1 What is a Time Series?

A time series is a realization of a sequence of a variable indexed by time. The notation we will use to denote this is $x_t$, $t = 1, 2, \ldots, T$. A variable is said to be “random” if its realizations are stochastic. Unlike cross-sectional data, time series data can typically not be modeled as independent across observations (i.e. independent across time). Rather, time series are persistent in the sense that observations of random variables are typically positively correlated across time. Most of what we do in macro involves variables with such dependence across time.

Before discussing the unique issues presented by time series data, we quickly review expectations and first and second moments.

2 Expectations

The expected value of $x_t$ is denoted by $E(x_t)$ and is the weighted average of possible realizations of $x_t$, where the weights are the probabilities of each realization:

$$E(x_t) = \sum_{t=1}^{T} x_t p(x_t)$$ (1)

This is a linear operator. As such, it has the following properties, where $a$ is a constant:

$$E(a) = a$$ (2)

$$E(ax_t) = aE(x_t)$$ (3)

$$E(x_t + y_t) = E(x_t) + E(y_t)$$ (4)

Non-linear operators cannot “go through” an expectation operator:
\begin{align*}
  E(x_t y_t) & \neq E(x_t)E(y_t) \\
  E(g(x_t)) & \neq g(E(x_t))
\end{align*}

We are often interested in \textit{conditional} expectations, which are expectations taken conditional on some information. Letting $I_t$ denote the information set available at time $t$, the conditional expectation of a random variable can be written: $E(x_t \mid I_t)$. It is common, however, to use the shorthand notation $E_t(x_t)$ to refer to the expectation of $x_t$ conditional on information available at time $t$.

For two arbitrary random variables $y$ and $z$, the Law of Iterated Expectations says that $E(y) = E(E(y \mid z))$. In words, the unconditional expectation of the conditional expectation of $y$ conditional on $z$ is equal to the unconditional expectation of $y$. This has the following implication for a time series:

\begin{equation}
  E_t(E_{t+1}(x_{t+2})) = E_t x_{t+2}
\end{equation}

In other words, your current best guess of your best guess next period of the realization of $x$ two periods from now is equal to your current best guess of $x$ two periods from now.

### 3 Higher Moments

The expectation is sometimes called the “first moment”. We are often interested in “higher” moments, and in particular second moments. The variance is defined as:

\begin{equation}
  \text{var}(x_t) = E(x_t - E(x_t))^2
\end{equation}

In words, the variance is equal to the expected (or average) squared deviation of $x_t$ about its mean. The standard deviation is the square root of the variance. The variance can also be written:

\begin{equation}
  \text{var}(x_t) = E(x_t^2) - (E(x_t))^2
\end{equation}

For mean zero random variables (such as white noise processes; see below) the variance will just be equal to $E(x_t^2)$. The following are properties of variance:

\begin{align*}
  \text{var}(a) & = 0 \\
  \text{var}(ax_t) & = a^2 \text{var}(x_t) \\
  \text{var}(x_t + y_t) & = \text{var}(x_t) + \text{var}(y_t) + 2\text{cov}(x_t, y_t)
\end{align*}

The covariance is a measure of the linear relationship between two random variables:
\[ \text{cov}(x_t, y_t) = E((x_t - E(x_t))(y_t - E(y_t))) = E(x_t y_t) - E(x_t)E(y_t) \] (13)

We say that two random variables are independent if knowing the realization of one of the variables does not alter one’s expectation for the other. Mathematically, this means that \( E(x_t \mid y_t) = E(x_t) \). If two variables are independent, then \( \text{cov}(x_t, y_t) = 0 \). The converse is not true – the covariance being zero does not imply two series are independent, since the dependence could be non-linear.

The following are some properties of covariance:

\[
\begin{align*}
\text{cov}(x_t, c) &= 0 \\
\text{cov}(x_t, x_t) &= \text{var}(x_t) \\
\text{cov}(x_t, y_t) &= \text{cov}(y_t, x_t) \\
\text{cov}(ax_t, by_t) &= ab\text{cov}(x_t, y_t)
\end{align*}
\]

The units of covariance depend on the units of the underlying series. So, for example, \( \text{cov}(x_t, y_t) > \text{cov}(z_t, v_t) \) does not imply that \( x \) and \( y \) are “more” strongly related than \( z \) and \( v \). The correlation coefficient provides a means by which to make such statements. It is equal to the covariance divided by the product of the standard deviations, and is bound between -1 and 1:

\[ \text{corr}(x_t, y_t) = \frac{\text{cov}(x_t, y_t)}{\sqrt{\text{var}(x_t)}\sqrt{\text{var}(y_t)}} \] (14)

4 Markov Processes

There are two common ways to model time series: as Markov processes or as autoregressive moving average processes built on white noise. The so-called “Markov Property” says that the current state of the system (i.e. \( x_t \)) is a sufficient statistic to form forecasts about the future of the system. In other words, knowing \( x_{t-j} \) for \( j > 0 \) provides no additional information on future values \( x_{t+j} \) for \( j > 0 \) that \( x_t \) does not.

Let \( \bar{x} \) be a \( n \times 1 \) vector of possible realizations of \( x_t \). Let \( P \) be \( n \times n \) matrix known as a probability or transition matrix. Its elements are the probabilities of transitioning from state \( i \) to state \( j \) between periods \( t \) and \( t + 1 \). Hence:

\[ P_{i,j} = \text{prob}(x_{t+1} = \bar{x}_j \mid x_t = \bar{x}_i) \] (15)

In words, the row tells you the current state, and the column tells you probabilities of transitioning to each possible state in the next period. As such, the rows must sum to one (i.e. the system has to transition to some value next period).
5 ARMA Processes

In contrast to Markov processes, which are discrete, autoregressive moving average (ARMA) processes are typically continuous. The building block of an ARMA process is a white noise process. A white noise process has zero mean, constant variance, and is uncorrelated across time. Let $\varepsilon_t$ be a white noise process. This means it has the following properties:

$$
E(\varepsilon_t) = 0
$$
$$
\text{var}(\varepsilon_t) = \sigma^2
$$
$$
\text{cov}(\varepsilon_t, \varepsilon_{t+j}) = 0 \quad \forall j
$$

Given a white noise process, an ARMA($p$, $q$) process is:

$$
x_t = a + \rho_1 x_{t-1} + \rho_2 x_{t-2} + \ldots + \rho_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \quad (16)
$$

Use of lag operators makes working with ARMA processes significantly easier. Formally, the lag operator, $L$, is defined such that $L^j x_t = x_{t-j}$. Hence, $L x_t = x_{t-1}$, $L^2 x_t = x_{t-2}$, and so on. We can write a lag polynomial of order $p$ as:

$$
a(L) = a_0 L^0 + a_1 L^1 + \ldots + a_p L^p \quad (18)
$$

As such, we can write an ARMA($p,q$) process as:

$$
a(L)x_t = b(L)\varepsilon_t \quad (19)
$$

Using the notation above, $a(L) = 1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_p L^p$ and $b(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q$.

The nice thing about ARMA processes is that they are not unique. In particular, we can (usually, at least) “invert” them to express the AR component as MA terms and vice versa. We’ll show this via an example. Consider an AR(1) process:

$$
x_t = \rho x_{t-1} + \varepsilon_t \quad (20)
$$

This can be written in lag operator notation as:

$$
(1 - \rho L)x_t = \varepsilon_t \quad (21)
$$

In terms of the notation above, $a(L) = (1 - \rho L)$, a lag polynomial of order 1. We can invert this as follows:
\[ x_t = (1 - \rho L)^{-1} \varepsilon_t \]  \hspace{1cm} (22)

Now why is this helpful? Recall a certain trick for infinite sums. Suppose that \(| \alpha | < 1\). Let:

\[ S = 1 + \alpha + \alpha^2 + \ldots \]  \hspace{1cm} (23)

Manipulate and solve:

\[ \alpha S = \alpha + \alpha^2 + \alpha^3 + \ldots \]

\[ S - \alpha S = 1 \]

\[ S = \frac{1}{1 - \alpha} \]

I can make this substitution because I assumed \(| \alpha | < 1\). Letting \(\rho L\) play the role of \(\alpha\), I see that \((1 - \rho L)^{-1} = 1 + \rho + \rho^2 + \ldots\). Therefore, I can write the AR(1) as an MA(\(\infty\)), with \(b(L) = a(L)^{-1}\):

\[ x_t = b(L) \varepsilon_t = (1 + \rho L + \rho^2 L^2 + \ldots) \varepsilon_t \]  \hspace{1cm} (24)

\[ x_t = \sum_{j=0}^{\infty} \rho^j \varepsilon_{t-j} \]  \hspace{1cm} (25)

We can also go from MA(1) to AR(\(\infty\)) as well. Suppose we have:

\[ x_t = \varepsilon_t + \theta \varepsilon_{t-1} \]  \hspace{1cm} (26)

\[ x_t = (1 + \theta L) \varepsilon_t \]  \hspace{1cm} (27)

\[ (1 + \theta L)^{-1} x_t = \varepsilon_t \]  \hspace{1cm} (28)

By the same logic used above, \((1 + \theta L)^{-1} = 1 - \theta L - \theta^2 L^2 - \ldots\). Hence, we can write the MA(1) process as:

\[ \sum_{j=0}^{\infty} -\theta^j x_{t-j} = \varepsilon_t \]  \hspace{1cm} (29)

This is an important result. It means that we can estimate MA processes by approximating them as AR processes. As long as the number of lags, \(p\), is sufficiently high and the “roots” of \(b(L)\) are not too close to one (basically that means that the coefficients decay from one pretty quickly), estimating an AR process will provide a reliable approximation to the MA process. We care about this because dynamic economic models often have MA representations, and MA terms are “hard” to estimate, whereas AR processes can be estimated consistently via OLS.

An “impulse response function” is the change in the current and expected future values of a
random variable conditional on the realization of uncertainty at some point in time. Formally:

\[ \text{IRF} = E_t x_{t+j} - E_{t-1} x_{t+j} \forall j \geq 0 \quad (30) \]

Suppose that we have an ARMA process, which we can write as \( x_t = b(L) \varepsilon_t \). Applying the expectations operator to the process, we have \( E_t x_{t+j} = b(L) E_t \varepsilon_{t+j} = b(L) \varepsilon_t \). This follows from the fact that \( E_t \varepsilon_{t+j} = 0 \forall j > 0 \). By the same logic, then \( E_{t-1} \varepsilon_{t+j} = 0 \forall j \geq 0 \). Hence, the impulse response function, using the definition above, is just \( b(L) \varepsilon_t \). In other words, the MA representation and the impulse response function are the same thing.

This presentation has all been in terms of scalars. ARMA processes apply equally well to vectors and matrices. The typical notation for vectors and matrices is capital letters; e.g. \( X_t \) is a vector, and \( A(L) \) is matrix lag polynomial. But the basic stuff is all the same.

6 Econometrics

Suppose we estimate a classical linear regression, where \( Y_t \) is a \( T \times 1 \) vector and \( X_t \) is \( T \times k \), which includes a constant:

\[ Y_t = X_t \beta + \varepsilon_t \quad (31) \]

The OLS estimator is:

\[
\hat{\beta} = (X'_t X_t)^{-1} X'_t Y_t \\
\hat{\varepsilon}_t = Y_t - X_t \hat{\beta} \\
\sigma^2 = \frac{1}{T-k} \hat{\varepsilon}'_t \hat{\varepsilon}_t \\
\text{var}(\hat{\beta}) = \sigma^2 (X'_t X_t)^{-1}
\]

The fact that the observations are indexed by \( t \) as opposed to \( i \) presents no special problems per se, as long as all of the variables are stationary, in a sense to be defined below. But what are the properties of the OLS estimator here?

We say that an estimator is unbiased if \( E(\hat{\beta}) = \beta \). This requires that \( \varepsilon_t \) and \( X_t \) are independent. In a time series context, this assumption is likely to be violated. To be independent, it must be the case that \( E(X_{t+j} | \varepsilon_t) = E(X_{t+j}) \). This is unlikely to happen, particularly if a lagged dependent variable appears in \( X_t \) (which would happen if you were estimating an AR term). This is because a positive realization of \( \varepsilon_t \) likely means, other things being equal, that you expect more positive realizations of \( Y_t \) in the future. If lags of \( Y_t \) are in \( X_t \), then the independence assumption will fail.

What about consistency? An estimator is consistent if, \( \lim_{T \to \infty} E(\hat{\beta}) = \beta \) and \( \lim_{T \to \infty} \text{var}(\hat{\beta}) = 0 \). This only requires that \( X_t \) is uncorrelated with the current realization of \( \varepsilon_t \), which is likely to hold. Hence, OLS slope estimates of time series regressions with stationary variables will typically
produce consistent estimates, though biased. This of course assumes that there are not other sources of endogeneity. Construction of “correct” standard errors is a little more complicated; but we won’t worry about that for now.

7 Stationarity, Trends, and Unit Roots

Many macroeconomic series have trends – for example, output. A trend means that a series grows or declines over time. This presents some unique challenges.

A related concept is that of stationarity. We say that a time series is stationary if it has a time-invariant mean, variance, and autocovariance. Formally:

\[ E(x_t) = E(x_{t-j}) = \mu \quad \forall \quad j \]  
\[ \text{var}(x_t) = \text{var}(x_{t-j}) = \sigma^2 \quad \forall \quad j \]  
\[ \text{cov}(x_t, x_{t-j}) = \gamma_j \quad \forall \quad t \]  

A series that has a trend clearly is not going to be stationary according to our definition. Why? If a series is growing or declining at a non-zero rate, then its mean is going to be different depending on when you look at it – i.e. the unconditional expectation of GDP in the 1950s is much different than the unconditional mean of GDP in the 1990s. That being said, a series that can also be non-stationary without having a trend (a random walk). We’ll return to this in a minute.

This all matters for a couple of reasons. Macroeconomists are often interested in “business cycle” statistics – things like standard deviations (volatility) and correlations (co-movement). These statistics aren’t well defined for non-stationary series. Secondly, standard econometric inference (i.e. construction of standard errors and hypothesis tests) breaks down when series become non-stationary. As such, we are often interested in transformations that render data stationary. Unfortunately, what is the appropriate transformation depends on what drives the non-stationarity.

Conventional wisdom in macro (at least until the 1980s) was that most macro series were trend stationary. A trend stationary series features some kind of deterministic time trend plus a stochastic component which is stationary. An example specification for log output, say, \( y_t \), would be:

\[ y_t = \Delta t + u_t \]  
\[ u_t = \rho u_{t-1} + \varepsilon_t \quad 0 < \rho < 1 \]  

Here \( \Delta t \) is the trend specification, with \( a \) equal to the trend growth rate. \( u_t \) represents deviations about the trend (i.e. the “cycle”), and \( \varepsilon_t \) is a random business cycle disturbance. These shocks have effects on \( y_t \) which eventually die out. If \( a = 0 \), then \( y_t \) would be stationary.

Another specification that is very different but turns out to the be nearly observationally equivalent is the random walk with drift. Take the same process but let \( \rho \) = 1. Make this substitution
into (35) and simplify:

\[ y_t = at + u_{t-1} + \varepsilon_t \]

\[ y_t = at + (y_{t-1} - a(t - 1)) + \varepsilon_t \]

⇒

\[ y_t = a + y_{t-1} + \varepsilon_t \] (37)

The second step follows by lagging (35) one period and solving for \( u_{t-1} \). Here, \( a \) is still the trend growth rate (remember we are working in logs, so first differences are growth rates). (37) is called a random walk with drift. It is a special case of (35)-(36) when \( \rho = 1 \).

The non-stochastic version of the deterministic trend and random walk with drift processes are identical (up to a constant): \( y_t^{ns} = at \). But the series have very different economic implications. For the deterministic trend case, stochastic shocks have transitory effects. For the random walk with drift case, shocks have effects that don’t die out. Even if \( a = 0 \) here, the random walk series is still non-stationary. Sometimes this kind of process is called a “stochastic trend” model, as opposed to “deterministic trend” above.

Formally, a process is said to have a “unit root” if one of the roots of its characteristic equation is one. For the specification given above it means that the AR(1) coefficient is 1. A good rule of thumb for more complicated processes is this: if stochastic disturbances don’t have effects that eventually die out, then you have a unit root.

The process for producing a stationary series from either of these series with trends is different. For the deterministic trend case, you would (i) estimate an OLS regression of a time dummy and (ii) take the residuals. You would be left with \( u_t \), which is stationary. For the random walk with drift case, removing a deterministic trend would not render the series stationary (verify this on your own). Here, you would want to first difference the series – i.e. construct \( \Delta y_t = y_t - y_{t-1} \). First differencing the trend stationary series will yield a stationary series, but it won’t correspond exactly to \( u_t \). Hence, if you think the series is a random walk with drift, difference it. If you think the series is trend stationary, remove a deterministic time trend (e.g. linear, quadratic, etc.).

8 Unit Root Econometrics

Unit roots are problematic for econometrics because the usual assumptions about error terms break down. If you have a deterministic trend series, you would remove the deterministic trend and could do econometrics (OLS, etc..) on the cycle component just as in Section 6 above. But if you think your series may have a stochastic trend, then removing the deterministic trend doesn’t remove the unit root.

Not removing trends can be problematic. Suppose that we have two independent random walks (say \( x_t \) and \( z_t \)). Regressing \( x_t \) on \( z_t \) will tend to produce spurious results, in the sense that you
will find regression coefficients that are “significantly” different from zero when doing conventional OLS-based hypothesis tests. Put differently, you’re likely to find relationships in the data where there truly are none. What should one do to deal with this issue? You should either detrend or difference each series (whichever is appropriate), provided the series are not cointegrated, which we come back to in a minute. For example, if both \( x_t \) and \( z_t \) have stochastic trends, the appropriate thing to do is to difference each series, and then regress \( \Delta x_t \) on \( \Delta z_t \). This is known as the “spurious regression” problem.

Because of all this, it is important to pre-test variables to see if there are unit roots, so that we know how to render them stationary. In practice, this is difficult to do. The intuition is straightforward. \( x_t = x_{t-1} + \varepsilon_t \) is non-stationary, whereas \( x_t = 0.9999x_{t-1} + \varepsilon_t \) is stationary. In finite samples, it’s essentially impossible to differentiate 0.9999 from 1.

But we proceed anyway. Much of what follows is based on Dickey and Fuller (1979). Take the process given by (35)-(36), but allow for the possibility that \( \rho = 1 \). Substitute (36) into (35):

\[
y_t = at + \rho u_{t-1} + \varepsilon_t
\]

Now eliminate \( u_{t-1} \) by lagging this equation one period:

\[
y_t = at + \rho(y_{t-1} - a(t-1)) + \varepsilon_t
\]

Now first difference the equation (i.e. subtract \( y_{t-1} \) from both sides) and simplify:

\[
\Delta y_t = \rho a + (1 - \rho)at + (\rho - 1)y_{t-1} + \varepsilon_t
\]  \hspace{1cm} (38)

Consider running the following regression based on (38): \( \Delta y_t = \alpha_0 + \beta t + \gamma y_{t-1} + \varepsilon_t \). The null hypothesis that the series is stationary about a deterministic trend corresponds to \( \gamma < 0 \) and \( \beta \neq 0 \). The null hypothesis that the series follows a random walk with drift corresponds to \( \gamma = 0 \) and \( \beta = 0 \). The null hypothesis that the series follows a random walk with no drift corresponds to \( \gamma = \beta = \alpha_0 = 0 \).

You can estimate this regression by OLS under the null of a unit root (i.e. under the null that \( \gamma = 0 \)), but inference is not standard. Dickey and Fuller (1979) did some Monte Carlo experiments to numerically construct critical values for a \( t \) test (testing the hypothesis that \( \gamma = 0 \)) and \( F \) tests (testing joint hypotheses that \( \gamma = \beta = 0 \) or \( \gamma = \beta = \alpha_0 = 0 \)). The critical values are much more “stringent” than under normal circumstances; for example, to reject \( \gamma = 0 \) at 95 percent significance you would need a \( t \) statistic of -3.45.

If one is interested in estimating regressions on non-stationary variables, one should proceed as follows:

- Pre-test the variables using something like a Dickey-Fuller test to see if the series are (i) stationary, (ii) are stationary about a deterministic trend, or (iii) are unit roots (i.e. have
stochastic trends)

- If (i), estimate regressions and do inference as normal
- If (ii), then fit deterministic trends to the series and then estimate regressions on the detrended series and do inference as normal
- If (iii), first difference the series and then estimate regressions and do inference as normal

9 VARs

Vector autoregressions, or VARs, are the multivariate generalization of univariate ARMA processes. They are widely used as tools both for prediction and for model building and evaluation.

Formally, a VAR($p$) is a vector autoregression where there are $p$ autoregressive lags of each variable in each equation. Consider two variables, $x_t$ and $z_t$. Assume for now that these are stationary. In scalar form, the VAR can be written:

$$x_t = a_x + \sum_{j=1}^{p} \beta_{x,j}^1 x_{t-j} + \sum_{j=1}^{p} \beta_{x,j}^2 z_{t-j} + e_{x,t}$$

$$z_t = a_z + \sum_{j=1}^{p} \beta_{z,j}^1 x_{t-j} + \sum_{j=1}^{p} \beta_{z,j}^2 z_{t-j} + e_{z,t}$$

We could equivalently write this in matrix form, letting $Y_t = [x_t \ z_t]'$ and $e_t = [e_{x,t} \ e_{y,t}]'$:

$$A(L)Y_t = e_t \quad (39)$$

Here $A(L)$ is a matrix lag polynomial of order $p$: $A(L) = I - A_1 L - A_2 L^2 \ldots A_p L^p$, where, for example, $A_1$ is:

$$
\begin{pmatrix}
\beta_{x,1}^1 & \beta_{x,2}^1 \\
\beta_{z,1}^1 & \beta_{z,2}^1
\end{pmatrix}
$$

This system can be estimated consistently and efficiently by OLS. The reason for consistency is that all of the variables on the right hand sides of the equation are dated $t - 1$ or earlier, whereas the error terms are dated $t$. The error terms are sometimes called “innovations” – they are the forecast errors of a variable conditional on observing its past values and the past values of other variables. As such, they should not be predictable, meaning they should be uncorrelated with the right hand side variables. OLS turns out to be efficient because the variables on the right hand side of each equation are the same (otherwise seemingly unrelated regressions would be efficient). In terms of choosing how many lags to include on the right hand side, a year’s worth is relatively common – hence 4 lags for quarterly data, 12 for monthly, 1 for annual. You can use lag selection criteria – like the AIC or BIC – to come up with statistical tests for the optimal number of lags.
As a general matter, \( e_{x,t} \) and \( e_{z,t} \) will be correlated. For example, suppose that the variables in the VAR are stock prices and consumption. Whatever causes stock prices to rise (say a positive \( e_{x,t} \)) will probably cause consumption to rise, too (so \( e_{z,t} \) would also go up). Hence, there is no “structural” interpretation to these innovations. They are affected by something deeper. For the purposes of forecasting, one probably doesn’t care what is driving the innovations. For the purposes of economic analysis, one does. A VAR like the one above is often called “reduced form”, because one cannot give a structural interpretation to the innovations.

Structural VAR analysis presumes that the innovations in the VAR equations are driven by deeper structural shocks. In particular, using the example from above, let \( \epsilon_t = [\epsilon_{1,t} \quad \epsilon_{2,t}]' \). These are “structural shocks”, which are, by definition, uncorrelated with one another. Assume that there is a linear mapping between these structural shocks and the reduced form innovations:

\[
e_t = B\epsilon_t
\]

Let \( \Sigma_e \) be the variance-covariance matrix of the reduced form innovations (i.e. \( \Sigma_e = E(e_t e_t') \)). Manipulating the equation above, we have:

\[
E(e_t e_t') = BE(\epsilon_t \epsilon_t')B'
\]

Because they are uncorrelated, the off-diagonal elements of \( E(\epsilon_t \epsilon_t') \) are zero. We can normalize the variance of each structural shock to be unity, which means that \( E(\epsilon_t \epsilon_t') = I \). The above equation then becomes:

\[
\Sigma_e = BB'
\]

This turns out to be a system of equations that, without some assumptions, is under-determined. For a two variable system, for example, there are 4 unique elements of \( B \), and hence four unique elements of \( BB' \). But there are only three unique elements of \( \Sigma_e \), since a variance covariance matrix is symmetric. Hence, without imposing an assumption on \( B \), it cannot be identified. As a general matter, if there are \( n \) total variables in the VAR, then there are going to be \( n \) innovations. The variance-covariance matrix will have \( n^2 \) elements. But only \( \frac{n^2+n}{2} \) of these will be unique. Since \( BB' \) will have \( n^2 \) elements, you need to impose \( n^2 - \frac{n^2+n}{2} \) restrictions on \( B \) to have the system be determined, which is \( \frac{n(n-1)}{2} \) restrictions. In the case of \( n = 2 \) considered above, this is just one restriction.

The restrictions used hopefully come from economic theory. The most common restrictions are recursive. This imposes timing assumptions – some shocks only affect some variables with a delay. Put differently, some of the elements of \( B \) are zero. Suppose that one believes \( \epsilon_{2,t} \) doesn’t affect \( x_t \) immediately in our two variable example. This would mean that the (1, 2) element of \( B \) would
be restricted to be zero. Given this restriction, the remaining 3 elements of \( B \) could be identified from the variance-covariance matrix of residuals.

Are these kind of recursive restrictions likely to be “true” in the real world? Maybe. It is common in the monetary policy literature that monetary policy shocks affect the economy with a delay. Suppose that you estimated a VAR with the fed funds rate and real GDP growth: 

\[
Y_t = \begin{bmatrix} i_t & \Delta y_t \end{bmatrix}'.
\]

You want to interpret \( \epsilon_{1,t} \) as a monetary policy shock, and \( \epsilon_{2,t} \) as a “supply” shock. Under this timing assumption, the \((2, 1)\) element of \( B \) would be zero – GDP growth doesn’t react to \( \epsilon_{1,t} \). In contrast, the fed funds rate is allowed to react to the supply shock immediately (i.e. the \((1, 2)\) element of \( B \) is not restricted to be zero).

Given the identification of the structural VAR, one is typically interested in computing impulse responses and variance decomposition. We are typically not interested in particular coefficient estimates – we are interested in the dynamic responses of the variables to the shocks. The impulse response function is just the vector moving average representation. Using lag operator notation, we can write the structural VAR system as:

\[
A(L)Y_t = B\epsilon_t \tag{41}
\]

Inverting the AR component and defining \( C(L) = A(L)^{-1}B \) to be the matrix polynomial of structural moving average coefficients, we have:

\[
Y_t = C(L)\epsilon_t \tag{42}
\]

To find the impulse response function of, say, \( x_t \) to \( \epsilon_{1,t} \), we would set \( \epsilon_{1,t} = 1, \epsilon_{2,t} = 0 \), and all subsequent shocks are zero in expectation. The impulse response on impact would be \( C_{1,1}(0) \), the response after two periods would be \( C_{1,1}(1) \), and so on. We could do the same for variable 2. Our generic definition would be that \( C_{i,j}(h) \) is the impulse response of variable \( i \) to shock \( j \) at horizon \( h \). The matrix \( B \) governs the impulse responses of the variables to the shock on impact – for this reason it is sometimes called the impact matrix or the matrix of impact multipliers.

The forecast error of a variable at time \( t \) is the change in the variable that couldn’t have been forecast between \( t - 1 \) and \( t \). This is due to the realization of the structural shocks in the system, \( \epsilon_t \). We can compute the forecast error over many different horizons, \( h \). The forecast error variance at horizon \( h = 0 \) for each variable is:

\[
E_t x_t - E_{t-1} x_t = C_{1,1}(0)\epsilon_{1,t} + C_{1,2}(0)\epsilon_{2,t} \nonumber \\
E_t z_t - E_{t-1} z_t = C_{2,1}(0)\epsilon_{1,t} + C_{2,2}(0)\epsilon_{2,t} \nonumber
\]

The forecast error variances are just the squares of the forecast errors (since the mean forecast error is zero). Let \( \Omega_i(h) \) denote the forecast error variance of variable \( i \) at horizon \( h \). Then at \( h = 0 \), this is simply:
\[ \Omega_1(0) = C_{1,1}(0)^2 + C_{1,2}(0)^2 \]
\[ \Omega_2(0) = C_{3,1}(0)^2 + C_{3,2}(0)^2 \]

The above follows from the assumptions that the shocks have unit variance and are uncorrelated. The forecast error of the variables at horizons \( h = 1 \) is:

\[
E_t x_{t+1} - E_{t-1} x_{t+1} = C_{1,1}(0)\epsilon_{1,t+1} + C_{1,2}(0)\epsilon_{2,t+1} + C_{1,1}(1)\epsilon_{1,t} + C_{1,2}(1)\epsilon_{2,t}
\]
\[
E_t z_{t+1} - E_{t-1} z_{t+1} = C_{2,1}(0)\epsilon_{1,t+1} + C_{2,2}(0)\epsilon_{2,t+1} + C_{2,1}(1)\epsilon_{1,t} + C_{2,2}(1)\epsilon_{2,t}
\]

The forecast error variances are then:

\[
\Omega_1(1) = C_{1,1}(0)^2 + C_{1,2}(0)^2 + C_{1,1}(1)^2 + C_{1,2}(1)^2
\]
\[
\Omega_2(1) = C_{3,1}(0)^2 + C_{3,2}(0)^2 + C_{1,1}(1)^2 + C_{1,2}(1)^2
\]

To go to more periods, we can then define the forecast error variances recursively as follows:

\[
\Omega_i(0) = C_{i,1}(0)^2 + C_{i,2}(0)^2
\]
\[
\Omega_i(1) = C_{i,1}(1)^2 + C_{i,2}(1)^2 + \Omega_i(0)
\]
\[
\vdots
\]
\[
\Omega_i(h) = C_{i,1}(h)^2 + C_{i,2}(h)^2 + \Omega_i(h - 1)
\]

More generally, for a \( n \) variable system, the total forecast error variance of variable \( i \) at horizon \( h \) in a \( n \) variable system is:

\[
\Omega_i(h) = \sum_{k=0}^{h} \sum_{j=1}^{n} C_{i,j}(k)^2
\]

A forecast error variance decomposition – or just variance decomposition for short – is a way to quantify how important each shock is in explaining the variation in each of the variables in the system. It is equal to the fraction of the forecast error variance of each variable due to each shock at each horizon. Let \( \omega_{i,j}(h) \) be the forecast error variance of variable \( i \) due to shock \( j \) at horizon \( h \).
This is:

$$\omega_{i,j}(h) = \sum_{k=0}^{h} C_{i,j}(k)$$

The fraction of the forecast error variance of variable $i$ due to shock $j$ at horizon $h$, denoted $\phi_{i,j}(h)$, is then the above divided by the total forecast error variance:

$$\phi_{i,j}(h) = \frac{\omega_{i,j}(h)}{\Omega_i(h)} = \frac{\sum_{k=0}^{h} C_{i,j}(k)^2}{\sum_{k=0}^{h} \sum_{j=1}^{n} C_{i,j}(k)^2}$$

(44)

9.1 An Example: Cochrane (1994)

Cochrane (1994) estimates a two variable VAR featuring real non-durable plus services consumption and real GDP. Ignoring for the moment that there are trends in these series (we’ll come back to this in a little bit), let’s estimate the following system with 2 lags (written out in “long hand” for your convenience):

$$c_t = a_c + \beta_{c,1} c_{t-1} + \beta_{c,2} c_{t-2} + \beta_{c,3} y_{t-1} + \beta_{c,4} y_{t-2} + \epsilon_{c,t} \quad (45)$$

$$y_t = a_y + \beta_{y,1} c_{t-1} + \beta_{y,2} c_{t-2} + \beta_{y,3} y_{t-1} + \beta_{y,4} y_{t-2} + \epsilon_{y,t} \quad (46)$$

Using data from 1947q1 to 2010q3, I get the following estimates (I ignore standard errors completely for now):

$$c_t = 0.0132 + 1.172c_{t-1} + 0.109y_{t-1} - 0.179c_{t-2} - 0.104y_{t-1} + \epsilon_{c,t}$$

$$y_t = 0.0789 + 0.725c_{t-1} + 1.139y_{t-1} - 0.611c_{t-2} - 0.256y_{t-1} + \epsilon_{y,t}$$

I get the following estimate for the variance of the covariance matrix:

$$\Sigma_e = \begin{pmatrix} 0.0000224 & 0.0000168 \\ 0.0000168 & 0.0000722 \end{pmatrix}$$

The simple intuition from the permanent income hypothesis tells us that consumption is equal to permanent income, while transitory income equals the difference between actual and permanent income. Cochrane wanted to identify a permanent income shock, call it $\epsilon_{1,t}$, and a transitory income shock, $\epsilon_{2,t}$. The intuition from the permanent income hypothesis tells us that consumption should not respond to the transitory income shock – this gives us the restriction that $B_{1,2} = 0$. Using this, we solve the system of equations for $B$. In MATLAB, if you type “chol(matrix name)”, it
will return an upper triangular Choleski factorization of this variance-covariance matrix. For this reason, recursive identifications are sometimes called Choleski decompositions. We want it to be lower triangular, so take the transpose of what that gives you and it will give you \( B \). I get:

\[
B = \begin{pmatrix}
0.0047 & 0 \\
0.0035 & 0.0077
\end{pmatrix}
\]

Given this estimate of \( B \), I can form \( C(L) \) and compute the impulse responses to each shock. These are given below:

![Response of Consumption and GDP](image)

These impulse responses are broadly consistent with the theory – the second shock is in fact transitory looking and consumption doesn’t really ever respond to it. The variance decomposition reveals:

<table>
<thead>
<tr>
<th>Variable and Horizon</th>
<th>Due to Permanent Shock</th>
<th>Due to Transitory Shock</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consumption</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( h = 0 )</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( h = 4 )</td>
<td>0.975</td>
<td>0.025</td>
</tr>
<tr>
<td>( h = 10 )</td>
<td>0.978</td>
<td>0.022</td>
</tr>
<tr>
<td>( h = 20 )</td>
<td>0.985</td>
<td>0.015</td>
</tr>
<tr>
<td><strong>GDP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( h = 0 )</td>
<td>0.175</td>
<td>0.825</td>
</tr>
<tr>
<td>( h = 4 )</td>
<td>0.505</td>
<td>0.495</td>
</tr>
<tr>
<td>( h = 10 )</td>
<td>0.633</td>
<td>0.367</td>
</tr>
<tr>
<td>( h = 20 )</td>
<td>0.727</td>
<td>0.273</td>
</tr>
</tbody>
</table>

The numbers in this table again conform to the basic theory. Consumption is basically completely explained (at all horizons) by the permanent shock. A significant fraction of the movements
in GDP at short horizons are due to the transitory shock; at longer horizons it’s mostly the permanent shock.

While recursive assumptions on $B$ are common, these are not the only restrictions that can be used to identify structural shocks and impulse responses. You could, for example, impose that variables move in equal and opposite directions to some shock (imposing that some element of $B$ equals the negative of another element). Anything like this is “valid”; but you would hope that it corresponds close to economic theory.

Another identification strategy, which we won’t discuss in detail, is that of Blanchard and Quah (1989). They impose that “demand” shocks have no permanent effect on output, whereas “supply” shocks do. This amounts to imposing a restriction on $B$ such that the impulse response of output to the demand shock approaches zero in the limit. This technique requires taking stands on units roots in the variables and deciding how to transform them before putting them into the VAR. We turn to that issue next.

10 Cointegration

Many macro variables have trends. In Section 8 I said that we would want to get rid of the trend (either by removing a deterministic trend or by first differencing) before doing econometrics. But in the VAR example above, I estimated the VAR in the levels of the variables. What gives? Wouldn’t that give rise to the spurious regression problem?

Not if the variables are cointegrated. Two variables are said to be cointegrated if they are each unit root processes, but if a linear combination of them is stationary. In such a case first differencing is not appropriate.

Take two unit root processes (abstract from constants for now):

\[ x_t = x_{t-1} + \varepsilon_t \]  
\[ z_t = z_{t-1} + \nu_t \]  

Suppose that $x_t - \theta z_t$ is, however, stationary. We say then that $x_t$ and $z_t$ are cointegrated, with cointegrating vector $[1 \ \theta]$. Suppose that we estimate the following regression (again ignoring constants):

\[ x_t = \beta z_t + e_t \]  

OLS, in a large enough sample, will pick $\hat{\beta} = \theta$. That is, OLS will produce a consistent estimate the cointegrating vector, even though both variables are non-stationary. Furthermore, and perhaps oddly, OLS turns out to the “super” consistent in this case, which means that the OLS estimate of $\beta$ converges to $\theta$ faster than it would if the series were stationary. In fact, OLS is going to produce good estimates even if $z_t$ is correlated with $e_t$, so that under normal circumstances we would have
an endogeneity problem. What’s the intuition for this? Recall that OLS tries to minimize the
sum of squared residuals. Hence OLS is trying to minimize \( x_t - \hat{\beta}z_t \). If it picks something other
than \( \hat{\beta} = \theta \), then the residuals are non-stationary, which means that they will get arbitrarily big or
small. Hence, if the sample size is big enough, OLS will hone in on \( \hat{\beta} = \theta \).

When variables are cointegrated, first differencing them is not appropriate. To see this, suppose
that you estimate:

\[
\Delta x_t = \gamma \Delta z_t + u_t \tag{50}
\]

What does \( \gamma \) measure? To see this, start with the true process, (49). Subtract \( x_{t-1} \) from both sides:

\[
\Delta x_t = -x_{t-1} + \theta z_t + e_t
\]

Now add and subtract \( \theta z_{t-1} \) from the right hand side:

\[
\Delta x_t = -x_{t-1} + \theta z_{t-1} + \theta z_t - \theta z_{t-1} + e_t
\]

Simplify:

\[
\Delta x_t = -(x_{t-1} - \theta z_{t-1}) + \theta \Delta z_t + e_t
\]

If you estimate (50), you will not get a consistent estimate of \( \theta \). This is because there is an
omitted term in the error, and that term is \( -(x_{t-1} - \theta z_{t-1}) \). \( \Delta z_t \) is correlated with this, and so
you have a bias.

The representation above is called an “error correction” representation. The term \( (x_{t-1} - \theta z_{t-1}) \)
is the “error”, or deviation from the long run equilibrium. Intuitively, if \( x_t - \theta z_t \) is stationary, then
\( x_t > \theta z_t \) means that \( x_t \) must be expected to fall so as to restore equilibrium over time. Regressing
the first difference of \( x_t \) on the first difference of \( z_t \) ignores this “long run” relationship, and thus
introduces a bias.

A vector error correction model (or VECM) is just like a VAR in first differences, except that
it includes an error correction term. Continuing with the two variable example, a VECM\((p)\) would be (ignoring the constant):

\[
\Delta x_t = \gamma \Delta z_t + u_t
\]
\[
\Delta x_t = \sum_{j=1}^{p} \beta_{1,j}^x \Delta x_{t-j} + \sum_{j=1}^{p} \beta_{2,j}^x \Delta z_{t-j} + \gamma_x (x_{t-1} - \theta z_{t-1}) + e_{x,t}
\]

\[
\Delta z_t = \sum_{j=1}^{p} \beta_{1,j}^z \Delta x_{t-j} + \sum_{j=1}^{p} \beta_{2,j}^z \Delta z_{t-j} + \gamma_z (x_{t-1} - \theta z_{t-1}) + e_{y,t}
\]

You can estimate a VECM in the following way. It is a two step procedure and goes as follows: (i) regress \(x_t\) on \(z_t\) to get \(\theta\). Then run a VAR in first differences, but include the first lag of \(x_t - \theta z_t\) on the right hand side in both equations. Then you just estimate by OLS again – this will be consistent since all the regressors are dated \(t-1\) or earlier and is also efficient since all right hand side variables are the same in each equation. Then you can impose restrictions on the relationship between structural and reduced form shocks just as above, and do impulse response and variance decomposition analysis just as above.

This two step procedure is a little problematic – if you make a mistake in the first step then the estimated VECM model has no meaning. Hence, a lot of people (myself included) advocate estimating VARs in levels – even if the variables are non-stationary. From the VECM relationship above, you can see that it implies a VAR(\(p+1\)) in levels. Thus, estimating a VAR in levels (with enough lags) is (asymptotically) the same as estimating the VECM model. Given that the variables are cointegrated, this is going to be fine, and you don’t have to worry about the spurious regression problem. The construction of standard errors on the coefficients will be a little non-standard, but you are typically not worried about significance of individual coefficients in conventional VAR analysis.

Hence, going back to Section 8 “unit root econometrics”, we need to add the following caveat. You should first test to see whether or not your variables are stationary. If they are non-stationary, what is the better trend representation: deterministic time trend or stochastic trend? If the variables are stationary then estimate a VAR in the levels of the variables. If the variables are stationary about a deterministic trend, then estimate the VARs including a time trend in both equations. If the variables have stochastic trends, you need to test to see whether or not they are cointegrated. A test for cointegration is conceptually simple: regress one on the other and look at the residuals. If the residuals are stationary, then the series are cointegrated. If the residuals are non-stationary, then the series are not cointegrated. If the series are not cointegrated, you should estimate a VAR in first differences. If the series are cointegreated, you should either estimate a VECM model or a VAR in the levels of the variables without a deterministic time trend.

How common is cointegration in the world? It turns out that it is a very common feature of many models and appears to show up in the world. For example, in a neoclassical growth model, the “great ratios” are all stationary – this means that \(\frac{c}{y}\) and \(\frac{I}{y}\) are stationary. More concretely, it means that GDP and its components are likely cointegrated, with cointegrating vectors \([1 \ -1]\). A lot of VARs are likely to include these kinds of variables, and thus you want to be weary of simply differencing everything. As mentioned above, a common practice is to simply estimate the VARs
in levels. I am a fan of this practice, though it is not suitable for all purposes (for example, the Blanchard and Quah (1989) assumptions require you to first difference output).

11 Bootstrapping

A convenient and popular way of constructing standard errors is via the bootstrap. The basic idea behind the bootstrap is as follows. Estimate a time series model and get the residuals. Then construct a large number of “fake” data sets by re-sampling (with replacement) from the observed residuals and using the estimated coefficients. Then re-apply your econometric procedure to the fake data and generate an entire distribution of estimates. Then construct standard errors using the bootstrap simulated distributions of coefficients.

I’ll show this through an example. Suppose that we have a time series process:

\[ x_t = 0.9x_{t-1} + \varepsilon_t \quad \varepsilon_t \sim N(0,1) \]  

(51)

I generated 1000 different data sets of 200 observations each using this as the data generating process. Then on each data set I estimated an AR(1) process. My average OLS estimated of the AR coefficient was 0.88. This is ever so slightly downward biased, which is normal in time series data, but if I were to increase the sample size the bias would go away. My average OLS estimate of the standard error of this estimate is 0.034. But the standard deviation of the distribution of estimates is actually 0.037. In other words, the OLS standard errors are too small.

Now I do a bootstrap. For each simulation of the data, I estimate via OLS and get an estimate of \( \hat{\rho} \) as well as a time series of residuals, \( \hat{\varepsilon}_t \). Then, I create \( N \) different re-sampled versions of \( \hat{\varepsilon}_t \) by drawing randomly (with replacement) from the observed empirical distribution of \( \hat{\varepsilon}_t \). You can use the “boostrp” command in Matlab to do this. Then, for each of these \( N \) different bootstrap samples, I construct another fake series of \( x_t \)s using my estimated \( \rho \). Then I re-estimate an AR(1) on the bootstrap sample and save it. I repeat this \( N \) times – i.e. once for each bootstrap sample. Then I look at the standard deviation of the estimated \( \hat{\rho} \) across each bootstrap sample. When I do this \( N = 300 \) times, my bootstrap standard deviation of \( \hat{\rho} \) comes out to be 0.038. This is much closer to the true standard deviation of \( \rho \) than is the OLS estimate of the standard error.

You can use bootstrapping in a variety of different contexts. For example, you can use it to construct confidence intervals for impulse response functions in VARs. You would estimate the VAR and then construct \( N \) bootstrap samples of the VAR residuals. Then using the estimated AR coefficients, you would create \( N \) different samples of the variables in the VAR using the bootstrapped simulations of the errors. Then on each simulated data set, you would run the VAR and construct impulse responses. You would save the impulse responses. Then you would take the standard deviation (or percentiles) of the bootstrap distributions of the estimated impulse responses.
12 Filtering

A related concept to first differencing and detrending data is that of filtering. The basic premise is to break a series down into two components: “trend” and “cycle”:

\[ x_t = x^T_t + x^C_t \]  

(52)

The basic approach is to come up with an estimate of the “trend” component, and then subtract that off from the actual series so as to get the “cycle” component. We’ve already seen one way to do this – estimate a linear time trend to get \( x^T_t \). But more generally, we may want to allow the trend to move around.

The Hodrick-Prescott filter (HP filter) is very common in empirical macro and does just this; it also has as a special case the linear time trend. Formally, let \( \lambda \) be an exogenous constant chosen by the researcher in advance. The HP filter chooses a sequence of trend, \( x^T_t \), to solve the following minimization problem:

\[
\min_{x^T_t} \sum_{t=1}^{T} (x_t - x^T_t)^2 + \lambda \sum_{t=2}^{T-2} \left( (x^T_{t+1} - x^T_t) - (x^T_t - x^T_{t-1}) \right)^2
\]  

(53)

The first part is the cycle component (i.e. \( x_t - x^T_t \)). It represents a penalty for large cyclical components. The second part is that change in the change of the trend. Basically, this part is a penalty for the trend not being smooth. In words, then, the HP filter finds a trend that minimizes cyclical fluctuations subject to a penalty for the trend itself moving around.

Consider a couple of different possible values of \( \lambda \). If \( \lambda = 0 \), then there is no penalty to the trend jumping around, and you would set \( x^T_t = x_t \) – in other words, the trend would be the actual series and there would be no cyclical component. As \( \lambda \to \infty \), you will want \( (x^T_{t+1} - x^T_t) - (x^T_t - x^T_{t-1}) = 0 \). This means that the change in the trend is constant, which means that you would pick out a linear time trend. For intermediate values of \( \lambda \), the trend will move around some, but there will still be a cyclical component. For quarterly data, it is common to use \( \lambda = 1600 \). For annual data, people typically use \( \lambda = 100 \). For monthly data, it is common to use \( \lambda = 14400 \).

The bandpass filter is another popular statistical filter. It is aimed to isolate cycles with different “periodicities”, where periodicity measures the amount of time it takes for a cycle to complete. To study the properties of this filter more fully, one needs to use frequency domain and spectral analysis, which is beyond the aim of the course. Business cycle frequencies are typically defined as having periodicities between 6-32 quarters (1.5 to 8 years). Hence a bandpass filter with periodicities between 6 and 32 quarters can be used to isolate the business cycle component of the data. In practice this turns out to be pretty similar to the HP filter.