ECO 434H

Forecasting Methods in Macroeconomics and Finance

Lecture notes

by Paul Karapanagiotidis

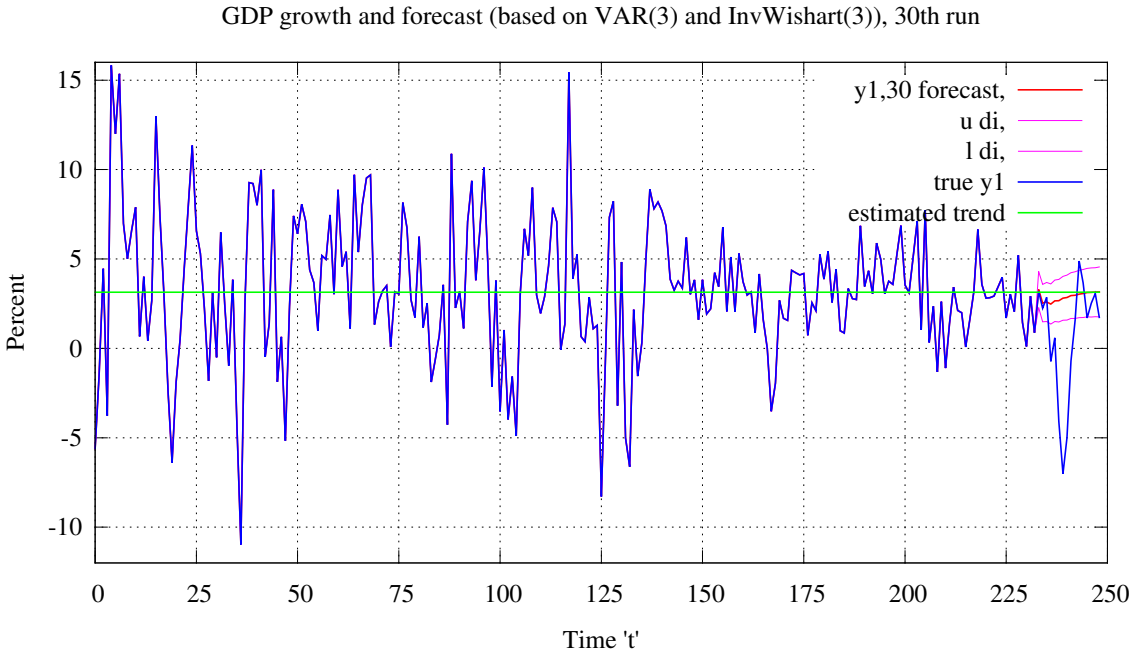
January-April, 2013

- Review the course outline.

- **The importance of time series analysis**

  - The study of time series analysis is important in many fields:
    - **engineering** - the "filtering" of a noisy electro-magnetic signal (e.g. car stereo equalizer or radio tuner).
    - **medicine** - incidence of particular disease in a given geographical area over time.
    - **meteorology** - monitoring and predicting weather conditions.
    - **economics** - measuring economic fundamentals like the GDP growth rate (see Figure 1).

Figure 1: US GDP growth forecast, 1948-2011



  - The basic idea is that the past tells us something about the future. Or put another way, the past repeats itself. That is, there are patterns in the past that will occur again in the future.
  - We wish to construct a mathematical probability model that most accurately "fits" the given historical data and which provides reasonable forecasts of the future based on these patterns "hidden" in the data.

Definition 1. *Stochastic process:*
Given a probability space $(\Omega, \mathcal{F}, P)$ and a measurable space $(S, \mathcal{S})$, an S-valued stochastic process is a collection of S-valued random variables on Ω indexed by a totally ordered set T (i.e. time). That is a stochastic process X is a collection $X_t : t \in T$ where each X_t is an S-valued random variable on Ω. The space S is called the state space of the process.

- ◇ The above definition is basically a fancy way of saying that we can always consider any time series, X_t, to be the result of some random occurrences across time.
  - ∗ Of course, the "true" process in question is actually deterministic,[1] but usually extremely complicated, and so it is easier to model it using the mathematics of randomness.
- ◇ Typically we will also impose a *parametric model* on the probability distribution of the process across time (notice that **Definition 1** above says nothing about the probability law P that defines X_t.)
  - ∗ That is, we will write down some equations that we believe describe the "true" behaviour of the process, and these equations will then define the probability distribution of the process.
  - ∗ We call this assumed "true behaviour" the Data Generating Process (or DGP). For example, we may assume that:

$$x_t = ax_{t-1} + \epsilon_t, \quad \text{where} \quad \epsilon_t \sim N(0,1) \tag{1}$$

    This means that x_t is always a function of its past value x_{t-1} plus some random variable ϵ_t which is Normally distributed with mean 0 and variance 1.
  - ∗ We can see that if we recursively substitute back in time, this model implies that $E[x_t] = aE[x_{t-1}] + 0 = a(aE[x_{t-2}] + 0) + 0 = \ldots$, etc, so that given some starting value of the process x_0 we have $E[x_t] = a^t E[x_0]$.
  - ∗ The same type of analysis can be applied to determine the other moments (e.g. both the conditional and unconditional variance, skewness, kurtosis, etc) which define the probability distribution of the stochastic process x_t.
  - ∗ We will talk more about what a "moment" of the process is in the probability review section below, but the basic idea is that the "moments" of the process define its probability density function.
  - ∗ Of course, our assumed DGP is *never correct*. It is simply a matter of *how well* it describes the underlying relationships we are trying to model, that are hidden in the data. Figuring out how well it describes these relationships is also not as easy as it first seems since we need to determine some measure of performance. This course is all about how we model different types of macroeconomic and financial time-series data, and ways we can go about testing model performance.

- ○ **Historical context**
- ○ *The Dynamics of Keynesian Theory: The Klein-Goldberger Model of the 1950's*

[1] Excluding exotic phenomenon like quantum particles, for example.

◇ The way we used to model the economy was through massive systems of simultaneous equations models. For example, consider the Keynesian model of the closed economy from your first year macro class:

The IS curve:

$$Y_t = C_t + I_t + G_t \tag{2a}$$
$$C_t = \alpha_{IS} + \beta_{IS}(Y_t - T_t) \tag{2b}$$
$$I_t = \gamma_{IS} - \delta_{IS} r_t \tag{2c}$$

The LM curve:

$$M_t/P_t = RM_t = L(r_t, Y_t) = \alpha_{LM} Y_t - \beta_{LM} r_t \tag{2d}$$

◇ Thousands of equations like this that defined the relationships between different sectors of the economy were each estimated by OLS. The results were then used to predict the economy or to inform policy decisions. Needless to say it didn't work very well – one of the reasons being that the parameters (e.g. α_{IS}) were fixed after being estimated.

◇ Models such as the Klein-Goldberger above are considered *structural* models, since they try to model structural relationships between economic variables that are informed by theory.

- ∗ The opposite of structural models are *reduced form* models.
- ∗ The reduced form model is where we write each of the endogenous variables in terms of only the exogenous variables of the model (in the above model, we assume that $G_t = T_t$ and so G_t and RM_t are the exogenous variables.)
- ∗ For example, if we combine the equations in (2) we get:

$$Y_t = \alpha_{IS} + \beta_{IS}(Y_t - T_t) + \gamma_{IS} - \delta_{IS} r_t + G_t \quad \text{where} \quad G_t = T_t \tag{3a}$$
$$= (\alpha_{IS} + \gamma_{IS}) + \beta_{IS} Y_t + G_t(1 - \beta_{IS}) - \delta_{IS} r_t \tag{3b}$$
$$M_t/P_t = RM_t = L(r_t, Y_t) = \alpha_{LM} Y_t - \beta_{LM} r_t \tag{3c}$$

so the IS equation can be rewritten as $Y_t = f_1(r_t, G_t)$, that is, Y_t as some linear function of r_t and G_t, and the LM equation can be written as $Y_t = f_2(r_t, RM_t)$, that is, as a linear function of r_t and RM_t (the interest rate and real money supply). If we combine the two equations, we can solve for the reduced forms, $Y_t = g_1(G_t, RM_t)$ and $r_t = g_2(G_t, RM_t)$.

◇ It is very important to note, that when we write out an equation for the parametric model, like in equation (1) above, we are really writing out the reduced form of some structural relationship we do not know (where x_{t-1} and ϵ_t are the exogenous variables.) We will discuss this fact in more detail later in the 3rd part of the course when we consider the VAR(p) models which are propular in Macroeconomics.

- ∗ While we can usually go from the structural form to the reduced form as we did above (given certain mathematical conditions), going from the reduced form to the structural form is not so easy. That is, suppose I estimated (1), and determined the parameter a. It is not clear what structural model this reduced form implies.

◇ Either way, modern econometric time series methods make use of models such as those defined in equation (1) all the time.

- ∗ Even though it says nothing about the theory behind the economic relationship between x_t and x_{t-1}, (1) is still an extremely useful model for forecasting purposes since it models the future as a random function of the past.

- In fact, nearly all the models we will discuss in this course will say nothing theoretical about the economic relationships between variables, as you will see.

- **Time domain vs. Frequency domain analysis**

  - When we write down an equation like that in (1), we are saying that x_t is some random function of x_{t-1} and ϵ_t in terms of the **time** subscript t. Functions such as this are called *stochastic difference equations*. [2]
  - However, we know from Fourier Analysis that any time series can be analyzed in either the time-domain or the frequency-domain.
    - What does this mean? It means that any function of time can always be rewritten as a sum of $Sine(\omega t)$ and $Cosine(\omega t)$ functions (where $1/\omega$ denotes the periodicity of these functions); that is, any function of time can be written as:

$$f(t) = \sum_{\omega=-\infty}^{\infty} a_\omega Cos(2\pi t\omega) + b_\omega Sin(2\pi t\omega)$$

    - Don't believe me? Take for example the repeating function $f(t) = t^3$ where $t \in (0, 2)$ (see Figure 7, where K is the number of terms in the sum when we use less than infinity).

[2] If we were in continuous time the analog to (1) is the stochastic differential equation, $dx_t = ax_t dt + \sigma_t dW_t$, where W_t is a Wiener process. This is also known as the Ornstein-Uhlenbeck process.

Figure 2: K=100

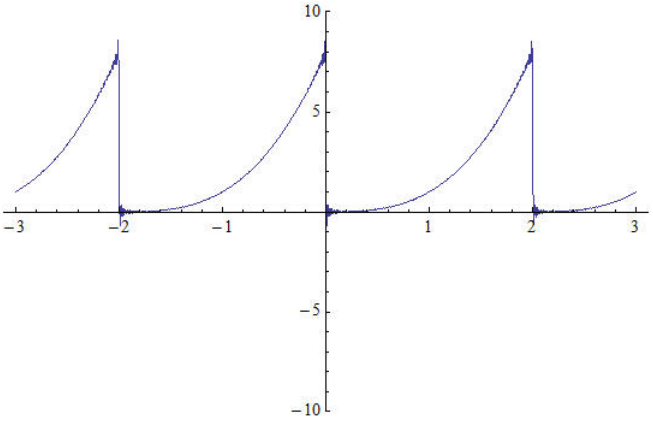


Figure 3: K=30

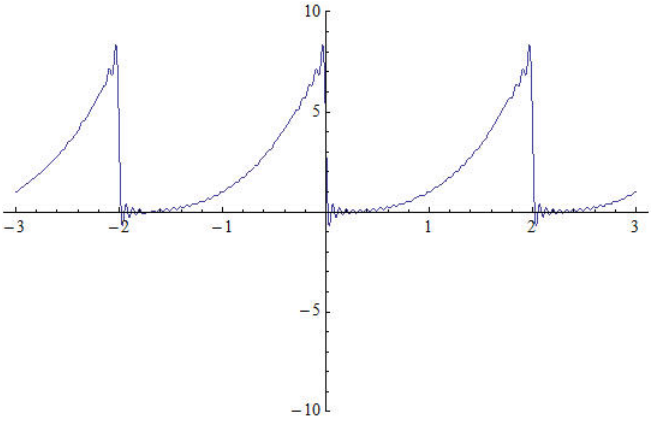


Figure 4: K=10

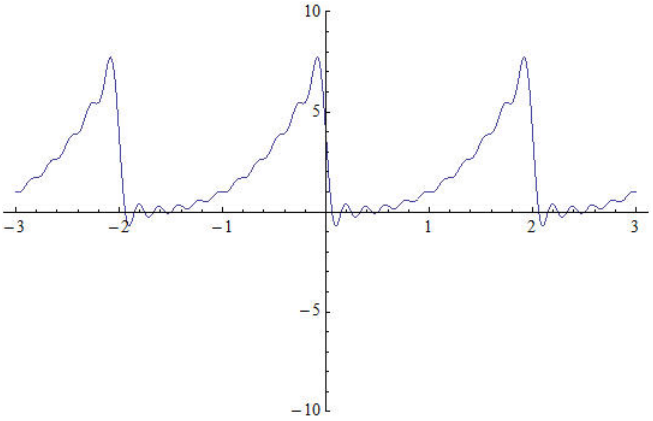
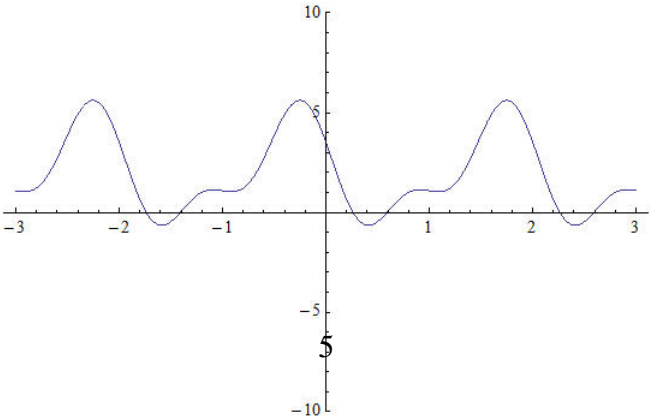


Figure 5: K=2



- ◇ Of course, it is the coefficients a_ω and b_ω that determine what function we get. They are the heart of the magic!
- ◇ Therefore, if we do not include enough of the "high" frequency terms in the sum, we do not get the sharp edges of $f(t)$ being represented correctly.
- ◇ How is this relevant to economic time-series? Well the frequency approach decomposes the time-series by frequency so that we can analyze which coefficients a_ω and b_ω have the most "weight." For example, if $a_{3/12}$ and $b_{3/12}$ were large, given monthly data, this might suggest some pattern that repeats every 3 months (seasonality). That is, the functions $Sin(2\pi\frac{3}{12}t)$ and $Cos(2\pi\frac{3}{12}t)$ are important in describing the movement in $f(t)$.
- ◇ While we will not deal with the frequency domain analysis in this class, it is important to know that it exists! It really is one-half of the story, so to speak.

- **Review of Probability Theory**

Definition 2. *Cumulative Distribution Function (CDF):*
For every real number x, the CDF of the random variable X is given by $F_X(x) = P(X \leq x)$, where $P(X \leq x)$ denotes the probability that $X \leq x$.

Definition 3. *Probability density function (PDF):*
For every real number x, the PDF of the random variable X is given by some function $f_X(x)$ such that $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and $P(a \leq X \leq b) = \int_a^b f_X(x)dx$.

Corollary 1. *If $f_X(x)$ is continuous at x, then the PDF represents the derivative of the CDF. That is, $f_X(x) = dF_X(x)/dx$.*

Corollary 2. *We have that $P(a \leq X \leq b) = F_X(b) - F_X(a)$.*

Definition 4. *Bayes Theorem:*
We have that $f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y) = f_{Y|X}(y|x)f_X(x)/f_Y(y)$, where $f_{X|Y}(x|y)$ is called the conditional probability density function of X, conditioned on the random variable Y.

Definition 5. *Moments about the origin (uncentered):*
We call $E[X^p] = \int_{-\infty}^{\infty} x^p f_X(x)dx$, the p'th moment about the origin.

Definition 6. *Centered moment:*
We call $E[(X - E[X])^p] = \int_{-\infty}^{\infty}(x - E[X])^p f_X(x)dx$, the p'th centered moment.

- It turns out that the uncentered and centered moments tell us a lot about the probability distribution of X. In fact, the whole notion of a "moment" is a way, loosely speaking, of quantifying the relationship between a set of points.

- ◇ For example, the 1st uncentered moment, $E[X]$, is called the *mean* of the distribution.
- ◇ Moreover, the second centered moment is called the *variance*, $E[(X - E[X])^2]$. Of course, if the mean is equal to zero, then the variance is the 2nd uncentered moment, $E[X^2]$. The variance tells us about how much the values of the distribution are dispersed evenly around the mean.
- ◇ There is a nice trick you can use if it is sometimes difficult to find the 2nd centered moment, but easier to find the uncentered ones. It can be shown that $E[(X-E[X])^2] = E[X^2]-E[X]^2$.
- ◇ The third central moment, $E[(X - E[X])^3]$ is called the skewness (or the standardized skewness as $E[(X - E[X])^3]/(E[(X - E[X])^2])^{3/2}$). The skewness tells about any asymmetry in the distribution on the left or right sides of the mean.
- ◇ Finally, we have the kurtosis, $E[(X - E[X])^4]/(E[(X - E[X])^2])^2$, which tells us about how much "weight" the tails of the probability distribution carry.
- ◇ If we subtract 3 from the kurtosis, we get "excess" kurtosis, which for a Normal distribution is 0. That is, the standard Normal distribution with mean 0 and variance 1 has kurtosis 3. Be careful which value you are using! Some software uses the regular version and other's use the excess kurtosis – read the documentation!
- ◇ If a distribution has excess kurtosis greater than zero, we call the distribution "fat tailed", "heavy tailed," or "leptokurtic." What this means is that outliers are more commonly occurring than with the Normal distribution. That is, there are more often cases of "rare" events that take place in the tails of the distribution. This will be important when we talk about models in Finance later.

Definition 7. *Conditional moments:*
We call $E_{X|Y}[(X - E_{X|Y}[X])^p] = \int_{-\infty}^{\infty}(x - E_{X|Y}[X])^p f_{X|Y}(x|y)dx$*, the p'th conditional centered moment and* $E_{X|Y}[X^p] = \int_{-\infty}^{\infty} x^p f_{X|Y}(x|y)dx$*, the conditional uncentered moment (about the origin).*

- ○ Note that sometimes you see the conditional moment denoted as $E_{X|Y}[X] \equiv E_X[X|Y] \equiv E[X|Y]$, but I like to be very specific in notation so that there is no confusion which underlying density $f_{X|Y}(x|y)$ we are integrating across. Sometimes authors do not specify any subscripts on $E[X|Y]$ and this can lead to much confusion, if we do not know what density to integrate with.
- ○ When are talking about a time-series, X_t, we often denote the conditional densities and expectations by t instead of by X and Y. That is, $E_{X_t|X_{t-1}}[X_t] \equiv E_{t-1}[X_t]$, is the conditional first uncentered moment of X_t conditioned on only the last value that occurred, X_{t-1}. Note that we must use the conditional density $f_{t-1}(x_t) \equiv f(x_t|x_{t-1})$, when integrating.
- ○ We will also sometimes speak of the "information set." The information set in the above case is simply X_{t-1}, and we may denote it as $F_{t-1} = \{X_{t-1}\}$. Of course, the information set can include many past values of X_τ, where $\tau = \{t-1, t-2, \ldots, t_0\}$ for example–it need not only include the last value! We will see more on this later.

Corollary 3. *From Bayes Theorem we have that if* F_{t-1} *is an information set, such that* $F_{t-1} = \{X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_0\}$ *then we can write the conditional density of* X_t*, conditioned on the information set, as:*

$$f(x_t|F_{t-1}) = f(F_{t-1}, x_t)/f(F_{t-1}) = f(F_t)/f(F_{t-1}) = f(F_{t-1}|x_t)f(x_t)/f(F_{t-1})$$

Corollary 4. *Factorization of the joint density:*
We can always factorize a joint density $f_{X,Y}(x, y)$ as $f_{X|Y}(x|y)f_Y(y)$.

Corollary 5. *The above implies that since F_{t-1} is a set, we can factorize over and over. That is:*

$$f(x_t, F_{t-1}) = f(x_t|F_{t-1})f(F_{t-1}) = f(x_t|F_{t-1})f(x_{t-1}, F_{t-2}) = f(x_t|F_{t-1})f(x_{t-1}|F_{t-2})f(F_{t-2})$$
$$\cdots = f(x_t|F_{t-1})f(x_{t-1}|F_{t-2})\dots f(x_0)$$

- This type of factorization will be important later when we talk about how models such as the one defined by the expression in (1) for example, imply both a conditional distribution (conditional on x_{t-1} in (1)) and an unconditional distribution.

Definition 8. *Law of iterated expectations (sometimes called "Tower property of expectations"):*
We have that $E_Y[E_X[X|Y]] = E_X[X]$. And if we condition on some other variable, Z we have: $E_Y[E_X[X|Y]|Z] = E_X[X|Z]$.

- What this is really saying is that since $E_X[X|Y]$ is random (for any outcome of Y we have a different expectation), if we take the expectation again, with respect to Y then this is just the same as taking the expectation across all values of X, irrespective of Y.
- In terms of a time-series process X_t, we have the equivalent representation where $X \equiv X_{t+\tau+v}, Y \equiv X_{t+\tau}, Z \equiv X_t$ (and where I now use the alternative notation $E_t[X_{t+1}] \equiv E[X_{t+1}|X_t]$):

$$E_t[E_{t+\tau}[X_{t+\tau+v}]] = E_t[X_{t+\tau+v}]$$

- So now we see that taking the expectation of an expectation at a later time is the same as just taking the expectation at the earlier time. That is, intuitively, our "best guess" at time t of our best guess at time $t + \tau$ is just our best guess at time t (confusing?).
- Some other useful properties are:
  - $E_t[X_t] = X_t$. That is, the best guess at time t is just the value at time t.
  - Also, $E_t[a + bX_{t+\tau}] = a + bE_t[X_{t+\tau}]$, so expectation is a linear operator; that is, it "passes through" linear functions. However, $E_t[a + bX_{t+\tau}^2] \neq a + b(E_t[X_{t+\tau}])^2$!
  - $E_t[X_tX_{t+\tau}] = X_tE_t[X_{t+\tau}]$, that is, we can treat X_t as a constant if we are conditioning on time t information, F_t.
  - Finally, remember that X_t and $X_{t+\tau}$ are different random variables; as such $E[X_tX_{t+\tau}] \neq E[X_t]E[X_{t+\tau}]$ unless X_t and $X_{t+\tau}$ are independent. Generally, $E[X_tX_t] = E[X_t^2] \neq E[X_t]E[X_t]$ so we always "group" random variables with the same time subscripts together before determining expectations.

- **The theoretical model vs. the sampling distribution**

- Finally, we need to always keep in mind the difference between the theoretical model, or DGP, that we assume on the data, and the sampling distributions.

- For example, consider the random variable X. We will assume the DGP is $X \sim N(\mu, \sigma^2)$. Now suppose we calculate the mean $\hat{X} = \sum_{t=1}^{T} X_t/T$ from some finite sample of size T. Of course, $\hat{X}$ will have its own distribution completely different from X since it is a random variable too for every sample we take. In this case, it has the distribution $\hat{X} \sim N(\mu, \sigma^2/T)$. The distribution of $\hat{X}$ is called the "sampling distribution" of the estimator $\hat{X}$.

- As another example, consider the random variable Y. We will assume that Y has a DGP that is conditional on X, so let's say for example $Y|X \sim N(a + bX, \sigma^2)$. Of course, for every X, $Y = a + bX + \sigma Z$ where Z has unconditional distribution $Z \sim N(0, 1)$.

  - Now suppose we want to estimate b. Since the DGP (or theoretical model) we chose is linear, we can use OLS. Using OLS will get us some estimate $\hat{\beta}$.
  - The point is that $\hat{\beta}$ is a random variable for any sample of X and Y.
  - Therefore, $\hat{\beta}$ has its own sampling distribution that is totally different (but based on) the distributions of X and Y.
  - In this case we know from the results on the "simple regression" model version of OLS (i.e. the univariate X version) that the estimator of b is given as:

$$\hat{\beta} = \frac{\hat{Cov}(X, Y)}{\hat{Var}(X)}$$

    And the estimate of a is $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, where $\bar{x}$ is the sample mean of X.
  - Therefore, the estimate of b is the sample covariance of X and Y divided by the sample variance of X.
  - It can be shown that the sampling distribution of $\hat{\beta}$ is such that $E[\hat{\beta}] = b$ (we say it is an unbiased estimator) and that $Var(\hat{\beta}) = \sigma^2/(\sum_{n=1}^{N} x_n^2)$. Moreover, as $N \to \infty$ the sampling distribution of $\hat{\beta}$ approaches the Normal distribution.

- **Part 1: ARIMA models**

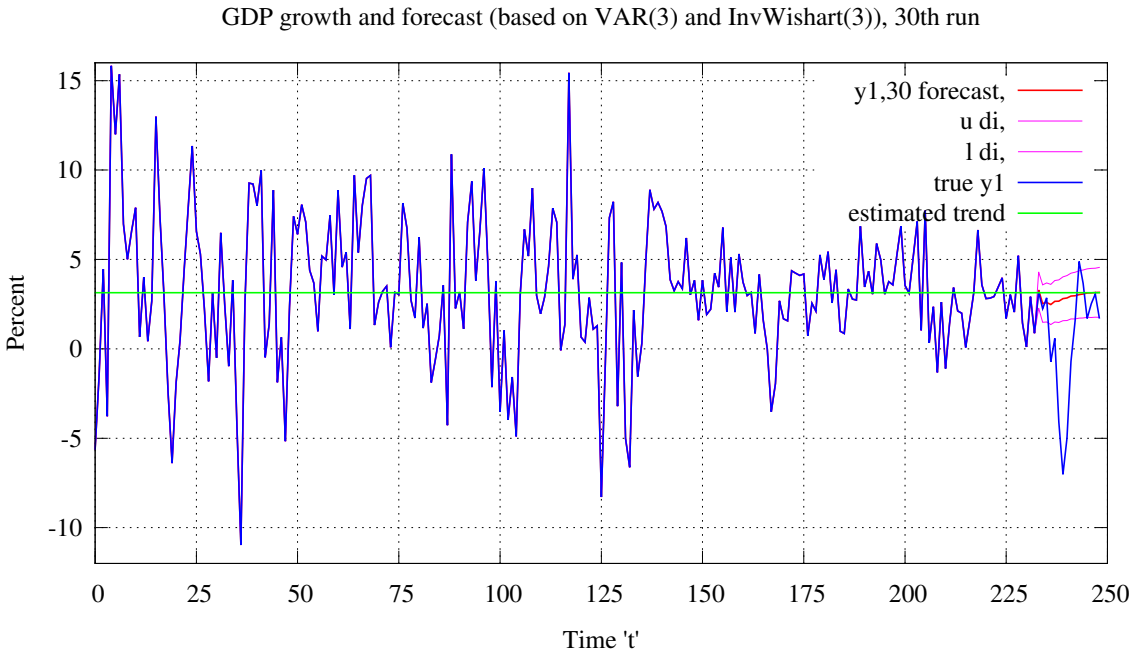
- **Stationarity and white noise**

- Consider the time series from the introduction discussion, which i've redisplayed below:

- What can we say about this time series?

  - Well for one thing, it looks like if we took the **sample mean** of the entire series it would match up approximated with the green line.
  - Moreover, for most parts of the series, say if we took the segment between 25 and 50, the mean is the same as it is between any other segment, say between 150 and 175.

Figure 6: US GDP growth forecast, 1948-2011



- ◇ That is, the series doesn't just seem to wander up or down indefinitely so that the mean is growing or declining as a function of time (at least approximately).
- ◇ Moreover, let us not only consider taking the sample mean, but let us also imagine taking the **sample variance**. Does it look like the sample variance would change if we took it across different segments of the sample?

- ○ Intuitively, what we are asking when we say "does this time series' moments look like they are relatively fixed over time" is the quesiton of whether a time-series is what we call *stationary*.

- ○ A series is in some sense "as stationary as its moments are fixed" so to the extent that more moments are "fixed" we can say the series is more or less stationary.

  - ◇ Bear in mind I am speaking loosely here since even if an infinite number of moments are both finite and constant across time, this doesn't necessarily imply that the distribution itself is constant across time.

- ○ **Also it is important to remember two things:**

  1. We are talking about the *unconditional* moments here–NOT the conditional moments. This distinction is important, since the conditional moments are still allowed to change in a stationary series.
     - ◇ Recall the factorization I spoke of in corollary 5.
     - ◇ In that case, we wrote out the joint PDF as a factorization of conditional densities: $f(x_t, F_{t-1}) = f(x_t, x_{t-1}, \ldots, x_0) = f(x_t|F_{t-1})f(x_{t-1}|F_{t-2})\ldots f(x_0)$.
     - ◇ Therefore, we can allow for the joint density to be constant for any series X_τ, $\tau = t, t-1, t-2, \ldots, 0$, as time changes, while the conditional densities themselves are all different for each time τ.

- ◇ This is very important since we often speak of writing out expressions for our model's DGP (the theoretical model) which define the conditional moments and which we usually have to solve for the unconditional ones (we will do this shortly–also recall what we talked about above in relation to equation (1)).

2. Even though I used the example of taking the sample moments at different points in time, recall our earlier discussion about the difference between the theoretical model and the sampling distributions.
   - ◇ When I speak of fixed moments through time, what I really mean is a *fixed DGP* or theoretical model. Since we never really know what the true DGP is, we are always just making a "guess" or an "assumption" on what we think it is. We then test the assumption by means of sampling distribution statistics (like testing the sample mean and sample variance to see if they are constant), but we must always remember that the concept of stationarity of the DGP is an assumption we place on the *theoretical model*.

◦ So, with all that said, let us now present the official definitions of a stationary process X_t. Note there are two version usually used, the first is called the strict (or strong form) stationarity and the second is called weak stationarity. In this course we will primarily be concerned with the weak form.

Definition 9. *Strictly stationary series:*
Given a strictly stationary series X_t, we have that:

$$F(\{X_t\}) = F(\{X_{t+\tau}\}), \quad \forall t, \tau \in \mathbb{N}$$

where $F(\cdot)$ is the cumulative distribution function (CDF) of the stochastic process $\{X_t\}$.

◦ So what this definition is saying is that the CDF is not changing for any theoretical stochastic process (i.e. times series) $\{X_t\}$.

◦ The reason why this is framed in terms of the CDF, and not the PDF, is because for some processes the derivative of $F(\cdot)$ does not exist.

Definition 10. *Weakly stationary series:*
Given a weakly stationary series X_t, we have that:

- ◇ $E[X_t] = \mu < \infty$
- ◇ $E[(X_t - \mu)^2] = \sigma^2 = R_X(0) < \infty$
- ◇ $E[(X_t - \mu)(X_{t-s} - \mu)] = R_X(s)$ *only depends on s.*

◦ So weak stationarity says the following:

- ◇ The mean of every X_τ, $\tau = t, t-1, t-2, \ldots, 0$ must be the same across time and finite. This is the same as saying that the mean of X_t must not be a function of t.
- ◇ The variance of every X_τ, $\tau = t, t-1, t-2, \ldots, 0$ must be the same across time and finite. Again, this is the same as saying that the variance of X_t must not be a function of t.
- ◇ Finally, the autocovariance $E[(X_t - \mu)(X_{t-s} - \mu)] = R_X(s)$ must only be a function of s, not a function of t.

- ∗ But you might be asking, what is the autocovariance and how does it differ from the variance? Well the answer is simple. The autocovariance, $R_X(s)$ is just the variance, but not between X_t and itself (which would be $R_X(0)$), but between X_t and X_{t-s}. Remember, that I said earlier that even though X_t and X_{t-s} (for any $s \in \mathbb{Z}$ but not $s = 0$) are part of the same stochastic series $\{X_t\}$, they are *different* random variables.
- ∗ Intuitively, what the autocovariance condition of weak stationarity is saying is that no matter which lag we choose s, the covariance between X_t and X_{t-s} should be constant across time t.

◦ So now that we have that our of the way, we can also define the normalized version of the autocovariance, the autocorrelation:

Definition 11. *The autocorrelation function:*
The autocorrelation function of the stochastic process X_t is defined as:

$$\rho_s = \frac{E[(X_t - \mu)(X_{t-s} - \mu)]}{E[(X_t - \mu)^2]} = \frac{R_X(s)}{R_X(0)}$$

◦ It should be noted that nearly all of time-domain analysis (as opposed to frequency-domain analysis I spoke of early) is dependent in some way on the autocovariance (autocorrelation) function, since it is the autocorrelations that tell us how the series "repeats" over time. It is essentially the autocovariance (autocorrelation) that defines the "hidden" patterns in the data, at least in the time-domain.

◦ Remember, correlation is a number between -1 and 1. That is, $-1 \leq \rho_s \leq 1$, while covariance does not have such bounds.

◦ Note, that we can estimate the autocorrelation function the same way we estimate the covariance, by finite sample equivalents:

Definition 12. *Sample autocorrelation function:*
The sample autocorrelation function is defined as:

$$\hat{\rho}_s = \frac{\sum_{t=s+1}^{T}(x_t - \bar{x})(x_{t-s} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}$$

where $\bar{x}$ is the sample mean of X_t.

◦ In R, the command for the sample autocorrelation function is ACF().

◦ There is another version of the autocorrelation function that will also be helpful when we choose the type of ARIMA model we want to use. It is called the "partial autocorrelation" function and is described in the next definition below.

Definition 13. *Partial autocorrelation function:*
The partial autocorrelation function of the stochastic process X_t is defined as:

$$\rho_{\Phi,s} = \frac{E_\Phi[(X_t - E_\Phi[X_t])(X_{t-s} - E_\Phi[X_{t-s}])]}{E_\Phi[(X_t - E_\Phi[X_t])^2]}$$

where $\Phi = \{X_{t-1}, X_{t-2}, \ldots, X_{t-(s-1)}\}$ *denotes the intermediate values of* X_{t-j} *with* $1 \leq j \leq s-1$.

- So what the partial autocorrelation is doing is not as mysterious as it seems. The expression is like the "conditional" version of the autocorrelation expression above, where we are conditioning on this set Φ. What does Φ represent? It represents all the random variables "between" X_t and X_{t-s}. If we consider this fact, then we can see that $X_t - E_\Phi[X_t]$ is just the "residual" that remains after we subtract from X_t its conditional mean, conditioned on Φ.

- In R, the command for the sample partial autocorrelation function is PACF().

- **Some further comments...**

  - Finally, before we move on to the concept of white noise, let me make a brief comment on why stationarity is a useful assumption to make. Think about how we deal with estimating the properties of the theoretical model. We typically will take some finite sample of the random variable and then estimate sample moments, etc.
  - For example, say we want to establish the average age of the Canadian population. We might take a random sample from the cross-section of Canadian population. We can easily ask N people their age, and as long as the sample is truly random the average of this sample will converge to the "true" average age of the Canadian population.
  - The problem inherent with time-series, however, is how do we draw samples over and over at any given time t? Well we cannot do this the same way we do in a cross section since time is always moving forward and we can really only grab one sample at any given time. While there are ways around this, the easiest way to deal with the problem is just to assume that the series is stationary. That way, we can sample over and over from *each time* $t = 1, 2, \ldots, T$ and this entire sample of size T can then be used to estimate the moments of the distribution of X_t.

- **White noise process**

- Since we now have a tool at our disposal, the autocorrelation function, to quantify the nature of repeating patterns in the historical time-series, we can now talk about a particular type of time-series process which exhibits no patterns what so ever. This is what we call, white noise.

Definition 14. *Strict white noise (SWN):*
A strict white noise process X_t *is mean zero and i.i.d. (that is, independent and identically distributed).*

Definition 15. *Weak white noise (WWN):*
A weak white noise process X_t *is such that:*

(i) $E[X_t] = 0$
(ii) $E[(X_t - \mu)^2] = \sigma^2 = R_X(0) < \infty$
(iii) $E[(X_t - \mu)(X_{t-s} - \mu)] = R_X(s) = 0, \forall s \in \mathbb{Z}$

- That is, the weak white noise process exhibits an ACF (autocorrelation function) that spikes at $s = 0$ with a value of 1 and is zero everywhere else.

- Note, that a strict white noise is also a weak white noise, but not necessarily the other way around.
- The importance of a white noise should not be underestimated. What a strict white noise process implies is that the past can tell us nothing about the future. That is, knowing the past values of the series X_t (if X_t is a SWN) tell us nothing about its future values. Therefore, both the conditional and unconditional means are the same, $E[X_t] = 0$.
- Moreover, while a WWN may exhibit no autocorrelation, this **does not mean** the series is unpredictatable. In fact, since we have no idea if X_t (as a WWN) is actually independent (or i.i.d.) it is possible that the conditional density is not equal to the unconditional, that is $f(X_t|F_{t-1}) \neq f(X_t)$ necessarily.[3]
- Interestingly, the name "white noise" comes from engineering and the frequency-domain approach to time-series analysis. We say that something is "white noise" if it exhibits all of the frequencies of the spectrum equally.
  - That is, in the Fourier series we talked about earlier, all the "weights" a_ω and b_ω are the same.
  - The analogy comes from white light which includes all colours of the frequency spectrum in equal proportion.
  - If you recall, in the old days a scrambled "greyish" signal would appear on your television if the TV station went off-line and so we called this "white noise." Therefore, what we really meant is that the signal we were seeing on the television was completely random! That is, it exhibited no inherent patterns (i.e. autocorrelations) over time.
  - Conversely then if a signal is not white noise and does exhibit autocorrelations, $R_X(s)$, that are non-zero for some $s \neq 0$ then we call this "coloured noise."

- **The AR(p) and MA(q) models**
- Before talking about what the ARIMA model is, we first need to consider both the autoregressive process of order p (the AR(p) model) and the moving average process of order q (the MA(q) model). These are the standard building blocks of the ARIMA(p,d,q).

  Definition 16. *The AR(p) model:*
  A stochastic process X_t follows an autoregressive model of order p (an AR(p) model) if X_t can be expressed as:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \epsilon_t, \quad \textit{where} \quad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

- So, we are saying that if we assume that the theoretical model for X_t is AR(p), for some lag order p, the above expression defines the process. That is to say it defines its moments, of which we will now derive the first two.

[3] In fact, while the series is not *linearly* predictable, this does not mean it can't be predicted by a non-linear function. For example, the ARCH model, which we will talk about in the next section is WWN, but it can be predicted by a non-linear function.

- **First uncentered moment of AR(p)**:

  Proving the moments of the AR(p) is much easier when we assume that X_t is weakly stationary, so I will make that assumption below. Of course, stationarity implies that $E[X_t] = E[X_s]$ for all t and s, so we have:

$$\begin{aligned} E[X_t] &= E[\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \epsilon_t] \\ &= \alpha_1 E[X_{t-1}] + \alpha_2 E[X_{t-2}] + \cdots + \alpha_p E[X_{t-p}] + \mu_\epsilon \\ &= E[X_t] \sum_{j=1}^{p} \alpha_j + \mu_\epsilon \end{aligned} \tag{4a}$$

$$\Rightarrow E[X_t] = \frac{\mu_\epsilon}{1 - \sum_{j=1}^{p} \alpha_j} \quad \text{where} \quad \mu_\epsilon = 0 \tag{4b}$$

- Notice that we are deriving the *unconditional* 1st uncentered moment here.

- The *conditional* 1st uncentered moment is given as:

$$E[X_t | F_{t-1}] = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} \tag{5}$$

- where F_{t-1} is the information set we discussed above, that is $F_{t-1} = \{X_{t-1}, X_{t-2}, \ldots, X_0\}$. Notice that in the AR(p), the conditional 1st uncentered moment only depends on information up to X_{t-p}.

- **Second centered moment of AR(p)**:

  Again, we assume that X_t is stationary. Instead of going with $E[(X_t - E[X_t])^2]$ I will work with $Var(\cdot)$ which avoids having to deal with the square. Since the variance is an integral, it is a linear operator and passes through linear functions, except we square any constants in front of random variables.

$$\begin{aligned} Var[X_t] &= Var[\alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \epsilon_t] \\ &= \alpha_1^2 Var[X_{t-1}] + \alpha_2^2 Var[X_{t-2}] + \cdots + \alpha_p^2 Var[X_{t-p}] \\ &\quad + 2 \sum_{k=1}^{p} \sum_{j=k+1}^{p} \alpha_k \alpha_j Covar[X_{t-k} X_{t-j}] + \sigma_\epsilon^2 \\ &= Var[X_t] \sum_{j=1}^{p} \alpha_j^2 + 2 \sum_{k=1}^{p} \sum_{j=k+1}^{p} \alpha_k \alpha_j E[X_{t-k} X_{t-j}] + \sigma_\epsilon^2, \quad \text{since} \quad E[X_t] = 0 \end{aligned} \tag{6a}$$

$$\Rightarrow Var[X_t] = \frac{2 \sum_{k=1}^{p} \sum_{j=k+1}^{p} \alpha_k \alpha_j E[X_{t-k} X_{t-j}] + \sigma_\epsilon^2}{1 - \sum_{j=1}^{p} \alpha_j^2} \tag{6b}$$

- where the result follows from the fact that $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + ab Covar[X, Y]$.

- The above expression gives the *unconditional* variance of the AR(p) process.

- Notice that the $2 \sum_{k=1}^{p} \sum_{j=k+1}^{p} \alpha_k \alpha_j E[X_{t-k} X_{t-j}]$ term represents the sum of the autocovariance functions $R_X(s)$, and it is not clear at this point what the autocovariance of the AR(p) is.

- ◇ Needless to say, the unconditional variance, $R_X(0)$ of the AR(p) is a complicated expression to deal with.

- ○ However, if we assume instead an AR(1) model, then the unconditional variance is given as $R_X(0) = \sigma_\epsilon^2/(1-\alpha_1^2)$.

- ○ The *conditional* 2nd centered moment is given as:

$$Var[X_t|F_{t-1}] = \sigma_\epsilon^2 \tag{7}$$

since all the X_{t-1} to X_{t-p} values are conditioned on and have zero variance.

- ○ Therefore, all the conditional variance is generated through the residual ϵ_t.

- ○ **Summary**

- ○ So to summarize, we have established that the properties of the weakly stationary AR(1) model are such that:

  - ◇ the unconditional distribution of the AR(1) is $X_t \sim N(\mu_\epsilon/(1-\alpha_1), \sigma_\epsilon^2/(1-\alpha_1^2))$ where $\mu_\epsilon = 0$.
  - ◇ the conditional distribution is $X_t|F_{t-1} \sim N(\alpha_1 X_{t-1}, \sigma_\epsilon^2)$.
  - ◇ the distributions are Normal because a linear sum of Normal random variables is also Normal.

  I tried to derive the variance of the AR(p) if only to demonstrate how complicated these types of derivations can become.

Definition 17. *The MA(q) model:*
A stochastic process X_t follows a moving average model of order q (an MA(q) model), if X_t can be expressed as:

$$X_t = \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q} + \epsilon_t, \quad \textit{where} \quad \epsilon_t \sim N(0, \sigma_\epsilon^2)$$

- ○ So, again we are saying that if we assume that the theoretical model for X_t is MA(q), for some lag order q, the above expression defines the process. That is to say it defines its moments, of which we will now derive the first two.

- ○ **First uncentered moment of MA(q)**:

$$\begin{aligned} E[X_t] &= E[\beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q} + \epsilon_t] \\ &= \beta_1 E[\epsilon_{t-1}] + \beta_2 E[\epsilon_{t-2}] + \cdots + \beta_q E[\epsilon_{t-q}] + \mu_\epsilon \\ &= \mu_\epsilon \sum_{j=1}^{q} \beta_j + \mu_\epsilon \end{aligned} \tag{8a}$$

$$\Rightarrow E[X_t] = \mu_\epsilon(1 + \sum_{j=1}^{p} \beta_j) \quad \text{where} \quad \mu_\epsilon = 0 \tag{8b}$$

- Notice that we are deriving the *unconditional* 1st uncentered moment here.

- The *conditional* 1st uncentered moment is given as:

$$E[X_t|F_{t-1}] = \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q} \tag{9}$$

- where F_{t-1} is still the information set we discussed above, that is $F_{t-1} = \{X_{t-1}, X_{t-2}, \ldots, X_0\}$. Again notice that in the MA(q), the conditional 1st uncentered moment only depends on information up to X_{t-q}. The reason why we use a set of X_τ's as information here is because the X_τ's are functions of the ϵ_τ's by virtue of the MA(q) DGP. That is, if we know X_{t-1} say, then we also know ϵ_{t-1}, ϵ_{t-2}, up to ϵ_{t-q}.

- **Second centered moment of MA(q)**:

  Again, instead of working with $E[(X_t - E[X_t])^2]$ I will work with $Var(\cdot)$ which avoids having to deal with the square. Note that this time around, however, because the autocovariances $R_\epsilon(s)$ are zero for all values of $s \neq 0$ (that is, ϵ_t is a WWN [4]) it is much easier to derive the variance of the MA(q) than it was to try and derive it for the AR(p).

$$\begin{aligned} Var[X_t] &= Var[\beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q} + \epsilon_t] \\ &= \beta_1^2 Var[\epsilon_{t-1}] + \beta_2^2 Var[\epsilon_{t-2}] + \cdots + \beta_q^2 Var[\epsilon_{t-q}] + 2\sum_{k=1}^{q}\sum_{j=k+1}^{q}\beta_k\beta_j Covar[\epsilon_{t-k}\epsilon_{t-j}] + \sigma_\epsilon^2 \\ &= \sigma_\epsilon^2\sum_{j=1}^{q}\beta_j^2 + \sigma_\epsilon^2, \quad \text{since} \quad R_\epsilon(s) = 0 \quad \forall s \neq 0 \end{aligned} \tag{10a}$$

$$\Rightarrow Var[X_t] = \sigma_\epsilon^2(1 + \sum_{j=1}^{q}\beta_j^2) \tag{10b}$$

- The above expression gives the *unconditional* variance, $R_X(0)$, of the MA(q) process.

- So, if we assume instead an MA(1) model, then the unconditional variance is given as $R_X(0) = \sigma_\epsilon^2(1 + \beta_1^2)$.

- The *conditional* 2nd centered moment is given as:

$$Var[X_t|F_{t-1}] = \sigma_\epsilon^2 \tag{11}$$

  since all the ϵ_{t-1} to ϵ_{t-q} values are conditioned on and have zero variance.

- Therefore, all the conditional variance is generated through the residual ϵ_t.

- **Summary**

[4] In fact, it is also SWN, since if a variable Z_t is Normally distributed and uncorrelated, this implies it is independent or i.i.d.

- So to summarize, we have established that the properties of the weakly stationary MA(q) model are such that:
  - the unconditional distribution of the MA(q) is $X_t \sim N(\mu_\epsilon(1 + \sum_{j=1}^{p} \beta_j), \sigma_\epsilon^2(1 + \sum_{j=1}^{p} \beta_j^2))$ where $\mu_\epsilon = 0$.
  - the conditional distribution is $X_t|F_{t-1} \sim N(\beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q}, \sigma_\epsilon^2)$.
  - the distributions are Normal because a linear sum of Normal random variables is also Normal.

- **The AR(1) and MA(q) autocorrelation functions**

- So, to this point we have derived the 1st uncentered moments (the conditional and unconditional means) and 2nd centered moments (the conditional and unconditional variances) of both the AR(p) and MA(q) processes (assuming weak stationarity).

- We will now derive the autocorrelation functions for both the AR(1) and MA(q) processes. The reason why I am avoiding the AR(p) case is because as we noticed earlier, the lagged regressors in the AR(p) are correlated with each other (which isn't the case in the MA(q)), which makes it is difficult to work with.

- **Autocorrelation of the AR(1)**

- The derivation is as follows, were again we assume that the mean, μ_ϵ, of ϵ_t is zero (and so we have that the unconditional mean of X_t, $E[X_t] = 0$). First we have that:

$$E[X_t X_{t-s}] = R_X(s) = E[(\alpha_1 X_{t-1})(\alpha_1 X_{t-1-s})] \tag{12}$$

Moreover, if we take the AR(1) and recursively substitute in each lagged value of X_{t-1} over and over we get:

$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + \epsilon_t \\ &= \alpha_1(\alpha_1 X_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= \alpha_1^2 X_{t-2} + \alpha_1\epsilon_{t-1} + \epsilon_t \\ &= \alpha_1^2(\alpha_1 X_{t-3} + \epsilon_{t-2}) + \alpha_1\epsilon_{t-1} + \epsilon_t \end{aligned} \tag{13a}$$

And so on, until at the end when we reach $X_0 = \epsilon_0$ we get:

$$X_t = \alpha_1^t X_0 + \alpha_1^{t-1}\epsilon_{t-(t-1)} + \cdots + \alpha_1^2\epsilon_{t-2} + \alpha_1\epsilon_{t-1} + \epsilon_t \tag{14}$$

So subbing (14) into (12), and shifting forward in time s periods (the autocovariance is independent of t since we are assuming stationarity) we have:

$$E[X_{t+s}X_t] \tag{15a}$$

$$= E[(\alpha_1^{t+s} X_0 + \alpha_1^{t-1+s}\epsilon_{t-(t-1)} + \cdots + \alpha_1\epsilon_{t-1+s} + \epsilon_{t+s})(\alpha_1^t X_0 + \alpha_1^{t-1}\epsilon_{t-(t-1)} + \cdots + \alpha_1\epsilon_{t-1} + \epsilon_t)] \tag{15b}$$

So what we have done here is replaced the values $\alpha_1 X_{t-1}$ and $\alpha_1 X_{t-s}$ by the random variables ϵ_t. This makes it easier to take cross products and determine the autocovariance. Therefore, multiplying (15b) out we get:

$$\begin{aligned} E[X_{t+s}X_t] &= \sigma_\epsilon^2(\alpha_1^s + \alpha_1^{s+2} + \alpha_1^{s+4} + \cdots + \alpha_1^{s+2t}) \\ &= \sigma_\epsilon^2 \alpha_1^s \frac{1-\alpha_1^{2(t+1)}}{1-\alpha_1^2} \end{aligned} \tag{16a}$$

Keeping in mind that all the cross product expectations are zero since ϵ_t is WWN. It turns out that assuming that $|\alpha_1| < 1$ is enough to ensure stationarity of the AR(1) process (more on this later). Moreover, if we let X_0 trail off into the infinite past we get:

$$E[X_{t+s}X_t] = \sigma_\epsilon^2 \alpha_1^s \frac{1}{1-\alpha_1^2} \tag{17}$$

However, we recognize $\sigma_\epsilon^2 \frac{1}{1-\alpha_1^2} = R_X(0)$ and so we have that:

$$\rho_s = \frac{E[X_{t+s}X_t]}{E[X_t X_t]} = \frac{R_X(s)}{R_X(0)} = \alpha_1^s \tag{18}$$

- And so the autocorrelation function of the AR(1) declines geometrically to zero as $s \to \infty$ if $|\alpha_1| < 1$.

- **Autocorrelation of the MA(q)**

- Again, as before we write:

$$E[X_{t+s}X_t] = E[(\beta_1\epsilon_{t-1+s} + \beta_2\epsilon_{t-2+s} + \cdots + \beta_q\epsilon_{t-q+s} + \epsilon_{t+s})(\beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q} + \epsilon_t)] \tag{19}$$

So this time we do not need to recursively substitute anything which makes the job simpler. Again, expectation of cross products are zero and so we have:

$$E[X_{t+s}X_t] = \begin{cases} \sigma_\epsilon^2(\beta_s + \beta_{s+1}\beta_1 + \beta_{s+2}\beta_2 + \cdots + \beta_{s+q}\beta_q) & 0 \le s \le q \\ 0 & s > q \end{cases} \tag{20a}$$

So recalling that $R_X(0) = \sigma_\epsilon^2(1 + \sum_{j=1}^q \beta_j^2)$ and letting $\beta_0 = 1$ we have that:

$$\rho_s = \begin{cases} \sum_{j=0}^q \beta_{s+j}\beta_j / \sum_{j=0}^q \beta_j^2 & 0 \le s \le q \\ 0 & s > q \end{cases} \tag{21a}$$

Remember that β_{s+j} where $s + j > q$ are all zero!

- And so it is clear that the ACF of the MA(q) is zero for all $|s| > q$, since by symmetry we have that $R_X(s) = R_X(-s)$. That is for any real process X_t the ACF of both the AR(p) and MA(q) are symmetric about zero (they are *even* functions of s).

- **Lag polynomials**

- There is a nicer way to deal with the AR(p) and MA(q) processes which involves something call the *lag operator*.
- The lag operator is the operator L such that $LX_t = X_{t-1}$, $L^{-1}X_t = X_{t+1}$, and $L^jX_t = X_{t-j}$ (and $L^{-j}X_t = X_{t+j}$).
- Therefore we can write the AR(p) model as:

$$X_t\alpha(L) = \epsilon_t \tag{22}$$

  where $\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p$.
- And we can write the MA(q) model as:

$$X_t = \beta(L)\epsilon_t \tag{23}$$

  where $\beta(L) = 1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q$.
- **Stationary conditions and the lag polynomials**

  Definition 18. *Any lag polynomial can be factored as:*

$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p = (1 - \lambda_1 L)(1 - \lambda_2 L)(1 - \lambda_3 L)\dots(1 - \lambda_p L) \tag{24}$$

- **Example:**

  Take for example, the MA(3) process with lag polynomial $\beta(L) = 1 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3$.

  We have:

$$\begin{aligned}\beta(L) &= 1 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 \\ &= (1 - \lambda_1 L)(1 - \lambda_2 L)(1 - \lambda_3 L) \\ &= (1 - (\lambda_1 + \lambda_2)L + \lambda_1\lambda_2 L^2)(1 - \lambda_3 L) \\ &= 1 - (\lambda_1 + \lambda_2 + \lambda_3)L + ((\lambda_1 + \lambda_2)\lambda_3 + \lambda_1\lambda_2)L^2 - \lambda_1\lambda_2\lambda_3 L^3\end{aligned} \tag{25a}$$

  So we have that $\beta_1 = -(\lambda_1 + \lambda_2 + \lambda_3)$, $\beta_2 = ((\lambda_1 + \lambda_2)\lambda_3 + \lambda_1\lambda_2)$, and $\beta_3 = -\lambda_1\lambda_2\lambda_3$.

  Definition 19. *Discount factor representation of the geometric polynomial series:*
  Any geometrically declining polynomial can be represented as:

$$a/(1 - \lambda) = \textstyle\sum_{j=0}^{\infty} \lambda^j a = a(\lambda^0 + \lambda^1 + \lambda^2 + \dots)$$

  where $|\lambda| < 1$.

  That is, the above definition does not hold if $|\lambda| \geq 1$ since the polynomial will explode (that is, will approach ∞ as the λ^j terms just get bigger and bigger).

- Notice in the above definition that a is just some constant. If we allow a to be a random variable, say ϵ_t, then we can use this result to determine if the AR(p) or MA(q) processes are stationary.
- Note that we can write: $\epsilon_t/(1-\lambda L) = \sum_{j=0}^{\infty} \lambda^j L^j \epsilon_t = \sum_{j=0}^{\infty} \lambda^j \epsilon_{t-j}$.

Example: AR(1)

- We have $\alpha(L)X_t = (1-\alpha_1 L)X_t = \epsilon_t$.
- This implies that $X_t = \epsilon_t/(1-\alpha_1 L) = \sum_{j=0}^{\infty} \alpha_1^j L^j \epsilon_t = \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}$.
- It can be shown that if $|\alpha_1| < 1$ then the three conditions for stationarity from **Definition 10** all hold. You should try and think about this yourself; recall how we derived the 1st uncentered moment and 2nd centered moments for the AR(1) above (both unconditional moments of course).
- The actual proof is on pg.539-540 of Heij et. al.

Example: MA(q)

- We have $X_t = (1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q)\epsilon_t$.
- This implies that $X_t = \beta(L)\epsilon_t = \sum_{j=0}^{q} \beta_j L^j \epsilon_t = \sum_{j=0}^{q} \beta_j \epsilon_{t-j}$, where $\beta_0 = 1$.
- It is easy to see that if $q < \infty$, that is if q is a finite number, then the stationarity conditions always hold (as long as none of the β_j's are infinite!).
- If $q = \infty$, that is we have an MA(∞) model, then stationarity depends on if $\sum_{j=0}^{\infty} \beta_j^2$ converges.

Example: AR(p)

- Recall that we can always factorize the AR(p) polynomial as:

$$\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p = (1-\lambda_1 L)(1-\lambda_2 L)(1-\lambda_3 L)\ldots(1-\lambda_p L)$$

- This implies that $X_t = \epsilon_t / \prod_{k=1}^{p}(1-\lambda_k L)$.
- It can be shown, similar to how was done in the AR(1) case, that if *any* of the λ_k's are such that $|\lambda_k| \geq 1$ then the three conditions for stationarity are violated.
- If any of the $(1-\lambda_k L) = 0$ so that $\prod_{k=1}^{p}(1-\lambda_k L) = 0$, and we have that $|\lambda_k| < 1$, we must also have that $L = |1/\lambda_k| > 1$. [5] This is sometimes called the "root" of the lag polynomial, $1/\lambda_k$, being "outside the unit circle." (We reserve the unit circle terminology for the case where the roots are complex variables).
- That is, the roots of the lag polynomial being outside the unit circle imply that all the λ_k's are less than one in absolute value.
- See the Heij et.al. textbook Chapter 7, pg.539, for more details.

[5] Well L is really an operator so it can't take on a value $1/\lambda_k$ but we can treat it as a real valued variable for the sake of determining the properties of the lag polynomial.

- **Invertibility condition**

- You may have noticed something interesting when we represented the geometrically declining polynomial in its discounted series representation.

- Consider the AR(1) model $X_t\alpha(L) = X_t(1 - \alpha_1 L) = \epsilon_t$.

  - If $|\alpha_1| < 1$ we have $X_t = \epsilon_t/(1 - \alpha_1 L) = \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}$.
  - But by "inverting" the lag polynomial $\alpha(L)$ we have effectively created an MA(∞) from the AR(1)!
  - That is, we say that if $|\alpha_1| < 1$ then the AR(1) model is *invertible* into an MA(∞).
  - Likewise, given a similar condition on β_1 we have that the MA(1) is invertible into an AR(∞).

- Therefore, given our discussion above, the AR(1) models that are invertible are also stationary, however, stationary MA(q) models are not necessarily invertible! Confusing!

- The real importance of invertibility though is that it allows us to represent an MA(q) model as an AR(p) for forecasting purposes. That is, it is usually easier to forecast an AR(p) than an MA(q). More on this later.

- **ARIMA models**

- Finally! We have reached our goal. Despite the hype the ARIMA model is not so much different than the AR(p) or MA(q) models. In fact, it is simply a combination of the two.

- However, first we need to quickly discuss what an *integrated* process is.

Definition 20. *Integrated process of order 'd':*
An integrated process, X_t, of order 'd', is a process that must be differenced d times to be made stationary. We call these processes I(d) processes.

Definition 21. *Random walk:*
A random walk Z_t is a process such that $Z_t = \sum_{j=0}^{t} \epsilon_t$ where the ϵ_t's are i.i.d.

Corollary 6. *Random walk is an integrated process:*
A random walk is an $I(1)$ process since its first differences, $\Delta Z_t = \epsilon_t$ are stationary.

- And with those definitions at our disposal we can now define both the regular ARMA(p,q) and its generalization the ARIMA(p,d,q).

Definition 22. *ARMA(p,q) model:*
An autoregressive, moving average model of orders p and q (an ARMA(p,q) model) is such that:

$$\alpha(L)X_t = \beta(L)\epsilon_t$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, and $\alpha(L)$ is an AR(p) lag polynomial and $\beta(L)$ is an MA(q) lag polynomial.

- Notice that if $\alpha(L)$ is invertible, then we have that $X_t = \frac{\beta(L)}{\alpha(L)}\epsilon_t$ and so the ARMA(p,q) is invertible into an MA(l) model (we don't know the value of l yet!) and the same can be said if $\beta(L)$ is invertible, we can derive an AR(m) model.
- Therefore, all the analysis we used above to determine the moments of the AR(p) and MA(q) models can be reused here since we just replace the original lag polynomial with the new ratio of lag polynomials.
- Of course, we must solve for the values of the parameters in the new ratio lag polynomial and this is not always an easy task!

Example: ARMA(1,1)

- Say we have $X_t(1 - \alpha_1 L) = \epsilon_t(1 + \beta_1 L)$.
- This implies that if $|\alpha_1| < 1$ then we have $X_t = \dfrac{\beta(L)}{\alpha(L)}\epsilon_t = \dfrac{(1 + \beta_1 L)}{(1 - \alpha_1 L)}\epsilon_t$.
- Therefore, we have: $X_t = \displaystyle\sum_{j=0}^{\infty} \alpha_1^j L^j(\epsilon_t + \beta_1\epsilon_{t-1}) = \sum_{j=0}^{\infty} \gamma_j \epsilon_{t-j}$, where we can solve for the values of the γ_j parameters.
- So the ARMA(1,1) can be rewritten as an MA(∞). (Likewise, if $\beta(L)$ is invertible we can rewrite the ARMA(1,1) as an AR(∞)).
- Therefore we have the important fact that while the ARMA(1,1) only involves two parameters, α_1 and β_1, it "replicates" the behaviour of either an MA(∞) or an AR(∞). Therefore, it achieves what would have otherwise required an infinite amount of parameters for the price of only two!

Definition 23. *ARIMA(p,d,q) model:*
An autoregressive, moving average model of orders p and q, applied to the dth order integrated process (an ARIMA(p,d,q) model) is such that:

$$\alpha(L)\Delta^d X_t = \beta(L)\epsilon_t$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, and $\alpha(L)$ is an AR(p) lag polynomial and $\beta(L)$ is an MA(q) lag polynomial.

- So the ARIMA(p,d,q) is just an ARMA(p,q) applied to a series X_t that has been differenced d times.

Example: Random walk

- The random walk must be differenced once for it to be stationary. Therefore, the random walk has the ARIMA(0,1,0) form:

$$\alpha(L)\Delta Z_t = \beta(L)\epsilon_t$$

where $\alpha(L) = \beta(L) = 1$.

- **Model forecasts:**
- So now that we understand the AR(p), MA(q), and ARIMA(p,d,q) models, we might ask how we can use them to forecast a time series? That is, how might we use them to derive a "best guest" as to the future path of the time-series given what we know about the past.
- In order to judge the performance of forecasts, we need metric. Usually the metric chosen is the Minimum Mean Squared Error or MMSE.

  Definition 24. *The Mean Squared Error:*
  The MSE is defined as:

$$E[(X_t - \hat{X}_t)^2]$$

  where $\hat{X}_t$ is the forecasted value of X_t and we call $X_t - \hat{X}_t$ the forecast error.

  - ◇ It turns out that the MSE is really the variance of the forecast error.
  - ◇ So we want a forecast that minimizes the MSE, which is equivalent to minimizing the variance of the forecast error.
  - ◇ Remember, X_t is still a random variable! Moreover, so is $\hat{X}_t$ since our forecasts will be conditioned on time $t-1$ information, F_{t-1} (that is, we get a different forecast for any different past outcome).
  - ◇ So what we really want to do is minimize the squared forecast error *on average*.
- It turns out that in general, the forecast that minimizes the MSE is the *conditional 1st uncentered moment*, the conditional expectation (or conditional mean).
- That is, we choose our forecast to be $\hat{X}_t = E[X_t|F_{t-1}] = E_{t-1}[X_t]$.

- **Example: AR(1)**
- Suppose we wish to forecast the AR(1) model at time T, which is the last period in our sample of the past values $F_T = \{x_T, x_{T-1}, \ldots, x_0\}$.

- The MMSE forecast at time $T+1$ is therefore $\hat{X}_{T+1} = E_T[X_{T+1}] = E_T[\alpha_1 X_T + \epsilon_T] = \alpha_1 x_T$.
- The MMSE forecast at time $T+2$ is therefore $\hat{X}_{T+2} = E_T[X_{T+2}] = E_T[\alpha_1 X_{T+1} + \epsilon_{T+1}] = \alpha_1 \hat{X}_{T+1} = \alpha_1^2 x_T$.
  - ◇ Recall from pg.8 above that $E_T[E_{T+1}[X_{T+2}]] = E_T[X_{T+2}]$.
- And we can repeat this for any time period $T+\tau$ all the way until $\tau \to \infty$.

- It should become clear that we have for any period $T+\tau$, $\hat{X}_{T+\tau} = E_T[X_{T+\tau}] = \alpha_1^\tau x_T$.

- Therefore, we have that if the AR(1) process is stationary (that is, if $|\alpha_1| < 1$), then the forecast will converge to $E[X_t] = 0$ as $\tau \to \infty$. That is, for any stationary AR(1) model, the forecast (i.e. the expectation conditioned on time T information) will converge to the unconditional mean of the process, which in this case is 0.
- That is, as we forecast further and further out, the information we condition on at time T, F_T, becomes less and less informative, until eventually all we can say is that the best guess is the unconditional mean, $E[X_t] = 0$.
- It turns out this is a general result for all stationary AR(p), MA(q), and ARIMA(p,d,q) models.

- **The AR(1) forecast error:**
- Notice that we have that $X_{T+1} - \hat{X_{T+1}} = X_{T+1} - E_T[X_{T+1}] = X_{T+1} - \alpha_1 X_T = \epsilon_{T+1}$.
- So it turns out the forecast error at time $T+1$ is just ϵ_{T+1}.
- At time $T+2$ we have that, $X_{T+2} - \hat{X_{T+2}} = X_{T+2} - E_T[X_{T+2}] = X_{T+2} - \alpha_1 \hat{X}_{T+1} = \epsilon_{T+2} + \alpha_1 \epsilon_{T+1}$.
- And so the forecast error at time $T+2$ is $\epsilon_{T+2} + \alpha_1 \epsilon_{T+1}$.

- Therefore, repeating this process, we find that for any τ we have that the forecast error $X_{T+\tau} - \hat{X_{T+\tau}}$ is:

$$X_{T+\tau} - \hat{X_{T+\tau}} = \sum_{j=0}^{\tau-1} \alpha_1^j \epsilon_{T+\tau-j}$$

- **Comments:**
- See pg.81-89 of Enders book for more details and an example of the forecast functions for the ARMA(2,1).
- In general, the forecast functions of the AR(p), and ARMA(p,q) are more complicated than the AR(1).
- Notice that the variance of the AR(1) forecast error is increasing in the horizon τ. This is a general result that applies to MA(q) and ARMA(p,q) models as well.
  - ◇ That is, as we forecast further out, the uncertainty in the accuracy of our forecasts increases.
- Given an MA(q) model, we require knowledge of the innovations, or ϵ_t's in order to directly forecast. For example the one-step ahead forecast of the MA(q) is just the conditional mean (conditioned on time T information):

$$\hat{X}_{T+1} = \beta_1 \epsilon_T + \beta_2 \epsilon_{T-1} + \cdots + \beta_q \epsilon_{T-(q-1)}$$

  Therefore, if we do not know the innovations, and the MA(q) lag polynomial is invertible, it is often easier to simply transform the MA(q) into an AR(∞) and then cut off the lags at some large value, say z, and then forecast from the AR(z) instead since we have the values of the series x_τ, for $\tau = T, T-1, \ldots$ on hand directly.

- In practical cases, **you do not need to derive the forecast functions or forecast errors directly** since software does this for you. For example, in R, there are functions that estimate the ARIMA(p,d,q) model, provide the residuals (innnovations ϵ_t), and also forecast the model.

- Some good R libraries you should download include: *stats, tseries,* and *moments*.

- **The Box-Jenkins approach:**

- Box and Jenkins were the ones who popularized the use of the ARIMA(p,d,q) models back in the 1970's. At that time they suggested a very simplistic approach to modeling time-series that follows a three step procedure:

  1. First, the researcher should examine the time plot of the series both visually and using diagnostic tools such as the SACF (sample autocorrelation function) to try and *identify* the DGP of the model.
  2. Second, the researcher should *estimate* the parameters of the chosen model.
  3. Third, the researcher should use some form of *diagnostic checking* to determine how "well" the model both fits the data and forecasts out of sample.

- You should read the sections on the Box-Jenkins methodology in both Enders (Ch.2.8) and in Heij (Ch.7.2) for a more in depth discussion.

- **Identification stage:**

- Some important things to check during the identification stage are as follows:

  - Graph the time-series. Does it look stationary? What do the first-differences of the series look like? Does it look like there are any upward or downward trends that should be removed first?
  - Would some transformation of the time-series possibly make it stationary? E.g. taking logs?
  - Estimate the time-series moments over time: are they fixed?
  - Are there any outlier observations that may suggest an error in the data series?
  - Plot out the SACF or SPACF and try to determine what a reasonable ARIMA(p,d,q) specification may be based on their shapes.

- Assuming the model is stationary, the sample autocorrelation (SACF) can tell you a lot about which model to use.

- **Determining lag length with the SACF and PSACF functions:**

- Recall that we dicussed earlier the differences between the ACF and PACF functions. The PACF function is just the ACF where we have removed the effect of correlation between the variables X_t and X_{t-s}.

- Consider the MA(q) model. Since the explanatory variables $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_q$ are uncorrelated, there is no need to remove any correlation effect between them. Therefore, it is appropriate to look at their sample ACF (SACF).

- Below I have plotted examples of the SACF's and SPACF's of various series: for example, we have the SACF of a MA(3), the SPACF of an AR(2), and the SACF of the AR(2).

Figure 7: SACF of MA(3)

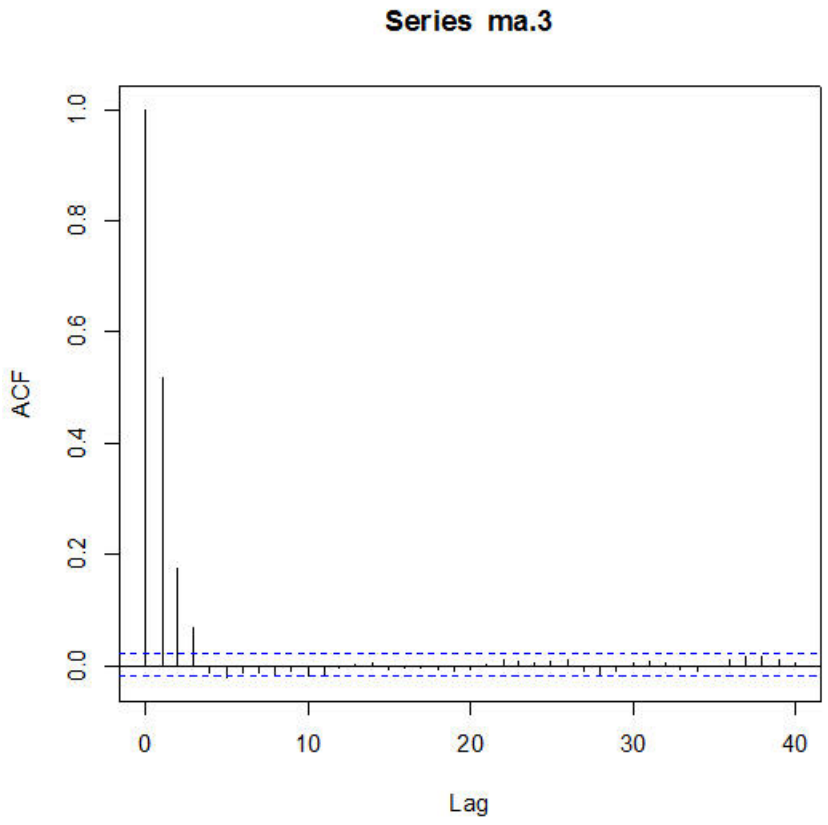


Figure 8: SPACF of AR(2)

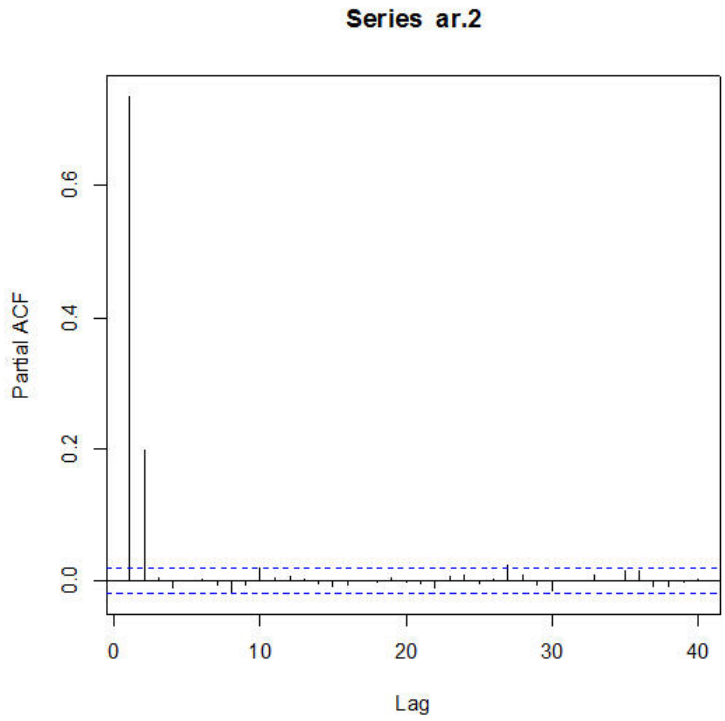
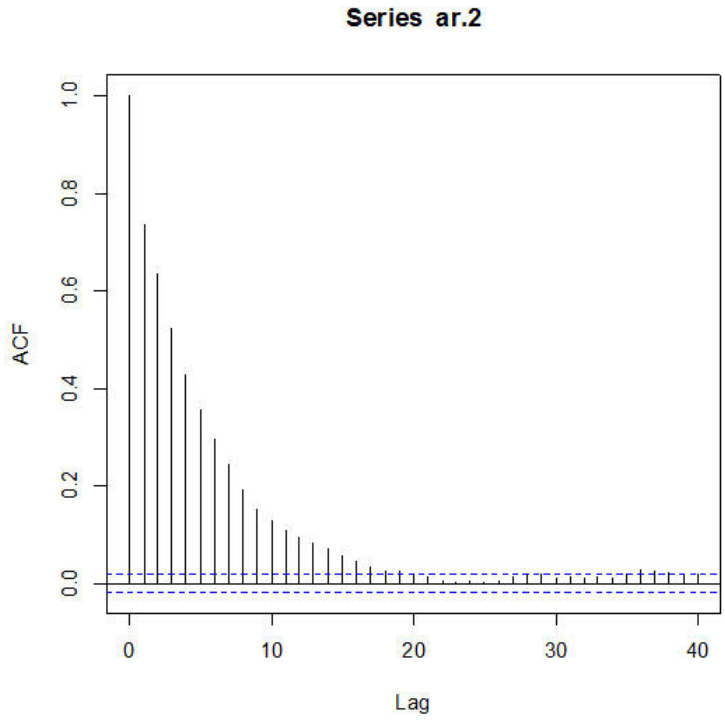


Figure 9: SACF of AR(2)



- Notice the odd feature of the R ACF function: It includes the lag value $s = 0$. However, the PACF starts at $s = 1$. Don't get mixed up!

- Also notice that the dotted blue lines are calculated automatically. They represent the region of statistical significance. Any vertical bars "outside" this region are statistically significant (that is, they most likely aren't due to randomness).

- Notice that for the MA(3) model the number of significant lags according to the SACF is 3. This is expected since the autocorrelation function of the MA(3) is 0 for all $s > q$. Recall our discussion above when we derived its expression.

- However, why did I use the SPACF for the AR(2) model? Well as we know from our discussion above on the autocorrelation function of the AR(1) model, the autocorrelation function dies off slowly. That is, the ACF of an AR(1) is $\rho_s = \alpha_1^s$. It turns out the same type of shape applies to the AR(2) model as well.

- This suggests that there exist autocorrelation *between* the values of the regressors in the AR(p) model (i.e. $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$). Since this is the case, it makes more sense to use the SPACF to determine lag length, since it accounts for this inter-value correlation.

- Therefore, when we apply the SPACF analysis to the AR(2) simulated data, we see that it suggests a lag length of 2 which is correct.

- Note, however, that since an ARMA(p,q) model is a mix of AR(p) and MA(q) models, SACF and SPACF analysis becomes more difficult.

- You should refer to pg.548-549 in Heij et. al. for more details.

- **Trends (side topic):**

  - The question of whether or not a trend exists and of what kind is complex. Generally we divide trends into two types: *deterministic* trends and *stochastic* trends.
  - Generally, if a trend exists then the DGP is non-stationary and direct application of ARIMA models can have misleading results since their theory assumes the series is stationary.
  - A deterministic trend in an AR(1), for example, can be represented as some function of time: $X_t = \gamma t + \alpha_1 X_{t-1} + \epsilon_t$. These types of trends can be removed by fitting $X_t = \beta t + u_t$ by OLS and then using the residuals as the new time-series in the AR(1).
  - An example of a stochastic trend is the *unit root* process: $X_t = X_{t-1} + \epsilon_t$ (that is, an AR(1) with $\alpha_1 = 1$). You may recognize this as the random walk we talked about earlier (hint: recursively substitute back to ϵ_0). Again, if we first difference the I(1) random walk, then we get a stationary I(0) process.
  - We must be very careful in choosing which trend type to use! Statistical tests have been developed to try and determine if a trend is deterministic or stochastic, for example the Dickey-Fuller unit root test.
  - However, keep in mind that trends themselves are only one type of non-stationarity, and other types exist (an obvious example is the explosive AR(1) with $|\alpha_1| > 1$). Another type of non-stationarity occurs if we assume that the parameters of our model depend themselves on time– for example, if we allow $\alpha_1(t)$ in our AR(1).
  - We won't deal much with trends in this course because the topic is quite vast. However, interested students are welcome to read section 7.3 in Heij et. al. for more details.

- **The estimation stage:**
- We won't talk too much about the estimation stage in this class. For the most part the R routines will handle this for you. However, you should know some basic facts.
- Consider the AR(p) model:
  - $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \epsilon_t$ where again, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$.
  - As we discussed, if this model is stationary, it is invertible into an MA(∞).
  - Therefore, from the MA(∞) representation, we can clearly see that the regressors of the AR(p) model, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, are uncorrelated with the residual, ϵ_t (since $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ are themselves functions of $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$).
  - We therefore say that the model satisfies the *orthogonality condition* and the AR(p) model can be estimated by OLS.
- However, now consider the MA(1) model:
  - $X_t = \epsilon_t + \beta_1 \epsilon_{t-1}$
  - If $|\beta_1| < 1$ then the model is invertible into an AR(∞).
  - However, the expression for the inverted model, the AR(∞) looks like this:

    $$X_t = \epsilon_t + \beta_1 X_{t-1} - \beta_1^2 X_{t-2} + \beta_1^3 X_{t-3} - \ldots$$

    (Try recursive substitution or see pg.543 in Heij et. al.)
  - But this model is a non-linear regression model and standard OLS cannot be applied. Therefore, a non-linear estimation method must be applied such as non-linear least squares (NLS) or maximum likelihood estimation (MLE).
  - This point will prove important later, when we talk about the difference between the ARCH and GARCH models.
- Since the ARMA(p,q) includes an MA(q) lag polynomial in its construction, the same non-linear estimation requirement applies to it as well.

- **The diagnostic stage:**
- The final stage in the Box-Jenkins methodology is the diagnostic stage.
- Once we have chosen a model (either an AR(p), MA(q), or ARMA(p,q)) in the identification stage, and have estimated it, we should then use some diagnostic tools to determine how well the model both "fits" the data and forecasts the future.
- In fact we should really try a number of different models and compare them according to the diagnostic tools.
- There exist a number of different diagnostic tools. Two of the most important include:

  1. model information criteria such as the AIC and SIC.
  2. tests of the model's residuals such as the Box-Pierce or Ljung-Box tests.

- **The information criteria of Akaike and Schwarz: the AIC and SIC**
- One way to determine how "well" a given model fits the data is to calculate an information criterion.
- These information criteria consider the tradeoff between how well the model fits versus how many parameters we include. Generally a model that fits well with less parameters is preferred.
- The **AIC criteria** (or Akaike Information Criterion) is given as:

$$AIC(k) = \ln s_k^2 + \frac{2k}{T}$$

where:

- k is the number of parameters (that is $k = p + q$ in an ARMA type model; note that information criteria can be used with many different parametric models!)
- s_k^2 is the maximum likelihood (MLE) estimate of the variance of the innovation, that is $\hat{\sigma_\epsilon^2}$.
- Finally, T is the sample size of our time-series.

- The **SIC criteria** (or Schwarz Information Criteria) is given as:

$$SIC(k) = \ln s_k^2 + \frac{k \ln T}{T}$$

- The SIC is sometimes also called the BIC or Bayes Information Criterion.

- Therefore, the **lower** is the information criterion the better!
- Note that R has built in functions that calculate the AIC and SIC.

- **The Box-Pierce and Ljung-Box tests**
- Another way to test the chosen models "fit" of the data is to consider the fact that if the model fits correctly it should "explain" most of the variation in the data.
- If a model explains all the variation in the data then the residuals of the model, the ϵ_t's, should be **white noise**.
  - ◇ That is, they should be *completely random* with no autocorrelation.
  - ◇ The reason being of course that if the model explains all the "hidden patterns" in the data, then the residuals should just represent the small random component that is unexplainable.
  - ◇ Therefore, if there is autocorrelation in the residuals then the model is missing something.
- Recall that the sample ACF function is given as:

$$\hat{\rho}_s = \frac{\sum_{t=s+1}^{T} \epsilon_t \epsilon_{t-s}}{\sum_{t=1}^{T} \epsilon_t^2}$$

since $\bar{\epsilon}_t = 0$.

- One way to test autocorrelation of the residuals is to use the SACF. If the residuals of the model are truly white noise then the SACF should should suggest that $\rho_0 = 1$ and $\rho_s = 0$ for all $s \neq 0$.
- Another way to test the white noise condition is to test if the sample autocorrelations, $\hat{\rho}_s$, are all zero *jointly*. That is, to test if they are all zero for $s \neq 0$ simultaneously.
- Two tests that do this are called the Box-Pierce and Ljung-Box test. They are both very similar and rely on the fact that the sum of the squared sample autocorrelations is distributed Chi-square.

- The **Box-Pierce test** statistic is given as:

$$BP(S) = T \sum_{s=1}^{S} \hat{\rho}_s^2$$

where $BP(S) \sim \chi^2(S)$ under the null hypothesis of white noise.

- Therefore, we reject white noise if the $BP(S)$ statistic is sufficiently large.
- Of course, we would like to test a large value of S, since this implies that the autocorrelations are jointly zero across more lags, but it is not reasonable to choose S too large in practical situations.

- The **Ljung-Box test** statistic is given as:

$$LB(S) = T \sum_{s=1}^{S} \frac{T+2}{T-s} \hat{\rho}_s^2$$

- And so the Ljung-Box test is just the Box-Pierce test but where the sample autocorrelations are weighted slightly differently, according to their lag values. That is, lags "further out" get more weight (since they are by construction estimated given a smaller sample pool).
- Finally, note that we must have that $S < T$ or else we cannot estimate.

- **Part 2: ARCH models**
- **Some mathematical tools:**
- Before we continue with the next section we need to briefly discuss some required math tools.
- Firstly we need a good definition of what exactly a financial "return" is.

Definition 25. *One-period simple return:*
We say that if an asset is held for one period from time $t-1$ then the simple gross return is defined as:

$$1 + R_t = P_t / P_{t-1}$$

where P_t is the price of the asset at time t.

Similarly, we also say that the simple net return is defined as:

$$R_t = (P_t - P_{t-1}) / P_{t-1}$$

Definition 26. *Multi-period simple return:*
We say that if an asset is held for k periods between dates $t-k$ and t, then the k-period simple gross return is defined as:

$$\begin{aligned} 1 + R_t(k) = P_t / P_{t-k} &= \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} \dots \frac{P_{t-(k-1)}}{P_{t-k}} \\ &= \prod_{j=0}^{k-1} (1 + R_{t-j}) \end{aligned}$$

Similarly, we also say that the k-period simple net return is defined as:

$$R_t(k) = (P_t - P_{t-k}) / P_{t-k}$$

Definition 27. *Continuously compounded return:*
The natural logarithm of the simple gross return of an asset is called the continuously compounded return, or log return:

$$r_t = \ln(1 + R_t) = \ln\left(\frac{P_t}{P_{t-1}}\right) = p_t - p_{t-1}$$

where $\ln P_t = p_t$.

Similarly, we also say that the k-period continuously compounded return is defined as:

$$r_t(k) = \ln(1 + R_t(k)) = \ln \prod_{j=0}^{k-1} (1 + R_{t-j}) = \sum_{j=0}^{k-1} r_{t-j} = p_t - p_{t-k}$$

- Therefore, dealing with continuously compounded returns is often *much* easier since the distribution of a sum of random variables is usually simpler to deal with than the distribution of their product.

- The next tool that may be useful is called the *change of variables* theorem. It allows us to talk about the distribution of some function of X_t, say $Y_t = f(X_t)$, in terms of the distribution of X_t.

Definition 28. *Change of variables:*
Let X_t be a continuous random variable with probability density function $f_X(X_t)$ defined over the support $c_1 < x_t < c_2$. Moreover, let $Y_t = u(X_t)$ be an invertible function of X_t with an inverse function $X_t = u^{-1}(Y_t)$. Then we have that the probability density function of Y_t is:

$$f_Y(Y_t) = f_X(u^{-1}(Y_t))|\frac{du^{-1}(Y_t)}{dY_t}|$$

defined over the support $u(c_1) < y_t < u(c_2)$.

- Note that the term on the RHS is the *absolute value* of the derivative of X_t with respect to Y_t but written in terms of Y_t.

Example:

- As an example, consider the bijective mapping $Y = X^2$ over the domain $0 < x < 1$ (we drop the time subscript for convenience). A bijective mapping means it is one-to-one and so it has an inverse.
- In this case the inverse function is $\sqrt{Y} = X$.
- Furthermore, suppose that X has p.d.f. $f_X(X) = 3X^2$.
- Taking the derivative dX/dY we have $\frac{1}{2}Y^{-1/2}$.
- Therefore using change of variables we have that $f_Y(Y) = 3\sqrt{Y}^2\frac{1}{2}Y^{-1/2} = \frac{3}{2}Y^{1/2}$.
- Note that if the mapping is *not* invertible (i.e. not one to one) more care has to be taken.

- Finally, the last tool is useful to know since it is widely encountered in finance. It is known as the *t-distribution* and defines the probability density function of a random variable with tails that are leptokurtic (that is, they are "fatter" than the Normal distribution).

Definition 29. *Standard t-distribution:*
We say that a random variable X_t has the standard t-distribution if its probability density function is defined as:

$$f_X(X_t) = \frac{1}{\sqrt{v\pi}}\frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{X_t^2}{v}\right)^{-\frac{v+1}{2}}$$

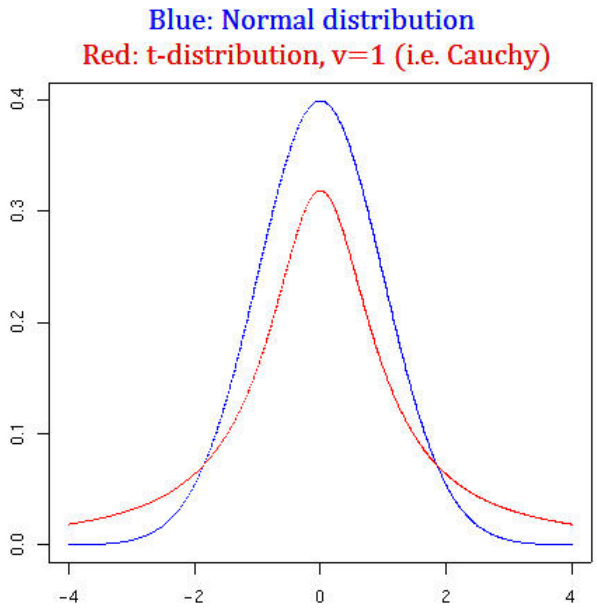
where $\Gamma(v) = \int_0^\infty t^{v-1}e^{-t}dt = (v-1)!$ *is the Gamma function and* $v > 0$ *is the degrees of freedom parameter (real valued).*

The distribution has:

- ◇ *mean and median of* 0 *if* $v > 1$*, and the mean is undefined if* $v = 1$.
- ◇ *variance of* $v/(v-2)$ *for* $v > 2$*,* ∞ *for* $1 < v \leq 2$*, and is undefined otherwise.*
- ◇ *skewness of* 0 *for* $v > 3$*, and is undefined otherwise.*
- ◇ *excess kurtosis of* $6/(v-4)$ *for* $v > 4$*, and* ∞ *for* $2 < v \leq 4$*, and is undefined otherwise.*

- Note that for $v = 1$ the t-distribution is the same as the Cauchy distribution which has "infinitely" fat tails. Moreover, as $v \to \infty$ the t-distribution converges on the Normal distribution.

Figure 10: "fat" tails



- So the closer is v to 1, the "fatter" the tails of the t-distribution. Fat tails mean that extremely large positive and negative values, while still rare, are drawn much more often than would be the case with the Normal distribution. See figure 10 above.
- In finance, the t-distribution is popular since it has been found that empirically, financial asset returns (e.g. stock price returns) have fatter tails than Normal. That is, extremely large positive or negative returns occur more often than under the Normal distribution (one-day stock market crashes, etc).

- **Stylized features of financial time-series**
- Since the early 1960's, a great deal of work has occurred in the field of empirical finance. Much of this work has been devoted to describing general characteristics observed in the data for financial asset returns. When a feature of the data is observed regularly and suggests a general pattern or behaviour, we call these features *stylized facts*.[6]
- What follows is a non-exhaustive list of the most common stylized facts observed in financial time series. Note that in the finance literature we often refer to the variance of returns as their *volatility* (although some authors use volatility to refer to the standard deviation of returns, we will apply it to the variance).

    1. The distribution of returns are leptokurtic (i.e. have fat tails). This implies that extreme, rare, events are more common than in the Normal distribution. This fact applies typically to the unconditional distribution of returns, although there is evidence it may also apply to the conditional distribution as well.

[6] One way of putting this is that we say that stylized facts are observed in the data in most cases or on average, although of course we might occasionally observe inconsistencies.

2. Asset returns are essentially random. That is they are white noise. Therefore, a reasonable model for the price level of an asset return is a random walk.
3. Returns are characterized by "bursts" of volatility. That is, volatility tends to "cluster" together over time. This is really the same as saying that the volatility process tends to follow an autocorrelated pattern and is **not** white noise. (We'll talk about what a volatility process is below).
    - ◇ Moreover, these volatility autocorrelations tend to die off very slowly. We say that the volatility has *long memory*.
    - ◇ Returns are less volatility in bull markets and more volatile in bear markets.
    - ◇ Trading volumes have memory the same way volatility does.
4. Past returns are negatively correlated with future volatility. This is sometimes called the *leverage effect*.
5. There are also other stylized features that describe the correlation *between* returns in multivariate models.

- ○ So these are probably the most common stylized features of financial time-series returns you will often encounter. In fact, the reason I bring them up is not only for their own sake but because it was Robert Engle in 1982 that derived a model for financial time-series that essentially captured the first three features in a very original and intuitive way.

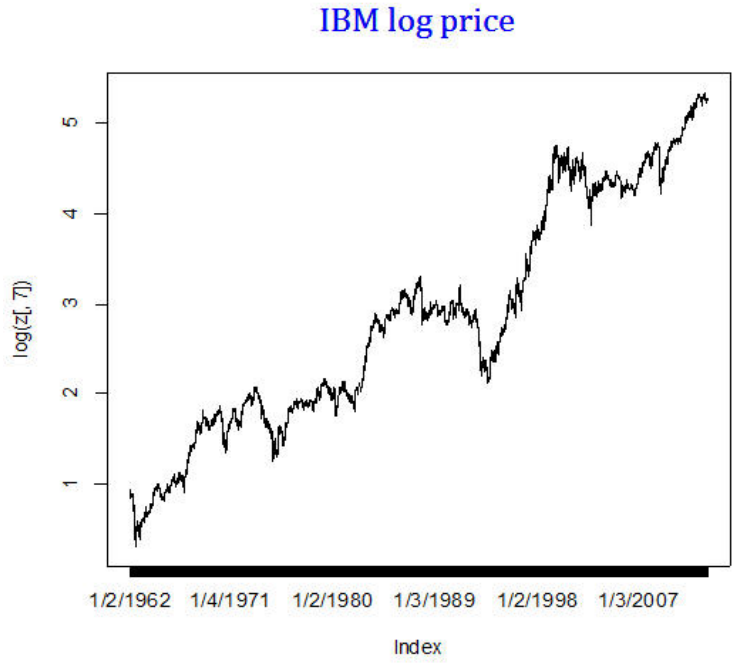
- ○ Engle called this model the ARCH(p) model which stands for "Autoregressive conditionally heteroskedasticity," which is just a fancy way of saying that the model captures the fact that the 2nd centered moment (i.e. the variance) is changing conditionally across time.

- ○ **Comments:**

- ○ However, before I reveal the ARCH(p) model, I'd like to make a few comments about the stylized features presented above.

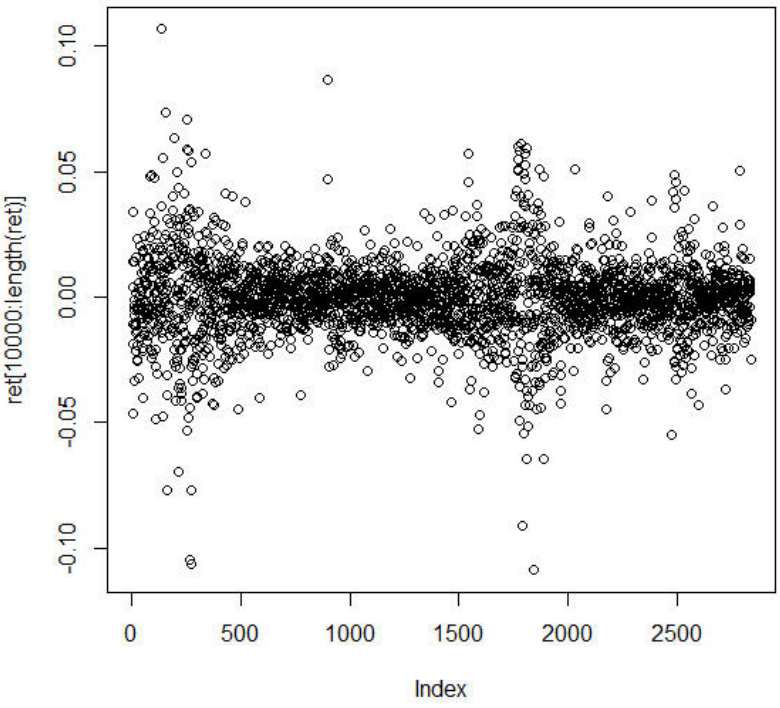
- ○ What follows is a plot of the IBM stock adjusted closing price (in natural logarithms) between 1962 and 2013 (**figure 11**).

Figure 11: IBM stock, adjusted daily closing price in logs, 1962-2013



- And in **figure 12** I have plotted the continuous compounded returns.

Figure 12: IBM continuously compounded returns, approximately last 3000 values of series

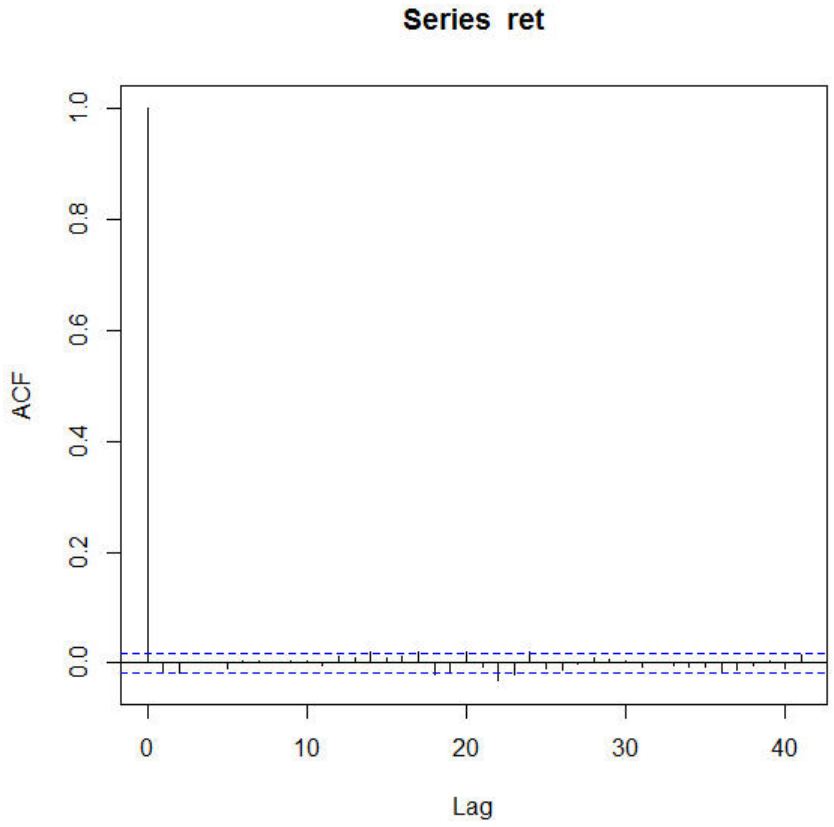


- Notice the **"volatility clustering"** in the returns series (**figure 12**). This suggests that the conditional volatility is changing across time.

- Moreover, if I was to take a histogram, say of the entire return series, it would suggest a shape with **"fatter" tails** than the Normal distribution. This suggests the unconditional distribution of returns is leptokurtic.

- Now let us focus on the ACF of the continuous compounded returns series (**figure 13**).

Figure 13: ACF of entire IBM returns series, approximately 13000 values

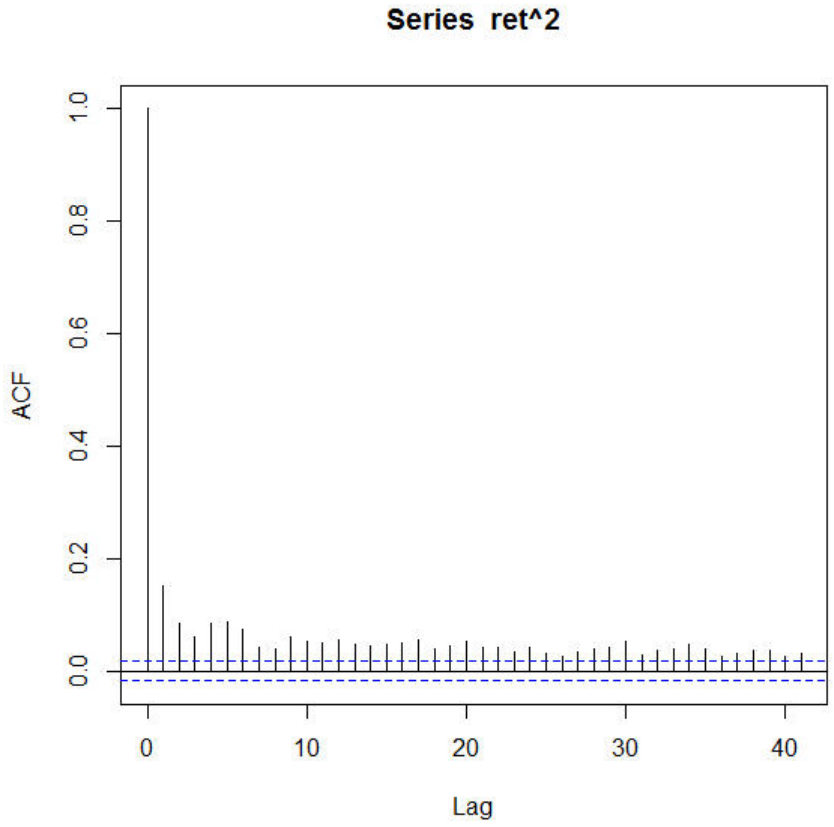


- Notice that it suggests it is **white noise**.

  - This suggests some validity to the EMH (or Efficient Markets Hypothesis) which is popular in Financial Economics. While we won't spend a lot of time on the EMH in this class, basically what it says is that since all traders incorporate all available information about prices into arbitrage opportunities, there should be very little if any information available left over to predict said prices. This suggests that prices are "informationally efficient" and should therefore follow a random walk (i.e. movements in prices should be completely unpredictable!).

- However, if we square the returns series (**figure 14**) and take the ACF again we get a completely different result!

Figure 14: ACF of entire IBM squared-returns series, approximately 13000 values



- We cannot discount the importance of this result.
- What it is telling us is that even though the prices themselves (i.e. returns) are completely unpredictable, the *volatility* (i.e variance) of these prices *is predictable*.
  - This has *huge* consequences for the pricing of anything whose price depends directly on the risk of an underlying asset (e.g. derivative products such as options).

- So let's take these two facts together (a) the unpredictability of returns and (b) the predictability of volatility and try and construct a reasonable model.

  1. First, if returns are unpredictable, we should have that the conditional mean of returns is equal to the unconditional mean of returns (i.e. the past tells us nothing) and we will assume that returns have a mean of zero. That is, we should have the following *levels* equation for our DGP:

     $$r_t = \epsilon_t \text{ where } \epsilon_t \sim WWN(0, \sigma_\epsilon^2), \text{ so that } E_{t-1}[r_t] = 0$$

     where r_t is our continuously compounded return.
  2. But let's also change the model above slightly so that the volatility σ_ϵ^2 depends, in some predictable fashion, on time. First let us rewrite the model so we can emphasize how σ_ϵ^2 now depends on time:

$$r_t = \sigma_{\epsilon,t} z_t = \epsilon_t \text{ where } z_t \sim i.i.d.(0,1)$$

Now, let us impose an AR(p) model not on the levels process, r_t, but on the 2nd conditional centered moment, $\sigma^2_{\epsilon,t}$. [7] That is, let:

$$r_t^2 = \sigma^2_{\epsilon,t} z_t^2 = (\alpha_0 + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_p r^2_{t-p}) z_t^2$$

(Notice that I have included a constant term in the AR(p) formulation. This is important and the reason will be given later).
The above expression implies that we also have:

$$\sigma^2_{\epsilon,t} = E_{t-1}[(r_t - E_{t-1}[r_t])^2] = E_{t-1}[r_t^2] = \alpha_0 + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_p r^2_{t-p}$$

Since $E_{t-1}[z_t^2] = 1$. Equivalently we have:

$$r_t^2 = \alpha_0 + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_p r^2_{t-p} + u_t$$

where $u_t \sim WWN(0, \sigma_u^2)$, and $u_t = r_t^2 - \sigma^2_{\epsilon,t}$.
And so the squared returns follow an AR(p) process! That is, we have found a perfectly suitable model to capture the autocorrelation exhibited in the squared returns, r_t^2, even though returns themselves, r_t, are unpredictable.

- Notice what the above suggests.
  - Returns *cannot* be a strong white noise since clearly $f(r_t|F_{t-1}) \neq f(r_t)$; that is, the unconditional distribution of returns is not equal to the conditional. Returns *do* in some way depend on the past, just not through the first moment of the distribution.
  - However, returns clearly *are* a weak white noise.
  - Moreover, the returns process is stationary, since the unconditional mean and variance are constant and finite (we'll show this next) and the autocovariance function only depends on the lag s.
  - Finally the model is now non-linear in the "basic" stochastic component, z_t. That is, if we were to substitute back in time recursively given $r_t = \sigma_{\epsilon,t} z_t = \sqrt{\alpha_0 + \alpha_1 r^2_{t-1} + \cdots + \alpha_p r^2_{t-p}} z_t$, we would have a highly non-linear function of z_t.

- So, given this broad discussion you may now be wondering what the point of it was. Well, it turns out what we have actually done is derive the ARCH(p) model!

- **The ARCH(p) model**

Definition 30. *The ARCH(p) model:*
The autoregressive conditionally heteroskedastic model of order p, (the ARCH(p) model) is given as follows:

$$r_t = \sigma_t z_t, \quad \textit{where} \quad z_t \sim N(0,1) \tag{26a}$$

$$E_{t-1}[r_t^2] = \sigma_t^2 = \alpha_0 + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_p r^2_{t-p} \tag{26b}$$

[7] Which in this case is really the 2nd uncentered moment since the conditional mean is zero.

Or sometimes it is put another, equivalent, way:

$$r_t = \epsilon_t, \quad \textit{where} \quad \epsilon_t | F_{t-1} \sim N(0, \sigma^2_{\epsilon,t}) \tag{27a}$$

$$\textit{and} \quad \sigma^2_{\epsilon,t} = \alpha_0 + \alpha_1 r^2_{t-1} + \alpha_2 r^2_{t-2} + \cdots + \alpha_p r^2_{t-p} \tag{27b}$$

$$\textit{with information set} \quad F_{t-1} = \{r^2_{t-1}, r^2_{t-2}, \dots\}$$

where in both cases $\alpha_0 > 0$.

- However, it is *very important* to note that, more generally, equations (26b) and (27b) are really describing the conditional moments of the *squared innovations* of the returns level process, not necessarily the squared-returns process, r^2_t, itself, although we have used that example up until now since we assumed the conditional mean of returns is zero.
  - Take for example the model $r_t = \mu + \epsilon_t = \mu + \sigma_t z_t$, where we have included the constant μ in the returns levels process. Of course, since z_t is i.i.d, μ is *both* the conditional and unconditional mean of r_t.
  - The corresponding ARCH(p) conditional variance equation is now:

$$E_{t-1}[(r_t - \mu)^2] = E_{t-1}[\epsilon^2_t] = \sigma^2_t = \alpha_0 + \alpha_1 \epsilon^2_{t-1} + \alpha_2 \epsilon^2_{t-2} + \cdots + \alpha_p \epsilon^2_{t-p}$$

- **The 1st and 2nd moments of the ARCH(1) model**
- First, recall from above that the squared returns in an ARCH(1) model can be written as:

$$r^2_t = \alpha_0 + \alpha_1 r^2_{t-1} + u_t$$

where $u_t \sim WWN(0, \sigma^2_u)$, and $u_t = r^2_t - \sigma^2_{\epsilon,t}$.

- Therefore, as per our discussion of the AR(1) model, if $|\alpha_1| < 1$ then the squared-returns process is stationary.
- That is, we have that:
  - $E[r^2_t] = \alpha_0/(1 - \alpha_1) < \infty$. That is, the unconditional mean of the squared-returns process exists and is finite.
    - But notice that the unconditional mean of the squared-returns process *is really just the unconditional variance of the returns process* (since the unconditional mean of the returns process is zero).
    - This is why we included a constant term $\alpha_0 > 0$; we don't want the unconditional variance to be equal to zero!

- Therefore, we can show that if $|\alpha_1| < 1$ then the ARCH(1) model for r_t is both weak stationary and a weak white noise.
- That is,
  - The unconditional 1st uncentered moment (the mean) is:

$$E[r_t] = 0 < \infty, \text{ since } E_{t-1}[r_t] = E_{t-1}[\sigma_t z_t] = \sigma_t E_{t-1}[z_t] = 0, \forall t$$

That is the unconditional mean is fixed at zero, since the conditional mean is zero for all values of t.

- ◇ The unconditional 2nd centered moment (the variance) is $E[r_t^2] = \alpha_0/(1-\alpha_1) < \infty$ as per above.
- ◇ The autocovariance function is:

$$R_r(s) = \begin{cases} \alpha_0/(1-\alpha_1) & s = 0 \\ 0 & s \neq 0 \end{cases} \text{ since } E[(\sigma_t z_t)(\sigma_{t-s} z_{t-s})] = 0 \text{ for all } s \neq 0,$$

since the z_t's are i.i.d.

- ○ **The 4th moment of the ARCH(1) model:**
- ○ So, so far we have shown that the ARCH(1) model satisfies the second and third stylized features of financial asset returns discussed above.
- ○ That is the ARCH(1) model implies that:
  - ◇ The returns proces r_t is a weak white noise and this implies that the price process is I(1) integrated (i.e. random walk).
  - ◇ The conditional volatility formulation captures the "volatility clustering" effects. That is, it captures the autocorrelated nature of the squared-returns process r_t^2.
- ○ Both of these results can also be shown to hold for the ARCH(p) model as well.
- ○ However, we have so far said nothing about the first stylized feature, or that of the "fat tailed" nature of financial asset return distributions.
- ○ Well as I mentioned earlier, the ARCH(p) model solves this problem as well since even though the conditional distribution of returns is Normal, the unconditional distribution of returns is fat tailed.
- ○ Let's show this, first by demonstrating that the conditional distribution of returns is Normal, which is straightforward:
  - ◇ From equation (34a) we have $r_t = \sigma_{\epsilon,t} z_t$ where $z_t \sim N(0,1)$. Therefore, this implies that $r_t | F_{t-1} \sim N(0, \sigma_{\epsilon,t}^2)$ which proves the result. Note that this is a direct application of the *change of variables* formula.
- ○ However, showing that the unconditional distribution of returns is **not** Normal is somewhat more complicated. Basically it relies on showing that the 4th centered unconditional moment (i.e. the Kurtosis) of r_t under the ARCH(p) model is greater than 3 (or in other words, the excess kurtosis is strictly positive).
- ○ The proof for the ARCH(1) case is as follows. Do not worry if you don't immediately understand all of it; other than the co-integration material, this is probably the most difficult proof you will see in this course. However, it is worthwhile to understand where the result is coming from.

- The proof involves two "tricks." The first one uses the fact that the conditional distribution of returns is Normal. Therefore, we have that:

$$K_{t-1} = \frac{E_{t-1}[(r_t - E_{t-1}[r_t])^4]}{E_{t-1}[(r_t - E_{t-1}[r_t])^2]^2} = \frac{E_{t-1}[r_t^4]}{E_{t-1}[r_t^2]^2} = 3 \tag{28a}$$

$$\Rightarrow E_{t-1}[r_t^4] = 3E_{t-1}[r_t^2]^2 = 3(\alpha_0 + \alpha_1 r_{t-1}^2)^2 \tag{28b}$$

  where K_{t-1} is the conditional Kurtosis, conditioned on time $t-1$ information. Since Normal distributions have Kurtosis of 3, the result follows.

- The second trick is to use the law of iterated expectations, **definition 8**:

$$E[r_t^4] = E[E_{t-1}[r_t^4]] = 3E[\alpha_0 + \alpha_1 r_{t-1}^2]^2 = 3E[\alpha_0^2 + 2\alpha_0\alpha_1 r_{t-1}^2 + \alpha_1^2 r_{t-1}^4] \tag{29a}$$

- Now, let $\mu_{4,t}$ be the 4th uncentered moment at time t. We have that:

$$\mu_{4,t} = 3(\alpha_0^2 + 2\alpha_0\alpha_1 E[r_{t-1}^2] + \alpha_1^2 \mu_{4,t-1}) \tag{30a}$$

- This is just a first-order difference equation in $\mu_{4,t}$. We know that $E[r_{t-1}^2] = \alpha_0/(1-\alpha_1)$, and so we just treat it as a constant:

$$(1 - 3\alpha_1^2 L)\mu_{4,t} = \mu_{4,t} - 3\alpha_1^2 \mu_{4,t-1} = 3(\alpha_0^2 + 2\alpha_0\alpha_1 (\alpha_0/(1-\alpha_1))) \tag{31a}$$

  Which suggests that a stationary solution for the 4th moment exists if $|3\alpha_1^2| < 1$ which implies that $|\alpha_1^2| < 1/3$. That is, as $|\alpha_1^2| \to 1/3$ the Kurtosis of the ARCH(1) model goes to infinity.

- Finally, if we have a stationary Kurtosis, we have:

$$(1 - 3\alpha_1^2 L)\mu_{4,t} = 3(\alpha_0^2 + 2\alpha_0\alpha_1 (\alpha_0/(1-\alpha_1))) \tag{32a}$$

$$\Rightarrow \mu_4 = \frac{3(\alpha_0^2 + 2\alpha_0\alpha_1 (\alpha_0/(1-\alpha_1)))}{(1 - 3\alpha_1^2)} \tag{32b}$$

  And it can therefore be shown (through algebraic manipulation) that the unconditional Kurtosis is greater than 3:

$$\Rightarrow K = \frac{\mu_4}{\mu_2^2} = \frac{3(\alpha_0^2 + 2\alpha_0\alpha_1 (\alpha_0/(1-\alpha_1)))}{(1 - 3\alpha_1^2)} \frac{(1-\alpha_1)^2}{\alpha_0^2} > 3 \tag{33a}$$

- **The GARCH(p,q) model**

- So now that we have the ARCH(p) model figured out, the GARCH(p,q) is not too difficult to understand since it is just a generalization of the ARCH(p) model the same way the ARMA(p,q) model is a generalization of the AR(p) model.

  Definition 31. *The GARCH(p,q) model:*
  The generalized autoregressive conditionally heteroskedastic model of orders p and q, (the GARCH(p,q) model) is given as:

$$r_t = \sigma_t z_t, \quad \textit{where} \quad z_t \sim N(0,1) \tag{34a}$$

$$E_{t-1}[r_t^2] = \sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \alpha_2 r_{t-2}^2 + \cdots + \alpha_p r_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \cdots + \beta_q \sigma_{t-q}^2 \tag{34b}$$

Or again, put the other, equivalent, way:

$$r_t = \epsilon_t, \quad \textit{where} \quad \epsilon_t | F_{t-1} \sim N(0, \sigma^2_{\epsilon,t}) \tag{35a}$$

$$\textit{and} \quad \sigma^2_{\epsilon,t} = \alpha_0 + \sum_{j=1}^{p} \alpha_j r^2_{t-j} + \sum_{j=1}^{q} \beta_j \sigma^2_{\epsilon,t-j} \tag{35b}$$

$$\textit{with information set} \quad F_{t-1} = \{r^2_{t-1}, r^2_{t-2}, \dots\}$$

where in both cases $\alpha_0 > 0$.

- Notice that all we have done is added the lagged values of $\sigma^2_{\epsilon,t}$ to the conditional variance equation, (34b).
  - ◇ However, this small change provides us a lot of freedom in specifying a very parsimonious model with few parameters just as an ARMA(p,q) frees us to do so as opposed to large lag AR(p)'s or MA(q)'s.
  - ◇ How is this so? Recall from our previous conversation that an ARMA(1,1) can be represented as either an AR(∞) or an MA(∞) by moving the lag polynomials around. (See **Example: ARMA(1,1)** under **definition 22** above).
- Notice that just as we could write the ARCH(p) model as an AR(p) on the squared-returns, $r^2_t = \sum_{j=1}^{p} \alpha_j r^2_{t-j} + u_t$, we can also write the GARCH(p,q) model as an ARMA(p,q) on the squared-returns.

- **Example: GARCH(1,1)**
- Take for example the GARCH(1,1). It can be shown that $\sigma^2_{\epsilon,t}$ can be rewritten as:

$$r^2_t = \alpha_0 + (\alpha_1 + \beta_1) r^2_{t-1} - \beta_1 \eta_{t-1} + \eta_t$$

where $\eta_t = r^2_t - \sigma^2_{\epsilon,t}$, which is an ARMA(1,1) on the squared-returns.

- Immediately we can see what the condition for stationarity of the GARCH(1,1) is, by writing the AR component as a lag polynomial:

$$\alpha(L) r^2_t = (1 - (\alpha_1 + \beta_1)L) r^2_t = \alpha_0 - \beta_1 \eta_{t-1} + \eta_t$$

and so in order to invert $\alpha(L)$ we require that $|(\alpha_1 + \beta_1)| < 1$.

- And so the unconditional variance of r_t is $E[r^2_t] = \alpha_0 / (1 - (\alpha_1 + \beta_1))$ (which is the same as the unconditional mean of the squared-returns process).

- **Estimation of the ARCH and GARCH models**
- Recall from our earlier discussion on the estimation stage of the Box-Jenkins approach to time-series analysis, that the AR(p) and MA(q) models (and thus the ARMA(p,q)) required different estimation strategies.

- ◇ At that point I explained that the while the AR(p) could be estimated by OLS, the MA(q) (and ARMA(p,q)) required a non-linear estimation technique such as maximum likelihood estimation (MLE).
- ◇ Furthermore, we also later discussed how an ARCH(p) model is like an AR(p) model on the squared returns (or squared innovations if the mean is not zero), and a GARCH(p,q) is like an ARMA(p,q) model imposed on the squared returns (or squared innovations).

○ Therefore, since I would like you to avoid using pre-programmed R routines to estimate the GARCH model on the 2nd assignment, we will now briefly review maximum likelihood estimation, since the estimation of the GARCH model parameters requires a non-linear technique such as MLE, just as estimation of an ARMA(p,q) does.

○ **Maximum likelihood estimation (MLE)**

○ Recall from **definition 4** that Bayes Theorem implies that:

$$f_{X|Y}(x|y) = f_{Y|X}(y|x)f_X(x)/f_Y(y)$$

○ While typically X and Y represent the random variables of our model, we can also think of the parameters of the GARCH(p,q) model as random variables too (in fact this really forms the basis of the "Bayesian approach" to statistics).

○ That is, we assume that the parameters of the GARCH(p,q) model have their own probability distributions. If this is the case, we can rewrite the above as:

$$f_{\Theta|Y}(\theta|y) = f_{Y|\Theta}(y|\theta)f_\Theta(\theta)/f_Y(y)$$

where $\theta = \{\alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q\}$ is the set of parameters, and $y = \{y_T, y_{T-1}, \ldots, y_1\}$ is the observed sample data series we have on hand (that is, y_t is the return at time t, or $y_t \equiv r_t$.)

○ Let us now focus on the term $f_{Y|\Theta}(y|\theta)$.

- ◇ Clearly since this is a pdf, if we integrate across all values of Y the integral must be 1. That is, the area under the pdf must equal 1.
- ◇ However, if we consider $f_{Y|\Theta}(y|\theta)$ rather as a function of θ, with Y held constant, we call this the *likelihood function* and we write it as $L_{\Theta|Y}(\theta|y)$.
- ◇ Of course now the likelihood function will not integrate to 1 across all values of θ but this is unimportant.
- ◇ What is important is that the maximum likelihood method of estimation basically says that if we maximize $L_{\Theta|Y}(\theta|y)$ with respect to θ, this will give us a reasonable estimate of the parameters of the model.

○ Just for reference, the Bayesian approach goes further and says we should actually specify (i.e. assume) some $f_\Theta(\theta)$ which is called the *prior* density, and then maximize the RHS of $f_{\Theta|Y}(\theta|y) \propto f_{Y|\Theta}(y|\theta)f_\Theta(\theta)$, where the LHS is called the *posterior* density (notice that the RHS is proportional to the LHS so they have the same max).

- ◇ That is, Bayesians believe that the *posterior* density better represents the "true" pdf of the parameters conditional on the data we observe, y.

- Of course the debate between frequentists who rely on maximizing $L_{\Theta|Y}(\theta|y)$ and Bayesians who maximize $f_{Y|\Theta}(y|\theta)f_{\Theta}(\theta)$ rages on, but for this class we will just maximize the likelihood (although you are welcome to maximize the posterior if you feel more Bayesian!)

- Either way, we now need to find some expression of the likelihood that we can maximize.

  - ◇ For complicated models like the GARCH(p,q) we will not be able to solve for an analytical solution (that is a solution that gives a close-form formula for the answer).
  - ◇ Rather we will solve the maximum of the likelihood function *numerically* which basically means using a computer program like R to do the hard work.
  - ◇ When you used the ARIMA() command in R for the 1st assignment, it was really solving for the parameter estimates according to a numerical optimization routine of the type we are about to talk about.

- An expression for the likelihood can be found by appealing to the result we discussed way back in **Corollary 5**. That is, we can factorize the joint likelihood into a product of conditional densities and then take logs. By doing so we turn the product into a sum which is easier to maximize.

  - ◇ Of course, since taking logs is a monotonic transformation the maximum of the log-likelihood is the same as that of the likelihood function.

- **Example: The GARCH(1,1) log-likelihood**

- Therefore, let us take the GARCH(1,1) model as an example:

- The likelihood function of the joint set of returns $r = \{r_T, r_{T-1}, \ldots, r_1\}$ is given by $L_{\Theta|R}(\theta|r)$. Of course this joint density can be factorized as:

$$\begin{aligned} L_{\Theta|R}(\theta|r) = f_{R|\Theta}(r|\theta) &= f_{R|\Theta}(r_T, r_{T-1}, \ldots, r_1|\theta) \\ &= f(r_T|F_{T-1}, \theta)f(r_{T-1}|F_{T-2}, \theta)\ldots f(r_2|F_1, \theta)f(r_1|\theta) \end{aligned} \tag{36a}$$

where as usual, $F_{t-1} = \{r_{t-1}, r_{t-2}, \ldots, r_1\}$.

Furthermore, we can now take logs:

$$\begin{aligned} LL_{\Theta|R}(\theta|r) &= \ln f_{R|\Theta}(r|\theta) = \ln f_{R|\Theta}(r_T, r_{T-1}, \ldots, r_1|\theta) \\ &= \ln f(r_T|F_{T-1}, \theta)f(r_{T-1}|F_{T-2}, \theta)\ldots f(r_2|F_1, \theta)f(r_1|\theta) \\ &= \sum_{t=2}^{T} \ln f(r_t|F_{t-1}, \theta) + \ln f(r_1|\theta) \end{aligned} \tag{37a}$$

So, since the GARCH(1,1) model is conditionally Normal we have:

$$\begin{aligned} LL_{\Theta|R}(\theta|r) &= \sum_{t=2}^{T} \ln f(r_t|F_{t-1},\theta) + \ln f(r_1|\theta) \\ &= -\frac{T-1}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\ln \sigma^2_{\epsilon,t} - \frac{1}{2}\sum_{t=2}^{T}\frac{r_t^2}{\sigma^2_{\epsilon,t}} + \ln f(r_1|\theta) \\ &= -\frac{T-1}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\ln\left(\alpha_0 + \alpha_1 r^2_{t-1} + \beta_1\sigma^2_{\epsilon,t-1}\right) - \frac{1}{2}\sum_{t=2}^{T}\frac{r_t^2}{\alpha_0 + \alpha_1 r^2_{t-1} + \beta_1\sigma^2_{\epsilon,t-1}} + \ln f(r_1|\theta) \end{aligned} \tag{38a}$$

where:

$$\ln f(r_1|\theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2_\epsilon - \frac{1}{2}\frac{r_1^2}{\sigma^2_\epsilon}$$
$$\text{and } \sigma^2_\epsilon = \frac{\alpha_0}{1-(\alpha_1+\beta_1)}$$

is the unconditional mean of the GARCH(1,1) process.

- See also Heij et. al. pg. 626-627 for more details.

- **Numerical optimizer**

- In R there are a few different numerical optimizer functions, for example *nlm()* and *optim()*.

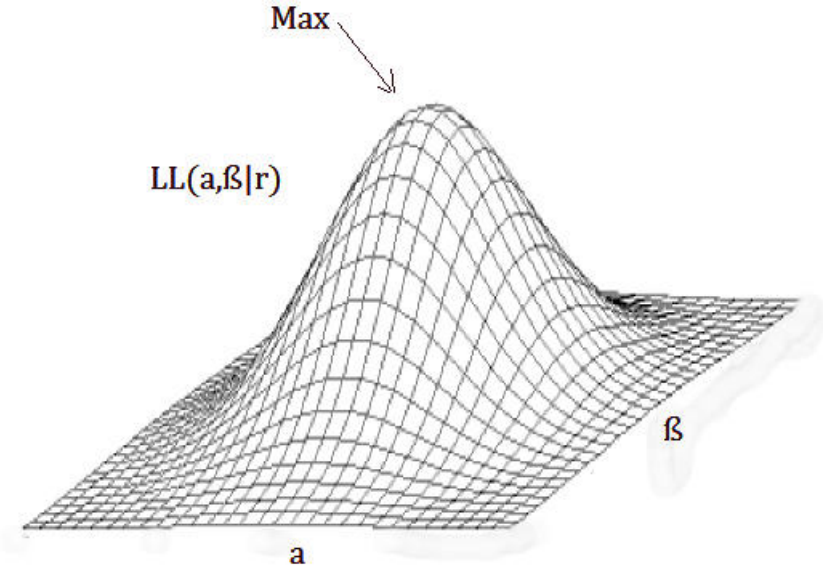
- Both routines can use the BFGS method which is a *quasi-Newton* method of numerical "hill-climbing optimization."

  - See http://en.wikipedia.org/wiki/BFGS_method for details.
  - Basically what a quasi-Newton hill climbing method does is exactly as it sounds. It chooses different parameter values θ_i over and over, checking to see if it has made the log-likelihood larger.
  - When it makes a change to θ_i and finds that the first derivative of the log-likelihood is zero and the second is negative, it stops, assuming it has found a maximum.
  - However, note that this could possibly be a local maxima if the log-likelihood is bi-modal, though usually it is uni-modal and this isn't a problem.

- The R routines typically do things backwards though: they minimize the negative of the log-likelihood which is equivalent to maximizing the log-likelihood. Be careful!

- Moreover, since the joint density will be small in value for large T, typically the log-likelihood is negative valued even at its maximum.

Figure 15: Log-likelihood surface in 2D, given parameters α_1 and β_1



- **Weaknesses of the ARCH and GARCH models**
- The following weaknesses of ARCH models follows from the Tsay textbook, pg.119:

1. The model assumes that positive and negative shocks have the same effect on volatility because volatility depends on the *squared* values of previous shocks.
   - By "shocks" here, we mean $\epsilon_t = r_t$. Remember, if our ARCH model contains a constant, then the "shock" is given by the difference between the return and that constant, that is $r_t - \mu = \epsilon_t = \sigma_{\epsilon,t}^2 z_t$. Of course given this model, the constant, μ, is the conditional and unconditional mean of r_t.
   - In practice it is well known that financial asset returns respond differently to positive and negative shocks.
   - Moreover, financial asset returns tend to be correlated with volatility; this is called the leverage effect.
   - Therefore, modifications have been proposed to account for this failing. For example, see the GARCH-M or E-GARCH models in the Tsay book.
2. The ARCH model is rather restrictive.
   - For example, $\alpha_1^2 < 1/3$ must hold for the ARCH(1) model to have a 4th unconditional moment that isn't infinite.
   - This constraint becomes even more complicated for higher order ARCH models.
3. The ARCH model does not "explain" what the sources of variation are in a financial time-series. Rather it merely provides a mechnical way to describe patterns in the conditional variance process. It says nothing about what causes these patterns.
4. ARCH models are likely to overpredict the volatility because they respond slowly to large isolated shocks in the returns series.

- While we won't go into too much detail here, you should know that the ARCH and GARCH models have also more recently (late 1980's) been modified to include a random component in the conditional variance equation. That is, we have:

$$\ln(\sigma^2_{\epsilon,t}) = \alpha_0 + \beta_1 \ln(\sigma^2_{\epsilon,t-1}) + \beta_2 \ln(\sigma^2_{\epsilon,t-2}) + \cdots + \beta_q \ln(\sigma^2_{\epsilon,t-q}) + \gamma_t$$

where $\gamma_t \sim N(0, \sigma^2_\gamma)$.

This model is called the *Stochastic Volatility* (SV) model and it adds a lot of flexibility to the ARCH/GARCH models. In fact, it looks a lot like a GARCH(0,q) model, but with the following modifications:

  - We have taken logs to ensure that the volatility remains positive.
  - We have added a random variable γ_t so that the volatility process is no longer a deterministic function of past returns, but is now an entirely stochastic process.

It turns out that adding these features makes the SV model much more difficult to estimate than the GARCH, however, as it requires a "filtering" method to extract the latent volatility. One way of doing this is through *state-space* methods.

For next class, you should review matrices, their inverses, determinants, and the multivariate Normal distribution. This will be useful for next week when we start discussing multivariate models like the VAR(p) model.

Those students who are weak in Linear Algebra should refer to Appendix A in Johnston and Dinardo, (2007), *Econometric Methods, 4th edition*, McGraw-Hill Press. While the other recommended texts may have good linear algebra reviews, the one in Johnston and Dinardo Appendix A is the best i've seen.

- **Part 3: Multivariate extensions for macro-econometrics**

- **Review of useful Linear Algebra results**

- Before we begin the next section of the course, you should already have a good grasp of basic Linear Algebra results like the use of matrices, their inverses, the notion of a determinant, and the multivariate Normal distribution.

- What follows are some specific results that we will need in discussing the multivariate extensions of the ARMA(p,q) class of models and their use in macro-econometrics.

Definition 32. *The covariance matrix:*
The covariance matrix of an $n \times 1$ vector-valued, mean zero, random variable $\boldsymbol{X}_t$, is given as:

$$E[\boldsymbol{X}_t \boldsymbol{X}'_t] = \boldsymbol{\Sigma}_X$$

where $\boldsymbol{\Sigma}_X$ is an $n \times n$ symmetric matrix, and the prime (') notation denotes vector or matrix transpose.

- **Example:**

- Say that $\boldsymbol{X}_t$ is 2×1. Then we have that the covariance matrix of $\boldsymbol{X}_t$ is given as:

$$\begin{bmatrix} E[X_1^2] & E[X_1 X_2] \\ E[X_2 X_1] & E[X_2^2] \end{bmatrix} = \boldsymbol{\Sigma}_X$$

- If $\boldsymbol{X}_t$ is not mean zero then we have the equivalent definition:

Definition 33. *The covariance matrix (non-zero mean):*
The covariance matrix of an $n \times 1$ vector-valued random variable $\boldsymbol{X}_t$ with mean $E[\boldsymbol{X}_t]$, is given as:

$$E[(\boldsymbol{X}_t - E[\boldsymbol{X}_t])(\boldsymbol{X}_t - E[\boldsymbol{X}_t])'] = E[\boldsymbol{X}_t \boldsymbol{X}_t'] - E[\boldsymbol{X}_t]E[\boldsymbol{X}_t]' = \boldsymbol{\Sigma}_X$$

where $\boldsymbol{\Sigma}_X$ is an $n \times n$ symmetric matrix.

Corollary 7. *The correlation matrix (non-zero mean):*
The correlation matrix of an $n \times 1$ vector-valued random variable $\boldsymbol{X}_t$ with mean $E[\boldsymbol{X}_t]$, is given as:

$$\boldsymbol{D}^{-1} \boldsymbol{\Sigma}_X \boldsymbol{D}^{-1}$$

where $\boldsymbol{\Sigma}_X$ is an $n \times n$ covariance matrix and $\boldsymbol{D}$ is the diagonal matrix of standard deviations. That is $\boldsymbol{D} = diag\{\sqrt{\boldsymbol{\Sigma}_{X,11}}, \sqrt{\boldsymbol{\Sigma}_{X,22}}, \ldots, \sqrt{\boldsymbol{\Sigma}_{X,nn}}\}$.

- **Note:** Of course there are also matrix analogs of the autocovariance function and autocorrelation functions (ACF). Just replace the above expressions with time period lag $t - s$ on the second X_t variable.

  We usually denote the matrix autocovariance at lag s as $\boldsymbol{\Gamma}(s)$. Therefore $\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma}_X$, and we have that the matrix autocorrelation function is $\boldsymbol{D}^{-1}\boldsymbol{\Gamma}(s)\boldsymbol{D}^{-1}$.

  See pg.390 in the Tsay textbook for more details.

- The following result generalizes the relationship between a scalar and a matrix. That is, in some sense it describes how multiplying a vector by a scalar might be equivalent to multiplication by a matrix.

Definition 34. *The eigenvalues and eigenvectors of a matrix:*
Let $\boldsymbol{A}$ be a square $n \times n$ matrix, $\mathbf{c}$ be an $n \times 1$ vector, and λ be a scalar. If we have that:

$$\boldsymbol{A}\mathbf{c} = \lambda \mathbf{c}, \quad \textit{for some} \quad \mathbf{c} \neq \mathbf{0}$$

Then we say that λ is an eigenvalue of the matrix $\boldsymbol{A}$ and $\mathbf{c}$ is an eigenvector.

- **Example:**

- As an example, let's solve for the eigenvalues of the matrix $\boldsymbol{A} = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$.

- The condition above requires that:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{c} &= \lambda \boldsymbol{c} \\ \Rightarrow (\boldsymbol{A} - \boldsymbol{I}\lambda)\boldsymbol{c} &= \boldsymbol{0}, \quad \text{for some} \quad \boldsymbol{c} \neq \boldsymbol{0}. \end{aligned} \tag{39a}$$

  But in order for $\boldsymbol{c} \neq \boldsymbol{0}$ we must have that the determinant of $(\boldsymbol{A} - \boldsymbol{I}\lambda)$ is equal to zero. Therefore, we can solve for the eigenvalues as the solutions to the *characteristic equation*:

$$\begin{aligned} det((\boldsymbol{A} - \boldsymbol{I}\lambda)) &= det\left(\begin{bmatrix} 4-\lambda & 2 \\ 2 & 1-\lambda \end{bmatrix}\right) = 0 \\ \Rightarrow (4-\lambda)(1-\lambda) - 4 &= (4-4) - \lambda(4+1) + \lambda^2 = 0 \end{aligned} \tag{40a}$$

  which has two solutions, $\lambda_1 = 5$ and $\lambda_2 = 0$. These are the eigenvalues of $\boldsymbol{A}$.

- For more complicated 2nd order polynomials you might have to use the quadratic formula:

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

  Therefore, we need to keep in mind that if $b^2 - 4ac < 0$ then the eigenvalue is a complex number! This case is not unusual.

- To find the eigenvectors, from $\boldsymbol{A}\boldsymbol{c} = \lambda \boldsymbol{c}$ we have:

$$\boldsymbol{A}\boldsymbol{c} = 5\boldsymbol{c} \Leftrightarrow \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ c_1 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ c_1 \end{bmatrix} \Rightarrow c_1 = 1/2 \tag{41a}$$

$$\boldsymbol{A}\boldsymbol{c} = 0\boldsymbol{c} \Leftrightarrow \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ c_2 \end{bmatrix} = 0 \begin{bmatrix} 1 \\ c_1 \end{bmatrix} \Rightarrow c_2 = -2 \tag{41b}$$

  where we have normalized the first element of each eigenvector to one. So the two eigenvectors are equal to $\boldsymbol{c}_1 = \begin{bmatrix} 1 & 1/2 \end{bmatrix}'$ and $\boldsymbol{c}_2 = \begin{bmatrix} 1 & -2 \end{bmatrix}'$.

- Notice that the eigenvectors are not unique! For any scalar $\alpha \in \mathbb{R}$ we have that $\alpha\mathbf{c}$ is also an eigenvector. Moreover, if there are $k \leq n$ distinct eigenvalues (that is k different eigenvalues) then there will also be $k \leq n$ linearly independent eigenvectors. However, when $k < n$ then $(n-k)$ of the n eigenvectors will be linear combinations of the others (that is they will be linearly dependent), since there will be $(n-k)$ repeated eigenvalues.

- Another useful concept when dealing with matrices is that of "positive-definiteness."
  - A positive-definite matrix is in some ways like the analog of a strictly positive scalar.
  - Note that just as the variance of a random variable should always be positive, we desire that our covariance matrices are always positive-definite.

Definition 35. *Positive-definite matrix:*
A necessary and sufficient condition for the real symmetric matrix $\boldsymbol{A}$ to be positive-definite is that all the eigenvalues of $\boldsymbol{A}$ are strictly positive.

- The next definition gives us a way of taking the "square root" of a square matrix. It is called the Cholesky decomposition and is as follows.

Definition 36. *Cholesky decomposition:*
If $\boldsymbol{A}$ is symmetric and positive-definite, a nonsingular, lower triangular matrix with strictly positive diagonal entries, $\boldsymbol{L}$ can be found such that $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$. We write $\boldsymbol{L} = \boldsymbol{A}^{1/2}$.

- **Example:**

- Let $\boldsymbol{C}$ be the correlation matrix, $\boldsymbol{D}^{-1}\boldsymbol{\Gamma}(0)\boldsymbol{D}^{-1} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

- The Cholesky decomposition is given as:

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{bmatrix}$$

- Another useful result that is similar to the Cholesky decomposition is known as the *diagonalization* of a square matrix.

Definition 37. *Matrix diagonalization:*
Given a (not necessarily symmetric) invertible (i.e. full rank) matrix $\boldsymbol{A}$ of dimension $n \times n$, with n distinct eigenvalues, we can write:

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1} \tag{42a}$$

$$\Leftrightarrow \boldsymbol{\Lambda} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \tag{42b}$$

where $\boldsymbol{P}$ has the n linearly independent eigenvectors of $\boldsymbol{A}$ as its columns and $\boldsymbol{\Lambda}$ is a diagonal matrix with the n distinct eigenvalues of $\boldsymbol{A}$ along its main diagonal.

Note that in the case where $\boldsymbol{A}$ is symmetric, we have that $\boldsymbol{P}^{-1} = \boldsymbol{P}'$. That is, $\boldsymbol{P}$'s inverse is its transpose so that $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}$. We call such matrices orthogonal.

- **The VARIMA(p) class of models**

- Now, with these tools we can tackle the multivariate extensions of the ARIMA class of models. We will start first with the VAR(p).

Definition 38. *The VAR(p) model:*
The vector autoregressive model of order p, (the VAR(p) model) is defined as:

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{\epsilon}_t, \quad \textit{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$$

where $\boldsymbol{X}_t$ is an $n \times 1$ vector, each of the autoregressive coefficient matrices $\boldsymbol{\Phi}_i$ are $n \times n$, and $\boldsymbol{\epsilon}_t$ is an $n \times 1$ random vector distributed multivariate Normal with $n \times 1$ mean vector $\mathbf{0}$, and $n \times n$ covariance $\boldsymbol{\Sigma}_\epsilon$.

- Just as was done for the AR(p) model, we can also solve for the first and second moments of the VAR(p) process.

- **First uncentered moment of the VAR(p)**

- We can use the same type of proof as was done above to solve for expression (4b), where we assume in advance that $\boldsymbol{X}_t$ is stationary:

$$\begin{aligned} E[\boldsymbol{X}_t] &= E[\boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{\epsilon}_t] \\ &= \boldsymbol{\Phi}_1 E[\boldsymbol{X}_{t-1}] + \boldsymbol{\Phi}_2 E[\boldsymbol{X}_{t-2}] + \cdots + \boldsymbol{\Phi}_p E[\boldsymbol{X}_{t-p}] + \boldsymbol{\mu}_{\boldsymbol{\epsilon}} \\ &= \left(\sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) E[\boldsymbol{X}_t] + \boldsymbol{\mu}_{\boldsymbol{\epsilon}} \end{aligned} \tag{43a}$$

$$\Rightarrow \boldsymbol{\mu}_{\boldsymbol{\epsilon}} = \left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right) E[\boldsymbol{X}_t] \tag{43b}$$

$$\Rightarrow E[\boldsymbol{X}_t] = \left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \right)^{-1} \boldsymbol{\mu}_{\boldsymbol{\epsilon}} \quad \text{where} \quad \boldsymbol{\mu}_{\boldsymbol{\epsilon}} = \mathbf{0} \tag{43c}$$

- So this gives us an expression for the unconditional mean of the VAR(p). Notice that the unconditional mean is a vector of length $n \times 1$ (in this case equal to 0 since the mean of $\boldsymbol{\epsilon}_t$ is zero).

- The conditional mean of the VAR(p) is likewise given as:

$$E[\boldsymbol{X}_t | F_{t-1}] = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} \tag{44}$$

- where F_{t-1} is the usual information set we discussed above, that is $F_{t-1} = \{\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \ldots, \boldsymbol{X}_0\}$. Notice that just as in the AR(p), the VAR(p) conditional 1st uncentered moment only depends on information up to $\boldsymbol{X}_{t-p}$.

Definition 39. *The VMA(q) model:*
The vector moving average model of order q, (the VMA(q) model) is defined as:

$$\boldsymbol{X}_t = \boldsymbol{\Theta}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Theta}_2\boldsymbol{\epsilon}_{t-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\epsilon}_{t-q} + \boldsymbol{\epsilon}_t, \quad \textit{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$$

where $\boldsymbol{X}_t$ is an $n \times 1$ vector, each of the moving average coefficient matrices $\boldsymbol{\Theta}_i$ are $n \times n$, and $\boldsymbol{\epsilon}_t$ is an $n \times 1$ random vector distributed multivariate Normal with $n \times 1$ mean vector $\mathbf{0}$, and $n \times n$ covariance $\boldsymbol{\Sigma}_\epsilon$.

- **First uncentered moment of VMA(q)**:

$$\begin{aligned} E[\boldsymbol{X}_t] &= E[\boldsymbol{\Theta}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Theta}_2\boldsymbol{\epsilon}_{t-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\epsilon}_{t-q} + \boldsymbol{\epsilon}_t] \\ &= \boldsymbol{\Theta}_1 E[\boldsymbol{\epsilon}_{t-1}] + \boldsymbol{\Theta}_2 E[\boldsymbol{\epsilon}_{t-2}] + \cdots + \boldsymbol{\Theta}_q E[\boldsymbol{\epsilon}_{t-q}] + \boldsymbol{\mu}_\epsilon \\ &= \left(\sum_{j=1}^{q} \boldsymbol{\Theta}_j\right) \boldsymbol{\mu}_\epsilon + \boldsymbol{\mu}_\epsilon \end{aligned} \tag{45a}$$

$$\Rightarrow E[\boldsymbol{X}_t] = \left(\boldsymbol{I} + \sum_{j=1}^{p} \boldsymbol{\Theta}_j\right) \boldsymbol{\mu}_\epsilon \quad \text{where} \quad \boldsymbol{\mu}_\epsilon = \mathbf{0} \tag{45b}$$

- So this gives us an expression for the unconditional mean of the VMA(q).

- The *conditional* mean is given as:

$$E[\boldsymbol{X}_t | F_{t-1}] = \boldsymbol{\Theta}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Theta}_2\boldsymbol{\epsilon}_{t-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\epsilon}_{t-q} \tag{46}$$

- Again notice that in the VMA(q), the conditional mean only depends on information up to $\boldsymbol{X}_{t-q}$.

- **Second centered moment of VMA(q)**:

  Again, instead of working with $E[(\boldsymbol{X}_t - E[\boldsymbol{X}_t])(\boldsymbol{X}_t - E[\boldsymbol{X}_t])']$ I will work with $Var(\cdot)$ which avoids having to deal with the outer product.

$$\begin{aligned} Var[\boldsymbol{X}_t] &= Var[\boldsymbol{\Theta}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Theta}_2\boldsymbol{\epsilon}_{t-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\epsilon}_{t-q} + \boldsymbol{\epsilon}_t] \\ &= \boldsymbol{\Theta}_1 Var[\boldsymbol{\epsilon}_{t-1}]\boldsymbol{\Theta}_1' + \boldsymbol{\Theta}_2 Var[\boldsymbol{\epsilon}_{t-2}]\boldsymbol{\Theta}_2' + \cdots + \boldsymbol{\Theta}_q Var[\boldsymbol{\epsilon}_{t-q}]\boldsymbol{\Theta}_q' \\ &\quad + 2\sum_{k=1}^{q}\sum_{j=k+1}^{q} \boldsymbol{\Theta}_k Covar[\boldsymbol{\epsilon}_{t-k}\boldsymbol{\epsilon}_{t-j}]\boldsymbol{\Theta}_j' + \boldsymbol{\Sigma}_\epsilon \\ &= \sum_{j=1}^{q} \boldsymbol{\Theta}_j\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Theta}_j' + \boldsymbol{\Sigma}_\epsilon, \quad \text{since} \quad \boldsymbol{\Gamma}_\epsilon(s) = 0 \quad \forall s \neq 0 \end{aligned} \tag{47a}$$

$$\Rightarrow Var[\boldsymbol{X}_t] = \sum_{j=1}^{q} \boldsymbol{\Theta}_j\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Theta}_j' + \boldsymbol{\Sigma}_\epsilon \tag{47b}$$

where the result follows from the fact that since $E[\boldsymbol{\epsilon}_t] = \mathbf{0}$, we have that:

$$Var[\boldsymbol{\Theta}_i\boldsymbol{\epsilon}_{t-i}] = E[\boldsymbol{\Theta}_i\boldsymbol{\epsilon}_{t-i}\,(\boldsymbol{\Theta}_i\boldsymbol{\epsilon}_{t-i})'] = \boldsymbol{\Theta}_i E[\boldsymbol{\epsilon}_{t-i}\boldsymbol{\epsilon}_{t-i}']\boldsymbol{\Theta}_i' = \boldsymbol{\Theta}_i\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Theta}_i'$$

- The above expression gives the *unconditional* variance, $\mathbf{\Gamma}_X(0)$, of the VMA(q) process.

- Of course, just as in the univariate case the *conditional* variance of the VMA(q) is simply:

$$Var[\boldsymbol{X}_t|F_{t-1}] = \mathbf{\Sigma}_\epsilon$$

since conditional on all the $\boldsymbol{\epsilon}_{t-j}$'s for $j = 1, \ldots, q$ the only variation stems from $\boldsymbol{\epsilon}_t$.

- **Lag polynomials**

- Just as in the univariate case, we can write the VAR(p) and VMA(q) in terms of lag polynomial operators.

- That is, we can write the VAR(p) model as:

$$\mathbf{\Phi}(L)\boldsymbol{X}_t = \boldsymbol{\epsilon}_t \tag{48}$$

where $\mathbf{\Phi}(L) = \boldsymbol{I} - \mathbf{\Phi}_1 L - \mathbf{\Phi}_2 L^2 - \cdots - \mathbf{\Phi}_p L^p$.

- And we can write the MA(q) model as:

$$\boldsymbol{X}_t = \mathbf{\Theta}(L)\boldsymbol{\epsilon}_t \tag{49}$$

where $\mathbf{\Theta}(L) = \boldsymbol{I} + \mathbf{\Theta}_1 L + \mathbf{\Theta}_2 L^2 + \cdots + \mathbf{\Theta}_q L^q$.

- **The VARIMA(p,q) model**

- The lag polynomial again gives us an easy way to represent the vector version of the ARIMA(p,q) model.

Definition 40. *The VARIMA(p,d,q) model:*
A vector autoregressive, moving average model of orders p and q, applied to the dth order integrated process (an VARIMA(p,d,q) model) is such that:

$$\mathbf{\Phi}(L)\Delta^d \boldsymbol{X}_t = \mathbf{\Theta}(L)\boldsymbol{\epsilon}_t$$

where $\boldsymbol{\epsilon}_t \sim MVN(0, \mathbf{\Sigma}_\epsilon)$, and $\mathbf{\Phi}(L)$ is an VAR(p) lag polynomial and $\mathbf{\Theta}(L)$ is an VMA(q) lag polynomial.

- **Invertibility condition for VAR(1)**

- Recall our earlier example of the *invertible* univariate AR(1) process: $X_t\alpha(L) = X_t(1 - \alpha_1 L) = \epsilon_t$.
  - ◇ If $|\alpha_1| < 1$ we have $X_t = \epsilon_t/(1 - \alpha_1 L) = \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}$.
  - ◇ So by "inverting" the lag polynomial $\alpha(L)$ we have effectively created an MA(∞) from the AR(1)!

- ◇ That is, we say that if $|\alpha_1| < 1$ then the AR(1) model is *invertible* into an MA(∞).
- ◇ Likewise, given a similar condition on β_1 we have that the MA(1) is invertible into an AR(∞).

- ○ However, one might ask what the analog of this result is in the multivariate case?
  - ◇ Notice that the inverse AR(1) lag polynomial expression $\alpha(L)^{-1} = 1/(1-\alpha_1 L) = \sum_{j=0}^{\infty} \alpha_1^j L^j$ only exists, that is converges to a finite value, if $|\alpha_1| < 1$.
  - ◇ That is, the values of α_1^j get smaller and smaller as $j \to \infty$.
  - ◇ Can we find some kind of analagous result for the VAR(1) model with matrix coefficient $\boldsymbol{\Phi}_1$?
- ○ It turns out we can, if we consider the diagonalization of the coefficient matrix $\boldsymbol{\Phi}_1$.
  - ◇ That is, we wish to find conditions such that $\boldsymbol{\Phi}(L)^{-1} = (\boldsymbol{I} - \boldsymbol{\Phi}_1 L)^{-1} = \sum_{j=0}^{\infty} \boldsymbol{\Phi}_1^j L^j$ converges to something finite.
  - ◇ Diagonalizing $\boldsymbol{\Phi}_1$ we get $\boldsymbol{\Phi}_1 = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$.
  - ◇ But this implies that $\boldsymbol{\Phi}_1^j = \left(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}\right)^j = \boldsymbol{P}\boldsymbol{\Lambda}^j\boldsymbol{P}^{-1}$.
  - ◇ Therefore, since $\boldsymbol{\Lambda}$ contains the eigenvalues of $\boldsymbol{\Phi}_1$ along its main diagonal, we have that:

    $$\boldsymbol{\Phi}_1^j \to 0 \quad \text{if.f.} \quad \boldsymbol{\Lambda}^j \to 0 \quad \text{as} \quad j \to \infty$$

    which requires that all of the eigenvalues of $\boldsymbol{\Phi}_1$ are less than 1 in absolute value.
- ○ Therefore we have that $\boldsymbol{\Phi}_1(L)^{-1} = (\boldsymbol{I} - \boldsymbol{\Phi}_1 L)^{-1} = \sum_{j=0}^{\infty} \boldsymbol{\Phi}_1^j L^j$ converges to something finite only if all of the eigenvalues of $\boldsymbol{\Phi}_1$ are less than 1 in absolute value.
- ○ This is the invertibility condition of the VAR(1) model. Therefore just as in the univariate case:
  - ◇ If the VAR(1) polynomial is invertible, we can write the VAR(1) as an infinite VMA(∞).
  - ◇ Moreover, invertibility of the VAR(1) implies stationarity of the VAR(1).
  - ◇ Invertibility of the VMA(1) lag polynomial means we can write the model as an infinite VAR(∞).
  - ◇ Stationarity of the VMA(q) does not necessarily imply invertibility.
  - ◇ The VARIMA(p,q) can be inverted into an infinite VAR(∞) or VMA(∞) given the correct invertibility condition.

- ○ **Estimation of the VAR(p) model**
- ○ The VAR(p) model includes $n(n+1)/2 + pn^2$ parameters (where n is the number of variables in $\boldsymbol{X}_t$). Therefore its estimation is subject to the so called "curse of dimensionality." That is, the number of parameters requiring estimation grows very large for even small lag values of p.
  - ◇ Since the covariance matrix $\boldsymbol{\Sigma}_\epsilon$ is symmetric, it has $n(n+1)/2$ distinct elements that require estimation.
  - ◇ Moreover, since each of the p VAR coefficient matrices $\boldsymbol{\Phi}_i$ have n^2 elements, together they account for pn^2 parameters that require estimation.

- We can avoid this problem by putting constraints on the p VAR coefficient matrices $\mathbf{\Phi}_i$ or the covariance matrix $\mathbf{\Sigma}_\epsilon$. For example we can assume that they are either diagonal or symmetric.
  - Of course, this represents a trade-off between estimation and model flexibility and fit.

- The VAR(p) model can either be estimated by MLE or by applying OLS individually to each of the n linear equations the VAR(p) represents.
- When the innovations $\boldsymbol{\epsilon}_t$ are multivariate Normal, that is $\boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$, we have that the likelihood function is given as:

$$\begin{aligned} LL_{\Theta|X}(\theta|x) &= \sum_{t=p+1}^{T} \ln f(\boldsymbol{x}_t|F_{t-1}, \theta) + \sum_{t=1}^{p} \ln f(\boldsymbol{x}_t|\theta) \\ &= -\frac{(T-p)n}{2} \ln 2\pi - \frac{T-p}{2} \ln det\left(\mathbf{\Sigma}_\epsilon\right) - \frac{1}{2} \sum_{t=p+1}^{T} \boldsymbol{\epsilon}_t \mathbf{\Sigma}_\epsilon^{-1} \boldsymbol{\epsilon}_t' + \sum_{t=1}^{p} \ln f(\boldsymbol{x}_t|\theta) \quad (50a) \end{aligned}$$

  where $\boldsymbol{\epsilon}_t = \boldsymbol{X}_t - \sum_{j=1}^{p} \mathbf{\Phi}_j \boldsymbol{X}_{t-j}$ and where we can just ignore the $\sum_{t=1}^{p} \ln f(\boldsymbol{x}_t|\theta)$ term and treat $\boldsymbol{X}_t$ for $t = 1, \ldots, p$ as fixed.
  - The alternative being of course to express their densities as unconditional (not depending on the past) with mean zero but having a complicated expression for the unconditional variance.
  - For large T, treating the first p values of $\boldsymbol{X}_t$ as fixed makes little difference to the MLE estimates of the VAR(p) parameters.

- **Impulse response functions (IRF)**
- The impulse response function (IRF) is a useful way of interpreting how some exogenous "shock" to a system feeds into it over time.
- More specifically, the impulse response function is the "time path" of the stochastic process x_t when we "shock" the process at some time τ so that $\epsilon_\tau = \gamma$, where γ is the shock size, but force all other $\epsilon_t = 0$ for $t \neq \tau$.
- That is, we wish to know how the process behaves given an initial shock at time τ and where all other innovations before or afterwards are zero.
- The notion of an impulse response function comes from the study of stochastic difference equations, of which the ARIMA(p,q) model is a special class.

Example: AR(1) impulse response function

- It turns out that the easiest way to find the impulse response function, $\phi(t)$, of the AR(1) model is to invert it into a MA(∞).
- What we really desire to do is to write out the process *only* in terms of the stochastic innovations, so that we can see how the process behaves as we set them all to zero (i.e. $\epsilon_t = 0$) except for our one shock at time τ.

- ◇ That is, consider the AR(1), $X_t = \alpha_1 X_{t-1} + \epsilon_t$. Inverting this model implies the MA(∞) representation $X_t = \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}$.
- ◇ Given the MA(∞) form it is easy to see that the impulse response function is $\phi(t) = \alpha_1^{t-\tau}\epsilon_\tau$, where $t \geq \tau$.

- That is, the time path of X_t given a shock at time τ, ϵ_τ, is:

1. ϵ_τ at time τ.
2. $\alpha_1\epsilon_\tau$ at time $\tau + 1$.
3. $\alpha_1^2\epsilon_\tau$ at time $\tau + 2$.
4. $\vdots$
5. $\alpha_1^t\epsilon_\tau$ at time $\tau + t$.

- Therefore, if the AR(1) model is stationary (i.e. $|\alpha_1| < 1$) we can see that shocks to the system should die out slowly over time.

Example: VAR(1) impulse response function

- As another similar example, consider the VAR(1), $\boldsymbol{X}_t = \boldsymbol{\Phi}_1\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t$. Inverting this model implies the VMA(∞) representation $\boldsymbol{X}_t = \sum_{j=0}^{\infty} \boldsymbol{\Phi}_1^j \boldsymbol{\epsilon}_{t-j}$.

- Given the VMA(∞) form it is easy to see that the impulse response function is $\phi(t) = \boldsymbol{\Phi}_1^{t-\tau}\boldsymbol{\epsilon}_\tau$, where $t \geq \tau$.

- That is, the time path of $\boldsymbol{X}_t$ given a shock at time τ, $\boldsymbol{\epsilon}_\tau$, is:

1. $\boldsymbol{\epsilon}_\tau$ at time τ.
2. $\boldsymbol{\Phi}_1\boldsymbol{\epsilon}_\tau$ at time $\tau + 1$.
3. $\boldsymbol{\Phi}_1^2\boldsymbol{\epsilon}_\tau$ at time $\tau + 2$.
4. $\vdots$
5. $\boldsymbol{\Phi}_1^t\boldsymbol{\epsilon}_\tau$ at time $\tau + t$.

- Therefore, if the VAR(1) model is stationary (i.e. if all the eigenvalues of $\boldsymbol{\Phi}_1$ are less than 1 in absolute value) we can see that shocks to the system should die out slowly over time.

Example: VMA(q) impulse response function

- As a third example, consider the VMA(q), $\boldsymbol{X}_t = \boldsymbol{\Theta}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Theta}_2\boldsymbol{\epsilon}_{t-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\epsilon}_{t-q} + \boldsymbol{\epsilon}_t$. This model need not be inverted since it is already written in a form that expresses $\boldsymbol{X}_t$ as a function of only the stochastic components.

- Given the VMA(q) form we can see that the impulse response function is:

$$\phi(t) = \begin{cases} \boldsymbol{\epsilon}_\tau & t = \tau \\ \boldsymbol{\Theta}_{t-\tau}\boldsymbol{\epsilon}_\tau & 1 \leq (t-\tau) \leq q \\ 0 & (t-\tau) > q \end{cases}$$

- That is, the time path of $\boldsymbol{X}_t$ given a shock at time τ, $\boldsymbol{\epsilon}_\tau$, is:
  1. $\boldsymbol{\epsilon}_\tau$ at time τ.
  2. $\boldsymbol{\Theta}_1\boldsymbol{\epsilon}_\tau$ at time $\tau + 1$.
  3. $\boldsymbol{\Theta}_2\boldsymbol{\epsilon}_\tau$ at time $\tau + 2$.
  4. $\vdots$
  5. $\boldsymbol{\Theta}_q\boldsymbol{\epsilon}_\tau$ at time $\tau + q$.
  6. Zero for all times $\tau + q + j$ where $j > 0$.
- Therefore, if the VMA(q) model is stationary (i.e. if all the coefficient matrices, $\boldsymbol{\Theta}_i$, are finite) we can see that shocks to the system should die out after exactly $q + 1$ time periods.

Example: Random walk, I(1), impulse response function

- As a final example, consider the AR(1), $X_t = \alpha_1 X_{t-1} + \epsilon_t$, but where $\alpha_1 = 1$. Intuitively, trying to invert this model implies that the MA(∞) representation $X_t = \epsilon_t/(1 - \alpha_1 L) = \sum_{j=0}^{\infty} \alpha_1^j \epsilon_{t-j}$ does not exist since the RHS will grow infinitely large (more formally, we must consider whether the moments of the RHS exist.)
- However, we can still consider the MA(∞) form to see how the impulse response function will behave. In this special case, $\phi(t) = \alpha_1^{t-\tau}\epsilon_\tau$, where $t \geq \tau$ but $\alpha_1 = 1$.
- Therefore, the time path of X_t given a shock at time τ, ϵ_τ, is now:
  1. ϵ_τ at time τ.
  2. ϵ_τ at time $\tau + 1$.
  3. ϵ_τ at time $\tau + 2$.
  4. $\vdots$
  5. ϵ_τ at time $\tau + t$.
- Therefore, if the AR(1) model has $|\alpha_1| = 1$ (technically we call this a "unit root"), it is not stationary and we can see that shocks to the system *never die out*. That is, they have a permanent effect on future values of X_t.

- **Macroeconometrics and the VAR(p) model**
- Since the early 1980's, VAR(p) models have become very popular in macroeconometrics.
- Some of the reasons include:
  - ◇ Most macroeconomic time series, such as GDP growth or measures of the inflation rate, typically exhibit a lot of autocorrelation. That is, they are **not** white noise like stocks are, and are therefore highly predictable given an AR(p) model on the levels process.
  - ◇ Moreover, many macro time series tend to be correlated *across* series – that is the GDP growth rate may be correlated with say, the unemployment rate.

- ◇ VAR(p) models are relatively easy to estimate and forecast, and can account for this cross-correlation.
- ◇ The Impulse Response Functions (IRF) for VAR(p)'s can be used to see how a single isolated exogenous "shock" to the economy may feed into the system and affect different macro variables across time.

- ○ Moreover, an econometrician named Sims in 1980 suggested that the VAR(p) model might actually be useful in informing economic theory.

- ○ That is:
  - ◇ Avoid starting with structural forms like those from the Keyesian model we discussed during the 1st class.
  - ◇ Start instead with a VAR(p). Throw in any macro variables you feel may be related.
  - ◇ Use the VAR(p) reduced form to derive a structural form as follows in the next section.
  - ◇ This way you let the data decide which variables are related instead of starting with theory.
  - ◇ You also avoid having to make the distinction between which variables are exogenous and those that are endogenous since the VAR(p) structural form assumes that all variables are endogenous.

- ○ **Recovering the structural form of a VAR(p)**

- ○ Take for example the following VAR(1) model:

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \tag{51a}$$

$$\text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon) \tag{51b}$$

and suppose that $\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 5/4 \end{bmatrix}$.

- ○ If $\boldsymbol{\Sigma}_\epsilon$ has a Cholesky decomposition, it can be rewritten as $\boldsymbol{\Sigma}_\epsilon^{1/2} \boldsymbol{I} \boldsymbol{\Sigma}_\epsilon^{1/2'}$.

- ○ Therefore, pre-multiply equation (51a) by $\boldsymbol{\Sigma}_\epsilon^{-1/2}$ so that we have:

$$\boldsymbol{\Sigma}_\epsilon^{-1/2} \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \boldsymbol{\Sigma}_\epsilon^{-1/2} \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \boldsymbol{\Sigma}_\epsilon^{-1/2} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \tag{52a}$$

So now we have that:

$$E[\boldsymbol{\Sigma}_\epsilon^{-1/2} \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t' \boldsymbol{\Sigma}_\epsilon^{-1/2'}] = \boldsymbol{I} \tag{53}$$

That is, we have reduced the covariance of the innovation to identity, so that $\boldsymbol{\epsilon}_t^* \sim MVN(\mathbf{0}, \boldsymbol{I})$ where $\boldsymbol{\epsilon}_t^* = \boldsymbol{\Sigma}_\epsilon^{-1/2} \boldsymbol{\epsilon}_t$.

- Finally, we can now derive a direct structural relationship between $x_{1,t}$ and $x_{2,t}$ since we have from equation (52a) that:

$$\begin{bmatrix} 1 & 0 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \boldsymbol{\epsilon}_t^* \tag{54a}$$

  So that from the second equation in the above system, (54a), we have:

$$x_{2,t} = \frac{1}{2}x_{1,t} - 0.7x_{1,t-1} + 0.95x_{2,t-1} + \epsilon_{2,t}^* \tag{55a}$$

  which gives us a structural relation between the endogenous variables $x_{2,t}$ and $x_{1,t}$, where $x_{2,t}$ is dependent on $x_{1,t}$.

- If we want to find the structural expression for $x_{1,t}$ dependent on $x_{2,t}$, we just start again from the beginning but where we switch the order of the equations so that the second equation becomes the first.

- That is, repeat all the steps above but starting with:

$$\begin{bmatrix} x_{2,t} \\ x_{1,t} \end{bmatrix} = \begin{bmatrix} 1.1 & -0.6 \\ 0.3 & 0.2 \end{bmatrix} \begin{bmatrix} x_{2,t-1} \\ x_{1,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{2,t} \\ \epsilon_{1,t} \end{bmatrix} \tag{56a}$$

$$\text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \tag{56b}$$

  and $\boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 5/4 & 1/2 \\ 1/2 & 1 \end{bmatrix}$.

- **Co-integration**

- The problem of trends in the data has always posed significant challenges to econometrics practitioners.

- Since trends in the data make the series non-stationary (recall our brief discuss above when we talked about the identification stage of the Box-Jenkins approach), we cannot apply all the models we have learned about in this course since they rely on the assumption that the series are in fact stationary.

- For review let's briefly discuss the topic of trending data again.

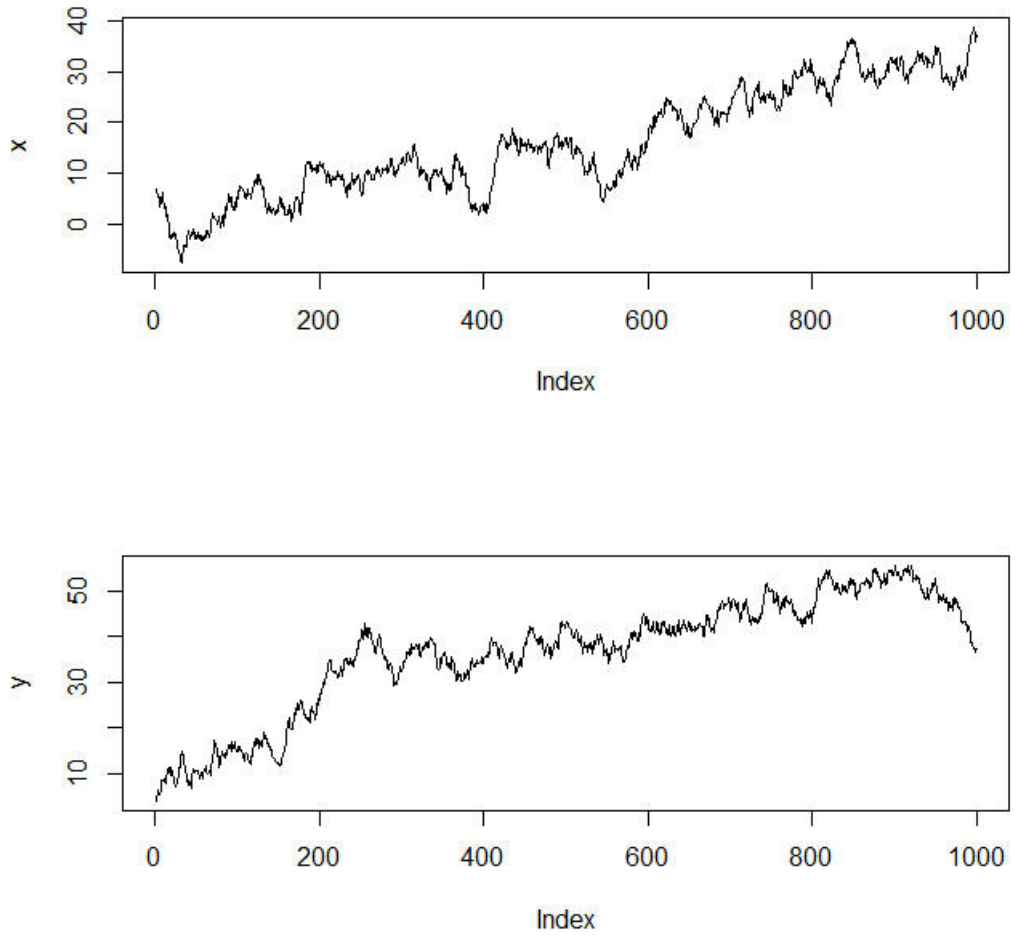
- **Trends (repeated again from ARIMA section):**

  - The question of whether or not a trend exists and of what kind is complex. Generally we divide trends into two types: *deterministic* trends and *stochastic* trends.

- ◇ Generally, if a trend exists then the DGP is non-stationary and direct application of ARIMA models can have misleading results since their theory assumes the series is stationary.
- ◇ A deterministic trend in an AR(1), for example, can be represented as some function of time: $X_t = \gamma t + \alpha_1 X_{t-1} + \epsilon_t$. These types of trends can be removed by fitting $X_t = \beta t + u_t$ by OLS and then using the residuals, u_t, as the new time-series in an AR(1).
- ◇ An example of a stochastic trend is the *unit root* process: $X_t = X_{t-1} + \epsilon_t$ (that is, an AR(1) with $\alpha_1 = 1$).
  - ∗ You may recognize this as the random walk we talked about earlier (hint: recursively substitute back to ϵ_0).
  - ∗ Recall, if we first difference the I(1) random walk, then we get a stationary I(0) process.
- ◇ We must be very careful in choosing which trend type to use! Statistical tests have been developed to try and determine if a trend is deterministic or stochastic, for example the Dickey-Fuller unit root test.

- ○ Consider the following plots:

Figure 16: Trendline challenge



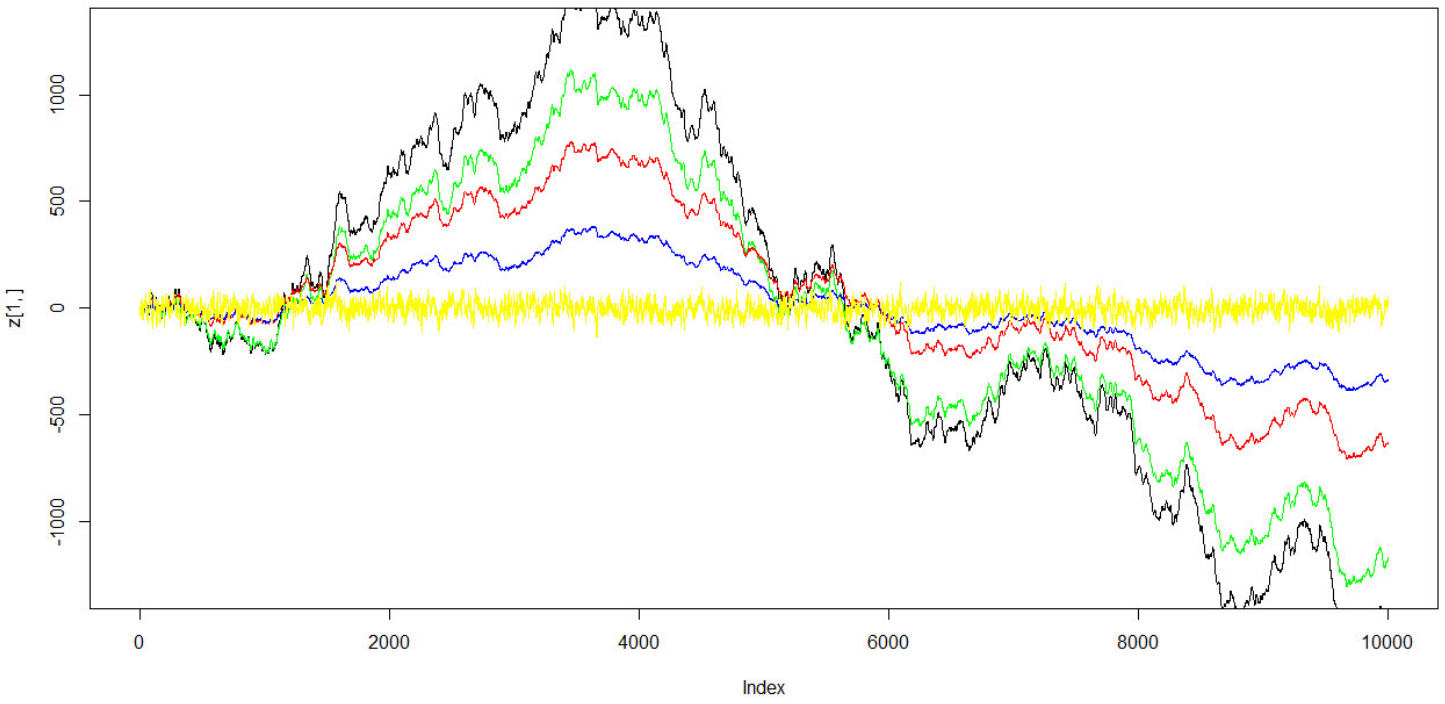
- Can you tell which of the two series, X_t and Y_t, follow either a *deterministic* trend or a *stochastic* trend?
- That is, the above plots were generated from either:
  - Random walk (stochastic trend):
    $$Z_t = Z_{t-1} + \epsilon_t, \text{ or equivalently, } Z_t = \sum_{j=0}^{t} \epsilon_{t-j}$$
  - An AR(1) with time trend (deterministic trend):
    $$Z_t = \gamma t + \alpha_1 Z_{t-1} + \epsilon_t$$
- If we have a deterministic trend, we can just subtract γt from the series and the process is then stationary. Since it is deterministic, we know exactly how to forecast it as well.
- However, if the trend is stochastic it is not clear that the particular path chosen by the series wasn't just "accidental." After all the trend is completely random so who can really say in which direction it will decide to move next?

- Either way, while we will skirt around most of these thorny issues in this class, it is important to note that historically it was suggested that we should take differences of processes that follow stochastic trends.
- That is, since a stochastic trend, Z_t, is I(1) we should difference it once, ΔZ_t, to make it stationary, or I(0), before applying the models we have talked about.
- Of course, a differenced series itself is not exactly the same as the original series, so this solution is somewhat unsatisfying.
- Generally then, we wish to avoid differencing if possible, or if we are going to difference, we wish to "retain" some of the original information stored in the series prior to differencing.
- The modern time-series literature has thus generated some new approaches, a couple of which turned out to be interrelated:
  1. The notion of co-integration and the VECM (vector error correction model) representation made famous by Clive Granger in (1981).
  2. The "Unobserved Components" methodology popularized by A.C. Harvey in the 1980's.
- We will discuss the 1st of these approaches in this class.
- The second one is more general and relies on decomposing the time series into a number of "components" that we do not directly see.
  - For example, we can decompose X_t as $X_t = \mu_t + Y_t + \epsilon_t$ where μ_t is a random walk, stochastic trend, and Y_t is say, an ARMA(p,q) process. Therefore, X_t, is I(1).
  - However, instead of taking differences, we actually *estimate* the entire stochastic trend along with the parameters of the ARMA(p,q) model.

- ◇ Of course, all we observe is the data for X_t, so in order to seperate what part is the stochastic trend and what part is the ARMA(p,q) component, we need something called a state-space representation and the notion of "signal extraction."
- ◇ These methods are highly computationally intensive and it is no surprise that their arrival coincided with the development of the modern computer.

- ○ **I(1) variables may be co-integrated**
- ○ Now, say that we have a few variables and we believe that they all follow stochastic trends. That is, we have a vector of variables $\boldsymbol{X}_t$ where each of the component variables, $x_{1,t}, x_{2,t}, \ldots$ are all I(1).
- ○ Consider the following plot:

Figure 17: Co-integrated series



- ○ Notice how all the series except the yellow one tend to move "together" across time?
  - ◇ This is what we mean by co-integrated series.
  - ◇ All of the lines except the yellow one are I(1) processes (they are the components of $\boldsymbol{X}_t$, $x_{1,t}, x_{2,t}, \ldots$).
  - ◇ Therefore, they all represent stochastic trends with no particular known direction or pattern.
- ○ However, the yellow line represents an I(0) process. Therefore, it is stationary.
  - ◇ How was the yellow series generated?

- ◇ It turns out that the yellow I(0) series is a linear combination of the other four, I(1), series.
- ◇ That is, I have somehow managed to construct some linear combination of I(1) series that is itself I(0)!

Definition 41. *Co-integrated series:*
Given some vector of n distinct I(1) series, $\boldsymbol{X}_t$, if there exists a vector $\mathbf{c}$ where $Z_t = \mathbf{c}'\boldsymbol{X}_t$ is itself I(0), we say that the vector $\mathbf{c}$ represents a co-integrating relation between the elements of $\boldsymbol{X}_t$.

- ○ Given this result, the rest of the course will devoted to:
  1. Identification of the co-integrating relations given some multivariate I(1) series, $\boldsymbol{X}_t$.
  2. Making use of these co-integrating relations in a formal model. The result will be the VECM (vector error correction) model.
  3. Estimating both the co-integrating relations and the parameters of the VECM model given a sample data set of size T.

- ○ **Identification of the co-integrating relations**

- ○ Despite how "intuitive" the concept seems, the mathematics of co-integration can be quite daunting so I will try to go through the process step by step.
  - ◇ You should make sure you understand where all the results are coming from before moving onto the next!

- ○ First, let us start with a typical VAR(1) model which we covered in the last section. Moreover, to simplify things we will assume that $\boldsymbol{X}_t$ is 2×1 – that is, $\boldsymbol{X}_t$ is a bivariate process.

- ○ That is, we have the model:

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \tag{57}$$

and all of the elements of $\boldsymbol{X}_t$ are I(1).

- ○ Using the diagonalization of a matrix theorem again, we can write (57) as:

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \\ \Rightarrow \boldsymbol{P}^{-1}\boldsymbol{X}_t &= \boldsymbol{\Lambda}\boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \boldsymbol{P}^{-1}\boldsymbol{\epsilon}_t \\ \Rightarrow \boldsymbol{Z}_t &= \boldsymbol{\Lambda}\boldsymbol{Z}_{t-1} + \boldsymbol{\epsilon}_t^* \end{aligned} \tag{58a}$$

where $\boldsymbol{Z}_t = \boldsymbol{P}^{-1}\boldsymbol{X}_t$ and $\boldsymbol{\epsilon}_t^* = \boldsymbol{P}^{-1}\boldsymbol{\epsilon}_t$. Recall that $\boldsymbol{P}$ has the eigenvectors of $\boldsymbol{\Phi}_1$ as its columns and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Phi}_1$ along its main diagonal.

- ○ Writing out (58a) in more detail we have:

$$\begin{aligned} \boldsymbol{Z}_t &= \boldsymbol{\Lambda}\boldsymbol{Z}_{t-1} + \boldsymbol{\epsilon}_t^* \\ \Rightarrow \begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} &= \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t}^* \\ \epsilon_{2,t}^* \end{bmatrix} \end{aligned} \tag{59a}$$

- We now have three cases:

  1. If both eigenvalues are less than one in absolute value, that is $|\lambda_1| < 1$ and $|\lambda_2| < 1$, then we have that there exist no co-integrating relations since both $z_{1,t}$ and $z_{2,t}$ represent stationary AR(1) processes (i.e. I(0) processes).
     - Of course, this case is ruled out by the fact that all the elements of $\boldsymbol{X}_t$ are I(1). That is, at least one eigenvalue *must* be equal to 1.
  2. If both of the eigenvalues are equal to 1 in absolute value, that is $|\lambda_1| = 1$ and $|\lambda_2| = 1$, then we again have there exist no co-integrating relations since both $z_{1,t}$ and $z_{2,t}$ represent unit-root AR(1) processes (i.e. I(1) processes).
     - This case implies that both $x_{1,t}$ and $x_{2,t}$ each follow *their own* individual stochastic trends; that is they don't "move together."
  3. Finally we have the case where one eigenvalue is less than one in absolute value, while the other is equal to 1. That is we have for example, $|\lambda_1| = 1$ and $|\lambda_2| < 1$. This case implies that there exists exactly one co-integrating relation between the elements of $\boldsymbol{X}_t$.

- **Aside:** Note that in all the above cases I've spoken of the "absolute value" of the eigenvalues. More generally what I really mean here is the *modulus* of the eigenvalues since they can quite often by complex valued! The modulus is like the analog of absolute value in the complex plane.

- How do we know that the 3rd case above is true? Well look again at the expression above in (59a).

  - Notice that what we have really done here is create a linear combination of $\boldsymbol{X}_t$ that is I(0).
  - Where have we done this? Look again at the second equation:

    $$z_{2,t} = \lambda_2 z_{2,t-1} + \epsilon^*_{2,t}$$

    Clearly $z_{2,t}$ is I(0), since $|\lambda_2| < 1$.
  - But we know that $z_{2,t} = \begin{bmatrix} p^*_{21} & p^*_{22} \end{bmatrix} \boldsymbol{X}_t$ where $\begin{bmatrix} p^*_{21} & p^*_{22} \end{bmatrix}$ is the second row of $\boldsymbol{P}^{-1}$.
  - Therefore, the second row of $\boldsymbol{P}^{-1}$, $\begin{bmatrix} p^*_{21} & p^*_{22} \end{bmatrix}$, represents the co-integrating relation.

- **The VECM(1) representation**

- Now, let's make use of the co-integrating relation $\begin{bmatrix} p^*_{21} & p^*_{22} \end{bmatrix}$ in a formal model.

- First, let's start again from the VAR(1) form:

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon), \tag{57}$$

  where all of the elements of $\boldsymbol{X}_t$ are I(1).

- Subtracting $\boldsymbol{X}_{t-1}$ from both sides we get:

$$\Delta \boldsymbol{X}_t = \boldsymbol{\Pi} \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \tag{60}$$

  where $\boldsymbol{\Pi} = \boldsymbol{\Phi}_1 - \boldsymbol{I}$. This is called the VECM representation.

- What we are eventually going to show is that $\boldsymbol{\Pi}$ embodies the co-integrating relation (not obvious yet).
  - Therefore, even though $\boldsymbol{X}_t$ is I(1), $\Delta\boldsymbol{X}_t$ is sort of "nudged" up or down according to the information stored in the co-integrating relation and towards the "long-run equilibrium" relationship it represents.
  - That is, instead of simply differencing $\boldsymbol{X}_t$ to make it I(0) we are going to incorporate the co-integrating relation information directly.
- So, how is this done? First, again using the diagonalization of a matrix theorem on (68) we get:

$$\Delta\boldsymbol{X}_t = \boldsymbol{P}\boldsymbol{\Lambda}^*\boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \tag{61a}$$

$$= \boldsymbol{P}\begin{bmatrix} (\lambda_1 - 1) & 0 \\ 0 & (\lambda_2 - 1) \end{bmatrix}\boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \tag{61b}$$

- Note that the eigenvalues of $\boldsymbol{\Pi} = \boldsymbol{\Phi}_1 - \boldsymbol{I}$ are equal to $\lambda_i - 1$ (where λ_i is an eigenvalue of $\boldsymbol{\Phi}_1$ as before) but the eigenvectors of both $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Pi}$ are the same.[8]
- Notice that from our earlier example that if $|\lambda_1| = 1$, we have:

$$\Delta\boldsymbol{X}_t = \boldsymbol{P}\begin{bmatrix} 0 & 0 \\ 0 & (\lambda_2 - 1) \end{bmatrix}\boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \tag{63}$$

  and so $\boldsymbol{\Pi}$ is a rank 1 matrix. That is, its columns are linearly dependent.

[8] Proof is as follows.

$$\boldsymbol{\Phi_1 c} = \lambda\boldsymbol{c} \tag{62a}$$

$$\Leftrightarrow \boldsymbol{\Phi_1 c} - \boldsymbol{c} = \lambda\boldsymbol{c} - \boldsymbol{c} \tag{62b}$$

$$\Leftrightarrow (\boldsymbol{\Phi_1} - \boldsymbol{I})\boldsymbol{c} = (\lambda - 1)\boldsymbol{c} \tag{62c}$$

$$\Leftrightarrow \boldsymbol{\Pi c} = (\lambda - 1)\boldsymbol{c} \tag{62d}$$

- This fact implies that we can now write $\mathbf{\Pi}$ in the following manner:

$$\begin{aligned}
\mathbf{\Pi} &= \boldsymbol{P} \begin{bmatrix} 0 & 0 \\ 0 & (\lambda_2 - 1) \end{bmatrix} \boldsymbol{P}^{-1} \\
&= \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & (\lambda_2 - 1) \end{bmatrix} \begin{bmatrix} p^*_{11} & p^*_{12} \\ p^*_{21} & p^*_{22} \end{bmatrix} && \text{(64a)} \\
&= \begin{bmatrix} 0 & p_{12}(\lambda_2 - 1) \\ 0 & p_{22}(\lambda_2 - 1) \end{bmatrix} \begin{bmatrix} p^*_{11} & p^*_{12} \\ p^*_{21} & p^*_{22} \end{bmatrix} && \text{(64b)} \\
&= \begin{bmatrix} p_{12}(\lambda_2 - 1) \\ p_{22}(\lambda_2 - 1) \end{bmatrix} \begin{bmatrix} p^*_{21} & p^*_{22} \end{bmatrix} && \text{(64c)} \\
&= \boldsymbol{\alpha}\boldsymbol{\beta}' && \text{(64d)}
\end{aligned}$$

  where $\boldsymbol{\alpha} = \begin{bmatrix} p_{12}(\lambda_2 - 1) \\ p_{22}(\lambda_2 - 1) \end{bmatrix}$ is called the "factor loadings" vector and $\begin{bmatrix} p^*_{21} & p^*_{22} \end{bmatrix}$ is the co-integration relation vector, as we discussed earlier.

- Therefore, in the bivariate VECM model if $|\lambda_1| = 1$ and $|\lambda_2| < 1$ so that $\mathbf{\Pi}$ is a rank 1 matrix, we can write the VECM model in (68) as:

$$\Delta \boldsymbol{X}_t = \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \tag{65}$$

- But recall that since $\boldsymbol{\beta}$ is the co-integrating relation, then $\boldsymbol{\beta}' \boldsymbol{X}_t = z_{2,t}$, the I(0) linear combination, so we have:

$$\begin{aligned}
\Delta \boldsymbol{X}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}' \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \\
&= \begin{bmatrix} p_{12}(\lambda_2 - 1) \\ p_{22}(\lambda_2 - 1) \end{bmatrix} z_{2,t-1} + \boldsymbol{\epsilon}_t, && \text{(66a)}
\end{aligned}$$

  and so the differenced elements of $\Delta \boldsymbol{X}_t$ can each be seen to be functions of the I(0) process $z_{2,t-1}$ but where they are each differently "nudged" by the loading factors in $\boldsymbol{\alpha}$.

  That is, the loading factors affect the "speed" of adjustment towards the long-run equilibrium suggested by $z_{2,t}$.

- Moreover, in writing out $z_{2,t}$ explicitly we can see how it represents the "equilibrium error" between

the two I(1) series, $x_{1,t}$ and $x_{2,t}$.

$$\begin{aligned} z_{2,t} &= \boldsymbol{\beta}' \boldsymbol{X}_t \\ &= \begin{bmatrix} p_{21}^* & p_{22}^* \end{bmatrix} \begin{bmatrix} x_{1,t} \\ \\ x_{2,t} \end{bmatrix} \\ &= p_{21}^* x_{1,t} + p_{22}^* x_{2,t} \end{aligned}$$

$$\Rightarrow p_{21}^* x_{1,t} = -p_{22}^* x_{2,t} + z_{2,t} \tag{67a}$$

where $z_{2,t}$ is like a stationary residual. That is, the individual I(1) series, $x_{1,t}$ and $x_{2,t}$, never wander "too far" from their long-run equilibrium relation defined in $\boldsymbol{\beta}$.

- **The VECM(p) representation**
- Throughout the entire earlier exposition we had assumed an underlying VAR(1) model.
- What if we want to work with a VAR(p) model instead? What is the VECM(p) representation?

- Take the **VAR(p) model**:

$$\boldsymbol{X}_t = \boldsymbol{\Phi}_1 \boldsymbol{X}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{X}_{t-p} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon), \tag{68}$$

where all of the elements of $\boldsymbol{X}_t$ are I(1).

- This can be rewritten in **VECM(p) model** form as:

$$\Delta \boldsymbol{X}_t = \boldsymbol{\Pi} \boldsymbol{X}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{X}_{t-1} + \boldsymbol{\Gamma}_2 \Delta \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{X}_{t-(p-1)} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon), \tag{69}$$

where all of the elements of $\Delta \boldsymbol{X}_t$ are I(0).
Moreover, we have that $\boldsymbol{\Pi} = -(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)$ and $\boldsymbol{\Gamma}_j = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$.

- Notice that the VECM(p) actually only includes $p-1$ lagged $\Delta \boldsymbol{X}_{t-i}$ terms, not p lags.

- **Proof of expression** (69)

1. Start with expression (68). It can be written as $\boldsymbol{\Phi}(L)\boldsymbol{X}_t = \boldsymbol{\epsilon}_t$, where $\boldsymbol{\Phi}(L) = \boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j L^j$.
2. We can rewrite $\boldsymbol{\Phi}(L)$ as $\boldsymbol{\Phi}(L) = \boldsymbol{\Phi}(1)L + \boldsymbol{\Phi}(L) - \boldsymbol{\Phi}(1)L$ so let $\boldsymbol{\Phi}(L) - \boldsymbol{\Phi}(1)L = (1-L)\boldsymbol{\Gamma}(L)$, where $\boldsymbol{\Gamma}(L) = \boldsymbol{I} - \sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j L^j$
3. Therefore, we have that $\dfrac{\boldsymbol{\Phi}(L) - \boldsymbol{\Phi}(1)L}{1-L} = \boldsymbol{\Gamma}(L) = \boldsymbol{I} - \displaystyle\sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j L^j$.
4. However, $\boldsymbol{\Phi}(L) - \boldsymbol{\Phi}(1)L = \boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j L^j - (\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)L = (\boldsymbol{I} - L) + \sum_{j=2}^{p} \boldsymbol{\Phi}_j (L - L^j)$.

5. Therefore from above, $\frac{\boldsymbol{\Phi}(L) - \boldsymbol{\Phi}(1)L}{1-L} = \boldsymbol{I} + \sum_{j=2}^{p} \boldsymbol{\Phi}_j \frac{(L - L^j)}{1-L}$.
6. But $\frac{(L - L^j)}{1-L} = L \sum_{i=0}^{j-2} L^i$.
7. Thus we have that $\boldsymbol{\Gamma}(L) = \boldsymbol{I} - \sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j L^j = \boldsymbol{I} + \sum_{j=2}^{p} \boldsymbol{\Phi}_j \frac{(L - L^j)}{1-L} = \boldsymbol{I} + \sum_{j=2}^{p} \boldsymbol{\Phi}_j L \sum_{i=0}^{j-2} L^i$.
8. Gathering terms on the RHS by L, L^2, ..., up to L^p we have that $\boldsymbol{\Gamma}_j = -\sum_{j=i+1}^{p} \boldsymbol{\Phi}_i$, for $j = 1, \ldots, (p-1)$ which gives us an expression for each of the $\boldsymbol{\Gamma}_j$ coefficient matrices in terms of the VAR(p) coefficient matrices.
9. Next, go back to (2) above, which implies that $\boldsymbol{\Phi}(L)\boldsymbol{X}_t = (\boldsymbol{\Phi}(1)L + \boldsymbol{\Phi}(L) - \boldsymbol{\Phi}(1)L)\boldsymbol{X}_t = (\boldsymbol{\Phi}(1)L + (1-L)\boldsymbol{\Gamma}(L))\boldsymbol{X}_t = \boldsymbol{\epsilon}_t$.
10. The above expression implies that $\boldsymbol{\Phi}(1)\boldsymbol{X}_{t-1} + \boldsymbol{\Gamma}(L)\Delta\boldsymbol{X}_t = \boldsymbol{\epsilon}_t$.
11. Finally we therefore have $\Delta\boldsymbol{X}_t = -\boldsymbol{\Phi}(1)\boldsymbol{X}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j \Delta\boldsymbol{X}_{t-j} + \boldsymbol{\epsilon}_t$ which proves the result.

- **Estimating the VECM(p) model**
- Suppose that we have a VECM(p) model but where the vector $\boldsymbol{X}_t$ is not bivariate. That is, suppose that $\boldsymbol{X}_t$ is $n \times 1$ so there are n individual I(1) variables included in it.
- All the theory above still applies. In determining the number of co-integration relations in this case, let's take a look back at the three cases again and consider the eigenvalues of the $\boldsymbol{\Pi}$ matrix.
- Recall from our earlier bivariate example VAR(1) example that if $|\lambda_1| = 1$ but $|\lambda_2| < 1$, we have:

$$\Delta\boldsymbol{X}_t = \boldsymbol{P} \begin{bmatrix} 0 & 0 \\ 0 & (\lambda_2 - 1) \end{bmatrix} \boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t \tag{63}$$

  and so $\boldsymbol{\Pi}$ was a rank 1 matrix since one of the eigenvalues of the $\boldsymbol{\Pi}$ matrix was now 0, and this situation implied there existed exactly one co-integrating relation.
- We can easily generalize this type of analysis to the VAR(p) (i.e. VECM(p)) case with n dimensional $\boldsymbol{X}_t$ as:

$$\Delta\boldsymbol{X}_t = \boldsymbol{\Pi}\boldsymbol{X}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j \Delta\boldsymbol{X}_{t-j} + \boldsymbol{\epsilon}_t \tag{70}$$

$$\Rightarrow \Delta\boldsymbol{X}_t = \boldsymbol{P} \begin{bmatrix} \gamma_1 & 0 & 0 & \ldots \\ 0 & \gamma_2 & 0 & \ldots \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & 0 & \gamma_n \end{bmatrix} \boldsymbol{P}^{-1}\boldsymbol{X}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Gamma}_j \Delta\boldsymbol{X}_{t-j} + \boldsymbol{\epsilon}_t \tag{71a}$$

where the γ_i's are the eigenvalues of the $\mathbf{\Pi}$ matrix.

- Be careful! The eigenvalues of $\mathbf{\Pi}$ are not in general the sum of the eigenvalues of the matrices in $-\mathbf{\Phi}(1) = -(\boldsymbol{I} - \sum_{j=1}^{p} \mathbf{\Phi}_j)$; that is we do not generally have $\gamma_i = \sum_{j=1}^{p} \lambda_{j,i} - 1$. [9]

- Nevertheless, we can still make use of the same type of analysis since for each element that is zero along the main diagonal, the matrix $\mathbf{\Pi}$ loses column rank. This suggests that a reasonable test of how many co-integrating relations exist is to test for the column rank of the $\mathbf{\Pi}$ matrix.

- Therefore we can be more formal. Specifically, we say that there are again three cases, analogous to the VAR(1) discussion above:

  1. $\gamma_i = 0$ for all $i = 1, \ldots, n$. In this case we have $Rank(\mathbf{\Pi}) = 0$ (i.e. $\mathbf{\Pi} = \mathbf{0}$) and there exist no co-integrating relations. In other words, $\boldsymbol{X}_t$ is driven by n different stochastic trends that each move indepedently of each other.
  2. $\gamma_i \neq 0$ for all $i = 1, \ldots, n$. In this case we have that $Rank(\mathbf{\Pi}) = n$ and there are no stochastic trends, so there can be no co-integrating relations between them.
  3. $\gamma_i = 0$ in $n - r$ cases, and $\gamma_i \neq 0$ for $r < n$ cases. We now have that $Rank(\mathbf{\Pi}) = r$ which implies there exist r co-integrating relations, and $n - r$ stochastic trends.

- It is the third case that is most interesting.

  - ◇ It suggests that r linearly independent co-integrating relations exist and that we can construct r many I(0) series by multiplying each co-integrating relation by $\boldsymbol{X}_t$.
  - ◇ That is, we have that there exist r vectors, $\boldsymbol{c}_i$ for $i = 1, \ldots, n - r$, where each $z_{i,t} = \boldsymbol{c}_i \boldsymbol{X}_t$ is I(0).

- The co-integrating relations in this case are therefore just the respective rows of the $\boldsymbol{P}^{-1}$ matrix.

  - ◇ That is, suppose that we have that $\gamma_i \neq 0$ for $i = 1, 5$ and 7. Therefore rows 1, 5, and 7 of the $\boldsymbol{P}^{-1}$ matrix are the co-integrating relations.
  - ◇ Since $Rank(\mathbf{\Pi}) = r$ we can factorize $\mathbf{\Pi}$ as follows:

    $$\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$$

    where now $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both $n \times r$ matrices. $\boldsymbol{\beta}$ includes all the relevant co-integrating relations as its columns, and $\boldsymbol{\alpha}$ contains the loading factors.

- **Estimation of the VECM(p)**

- Estimation of the VECM(p) in principle can be done by maximum likelihood (or OLS) in the same as we discussed for the VAR(p) model.

- Again, recall that the **VECM(p) model** has the form:

$$\Delta \boldsymbol{X}_t = \mathbf{\Pi}\boldsymbol{X}_{t-1} + \mathbf{\Gamma}_1 \Delta \boldsymbol{X}_{t-1} + \mathbf{\Gamma}_2 \Delta \boldsymbol{X}_{t-2} + \cdots + \mathbf{\Gamma}_{p-1} \Delta \boldsymbol{X}_{t-(p-1)} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \mathbf{\Sigma}_\epsilon), \tag{69}$$

[9] The only time this holds is the special case where all of the coefficient matrices in $-\mathbf{\Phi}(1)$ have the same exact eigenvectors since we have that for any i and j, $(\mathbf{\Phi}_i + \mathbf{\Phi}_j)\boldsymbol{c} = \mathbf{\Phi}_i \boldsymbol{c} + \mathbf{\Phi}_j \boldsymbol{c} = \lambda_i \boldsymbol{c} + \lambda_j \boldsymbol{c} = (\lambda_i + \lambda_j)\boldsymbol{c}$ in this case.

where all of the elements of $\Delta \boldsymbol{X}_t$ are I(0).

Moreover, we have that $\boldsymbol{\Pi} = -(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j)$ and $\boldsymbol{\Gamma}_j = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$.

- However, in this case $\boldsymbol{\Pi}$ may of course be of reduced rank if there exist co-integrating relations.
  - If this is the case, we wish to not only estimate the $\boldsymbol{\Gamma}_j$'s and $\boldsymbol{\Sigma}_\epsilon$ of the model, but we wish to *seperately identify* both the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$.
- A method of dealing with this type of situation is called *reduced rank regressions*.
- The following *Johansen method* for estimating both the rank of $\boldsymbol{\Pi}$ and the co-integrating relations $\boldsymbol{\beta}$ is based on the reduced rank regressions method.

- **Johansen method**
- The Johansen method is as follows, given a VECM(p) model:

1. Estimate the corresponding VAR(p) in first differences. That is, estimate:
$$\Delta \boldsymbol{X}_t = \boldsymbol{A}_1 \Delta \boldsymbol{X}_{t-1} + \boldsymbol{A}_2 \Delta \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{A}_{p-1} \Delta \boldsymbol{X}_{t-(p-1)} + \boldsymbol{e}_t \tag{72}$$
to obtain residuals $\boldsymbol{e}_t$.
2. Next, regression $\boldsymbol{X}_{t-1}$ on the lagged differences. That is, estimate:
$$\boldsymbol{X}_{t-1} = \boldsymbol{C}_1 \Delta \boldsymbol{X}_{t-1} + \boldsymbol{C}_2 \Delta \boldsymbol{X}_{t-2} + \cdots + \boldsymbol{C}_{p-1} \Delta \boldsymbol{X}_{t-(p-1)} + \boldsymbol{u}_t \tag{73}$$
to obtain residuals $\boldsymbol{u}_t$.
3. Compute the squares of the so called *canonical correlations* betweeb $\boldsymbol{u}_t$ and $\boldsymbol{e}_t$.
   - We will denote the squares of the canonical correlations by δ_i.
   - It turns out that the δ_i's can be obtained as the solution to what is called the *generalized eigenvalue problem*.[10]
   - That is, the δ_i's are the solutions that solve:
   $$det(\boldsymbol{S}_{ue}\boldsymbol{S}_{ee}^{-1}\boldsymbol{S}_{eu} - \delta \boldsymbol{S}_{uu}) = \boldsymbol{0}$$
   where $\boldsymbol{S}_{eu} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{e}_t \boldsymbol{u}_t'$.
   - Therefore, the δ_i's are the eigenvalues of the generalized eigenvalue problem where $\boldsymbol{S}_{eu} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{e}_t \boldsymbol{u}_t'$.
     - Notice that $\boldsymbol{S}_{eu}$ is the sample covariance matrix between the residuals $\boldsymbol{e}_t$ and $\boldsymbol{u}_t$.
     - It turns out that the eigenvalues δ_i are also the eigenvalues of $\boldsymbol{\Pi}$.
4. The eigenvectors, $\boldsymbol{c}_i$ associated with each eigenvalue of the generalized problem above, are therefore co-integrating relations (conditional on us determining the rank of $\boldsymbol{\Pi}$.)
5. Determining the rank of $\boldsymbol{\Pi}$:

[10]The generalized eigenvalue problem is just a more general version of $\boldsymbol{A}\boldsymbol{c} = \lambda \boldsymbol{c}$. Rather we now have $\boldsymbol{A}\boldsymbol{c} = \lambda \boldsymbol{B}\boldsymbol{c}$.

◇ Since the rank of $\mathbf{\Pi}$ is determined by the number of non-zero eigenvalues, δ_i, we can test the rank by means of the *trace* and *maximum eigenvalue* statistics which are given as follows:

$$f_{trace}(r) = -T \sum_{i=r+1}^{n} \ln\left(1 - \hat{\delta}_i\right) \tag{74a}$$

$$f_{max}(r, r+1) = -T \ln\left(1 - \hat{\delta_{r+1}}\right) \tag{74b}$$

where the eigenvalues are ordered such that $1 \geq \delta_n \geq \delta_{n-1} \geq \cdots \geq \delta_1 \geq 0$.

- ∗ Notice that as these statistics get larger, the less likely it is that the eigenvalues δ_i are all equal to zero.
- ∗ More specially, we have that the trace statistic tests the null hypothesis that the number of distinct co-integrating relations is less than or equal to r against a general alternative. Clearly, if all δ_i's are zero we have that the trace statistic should be zero.
- ∗ The maximum eigenvalue statistic tests the null that the number of co-integrating relations is r versus the alternative that it is $r + 1$.

◇ Critical values for these test statistics follow non-standard distributions. You can use the "urca" package in R to do these tests. The routines therein include the critical values in their output.

◇ Therefore, in testing for the number of co-integrating relations, you should start with the null hypothesis that $H_0 : r = 0$. If you are able to reject this, then try $H_0 : r = 1$, etc, until you can no longer reject.

◇ Once you have established the number of co-integrating relations, r, you obtain the co-integrating relations as the eigenvectors, $\boldsymbol{c}_i$, from the previous step, each associated with the $\delta_1, \delta_2, \ldots, \delta_r$ you have selected.

6. Finally, once you have the eigenvectors decided upon, you can restimate the VECM(p) by OLS or MLE given the following expression:

$$\Delta \boldsymbol{X}_t = \boldsymbol{\alpha} \boldsymbol{z}_{t-1} + \mathbf{\Gamma}_1 \Delta \boldsymbol{X}_{t-1} + \mathbf{\Gamma}_2 \Delta \boldsymbol{X}_{t-2} + \cdots + \mathbf{\Gamma}_{p-1} \Delta \boldsymbol{X}_{t-(p-1)} + \boldsymbol{\epsilon}_t \tag{75}$$

where $\boldsymbol{z}_{t-1} = \hat{\boldsymbol{\beta}}' \boldsymbol{X}_{t-1}$. Since you now "know" $\boldsymbol{\beta}$, the model is identified.