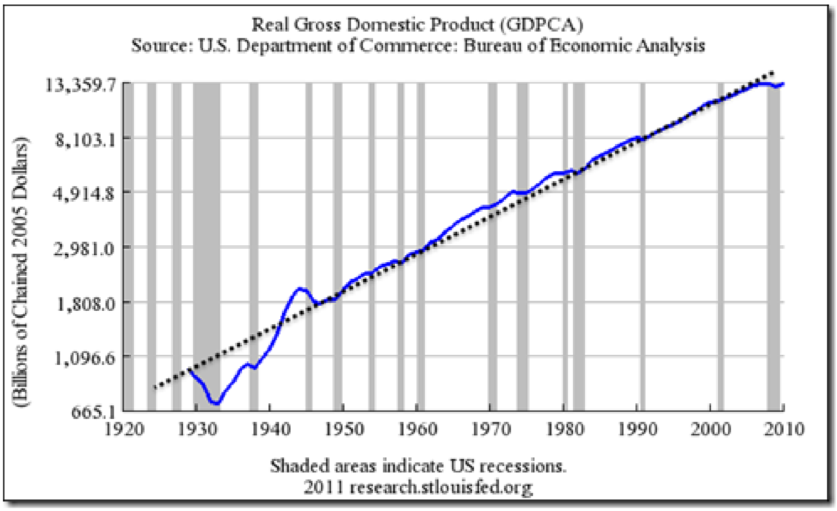
Notes on Time Series Econometrics

Diego Vilán [*]

Fall 2012

These notes cover topics in time series/macroeconometrics. Some of the topics included are: stochastic processes, with a focus on autoregressive and moving average models. Also covered are multivariate autoregressions (VARs), impulse response analysis and state-space modeling. As usual, it all begins with the data. Consider the following plot of US GDP:



The time series above can be regarded as a particular realization of a stochastic process. We will argue that, if the data generating process (DGP) satisfies certain conditions, then a single realization of the time series will be sufficient for us to understand the underlying nature of the existing relationship. This temporal homogeneity in the behavior of a series is usually referred to as stationarity.

These notes are organized as follows: First, some fundamental concepts and definitions are introduced. Next, univariate and multivariate stationary models are discussed. I believe this to be the appropriate starting point because the asymptotic theory behind these models is the same as in the standard linear regression model with which the reader might be familiar with. Additionally, these notes cover two special topics in stationary time series: State Space models and VARs with exogenous variables (VARX). Finally I cover non-stationary models.

[*] **DISCLAIMER**: I wrote these notes as a study aid for myself. They are work in progress and could be incomplete, inaccurate and even somewhat incorrect. Keep that in mind should you decide to use them. Comments and suggestions welcomed!

Contents

1 Fundamental Concepts

1.1 Stationarity

Stationarity is a probabilistically meaningful measure of regularity. This regularity may be exploited to estimate unknown parameters and characterize the dependence between observations across time. Broadly speaking, a process is said to be stationary if its probability distribution remains unchanged over time. The invariability of the structural relation (i.e.: the conditional distribution, conditional mean and variance) makes it possible to use historical data to estimate the underlying coefficients governing the DGP. It also allows us to forecast the future based on past observations.

In short, why do we care about stationary processes?

- Fundamentally because stationarity implies that the underlying relationships between variables described by the DGP is not changing over time. If the DGP would change frequently and in an unpredictable way, constructing a meaningful model of it would be almost impossible.
- Moreover, given the nature of time series observations, studying one realization of a stationary stochastic process will afford us an understanding of the existing relationship.
- **Forecasting:** If we know that the statistical relationship between a variable at different points in time is stable, this can help us forecast the variable in the future.
- Since we often think of economic equilibria as being characterized by stable, long-term relations, stationarity processes fit well that description.
- On a technical side, many standard results in econometrics (e.g.: CLT) often require stationarity. Nonstationarity will introduce several complications: estimators and test statistics will often have nonstandard asymptotic distributions.

Formal definitions:

Definition 1 *Strict stationary:*
A stochastic process $\{x_t\}$ is strictly stationary if the joint distribution of $\{x_t, x_{t+1}, ..., x_{t+h}\}$ depends only on h and not on t.

Strict stationarity requires that the joint distribution of a stochastic process does not depend on time and so the only factor affecting the relationship between two observations is the gap between them.

Strict stationarity is weaker than i.i.d. since a stationary process may be dependent. Nonetheless it is still often too strong of an assumption for most financial and macroeconomic time series. Consequently, we do not usually require strict stationarity but rather covariance stationarity.

Definition 2 *Weak or covariance stationary:*
A stochastic process $\{x_t\}$ is covariance stationary if:

$$\begin{aligned} E[x_t] &= \mu \quad \forall t \\ V[x_t] &= \sigma^2 < \infty \quad \forall t \\ E[(x_t - \mu)(x_{t-s} - \mu)] &= \gamma_j \quad \forall t \end{aligned}$$

Covariance stationarity requires that both the unconditional mean and variance are finite and do not change over time. Note that covariance stationarity only applies to unconditional moments, so a covariance process may still have varying conditional mean.

These two types of stationarity are related although neither nests the other. If a process is strictly stationary and has finite second moments, then it is covariance stationary. If a process is covariance stationary and the joint distribution of the studentized residuals does not depend on time, then the process is strictly stationary.

For purposes of these notes, stationarity will usually imply weak/ covariance stationarity.

Remark 1 *Most economic time series in their original form are usually not stationary. It is widely accepted that time series are often governed by four main components that can be associated to different types of temporal variations*[1]*:*

1. *A trend of long term movement*
2. *A seasonal effect*
3. *A cyclical component*
4. *A residual, irregular or random effect*

Nonstationary time series often undergo some transformation to become stationary (for e.g.: differencing or detrending). These notes will only briefly mention some of these topics in the appendix and focus almost exclusive on stationary processes.

1.2 Stochastic Process

We mentioned that a time series could be regarded as a particular realization of a stochastic process. But what is a stochastic process exactly?

Definition 3 *Stochastic process:*
Any process that generates a sequence of random variables $\{x\}$ that are ordered in an immutable fashion. One such way of ordering is by time, usually noted as $\{x_t\}$.

Alternative definition

Definition 4 *Stochastic process:*
Any process that is ordered (i.e.: is a sequence that is immutable) is a stochastic process. A time series is an example of a stochastic process.

Note that a stochastic process is a sequence of random variables, while a realization of such a process is a sequence of real numbers.

Associated with this stochastic process is its history:

[1] These four components are usually combined together either using an additive or multiplicative model.

$$H_t = \{x_t, x_{t-1}, x_{t-2}, ...\}$$

The expected value at time t of a particular random variable is:

$$E_t(x_s) = E(x_s|H_t)$$

or its expected value conditional on all information available at time t.

2 Stationary Univariate models

There are two special classes of stochastic processes that are particularly useful for Macroeconomics: (i) Markov Processes and (ii) Linear Stochastic Processes.

2.1 Markov Processes

2.1.1 Definition

A Markov process is a stochastic process which satisfies the following property:

$$P(x_{t+1}|H_t) = P(x_{t+1}|x_t)$$

where $P(x)$ is the probability distribution of x, and $H_t = \{x_{t-1}, x_{t-2}, ...\}$ is the history of realizations. The so called Markov-property establishes that the current state of the system (i.e.: x_t) is a sufficient statistic to form forecasts about the future of the system. In other words, knowing $x_{t-j} \quad \forall j > 0$ provides no additional information of future values of $x_t + j$ than knowing x_t does.

Definition 5 *Markov Chain:*
A Markov chain is a Markov process with a countable space. A Markov chain can be defined using three elements:

1. *An n-dimensional vector $\omega \in \mathrm{R}^{\mathrm{n}}$ defining the state space*
2. *An n-dimensional vector of initial probabilities π_0*

$$\pi_{0i} = P(x_0 = \omega_i)$$

3. *An n-by-n transition matrix:*

$$P = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{21} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \ldots & \vdots \\ p_{1N} & 0 & \cdots & p_{NN} \end{bmatrix}$$

where $p_{ij} = P(x_{t+1} = \omega_j | x_t = \omega_i)$. The row tells you the current state , the column tells you the probabilities of transitioning to each possible state in the next period.

In order for everything to be well defined, we also require:

$$\sum_{i=0}^{n} \pi_{0i} = 1$$
$$\sum_{j=0}^{n} p_{i,j} = 1$$

Also, let the n-vector π_t be defined as:

$$\pi_{ti} = P(x_t = \omega_i)$$

Note that π_t is not a random object.

Why might we find Markov chains convenient for modeling? Two main reasons:

1. Dynamic Programming problems are often easier to solve for a discrete state space and there is usually a Markov chain which is "close enough" to another stochastic process to serve as a more convenient approximation.
2. Several of the properties of a given Markov chain can be easily derived from its transition matrix.

Example:

Assume there exists two states of the world: high and low income. The vector of states would be:

$$x_t = \begin{pmatrix} H \\ L \end{pmatrix}$$

Lets assume that if income is high today, there is a 70 % chance it will continue to be high tomorrow and a 30 % chance it will be a low income state. Further, if we are today in a low income state there is a 40 % chance we will be in a high income tomorrow and a 60 % chance we will be in a low income state. This defines our transition matrix as:

$$P = \begin{pmatrix} 0.70 & 0.30 \\ 0.40 & 0.60 \end{pmatrix}$$

Finally, lets argue there is a 75 % chance that we will initially find ourselves in a high income state and 25 % in a low income one:

$$\pi_0 = \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix}$$

With this set-up, we can compute several useful probabilities:

1. The unconditional probability of being in a high state in period 1:
   P(starting in a high state at $t = 0$ and remaining in a high state at $t = 1$) = 0.75 x 0.70 = 0.525 +
   P(starting in a low state at $t = 0$ and transitioning into a high state at $t = 1$) = 0.25 x 0.40 = 0.1
   = 0.625
2. The unconditional probability of being in a low state in period 1:
   P(starting in a low state at $t = 0$ and remaining in a low state at $t = 1$) = 0.75 x 0.30 = 0.xxx +
   P(starting in a high state at $t = 0$ and transitioning into a high state at $t = 1$) = 0.25 x 0.60 = 0.xxx
   = 0.375
3. The unconditional probability for being in the high and low states in period 1:

$$\begin{aligned} \pi_1^{'} &= \pi_0^{'} P = (0.75, 0.25) \begin{pmatrix} 0.70 & 0.30 \\ 0.40 & 0.60 \end{pmatrix} \\ &= (0.75 * 0.7 + 0.25 * 0.4, 0.75 * 0.30 + 0.25 * 0.60) \\ &= (0.625, 0.375) \end{aligned}$$

4. The unconditional probability for being in the high and low states in period 2:

$$\begin{aligned} \pi_2^{'} &= \pi_1^{'}P = \pi_0^{'}PP \\ &= \pi_0^{'}P^2 \end{aligned}$$

5. The unconditional probability for being in the high and low states in period k:

$$\pi_k^{'} = \pi_0^{'}P^k$$

Under this framework expectations k periods ahead can easily be computed. Suppose that a high income state takes a value of 10, and a low income state takes a value of 5. Given this we can compute the conditional expectations of income in the next period:

1. If income today is <u>high</u>, the conditional expected income next period is:

$$P(x|H) = 0.70 * 10 + 0.30 * 5 = 8$$

2. If income today is <u>low</u>, the conditional expected income next period is:

$$P(x|L) = 0.40 * 10 + 0.60 * 5 = 7$$

2.1.2 Stationary distributions

The unconditional probability distributions evolve by:

$$\pi_{t+1}^{'} = \pi_t P$$

The unconditional distributions is called stationary or invariant if it satisfies:

$$\begin{aligned} \pi_{t+1} &= \pi_t \\ \pi^{'} &= \pi^{'}P \\ &\Rightarrow \pi^{'}(I - P) = 0 \\ &= (I - P^{'})\pi = 0 \end{aligned}$$

which implies that π is the eigenvector associated with the unit eigenvalue of $P^{'}$.

We can prove that there will always be at least one stationary distribution for any transition matrix. When will there be exactly one?

Proposition 1 *Let P be a Markov transition matrix such that there exists $n \geq 1$ such that $(P^n)_{ij} > 0 \; \forall i, j$. Then P is asymptotically stationary and has a unique stationary distribution.*

In other words, if every element of P is strictly positive, then for any initial distribution π_0 :

$$\lim_{t \to \infty} \pi_t = \pi$$

2.2 Linear (stationary) Stochastic Processes

2.2.1 White Noise

The simplest form of a covariance (or weakly) stationary process is the *white noise*[2] process.

Definition 6 *White noise:*
A stochastic process $\{\varepsilon_t\}$ is a white noise process if the following holds:

$$\begin{aligned} E(\varepsilon_t) &= 0 \quad \forall t \quad \textit{(zero mean)} \\ E(\varepsilon_t^2) &= \sigma^2 < \infty \quad \forall t \quad \textit{(constant \& finite variance)} \\ E(\varepsilon_t \varepsilon_\tau) &= 0 \quad \forall t \neq \tau \quad \textit{(serially uncorrelated)} \end{aligned}$$

In discrete time, white noise is a discrete process whose samples are regarded as a sequence of serially uncorrelated random variables with zero mean and finite variance.

It should be noted that although $\{\varepsilon_t\}$ are serially uncorrelated, they are not **necessarily** serially independent, since they are not **necessarily** normally distributed[3]. However, it is common to require that each sample has a Normal distribution with zero mean. In turn, this process is called Gaussian White Noise.

Definition 7 *Gaussian White Noise:*
A stochastic process $\{\varepsilon_t\}$ is a Gaussian White Noise process if the following holds:

$$\begin{aligned} E(\varepsilon_t) &= 0 \quad \forall t \\ E(\varepsilon_t^2) &= \sigma^2 < \infty \quad \forall t \\ E(\varepsilon_t \varepsilon \tau) &= E(e_t)E(e_\tau) \quad \forall t \neq \tau \\ \varepsilon_t &\sim N(0, \sigma^2) \end{aligned}$$

The classic example is to assume that the sample is independent and identically distributed and has identical probability distribution, with a normal distribution.

The Gaussian white noise process is important because it will be the foundation for most of the other stochastic process we are interested in. For example, we can use a white noise process to construct a wide range of MA, AR, and ARMA processes. We begin by considering the moving average (MA) model.

[2] Why white noise? This name comes from the engineering literature where certain instruments are able to analyze the light waves into components of various frequencies. The white light has the property that all frequencies enter equally. It turns out that with truly iid realizations of a random variable normally distributed as inputs, one obtains a flat spectrum for the light waves which is similar to that of white light. Just as all rainbow colors can be obtained from white color, **many stationary processes can be written as a linear combination of a white noise process.**

[3] Recall that zero correlation only implies independence for normally distributed random variables.

2.2.2 Moving Average (MA) Models

The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic term. A moving-average model is conceptually a linear regression of the current value of the series against current and previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to be mutually independent and to come from the same distribution.

Together with the autoregressive (AR) model, the moving-average model is a special case and key component of the more general ARMA and ARIMA models of time series, Both models have a more complicated structure, and are covered later in these notes.

Definition 8 *Moving Average of order 1: MA(1)*
The stochastic process $\{y_t\}$ is a moving average process of order 1 *if:*

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

where $\{\varepsilon_t\} \sim WN(N, \sigma^2)$ with the additional property that $E_{t-1}(\varepsilon_t) = 0^4$. The defining characteristic of the MA process in general, and MA(1) in particular, is that the current value of the observed series is expressed as a function of current and lagged unobservable innovations.

2.2.2.1 Unconditional and conditional means

The unconditional mean is:

$$\begin{aligned} E(y_t) &= E(\mu + \theta\varepsilon_{t-1} + \varepsilon_t) \\ &= \mu + \theta E(\varepsilon_{t-1}) + E(\varepsilon_t) \\ &= \mu + \theta 0 + 0 \\ &= \mu \end{aligned}$$

The conditional mean is:

$$\begin{aligned} E_{t-1}(y_t) &= E_{t-1}(\mu + \theta\varepsilon_{t-1} + \varepsilon_t) \\ &= \mu + \theta E_{t-1}(\varepsilon_{t-1}) + E_{t-1}(\varepsilon_t) \\ &= \mu + \theta\varepsilon_{t-1} + 0 \\ &= \mu + \theta\varepsilon_{t-1} \end{aligned}$$

The differences in the means reflect the persistence of the previous shocks in the current period. The variances can be similarly derived.

[4] This assumption follows from the fact that the innovation is unpredictable using the time $t-1$ information set. The process is called a moving average since it is built as a weighted average of a white noise process. The term $\{\varepsilon_t\}$ is often called an **innovation**.

2.2.2.2 Unconditional and conditional variance

The unconditional variance is:

$$\begin{aligned}
V(y_t) &= E[(\mu + \theta\varepsilon_{t-1} + \varepsilon_t - E[\mu + \theta\varepsilon_{t-1} + \varepsilon_t])^2)] \\
&= E[(\mu + \theta\varepsilon_{t-1} + \varepsilon_t - \mu)^2] \\
&= [\mu + \theta\varepsilon_{t-1} + \varepsilon_t - \mu)^2] \\
&= \theta^2 E[\varepsilon_{t-1}^2] + E[\varepsilon_t^2] + 2\theta E(\varepsilon_{t-1}\varepsilon_t)] \\
&= \sigma^2\theta^2 + \sigma^2 + 0 \\
&= \sigma^2(1 + \theta^2)
\end{aligned}$$

where $E(\varepsilon_{t-1}\varepsilon_t = 0)$ follows from the white noise assumption.

The conditional variance is:

$$\begin{aligned}
V_{t-1}(y_t) &= E_{t-1}[(\mu + \theta\varepsilon_{t-1} + \varepsilon_t - E_{t-1}[\mu + \theta\varepsilon_{t-1} + \varepsilon_t])^2)] \\
&= E_{t-1}[(\mu + \theta\varepsilon_{t-1} + \varepsilon_t - \mu - \theta\varepsilon_{t-1})^2] \\
&= E_{t-1}[\varepsilon_t^2] \\
&= \sigma_t^2
\end{aligned}$$

where σ_t^2 is the conditional variance of $\{\varepsilon_t\}$. White noise processes do not necessarily have to be homoskedastic, although if $\{\varepsilon_t\}$ is, then $V_{t-1}(y_t) = \sigma^2$. Like the mean, the unconditional variance and the conditional variance are different. The unconditional variance is unambiguously larger than the average conditional variance due to the extra variability introduced by the moving average term.

The autocovariances[5] can be derived as:

$$\begin{aligned}
\gamma_{1t} &= E[(y_t - E(y_t))(y_{t-1} - E(y_{t-1}))] \\
&= E[(\mu + \theta\varepsilon_{t-1} + \varepsilon_t - \mu)(\mu + \theta\varepsilon_{t-2} + \varepsilon_{t-1} - \mu)] \\
&= E[\theta\varepsilon_{t-1}^2 + \theta\varepsilon_t\varepsilon_{t-2} + \varepsilon_t\varepsilon_{t-1} + \theta^2\varepsilon_{t-1}\varepsilon_{t-2}] \\
&= \theta\sigma^2 + 0 + 0 + 0 \\
&= \theta\sigma^2
\end{aligned}$$

$$\begin{aligned}
\gamma_{2t} &= E[(y_t - E(y_t))(y_{t-2} - E(y_{t-2}))] \\
&= E[(\mu + \theta\varepsilon_{t-1} + \varepsilon_t - \mu)(\mu + \theta\varepsilon_{t-3} + \varepsilon_{t-2} - \mu)] \\
&= E[(\theta\varepsilon_{t-1} + \varepsilon_t)(\theta\varepsilon_{t-3} + \varepsilon_{t-2})] \\
&= E[\theta\varepsilon_{t-1}\varepsilon_{t-2} + \theta\varepsilon_{t-3}\varepsilon_t + \varepsilon_t\varepsilon_{t-2} + \theta^2\varepsilon_{t-1}\varepsilon_{t-3})] \\
&= \theta E[\varepsilon_{t-1}\varepsilon_{t-2}] + \theta E[\varepsilon_{t-3}\varepsilon_t] + E[\varepsilon_t\varepsilon_{t-2}] + \theta^2 E[\varepsilon_{t-1}\varepsilon_{t-3})] \\
&= 0 + 0 + 0 + 0 = 0
\end{aligned}$$

[5] For a refresher on autocovariances and autocorrelation refer to the appendix.

2.2.2.3 Autocovariances and autocorrelation functions

The MA(1) process has a non-zero autocorrelation at lag 1:

$$\begin{aligned}\rho &= \frac{\gamma_1}{\gamma_0} \\ &= \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} \\ &= \frac{\theta}{(1+\theta^2)}\end{aligned}$$

the value of the parameter depends on θ, but note that its max value is 0.5.

The MA(1) can be generalized into the class of MA(q) processes by including additional lagged errors.

Definition 9 *Moving Average of order q: MA(q)*
The stochastic process $\{y_t\}$ is a moving average process of order q if:

$$\begin{aligned}y_t &= \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + ... + \theta_q\varepsilon_{t-q} \\ &= \varepsilon_t + \sum_{q=1}^{Q}\theta_q\varepsilon_{t-q} \\ &= \Theta(L)\varepsilon_t\end{aligned}$$

where $\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q$ is a q-ordered lag-operator polynomial, $\{\theta_1, \theta_2, ..., \theta_q\}$ is a sequence of real numbers, and $\{\varepsilon_t\}$ is a white noise process with the additional property that $E_{t-1}(\varepsilon_t) = 0$.

The MA(q) process is a natural generalization of the MA(1) one. By allowing for more lags of the shock on the right side of the equation, the MA(q) process can capture richer dynamics.

Properties:

1. Unconditional mean: $E(y_t) = \mu \quad \forall t$
2. Unconditional variance: $V(y_t) = (1 + \sum_{q=1}^{Q}\theta_q^2)\sigma^2 \quad \forall t$
3. $$\gamma_{jt} = \begin{cases}\sigma^2\sum_{i=0}^{q-j}\theta_i\theta_{i+j} & \text{if } j \leq q \\ 0 & \text{if } j > q\end{cases}$$

   where $\theta_0 = 1$.

The potential longer memory of the MA(q) process emerges clearly in its autocorrelation function. In the MA(1) case, all autocorrelations beyond displacement 1 are zero, while in the MA(q) case all autocorrelations beyond displacement q are zero. This autocorrelation cutoff is a distinctive property of MA processes.

Definition 10 *Infinite Moving Average: MA(∞)*
The stochastic process $\{y_t\}$ is a moving average process of infinite order if:

$$\begin{aligned} y_t &= \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \\ &= \varepsilon_t + \sum_{q=1}^{\infty} \theta_q \varepsilon_{t-q} \end{aligned}$$

2.2.2.4 Stationarity conditions

Stability of a MA process:

- A MA(1) process is always covariance-stationary.
- A MA(q) process is covariance-stationary as long as $\theta_j < \infty \quad \forall j = 1, ..., q$.
  This implies that the MA(q) process will be stable as long as it is of finite order.
- A MA(∞) process is covariance-stationary if:

  $$\sum_{j=0}^{\infty} \theta_j^2 < \infty$$

  That is to say, if the process is square summable, then it is covariance-stationary.

2.2.2.5 Impulse response functions

The MA process is fundamental in analyzing the dynamics of economic systems, and is used to construct impulse response functions (IRFs). These respond to the question: how do shocks "today" affect this random variable "today" and into the future? In the case of a MA(1) process:

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

by construction a one unit shock to ε_t "today" changes the random variable by that same amount "today". "Tomorrow" this shock affects y_t by a factor of θ. Think that "tomorrow", ε_t becomes ε_{t-1} as we refer to the original innovation.

In a higher order MA(q) process, a shock "today" affects y_t for more periods. In particular, the number of periods into the future that a shock today has an effect is given by the order of the process.

2.2.2.6 Wold Decomposition Theorem

The MA process also plays an important role in the Wold Decomposition Theorem. The theorem states that any mean-zero covariance-stationary process can be written as an infinite moving average process plus a deterministic term.

$$y_t = \sum_{i=0}^{\infty} \varepsilon_{t-i} + k_t$$

In other words, any mean zero covariance-stationary process has a moving average representation. This will be particularly useful when working with autoregressive processes.

2.2.3 Autoregressive (AR) models

Another important class of stochastic, stationary process is the autoregressive process (AR).

Definition 11 *First Order Autoregressive Process:*
A first order autoregressive process -AR(1)- has dynamics which follow:

$$y_t = \mu + \phi_1 y_{t-1} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise process with the additional property that $E_{t-1}(\varepsilon_t) = 0$.

Unlike the MA(1) process, y appears on both sides of the equation. However, this is just a convenience and the process can be re-written to provide an expression that depends only on the disturbances ϵ_t and an initial condition y_0.

$$\begin{aligned}
y_t &= \mu + \phi_1 y_{t-1} + \varepsilon_t \\
&= \mu + \phi_1(\mu + \phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
&= \mu + \phi_1 \mu + \phi_1^2 y_{t-2} + \varepsilon_t + \phi_1 \varepsilon_{t-1} \\
&= \mu + \phi_1 \mu + \phi_1^2(\mu + \phi_1 y_{t-3} + \varepsilon_{t-2}) + \varepsilon_t + \phi_1 \varepsilon_{t-1} \\
&\vdots \\
&= \sum_{i=0}^{t-1} \phi_1^i \mu + \sum_{i=0}^{t-1} \phi_1^i \varepsilon_{t-i} + \phi_1^t y_0
\end{aligned}$$

Substituting backwards, a stationary AR(1) can be expressed as an MA(t) process. In many cases the initial condition is unimportant and the AR process can be assumed to have begun long ago in the past. In that case, as long as $|\phi_1| < 1$, taking $\lim_{t \to \infty} \phi^t y_0 \to 0$ implies the effect of the initial condition will be negligible.

Using the infinite history version of an AR(1) (and assuming that $|\phi_1| < 1$), then the above solution simplifies to:

$$\begin{aligned}
y_t &= \sum_{i=0}^{\infty} \phi_1^i \mu + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} + 0 \\
&= \frac{\mu}{1 - \phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}
\end{aligned}$$

where the identity $\sum_{i=0}^{\infty} = \frac{1}{1-\phi_1}$ is used.

This expression of an AR process is known as an infinite moving average representation -MA(∞)- and it is useful for deriving properties.

2.2.3.1 Unconditional and Conditional mean

The unconditional mean:

$$\begin{aligned} E(y_t) &= E\left[\frac{\mu}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}\right] \\ &= \frac{\mu}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i E(\varepsilon_{t-i}) \\ &= \frac{\mu}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i 0 \\ &= \frac{\mu}{1-\phi_1} \end{aligned}$$

As long as $\{y_t\}$ is covariance stationary (so that $E(y_t) = E(y_{t-1}) = \mu$) the unconditional mean may also be derived as:

$$\begin{aligned} E(y_t) &= E(\mu + \phi_1 y_{t-1} + \varepsilon_t) \\ &= \mu + \phi_1 E(y_{t-1}) + E(\varepsilon_{t-1}) \\ \mu &= \mu + \phi_1 \mu + 0 \\ \mu(1-\phi_1) &= \mu \\ E(y_t) &= \frac{\mu}{1-\phi_1} \end{aligned}$$

2.2.3.2 Unconditional and Conditional variance

The unconditional variance (assuming stationarity):

$$\begin{aligned} V(y_t) &= V(\mu + \phi_1 y_{t-1} + \varepsilon_t) \\ &= V(\mu) + V(\phi_1 y_{t-1}) + V(\varepsilon_t) + 2Cov(\phi_1 y_{t-1}, \varepsilon_t) \\ &= 0 + \phi_1^2 V(y_{t-1}) + \sigma^2 + 2 * 0 \\ &= \phi_1^2 V(y_t) + \sigma^2 \\ V(y_t) - \phi_1^2 (y_t) &= \sigma^2 \\ V(y_t) &= \frac{\sigma^2}{1-\phi_1^2} \end{aligned}$$

where $Cov(y_{t-1}, \varepsilon_t) = 0$ follows from the white noise assumption given that y_{t-1} is a function of $\varepsilon_{t-1}, \varepsilon_{t-2}, ...$

The conditional variance:

$$\begin{aligned} V_{t-1}(y_t) &= E_{t-1}[(\mu + \phi_1 y_{t-1} + \varepsilon_t)^2] \\ &= E_{t-1}(\varepsilon_t^2) \\ &= \sigma_t^2 \end{aligned}$$

The unconditional variance is again larger than the average conditional variance and the variance explodes as $|\phi_1|$ approaches 1 or -1.

2.2.3.3 Autocovariances and autocorrelation functions

The jth autocovariance is:

$$\gamma_{jt} = \sigma^2 \frac{\phi_1^j}{1 - \phi_1^2}$$

The jth autocorrelation is:

$$\gamma_j = \frac{\gamma_j}{\gamma_0} = \phi^j$$

The AR(1) can be extended to the AR(p) class by including additional lags of y_t. A second order autoregressive process -AR(2)-:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

while a p-order autoregressive process -AR(p)-:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} \varepsilon_t$$

2.2.3.4 Stationarity conditions

$\boxed{\text{Stability of an AR process:}}$

The stability conditions for univariate AR processes are:

1. An AR(1) process

$$y_t \quad = \quad \mu + \phi_1 y_{t-1} + \varepsilon_t$$

is covariance-stationary if $|\phi| < 1$

2. An AR(p) process

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} \varepsilon_t$$

is covariance-stationary if the roots of the:

   (a) Characteristic polynomial: $0 = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p$ lie <u>outside</u> the unit circle
   (b) Inv characteristic polynomial: $0 = z^p - \phi_1 z^{p-1} - \phi_2 z^{p-2} - \ldots - \phi_p$ lie <u>inside</u> the unit circle

3. An AR(∞) process

$$y_t = \mu + \sum_{i=1}^{\infty} \phi_i y_{t-i} \varepsilon_t$$

is covariance-stationary if xx

2.2.3.5 Special: Random walk

COMPLETE

2.2.4 Autoregressive Moving Average (ARMA) models

Autoregressive moving average (ARMA) processes will form the core of time-series analysis. The ARMA class can be decomposed into two smaller classes: autoregressive (AR) and moving average (MA) processes.

2.2.4.1 Unconditional and Conditional mean

2.2.4.2 Unconditional and Conditional variance

2.2.4.3 Invertibility I: From ARMA to MA representation

2.2.4.4 Invertibility II: From ARMA to AR representation

2.2.4.5 Autocovariances

2.2.5 The Box-Jenkins approach

2.2.6 Model Selection criteria

3 Stationary Multivariate models

Multivariate time-series analysis extends many of the ideas of univariate time series to systems of equations. Perhaps the most successful model is the vector autoregression (VAR) model, a direct and natural extension of the univariate autoregression one. Most results that apply to univariate time-series can be directly ported to multivariate time series with a minimal change of notation and the use of linear algebra.

3.1 Vector Autoregressive (VAR) models

There are several circumstances in which, in order to capture the relationship between variables in a dynamic system (like an economy), a single equation model may not be enough. We then need a system of equations to accurately describe the data generating process.

3.1.1 Vector Autoregressive process

A VAR is an econometric model used to capture the linear interdependencies among multiple time series. VAR models generalize the univariate autoregressive model (AR model) by allowing for more than one evolving variable. All variables in a VAR enter the model in the same way: each variable has an equation explaining its evolution based on its own lags and the lags of the other model variables.

A p-th order VAR, denoted VAR(p), is:

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t \tag{1}$$

where the l-periods back observation y_{t-l} is called the l-th lag of y, c is a k x 1 vector of constants (intercepts), A_i is a time-invariant k x k matrix and e_t is a k x 1 vector of error terms satisfying:

$$\begin{aligned} E(e_t) &= 0 \quad \forall t \\ E(e_t e_t^{'}) &= \Omega \\ E(e_t e_{t-k}^{'}) &= 0 \quad \forall k \neq 0 \end{aligned}$$

In words:

i Every error term has zero mean

ii The contemporaneous covariance matrix of error terms is Ω (a k x k positive semidefinite matrix)

iii There is no correlation across time; in particular, no serial correlation in individual error terms

In short, the vector e_t represents the multivariate equivalent of the univariate white noise process in the AR model.

Estimation:
Lets assume that (1) is a bivariate system. We then have:

$$\begin{aligned} x_t &= a_{10} + a_{11} x_{t-1} + a_{12} z_{t-1} + e_{1t} && (2) \\ z_t &= a_{20} + a_{21} x_{t-1} + a_{22} z_{t-1} + e_{2t} && (3) \end{aligned}$$

or more compactly:

$$\begin{aligned} y_t &= A_0 + A_1 y_{t-1} + e_t \qquad (4)\\ &: \quad y_t = \begin{pmatrix} x_t \\ z_t \end{pmatrix}\\ &: \quad A_0 = \begin{pmatrix} a_{10} \\ a_{20} \end{pmatrix}\\ &: \quad A_1 = \begin{pmatrix} a_{11} & a_{11} \\ a_{21} & a_{22} \end{pmatrix}\\ &: \quad e_t = \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \end{aligned}$$

If the RHS of (4) contains only predetermined variables, and the error terms are assumed to be serially uncorrelated with constant variance, then each equation in the system can be estimated using OLS. As long as there are identical regressors in each equation, OLS provides consistent and asymptotically efficient estimates. Further, even if errors are correlated across equations, seemingly unrelated regressions (SUR) does not add to the efficiency of the estimation procedure since all regressions have identical RHS variables. If, however, some VAR equations have regressors not included in the others, SUR does provide efficient estimates of the VAR coefficients [6].

VAR models major contribution

In the 1960s, macroeconometric hypothesis testing and forecasts were conducted using large-scale models usually consisting of a system of equations involving structural and behavioral relationships that were estimated one at a time. However, these models were rarely able to capture the dynamic feedback mechanism (reverse causality) as they traditionally would contain explanatory/ independent variables and explained/dependent ones.

The example above (1) reflects the natural generalization of the univariate autoregressive models, by allowing for more than one evolving variable. All variables in a VAR enter the model in the same way: each variable has an equation explaining its evolution based on its own lagged values, the lagged values of the other model variables, and an error term. Because the LHS variables only depend on lagged (or predetermined) values, this specification can be estimated via OLS.

This model, however, does not capture the main strength of the VAR methodology which resides in modeling contemporaneous feedback relationships in an system of equations. In other words, in a VAR variables could be modeled as having contemporaneous influence on each other rendering OLS estimation unfeasible. This is mostly useful when one wishes to solve a reserve causality problem, or simply when we are unsure that a variable is actually exogenous, and so the natural solution is to treat each variable symmetrically. This has lead to different approaches in the identification strategy of VARs which can be roughly defined by the starting point: either a structural or a reduced VAR model.

3.1.2 Reduced form v.s. structural VAR models

A VAR can be quite useful in examining the statistical relationship among a set of economic variables. Moreover, the resulting estimates can be employed for forecasting or variance decomposition purposes. And although these are very important uses, the VAR approach is generally devoided of much

[6] For a refresher on SUR, please refer to my notes on cross-section econometrics

economic content. The sole role of the economist is to suggest the appropriate variables to include in the VAR and from that point onwards the procedure is almost mechanical. Since there is so little economic input in a VAR, there is usually little economic content and interpretation in its results.

There are, nonetheless, two principal ways in which a VAR model could offer a greater degree of economic insight. One could begin with a structural model where the relationship amongst variables is based on economic theory. Alternatively, the researcher could choose to being by estimating a reduced form VAR and seek to recover the underlying structural model described by his or her choice of identifying assumptions.

Naturally, both approaches are intimately related and the starting point choice depends on the researcher's objectives. Beginning from a structural VAR imposes more theoretical discipline and delivers greater economic insights. Starting from a reduced-form VAR is most useful when forecasting or variance decomposition are the main objectives.

In what follows I begin by providing an example of the estimation and identification of a structural VAR with a pervasive simultaneity problem (reverse causality). We'll use a bivariate model with two endogenous variables x_t and z_t. In this framework, the path of $\{x_t\}$ will be affected by the current and past realizations of the $\{z_t\}$ sequence, and the path of $\{z_t\}$ is also affected by the current and past realizations of the $\{x_t\}$ sequence.

3.1.3 Structural VAR (SVAR) models

Structural vector autoregressive (SVAR) models were introduced in 1980 as an alternative to traditional large-scale macroeconometric ones when the theoretical and empirical support for these models became increasingly doubtful. VAR models were first proposed by Sims (1980) as an alternative to traditional large-scale dynamic simultaneous equation models. Sims' research program stressed the need to dispense with ad hoc dynamic restrictions in regression models and to discard empirically implausible exogeneity assumptions. He also stressed the need to model all endogenous variables jointly rather than one equation at a time.

Structural interpretations of VAR models require additional identifying assumptions that must be motivated based on institutional knowledge, economic theory, or other extraneous constraints on the model responses. Only after decomposing reduced-form errors into structural shocks that are mutually uncorrelated and have an economic interpretation can we assess the causal effects of these shocks on the model variables.

.

3.1.3.1 Origins of SVAR modeling

The original meaning of a "structural" model in econometrics is explained in an article by Hurwicz (1962)[7]. A model is structural if it allows us to predict the effect of "interventions" — deliberate policy actions, or changes in the economy or in nature of known types. To make such a prediction, the model must tell us how the intervention corresponds to changes in some elements of the model (parameters, equations, observable or unobservable random variables), and it must be true that the changed model is an accurate characterization of the behavior being modeled after the intervention.

In the traditional simultaneous equations models that Hurwicz had in mind, the intervention was ordinarily taken to correspond to changing the parameters in an equation or block of equations in the model. The simplest conceptual example, corresponding to the monetary VAR literature, is where one block of equations describes policy behavior and another describes private sector behavior. The model is claimed to be structural because one set of policy equations can be replaced by another, while leaving the private sector equations unchanged, to obtain a prediction about the behavior of the economy with the new monetary policy.

However, there is no need for the intervention to correspond to changing an equation. In a model derived from a general equilibrium, for example, the natural parameters of the model (from utility functions, production functions, policy makers' objective functions) are likely to appear in many equations of the model. Such a model will claim to be structural relative to changes in at least some of these natural parameters — policy makers' objective functions, for example. One way to describe the Lucas critique of econometric policy advice is to say that he pointed out that parameters characterizing monetary policy behavior are likely to appear, via expectations, in many equations of the model, not just in the "policy equations". Thus an attempt to predict the effects of a policy change by changing only the policy equation, holding other equations in the model fixed, will fail, because the other equations will in fact change when the policy changes.

[7] This sub-section is non-technical and provides a bit of history. The reader might chose to skip it if not interested. It very closely follows the introduction of Sims (2002)

3.1.3.2 SVAR basic set-up:

The main conceptual difference embedded in a SVAR, is that behind the system of equations there is a sound theoretical hypothesis. Currently, we usually associate a SVAR with the solution to a micro-founded model, possibly a DSGE. This framework provides the necessary theoretical underpinning upon which the SVAR is derived from.

In view of the above, consider the following SVAR where variables affect each other contemporaneously. We begin with the following bivariate system:

$$\begin{aligned} x_t &= \gamma_{10} - b_{12} z_t + \gamma_{11} x_{t-1} + \gamma_{12} z_{t-1} + \varepsilon_{xt} && (5) \\ z_t &= \gamma_{20} - b_{21} x_t + \gamma_{21} x_{t-1} + \gamma_{22} z_{t-1} + \varepsilon_{zt} && (6) \\ &: \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{pmatrix} \sim \text{iid}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{pmatrix} \right) && (7) \end{aligned}$$

In matrix form, the model becomes:

$$\begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{pmatrix} \quad (8)$$

or, more compactly:

$$\begin{aligned} By_t &= \Gamma_0 + \Gamma_1 y_{t-1} + \varepsilon_t && (9) \\ &: \; y_t = \begin{pmatrix} x_t \\ z_t \end{pmatrix} \\ &: \; E(\varepsilon_{xt} \varepsilon_{zt}^{'}) = \Sigma_\varepsilon = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{pmatrix} \end{aligned}$$

Note that the error terms ε_t (structural shocks) satisfy the conditions (i) - (iii) in the definition above, with the particularity that all the elements off the main diagonal of the covariance matrix are zero. That is, the structural shocks are uncorrelated. The structure of the system incorporates feedback because y_t and z_t are allowed to affect each other contemporaneously. For example, $-b_{12}$ represents the contemporaneous effect of a unit change of z_t on x_t, while γ_{12} is the effect of a unit change in z_{t-1} on x_t.

3.1.3.3 Estimation and Identification:

Consider the structural bivariate system presented above:

$$\begin{aligned} x_t &= \gamma_{10} - b_{12} z_t + \gamma_{11} x_{t-1} + \gamma_{12} z_{t-1} + \varepsilon_{xt} \\ z_t &= \gamma_{20} - b_{21} x_t + \gamma_{21} x_{t-1} + \gamma_{22} z_{t-1} + \varepsilon_{zt} \\ &: \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{pmatrix} \sim \text{iid}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{pmatrix} \right) \end{aligned}$$

Due to the feedback dynamics inherent in this SVAR model, estimating the system directly via OLS would yield inconsistent results. This is because OLS requires regressors to be uncorrelated with the error term and in the system above z_t is correlated with the error term ε_{xt} and x_t is correlated with the error term ε_{zt}. There is a pervasive simultaneity bias.

Note, however, that there is no such problem in estimating the VAR model in (4). OLS provides consistent estimates of the 2 elements of A_0 and the four elements of A_1. Further, obtaining the residuals from the 2 regressions, it is possible to calculate estimates of the variance of e_{1t} and e_{2t} as well as the covariance between them.

The main issue is whether we can recover all the information present in the original system, given by (5) and (6). In other words, is the primitive system identifiable given the OLS estimates of the VAR model in the form (4)? The short answer to this question is no. The reason is evident if we compare the number of parameters of the primitive system, with those recovered from the estimated reduced form model:

Estimating (4) yields nine parameter values:

(a) Six coefficient estimates: $a_{10}, a_{20}, a_{11}, a_{12}, a_{21}, a_{22}$

(b) Calculated values of: var(e_{1t}), var(e_{2t}), and cov(e_{1t}, e_{2t})

However, the SVAR model (5) and (6) contains a total of ten parameters:

(a) 2 intercepts: γ_{10}, γ_{20}

(b) 4 autoregressive: $\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22},$

(c) 2 feedback coefficients: b_{12}, b_{21}

(d) 2 standard deviations: σ_x, σ_z

In all, the primitive model contains 10 parameters whereas the reduced form VAR estimation yields only 9 parameters. Structural VAR methodology proceeds by making identifying assumptions that allow this system of equations to be estimated consistently.

Identifying restrictions I: Recursive or zero short-run restrictions

The most common set of identification restrictions used in SVAR models are recursive ones, which require certain variables not to respond within period to structural shocks. Frequently these restrictions are also called zero, or short-run restrictions. Suppose, for example, you are willing to impose a restriction on the primitive system (following Sims (1980)) such that the coefficient b_{21} equals zero. This implies that z_t has a contemporaneous effect on x_t but x_t does not contemporaneously affect z_t. Hence x_t only affects the z_t with a one period lag.

Imposing this restriction and rewriting (5) and (6) yields:

$$\begin{aligned} x_t &= \gamma_{10} - b_{12}z_t + \gamma_{11}x_{t-1} + \gamma_{12}z_{t-1} + \varepsilon_{xt} && (10) \\ z_t &= \gamma_{20} + \gamma_{21}x_{t-1} + \gamma_{22}z_{t-1} + \varepsilon_{zt} && (11) \end{aligned}$$

This restriction (usually suggested by theory or a particular economic model) results in an exactly identified system. Imposing $b_{21} = 0$ means that B^{-1} is given by the following lower triangular matrix[8]:

[8] Recall that:

$$\begin{aligned} B^{-1} &= \frac{1}{\Delta}\begin{pmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{pmatrix} \\ \Delta &= det(B) = 1 - b_{12}b_{21} \end{aligned}$$

$$B^{-1} = \begin{pmatrix} 1 & -b_{12} \\ 0 & 1 \end{pmatrix}$$

Pre-multiplication of the primitive system by B^{-1} yields:

$$\begin{aligned} \begin{pmatrix} x_t \\ z_t \end{pmatrix} &= \begin{pmatrix} 1 & -b_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \end{pmatrix} + \begin{pmatrix} 1 & -b_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & -b_{12} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{pmatrix} \qquad (12) \\ &= \begin{pmatrix} \gamma_{10} & -b_{12}\gamma_{20} \\ 0 & \gamma_{20} \end{pmatrix} + \begin{pmatrix} \gamma_{11} - b_{12}\gamma_{21} & \gamma_{12} - b_{12}\gamma_{22} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{xt} - b_{12}\varepsilon_{zt} \\ \varepsilon_{zt} \end{pmatrix} \end{aligned}$$

Which can be rewritten as:

$$\begin{aligned} x_t &= a_{10} + a_{11}x_{t-1} + a_{12}z_{t-1} + e_{1t} && (13) \\ z_t &= a_{20} + a_{21}x_{t-1} + a_{22}z_{t-1} + e_{2t} && (14) \\ : & \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim \text{iid} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Var(e_{1t}) & Cov(e_{1t}, e_{2t}) \\ Cov(e_{2t}, e_{1t}) & Var(e_{2t}) \end{pmatrix} \right) && (15) \end{aligned}$$

where:

$$\begin{aligned} a_{10} &= \gamma_{10} - b_{12}b_{20} \\ a_{20} &= \gamma_{20} \\ a_{11} &= \gamma_{11} - b_{12}\gamma_{21} \\ a_{12} &= \gamma_{12} - b_{12}\gamma_{22} \\ a_{21} &= \gamma_{21} \\ a_{22} &= \gamma_{22} \\ e_{1t} &= \varepsilon_{xt} - b_{12}\varepsilon_{zt} \\ e_{2t} &= \varepsilon_{zt} \\ Var(e_{1t}) &= \sigma_x^2 + b_{12}^2\sigma_z^2 \\ Var(e_{2t}) &= \sigma_z^2 \\ Cov(e_{1t}, e_{2t}) &= -b_{12}\sigma_z^2 \end{aligned}$$

Estimating the system using OLS yields 9 parameter estimates that can be substituted into the 9 equations above in order to simultaneously solve for: $\gamma_{10}, \gamma_{20}, b_{12}, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \sigma_x^2, \sigma_z^2$. Note that the estimates of the ε_{xt} and ε_{zt} can also be recovered.

Decomposing the residuals in this triangular fashion is equivalent to performing a Choleski decomposition as we will see in the next section. Further details in the following section. In a n-variable VAR, B is a n x n matrix. Exact identification requires that $\frac{(n^2-n)}{2}$ restrictions be placed on the relationship between the regression residuals and the structural innovation. Since the Choleski decomposition is triangular, it forces exactly $\frac{(n^2-n)}{2}$ values of the B matrix equals to zero.

Identifying restrictions II: Blanchard and Quah (1989) or long-run restrictions

Complete

Identifying restrictions III: Linear restrictions

Complete

Identifying restrictions IV: Sign restrictions

Complete

3.1.4 Reduced-form VAR models

A reduced-form VAR expresses each variable as a linear function of its own past values, the past values of all the other variables being considered and a serially uncorrelated error term. Continuing with the example above, premultiplication by B^{-1} allows us to obtain the VAR in reduced form:

$$\begin{aligned} B^{-1}By_t &= B^{-1}\Gamma_0 + B^{-1}\Gamma_1 y_{t-1} + B^{-1}\varepsilon_t && (16) \\ \Rightarrow y_t &= A_0 + A_1 y_{t-1} + e_t && (17) \\ &: A_0 = B^{-1}\Gamma_0 \\ &: A_1 = B^{-1}\Gamma_1 \\ &: e_t = B^{-1}\varepsilon_t \end{aligned}$$

Define:
a_{i0} as an element of vector A_0
a_{ij} as an element in row i and column j of the matrix A_1
e_{it} as an element of vector e_t

Where (9) is called a structural VAR or primitive system, and (17) is referred to as a reduced-form VAR. Note that the error terms e_{1t} and e_{2t} in the reduced form VAR, are composites of the 2 structural shocks ε_{xt} and ε_{zt}:

$$\begin{aligned} e_{1t} &= \frac{(\varepsilon_{yt} - b_{12}\varepsilon_{zt})}{1 - b_{12}b_{21}} \\ e_{2t} &= \frac{(\varepsilon_{zt} - b_{21}\varepsilon_{yt})}{1 - b_{12}b_{21}} \end{aligned}$$

3.1.4.1 Properties of reduced-form error terms

Since $\{\varepsilon_{yt}\}$ and $\{\varepsilon_{zt}\}$ are white noise processes, it follows that both e_{1t} and e_{2t} have zero mean, constant variances and are individually serially uncorrelated:

(1) Expected Value:

$$E(e_{1t}) = E\left[\frac{(\varepsilon_{xt} - b_{12}\varepsilon_{zt})}{1 - b_{12}b_{21}}\right] = 0$$

(2) Variance:

$$E(e_{1t}^2) = E\left[\frac{(\varepsilon_{xt} - b_{12}\varepsilon_{zt})}{1 - b_{12}b_{21}}\right]^2 = \frac{\sigma_y^2 + b_{12}\sigma_z^2}{(1 - b_{12}b_{21})^2}$$

(3) Autocorrelation:

$$E(e_{1t}e_{1t-i}) = E\left[\frac{(\varepsilon_{xt} - b_{12}\varepsilon_{zt})(\varepsilon_{xt-i} - b_{12}\varepsilon_{zt-i})}{(1 - b_{12}b_{21})^2}\right] = 0 \quad \forall i \neq 0$$

Similarly, we can demonstrate that e_{2t} is a stationary process with zero mean, constant variance and all autocovariances equal to zero satisfying all the VAR requirements. Note however, that e_{1t} and

e_{2t} will be correlated. The covariance of the two terms comes down to:

$$\begin{aligned} E(e_{1t}e_{2t}) &= E\left[\frac{(\varepsilon_{xt} - b_{12}\varepsilon_{zt})(\varepsilon_{zt} - b_{12}\varepsilon_{yt})}{(1 - b_{12}b_{21})^2}\right] \\ &= \frac{-(b_{21}\sigma_x^2 + b_{12}\sigma_z^2)}{(1 - b_{12}b_{21})^2} \end{aligned} \tag{18}$$

Only in the special case where $b_{12} = b_{21} = 0$ (i.e.: there are no contemporaneous effects of y_t on z_t and z_t on y_t) the shocks will be uncorrelated.

It is useful to define the matrix of variance-covariance of the e_{1t} and e_{2t} shocks:

$$\begin{aligned} \Sigma_e &= \begin{pmatrix} \text{var}(e_{1t}) & \text{cov}(e_{1t}, e_{2t}) \\ \text{cov}(e_{2t}, e_{1t}) & \text{var}(e_{2t}) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \\ &= B^{-1}E[\varepsilon_t\varepsilon_t']B^{-1'} \\ &= B^{-1}\Sigma_\varepsilon B^{-1'} \\ &= \frac{1}{\Delta^2}\begin{pmatrix} \sigma_x^2 + b_{12}^2\sigma_z^2 & -(b_{21}\sigma_x^2 + b_{12}\sigma_z^2) \\ -(b_{21}\sigma_x^2 + b_{12}\sigma_z^2) & \sigma_z^2 + b_{21}^2\sigma_x^2 \end{pmatrix} \end{aligned}$$

To sum:
A SVAR is different from a reduced-form VAR in at least two ways:

1. Variables do not affect each other contemporaneously

2. Shocks are not correlated with each other

3.1.4.2 Estimation and Identification:

Assume the researcher decides to begin from a reduced-form VAR as in (17):

$$\begin{aligned} y_t &= A_0 + A_1 y_{t-1} + e_t \\ &: A_0 = B^{-1}\Gamma_0 \\ &: A_1 = B^{-1}\Gamma_1 \\ &: e_t = B^{-1}\varepsilon_t \\ &: \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim (0, \Sigma_e) \end{aligned}$$

where the underlying SVAR (9) is:

$$\begin{aligned} By_t &= \Gamma_0 + \Gamma_1 y_{t-1} + \varepsilon_t \\ &: \begin{pmatrix} \varepsilon_x \\ \varepsilon_z \end{pmatrix} \sim i.i.d. \, (0, \Sigma_\varepsilon) \end{aligned}$$

Since there is no simultaneity problem here, this reduced-form VAR (17) can be estimated directly via OLS. This will provide the researcher with an estimate of all six parameter values ($a_{10}, a_{20}, a_{11}, a_{12}, a_{21}, a_{22}$) as well as three calculated values for the variance-covariance matrix (var(e_{1t}),var(e_{2t}), cov(e_{1t}, e_{2t})). The structural primitive system, however, is not yet identified as there are more unknowns than equations. Consider the variance-covariance matrix estimated in reduced-form:

$$\begin{aligned} \Sigma_e &= B^{-1}\Sigma_\varepsilon B^{-1'} \\ &= B^{-1} \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{pmatrix} B^{-1'} \\ &= \underbrace{B^{-1} \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_z \end{pmatrix}}_{P^{-1}} \underbrace{\begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_z \end{pmatrix} B^{-1'}}_{P^{-1'}} \\ &= P^{-1} P^{-1'} \end{aligned}$$

which implies that:

$$\underbrace{\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}}_{\text{3 values}} = \underbrace{\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}^{-1}}_{\text{4 unknowns}} \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}^{-1'}$$

and hence the primitive system is not identified.

Variable ordering:
From the reduced form estimation A_0, A_1, Σ_e are identified. If B^{-1} would be known, then the structural parameters $\Gamma_0, \Gamma_1, \Sigma_\varepsilon$ could also be recovered, and so the identification of the structural model comes down to finding B^{-1}. In order to do so, one would need to impose some sort of identification restriction similar to what was done in the previous section.

Having started from a reduced-form framework, a popular identification strategy consists of imposing a particular structure on the matrix B (or equivalently B^{-1}) above. Notably, we know from the

Choleski decomposition of a Hermitian, positive-definite matrix "M", that this matrix could be written as the product of a lower triangular matrix "C" and its conjugate transpose[9]:

$$M = CC'$$

In turn, assuming a particular structure on B will allow use to use the above decomposition to identify the structural system. Specifically, if we assume that B is a lower triangular matrix $\widetilde{B}$, then we could re-write our previous reduced-form model (17) as:

$$\begin{aligned}
y_t &= \widetilde{A}_0 + \widetilde{A}_1 y_{t-1} + e_t \\
&: \widetilde{A}_0 = \widetilde{B}^{-1}\Gamma_0 \\
&: \widetilde{A}_1 = \widetilde{B}^{-1}\Gamma_1 \\
&: e_t = \widetilde{B}^{-1}\varepsilon_t \\
&: \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim \left(0, \widetilde{\Sigma}_e\right)
\end{aligned} \tag{19}$$

or less compactly:

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} \widetilde{a}_{10} \\ \widetilde{a}_{20} \end{pmatrix} + \begin{pmatrix} \widetilde{a}_{11} & \widetilde{a}_{21} \\ \widetilde{a}_{21} & \widetilde{a}_{22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \widetilde{b}_{11} & 0 \\ \widetilde{b}_{21} & \widetilde{b}_{22} \end{pmatrix} \begin{pmatrix} \varepsilon_x \\ \varepsilon_z \end{pmatrix}$$

where:

$$\begin{aligned}
\widetilde{\Sigma}_e &= \widetilde{B}^{-1}\Sigma_\varepsilon \widetilde{B}^{-1'} \\
&= \widetilde{B}^{-1} \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_z^2 \end{pmatrix} \widetilde{B}^{-1'} \\
&= \underbrace{\widetilde{B}^{-1} \begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_z \end{pmatrix}}_{\widetilde{P}^{-1}} \underbrace{\begin{pmatrix} \sigma_x & 0 \\ 0 & \sigma_z \end{pmatrix} \widetilde{B}^{-1'}}_{P^{-1'}} \\
&= \widetilde{P}^{-1}\widetilde{P}^{-1'}
\end{aligned}$$

Following Choleski, we then have that:

$$\underbrace{\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}}_{\text{3 values}} = \underbrace{\begin{pmatrix} \widetilde{p}_{11} & 0 \\ \widetilde{p}_{21} & \widetilde{p}_{22} \end{pmatrix}^{-1}}_{\text{3 unknowns}} \begin{pmatrix} \widetilde{p}_{11} & 0 \\ \widetilde{p}_{21} & \widetilde{p}_{22} \end{pmatrix}^{-1'}$$

and so the system is identified. In other words, a Choleski factorization allows for identification as it generates a lower triangular decomposition of the variance-covariance matrix, which happens to be exactly the mapping between innovations and structural shocks previously described. With the primitive system identified, we can give the estimation results a more palpable economic interpretation.

Intuition

Economically, what does imposing a lower triangular structure on B mean? It means that the first variable doesn't respond within period to the last n-1 structural shocks, the second variable doesn't respond withing the period to the last n-2 structural shocks and so on. In other words, variables

[9] For a refresher on matrix factorization methods please refer to the appendix.

ordered first are not allowed to react to other shocks within period; variables ordered last are allowed to respond to all shocks within the period. This is often referred to as "casusal ordering" or simply "ordering" of variables. In the case of our identified var (19) we have:

$$\begin{aligned} x_t &= ... + \widetilde{b}_{11}\varepsilon_x \\ z_t &= ... + \widetilde{b}_{21}\varepsilon_x + \widetilde{b}_{22}\varepsilon_z \end{aligned}$$

where ε_x affects contemporaneously all variables, but ε_z only affects contemporaneously z_t. This exemplifies how the ordering of the variables matter: the variable placed on top is only affected by a shock to itself, whereas each variable afterwards is affected by all variables with a delay.

In some sense, variables are ordered by degree of relative exogeneity. Doing so implies some variables in the VAR are more exogenous than others and hence are more likely to influence the rest of the variables but not be influenced by them. The variable placed on top is considered to be the most exogenous. Note that "exogenous" is being used loosely here, as all variables are endogenous in the dynamic sense of a VAR.

This decomposition forces a potentially important asymmetry on the model, since it will imply some variables have differential effects. In essence the variable ordering will change the order of the coefficients in the B^{-1} matrix thereby mimicking the identification restriction used in the previous section. It should also be noted that the importance of ordering depends on the magnitude of the correlation between e_{1t} and e_{2t}. If the correlation would be zero, the ordering would be immaterial. At the other extreme, if the correlation is perfect, there is a single shock that affects both variables.

3.1.5 Impulse responses:

Just as an autoregression has a moving average representation, a vector autoregression can be written as a vector moving average (VMA). The VMA representation allows you to trace out the time path of the various shocks on the variables contained in the VAR system. For illustration purposes, consider the following two-variable reduced-form VAR as (17):

$$\begin{aligned} x_t &= a_{10} + a_{11}x_{t-1} + a_{12}z_{t-1} + e_{1t} \\ z_t &= a_{20} + a_{21}x_{t-1} + a_{22}z_{t-1} + e_{2t} \end{aligned}$$

Writing the system in matrix form:

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} a_{10} \\ a_{20} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix}$$

or

$$y_t = A_0 + A_1 y_{t-1} + e_t \tag{20}$$

$$: \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim \text{iid} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right) \tag{21}$$

As previously mentioned, such a model can be estimated via OLS, providing estimates for all parameter values. In order to find the VMA representation we proceed as follows. Using .... XXX we can obtain:

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \bar{z} \end{pmatrix} + \sum_{i=0}^{\infty} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^i \begin{pmatrix} e_{1t-i} \\ e_{2t-i} \end{pmatrix}$$

The equation above expresses x_t and z_t in terms of $\{e_{1t}\}$ and $\{e_{2t}\}$ sequences. However, it is more insightful to write the above in terms of the sequence of underlying structural shocks $\{\varepsilon_{xt}\}$ and $\{\varepsilon_{zt}\}$. The vector of errors can be written as:

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} = \frac{1}{1 - b_{12}b_{21}} \begin{pmatrix} 1 & -b_{12} \\ -b_{12} & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{pmatrix}$$

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \bar{z} \end{pmatrix} + \frac{1}{1 - b_{12}b_{21}} \sum_{i=0}^{\infty} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^i \begin{pmatrix} 1 & -b_{12} \\ -b_{12} & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{yt-i} \\ \varepsilon_{zt-i} \end{pmatrix}$$

To simplify the notation, let:

$$\phi_i = \frac{A_1^i}{1 - b_{12}b_{21}} \begin{pmatrix} 1 & -b_{12} \\ -b_{12} & 1 \end{pmatrix}$$

Hence, the moving average representation can be written in terms of $\{\varepsilon_{yt}\}$ and $\{\varepsilon_{zt}\}$ sequences:

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \bar{z} \end{pmatrix} + \sum_{i=0}^{\infty} \begin{pmatrix} \phi_{11}(i) & \phi_{12}(i) \\ \phi_{21}(i) & \phi_{22}(i) \end{pmatrix} \begin{pmatrix} \varepsilon_{yt-i} \\ \varepsilon_{zt-i} \end{pmatrix}$$

or more succinctly:

$$x_t = \mu + \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i}$$

The coefficients of ϕ_i can be used to generate the effects of $\{\varepsilon_{yt}\}$ and $\{\varepsilon_{zt}\}$ shocks on the entire time paths of the $\{y_t\}$ and $\{z_t\}$ sequences.

It should be clear that the four elements of $\phi_{jk}(0)$ are the impact multipliers. For example, the coefficient $\phi_{12}(0)$ is the instantaneous impact of a one-unit change in ε_{zt} on y_t. In the same way, the elements $\phi_{11}(1)$ and $\phi_{12}(1)$ are the one-period responses of unit changes in ε_{yt-1} and ε_{zt-1} on y_t respectively. Updating by one period would imply that $\phi_{11}(1)$ and $\phi_{12}(1)$ also represent the unit changes in ε_{yt} and ε_{zt} on y_{t+1}.

The accumulated effects of unit impulses on ε_{yt} can be obtained by the appropriate summation of the coefficients of the impulse response functions. For example, after n periods, the cumulated sum f the effects of ε_{zt} on the $\{y_t\}$ sequence is:

$$\sum_{i=0}^{n} \phi_{12}(i)$$

As n approaches infinity we obtain the long-run multiplier.

$$\sum_{i=0}^{\infty} \phi_{jk}(i)$$

The four set of coefficients $\phi_{11}(i), \phi_{12}(i), \phi_{21}(i), \phi_{22}(i)$ are referred to as the impulse response functions. Plotting the impulse response functions (i.e.: plotting the coefficients of XX against i) is a useful way to visually represent the dynamic behavior of the $\{x_t\}$ and $\{z_t\}$ series in response to the structural shocks.

3.1.6 Confidence Interval

3.1.7 Variance decomposition

3.1.8 VAR order

3.1.9 Stability and Stationarity

In the univariate autoregressive model: XXX the stability condition was..

3.2 Special Topics

3.2.1 State Space models

In many applications, the driving forces behind the evolution of variables are either not observable (at least partially), or precisely measurable. For example, on an individual level, a person's income may depend on his or hers intelligence, social skills, family connections, and so on. Naturally, these factors are all difficult to observe directly and accurately measure. Similarly, on a more aggregate level, economic theory suggests that macroeconomic variables such as output growth are driven by unobservable factors such as technological progress or human capital accumulation.

When explanatory variables are not directly observable or precisely measured, standard VAR models can no longer be applied to study the evolution of the endogenous variables. However, it is possible to extend the VAR framework to analyze scenarios with unobservable explanatory variables using state space models.

3.2.1.1 Definition

State Space models allow the researcher to model an observed time series, $\{y_t\}_{t=1}^{T}$, as being explained by a vector of (possibly unobserved) variables (usually state variables), which are driven by a stochastic process. The nature of this stochastic process will condition the nature of the state space model. For example, a basic linear state space model takes the following form:

$$\begin{aligned} y_t &= Hz_t + v_t : \qquad v_t \sim N(0, \Sigma_v) && (22) \\ z_t &= Bz_{t-1} + w_t : \qquad w_t \sim N(0, \Sigma_w) && (23) \end{aligned}$$

Given the assumption on the error terms, the above is usually referred to as a linear Gaussian state space model or dynamic linear model. In this model we do not observe the state vector z_t directly, but rather a linear transformation of it with added noise. We assume, for simplicity, that z_0 is known, and that $\{v_t\}$ and $\{w_t\}$ are uncorrelated. Furthermore, we also assume that the process starts with a normal vector z_0 such that $z_0 \sim N(\mu_0, \Sigma_0$ and that v_t ...

The <u>first</u> equation, called the observation or measurement equation, describes the relationship between the observed time series y_t and the (possibly unobserved) state z_t. In general it is assumed that the data y_t are measured with uncertainty/error, which is reflected in the stochastic term that enters (22). The standard approach is to model v_t as a Gaussian error term, $v_t \sim N(0, \Sigma_v)$.

The <u>second</u> equation, the state or transition equation, describes the evolution of the state variables as being driven by the stochastic process of innovations w_t. Typically we assume this innovations are normally distributed such that $w \sim N(0, \Sigma_w)$.

In general, the state space model is characterized by two principles. First, there is a hidden or latent process z_t called the state process. The state is assumed to be a Markov process; this means that the future $\{z_s : s > t\}$, and past $\{z_s : s < t\}$ are independent conditional on the present z_t. The second condition is that the observations y_t are independent given the states z_t. This means that the dependence among the observations is generated by the states.

Note that the State space model above could be formulated in a much more general fashion. For example, matrices H and B could depend explicitly on time, or one could introduce constants in the specification[10]. In fact, the popularity of State space models derives partly from the fact that they offer a unified approach to a wide range of models and techniques such as dynamic regressions, ARIMA, unobserved component models, latent variables models, etc. Nonetheless, mostly for notational simplicity, we will discuss state space models using the basic specification described above.

3.2.1.2 Examples

The linearized solution to a Real Business Cycle model could written in state space form as follows:

$$qqq$$

3.2.1.3 Estimation:

In most situations, the system's matrices H and B together with the variances Σ_v and Σ_w are unknown and must be estimated. Obviously, whenever the explanatory variables are not observable, OLS estimation is not the way to go. However, even in this case, one can apply likelihood based inference, since the so-called Kalman filter allows to construct the likelihood function associated with a state space model.

Filtering:
Assume that we observe data $\{y_t\}_{t=1}^T$, that are to be described by the model XX and XX. Assume that reasonable (yet not necessarily true) values for the model's parameters are available, and equal to $H^*, B^*, \Sigma_v^*, \Sigma_w^*,$. Let δ summarize these values $\delta = \{H^*, B^*, \Sigma_v^*, \Sigma_w^*, \}$. Let the sample density (or likelihood) function associated with a state space model for given parameters δ be denoted by $f(y_1, y_2, ..., y_T; \delta)$. By Bayes theorem, we acn factor the likelihood as:

$$\begin{aligned} f(y_1, y_2, ..., y_T; \delta) &= f(y_1, \delta) f(y_2|y_1, \delta) f(y_3|y_2, y_1, \delta) .... f(y_T|y_{T-1}, ..., y_1, \delta) && (24) \\ &= f(y_t|y^{t-1}, \delta) && (25) \end{aligned}$$

where $y^0 = 0$, and $y^{t-1} = (y_1, y_2, ..., y_{t-1})$ for $t \geq 2$. The log-likelihood function is thus given by:

$$\ln L(y^T, \delta) = \sum_{t=1}^{T} \ln f(y_t|y^{t-1}, \delta) \tag{26}$$

Naturally, to construct the likelihood function, we need to derive the densities:

$$f(y_t|y^{t-1}) : t = 1, 2, ..., T \tag{27}$$

We can achieve this by using filtering techniques. In particular when the system is linear and errors are Gaussian we can use the Kalman filter. The Kalman filter is a recursive procedure that involves the following steps:

1. Initialization

2. Prediction

[10] In some engineering specifications in fact, equation (1) has no error term reflecting uncertainty, hence making it deterministic; and equation's (2) innovation is referred to as the system's impulse or driver.

3. Correction
4. Likelihood construction

Initialization:

Let $x_{t|s}$ denote the prediction of variable x at time t, conditional upon the information available at time s. The Kalman is initialized by deriving the best predictor of the initial state, $z_{0|0}$, and an estimate of its covariance matrix: $\Sigma^z_{0|0} = E[(z_0 - z_{0|0})(z_0 - z_{0|0})']$. If the process is stationary, this is straightforward as one can build on the steady state of the system. More precisely one can set up $z_0|0 = z^*$ and $\Sigma^z_{0|0} = \Sigma^*$ such that:

1. $z^* = Bz^*$
2. $\Sigma^* = B\Sigma^* B' + \Sigma_w = [I - B \otimes B]^{-1} vec(\Sigma_w)$

Prediction:
Setting $t = 1$ yields: $z_{0|0} = z_{t-1|t-1}$ and $\Sigma^z_{0|0} = \Sigma^z_{t-1|t-1} =$. Together with the transition equation to compute:

$$z_{t|t-1} = Bz_{t-1|t-1} \tag{28}$$

$$\Sigma^z_{t|t-1} = B\Sigma^z_{t-1|t-1}B' + \Sigma_w \tag{29}$$

We can the use:

$$\begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} y_t \\ z_t \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{pmatrix} \tag{30}$$

3.2.2 VARX models

To be completed.

3.2.2.1 Definition

3.2.2.2 Examples

3.2.2.3 Estimation

3.2.3 Forecasting

To be completed.

3.2.3.1 Definition

3.2.3.2 Examples

4 Nonstationary Univariate models

Most macroeconomic time series are not stationary in their raw form. A common approach to analyze a time series is to decompose its evolution in 4 main components:

$$\begin{aligned} Y_t &= T_t + C_t + S_t + e_t \\ &: T_t \text{ is trend (long-run behavior)} \\ &: C_t \text{ is cycle (short-run behavior)} \\ &: S_t \text{ is seasonality (recurrent behavior associated with certain frequencies)} \\ &: e_t \text{ is noise (what we cannot explain)} \end{aligned}$$

Nonstationary time series present some particular difficulties as standard inference techniques often fails when a process depends explicitly on t. In turn, stationarity can be violated by the following:

- Seasonality
- Deterministic Trend (time trend)
- Stochastic Trend (unit roots)
- Structural breaks

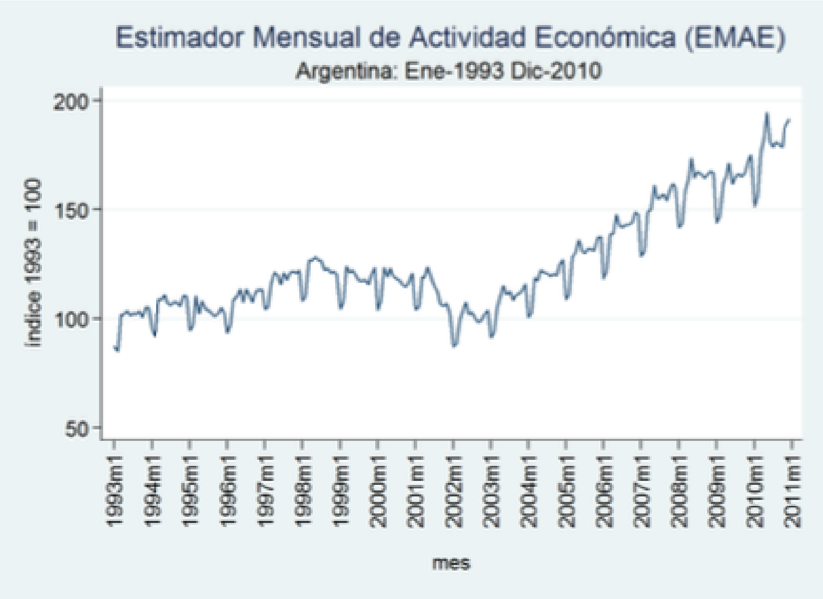
Each type has unique features. Seasonality is technically a form of deterministic trend, although their analysis is sufficiently similar to stationary time series that little is lost in treating a seasonal time series as if it were stationarity. Process with deterministic trends have unconditional means which depend on time and unit roots processes have unconditional variances that grow over time. Structural breaks are an encompassing class which may result in either or both mean and variance exhibiting time dependence.

4.1 Seasonality

Seasonality refers to the recurrent behavior associated to specific frequencies (monthly, quarterly, annual, etc). A few examples:

1. Increase in airfare prices due to peak travel season
2. Decrease in ice cream consumption during the winter months
3. Increase in payrolls due to bonus season
4. Increase in personal spending during holiday season

A time series with seasonality would look something like this:



A simple way to capture this is to work with dummy variables. Take for example the following model:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 D_{1,t} + \beta_2 D_{2,t} + \beta_3 D_{3,t} + u_t \\ &: D_{j,t} = 1 \quad \forall j = 1,2,3\ldots \quad \text{if the observation corresponds to quarter j} \\ &: D_{j,t} = 0 \quad \forall j = 1,2,3\ldots \quad \text{if the observation does not corresponds to quarter j} \end{aligned}$$

This can be estimated by OLS, and the residuals will correspond to the series adjusted for seasonality.

Another alternative is simply to work with the inter-annual change value of the variable for each period. For the quarterly case this would imply $y_t - y_{t-4}$.

4.2 Trends

A trending mean is a common violation of stationarity. There are two popular models for modeling trends in nonstationary series.

4.2.1 Deterministic Trend

A first alternative is to characterize the trend as a deterministic (i.e.: not stochastic) function. In other words, we simply assume that the trend is a function of time. In this case we say that the process has a **deterministic trend**, or that the process is **trend stationary**.

A trend stationary process is composed of a deterministic trend plus a stochastic component which is stationary. Common specifications are[11].

$$\begin{aligned}
\text{Linear Trend:} \quad y_t &= \beta_1 t + u_t \\
\text{Quadratic Trend:} \quad y_t &= \beta_1 t + \beta_2 t^2 + u_t \\
\text{Log-linear Trend:} \quad \log(y_t) &= \beta_1 t + u_t \\
\text{Log-linear Quadratic Trend:} \quad \log(y_t) &= \beta_1 t + \beta_2 t^2 + u_t \\
&: \quad u_t = \rho u_{t-1} + \varepsilon_t \quad 0 < \rho < 1
\end{aligned}$$

where $\beta_1 t$ is the trend specification, with β_1 equal to the trend growth rate, u_t represents deviations from the trend (i.e.: the cycle) and ε_t is a random business cycle disturbance. These shocks all have effects on y_t which eventually die out.

Estimation: xxxx

4.2.2 Stochastic Trend

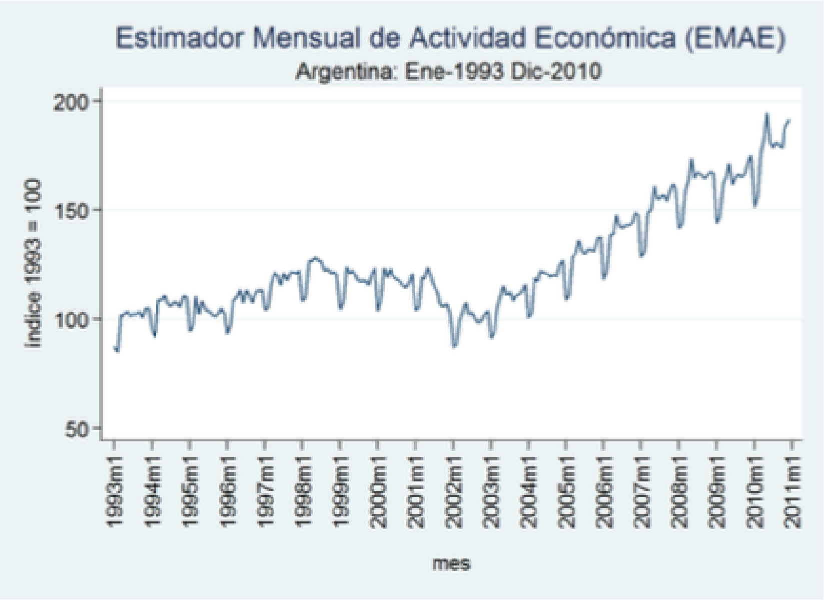
A second option is very different but turns out to be nearly observationally equivalent (especially in small samples) is the random walk with drift. Take the same process as above but assume $\rho = 1$:

$$\begin{aligned}
y_t &= \beta_1 t + u_t \\
&: \quad u_t = u_{t-1} + \varepsilon_t \\
\text{replacing } u_t \text{ yields} & \\
y_t &= \beta_1 t + u_{t-1} + \varepsilon_t \\
\text{lagging } y_t \text{ 1 period yields} & \\
y_{t-1} &= \beta_1 (t-1) + u_{t-1} \\
\Rightarrow u_{t-1} &= y_{t-1} - \beta_1 (t-1) \\
\text{replacing in original} & \\
y_t &= \beta_1 t + y_{t-1} - \beta_1 (t-1) + \varepsilon_t \\
&= \beta_1 + y_{t-1} + \varepsilon_t
\end{aligned}$$

[11] Note that you can always add a constant to each of these specifications. In that case, for example, the linear trend would be specified as: $y_t = \beta_0 + \beta_1 t + u_t$

Formally, a process is said to have a unit root if at least one of the roots of its characteristic equation is 1. For an AR(1) specification as the one above, this implies that the coefficient of autoregression is equal to 1. Unit roots are problematic for econometrics because the usual assumptions about error terms break down. If you have a deterministic trend you could do the OLS on the cycle component, but if the series has a stochastic trend, removing the deterministic trend doesn't remove the unit root.

In finite samples it is usually very difficult to distinguish when a series has a unit root, this is called the "near observation equivalence" problem. For example, it is hard to observationally determine whether figure xxx contains a time trend or a random walk with drift?



Further, the distinction between a deterministic and stochastic trend has important implications for the long-term behavior of a process:

- Time series with a deterministic trend always revert to the trend in the long run (the effects of shocks are eventually eliminated). Forecast intervals have constant width.
- Time series with a stochastic trend never recover from shocks to the system (the effects of shocks are permanent). Forecast intervals grow over time.
- Unfortunately, for any finite amount of data there is a deterministic and stochastic trend that fits the data equally well (Hamilton, 1994).
- Unit root tests are frequently used for assessing the presence of a stochastic trend in an observed series.

4.2.3 Testing for unit roots

For the reasons above mentioned, it is important to pre-test variables, detect the right type of trend present (if any), and establish the appropriate procedure for producing a stationary series. In short, there are (at least) four reasons to test for the presence of units roots:

1. To test whether a series is stationary

2. To determine the duration of shocks affecting a variable
3. To test the type of trend present in the data
4. To avoid spurious regressions

Much of what follows is based on Dickey and Fuller (1979). We begin from the model with a linear time trend:

$$\begin{aligned} y_t &= at + u_t \\ &: \quad u_t = \rho u_{t-1} + \varepsilon_t \\ &\Rightarrow \quad y_t = at + \rho u_{t-1} + \varepsilon_t \end{aligned}$$

Lag original one period yields:

$$\begin{aligned} y_{t-1} &= a(t-1) + u_{t-1} \\ \text{replace into above:} & \\ y_t &= at + \rho(y_{t-1} - a(t-1)) + \varepsilon_t \\ &= at + \rho y_{t-1} - \rho at + \rho a + \varepsilon_t \\ \text{take 1st difference:} & \\ y_t - y_{t-1} &= (1-\rho)at + \rho a + (\rho - 1)y_{t-1} + \varepsilon_t \end{aligned}$$

$$\text{Let:} \quad \Delta y_t = y_t - y_{t-1}, \gamma = \rho - 1, \beta = (1-\rho)a, \alpha = \rho a.$$

The model to estimate becomes:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \varepsilon_t$$

- The null that a series is stationary around a deterministic trend corresponds to $|\gamma| < 0$ and $\beta \neq 0$.

$$\begin{aligned} \text{If} \quad |\gamma| < 0 \quad &\Rightarrow \quad |\rho| < 1 \\ \text{If} \quad \beta \neq 0 \quad &\Rightarrow \quad (1-\rho)a \neq= 0 \Rightarrow a \neq 0 \end{aligned}$$

- The null hypothesis that the series follows a RW (no drift) corresponds to $\gamma = \alpha = \beta = 0$.

$$\begin{aligned} \text{If} \quad \gamma = 0 \quad &\Rightarrow \quad \rho = 1 \\ \text{If} \quad \alpha = 0 \quad &\Rightarrow \quad \rho a = 0 \Rightarrow a = 0 \\ \text{If} \quad \beta = 0 \quad &\Rightarrow \quad (1-\rho)a = 0 \end{aligned}$$

- The null hypothesis that the series follows a RWD corresponds to $\gamma = 0$ and $\alpha \neq 0$.

$$\begin{aligned} \text{If} \quad \gamma = 0 \quad &\Rightarrow \quad \rho = 1 \\ \text{If} \quad \alpha \neq 0 \quad &\Rightarrow \quad \rho a \neq 0 \Rightarrow a \neq 0 \\ \text{The above implies} \quad \beta = 0 & \end{aligned}$$

You can estimate this regression by OLS under the null of a unit root ($\gamma = 0$), but inference is not standard. Since the test is done over the residual term rather than raw data, it is not possible to use standard t-distribution to provide critical values. Dickey and Fuller (1979) did some Monte Carlo experiments to numerically construct critical values for a t test (testing the hypothesis that $\gamma = 0$ and F tests (for testing joint hypothesis ($\gamma = \beta = 0$). Therefore this statistic t has a specific distribution simply known as the Dickey–Fuller table.

The intuition behind the test is as follows. If the series y is stationary (or trend stationary), then it has a tendency to return to a constant (or deterministically trending) mean. Therefore large values will tend to be followed by smaller values (negative changes), and small values by larger values (positive changes). Accordingly, the level of the series will be a significant predictor of next period's change, and will have a negative coefficient. If, on the other hand, the series is integrated, then positive changes and negative changes will occur with probabilities that do not depend on the current level of the series; in a random walk, where you are now does not affect which way you will go next.

To summarize, the procedure is as follows:

1. Pre-test variables using something like a Dickey-Fuller test to see if the series are (i) stationary, (ii) trend stationary, or (iii) are unit roots (i.e.: have stochastic trends)
2. If (i), estimate regressions and do inference as normal
3. If (ii), fit a deterministic trend to the series and then estimate regressions on the detrended data and do inference as usual
4. If (iii), first difference the series and then estimate regressions and do inference as usual.

4.3 Achieving Stationarity

Economic time series often violate our assumption of covariance stationarity. In particular, their mean is typically changing over time[12]. One way of dealing with this type of nonstationarity is by using stationary-inducing transformations of the data[13].

The method for producing a stationary series from processes with either type of trends is different. For the deterministic trend case you would estimate the trend and remove it from the series. For the stochastic case (RWD), removing a deterministic trend does not render the series stationary. In this case you want to first difference the series. First differencing the.

If you have a deterministic trend series, you would remove the trend and could do OLS on the cycle component. But if you think your series may have a stochastic trend, then removing the deterministic trend does not remove the unit root.

4.3.1 Removing deterministic trends:

A first alternative to induce stationarity for processes that grow over time is to remove a deterministic trend from their logged values. For example, removing a linear trend means taking the residuals from the following regression:

$$u_t = y_t - \hat{\mu} - \alpha t$$

In other words, u_t would be the resulting stationary series. One could also add quadratic, cubic, etc. terms to this regression to remove non-linear trends.

4.3.2 Removing stochastic trends

An alternative approach is to take first differences after taking logs of the data. In turn, we define a new variable Δy_t:

$$\Delta y_t = y_t - y_{t-1}$$

This transformation almost always induces stationarity for processes that have means (in log levels) that change over time in a systematic way.

Note that because the data has been logged, Δy_t measures the growth rate of the variable. This is because the log-difference transformation of a variable represents that variable in terms of its growth rates. Most growth rates of economic variables are stationary.

4.3.3 Filtering

A third alternative is to use some sort of filter to remove the nonstationarities. Examples of such filters are the Hodrick-Prescott (HP) filter or some band-pass filer. A thorough discussion of these methods

[12] For example, the average value of U.S. GDP in the 2000s is much higher than the average value that at the dawn of the 20th century.

[13] A preliminary transformation often used is to take logs of the time series. This is useful because most macro data typically grow in an exponential fashion and if a series grows with an exponential trend, taking logs will make the trend linear. Moreover, taking logs also deals with certain types of heteroskedasticity.

requires getting into the frequency domain (rather than time domain). Intuitively, however, we would take the stationary component as:

$$u_t = y_t - y_t^{HP}$$

where y_t^{HP} represents the trend component of the time series as identified by the HP filter.

4.4 Cointegration (multivariate)

I mentioned before that not removing trends could be problematic. Suppose I have two independent random walks (say x_t and z_t). Regressing xx on xx will tend to produce spurious results, in the sense that you will find coefficients that are "significantly" different from zero when doing conventional hypothesis testing. In other words, you are likely to find relationships in the data where truly there are none. Because of this I mentioned that one would want to get rid of the trend (either by removing a deterministic trend or by first differencing for example) before doing any econometrics.

There is, however, a very important caveat to this prescriptions. This occurs when if two variables are cointegrated, which means that each follow a unit root process, but a linear combination of them is stationary. In such a case differencing is not appropriate.

Take two unit root process (I omit the constant):

$$x_t = x_{t-1} + \varepsilon_t$$
$$z_t = z_{t-1} + v_t$$

Assume that $x_t = \gamma z_t$ is stationary. We say then that x_t and z_t are cointegrated with coitegrating vector $[1, \gamma]$. Suppose that we estimate the following regression (ignoring constants):

$$x_t = \beta z_t + \varepsilon_t$$

Doing OLS in a large enough sample will most likely pick $\hat{\beta} = \gamma$. In other words, OLS will produce a consistent estimate of the cointegrating vector even though both variables are non-stationary. This happens because OLS tries to minimize the sum of squared residuals. Hence OLS is trying to minimize $x_t - \hat{\beta} z_t$. If it chooses something other than $\hat{\beta} = \gamma$, then the residuals are non-stationary, which means they will get arbitrarily big or small. Hence, the spurious regression problem does not apply if variables are cointegrated with one another and therefore differencing them is not appropriate.

5 Appendix

5.1 Probability and Statistical Preliminaries

Let x and y be random variables characterized by a probability distribution and a be a constant term.

Definition 12 *1st central moment (Mean):*

$$E(x) = \mu$$

Definition 13 *2nd central moment (Covariance):*

$$E\left[(x - E(x)(y - E(y)\right] = E(xy) - E(x)E(y)$$

The variance is just a special case of the covariance when the two variables are identical. That is to say the Variance is the Covariance of the random variable. with itself:

$$\begin{aligned} Cov(x,x) &= E\left[(x - E(x)(x - E(x)\right] \\ &= E[(x - E(x))^2] \\ &= E(x^2) - [E(x)]^2 \\ &= Var(x) \end{aligned}$$

Properties:

$$\begin{aligned} Var(a) &= 0 \\ Var(x) &= \sigma^2 \\ Var(a + x) &= Var(x) + Var(a) = \sigma^2 \\ Var(ax) &= a^2 Var(x) = a^2\sigma^2 \\ Var(x + y) &= Var(x) + Var(y) + 2Cov(x,y) \end{aligned}$$

Additional properties:

$$\begin{aligned} Cov(x,a) &= 0 \\ Cov(x,y) &= Cov(y,x) \\ Cov(ax + by) &= abCov(x,y) \\ Cov(a + x, b + y) &= Cov(x,y) \end{aligned}$$

Definition 14 *Statistical Independence:*
Two random variables x and y are statistical independent iff:

$$\begin{aligned} E(xy) &= E(x)E(y) \\ &\Rightarrow Cov(x,y) = 0 \end{aligned}$$

Note: Statistical independence implies no correlation between x and y (i.e.: $Cov(x, y) = 0$). However $Cov(x, y) = 0$ does not imply statistical independence unless x and y are normally distributed.

Definition 15 *Autocovariance:*
The variance of a random variable at difference points in time.

$$\begin{aligned} \gamma_{jt} &= E[(x_t - E(x_t))(x_{t-j} - E(x_{t-j})] \\ &= E(x_t x_{t-j}) - E(x_t)E(x_{t-j}) \end{aligned}$$

If $\{x_t\}_{t=0}^{\infty}$ *has zero mean then:*

$$\gamma_{jt} = E(x_t x_{t-j})$$

Why do we care about autocovariances? Two reasons:

1. Forecasting: helps us understand a variable's evolution over time.
2. Modeling: helps us understand how persistent a process is. Models should reflect this persistence.

Just as it is useful to normalize covariances by dividing by the respective variables' standard deviations, it is also useful to normalize autocovariances.

Definition 16 *(Autocorrelation)*
Insert text

5.2 Difference Equations

Definition 17 *Difference equation:*
A difference equation expresses the value of a variable as a function of its own lagged values, time and other variables.

Consider the following difference equation given by:

$$y_t = \mu + \sum_{i=1}^{n} \phi_i y_{t-i} + x_t \tag{31}$$

Note:

1. The order of the difference equation is given by the number of lags (i.e.: the value of n)
2. The term x_t is called the driving or forcing process and μ is a constant.
3. From an appropriate choice of the forcing process, we can obtain a wide variety of important macroeconomic models (see stochastic processes).
4. This n-th order linear difference equation is a special case where the coefficients a_i are constant. Economic theory may indicate how the various ϕ_i are function of variables in the economy. As long as they do not depend on any values of y_t or x_t, we can regard them as parameters.

5.2.1 Classification of difference equations

A difference equation can be classified according to:

1. Order
2. Linearity/nonlinear
3. Autonomonous/ nonautonomous
4. Stochastic/deterministic

Order. The order of a difference equation is determined by the highest order of difference contained in the equation. An *nth* order difference equation contains variables at most *n* periods apart.

Autonomous. A difference equation is said to be autonomous if it does not depend on time explicitly; otherwise it is non autonomous. For example:

$$y_{t+1} = 2y_t + 3t$$

is nonautonomous because it depends explicitly on the variable t. On the other hand:

$$y_{t+1} = 2y_t + 3$$

is autonomous because it does not depend explicitly on t. **Autonomous difference equations are most prevalent in economics.**

Linear or nonlinear. A difference equation is nonlinear if it involves any nonlinear terms in y_t, y_{t+1} and so on. For example:

$$y_{t+1} = 2y_t^2 + 3$$

is a nonlinear, autonomous, first order difference equation.

Stochastic or deterministic. A difference equation is stochastic if the forcing process is itself a random variable which follows a stochastic process. Alternatively, for a deterministic difference equation, the driving process is either a constant value, or a known sequence of values. One could think of this deterministic component as being captured by the intercept of the equation.

5.2.2 Solution Concept

The concept of a solution to a difference equation is different from the concept of a solution to an algebraic equation where the solution is a variable or a number.

A solution to a linear difference equation expresses the value of y_t as a function of the forcing process $\{x_t\}$, t and possibly some known initial condition (y_0) for the $\{y_t\}$ sequence. A solution to a difference equation is itself a function that makes the difference equation true. In other words, given the difference equation (31), we seek to find the primitive function $f(t)$ that satisfies the difference equation for all permissible values of t and $\{x_t\}$. In a time series context, the solution will be a sequence of values that satisfy the difference equation **at all points in time**.

A corollary of the above is that the substitution of a solution into the difference equation itself must results in an identity. Consider the following examples:

Example 1 *Deterministic difference equation:*

Consider the difference equation

$$y_t = y_{t-1} + 2 \tag{32}$$

A potential solution to this difference equation would be:

$$y_t = 2t + c \tag{33}$$

where c is an arbitrary constant. If (33) is a solution to (32) it must hold for all values of t. As such, for t and t_{-1} we would have:

$$\begin{aligned} y_t &= 2t + c \\ y_{t-1} &= 2(t-1) + c \end{aligned}$$

Substitute the above into (32) yields:

$$2t + c = 2(t-1) + c + 2 \tag{34}$$

which verifies that (34) is an identity and hence (33) a solution to (32).

Example 2 *Stochastic difference equation:*

Consider the following law of motion for investment

$$i_t = 0.7 i_{t-1} + \varepsilon_t \tag{35}$$

The proposed solution in this case is:

$$i_t = \sum_{i=0}^{\infty} (0.7)^i \varepsilon_{t-i} \tag{36}$$

To check the validity of this solution, we iterate (35) one period backwards:

$$i_{t-1} = \sum_{i=0}^{\infty} (0.7)^i \varepsilon_{t-1-i} \tag{37}$$

Substitute (36) and (37) into (35) yields:

$$\varepsilon_t + 0.7\varepsilon_{t-1} + (0.7)^2 \varepsilon_{t-2} + (0.7)^3 \varepsilon_{t-3} + ... = 0.7[\varepsilon_{t-1} + 0.7\varepsilon_{t-2} + (0.7)^2 \varepsilon_{t-3} + ...] + \varepsilon_t \tag{38}$$

The two sides of (38) are identical, which proves that (36) is a solution to (35).

5.2.3 Iterative Solutions

A solution to a linear difference equation expresses the linear difference equation as a function of $\{x_t\}$, time and possibly some initial condition y_0. How to find the particular solution will depend on whether this initial condition is actually known. Two useful approaches are:

1. Iterate Forward (with an initial condition)
2. Iterate Backwards (without an initial condition)

(a) Solution by iteration with an initial condition:

If the value of y in some specific period is known, a direct method of solution is to iterate **forward** from that period onwards to obtain the subsequent time path of the entire $\{y\}$. Refer to this known value of y as the initial condition y_0. To illustrate this technique we'll use a first order difference equation:

(i) Deterministic Case: consider the following difference equation

$$y_t = a_1 y_{t-1} + b \tag{39}$$

Lets say a value y_0 is known at $t = 0$. Then at $t = 1$ equation (39) implies that:

$$y_1 = a_1 y_0 + b \tag{40}$$

At $t = 2$:

$$\begin{aligned} y_2 &= a_1 y_1 + b \\ &= a_1(b + a_1 y_0) + b \\ &= a_1^2 y_0 + b(a_1 + 1) \end{aligned}$$

At $t = 3$:

$$\begin{aligned} y_3 &= a_1 y_2 + b \\ &= a_1[a_1^2 y_0 + b(a_1 + 1)] + b \\ &= a_1^3 y_0 + b(a_1^2 + a_1 + 1) \end{aligned}$$

Keep iterating the until we obtain:

$$y_t = a_1^t y_0 + b(a_1^{t-1} + a_1^{t-2} + ... + a_1 + 1)$$

Note that the above expression between brackets can be also written as:

$$1 + a_1 + a_1^2 + ... + a_1^{t-1} = \begin{cases} \sum_{j=0}^{t-1} a_1^j & \text{if } a_1 \neq 1 \\ t & \text{if } a_1 = 1 \end{cases}$$

Making the solution to the difference equation:

$$y_t = \begin{cases} a_1^t y_0 + b \sum_{j=0}^{t-1} a_1^j & \text{if } a_1 \neq 1 \\ y_0 + bt & \text{if } a_1 = 1 \end{cases}$$

Remark 2 *There is only one solution that satisfies both the difference equation and an initial condition. However, there is, in general, an infinite number of solutions to a linear, first-order difference equation.*

The general solution would be given by:

$$y_t = \begin{cases} C a_1^t y_0 + a_0 \left(\frac{1-a_1^t}{1-a_1} \right) & \text{if } a_1 \neq 1 \\ C + a_0 t & \text{if } a_1 = 1 \end{cases}$$

where C stands for an arbitrary constant. In other words, the presence of an initial condition eliminates the arbitrariness of C.

(ii) Stochastic Case: consider the following difference equation

$$y_t = \mu + \phi y_{t-1} + \varepsilon_t \tag{41}$$

where $\{\varepsilon_t\}$ follows some kind of stochastic process, reflecting the uncertainty about the value of y_t. Given a known value y_0 of $\{y_t\}$, it follows that y_1 will be given by:

$$y_1 = \mu + \phi y_0 + \varepsilon_1 \tag{42}$$

Similarly, y_2 must be:

$$\begin{aligned} y_2 &= \mu + \phi y_1 + \varepsilon_2 \\ &= \mu + \phi[\mu + \phi y_0 + \varepsilon_1] + \varepsilon_2 \\ &= \mu + \mu\phi + (\phi)^2 y_0 + \phi\varepsilon_1 + \varepsilon_2 \end{aligned}$$

Repeated iterations yield:

$$y_t = \mu \sum_{i=0}^{t-1} \phi^i + \phi^t y_0 + \sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i} \tag{43}$$

Assuming $|\phi| < 1$, $\lim_{t\to\infty} \phi^t y_0 = 0$ then (43) converges to:

$$y_t = \frac{\mu}{(1-\phi)} + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i} \tag{44}$$

Remark 3 *Should the difference equation miss a constant term μ, then the sequence would converge to:*

$$y_t = \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i} \tag{45}$$

(b) Solution by iteration without an initial condition:

When no initial condition is given (or the series is assumed to be infinite), the solution can be found by iterating backwards. Consider the following first-order difference equation:

$$\begin{aligned} y_t &= \mu + \phi_1 y_{t-1} + \varepsilon_t \\ y_{t-1} &= \mu + \phi_1 y_{t-2} + \varepsilon_{t-1} \\ &\Rightarrow y_t = \mu + \phi_1\mu + \phi_1^2 y_{t-2} + \varepsilon_t + \phi_1\varepsilon_{t-1} \\ y_{t-2} &= \mu + \phi_1 y_{t-3} + \varepsilon_{t-2} \\ &\Rightarrow y_t = \mu + \phi_1\mu + \phi_1^2(\mu + \phi_1 y_{t-3} + \varepsilon_{t-2}) + \varepsilon_t + \phi_1\varepsilon_{t-1} \\ &= \mu + \phi_1\mu + \phi_1^2\mu + \phi_1^3 y_{t-3} + \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2\varepsilon_{t-2} \end{aligned}$$

which leads to the approximate solution:

$$y_t = \mu \sum_{j=0}^{t-1} \phi^j + \phi^j y_{t-j} + \sum_{j=0}^{t-1} \phi_1^j \varepsilon_{t-j}$$

To understand the behavior of this solution, it is necessary to understand what happens in the limit. If $|\phi_1| < 1$, taking $\lim_{j\to\infty}$ the above expression simplifies to:

$$\begin{aligned} y_t &= \mu \sum_{j=0}^{\infty} \phi^j + \lim_{j\to\infty} \phi^j y_{t-j} + \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j} \\ &= \frac{\mu}{1-\phi_1} + \sum_{j=0}^{\infty} \phi_1^j \varepsilon_{t-j} \end{aligned}$$

given that $\lim_{j\to\infty} \phi^j y_{t-j} = 0$, and $\sum_{j=0}^{\infty} \phi^j = \frac{1}{1-\phi_1}$ for $|\phi_1| < 1$.

The expression above is the solution to this problem with an infinite history. The solution concept is important because it details the relationship between observations in the distant past and the current observations.

Given that when $|\phi_1| < 1$, the expression $\lim_{j\to\infty} \phi^j y_{t-j}$ converges to zero, this implies that the observations arbitrarily fair in the past have no influence on the value of y_t today. Conversely, when $|\phi_1| > 1$ the system is said to be nonconvergent since ϕ_1^t as t grows large and values in the past are not only important, but will in fact dominate when determining the current value of y_t. In the special case which $\phi_1 = 1$, which is a random walk when $\{\varepsilon_t\}$ is a white noise process, the influence of a single ε_t never diminishes.

5.2.4 Steady State and Convergence

An important property of **autonomous** difference equations is that they **often** have a steady state. A steady state is the value of y at which the dynamic system becomes stationary. That is to say, y_{t+1} takes the same value as y_t for all values of t.

Definition 18 *Steady State: The steady state or stationary value in a difference equation is defined as the value of y at which the system comes to rest. This implies that $y_{t+1} = y_t = \bar{y}\ \forall t$.*

Theorem 1 *Existence of steady state:*
In a linear, autonomous, first order difference equation there always exist a steady state as long as $a \neq 1$.

Example 3 *Finding the steady state:*
Consider the deterministic difference equation (39). Its steady state would be:

$$\begin{aligned} \bar{y} &= a_1\bar{y} + b \\ \Rightarrow \bar{y} &= \frac{b}{1-a_1} : a_1 \neq 1 \end{aligned}$$

5.2.5 Convergence to steady state

If y ever becomes equal to its steady-state value, it will remain at that value for all successive time periods. The key question is: if y begins at any arbitrary value different from its steady state, will it always tend to converge towards this point? To answer this we need to study the solution to the difference equation.

Theorem 2 *Convergence of first-order difference equations:*
A linear, first order difference equation will converge to its steady state value (ss) if and only if $|\phi_1|$ <1.

While convergence is guaranteed if $|a_1| < 1$, the path that y_t will take over time is vary depending on the sign of a_1. In turn, we'll have that:

1. If $0 < \phi_1 < 1$, y_t converges monotonically towards the ss
2. If $-1 < \phi_1 < 1$, y_t converges in an oscillating path towards the ss

For all other cases the divergence will be:

1. If $\phi_1 > 1$, y_t diverges exponentially away from the ss
2. If $-\phi_1 < -1$, y_t diverges in an oscillating path away from the ss

Key things to remember

(1) How to find an iterative solution to a difference equation:

- If y_0 is known you iterate forwards. If y_0 is unknown, you iterate backwards.

(2) Does a steady state exist? If so, how to find it.

- A linear, autonomous, first-order difference equation always has a steady state so long as $a_1 \neq 1$.

(3) Does the difference equation converge to its steady state? If so, in which way?

- For a first-order difference equation, convergence to the steady state will depend on the sign and magnitude of the coefficient on the lagged variable (a_1).

When the order is greater than 1, examining the stability of the system is somewhat more involved. The key to understanding the convergence of linear difference equations is the study of its homogeneous portion. Consider the following second order difference equation:

$$\begin{aligned} y_t &= \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \\ &: \ \varepsilon_t \sim WN(0, \sigma^2) \quad \forall t \end{aligned}$$

The homogeneous portion is defined as the term involving only y:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2}$$

The intuition behind studying this portion of the system is that, given the sequence of $\{\varepsilon_t\}$, all the dynamics and the stability of the system are determined by the relationship between contemporaneous y_t and its lagged values. This allows for the determination of the parameter values where the system is stable.

The above equation can be rewritten as:

$$\begin{aligned} y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} &= 0 \\ \text{or, alternatively:} & \\ (1 - \phi_1 L - \phi_2 L^2) y_t &= 0 \end{aligned}$$

For any variable z, consider the following equation based on the rearrangement above:

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 = 0$$

This expression is known as the **characteristic equation** or **characteristic polynomial** of the difference equation. The solution to the equation will be stationary iff all roots of the characteristic polynomial are greater than 1 in absolute value. That is to say, the roots lie outside the unit circle.

It's worth noting that some authors define the above polynomial as:

$$\phi(z) = z^2 - \phi_1 z - \phi_2 = 0$$

This expression is usually referred to as **reverse characteristic equation**. Under this specification the process is stationary iff all roots of the characteristic polynomial are less than 1 in absolute value. That is to say, if all the roots lie inside the unit circle. Consider the case of an AR(2) process:

1. If both roots are smaller than 1 in absolute value, then there are three interesting cases:
   (a) Both roots are real and positive: the system will converge exponentially.
   (b) Both roots are real, but the same: since there is only 1 root, the stability depends on that one value.
   (c) Both roots distinct are imaginary, or real and at least one negative: the system will oscillate and converge.
2. If one or both roots are greater than 1 in absolute value, the system is divergent.

The above expression and can be obtained via two methods. One option implies employing a forward, not a backward operator. Again, the homogeneous part of the AR(2) process would be:

$$\begin{aligned} y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} &= 0 \\ \text{rearranging:} & \\ (F^2 - \phi_1 F - \phi_2) y_{t-2} &= 0 \\ \text{and the characteristic polynomial looks like:} & \\ z^2 - \phi_1 z - \phi_2 &= 0 \end{aligned}$$

where F is a forward operator.

An alternative route to reach the above polynomial is based on the conditions to invert a matrix. Again, consider the same second-order difference equation:

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} \quad = \quad 0$$

Re-write the second order difference equation as a 1st order vector difference equation:

$$\underbrace{\begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix}}_{\zeta_t} = \underbrace{\begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}}_{F} \underbrace{\begin{pmatrix} y_{t-1} \\ y_{t-2} \end{pmatrix}}_{\zeta_{t-1}} + \underbrace{\begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix}}_{v_t}$$

In turn, given:

$$\begin{array}{ccccccc} \zeta_t & = & F & \zeta_{t-1} & + & v_t \\ \text{(2 x 1)} & & \text{(2 x 2)} & \text{(2 x 1)} & & \text{(2 x 1)} \end{array}$$

We can use the insights from a first-order difference equation to study the behavior of this VAR(1). Iterating j-periods out yields:

$$\begin{aligned} \zeta_{t+j} &= F^{j+1}\zeta_{t-1} + F^j v_t + ... + F v_{t+j-1} + v_t \\ &: \quad F^j = \text{F x F x F x ... x F (j times)} \end{aligned}$$

which in our particular example would be [14]:

$$\begin{pmatrix} y_{t+j} \\ y_{t+j-1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}^{j+1} \begin{pmatrix} y_{t-1} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix}^{j} \begin{pmatrix} \varepsilon_t \\ 0 \end{pmatrix} + ... + \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_{t+j-1} \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_{t+j} \\ 0 \end{pmatrix}$$

Stationarity requires that:

$$\lim_{j \to \infty} F^j = 0$$

This condition will be satisfied provided <u>all eigenvalues</u> of F have modulus <u>less than 1</u>.

λ is an eigenvalue of F and x a eigenvector if:

$$\begin{aligned} Fx &= \lambda x \\ &\Rightarrow (F - \lambda I)x = 0 \\ &\Rightarrow (F - \lambda I) \quad \text{is singular} \\ &\Rightarrow \det(F - \lambda I) = 0 \end{aligned}$$

[14] Note that:

$$F^2 = \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \phi_1^2 + \phi_2 & \phi_1\phi_2 \\ \phi_1 & \phi_2 \end{pmatrix}$$

In our example:

$$\begin{aligned}
\det(F - \lambda I) &= \det\left[\begin{pmatrix}\phi_1 & \phi_2 \\ 1 & 0\end{pmatrix} - \begin{pmatrix}\lambda & 0 \\ 0 & \lambda\end{pmatrix}\right] \\
&= \det\begin{pmatrix}\phi_1 - \lambda & \phi_2 \\ 1 & -\lambda\end{pmatrix} \\
&= \lambda^2 - \phi_1\lambda - \phi_2
\end{aligned}$$

The roots λ_1 and λ_2 of the expression above are the eigenvalues of F. It turns out that these eigenvalues are also the solution to the *reverse* characteristic equation:

$$z^2 - \phi_1 z - \phi_2 \quad = \quad 0$$

5.2.6 Equivalent approaches

The conditions stated above are essentially the inverse of each other. The stability condition for the characteristic polynomial

$$1 - \phi_1 z - \phi_2 z^2 \quad = \quad 0$$

is that all roots lie <u>outside</u> the unit circle. While the stability condition for the *reverse* characteristic equation

$$\lambda^2 - \phi_1\lambda - \phi_2 \quad = \quad 0$$

is that all roots lie <u>inside</u> the unit circle. By the fundamental theorem of algebra we know both of these polynomials have two roots. In turn:

$$\begin{aligned}
1 - \phi_1 z - \phi_2 z^2 &= (1 - \lambda_1 z)(1 - \lambda_2 z) \\
\text{so that} & \\
&\Rightarrow \quad z_1 = \frac{1}{\lambda_1}, \quad z_2 = \frac{1}{\lambda_2}
\end{aligned}$$

are the roots of the characteristic equation. The values λ_1 and λ_2 are the roots of the eigenvalues of F.

Result: The inverses of the roots of the characteristic equation

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p \quad = \quad 0$$

are the eigenvalues of the companion matrix F. The *reverse* characteristic equation

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \ldots - \phi_p\lambda - \phi_p \quad = \quad 0$$

is the same polynomial equation used to find the eigenvalues of F.

5.2.7 Understanding the vector stability condition:

To see why $|\lambda_i| < 1$ implies $\lim_{j\to\infty} F^j = 0$ consider the AR(2) with real-valued eigenvalues. By the eigen-decomposition of a squared matrix we have that:

$$\begin{aligned} F &= T\Lambda T^{-1} \\ &: \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \\ &: \quad T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} \\ &: \quad T^{-1} = \begin{pmatrix} t^{11} & t^{12} \\ t^{21} & t^{22} \end{pmatrix} \end{aligned}$$

Then

$$\begin{aligned} F^j &= (T\Lambda T^{-1})\text{x ... x}(T\Lambda T^{-1}) \\ &= T\Lambda^j T^{-1} \end{aligned}$$

Iterating forward yields:

$$\lim_{j\to\infty} F^j = \lim_{j\to\infty} T\Lambda^j T^{-1} = 0$$

provided that $|\lambda_1| < 1$ and $|\lambda_2| < 1$.

Note that:

$$\begin{aligned} F &= T\Delta^j T^{-1} \\ &= \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix} \begin{pmatrix} \lambda_1^j & 0 \\ 0 & \lambda_2^j \end{pmatrix} \begin{pmatrix} t^{11} & t^{12} \\ t^{21} & t^{22} \end{pmatrix} \end{aligned}$$

5.2.8 Aside: Eigenvalues and the characteristic equation:

Eigenvectors and eigenvalues are numbers and vectores associated to sqaure matrices, and together they provide the eigen-decomposition of a matrix which analyzes the its structure. Eigenvectors and eigenvalues are also referred to as characteristic vectors and latent roots or characteristic equation (in German,"eigen" means "specific of" or "characteristic of"). The set of eigenvalues of a matrix is also called its spectrum, so often the eigenvalue decomposition of a matrix is called spectral decompisition.

There are several ways to define eigenvectors and eigenvalues, the most common approach defines an eigenvector of the matrix A as a vector x that satisfies the following equation:

$$Ax = \lambda x$$

when rewritten, the equation becomes:

$$(A - \lambda I)x = 0$$

where λ is a scalar called the eigenvalue associated to the eigenvector.

In a similar manner, we can also state that a vector x is an eigenvector of matrix A if only the length of the vector (but not its direction) is changed when it is multiplied by the matrix. Take for example the following matrix:

$$A = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$$

has eigenvectors:

$$x_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

with eigenvalue $\lambda_1 = 4$, and

$$x_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

with eigenvalue $\lambda_2 = -1$. We can verify that only the length of x_1 and x_2 is changed when one of these two vectors is multiplied by the matrix A:

$$\begin{aligned} \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} &= \begin{pmatrix} 12 \\ 8 \end{pmatrix} \\ 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix} &= \begin{pmatrix} 12 \\ 8 \end{pmatrix} \\ &\text{hence:} \\ \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} &= 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix} \end{aligned}$$

Similarly:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -1 \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Traditionally, we put together the set of eigenvectors of A in a matrix denoted U. Each column of U is a eigenvector of A. The eigenvalues are stored in a diagonal matrix denoted Δ, where the diagonal elements gives the eigenvalues (and all other values are zeros). We can then reqrite the original equation as:

$$\begin{aligned} A &= U \Delta U^{-1} \\ &\text{or also:} \\ A &= U \Delta U^{-1} \end{aligned}$$

It is important to note that not all square matrices have eigenvalues.

5.2.9 Matrix Factorization

A matrix factorization is the decomposition of a matrix into a product of matrices. There are many different matrix factorizations; each finds use among a particular class of problems.

5.2.9.1 LU decomposition

In numerical analysis and linear algebra, LU decomposition (also called LU factorization) factors a matrix as the product of a lower triangular matrix "L", and an upper triangular matrix "U". The method is mainly applied to square matrices, although it can also be applied to non-squared one as well as non invertible matrices.

Consider the following square and symmetric matrix A:

$$A = \begin{pmatrix} 4 & 4 & 6 \\ 4 & 13 & 15 \\ 6 & 15 & 43 \end{pmatrix}$$

This method consists of performing Gaussian elimination on matrix A, until a upper triangular matrix "U" is obtained. Next employ the elementary matrices which keep track of the Gaussian row operations to obtain the lower triangular "L". Keeping the above case in mind:

Gaussian Elimination	Elementary Matrix
$A = \begin{pmatrix} 4 & 4 & 6 \\ 4 & 13 & 15 \\ 6 & 15 & 43 \end{pmatrix} \underbrace{\rightarrow}_{R2-R1} \begin{pmatrix} 4 & 4 & 6 \\ 0 & 9 & 9 \\ 6 & 15 & 43 \end{pmatrix}$	$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{l_1}$
$\begin{pmatrix} 4 & 4 & 6 \\ 0 & 9 & 9 \\ 6 & 15 & 43 \end{pmatrix} \underbrace{\rightarrow}_{R3-\frac{3}{2}R1} \begin{pmatrix} 4 & 4 & 6 \\ 0 & 9 & 9 \\ 0 & 9 & 34 \end{pmatrix}$	$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{3}{2} & 0 & 1 \end{pmatrix}}_{l_2}$
$\begin{pmatrix} 4 & 4 & 6 \\ 0 & 9 & 9 \\ 0 & 9 & 34 \end{pmatrix} \underbrace{\rightarrow}_{R3-R2} \begin{pmatrix} 4 & 4 & 6 \\ 0 & 9 & 9 \\ 0 & 0 & 25 \end{pmatrix}$	$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}}_{l_3}$

which implies that pre-multiplication by l_3, l_2 and l_1 of A yields the upper triangular matrix U. In this sense, the steps of the Gaussian elimination are captured by the elementary matrices above. Now lets define:

$$\begin{aligned} l &= l_3 l_2 l_1 \\ l^{-1} &= L \\ &\Rightarrow \; L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{3}{2} & 1 & 1 \end{pmatrix} \end{aligned}$$

Given the fact that matrix "L" above captures all of the Gaussian elimination steps, this will usually imply that it will have "1" along its diagonal. Hence the LU decomposition of the matrix A is complete:

$$\underbrace{\begin{pmatrix} 4 & 4 & 6 \\ 4 & 13 & 15 \\ 6 & 15 & 43 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{3}{2} & 1 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 4 & 4 & 6 \\ 0 & 9 & 9 \\ 0 & 0 & 25 \end{pmatrix}}_{U}$$

5.2.9.2 LDU decomposition

In the example above, the "L" matrix captures all the Gaussian elimination steps, and hence will usually have "1" along its diagonal. With the LDU decomposition we strive to achieve a similar structure of the "U" matrix by factoring the matrix pivots so that "U" also has "1" in its diagonal. The basic idea will be to use the pivots from matrix "U" to factorize that matrix and obtain a new diagonal matrix "U". Following the example above:

$$\underbrace{\begin{pmatrix} 4 & 4 & 6 \\ 4 & 13 & 15 \\ 6 & 15 & 43 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{3}{2} & 1 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 25 \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 1 & 1 & \frac{3}{2} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{U}$$

where basically matrix "D" contains the three pivots from the original matrix "U", and the new matrix "U" has each of its rows divided by one of the pivots now contained in matrix "D". In turn, the new matrix "U" does have ones along its diagonal, while matrix "L" remains the same as before.

Note that because matrix "A" is symmetric, then "U" is now the transpose of "L". In turn:

$$A = LDL'$$

5.2.9.3 Choleski decomposition

The Choleski decomposition aims to reduce the above factorization into just two matrices. It aims to do so by factoring out the diagonal matrix "D" above, and enhancing both "L" and "U" matrices. Since "D" is a diagonal matrix, a typical way of doing this is by taking the square root of the matrix. In turn:

$$\begin{aligned} \underbrace{\begin{pmatrix} 4 & 4 & 6 \\ 4 & 13 & 15 \\ 6 & 15 & 43 \end{pmatrix}}_{A} &= \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{3}{2} & 1 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}}_{\sqrt{D}} \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}}_{\sqrt{D}} \underbrace{\begin{pmatrix} 1 & 1 & \frac{3}{2} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{U} \\ &= \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 2 & 3 & 0 \\ 3 & 3 & 5 \end{pmatrix}}_{P} \underbrace{\begin{pmatrix} 2 & 2 & 3 \\ 0 & 3 & 3 \\ 0 & 0 & 5 \end{pmatrix}}_{P'} \end{aligned}$$

References:

- J. H. Stock and M.W. Watson, 2003. Introduction to Econometrics.
- J. H. Stock and M.W. Watson, 2001. Vector Autoregression. Journal of Economic Perspectives 15, 101-115.
- W. H. Greene, 2000. Econometric Analysis.
- J.D., Hamilton 1994. Time Series Analysis.
- W. Enders, 2004. Applied Econometric Time Series.