-1}$. An alternative involves assuming a distribution for X_t to estimate $E\left[\hat{Q}_{XX}^{-1}\right]$. Note that as $T \to \infty$ and assuming $Var(X_t) \neq 0$, the variance of the regression coefficient approaches zero, making it a consistent estimator under these assumptions.

While imposing those assumptions is a common way to estimate the regression coefficient in cross-sectional data, inspection of the data-generating process reveals that these assumptions may not hold, even if $\rho\phi\gamma = 0$. Consequently, a different approach is necessary, outlined in the following lemma:

Lemma 1 (Variance Approximation of Regression Coefficient)**.**
If there exists λ and δ s.t.,
(a) $\lim_{T\to\infty} E\left[T^{\lambda}\hat{Q}_{XX}\right] > 0$,
(b) $\lim_{T\to\infty} Var\left(T^{\lambda}\hat{Q}_{XX}\right) = 0$, *and*
(c) $\lim_{T\to\infty} Var\left(T^{\lambda-\delta}\hat{Q}_{Xu}\right) > 0$,
then $Var(\hat{\beta}) = \frac{Var(\hat{Q}_{Xu})}{E[\hat{Q}_{XX}]^2} + o\left(T^{-2\delta}\right) = O\left(T^{-2\delta}\right)$.

Proof. If the convergences in (a) and (c) are towards constants a and c respectively, we find that the first term of $\text{Var}(\hat{\beta})$ converges to a/c^2 if it is scaled by $T^{2\delta}$:

$$\lim_{T\to\infty} T^{2\delta}\left(\frac{Var(\hat{Q}_{Xu})}{E[\hat{Q}_{XX}]^2}\right) = \frac{\lim_{T\to\infty} Var\left(T^{\lambda-\delta}\hat{Q}_{Xu}\right)}{\lim_{T\to\infty} E\left[T^{\lambda}\hat{Q}_{XX}\right]^2} = \frac{a}{c^2}$$

Thus, the first term of $Var(\hat{\beta})$ is $O(T^{-2\delta})$. The second term is defined as:

$$Var\left(\frac{\hat{Q}_{Xu}}{\hat{Q}_{XX}}\right) - \frac{Var(\hat{Q}_{Xu})}{E[\hat{Q}_{XX}]^2} = T^{2\delta}\left[Var\left(\frac{T^{\lambda-\delta}\hat{Q}_{Xu}}{T^{\lambda}\hat{Q}_{XX}}\right) - \frac{Var\left(T^{\lambda-\delta}\hat{Q}_{Xu}\right)}{E\left[T^{\lambda}\hat{Q}_{XX}\right]^2}\right]$$

Given that $T^{\lambda}\hat{Q}_{XX}$ becomes deterministic as T increases, while the variance of $T^{\lambda-\delta}\hat{Q}_{Xu}$ remains finite, the discrepancy between the variance of $\hat{\beta}$ and its simplified expression diminishes. In particular, Chebyshev's inequality ensures that $T^{\lambda}\hat{Q}_{XX}$ converges to its expected value as $T \to \infty$, so that the term in brackets goes to zero, making it $o(1)$. After scaling the $o(1)$ expression with $T^{2\delta}$, it becomes $o(T^{-2\delta})$. Finally, since $O(T^{-2\delta}) + o(T^{-2\delta}) = O(T^{-2\delta})$, we have proven the lemma. □

This lemma demonstrates that the expression $Var(\hat{Q}_{Xu})/E[\hat{Q}_{XX}]^2$ serves as a reliable approximation of the variance $Var(\hat{\beta})$ for sufficiently large T.

Lemma 1 relates to the concept of asymptotic variance, defined as follows:

$$V_{\hat{\beta}} = \lim_{T\to\infty} Var\left(T^{\delta}\hat{\beta}\right) = \lim_{T\to\infty} T^{2\delta}\left(\frac{Var(\hat{Q}_{Xu})}{E[\hat{Q}_{XX}]^2}\right) = \frac{\lim_{T\to\infty} Var\left(T^{\lambda-\delta}\hat{Q}_{Xu}\right)}{\lim_{T\to\infty} E\left[T^{\lambda}\hat{Q}_{XX}\right]^2}$$

where the second equation follows from Lemma 1. The $o(T^{-2\delta})$ term from Lemma 1 drops out because $T^{-2\delta}o(T^{-2\delta}) = o(1) \to 0$ as $T \to \infty$. Estimating $Var(\hat{Q}_{Xu})/E[\hat{Q}_{XX}]^2$ parallels the estimation of the asymptotic variance $\hat{V}_{\hat{\beta}}$ and then scaling it by $T^{-2\delta}$, although this approximation might not be exact as it could include terms dependent on T, whereas $\hat{V}_{\hat{\beta}}$ does not.

Let's use Lemma 1 to derive the following expressions for the variance of $\hat{\beta}$ depending on the parameters ρ, ϕ, and γ of the data generating process:

$$Var(\hat{\beta}) = \begin{cases} \frac{1}{T}\frac{\sigma_{\epsilon\epsilon}\left(1-\rho^2\right)(1+\rho\phi)}{\sigma_{\eta\eta}(1-\phi^2)(1-\rho\phi)} + o\left(T^{-1}\right) & \text{if } |\rho| < 1,\ |\phi| < 1,\ \rho\phi\gamma = 0 \\ \frac{1}{T^2}\frac{2\left(\sigma_{\epsilon\epsilon}+\sigma_{\eta\eta}\gamma^2\phi^2\right)}{\sigma_{\eta\eta}(1-\phi)^2} + o\left(T^{-2}\right) & \text{if } \rho = 1,\ |\phi| < 1,\ \rho\phi\gamma = 0 \\ \frac{\sigma_{\epsilon\epsilon}(1+\rho)^2+\sigma_{\eta\eta}\gamma^2}{2\sigma_{\eta\eta}} + o(1) & \text{if } |\rho| < 1,\ \phi = 1,\ \rho\phi\gamma = 0 \\ \frac{\sigma_{\epsilon\epsilon}}{6\sigma_{\eta\eta}^3} + o(1) & \text{if } \rho = 1,\ \phi = 1,\ \gamma = 0 \end{cases}$$

The unbiased estimator $\hat{\beta}$ is consistent if both dependent and explanatory variables are stationary. It becomes superconsistent if they are cointegrated, meaning the only reason Y_t is non-stationary is its dependence on X_t. In cases where the dependent variable exhibits a stochastic trend independent of the explanatory variable, the unbiased estimator $\hat{\beta}$ becomes inconsistent. This means that even with an increasing amount of data, an estimated value different from $\beta = \gamma$ is very likely.

Note that the scenario where $\rho = \phi = 1$ and $\gamma = 0$ is more problematic than the one where the explanatory variable is stationary ($|\rho| < 1$). This is because standard regression analyses typically compute the variance of the estimator assuming conditions like strong ZCM, homoskedasticity, no serial correlation, and stationarity. In the latter scenario, as $\text{Var}(u_t)$ increases with sample size, the estimated variance of $\hat{\beta}$ also rises, leading to insignificant t-values. However, in the former case where $\rho = \phi = 1$ and $\gamma = 0$, both $\hat{Q}_{XX}$ and $\text{Var}(u_t)$ increase at the same rate. Consequently, the estimated variance using the standard formula decreases with sample size, falsely suggesting that the coefficient is significant, even if it is not. This leads to what is known as **spurious regression**, which can result in misleading statistical inferences.

7.4 Vector Error Correction Model (VECM)

Consider an n-dimensional $I(1)$ stochastic process $\{Y_t\}$ and assume it follows a VAR(p):

$$\Phi(L)Y_t = u_t, \qquad \Phi(L) = I - \Phi_1 L - \cdots - \Phi_p L^p$$

This model implies that looking at the last p lags is sufficient not only to predict short-run fluctuations but also to capture all stochastic trends. For instance, if one of the variables is output, then the last p years of observations are deemed sufficient to predict future output. This assumption would not be very restrictive if the stochastic process was $I(0)$, because, by definition of stationarity, events from a long time ago become irrelevant. However, since $\{Y_t\}$ is $I(1)$, there are stochastic trends, meaning that past events aren't discounted,

and events that happened 100 years ago are still relevant today. Consequently, the model implies that the information contained in the last p years contains all relevant information, including information from events that happened 100 years ago.

Alternatively, one could model a VAR for the differenced series $\{\Delta Y_t\}$ and then predict ΔY_{t+h}. Such predictions are likely to be accurate even if the p lags do not capture all the information, as realizations from the distant past become less relevant in stationary time series. In this case, predictions for the level variables Y_{t+h} can be obtained by summing the differences: $Y_t + \Delta Y_{t+1} + \ldots + \Delta Y_{t+h}$. However, in doing so, the missed information from the distant past accumulates, leading to a significant bias when h is large. Hence, estimating a VAR for the differenced series likely produces good predictions for ΔY_{t+h} but bad predictions for Y_{t+h}.

We demonstrate below that even when the level variables Y_t follow a VAR, estimating a VAR in differences and cumulating the variables results in a substantial bias. One exception is when the VAR coincidentally has as many independent stochastic trends as variables, in which case the VAR(p) for Y_t collapses to a VAR($p-1$) for ΔY_t, as was the case in Section 4.1 where the ARMA(p, q) model for an $I(1)$ process collapsed to an ARMA($p-1, q$) for its first difference. In general, there may be fewer stochastic trends than variables, indicating that some of the disturbances in the model have only temporary effects. In that case, some of the variables are cointegrated (meaning there are disturbances in the system that cause temporary deviations from a steady state, rather than stochastic trends), and predictions for Y_{t+h} can be performed using a **vector error correction model (VECM)**. However, in either case, we would need to impose the strong assumption that the $I(1)$ process Y_t follows a non-stationary VAR.

Let's derive the VECM from the VAR model. As we rewrote the AR model with a stochastic trend to perform the Dickey Fuller test in Section 4.3, we can rewrite the VAR

model by defining a lag polynomial $\Gamma(L)$ such that $\Phi(L) = \Gamma(L)(I - L) + \Phi(1)L$:

$$\Gamma(L)\Delta Y_t = \Pi Y_{t-1} + u_t, \qquad \Gamma(L) = I - \Gamma_1 L - \cdots - \Gamma_{p-1}L^{p-1}$$
$$\Gamma_l = -(\Phi_{l+1} + \cdots + \Phi_p)$$
$$\Pi = -(I - \Phi_1 - \cdots - \Phi_p) = -\Phi(1)$$

where

$$\begin{aligned}
\Gamma(L) &= [\Phi(L) - \Phi(1)L](I - L)^{-1} \\
&= \left[I - \Phi_1 L - \Phi_2 L^2 - \cdots - \Phi_p L^p - \Phi(1)L\right]\left(I + L + L^2 + \cdots\right) \\
&= I + (I - \Phi_1 - \Phi(1))L + \cdots + (I - \Phi_1 - \Phi_2 - \cdots - \Phi_{p-1} - \Phi(1))L^{p-1} \\
&\quad + \underbrace{(I - \Phi_1 - \cdots - \Phi_p - \Phi(1))}_{=0}\left(L^p + L^{p+2} + L^{p+3} + \cdots\right) \\
&= I_n - \Gamma_1 L - \cdots - \Gamma_{p-1}L^{p-1}, \qquad \Gamma_l = \Phi_{l+1} + \Phi_{l+2} + \cdots + \Phi_p
\end{aligned}$$

For those not comfortable with the above derivation, consider the VAR:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + u_t$$

Using $Y_t = Y_{t-1} + \Delta Y_t$ and $Y_{t-k} = Y_{t-1} - \sum_{s=1}^{k-1} \Delta Y_{t-k}$, for $k \geq 2$, we have:

$$\Delta Y_t = \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \cdots + \Gamma_{p-1}\Delta Y_{t-(p-1)} + u_t$$

where Π and Γ_l can be derived by comparing the two systems.

In the univariate case, the only way Y_{t-1} has a stochastic trend but ΔY_t doesn't is if $\Pi = 0$. However, in the multivariate case, it's possible that there are cointegrated relationships $\beta' Y_t$ among Y_t, so that ΔY_t may depend on $\beta' Y_{t-1}$ and thus depends on Y_{t-1}, but still be stationary. Hence, Π does not need to be zero for ΔY_t to be stationary.

However, note that the matrix Π cannot have full rank, as this would imply that both the level and differenced variables are either stationary or non-stationary. Given that a full

rank would make the matrix Π invertible, one could express the level variables in terms of the differenced variables and the stationary shocks. Consequently, the level variables could only be $I(1)$ if the differenced variables are $I(1)$ as well.

If it does happen that Y_t is $I(1)$, ΔY_t is $I(0)$, and Π as full rank, one would have to conclude that the assumption of Y_t following a non-stationary VAR(p) - and thus a VECM($p-1$) - is violated. In such a situation, a possible remedy could be to increase the lag order p until Π no longer has full rank. This stems from Wold's decomposition theorem, according to which there exists a VMA(∞) and and thus a VAR(∞) representation for ΔY_t, leading to a rank of Π equal to zero when the lag order approaches infinity. Thus, it's possible that the rank eventually decreases as the lag order increases.

It turns out that the rank of Π reveals the number of cointegrated relationships. For example, suppose there are r cointegrated relationships, and let's combine those into an $n \times r$ matrix β so that $\beta' Y_t$ is a r-dimensional vector of stationary processes. Then we have that $\text{rank}(\Pi) = r$ because there are r linearly independent ways ΔY_t can depend on Y_{t-1} and still be stationary, given that there are r cointegrated relationships. Therefore, if Π had a rank larger than r, ΔY_t would also depend on the part of Y_{t-1} that is not cointegrated, resulting in a non-stationary ΔY_t.

To extract the cointegrated relationships β, we decompose the $n \times n$ matrix Π of rank r into two $n \times r$ matrices α and β:

$$\Pi = \alpha\beta'$$

This so-called **rank factorization** always exists, but it is not unique. One can redefine the matrices $\alpha^* = \alpha M'$ and $\beta^* = \beta M^{-1}$, where M is any non-singular $r \times r$ matrix, and still obtain $\Pi = \alpha^* \beta^{*'}$.

Replacing Π with $\alpha\beta'$ gives us the **vector error correction model (VECM)**:

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \cdots + \Gamma_{p-1} \Delta Y_{t-(p-1)} + u_t$$

where $\beta' Y_t$ are r stationary processes, representing temporary deviations from cointegrated

relationships. The term $\beta' Y_t$ is sometimes referred to as the **error correction term (ECT)**, α is known as the **loading matrix**, and $\beta = \begin{bmatrix} \beta_1 & \dots & \beta_r \end{bmatrix}$ is the **cointegration matrix**, where each column β_j represents a different cointegrating relationship.

If there are no deviations from the cointegrating relationships, then $\beta' Y_t = 0$. Since these deviations are stationary and therefore only temporary, we have that the predicted deviations in the long run are zero, i.e. $\lim_{h \to \infty} E_t \left[\beta' Y_{t+h} \right] = 0$. Note that increasing the size of the loading matrix α makes deviations $\beta' Y_t$ from the cointegrating relationship less persistent, resulting in the variables reverting back to the long-run equilibrium faster.

Because α and β are not unique, it is common to apply some normalizations. If the variables in Y_t are ordered in a way that the first r variables contribute to at least one cointegrating relationship, and if combined they contribute to all r cointegrated relationships, then there exists a non-singular $r \times r$ matrix M such that $\beta^* = \beta M^{-1} = \begin{bmatrix} I_r \\ \gamma \end{bmatrix}$, where γ is an $(n-r) \times r$ matrix. For example, M may consist of the first r rows of β. This normalization allows for the following **triangular representation of a cointegrated system**:

$$\begin{aligned} Y_t^{(1)} &= -\gamma' Y_t^{(2)} + Z_t^{(1)} \\ \Delta Y_t^{(2)} &= Z_t^{(2)} \end{aligned}$$

where $Z_t^{(1)} = \beta^{*\prime} Y_t$ captures all the cointegrated relationships, and $Y_t^{(2)}$ represents the $I(1)$ stochastic trends. $Z_t = \begin{bmatrix} Z_t^{(1)\prime} & Z_t^{(2)\prime} \end{bmatrix}'$ is a stationary process as it only depends on stationary cointegrated relationships and differenced $I(1)$ processes.

Another useful representation of a cointegrated system is its Beveridge-Nelson decomposition, which rewrites the system as an infinite MA process, where some of the shocks are integrated. The Beveridge-Nelson decomposition of a VECM is often referred to as the **Granger Representation Theorem**, introduced by Johansen (1995), Theorem 4.2. To derive this representation, let $M_\perp$ be the orthogonal complement of the $n \times r$ matrix M with $\text{rank}(M) = r$, such that $M' M_\perp = 0$. If M is a nonsingular square matrix, then $M_\perp = 0$, and if $r = 0$, we define $M_\perp = I_n$.

In the triangular representation, we defined Z_t such that $Z_t^{(1)} = \lambda' Z_t = \beta^{*\prime} Y_t$ captures the stationary cointegrated relationships, and $Z_t^{(2)} = \lambda'_\perp Z_t = \lambda'_\perp \Delta Y_t$ represents the first difference of the stochastic trends, where $\lambda' = \begin{bmatrix} I_r & 0 \end{bmatrix}$, and $\lambda'_\perp = \begin{bmatrix} 0 & I_{n-r} \end{bmatrix}$. Unlike in the triangular representation, the goal here is to find a stationary stochastic process Z_t that captures all cointegrated relationships and stochastic trends without requiring any specific ordering of Y_t. We can achieve that by defining Z_t using $\lambda = \beta$:

$$\left.\begin{aligned} \beta' Z_t &= \beta' Y_t \\ \beta'_\perp Z_t &= \beta'_\perp \Delta Y_t \end{aligned}\right\} \quad \Rightarrow \quad Z_t = \beta \left(\beta'\beta\right)^{-1} \beta' Y_t + \beta_\perp \left(\beta'_\perp \beta_\perp\right)^{-1} \beta'_\perp \Delta Y_t$$

where $\beta' Z_t$ captures the cointegrated relationships, and $\beta'_\perp Z_t$ captures the stochastic trends.

Since Z_t is stationary, it has a VMA(∞) representation by Wold's decomposition theorem. The goal is to derive this representation in terms of the VECM residuals u_t and then compute the Beveridge-Nelson decomposition of Y_t by integrating the VMA representation of Z_t.

To express Z_t as a function of Y_t, define $\overline{\beta}$ and Q as follows:

$$\overline{\beta} = \beta \left(\beta'\beta\right)^{-1}, \quad Q = \begin{bmatrix} \beta' \\ \overline{\beta}'_\perp \end{bmatrix}, \quad Q^{-1}Q = QQ^{-1} = I_r \quad \Rightarrow \quad Q^{-1} = \begin{bmatrix} \overline{\beta} & \beta_\perp \end{bmatrix}$$

which then implies the following relationship between Z_t and Y_t:

$$\begin{aligned} Z_t &= \overline{\beta}\beta' Y_t + \beta_\perp \overline{\beta}'_\perp \Delta Y_t \\ &= Q^{-1} P\left(L\right) Y_t, \qquad\qquad P\left(L\right) = \begin{bmatrix} \beta' \\ \left(1-L\right)\overline{\beta}'_\perp \end{bmatrix} = \begin{bmatrix} I_r & 0 \\ 0 & \left(1-L\right) I_{n-r} \end{bmatrix} Q \end{aligned}$$

where $P\left(L\right)$ is a lag polynomial.

Next, to relate Z_t to the VECM residuals u_t, let's rewrite the VECM so that the lag

polynomial includes $P(L)$:

$$\begin{aligned}
\Psi(L) Y_t = u_t, \qquad \Psi(L) &= \Gamma(L)(1-L) - \alpha\beta' L \\
&= (\Gamma(L)(1-L) - \alpha\beta' L) Q^{-1} Q \\
&= (\Gamma(L)(1-L) - \alpha\beta' L) \begin{bmatrix} \overline{\beta} & \beta_\perp \end{bmatrix} Q \\
&= \begin{bmatrix} \Gamma(L)\overline{\beta}(1-L) - \alpha L & \Gamma(L)\beta_\perp \end{bmatrix} \begin{bmatrix} \beta' \\ (1-L)\overline{\beta}'_\perp \end{bmatrix} \\
&= M(L) P(L)
\end{aligned}$$

and therefore,

$$\begin{aligned}
M(L) P(L) Y_t &= u_t \\
M(L) Q Q^{-1} P(L) Y_t &= u_t \\
B(L) Z_t = u_t, \qquad B(L) &= M(L) Q \\
&= \Gamma(L)\overline{\beta}\beta'(1-L) - \alpha\beta' L + \Gamma(L)\beta_\perp \overline{\beta}'_\perp
\end{aligned}$$

Thus Z_t follows a stable VAR(p) process with the same residual process u_t as Y_t. Because it's stable, it has an MA(∞) representation:

$$Z_t = \Theta(L) u_t, \qquad \Theta(z) = B(z)^{-1} = Q^{-1} M(z)^{-1}, \quad \forall |z| \leq 1$$

where $\Theta(L)$ is a lag polynomial of infinite order.

Since Y_t depends on the integral of Z_t, let's compute the **Beveridge-Nelson decomposition** of the integral of Z_t (see Section 4.2 for the univariate version of this Beveridge-

Nelson decomposition):

$$\begin{aligned}
\sum_{s=1}^{t} Z_s &= \sum_{s=1}^{t} \Theta(L) u_s \\
&= \sum_{s=1}^{t} \left(\Theta^*(L)(1-L) + \Theta(1)\right) u_s, \qquad \Theta^*(L) = \left(\Theta(L) - \Theta(1)\right)(I_n - L)^{-1} \\
&= \Theta(1) \sum_{s=0}^{t} u_s + \Theta^*(L) \sum_{s=1}^{t} \Delta u_s \\
&= \Theta(1) \sum_{s=0}^{t} u_s + \Theta^*(L)(u_t - u_0)
\end{aligned}$$

where $\Theta^*(L) u_0 + \Theta(1) \sum_{s=0}^{t} u_s$ captures the permanent effects of the shocks, and $\Theta^*(L) u_t$ captures the transitory effects.

Next, let's write ΔY_t in terms of Z_t:

$$\begin{aligned}
\Delta Y_t &= Q^{-1} Q \Delta Y_t \\
&= \overline{\beta}\beta' \Delta Y_t + \beta_\perp \overline{\beta}'_\perp \Delta Y_t \\
\left[Z_t = \overline{\beta}\beta' Y_t + \beta_\perp \overline{\beta}'_\perp \Delta Y_t\right] \qquad &= \overline{\beta}\beta' \Delta Y_t + Z_t - \overline{\beta}\beta' Y_t \\
&= \overline{\beta}\beta' (Y_t - Y_{t-1}) + Z_t - \overline{\beta}\beta' Y_t \\
[\beta' Z_t = \beta' Y_t] \qquad &= Z_t - \overline{\beta}\beta' Z_{t-1}
\end{aligned}$$

Thus, multiplying both sides by $\beta_\perp \overline{\beta}'_\perp$ reveals that $\beta_\perp \overline{\beta}'_\perp \Delta Y_t = \beta_\perp \overline{\beta}'_\perp Z_t$, and therefore:

$$\Delta Y_t = \overline{\beta}\beta' \Delta Y_t + \beta_\perp \overline{\beta}'_\perp Z_t$$

Finally, let's write Y_t in terms of Z_t and thus u_t by integrating ΔY_t:

$$\begin{aligned}
Y_t &= Y_0 + \sum_{s=1}^{t} \Delta Y_s \\
&= Y_0 + \sum_{s=1}^{t} \left(\overline{\beta} \beta' \Delta Y_s + \beta_\perp \overline{\beta}'_\perp Z_s \right) \\
&= Y_0 + \overline{\beta} \beta' \left(Y_t - Y_0 \right) + \beta_\perp \overline{\beta}'_\perp \sum_{s=1}^{t} Z_s \\
&= Y_0 + \overline{\beta} \beta' Z_t - \overline{\beta} \beta' Y_0 + \beta_\perp \overline{\beta}'_\perp \Theta \left(1 \right) \sum_{s=0}^{t} u_s + \beta_\perp \overline{\beta}'_\perp \Theta^* \left(L \right) \left(u_t - u_0 \right) \\
&= Y_0 - \overline{\beta} \beta' Y_0 - \beta_\perp \overline{\beta}'_\perp \Theta^* \left(L \right) u_0 + \beta_\perp \overline{\beta}'_\perp \Theta \left(1 \right) \sum_{s=0}^{t} u_s + \left(\beta_\perp \overline{\beta}'_\perp \Theta^* \left(L \right) + \overline{\beta} \beta' \Theta \left(L \right) \right) u_t
\end{aligned}$$

which is the Beveridge-Nelson decomposition of Y_t.

Finally, let's use the same notation for the Beveridge-Nelson decomposition as in the **Granger Representation Theorem**, by Johansen (1995), Theorem 4.2 (see also Lütkepohl (2006), p. 244 - 256):

$$Y_t = Y_0^* + \Sigma \sum_{s=0}^{t} u_s + \Sigma^* \left(L \right) u_t$$

where

$$\begin{aligned}
Y_0^* &= Y_0 - \overline{\beta} \beta' Y_0 - \beta_\perp \overline{\beta}'_\perp \Theta^* \left(L \right) u_0 \\
\Sigma^* \left(L \right) &= \beta_\perp \overline{\beta}'_\perp \Theta^* \left(L \right) + \overline{\beta} \beta' \Theta \left(L \right)
\end{aligned}$$

and

$$
\begin{aligned}
\Sigma &= \beta_\perp \overline{\beta}'_\perp \Theta(1) \\
&= \beta_\perp \overline{\beta}'_\perp B(1)^{-1} \\
&= \beta_\perp \overline{\beta}'_\perp Q^{-1} M(1)^{-1} \\
&= \beta_\perp \overline{\beta}'_\perp \begin{bmatrix} \overline{\beta} & \beta_\perp \end{bmatrix} \begin{bmatrix} -\alpha & \Gamma(1)\beta_\perp \end{bmatrix}^{-1} \\
&= \beta_\perp \begin{bmatrix} 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} -\alpha & \Gamma(1)\beta_\perp \end{bmatrix}^{-1} \\
&= \beta_\perp \begin{bmatrix} 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} (\alpha'\alpha)^{-1}\alpha' \left(\Gamma(1)\beta_\perp \left(\alpha'_\perp \Gamma(1)\beta_\perp\right)^{-1} \alpha'_\perp - I_n\right) \\ \left(\alpha'_\perp \Gamma(1)\beta_\perp\right)^{-1} \alpha'_\perp \end{bmatrix} \\
&= \beta_\perp \left(\alpha'_\perp \Gamma(1)\beta_\perp\right)^{-1} \alpha'_\perp
\end{aligned}
$$

where $Y_0^* + \Sigma \sum_{s=0}^{t} u_s$ captures the permanent effects and $\Sigma^*(L)u_t$ captures the temporary effects.

A key insight of the Granger Representation Theorem is that $\beta_\perp$ and thus Σ have rank $n-r$, representing the $n-r$ stochastic trends. This implies that the permanent effects are restricted, allowing for no more than $n-r$ independent movements in the long-run. On the other hand, in the short-run, there are n independent movements arising from both the $n-r$ stochastic trends and the r temporary disturbances of the r cointegrated relationships.

7.5 Deterministic Trends in VECMs

In Section 4.2, we discovered that adding a deterministic trend $\delta^k(t)$ that follows a polynomial of degree k to an ARIMA(p,d,q) model causes the deterministic trend to become a polynomial of degree $k+d$, because the stochastic trends cause additional deterministic movements. In this section, we will show that something similar happens when adding a deterministic trend $\delta^k(t)$ to a VECM. However, since there are only $n-r$ stochastic trends in a VECM with r cointegrated relationships, we end up with $n-r$ deterministic trends that are polynomials of degree $k+1$, and r deterministic trends that remain polynomials of degree k. We will discuss under what circumstances this assumption makes sense and how to include a deterministic trend into a VECM so that all deterministic trends are

polynomials of the same degree.

Suppose the $I(1)$ stochastic process $\{Y_t\}$ follows a VAR(p) model with deterministic trends:

$$\begin{aligned}
\text{VAR}: \quad & \Phi(L) Y_t = \delta^k(t) + u_t, \\
\text{VECM}: \quad & \Psi(L) Y_t = \delta^k(t) + u_t, \qquad \Psi(L) = \Gamma(L)(1-L) - \alpha\beta' L \\
& \qquad\qquad\qquad\qquad \Gamma(L) = (\Phi(L) - \Phi(1) L)(I-L)^{-1} \\
& \qquad\qquad\qquad\qquad \alpha\beta' = \Phi(1)
\end{aligned}$$

where $\delta^k(t) = \delta_0 + \delta_1 t + \ldots + \delta_k t^k$ is a polynomial of degree k with $n \times 1$ vectors δ_i, for $i = 0, \ldots, k$. Here, $\Phi(L)$ and $\Psi(L)$ are lag polynomials representing the non-stationary VAR and VEC models, respectively.

This results in the following Beveridge-Nelson representation:

$$\begin{aligned}
Y_t &= Y_0^* + \beta_\perp \overline{\beta}'_\perp \sum_{s=0}^{t} \delta^k(s) + \overline{\beta}\beta'\delta^k(t) + \Sigma \sum_{s=0}^{t} u_s + \Sigma^*(L) u_t \\
&= Y_0^* + \beta_\perp \theta^{k+1}(t) + \beta\kappa^k(t) + \Sigma \sum_{s=0}^{t} u_s + \Sigma^*(L) u_t
\end{aligned}$$

where $\theta^{k+1}(t) = \theta_0 + \theta_1 t + \ldots + \theta_{k+1} t^{k+1}$ is a polynomial of degree $k+1$ with $(n-r) \times 1$ vectors θ_i, and $\kappa^k(t) = \kappa_0 + \kappa_1 t + \ldots + \kappa_k t^k$ is a polynomial of degree k with $r \times 1$ vectors κ_i. Hence, the system is driven by $n-r$ deterministic trends that are more sophisticated due to the $n-r$ stochastic trends, and r trends that are less sophisticated.

For example, if we have $\delta^0(t) = \delta_0$, then all n variables can have different means, but they can only depend on $n-r$ distinct linear trends. This assumption makes sense when the linear trends are orthogonal to the cointegrated relationships, meaning that the deviations from the cointegrated relationship converge to a constant in the long run, rather than converging to a linear trend. However, if the cointegrated relationships exhibit a linear divergence, then such trends need to be included in the cointegrated relationships, resulting in n different linear trends rather than $n-r$. This makes sense, for example, if output and

technology are cointegrated, but unlike technology, output experiences an additional 1% growth each year due to deterministic population growth.

We can model the deterministic trends independently of the stochastic trends by assuming that the detrended series follows a VECM without deterministic trend:

$$\Gamma(L)\,\Delta X_t = \alpha\beta' X_{t-1} + u_t, \qquad X_t = Y_t - \mu^k(t)$$

where $\mu^k(t) = \mu_0 + \mu_1 t + \ldots + \mu_k t^k$ is a polynomial of degree k with $n \times 1$ vectors μ_i, for $i = 0, \ldots, k$.

This results in the following VECM for Y_t:

$$\begin{aligned}
\Gamma(L)\,\Delta Y_t &= \Gamma(L)\,\Delta\mu^k(t) + \alpha\beta'\left(Y_{t-1} - \mu^k(t-1)\right) + u_t \\
&= \Gamma(L)\,\Delta\mu^k(t) - \alpha\beta'\sum_{s=0}^{k-1}\mu_s\,(t-1)^s + \alpha\begin{bmatrix}\beta' & -\beta'\mu_k\end{bmatrix}\begin{bmatrix}Y_{t-1}\\ (t-1)^k\end{bmatrix} + u_t \\
&= \delta^{k-1}(t) + \alpha\lambda'\begin{bmatrix}Y_{t-1}\\ t^k\end{bmatrix} + u_t
\end{aligned}$$

where $\lambda' = \begin{bmatrix}\beta' & \eta\end{bmatrix}$ is a $r \times (n+1)$ dimensional matrix with an unrestricted r-dimensional vector $\eta = -\beta'\mu_k$, and $\delta^{k-1}(t)$ is an unrestricted polynomial of degree $k-1$. Note that the degree of $\delta^{k-1}(t)$ is $k-1$ because differencing a polynomial of degree k results in reducing the degree by one, i.e., $\Delta\mu^k(t)$ is a polynomial of degree $k-1$.

The r cointegrated relationships now also include t^k so that the deterministic deviations from the cointegrated relationships become as sophisticated as the deterministic components of the stochastic trends. The result is a Beveridge-Nelson representation with an unrestricted polynomial of degree k, unlike the Beveridge-Nelson representation with two polynomials of varying degrees when the deterministic trend is included directly in the VECM.

7.6 Testing for Rank of Cointegration

If the number of cointegrating relationships r isn't known, it can be determined by estimating the VECM and then testing the rank of Π. It's more common to assume that the detrended series follow a VECM, hence, the rank of Π is tested based on the following VECM for Y_t, as per Section 7.5:

$$\Gamma(L)\Delta Y_t = \delta^{k-1}(t) + \Pi^+ Y_{t-1}^+ + u_t, \qquad \Pi^+ = \alpha\lambda', \quad Y_{t-1}^+ = \begin{bmatrix} Y_{t-1} \\ t^k \end{bmatrix}$$

However, if the deterministic trend is included directly in the VECM, being orthogonal to the cointegrating relationships, then we'd have $\Pi^+ = \alpha\beta'$, and $Y_{t-1}^+ = Y_{t-1}$, according to Section 7.5.

A smaller rank of Π^+ implies that the model is more restricted than a higher rank. A **likelihood ratio (LR) test**, which tests a more restricted model against a more general model, can be conducted. The problem is that the distribution of the LR statistic depends on the rank of Π^+ in the general model, hence, we have to specify the possible ranks of the alternative hypothesis.

Two different types of hypotheses are common in the related literature. First, the **trace statistic** $\lambda_{LR}(r_0, n)$ tests whether Π^+ has a rank of r_0 against a rank larger than r_0:

$$H_0: \ \text{rank}\left(\Pi^+\right) = r_0, \qquad H_0: \ r_0 < \text{rank}\left(\Pi^+\right) \leq n$$

and the **maximum eigenvalue statistic** $\lambda_{LR}(r_0, r_0 + 1)$ tests whether Π^+ has a rank of r_0 against a rank of $r_0 + 1$:

$$H_0: \ \text{rank}\left(\Pi^+\right) = r_0, \qquad H_0: \ \text{rank}\left(\Pi^+\right) = r_0 + 1$$

The strategy to determine the rank of Π^+ is to start at $r_0 = 0$, and then increment r_0 by one until the null hypothesis cannot be rejected for the first time. The cointegrating rank

is then chosen accordingly. This strategy works for both the maximum eigenvalue and the trace test.

7.7 Structural Vector Error Correction Model (SVECM)

In Section 6.4, we discussed how the exploration of causal relationships in VARs requires the identification of an impulse vector, which shifts the variables in a meaningful way upon impact. The impulse response function (IRF) then measures the causal effect of the exogenous event that triggered the impulse.

Similar to Section 6.4, we can identify impulse vectors by decomposing the residuals of the VECM into uncorrelated shocks:

$$u_t = A\epsilon_t$$

where A is the impact matrix, and its columns represent the impulse vectors for each of the n shocks. The resulting system of equations, where the VECM residuals u_t are replaced with $A\epsilon_t$, is referred to as a **structural vector error correction model (SVECM)**.

As in the VAR, the VECM identifies the residual covariance matrix $\Sigma = \text{Cov}(u_t)$, which is an $n \times n$ symmetric matrix. Thus, we require an additional $\frac{n(n-1)}{2}$ equations to identify the $n \times n$ impact matrix A, which is not symmetric.

In the SVECM literature, a common approach is to partition the n shocks into transitory and permanent shocks. The transitory shocks cause temporary deviations from the cointegrated relationships, while the permanent shocks contribute to the stochastic trends. From the Beveridge-Nelson representation, we can express the relationship between the shocks and the stochastic trends as follows:

$$\Sigma \sum_{s=0}^{t} u_s = \Sigma A \sum_{s=0}^{t} \epsilon_s$$

If there are r cointegrated relationships, then there are $n - r$ stochastic trends. Therefore,

we can assume that only the first $n-r$ shocks contribute to the stochastic trends, while the remaining r shocks only cause temporary deviations in the cointegrated relationships. This implies that the last r columns of $\mathbf{\Sigma}A$ are all zero. Although this results in nr zero elements in $\mathbf{\Sigma}A$, we only obtain $(n-r)r$ independent zero restrictions, because the matrix Σ has a rank of $n-r$, meaning that there are only $n-r$ independent rows in Σ and ΣA.

To understand why we only get $(n-r)r$ independent restrictions, let's perform the rank decomposition of $\Sigma = \lambda\kappa'$, where λ and κ are $n \times (n-r)$ matrices. We can then choose A such that the last r columns of the $(n-r) \times n$ matrix $\kappa' A$ are zero, which results in $(n-r)r$ zero restrictions. This implies that the last r columns of $\Sigma A = \lambda\kappa' A$ are zero as well. Hence, we need $\frac{n(n-1)}{2} - (n-r)r$ additional restrictions for identification.

It is important to note that the partition into transitory and permanent shocks does not always make sense. For example, in a VECM with only one stochastic trend resulting from the cumulative property of technological progress, multiple shocks may independently shift technology. In this case, each shock can have its own long-run effect, but the long-run effect of the shocks on the other variables will always be the same function of their overall effect on technology.

8 References

References

Beveridge, Stephen, and Charles R Nelson. 1981. "A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'." *Journal of Monetary Economics*, 7(2): 151–174.

Dickey, David A, and Wayne A Fuller. 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association*, 74(366a): 427–431.

Granger, Clive W. J., and Paul Newbold. 1974. "Spurious Regressions in Econometrics." *Journal of Econometrics*, 2(2): 111–120.

Johansen, Søren. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford University Press.

Lütkepohl, H. 2006. *New Introduction to Multiple Time Series Analysis.* Springer Berlin Heidelberg.

White, Halbert. 2000. *Asymptotic Theory for Econometricians: Revised Edition.* Emerald Group Publishing Limited.