Graduate Macro Theory II: Preliminaries

Eric Sims
University of Notre Dame

Spring 2017

1 Introduction

This set of notes describes some "preliminaries" before we get too far into the course. Macroeconomics is both dynamic and stochastic. Dynamic means that we have to keep track of variables across time, and stochastic means that we need to study expectations. Hence, I will first start by saying a few things about notation and expectations operators. Then we'll say something about the stochastic processes that we assume in macro – I'll talk about Markov processes and ARMA processes. Then we'll discuss a couple of tools that I'll discuss in the context of ARMA models, but which are more broadly applicable and will be used when we study economic models – impulse response functions and variance decompositions. I'll also give a brief discussion of "filtering" which is important in that we typically need to transform real-world data (because of trends) before analyzing them in the context of an economic model. I'll conclude with a discussion of the Lucas Critique, which forms the basis for much of modern macroeconomics.

2 Notation and Expectations

A variable is a realization of something that can change (either deterministically or stochastically). Endogenous variables are variables whose values are determined "inside" of a model (through decision rules derived from optimization problems and imposition of an equilibrium concept). Exogenous variables are variables whose values are determined "outside" of a model – these are taken as given. We take exogenous variables, feed them into the model, and the realizations of the endogenous variables are the outcome of the model. Parameters are values that govern relationships in a model (e.g. how much curvature there is in a utility function, how heavily agents discount future utility flows, etc.). We think of parameters as exogenous and fixed, though one could conceive of models in which parameters change stochastically (at which point the parameters become exogenous variables). I will try to use Latin letters (e.g. X, Y) to denote variables, and Greek letters (e.g. α, β) to denote parameters. I will try, but I will probably fail at some point.

We also encounter the terms "state" and "control" in describing variables in a macroeconomic model. Exogenous variables are always state variables, but endogenous variables can be either controls or states. Loosely, "control" variables are variables whose values are chosen in a model and are free to "jump" in response to new information. State variables are variables whose values agents need to know to make decisions. These are variables that are either exogenous (a productivity term, government spending) or endogenous (capital stocks, stocks of assets, etc.). State variables are predetermined with respect to control variables: you need to know the states to choose the controls. As I said, some states are endogenous in the sense that current actions can affect *future* values of the state, but the current value of the states is known.

Macroeconomic models and data are dynamic – we observe realizations of variables at points in time. Most of macro (with the exception of some growth models and asset pricing models) models time as discreet. Let X_t denote some variable (either endogenous or exogenous, control or state). This notation refers to the realization of the variable at date t. X_{t-1} refers to the realized value one period before t, X_{t+k} the realized value k periods subsequent to t, and so on. Think of these time indexes as representing integers. People flip back and forth on timing notations, and I do the same (sorry). Sometimes, we think of period 0 as being the "present" and move time forward; then X_t for $t = 0, 1, 2, ...$ represents realizations of the variable in the present (period 0) and going forward. Other times, we think of period t as being the present; here we have X_{t+k} as representing realizations of X either moving forward ($k > 0$), backward ($k < 0$), or in the present ($k = 0$).

Macroeconomic models are stochastic, in the sense that there is randomness in the realization of variables. The stochastic nature of macro models comes in from exogenous variables, which we typically model as having a random component. Because these models are stochastic, and because agents are forward-looking, we need to worry about expectations. $E(X_t)$ refers to the *unconditional* expectation of X_t. By unconditional I mean knowing nothing about the current state of the system. $E_t X_{t+k}$ refers to expectation of future realizations of X *conditional* on all information known at time t. By convention we have that $E_t X_t = X_t$: since X_t is known in period t, there is no uncertainty over its realization. Also, $E_t X_{t-k} = X_{t-k}$.

For two arbitrary random variables, Y and Z, the Law of Iterated Expectaions says that $E(Y) = E(E(Y \mid Z))$. In words, this says that the unconditional expectation of a conditional exepctation is the unconditional expectation. This has the following implication for time series: $E_t(E_{t+1}(X_{t+2})) = E_t X_{t+2}$. In other words, your best guess conditional on today's information (where "today" is taken to be period t) of your best guess conditional on tomorrow's information ("tomorrow" being period $t + 1$) of a variable two periods out from now is just your best guess based on today's information.

Rational expectations moves beyond simple expected value and imposes some more structure. This dates back to Muth and Lucas. Rational expectations says that expectations of future realizations of relevant variables are (i) correct on average and (ii) the forecast errors are unpredictable given available information. In other words, agents have *model consistent expectations* in the sense that they (i) know the model generating endogenous variables and (ii) use this knowledge to make forecasts. This does *not* imply that agents do not make forecast errors. Let $E_t X_{t+k}$ be the forecast of X_t k periods from now conditional on available information at time t. The forecast error is

$u_{t+k} = X_{t+k} - E_t X_{t+k}$: just the realized value minus the expected value. In general, u_{t+k} will not be zero, but it ought to be zero on average; i.e. the unconditional expectation of it should be zero: $E(u_{t+k}) = 0$. The logic behind this is pretty simple: if you were making forecast errors on average, you aren't forming expectations optimally. Furthermore, the covariance of the forecast error with anything known at the time the forecast is made should be zero: $\text{cov}(u_{t+k}, Z_t) = 0$ where Z_t represents anything known at t. So rational expectations says that your forecasts are right on average and are unpredictable. Another way to think about this is that expectations are "optimal" in some sense – if you were wrong on average or predictably wrong (and being wrong mattered), you couldn't be forming expectations optimally. Rational expectations is widely used in empirical work, in that it implies restrictions that can be used in econometrics. Note that rational expectation does not necessarily rule out informational frictions: we could restrict the information that agents have available to them. This may give rise to them appearing to violate rational expectations (their forecast errors are predictable), but only if one conditions on more information than the agents have at the time they make the forecast.

3 Stochastic Processes

As noted above, most macro models are driven by shocks to exogenous processes. We need to specify properties of the stochastic processes that these exogenous states follow.

The two most common ways to model a stochastic process are as a Markov process (discreet outcomes) or as an autoregressive moving average (ARMA) process. The so-called "Markov Property" says that the current state of a system is a sufficient statistic to forecast future values of the state; e.g. once you know S_t (the current state), knowing S_{t-k} for $k > 0$ doesn't tell you anything about the expected evolution of the state going forward.

Let $\bar{S}$ be a $N \times 1$ vector of possible realizations of some exogenous state, call it s_t. Let P be a $N \times N$ probability (or transition) matrix. Its elements are the the probabilities of transition from state i to state j between periods t and $t+1$. Hence:

$$P_{i,j} = \text{prob}\left(s_{t+1} = s_j \mid s_t = s_i\right)$$

Here i and j are particular discreet realizations in $\bar{S}$. In other words, the rows (the i index) refer to the current state, and the columns (the j index) tell you the probability of going to each possible other state in $t+1$, given that you are sitting in state i. All rows must sum to one (i.e. the system will transition to some possible realization with probability 1 in the next period). Kind of naturally, the larger are the elements on the diagonal, the more persistent is the process.

ARMA processes are continuous process which are built off of white noise processes. A white noise process, which I denote here by ε_t, has the properties that it is mean zero (i.e. $E(\varepsilon_t) = 0$); has a known and time-invariant variance (e.g. $\text{var}(\varepsilon_t) = \sigma^2$), and the realizations of the white noise process are uncorrelated at all leads and lags (e.g. $\text{cov}(\varepsilon_t, \varepsilon_{t+j}) = 0 \ \forall \ j$).

An ARMA(p,q) process can be written:

$$s_t = a + \rho_1 s_{t-1} + \rho_2 s_{t-2} + \cdots + \rho_p s_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots \theta_q \varepsilon_{t-q}$$

Here, the p refers to the autoregressive lag order (the number of lagged dependent variables) and q the moving average lag order (the number of lagged white noise terms). ARMA processes are not unique – under certain conditions, you can go back and forth between an MA process and an AR process. This means that you can often approximate the MA terms with a pure AR process with sufficiently many lags. So, much of the time you can approximate any ARMA process with a sufficiently long-lived AR(p) process.

An AR(p) process does not formally have the Markov property discussed above – knowing s_t is not sufficient to forecast s_{t+1}, you'd also need to know s_{t-1}, s_{t-2} and so on. But it turns out you can redefine the state in such a way as to write an AR(p) as a VAR(1) (where the V stands for vector). In particular, suppose I have a process (here I ignore any constant):

$$s_t = \rho_1 s_{t-1} + \rho_2 s_{t-2} + \rho_3 s_{t-3} + \ldots \rho_p s_{t-p} + \varepsilon_t$$

Then define a vector:

$$\mathbf{s}_t = \begin{bmatrix} s_t \\ s_{t-1} \\ s_{t-2} \\ \vdots \\ s_{t-p+1} \end{bmatrix}$$

Then we can write:

$$\begin{bmatrix} s_t \\ s_{t-1} \\ s_{t-2} \\ \vdots \\ s_{t-p+1} \end{bmatrix} = \begin{pmatrix} \rho_1 & \rho_2 & \rho_3 & \ldots & \rho_p \\ 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \begin{bmatrix} s_{t-1} \\ s_{t-2} \\ s_{t-3} \\ \vdots \\ s_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Or just:

$$\mathbf{s}_t = \Lambda \mathbf{s}_{t-1} + \varepsilon_t$$

Where Λ is the matrix of coefficients described above. With the state so redefined, this process again has the Markov property. As we will see, having an AR(1) structure makes computation of impulse responses and variance decompositions particularly straightforward.

4 Impulse Response Functions and Variance Decompositions

An impulse response function is defined as the change in the current and expected values of a variable (or a vector of variables) conditional on the realization of a shock at a point in time. We think of a shock as an innovation (or surprise movement) in an exogenous variable, where the innovation is modeled as white noise. This definition does not depend on any particular process for the exogenous or endogenous variables. Suppose that X_t is a univariate process of some sort built off white noise shocks. The impulse response function is defined as:

$$\text{IRF}(h) = E_t X_{t+h} - E_{t-1} X_{t+h} \mid \varepsilon_t = e, \quad h \geq 0$$

Here, e is some particular realization of ε_t. It is common to calculate impulse response functions to one standard deviation shocks, in which case $e = \sigma$, where σ is the standard deviation of ε_t. But you can calculate an impulse response function for any sized (or signed) shock you want. If the underlying process is linear, it won't matter – there will be no dependence of the impulse response function on the size of the shock (other than for scaling the response up or down), and the sign will only affect the sign of the response. For a non-linear process this need not be the case. The impulse response function is defined for different forecast horizons, h. $h = 0$ is said to the be "impact" horizon.

As an example, suppose we have an AR(1) process:

$$X_t = 0.9 X_{t-1} + \varepsilon_t$$

Suppose that $\varepsilon_t \sim N(0, 1)$. Then the impulse response function can be computed by using this process to calculate expected values. At time $t - 1$, the expected value of ε_t is 0. Hence:

$$\begin{aligned}
E_{t-1} X_t &= 0.9 X_{t-1} \\
E_{t-1} X_{t+1} &= 0.9^2 X_{t-1} \\
E_{t-1} X_{t+2} &= 0.9^3 X_{t-1} \\
&\vdots \\
E_{t-1} X_{t+h} &= 0.9^{h+1} X_{t-1}
\end{aligned}$$

Now compute expected values conditional on the realization of a shock at time t of 1 (equal to the standard deviation):

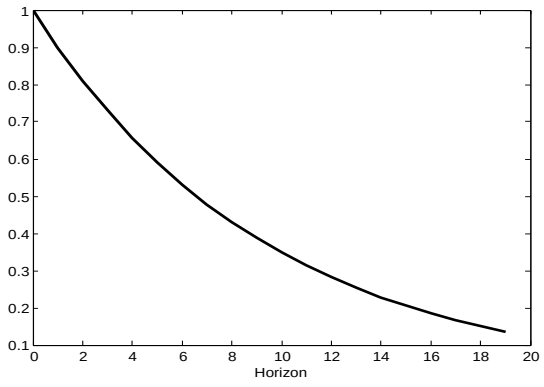
$$
\begin{aligned}
E_t X_t &= 0.9 X_{t-1} + 1 \\
E_t X_{t+1} &= 0.9^2 X_{t-1} + 0.9 \\
E_t X_{t+2} &= 0.9^3 X_{t-1} + 0.81 \\
&\vdots \\
E_t X_{t+h} &= 0.9^{h+1} X_{t-1} + 0.9^h
\end{aligned}
$$

Taking the difference between these two yields:

$$
\begin{aligned}
E_t X_t - E_{t-1} X_t &= 1 \\
E_t X_{t+1} - E_{t-1} X_{t+1} &= 0.9 \\
E_t X_{t+2} - E_{t-1} X_{t+2} &= 0.81 \\
&\vdots \\
E_t X_{t+h} - E_{t-1} X_{t+h} &= 0.9^h
\end{aligned}
$$

Above I have shown an impulse response function for an exogenous variable. We can also do this for an endogenous variable, Y_t. Suppose that Y_t depends on X_t in some way; a particularly simple assumption is that this dependence is linear, e.g. $Y_t = \beta X_t$. We would proceed with the construction of the IRF of Y_t in the same way – compute the forecasts conditional on the realization of a shock at time t, and compare them to the forecasts without that realization. For this example, the impulse response of Y_t would simply be proportional to the impulse response of X_t, with the constant of proportionality equal to β.

It is common to plot impulse responses as a way to analyze them. For this process, we'd have:



A useful insight is that an impulse response function is essentially just a plot of the coefficients of the moving average representation of a time series. Take our AR(1) example from above. To

recover the moving average representation, iteratively substitute as follows:

$$\begin{aligned} X_t &= 0.9X_{t-1} + \varepsilon_t \\ X_t &= 0.9\left(0.9X_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t = 0.81X_{t-2} + 0.9\varepsilon_{t-1} + \varepsilon_t \\ X_t &= 0.81\left(0.9X_{t-3} + \varepsilon_{t-2}\right) + 0.9\varepsilon_{t-1} + \varepsilon_t = 0.9^3X_{t-3} + 0.81\varepsilon_{t-2} + 0.9\varepsilon_{t-1} + \varepsilon_t \end{aligned}$$

If you keep doing this, since $0.9^k \to 0$ for k sufficiently big, you get:

$$X_t = \varepsilon_t + 0.9\varepsilon_{t-1} + 0.81\varepsilon_{t-2} + 0.9^3\varepsilon_{t-3} + \ldots$$

Or in summation notation:

$$X_t = \sum_{j=0}^{\infty} 0.9^j \varepsilon_{t-j}$$

Comparing this to what we had above, we see that the impulse response function is just the moving average representation "moved forward." To see this, do what we did above by taking expectations going forward:

$$\begin{aligned} E_{t-1}X_t &= 0.9\varepsilon_{t-1} + 0.81\varepsilon_{t-2} + 0.9^3\varepsilon_{t-3} + \ldots \\ E_{t-1}X_{t+1} &= 0.81\varepsilon_{t-1} + 0.9^3\varepsilon_{t-2} + \ldots \\ E_{t-1}X_{t+2} &= 0.9^3\varepsilon_{t-1} + 0.9^4\varepsilon_{t-2} + \ldots \end{aligned}$$

Above, the ε_t term doesn't show up, since $E_{t-1}\varepsilon_t = 0$. Now do this conditional on the realization of a value of ε_t of 1 in period t (same as we did above):

$$\begin{aligned} E_tX_t &= 1 + 0.9\varepsilon_{t-1} + 0.81\varepsilon_{t-2} + 0.9^3\varepsilon_{t-3} + \ldots \\ E_tX_{t+1} &= 0.9 + 0.81\varepsilon_{t-1} + 0.9^3\varepsilon_{t-2} + \ldots \\ E_tX_{t+2} &= 0.81 + 0.9^3\varepsilon_{t-1} + 0.9^4\varepsilon_{t-2} + \ldots \end{aligned}$$

Then take the difference to construct the impulse response function:

$$\begin{aligned} E_tX_t - E_{t-1}X_t &= 1 \\ E_tX_{t+1} - E_{t-1}X_{t+1} &= 0.9 \\ E_tX_{t+2} - E_{t-2}X_{t+2} &= 0.81 \end{aligned}$$

This is exactly the same thing we found above. In other words, the impulse response at horizon h is just the MA coefficient at lag h. Ultimately, what we're often interested in in macro is exactly this moving average representation, which tells us how primitive shocks (the εs) affect variables at different horizons. Estimating and working with moving average terms is difficult (requires distributional assumptions and maximum likelihood, since the εs are directly observed). Estimating and working with AR(p) processes is relatively straightforward – most of the time, OLS will be a consistent way to estimate these processes. So this tells us that we can recover a moving average representation of a time series by estimating an AR model and constructing impulse response functions – the coefficients of the impulse response function at different forecast horizons (i.e. different h) are just the MA coefficients at the same lag.

Of course, in what I did above I (implicitly) assumed that I could recover the moving average representation from the AR process. Sometimes this isn't possible, in which case we would say we have a non-invertible AR process. In multivariate frameworks there are some interesting economic mechanisms that can give rise to this in an economic model. Don't worry about this unless told otherwise. Also, in what I did above, I assumed a simple AR(1) process, where mechanically constructing impulse responses was pretty easy. You could have much more complicated processes and the same general definition of an impulse response function is the same. An AR(p) process may look nasty to compute an impulse response function by hand, but if you remember above, you can write an AR(p) process as a VAR(1), which makes it quite straightforward to compute the impulse response function. You may also encounter non-linear models, which complicates things (but doesn't change the definition of an impulse response function). To compute impulse response functions in a non-linear model, we'll use what is called "generalized impulse response functions." Here what you do is in essence simulate data from the non-linear model out to forecast horizons h. You do this a bunch of times. Then you simulate data from the same model, but this time condition on the realization of a particular shock in the first period, and then simulate data from the non-linear model out to forecast horizon h. Then you average both of these simulations across the number of times you did. This gives you the $E_t X_{t+k}$ and the $E_{t-1} X_{t+k}$, and then compute the difference. This conceptually gives you the difference in the expected values of future values of the model conditional on a realization of a shock in period t. We'll revisit this issue later.

We can also construct impulse responses for multivariate processes. Suppose we have a 2×1 vector of variables, $\mathbf{X}_t$, that obeys a vector AR(1), with two uncorrelated white noise processes (e.g. "shocks") buffeting it:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

The matrix $\mathbf{B}$ is 2×2. If the off-diagonal elements of both it and $\mathbf{A}$ are zeros, then this is just two independent AR(1) processes. But non-zero off-diagonal elements allow for some more interesting feedback, both from shocks and from lags of variables. Conceptually, the impulse response function is the same as before – the displacement of forecasts of the now vector of variables conditional on the realization of a shock at time t. But since there are now two shocks, there will be two different

impulse response functions – one conditional on each shock. The impulse response function will also be a vector – showing how both elements of X_t react over time. Since I've written this as a vector AR(1) process, the impulse response functions look basically the same as in the scalar case. Conditioning on one unit shocks to $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ (and setting the other equal to zero), we have:

$$\text{IRF}_1(h) = \mathbf{A}^{h-1}\mathbf{B}(:,1)$$

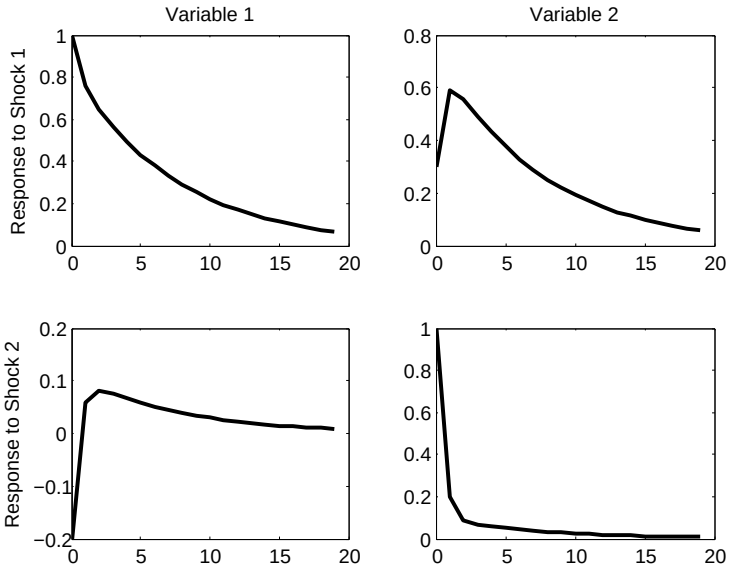
$$\text{IRF}_2(h) = \mathbf{A}^{h-1}\mathbf{B}(:,2)$$

To compute these, we just condition on the different columns of the "impact matrix." As an example, suppose that the matrixes are:

$$\mathbf{A} = \begin{pmatrix} 0.7 & 0.2 \\ 0.5 & 0.3 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 1 & 0.3 \\ -0.2 & 1 \end{pmatrix}$$

Below are the impulse responses of variables "1 and 2" (corresponding to the rows of $\mathbf{X}_t$) to shocks "1 and 2" (corresponding to the rows of the shock vector).



A related concept to impulse response functions is a variance decomposition. A variance decomposition tells you what fraction of the forecast error variance of a variable (say X_t) is due to different shocks, potentially at different horizons. Naturally, a variance decomposition is not a particularly interesting construct for a univariate model with one shock (one white noise process buffeting it) – that one shock explains all of the forecast error variance at all horizons. I will use the terminology that an *unconditional* variance decomposition tells you how much of the variance

a particular shock explains in an unconditional sense, while a conditional variance decomposition tells you how much of the forecast error variance of a variable is explained by a shock at a particular forecast horizon.

For simplicity, suppose that we have a univariate process that we can write in MA form. Let's write it:

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3}$$

The variance of X_t is just (using properties of variance, and the fact that the variance of the white noise process is constant over time):

$$\text{var}(X_t) = \left(1 + \theta_1^2 + \theta_2^2 + \theta_3^2\right) \sigma^2$$

Now, suppose that X_t were hit by two different white noise processes, $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$. These are independent, each with variance σ_i^2 for $i = 1, 2$. Let's write the process:

$$X_t = \varepsilon_{1,t} + \theta_1 \varepsilon_{1,t-1} + \theta_2 \varepsilon_{1,t-2} + \theta_3 \varepsilon_{1,t-3} + \varepsilon_{2,t} + \alpha_1 \varepsilon_{2,t-1} + \alpha_2 \varepsilon_{2,t-2} + \alpha_3 \varepsilon_{2,t-3}$$

The total variance of X_t is then:

$$\text{var}(X_t) = \left(1 + \theta_1^2 + \theta_2^2 + \theta_3^2\right) \sigma_1^2 + \left(1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2\right) \sigma_2^2$$

What I call the unconditional variance decomposition then just the share of the total variance due to each shock. We have:

$$\text{var}(X_t \mid \varepsilon_{1,t}) = \frac{\left(1 + \theta_1^2 + \theta_2^2 + \theta_3^2\right) \sigma_1^2}{\left(1 + \theta_1^2 + \theta_2^2 + \theta_3^2\right) \sigma_1^2 + \left(1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2\right) \sigma_2^2}$$

$$\text{var}(X_t \mid \varepsilon_{2,t}) = \frac{\left(1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2\right) \sigma_2^2}{\left(1 + \theta_1^2 + \theta_2^2 + \theta_3^2\right) \sigma_1^2 + \left(1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2\right) \sigma_2^2}$$

Naturally, the shares must sum up to 1. This exercise gives you an idea of how important each shock is in accounting for volatility in X_t: this will depend on the "magnitudes" of the shocks (the variances) as well as the coefficients. Since the variance decomposition is just a function of the MA coefficients, it doesn't contain any different information than the impulse response functions – it just is a different way to view the model.

Computing the variance (and hence the variance share) is easy here since I used a simple MA process. For a more complicated model (potentially non-linear), I could compute the unconditional variance decomposition via simulation. I could simulate the model for many periods and calculate the variance of X_t. Then I would simulate the model "turning off" shock 1 (setting the variance to zero), and calculate the variance of X_t. Then I would do the same for shock 2. Then I would take the shares.

To compute variance decompositions at different forecast horizons, we need to compute forecast

errors. Define the forecast error at horizon h as $X_{t+h} - E_{t-1}X_{t+h}$. In other words, this is the difference between the realized value and the expected value before observing shocks in period t. Take the MA process driven by the two shocks above. The forecast error at horizon $h = 0$ can be constructed as:

$$\begin{aligned} E_{t-1}X_t &= \theta_1\varepsilon_{1,t-1} + \theta_2\varepsilon_{1,t-2} + \theta_3\varepsilon_{1,t-3} + \alpha_1\varepsilon_{2,t-1} + \alpha_2\varepsilon_{2,t-2} + \alpha_3\varepsilon_{2,t-3} \\ X_t &= \varepsilon_{1,t} + \theta_1\varepsilon_{1,t-1} + \theta_2\varepsilon_{1,t-2} + \theta_3\varepsilon_{1,t-3} + \varepsilon_{2,t} + \alpha_1\varepsilon_{2,t-1} + \alpha_2\varepsilon_{2,t-2} + \alpha_3\varepsilon_{2,t-3} \end{aligned}$$

The difference is just:

$$X_t - E_{t-1}X_t = \varepsilon_{1,t} + \varepsilon_{2,t}$$

Now compute the forecast error horizon at $h = 1$. To do this, we have:

$$\begin{aligned} E_{t-1}X_{t+1} &= \theta_2\varepsilon_{1,t-1} + \theta_3\varepsilon_{1,t-2} + \alpha_2\varepsilon_{2,t-1} + \alpha_3\varepsilon_{2,t-2} \\ X_{t+1} &= \varepsilon_{1,t+1} + \theta_1\varepsilon_{1,t} + \theta_2\varepsilon_{1,t-1} + \theta_3\varepsilon_{1,t-2} + \varepsilon_{2,t+1} + \alpha_1\varepsilon_{2,t} + \alpha_2\varepsilon_{2,t-1} + \alpha_3\varepsilon_{2,t-2} \end{aligned}$$

The difference is:

$$X_{t+1} - E_{t-1}X_{t+1} = \varepsilon_{1,t+1} + \theta_1\varepsilon_{1,t} + \varepsilon_{2,t+1} + \alpha_1\varepsilon_{2,t}$$

Similarly, if you do this for horizon $h = 2$, you get:

$$X_{t+2} - E_{t-1}X_{t+2} = \varepsilon_{1,t+2} + \theta_1\varepsilon_{1,t+1} + \theta_2\varepsilon_{1,t} + \varepsilon_{2,t+1} + \alpha_1\varepsilon_{2,t+1} + \alpha_2\varepsilon_{2,t}$$

For any horizon $h \geq 3$, the forecast error is just the process, since the $t-1$ forecast 3 or more periods out will just be zero:

$$X_{t+h} - E_{t-1}X_{t+h} = \varepsilon_{1,t+h} + \theta_1\varepsilon_{1,t+h-1} + \theta_2\varepsilon_{1,t+h-2} + \theta_3\varepsilon_{1,t+h-3} + \varepsilon_{2,t+h} + \alpha_1\varepsilon_{2,t+h-1} + \alpha_2\varepsilon_{2,t+h-2} + \alpha_3\varepsilon_{2,t+h-3}$$

Now take the variance of the forecast error at each horizon:

$$\begin{aligned} h = 0: &\quad \text{var}(X_t - E_{t-1}X_t) = \sigma_1^2 + \sigma_2^2 \\ h = 1: &\quad \text{var}(X_{t+1} - E_{t-1}X_{t+1}) = (1 + \theta_1^2)\sigma_1^2 + (1 + \alpha_1^2)\sigma_2^2 \\ h = 2: &\quad \text{var}(X_{t+2} - E_{t-1}X_{t+2}) = (1 + \theta_1^2 + \theta_2^2)\sigma_1^2 + (1 + \alpha_1^2 + \alpha_2^2)\sigma_2^2 \\ h \geq 3: &\quad \text{var}(X_{t+h} - E_{t-1}X_{t+h}) = \left(1 + \theta_1^2 + \theta_2^2 + \theta_3^2\right)\sigma_1^2 + \left(1 + \alpha_1^2 + \alpha_2^2 + \alpha_3^2\right)\sigma_2^2 \end{aligned}$$

The variance decomposition is again just the shares, but now at difference forecast horizons:

$$\text{var}(X_t - E_{t-1}X_t \mid \varepsilon_1) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$$
$$\text{var}(X_t - E_{t-1}X_t \mid \varepsilon_2) = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

$$\text{var}(X_{t+1} - E_{t-1}X_{t+1} \mid \varepsilon_1) = \frac{(1+\theta_1^2)\sigma_1^2}{(1+\theta_1^2)\sigma_1^2 + (1+\alpha_1^2)\sigma_2^2}$$
$$\text{var}(X_{t+1} - E_{t-1}X_{t+1} \mid \varepsilon_2) = \frac{(1+\alpha_1^2)\sigma_2^2}{(1+\theta_1^2)\sigma_1^2 + (1+\alpha_1^2)\sigma_2^2}$$

$$\text{var}(X_{t+2} - E_{t-1}X_{t+2} \mid \varepsilon_1) = \frac{(1+\theta_1^2+\theta_2^2)\sigma_1^2}{(1+\theta_1^2+\theta_2^2)\sigma_1^2 + (1+\alpha_1^2+\alpha_2^2)\sigma_2^2}$$
$$\text{var}(X_{t+2} - E_{t-1}X_{t+2} \mid \varepsilon_2) = \frac{(1+\alpha_1^2+\alpha_2^2)\sigma_2^2}{(1+\theta_1^2+\theta_2^2)\sigma_1^2 + (1+\alpha_1^2+\alpha_2^2)\sigma_2^2}$$

$$h \geq 3: \quad \text{var}(X_{t+h} - E_{t-1}X_{t+h} \mid \varepsilon_1) = \frac{\left(1+\theta_1^2+\theta_2^2+\theta_3^2\right)\sigma_1^2}{\left(1+\theta_1^2+\theta_2^2+\theta_3^2\right)\sigma_1^2 + \left(1+\alpha_1^2+\alpha_2^2+\alpha_3^2\right)\sigma_2^2}$$
$$h \geq 3: \quad \text{var}(X_{t+h} - E_{t-1}X_{t+h} \mid \varepsilon_2) = \frac{\left(1+\alpha_1^2+\alpha_2^2+\alpha_3^2\right)\sigma_2^2}{\left(1+\theta_1^2+\theta_2^2+\theta_3^2\right)\sigma_1^2 + \left(1+\alpha_1^2+\alpha_2^2+\alpha_3^2\right)\sigma_2^2}$$

For this particular process, we see that the conditional variance decomposition at forecast horizons greater than or equal to 3 is exactly the same as what I defined as the unconditional forecast error variance decomposition above. This occurs because I wrote down an MA process with a limited number of terms. For a more general process, the unconditional and conditional variance decompositions will be the same only in the limit, as $h \to \infty$.

In practice, how might one compute variance decompositions for more complicated processes? As we see in these formulas, the variance decomposition just depends on the moving average terms on the variance of the shocks. Since the impulse response functions are just the moving average terms, we can get the variance decomposition by first computing the impulse response functions to one standard deviation shocks (it's important to do one standard deviation shocks to get this to work). The total forecast error variance at horizon h will then just be the sum of squared impulse responses to *all shocks* up to that horizon. With one standard deviation shocks, in the simple MA process I gave above, the impulse response at horizon $h = 0$ is just σ_1 to shock 1 and σ_2 to shock 2. At $h = 1$, the impulse responses are $\theta_1\sigma_1$ and $\alpha_1\sigma_2$. Now square the impulse responses and sum them – you get $(1+\theta_1^2)\sigma_1^2$ and $(1+\alpha_1^2)\sigma_2^2$. The total variance is just the sum of these (the denominator above). To compute the variance decomposition, just take the ratio of the sum of squared impulse responses to one of the shocks (either 1 or 2), and divide by the total variance. Do

this for each horizon. At each horizon, the total forecast error variance is just the sum of squared impulse responses up to that horizon summed across both shocks. To get the contribution of one shock or another at that horizon, take the ratio of the sum of squared impulse responses to one of the shocks to the total forecast variance, and you have exactly the expressions above.

5 Logs

When working with macroeconomic data, most of the time you'll want to work with the natural logs of the series. This will also be true in a lot of economic models, where we'll take approximations to logs of the variables (i.e. what we call log-linearizing).

Why do we like logs? Logs put give things percentage interpretations and log differences can be interpreted as percentage differences. Suppose that a series is growing at a constant rate, x. This means:

$$X_{t+1} = (1+x)X_t$$

Take logs:

$$\ln X_{t+1} = \ln(1+g) \ln X_t$$

Now here's a very useful fact: the log of one plus a number is approximately the number. To see this, let's take a first order Taylor Series approximation of $\ln(1+x)$. Recall, the generic definition of a first order Taylor approximation is:

$$f(x) \approx f(x^*) + f'(x^*)(x - x^*)$$

Here x is a variable and x^* is a particular realization of that variable. Let's apply this to $\ln(1+x)$ by taking an approximation about the point $x^* = 0$:

$$\ln(1+x) \approx \ln(1+0) + \frac{1}{1+0}(x-0) = x$$

In other words, the log of one plus a small number is approximately the small number. This approximation is very good for x small (growth rates are small). Applying it above, it means that:

$$\ln X_{t+1} - \ln X_t \approx x$$

This means that we can interpret the log difference across time as the growth rate. This is very useful. If a series grows at an approximate constant rate, then the time series plot will be linear in the log (but would be exponential in the level).

Putting things in logs also makes interpretation of moments and other things "scale-free." In particular, if you first take the log of a series and then calculate a standard deviation, the standard deviation has units which can be interpreted as a percentages. This is useful if we want to compare

the volatility of two series with very different means. For example, consumption is a much bigger fraction of total income than is investment. In percentage/log terms, investment is quite a bit more volatile than consumption. But in the levels, consumption looks pretty volatile relative to investment, because it's jumping around a much bigger mean.

6 Filtering

Many macroeconomic data have trends – consumption, output, etc. are rising across time. As you will see, in studying business cycles (the focus of this course) we often want to look at "second moments" – standard deviations, correlations, autocorrelations, etc. Second moments like this may be ill-defined if the variables under consideration are non-stationary (which means they don't have well-defined and/or time-invariant first or second moments).

I will not discuss it in depth here, but to "detrend" a series is to do a transformation that breaks a series into two components, a non-stationary trend component and a stationary cycle component. Detrending may result in the cycle component being stationary, but it may not depending on the series in question and the detrending method used. Suppose that one wants to isolate the stationary component of a time series as the cycle component. How one should detrend a series depends on the source of the non-stationarity. Loosely, there are two models of non-stationary: deterministic and stochastic trends. Suppose that a variable follows a process like this:

$$X_t = X_t^\tau + \widehat{X}_t$$

Here, X_t^τ is the deterministic trend component; by deterministic I mean that it grows deterministically with time; a simple example would be something like bt, where t is a time index and b is a coefficient. $\widehat{X}_t$ is a stochastic component. If the stochastic component is mean-reverting, we say that the series has a deterministic trend (essentially meaning that a series will tend to return to its trend line after a stochastic shock). If the stochastic component is not mean-reverting, we say that the series has a stochastic trend (meaning that the series meanders without necessarily returning to a trend line – if shocks are permanent, so the stochastic component is not mean-reverting, there will be no tendency for the series to revert to trend). To render a deterministic trending series stationary you estimate a deterministic trend (regress the variable on deterministic time indexes) and remove the trend (take the residual). To render a stochastic trend series stationary you typically first difference the series (i.e. if the underlying series is in logs, which it almost always should be, the first difference is the approximate growth rate, so put the series in growth rates).

A related concept to detrending is filtering. Filtering will not necessarily render a non-stationary series stationary, and filtering can always be applied to a series even if it isn't (or isn't suspected to be) non-stationary. The basic idea is to suppose a series has two components: a "smooth" component and a "cycle" component:

$$X_t = X_t^s + X_t^c$$

The basic idea is to use some criterion to come up with a "smooth" component, and then attribute the rest to the "cycle" component. One obvious candidate would be something like a moving average – at each point of observations, take a moving average of X_t in a rolling window (the rolling window could take many forms, but suppose it's two sided, so you're taking the average from X_{t-k} to X_{t+k} for some k). This will produce a "smoothed" version of X_t, and subtracting this off from actual X_t would yield the "cycle" component.

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

$$\min_{X_t^s} \quad \sum_{t=1}^{T} (X_t - X_t^s)^2 + \lambda \sum_{t=2}^{T-2} \left((X_{t+1}^s - X_t^s) - (X_t^s - X_{t-1}^s) \right)^2 \tag{1}$$

The basic idea of this is to minimize the squared deviations about the smoothed component, subject to a penalty for the smoothed component moving too much. The magnitude of the penalty is governed by the parameter λ. If $\lambda = 0$, then the solution is simple $X_t^s = X_t$. As λ gets bigger, you will not allow X_t^s to move as much as the actual series, and so will start picking up more interesting "cycle" dynamics. In the limiting case where $\lambda \to \infty$, you can show that the HP filters reduces to removing a linear time trend from a series – in this case, the smoothed series must be a line, so if the series is trending up the smoothed series will be a straight line. It is common in empirical work in macro to use a value of $\lambda = 1600$ for quarterly frequency data.

What we'll often be doing is comparing moments (standard deviations, correlations, etc) from the data to moments generated from a model. An important point is to always treat the actual data and data generated from a model the same way. Hence, if you HP filtered actual data and use those to compute moments, you should apply an HP filter to model generated data before calculating moments there.

7 Lucas Critique

The Lucas Critique (from Lucas, 1976) is an important philosophical point that forms the basis of much of modern macroeconomics. From Keynes until the mid-1970s, macroeconomics looked quite different than it does now. On the theoretical side, people used variants of a textbook IS-LM model. That model did not take agent optimization, dynamics, or expectations formation very seriously. On the empirical side, people used "large scale" macroeconometric models. These were essentially systems of simultaneous equations featuring aggregate variables – many of the larger models would feature hundreds of variables. The design of these macroeconometric models was based on fit and forecasting, with little attention paid to any underlying theory or actual economics.

The essential gist of Lucas' Critique is that it is fraught with hazard to try to predict the effects

of a policy change based on correlations (or regression coefficients) based on historical data. We say that a parameter is "structural" if it is invariant to the rest of the economic environment, and in particular the policy environment. A parameter is "reduced form" if it is not invariant to the environment, or more generally if that parameter cannot be mapped back into some economic primitive. I'll consider two examples to make this point.

7.1 Simple Consumption Saving Model

Consider a very simple two period consumption saving model with a fixed real interest rate and no uncertainty. The household takes income flows to be exogenous. It solves the following problem:

$$\max_{C_t, C_{t+1}} \quad \frac{C_t^{1-\sigma} - 1}{1-\sigma} + \beta \frac{C_{t+1}^{1-\sigma} - 1}{1-\sigma}$$

s.t.

$$C_t + \frac{C_{t+1}}{1+r} = Y_t + \frac{Y_{t+1}}{1+r}$$

The first order condition, or Euler equation, is:

$$C_t^{-\sigma} = \beta(1+r)C_{t+1}^{-\sigma}$$

There are two structural parameters here – β and σ, which govern how heavily you discount future utility flows and how much curvature there is in the utility function. Let's assume that $\sigma = 1$ (which means the utility function collapses to $\ln C_t$ via L'Hopital's rule). We can then derive a consumption function that looks like:

$$C_t = \frac{1}{1+\beta}\left(Y_t + \frac{Y_{t+1}}{1+r}\right)$$

Here the "marginal propensity to consume" (or MPC) is the partial derivative of C_t with respect to Y_t, which is $\frac{1}{1+\beta}$. This is just a transformation of a structural parameter, and so we could consider the MPC itself to actually be structural.

Now, suppose an econometrician estimates a regression of consumption on income:

$$C_t = \alpha + \gamma Y_t + u_t$$

This regression is misspecified in the sense that it omits Y_{t+1} – this is in the error term. If current income is uncorrelated with future income, Y_t would be uncorrelated with the error term, and we would get $\gamma = \frac{1}{1+\beta}$ (at least in a large enough sample). But what if current income is correlated with future income (i.e. income is persistent)? Then there is an omitted variable; Y_t will be positively correlated with the error term, which will mean that you will get an upward-biased estimate of γ.

Suppose that in the past changes in income have been very persistent – meaning that when Y_t changes, Y_{t+1} changes by almost the same amount. The consumption function derived from

the theory would suggest that consumption would then react roughly one-for-one with changes in income. Suppose that an econometrician goes and estimates this equation and comes back with a very large estimate of γ (let's say near 1). He then goes to a policy advisor and says "The MPC is near 1. If we give people more income (say, through a tax cut), then they will spend virtually all of it, and there will be large stimulative effects on overall economic activity!" So the policymaker says "Okay, let's cut people's taxes for one period by one dollar." The theory tells us that raising people's income for one period (a tax cut would effectively do that) will cause them to increase their consumption by only $\frac{1}{1+\beta}$: if β is near 1, then loosely people would increase their consumption by 1/2 of the tax cut. This is smaller than the results estimated from the regression, which suggest that the MPC is much higher and close to 1. In this example, using the correlation between income and consumption estimated in past data (when income changes were very persistent) is not informative about what will happen if you consider a temporary change in income.

7.2 Phillips Curve Model

Consider another example, which was really the thing that Lucas was criticizing. As we will see later in the course, it is possible to derive a "Phillips curve" which shows some relationship between economic activity, inflation, and expected inflation:

$$\pi_t = \theta(u_t - u^N) + \beta E_t \pi_{t+1}$$

Above, θ is a coefficient, β is a discount factor (same as in the previous example), π_t is inflation, $E_t\pi_{t+1}$ is expected inflation, u_t is the unemployment rate, and u^N is the "natural rate" of unemployment (which I here assume to be time-invariant). θ and β are structural parameters.

Particularly before rational expectations (also attributed to Lucas), people didn't know how to treat expectations seriously; and indeed, many models were static and so had no role for expectations of what was going to happen in the future. Suppose an econometrician estimated the following regression:

$$\pi_t = \xi(u_t - u^N) + \epsilon_t$$

As in the above example, this regression is misspecified relative to the theory – the error term includes expected future inflation. But suppose that in historical data expected inflation was pretty stable. This would mean there wouldn't be much bias in the coefficient estimate, and we would expect that an estimate ξ would be close to the true θ. Suppose that the true $\theta < 0$: there is a negative relationship between inflation and unemployment. One would be tempted to conclude that raising inflation would lead to a reduction in unemployment. So the econometrician goes to the policymaker and says "Let's raise inflation and this will result in lower unemployment!" But will it?

It will, but only to the extent to which higher inflation doesn't get incorporated into higher inflation expectations. If people are paying attention, they will expected more inflation – $E_t\pi_{t+1}$

will rise, which means u_t won't fall by as much as the simple regression would have predicted. Again, using past correlations to predict the effects of a policy change may very well be misleading.

7.3 So should we do econometrics?

The conclusion of the Lucas critique is that we need to take economic theory seriously – correlations (or regression coefficients) estimated in the data may not be policy-invariant, and therefore may not be useful in thinking about "counterfactuals" where we think of what would happen under alternative policy regimes.

Some people (incorrectly) interpret the Lucas Critique as saying we shouldn't do econometrics at all in macro. This is too strong. The Lucas Critique tells us that we need to take theory seriously when doing econometrics; and when we do econometrics without theory (e.g. reduced form econometrics), be honest and open about the potential misgivings. In both of the examples I gave you above, we actually have regression specifications implied by the theory – it's just that in the regressions I considered running, there was an omitted variable. "Theory" doesn't tell us values of structural parameters like β or θ – that's what econometrics is for. But theory might tell us what kind of econometric models to run, what kind of restrictions we can impose, etc. Then once we have good estimates of the structural parameters, we can use the model to to consider the effects of different policies.

It is actually here where the implications of rational expectations can be useful. Consider the two period consumption model (this time, make it stochastic so that the point is clearer). The theory tells us to run a regression like:

$$C_t = \alpha_1 Y_t + \alpha_2 E_t Y_{t+1} + \epsilon_t$$

The problem here is that we don't necessarily observe $E_t Y_{t+1}$. Rational expectations tells us how to get around this, however. In particular, rational expectations tells us that $E_t Y_{t+1} = Y_{t+1} + u_{t+1}$, where u_{t+1} is (i) mean zero and (ii) uncorrelated with anything known at date t or earlier. So rational expectations tells us that we can run the following regression:

$$C_t = \alpha_1 Y_t + \alpha_2 Y_{t+1} + v_t$$

Now v_t is a composite error term, equal to $\epsilon_t + \alpha_2 u_{t+1}$. Y_{t+1} is correlated with u_{t+1}, so OLS won't work here. But rational expectations tells us that we can instrument for Y_{t+1} with anything known at date t or earlier – rational expectations tells us that the forecast error, u_{t+1}, is uncorrelated with anything dated t or earlier, making anything dated t or earlier valid instruments. We could do a similar exercise for the Phillips Curve equation, including realized future inflation on the right hand side and instrumenting for it with something known at time t or earlier. In other words, taking rational expectations seriously often gives us a "theory of the error term" in regression models and therefore guides us on how to deal with that error term.